Spring 5-15-2019

The Evolutionary and Functional Roles of Synonymous Codon Usage in Eukaryotes

Zhen Peng

Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST. LOUIS

Division of Biology and Biomedical Sciences
Evolution, Ecology and Population Biology

Dissertation Examination Committee:
Yehuda Ben-Shahar, Chair
Barak Cohen
Ian Duncan
Kenneth Olsen
David Queller
Hani Zaher

The Evolutionary and Functional Roles of Synonymous Codon Usage in Eukaryotes

by
Zhen Peng (彭镇)

A dissertation presented to
The Graduate School
of Washington University in
partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

May 2019
St. Louis, Missouri

Table of Contents

List of Figures

List of Tables

List of Abbreviations

SNP: single-nucleotide polymorphism

GWAS: genome-wide association study

GO: gene ontology

PFCC: putatively functional codon cluster

CAI: codon adaptation index

TCAI: transformed codon adaptation index

RLI: relative location index

Nav: voltage-gated sodium channel α subunit

DIII-IV linker: the cytoplasmic loop linking the Domain III and Domain IV of the voltage-gated sodium channel α subunit.

Acknowledgments

I thank my thesis committee members and dissertation defense committee members Drs. Yehuda Ben-Shahar, Barak Cohen, Ian Duncan, Kenneth Olsen, David Queller, and Hani Zaher. I thank Dr. Alan Templeton, Dr. Allan Larson, and members of the Ben-Shahar and Zaher laboratories for helpful discussions. I also thank Paula Kiefel and Dianne Duncan for technical assistance with generating transgenic animals and imaging.

This work was supported by the National Institutes of Health (grant numbers R21NS089834 to Yehuda Ben-Shahar, R01GM112641 to Hani Zaher) and the National Science Foundation (grant numbers 1545778 and 1707221 to Yehuda Ben-Shahar). Zhen Peng is supported by the National Institutes of Health training program (grant number T32HG000045).

Zhen Peng (彭镇)

Washington University in St. Louis

May 2019

Dedicated to Dr. Yang Zhong (钟扬).

ABSTRACT OF THE DISSERTATION

The Evolutionary and Functional Roles of Synonymous Codon Usage in Eukaryotes

by

Zhen Peng (彭镇)

Doctor of Philosophy in Biology and Biomedical Sciences

Evolution, Ecology and Population Biology

Washington University in St. Louis, 2019

Professor Yehuda Ben-Shahar, Chair

Most amino acids are encoded by multiple synonymous codons. Although alternative usage of synonymous codons does not affect the amino acid sequences of proteins, researchers have been reporting evidence for functional synonymous codon usage at the species- and gene-specific levels for over four decades. It has been shown that variations in synonymous codon usage can affect phenotypes through diverse mechanisms such as shaping translation efficiency and mRNA stability. On the other hand, the common view that cellular and organismal phenotypes are primarily determined by proteins whose functions are primarily determined by amino acid sequences, often drives the assumption that synonymous mutations are evolutionarily neutral. Consequently, this assumption has been used extensively in evolutionary biology, population genetics, and structural biology. One explanation of the apparent contradiction between the empirical findings, which indicate that synonymous mutations can affect related phenotypes, and the theoretical models, which stipulate that synonymous mutations are neutral, is that neutral synonymous mutations represent the general rule while non-neutral synonymous mutations represent the rare exceptions. In my thesis, I examined this explanation by applying computational and experimental approaches, which indicated that: 1) Non-neutral synonymous mutations significantly affect a considerable proportion of protein-coding genes; 2) Gene-specific codon usage patterns, such as the preference for a specific combination of rare codons, are possibly associated with specific gene functions, such as enhancing tissue-specific gene expression; 3) Some protein-coding genes include codon clusters whose codon usage patterns cannot be explained by selection-independent processes, and thus such codon clusters seem to serve as domains affecting protein functions. Together, these data suggest that synonymous mutations should not be *a priori* considered neutral. Furthermore, my studies suggest that the biochemical functions of at least some proteins are not only shaped by the constituent amino acid residues but also by codon usage biases at the gene-specific and sub-genic levels. In conclusion, my thesis work suggests that many of the commonly used approaches for analyzing the selection on protein-coding DNA sequences, which rely on the assumption that synonymous mutations are generally neutral, may generate biased results. Furthermore, my studies indicate that selection on gene-specific codon usage bias has evolved to serve diverse biological functions, which are still mostly uncharacterized.

Chapter 1: Introduction

1.1 Evidence for Functional Synonymous Codon Usage

A single amino acid can be encoded by different synonymous codons. Although mutations that cause the switch between synonymous codons, namely synonymous mutations, do not affect protein sequences, species- and/or gene-specific codon usage patterns often exhibit non-random biases [1–4]. Such biased patterns of synonymous codon usage, or codon usage biases, suggest that although different synonymous codons encode the same amino acid, they could potentially have different biological functions [1,5,6]. In fact, researchers have been reporting evidence for functional synonymous codon usage for more than four decades [7,8,1,9,3,4,10–12,2,13–31,6,32–37].

A prominent theory explaining the mechanism by which synonymous codon usage can affect protein functions is the translational selection theory [1–3,6,12,31,38]. Empirical studies done in prokaryotes have shown that the codons preferred by highly expressed genes are usually associated with abundant tRNAs, and mRNAs with enriched codons of such a type are indeed translated faster on average [1,2,38]. Thus, the frequently used synonymous codons are termed “optimal codons”, while the rarely used codons are termed “suboptimal codons” or “non-optimal codons” [1,3,16,30,38,39]. Consequently, it has been hypothesized that natural selection on synonymous mutations may result in adaptive synonymous codon usage that optimizes translation efficiency, which is referred to as the “translational selection theory” [1]. Empirical and theoretical data suggest that translational selection is especially strong for highly expressed housekeeping genes such as actins and tubulins [1,31]. In addition, the concept of adaptive synonymous codon usage also led to the proposal of codon adaptation index (CAI) [3], which measures how well the codon usage bias of a given protein-coding sequence is adapted to the putatively most optimal synonymous codon usage that is usually inferred from the codon usage of a set of highly-expressed housekeeping genes or the whole genome. CAI has been widely used in directing codon optimization, in which the codon usage of transgenes is designed to be adapted to the codon usage biases of host organisms [16,40,41].

Biased synonymous codon usage has also been shown to play functional roles through other mechanisms. First, mRNA splicing is regulated by the specific nucleotide sequences at the intron-exon boundaries [42–44], and therefore, synonymous mutations in these regions could dramatically affect splicing, which can lead to frame shifts and aberrant protein products [20,22,23]. Second, the accessibility of miRNAs to mRNAs is affected by synonymous codon usage [27]. Third, synonymous changes at the DNA level could affect epigenetic regulation of gene action [45,46]. Fourth, “suboptimal” codons could have important biological consequences via their direct impact on the deceleration of translation rates, which subsequently could impact co-translational processes such as protein folding, post-translational modifications (e.g., phosphorylation), and subcellular localization [5,24,30,47,48].

1.2 The Neutrality of Synonymous Mutations Is Important for Evolutionary Biology, Population Genetics, and Structural Biology

Even though numerous studies have suggested that synonymous mutations could affect protein functions and associated phenotypes, the assumption that synonymous mutations are mostly neutral is still frequently used in evolutionary biology and population genetics [49–146], which presents an apparent conundrum.

One possible explanation for this conundrum is that neutral synonymous mutations represent the general rule while non-neutral synonymous mutations are the rare exceptions [147]. The logic underlying this explanation is “amino-acid determinism”, which stipulates that the phenotypes are primarily determined by protein functions, and the properties of proteins are primarily determined by the amino acid sequences. Therefore, because synonymous mutations do not change the amino acid sequences of a protein, they should have minimal impacts on the biochemical functions of the protein and its associated phenotypes [88,147–149].

In evolutionary biology, the assumption that synonymous mutations are generally neutral has been the foundation of multiple commonly used methods to detect signatures of natural selection, to estimate the rate of evolution, and to classify the types of natural selection [70,88,101,103,107,137]. One example is the dN/dS method [103]. This method is used to analyze the influence of natural selection on the evolution of the aligned protein-coding homologs during a certain period of time [70,103]. dN denotes the number of nonsynonymous mutations per nonsynonymous site and dS denotes the number of synonymous mutations per synonymous site. Thus, dS serves as the estimate of the influence of neutral evolution while dN serves as the estimate of the combined effect of neutral evolution and natural selection on the protein-coding genes. If dN/dS is near 1, natural selection on the protein-coding genes is absent or very weak. If dN/dS is dramatically higher than 1, amino acid substitutions are generally beneficial and thus are favored by natural selection; such a case is usually regarded as an example of positive selection. If dN/dS is dramatically lower than 1, amino acid substitutions are generally deleterious and thus natural selection keeps the conserveness of amino acid sequences; such a case is usually regarded as an example of negative selection or purifying selection [70]. Consequently, the validity of the dN/dS method relies on whether dS truly represents the influence of neutral evolution. Similarly, the McDonald-Kreitman test (MK test) and the derivatives of the dN/dS method and MK test also depend on the assumption that synonymous mutations are generally neutral. These methods are still frequently applied in evolutionary biology studies [49–69,71–87,89–100,102,104–106,108–136,138–145].

In population genetics, synonymous single-nucleotide polymorphisms (SNPs) are usually treated as functionally neutral genetic variants, while nonsynonymous SNPs are usually treated as the candidate functional variants [150,151]. Consequently, in many genome-wide association studies (GWAS) aimed at detecting genetic variants underlying specific traits ranging from the mass of seeds to the susceptibility to a specific disease, nonsynonymous SNPs are usually primarily or exclusively focused on [152–157]. Thus, whether synonymous SNPs are truly functionally neutral determines the validity of excluding synonymous SNPs from the analyses.

Besides, structural biology studies that are aimed at determining the three-dimension structures of proteins are also affected by the assumed neutrality of synonymous mutations. In practice, structural biologists mostly focus on analyzing amino-acid-level information by techniques such as X-ray crystallography and cryogenic electron microscopy, in order to determine the structures of proteins [158,159]. Such practice relies on two assumptions: first, the native structures of proteins are at most slightly affected by sample preparation; second, nucleic-acid-level information, as long as it does not affect amino acid sequences, has minimal impacts on protein structures. Thus, the validity of the second assumption essentially depends on the neutrality of synonymous mutations.

1.3 The Possible Impacts of Prevalent Non-neutral Synonymous Mutations on the Relevant Research

The explanation that neutral synonymous mutations represent a general rule while non-neutral synonymous mutations are rare exceptions seems able to reconcile the contradiction between the empirical evidence for functional synonymous codon usage and the frequent usage of synonymous mutations as proxies for neutral mutations. Nevertheless, this explanation has not been rigorously tested, possibly for two major reasons. First, although multiple empirical studies have reported evidence for non-neutral synonymous mutations, they are still quantitatively anecdotal compared to the entire set of known genes and species. Therefore, they are not enough to disprove or verify the generality of neutral synonymous mutations. Second, although multiple computational studies tried to assess the prevalence of non-neutral synonymous mutations [7,10–13,29,36], their methods had at least one of the two key weaknesses that hindered the generalization of their results. One weakness was that some methods were only applicable to few species with high-quality genetic variation data at the level of population or phylogeny. The other weakness was that some methods had limited statistical power because they either focused exclusively on the four-fold degenerate codons or treated codons with different degrees of degeneracy separately. Thus, the lack of definitive tests to confirm the generality of neutral synonymous mutations has left a critical gap in the relevant research.

If the prevalence of neutral synonymous mutations is actually not order(s) of magnitude larger than that of non-neutral synonymous mutations, it is likely that the frequently adopted assumption that synonymous mutations are neutral has been introducing systematic biases into relevant studies. First, prevalent non-neutral synonymous mutations mean that the rate of synonymous substitutions may not be a good proxy for the rate of neutral evolution. For example, the dS parameter in the dN/dS-type methods likely misestimates the impact of neutral evolution. If synonymous mutations are generally under purifying selection, dS will underestimate the rate of neutral evolution, which will result in overestimated impact of positive selection on the amino acid sequences. Actually, it has been shown that even weak selection on synonymous codon usage can strongly bias the results of the methods that are based on the neutrality of synonymous mutations [37]. Second, prevalent non-neutral synonymous mutations indicate that excluding synonymous SNPs from GWAS may not be appropriate practice. If a trait is actually only associated with synonymous SNPs, such practice will generate false-negative results. If a trait is indeed associated with both synonymous and nonsynonymous SNPs, such practice will misestimate the significance and effect sizes of the detected functional nonsynonymous SNPs. Third, prevalent non-neutral synonymous mutations suggest that there must be important mechanisms, other than the amino acid sequences, that regulate the functions of proteins and relevant phenotypes. This will undermine the generality and applicability of the research paradigms that rely on amino-acid determinism, which indicates that researchers should not simply infer the structures and functions of proteins or sub-protein segments only from amino-acid-level information. For example, the native structure of a protein or a sub-protein segment may be highly dynamic such that the structure had better be described as a spectrum rather than a single representative three-dimension model [160,161]. If such a dynamic feature is associated with specific synonymous codon usage patterns, ignoring nucleic-acid-level information would prevent structural biologists from depicting the spectrum of protein structures.

Consequently, it is necessary and timely to assess to what extent synonymous mutations are neutral. If synonymous mutations are indeed generally neutral, methods based on the general neutrality of synonymous mutations can be legitimately used as the default with little, if any, risks of generating biased results. However, if the neutrality of synonymous mutations is not as general as previously assumed [88,147–149], researchers may need to seriously consider updating their paradigms by incorporating non-neutral synonymous mutations as a regular source of functional genetic variants, and further investigation should be done to reveal the exact roles played by synonymous codon usage under specific biological contexts.

1.4 Disproving the Generality of Neutral Synonymous Mutations and Amino-acid Determinism

In my thesis, I used gene-specific and sub-genic codon usage biases to investigate the impacts of non-neutral synonymous mutations on molecular evolutionary biology in general and on some specific biological functions of synonymous codon usage in particular. With publicly available databases of protein-coding sequences in diverse species and annotations of gene functions in frequently used model organisms [162–166], and with convenient genetic manipulation of fruit flies [167,168], gene-specific codon usage biases can serve as ideal materials that balance the breadth and depth of computational and experimental analyses on the evolutionary and functional roles of synonymous mutations. On the other hand, sub-genic codon usage biases provide the opportunity to conduct higher-resolution analyses on the relationship between codon usage biases and structural units of proteins [169], which supplements the analyses on gene-specific synonymous codon usage.

In Chapter 2, by developing a widely applicable statistical method to detect signatures of natural selection on gene-specific codon usage biases and applying it to diverse eukaryotic species, I found that non-neutral synonymous mutations significantly affect a considerable proportion of protein-coding genes. Thus, this result refutes the claim that non-neutral synonymous mutations are rare exceptions to the putatively general rule that synonymous mutations are neutral.

Furthermore, I showed that gene-specific codon usage patterns are associated with specific gene functions. I experimentally showed that a combination of rare codons for specific amino acids is involved in regulating tissue-specific gene expression in *Drosophila melanogaster*. Thus, these results suggest that the relationship between codon usage patterns and gene functions in multicellular eukaryotes is more complex and context-dependent than currently assumed.

In Chapter 3, by developing a statistical method to detect sub-genic regions with characteristic codon usage patterns that cannot be explained by selection-independent processes and applying it to the *D. melanogaster* genome, I identified multiple putatively functional codon clusters (PFCCs). I found that although some of these PFCCs are associated with protein domains, which are predicted from their amino acid sequences, the majority of the PFCCs are not associated with any known protein domains. Thus, it is highly likely that some functional units of proteins are encoded by codon usage patterns instead of amino acid sequences. In this regard, amino-acid determinism is at least partially flawed.

Together, my results suggest that synonymous mutations are not generally neutral, and that the properties of proteins are not exclusively determined by the constituent amino acid residues.

Chapter 2: Natural Selection on Eukaryotic Gene-Specific Codon Usage Bias Has Broad Evolutionary and Functional Implications for the Regulation of Gene Action

2.1 Abstract

Because they do not affect protein sequences, synonymous mutations are often used as proxies for neutral mutations in tests for signatures of natural selection on protein-coding DNA sequences. Yet, numerous studies have also indicated that synonymous mutations can have dramatic effects on phenotypes. One hypothesis that might explain these seemingly contradictory interpretations of the biological significance of synonymous mutations is that the majority of synonymous mutations are indeed neutral, while non-neutral synonymous mutations are the rare exceptions. However, due to the lack of broadly applicable approaches for estimating the prevalence of non-neutral synonymous mutations across sequenced genomes, and for predicting specific biological functions of gene-specific codon usage biases, this important hypothesis has not been rigorously tested. Here we used computational and empirical approaches to demonstrate that signatures of natural selection on gene-specific codon usage bias are common in eukaryotic genomes, which necessitates reconsidering the frequently adopted assumption that synonymous mutations are generally neutral. As a proof of principle, we show that in *Drosophila melanogaster*, selection on the increased usage of a specific combination of rare codons plays an important role in enhancing translation specifically in the male reproductive system. Together, these data indicate that synonymous mutations should not be *a priori* assumed to be neutral, and

that the relationship between codon usage patterns and gene functions in multicellular eukaryotes is more complex and context-dependent than currently assumed.

2.2 Introduction

Ever since the “neutral theory of molecular evolution” was proposed [149], synonymous mutations in protein-coding DNA sequences have been generally considered evolutionarily neutral because they do not affect amino acid sequences, and therefore, should have a minimal impact on protein functions and associated phenotypes [88,147–149]. Consequently, some quantitative and statistical approaches such as the dN/dS method, the McDonald-Kreitman test, and their derivatives, which are commonly used in molecular evolutionary research for detecting signatures of natural selection, identifying types of natural selection, and estimating rates of the molecular evolution of protein-coding genes, often use synonymous mutations as proxies for neutral mutations [49–146,170]. Yet, numerous studies have also shown that, in contrast to the assumption of neutrality, some synonymous mutations could also affect phenotypic outcomes, possibly via impacting mRNA secondary structures and stability, splicing, miRNA binding, epigenetic modifications, and translation efficiency [7,8,1,9,3,4,10–12,2,13–31,6,32–37].

The contradictory interpretations of the biological significance of synonymous mutations could be explained by two hypotheses. The first hypothesis is that natural selection on synonymous mutations is likely generally weak, so that the impact of non-neutral synonymous mutations on the results generated by the neutral-synonymous-mutation-based computational methods should also be weak [147]. However, it has been recently shown mathematically that weak selection on synonymous mutations can strongly bias the results generated by the neutral-synonymous-mutation-based methods [37]. Therefore, this first hypothesis does not hold. The second hypothesis is that most synonymous mutations are indeed neutral, while non-neutral synonymous mutations are the rare exceptions [147]. However, empirical and theoretical studies that directly estimate the actual prevalence of non-neutral synonymous mutations in eukaryotic genomes are rare, possibly for two major reasons. First, although multiple empirical studies have investigated the impact of specific synonymous mutations on the associated functions and phenotypes, they represent a very small fraction of the overall number of known protein-coding genes, and therefore, are not enough for quantitatively estimating the prevalence of non-neutral synonymous mutations. Second, previous studies that used theoretical and computational approaches for testing whether synonymous mutations are generally neutral have had limited power because they either focused solely on four-fold degenerate codons, or had treated codons with different degrees of degeneracy independently [7,10,12,29,36]. In addition, the applicability of some of the computational approaches for identifying signatures of selection on synonymous mutations has been restricted to the few species with high-quality genetic variation data at the levels of populations and/or phylogeny [11,13,29,36]. Therefore, it is timely and important to develop new powerful and broadly applicable quantitative approaches for detecting signatures of natural selection on synonymous mutations from currently available genomic data.

