University of Nevada, Reno

Differences in the Protein Evolutionary Rates of *Arabidopsis* Species and Codon Usage Biases in the Tissues of *Drosophila melanogaster*

A thesis submitted in partial fulfillment
of the requirements for the degree
of Master of Science in Biology

by

Bryan L Payne

Dr. David Alvarez-Ponce/Thesis Advisor

May, 2018

THE GRADUATE SCHOOL

We recommend that the thesis
prepared under our supervision by

BRYAN L PAYNE

Entitled

Differences In The Protein Evolutionary Rates Of Arabidopsis Species And Codon Usage Biases In The Tissues Of Drosophila Melanogaster

be accepted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE

David Alvarez-Ponce, Advisor

Guy Hoelzer, Committee Member

Marjorie Matocq, Graduate School Representative

David W. Zeh, Ph.D., Dean, Graduate School

May, 2018

Abstract

Protein evolution in an organism or population is determined by a host of phenomena that affect the overall rate of change in the genome. Selection is thought to have a decreased effect in self-fertilizing plants due to decreased effective population size. *Arabidopsis thaliana* transitioned to self-fertilization while its congeners *A. lyrata* and *A. halleri* have retained obligatory out-crossing. The rate of protein change, measured as the nonsynonymous to synonymous divergence ratio (d_N/d_S), showed an increase in evolutionary rate in *A. thaliana* compared to *A. lyrata* and *A. halleri*, likely due to self-fertilization. Preferential codon usage is affected by translational selection, decreasing the rate of synonymous substitution. Preferred codons tend to correspond to the most abundant tRNAs. Codon usage biases differ amongst species and can be different among the tissues of an organism if relative tRNA abundances differ between the different tissues. Previous studies have found that the differences in codon usage biases may also be attributed to GC content and not only to differences in tissue specific protein expression. In the genome of *Drosophila melanogaster,* I have found that patterns of codon usage are different amongst proteins expressed in different tissues. Using randomized datasets, I show that these differences are always explained by which tissue these genes are expressed in, and are not due to other confounding properties of these proteins, such as GC content, protein length, or protein expression levels.

Acknowledgements

First, I would like to thank my wife, Georgia, for helping me balance life and work. Thank you for being the most stable person as I took the crazy journey that is graduate school. I literally could not have done this without your support. I also could not have done this without the additional support of the rest of my friends and family.

Secondly, I would like to thank Dr. David Alvarez-Ponce for the excellent guidance and mentorship over the last two years. You have encouraged me to think more critically, challenged me to do better, and helped me to achieve this. To my committee members, Guy Hoelzer and Marjorie Matocq, thank you for support and expertise in my undergraduate and graduate career.

Contents

List of Tables

List of Figures

Chapter 1: Introduction

The natural selection and evolution in an organism or population are affected by many phenomena. Classical population level factors that affect evolution include migration, genetic drift, selection, and mutation. Rates of protein evolution are also governed by effects within an organism that curb mutation and its effects on translated proteins. In this work, the genomes of members of the *Arabidopsis* genus were compared to survey the effect of reduced efficacy of selection due to reduced effective population size in *Arabidopsis thaliana*. In the second chapter, gene expression data from *Drosophila melanogaster* was used to study codon usage within genes expressed in different tissues of an organism, which in turn can have a marked effect on protein evolution.

The rate of nonsynonymous mutations is determined as the number of mutations that cause changes in the amino acid per nonsynonymous site in the protein sequence and is denoted by d_N. Inversely, d_S is the rate of synonymous mutations per synonymous site, or mutations that do not change the amino acid of the translated protein. The ratio of d_N and d_S is the rate at which nonsynonymous changes are fixed relatively to synonymous changes and provides a context for studying selection in a protein coding sequence. When d_N/d_S is greater than 1, selection is promoting fixation of nonsynonymous mutations and the gene is undergoing positive selection; a d_N/d_S of 1 is the described as neutral evolution (e.g., as expected in a pseudogene); and a d_N/d_S that is less than 1 is where selection is preventing the accumulation of nonsynonymous mutations, or purifying selection. The efficacy of selection, and therefore the rate of protein change, is subject to a cascade of factors, one of which is effective population size (for review, see (Alvarez-Ponce, 2014)).

Effective population size or N_e has long been proven, in multiple species, to have a large effect on evolutionary rates (Haudry et al. 2008; Escobar et al. 2010; Qiu et al. 2011; Ness et al. 2012; Slotte et al. 2013). The efficacy of selection diminishes with decreases in the effective population sizes of an organism, allowing for the accumulation of deleterious alleles (Ohta 1973; Kimura 1983; Charlesworth and Wright 2001). A cause of reduced effective population size in plants is the transition from out-crossing to self-fertilization (Pollak 1987). The introduction of self-compatibility is thought to have happened in *A. thaliana* between 100,000 and 1,000,000 years ago, providing an opportunity to examine these effects in an organism with high availability of genomic data (Charlesworth and Vekemans 2005; Bechsgaard et al. 2006; Durvasula et al. 2017).

The closely related *Arabidopsis lyrata* and *Arabidopsis halleri* retain the ancestral state of obligate out-crossing. Previous works in *Arabidopsis* have tried to characterize a relaxation of purifying selection, and therefore increased d_N/d_S ratios, by comparing the genome of *A. thaliana* to relatively few orthologous proteins in *A. lyrata* (Wright et al. 2002; Foxe et al. 2008). Wright and collaborators found that *A. thaliana* was not evolving faster in 13 proteins (Wright et al. 2002). Foxe and collaborators found increased d_N/d_S in ~ 600 and73 proteins in *A. thaliana* and *A. lyrata* respectively, however these results may have been biased by preferential inclusion of highly expressed genes in the *A. lyrata* dataset, which evolve slower than lowly expressed proteins (Foxe et al. 2008). Because these results were limited by the data available at the time, genomic studies in other species and genera have allowed further exploration on the effects of self-fertilizing on protein evolution.

Studies in *Triticeae* (Haudry et al. 2008; Escobar et al. 2010), *Capsella* (Qiu et al. 2011; Slotte et al. 2013), and *Mimulus* (Brandvain et al. 2014), among others, have

found that increases in d_N/d_S are detectable in self-fertilizing species with large datasets (for review see (Shimizu and Tsuchimatsu 2015)). With the recent release and annotation of genomes for *A. lyrata* and *A. halleri* a comparison of the patterns of evolution of these out-crossing species with those of *A. thaliana* provides an excellent context to study the effects of reduced effective population size on genetic evolution (Hu et al. 2011; Briskine et al. 2016). I determined differences in d_N/d_S in the self-fertilizing *A. thaliana* and the out-crossing *A. lyrata* and *A. halleri* by finding orthologous proteins in *A. thaliana*, *Capsella rubella* (as the outgroup), and either *A. halleri* or *A. lyrata*. The rates of evolution in *A. thaliana* were compared to those of its sister taxa to explore the effects of self-fertilization on selection and evolution.

Protein evolution can also be affected by several factors within an organism as well. Codon usage bias is categorized as the cell's preference to use particular codons for encoding amino acids (for review, see (Hanson and Coller 2017). Preferred codons are thought to be favored by selection, decreasing the likelihood of fixation of synonymous mutations (Akashi 1994; Yang and Nielsen 2008). Preferred codons are more efficiently translated within the cell by their corresponding tRNA, which are found in higher abundances than their non-preferred alternatives (Moriyama and Powell 1997; Hanson and Coller 2017). Biases within an organism are also correlated with differences in expression patterns, often attributed to tRNA abundances (Lu et al. 2006; Olivares-Hernández et al. 2011).

Differences in tRNA abundances and codon usage have been noted across many species (Muto and Osawa 1987; Kanaya et al. 2001; Rocha 2004; Vicario et al. 2007). Human tissues have been noted to have different codon usage profiles than the codons preferred in the whole organism (Dittmar et al. 2006). The effects of the tissue on codon usage have been disputed in studies of human tissues using small numbers of

tissue specific proteins (Plotkin et al. 2004; Sémon et al. 2006). In a study of less than 200 tissue-specific genes, Plotkin *et al.* found that these differences were determined by the tissue that the protein was expressed in (Plotkin et al. 2004). Sémon *et al.* disputed that only 2.3% of this variability was due to differences among tissues, and that this result was mostly due to differences in GC content, using a dataset of 2126 proteins (Sémon et al. 2006). In *Arabidopsis*, however, Camiolo *et al.* found that codon usage biases were significantly linked to tissue specificity (Camiolo et al. 2012).

The genome of *Drosophila melanogaster* has been the subject of extensive studies in tissue specific protein expression and codon usage bias, and these data were leveraged to explore the codon usage biases in each of these tissues (Chintapalli et al. 2007; Vicario et al. 2007). To determine if the codon usage biases in tissue-specific proteins were due to the tissue that they are expressed in, and not to potentially confounding factors, permutational analyses of variance and correspondence analyses were performed (Plotkin et al. 2004; Sémon et al. 2006). Permutational analyses showed significant results in *Arabidopsis* and correspondence analyses were used to disprove Plotkin et al.'s results in humans (Plotkin et al. 2004; Sémon et al. 2006; Camiolo et al. 2012)

The following chapters explore and determine that previous analyses regarding evolutionary rates in *A. thaliana* and tissue specific codon usage biases may have been influenced by a lack of available data. Rates of nonsynonymous and synonymous substitutions of orthologous genes in *A. thaliana*, *A. lyrata*, and *A. halleri* show that evolutionary rates are indeed increased in *A. thaliana*. Codon usage biases in proteins expressed exclusively in the different tissues of *D. melanogaster* can be attributed to the tissue that the proteins are expressed in. Together these studies provide additional

evidence that the process of evolution is complex and is affected by subtle factors detectable through large amounts of genetic information.

Chapter 2: Higher rates of protein evolution in the self-fertilizing plant *Arabidopsis thaliana* than in the out-crossers *Arabidopsis lyrata* and *Arabidopsis halleri*

Abstract

The common transition from out-crossing to self-fertilization in plants decreases effective population size. This is expected to result in a reduced efficacy of natural selection and in increased rates of protein evolution in selfing plants compared to their outcrossing congeners. Prior analyses, based on a very limited number of genes, detected no differences between the rates of protein evolution in the selfing *Arabidopsis thaliana* compared to the out-crosser *Arabidopsis lyrata.* Here, we re-evaluate this trend using the complete genomes of *A. thaliana*, *A. lyrata*, *Arabidopsis halleri* and the outgroups *Capsella rubella* and *Thellungiella parvula*. Our analyses indicate slightly but measurably higher nonsynonymous divergences (d_N), synonymous divergences (d_S) and d_N/d_S ratios in *A. thaliana* compared with the other *Arabidopsis* species, indicating that purifying selection is indeed less efficacious in *A. thaliana*.

Introduction

In plants, the transition from out-crossing to self-fertilization is quite common, and is generally seen as a dead-end due to accumulation of deleterious mutations (Stebbins 1957). Population genetics theory predicts that selfing organisms will have a lower effective population size (N_e) than their outcrossing congeners with the same population size (Pollak 1987). Reduced N_e is expected to result in a reduced efficacy of natural selection (Charlesworth et al. 1993; Charlesworth and Wright 2001), thus allowing the fixation of slightly deleterious mutations (Ohta 1973). As a result, selfing organisms are expected to exhibit accelerated rates of protein evolution (Kimura 1983; Charlesworth and Wright 2001) and less codon usage bias (Qiu et al. 2011). These predictions are supported by some empirical evidence: natural selection is reduced in selfing species of the family Triticeae (Haudry et al. 2008; Escobar et al. 2010) and the genera *Capsella* (Johnston et al. 2008; Qiu et al. 2011; Slotte et al. 2013), *Eichhornia* (Ness et al. 2012), *Collinsia* (Hazzouri et al. 2013) and *Mimulus* (Brandvain et al. 2014) (for review, see (Hough et al. 2014) and (Shimizu and Tsuchimatsu 2015)). In addition, an analysis of polymorphism data for a number of plant species revealed a weak increase in the nonsynonymous to synonymous polymorphism ratio (π_a/π_s) of selfers (Glémin et al. 2006).

The plant *Arabidopsis thaliana* is thought to have shifted to self-fertilization 150,000-1,000,000 years ago (Charlesworth and Vekemans 2005; Bechsgaard et al. 2006). In agreement with the predicted reduction in the efficacy of natural selection, this species exhibits less codon bias than the out-crosser *Arabidopsis lyrata* (Qiu et al. 2011). However, analysis of 16 genes did not detect significantly higher rates of protein evolution in *A. thaliana* compared to *A. lyrata* (Wright et al. 2002). In addition, comparison of 13 pairs of orthologous genes in these two species revealed no differences in the ratios of nonsynonymous to synonymous polymorphisms or in the ratios of nonsynonymous to

synonymous fixations (Foxe et al. 2008). A comparison of 675 *A. thaliana* and 73 *A. lyrata* non-orthologous genes found higher ratios of nonsynonymous to synonymous polymorphisms and higher ratios of nonsynonymous to synonymous fixations in *A. thaliana* (Foxe et al. 2008); however, these results may have been affected by biases in the dataset – e.g., 7 of the *A. lyrata* genes were chosen due to their high levels of expression, and highly expressed genes tend to evolve under strong purifying selection (Pál et al. 2001; Drummond et al. 2005).

These analyses, in any case, were limited by the very small amount of genomic information available at the time. Here, we revisit the prediction that *A. thaliana* should exhibit faster rates of protein evolution than *A. lyrata* or *Arabidopsis halleri* taking advantage of the now completely sequenced genomes of *A. thaliana* (2000), *A. lyrata* (Hu et al. 2011), *A. halleri* (Briskine et al. 2016) and the outgroup *Capsella rubella* (Slotte et al. 2013). *A. thaliana* diverged 6-13 MYA from the *A. lyrata*/*A. halleri* clade (Beilstein et al. 2010) and 8-14 MYA from *C. rubella* (Koch and Kiefer 2005) (Fig. 2.1).

Results

For each *C. rubella* gene, we identified the most likely orthologs in *A. thaliana* and *A. lyrata*. For each of the 18,107 identified trios, protein sequences were aligned, and the resulting alignments were used to guide the alignment of the corresponding coding sequences (CDSs). To reduce the impact of annotation errors, we removed all alignments for which >5% of positions included gaps. For each of the resulting 12,994 alignments, PAML (free-ratios model; (Yang 2007)) was used to estimate the nonsynonymous divergence (d_N), synonymous divergence (d_S) and the nonsynonymous to synonymous divergence ratio ($\omega = d_N/d_S$) in each of the branches of the phylogeny (Fig. 2.1). The ratio d_N/d_S is expected to be lower than 1 when nonsynonymous mutations are under purifying selection (with values closer to 0 indicating stronger selection), equal to 1 when protein

sequences evolve neutrally, and higher than 1 for genes under positive selection (for review, see (Alvarez-Ponce 2014)).

In the *A. thaliana* branch, the median of the values estimated by the free-ratios model were $d_N = 0.0108$, $d_S = 0.0757$ and $d_N/d_S = 0.1427$, and the mean values were $d_N = 0.0133$, $d_S = 0.0805$ and $d_N/d_S = 0.1865$. In the *A. lyrata* branch, the median values were $d_N = 0.0085$, $d_S = 0.0612$ and $d_N/d_S = 0.1389$, and the mean values were $d_N = 0.0107$, $d_S = 0.0667$ and $d_N/d_S = 0.1880$ (Table S2.1; Fig. 2.2). A Mann-Whitney *U* test showed significant differences in the d_N ($P = 1.964 \times 10^{-119}$), d_S ($P < 10^{-300}$) and d_N/d_S ($P = 0.0127$) of both species. In 8572 of the cases, d_N was higher in *A. thaliana* than in *A. lyrata*, and in 4396 of the cases d_N was higher in *A. lyrata*, indicating that rates of protein sequence evolution are often higher in *A. thaliana* (binomial test, $P = 1.20 \times 10^{-324}$). In 8938 of the cases, d_S was higher in *A. thaliana*, and in 4055 of the cases d_S was higher in *A. lyrata*, indicating faster rates of evolution of synonymous sites in *A. thaliana* (binomial test, $P = 4.94 \times 10^{-324}$); these results are consistent with prior studies reporting higher mutation rates in *A. thaliana* (Yang et al. 2013). A small proportion of comparisons, d_N and d_S were equal between branches. In 6625 of the cases, d_N/d_S was higher in *A. thaliana*, and in 6161 of the cases d_N/d_S was higher in *A. lyrata*, indicating that purifying selection on protein sequences is often less effective in *A. thaliana* (binomial test, $P = 4.22 \times 10^{-5}$). Differences in d_N/d_S were more pronounced when analyses were restricted to genes that are highly expressed in *A. thaliana* (Table S2.2).

For each alignment, Tajima's relative rate test (Tajima 1993) was used to contrast whether the number of substitutions accumulated in *A. thaliana* and *A. lyrata* was significantly different. Statistically significant differences were detected in 1363 and 1333 genes for synonymous and nonsynonymous sites, respectively. Of the 1363 genes with significant differences in synonymous rates of evolution, there were more unique

synonymous changes in *A. thaliana* in 1222 genes compared to 141 genes where *A. lyrata* had more unique synonymous changes. Of the 1333 genes with an asymmetry in the number of nonsynonymous sites, *A. thaliana* and *A. lyrata* had more unique changes in 1077 and 256 cases, respectively (Table 2.1).

For each of the 12,994 alignments, we compared the likelihood of the free-ratios model (in which each of the three branches exhibits an independent d_N/d_S ratio) vs. that of a 2-ratios model (one d_N/d_S ratio for *A. thaliana* and *A. lyrata*, and another for *C. rubella*). The free-ratios model fit the data significantly better in 907 of the alignments (likelihood ratio test, $P < 0.05$), indicating that the d_N/d_S ratio is significantly different in *A. thaliana* and *A. lyrata*. In 477 of the 907 cases where the free-ratio model fit better than the two-ratio model, d_N/d_S was higher for *A. thaliana*, and in 430 of the cases d_N/d_S was higher for *A. lyrata;* these numbers were not significantly different from the 50%:50% (453.5:453.5) expected by chance (binomial test, $P = 0.166$).

Given that *A. thaliana* and *A. lyrata* are very closely related, some gene alignments may not contain sufficient information (in terms of number of substitutions) to accurately infer the strength of purifying selection acting on each branch. In order to increase the power of our analyses, we combined all 12,994 alignments into a single concatenome containing 17.8 million base pairs and repeated our analyses on it. The *A. thaliana* lineage exhibited higher d_N, d_S and d_N/d_S values (0.0127, 0.0759 and 0.1671, respectively) (Table S2.3) than the *A. lyrata* branch (0.0102, 0.0622, 0.1644). These values are comparable to the mean values resulting from analysis of individual alignments. The free-ratios model fit the data significantly better than the 2-ratios model ($2\Delta\ell = -10.213$, $P = 0.0014$), showing that d_N/d_S is significantly higher in *A. thaliana*, even though the differences are small. Tajima's relative rate test (Tajima 1993) revealed an excess of synonymous and nonsynonymous changes in *A. thaliana* compared to *A. lyrata* ($\chi^2 = 4369.4$ and 2207.0, P

<< 0.001 and P << 0.001, respectively). The *A. thaliana* concatenome contained 249,860 synonymous 163,355 nonsynonymous substitutions that were not present in *A. lyrata*, and the *A. lyrata* concatenome contained 205,266 unique synonymous substitutions and 137,578 unique nonsynonymous substitutions.

It is expected that the evolution of selfing in *A. thaliana* may have resulted in pseudogenization of, or at least relaxation of purifying selection in, genes involved in out-crossing. If these represent a sufficiently large number of genes, this effect alone, rather than a reduction of N_e, might conceivably explain the higher average rates of protein evolution observed in *A. thaliana*. To discard this possibility, we repeated our analyses separately for genes of different functional categories according to KOG categories for eukaryotes. For all 23 KOG categories represented in the dataset, the number of genes with higher d_N and d_S values in *A. thaliana* was significantly higher than the number of genes with higher d_N and d_S values in *A. lyrata*. For 19 of the categories, the number of genes for which d_N/d_S was higher in *A. thaliana* was higher than the number of genes for which d_N/d_S was higher in *A. lyrata*. For only 3 categories there were more genes with a higher d_N/d_S in *A. lyrata* (binomial test, $P = 0.0009$; Table 2.2). These results indicate that the higher d_N, d_S and d_N/d_S values observed in *A. thaliana* represent a generalized trend, not specific to certain functional categories.

Throughout the current work we have reported the comparison of the *A. thaliana* reference genome from the TAIR 10 release, a composite genome from 11 Columbia ecotype (Col-0) individuals, with that of *A. lyrata*, using *C. rubella* as outgroup. Nonetheless, equivalent results were obtained using another 18 *A. thaliana* accessions instead of the reference one (Tables S2.4, S2.5), using *A. halleri* (Briskine et al. 2016) instead of *A. lyrata* (Tables S2.6, S2.7, and S2.8) or using the outcrossing and more

distantly related *Thellungiella parvula* (Dassanayake et al. 2011) as outgroup instead of the selfing and closely related *C. rubella* (Tables S2.9, S2.10).

Discussion

In summary, all our genome-wide analyses converge to show that, as expected from the reduced N_e due to selfing, proteins evolved faster in *A. thaliana* than in *A. lyrata* or *A. halleri*. Such protein sequence evolution acceleration is likely due to the combination of faster mutation rates in *A. thaliana* (supported by high d_S values and by prior results; (Yang et al. 2013) and by a weaker efficacy of natural selection on nonsynonymous mutations (supported by high d_N/d_S ratios). Prior analyses based on a handful of orthologous genes failed to detect differences in d_N and d_N/d_S between *A. thaliana* and *A. lyrata*, most likely because of limited statistical power (Wright et al. 2002; Foxe et al. 2008). Indeed, the differences that we detected are subtle, consistent with the fact that *A. thaliana* has been selfing for a relatively short amount of time (150,000-1,000,000 years; (Charlesworth and Vekemans 2005; Bechsgaard et al. 2006); Tang et al. 2007, Durvasula et al 2017) compared to the time of divergence between *A. thaliana* and the *A. lyrata*/*A. halleri* clade (7-13 MY; (Beilstein et al. 2010); Hohmann et al. 2015). Our analyses have compared the patterns of evolution of the *A. thaliana* lineage (the branch connecting *A. thaliana* and the most recent common ancestor of *A. thaliana* and *A. lyrata*) and the *A. lyrata* and *A. halleri* lineages (the branches connecting *A. lyrata* or *A. halleri* and the most recent common ancestor of *A. thaliana* and *A. lyrata*), and plants in the *A. thaliana* lineage have been selfing for only 1-17% of the length of the branch.

