Codon and Amino Acid Usage Are Shaped by Selection Across Divergent Model Organisms of the Pancrustacea

Carrie A. Whittle* and Cassandra G. Extavour*,†
*Department of Organismic and Evolutionary Biology and †Department of Molecular and Cellular Biology, Harvard University, Cambridge, Massachusetts 02138
ORCID ID: 0000-0003-2922-5855 (C.G.E.)

ABSTRACT
In protein-coding genes, synonymous codon usage and amino acid composition correlate to expression in some eukaryotes, and may result from translational selection. Here, we studied large-scale RNA-seq data from three divergent arthropod models, including cricket (Gryllus bimaculatus), milkweed bug (Oncopeltus fasciatus), and the amphipod crustacean Parhyale hawaiensis, and tested for optimization of codon and amino acid usage relative to expression level. We report strong signals of AT3 optimal codons (those favored in highly expressed genes) in G. bimaculatus and O. fasciatus, whereas weaker signs of GC3 optimal codons were found in P. hawaiensis, suggesting selection on codon usage in all three organisms. Further, in G. bimaculatus and O. fasciatus, high expression was associated with lowered frequency of amino acids with large size/complexity (S/C) scores in favor of those with intermediate S/C values; thus, selection may favor smaller amino acids while retaining those of moderate size for protein stability or conformation. In P. hawaiensis, highly transcribed genes had elevated frequency of amino acids with large and small S/C scores, suggesting a complex dynamic in this crustacean. In all species, the highly transcribed genes appeared to favor short proteins, high optimal codon usage, specific amino acids, and were preferentially involved in cell-cycling and protein synthesis. Together, based on examination of 1,680,067, 1,667,783, and 1,326,896 codon sites in G. bimaculatus, O. fasciatus, and P. hawaiensis, respectively, we conclude that translational selection shapes codon and amino acid usage in these three Pancrustacean arthropods.

KEYWORDS
Gryllus bimaculatus
Oncopeltus fasciatus
Parhyale hawaiensis
optimal codons
translational selection
Crickets have historically been key systems for research in functional
genetics, neurobiology, and developmental biology but have more re-
cently emerged as an important model for evolutionary biology, in-
cluding topics in sexual selection and speciation (Garcia-Gonzalez and
Simmons 2010; Huber et al. 1989; Zak et al. 2014). The two-spotted
cricket *G. bimaculatus*, like other crickets, branches basally to the Dip-
tera, Hymenoptera, and Coleoptera and has a large genome (~1.7 Gb)
predicted to contain more than 19,000 unique gene sequences (Zeng
et al. 2013). An additional basally branching hemimetabolous insect,
the milkweed bug *O. fasciatus*, provides further opportunities for mo-
lecular evolutionary research (Butt 1949; Liu and Kaufman 2009a,b).
Specifically, more than 10,000 unique genes, including those involved
in an array of developmental, signaling, and housekeeping functions,
have been identified, allowing comparative genomics in this taxon
(Ewen-Campen et al. 2011). The amphipod crustacean *P. hawaiiensis*
belongs to the crustacean class Malacostraca, which is one of the crus-
tacean groups most closely related to the Hexapods (Insecta, Collem-
bola, Proptera, and Diplura) (Regier et al. 2010) and is common in
intertidal habitats worldwide. It has extensive genetic and developmen-
tal tools available (Kontarakis et al. 2011; Pavlopoulos and Averof 2005;
Rehm et al. 2008), yet remarkably is one of the only crustaceans besides
*Daphnia pulex*, with whole genome or transcriptome datasets available
(Colbourne et al. 2011; Zeng and Extavour 2012; Zeng et al. 2011).
Currently, few studies are available regarding the genomic traits and
evolution in nontraditional arthropod models, including the evolution
of synonymous codon usage and amino acid composition in protein-
coding DNA.

Synonymous codons are not used randomly in the genome. Biases in
synonymous codon usage may arise from mutational pressure (Osawa
et al. 1988; Sharp et al. 1995; Sueoka 1988) or from selective forces
favoring efficient and accurate translation (Duret 2000; Duret and
Mouchiroud 1999; Stootski and Eyre-Walker 2007). Findings that
codon usage biases are correlated with transfer RNA (tRNA) abundance
and/or gene copy number (Duret 2000; Ikemura 1981; Ikemura 1985)
and with gene expression levels in numerous taxonomic groups (e.g.,
Escherichia coli, Saccharomyces, Caenorhabditis, Drosophila, Arabidopsis,
Silene, Populus) support the theory that selection favors specific cod-
ons that promote efficient and accurate translation of genes that are
expressed at high levels (Cutter et al. 2006; Duret and Mouchiroud
1999; Ingvarsson 2008; Qiu et al. 2011; Sharp and Li 1986; Sharp et al.
1986). Thus, an effective method to identify codons favored by selec-
tion (optimal codons) is to compare codon usage per amino acid in
highly expressed genes and lowly expressed genes (Cutter et al. 2006;
Duret and Mouchiroud 1999; Ingvarsson 2008; Wang et al. 2011;
Whittle et al. 2007). In this regard, the presence/absence of optimal
codons and factors linked to their usage may reveal whether selection
for translational efficiency and/or accuracy plays a major role in an
organism’s genome evolution.

In arthropods, nearly all information to date about optimal codons is
from higher insects: optimal codons have been identified in species of
*Drosophila* and *Anopheles*, and other taxa such as *Tribolium castaneum*,
*Apis mellifera*, and *Nasonia vitripennis* (Behura and Severson 2012,
2013; Duret and Mouchiroud 1999; Williford and Demuth 2012;
C. A. Whittle et al., unpublished data), although no or weak genome-
wide trends were evident in *Bombyx mori* (Jia et al. 2015; C. A. Whittle
et al., unpublished data). The bias favors GC3 codons in some organ-
isms, including *D. melanogaster* (Duret and Mouchiroud 1999), but
AT3 is favored in others (Behura and Severson 2012; C. A. Whittle
et al., unpublished data), indicating marked divergence in the pres-
ence/absence of optimal codons and types of optimal codons among
these insects. Nevertheless, further data are needed about the dynamics
of codon usage in arthropods outside of the higher insects, such as
hemimetabolous insects and crustaceans, to answer a number of out-
standing questions, including, for example whether optimal codons exist
in these organisms, and if so, how they are related to traits such as amino
acid composition (Akashi 2003; Cutter et al. 2006; Williford and
Demuth 2012).

The amino acid composition of proteins also appears to change with
increasing expression levels (Akashi 2003; Cutter et al. 2006; Duret
2000; Williford and Demuth 2012). Several reports have indicated that
amino acid frequency in proteins encoded by highly expressed genes
correlates to the most abundant tRNA in an organism, consistent with
selection for speed and accuracy of translation (Akashi 2003; Duret
2000), Further, studies in some eukaryotes, including yeast, *T. castaneum*,
and *Caenorhabditis elegans*, have indicated that amino acids in highly
expressed genes tend to be less metabolically costly (Cutter et al. 2006;
Raiford et al. 2008; Williford and Demuth 2012). For instance, highly
transcribed genes contain fewer amino acids of large size and complexity
(S/C scores; Dufon 1997), and favor low cost amino acids (Akashi 2003;
Cutter et al. 2006; Williford and Demuth 2012). Thus, in those studies,
abundant proteins, presumably encoded by highly transcribed genes,
tended to be comprised of smaller less complex amino acids, which is
not only predicted to minimize the biochemical energy costs of synthesis,
but might also contribute to the stability of the protein structure and its
conformation (Cutter et al. 2006; Dufon 1997; Williford and Demuth
2012). Expanding this molecular evolutionary research to non-traditional
model organisms will help elucidate the breadth of this phenomenon
and the role of evolution of amino acid preferences in promoting translational
efficiency.

In the present investigation, we study optimal codon and amino acid
usage based on recently available large-scale RNA-seq data from three
emerging models of arthropods, *G. bimaculatus*, *O. fasciatus*, and
*P. hawaiiensis*. The RNA-seq data are available at the Assembled Search-
able Giant Arthropod Read Database (i.e., ASGARD) (Ewen-Campen
Using these data, we provide strong evidence that codon usage and
amino acid frequency has been optimized in highly expressed genes
each of these organisms, with the strongest signals observed in
*G. bimaculatus* and *O. fasciatus*, and weaker, yet significant effects
detected in *P. hawaiiensis*. Further, using *D. melanogaster* as a reference, we
show that highly expressed coding sequence (CDS) are shorter, consist-
tent with translational selection, and are enriched for genes involved in
cell-cycle processes and protein synthesis in all three organisms. To-
gether, the results are consistent with a history of selection on synon-
ymous codon usage and on amino acid frequency in highly transcribed
genes to promote translational efficiency and accuracy across Pancrus-
taceae genome evolution.