Here we developed a statistical approach for detecting gene-specific signatures of natural selection on synonymous mutations, which integrates information from all degenerate codons in any native protein-coding sequence that uses the standard genetic code. Using this approach, we screened the sequenced genomes of 40 species from diverse eukaryotic clades, and found that the majority of them have numerous protein-coding genes that carry statistically significant signatures of natural selection on synonymous mutations. Although the observed level of impact of selection on gene-specific codon usage bias varies dramatically across species, these conservative estimates inevitably disprove the assumption that synonymous mutations are generally neutral. Furthermore, by exploiting the fruit fly *Drosophila melanogaster* as a model, we demonstrate that different groups of protein-coding genes have likely been selected for the increased usage of different combinations of specific codons, and that preference for specific codon combinations might be associated with specific categories of gene functions. Particularly, we show that a specific combination of rare codons, which are often referred to as "suboptimal codons" because they are preferentially recognized by tRNAs with small gene copy numbers [171–173], play an important role in enhancing protein expression specifically in the male reproductive system. Together, these findings suggest that both neutral and non-neutral synonymous mutations are prevalent in eukaryotic genomes, and that the functional roles of gene-specific synonymous codon usage are diverse and cannot be simply predicted from whole-genome codon usage frequencies or tRNA gene copy numbers. Thus, it is highly likely that methods based on the assumed neutrality of synonymous mutations have already generated systematic biases in genetics and evolutionary biology. Also, the view that the optimality of a codon is intrinsically associated with its whole-genome usage frequency or the copy number of its cognate tRNA genes, may overlook diverse regulatory roles played by synonymous codon usage.

2.3 Results

2.3.1 Gene-specific Signatures of Natural Selection on Synonymous Mutations Are Common Across Eukaryotes

Testing the hypothesis that non-neutral synonymous mutations are rare exceptions requires a broadly applicable method for statistically identifying signatures of natural selection on synonymous mutations across protein-coding genes throughout genomes. Mathematically, the previously published "selection-mutation-drift model", which stipulates that observed codon usage patterns are the result of an interplay between natural selection, mutational bias, and genetic drift, could be used for identifying such signatures [174]. However, applying this specific model to many genes across diverse species requires prior empirical information about population-level genetic variations, temporal and spatial gene expression patterns, and related estimates of fitness [175–180]. Since such data are not available for most species, we developed an alternative statistical approach that is solely based on the analyses of DNA sequences of protein-coding genes from publicly available reference genomic data. In contrast to the previously published "selection-mutation-drift model", which evaluates the relative contribution of natural selection to the observed codon usage bias by estimating the selection coefficient for each possible nucleotide substitution, our approach is based on the statistical rejection of the null hypothesis that "synonymous mutations are neutral". Although this approach does not provide specific quantitative estimates for the level of selection on each individual gene, it serves as an effective method for identifying specific genes whose biased codon usage patterns have been impacted by natural selection.

Our statistical model is based on several key assumptions: 1) Reference genomic data represent a "wild type" genome. 2) Observed gene-specific codon usage patterns are at equilibrium. 3) For each codon, no more than one nucleotide substitution per generation is possible. 4) Nonsynonymous mutations are more likely to be deleterious than synonymous mutations, and therefore, the probabilities of observed nonsynonymous substitutions are negligible relative to those of synonymous ones. 5) The contributions of individual alleles of a single gene to fitness are additive. 6) Gene-specific mutational bias, which is assumed to be shaped by both genome-scale mutational bias and local forces such as gene conversions [181], is independent of that of any other gene. Thus, this assumption simplifies the estimation of the mutational bias of each focal gene by using its specific codon counts. 7) For each individual open reading frame, the mutation rate for each of the 12 possible nucleotide substitutions (A-to-T, T-to-A, *etc.*) is constant [182], and therefore, the possible effects of adjacent nucleotides on mutation rates of each focal nucleotide [183] are not considered. It should be noted that in our model, “mutation rates” refer to the rates of *de novo* mutations rather than site-specific or position-specific substitution rates. 8) The ratio of the lowest to highest mutation rates for all possible nucleotide substitutions above is at least 1/100, which represents a relatively relaxed constraint on gene-specific mutational bias [184,185].

Next, we used the previously published “selection-mutation-drift model” [174,175] to describe the relationships between the various evolutionary molecular processes that may have shaped observed gene-specific codon usage biases (Equation (2.1)).

$$\begin{aligned} &\sum_i \frac{2s_{pqi\to pqr}}{1-e^{-2Ns_{pqi\to pqr}}} \times x_{pqi} \times \mu_{i\to r} + \sum_j \frac{2s_{jqr\to pqr}}{1-e^{-2Ns_{jqr\to pqr}}} \times x_{jqr} \times \mu_{j\to p} \\ &= \sum_i \frac{2s_{pqr\to pqi}}{1-e^{-2Ns_{pqr\to pqi}}} \times x_{pqr} \times \mu_{r\to i} + \sum_j \frac{2s_{pqr\to jqr}}{1-e^{-2Ns_{pqr\to jqr}}} \times x_{pqr} \times \mu_{p\to j} \end{aligned} \tag{2.1}$$

Variables *p*, *q*, and *r* denote the specific nucleotide identities (A, C, G, or T) of the first, second, and third positions respectively in each codon. Codons synonymous to codon *pqr* that vary at either the first or third position are denoted by codons *jqr* or *pqi* respectively. Therefore, $s_{pqi\to pqr}$, for example, denotes the selection coefficient of the *pqi*-to-*pqr* mutation. *N* denotes effective chromosomal population size [175]. x_{pqr} denotes the count of codon *pqr* in a focal gene, and $\mu_{i\to r}$ denotes an estimate for the *i*-to-*r* mutation rate. Consequently, the probability that an *i*-to-*r* mutation in the focal gene is fixed is $\frac{2s_{pqi\to pqr}}{1-e^{-2Ns_{pqi\to pqr}}}$ [175]. The amino acid serine represents a unique case because it is encoded by six synonymous codons that belong to two independent codon groups, which are not interchangeable via single synonymous substitutions. Therefore, in our model, we treat the Ser codon groups AGC, AGT and TCA, TCC, TCG, TCT as if they were encoding two independent amino acids [178].

Since the null hypothesis of our model specifies that all synonymous mutations are neutral, the selection coefficient of each synonymous mutation should have a zero value, which simplifies Equation (2.1) as follows:

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

$$\Rightarrow$$

$$\sum_i \frac{1}{N} \times x_{pqi} \times \mu_{i\to r} + \sum_j \frac{1}{N} \times x_{jqr} \times \mu_{j\to p} = \sum_i \frac{1}{N} \times x_{pqr} \times \mu_{r\to i} + \sum_j \frac{1}{N} \times x_{pqr} \times \mu_{p\to j} \quad (2.2)$$

Furthermore, since each term in Equation (2.2) includes the factor $\frac{1}{N}$, Equation (2.2) could be further simplified as:

$$\sum_i x_{pqi} \times \mu_{i\to r} + \sum_j x_{jqr} \times \mu_{j\to p} = \sum_i x_{pqr} \times \mu_{r\to i} + \sum_j x_{pqr} \times \mu_{p\to j} \quad (2.3)$$

Subsequently, if the null hypothesis that all synonymous mutations are neutral is true then the impacts of natural selection, represented by selection coefficient *s*, and genetic drift, represented by the effective chromosomal population size *N*, are canceled as shown above, and therefore, the observed codon usage pattern could be explained solely by local mutational bias. Canceling the effect of genetic drift also allows us to apply our model to single focal genes, as has been previously suggested [7,10,12]. In contrast, if the null hypothesis is rejected then we must accept the alternative hypothesis, which indicates that natural selection has had a significant impact on the observed gene-specific codon usage bias.

To apply our derived statistical model to actual genomic data, we initially generated a dataset of expected codon counts for gene-specific codon usage pattern of a focal reference gene, assuming that it is determined solely by gene-specific mutational bias. By applying Equation (2.3) to the open reading frame of a focal gene, gene-specific μ values were estimated such that the χ^2 statistic calculated from the expected and observed codon counts for the focal gene was minimized. We then tested whether the model-generated (expected) and the actual (observed) gene-specific codon usage patterns are similar by using a χ^2 test (df = 40) [186–188]. Subsequently, all genes with observed codon usage patterns that were significantly different from the expected patterns generated by the model were classified as genes that carry signatures of natural selection on synonymous mutations. Because detection of selection signatures by our approach is strongly associated with differential usage of nonsynonymous codons that end with the same nucleotide, our approach effectively filters out the interference from local gene conversions since gene conversions make the usage of codons ending with the same nucleotide change in a similar way during evolution. Besides, as this approach only requires genetic code and the nucleotide sequences of open reading frames as input, in principle, it can be applied to any native protein-coding genes in any species whose genetic code is known. The major cost of the broad applicability of our method is the ability to quantitatively estimate the strength of selection on synonymous mutations. Nevertheless, we think that the cost is affordable because it has been theoretically shown that even weak selection on synonymous mutations can strongly bias the results of the methods based on the general neutrality of synonymous mutations [37]. Therefore, if the selection on synonymous mutations is strong enough to allow the statistical detection of selection signatures by our method, such signatures should indicate that it is inappropriate to use synonymous mutations as proxies for neutral mutations.

Next, we used our approach to analyze codon usage bias patterns in diverse eukaryotic species. We found that after correcting for false discovery rate (FDR = 0.05) [189], in 35 out of 40 eukaryotic genomes analyzed, at least 10% of the protein-coding genes carried significant signatures of natural selection on synonymous mutations, independent of whether they are unicellular or multicellular (Figure 2.1). In 9 species, including *Homo sapiens* and frequently used model organisms *Dictyostelium discoideum*, *Chlamydomonas reinhardtii*, *Mus musculus*, *Danio rerio*, and *Drosophila melanogaster*, the percentages of protein-coding genes carrying selection signatures were even over 50%. Although these estimates might be affected, at least in part, by the current variable states of sequence annotation qualities across the different publicly available genomes, a closer look at the very-well annotated genomes of *Drosophila melanogaster*, *Arabidopsis thaliana*, *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, and *Homo sapiens* [190] revealed that the percentages of protein-coding genes carrying selection signatures in these species still varied between 14% and 64% (Figure 2.1). These results indicate that the overall impact of natural selection on eukaryotic gene-specific codon usage bias is significantly broader than would be expected if synonymous mutations were mostly neutral. Our results also suggest that the relative impact of selection on gene-specific codon usage bias is not constant across the eukaryotic phylogeny, and therefore, might be the result of diverse, clade-specific selective forces.

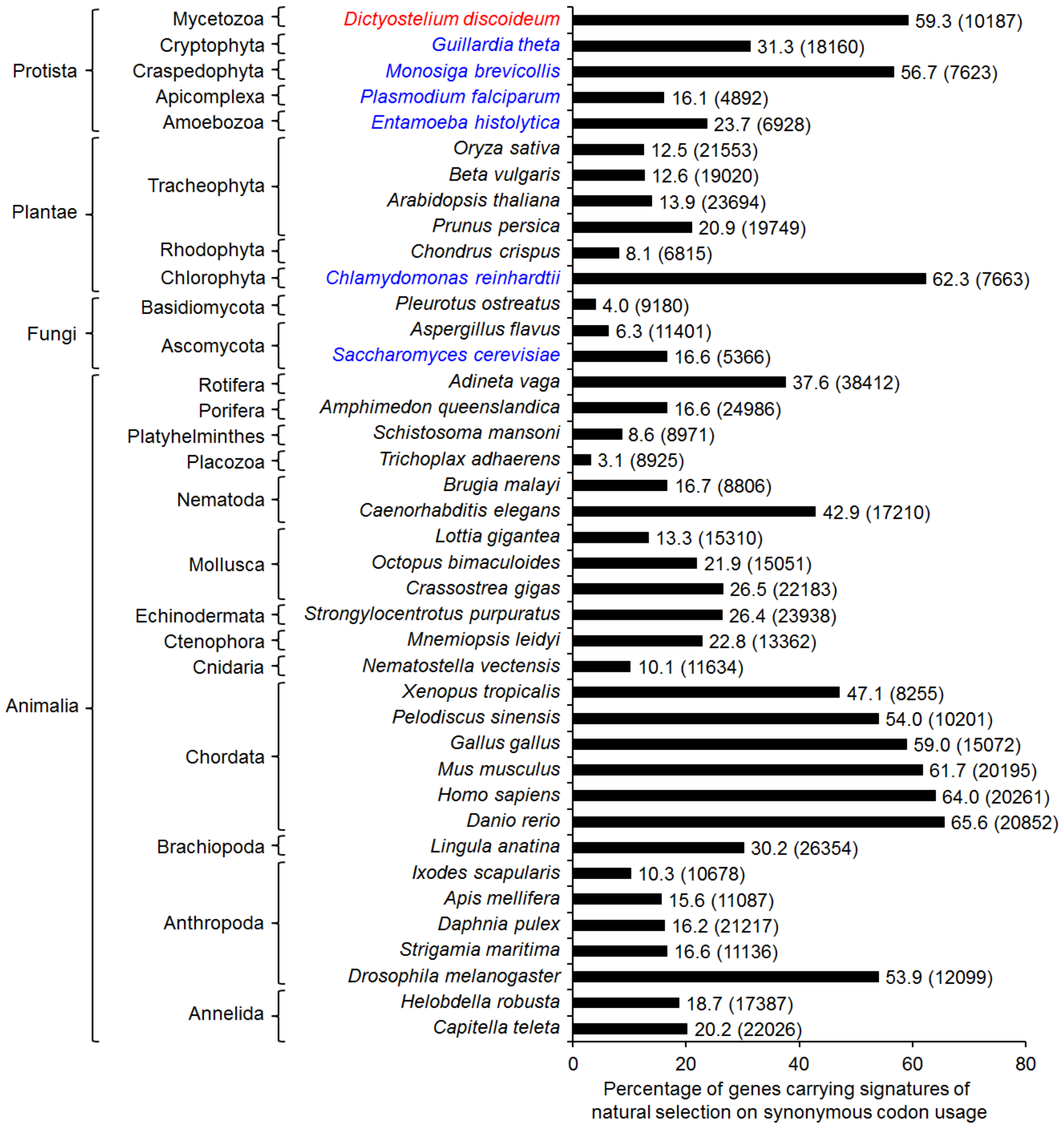

Figure 2.1: Signatures of natural selection on gene-specific codon usage bias in eukaryotes. For shown species, all annotated protein-coding genes that passed data pre-processing filters were included. Total numbers of protein-coding genes analyzed for these species are shown in the parentheses. Percentages of genes carrying signatures of selection on codon usage bias are corrected by false discovery rate (FDR=0.05). Species are stacked by phylum and kingdom. Blue, unicellular species; black, multicellular species; red, both unicellular and multicellular forms exist.

2.3.2 The Heterogeneous Impact of Natural Selection on Gene-specific Codon Usage Bias Is Likely Driven by Diverse Biological Functions

Although our statistical approach is able to provide quantitative estimates for the proportion of protein-coding genes whose codon usage patterns have been biased by natural selection in a particular genome (Figure 2.1), it does not provide qualitative information in terms of which specific codons may have contributed more to the overall observed signature in each focal gene, nor what might be the biological functions of specific biased codon usage patterns. Consequently, we next used a clustering approach to determine whether the genes we have identified in our initial screen share similar codon usage patterns, or alternatively, represent a heterogeneous population comprised of multiple gene clusters, each defined by a unique pattern of biased codon usage that may support specific biological functions.

Due to its well-annotated genome, wealth of available genetic and phenotypic data, and the high prevalence of gene-specific signatures of natural selection on synonymous codon usage, we chose to use *D. melanogaster* as a model for further analyses of the possible biological roles of gene-specific codon usage biases. We used a hierarchical clustering analysis to identify groups of *Drosophila* genes that share similar codon usage patterns, which classified the genes into two major clusters (Clusters A and B; Figure 2.2). Further analysis of Cluster A revealed that these genes prefer the rare codons Lys-AAA, Glu-GAA, Gln-CAA, Phe-TTT, Tyr-TAT, and His-CAT, relative to their usage in Cluster B genes. Genes in Cluster B could be further divided into several smaller subclusters defined by the selective usage of specific combinations of codons. Together, these data indicate that in *D. melanogaster*, different functional classes of genes may have been shaped by selection for different patterns of biased codon usage, and therefore, different codon combinations might correspond to different biological functions. We also observed species-specific heterogeneous codon usage patterns across genes in the genomes of *A.*

thaliana, *S. cerevisiae*, *C. elegans*, and *H. sapiens* (Figures 2.3-2.6), which indicates that selection for the increased usage of specific combinations of codons might represent a fundamental element in genome architecture.

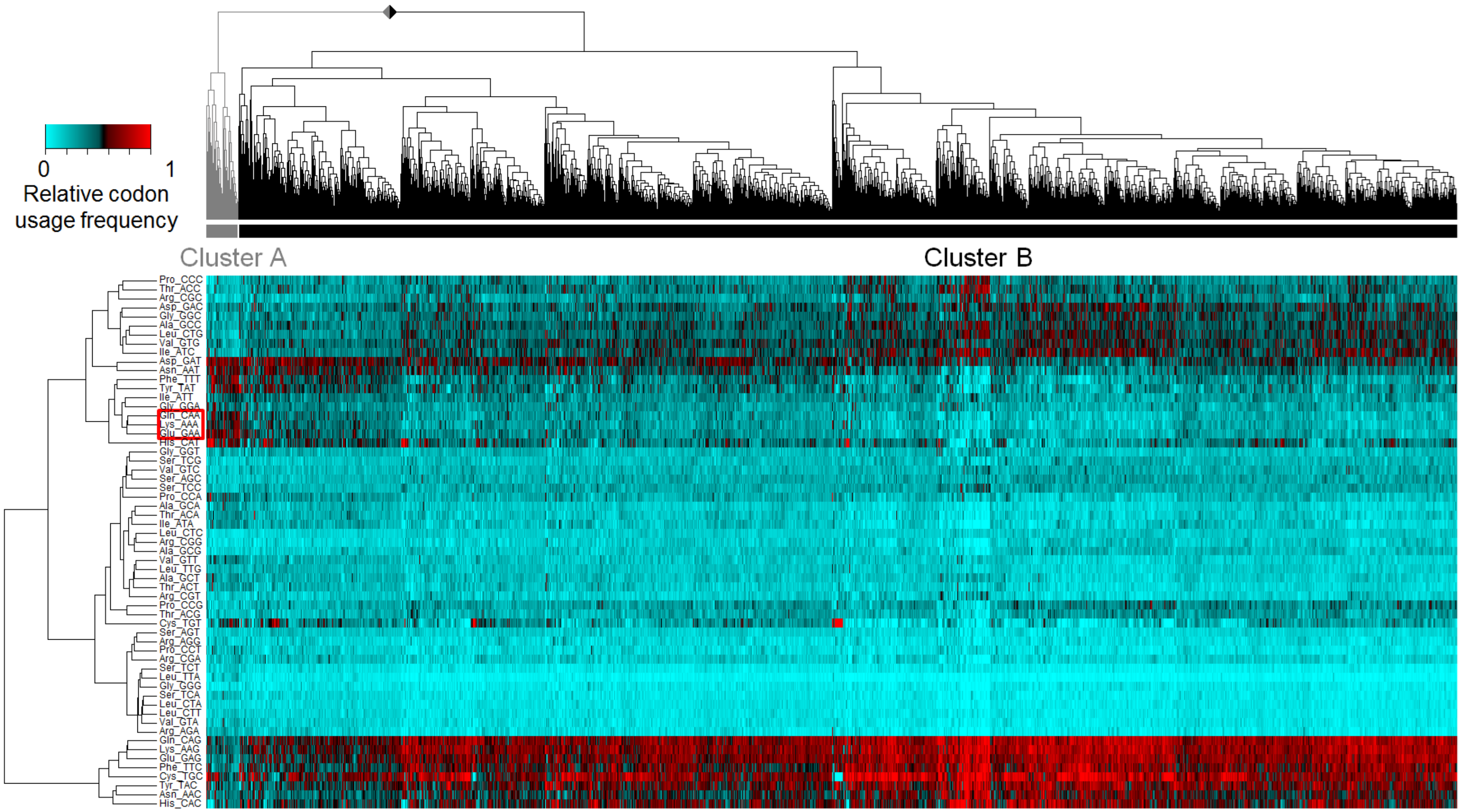

Figure 2.2: Clustered codon usage patterns across genes carrying signatures of selection on gene-specific codon usage bias in *D. melanogaster*. Hierarchical clustering of all genes (horizontal axis) identified as carrying significant signatures of selection on gene-specific codon usage bias. The relative usage frequencies of each codon in each gene are color-coded (vertical axis). Euclidean distances were used to measure dissimilarities between codons. Spearman's correlation coefficients were used to measure similarities between genes. Complete linkage was used as the clustering criterion. Clusters A (grey bar) and B (black bar) represent the two major gene groups identified.