In addition to the recent transition to selfing of *A. thaliana*, other scenarios may account for the small magnitude of differences observed between the rates of protein evolution of *A. thaliana* and *A. lyrata*. First, most proteins are under strong purifying selection in both species, in agreement with prior observations (Wright et al. 2002; Foxe

et al. 2008. Yang and Gaut 2011), thus hindering the detection of strong differences. Second, selfing increases homozygosity, thus exposing recessive alleles to selection, which can reduce rates of protein evolution (see Glémin 2007). Last, population genetics analyses indicate that the N_e of *A. lyrata* may have also been reduced within the last 100,000 years (Mattila et al. 2017); this might have increased the rates of protein evolution in this species, thus attenuating the differences between *A. thaliana* and *A. lyrata*.

Finally, it should be noted that the fast rates of protein evolution observed in *A. thaliana* might be due to peculiarities of the biology of this species other than selfing. In particular, *A. thaliana* switched to an annual life history, whereas *A. lyrata* is perennial. Annual plants tend to evolve faster than perennial plants (Smith and Donoghue 2008; Lanfear et al. 2013; Gaut et al. 2011), which might account for the higher rates of synonymous evolution observed in *A. thaliana*. However, annual plants exhibit lower nonsynonymous to synonymous polymorphism ratios (Chen et al. 2017), and thus the annual life history of *A. thaliana* may not explain the observed d_N/d_S ratios observed in this species.

Methods

For each *C. rubella* gene, the longest encoded protein was chosen for analysis and orthologs in *A. thaliana* and *A. lyrata* were identified using a best reciprocal hit approach (BLASTP, *E*-value < 10^{-10}). Only genes for which orthologs could be identified in both *Arabidopsis* species were retained. Trios of orthologous protein sequences were aligned using PRANK v.140603 (Löytynoja and Goldman 2005), and the resulting alignments were used to guide the alignments of the CDSs using an in-house script. Alignments which contained less than 5% gaps were retained for analyses. For each alignment, the codeml program of PAML v. 4.9 (Yang 2007) was used to estimate d_N, d_S and d_N/d_S in each of the three branches (free-ratios model) and in the *A. thaliana*/*A. lyrata*

branch and the *C. rubella* branch separately (2-ratios model). Values of d_N/d_S above 10 were removed from mean calculations, in order to prevent the bias introduced by these outliers, which represent artifacts due to the presence of very few mutations in the relevant lineages. The fit of both nested models was compared using a likelihood ratio test, assuming that twice the difference between the log-likelihoods of both models ($2\Delta\ell$) follow a χ^2 distribution with one degree of freedom (Huelsenbeck and Crandall 1997). Tajima's relative rate tests (Tajima 1993) were conducted using in-house scripts. *A. thaliana* genes were classified into different eukaryotic orthologous groups (KOG) categories using the eggNOG database v4.5.1 (Huerta-Cepas et al. 2015). Data for the 18 accessions of *A. thaliana* were obtained from the 1000 genomes project (Gan et al. 2011). *A. thaliana* gene expression data were obtained from Schmid et al. (Schmid et al. 2005) and processed as in Alvarez-Ponce and Fares (Alvarez-Ponce and Fares 2012). All our alignments and scripts are available upon request.

Acknowledgements

We are grateful to Julio Rozas for helpful discussion. This work was supported by funds from the University of Nevada, Reno.

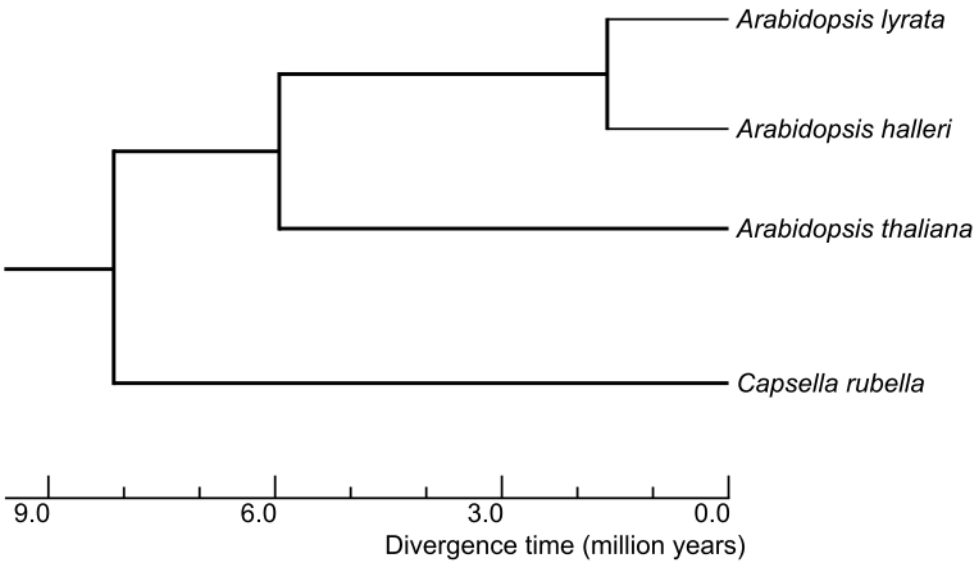

Figure 2.1. Phylogenetic relationships among the species used in the current study. The tree topology and divergence times were obtained from Hohmann et al. (Hohmann et al. 2015).

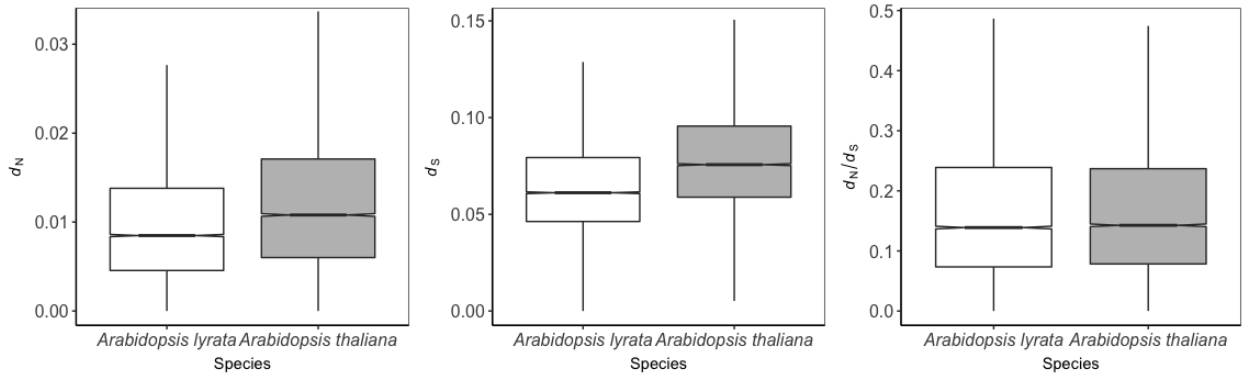

Figure 2.2. Distribution of dN, dS and dN/dS values in the A. thaliana and A. lyrata branches. Values above the 90th percentile are not represented.

Table 2.1. Tajima's relative rate tests

	All substitutions	Synonymous substitutions	Nonsynonymous substitutions
Unique substitutions in *A. thaliana*	413,215	249,860	163,355
Unique substitutions in *A. lyrata*	343,062	205,266	137,578
Genes where *A. thaliana* had more substitutions	9203	8701	7575
Genes where *A. lyrata* had more substitutions	3207	3424	4102
Genes where $P < 0.05$	2008	1363	1333
Genes where $P < 0.05$ and *A. thaliana* had more substitutions	1824	1222	1077
Genes where $P < 0.05$ and *A. lyrata* had more substitutions	184	141	256
χ^2 value for concatenome	6507.5	4369.4	2208.0
P-value for concatenome	<< 0.001***	<< 0.001***	<< 0.001***

***, $P < 0.001$.

Table 2.2. Analyses of evolutionary rates in different KOG categories

Category[a]	Genes with higher d_N in *A. thaliana*	Genes with higher d_N in *A. lyrata*	Genes with higher d_S in *A. thaliana*	Genes with higher d_S in *A. lyrata*	Genes with higher d_N/d_S in in *A. thaliana*	Genes with higher d_N/d_S in in *A. lyrata*	d_N *P*-value[b]	d_S *P*-value[b]	d_N/d_S *P*-value[b]
A	245	99	248	97	175	163	2.03×10^{-15} ***	2.2×10^{-16} ***	0.550
B	77	32	78	32	62	45	1.94×10^{-5} ***	1.36×10^{-5} ***	0.122
C	256	138	257	137	209	174	2.89×10^{-9} ***	1.53×10^{-9} ***	0.082
D	121	50	124	48	85	84	5.64×10^{-8} ***	6.13×10^{-9} ***	1.000
E	189	113	216	86	152	150	1.44×10^{-5} ***	4.63×10^{-14} ***	0.954
F	60	32	60	32	53	36	0.005 **	0.0046 **	0.089
G	500	308	566	245	389	422	1.47×10^{-11} ***	$< 2.2 \times 10^{-16}$ ***	0.261
H	116	74	130	60	94	94	0.003 **	4.16×10^{-7} ***	1.000
I	225	125	243	107	184	166	9.99×10^{-8} ***	2.69×10^{-13} ***	0.364
J	214	114	201	127	178	135	3.63×10^{-8} ***	5.20×10^{-5} ***	0.017 *
K	697	362	720	344	541	511	$< 2.2 \times 10^{-16}$ ***	$< 2.2 \times 10^{-16}$ ***	0.371
L	166	68	161	73	119	113	1.23×10^{-10} ***	8.70×10^{-9} ***	0.743
M	104	67	121	50	78	92	0.006 **	5.64×10^{-8} ***	0.319
O	710	317	743	287	507	494	$< 2.2 \times 10{-}16$ ***	$< 2.2 \times 10^{-16}$ ***	0.704
P	340	167	367	140	262	243	1.22×10^{-14} ***	$< 2.2 \times 10^{-16}$ ***	0.423
Q	227	115	241	101	189	153	1.40×10^{-9} ***	2.45×10^{-14} ***	0.058
S	2344	1213	2405	1156	1784	1751	$< 2.2 \times 10^{-16}$ ***	$< 2.2 \times 10^{-16}$ ***	0.590
T	673	335	736	272	517	479	$< 2.2 \times 10^{-16}$ ***	$< 2.2 \times 10^{-16}$ ***	0.241
U	353	182	389	150	269	241	1.24×10^{-13} ***	$< 2.2 \times 10^{-16}$ ***	0.232
V	52	28	54	26	44	35	0.001 **	0.002 **	0.368
W	48	26	51	23	34	40	0.014 *	0.001 **	0.561
Y	22	3	19	6	17	8	1.57×10^{-4} ***	0.015 *	0.108
Z	159	58	155	63	119	92	4.64×10^{-12} ***	3.94×10^{-10} ***	0.073

[a]Category functions: A, RNA processing and modifications; B, chromatin structure and dynamics; C, energy production and conversion; D, cell cycle control, cell division, chromosome partitioning; E, amino acid transport and metabolism; F, nucleotide transport and metabolism; G, carbohydrate transport and metabolism; H, coenzyme transport and metabolism; I, lipid transport and metabolism; J, translation, ribosomal structure and biogenesis; K, transcription; L, replication, recombination, and repair; M, cell wall, cell membrane and envelope biogenesis; O, posttranslational modification; P, inorganic ion transport and metabolism; Q, secondary metabolite biosynthesis; S, function unknown; T, signal transduction; U, intracellular trafficking, secretion, and vesicular transport; V, defense mechanisms; W, extracellular structures; Y, nuclear structure; Z, cytoskeleton (Tatusov et al. 2003).

[b]*P*-values determined using a binomial test comparing the total number of genes where d_N/d_S was higher in *A. thaliana* and in *A. lyrata*. *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$.

Chapter 2: Codon usage differences among genes expressed in different tissues of *Drosophila melanogaster*

Abstract

Patterns of codon usage are affected by both mutational biases and translational selection. The frequency at which each codon is used in the genome is directly linked to the cellular concentrations of their corresponding tRNAs, which favors optimal translation. Codon usage patterns are known to vary among organisms, due to differences in mutational biases and relative tRNA abundances. Given the potential that tRNA abundances vary across different tissues, it is possible that genes expressed in different tissues are subject to different translational selection regimes, and thus differ in their patterns of codon usage. These differences, however, are poorly understood, having been studied only in Arabidopsis and in human. *Drosophila melanogaster* is an ideal model organism to study tissue-specific codon adaptation, given its large effective population size and lack of isochores. Here, we examine 2046 genes, each expressed specifically in one tissue of *D. melanogaster*. We show that genes expressed in different tissues exhibit significant differences in their patterns of codon usage, and that these differences are only partially due to differences in GC content, expression levels or protein lengths. Interestingly, these differences are stronger when the analysis is restricted to highly expressed genes. Our results strongly suggest that genes expressed in different tissues are subject to different regimes of translational selection, probably owing to tissues differing in their relative tRNA abundances.

Introduction

Groups (or families) of synonymous codons encode the same amino acid, but are used at largely different frequencies in any genome, a phenomenon known as codon usage bias. Codon bias is affected by both genome nucleotide composition (mutational biases) and translational selection (Sharp et al. 1993a). The frequency at which each codon is used by a given genome positively correlates with the cellular concentrations of the corresponding tRNAs, and genes expressed at high levels tend to exhibit increased frequencies of preferred codons (Ikemura 1981, 1982). High tRNA abundances for these codons result in faster and more accurate translation, which makes these codons preferred by natural selection (Ikemura 1982; Andersson and Kurland 1990; Dong et al. 1996; Rocha 2004). The patterns of codon usage vary among organisms (Kanaya et al. 2001; Duret 2002; Basak and Ghosh 2006; Vicario et al. 2007; Hassan et al. 2009; Du et al. 2014), as expected from the fact that different organisms exhibit different relative tRNA abundances and nucleotide compositions (Muto and Osawa 1987; Kanaya et al. 2001; Rocha 2004; Goodenbour and Pan 2006). Transfer RNA abundances can also differ among the different tissues of an organism (Dittmar et al. 2006), raising the possibility that different patterns of codon usage may be selected in different tissues. However, very few studies, restricted to human and *Arabidopsis*, have explored this possibility, producing controversial results.

Using a limited dataset (n < 200 genes), Plotkin *et al.* (Plotkin et al. 2004) found significant differences among genes expressed in six human tissues, which they attributed to genes being adapted to the tRNA pools of the tissue in which they are expressed. In contrast, using internal correspondence analysis and a larger dataset (n = 2126), Sémon *et al.* (Semon et al. 2006) found that the fraction of the variability of codon usage attributed to tissue specificity was very small (~2.3%), and mostly due to differences in the GC

content of genes expressed in the different tissues. However, Dittmar *et al.* (Dittmar et al. 2006) observed significant differences in the relative abundances of tRNAs among different human tissues, with preferred codons usually corresponding to the most abundant tRNAs, in line with the results and interpretation of Plotkin *et al.* (Plotkin et al. 2004). Finally, Camiolo *et al.* (Camiolo et al. 2012) found that genes expressed in different tissues of *Arabidopsis thaliana* significantly differed in their patterns of codon usage, even after controlling for differences in GC content and expression levels.

The relative importance of translational selection versus nucleotide composition in shaping codon usage is expected to depend on the effective population size (N_e). In organisms with large N_e, natural selection is more effective at driving slightly advantageous mutations to fixation and at removing slightly deleterious mutations, such as synonymous mutations (Kimura et al. 1963; Kimura 1968; Kimura 1983). N_e has been estimated to be significantly higher for *D. melanogaster* (1,000,000–5,000,000 individuals (Wagner 2005; Shapiro et al. 2007), than for *A. thaliana* (250,000–400,000 individuals (Moore and Purugganan 2003; Cao et al. 2011) or humans (~10,000 individuals (Yu et al. 2004). This, together with the fact that *D. melanogaster* is the best characterized multicellular organism in terms of codon bias (Vicario et al. 2007), and the absence of isochores, stretches of uniform GC content in the genome, in this organism (Nekrutenko and Li 2000; Oliver et al. 2001), makes it ideal to characterize the differences in codon usage among tissues.

Here, we describe significant differences in the patterns of codon usage of genes expressed in 16 *D. melanogaster* adult tissues. Multivariate analyses indicate that the differences are small but significant and only partially due to differences in GC content. The differences were stronger when analyses were restricted to highly expressed genes. Our results indicate different patterns of translational selection among genes expressed

in different tissues of *Drosophila*, potentially due to adaptation to different tRNA abundances.

Materials and Methods

Genomic data

We downloaded all *D. melanogaster* coding sequences (CDSs) from Ensembl BioMart, version 83 (Flicek et al. 2013; Kersey et al. 2016). If a gene had multiple CDSs, then the longest one was chosen for analysis. After filtering, we retained 13,905 *D. melanogaster* CDSs.

Gene expression data

For each *D. melanogaster* protein-coding gene, the mRNA abundances in the whole adult body and in 16 adult nonredundant tissues/organs (adult carcass, brain, crop, eyes, fat body, head, heart, hindgut, male accessory glands, midgut, ovaries, salivary glands, testes, thoracoabdominal ganglia, tubules, and virgin spermatheca) were obtained from the FlyAtlas database (Chintapalli et al. 2007). Probes were mapped to genes using the Affymetrix annotation file “Drosophila 2”, version 35. We discarded from our analysis those probes that matched multiple genes. If a gene mapped to multiple probes, we used the probe with the highest mRNA signal in the whole fly. After filtering, a total of 13,088 *D. melanogaster* genes with available mRNA abundance data were retained for our study. Messenger RNA abundances were averaged across 4 biological replicates.

We used this gene expression data to obtain a list of tissue-specific genes. A gene was considered to be expressed in a certain tissue/organ if it was detectable in at least 3 out of the 4 biological replicates (as in ref. (Chakraborty and Alvarez-Ponce 2016)). Genes expressed only in one out of the 16 tissues/organs were considered as tissue-specific genes. Using these criteria, we identified a total of 2,046 *D. melanogaster* tissue-specific genes.

Data analysis

We processed our data using several in-house PERL scripts. Data analysis, including generation of plots and statistical tests, were conducted using R (R-Core-Team 2013). Codon frequencies and relative synonymous codon usage (RSCU) values for each gene were calculated using the "Bio::Tools::CodonOptTable" module of the BioPerl package. We used the seqinr (Charif et al. 2017) and ade4 (Dray et al. 2016) packages to perform correspondence analysis in R. Additionally, we used the pipeline of Sémon *et. al.* (Semon et al. 2006) to perform the internal correspondence analysis (Cazes et al. 1988). We also used the vegan package (https://cran.r-project.org/web/packages/vegan/) to perform PERMANOVA and PERMANCOVA analyses in R. Expression levels were log-transformed for our correspondence and PERMANCOVA analyses to improve normality. Protein lengths were log-transformed for our PERMANCOVA analyses.

Results

Patterns of codon usage in D. melanogaster

We first conducted a codon usage analysis based on 13,088 *D. melanogaster* nucleus-encoded protein-coding genes whose expression level is available in the FlyAtlas database (Chintapalli et al. 2007). We first counted how many times each codon is used. The most frequent codon within each synonymous family were: GCC (Ala), CGC (Arg), AAC (Asn), GAU (Asp), UGC (Cys), CAG (Gln), GAG (Glu), GGC (Gly), CAC (His), AUC (Ile), CUG (Leu), AAG (Lys), UUC (Phe), CCC (Pro), AGC (Ser), ACC (Thr), UAC (Tyr), GUG (Val), and UAA (Stop). AUG and UGG are the only codons coding for Met and Trp, respectively (Table S3.1).

The most frequently used codons are not necessarily the preferred ones (favored by natural selection). To identify the preferred codon in each of the 18 multi-codon synonymous families, we compared the patterns of codon usage of highly and lowly

expressed genes. First, we identified the most highly expressed (10% top expression), and the least expressed (10% bottom expression). Second, we compared the relative synonymous codon usage (RSCU) of each codon among highly and lowly expressed genes. We considered a codon as preferred if its RSCU value was significantly higher in the highly expressed gene set (Mann-Whitney *U* test) after controlling for the false discovery rate associated with multiple testing using the Benjamini and Hochberg approach (Benjamini and Hochberg 1995) ($q < 0.05$). We identified a total of 22 preferred codons (excluding the three termination codons, the one coding for Met, and the one coding for Trp): UUC (Phe), CUG (Leu), AUC (Ile), GUC and GUG (Val), UAC (Tyr), CAC (His), CAG (Gln), AAC (Asn), AAG (Lys), GAC (Asp), GAG (Glu), UCC and UCG (Ser), CCC (Pro), Thr (ACC), GCC (Ala), UGC (Cys), CGU and CGC (Arg), and GGU and GGC (Gly) (Table 1). Of note, most of these codons end in G or C, with the exception of GGU and CGU.

Codon usage differences among genes expressed in different tissues of D. melanogaster

For each of the 13,088 *D. melanogaster* protein-coding genes, we retrieved their levels of expression (mRNA abundances) in the whole adult body, and in 16 individual adult tissues, from the FlyAtlas database (Chintapalli et al. 2007). This information was used to identify a total of 2046 genes that are expressed in only one tissue (19 in the adult carcass, 77 in the brain, 22 in the crop, 44 in the eyes, 23 in the fat body, 47 in the head, 15 in the heart, 30 in the hindgut, 116 in the male accessory glands, 133 in the midgut, 84 in the ovaries, 10 in the salivary glands, 1364 in the testes, 10 in the thoracoabdominal ganglia, 28 in the tubules, and 24 in the virgin spermatheca).