MATERIALS AND METHODS

Transcriptomes Under Study

Here, we studied RNA-seq data from ASGARD derived from ovaries,
multi-stage embryos and/or postembryonic tissues for *G. bimaculatus*,
*O. fasciatus*, and *P. hawaiiensis* (Supporting Information, Table S1)
(Ewen-Campen et al. 2011; Zeng et al. 2011; Zeng and Extavour
2012; Zeng et al. 2013). Multistage embryos and ovaries comprise
highly complex tissues, expressing a large component of genes in the
genome (Li et al. 2014); thus, genes highly transcribed in these tissues
are apt to be an effective tool to reveal genome-wide optimization of
codon usage and amino acids in an organism (c.f. Subramanian and
Kumar 2004). The next-generation sequencing data are described in
Table S1. We identified all assembled transcripts (Zeng and Extavour
without isoforms, to allow accurate mapping of reads to a specific single CDS to quantify expression. For this gene set, we extracted all coding regions with a start codon, and lacking any unknown sites or internal stop codons. To quantify expression levels, we mapped reads to the CDS and calculated reads per million (RPM) using the non-normalized libraries that comprise the largest (or only) dataset per species (Table S1; where read number directly reflects abundance; average read length was 349, 297, and 400 bp per species, respectively) (Zeng and Extavour 2012). RPM was defined as the number of read matches to a CDS/Total number of reads matching all CDS \( \times 1,000,000 \); see Identification of optimal codons for reads per kilobase million, RPM). Tools and software for measuring expression and codon and amino acid usage in the various datasets are described in File S3.

**Data availability**

Data are available at ASGARD (Ewen-Campen et al. 2011; Zeng et al. 2011; Zeng and Extavour 2012; Zeng et al. 2013). Details are provided in Table S1.

**RESULTS**

For our analyses, the number of CDS examined after excluding genes with isoforms, unknown sites, or internal stop codons for *G. bimaculatus*, *O. fasciatus*, and *P. hawaiensis* was 5284, 6161, and 6731, which spanned 1680,067, 1667,783, and 1,326,896 codons, respectively. The nucleotide composition varied across all CDS among organisms: it was AT-rich for *G. bimaculatus* (AT \( = 0.6177 \pm 0.006 \); with similar levels of A and T) and *O. fasciatus* (0.6520 \( \pm 0.006 \); similar levels of A and T) and AT was mildly higher than GC content in *P. hawaiensis* (0.5529 \( \pm 0.0417 \)).

We first asked whether we could detect differences in codon usage relative to expression levels in all three study species. To study synonymous codon usage, we first determined the GC3 content among the 5% highest and lowest expressed CDS and detected striking differences consistent with selection on codon usage among these two categories. A highly effective method to identify codon preferences is to compare these preferences between genes expressed at extremely high or extremely low levels (Cutter et al. 2006; Duret and Mouchiroud 1999; Ingvarsson 2008; Wang et al. 2011; Whittle et al. 2007, 2011a). As shown in Figure 1, we found that GC3 was statistically significantly lower in CDS with the 5% highest RPM (mean \( = 0.307 \pm 0.003 \)) than in the 5% of CDS with the lowest RPM (mean \( = 0.334 \pm 0.006 \)) in *G. bimaculatus* (t-test \( P = 2.0 \times 10^{-7} \)) and *O. fasciatus* (mean 5% highest = 0.306 \( \pm 0.004 \) and mean 5% lowest = 0.349 \( \pm 0.007 \), respectively, \( P = 1.5 \times 10^{-7} \)). These trends indicate that AT3 codons are more common under high transcription in these insects. For *P. hawaiensis*, GC3 was significantly elevated in highly expressed genes compared with lowly expressed genes (mean 5% highest = 0.507 \( \pm 0.005 \) and mean 5% lowest = 0.473 \( \pm 0.005 \)) suggesting selection for GC3 codons in highly expressed genes in this crustacean. For all genes per species, we then assessed the effective number of codons (ENC), wherein values range from 20 (when one codon is exclusively used to code for a given amino acid) to 61 (all codons used equally), and lower values denote greater biases in codon usage in a gene (Wright 1990).

We found that AT3 was strongly negatively correlated with ENC in *G. bimaculatus* (Spearman’s Rank \( r = -0.65, P = 2.0 \times 10^{-6} \)) and *O. fasciatus* (\( r = -0.68, P = 2.0 \times 10^{-7} \)), thus concurring with AT3 codon preferences in those species. In contrast, no correlation was evident for GC3 and ENC across all genes from *P. hawaiensis* (\( P > 0.05 \)), but a negative correlation evident in the uppermost expressed gene set (\( r = -0.20, P < 2.0 \times 10^{-4} \) for the uppermost 5%), implying marked GC3 favoritism within this dataset. Given these findings, we next assessed synonymous codon usage in every amino acid of the analyzed CDS, to test the hypothesis that AT3 in the insects, and GC3 in the amphipod, were optimal codons in these species.

**Identification of optimal codons**

Although there is some variation in approaches to identify optimal codons (Hershberg and Petrov 2009, 2012; Wang et al. 2011), here we identified the optimal codon list in each amino acid in *G. bimaculatus*, *O. fasciatus*, and *P. hawaiensis* using relative synonymous codon usage (RSCU) relative to gene expression (Cutter et al. 2006; Duret and Mouchiroud 1999; Ingvarsson 2008; Qiu et al. 2011; Wang et al. 2011; Whittle et al. 2007; Whittle et al. 2011a). RSCU measures the observed frequency of a particular codon relative to the expected frequency if all synonymous codons were used equally. RSCU values greater than 1 indicate preferential usage, and higher values among codons within a synonymous codon family denote increased usage (Sharp and Li 1986). Optimal codons were defined as those having a statistically significant and positive ARSCU = RSCU_Mean_Highly_Expressed_CDS - RSCU_Mean_Low_Expressed_CDS when more than one codon matched this criterion per amino acid, the codon with the largest ARSCU amino acid was defined as the primary optimal codon (Cutter et al. 2006; Duret and Mouchiroud 1999; Ingvarsson 2008; Whittle et al. 2011a). The ARSCU for all amino acids is shown in Table 1, and the mean RSCU and standard errors per expression class are shown in Table S2. We report that 17 of the 18 amino acids with synonymous codons had an optimal codon with a statistically significant and positive ARSCU in *G. bimaculatus*. For *O. fasciatus*, we identified a total of 16 optimal codons. Both species favored AT3 codons in highly expressed genes. For instance, in *G. bimaculatus*, 15 of 17 optimal codons ended in AT, whereas 16 of 16 optimal codons ended in AT for *O. fasciatus*. In fact, for each of the eight amino acids encoded by four or more codons (e.g., Ala, Arg, Gly, Leu, Pro, Ser, Thr, and Val), the optimal codon always ended in T in *O. fasciatus*, whereas six of eight (the two exceptions being Leu and Pro) ended in T for *G. bimaculatus*. Thus, there are marked preferences toward specific types of codons (T3) even for amino acids wherein an assortment of synonymous codons exist. The bias toward optimal codons ending in AT was also observed for amino acids with two or three synonymous codons (Table 1). As *G. bimaculatus* and *O. fasciatus* each have inherently AT rich transcriptomes (see first paragraph of the Results section), codon}
usage is likely partly influenced by the inherent genome composition, which might be expected to contain AT-rich codons. Despite this, our data show that the usage of AT3 codons is enhanced markedly in highly expressed genes (not expected under mutational pressure; see also Optimal codon usage is shaped by selection), and thus is consistent with selection favoring these optimal codons in genes that experience high rates of transcription.

For *P. hawaiiensis*, 13 amino acids also were found to have an optimal codon that was favored in highly expressed genes (Table 1). However, in contrast to *G. bimaculatus* and *O. fasciatus*, the majority (12 of 13) of optimal codons were GC3 codons (Table 1). Notably, the absolute value of ∆RSCU was of markedly lower magnitude for *P. hawaiiensis* (mean and standard error, 0.0729 ± 0.0062) than for *G. bimaculatus* (0.1060 ± 0.0110) and *O. fasciatus* (0.1158 ± 0.0111) (*P* < 0.016 for *t*-tests). This finding may indicate that selective forces acting on optimal codons are weaker in the amphibian than in the two insects. Nevertheless, collectively the transcript data for all three invertebrate species show evidence of favoritism toward specific synonymous codons in highly expressed genes.

It is worth noting that genes with longer assembled CDS in the aforementioned analyses could have a greater RPM using next-generation sequencing data due to their greater size, and thus not solely result from high expression. Accordingly, as described in File S1, we repeated our analyses of RSCU, which includes length when measuring expression level, and found the same optimal codon lists to that obtained from RPM, with the exception that P-values were weaker. In sum, we show that RPM provides the most rigorous method to identify optimal codons (File S1), and thus use these codon lists for all analyses.

Another complementary method used to further confirm optimal codon lists is to compare codon usage between ribosomal protein genes (RPGs), which typically are highly expressed, and poorly expressed genes (Heger and Ponting 2007; Wang et al. 2011). Thus, we repeated our analyses of RSCU using RSCU_RPGs = RSCU_RPGs − RSCU_CDS with Lowest 5% Expression (File S2) and found that the results of these analyses also support the existence of AT3 optimal codons in *G. bimaculatus* and *O. fasciatus*, and GC3 codons for *P. hawaiiensis*.