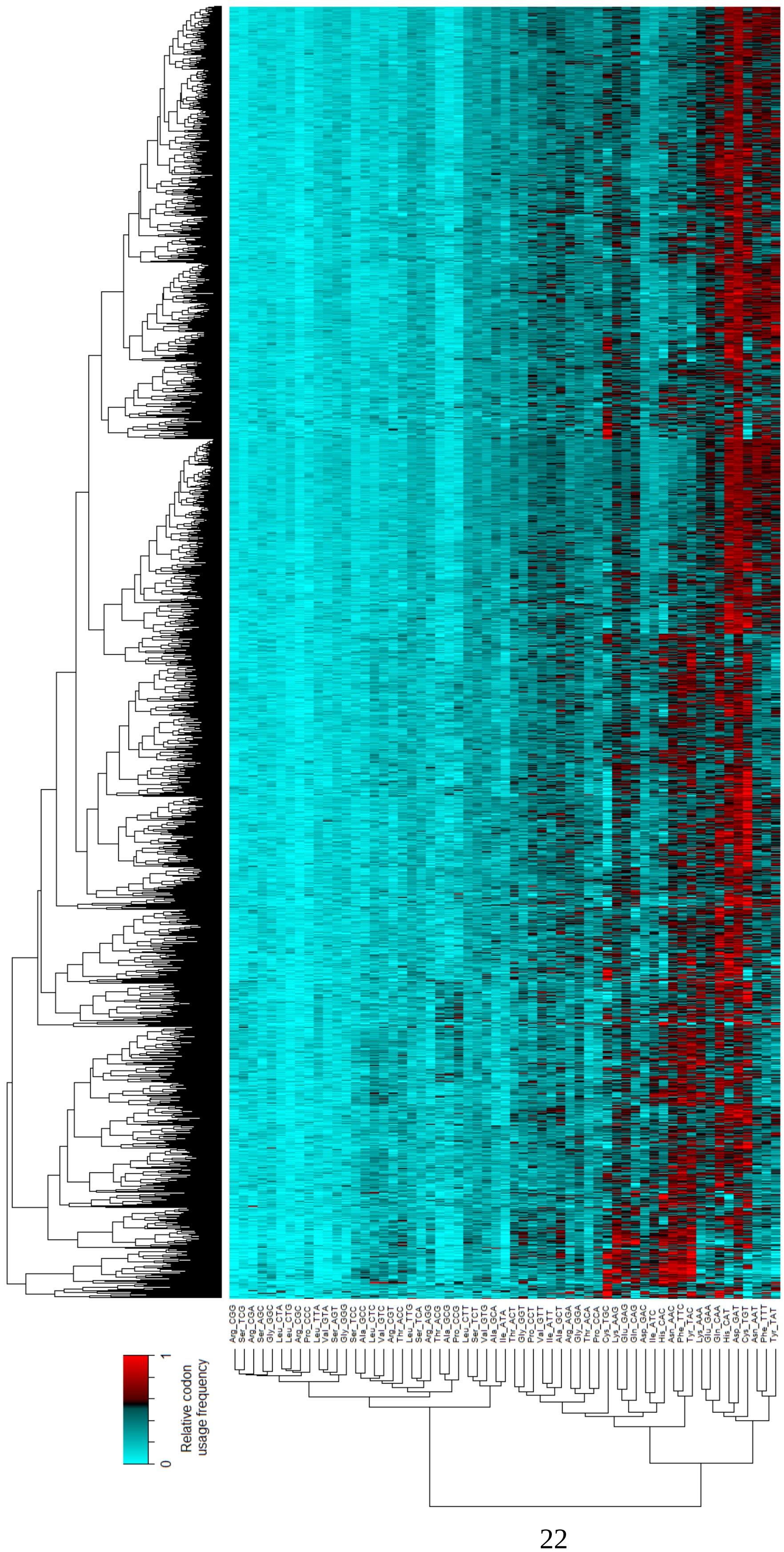

Figure 2.3: Clustered codon usage patterns across genes carrying signatures of selection on gene-specific codon usage bias in *A. thaliana*. Hierarchical clustering of all genes identified as under selection for gene-specific codon usage bias in *A. thaliana*. Analyses as in Figure 2.2.

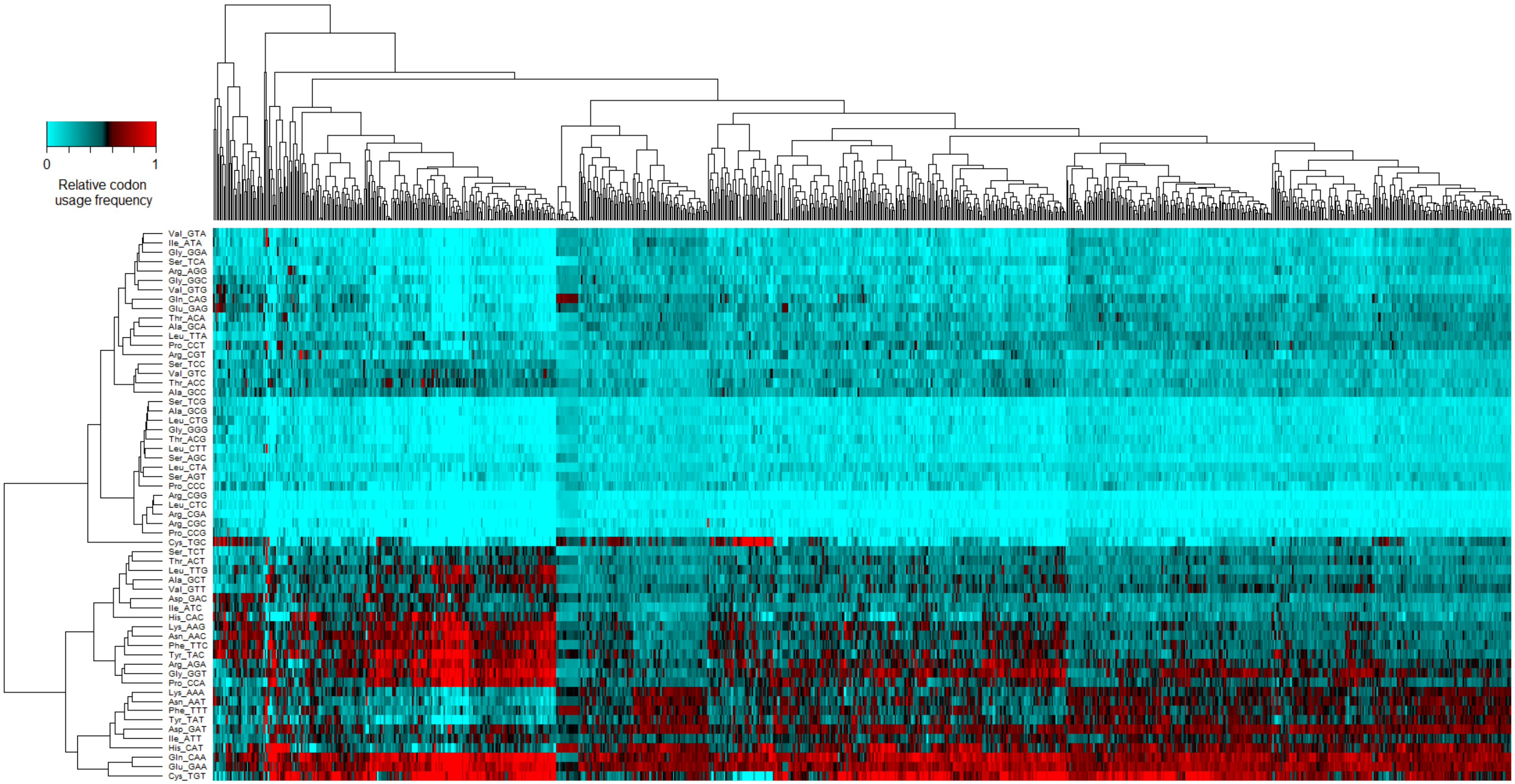

Figure 2.4: Clustered codon usage patterns across genes carrying signatures of selection on gene-specific codon usage bias in *S. cerevisiae*. Hierarchical clustering of all genes identified as under selection for gene-specific codon usage bias in *S. cerevisiae.* Analyses as in Figure 2.2.

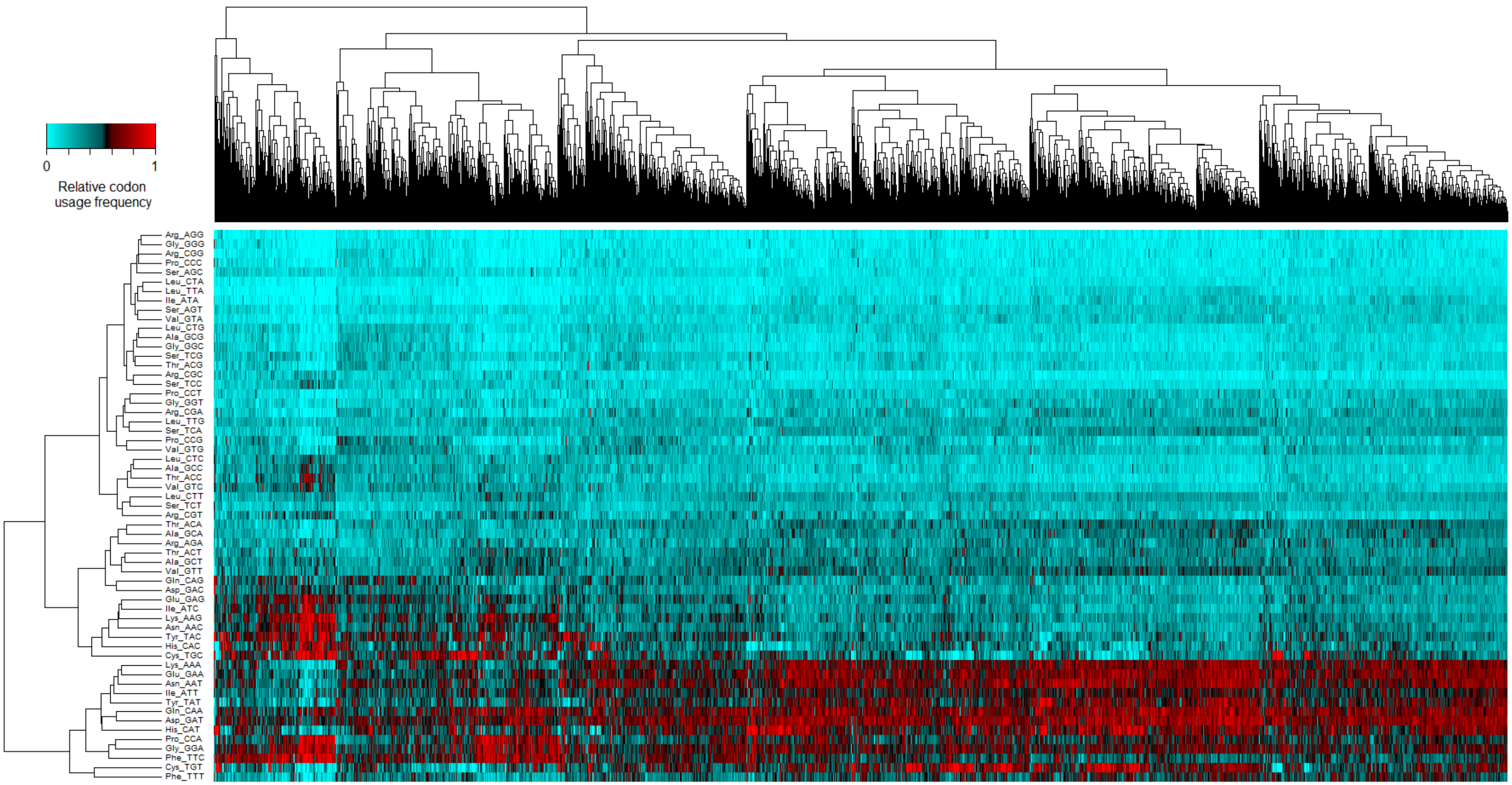

Figure 2.5: Clustered codon usage patterns across genes carrying signatures of selection on gene-specific codon usage bias in *C. elegans*. Hierarchical clustering of all genes identified as under selection for gene-specific codon usage bias in *C. elegans*. Analyses as in Figure 2.2.

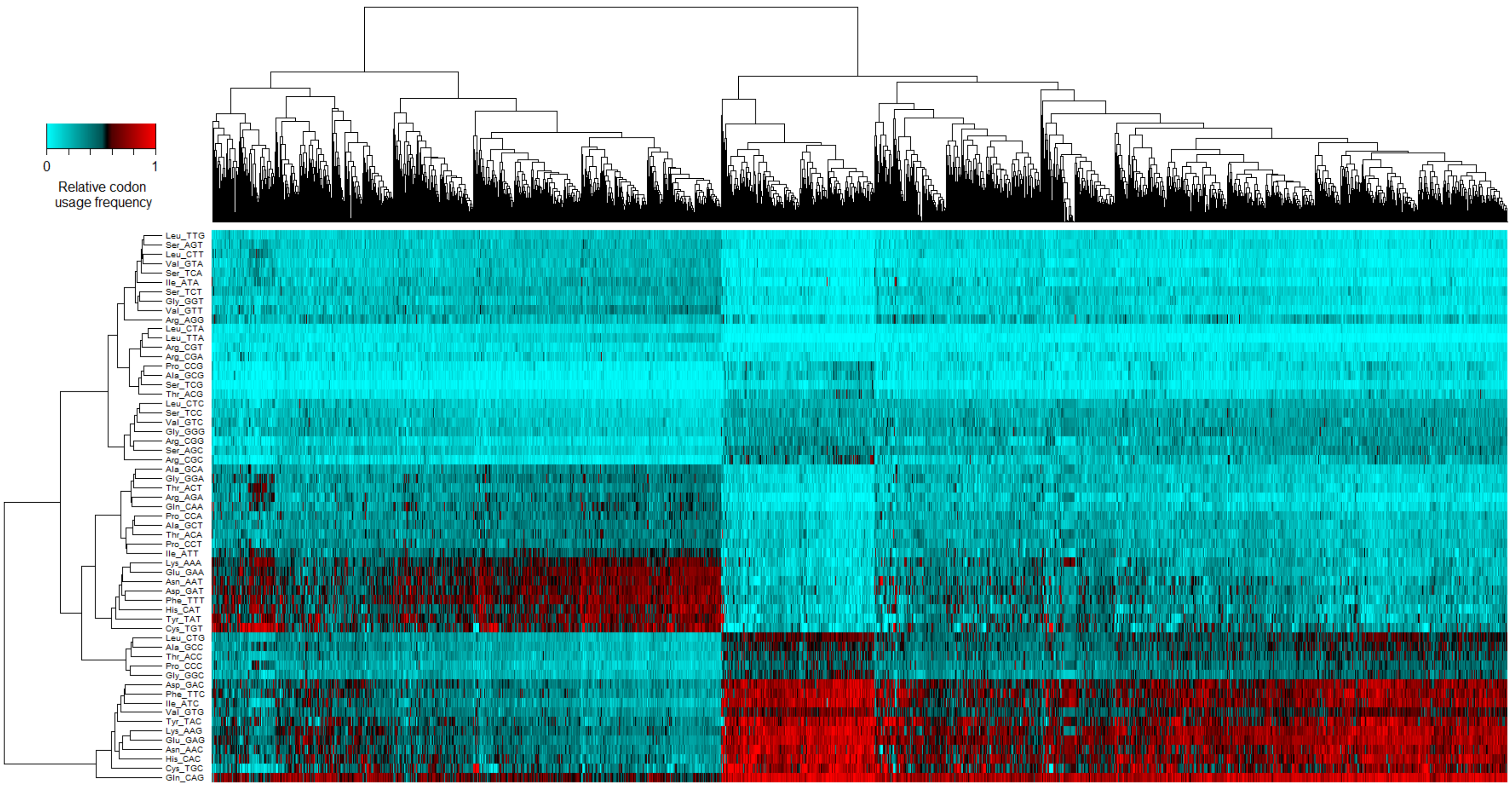

Figure 2.6: Clustered codon usage patterns across genes carrying signatures of selection on gene-specific codon usage bias in *H. sapiens*. Hierarchical clustering of all genes identified as under selection for gene-specific codon usage bias in *H. sapiens*. Analyses as in Figure 2.2.

Next, we hypothesized that clustered genes with similar codon usage patterns might share similar biological functions. To test this hypothesis, we first analyzed publicly available tissue-specific gene expression data in *D. melanogaster* [162], which revealed that genes that are preferentially expressed in the male accessory glands are significantly overrepresented in Cluster A and underrepresented in Cluster B (Table 2.1). Similarly, analysis of previously published data on sexually dimorphic genes [163,164] revealed that genes with male-enriched expression are significantly overrepresented in Cluster A and underrepresented in Cluster B (Table 2.2), which is consistent with previous observation that in *D. melanogaster*, non-sexually-dimorphic and female-enriched genes exhibit stronger usage bias towards common codons than genes with male-enriched expression [191]. Together, these data suggest that a specific combination of rare codons contributes to functions of some genes associated with the male reproductive system.

Table 2.1: Tissue-specific expression patterns of genes in Clusters A and B. Genes included in Clusters A and B are as in Figure 2.2. Significant over- or under-representation is shown in bold. N/A means that no tissue-specific genes were found in the entire genome or the focal gene cluster. Although present in the FlyAtlas database, we did not identify tissue-enriched genes in the thoracic-abdominal ganglion, virgin female spermatheca, inseminated female spermatheca, and adult fat body, and therefore, are not included here. Fold enrichment is relative to the entire genome. Bonferroni correction is applied, *p*< 0.05/20 = 0.0025.

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

Table 2.2: Sex-biased expression patterns of genes in Clusters A and B. Genes included are as in Table 2.1. Significant over- or under-representation is shown in bold. Fold enrichment is relative to the entire genome. Bonferroni correction is applied, $p < 0.05/4 = 0.0125$.

Gene cluster	Sex	Fold enrichment	*p*-value
Cluster A	Male	**1.66**	**<0.001**
	Female	1.67	0.091
Cluster B	Male	**0.61**	**<0.001**
	Female	1.07	0.06

To identify additional biological functions that might be affected by selection on gene-specific codon usage biases, we next used Gene Ontology (GO) analyses [192]. The most robust GO categories identified indicated that genes that encode extracellular matrix proteins are overrepresented, and genes that encode cytoplasmic proteins are underrepresented, in Cluster A (Table 2.3). We also found that genes annotated as encoding odorant-binding proteins, a class of secreted proteins, as well as other extracellular space proteins, are underrepresented in Cluster B (Table 2.3). Together, these data suggest that one possible common function for at least some of the genes in Cluster A is that they encode secreted proteins. These findings are in agreement with previous analyses of codon usage patterns in other eukaryotes, which revealed a similar trend of increased usage of rare codons in extracellular proteins [193,194]. Further analyses of the heterogeneous Cluster B revealed additional enriched GO terms, which may be associated with specific codon usage patterns of subclusters within Cluster B (Table 2.3 and Figure 2.2).