For each of these 16 gene sets, we computed the frequencies at which the different codons were used. In the majority of cases, the most frequent codon was the same as

that for the entire gene set. However, a number of differences existed. In the hindgut, male accessory glands, testes, thoracoabdominal ganglia and virgin spermatheca, AAU is the most frequently used codon to code for Asn, instead of AAC (the most commonly used codon genome-wide to code for Asn). Similarly, the most frequent codon for Cys is UGC, except for genes expressed in the salivary glands, which tend to use UGU. Glu is often encoded by GAG, except for genes expressed in the male accessory glands and the virgin spermatheca, which tend to use GAA. In general, Gly is most frequently encoded by GGC; however, genes expressed in the carcass, head, male accessory glands, salivary glands, and virgin spermatheca tend to use GGA. Genes expressed in all tissues prefer CAC to encode His, except those expressed in the male accessory glands, which use CAU more often. Ile is often encoded by AUC, except among genes expressed in the male accessory glands, which tend to use AUU. Phe is generally encoded by UUC, but genes expressed in the male accessory glands use more frequently UUU. The most commonly used codon to encode Pro is CCC, but genes expressed in the crop, male accessory glands, salivary glands and virgin spermatheca prefer CCA, and those expressed in the brain prefer CCG. The most used codon to encode Ser is AGC; however, genes expressed in the carcass, crop, head, hindgut, midgut, salivary glands, testes, and tubules prefer UCC, and genes expressed in the virgin spermatheca prefer AGU. Finally, Tyr is generally encoded by UAC, but genes expressed in the salivary glands use more frequently UAU (Table S3.1).

Most of these differences represent significant departures from the frequencies at which codons are used in the entire genome (χ^2 test, $P < 0.05$; Table S2). For each tissue (16 tissues) and for each family of synonymous codons with more than one codon (18 codons after excluding those encoding Met and Trp) (i.e., 16 tissues × 18 amino acids = 288 contrasts), we used a χ^2 test to compare the frequencies at which the different codons are used in that tissue with the frequencies at which the codons are used in the overall

genome. For instance, the *D. melanogaster* proteome contains a total of 338,998 asparagines, of which 156,904 are encoded by AAU and 182,094 are encoded by AAC; the male accessory glands proteome contains a total of 2376 asparagines, of which 1392 are encoded by AAU and 984 are encoded by AAC; the frequencies at which codons are used are significantly different in both gene sets (χ^2 = 145.14, 1 degree of freedom, P = 1.88×10^{-33}). Out of the 288 contrasts, 168 were significant (χ^2 test, $P < 0.05$; Table S3.2). Genes expressed in the male accessory glands were the ones with the highest number of significant differences: contrasts were significant for all 18 amino acids (Table S3.2). Among genes expressed in other tissues, significant differences were observed in the heart (in 3 amino acids), crop (in 5 amino acids), fatbody (in 5 amino acids), thoracicoabdominal ganglia (in 6 amino acids), hindgut (in 7 amino acids), adult carcass (in 9 amino acids), head (in 9 amino acids), tubules (in 10 amino acids), ovaries (in 11 amino acids), salivary glands (in 11 amino acids), brain (in 13 amino acids), eyes (in 14 amino acids), midgut (in 14 amino acids), virgin spermatheca (in 15 amino acids), male accessory glands (in 17 amino acids), and testes (in 18 amino acids). The number of amino acids with significant differences positively correlates with the number of genes expressed in each tissue (Spearman ρ = 0.689, P = 0.003), suggesting that our contrasts are to some extent limited by statistical power.

For the three tissues in which a larger number of genes are expressed (testes, midgut and male accessory glands) we determined the set of preferred codons by comparing the most highly expressed (top 20%) and the least expressed (bottom 20%) genes. A total of 14 preferred codons were identified among genes expressed in the midgut (Table S3.3); all of these codons were previously identified as preferred codons in our analyses of the entire *D. melanogaster* genome (Table 3.1). Four codons (CAG, AAG, CCC, and CGU) were identified as preferred among genes expressed in testes (Table

S3.4); again, all of these codons were previously identified as preferred codons in our analyses of the entire genome (Table 3.1). Our analysis of genes expressed in male accessory glands did not identify any preferred codon (Table S3.5).

Codon usage is strongly correlated with GC content at the third codon positions (GC3) (Sueoka and Kawanishi 2000; Wan et al. 2004) and GC3 content varies among genes expressed in different tissues (ranging from 51.03% in the salivary glands to 66.48% in the eyes; Kruskal-Wallis test, $P = 3.13\times10^{-23}$; Table 3.2). Together, these observations raise the possibility that the observed differences in codon usage among genes expressed in different tissues may be due to differences in GC content. In order to discard this possibility, for each tissue, we generated a list of genes with a distribution of GC3 virtually identical to that of the genes expressed in the tissue. For that purpose, for each of the genes expressed in the tissue of interest, we randomly selected a gene not expressed specifically in the tissue with a very similar GC3 content (± 1%). Two lines of evidence indicate that our observations are not explained (at least entirely) by GC content. First, many of the tissue-specific deviations from the codon preferences of the entire genome (i.e., many of the cases in which one codon is preferred in general, but another codon is preferred among genes expressed in a certain tissue) are not observed in the randomized dataset (Table S3.6). For instance, as mentioned above, in the original dataset AAC is the most commonly used codon to encode Asn, except for genes expressed in the hindgut, male accessory glands, testes, thoracoabdominal ganglia, and virgin spermatheca, in which AAU is preferred (Table S3.1). In the randomized dataset, AAC is the most commonly used codon, except for the gene sets matching the GC3 content of genes expressed in the heart, male accessory glands, and virgin spermatheca (Table S3.6). Second, in 210 out of the 288 cases the frequencies at which codons are used in each tissue significantly differ (χ^2 test, $P < 0.05$) from the frequencies at which

codons are used in the randomized datasets corresponding to the same tissue (Table S3.7); we would not expect this to be the case if codon preference differences were only dictated by GC3.

Codon usage is known to be highly affected by gene expression and by protein length (Duret and Mouchiroud 1999; Powell and Dion 2015), and genes expressed in different tissues differ in their levels of expression (Kruskal-Wallis test, $P = 2.93\times10^{-83}$) and in the length of their encoded products (Kruskal-Wallis test, $P = 1.56\times10^{-18}$; Table 3.2). Therefore, we repeated our analyses using these variables for randomized datasets (instead of GC3) as controlling variables. Similar results were obtained, indicating that our observations are not due to expression levels or protein lengths either (Tables S3.8 and S3.9).

Correspondence analysis and internal correspondence analysis

We used correspondence analysis (Grantham et al. 1980) to visualize codon usage differences among genes expressed in the different tissues. Correspondence analysis is a multivariate analysis method that summarizes the information from a high-dimensional space into a low-dimensional space while losing as little information as possible (Lobry and Chessel 2003). In our case, we only considered the two main axes and plotted the centroids of the cluster for each tissue in Figure 3.1. Consistent with the analyses described in the previous section, we found that genes expressed in different tissues exhibit different codon usage patterns (Figure 3.1).

This analysis, however, does not allow us to distinguish between the differences in codon usage due to different amino acid usage (proteins expressed in different tissues tend to use different amino acids) or to differences in the usage of synonymous codons (different codons being preferred to encode a certain amino acid) (Semon et al. 2006). Therefore, we next used internal correspondence analysis (Cazes et al. 1988; Semon et

al. 2006). This technique is basically a double within-between-correspondence analysis, which allows us to partition the variance of codon usage into different components (Lobry and Chessel 2003). We used the pipeline of Sémon et al (Semon et al. 2006) to partition the codon usage variability into four components: within tissues within amino acids, within tissues between amino acids, between tissues within amino acids, and between tissues between amino acids. Interestingly, we found that 51.8% of the total variability in codon usage is due to variability in synonymous codon usage (Figure 3.2g), but only 2.2% of the variation in synonymous codon usage is due to tissue specificity. To assess the statistical significance of this value (2.2%), we generated 1000 randomized datasets and performed internal correspondence analysis in each of them. Each randomized dataset was generated by randomly assigning each gene to one of the 16 studied tissues, keeping the same number of genes in each tissue as in the original dataset. All of the 1000 randomized datasets exhibited a lower value compared to the observed one, indicating that the observed value is higher than expected by chance (expected $value_{median}$ = 0.6%, expected $value_{mean}$ = 0.75%; $P < 0.001$; Figure 3.3).

We repeated this analysis by controlling for GC3. For that purpose, each of the randomized datasets was generated by selecting, for each of the genes in our dataset, a gene with a very similar GC3 (± 2%) not expressed specifically in the same tissue. Similar results were obtained (expected $value_{median}$ = 1.5%, expected $value_{mean}$ = 1.58%; P = 0.037; Fig. 3.3). These results indicate that the variation of tissue-specific codon usage is small but significant, and not due to GC content. Also similar results were obtained when using expression level (expected $value_{median}$ = 0.7%, expected $value_{mean}$ = 1.13%; $P <$ 0.001; Figure 3.3) or protein length (expected $value_{median}$ = 0.7%, expected $value_{mean}$ = 0.69%; $P < 0.001$; Figure 3.3) as controlling variables, indicating that these factors are not the sole cause of the observed differences among tissues either.

We also repeated the internal correspondence analysis restricting our analyses to highly expressed genes. Highly expressed genes are expected to be subject to strong translational selection and thus are expected to exhibit stronger differences if these are due to translational selection. We repeated our analyses on genes whose log-expression levels in their tissue of expression were 25% (*N* = 1298), 50% (*N* = 901), 75% (*N* = 385), and 100% (*N* = 105) over the average expression level. Interestingly, we observed that the variation in synonymous codon usage between the genes expressed in different tissues increases as we increase the expression cut-off (Table 3.3), suggesting that the observed trend is due to translational selection.

The number of genes expressed in certain tissues is very small (Table S3.1). In order to discard the possibility that this might be inflating the observed differences among tissues, we repeated our analyses after removing all tissues in which less than 30 genes were expressed. In this case, the fraction of the variability explained by tissue and not by amino acid differences was 2.1%, i.e. similar to the fraction estimated from the entire dataset.

Permutational multivariate analysis of variance

For further investigation of codon usage differences among genes expressed in different tissues, we used permutational multivariate analysis of variance (PERMANOVA) (Anderson 2001), a permutation-based extension of multivariate analysis of variance. In order to reduce the intrinsic correlation among the RSCU values corresponding to any set of synonymous codons, we generated 1000 randomized versions of our dataset. In each randomization, one codon per amino acid was randomly chosen, and all its RSCU values were removed (18 columns in total). All randomized datasets were analyzed using PERMANOVA (with 999 permutations). In all 1000 cases, the effect of tissue on codon usage was statistically significant ($P < 0.05$; average pseudo-F ratio = 3.94; Table 3.5).

Given the possibility that the observed results may be affected by the strong variation in the number of genes expressed in the different tissues (Shaw and Thomas 1993) we repeated our analyses on a second set of 1000 randomized versions of our dataset, each obtained using a double randomization technique: first, one codon per amino acid was removed (as above), and second, 10 genes from each tissue were selected for analysis. In this case, we observed a significant association between codon usage and tissue of expression in 891 of the randomized datasets (i.e., 89.1% of cases, average pseudo-*F* ratio = 1.42; Table 3.5).

In order to discard the possibility that the observed results might be a by-product of co-variation of both codon usage and tissue specificity with GC3, gene expression or protein length, we repeated our analyses using PERMANCOVA, a non-parametric version of ANCOVA, using the three confounding variables as covariates. Using internal single randomization, the tissue of expression had a significant effect on codon usage in all 1000 randomized datasets (average pseudo *F*-ratio = 7.05). Using internal double randomization, tissue had a significant effect in 1000 of the randomized datasets (100%; average pseudo *F*-ratio = 1.97). These results strongly indicate that the effect of tissue of expression on codon usage is not due to GC3, expression level or protein length. Indeed, the effect of tissue was stronger after controlling for these factors (compare Tables 3.5 and 3.6).

Discussion

We analyzed the patterns of codon usage of 2046 genes, each expressed exclusively in one *D. melanogaster* tissue/organ. We observed significant differences among the genes expressed in the different tissues. Our multivariate analyses showed that codon usage differences are not due to differences in GC content, expression level or protein length. This strongly suggests that the observed differences codon usage

reflect, at least in part, adaptation to different relative tRNA concentrations in different tissues.

Unfortunately, tRNA abundance data for the different tissues of *D. melanogaster* are not available at the moment. Therefore, our expectation that the codons preferred in each tissue are the ones with more abundant tRNAs in that tissue cannot be tested directly. Methodologies to directly sequence tRNA are under development (Smith et al. 2015), thus it may eventually be possible to directly test our expectation. Consistent with our hypothesis that our observations are due at least in part to translational selection, the differences in codon usage among genes expressed in different tissues are more pronounced among highly expressed genes (which are expected to be subject to stronger translational selection).

Plotkin *et al.* (Plotkin et al. 2004) and Camiolo *et al.* (Camiolo et al. 2012) also found differences in the patterns of codon usage of genes expressed in different tissues of human and *A. thaliana*, respectively. In *A. thaliana*, the differences were shown to be independent of GC content and expression level. In contrast, in human the differences appear to be largely due to differences in GC content rather than to translational selection (Semon et al. 2006).

At least two factors may account for the differences observed between human and *D. melanogaster* and *A. thaliana*. First, humans have a much lower N_e than the other two species (humans: ~10,000 individuals; *D. melanogaster*: 1,000,000–5,000,000 individuals; *A. thaliana*: 250,000–400,000 individuals (Moore and Purugganan 2003; Yu et al. 2004; Wagner 2005; Shapiro et al. 2007; Cao et al. 2011), which is expected to reduce the efficacy of translational selection (Kimura 1983; Charlesworth 2009). Second, mammalian genomes exhibit a strong isochoric structure which produces strong regional variation in GC content (Bernardi 1989, 2000), and genes co-expressed in specific tissues

tend to cluster next to each other in the genome (Lercher et al. 2002), making them likely to exhibit similar GC contents. The *D. melanogaster* and *A. thaliana* genomes, however, do not exhibit an isochoric structure (Thiery et al. 1976; Jabbari and Bernardi 2000; Nekrutenko and Li 2000; Oliver et al. 2001).

Our observations have implications for heterologous gene expression. If a transgene is to be expressed in a specific fly tissue, using the optimal patterns of codon usage specific for that tissue will likely result in optimal protein translation. In addition, it is expected that the patterns of codon usage of viruses resemble the patterns of codon usage of the genes expressed in the tissues infected by the virus.

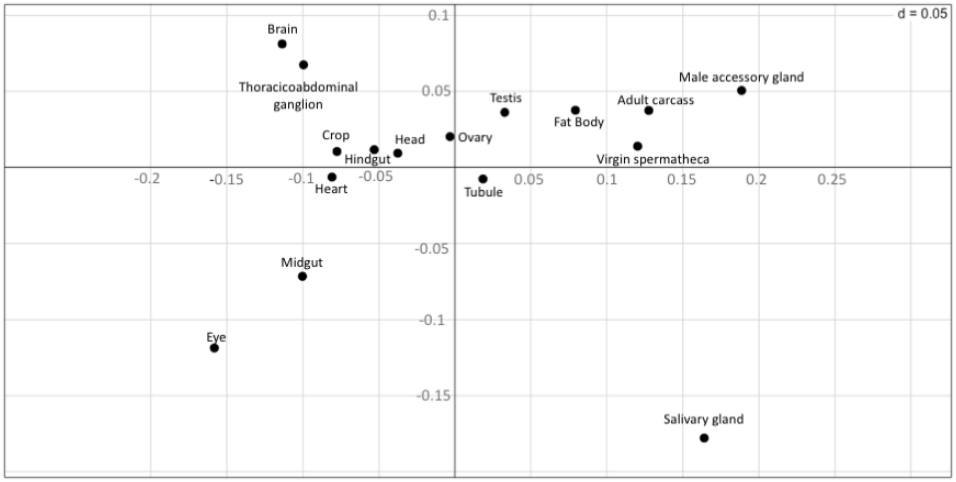

Figure 3.1. Position of tissues along the first two major axes of the correspondence analysis based on the centroid of codon usage values. The vertical axis represents principal component 1 and the horizontal axis corresponds to principal component 2.

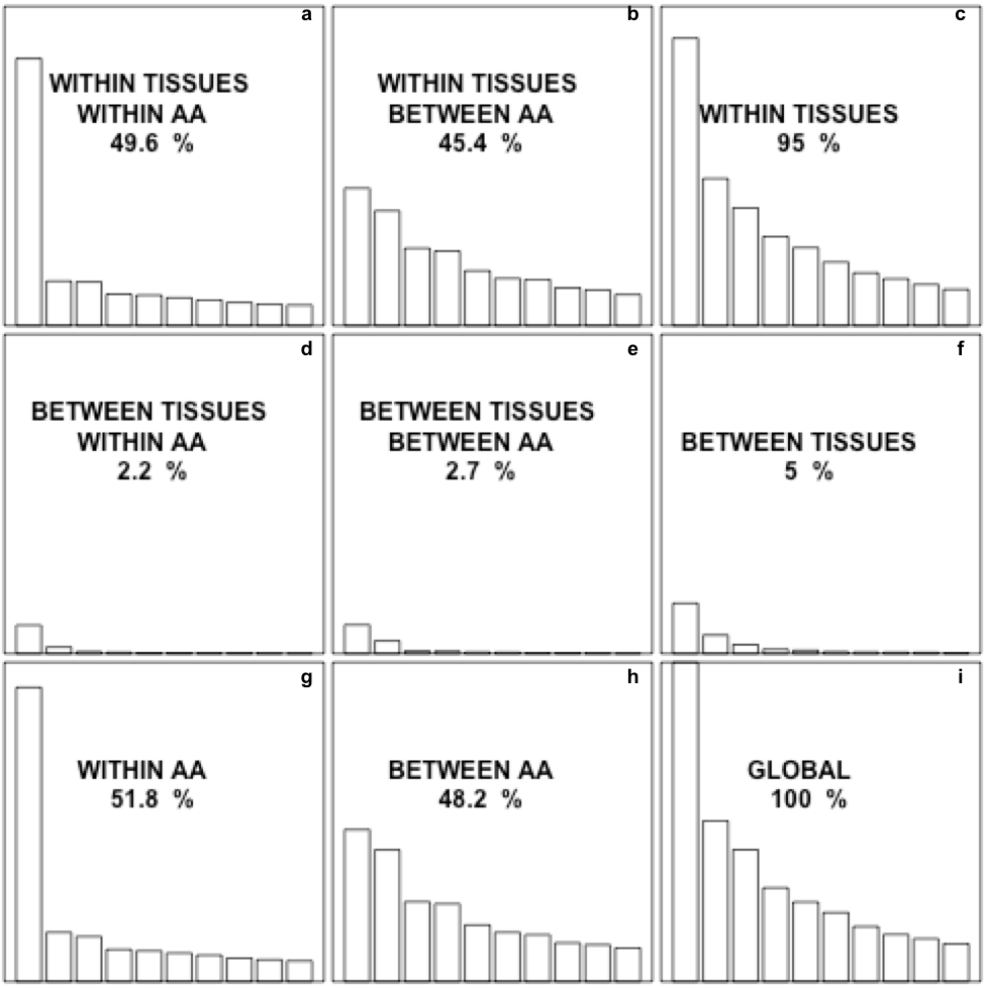

Figure 3.2. Contribution to the global codon usage variability of synonymous, nonsynonymous, between-tissues, and within-tissues effects. The eigenvalue for a given factor is proportional to the fraction of the variability in codon usage that is accounted for by that factor. The total contribution to the variance of each component is indicated. All the graphs are on the same scale to allow direct visual comparison. In each graph, only the first 10 eigenvalues are represented. The fraction of the global variability due to synonymous codon usage (a, d, g) is higher than the fraction explained by nonsynonymous codon usage (b, e, h). The fraction explained by the differences in codon usage within tissues (a, b, c) is much higher than the fraction explained by the differences between tissues (d, e, f).

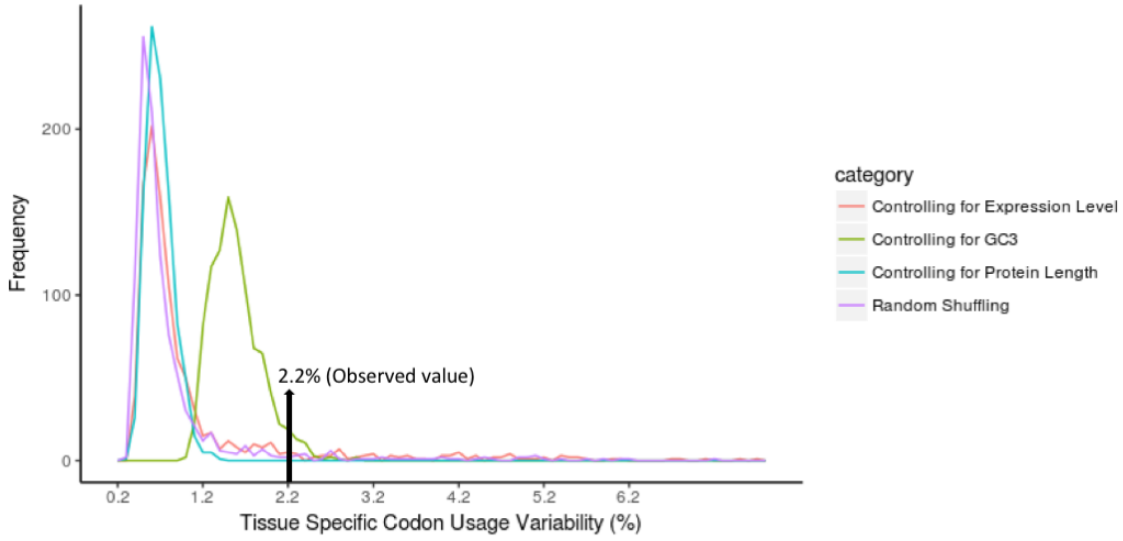

Figure 3.3. Distribution of tissue-specific codon usage variation in 1000 randomized datasets.