**Genome-wide optimal codon usage**

Using the optimal codon lists in Table 1, we calculated the frequency of optimal codons (Fop) (Ikekura 1981) for each CDS under study and found that this parameter was highly statistically significantly and positively correlated to RPM across all CDS for *G. bimaculatus* (Spearman

### Table 1, continued

<table>
<thead>
<tr>
<th>Glyrurus</th>
<th>Oncopeltus</th>
<th>Parhyale</th>
</tr>
</thead>
<tbody>
<tr>
<td>Val</td>
<td>GTC</td>
<td>−0.141</td>
</tr>
<tr>
<td>Val</td>
<td>GAT</td>
<td>−0.126</td>
</tr>
<tr>
<td>Val</td>
<td>GTG</td>
<td>+0.037</td>
</tr>
</tbody>
</table>

The codon identified as the primary optimal codon for each amino acid is in boldface. RSCU, relative synonymous codon usage.

a Asterisks indicate P value using t-tests where "*P* < 0.05, **P* < 0.001. Codons with '0.05 > *P* > 0.1 are also indicated and considered putative optimal codons. The means and standard errors for highly and for lowly expressed CDS are provided in Table S2. Species are abbreviated using their genus name.

b For the amino acid Pro in *G. bimaculatus*, CCA was selected as the optimal codon due to the fact that it had a lower *P* value than CCT, although both exhibit signals of being optimal codons.
Rank R = 0.23, P < 10^{-15}), O. fasciatus (R = 0.17, P < 10^{-15}), and P. hawaiensis (R = 0.09, P < 10^{-15}). The R values were each < 0.3, suggesting a moderate, yet highly significant, association between Fop and RPM. We subsequently binned each CDS into one of three distinct RPM categories, namely low (below the 5th percentile), moderate (between the 5th and the 95th percentile), and high (above the 95th percentile). This approach revealed an unambiguous shift in Fop, increasing progressively from the low, moderate to high RPM classes for G. bimaculatus, O. fasciatus, and P. hawaiensis (Figure 2), a trend consistent with a strong connection between transcription rates and Fop. Thus, although selection on codon usage has greatest effects in the uppermost expression levels, it also shapes codon usage (albeit to a lesser extent) in moderately expressed genes in those taxa. The weakest effect appears to be for moderately expressed genes in P. hawaiensis, which was only modestly higher than the low expression class. Thus, optimization of codon usage for the moderate expression level classes may be mild in this taxon given the relatively low absolute value of the genome-wide R value above, the similarity of Fop to the lowest class (Figure 2), and the fact that ENC showed an effect only in the highest expression category (see Identification of optimal codons). Given these trends, it is evident that the method of contrasting of expression among the CDS with the 5% highest and lowest RPM used herein, was advantageous in revealing the list of optimal codons. This agrees with prior research indicating that comparison of codon usage among the highest and lowest expressed CDS (known as the comparison method) provides a more effective tool (not weakened by mild correlations in the mid-ranges of expression) for revealing optimal codons than broad correlations between codon usage and expression across all CDS (known as the correlation method), at least for these particular organisms (Hershberg and Petrov 2009; Wang et al. 2011).

We also report that strong positive correlations were found between Fop and AT3 for G. bimaculatus (Spearman R = 0.62, P < 10^{-15}) and O. fasciatus (R = 0.82, P < 10^{-15}) and between Fop and GC3 (R = 0.64, P < 10^{-15}) for P. hawaiensis (Figure S1). This affirms the strong link between these traits, even in the lattermost taxon where only 12 GC3 optimal codons were identified.

**Optimal codon usage varies among amino acids**

The strength of bias in codon usage varied among amino acids. Specifically, the degree of biased codon usage in the 5% most highly and least highly expressed genes was strongest for the highly degenerate amino acids. For G. bimaculatus, the three amino acids with six synonymous codons tended to have the largest ΔRSCU among highly and lowly expressed genes: Arg (+0.277), Leu (+0.300), and Ser (+0.136), but the remaining amino acids showed no trends with respect to degeneracy. Similarly, for the taxon G. bimaculatus, the highest ΔRSCU occurred for Leu (+0.146) and Ile (+0.136), but the remaining amino acids showed no trends with respect to degeneracy. This may reflect the fact that optimal codon usage and ΔRSCU were weakest in this taxon, making differences among amino acids accordingly less marked.

**Optimal codon usage is shaped by selection**

It has been observed that although codons with elevated frequency in highly expressed genes typically result from selection, they also could result from mutational biases (Comeron 2004). For instance, this might occur if high transcription rates lead to a bias toward specific mutations (e.g., C to T) in the DNA strand, which has been observed in *Escherichia coli* (Beletskii and Bhagwat 1996, 1998) and/or from biases arising during transcriptional-coupled repair such as those reported in mammals (Green et al. 2003). Traditionally, one method used to exclude mutational bias is to compare GC content of introns vs. GC3 frequency in CDS, where the absence of a relationship excludes mutational bias (Bachtrog 2007; Comeron 2004; Qiu et al. 2011; Willford and Demuth 2012). As the RNA-seq data we have studied here does not contain introns, we cannot test for putative mutation bias using this method. Nonetheless, several features of our data point toward selection as a significant factor in the evolution of optimal codons in these taxa, rather than an expression-induced mutational bias. First, for G. bimaculatus and O. fasciatus the vast majority of codons defined as optimal in Table 1 end in T. However, examining amino acids with fourfold degenerate sites (Ala, Gly, Pro, Thr, Val) which can end in A, G, C, or T, it is evident that there is substantial variation in the ΔRSCU of the optimal codon across amino acids. This finding is not consistent with mutational bias, as all possible types of mutations to T (G to T, C to T, A to T) can occur at the degenerate site. If any neutral mutational bias to T3 were responsible for the evolution of optimal codons for these amino acids, then such bias should affect all five of these amino acids similarly and thus should lead to similar ΔRSCU values. However, these values in
fact ranged from $+0.139$ to $+0.279$ for *G. bimaculatus* and from $+0.194$ to $+0.370$ for *O. fasciatus* (Table 1). Second, the very large positive ΔRSCU values for four- and six-fold degenerate amino acids (as compared to two-fold) are consistent with selection (McVean and Vieira 2001; Cutter et al. 2006) rather than mutational bias. Third, for the six-fold degenerate amino acid Ser, which has two codons ending in T (TCT, AGT), we only observed a strong signal of optimal codons for TCT and not AGT in *G. bimaculatus*; this was also the case for *O. fasciatus*. The presence of a C to T mutational bias would be expected to result in both TCT (from TCC to TCT mutations) and AGT (from AGC to AGT mutations) having a statistically significant and positive ΔRSCU. However, we observed no such effect at the AGT codon (+0.024 for *G. bimaculatus* and $-0.081$ for *O. fasciatus*, neither of which were significant; Table 1). We note that the caveat for this particular finding for Ser is that it excludes solely a C to T mutational bias in these species. Collectively the aforementioned results are suggestive that factors other than mutational bias, namely selective pressures, contribute toward shaping the optimal codons for *G. bimaculatus* and *O. fasciatus*. Nonetheless, we do not exclude that some mutational pressures, such as neighboring-nucleotide effects, could contribute to nucleotide composition (Hodgkinson and Eyre-Walker 2011). Finally, for *P. hawaiensis*, where the majority of optimal codons ended in GC, the ΔRSCU for the optimal codon identified for amino acids with four synonymous codons varied extensively ($+0.028$ to $+0.114$), whereas one ended in A. Both of these observations are inconsistent with mutational bias as an underlying cause of optimal codons. In sum, our data point toward selection as a factor shaping codon usage in all three arthropods.

**Optimal codon usage correlates to amino acid size and complexity**

We assessed the cost of protein synthesis using the Dufton (1997) methodology, wherein each amino acid is assigned a size complexity score (S/C) based on its molecular weight and complexity. The S/C score reflects the chemical energy investment, as well as the costs of stability in the protein’s final conformation (Dufton 1997; Williford and Demuth 2012). The S/C scores are listed in Table S3. For each species, we calculated the proportion of amino acids per protein with high S/C ($>40$; PrHigh SC), which include Cys, Phe, His, Met, Arg, Trp, and Tyr. We found that PrHigh SC was inversely correlated with Fop for *G. bimaculatus* (Spearman $R = -0.09$, $P < 10^{-13}$) and *O. fasciatus* (Spearman $R = -0.12$, $P < 10^{-13}$) (see below in this section for *P. hawaiensis*). Given that codon adaptation indices, such as Fop, correlate with expression and can be used as a proxy to measure expression levels of genes across the genome (Coghlan and Wolfe 2000; Drummond et al. 2005; Popescu et al. 2006; Wall et al. 2005; Williford and Demuth 2012), we can infer that higher gene expression is linked to reduced S/C score in these insects. The R values are similar to those reported for expression and mean S/C in the flour beetle *T. castaneum* (Williford and Demuth 2012) and suggest that selection for reduced S/C is not exclusive to highly expressed genes but rather exhibits a gradual decline with reduced transcription rates. Nonetheless, our data show that the relationship is not particularly strong as revealed by the absolute value of R ($=0.12$), possibly indicating that the selection pressure on optimal codon usage relative to S/C varies among genes. In other words, S/C may be under relatively greater selective pressure in some genes than Fop and vice-versa.