Table 2.3: Gene ontology analysis of gene clusters defined by codon usage patterns. Clusters A and B are defined by their codon usage patterns, and only genes carrying signatures of natural selection on synonymous codon usage are included, as shown in Figure 2.2. Molecular function, cellular component, and biological process GO terms are analyzed. Only significant over- or under-representations are shown. Fold enrichment is relative to the entire genome. We did not obtain informative results when analyzing biological process GO terms of Cluster A genes.

Gene cluster	Annotation category	GO term	Fold enrichment	*p*-value
Cluster A	Molecular function	extracellular matrix structural constituent (GO:0005201)	24.32	<0.001
	Cellular component	extracellular matrix (GO:0031012)	5.55	0.002
		cytoplasm (GO:0005737)	0.54	0.023
Cluster B	Molecular function	organic anion transmembrane transporter activity (GO:0008514)	1.54	0.006
		cofactor binding (GO:0048037)	1.5	0.005
		ATP binding (GO:0005524)	1.47	<0.001
		ATPase activity, coupled (GO:0042623)	1.44	<0.001
		active transmembrane transporter activity (GO:0022804)	1.43	0.006
		phosphotransferase activity, alcohol group as acceptor (GO:0016773)	1.34	0.018
		cation transmembrane transporter activity (GO:0008324)	1.33	0.017
		protein binding (GO:0005515)	1.18	<0.001
		odorant binding (GO:0005549)	0.43	<0.001
	Cellular component	apical part of cell (GO:0045177)	1.56	0.003
		cell cortex (GO:0005938)	1.53	0.002
		supramolecular fiber (GO:0099512)	1.47	0.015
		integral component of plasma membrane (GO:0005887)	1.37	<0.001
		cytoskeletal part (GO:0044430)	1.29	0.018

		cytosol (GO:0005829)	1.23	0.032
		endomembrane system (GO:0012505)	1.17	0.024
		nucleus (GO:0005634)	1.12	0.001
		macromolecular complex (GO:0032991)	1.11	0.006
		extracellular space (GO:0005615)	0.76	0.023
	Biological process	regulation of small GTPase mediated signal transduction (GO:0051056)	1.56	0.036
		axon guidance (GO:0007411)	1.51	<0.001
		actin cytoskeleton organization (GO:0030036)	1.47	0.004
		photoreceptor cell differentiation (GO:0046530)	1.47	0.038
		anion transport (GO:0006820)	1.46	0.024
		compound eye morphogenesis (GO:0001745)	1.45	<0.001
		central nervous system development (GO:0007417)	1.45	<0.001
		muscle structure development (GO:0061061)	1.44	0.002
		open tracheal system development (GO:0007424)	1.43	0.007
		imaginal disc-derived wing morphogenesis (GO:0007476)	1.43	<0.001
		regulation of developmental growth (GO:0048638)	1.41	0.012
		transmembrane transport (GO:0055085)	1.41	<0.001
		embryonic morphogenesis (GO:0048598)	1.4	0.029
		regulation of localization (GO:0032879)	1.4	<0.001
		cell migration (GO:0016477)	1.4	0.006
		carboxylic acid metabolic process (GO:0019752)	1.4	<0.001
		regulation of cellular component biogenesis (GO:0044087)	1.39	0.008

		regulation of anatomical structure morphogenesis (GO:0022603)	1.38	0.027
		negative regulation of cell communication (GO:0010648)	1.37	0.013
		negative regulation of signaling (GO:0023057)	1.37	0.013
		negative regulation of transcription, DNA-templated (GO:0045892)	1.37	0.023
		regulation of nervous system development (GO:0051960)	1.36	0.013
		ovarian follicle cell development (GO:0030707)	1.35	0.016
		regionalization (GO:0003002)	1.33	<0.001
		regulation of cell differentiation (GO:0045595)	1.33	0.025
		anatomical structure formation involved in morphogenesis (GO:0048646)	1.33	<0.001
		regulation of organelle organization (GO:0033043)	1.32	0.02
		response to abiotic stimulus (GO:0009628)	1.31	0.007
		regulation of biological quality (GO:0065008)	1.3	<0.001
		cell fate commitment (GO:0045165)	1.28	0.019
		behavior (GO:0007610)	1.28	0.001
		organic substance transport (GO:0071702)	1.27	0.002
		positive regulation of cellular process (GO:0048522)	1.24	<0.001
		macromolecule localization (GO:0033036)	1.23	0.023
		organic substance catabolic process (GO:1901575)	1.21	0.039
		phosphate-containing compound metabolic process (GO:0006796)	1.21	0.02
		signal transduction (GO:0007165)	1.21	<0.001
		cellular component assembly (GO:0022607)	1.2	0.012

		cellular macromolecule metabolic process (GO:0044260)	1.11	0.022
		organonitrogen compound metabolic process (GO:1901564)	1.11	0.016

2.3.3 Biased Gene-specific Codon Usage Contributes to the Regulation of Spatial Protein Expression Patterns

Our analyses revealed that Cluster A, defined by the increased usage of a specific combination of several rare codons, includes genes with putative male-biased functions (Tables 2.1-2.2). One possible explanation for the observed association between a specific codon usage pattern and tissue-enriched expression pattern is that the selected specific rare synonymous codons match tissue-specific tRNA pools, and thus the preferential usage of these specific rare codons seems to enhance protein translation efficiency in specific tissues or cell types [48,195,196]. Therefore, we next tested the hypothesis that the specific pattern of codon usage bias exhibited by Cluster A genes contributes to increased translation efficiency in the male reproductive system.

Our clustering analysis revealed that Cluster A genes preferentially use the rare codons Lys-AAA, Gln-CAA, Glu-GAA, Phe-TTT, Tyr-TAT, and His-CAT (Figure 2.2). However, only the first three rare codons are recognized by exactly matching tRNA anticodons, while the latter three rare codons share the same tRNAs with their more commonly used synonymous codons with a mismatch (wobble) at the third codon position [171,197]. Therefore, because the hypothetical impact of codon usage bias on the spatial regulation of protein translation depends on the increased availability of specific tRNAs in specific tissues and cell types, we next analyzed the specific usage of rare codons AAA, CAA, and GAA in all genes that show enriched expression in the male reproductive system, independent of whether these genes have passed the initial statistical threshold for the detection of selection on gene-specific codon usage patterns. We found that these specific rare synonymous codons are indeed overrepresented in male-reproductive-system-specific genes (Figure 2.7).

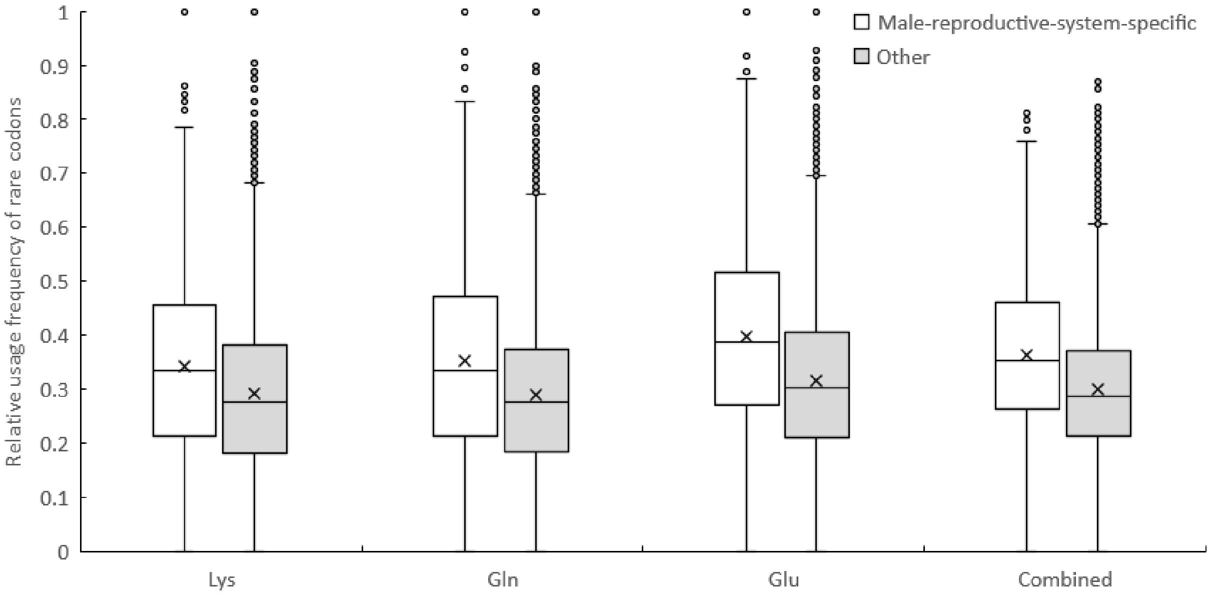

Figure 2.7: Rare codons Lys-AAA, Gln-CAA, and Glu-GAA are overrepresented in protein-coding sequences of genes with enriched expression in the male reproductive system. The relative usage frequencies of Lys-AAA, Gln-CAA, Glu-GAA, and the combined relative usage frequency of these codons weighed by the amino acid composition, were calculated for all genes with valid records in the FlyAtlas tissue-specific transcriptomic database. 1284 male-reproductive-system-specific genes and 9822 other genes were included. Data are presented in a box plot, where means are shown by crosses and medians are shown by solid lines in the middle of boxes. The relative usage frequencies and the combined relative usage frequency of the three rare codons are all significantly higher for male-reproductive-system-specific genes than for other genes (Mann-Whitney test, $p < 0.001$).

Next, we hypothesized that if the selection on an increased relative usage of rare codons AAA, CAA, and GAA serves a biological function that is specific to the male reproductive system then it should depend, at least in part, on the increased expression of the specific rare tRNAs that bind these specific codons. In accordance with this hypothesis, northern blot analyses of tissue-specific tRNA expression revealed that $tRNA^{Lys}_{TTT}$ is enriched in the male reproductive organs, further supporting the hypothesis that spatial regulation of some tRNA genes that correspond to rare codons could contribute to tissue-specific increase in protein translation efficiency (Figure 2.8A-C). The expression patterns of the tRNAs that match Glu-GAA and Gln-CAA were not investigated because their sequences are almost identical to their common tRNA counterparts, which does not allow their independent detection by hybridization probes [198].

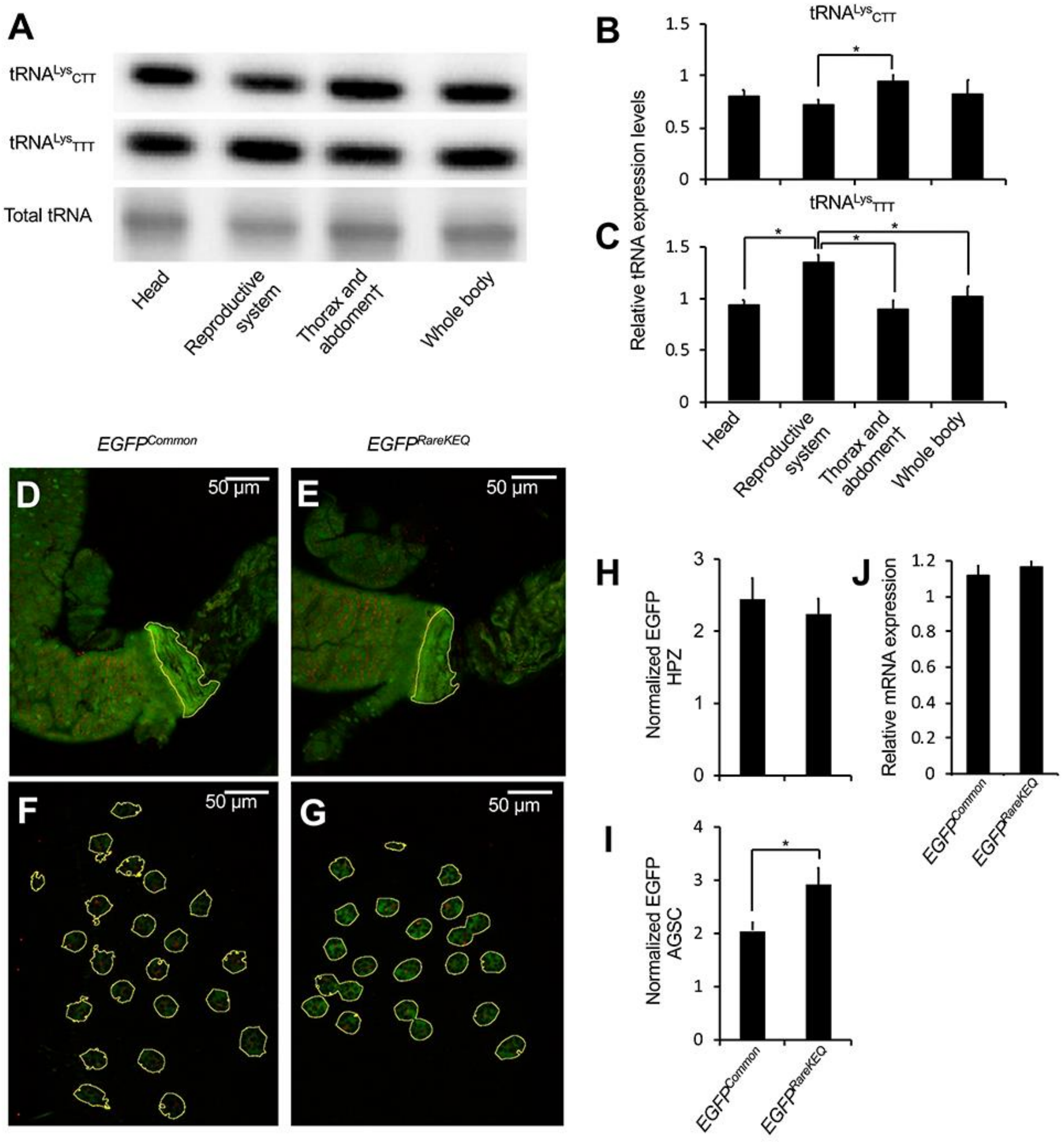

Figure 2.8: Gene-specific codon usage bias affects spatial protein expression in *D. melanogaster*. (**A**) Representative northern blots and (**B**-**C**) summary data of the relative transcript abundance of $\text{tRNA}^{\text{Lys}}_{\text{CTT}}$ (common) and $\text{tRNA}^{\text{Lys}}_{\text{TTT}}$ (rare) across male tissues. *, $p<0.05$ (n = 4, ANOVA followed by SNK *post hoc* tests). Error bars denote standard deviation. †: Reproductive system is excluded. (**D**-**E**) Representative images showing EGFP and RFP expression in the HPZ (yellow lines mark tissue boundaries). (**F**-**G**) Representative images showing EGFP and RFP expression in the AGSC (yellow outlines surround secondary cells). (**H**-**I**) Summary data of normalized EGFP signals in HPZ and AGSC. *, $p<0.01$ (n=5, two-tailed unpaired Student's *t*-test). Error bars denote standard deviation. (**J**) Real-time qRT-PCR mRNA expression data of $EGFP^{Common}$ and $EGFP^{RareKEQ}$ (two-tailed unpaired Student's *t*-test, n = 4, NS). Error bars denote SEM.

Observing increased expression of the rare tRNA that matches the rare codon Lys-AAA in the male reproductive system further suggested that some male-enriched *D. melanogaster* genes have evolved a biased codon usage pattern that restricts their efficient translation in a tissue-specific manner. We tested this hypothesis by generating transgenic flies that express different alleles of *EGFP* under the control of the UAS-GAL4 binary expression system [168]. One allele was comprised of common codons for all amino acids (*EGFP*^{Common}), while the other allele differed from *EGFP*^{Common} only by using rare codons AAA, GAA, and CAA for amino acids Lys, Glu, and Gln respectively (*EGFP*^{RareKEQ}), which represent about 18% of the total residues in EGFP. Each *EGFP* allele was also co-expressed with an *RFP* reporter encoded by common codons for all residues, which served as a transgene expression control. Both transgenes were driven by the ubiquitous *Act*5*C*-GAL4 driver [199]. To evaluate the effect of codon usage bias on the spatial pattern of both *EGFP* alleles *in vivo*, we compared the two genotypes by measuring the RFP-normalized EGFP signals in the accessory gland secondary cells (AGSCs), which are responsible for secreting seminal proteins, and in the hindgut proliferation zone (HPZ), which harbors gut epithelial stem cells [200]. We found that the signal from the *EGFP*^{RareKEQ} allele in the AGSCs was significantly higher than that of the *EGFP*^{Common} allele. In contrast, both alleles produced similar signals in the HPZ (Figure 2.8D-I). Because both alleles were expressed by the same GAL4 driver and the baseline mRNA expression levels of both alleles are similar (Figure 2.8J), these results suggest that this specific combination of rare codons enhances protein translation in the *D. melanogaster* male reproductive system. Together, these data indicate that in contrast to the widely accepted assumption that “optimal” translation rates depend on the selective usage of common synonymous codons, the actual optimal translation of some proteins associated with the male reproductive system depends on the selective usage of a specific combination of rare codons.

2.4 Discussion

The computational and empirical data we present here indicate that, in spite of variation in the proportion of protein-coding genes affected by selection on gene-specific codon usage bias across species, non-neutral synonymous mutations are likely much more common in eukaryotic genomes than currently assumed. Thus, the results of our unbiased and broadly applicable statistical approach suggest that gene-specific codon usage bias is a fundamental organizational principle of eukaryotic genomes that is sensitive to natural selection, and therefore, represents an important component of the genotype-phenotype axis on the developmental and physiological timesclaes, as well as the molecular evolution of genomes in diverse eukaryotic clades and biological contexts. Furthermore, our studies indicate that many of the previous reports about the phenotypic consequences of synonymous mutations are unlikely to represent anecdotal cases. Instead, they signify a fundamental aspect of the spatial regulation of eukaryotic gene activity.

Our genome-scale data also suggest that it should no longer be assumed *a priori* that most synonymous mutations are neutral. Instead, tests for selection on protein-coding genes should adjust their parameters according to the statistical probabilities of synonymous mutations being non-neutral [201]. Thus, common estimates for rates of molecular evolution of protein-coding genes, which rely on the assumption that synonymous mutations are neutral, are likely overestimating or underestimating the impact of natural selection. Consequently, the interpretations of the possible associations between specific genetic and phenotypic variations in particular [202], and genome evolution in general [37,203], could be biased. This is especially important in the context of the commonly used tests for identifying molecular signatures of selection in specific protein-coding genes, and for classifying specific modes of natural selection (e.g., “positive” *versus* “negative” selection) [70,101,107,157,204,205].

As currently framed, our quantitative approach for identifying gene-specific signatures of natural selection on synonymous mutations is based on a set of assumptions that enable its broad application to any genomes with available annotations of protein-coding sequences, so that the approach can assess the generality of neutral synonymous mutations. However, we acknowledge that some of these assumptions might not represent the correct biological context for all genes under all possible conditions. We also anticipate that the assumption that adjacent nucleotides have no effect on the mutation rates of focal nucleotides could increase false-positive rates [206]. However, we foresee that this possible source of false-positives would be compensated for by the assumption that the mutational bias of each focal gene is independent, which mathematically minimizes the χ^2 statistic calculated from the observed and expected codon usage patterns according to Equation (2.3). The risk for false-positive outcomes is further alleviated by the conservative use of relaxed constraints for estimating μ values [184,185]. Nevertheless, false-positive outcomes produced by our method could be further reduced by having more precise estimates of the μ parameters, which could be achieved by using population-level genetic variation data, sequences of short introns in or near the focal gene [207], more specific information about species-specific mutation rates, and how mutation rates might be impacted by the identities of adjacent nucleotides.