Table 3.1. Preferred and unpreferred codons in *D. melanogaster*

Amino acid	Codon	High expression (average RSCU)	Low expression (average RSCU)	*P*-value (RSCU)	*q*-value	Amino acid	Codon	High expression (average RSCU)	Low expression (average RSCU)	*P*-value (RSCU)	*q*-value
Phe	UUU	0.54	0.80	3.7×10^{-60}	2.6×10^{-59}	Ser	UCU	0.58	0.50	3.9×10^{-1}	4.1×10^{-1}
	UUC*	1.46	1.20	2.6×10^{-60}	1.7×10^{-59}		**UCC***	1.76	1.43	8.6×10^{-22}	1.3×10^{-21}
Leu	UUA	0.23	0.35	1.0×10^{-47}	4.0×10^{-47}		UCA	0.43	0.57	1.7×10^{-29}	3.3×10^{-29}
	UUG	1.00	1.17	6.0×10^{-16}	8.6×10^{-16}		**UCG***	1.28	1.10	1.4×10^{-9}	1.27410^{-9}
Leu	CUU	0.56	0.65	1.7×10^{-12}	1.5×10^{-12}	Pro	CCU	0.49	0.55	3.1×10^{-9}	3.8×10^{-9}
	CUC	0.95	0.92	9.0×10^{-1}	9.0×10^{-1}		**CCC***	1.75	1.34	7.0×10^{-42}	1.9×10^{-41}
	CUA	0.39	0.56	1.8×10^{-36}	4.5×10^{-36}		CCA	0.80	1.06	7.5×10^{-30}	1.5×10^{-29}
	CUG*	2.87	2.36	1.0×10^{-38}	2.6×10^{-38}		CCG	0.96	1.06	1.2×10^{-6}	1.4×10^{-6}
Ile	AUU	0.96	1.02	8.1×10^{-4}	1.0×10^{-3}	Thr	ACU	0.67	0.73	1.8×10^{-5}	2.1×10^{-5}
	AUC*	1.70	1.37	1.1×10^{-47}	4.0×10^{-47}		**ACC***	1.95	1.55	1.4×10^{-35}	2.9×10^{-35}
	AUA	0.34	0.61	1.9×10^{-80}	8.1×10^{-79}		ACA	0.58	0.77	6.8×10^{-28}	1.2×10^{-27}
Met	AUG	-	-	-	-		ACG	0.80	0.95	1.6×10^{-13}	2.1×10^{-13}
Val	GUU	0.73	0.81	3.8×10^{-8}	2.7×10^{-8}	Ala	GCU	0.80	0.82	5.7×10^{-1}	5.9×10^{-1}
	GUC*	1.08	0.96	3.6×10^{-10}	4.6×10^{-10}		**GCC***	2.12	1.74	8.6×10^{-49}	4.0×10^{-48}
	GUA	0.34	0.44	2.8×10^{-19}	4.1×10^{-19}		GCA	0.52	0.73	1.9×10^{-43}	6.1×10^{-43}
	GUG*	1.85	1.80	1.4×10^{-2}	1.7×10^{-2}		GCG	0.56	0.72	3.1×10^{-26}	5.3×10^{-26}
Tyr	UAU	0.58	0.81	3.6×10^{-47}	1.3×10^{-46}	Cys	UGU	0.46	0.64	3.5×10^{-25}	5.8×10^{-25}
	UAC*	1.42	1.19	5.7×10^{-47}	1.9×10^{-46}		**UGC***	1.54	1.37	5.3×10^{-25}	8.6×10^{-25}
STOP	UAA	-	-	-	-	STOP	UGA	-	-	-	-
STOP	UAG	-	-	-	-	Trp	UGG	-	-	-	-
His	CAU	0.71	0.84	2.0×10^{-14}	2.8×10^{-14}	Arg	**CGU***	1.33	0.86	4.0×10^{-31}	8.2×10^{-31}
	CAC*	1.29	1.16	3.5×10^{-14}	4.8×10^{-14}		**CGC***	2.47	1.62	2.3×10^{-76}	2.5×10^{-76}
Gln	CAA	0.51	0.67	9.1×10^{-36}	2.1×10^{-35}		CGA	0.62	0.98	1.8×10^{-57}	1.0×10^{-58}
	CAG*	1.49	1.33	1.2×10^{-35}	2.7×10^{-35}		CGG	0.60	0.89	5.8×10^{-41}	1.5×10^{-40}
Asn	AAU	0.71	0.97	2.8×10^{-61}	2.8×10^{-60}	Ser	AGU	0.58	0.97	2.7×10^{-80}	8.1×10^{-79}
	AAC*	1.329	1.03	2.5×10^{-61}	2.8×10^{-60}		AGC	1.37	1.42	1.1×10^{-2}	1.4×10^{-2}
Lys	AAA	0.45	0.69	4.5×10^{-61}	3.5×10^{-60}	Arg	AGA	0.41	0.75	3.1×10^{-69}	4.6×10^{-69}
	AAG*	1.55	1.31	4.7×10^{-61}	3.5×10^{-60}		AGG	0.57	0.89	2.3×10^{-54}	1.1×10^{-53}
Asp	GAU	0.97	1.14	4.3×10^{-29}	8.02×10^{-29}	Gly	**GGU***	0.90	0.84	1.1×10^{-3}	1.2×10^{-3}
	GAC*	1.03	0.86	4.9×10^{-29}	8.84×10^{-29}		**GGC***	1.81	1.54	9.6×10^{-25}	1.5×10^{-24}
Glu	GAA	0.54	0.73	3.1×10^{-43}	8.84×10^{-42}		GGA	1.10	1.29	1.7×10^{-20}	2.5×10^{-20}
	GAG*	1.46	1.27	3.0×10^{-43}	8.84×10^{-42}		GGG	0.19	0.34	3.7×10^{-48}	1.8×10^{-47}

Preferred codons (those for which RSCU is significantly higher for highly expressed genes) are marked with an asterisk and in bold face. *P*-values correspond to the Mann-Whitney *U* test and *q* values indicate FDR correction using the Benjamini and Hochberg approach.

Table 3.2. GC3, expression levels, and protein lengths in different tissues
Variation of GC3, expression level and protein length among genes expressed in different tissue is significant (Kruskal-Wallis test, $P < 0.05$).

Tissue	GC3		Tissue specific expression level		Protein length	
	Median	Mean	Median	Mean	Median	Mean
All genes	0.66	0.65	69.2	196.8	409.00	554.59
Adult carcass	0.54	0.57	38.1	235.1	199.00	279.32
Brain	0.66	0.67	18.5	23.6	652.00	886.91
Crop	0.65	0.63	29.2	51.1	472.00	482.73
Eyes	0.67	0.67	23.4	54.6	229.50	360.91
Fat body	0.60	0.59	11.1	12.3	352.00	492.48
Head	0.62	0.63	19.1	178.4	398.00	552.83
Heart	0.63	0.65	40.2	46.2	362.00	448.67
Hingut	0.62	0.63	59.9	366.4	388.00	421.27
Male accessory glands	0.52	0.54	1996.7	2741.4	344.50	399.33
Midgut	0.66	0.66	464.9	1593.6	335.00	429.83
Ovaries	0.61	0.61	131.7	304.9	415.00	537.60
Salivary glands	0.51	0.52	164.5	373.7	229.00	302.40
Testes	0.60	0.51	411.0	590.6	295.50	404.57
Thoracicoabdominal ganglia	0.67	0.66	26.4	99.3	429.00	457.60
Tubules	0.60	0.62	1227.3	1849.7	414.00	435.14
Virgin spermatheca	0.53	0.5	15.5	1343.3	250.00	280.83

Table 3.3. Internal correspondence analysis with different cut-offs

Cut-off (percent over mean expression level)	Variation in synonymous codon usages between the tissues
25%	2.5%
50%	3.3%
75%	5.6%
100%	13.9%

Different cut-offs were used to quantify the variation in synonymous codon usage between genes highly expressed in the different tissues.

Table 3.4. PERMANOVA results

Variable	Internal single randomization		Internal double randomization	
	Average pseudo-*F*	**No. of datasets with significance at *P* < 0.05**	**Average pseudo-*F***	**No. of datasets with significance at *P* < 0.05**
Tissue	3.94	1000 (100%)	1.42	891 (89.10%)

We generated 1000 random datasets to calculate the average pseudo-*F* value, and for each dataset we used 999 permutations to assess the significance of the observed pseudo-*F* value.

Table 3.5. PERMANCOVA results

Variable	Internal single randomization		Internal double randomization	
	Average pseudo-*F*	**No. of datasets with significance at *P* < 0.05**	**Average pseudo-*F***	**No. of datasets with significance at *P* < 0.05**
Tissue	4.78	1000 (100%)	1.97	1000 (100%)
GC3	397.75	1000 (100%)	36.26	994 (99.4%)
Expression level*	7.05	1000 (100%)	1.97	556 (56.6%)
Protein length*	12.41	1000 (100%)	4.06	210 (21.0%)

We generated 1000 random datasets to calculate the average pseudo-*F* value, and for each dataset we used 999 permutations to assess the significance of the observed pseudo-*F* value. * Data were normalized using logarithmic transformation using base 10.

Chapter 4: Summary and Conclusions

Increased rates of evolution in *A. thaliana* and differences in codon usage biases in the tissues of *D. melanogaster* were found to be small but significant and robust in my data. The effects of self-fertilization on d_N/d_S were not measurable when limited by the number of genes in *A. lyrata* (Wright et al. 2002; Plotkin et al. 2004; Sémon et al. 2006; Foxe et al. 2008). The analyses of the genomes of *A. lyrata* and *A. halleri* showed that these small differences were consistently significant. Codon usage differences in the tissues of *D. melanogaster* were also small, but were found to be consistent when tested with Monte Carlo methods and investigations of the effects of GC content, length, and expression level.

Values of d_N, d_S and d_N/d_S were found to be significantly higher in *A. thaliana* when compared to orthologous sequences in *A. lyrata* and *A. halleri* (Table S2.1). *A. thaliana* was found to have higher d_N/d_S in a significantly higher number of genes than either *A. lyrata* or *A. halleri.* 907 orthologous genes were determined to be evolving at significantly different rates using a likelihood ratio test, and in these genes, the differences in evolutionary rate proved to be marginally higher in *A. thaliana.* These results were also found in the analyses of 18 different accessions of the *Arabidopsis thaliana* genome (Tables S2.4 and S2.5). Genes were separated into functional categories using KOG classification to rule out the potential effect of pseudogenization of genes pertaining to outcrossing. *A. thaliana* was found to have higher d_N/d_S in 19 of these 23 categories (Table 2.2). Many nonsynonymous and synonymous mutations were found to be exclusive to the genome of *A. thaliana*, and for both nonsynonymous and synonymous mutations these differences were found to be significant using Tajima's test (Table 2.1).

It is possible that these findings may be indicative of factors in the evolutionary history of the *Arabidopsis* family other than differences in reproductive methods. *Arabidopsis thaliana* diverged from the *Arabidopsis lyrata/Arabidopsis halleri* clade only 6-13 MYA and transitioned to self-fertilization 100,000 – 1M YA (Beilstein et al. 2010; Hohmann et al. 2015). The short evolutionary time between these events may have limited the accumulation of mutations when comparing these genomes. My results have shown that the *Arabidopsis* genomes are all undergoing strong purifying selection showing that this time difference may not be long enough to expose significant differences in the genome, although the concatenated genomes were shown to be evolving at significantly different rates. High purifying selection also may have had purged recessive, deleterious mutations in *A. thaliana* due to increased homozygosity due to low effective population size (Pollak 1987; Glémin 2007). *A. lyrata* is also thought to have recently gone through a significant population bottleneck, which may have decreased the efficacy of selection in this species (Mattila et al. 2017).

Evolutionary differences in annual plants, such as *A. thaliana*, and perennials, such as *A. lyrata*, have been noted (Gaut et al. 2011). Annual plants tend to have an increase in synonymous substitution rates but decreased d_N/d_S (Smith and Donoghue 2008). While this does not explain increased d_N/d_S in *A. thaliana* it may explain why differences were not very pronounced. While these extra-genomic events may convolute the life-history of these species, the results show a robustness to the increased evolutionary rates in *A. thaliana*.

The genome of *Drosophila melanogaster* shows a significant difference in codon usage bias in genes exclusively expressed in different tissues. Microarray based expression data was used to determine genes that were expressed in exclusively one tissue. It was determined that the globally preferred codons were not always preferred in

tissue specific genes (Table S3.1), and where these codons were exclusive to a tissue, the frequencies were significantly different using X^2 tests (Table S3.2). It was found using a randomized dataset that differences in codon usage bias were not wholly explained by GC content, protein length, or expression level (Tables S3.7-9). A small amount of these biases could be explained by the tissue of origin for a gene using correspondence analyses (2.2%, Figure 3.2). Monte Carlo simulations with randomly selected genes and genes selected randomly to recreate GC, protein length, and expression distributions proved that differences were due to the tissue that proteins were expressed in and not due to random variation in codon usage or other protein factors (Figure 3.3). PERMANOVA analyses with randomized datasets found that variance in codon usages were always explained by tissue differences (Table 3.4). Analyses of covariance found that GC content, protein length, and gene expression did not significantly explain these variances, especially when controlling for the number of genes within tissues (Table 3.5).

Studying proteins that were expressed in their respective tissues more than the global average expression level amplified the effects of tissue exclusivity on codon usage (Table 3.3). This result was peculiar due to finding that protein expression levels did not always significantly predict the differences in codon usage, however higher usage of preferred codons has been found in highly expressed proteins (Bennetzen and Hall 1982). tRNA abundances were not available for this analyses, which could have provided a correlation between preferred codons in tissues to cognate tRNA abundances in the cell.

As I have shown additional evidence that effective population size effects selection, the difference in the effect of tissue of expression on codon preferences between *D. melanogaster* and humans (stronger in *D. melanogaster*) may be due to

overall higher effective population sizes in *Drosophila* (Yu et al. 2001; Shapiro et al. 2007). Preferred codons are favored by selection, and decreased effective population size in humans may have stifled prominent differences in codon usage bias within tissues (Sharp et al. 1993b). A future side-by-side analysis replicating work from Plotkin *et al.* and Sémon *et al.* to study the correlation between selection and codon usage could aid in explaining the differences of these findings. This future work could also explore the effects of isochoric structures in humans compared to the lack of isochoric structure in both *Drosophila* and *Arabidopsis.*

Previous studies were the number of available genes were often less than 1000 often showed that these results were either not measurable or highly correlated with other factors (Wright et al. 2002; Plotkin et al. 2004; Foxe et al. 2008). Rigorous randomization and controls in both experiments concluded that these subtle effects of evolution can be quite robust with sufficiently large datasets.

References

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

Alvarez-Ponce, D. 2014. Why proteins evolve at different rates: The determinants of proteins' rates of evolution. Pp. 126-178 *in* M. A. Fares, ed. Natural Selection: Methods and Applications. CRC Press (Taylor & Francis), London.

Alvarez-Ponce, D. and M. A. Fares. 2012. Evolutionary Rate and Duplicability in the Arabidopsis thaliana Protein–Protein Interaction Network. Genome Biol Evol 4:1263-1274.

Anderson, M. J. 2001. A new method for non-parametric multivariate analysis of variance. 26:32–46.

Andersson, S. G. E. and C. G. Kurland. 1990. CODON PREFERENCES IN FREE-LIVING MICROORGANISMS. Microbiol Rev 54:198-210.

Basak, S. and T. C. Ghosh. 2006. Temperature adaptation of synonymous codon usage in different functional categories of genes: A comparative study between homologous genes of Methanococcus jannaschii and Methanococcus maripaludis. Febs Letters 580:3895-3899.

Bechsgaard, J. S., V. Castric, D. Charlesworth, X. Vekemans, and M. H. Schierup. 2006. The Transition to Self-Compatibility in Arabidopsis thaliana and Evolution within S-Haplotypes over 10 Myr. Mol Biol Evol 23:1741-1750.

Beilstein, M. A., N. S. Nagalingum, M. D. Clements, S. R. Manchester, and S. Mathews. 2010. Dated molecular phylogenies indicate a Miocene origin for Arabidopsis thaliana. Proc Natl Acad Sci U S As 107:18724-18728.

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

Bennetzen, J. L. and B. D. Hall. 1982. Codon selection in yeast. J Biol Chem 257:3026-3031.

Bernardi, G. 1989. THE ISOCHORE ORGANIZATION OF THE HUMAN GENOME. Annu Rev of Genet 23:637-661.

Bernardi, G. 2000. Isochores and the evolutionary genomics of vertebrates. Gene 241:3-17.

Brandvain, Y., A. M. Kenney, L. Flagel, G. Coop, and A. L. Sweigart. 2014. Speciation and Introgression between Mimulus nasutus and Mimulus guttatus. PLoS Genet 10:e1004410.

Briskine, R. V., T. Paape, R. Shimizu-Inatsugi, T. Nishiyama, S. Akama, J. Sese, and K. K. Shimizu. 2016. Data from: Genome assembly and annotation of Arabidopsis halleri, a model for heavy metal hyperaccumulation and evolutionary ecology. Dryad Data Repository.

Camiolo, S., L. Farina, and A. Porceddu. 2012. The Relation of Codon Bias to Tissue-Specific Gene Expression in Arabidopsis thaliana. Genetics 192:641-649.

Cao, J., K. Schneeberger, S. Ossowski, T. Gunther, S. Bender, J. Fitz, D. Koenig, C. Lanz, O. Stegle, C. Lippert, X. Wang, F. Ott, J. Muller, C. Alonso-Blanco, K. Borgwardt, K. J.

Schmid, and D. Weigel. 2011. Whole-genome sequencing of multiple Arabidopsis thaliana populations. Nat Genet 43:956-963.

Cazes, P., D. Chessel, and S. Doledec. 1988. L'analyse des correspondances internes d'un tableau partitionné : son usage en hydrobiologie. Revue de Statistique Appliquée 36:39-54.

Chakraborty, S. and D. Alvarez-Ponce. 2016. Positive Selection and Centrality in the Yeast and Fly Protein-Protein Interaction Networks. Biomed Res Int.

Charif, D., O. Clerc, C. Frank, J. R. Lobry, A. Necsulea, L. Palmeira, S. Penel, and G. Perrière. 2017. Package 'seqinr'. Biological Sequences Retrieval and Analysis.

Charlesworth, B. 2009. Effective population size and patterns of molecular evolution and variation. Nat Rev Genet 10:195-205.

Charlesworth, B., M. T. Morgan, and D. Charlesworth. 1993. The effect of deleterious mutations on neutral molecular variation. Genetics 134:1289-1303.

Charlesworth, D. and X. Vekemans. 2005. How and when did Arabidopsis thaliana become highly self-fertilising. BioEssays 27:472-476.

Charlesworth, D. and S. I. Wright. 2001. Breeding systems and genome evolution. Curr Opin Genet Dev 11.

Chintapalli, V. R., J. Wang, and J. A. T. Dow. 2007. Using FlyAtlas to identify better Drosophila melanogaster models of human disease. Nature Genetics 39:715-720.

Dittmar, K. A., J. M. Goodenbour, and T. Pan. 2006. Tissue-specific differences in human transfer RNA expression. PLoS Genet 2:2107-2115.

Dong, H. J., L. Nilsson, and C. G. Kurland. 1996. Co-variation of tRNA abundance and codon usage in Escherichia coli at different growth rates. J Mol Biol 260:649-663.

Dray, S., A.-B. Dufour, J. Thioulouse, T. Jombart, S. Pavoine, J. R. Lobry, S. Ollier, and A. Siberchicot. 2016. Package 'ade4'. Analysis of Ecological Data : Exploratory and Euclidean Methods in Environmental Sciences.

Du, J., S. Z. Dungan, A. Sabouhanian, and B. S. Chang. 2014. Selection on synonymous codons in mammalian rhodopsins: a possible role in optimizing translational processes. BMC Evol Biol 14:96.

Duret, L. 2002. Evolution of synonymous codon usage in metazoans. Curr Opin Genet Dev 12:640-649.

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

Durvasula, A., A. Fulgione, R. M. Gutaker, S. I. Alacakaptan, P. J. Flood, C. Neto, T. Tsuchimatsu, H. A. Burbano, F. X. Picó, C. Alonso-Blanco, and A. M. Hancock. 2017. African genomes illuminate the early history and transition to selfing in Arabidopsis thaliana. Proc Natl Acad Sci U S As.

Escobar, J. S., A. Cenci, J. Bolognini, A. Haudry, S. Laurent, J. David, and S. Glémin. 2010. An integrative test of the dead-end hypothesis of selfing evolution in Triticeae (Poaceae). Evolution 64:2855-2872.

Flicek, P., I. Ahmed, M. R. Amode, D. Barrell, K. Beal, S. Brent, D. Carvalho-Silva, P. Clapham, G. Coates, S. Fairley, S. Fitzgerald, L. Gil, C. Garcia-Giron, L. Gordon, T.

Hourlier, S. Hunt, T. Juettemann, A. K. Kaehaeri, S. Keenan, M. Komorowska, E. Kulesha, I. Longden, T. Maurel, W. M. McLaren, M. Muffato, R. Nag, B. Overduin, M. Pignatelli, B. Pritchard, E. Pritchard, H. S. Riat, G. R. S. Ritchie, M. Ruffier, M. Schuster, D. Sheppard, D. Sobral, K. Taylor, A. Thormann, S. Trevanion, S. White, S. P. Wilder, B. L. Aken, E. Birney, F. Cunningham, I. Dunham, J. Harrow, J. Herrero, T. J. P. Hubbard, N. Johnson, R. Kinsella, A. Parker, G. Spudich, A. Yates, A. Zadissa, and S. M. J. Searle. 2013. Ensembl 2013. Nucleic Acids Res 41:D48-D55.

Foxe, J. P., V.-u.-N. Dar, H. Zheng, M. Nordborg, B. S. Gaut, and S. I. Wright. 2008. Selection on Amino Acid Substitutions in Arabidopsis. Mol Biol Evol 25:1375-1383.

Gan, X., O. Stegle, J. Behr, J. G. Steffen, P. Drewe, K. L. Hildebrand, R. Lyngsoe, S. J. Schultheiss, E. J. Osborne, V. T. Sreedharan, A. Kahles, R. Bohnert, G. Jean, P. Derwent, P. Kersey, E. J. Belfield, N. P. Harberd, E. Kemen, C. Toomajian, P. X. Kover, R. M. Clark, G. Ratsch, and R. Mott. 2011. Multiple reference genomes and transcriptomes for Arabidopsis thaliana. Nature 477:419-423.

Gaut, B., L. Yang, S. Takuno, and L. E. Eguiarte. 2011. The Patterns and Causes of Variation in Plant Nucleotide Substitution Rates. Annu Rev Ecol Evol Syst 42:245-266.

Glémin, S. 2007. Mating Systems and the Efficacy of Selection at the Molecular Level. Genetics 177:905-916.

Goodenbour, J. M. and T. Pan. 2006. Diversity of tRNA genes in eukaryotes. Nucleic Acids Res 34:6137-6146.

Grantham, R., C. Gautier, and M. Gouy. 1980. CODON FREQUENCIES IN 119 INDIVIDUAL GENES CONFIRM CONSISTENT CHOICES OF DEGENERATE BASES ACCORDING TO GENOME TYPE. Nucleic Acids Res 8:1893-1912.