We further studied the relationship between expression and the frequency of each of the individual 20 amino acids. For this, we used Fop as a proxy for the relative expression level at the genome-wide level (Coghlan and Wolfe 2000; Drummond et al. 2005; Popescu et al. 2006; Wall et al. 2005), which is apt to be less noisy than RPM or RPKM at intermediate expression levels (outside the upper and lower 5%; see Identification of optimal codons). In addition, we wished to assess the relationship between Fop and amino acid in and of itself. The Spearman rank correlations between Fop and amino acid frequency per CDS at the genome-wide level are shown in Table 2. The results showed that for *G. bimaculatus*, 15 amino acids exhibited a statistically significant correlation. The negative R values, which indicate amino acids used rarely in genes with high Fop/expression level, include four of the six amino acids with high S/C scores (>40, Table S3) namely Arg, Met, His, and Trp (the latter two are nonsignificant), consistent both with the inverse relationship described above between PrHigh SC and Fop (see Optimal codon usage correlates to amino acid size and complexity), and suggest selection against these amino acids in highly expressed genes. The positive correlations indicate that the more frequently an amino acid appeared in a CDS sequence, the more likely it was to display elevated optimal codon usage. Accordingly, the amino acids most favored under high expression (with R > 0.269, P < 10^{-15}) included Glu and Asp, which have moderate S/C scores (between 32.7 and 36.5, Table S3). The amino acids Asn, Lys, and Ile also exhibited substantial positive correlations with frequency in CDS and expression levels (R = 0.169, 0.151, and 0.115, respectively, P < 10^{-15}), and have moderate or low S/C scores (33.7, 30.1, and 16.0, respectively, Table S3). The two amino acids with the largest positive R values with respect to Fop, namely Glu and Asp, and those with the most negative values, Arg and Thr, are illustrated in Figure 3A (Fop values are binned into four distinct categories: Fop < 0.3, >3 Fop < 0.4, >0.4 Fop < 0.5, and Fop >0.5). These results further confirm the striking shifts in amino acid frequency under high Fop/expression. In summary, it is evident that specific amino acids are preferred under high expression in *G. bimaculatus*, and these tend to be of moderate or low size and complexity.

In *O. fasciatus*, 16 amino acids exhibited a statistically significant change in frequency with Fop/expression level (Table 2). The two amino acids with the strongest negative (Arg, Gly) and strongest positive (Asn, Lys) R values are illustrated in Figure 3B. The strongest negative correlation was for Arg (R = $-0.252$, P < 10^{-15}), which also has among the largest S/C scores (56.3), implying losses of this metabolically expensive amino acid in CDS expressed at high levels (Table 2, Figure 3B). However, Gly and Ala also exhibited substantial inverse correlations (between $-0.188$ and $-0.183$, P < 10^{-15}) and have low S/C scores (between 1 and 4.8), implying reduced frequency of these amino acids despite their very low metabolic costs. The amino acids exhibiting the strongest positive correlations with Fop/expression levels were Asn, Lys, Asp, and Glu (R between 0.211 and 0.287, P < 10^{-15}, Table 2, Figure 3B). Remarkably, each of these amino acids has S/C scores within a narrow range from 30.1 to 36.5. Thus, similar to *G. bimaculatus*, genes expressed at high levels appear to favor not amino acids with the lowest S/C scores, but rather a narrow range of moderately sized amino acids, potentially indicating an advantage to maintaining an intermediate protein size under high expression, rather than the smallest possible size. Thus, this explains the negative association between PrHigh SC and Fop reported previously (see preceding paragraph in this section), i.e., an average decline in use of amino acids with extreme S/C in highly expressed genes, while indicating this decline does not necessarily favor the lowest S/C amino acids in these invertebrates. Notably, eight of the nine amino acids with positive R in *G. bimaculatus* also had positive R in *O. fasciatus*, implying similar preferences for specific amino acids among these divergent insects from different orders (Orthoptera and Hemiptera, respectively). Furthermore, the four amino acids with strongest positive R values (Lys, Asn, Asp and Glu) were identical among these two arthropods.
(Table 2), implying a shared preference for intermediate sized amino acids under high expression. In the beetle *T. castaneum* (Coleoptera), it has been reported that amino acids with moderate and with low scores were prefered under high expression (e.g., Glu, Asp, Lys, Val, Ala, and Gly) (Williford and Demuth 2012), whereas our data for *G. bimaculatus* and *O. fasciatus* reveal a strong preference for intermediate sized amino acids (excluding the lowest classes), suggesting species-specific effects on amino acid preferences.

Together, these results from individual amino acids in *G. bimaculatus* and *O. fasciatus* indicate that the moderate negative correlation between $P_{S/C}$ and Fop is due to a shift toward increased usage of intermediate cost amino acids, coupled with a decline in both very high and low cost amino acids under high expression. It is worth noting that degeneracy was sometimes lower for higher cost amino acids (e.g., Phe vs. Ser, Table S3), which could conceivably contribute to an inverse correlation between $P_{S/C}$ and Fop (since codon bias was reduced for low degeneracy amino acids, Table 1). However, our results in Table 2 show that degeneracy effects on Fop did not underlie the $P_{S/C}$ and Fop negative correlation. For example, for *G. bimaculatus*, the two-fold degenerate amino acids Phe (S/C score 44.0) and Cys (57.2) exhibited positive (rather than negative) correlations with expression while highly degenerate amino acids such as Ser (S/C score 17.8) exhibited a large negative correlation (rather than positive) with transcription/Fop level (Table 2). We therefore propose that our findings with respect to $P_{S/C}$ cannot result from an artificial relationship between Fop and amino acid usage. Rather, our data suggest that the reduced usage of expensive amino acids is primarily caused by a greater frequency of moderate cost amino acids under high expression.

For *P. hawaiensis*, the relationship between Fop and $P_{S/C}$ was weakly positive (Spearman’s R = 0.050, $P < 10^{-15}$), consistent with a mild tendency for higher protein size under high expression, and implying a more complex dynamic between expression and protein size in this taxon. This is supported by the relationships between Fop/ expression level and amino acid frequencies per gene. Specifically, Fop/expression level was inversely correlated with the frequency of several amino acids per CDS, including Ser, Cys, Thr, and Leu (R between $-0.377$ and $-0.102$, $P < 10^{-15}$). In turn, the strongest positive correlations were found for Asp, Tyr, Gly, Lys, Glu and Met (R between $0.103$ and $0.195$, $P < 10^{-15}$). Thus, high Fop/expression level in *P. hawaiensis* genes favors use of a distinct set of amino acids as compared to *G. bimaculatus* and *O. fasciatus*. As shown in Figure 3C, when binned into four distinct classes of Fop values, the amino acids with the most positive (Asp) and negative (Ser) R values exhibited a clear link to Fop/expression level, but those with the second largest value (Tyr and Cys, respectively) exhibited a much weaker effect. Further, some amino acids that were more frequent under high Fop/expression level showed no clear trend in S/C score. For example, Tyr (with positive R) has among the highest S/C scores (57), whereas Gly (also with positive R) exhibits the lowest S/C score (1) (Table 2 and Table S3). This might indicate that a balance of large and small amino acids are preferred under elevated expression in *P. hawaiensis*, that codon and amino acid usage are subject to much weaker selection outside the highest expressed gene set (see Genome-wide optimal codon usage), or that factors other than protein size play a predominant role in shaping amino acid frequency in this organism.

### Codon usage and CDS lengths in *D. melanogaster* orthologs support translational selection

We evaluated the role of CDS length and GO annotation category in our findings using the well-studied and annotated model system *D. melanogaster* (Graveley et al. 2011; St Pierre et al. 2014) as a reference system. First, we determined whether genes exhibiting optimal codon usage in each of our three arthropods under study also exhibit optimal codon usage in their orthologs in *D. melanogaster*, a species previously shown to have GC3 optimal codons (Duret and Mouchiroud 1999). Thus, GC3 provides an effective measure of optimal codon usage in this taxon. Note that we did not cross compare the CDS sets of *G. bimaculatus*, *O. fasciatus* and *P. hawaiensis* since, given the *de novo* assembled transcriptomes that we used, apparent losses of orthologs among each pair might result from any one of expression below.