The specific biological reasons for why different combinations of codons have been selected for in the context of individual genes in various species remain mostly unknown. Nevertheless, our data indicate that in multicellular organisms, one of the reasons may be that natural selection optimizes codon usage patterns to match specific spatial constraints on gene function. We found that gene-specific codon usage bias is not uniform across the genome. Instead, our data suggest that different groups of genes in the same species often exhibit enrichment for specific combinations of rare or common codons, which suggests that different combinations of codons may have been selected for diverse biological reasons. In this regard, we show that at least one group of genes with enriched expression in the reproductive system of male *D. melanogaster* preferentially use the rare codons Lys-AAA, Gln-CAA, and Glu-GAA. By using allelic variants of a reporter gene, we show that the selective usage of these three rare codons is sufficient for generating spatially biased protein expression patterns. Therefore, "optimal" codon usage patterns of individual genes do not necessarily require the use of the most common codons for all amino acids; instead, some rare codons may serve as "facultative optimal codons", as the tRNAs perfectly matching them could be relatively abundant in specific cells and/or during a specific period of time. Although our findings were not formally stipulated by the "translational selection theory", they are consistent with its assertion that natural selection on codon usage bias can optimize translation by matching gene-specific codon usage patterns to cellular tRNA pools [1,38,195,196]. Consequently, our study provides a broader context to this fundamental evolutionary theory by emphasizing the possible role of gene-specific codon usage bias in the spatial regulation of proteins in multicellular eukaryotes. However, regulation of spatial protein expression cannot explain all cases of selection on gene-specific codon usage bias because genomes of some unicellular eukaryotes, such as the choanoflagellate *Monosiga brevicollis* and the green algae *Chlamydomonas reinhardtii*, also include a large proportion of protein-coding genes that exhibit signatures of selection on their codon usage patterns (Figure 2.1). While we do not understand yet the role of selection on heterogeneous gene-specific codon usage bias in unicellular eukaryotes, it is possible that in these organisms it contributes to the temporal regulation of gene expression and/or phenotypic responses to external stimuli.

Our studies also highlight the importance of understanding the relationships between tRNAs and protein-coding genes. Specifically, our studies demonstrate that for at least some *D. melanogaster* protein-coding genes, efficient tissue-specific translation requires an interaction between an increased usage of a specific set of rare codons, and the increased expression of their matching rare tRNAs. This is likely the result of the co-evolution between enriched expression of tRNAs and the usage patterns of their cognate codons in genes that require efficient translation in specific tissues. However, the mechanism that enables this observed co-evolution remains unknown. Although previous studies have argued that tRNA gene copy number is likely the primary mechanism that regulates the relative levels of individual tRNAs in cellular pools [38], variations in gene copy number alone cannot explain the tissue-specific expression patterns of some tRNAs observed by us (Figure 2.8A-C) and others [196,208]. Therefore, although tRNAs are thought to be exclusively transcribed by the constitutively-active RNA Pol III complex, there must be additional molecular mechanisms that enable the temporal and spatial regulation of some unique tRNAs at the transcriptional and/or post-transcriptional levels [209–212].

Although the majority of data presented here are from studies in *Drosophila*, we also show that many human protein-coding genes carry signatures of selection on their biased and heterogeneous codon usage patterns as well (Figure 2.1 and Figure 2.6). In addition, previous meta-analyses of human genome-wide association studies (GWAS) suggested that synonymous and nonsynonymous SNPs have similar likelihoods and effect sizes in terms of association with disease phenotypes [202]. Therefore, our findings that many protein-coding genes are under selection for codon usage bias could have broad implications for studies of genetic variants underlying quantitative phenotypes in human populations. Specifically, our analyses suggest that synonymous SNPs are not necessarily neutral as often assumed [152–157], and therefore, are likely to contribute to overall trait variations by directly impacting the functions of specific genes and their associated phenotypes in health and disease [213]. Consequently, as was first stated by Darwin in the *On the Origin of Species*, "*Variations neither useful nor injurious would not be affected by natural selection*", a genuine "neutral" mutation in a protein-coding DNA sequence should be defined by its impact on associated phenotypes in the context of fitness, independent of whether it is synonymous or nonsynonymous.

2.5 Materials and Methods

2.5.1 Genomic and Transcriptomic Data

Protein-coding DNA sequences were from Ensembl 89 (https://www.ensembl.org/) [166]. Reference coding sequences included in the analysis of gene-specific codon usage bias were chosen according to the following criteria: 1) The sequence length is a multiple of three. 2) The sequence uses standard genetic code. 3) For each gene, only the longest mRNA isoform was used for analysis. If there were multiple isoforms of the same length, then the first record shown in the FASTA file was used. 4) The encoded protein includes all 19 amino acids that have degenerate codons. For the amino acid serine, the two-fold and four-fold degenerate codon groups were treated as if they encoded two different amino acids. Transcriptomic data were from the FlyAtlas microarray database [162] and the modENCODE RNA-seq database [163,164].

2.5.2 Estimating μ and Expected Codon Counts

The relationships between μ values and codon counts are described by Equation (2.3) (Results section). Based on the standard genetic code and all possible combinations for synonymous nucleotide substitutions, we classified all degenerate codons into six categories. For each category, we used Equation (2.3) to generate a homogeneous linear equation system. Subsequently, we treated one of the *x* variables as a known parameter and analytically solved all other *x* variables. Finally, these solutions were used to calculate the expected codon counts of a protein-coding gene.

Category one – Codons for Arginine:

$$\begin{cases} x_{AGA} \times \mu_{A\to C} + x_{AGA} \times \mu_{A\to G} = x_{CGA} \times \mu_{C\to A} + x_{AGG} \times \mu_{G\to A} \\ x_{AGG} \times \mu_{A\to C} + x_{AGG} \times \mu_{G\to A} = x_{CGG} \times \mu_{C\to A} + x_{AGA} \times \mu_{A\to G} \\ x_{CGA} \times \mu_{C\to A} + x_{CGA} \times (\mu_{A\to C} + \mu_{A\to G} + \mu_{A\to T}) \\ \qquad\qquad = x_{AGA} \times \mu_{A\to C} + x_{CGC} \times \mu_{C\to A} + x_{CGG} \times \mu_{G\to A} + x_{CGT} \times \mu_{T\to A} \\ x_{CGG} \times \mu_{C\to A} + x_{CGG} \times (\mu_{G\to A} + \mu_{G\to C} + \mu_{G\to T}) \\ \qquad\qquad = x_{AGG} \times \mu_{A\to C} + x_{CGA} \times \mu_{A\to G} + x_{CGC} \times \mu_{C\to G} + x_{CGT} \times \mu_{T\to G} \\ x_{CGC} \times (\mu_{C\to A} + \mu_{C\to G} + \mu_{C\to T}) = x_{CGA} \times \mu_{A\to C} + x_{CGG} \times \mu_{G\to C} + x_{CGT} \times \mu_{T\to C} \\ x_{CGT} \times (\mu_{T\to A} + \mu_{T\to C} + \mu_{T\to G}) = x_{CGA} \times \mu_{A\to T} + x_{CGC} \times \mu_{C\to T} + x_{CGG} \times \mu_{G\to T} \end{cases} . \quad (2.4)$$

Category two – Codons for Leucine:

$$\begin{cases} x_{TTA} \times \mu_{T\to C} + x_{TTA} \times \mu_{A\to G} = x_{CTA} \times \mu_{C\to T} + x_{TTG} \times \mu_{G\to A} \\ x_{TTG} \times \mu_{T\to C} + x_{TTG} \times \mu_{G\to A} = x_{CTG} \times \mu_{C\to T} + x_{TTA} \times \mu_{A\to G} \\ x_{CTA} \times \mu_{C\to T} + x_{CTA} \times (\mu_{A\to C} + \mu_{A\to G} + \mu_{A\to T}) \\ \qquad\qquad = x_{TTA} \times \mu_{T\to C} + x_{CTC} \times \mu_{C\to A} + x_{CTG} \times \mu_{G\to A} + x_{CTT} \times \mu_{T\to A} \\ x_{CTG} \times \mu_{C\to T} + x_{CTG} \times (\mu_{G\to A} + \mu_{G\to C} + \mu_{G\to T}) \\ \qquad\qquad = x_{TTG} \times \mu_{T\to C} + x_{CTA} \times \mu_{A\to G} + x_{CTC} \times \mu_{C\to G} + x_{CTT} \times \mu_{T\to G} \\ x_{CTC} \times (\mu_{C\to A} + \mu_{C\to G} + \mu_{C\to T}) = x_{CTA} \times \mu_{A\to C} + x_{CTG} \times \mu_{G\to C} + x_{CTT} \times \mu_{T\to C} \\ x_{CTT} \times (\mu_{T\to A} + \mu_{T\to C} + \mu_{T\to G}) = x_{CTA} \times \mu_{A\to T} + x_{CTC} \times \mu_{C\to T} + x_{CTG} \times \mu_{G\to T} \end{cases} . \quad (2.5)$$

Category three – Four-fold degenerate codons:

$$\begin{cases} x_{pqA} \times (\mu_{A\to C} + \mu_{A\to G} + \mu_{A\to T}) = x_{pqC} \times \mu_{C\to A} + x_{pqG} \times \mu_{G\to A} + x_{pqT} \times \mu_{T\to A} \\ x_{pqC} \times (\mu_{C\to A} + \mu_{C\to G} + \mu_{C\to T}) = x_{pqA} \times \mu_{A\to C} + x_{pqG} \times \mu_{G\to C} + x_{pqT} \times \mu_{T\to C} \\ x_{pqG} \times (\mu_{G\to A} + \mu_{G\to C} + \mu_{G\to T}) = x_{pqA} \times \mu_{A\to G} + x_{pqC} \times \mu_{C\to G} + x_{pqT} \times \mu_{T\to G} \\ x_{pqT} \times (\mu_{T\to A} + \mu_{T\to C} + \mu_{T\to G}) = x_{pqA} \times \mu_{A\to T} + x_{pqC} \times \mu_{C\to T} + x_{pqG} \times \mu_{G\to T} \end{cases} . \quad (2.6)$$

Category four – Codons for Isoleucine:

$$\begin{cases} x_{ATA} \times (\mu_{A\to C} + \mu_{A\to T}) = x_{ATC} \times \mu_{C\to A} + x_{ATT} \times \mu_{T\to A} \\ x_{ATC} \times (\mu_{C\to A} + \mu_{C\to T}) = x_{ATA} \times \mu_{A\to C} + x_{ATT} \times \mu_{T\to C} \\ x_{ATT} \times (\mu_{T\to A} + \mu_{T\to C}) = x_{ATA} \times \mu_{A\to T} + x_{ATC} \times \mu_{C\to T} \end{cases}. \quad (2.7)$$

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

$$x_{pqC} \times \mu_{C\to T} = x_{pqT} \times \mu_{T\to C}. \quad (2.8)$$

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

$$x_{pqA} \times \mu_{A\to G} = x_{pqG} \times \mu_{G\to A}. \quad (2.9)$$

The above equations are analytically solved by SymPy [214] and the results are shown in Appendix 1.2. Using these solutions, with a given set of μ values and counts of amino acid residues, we can calculate the expected counts of synonymous codons. For example, the expected counts of codons for Lys can be calculated by

$$\begin{cases} Ex_{AAA} = y_{Lys} \times \frac{x_{AAA}}{x_{AAA}+x_{AAG}} = y_{Lys} \times \frac{x_{AAA}}{x_{AAA}+x_{AAA}\times\frac{\mu_{A\to G}}{\mu_{G\to A}}} = y_{Lys} \times \frac{1}{1+\frac{\mu_{A\to G}}{\mu_{G\to A}}} \\ Ex_{AAG} = y_{Lys} \times \frac{x_{AAG}}{x_{AAA}+x_{AAG}} = y_{Lys} \times \frac{x_{AAA}\times\frac{\mu_{A\to G}}{\mu_{G\to A}}}{x_{AAA}+x_{AAA}\times\frac{\mu_{A\to G}}{\mu_{G\to A}}} = y_{Lys} \times \frac{\frac{\mu_{A\to G}}{\mu_{G\to A}}}{1+\frac{\mu_{A\to G}}{\mu_{G\to A}}} \end{cases}, \quad (2.10)$$

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

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

$$\Theta = (\mu_{A\to C}, \mu_{C\to A}, \mu_{A\to G}, \mu_{G\to A}, \mu_{A\to T}, \mu_{T\to A}, \mu_{C\to G}, \mu_{G\to C}, \mu_{C\to T}, \mu_{T\to C}, \mu_{G\to T}, \mu_{T\to G}). \quad (2.11)$$

For a gene, to estimate μ values, we try to minimize the value of $\chi^2(\Theta)$ by changing the elements in Θ, using the sequential least squares programming (SLSQP) algorithm [215].

Since equations generated from Equation (2.3) form a system of homogeneous linear equations, it is meaningless to infer exact μ values from the minimization of $\chi^2(\Theta)$; rather, only the relative magnitudes of different μ values are important for calculating the expected codon counts later. As we used the assumption that the lowest μ value is at least 1/100 of the highest, we set the range of μ values between 0.001 and 0.1 during the minimization of $\chi^2(\Theta)$. For each gene, the minimum $\chi^2(\Theta)$ is used to calculate p-value. Since we assumed that reference genomic data represent a "wild type" genome, the procedure mentioned above was applied to each individual reference gene.

2.5.3 Codon Usage Heatmaps

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

$$f_{gd} = \frac{n_{gd}}{y_{ga}}, \text{(2.12)}$$

where n_{gd} is the count of d in g, and y_{ga} is the count of amino acid a encoded by d and its synonymous codons in g. It should be noted that codons for Ser are not treated as two codon groups in codon usage heatmaps. As the mechanism of recognizing stop codons is fairly different from recognizing other codons [216], and methionine and tryptophan respectively have only one codon, the analysis is restricted to the other 59 codons. Therefore, a 59-dimension vector B_g is used to describe the codon usage pattern of g,

$$B_g = \left(f_{g1}, f_{g2}, f_{g3}, \cdots, f_{gd}, \cdots, f_{g59}\right)^T. \text{(2.13)}$$

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

$$H_d = \left(f_{1d}, f_{2d}, f_{3d}, \cdots, f_{gd}, \cdots, f_{Md}\right). \text{ (2.14)}$$

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

The “gplots” package [217] in R was used to generate heatmaps. The method of hierarchical clustering was complete linkage with Euclidean distance measuring the dissimilarities between codons and Spearman’s correlation coefficient measuring the similarities between genes.

2.5.4 Identifying Genes with Tissue-specific Expression Patterns

Mean adult gene expression data from the FlyAtlas database were used to identify genes with tissue-specific expression patterns. A gene was classified as tissue-specific if its average mRNA level in a specific tissue was at least ten-fold to the tissue with the second highest mRNA level. To reduce redundancy, “Head” expression values were excluded from the analysis. Specifically, since both “Brain” and “Thoracic-abdominal ganglion” are parts of the central nervous system (CNS), and many CNS-specific genes are expressed in both places, we generated a merged “Adult central nervous system” category that included the highest mRNA level across these two original FlyAtlas categories for each analyzed gene.

2.5.5 Identifying Genes with Sex-biased Expression Patterns

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

2.5.6 Hypergeometric Tests for Gene Set Enrichment

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

2.5.7 Gene Ontology Analysis

Gene ontology (GO) analysis [192] of *D. melanogaster* genes was performed by using online tools (http://geneontology.org/). Category enrichments were determined by comparing term frequencies between each gene cluster and the whole genome, followed by a Bonferroni correction.

2.5.8 Animals

Fruit flies (*D. melanogaster*) were raised on corn syrup-soy food (Archon Scientific) at 25°C and 60% relative humidity with a 12-hour light/dark cycle. Custom gene synthesis was used to generate the cDNAs encoding EGFP and mCherry fluorescent proteins (IDT Inc., Iowa City IA). See Appendix 1.1 for sequences of the $mCherry^{Common}$, $EGFP^{Common}$, and $EGFP^{RareKEQ}$ alleles. Transgenic animals that express each allele under UAS control were generated by cloning each cDNAs into the EcoRI/NotI cloning sites of the pUASTattB plasmid [167]. Since the $EGFP^{RareKEQ}$ allele contains one internal EcoRI site, digestion time was shortened to less than 20 minutes to allow incomplete digestion. The UAS-RFP^{Common} transgene was inserted into a chromosome II landing site (Bloomington #24483), and the UAS-$EGFP^{Common}$ and UAS-$EGFP^{RareKEQ}$ transgenes were inserted into the same chromosome III landing site (Bloomington #24749) by using the ΦC31 integrase approach [167]. Double homozygotic lines UAS-RFP^{Common}; UAS-$EGFP^{Common}$ and UAS-RFP^{Common}; UAS-$EGFP^{RareKEQ}$ were generated and then crossed to *Act*5*C*-GAL4/*SM*6 (Duncan lab, Washington University). Unless specified, the wild type Canton-S strain was used in all molecular analyses.

2.5.9 Analyses of tRNA Gene Expression

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

2.5.10 Real-time qRT-PCR

The mRNA expression levels of reporter genes in whole four-day-old male flies were quantified by using real-time qRT-PCR, following previously published methods [218,219]. For *RpL*32, the forward primer was CACCAAGCACTTCATCCG, and the reverse primer was TCGATCCGTAACCGATGT. For $EGFP^{Common}$ and $EGFP^{RareKEQ}$, the forward primers were respectively AACTTCAAGATCCGCCACAAC and AACTTCAAAATCCGCCACAAC, while these alleles shared the same reverse primer GTGCTCAGGTAGTGGTTATCG.

2.5.11 Quantitative Reporter Gene Imaging

Male reproductive and gut tissues from four-day-old adult male flies that express either the $EGFP^{Common}$ or $EGFP^{RareKEQ}$ were dissected in chilled PBS and mounted for imaging on a Nikon A1Si laser scanning confocal microscope with a 20X oil objective (n=5 samples per genotype). All images were taken within 10 minutes of dissection. Single plane fluorescent images of the AGSC were used to estimate EGFP expression levels of each allele. Similar images of the HPZ were used as generic tissue controls. The NIS-Element Ar software was used to capture EGFP and RFP signals and generate channel-merged images. Normalized EGFP signals from each image were quantified using the Fiji image processing software [220]. The tissue-specific effect of the EGFP codon usage was analyzed by comparing the normalized EGFP signals in either the AGSC or HPZ between genotypes with a two-tailed unpaired Student's *t*-test.