Hanson, G. and J. Coller. 2017. Codon optimality, bias and usage in translation and mRNA decay. Nat Rev Mol Cell Biol. 19:20.

Hassan, S., V. Mahalingam, and V. Kumar. 2009. Synonymous codon usage analysis of thirty two mycobacteriophage genomes. Adv Bioinformatics:316936.

Haudry, A., A. Cenci, C. Guilhaumon, E. Paux, S. Poirier, S. Santoni, J. David, and S. Glémin. 2008. Mating system and recombination affect molecular evolution in four Triticeae species. Genet Res 90:97-109.

Hohmann, N., E. M. Wolf, M. A. Lysak, and M. A. Koch. 2015. A Time-Calibrated Road Map of Brassicaceae Species Radiation and Evolutionary History. Plant Cell.

Hu, T. T., P. Pattyn, E. G. Bakker, J. Cao, J.-F. Cheng, R. M. Clark, N. Fahlgren, J. A. Fawcett, J. Grimwood, H. Gundlach, G. Haberer, J. D. Hollister, S. Ossowski, R. P. Ottilar, A. A. Salamov, K. Schneeberger, M. Spannagl, X. Wang, L. Yang, M. E. Nasrallah, J. Bergelson, J. C. Carrington, B. S. Gaut, J. Schmutz, K. F. X. Mayer, Y. Van de Peer, I. V. Grigoriev, M. Nordborg, D. Weigel, and Y.-L. Guo. 2011. The Arabidopsis lyrata genome sequence and the basis of rapid genome size change. Nature genetics 43:476-481.

Huelsenbeck, J. P. and K. A. Crandall. 1997. Phylogeny estimation and hypothesis testing using maximum likelihood. Annu Rev of Ecol and Syst. 28:437-466.

Huerta-Cepas, J., D. Szklarczyk, K. Forslund, H. Cook, D. Heller, M. C. Walter, T. Rattei, D. R. Mende, S. Sunagawa, and M. Kuhn. 2015. eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. Nucleic Acids Res 44:D286-D293.

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 ESCHERICHIA-COLI TRANSLATIONAL SYSTEM. J Mol Biol 151:389-409.

Ikemura, T. 1982. Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes. Differences in synonymous codon choice patterns of yeast and Escherichia coli with reference to the abundance of isoaccepting transfer RNAs. J Mol Biol 158:573-597.

Jabbari, K. and G. Bernardi. 2000. The distribution of genes in the Drosophila genome. Gene 247:287-292.

Johnston, M. O., E. Porcher, P.-O. Cheptou, C. G. Eckert, E. Elle, M. A. Geber, S. Kalisz, J. K. Kelly, D. A. Moeller, and M. Vallejo-Marin. 2008. Correlations among fertility components can maintain mixed mating in plants. The Am Nat 173:1-11.

Kanaya, S., Y. Yamada, M. Kinouchi, Y. Kudo, and T. Ikemura. 2001. Codon usage and tRNA genes in eukaryotes: correlation of codon usage diversity with translation efficiency and with CG-dinucleotide usage as assessed by multivariate analysis. J Mol Evol 53:290-298.

Kersey, P. J., J. E. Allen, I. Armean, S. Boddu, B. J. Bolt, D. Carvalho-Silva, M. Christensen, P. Davis, L. J. Falin, C. Grabmueller, J. Humphrey, A. Kerhornou, J. Khobova, N. K. Aranganathan, N. Langridge, E. Lowy, M. D. McDowall, U. Maheswari, M. Nuhn, C. K. Ong, B. Overduin, M. Paulini, H. Pedro, E. Perry, G. Spudich, E. Tapanari, B. Walts, G. Williams, M. Tello-Ruiz, J. Stein, S. Wei, D. Ware, D. M. Bolser, K. L. Howe, E. Kulesha, D. Lawson, G. Maslen, and D. M. Staines. 2016. Ensembl Genomes 2016: more genomes, more complexity. Nucleic Acids Res 44:D574-D580.

Kimura, M. 1968. Evolutionary rate at the molecular level. Nature 217:624-626.

Kimura, M. 1983. The Neutral Theory of Molecular Evolution. Cambridge University Press, Cambridge.

Kimura, M., T. Maruyama, and J. F. Crow. 1963. The Mutation Load in Small Populations. Genetics 48:1303-1312.

Koch, M. A. and M. Kiefer. 2005. Genome evolution among cruciferous plants: a lecture from the comparison of the genetic maps of three diploid species—Capsella rubella, Arabidopsis lyrata subsp. petraea, and A. thaliana. Am J Bot 92:761-767.

Lercher, M. J., A. O. Urrutia, and L. D. Hurst. 2002. Clustering of housekeeping genes provides a unified model of gene order in the human genome. Nature Genetics 31:180-183.

Lobry, J. R. and D. Chessel. 2003. Internal correspondence analysis of codon and amino-acid usage in thermophilic bacteria. J Appl Genet 44:235-261.

Lu, P., C. Vogel, R. Wang, X. Yao, and E. M. Marcotte. 2006. Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. Nature Biotechnology 25:117.

Löytynoja, A. and N. Goldman. 2005. An algorithm for progressive multiple alignment of sequences with insertions. Proc Natl Acad Sci U S A 102:10557-10562.

Mattila, T. M., J. Tyrmi, T. Pyhäjärvi, and O. Savolainen. 2017. Genome-wide analysis of colonization history and concomitant selection in Arabidopsis lyrata. Mol Biol Evol.

Moore, R. C. and M. D. Purugganan. 2003. The early stages of duplicate gene evolution. Proc Natl Acad Sci U S A 100:15682-15687.

Moriyama, E. N. and J. R. Powell. 1997. Codon usage bias and tRNA abundance in Drosophila. J Mol Evol 45.

Muto, A. and S. Osawa. 1987. The guanine and cytosine content of genomic DNA and bacterial evolution. Proc Natl Acad Sci U S A 84:166-169.

Nekrutenko, A. and W. H. Li. 2000. Assessment of compositional heterogeneity within and between eukaryotic genomes. Genome Res 10:1986-1995.

Ness, R. W., M. Siol, and S. C. Barrett. 2012. Genomic consequences of transitions from cross-to self-fertilization on the efficacy of selection in three independently derived selfing plants. BMC genomics 13:611.

Ohta, T. 1973. Slightly Deleterious Mutant Substitutions in Evolution. Nature 246:96.

Olivares-Hernández, R., S. Bordel, and J. Nielsen. 2011. Codon usage variability determines the correlation between proteome and transcriptome fold changes. BMC Systems Biology 5:33.

Oliver, J. L., P. Bernaola-Galvan, P. Carpena, and R. Roman-Roldan. 2001. Isochore chromosome maps of eukaryotic genomes. Gene 276:47-56.

Plotkin, J. B., H. Robins, and A. J. Levine. 2004. Tissue-specific codon usage and the expression of human genes. Proc Natl Acad Sci U S A 101:12588-12591.

Pollak, E. 1987. On the Theory of Partially Inbreeding Finite Populations. I. Partial Selfing. Genetics 117:353.

Powell, J. and K. Dion. 2015. Effects of Codon Usage on Gene Expression: Empirical Studies on Drosophila. J Mol Evol 80:219-226.

Qiu, S., K. Zeng, T. Slotte, S. Wright, and D. Charlesworth. 2011. Reduced Efficacy of Natural Selection on Codon Usage Bias in Selfing Arabidopsis and Capsella Species. Genome Biol Evol 3:868-880.

R-Core-Team. 2013. R: A language and environment for statistical computing. R Foundation for Statistical Computing. 2013 Vienna, Austria.

Rocha, E. P. C. 2004. Codon usage bias from tRNA's point of view: Redundancy, specialization, and efficient decoding for translation optimization. Genome Res 14:2279-2286.

Schmid, M., T. S. Davison, S. R. Henz, U. J. Pape, M. Demar, M. Vingron, B. Scholkopf, D. Weigel, and J. U. Lohmann. 2005. A gene expression map of Arabidopsis thaliana development. Nat Genet 37:501-506.

Semon, M., J. R. Lobry, and L. Duret. 2006. No evidence for tissue-specific adaptation of synonymous codon usage in humans. Mol Biol Evol 23:523-529.

Sémon, M., J. R. Lobry, and L. Duret. 2006. No Evidence for Tissue-Specific Adaptation of Synonymous Codon Usage in Humans. Mol Biol Evol 23:523-529.

Shapiro, J. A., W. Huang, C. Zhang, M. J. Hubisz, J. Lu, D. A. Turissini, S. Fang, H. Y. Wang, R. R. Hudson, R. Nielsen, Z. Chen, and C. I. Wu. 2007. Adaptive genic evolution in the Drosophila genomes. Proc Natl Acad Sci U S A 104:2271-2276.

Sharp, P. M., M. Stenico, J. F. Peden, and A. T. Lloyd. 1993a. CODON USAGE - MUTATIONAL BIAS, TRANSLATIONAL SELECTION, OR BOTH. Biochem Soc Trans 21:835-841.

Sharp, P. M., M. Stenico, J. F. Peden, and A. T. Lloyd. 1993b. Codon usage: mutational bias, translational selection, or both? Portland Press Limited.

Shaw, R. G. and M.-O. Thomas. 1993. ANOVA for unbalanced data: an overview. 74:1638-1645.

Shimizu, K. K. and T. Tsuchimatsu. 2015. Evolution of Selfing: Recurrent Patterns in Molecular Adaptation. Annu Rev Ecol Evol Syst 46:593-622.

Slotte, T., K. M. Hazzouri, J. A. Agren, D. Koenig, F. Maumus, Y.-L. Guo, K. Steige, A. E. Platts, J. S. Escobar, L. K. Newman, W. Wang, T. Mandakova, E. Vello, L. M. Smith, S. R. Henz, J. Steffen, S. Takuno, Y. Brandvain, G. Coop, P. Andolfatto, T. T. Hu, M. Blanchette, R. M. Clark, H. Quesneville, M. Nordborg, B. S. Gaut, M. A. Lysak, J. Jenkins, J. Grimwood, J. Chapman, S. Prochnik, S. Shu, D. Rokhsar, J. Schmutz, D. Weigel, and S. I. Wright. 2013. The Capsella rubella genome and the genomic consequences of rapid mating system evolution. Nat Genet 45:831-835.

Smith, A. M., R. Abu-Shumays, M. Akeson, and D. L. Bernick. 2015. Capture, Unfolding, and Detection of Individual tRNA Molecules Using a Nanopore Device. Front Bioeng Biotechnol 3:91.

Smith, S. A. and M. J. Donoghue. 2008. Rates of Molecular Evolution Are Linked to Life History in Flowering Plants. Science 322:86-89.

Stebbins, G. L. 1957. Self Fertilization and Population Variability in the Higher Plants. The Am Nat 91:337-354.

Sueoka, N. and Y. Kawanishi. 2000. DNA G+C content of the third codon position and codon usage biases of human genes. Gene 261:53-62.

Tajima, F. 1993. Simple methods for testing the molecular evolutionary clock hypothesis. Genetics 135:599.

The *Arabidopsis* Genome Initative. 2000. Anlaysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature. 408:796-815.

Thiery, J. P., G. Macaya, and G. Bernardi. 1976. ANALYSIS OF EUKARYOTIC GENOMES BY DENSITY GRADIENT CENTRIFUGATION. J Mol Biol 108:219-235.

Vicario, S., E. N. Moriyama, and J. R. Powell. 2007. Codon usage in twelve species of Drosophila. BMC Evol Biol 7.

Wagner, A. 2005. Energy constraints on the evolution of gene expression. Mol Biol Evol 22:1365-1374.

Wan, X. F., D. Xu, A. Kleinhofs, and J. Z. Zhou. 2004. Quantitative relationship between synonymous codon usage bias and GC composition across unicellular genomes. BMC Evol Biol 4.

Wright, S. I., B. Lauga, and D. Charlesworth. 2002. Rates and Patterns of Molecular Evolution in Inbred and Outbred Arabidopsis. Mol Biol Evol 19:1407-1420.

Yang, Y.-F., T. Zhu, and D.-K. Niu. 2013. Association of intron loss with high mutation rate in Arabidopsis: implications for genome size evolution. Genome Biol Evol 5:723-733.

Yang, Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 24:1586-1591.

Yang, Z. and R. Nielsen. 2008. Mutation-Selection Models of Codon Substitution and Their Use to Estimate Selective Strengths on Codon Usage. Mol Biol Evol 25:568-579.

Yu, A., C. Zhao, Y. Fan, W. Jang, A. J. Mungall, P. Deloukas, A. Olsen, N. A. Doggett, N. Ghebranious, K. W. Broman, and J. L. Weber. 2001. Comparison of human genetic and sequence-based physical maps. Nature 409.

Yu, N., M. I. Jensen-Seaman, L. Chemnick, O. Ryder, and W. H. Li. 2004. Nucleotide diversity in gorillas. Genetics 166:1375-1383.

Appendix A

Table S1.1. Estimation of evolutionary rates in *Arabidopsis thaliana* and *Arabidopsis lyrata*.

	Arabidopsis thaliana	*Arabidopsis lyrata*	Binomial *P*-value
Median d_N	0.0108	0.0085	-
Median d_S	0.0757	0.0612	-
Median d_N / d_S	0.1423	0.1383	-
Mean d_N	0.0133	0.0107	-
Mean d_S	0.0805	0.0667	-
Mean d_N / d_S	0.1865	0.188	-
Genes with higher d_N	8572	4396	1.20×10^{-324} ***
Genes with higher d_S	8938	4055	4.94×10^{-324} ***
Genes with higher d_N / d_S	6625	6161	4.22×10^{-5} ***

*, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$

Table S1.2. Estimation of evolutionary rates on expression sorted genes in *A. thaliana* and *A. lyrata.*

	Low Expression[a]	Medium Expression[b]	High Expression[c]
Number of genes with higher d_N in *A. thaliana*	2079	2060	2024
Number of genes with higher d_N in *A. lyrata*	1028	1045	1070
Binomial test *P*-value for d_N	1.08×10^{-80} ***	2.5×10^{-75} ***	7.54×10^{-67} ***
Number of genes with higher d_S in *A. thaliana*	2132	2195	2098
Number of genes with higher d_S in *A. lyrata*	977	913	1009
Binomial test *P*-value for d_N	2.04×10^{-97} ***	2.52×10^{-120} ***	1.26×10^{-86} ***
Number of genes with higher d_N/d_N in *A. thaliana*	1576	1616	1582
Number of genes with higher d_N/d_S in *A. lyrata*	1522	1468	1411
Binomial test *P*-value for d_N/d_S	0.3410	0.0081 **	0.0019 *

[a] Expression < 0.48 PPT

[b] Expression levels > 0.49 PPT and < 6.9 PPT

[c] Expression levels > 6.9 PPT

*, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$.

Table S1.3. Evolutionary rates of *Arabidopsis thaliana* and *Arabidopsis lyrata* concatenomes.

Branch	d_N	d_S	d_N / d_S	*P*- value[a]
Capsella rubella	0.0293	0.1819	0.1611	-
Arabidopsis thaliana	0.0127	0.0759	0.1671	0.0014 **
Arabidopsis lyrata	0.0102	0.0622	0.1644	

[a]*P*-value was determined comparing a 2-ratio model and a free-ratios model using the likelihood ratio test.

*, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$.

Table S1.4. Estimation of evolutionary rates in alternate strains of Arabidopsis thaliana.

	A. thaliana	*A. lyrata*	*P*-value[a]	Significant *A. thaliana*	Significant *A. lyrata*	Significant Genes *P*-value[a]
Higher d_N in Bur0 comparison	8367	4251	$< 2.2 \times 10^{-16}$ ***	479	399	0.0076 **
Higher d_S in Bur0 comparison	8702	3935	$< 2.2 \times 10^{-16}$ ***	527	352	3.91×10^{-5} ***
Higher d_N / d_S in Bur0 comparison	6438	5997	7.94×10^{-5} ***	467	411	0.0634
Higher d_N in Can0 comparison	8255	4186	$< 2.2 \times 10^{-16}$ ***	474	409	0.031 *
Higher d_S in Can0 comparison	8601	3865	$< 2.2 \times 10^{-16}$ ***	537	346	1.38×10^{-10} ***
Higher d_N / d_S in Can0 comparison	6341	5921	1.54×10^{-4} ***	460	422	0.2128
Higher d_N in Ct1 comparison	8288	4201	$< 2.2 \times 10^{-16}$ ***	2881	1679	$< 2.2 \times 10^{-16}$ ***
Higher d_S in Ct1 comparison	8632	3882	$< 2.2 \times 10^{-16}$ ***	3088	1477	$< 2.2 \times 10^{-16}$ ***
Higher d_N / d_S in Ct1 comparison	6363	5948	.00019 ***	2285	2226	0.3878
Higher d_N in Edi0 comparison	8367	4246	$< 2.2 \times 10^{-16}$ ***	480	391	0.0028 **
Higher d_S in Edi0 comparison	8710	3929	$< 2.2 \times 10^{-16}$ ***	519	352	1.69×10^{-8} ***
Higher d_N / d_S in Edi0 comparison	6430	6011	1.78×10^{-4} ***	463	407	0.0622
Higher d_N in Hi0 comparison	8275	4329	$< 2.2 \times 10^{-16}$ ***	489	411	0.0102 *
Higher d_S in Hi0 comparison	8699	3928	$< 2.2 \times 10^{-16}$ ***	542	358	9.31×10^{-10} ***
Higher d_N / d_S in Hi0 comparison	6414	6012	3.21×10^{-4} ***	475	424	0.0953
Higher d_N in Kn0 comparison	8367	4209	$< 2.2 \times 10^{-16}$ ***	467	400	0.02493901 *
Higher d_S in Kn0 comparison	8728	3876	$< 2.2 \times 10^{-16}$ ***	525	343	7.02×10^{-10} ***
Higher d_N / d_S in Kn0 comparison	6407	5986	1.61×10^{-4} ***	452	415	0.2214
Higher d_N in Ler0 comparison	8343	4251	$< 2.2 \times 10^{-16}$ ***	494	413	0.0079 **
Higher d_S in Ler0 comparison	8737	3876	$< 2.2 \times 10^{-16}$ ***	536	372	5.84×10^{-8} ***
Higher d_N / d_S in Ler0 comparison	6403	6015	5.14×10^{-4} ***	482	425	0.0629
Higher d_N in Mt0 comparison	8335	4271	$< 2.2 \times 10^{-16}$ ***	485	399	0.0042 **
Higher d_S in Mt0 comparison	8708	3921	$< 2.2 \times 10^{-16}$ ***	516	369	8.72×10^{-7} ***
Higher d_N / d_S in Mt0 comparison	6450	5978	2.38×10^{-5} ***	477	407	0.0202 *
Higher d_N in No0 comparison	8338	4260	$< 2.2 \times 10^{-16}$ ***	469	414	0.0691
Higher d_S in No0 comparison	8710	3911	$< 2.2 \times 10^{-16}$ ***	530	353	2.91×10^{-9} ***
Higher d_N / d_S in No0 comparison	6396	6019	7.38×10^{-4}***	459	423	0.2386
Higher d_N in Oy0 comparison	8350	4233	$< 2.2 \times 10^{-16}$ ***	495	414	0.0079 **
Higher d_S in Oy0 comparison	8686	3921	$< 2.2 \times 10^{-16}$ ***	544	365	3.17×10^{-9} ***
Higher d_N / d_S in Oy0 comparison	6442	5968	2.17×10^{-5} ***	483	425	0.0585
Higher d_N in Po0 comparison	7643	4133	$< 2.2 \times 10^{-16}$ ***	449	378	0.0149 *
Higher d_S in Po0 comparison	7957	3844	$< 2.2 \times 10^{-16}$ ***	482	345	2.14×10^{-6} ***
Higher d_N / d_S in Po0 comparison	5996	5616	4.36×10^{-4} ***	442	384	0.0473 *
Higher d_N in Rsch4 comparison	7880	4018	$< 2.2 \times 10^{-16}$ ***	448	393	0.0625
Higher d_S in Rsch4 comparison	8200	3722	$< 2.2 \times 10^{-16}$ ***	512	329	2.97×10^{-10} ***
Higher d_N / d_S in Rsch4 comparisoı	6079	5651	8.05×10^{-5} ***	437	403	0.2549
Higher d_N in Sf2 comparison	8317	4250	$< 2.2 \times 10^{-16}$ ***	475	413	0.0406 *
Higher d_S in Sf2 comparison	8641	3948	$< 2.2 \times 10^{-16}$ ***	537	351	4.65×10^{-10} ***
Higher d_N / d_S in Sf2 comparison	6396	5995	3.26×10^{-4} ***	461	427	0.2681
Higher d_N in Tsu0 comparison	8320	4284	$< 2.2 \times 10^{-16}$ ***	480	414	0.0297 *
Higher d_S in Tsu0 comparison	8698	3934	$< 2.2 \times 10^{-16}$ ***	538	357	1.57×10^{-9} ***
Higher d_N / d_S in Tsu0 comparison	6405	6030	7.96×10^{-4} ***	468	426	0.1703
Higher d_N in Wil2 comparison	8345	4214	$< 2.2 \times 10^{-16}$ ***	479	391	0.0032 **
Higher d_S in Wil2 comparison	8727	3857	$< 2.2 \times 10^{-16}$ ***	515	355	6.47×10^{-8} ***
Higher d_N / d_S in Wil2 comparison	6382	6006	7.53×10^{-4} ***	465	404	0.0418 *
Higher d_N in Ws0 comparison	8335	4262	$< 2.2 \times 10^{-16}$ ***	481	394	0.0036 **
Higher d_S in Ws0 comparison	8709	3915	$< 2.2 \times 10^{-16}$ ***	518	357	5.83×10^{-8} ***
Higher d_N / d_S in Ws0 comparison	6424	6001	1.53×10^{-4} ***	467	407	0.0459 *
Higher d_N in Wu0 comparison	8344	4263	$< 2.2 \times 10^{-16}$ ***	479	419	0.0489 *
Higher d_S in Wu0 comparison	8722	3907	$< 2.2 \times 10^{-16}$ ***	531	367	4.93×10^{-8} ***
Higher d_N / d_S in Wu0 comparison	6429	5998	1.14×10^{-4} ***	470	427	0.1608
Higher d_N in Zu0 comparison	8317	4289	$< 2.2 \times 10^{-16}$ ***	483	402	0.0071 **
Higher d_S in Zu0 comparison	8751	3875	$< 2.2 \times 10^{-16}$ ***	530	355	4.42×10^{-9} ***
Higher d_N / d_S in Zu0 comparison	6387	6035	1.63×10^{-3} ***	469	415	0.0746

[a]*P*-values determined using a binomial test comparing the number of genes where d_N / d_S was higher in A. thaliana vs. the number of genes where d_N / d_S was higher in *A. lyrata*.

*, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$.

Table S1.5. Additional *A. thaliana* concatenome d_N / d_S analyses.

Strain-id	*C. rubella* d_N	*A. thaliana* d_N	*A. lyrata* d_N	*C. rubella* d_S	*A. thaliana* d_S	*A. lyrata* d_S	*C. rubella* d_N / d_S	*A. thaliana* d_N / d_S	*A. lyrata* d_N / d_S	*P*-value[a]
Bur0	0.02920	0.01269	0.01016	0.18192	0.07610	0.06214	0.1605	0.1667	0.1635	2.04×10^{-4} ***
Can0	0.02903	0.01259	0.01010	0.18191	0.07600	0.06203	0.1596	0.1657	0.1628	8.81×10^{-4} ***
Col0	0.02933	0.01268	0.01022	0.18189	0.07588	0.06216	0.1612	0.1671	0.1644	0.0014 **
Ct1	0.02912	0.01262	0.01014	0.18203	0.07606	0.06207	0.1600	0.1659	0.1633	0.0024 **
Edi0	0.02914	0.01264	0.01013	0.18190	0.07605	0.06207	0.1602	0.1662	0.1633	4.82×10^{-4} ***
Hi0	0.02915	0.01254	0.01017	0.18193	0.07561	0.06212	0.1602	0.1659	0.1637	0.0121 *
Kn0	0.02917	0.01268	0.01020	0.18178	0.07624	0.06221	0.1605	0.1663	0.1640	0.0077 **
Ler0	0.02916	0.01261	0.01013	0.18199	0.07601	0.06193	0.1603	0.1659	0.1636	0.0057 **
Mt0	0.02917	0.01266	0.01015	0.18184	0.07614	0.06212	0.1604	0.1663	0.1634	7.74×10^{-4} ***
No0	0.02915	0.01264	0.01017	0.18182	0.07602	0.06209	0.1603	0.1662	0.1637	0.0037 **
Oy0	0.02915	0.01261	0.01010	0.18185	0.07595	0.06205	0.1603	0.1661	0.1628	1.54×10^{-4} ***
Po0	0.02918	0.01237	0.01011	0.18184	0.07476	0.06202	0.1604	0.1654	0.1630	0.0063 **
Rsch4	0.02924	0.01265	0.01021	0.18207	0.07591	0.06221	0.1606	0.1666	0.1641	0.0052 **
Sf2	0.02919	0.01263	0.01013	0.18199	0.07590	0.06200	0.1604	0.1664	0.1633	3.37×10^{-4} ***
Tsu0	0.02916	0.01262	0.01014	0.18186	0.07600	0.06207	0.1604	0.1660	0.1633	0.0017 **
Wil2	0.02917	0.01260	0.01013	0.18191	0.07603	0.06200	0.1603	0.1658	0.1634	0.0058 **
Ws0	0.02920	0.01267	0.01019	0.18192	0.07613	0.06217	0.1605	0.1664	0.1639	0.0034 **
Wu0	0.02919	0.01266	0.01018	0.18188	0.07606	0.06217	0.1605	0.1664	0.1637	0.0017 **
Zu0	0.02920	0.01267	0.01018	0.18177	0.07615	0.06205	0.1606	0.1664	0.1641	0.0081 **

[a]*P*-value was determined comparing a 2-ratio model and a free-ratios model using the likelihood ratio test.

*, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$.

Table S1.6. Rates of evolution in *Arabidopsis thaliana* and *Arabidopsis halleri.*

	Arabidopsis thaliana	*Arabidopsis halleri*	Binomial *P*-value
Median d_N	0.0107	0.0080	-
Median d_S	0.0754	0.0591	-
Median d_N / d_S	0.1419	0.1342	-
Mean d_N	0.0130	0.0102	-
Mean d_S	0.0863	0.0703	-
Mean d_N / d_S	0.1837	0.189	-
Genes with higher d_S	8582	3441	4.94×10^{-394} ***
Genes with higher d_N	8270	3732	4.94×10^{-394} ***
Genes with higher d_N / d_S	6183	5642	6.81×10^{-7} ***

*, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$.

Table S1.7. Evolutionary rates of *Arabidopsis thaliana* and *Arabidopsis halleri* concatenomes.

Organism	d_N	d_S	d_N / d_S	*P*-value[a]
Capsella rubella	0.0289	0.1812	0.1597	-
Arabidopsis thaliana	0.0126	0.0757	0.1670	0.0002
Arabidopsis halleri	0.0098	0.0602	0.1633	

[a]*P*-value was determined comparing a 2-ratio model and a free-ratios model using the likelihood ratio test.

*, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$.

Table S1.8. Tajima's relative rate tests in *Arabidopsis thaliana* and *Arabidopsis halleri.*

	All substitutions	Synonymous substitutions	Nonsynonymous substitutions
Unique substitutions in *A. thaliana*	385337	232944	152393
Unique substitutions in *A. halleri*	310678	186707	123886
Genes where *A. thaliana* had more substitutions	8829	8342	7342
Genes where *A. halleri* had more substitutions	2645	2928	3433
Genes where $P < 0.05$	2123	1402	1375
Genes where $P < 0.05$ and *A. thaliana* had more substitution	1961	1282	1143
Genes where $P < 0.05$ and *A. halleri* had more substitutions	162	120	232
χ^2 value for concatenome	8008.4	5094.4	2941.4
P-value for concatenome	<< 0.001 ***	<< 0.001 ***	<< 0.001 ***

*, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$.

Table S1.9. Estimation of evolutionary rates in *A. thaliana, A. lyrata,* and *A. halleri* using *T. parvula* as outgroup.

	Arabidopsis thaliana	*Arabidopsis lyrata*	Binomial *P*-value
Number of genes with higher d_N	8760	3660	$< 2.2 \times 10^{-16}$ ***
Number of genes with higher d_S	8352	4050	$< 2.2 \times 10^{-16}$ ***
Number of genes with higher d_N / d_S	6236	5975	0.019 *
	Arabidopsis thaliana	*Arabidopsis halleri*	Binomial *P*-value
Number of genes with higher d_N	8455	3141	$< 2.2 \times 10^{-16}$ ***
Number of genes with higher d_S	8088	3489	$< 2.2 \times 10^{-16}$ ***
Number of genes with higher d_N / d_S	5841	5557	0.008 **

*, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$.

Table S1.10. Evolutionary rates of *T. parvula, A. lyrata, and A. halleri.*

Organism	d_N	d_S	d_N / d_S	*P*-value[a]
Thellungiellia parvula	0.0443	0.2557	0.1733	-
Arabidopsis thaliana	0.0120	0.0765	0.1575	0.0668
Arabidopsis lyrata	0.0095	0.0604	0.1565	
Thellungiellia parvula	0.0441	0.2556	0.1727	-
Arabidopsis thaliana	0.0121	0.0766	0.1582	0.1796
Arabidopsis halleri	0.0091	0.0583	0.1565	

[a]*P*-value was determined comparing a 2-ratio model and a free-ratios model using the likelihood ratio test.

*, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$.

Table S3.1. Frequency of codons used in *D. melanogaster* genes expressed in different tissues

Tissue name	Number of genes	Average GC3 (%)	Alanine GCU	Alanine GCC	Alanine GCA	Alanine GCG	Asparagine AAU	Asparagine AAC
Adult carcass	19	56.97	119	**206**	90	68	91	**115**
Brain	77	66.59	828	**2505**	1068	1225	1422	**1722**
Crop	22	63.22	175	**373**	138	152	233	**240**
Eyes	44	67.26	261	**618**	160	203	291	**397**
Fat body	23	58.94	140	**354**	134	124	271	**293**
Head	47	62.91	366	**925**	301	295	546	**558**
Heart	15	65.14	85	**258**	89	114	137	**142**
Hindgut	30	63.16	203	**425**	161	173	**309**	270
Male accessory glands	116	53.71	697	**1142**	591	421	**1392**	984
Midgut	133	65.52	1003	**1978**	684	581	1096	**1603**
Ovaries	84	61.21	661	**1331**	655	543	949	**1039**
Salivary glands	10	52.27	50	**87**	37	36	52	**59**
Testes	1364	60.79	7121	**15058**	6495	6279	**13399**	12052
Thoracicoabdominal ganglia	10	65.57	51	**156**	59	82	**107**	94
Tubules	28	62.05	185	**419**	126	148	256	**260**
Virgin spermatheca	24	56.26	97	**168**	96	61	**168**	147
All genes	13088	64.60	104207	**234116**	93459	99827	156904	**182094**

Tissue name	Arginine CGU	Arginine CGC	Arginine CGA	Arginine CGG	Arginine AGA	Arginine AGG	Aspartic acid GAU	Aspartic acid GAC
Adult carcass	30	**69**	34	32	21	25	**121**	117
Brain	537	**1218**	639	704	346	428	**1690**	1523
Crop	65	**160**	67	101	70	63	**239**	200
Eyes	132	**306**	102	124	55	110	**421**	342
Fat body	83	**170**	108	101	96	81	**315**	237
Head	176	**349**	228	213	122	177	**660**	509
Heart	55	**94**	67	42	38	54	**172**	148
Hindgut	97	**190**	119	97	55	95	**315**	249
Male accessory glands	312	**499**	441	355	422	387	**1398**	913
Midgut	393	**732**	334	330	273	320	**1631**	1408
Ovaries	341	**655**	403	344	324	327	**1452**	996
Salivary glands	45	**52**	31	18	27	25	**110**	47
Testes	5186	**8829**	5684	5437	4480	5284	**16585**	12202
Thoracicoabdominal ganglia	36	**74**	48	41	17	34	97	**108**
Tubules	71	**136**	109	82	83	95	**340**	197
Virgin spermatheca	47	**76**	51	52	40	54	**176**	128
All genes	62237	**124917**	63382	60805	40216	48768	**202480**	172600

For each amino acid and tissue, the most frequently used codon is marked in bold face.

Table S3.1 Cont.

Tissue name	Cysteine UGU	Cysteine UGC	Glutamic acid GAA	Glutamic acid GAG	Glutamine CAA	Glutamine CAG	Glycine GGA	Glycine GGG
Adult carcass	29	**58**	94	**141**	69	**111**	**164**	45
Brain	348	**900**	1076	**2678**	1128	**2982**	1323	369
Crop	61	**118**	169	**330**	124	**341**	281	52
Eyes	90	**237**	240	**703**	235	**699**	310	56
Fat body	85	**163**	277	**421**	147	**283**	206	48
Head	144	**313**	474	**999**	422	**849**	**775**	104
Heart	54	**109**	135	**272**	85	**283**	110	37
Hindgut	90	**185**	217	**434**	147	**397**	274	60
Male accessory glands	339	**563**	**1417**	1324	976	**1068**	**947**	248
Midgut	372	**920**	1140	**2152**	660	**1686**	1271	257
Ovaries	284	**653**	1056	**1714**	813	**1580**	874	219
Salivary glands	**18**	16	85	**108**	42	**79**	**108**	17
Testes	4244	**8035**	14733	**22474**	9092	**16643**	8801	2863
Thoracicoabdominal ganglia	24	**58**	66	**183**	52	**158**	80	21
Tubules	105	**234**	249	**373**	144	**285**	271	64
Virgin spermatheca	46	**96**	**153**	152	104	**160**	**199**	39
All genes	41563	**96489**	162092	**304644**	117055	**258636**	128408	34003

Tissue name	Glycine GGU	Glycine GGC	Histidine CAU	Histidine CAC	Isoleucine AUU	Isoleucine AUC	Isoleucine AUA	Leucine UUA
Adult carcass	75	122	43	**65**	113	**119**	77	36
Brain	959	**2194**	889	**1292**	917	**1424**	567	231
Crop	197	**285**	148	**197**	184	**256**	115	43
Eyes	228	**472**	280	**413**	212	**391**	87	40
Fat body	148	**262**	134	**187**	183	**233**	134	83
Head	398	673	220	**305**	539	**672**	291	128
Heart	83	**163**	105	**130**	93	**119**	36	19
Hindgut	209	**394**	153	**206**	196	**333**	146	50
Male accessory glands	555	794	**624**	610	**1059**	915	736	470
Midgut	804	**1448**	526	**916**	1066	**1566**	455	215
Ovaries	550	**973**	529	**739**	787	**950**	456	248
Salivary glands	78	80	28	**43**	40	**69**	30	24
Testes	6394	**10426**	5993	**7392**	10456	**12091**	6811	3677
Thoracicoabdominal ganglia	67	**124**	53	**71**	64	**108**	41	32
Tubules	169	**304**	100	**151**	267	**359**	167	62
Virgin spermatheca	106	179	66	**88**	151	**164**	109	75
All genes	93935	**182351**	77619	**113645**	122884	**161060**	72457	34520

For each amino acid and tissue, the most frequently used codon is marked in bold face.

Table S3.1 Cont.

Tissue name	Leucine UUG	Leucine CUU	Leucine CUC	Leucine CUA	Leucine CUG	Lysine AAA	Lysine AAG	Phenylalanine UUU
Adult carcass	91	47	67	34	**153**	98	**177**	93
Brain	1031	403	949	476	**2702**	819	**2014**	746
Crop	152	97	166	92	**418**	123	**375**	136
Eyes	224	116	199	88	**565**	183	**631**	140
Fat body	211	118	150	99	**397**	222	**382**	188
Head	473	262	423	242	**999**	332	**887**	538
Heart	111	51	116	56	**260**	89	**203**	84
Hindgut	214	90	208	107	**502**	145	**336**	173
Male accessory glands	965	693	612	540	**1285**	1374	**1553**	**992**
Midgut	1002	591	781	440	**2034**	688	**1787**	823
Ovaries	782	413	610	419	**1675**	756	**1510**	719
Salivary glands	54	38	37	26	**79**	59	**77**	60
Testes	10583	6175	7597	5271	**18414**	12529	**22311**	9820
Thoracicoabdominal ganglia	57	48	58	37	**199**	54	**194**	59
Tubules	250	166	207	123	**476**	193	**339**	251
Virgin spermatheca	135	77	91	56	**206**	117	**150**	127
All genes	121202	68183	100596	62830	**275746**	126766	**279280**	101381

Tissue name	Phenylalanine UUC	Proline CCU	Proline CCC	Proline CCA	Proline CCG	Serine UCU	Serine UCC	Serine UCA
Adult carcass	**130**	54	**107**	89	69	38	**119**	68
Brain	**1315**	381	1311	984	**1362**	340	1535	570
Crop	**220**	90	214	**181**	157	51	**243**	85
Eyes	**304**	131	**452**	257	263	85	343	78
Fat body	**241**	109	**272**	187	163	94	226	110
Head	**728**	191	**428**	296	384	131	**451**	166
Heart	**130**	69	**176**	112	159	41	129	47
Hindgut	**265**	84	**248**	185	196	86	**313**	100
Male accessory glands	956	471	665	**848**	561	419	695	568
Midgut	**1515**	473	**961**	620	631	446	**1318**	380
Ovaries	**971**	357	**839**	722	670	383	867	416
Salivary glands	**72**	26	47	**62**	33	38	**62**	27
Testes	**11860**	4300	**8832**	8147	7918	4357	**9826**	4794
Thoracicoabdominal ganglia	**104**	32	**106**	66	97	28	87	39
Tubules	**404**	61	**169**	119	96	118	**265**	75
Virgin spermatheca	**151**	73	101	**115**	80	76	107	65
All genes	**156360**	53267	**129444**	102619	114414	52658	141467	59411

For each amino acid and tissue, the most frequently used codon is marked in bold face.

Table S3.1 Cont.

Tissue name	Serine UCG	Serine AGU	Serine AGC	Threonine ACU	Threonine ACC	Threonine ACA	Threonine ACG	Tyrosine UAU
Adult carcass	95	57	93	76	**111**	61	59	125
Brain	1476	997	**1872**	532	**1559**	832	1238	736
Crop	197	151	226	104	**270**	138	162	141
Eyes	242	154	**356**	104	**317**	117	176	219
Fat body	175	162	**260**	151	**264**	128	189	139
Head	403	280	425	265	**583**	261	345	391
Heart	116	98	**145**	60	**194**	65	111	84
Hindgut	248	180	257	96	**274**	104	186	191
Male accessory glands	508	554	**698**	630	**685**	676	523	789
Midgut	717	654	1128	835	**1661**	688	678	803
Ovaries	631	535	**914**	591	**973**	780	641	579
Salivary glands	36	49	58	41	**61**	31	25	**48**
Testes	7782	6613	8906	6020	**10579**	5745	6380	7513
Thoracicoabdominal ganglia	93	52	**106**	50	**98**	46	68	54
Tubules	163	142	188	158	**306**	119	137	162
Virgin spermatheca	83	**137**	127	105	**139**	88	74	121
All genes	117862	87562	**146922**	73345	**153057**	84113	104110	82744

Tissue name	Tyrosine UAC	Valine GUU	Valine GUC	Valine GUA	Valine GUG	STOP UGA	STOP UAA	STOP UAG
Adult carcass	**132**	82	86	55	**170**	5	6	**8**
Brain	**1149**	638	917	329	**1847**	21	**34**	22
Crop	**245**	128	135	54	**315**	5	6	**11**
Eyes	**380**	168	265	66	**538**	10	**19**	15
Fat body	**196**	125	164	70	**293**	7	**9**	7
Head	**602**	322	357	145	**761**	9	**21**	17
Heart	**135**	74	85	42	**167**	4	3	**8**
Hindgut	**307**	144	197	71	**415**	10	**12**	8
Male accessory glands	**833**	625	578	405	**1037**	22	**59**	35
Midgut	**1526**	746	964	411	**1724**	21	**74**	38
Ovaries	**862**	543	559	309	**1237**	19	**39**	26
Salivary glands	46	42	43	24	**69**	1	4	**5**
Testes	**9695**	6646	7449	4066	**14148**	327	**560**	477
Thoracicoabdominal ganglia	**86**	47	53	32	**147**	3	**4**	3
Tubules	**304**	157	183	75	**393**	6	**11**	**11**
Virgin spermatheca	**139**	117	110	74	**173**	6	**16**	2
All genes	**133418**	81904	97974	47512	**200302**	3237	**5275**	4576

For each amino acid and tissue, the most frequently used codon is marked in bold face.

Table S3.2 Comparison of patterns of codon usage for genes expressed in different tissues vs. the entire genome

Tissue name	Alanine	Arginine	Asparagine	Aspartic acid	Cysteine	Glutamine	Glutamic acid	Glycine	Histidine
Adult carcass	**6.00E-03**	9.95E-01	5.92E-01	3.65E-01	5.90E-01	**4.57E-02**	1.04E-01	**3.49E-08**	9.49E-01
Brain	**3.35E-22**	**1.39E-07**	2.42E-01	1.20E-01	9.07E-02	**2.55E-07**	**5.08E-15**	**2.28E-06**	8.82E-01
Crop	6.78E-01	**3.09E-03**	2.10E-01	8.83E-01	2.82E-01	**4.10E-02**	7.20E-01	**1.07E-04**	4.11E-01
Eyes	**1.32E-06**	**1.01E-04**	**3.95E-02**	5.29E-01	3.36E-01	**8.54E-05**	**2.55E-09**	**1.45E-02**	9.54E-01
Fat body	2.53E-01	**2.63E-04**	4.23E-01	1.57E-01	1.73E-01	1.92E-01	**6.72E-03**	5.98E-01	7.13E-01
Head	**3.08E-05**	**4.07E-03**	**3.66E-02**	9.39E-02	5.47E-01	1.23E-01	**4.20E-02**	**3.22E-24**	5.66E-01
Heart	5.44E-02	7.05E-02	3.75E-01	9.80E-01	4.50E-01	**1.02E-03**	5.42E-01	6.51E-01	2.25E-01
Hindgut	6.10E-01	1.42E-01	**7.20E-04**	3.95E-01	3.78E-01	**4.14E-02**	4.78E-01	4.55E-01	4.64E-01
Male accessory glands	**2.24E-18**	**3.51E-56**	**1.88E-33**	**3.32E-10**	**1.11E-06**	**4.04E-59**	**6.07E-78**	**5.79E-29**	**9.56E-13**
Midgut	**7.91E-23**	**4.16E-03**	**3.34E-09**	7.46E-01	3.14E-01	**1.61E-03**	9.16E-01	**1.84E-08**	**1.58E-03**
Ovaries	**5.66E-06**	**1.31E-09**	1.99E-01	**1.19E-07**	9.23E-01	**3.09E-03**	**1.85E-04**	**2.33E-06**	4.25E-01
Salivary glands	4.72E-01	**8.24E-03**	9.80E-01	**7.35E-05**	**6.62E-03**	4.56E-01	**8.26E-03**	**1.05E-05**	9.40E-01
Testes	**1.48E-11**	**5.16E-218**	**1.61E-99**	**6.31E-38**	**2.15E-29**	**1.47E-50**	**9.20E-94**	**7.57E-93**	**1.45E-24**
Thoracicoabdominal ganglia	**3.32E-02**	3.59E-01	**5.65E-02**	6.52E-02	9.63E-01	5.38E-02	**7.80E-03**	8.38E-01	6.90E-01
Tubules	**1.39E-02**	**3.64E-07**	1.40E-01	**1.70E-05**	7.74E-01	3.06E-01	**6.22E-03**	**4.26E-02**	8.61E-01
Virgin spermatheca	**2.01E-03**	**1.69E-02**	**1.41E-02**	1.89E-01	6.16E-01	**4.76E-03**	**2.12E-08**	**1.09E-04**	6.22E-01

Tissue name	Isoleucine	Leucine	Lysine	Phenylalanine	Proline	Serine	Threonine	Tyrosine	Valine
Adult carcass	**3.60E-02**	**8.92E-03**	1.30E-01	5.13E-01	**2.55E-02**	**4.19E-03**	**4.67E-03**	**7.94E-04**	1.87E-01
Brain	**1.48E-04**	**9.87E-27**	**8.21E-03**	**3.62E-03**	**1.37E-18**	**1.84E-35**	**3.28E-20**	5.06E-01	**6.09E-08**
Crop	8.05E-01	1.70E-01	**1.97E-03**	7.00E-01	1.17E-01	**4.42E-03**	2.52E-01	5.12E-01	1.10E-01
Eyes	**4.04E-10**	**2.92E-04**	**8.90E-08**	**8.94E-04**	**2.49E-08**	**3.28E-08**	**1.50E-04**	4.10E-01	**1.45E-07**
Fat body	6.05E-02	**6.67E-04**	**3.81E-03**	6.38E-02	**7.46E-04**	2.44E-01	8.19E-02	2.48E-01	5.69E-01
Head	4.47E-01	2.15E-01	**2.92E-03**	**2.28E-02**	8.54E-02	**1.18E-02**	**2.55E-02**	4.96E-01	6.57E-02
Heart	7.38E-02	**1.98E-02**	8.34E-01	1.00E+00	2.11E-01	2.67E-01	**6.98E-04**	1.00E+00	9.45E-01
Hindgut	**1.15E-02**	**8.51E-03**	6.46E-01	9.85E-01	4.08E-01	**1.09E-02**	**8.20E-04**	1.00E+00	**3.92E-02**
Male accessory glands	**4.35E-35**	**1.06E-125**	**1.26E-75**	**1.02E-25**	**1.14E-38**	**1.90E-60**	**3.11E-46**	**8.28E-18**	**1.72E-22**
Midgut	**1.02E-15**	**9.50E-06**	**2.49E-04**	**4.26E-05**	**5.59E-17**	**1.75E-21**	**2.04E-37**	**1.62E-04**	**7.52E-03**
Ovaries	2.01E-01	8.33E-02	**2.89E-02**	**7.32E-03**	**6.59E-03**	**1.48E-05**	**3.54E-19**	1.43E-01	7.02E-02
Salivary glands	3.63E-01	**5.39E-04**	**2.99E-03**	1.77E-01	**2.21E-03**	**3.45E-03**	**8.57E-03**	**1.45E-02**	1.49E-01
Testes	**4.21E-56**	**1.08E-258**	**1.04E-88**	**1.80E-78**	**5.79E-38**	**6.32E-107**	**1.09E-62**	**9.88E-52**	**1.50E-37**
Thoracicoabdominal ganglia	2.50E-01	**1.70E-02**	**1.67E-03**	4.59E-01	1.44E-01	3.05E-01	7.18E-01	1.00E+00	1.93E-01
Tubules	8.36E-01	**4.19E-03**	**1.34E-02**	6.21E-01	**6.45E-03**	**4.37E-07**	**2.09E-06**	1.30E-01	3.83E-01
Virgin spermatheca	**6.17E-03**	**5.84E-14**	**1.20E-05**	**3.53E-02**	**1.57E-05**	**7.11E-12**	**2.30E-05**	**7.40E-03**	**5.71E-06**

P-values correspond to the Chi-squared test. *P*-values < 0.05 are marked with bold face.