### Table 2 The Spearman rank correlations between the frequency of each amino acid per CDS and the Fop

<table>
<thead>
<tr>
<th>Amino acid</th>
<th>Gryllus, $R$</th>
<th>P</th>
<th>Oncopeltus, $R$</th>
<th>P</th>
<th>Parhyale, $R$</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arg</td>
<td>-0.198</td>
<td>&lt;0.001</td>
<td>Arg</td>
<td>-0.252</td>
<td>&lt;0.001</td>
<td>Ser</td>
</tr>
<tr>
<td>Thr</td>
<td>-0.123</td>
<td>&lt;0.001</td>
<td>Gly</td>
<td>-0.188</td>
<td>&lt;0.001</td>
<td>Cys</td>
</tr>
<tr>
<td>Pro</td>
<td>-0.122</td>
<td>&lt;0.001</td>
<td>Ala</td>
<td>-0.183</td>
<td>&lt;0.001</td>
<td>Thr</td>
</tr>
<tr>
<td>Ser</td>
<td>-0.110</td>
<td>&lt;0.001</td>
<td>Pro</td>
<td>-0.161</td>
<td>&lt;0.001</td>
<td>Leu</td>
</tr>
<tr>
<td>Ala</td>
<td>-0.093</td>
<td>&lt;0.001</td>
<td>Leu</td>
<td>-0.183</td>
<td>&lt;0.001</td>
<td>Arg</td>
</tr>
<tr>
<td>Leu</td>
<td>-0.083</td>
<td>&lt;0.001</td>
<td>Thr</td>
<td>-0.099</td>
<td>&lt;0.001</td>
<td>Pro</td>
</tr>
<tr>
<td>Gly</td>
<td>-0.061</td>
<td>&lt;0.001</td>
<td>Met</td>
<td>-0.061</td>
<td>&lt;0.001</td>
<td>Val</td>
</tr>
<tr>
<td>Met</td>
<td>-0.049</td>
<td>&lt;0.001</td>
<td>Thr</td>
<td>-0.059</td>
<td>&lt;0.001</td>
<td>Asn</td>
</tr>
<tr>
<td>Gln</td>
<td>-0.019</td>
<td>0.155</td>
<td>Val</td>
<td>-0.054</td>
<td>&lt;0.001</td>
<td>His</td>
</tr>
<tr>
<td>Trp</td>
<td>-0.008</td>
<td>0.571</td>
<td>Ser</td>
<td>-0.038</td>
<td>0.003</td>
<td>Ile</td>
</tr>
<tr>
<td>His</td>
<td>-0.003</td>
<td>0.850</td>
<td>Gln</td>
<td>-0.032</td>
<td>0.012</td>
<td>Ala</td>
</tr>
<tr>
<td>Cys</td>
<td>0.029</td>
<td>0.033</td>
<td>His</td>
<td>-0.200</td>
<td>0.110</td>
<td>Gln</td>
</tr>
<tr>
<td>Val</td>
<td>0.040</td>
<td>0.003</td>
<td>Cys</td>
<td>0.035</td>
<td>0.006</td>
<td>Trp</td>
</tr>
<tr>
<td>Phe</td>
<td>0.050</td>
<td>&lt;0.001</td>
<td>Tyr</td>
<td>0.059</td>
<td>&lt;0.001</td>
<td>Phe</td>
</tr>
<tr>
<td>Tyr</td>
<td>0.079</td>
<td>&lt;0.001</td>
<td>Phe</td>
<td>0.085</td>
<td>&lt;0.001</td>
<td>Met</td>
</tr>
<tr>
<td>Ile</td>
<td>0.115</td>
<td>&lt;0.001</td>
<td>Ile</td>
<td>0.198</td>
<td>&lt;0.001</td>
<td>Glu</td>
</tr>
<tr>
<td>Lys</td>
<td>0.151</td>
<td>&lt;0.001</td>
<td>Glu</td>
<td>0.211</td>
<td>&lt;0.001</td>
<td>Lys</td>
</tr>
<tr>
<td>Asn</td>
<td>0.169</td>
<td>&lt;0.001</td>
<td>Glu</td>
<td>0.226</td>
<td>&lt;0.001</td>
<td>Gly</td>
</tr>
<tr>
<td>Asp</td>
<td>0.269</td>
<td>&lt;0.001</td>
<td>Lys</td>
<td>0.272</td>
<td>&lt;0.001</td>
<td>Tyr</td>
</tr>
<tr>
<td>Glu</td>
<td>0.270</td>
<td>&lt;0.001</td>
<td>Asn</td>
<td>0.287</td>
<td>&lt;0.001</td>
<td>Asp</td>
</tr>
</tbody>
</table>

Species are abbreviated using their genus name. CDS, coding sequence; Fop, frequency of optimal codons.  
* Amino acids are listed from the most negative to positive R values for each species.
detection thresholds, assembly methods, or true gene loss, thereby minimizing gene sets; this ambiguity would not occur in *D. melanogaster*, which has a complete annotated gene list. Thus, we independently contrasted each gene set to their orthologs in *D. melanogaster*. The goal of this analysis was to determine whether orthologs to highly expressed genes in each of the three species under study, also

**Figure 3** The two amino acids with the largest positive (left) and negative (right) correlation to Fop/expression level in (A) *G. bimaculatus*; (B) *O. fasciatus*; and (C) *P. hawaiensis*. Fop was binned into four categories as shown. Spearman R correlations in Table 2 were calculated with the use of all (unbinned) data points.
exhibited elevated optimal codon usage in a divergent taxon. The number of D. melanogaster orthologs matching the CDS list from G. bimaculatus, O. fasciatus and P. hawaiensis was 3960 (74.9% of the CDS set under study), 4190 (68.0%) and 2822 (41.9%), respectively. Using the D. melanogaster orthologs of those CDS from the low (lowest 5% RPM), moderate (>5% RPM < 95%), and high (highest 5% RPM) categories in each of our three study species, we measured the ENC. We then evaluated the ENCs and GC3 (optimal codon) values across the three expression level categories. As shown in Figure 4A, the D. melanogaster orthologs of the G. bimaculatus genes in the highest expression category showed markedly lower ENC and higher GC3 than those from the low or moderate categories (P-value ranked analysis of variance [ANOVA] were 2.2 × 10⁻¹³ and 2.2 × 10⁻¹³, respectively, Dunn’s post hoc tests P < 0.05). This confirms optimization of codon usage in the highly expressed gene set across these divergent organisms, bias toward AT3 codons in G. bimaculatus and GC3 in D. melanogaster (note: higher expression in D. melanogaster is evident based on elevated optimal codon usage (defined here as GC3), (Cogshall and Wolfe 2000; Drummond et al. 2005; Popescu et al. 2006; Wall et al. 2005; Williford and Demuth 2012). For O. fasciatus, identical trends were detected (Figure 4B), wherein D. melanogaster orthologs of the high expression dataset had lower ENC and higher GC3 than those from the low and moderate classes (P-ranked ANOVA = 4.2 × 10⁻¹¹ and 6.9 × 10⁻¹⁷, respectively, Dunn’s post hoc tests P < 0.05). Finally, for P. hawaiensis, lower ENC and higher GC3 was also observed for orthologs matching the high expression category genes, than for the low and moderate expression categories (P = 5.8 × 10⁻⁷ and 3.3 × 10⁻⁴, respectively; Dunn’s post hoc tests P < 0.05) (Figure 4C), indicating a GC3 bias in each of these arthropods. Collectively, these results demonstrate that enhanced optimal codon usage in the high expression CDS dataset of G. bimaculatus, O. fasciatus and of P. hawaiensis is shared with their orthologs from the divergent relative D. melanogaster.

We next evaluated the relationship between CDS length and gene expression level for each arthropod under study, using lengths obtained from D. melanogaster orthologs. This is an effective approach as protein lengths tend to be highly conserved in eukaryotes (Wang et al. 2005), and was used since transcriptome assembly may not always yield full CDS lengths. Our objective was to determine whether high expression is linked to shorter/longer protein lengths. As shown in Figure S2, the length (number of codons) of the D. melanogaster orthologs matching CDS from G. bimaculatus, O. fasciatus and in P. hawaiensis were shortest for the high expression category, intermediate for moderate class, and longest for the lowest expression category for each of the species under study (ranked ANOVA P < 2.9 × 10⁻⁹ for all three species). This pattern indicates that translational selection favors short proteins as a function of expression (Figure 4). Taken together, our collective findings indicate that highly expressed genes encode short proteins (Figure 4), exhibit high Fop (Table 1 and Figure 2), and preferentially use specific amino acids (Table 2), each of which promote cost-efficient translation.

Given that highly expressed genes tend to be shorter than and lower expressed genes (Figure S2), we considered the role of genetic-linkage in our study. Selective sweeps or background selection can lead to fixation of linked nonoptimal codons, which may have greater effects in longer genes, and could reduce Fop (Comeran et al. 1999; Loewe and Charlesworth 2007; Whittle et al. 2011b). We thus compared Fop of long genes (>500 codons) from the high expression (highest 5% expression) vs. those from the low expression (lowest 5%) category for each species under study. We found that, consistent with the results of our analysis for all highly expressed genes regardless of length (Table 1 and Figure 2), Fop was statistically significantly elevated in the high-compared with the low-expression category for long genes in G. bimaculatus (MeanHigh = 0.483 ± 0.003, MeanLow = 0.441 ± 0.007 MWU P < 10⁻¹⁵), O. fasciatus (MeanHigh = 0.497 ± 0.005, MeanLow = 0.426 ± 0.008 MWU P < 10⁻¹⁵), and in P. hawaiensis (MeanHigh = 0.409 ± 0.013, MeanLow = 0.348 ± 0.009 MWU P = 0.0002). This argues against a pervasive role of genetic linkage, which, were it the major factor shaping codon use evolution, would be expected to yield similar Fop in long genes from both the high and low expression categories. Thus, long genes appear most influenced by their expression level, and are under greater selective pressure on codon usage when highly transcribed. Accordingly, we propose that genetic linkage, if it has an effect on Fop in these organisms, is largely restricted to the lower expression classes, where CDS are on average markedly longer (Figure S2) and purifying selective pressures are apt to be weaker due to their low transcription (Subramanian and Kumar 2004).