Chapter 3: Codon Clusters with Biased Synonymous Codon Usage Represent Hidden Functional Domains in Protein-coding DNA Sequences

3.1 Abstract

Protein-coding DNA sequences are thought to primarily affect phenotypes via the amino acid sequences they encode. Yet, emerging data suggest that, although they do not affect protein sequences, synonymous mutations can cause phenotypic changes. Previously, we have shown that signatures of selection on gene-specific codons usage bias are common in genomes of diverse eukaryotic species. Thus, synonymous codon usage, just as amino acid usage pattern, is likely a regular target of natural selection. Consequently, here we propose the hypothesis that at least for some protein-coding genes, codon clusters with biased synonymous codon usage patterns might represent “hidden” nucleic-acid-level functional domains that affect the action of the corresponding proteins via diverse hypothetical mechanisms. To test our hypothesis, we used computational approaches to identify over 3,000 putatively functional codon clusters (PFCCs) with biased usage patterns in about 1,500 protein-coding genes in the *Drosophila melanogaster* genome. Specifically, our data suggest that these PFCCs are likely associated with specific categories of gene function, including enrichment in genes that encode membrane-binding and secreted proteins. Yet, the majority of the PFCCs that we have identified are not associated with previously annotated functional protein domains. Although the specific functional significance of the majority of the PFCCs we have identified remains unknown, we show that in the highly conserved family of voltage-gated sodium channels, the existence of rare-codon cluster(s) in the

nucleic-acid region that encodes the cytoplasmic loop that constitutes inactivation gate is conserved across paralogs as well as orthologs across distant animal species. Together, our findings suggest that codon clusters with biased usage patterns likely represent “hidden” nucleic-acid-level functional domains that cannot be simply predicted from the amino acid sequences they encode. Therefore, it is likely that on the evolutionary timescale, protein-coding DNA sequences are shaped by both amino-acid-dependent and codon-usage-dependent selective forces.

3.2 Introduction

In general, it is assumed that the primary function of a protein-coding sequence is to encode a specific sequence of amino acids whose biochemical properties determine the structure and functions of the encoded peptide. However, emerging data indicate that synonymous mutations, which do not affect amino acid sequences, can still have dramatic phenotypic impacts [6,21,30]. Thus, it has been hypothesized that some important factors affecting protein structures and functions are not simply encoded by amino acid residues but by nucleic-acid-level information, such as codon usage bias [6,172]. Therefore, just as a sequence of amino acids with a specific order and/or specific biochemical properties can form a protein domain that performs specific functions, it is also possible that a sequence of codons with a specific codon usage pattern could serve as a nucleic-acid-level domain that affects the functions of the mature protein.

Based on the hypothesis that codon-usage-encoded domains can affect protein functions, researchers have identified rare-codon clusters, characterized by enriched whole-genome rare codons in relatively short regions within protein-coding sequences, that possibly decelerate translation and thus modify protein functions by affecting co-translational folding and/or modifications of nascent peptide chains [24,172,221–224]. Nevertheless, if functional codon clusters do exist, local deceleration of translation may not be the only mechanism through which they affect protein functions. It is also possible that functional codon clusters could correspond to locally accelerated translation, a specific combination of translationally decelerated and accelerated regions, specific RNA secondary structures [8,225], and accessibility of miRNAs [27]. Thus, for generally investigating codon clusters as functional domains that may be "hidden" from the amino acid sequences, exclusive focus on rare-codon clusters may lead to biased results. Therefore, it is necessary to develop statistical methods that generally detect putatively functional codon clusters (PFCCs), no matter what specific codons they prefer or through what mechanisms they may affect protein functions.

Consequently, to identify PFCCs, we developed a conservative statistical approach and applied it to the *Drosophila melanogaster* genome with approximately 14,000 protein-coding genes, which yielded over 3,000 PFCCs in about 1,500 genes. Interestingly, some of these PFCCs strongly prefer common codons while some others adopt complex codon usage patterns that cannot be simply described as preference for common or rare codons, which has not been reported before. Furthermore, we found that genes encoding transmembrane proteins are more likely to bear PFCCs. However, only a small proportion of the identified PFCCs are associated with the coding sequences of transmembrane helices, which suggests that PFCCs are either associated with other types of protein domains that are overrepresented in transmembrane proteins or not necessarily associated with amino-acid-encoded domains. We further found that the majority of the identified PFCCs are not associated with established protein domains in the Pfam database [169]. These data suggest that most PFCCs likely encode "hidden" nucleic-acid-level functional domains that cannot be predicted solely from amino acid sequences. The rationale for this inference is as follows: first, Pfam is a well-established database of conserved protein domains that have undergone strong natural selection; second, the PFCCs can be identified only when natural selection on local codon usage patterns is strong enough to generate statistically detectable signals; third, if the major impacts of PFCCs on gene functions are mediated by amino-acid-encoded protein domains, most PFCCs are expected to be associated with amino-acid-encoded domains that have undergone strong natural selection; fourth, the actual observation contradicts the expectation, and thus the functions of PFCCs should not be strongly associated with amino-acid-encoded domains. Finally, by implementing comparative analysis between homologs, we showed that the family of voltage-gated sodium channels likely evolved conserved preference for rare codons in a region responsible for the channel inactivation. Together, our data suggest that similar to amino acid sequences, codon clusters can also encode diverse functional domains, which provides an additional level of regulation over the structures, modifications, and functions of proteins.

3.3 Results

3.3.1 Identifying Putatively Functional Codon Clusters (PFCCs)

If the synonymous codon usage of a codon cluster does not perform specific functions, it should not be affected by natural selection and thus it can be explained by the background codon usage frequencies, which is mainly determined by mutations and genetic drift [174,226]. For example, if a gene locates in a GC-enriched chromosomal region that has resulted from GC-biased mutations, it is expected that the background codon usage is biased towards GC-ended codons; thus, if a sub-genic region is not significantly affected by natural selection on codon usage, its synonymous codon usage should also be biased towards GC-ended codons. Therefore, if the codon usage pattern of a codon cluster cannot be explained by the background codon usage frequencies, it should be significantly affected by natural selection; thus, such a codon cluster is by definition a PFCC. To identify PFCCs, first we needed to choose background codon usage frequencies. Previous studies on synonymous codon usage usually used the whole-genome codon usage frequencies as the background [221–224]. Nevertheless, our recent study [227] showed that gene-specific codon usage pattern can be fairly different from the whole-genome one. Thus, even if the synonymous codon usage of a codon cluster cannot be explained by whole-genome codon usage, it may still be adequately explained by gene-specific codon usage, and *vice versa*. Therefore, to filter out the interference from the discrepancy between whole-genome and gene-specific codon usage patterns so that PFCCs are conservatively identified, neither whole-genome nor gene-specific codon usage frequencies should be able to explain the codon usage pattern of a PFCC. Based on the aforementioned logic, we developed a statistical approach to scan protein-coding sequences in order to identify PFCCs (see 3.5.2 Identifying PFCCs).

By applying the approach to 13,821 protein-coding genes from the reference *D. melanogaster* genome, we identified 3,050 PFCCs in 1,445 genes (Appendix 3.1). This result indicates that PFCCs do exist, and they impact at least 10% of protein-coding genes in the *D. melanogaster* genome.

3.3.2 Codon Usage Patterns of PFCCs Are Diverse

In principle, the codon usage patterns of PFCCs can deviate from the background codon usage frequencies for various non-mutually exclusive biological reasons. First, the enrichment of rare codons in a PFCC might decelerate translation [224]. Second, it is possible that the enrichment of common codons in a PFCC could accelerate translation. Third, PFCCs with more complex codon usage patterns, which cannot be simply described as the preference for common or rare codons, might serve important functions by affecting mRNA secondary structure [8,225], miRNA accessibility [27], or epigenetic modifications [46]. Thus, classifying the identified

PFCCs by their codon usage patterns could be informative for assessing how PFCCs may affect protein functions.

Codon adaptation index (CAI) [3] has been widely used to describe a protein-coding sequence's propensity of using common codons. In general, a higher CAI indicates stronger preference for common codons and/or avoidance of rare codons. However, directly using CAI as the index to classify PFCCs could lead to biased results, especially when common codons are not enriched in the PFCCs. This is because the differences between usage frequencies of the synonymous codons for some amino acids are much larger than those of other amino acids. Thus, even if two codon clusters both strictly use rare codons, they could have very different CAIs depending on the amino acid sequences. To circumvent such a weakness of CAI, we propose to use a transformed CAI (TCAI) to describe the general codon usage pattern of a PFCC.

TCAI is calculated as below. For a PFCC, the corresponding amino acid sequence and the background codon usage pattern – either the whole-genome or gene-specific codon usage pattern – are used to randomly generate 10,000 "pseudo-clusters" of codons that encode exactly the same amino acid sequences as what is encoded by the PFCC. Thus, on average, the overall codon usage patterns of these pseudo-clusters should follow the background codon usage pattern. Then the CAIs of all pseudo-clusters are calculated, and TCAI is defined as the result of subtracting the proportion of pseudo-clusters whose CAIs are higher than the CAI of the PFCC from the proportion of pseudo-clusters whose CAIs are lower than the CAI of the PFCC. Thus, TCAI varies between -1 and 1. TCAI = -1 means that the CAIs of all pseudo-clusters are higher than that of the PFCC, suggesting that the PFCC strongly prefers rare codons; in contrast, TCAI =1 suggests that the PFCC strongly prefers common codons. Thus, TCAI effectively suppresses the interference from different levels of codon usage biases for different amino acids.

We calculated TCAIs for all identified PFCCs, either by using whole-genome (Figure 3.1A) or gene-specific (Figure 3.1B) codon usage pattern as the background. The distribution of TCAI values (Figure 3.1) indicates that most PFCCs are rare-codon clusters, while common-codon clusters do exist as shown by a small peak in the rightmost part of the histograms. More interestingly, there are also some codon clusters whose TCAI values are intermediate, suggesting that their codon usage patterns are more complex and cannot be simply described by strong preference for common or rare codons. The preponderance of rare-codon clusters may be explained by two reasons that are not mutually exclusive. First, the preponderance may represent the fact that rare-codon clusters are biologically more important than other types of functional codon clusters. Second, the preponderance may also be partly an artifact of technically easier detection of enriched rare codons in a short nucleotide sequence. Nonetheless, it was undoubtedly confirmed that there are different types of codon clusters in terms of synonymous codon usage patterns.

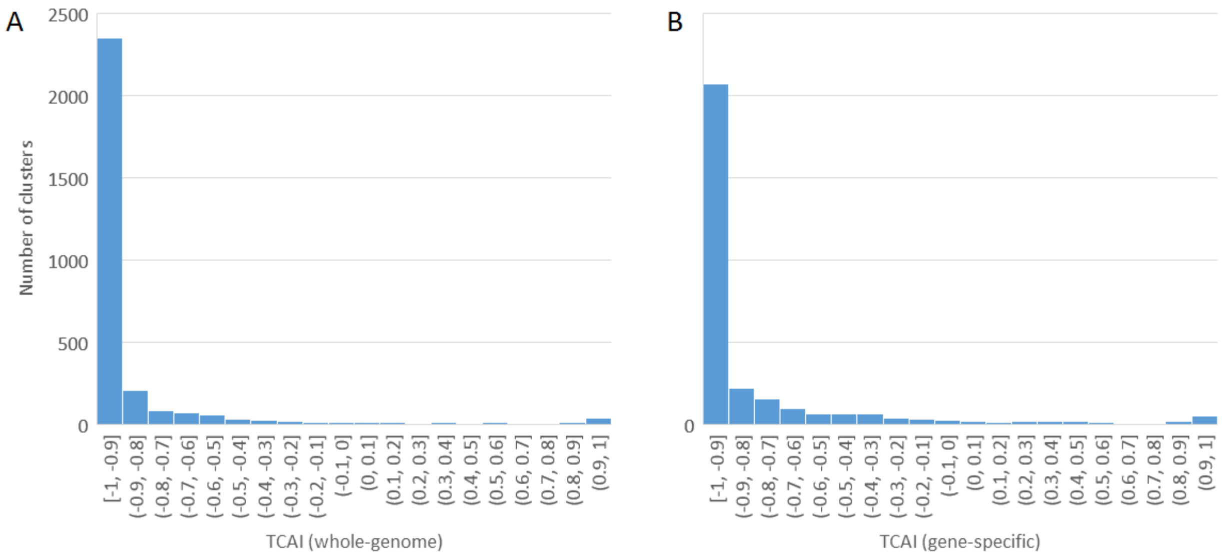

Figure 3.1: Distribution of TCAI values. TCAI values were calculated by using the whole-genome (**A**) or gene-specific (**B**) codon usage patterns as the background codon usage. The TCAI of a rare-codon cluster is near -1, while that of a common-codon cluster is near 1.

We also noted that although the distribution patterns shown in Figure 3.1A and Figure 3.1B are qualitatively similar, the actual values of corresponding columns in the histograms are quantitatively different. This suggests the possibility that a PFCC could be assigned to different types of codon clusters, depending on which background codon usage pattern is used. Such a possibility may interfere the interpretations of the putative functions of the PFCC. For example, a rare-codon cluster in terms of whole-genome codon usage may be classified as a common-codon cluster in terms of gene-specific codon usage, and thus it could be unclear whether the PFCC may decelerate or accelerate translation. In order to assess the influence of the discrepancy between whole-genome and gene-specific codon usage patterns on the classification of PFCCs, we used a scatter plot to examine the relationship between whole-genome TCAI and gene-specific TCAI (Figure 3.2). The data points were then clustered by K-mean clustering to seven types (K=7).

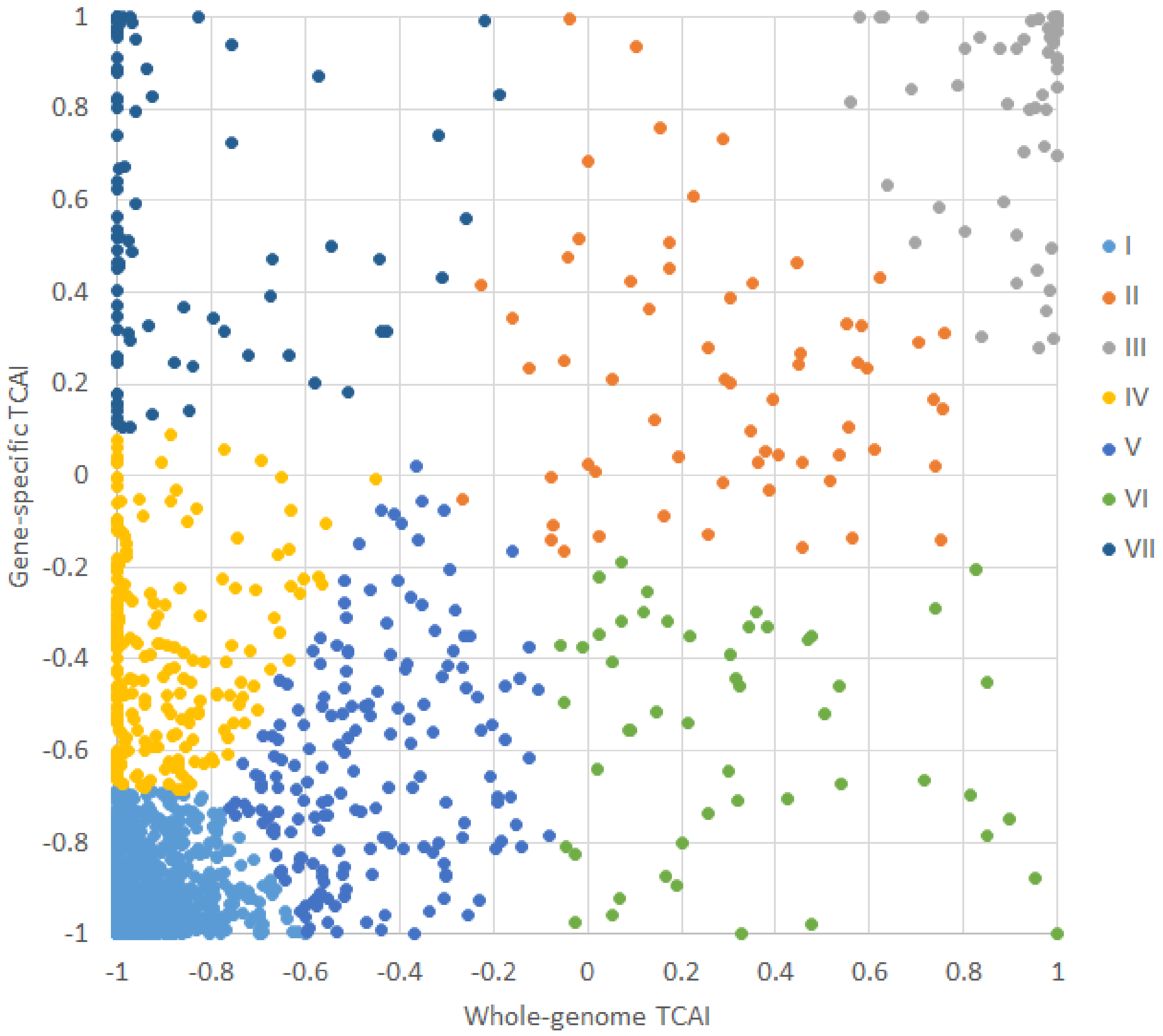

Figure 3.2: Influence of the discrepancy between whole-genome and gene-specific codon usage patterns on classifying PFCCs. Codon usage patterns of identified PFCCs were described by TCAI. Since gene-specific and whole-genome-level TCAI values for the same codon cluster could be different, we plotted the gene-specific TCAI against whole-genome TCAI for all codon clusters and then classified codon clusters by K-mean clustering (K=7).

We found that most codon clusters have similar whole-genome and gene-specific TCAI (Figure 3.2, types I-V). However, some common-codon clusters in terms of whole-genome TCAI were classified as rare-codon clusters in terms of gene-specific TCAI (Figure 3.2, type VI), and *vice versa* (Figure 3.2, type VII). This result suggests that due to the discrepancy between whole-genome and gene-specific codon usage patterns, it is difficult to predict the exact biological roles of some identified PFCCs. For example, in our previous study, we showed that some whole-genome rare codons can be translationally optimal for tissue-specific genes [227]. Thus, a rare-codon cluster in terms of whole-genome codon usage, which would be naïvely considered as a "decelerating codon cluster", might be a common-codon cluster in terms of gene-specific codon usage, which could actually serve as an "accelerating codon cluster". Therefore, although PFCCs can be detected by statistical approaches proposed by us and others [223,224], to computationally predict the candidate functional roles of these codon clusters may require extra information such as tRNA expression profile and better tools for predicting the secondary and tertiary structures of RNAs.

To summarize, PFCCs are diverse according to their codon usage patterns. Rare-codon clusters, whose main function is presumably decelerating translation [223,224], seem to be the majority of PFCCs. There are also other types of PFCCs, including common-codon clusters and PFCCs with more complex codon usage patterns, which likely have functions other than decelerating translation. Nonetheless, the discrepancy between whole-genome and gene-specific codon usage patterns makes it hard to predict the possible functions of the PFCCs whose whole-genome TCAI and gene-specific TCAI are dramatically different.

3.3.3 PFCC Distribution Is Not Restricted to Specific Regions of Protein-Coding Sequences

Except for the codon usage patterns of PFCCs, the locations of PFCCs in protein-coding sequences may also provide hints to the possible functions of PFCCs. Previous studies have shown that a potential important function of codon clusters is that N-terminal rare-codon clusters could affect secretion of proteins [48,193,194], possibly via interaction with the nascent chains of signal peptides [48,193]. Therefore, we next tested if the PFCCs detected by our approach tend to locate near the N-terminus; if they do, it could suggest that PFCCs are likely associated with secretion of proteins.

To measure how close a PFCC-encoded region is to the N-terminus, we defined the relative location index (RLI) of a PFCC as the ratio of the distance between the midpoint of the PFCC-encoded region and the N-terminus to the length of the entire protein. Thus, a small RLI means that the PFCC-encoded region is close to the N-terminus. We then plotted the distribution of PFCCs against their RLIs (Figure 3.3A). We found that although the density of PFCCs is apparently higher in the N-terminal region, the distribution of PFCCs is not restricted to this region (Figure 3.3A). As we have assigned these codon clusters to seven types (Figure 3.2), we also examined if some specific types of PFCCs exhibit skewed distribution towards the N-terminal region (Figure 3.3B-H). As expected, type I codon clusters, which can be described as rare-codon clusters, exhibit slight enrichment near the N-terminus (Figure 3.3B). To our surprise, type III codon clusters, which can be described as common-codon clusters, exhibit relatively strong enrichment near the N-terminus (Figure 3.3D). Other types of codon clusters do not exhibit clear enrichment near the N-terminus. We also performed a gene ontology (GO) analysis [165,228] (http://geneontology.org/) on the genes carrying N-terminal codon clusters ($RLI < 0.1$) to see if the genes encoding secreted proteins are enriched. We found that not only some extracellular matrix structural constituents, mostly mucins, are enriched, but also proteins associated with plasma membrane or transcription-level regulation are enriched (Appendix 3.2).