Table S3.3. Preferred and unpreferred codons in midgut-specific genes of *D. melanogaster*

Amino acid	Codon	High expression (average RSCU)	Low expression (average RSCU)	*P* value	Amino acid	Codon	High expression (average RSCU)	Low expression (average RSCU)	*P* value
Phe	UUU	0.460	0.859	**6.11E-03**	Ser	UCU	0.438	0.556	1.00E+00
	UUC*	1.540	1.141	**5.56E-03**		**UCC***	2.473	1.672	**9.23E-03**
Leu	UUA	0.159	0.388	**3.08E-03**		UCA	0.195	0.468	**2.06E-02**
	UUG	1.118	1.073	9.88E-01		UCG	0.732	0.937	9.53E-02
Leu	CUU	0.420	0.673	**3.38E-02**	Pro	CCU	0.696	0.586	5.14E-01
	CUC	0.895	1.144	1.12E-01		**CCC***	2.073	1.432	**1.75E-02**
	CUA	0.412	0.687	**1.95E-02**		CCA	0.721	1.011	1.30E-01
	CUG*	2.997	2.036	**1.10E-02**		CCG	0.510	0.972	**7.00E-03**
Ile	AUU	0.993	1.134	3.72E-01	Thr	ACU	0.625	1.010	**1.36E-02**
	AUC*	1.855	1.329	**2.23E-02**		**ACC***	2.834	1.331	**2.25E-05**
	AUA	0.152	0.537	**1.52E-04**		ACA	0.249	0.614	**7.20E-03**
Met	AUG	-	-	-		ACG	0.292	1.045	**3.06E-05**
Val	GUU	0.780	0.744	7.28E-01	Ala	GCU	1.052	0.851	2.89E-01
	GUC*	1.248	0.796	**1.30E-02**		GCC	2.262	1.903	9.90E-02
	GUA	0.266	0.499	6.07E-02		GCA	0.383	0.708	**6.71E-04**
	GUG	1.705	1.961	1.30E-01		GCG	0.303	0.538	5.02E-02
Tyr	UAU	0.531	0.779	**2.22E-02**	Cys	UGU	0.466	0.725	1.23E-01
	UAC*	1.469	1.221	**2.13E-02**		UGC	1.534	1.275	1.15E-01
STOP	UAA	-	-	-	STOP	UGA	-	-	-
STOP	UAG	-	-	-	Trp	UGG	-	-	-
His	CAU	0.608	0.718	2.43E-01	Arg	CGU	1.473	0.939	6.45E-02
	CAC	1.392	1.282	2.25E-01		**CGC***	3.069	1.350	**8.19E-04**
Gln	CAA	0.384	0.660	8.52E-02		CGA	0.243	0.982	**6.91E-06**
	CAG	1.616	1.340	8.24E-02		CGG	0.305	0.726	**2.53E-03**
Asn	AAU	0.506	1.029	**2.82E-03**	Ser	AGU	0.628	0.878	1.02E-01
	AAC*	1.494	0.971	**3.11E-03**		AGC	1.534	1.489	8.80E-01
Lys	AAA	0.352	0.673	**4.98E-03**	Arg	**AGA***	0.468	1.151	**2.01E-03**
	AAG*	1.648	1.327	**4.98E-03**		AGG	0.443	0.851	**3.79E-02**
Asp	GAU	1.004	1.082	6.50E-01	Gly	GGU	0.983	0.826	4.67E-01
	GAC	0.996	0.918	6.07E-01		GGC	1.666	1.401	1.64E-01
Glu	GAA	0.426	0.817	**3.39E-03**		GGA	1.235	1.435	4.31E-01
	GAG*	1.574	1.183	**3.74E-03**		**GGG***	0.116	0.338	**8.89E-03**

Preferred codons (those for which RSCU is significantly higher for highly expressed genes) are marked with an asterisk and in bold face. *P*-values correspond to the Mann-Whitney *U* test

Table S3.4. Preferred and unpreferred codons in testes-specific genes of *D. melanogaster*

Amino acid	Codon	High expression (average RSCU)	Low expression (average RSCU)	*P* value	Amino acid	Codon	High expression (average RSCU)	Low expression (average RSCU)	*P* value
Phe	UUU	0.817	0.854	8.09E-01	Ser	UCU	0.679	0.572	4.81E-01
	UUC	1.183	1.146	8.24E-01		UCC	1.631	1.371	1.90E-01
Leu	UUA	0.192	0.424	**7.10E-03**		UCA	0.416	0.590	1.16E-01
	UUG	0.971	1.221	1.74E-01		UCG	1.149	1.141	7.13E-01
Leu	CUU	1.114	0.687	9.95E-01	Pro	CCU	0.676	0.550	1.83E-01
	CUC	0.785	0.851	5.87E-01		**CCC***	1.716	1.295	**1.67E-04**
	CUA	0.887	0.645	9.29E-01		CCA	0.743	1.085	**1.19E-02**
	CUG	2.051	2.171	9.92E-01		CCG	0.865	1.070	1.02E-01
Ile	AUU	0.924	0.994	2.89E-01	Thr	ACU	1.101	0.790	6.25E-01
	AUC	1.457	1.300	3.08E-01		ACC	1.579	1.510	4.29E-01
	AUA	0.619	0.706	3.29E-01		ACA	0.734	0.781	**3.06E-02**
Met	AUG	-	-	-		ACG	0.587	0.918	**2.17E-02**
Val	GUU	0.700	0.804	5.00E-01	Ala	GCU	0.687	0.818	1.29E-01
	GUC	1.102	0.914	4.82E-01		GCC	1.926	1.707	2.48E-01
	GUA	0.491	0.469	9.63E-01		GCA	0.800	0.754	3.87E-01
	GUG	1.708	1.812	1.00E+00		GCG	0.587	0.721	8.78E-02
Tyr	UAU	0.839	0.862	4.50E-01	Cys	UGU	0.636	0.621	8.39E-01
	UAC	1.161	1.138	4.70E-01		UGC	1.364	1.379	7.85E-01
STOP	UAA	-	-	-	STOP	UGA	-	-	-
STOP	UAG	-	-	-	Trp	UGG	-	-	-
His	CAU	0.621	0.904	**2.71E-02**	Arg	**CGU***	1.647	0.826	**4.73E-03**
	CAC	1.379	1.096	**2.91E-02**		CGC	1.357	1.454	7.92E-01
Gln	CAA	0.509	0.727	**9.02E-03**		CGA	1.112	1.015	6.05E-01
	CAG*	1.491	1.273	**8.76E-03**		CGG	1.009	0.897	8.61E-01
Asn	AAU	0.983	1.000	7.46E-01	Ser	AGU	0.835	1.011	1.05E-01
	AAC	1.017	1.000	7.37E-01		AGC	1.290	1.314	6.54E-01
Lys	AAA	0.572	0.764	**1.20E-02**	Arg	AGA	0.388	0.859	**3.87E-03**
	AAG*	1.428	1.236	**1.21E-02**		AGG	0.486	0.949	**2.29E-03**
Asp	GAU	1.049	1.152	1.42E-01	Gly	GGU	0.778	0.784	8.01E-01
	GAC	0.951	0.848	1.55E-01		GGC	1.706	1.463	4.19E-01
Glu	GAA	0.944	0.829	7.44E-01		GGA	1.318	1.346	6.83E-01
	GAG	1.056	1.171	7.47E-01		GGG	0.198	0.407	**3.21E-03**

Preferred codons (those for which RSCU is significantly higher for highly expressed genes) are marked with an asterisk and in bold face. *P*-values correspond to the Mann-Whitney *U* test

Table S3.5. Preferred and unpreferred codons in male accessory glands-specific genes of *D. melanogaster*

Amino acid	Codon	High expression (average RSCU)	Low expression (average RSCU)	*P* value	Amino acid	Codon	High expression (average RSCU)	Low expression (average RSCU)	*P* value
Phe	UUU	0.000	0.997	1.32E-01	Ser	UCU	0.947	0.513	4.16E-01
	UUC	2.000	1.003	1.32E-01		UCC	1.579	1.170	4.00E-01
Leu	UUA	0.621	0.479	5.33E-01		UCA	0.000	0.791	1.32E-01
	UUG	1.034	1.337	5.62E-01		UCG	1.895	1.047	2.67E-01
Leu	CUU	1.034	0.714	2.67E-01	Pro	CCU	1.200	0.443	1.29E-01
	CUC	0.828	0.711	5.62E-01		CCC	1.600	1.243	2.97E-01
	CUA	0.414	0.640	8.00E-01		CCA	0.400	1.436	2.03E-01
	CUG	2.069	2.119	9.33E-01		CCG	0.800	0.878	1.00E+00
Ile	AUU	1.227	1.107	7.28E-01	Thr	ACU	1.333	0.675	2.67E-01
	AUC	0.955	1.237	6.67E-01		ACC	1.333	1.614	8.00E-01
	AUA	0.818	0.657	7.28E-01		ACA	0.667	1.000	2.47E-01
Met	AUG	-	-	-		ACG	0.667	0.712	1.00E+00
Val	GUU	1.091	0.810	6.67E-01	Ala	GCU	0.667	0.912	7.28E-01
	GUC	0.909	0.948	1.00E+00		GCC	2.167	1.751	6.67E-01
	GUA	0.364	0.496	1.00E+00		GCA	0.667	0.772	9.33E-01
	GUG	1.636	1.746	9.33E-01		GCG	0.500	0.565	8.17E-01
Tyr	UAU	0.800	0.807	9.08E-01	Cys	UGU	0.667	0.812	1.00E+00
	UAC	1.200	1.193	9.08E-01		UGC	1.333	1.188	1.00E+00
STOP	UAA	-	-	-	STOP	UGA	-	-	-
STOP	UAG	-	-	-	Trp	UGG	-	-	-
His	CAU	1.000	0.946	7.28E-01	Arg	CGU	0.545	0.622	9.08E-01
	CAC	1.000	1.054	8.00E-01		CGC	0.000	1.238	1.33E-01
Gln	CAA	1.500	0.789	1.31E-01		CGA	0.545	1.208	2.47E-01
	CAG	0.500	1.211	1.32E-01		CGG	1.091	0.986	8.17E-01
Asn	AAU	1.400	1.162	6.67E-01	Ser	AGU	1.263	1.027	6.67E-01
	AAC	0.600	0.838	6.67E-01		AGC	0.316	1.453	2.67E-01
Lys	AAA	1.091	0.693	4.00E-01	Arg	AGA	3.273	0.816	1.32E-01
	AAG	0.909	1.307	4.00E-01		AGG	0.545	1.129	1.33E-01
Asp	GAU	1.250	1.233	1.00E+00	Gly	GGU	0.762	0.969	4.17E-01
	GAC	0.750	0.767	1.00E+00		GGC	1.714	1.378	2.97E-01
Glu	GAA	0.600	1.105	4.00E-01		GGA	1.333	1.348	1.00E+00
	GAG	1.400	0.895	4.00E-01		GGG	0.190	0.305	4.17E-01

Preferred codons (those for which RSCU is significantly higher for highly expressed genes) are marked with an asterisk and in bold face. *P*-values correspond to the Mann-Whitney *U* test

Table S3.6. Frequency of codons used in random sets of D. melanogaster genes with similar GC3

Tissue name	Number of genes	Average GC3 (%)	Alanine GCU	Alanine GCC	Alanine GCA	Alanine GCG	Arginine CGU	Arginine CGC
Adult carcass	19	57.08	152	**217**	172	83	104	**143**
Brain	77	66.52	516	**1232**	392	528	289	**642**
Crop	22	63.14	210	**495**	150	182	133	**236**
Eyes	44	67.08	286	**970**	295	421	276	**553**
Fat body	23	58.90	118	**245**	112	114	87	**160**
Head	47	63.02	333	**717**	297	310	192	**385**
Heart	15	65.30	170	**311**	140	131	99	**159**
Hindgut	30	63.31	300	**690**	263	290	174	**327**
Male accessory glands	116	53.63	1421	**1942**	1278	856	708	**1037**
Midgut	133	65.56	907	**2416**	852	1004	637	**1193**
Ovaries	84	61.23	855	**1556**	802	662	425	**767**
Salivary glands	10	52.38	66	**71**	42	31	32	**43**
Testes	1364	61.18	12680	**24221**	11005	10567	7170	**12709**
Thoracicoabdominal ganglia	10	65.34	83	**199**	67	76	49	**112**
Tubules	28	62.05	250	**560**	233	256	168	**334**
Virgin spermatheca	24	56.19	170	**307**	137	151	100	**156**
All genes	2046	61.37	18517	**36149**	16237	15662	10643	**18956**

Tissue name	Arginine CGA	Arginine CGG	Arginine AGA	Arginine AGG	Asparagine AAU	Asparagine AAC	Aspartic acid GAU	Aspartic acid GAC
Adult carcass	92	61	88	70	**234**	205	**316**	189
Brain	313	294	158	243	655	**826**	**959**	782
Crop	111	84	71	98	**320**	317	**458**	372
Eyes	248	202	139	174	572	**716**	**781**	760
Fat body	91	83	59	70	**304**	223	**306**	204
Head	200	168	138	156	507	**540**	**650**	629
Heart	90	67	30	70	178	**242**	**317**	225
Hindgut	189	168	96	158	410	**526**	**614**	542
Male accessory glands	876	526	681	595	**2332**	1786	**2613**	1625
Midgut	545	513	310	393	1525	**1787**	**2005**	1711
Ovaries	409	388	327	346	1288	**1352**	**1410**	1092
Salivary glands	42	27	39	28	**126**	88	**112**	72
Testes	6968	6140	4824	5447	18166	**19450**	**23594**	17992
Thoracicoabdominal ganglia	40	41	31	29	114	**138**	**151**	116
Tubules	193	169	151	169	438	**498**	**581**	427
Virgin spermatheca	100	78	67	74	244	**247**	**324**	217
All genes	10507	9009	7209	8120	27413	**28941**	**35191**	26955

For each amino acid and tissue, the most frequently used codon is marked in bold face.

Table S3.6. Cont.

Tissue name	Cysteine UGU	Cysteine UGC	Glutamine CAA	Glutamine CAG	Glutamic acid GAA	Glutamic acid GAG	Glycine GGU	Glycine GGC
Adult carcass	78	**102**	174	**280**	285	**416**	128	**215**
Brain	264	**687**	501	**1292**	742	**1452**	477	**872**
Crop	95	**229**	168	**435**	278	**620**	189	**387**
Eyes	202	**524**	349	**1008**	545	**1332**	334	**755**
Fat body	90	**157**	152	**201**	260	**318**	139	**227**
Head	148	**304**	352	**658**	583	**959**	302	**549**
Heart	50	**120**	162	**402**	167	**391**	127	**279**
Hindgut	129	**239**	318	**710**	471	**987**	252	**541**
Male accessory glands	552	**827**	1781	**2228**	2705	**2594**	1175	1522
Midgut	546	**1217**	1003	**2330**	1546	**2868**	954	**1900**
Ovaries	250	**549**	1103	**1930**	1324	**2032**	688	**1192**
Salivary glands	17	**32**	72	**117**	**86**	74	59	88
Testes	4809	**9852**	14257	**27673**	20115	**31906**	10952	**19063**
Thoracicoabdominal ganglia	31	**57**	64	**228**	98	**264**	61	**146**
Tubules	107	**196**	376	**720**	467	**826**	248	**436**
Virgin spermatheca	52	**119**	198	**375**	235	**353**	216	**300**
All genes	7420	**15211**	21030	**40587**	29907	**47392**	16301	**28472**

Tissue name	Glycine GGA	Glycine GGG	Histidine CAU	Histidine CAC	Isoleucine AUU	Isoleucine AUC	Isoleucine AUA	Leucine UUA
Adult carcass	178	43	94	**120**	**195**	184	132	73
Brain	627	153	332	**455**	513	**789**	317	99
Crop	243	60	173	**283**	261	**393**	124	56
Eyes	411	121	284	**474**	457	**682**	241	97
Fat body	153	65	97	**119**	224	**272**	134	52
Head	353	95	268	**316**	383	**518**	302	158
Heart	194	33	110	**134**	142	**226**	84	31
Hindgut	357	112	246	**351**	381	**505**	225	95
Male accessory glands	**1717**	390	930	**941**	**1684**	1309	1183	876
Midgut	1240	260	740	**1185**	1145	**1617**	664	307
Ovaries	987	243	555	**769**	**957**	953	552	341
Salivary glands	**144**	18	38	**41**	69	53	46	28
Testes	15194	3591	8256	**11065**	13736	**16205**	8248	4089
Thoracicoabdominal ganglia	100	21	56	**94**	79	**135**	50	14
Tubules	339	96	235	**287**	330	**382**	210	93
Virgin spermatheca	226	40	132	**169**	**217**	191	130	70
All genes	22463	5341	12546	**16803**	20773	**24414**	12642	6479

For each amino acid and tissue, the most frequently used codon is marked in bold face.

Table S3.6. Cont.

Tissue name	Leucine UUG	Leucine CUU	Leucine CUC	Leucine CUA	Leucine CUG	Lysine AAA	Lysine AAG	Phenylalanine UUU
Adult carcass	224	136	107	101	**252**	300	**396**	157
Brain	560	266	445	259	**1298**	495	**1392**	397
Crop	264	155	225	123	**614**	204	**555**	195
Eyes	471	206	391	251	**1184**	372	**1098**	370
Fat body	184	95	134	117	**334**	182	**327**	199
Head	362	231	312	228	**836**	469	**989**	392
Heart	175	109	139	93	**429**	142	**321**	136
Hindgut	331	205	275	171	**795**	365	**843**	334
Male accessory glands	1445	1062	823	893	**1900**	2175	**2642**	**1464**
Midgut	1204	656	985	632	**2872**	1069	**2652**	979
Ovaries	828	494	570	502	**1580**	1074	**1729**	698
Salivary glands	57	42	41	35	**84**	82	**101**	66
Testes	13591	7944	10249	7192	**27253**	15129	**29741**	11310
Thoracicoabdominal ganglia	86	52	85	26	**221**	60	**196**	82
Tubules	335	201	263	171	**723**	377	**722**	288
Virgin spermatheca	206	112	159	118	**381**	162	**263**	211
All genes	20323	11966	15203	10912	**40756**	22657	**43967**	17278

Tissue name	Phenylalanine UUC	Proline CCU	Proline CCC	Proline CCA	Proline CCG	Serine UCU	Serine UCC	Serine UCA
Adult carcass	**182**	103	**159**	179	128	95	176	96
Brain	**761**	211	**606**	434	561	178	**675**	242
Crop	**367**	108	**271**	158	230	96	**297**	89
Eyes	**698**	144	497	330	**502**	122	526	166
Fat body	**252**	68	**145**	149	142	53	**171**	102
Head	**575**	165	**404**	282	392	179	**399**	193
Heart	**221**	76	**177**	141	179	79	**220**	68
Hindgut	**497**	156	**395**	283	331	132	**372**	156
Male accessory glands	1286	918	1269	**1552**	1101	936	**1424**	1032
Midgut	**1664**	415	**1174**	906	1050	466	1400	531
Ovaries	**929**	461	**936**	881	766	517	999	489
Salivary glands	**104**	33	58	**59**	46	38	50	34
Testes	**15905**	6534	**14134**	12599	11814	7017	15564	7268
Thoracicoabdominal ganglia	**122**	29	**96**	97	93	37	**116**	34
Tubules	**420**	170	**355**	333	267	161	**375**	160
Virgin spermatheca	**231**	84	213	**167**	163	70	176	90
All genes	**24214**	9675	**20889**	18550	17765	10176	22940	10750

For each amino acid and tissue, the most frequently used codon is marked in bold face.

Table S3.6. Cont.