**Functional ontology annotation shows highly transcribed CDS are involved in cell-cycling and translation**

We wished to determine whether highly expressed genes with elevated optimal codon usage were preferentially associated with particular functions, biological processes or cellular components. To this end, we used D. melanogaster orthologs to assess GO annotation (Ashburner et al. 2000) of the CDS in all three arthropods under study. We clustered genes from the highest expression category (above the 95th percentile) in each species into functional groups using the procedures in DAVID (Huang da et al. 2009a,b). The results in Table 3 show that high expression was associated with enrichment of genes involved in cell-cycle processes (e.g., spindle organization, microtubule organization, proteosomes), translation (ribosome synthesis, RPGs) and DNA/RNA binding in all three arthropods under study, consistent with roles in rapid cell divisions in the embryo and/or reproductive tissues. Thus, the optimization of codon and amino acid usage observed in these organisms could be instrumental in minimizing the protein biosynthetic costs of cell division and protein turnover in the embryo and sexual tissues (Table S1). Remarkably, translational efficiency includes optimization of codons and amino acids not only in genes involved in cell division, but also in genes controlling protein synthesis itself (RPGs). This latter finding is concordant with extensive literature showing that protein synthesis genes are typically among the highest expressed in an organism (see for example Sharp and Li 1987; Wang et al. 2011).

For the CDS pooled from the moderate and low expression categories, we found enrichment of certain DNA/RNA and protein synthesis genes, primarily from the upper expression levels in the moderate class. However, the CDS with moderate and low expression levels spanned a vast range of functionalities rather than showing significant enrichment for any specific GO category compared with the high expression category (Table S4). Such a distribution of GO categories is expected for highly complex eukaryotic embryo/sexual tissues, which are typically comprised of an array of distinct cell types and transcripts (Combs and Eisen 2013; Diez-Roux et al. 2011; Donouhe and Extavour 2015; Ewen-Campen et al. 2011; Gilboa 2015; Jankovics et al. 2014). Although the moderate/low expressed genes are clearly biologically relevant, our data indicate they are less apt to influence codon or amino acid optimization (Figure 2, Figure 3, and Table 2) and have longer CDS lengths (Figure S2), likely due to their reduced transcriptional and translational levels, and thus biosynthetic costs. Together, the consistency in functionalities of the highly expressed dataset in all three arthropods studied herein reveals that optimization of codon and
amino acid usage is a shared feature of highly expressed genes, particularly those involved in cell division and translation, in these divergent arthropods.

**DISCUSSION**

**Optimal codon and amino acid usage**

Major codon and amino acid preferences under high expression point toward adaptation for translational accuracy and/or efficiency (Akashi 2001; Akashi 2003). Here, we identified clear preferences in codon usage and in amino acid frequency relative to expression in three emerging models of arthropods, *G. bimaculatus*, *O. fasciatus*, and *P. hawaiensis*. With respect to optimal codons, preferences have been previously reported in bacteria, fungi such as yeast and Neurospora, and some animals including *D. melanogaster* and *C. elegans*, but remain rare in arthropods outside of the higher insects (Akashi 2003; Duret and Mouchiroud 1999; Whittle et al. 2011a). Our results revealed AT3 optimal codons across the vast majority of amino acids in *G. bimaculatus* and *O. fasciatus*, whereas weaker yet significant GC3 preferences were observed in *P. hawaiensis*. Optimal codons have been linked to rapid and accurate translation, the efficient use of ribosomes and a reduced cost of proofreading (Akashi 2001, 2003) and may also improve transcription (Trotta 2011). Thus, the stronger bias in codon usage in highly expressed genes reported here (Table 1) likely reflects the fitness effects of a higher number of transcription and translation events, and in this regard is similar to other invertebrates including *D. melanogaster*, *C. elegans*, and *T. castaneum* (Akashi 2001, 2003; Duret 2000; Williford and Demuth 2012). Given that selection coefficients among synonymous codons are believed to be small, optimal codons may not be as prevalent in organisms with smaller population sizes (Akashi 2003). Thus, our results are consistent with sufficiently large population sizes in each of these three arthropods such that Nes > 1, allowing optimization of codon usage in each organism (Comeron 2004; Cutter et al. 2006; Duret and Mouchiroud 1999).

Both mutation and selection can contribute to codon usage. For example, in some organisms such as humans, GC3 content is correlated to GC of adjacent noncoding DNA, suggesting mutational pressures, in addition to selection, play a significant role in its codon usage profiles (Clay et al. 1996; Comeron 2004). It is also conceivable that AT3 or GC3 biases correlate to expression because highly transcribed genes are localized to regions that are inherently AT- or GC-rich, respectively. Our present results revealed substantial variation in codon bias among amino acids with the same degeneracy, which combined with the fact that amino acids are likely arbitrarily distributed throughout the genome, supports a role of selection herein. Further, the finding that higher degeneracy is connected to greater biases in codon usage is also consistent with selection, and might indicate that the presence of multiple tRNAs leads to greater pressure for optimization. Nonetheless, future sequence data for introns or other noncoding DNA from all three species under study will help further elucidate the precise role of selection relative to mutation in shaping codon usage in these models of the Pancrustacea.

Selective pressure on amino acids for efficient protein synthesis is much less well studied than codon usage. Revealing preferences for specific amino acids in a range of organism could play a major role in advancing our understanding of the dynamics underlying protein evolution, and thus warrants greater attention. Although some amino acid changes might individually have minor effects on cell physiology or fitness, collectively these effects could explain proteome-wide patterns of amino acid composition and evolution (Akashi 2003). Indeed, amino acid size and complexity has been linked to expression in the invertebrates *D. melanogaster*, *C. elegans* and *T. castaneum* as well as yeast
Table 3 Functional clustering of the most highly expressed CDS (above the 95th percentile) for each of three arthropod species under study using their orthologs in *Drosophila melanogaster* and the gene ontology system DAVID (Huang da et al. 2009a,b)