Together, these data indicate that although N-terminal regions are more likely to harbor PFCCs, many PFCCs actually locate in other regions (Figure 3.3). They also suggest that although the function of a subset of the PFCCs may be explained by N-terminal rare-codon clusters' impact on secretion or signal peptides, such a function is unlikely a general role played by other PFCCs. For example, the codon clusters locating in the middle of genes should have little to do with signal peptides. Thus, PFCCs likely perform various biological functions that need further investigation.

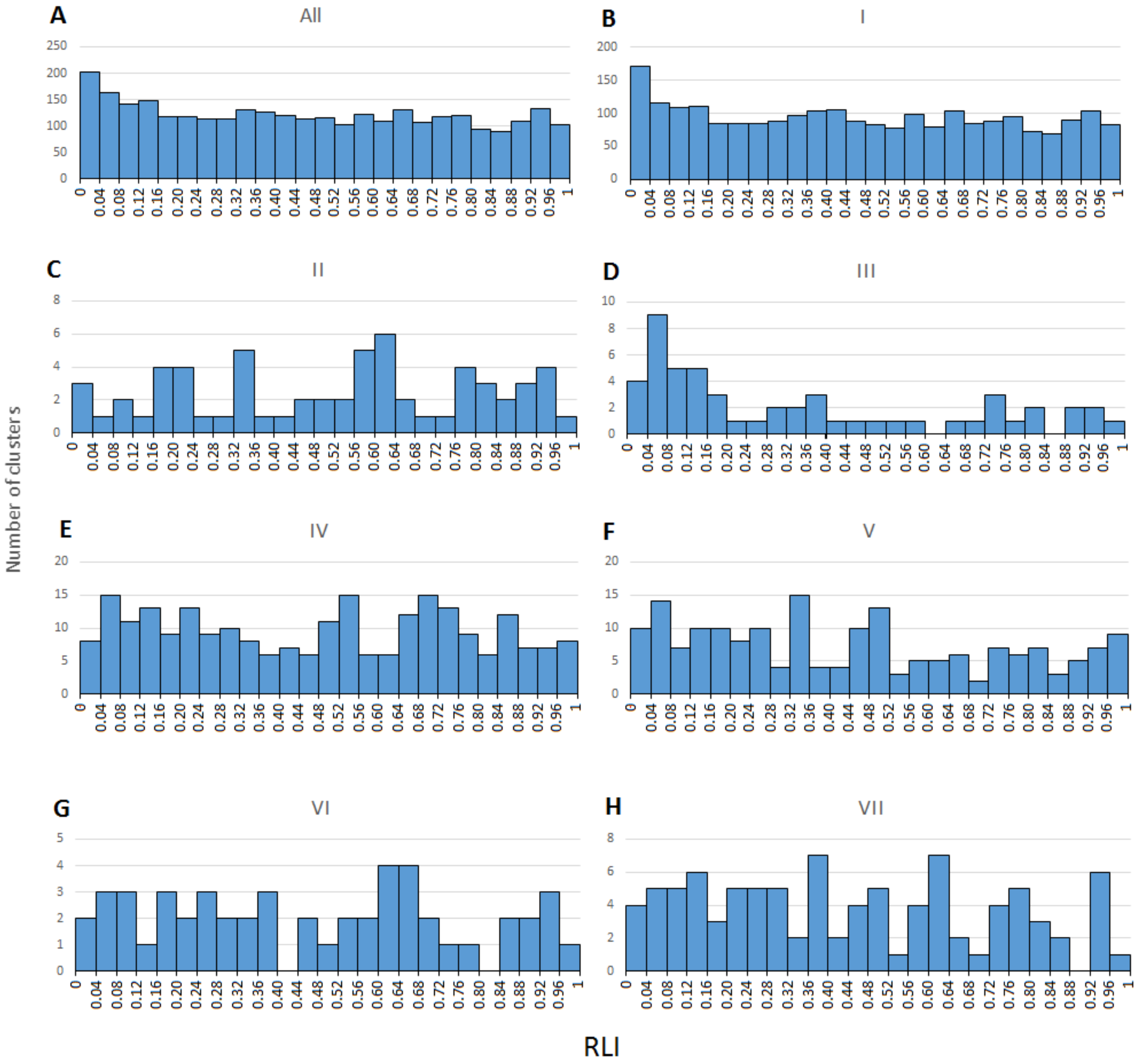

Figure 3.3: Spatial distribution of putatively functional codon clusters. For all identified PFCCs and each type of PFCCs shown in Figure 3.2, the distribution of PFCCs is plotted against the location coordinates, measured by RLI (RLI=0 means N-terminus; RLI=1 means C-terminus).

3.3.4 Specific Protein Functional Classes Are Overrepresented in Genes Carrying PFCCs While Most PFCCs Are Not Associated with Known Protein Domains

To further investigate the biological roles of PFCCs, we next performed GO analyses on the genes carrying PFCCs, in order to test the hypothesis that PFCCs are associated with various functional features of protein-coding genes. We found that in all 1445 genes that carry the PFCCs, genes encoding membrane-binding proteins and transcription-related proteins are overrepresented, while genes encoding ribosomal and mitochondrial proteins are underrepresented (Appendix 3.3). This result suggests that functional codon clusters might be associated with transmembrane domains, so we then tested if the amino acid sequences encoded by the PFCCs are near or overlapped with the transmembrane helices predicted by TMHMM [229]. Unexpectedly, we found that only about 6% of the PFCCs are near or overlapped with some transmembrane helices (Table 3.1, Appendix 3.4). Thus, there seems to be a discrepancy between the overrepresentation of transmembrane proteins in the genes carrying PFCCs and relatively few PFCCs that are near or overlapped with the sequences encoding transmembrane helices. Nevertheless, such a discrepancy could be explained by that PFCCs may be functionally more important for the non-transmembrane regions in transmembrane proteins. The discrepancy may also be explained by that transmembrane helices are less sensitive to the change in codon usage since the helices are strongly affected by the biochemical properties, such as hydrophobicity, of amino acid residues [229,230].

Table 3.1: Biased codon clusters overlap with transmembrane helices. A codon cluster is defined to be associated with a transmembrane helix if the distance between at least one amino acid residue of the helix and the closest residue encoded by the codon cluster does not exceed 20 amino acids.

	Association type		Number of clusters		
Clusters associated with transmembrane helices	1-to-1 association	cluster in helix	14	115	195
		helix in cluster	16		
		helix overlap left of cluster	21		
		helix overlap right of cluster	24		
		helix upstream to cluster	20		
		helix downstream to cluster	20		
	1-to-multiple association		80		
All clusters			3050		

If most PFCCs are not associated with transmembrane helices, then it is possible that PFCCs are associated with other types of protein domains. Consequently, we examined the association between PFCCs and annotated protein domains in the Pfam database [169,231]. We found that about 1/4 of the PFCCs are near or overlapped with some annotated Pfam protein domains, yet it is still unclear how the other 3/4 might influence protein functions (Table 3.2, Appendix 3.5). Among the PFCCs of which each is associated with only one Pfam protein domain, about 1/2 locate within protein domains (Table 3.2), which was consistent with what was recently reported by Chaney et al. [224]. These data suggest that although some PFCCs likely affect protein functions by modifying the co-translational processes concerning protein domains defined by amino acid sequences, the majority of PFCCs seem to be associated with unknown functional domains.

To summarize, although specific protein functional classes are overrepresented in the genes carrying PFCCs, most of the PFCCs are not associated with known protein domains defined by amino acid sequences. Therefore, PFCCs likely represent “hidden” nucleic-acid-level domains that regulate protein functions.

Table 3.2: Biased codon clusters overlap with Pfam domains. A codon cluster is defined to be associated with a Pfam domain if the distance between at least one amino acid residue of the Pfam domain and the closest residue encoded by the codon cluster does not exceed 20 amino acids.

	Association type		Number of clusters		
Clusters associated with Pfam domains	1-to-1 association	cluster in domain	299	584	746
		domain in cluster	3		
		domain overlap left of cluster	63		
		domain overlap right of cluster	75		
		domain upstream to cluster	58		
		domain downstream to cluster	86		
	1-to-multiple association		162		
All clusters			3050		

3.3.5 Voltage-gated Sodium Channels Include a Conserved Rare-codon Cluster Associated with the Inactivation Gate

To identify possible specific functions of some PFCCs, we next investigated PFCCs identified in the *D. melanogaster* voltage-gated sodium channel (Nav) genes as a proof of principle for the following reasons. First, Nav has multiple transmembrane domains [232–234] and we have shown that transmembrane proteins are associated with PFCCs (Appendix 3.3). Second, Nav is a well-characterized protein family in terms of its physiological roles and structure-function relationship. Third, the *D. melanogaster* genome harbors two Nav paralogs whose divergence was dated back to the origin of Bilateria, which allows us to identify the PFCCs with conserved codon usage patterns.

Each Nav α-subunit consists of four transmembrane domains (Domains I-IV) linked by cytoplasmic chains, plus an N-terminal and a C-terminal cytoplasmic chains. The inactivation gate, which is responsible for stopping the sodium influx during action potential, is formed by the cytoplasmic chain between Domain III (DIII) and Domain IV (DIV) that will be refer to as DIII-IV linker below [234]. In general, most invertebrates have two types of Nav, namely type 1 Nav (Nav1) and type 2 Nav (Nav2), while vertebrates have lost the Nav2 gene but have gained multiple Nav1 paralogs [234]. As aforementioned, *D. melanogaster* has two paralogs of Nav, namely *para*, the Dmel/Nav1, and *NaCP60E*, the Dmel/Nav2 [235,236].

Multiple PFCCs were identified in Dmel/Nav1 and Dmel/Nav2, but the PFCCs in Dmel/Nav1 and those in Dmel/Nav2 are not always homologous. Nonetheless, we found that both genes have PFCCs in the DIII-IV linkers (Figure 3.4). To assess the potential functions of these PFCCs, we then scanned the DIII-IV linkers with a 15-amino-acid sliding window and calculated TCAI for each window. We found that these PFCCs exhibit strong preference for rare codons

(Figure 3.5A, Dmel; Figure 3.5B, Dmel), suggesting that decelerating translation during the synthesis of the inactivation gate may be the key function of these PFCCs. We further scanned the DIII-IV linkers of Nav homologs in several other representative eukaryotic species, and found that the majority of them also have sub-regions preferring rare codons (Figure 3.5, TCAI < -0.8).

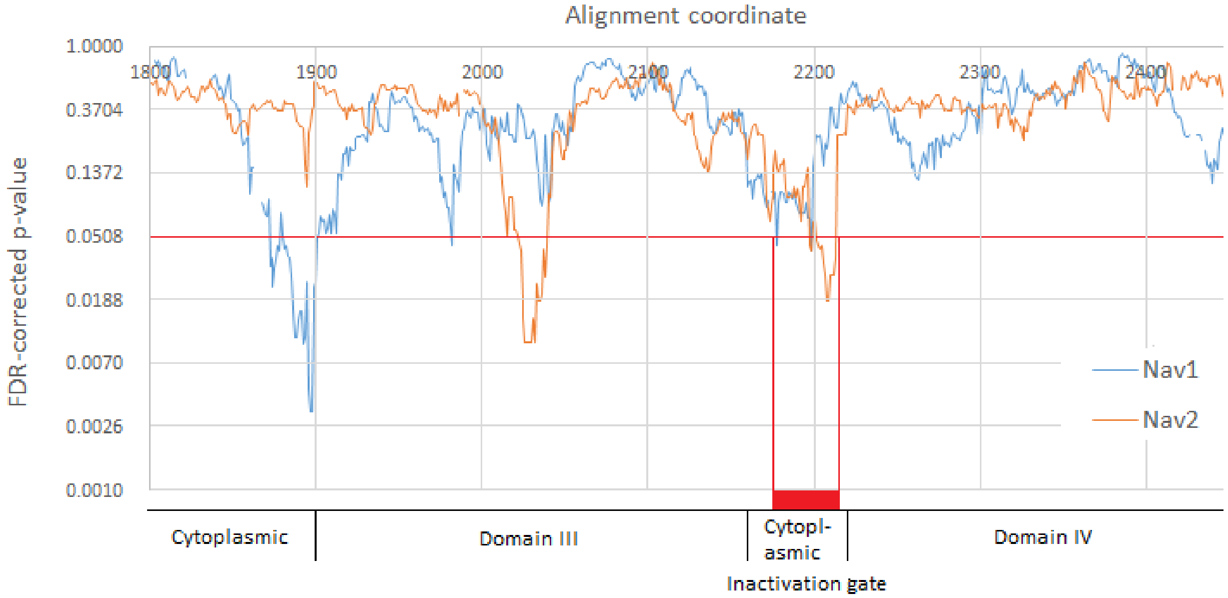

Figure 3.4: Identifying PFCCs in *D. melanogaster* Nav paralogs. Dmel/Nav1 and Dmel/Nav2 are aligned by amino acid sequences. *p*-values were corrected by FDR (FDR = 0.05), and those lower than the threshold indicate codon clusters whose codon usage patterns are significantly different from both whole-genome and gene-specific codon usage patterns. Both Dmel/Nav1 and Dmel/Nav2 have PFCCs in the DIII-IV linkers, shown by the red bar.

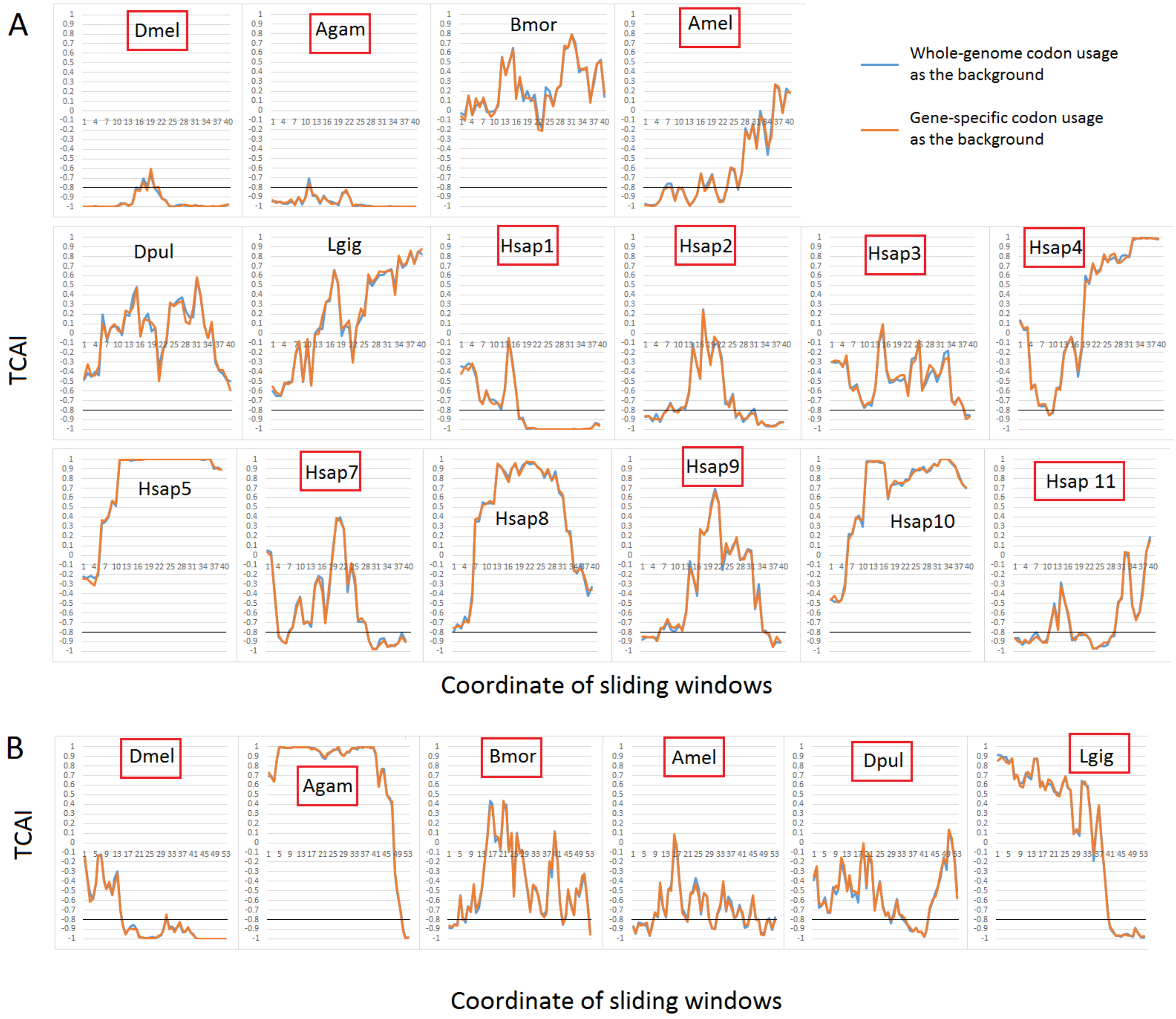

Figure 3.5: Nav paralogs generally bear rare-codon clusters in DIII-DIV linkers. (A) Nav1. **(B)** Nav2. Dmel: *Drosophila melanogaster*, fruit fly; Agam: *Anopheles gambiae*, malaria mosquito; Bmor: *Bombyx mori*, silkmoth; Amel: *Apis mellifera*, Western honey bee; Dpul: *Daphnia pulex*, water flea; Lgig: *Lottia gigantea*, owl limpet; Hsap: *Homo sapiens*, human. Homo sapiens has ten Nav1 paralogs but no Nav2. As suggested by Fig. 2, regions with TCAI < -0.8 are regarded as rare-codon clusters. Red boxes highlight the DIII-IV linkers carrying rare-codon clusters. Black lines: TCAI = 0.8. Blue curves: TCAI calculated by using whole-genome codon usage as the background. Orange curves: TCAI calculated by using gene-specific codon usage as the background.

Considering that the divergence between Nav1 and Nav2 was dated back to the origin of Bilateria [234], the conserved preference for rare codons in the DIII-IV linkers further support the hypothesis that the normal function of inactivation gate requires decelerated translation of this region. Decelerated translation is possibly critical for the correct folding pattern or phosphorylation of the DIII-IV linker [237–240]. In this regard, we hypothesize that synonymous mutations from rare codons to common codons in the DIII-IV linker could induce changes in the action potential through prolonged or shortened depolarization. Also, as some nonsynonymous mutations in the DIII-IV linker could cause cold-induced paralysis [241], it is possible that the synonymous mutations from rare codons to common codons in this region can cause similar phenotypes.

Furthermore, we noticed that not all DIII-IV linkers bear obvious rare-codon clusters (Figure 3.5A, Bmor, Dpul, Lgig, Hsap5, Hsap8, Hsap10). Therefore, it is possible that for some species, synonymous codon usage in the DIII-IV linker is less sensitive to natural selection, perhaps due to other mechanisms that compensate the effects of rare codons on protein folding. More interestingly, we found that among the Nav1 paralogs in human, some have rare-codon clusters in the DIII-IV linkers while others do not. We also found that among the paralogs with rare-codon clusters, the specific locations of rare-codon clusters can be different. These findings perhaps suggest that rare-codon clusters are associated with the division of labor between Nav1 paralogs. As Nav1 paralogs have differentiated tissue-specific expression profiles [242], one mechanism underlying the possible codon-usage-mediated division of labor may be that these paralogs adapt their DIII-IV linkers' codon usage patterns to tissue-specific tRNA pools [196,208,227], so that the corresponding protein-coding sequences are able to more finely regulate the function of inactivation gate.