Tissue name	Serine UCG	Serine AGU	Serine AGC	Threonine ACU	Threonine ACC	Threonine ACA	Threonine ACG
Adult carcass	**178**	140	173	123	**186**	138	118
Brain	573	354	606	300	**739**	341	536
Crop	214	156	258	129	**317**	139	191
Eyes	443	265	**590**	220	**609**	258	452
Fat body	136	126	**171**	104	**183**	134	125
Head	350	266	394	197	**482**	227	333
Heart	172	118	190	88	**202**	115	164
Hindgut	292	230	353	197	**452**	191	287
Male accessory glands	1170	1256	1407	1114	**1399**	1326	964
Midgut	1117	780	**1497**	646	**1455**	692	969
Ovaries	845	672	**1064**	616	**1037**	647	658
Salivary glands	47	50	**57**	48	**60**	36	39
Testes	12439	10182	**15571**	9076	**16377**	10350	11115
Thoracicoabdominal ganglia	99	71	107	43	**127**	60	76
Tubules	292	253	326	169	**347**	209	265
Virgin spermatheca	139	128	**207**	129	**207**	125	121
All genes	18506	15047	**22971**	13199	**24179**	14988	16413

Tissue name	Tyrosine UAU	Tyrosine UAC	Valine GUU	Valine GUC	Valine GUA	Valine GUG
Adult carcass	131	**142**	108	134	79	**213**
Brain	346	**633**	388	485	188	**996**
Crop	166	**279**	151	205	89	**442**
Eyes	312	**604**	238	369	145	**923**
Fat body	147	**199**	145	161	71	**273**
Head	314	**440**	256	299	156	**623**
Heart	115	**169**	112	131	64	**269**
Hindgut	225	**422**	224	316	134	**635**
Male accessory glands	1164	**1171**	1318	883	779	**1632**
Midgut	809	**1401**	743	950	387	**1944**
Ovaries	562	**803**	676	626	409	**1157**
Salivary glands	51	**55**	35	47	31	**78**
Testes	9436	**13758**	9773	10354	5680	**20954**
Thoracicoabdominal ganglia	70	**126**	59	80	24	**134**
Tubules	255	**360**	196	231	133	**499**
Virgin spermatheca	208	**201**	154	146	81	**271**
All genes	14311	**20763**	14576	15417	8450	**31043**

For each amino acid and tissue, the most frequently used codon is marked in bold face.

Table S3.7. Comparison of patterns of codon usage for genes expressed in different tissues in the real vs. randomized datasets controlling for GC3

Tissue name	Alanine	Arginine	Asparagine	Aspartic acid	Cysteine	Glutamine	Glutamic acid	Glycine	Histidine
Adult carcass	**3.43E-14**	2.60E-01	2.31E-01	7.12E-02	**6.44E-04**	**2.48E-16**	**1.04E-13**	**7.39E-29**	5.82E-01
Brain	**6.74E-07**	1.08E-01	5.57E-02	5.97E-01	4.24E-01	8.84E-01	**4.74E-02**	**1.76E-02**	2.89E-01
Crop	1.69E-01	**7.41E-09**	**3.52E-06**	4.49E-01	**1.70E-04**	4.35E-01	**1.59E-02**	**2.45E-10**	**2.74E-02**
Eyes	**1.28E-08**	**8.82E-05**	9.81E-01	1.96E-01	5.35E-01	**4.27E-02**	**1.29E-06**	**6.98E-06**	3.85E-01
Fat body	**3.38E-03**	**3.12E-14**	**1.33E-04**	**7.54E-03**	**1.18E-04**	**3.66E-07**	**6.89E-13**	6.28E-02	1.12E-01
Head	**3.49E-05**	**1.35E-04**	**1.84E-06**	**2.59E-02**	**1.07E-02**	**1.47E-05**	3.22E-01	**2.10E-20**	9.41E-02
Heart	**7.34E-04**	**4.19E-13**	**5.82E-06**	8.00E-01	**9.10E-04**	**1.62E-04**	6.78E-02	4.29E-01	**1.61E-03**
Hindgut	9.08E-02	**4.66E-04**	**1.33E-13**	7.44E-02	**1.76E-03**	6.24E-01	**4.95E-02**	**1.79E-02**	**3.98E-02**
Male accessory glands	**1.49E-16**	**1.49E-38**	**9.00E-28**	**2.05E-07**	**8.12E-08**	**7.96E-51**	**4.14E-62**	**1.50E-17**	**3.92E-10**
Midgut	**2.14E-12**	**4.00E-04**	2.37E-01	8.46E-01	2.13E-01	7.27E-01	**2.31E-03**	**2.28E-04**	2.35E-01
Ovaries	**1.82E-06**	**5.36E-10**	**3.06E-04**	**1.54E-05**	**4.77E-02**	**8.47E-07**	**3.26E-09**	**9.35E-04**	1.15E-01
Salivary glands	**6.66E-07**	**3.40E-32**	**2.62E-03**	**1.74E-35**	**1.76E-33**	**4.15E-08**	**1.67E-26**	**1.64E-34**	7.29E-01
Testes	**1.85E-02**	**3.40E-11**	**4.58E-12**	**2.12E-03**	**6.77E-05**	**2.64E-09**	**1.15E-12**	**5.62E-04**	**1.36E-03**
Thoracicoabdominal ganglia	**1.14E-07**	**3.69E-05**	**2.61E-13**	**1.31E-05**	1.40E-01	**1.83E-02**	**9.96E-05**	**4.42E-02**	**3.39E-02**
Tubules	**1.94E-04**	**1.83E-24**	**1.09E-06**	**3.03E-13**	**2.15E-02**	**3.98E-06**	**8.58E-14**	**1.16E-03**	5.72E-01
Virgin spermatheca	**2.99E-19**	**2.75E-15**	**1.58E-13**	**1.04E-03**	**2.96E-03**	**2.76E-19**	**8.41E-54**	**2.31E-13**	**2.90E-02**

Tissue name	Isoleucine	Leucine	Lysine	Phenylalanine	Proline	Serine	Threonine	Tyrosine	Valine
Adult carcass	**1.09E-10**	**6.76E-33**	**3.94E-14**	**9.39E-04**	**6.47E-22**	**2.53E-19**	**2.22E-16**	**1.43E-12**	**8.26E-11**
Brain	9.98E-02	**2.74E-02**	**1.84E-02**	9.71E-01	7.08E-02	**1.75E-04**	**1.46E-05**	8.57E-02	1.78E-01
Crop	1.42E-01	**2.12E-02**	3.04E-01	2.51E-01	**4.35E-11**	**2.28E-05**	1.71E-01	7.83E-01	**1.59E-02**
Eyes	**1.98E-12**	**1.93E-02**	**3.50E-03**	**2.58E-03**	**6.72E-12**	**5.49E-08**	**1.32E-02**	7.69E-01	**3.49E-06**
Fat body	**2.02E-06**	**2.93E-25**	**4.46E-17**	**4.24E-06**	**1.30E-15**	**3.20E-09**	**1.71E-04**	**1.91E-03**	**2.00E-04**
Head	2.55E-01	**2.72E-06**	3.14E-01	**1.49E-04**	**1.04E-03**	**4.53E-02**	2.25E-01	5.66E-02	**1.64E-02**
Heart	**5.10E-04**	**7.16E-03**	**3.15E-04**	7.18E-02	6.06E-02	**6.17E-04**	**1.86E-04**	1.85E-01	**2.12E-04**
Hindgut	**1.02E-04**	7.87E-02	**8.38E-04**	5.09E-02	**1.69E-03**	**1.14E-04**	**8.31E-03**	1.86E-01	3.04E-01
Male accessory glands	**1.09E-22**	**2.94E-94**	**4.79E-56**	**7.20E-19**	**8.58E-39**	**4.34E-50**	**8.85E-37**	**1.40E-12**	**2.12E-23**
Midgut	**1.07E-03**	**7.69E-08**	1.43E-01	5.07E-01	**5.44E-17**	**9.17E-15**	**6.70E-13**	4.01E-01	**1.42E-04**
Ovaries	**1.93E-02**	**4.13E-09**	**5.25E-09**	**1.33E-04**	**1.99E-08**	**5.04E-11**	**2.42E-16**	**1.81E-02**	**2.34E-04**
Salivary glands	**5.20E-05**	**1.54E-68**	**5.28E-40**	**2.30E-08**	**2.23E-44**	**2.22E-43**	**1.45E-25**	**3.90E-17**	**1.50E-20**
Testes	**9.37E-06**	**3.35E-28**	**5.86E-15**	**3.99E-08**	**3.82E-09**	**1.68E-13**	**5.63E-06**	**1.56E-05**	**1.54E-07**
Thoracicoabdominal ganglia	**4.42E-03**	**9.96E-22**	**4.06E-04**	9.71E-01	**4.68E-01**	1.46E-01	**2.84E-02**	1.48E-01	**4.44E-03**
Tubules	7.64E-02	**1.13E-16**	**8.69E-16**	2.21E-01	**4.29E-17**	**1.16E-24**	**9.36E-11**	5.01E-01	1.32E-01
Virgin spermatheca	**1.82E-11**	**2.07E-69**	**7.71E-42**	**1.03E-08**	**1.80E-37**	**1.21E-58**	**2.06E-23**	**3.21E-09**	**1.45E-31**

P-values correspond to the Chi-squared test. *P*-values < 0.05 are marked with bold face.

Table S3.8. Comparison of patterns of codon usage for genes expressed in different tissues in the real vs. randomized datasets controlling for expression level

Tissue name	Alanine	Arginine	Asparagine	Aspartic acid	Cysteine	Glutamine	Glutamic acid	Glycine	Histidine
Adult carcass	**4.48E-03**	8.14E-01	**3.98E-03**	**1.36E-03**	7.72E-01	**3.43E-03**	**3.03E-05**	3.18E-11	7.03E-01
Brain	**2.04E-06**	**1.55E-06**	**9.33E-03**	**6.44E-03**	**1.70E-02**	**4.29E-06**	**1.12E-07**	**7.31E-02**	6.94E-01
Crop	4.42E-01	1.96E-01	2.61E-01	7.44E-01	3.32E-01	**4.47E-02**	6.25E-01	**3.04E-03**	9.11E-01
Eyes	**1.11E-07**	**5.10E-07**	9.64E-02	9.89E-01	5.18E-01	**2.32E-03**	**3.38E-05**	**1.09E-03**	3.37E-01
Fat body	1.45E-01	7.89E-02	**3.62E-02**	4.32E-01	5.86E-01	6.35E-01	**2.46E-02**	7.67E-01	3.11E-01
Head	**2.75E-02**	**6.48E-04**	1.51E-01	4.73E-01	9.57E-01	1.71E-01	**1.58E-04**	**1.32E-05**	3.71E-01
Heart	7.82E-01	**1.11E-02**	2.46E-01	8.61E-01	2.44E-01	1.33E-01	2.90E-01	7.26E-01	5.39E-01
Hindgut	1.80E-01	**8.57E-03**	**4.22E-02**	1.00E+00	**9.94E-02**	1.14E-01	6.76E-01	6.35E-01	1.01E-01
Male accessory glands	**2.09E-20**	**5.32E-55**	**7.15E-26**	**8.02E-09**	**8.87E-10**	**4.09E-33**	**2.73E-49**	**1.65E-11**	**3.69E-05**
Midgut	**1.29E-05**	**1.83E-03**	**1.16E-04**	6.00E-02	**1.22E-03**	**4.36E-04**	9.92E-01	**4.63E-03**	**2.18E-02**
Ovaries	5.27E-02	**1.79E-09**	1.48E-01	**1.15E-09**	6.15E-01	**9.38E-03**	**3.62E-03**	**9.31E-03**	6.75E-01
Salivary glands	5.27E-01	**1.16E-03**	9.49E-01	**3.09E-06**	**1.70E-02**	6.44E-01	**3.97E-03**	**1.18E-04**	5.79E-01
Testes	**8.25E-48**	0.00E+00	**7.54E-142**	**1.36E-54**	**6.90E-38**	**3.49E-67**	**2.62E-148**	**1.33E-131**	**4.60E-28**
Thoracicoabdominal ganglia	1.31E-01	1.73E-01	**4.64E-02**	6.65E-02	9.68E-01	1.76E-01	1.00E+00	8.88E-01	3.50E-01
Tubules	**6.64E-04**	**1.59E-20**	**4.74E-07**	**2.05E-06**	2.77E-01	**1.82E-03**	**5.17E-07**	**2.44E-03**	1.25E-01
Virgin spermatheca	**1.10E-04**	**4.93E-04**	**4.12E-02**	1.07E-01	4.67E-01	**9.17E-06**	**1.34E-12**	**1.02E-02**	**4.48E-02**

Tissue name	Isoleucine	Leucine	Lysine	Phenylalanine	Proline	Serine	Threonine	Tyrosine	Valine
Adult carcass	1.73E-01	**6.17E-07**	**3.11E-05**	4.29E-01	**2.98E-04**	**3.40E-06**	**3.66E-07**	**1.00E-02**	**3.99E-02**
Brain	**2.29E-02**	**1.93E-14**	**3.90E-03**	**4.78E-02**	**3.64E-08**	**8.76E-10**	**1.42E-09**	8.57E-01	**3.82E-04**
Crop	2.01E-01	**1.72E-03**	**6.14E-04**	**1.12E-02**	8.79E-02	2.44E-01	7.37E-02	**2.45E-02**	2.39E-01
Eyes	**5.34E-07**	**3.82E-04**	**1.09E-09**	**7.38E-05**	**1.15E-05**	**1.42E-04**	**6.46E-05**	5.06E-01	**9.34E-07**
Fat body	**3.74E-02**	4.47E-01	1.00E+00	2.25E-01	6.28E-02	6.60E-01	**8.64E-03**	1.00E+00	8.67E-01
Head	5.81E-01	5.06E-01	**4.44E-03**	4.83E-01	**2.31E-03**	**1.38E-06**	**1.42E-19**	6.92E-01	**3.31E-02**
Heart	2.22E-01	5.31E-01	3.49E-01	1.37E-01	4.86E-01	9.34E-01	1.81E-01	2.72E-01	5.59E-01
Hindgut	3.50E-01	**6.04E-03**	7.54E-01	4.95E-01	1.32E-01	1.11E-01	**1.45E-02**	1.36E-01	1.06E-01
Male accessory glands	**1.84E-25**	**2.01E-50**	**1.11E-60**	**3.31E-13**	**2.09E-03**	**3.17E-23**	**3.89E-06**	**1.55E-12**	**2.97E-08**
Midgut	**1.08E-08**	**1.36E-10**	**1.09E-04**	**5.67E-06**	**3.22E-11**	**2.54E-18**	**5.36E-06**	1.60E-01	**2.50E-02**
Ovaries	3.13E-01	2.59E-04	**3.04E-05**	**1.73E-04**	**3.80E-03**	**2.01E-06**	**6.45E-10**	**4.92E-03**	1.37E-01
Salivary glands	3.49E-01	5.22E-02	**4.68E-03**	9.35E-01	**3.88E-03**	**5.50E-02**	6.43E-02	**3.30E-02**	1.32E-01
Testes	**3.28E-138**	**4.03E-273**	**6.51E-169**	**1.91E-131**	**7.94E-73**	**2.05E-93**	**1.94E-53**	**2.69E-66**	**5.23E-50**
Thoracicoabdominal ganglia	5.99E-01	**1.94E-03**	6.01E-01	6.17E-01	8.27E-01	4.00E-01	5.72E-01	5.96E-01	**4.42E-02**
Tubules	**1.75E-03**	**1.39E-05**	**1.00E-12**	**7.22E-04**	9.46E-01	**7.05E-05**	2.50E-01	8.87E-02	**1.10E-02**
Virgin spermatheca	**8.52E-05**	**6.05E-17**	**1.72E-10**	**1.42E-04**	**1.91E-07**	**3.74E-14**	**5.48E-04**	5.44E-02	**6.17E-11**

P-values correspond to the Chi-squared test. *P*-values < 0.05 are marked with bold face.

Table S3.9. Comparison of patterns of codon usage for genes expressed in different tissues in the real vs. randomized datasets controlling for protein length

Tissue name	Alanine	Arginine	Asparagine	Aspartic acid	Cysteine	Glutamine	Glutamic acid	Glycine	Histidine
Adult carcass	**3.73E-02**	7.30E-01	7.83E-01	4.25E-01	3.82E-01	**3.37E-02**	9.58E-01	**1.30E-03**	1.00E+00
Brain	**7.20E-17**	**4.01E-05**	**2.85E-02**	**3.02E-02**	1.00E-01	**7.70E-04**	**7.14E-11**	**6.18E-06**	1.27E-01
Crop	3.16E-01	2.39E-01	4.76E-01	2.97E-01	4.95E-01	1.26E-01	2.58E-01	3.22E-01	7.98E-01
Eyes	**4.30E-04**	**3.92E-07**	9.74E-02	1.90E-01	9.18E-01	1.53E-01	**3.12E-10**	**1.69E-03**	6.16E-01
Fat body	8.35E-01	**1.99E-05**	**7.10E-02**	**8.85E-02**	8.87E-01	**3.34E-04**	**1.68E-06**	**1.53E-02**	**2.30E-02**
Head	**1.42E-02**	**3.90E-05**	**1.25E-03**	**7.84E-03**	**2.40E-02**	**2.40E-06**	**1.59E-02**	**1.18E-13**	7.05E-02
Heart	2.01E-01	**1.97E-02**	3.04E-01	3.62E-01	1.92E-01	4.27E-01	9.21E-01	7.74E-01	**4.31E-02**
Hindgut	1.67E-01	3.43E-01	**2.03E-02**	9.07E-02	2.14E-01	4.17E-01	1.73E-01	**4.31E-02**	4.51E-01
Male accessory glands	**1.63E-13**	**1.71E-27**	**1.43E-19**	**9.98E-06**	**3.55E-06**	**1.82E-29**	**8.76E-39**	**3.31E-13**	**1.68E-11**
Midgut	**8.00E-10**	**3.35E-02**	**1.70E-08**	2.08E-01	6.05E-02	1.27E-01	6.57E-01	**2.33E-04**	**1.37E-02**
Ovaries	**5.17E-05**	**1.39E-02**	2.93E-01	**2.04E-05**	3.82E-01	3.40E-01	1.28E-01	**4.66E-03**	1.36E-01
Salivary glands	1.37E-01	**5.02E-03**	2.56E-01	**1.64E-04**	**1.42E-04**	5.20E-01	**3.54E-04**	**1.65E-02**	5.51E-01
Testes	**2.72E-17**	**1.47E-104**	**1.84E-75**	**6.66E-30**	**6.70E-22**	**1.72E-48**	**1.25E-62**	**5.59E-65**	**1.07E-21**
Thoracicoabdominal ganglia	**8.15E-03**	4.92E-01	**4.79E-03**	4.89E-01	1.87E-01	3.44E-01	7.44E-02	3.87E-01	1.08E-01
Tubules	6.21E-01	**2.93E-03**	5.76E-01	**1.37E-03**	9.89E-01	4.35E-01	**6.27E-03**	4.33E-01	4.07E-01
Virgin spermatheca	5.94E-02	**8.55E-04**	**1.48E-02**	6.51E-01	5.86E-01	**8.37E-06**	**6.96E-09**	**6.94E-02**	4.60E-01

Tissue name	Isoleucine	Leucine	Lysine	Phenylalanine	Proline	Serine	Threonine	Tyrosine	Valine
Adult carcass	1.26E-01	3.70E-01	1.80E-01	1.00E+00	**1.31E-02**	2.22E-01	**1.23E-02**	9.48E-02	2.49E-01
Brain	**1.12E-06**	**1.17E-22**	**2.18E-04**	**5.52E-03**	**6.28E-12**	**5.52E-20**	**2.54E-09**	9.09E-01	**2.41E-10**
Crop	1.68E-01	**2.99E-02**	**1.14E-04**	**2.55E-02**	2.45E-01	2.74E-01	9.60E-01	6.48E-01	4.19E-01
Eyes	**4.05E-07**	**9.80E-04**	**1.35E-09**	**1.79E-04**	**4.60E-04**	**5.15E-06**	**1.36E-04**	5.22E-01	**4.16E-04**
Fat body	1.03E-01	**5.55E-12**	**3.22E-07**	**2.04E-02**	1.52E-01	**3.45E-02**	**1.01E-02**	1.15E-01	**7.05E-03**
Head	**3.34E-03**	**5.41E-04**	2.70E-01	**2.84E-03**	4.10E-01	**1.19E-02**	2.94E-01	1.59E-01	**1.03E-03**
Heart	1.89E-01	6.91E-01	6.34E-01	7.03E-01	7.51E-01	6.10E-01	3.51E-01	3.60E-01	9.17E-01
Hindgut	**4.39E-03**	1.85E-01	9.43E-01	2.39E-01	1.09E-01	3.68E-01	6.06E-01	3.54E-01	6.91E-01
Male accessory glands	**3.34E-22**	**3.53E-60**	**1.71E-25**	**3.78E-12**	**8.75E-25**	**6.69E-31**	**1.53E-30**	**3.93E-11**	3.14E-11
Midgut	**3.09E-08**	**2.25E-06**	**3.55E-03**	**2.78E-03**	**1.51E-12**	**4.15E-08**	**1.28E-12**	3.33E-01	**1.12E-02**
Ovaries	7.32E-01	**4.50E-03**	1.96E-01	**4.30E-02**	9.25E-02	**4.93E-02**	**2.02E-06**	**1.21E-02**	7.20E-01
Salivary glands	5.10E-01	**2.45E-04**	**1.69E-05**	5.34E-01	**1.34E-02**	**3.79E-02**	**3.07E-02**	8.37E-02	4.75E-01
Testes	2.02E-51	**3.18E-160**	**3.31E-59**	**2.30E-59**	**3.11E-34**	**9.48E-112**	**3.20E-56**	**1.68E-50**	**1.98E-34**
Thoracicoabdominal ganglia	1.30E-01	**1.02E-02**	**1.38E-02**	4.41E-01	2.51E-01	4.61E-01	8.75E-01	1.00E+00	6.51E-01
Tubules	7.45E-01	2.25E-01	**4.42E-02**	9.24E-01	**8.20E-03**	**1.20E-03**	**8.57E-04**	4.82E-01	4.91E-01
Virgin spermatheca	**3.92E-02**	**1.47E-06**	**6.36E-05**	4.29E-01	**1.59E-03**	**1.02E-09**	**8.99E-04**	**2.06E-02**	**3.24E-06**

P-values correspond to the Chi-squared test. *P*-values < 0.05 are marked with bold face.