<table>
<thead>
<tr>
<th>Enrichment Score</th>
<th>P Valuea</th>
<th>Enrichment Score: 13.46</th>
<th>P Value</th>
<th>Enrichment Score: 10.7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gryllus bimaculatus</td>
<td>6.20E-5</td>
<td>Mitotic spindle organization</td>
<td>5.70E-18</td>
<td>Ribonuleoprotein</td>
</tr>
<tr>
<td>Proteasome regulatory particle complex</td>
<td>8.00E-5</td>
<td>Mitotic spindle organization</td>
<td>7.70E-14</td>
<td>Ribosomal protein</td>
</tr>
<tr>
<td>Enrichment Score: 3.9</td>
<td>3.70E-4</td>
<td>Mitotic spindle organization</td>
<td>3.80E-17</td>
<td>Structural constituent of ribosome</td>
</tr>
<tr>
<td>Spindle organization</td>
<td>6.30E-6</td>
<td>Microtubule organizing protein</td>
<td>2.40E-16</td>
<td>Ribosomal subunit</td>
</tr>
<tr>
<td>Cell-cycle process</td>
<td>3.80E-5</td>
<td>Microtubule-based process</td>
<td>1.20E-14</td>
<td>Ribosome</td>
</tr>
<tr>
<td>Mitotic spindle organization</td>
<td>7.10E-5</td>
<td>Cell cycle</td>
<td>2.10E-12</td>
<td>Cytosolic ribosome</td>
</tr>
<tr>
<td>Mitotic cell cycle</td>
<td>8.10E-5</td>
<td>M phase</td>
<td>8.60E-12</td>
<td>Ribosome</td>
</tr>
<tr>
<td>Cell cycle</td>
<td>1.20E-4</td>
<td>Microtubule cytoskeleton</td>
<td>1.90E-11</td>
<td>Structural molecule activity</td>
</tr>
<tr>
<td>Microtubule cytoskeleton organization</td>
<td>2.20E-4</td>
<td>Cell-cycle phase</td>
<td>2.00E-11</td>
<td>Large ribosomal subunit</td>
</tr>
<tr>
<td>Microtubule-based process</td>
<td>4.40E-4</td>
<td>Cell-cycle phase</td>
<td>2.50E-11</td>
<td>Cytosolic part</td>
</tr>
<tr>
<td>M phase</td>
<td>5.70E-4</td>
<td>Cell cycle</td>
<td>2.10E-12</td>
<td>Cytosolic ribosome</td>
</tr>
<tr>
<td>Cell-cycle phase</td>
<td>9.30E-4</td>
<td>M phase</td>
<td>8.60E-12</td>
<td>Ribosome</td>
</tr>
<tr>
<td>Proteasome complex</td>
<td>3.70E-4</td>
<td>Ribosomal protein</td>
<td>1.30E-8</td>
<td>Mitotic spindle organization</td>
</tr>
<tr>
<td>Proteasome</td>
<td>1.00E-5</td>
<td>Ribosome</td>
<td>4.10E-8</td>
<td>Microtubule cytoskeleton organization</td>
</tr>
<tr>
<td>Proteasome</td>
<td>3.20E-3</td>
<td>Cytosolic large ribosomal subunit</td>
<td>9.50E-8</td>
<td>Spindle organization</td>
</tr>
<tr>
<td>Enrichment Score: 3.44</td>
<td>1.60E-5</td>
<td>Structural constituent of ribosome</td>
<td>1.80E-6</td>
<td>Microtubule-based process</td>
</tr>
<tr>
<td>Ribosomal subunit</td>
<td>5.10E-5</td>
<td>M phase</td>
<td>6.90E-5</td>
<td>Mitotic cell cycle</td>
</tr>
<tr>
<td>Nucleosome-binding</td>
<td>6.90E-5</td>
<td>Cell-cycle phase</td>
<td>3.70E-5</td>
<td>Cytoskeleton organization</td>
</tr>
<tr>
<td>Adenyl nucleotide binding</td>
<td>1.40E-3</td>
<td>Cell-cycle process</td>
<td>1.40E-4</td>
<td>Cell cycle</td>
</tr>
<tr>
<td>Purine nucleoside binding</td>
<td>9.70E-4</td>
<td>Enrichment Score: 4.44</td>
<td>5.40E-8</td>
<td>Cell cycle</td>
</tr>
<tr>
<td>ATP binding</td>
<td>9.80E-4</td>
<td>Nucleotide-binding</td>
<td>5.40E-8</td>
<td>Cell cycle</td>
</tr>
<tr>
<td>Adenyl ribonucleotide binding</td>
<td>1.00E-3</td>
<td>Atp-binding</td>
<td>1.70E-7</td>
<td>Enrichment Score: 3.63</td>
</tr>
<tr>
<td>Nucleosome-binding</td>
<td>1.10E-3</td>
<td>Nucleotide binding</td>
<td>7.40E-6</td>
<td>Transit peptide</td>
</tr>
<tr>
<td>Purine nucleotide binding</td>
<td>1.70E-3</td>
<td>Purine nucleotide binding</td>
<td>5.90E-5</td>
<td>Transit peptide/mitochondrion</td>
</tr>
<tr>
<td>Ribonucleotide binding</td>
<td>1.90E-3</td>
<td>Purine ribonucleotide binding</td>
<td>7.50E-5</td>
<td>Enrichment Score: 3.26</td>
</tr>
<tr>
<td>Purine ribonucleotide binding</td>
<td>1.90E-3</td>
<td>Ribonucleotide binding</td>
<td>7.50E-5</td>
<td>Proteasome</td>
</tr>
<tr>
<td>Enrichment Score: 3.31</td>
<td>1.20E-4</td>
<td>ATP binding</td>
<td>2.10E-4</td>
<td>Proteasome</td>
</tr>
<tr>
<td>PINT</td>
<td>1.90E-4</td>
<td>Adenyl ribonucleotide binding</td>
<td>2.50E-4</td>
<td>Proteasome</td>
</tr>
<tr>
<td>Proteasome component region PCI</td>
<td>5.40E-3</td>
<td>Adenyl nucleotide binding</td>
<td>4.30E-4</td>
<td>Proteasome complex</td>
</tr>
<tr>
<td>Domain:PCI</td>
<td>5.40E-4</td>
<td>Nucleoside binding</td>
<td>4.80E-4</td>
<td>Enrichment Score: 2.59</td>
</tr>
<tr>
<td>Enrichment Score: 2.96</td>
<td>5.40E-4</td>
<td>Nucleoside binding</td>
<td>5.50E-4</td>
<td>PINT</td>
</tr>
<tr>
<td>Chaperone</td>
<td>5.60E-4</td>
<td>Nucleoside binding</td>
<td>5.50E-4</td>
<td>Proteasome</td>
</tr>
<tr>
<td>Domain:PCI</td>
<td>5.40E-4</td>
<td>Chaperone</td>
<td>8.0E-4</td>
<td>Chaperonin TCP-1</td>
</tr>
<tr>
<td>Chaperonin TCP-1</td>
<td>7.4E-4</td>
<td>Chaperone</td>
<td>8.0E-4</td>
<td>Chaperonin TCP-1</td>
</tr>
<tr>
<td>Chaperonin-containing T-complex</td>
<td>1.40E-3</td>
<td>Chaperonin-containing T-complex</td>
<td>2.70E-5</td>
<td>Chaperonin TCP-1</td>
</tr>
<tr>
<td>Chaperonin Cpn60/TCP-1</td>
<td>2.50E-3</td>
<td>Chaperonin-containing T-complex</td>
<td>7.60E-5</td>
<td>PIRSF002584: molecular chaperone t-complex</td>
</tr>
<tr>
<td>Ribosomal protein</td>
<td>9.80E-5</td>
<td>Chaperonin Cpn60/TCP-1</td>
<td>4.00E-4</td>
<td>Chaperon</td>
</tr>
<tr>
<td>Ribonucleoprotein</td>
<td>1.00E-4</td>
<td>PIRSF002584: molecular chaperone t-complex</td>
<td>2.70E-5</td>
<td>Proteasome</td>
</tr>
<tr>
<td>Cytosolic large ribosomal subunit</td>
<td>2.20E-4</td>
<td>Ribosomal protein</td>
<td>7.60E-4</td>
<td>Enrichment Score: 3.91</td>
</tr>
<tr>
<td>Cytosolic ribosome</td>
<td>4.30E-4</td>
<td>Proteasome</td>
<td>3.60E-6</td>
<td>Proteasome</td>
</tr>
<tr>
<td>Structural constituent of ribosome</td>
<td>2.00E-3</td>
<td>Proteasome complex</td>
<td>7.00E-4</td>
<td>Proteasome</td>
</tr>
<tr>
<td>Ribosome</td>
<td>2.50E-3</td>
<td>Proteasome</td>
<td>7.60E-4</td>
<td>Proteasome</td>
</tr>
<tr>
<td>Large ribosomal subunit</td>
<td>3.70E-3</td>
<td>Proteasome</td>
<td>2.40E-3</td>
<td>Proteasome</td>
</tr>
<tr>
<td>Ribosomal subunit</td>
<td>1.10E-2</td>
<td>Proteasome regulatory particle</td>
<td>3.00E-3</td>
<td>Proteasome</td>
</tr>
<tr>
<td>Ribosome</td>
<td>2.60E-2</td>
<td>Proteasome regulatory particle</td>
<td>3.00E-3</td>
<td>Proteasome</td>
</tr>
<tr>
<td>Structural molecule activity</td>
<td>8.20E-2</td>
<td>Proteasome accessory complex</td>
<td>3.50E-3</td>
<td>Proteasome</td>
</tr>
</tbody>
</table>

**Notes:**
- CDS, coding sequence.
- P values represent a modified Fisher’s test, wherein lower values indicate greater enrichment.
- Functional categories with enrichment values > 2.5 are shown.

Value 2.5 are shown.

**Notes:**
- CDS, coding sequence.
- P values represent a modified Fisher’s test, wherein lower values indicate greater enrichment.
- Functional categories with enrichment values > 2.5 are shown.
Our findings extend these conclusions to include \textit{G. bimaculatus} and \textit{O. fasciatus}. However, our results differ from those of other eukaryotes in that we report that highly expressed genes exhibit a prevalence of amino acids with moderate size and complexity, which does not include the smallest classes, suggesting a benefit to utilizing mid-range amino acids, rather than using the smallest in these insects. It is feasible that these amino acids with moderate size/complexity, while more metabolically costly than the smallest, confer an advantage to protein stability or activity.

For \textit{P. hawaiensis}, the dynamic between Fop/expression level and amino acid biosynthesis costs is less straightforward, as amino acids with both high and low S/C scores exhibited statistically significant positive correlations to Fop (Table 2 and Figure 3). This suggests that there is a selective advantage of a mixture of large and small amino acids on protein stability or function in this taxon. However, another interpretation is that only a few amino acids are optimized in this taxon. For instance, the strong positive correlation between Fop/expression and frequency of Asp, and the strong negative correlation for Ser (Table 2 and Figure 3), each demonstrate unambiguous signals of amino acid optimization. However, the weaker signals from the other amino acids (Figure 3) might indicate that the main or strongest effects of expression are limited to Asp and Ser. In addition, the weaker signal of optimal codon usage in \textit{P. hawaiensis}, which is very mild outside the highest expression class, suggests that a more complex dynamic underlies this phenomenon for this taxon. Further studies will be needed to ascertain the dynamics of codon and amino acid usage in this crustacean. Nevertheless, taken together, our data provides evidence of optimization of codon and amino acid usage in highly transcribed genes across all three of these distinct arthropods.

The fact that \textit{D. melanogaster} orthologs to CDS from the high expression category of \textit{G. bimaculatus} and of \textit{O. fasciatus} each had higher GC3 frequencies than orthologs from lower expression classes (Figure 4, A and B), confirms that optimal codon usage is retained across the same gene sets among divergent insects. Thus, despite the fact these various insects have evolved distinct sets of optimal codons, selection for a subset of codons that promote translational efficiency of highly expressed CDS appears to be a conserved trait. Further, our findings showing elevated usage of GC3 optimal codons in highly transcribed CDS of \textit{P. hawaiensis} and their orthologs in \textit{D. melanogaster} (Figure 2 and Figure 4) reveals similar evolution traits across the Pan-arthropods. Functional categorization (GO) analyses indicate that for all three species studied herein, the CDS from the highest Fop/expression category primarily included genes involved in cell-cycling and protein synthesis, such as RPGs (which are typically highly expressed) (Wang \textit{et al.} 2011) (Table 3), suggesting that such fundamental processes are especially prone to selection to reduce biosynthetic costs (Seligmann 2003; Swire 2007; Williford and Demuth 2012). Given that transcripts of cell-cycling and translation genes are known to be prevalent throughout embryogenesis and in later developmental phases in arthropods such as \textit{D. melanogaster} and other invertebrates including \textit{C. elegans} (Li \textit{et al.} 2014), optimization of codons and amino acids in these gene sets could potentially markedly reduce lifetime biosynthetic costs (Duret 2000; Seligmann 2003; Swire 2007; Williford and Demuth 2012) and enhance fitness in these organisms.

Our finding that highly expressed CDS encoded short proteins in \textit{G. bimaculatus}, \textit{O. fasciatus}, and in \textit{P. hawaiensis} (Figure S2) concurs with previous studies in eukaryotes such as \textit{D. melanogaster}, \textit{T. castaneum}, \textit{D. viridis}, \textit{Y. p. tenera}, \textit{C. elegans}, humans, yeast, and some plants (Akashi 2001, 2003; Comeron 2004; Ingvarsson 2007; Lemor \textit{et al.} 2005; Moriyama and Powell 1998; Urrutia and Hurst 2001; Whittle \textit{et al.} 2007; Williford and Demuth 2012). The phenomenon of shorter CDS in highly expressed genes might reflect the action of selection to minimize the protein length, which combined with high optimal codon usage and preferential amino acid usage observed herein (Figure 2 and Table 2), potentially reduces the costs of transcription, translation, and cellular transport, in these arthropods (Akashi 2001, 2003). With respect to translational costs, reductions in protein length of highly transcribed genes should have similar advantages to the usage of optimal codons and of reduced-cost amino acids in terms of biosynthesis, suggesting potential for co-evolution of these traits (Akashi 2001, 2003).

The fact that long CDS under high transcription had elevated Fop compared with their lowly transcribed counterparts supports the hypothesis that expression level, rather than genetic-linkage, is the primary factor shaping Fop differences among expression classes. Nevertheless, it is formally possible that genetic linkage may contribute toward the observed reduced Fop in the lower expression level datasets, as longer genes appear most prone to fixation of non-optimal codons (Betancourt and P Gratis 2002; Comeron \textit{et al.} 1999; Loewe and Charlesworth 2007; Whittle and Johannesson 2013). This linkage phenomenon might be most common in genes with weak expression levels, which would be expected to experience lower purifying selection pressures (Subramanian and Kumar 2004). Future data on amino acid substitutions and linked fixations of nonoptimal codons with respect to CDS length among related species in each taxonomic group (\textit{Gryllus, Oncopeltus,} and \textit{Parhyale}) will help further discern the role of genetic-interference on Fop (Betancourt and P Gratis 2002; Kim 2004; Whittle and Johannesson 2013). In addition, when available in the future, genomic data for all three of these arthropods will allow precise measures of CDS lengths in all genes under study (not only those with orthologs in \textit{D. melanogaster}), providing a means to test fine-scale gene-length effects on codon usage within expression classes, which to date have typically shown higher codon bias in shorter genes in eukaryotes (Akashi 2001, 2003; Cutter 2008; Duret and Mouchirioud 1999; Moriyama and Powell 1998; Whittle \textit{et al.} 2007, 2011a).

Together, it is evident that at a genome-wide level, high transcription is accompanied by a marked shift toward short protein lengths (Figure S2) and this shift co-occurs with increased usage of optimal codons (Figure 2) and specific amino acids (Table 2 and Figure 3). This points toward multifaceted aspects of translational selection, which affect several genomic traits (codons, amino acids, and CDS length) to yield efficient protein synthesis in these arthropods.

\textbf{Fop as a predictor of expression}

While Fop, or other codon adaptation indices, have been shown to reliably reflect the expression level in an organism (Coglan and Wolfe 2000; Drummond \textit{et al.} 2005; Popescu \textit{et al.} 2006; Wall \textit{et al.} 2005), it is worthwhile to note that selection at the protein level can influence optimal codon usage. For instance, at the interspecies level, proteins that evolve rapidly (displaying high nonsynonymous to synonymous substitution rates, \textit{dN/dS}) (Comeron \textit{et al.} 1999; Loewe and Charlesworth 2007; Whittle and Johannesson 2013; Yang 2007) often exhibit reductions in optimal codon usage. Adaptive evolution at the protein level may cause selective sweeps that fix linked non-optimal codon mutations, and relaxed selection on proteins often correlates to losses in optimal codons (Betancourt and P Gratis 2002; Kim 2004; Sella \textit{et al.} 2009; Whittle \textit{et al.} 2011b). Relaxed selection on proteins or synonymous codons may arise from factors such as reduced population size or lowered recombination rates over an organism’s evolutionary history (Betancourt and P Gratis 2002; Comeron and Kreitman 1998; Haddrill \textit{et al.} 2011; Sella \textit{et al.} 2009; Whittle \textit{et al.} 2011b). However, the relationship between Fop and expression
should not be uncoupled by these processes, but rather is likely maintained due to the fact that rates of protein evolution appear to be inversely correlated to expression level in various organisms including mice, fruit fly, and yeast (see for example Drummond et al. 2005; Lemos et al. 2005; Pal et al. 2001; Subramanian and Kumar 2004). The conservation of highly expressed genes at the protein sequence level may be explained by their essentiality, or indispensability, to fitness (Mank and Ellegren 2009). In addition, highly expressed genes might also exhibit signatures of selection on (otherwise) neutral amino acids if such residues are needed to promote translational efficiency, for example because they promote a small size or specific conformation (Akashi 2003). In this regard, Fop and expression level likely remain connected even for genes with rapid rates of protein evolution, such that Fop would decrease under rapid protein evolution, which is more apt to occur in genes with lowered expression. Given that we observed clear relationships between Fop and RPM, and between Fop and the S/C score and amino acid composition (Table 2 and Figure 3), we suggest that in these arthropods, optimal codon usage reflects expression level on a genome-wide level, and that optimal codon use has evolved to promote efficient translation. In this regard, the relationships observed between Fop and amino acid composition would directly reflect adaptation for fast and accurate translation in genes with high expression.

We conclude that highly expressed genes, in particular those involved in cell-cycling and protein synthesis (Table 3), exhibit a history of coevolution of optimal codons (Table 1 and Figure 2), the usage of specific amino acids (Table 2 and Figure 3) and shorter CDS length (Figure S2), which act in concert to promote accurate and cost efficient translation in these divergent arthropods. Our findings indicate that selection favors specific optimal codons and amino acids in three emerging models of arthropods, G. bimaculatus, O. fasciatus, and P. hawaiiensis. For both insects, a particularly strong signal of optimal codons was identified. Further, the marked connection between Fop and amino acid size and complexity suggests an adaptive strategy to promote efficient translation, which likely allows fast rates of accurate protein synthesis. Although weaker selective effects on codon usage, with fewer optimal codons, were evident in P. hawaiiensis compared with the two insects, this taxon also exhibits a significant connection among Fop and size/complexity, suggesting coevolution that appears to favor both small and large amino acids has also occurred in this crustacean. Thus, the dynamics of protein evolution may be more complex in this taxon, wherein protein stability or function in highly expressed genes may necessitate a wide range of amino acids. Further studies should explore the variation in amino acid and synonymous codon mutations and specific types of amino acid replacements at the population level (Akashi and Schaeffer 1997) to reveal whether selective variation can be observed at the intraspecies level in these arthropods. Furthermore, future availability of genomes from these three species will allow further study of other facets of translational selection, including the relationship between tRNA gene abundance and expression, which is particularly relevant to understanding the evolution of optimal codons and its contribution to functional genome evolution.

ACKNOWLEDGMENTS

We thank the anonymous reviewers for valuable comments on the manuscript.

LITERATURE CITED