3.4 Discussion

Here we show that clusters of codons with biased codon usage patterns may serve as nucleic-acid-level domains that affect gene functions, just as a sequence of amino acids with a specific order and/or specific biochemical properties can form a protein domain. We accomplished this by developing a conservative statistical approach to identify PFCCs in the *D. melanogaster* genome. We have identified over 3000 PFCCs, and most of them strongly prefer rare codons. Nevertheless, we also found that a small proportion of the PFCCs exhibit other patterns of codon usage, such as preference for common codons, which was not reported before. We showed that although the PFCCs are associated with specific protein functional classes including transmembrane proteins and transcription factors, most of them are not associated with known protein domains defined by amino acid sequences. As a proof-of-principle, we used the example of a rare-codon cluster associated with the inactivation gate of Nav to propose a hypothesis concerning how a PFCC could affect specific biochemical and physiological properties of a protein. Together, our results suggest that it is likely a general phenomenon that codon clusters with biased codon usage patterns serve as diverse “hidden domains” involved in regulating protein functions.

In this paper, based on a widely used codon usage index CAI [3], we proposed an alternative codon usage index TCAI (see 3.5.3 Calculating TCAI) that was used for classifying PFCCs. Compared to CAI, TCAI is better at describing the preference for rare codons. This is because when CAI is calculated, codon usage frequencies are all normalized to the frequencies of the most common synonymous codons. Thus, the CAI value of any codon cluster that strictly uses common codons will always be 1, while if two codon clusters that strictly use rare codons but have different amino acid sequences, they may have fairly different CAI values. However, by using the newly proposed TCAI, rare-codon clusters will have similar TCAI values that are -1 or very close to -1, while common-codon clusters keep TCAI values at 1 or near 1. Thus, TCAI is a good choice when researchers intend to identify rare-codon clusters.

In comparison to previous methods for detecting functional codon clusters [224], the method presented here is more conservative in terms of detecting rare-codon clusters due to the usage of both whole-genome and gene-specific codon usage patterns as the background codon usage. Yet, it is more powerful in terms of detecting other types of codon clusters due to a more relaxed assumption about the possible functional roles of codon clusters. The diverse codon usage patterns and locations of the PFCCs suggest that codon clusters may affect protein functions through various mechanisms. The major mechanism through which codon clusters regulate protein functions is possibly the deceleration of translation, as shown by the preponderance of rare-codon clusters in the identified PFCCs. However, we must admit that the preponderance of rare-codon clusters may be partly an artifact of technically easier detection of the preference for rare codons by our approach. To increase the power of codon-cluster-detection algorithms and more accurately assess the prevalence and importance of different types of codon clusters, researchers may need to incorporate phylogenetic analyses of homologous protein-coding genes in order to identify codon clusters with conserved codon usage patterns.

Consistent with previous reports [224], we found that some of the PFCCs are associated with known protein domains defined by amino acid sequences, which suggests that some codon clusters do have the potential to assist correct folding and modifications of protein domains. However, we also found that the majority of PFCCs are not associated with known protein domains [169,231], indicating that these PFCCs may carry necessary information for regulating protein functions and such information cannot be predicted from amino acid sequences. Thus, codon clusters could serve as “hidden domains” in protein-coding sequences. For example, some “free coiled regions” of proteins may not be actually “free”: their folding and modifications could be restricted by the codon usage patterns of the corresponding genomic regions. Further investigation into the codon clusters that may encode “hidden domains” could be important for biologists to better understand how genetic information directs the functions of proteins.

As we have shown by the example of rare-codon clusters in the DIII-IV linkers of Nav proteins, functional codon clusters may be important for some key functions of proteins. This could have important implications for molecular evolutionary studies and genetic engineering practice. For molecular evolutionary studies, codon clusters with critical functions suggest that synonymous sites in such functional codon clusters may bias the estimation of the rate of neutral evolution if researchers consider synonymous mutations as neutral mutations. Moreover, it is possible that the selective pressure on synonymous codon usage may be even stronger than that on nonsynonymous mutations, which could greatly interfere the results and inferences of the evolutionary analyses based on the comparison between synonymous and nonsynonymous sites. For genetic engineering practice, functional codon clusters suggest that when transgenes are designed, simple codon optimization [40], which generally uses common codons to encode amino acid residues, may not be the best choice to achieve desired structure and functions of the engineered proteins. Instead, the codon usage of different regions within a transgene may need to be more delicately controlled.

Together, our data support the broad existence of diverse and functional codon clusters that may affect protein functions and associated phenotypes through various mechanisms. In this regard, we suggest that functional codon clusters should be seriously considered if researchers are to thoroughly understand how genetic information is interpreted into functional, phenotypic, and evolutionary outputs *in vivo*.

3.5 Materials and Methods

3.5.1 Reference Protein-coding Sequences

Reference protein-coding sequences of *D. melanogaster* were downloaded from Ensembl 89 [166]. Protein-coding sequences fulfilling the following criteria were chosen. 1) The sequence length is a multiple of three. 2) The sequence uses standard genetic code. 3) For each gene, only the longest mRNA isoform was used; if there were multiple isoforms of the same length, then the first record shown in the FASTA file was used.

The protein-coding sequences of Nav1 and Nav2 in analyzed species can be found in Appendix 3.6.

3.5.2 Identifying PFCCs

Figure 3.6 depicts how to identify PFCCs in a protein-coding sequence. For a window W_i starting with the *i*th codon in a protein-coding sequence, the window size S is set to vary between 5 to 50 codons. For each window size, two χ^2 tests are performed by comparing the codon usage of the window respectively to whole-genome codon usage and gene-specific codon usage, and the higher p-value is selected as the representative p-value. Then the representative p-values are plotted against window sizes, which generates a p-S curve representing a function $p(S)$ that describes the relationship between p-value and window size (Figure 3.6A-D). If $p(S)$ is monotonic, the lowest p-value together with its corresponding S are selected as the representative p and S for W_i, namely p_i and S_i; otherwise the p-value and the S that correspond to the lowest stationary point of $p(S)$ are selected as p_i and S_i. For the focal protein-coding sequence, all p_i's are corrected by setting the false discovery rate (FDR) [189] to 0.05 so as to get the corrected p-

values $p_{i,corrected}$'s; then windows with $p_{i,corrected}$ values lower than the threshold 0.05 are detected as positive segments with unexpected codon usage patterns (Figure 3.6E). Finally, isolated positive segments, together with the codon clusters generated by merging overlapped positive segments, are detected as PFCCs.

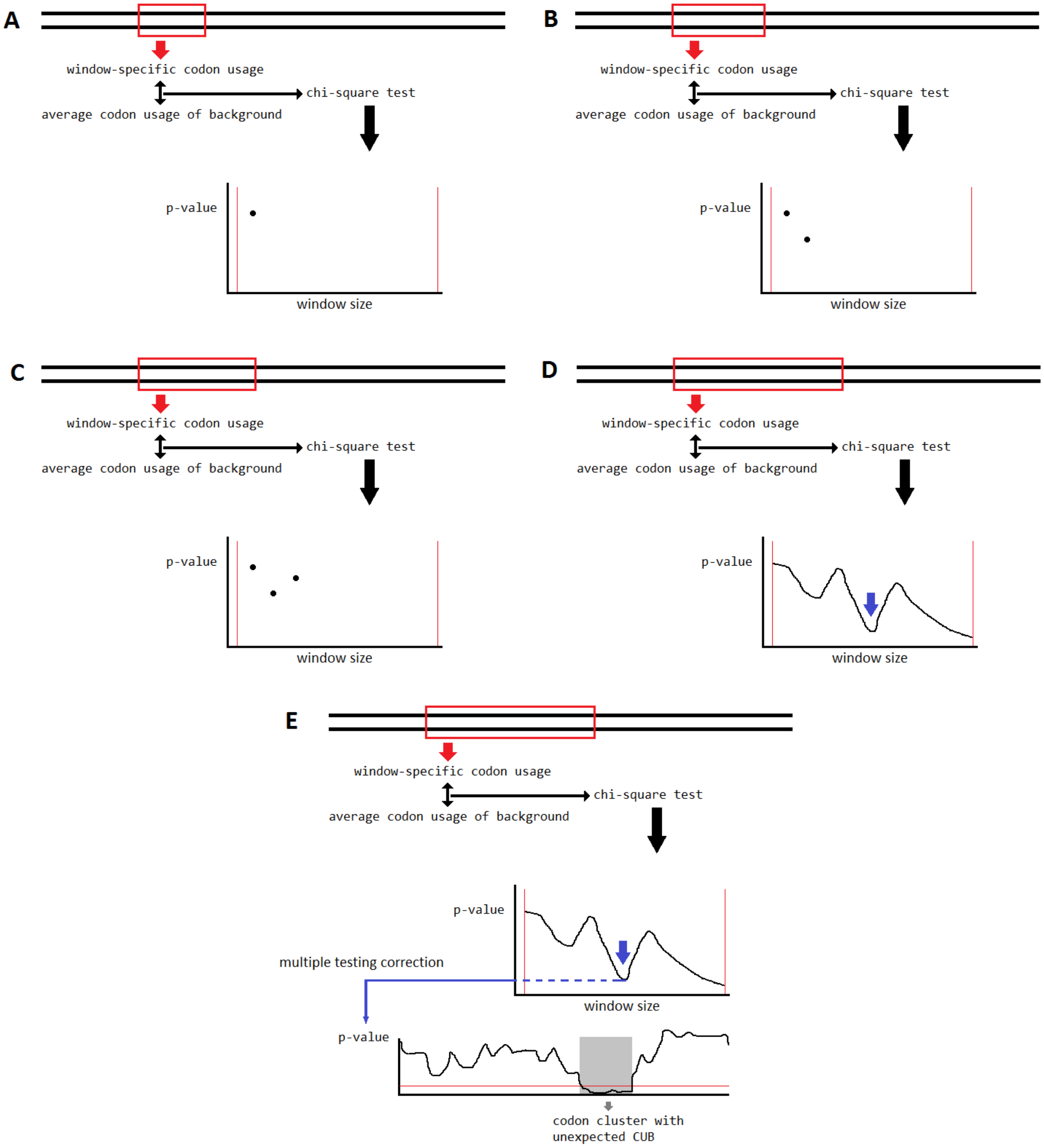
A
window-specific codon usage
chi-square test
average codon usage of background
p-value
window size
B
window-specific codon usage
chi-square test
average codon usage of background
p-value
window size
C
window-specific codon usage
chi-square test
average codon usage of background
p-value
window size
D
window-specific codon usage
chi-square test
average codon usage of background
p-value
window size
E
window-specific codon usage
chi-square test
average codon usage of background
p-value
multiple testing correction
window size
p-value
codon cluster with
unexpected CUB

Figure 3.6: Using sliding windows with adaptive sizes to identify PFCCs. (**A-D**) With a given start of the window, *p*-values for different window sizes are calculated. (**D**) The lowest stationary point on the *p-S* curve is picked to get the representative window size and *p*-value. (**E**) Windows with different starts are processed as described in (**A-D**), and then representative *p*-values are corrected by setting FDR=0.05. All representative *p*-values are plotted against the coordinates of the starts of windows in order to locate PFCCs.

3.5.3 Calculating TCAI

To calculate the TCAI of a given sequence of codons, the background relative codon usage frequencies need to be calculated first. For example, if a gene uses 10 AAA and 30 AAG to encode Lys, the gene-specific background relative codon usage frequencies of AAA and AAG will respectively be 10/(10+30)=0.25 and 10/(10+30)=0.75. Then the focal sequence of codons is translated to an amino acid sequence. The next step is to generate a pseudo-sequence of codons according to the amino acid sequence and the background relative codon usage frequencies. For example, assuming that the amino acid sequence is Lys-Lys and the background relative codon usage frequencies are 0.25 for AAA and 0.75 for AAG, the first Lys will have a 25% chance to be encoded by AAA and 75% chance to be encoded by AAG, and so will the second Lys. This step of pseudo-sequence generation is repeated for 10,000 times so that there will be 10,000 pseudo-sequence of codons, which represent the expected results if codons are used randomly to encode the amino acids. Then the CAIs [3] of all pseudo-sequences and the CAI of the actual codon sequence are calculated. Finally, TCAI is calculated by subtracting the proportion of pseudo-sequences whose CAIs are higher than the CAI of the corresponding actual sequences from the proportion of pseudo-sequences whose CAIs are lower than the CAI of the corresponding actual sequences.

When TCAI is -1, it means that none of the pseudo-sequences has a CAI lower than the actual sequence; thus, the actual sequence strongly prefers rare codons. In contrast, when TCAI is 1, the actual sequence strongly prefers common codons.

3.5.4 K-mean Clustering of PFCCs

K-mean clustering is done by using the online tool at http://scistatcalc.blogspot.com/2014/01/k-means-clustering-calculator.html. The number of clusters (i.e., K) is determined by the elbow method, according to https://pythonprogramminglanguage.com/kmeans-elbow-method/. Each input data point of K-mean clustering is specified by its gene-specific and whole-genome TCAIs.

3.5.5 Calculating RLI

For a protein-coding sequence with L codons, the RLI of a PFCC which starts at the ith codon and has a size of S_i codons is calculated as $(i + S_i / 2) / L$.

3.5.6 Searching for Transmembrane Helices

For a focal PFCC, the protein sequence from the first residue or the 150th residue upstream to the PFCC-encoded region, whichever is closer to the PFCC-encoded region, to the last sense codon or the 150th codon downstream to the PFCC-encoded region, whichever is closer to the PFCC-encoded region, is input to TMHMM [229] in order to search for transmembrane helices near or overlapped with the PFCC-encoded region. The coordinates of identified transmembrane helices are recorded.

3.5.7 Searching for Pfam Protein Domains

For a focal PFCC, the protein sequence from the first residue or the 150th residue upstream to the PFCC-encoded region, whichever is closer to the PFCC-encoded region, to the last sense codon or the 150th codon downstream to the PFCC-encoded region, whichever is closer to the PFCC-encoded region, is input to the hmmscan program of HMMER [231] on https://www.ebi.ac.uk/Tools/hmmer/search/hmmscan in order to search for Pfam protein domains [169] near or overlapped with the PFCC-encoded region. The coordinates and names of identified Pfam domains are recorded.

3.5.8 Classifying Association between PFCCs and Protein Domains

The association between a PFCC and a protein domain is classified to one of the following categories.

1) No association: The closest distance between the PFCC-encoded region and the protein domain is longer than 20 residues.

2) 1-to-multiple association: Multiple protein domains are overlapped with the region that starts from the 20th residue upstream to the PFCC-encoded region and ends at the 20th residue downstream to the PFCC-encoded region.

3) Cluster in domain: Only one protein domain is associated with the PFCC. The PFCC-encoded region locates within the protein domain.

4) Domain in cluster: Only one protein domain is associated with the PFCC. The protein domain locates within the PFCC-encoded region.

5) Domain overlap left of cluster: Only one protein domain is associated with the PFCC. The start of the protein domain is upstream to the PFCC-encoded region and the end of the protein domain locates within the PFCC-encoded region.

6) Domain overlap right of cluster: Only one protein domain is associated with the PFCC. The start of the protein domain locates within the PFCC-encoded region and the end of the protein domain is downstream to the PFCC-encoded region.

7) Domain upstream to cluster: Only one protein domain is associated with the PFCC. The end of the protein domain is upstream to the PFCC-encoded region.

8) Domain downstream to cluster: Only one protein domain is associated with the PFCC. The start of the protein domain is downstream to the PFCC-encoded region.

3.5.9 Alignment of Nav Homologs and Identification of DIII-IV Linkers

Nav orthologs were aligned by using MAFFT algorithm [243,244]. The annotated DIII-IV linkers of Dmel/Nav1 (https://www.uniprot.org/uniprot/P35500) and Dmel/Nav2 (https://www.uniprot.org/uniprot/Q9W0Y8) were used to locate the DIII-IV linkers of the Nav1 and Nav2 in other analyzed species. Dmel/Nav1 and Dmel/Nav2 were also aligned by using MAFFT algorithm (https://www.ebi.ac.uk/Tools/msa/mafft/) [243,244].

Chapter 4: Conclusions

4.1 Synonymous Mutations Are Not Intrinsically Associated with Neutral Mutations

It has been known for over four decades that synonymous mutations could be non-neutral as they can cause phenotypic changes through affecting physical and chemical properties of mRNA, translational machinery, co-translational processes, and epigenetic modifications. However, the reports of non-neutral synonymous mutations are mostly perceived by evolutionary biologists as rare exceptions to a seemingly convincing rule that synonymous mutations are generally neutral. Nonetheless, the putative generality of such a rule had not been rigorously tested.

In my thesis, by developing and applying a widely applicable statistical method to detect signatures of natural selection on gene-specific codon usage biases, I have shown that non-neutral synonymous mutations must not be rare exceptions. This is because the broadly existing signatures of natural selection on gene-specific codon usage biases contradict the putative generality of neutral synonymous mutations. In this regard, I think it is legitimate to claim that the known cases of non-neutral synonymous mutations are not a mere collection of anecdotal examples; rather, they should be the outcomes of a general rule that synonymous codon usage performs broad and critical biological functions in ontogenesis and phylogenesis. Using synonymous mutations as general proxies for neutral mutations will likely introduce systematic biases.

Nonetheless, the dissociation between synonymous and neutral mutations does not necessarily lead to the rejection of the neutral theory of molecular evolution. This is because the core statement of the neutral theory, that most mutations are evolutionarily neutral, is compatible with prevalent non-neutral synonymous mutations. By definition, synonymous mutations are unlikely to affect intergenic regions and introns, which could form the majority of genomes, especially for eukaryotes. Therefore, even if all synonymous mutations are not neutral, it is still possible that most mutations are neutral. It is a specific branch of the neutral theory that is likely significantly impacted by the broad existence of non-neutral synonymous mutations – that is, the methods that detect signatures of natural selection on protein-coding genes by assuming synonymous mutations as neutral mutations. Evolutionary biologists and population geneticists should rely more on the methods that are not based on this assumption, find better proxies for neutral mutations, or incorporate the probabilities of non-neutral synonymous mutations in order to assess the confidence intervals of their results.

4.2 Rare Codons Are Not Necessarily Translationally Suboptimal

For most biologists who admit that translational selection can result in broad impacts of synonymous codon usage on protein functions, translation efficiency of a codon is usually thought to be intrinsically linked to the copy number of its cognate tRNA genes and/or its usage frequency in the whole genome or in a set of highly-expressed housekeeping genes [1–3,31,38]. For example, rare codons should be recognized by low-copy-number tRNA anticodons, and they should be suboptimal for translation.

However, by combining computational and experimental approaches, I have shown that such a view oversimplifies the possible functions of biased synonymous codon usage in protein translation. Genome-wide rare codons could be recognized by locally enriched tRNAs, and thus the preference for rare codons may actually increase translation efficiency in specific cells and/or during a specific period of time. In this regard, I claim that codon optimality may not be simply inferred from whole-genome codon usage frequencies or tRNA gene copy numbers. Rather, codon optimality is context-dependent; therefore, more information, especially the actual cellular tRNA expression profiles, will be necessary to precisely infer cell- and/or tissue-specific optimal codon usage. This also means that the traditional strategy for optimizing the codon usage of transgenes, which mostly assumes that genome-wide common codons are optimal, may need to be revised. For example, if a genetic engineering project requires efficient tissue-specific expression of a transgene, it could be a better strategy to encode some amino acid residues of the transgenic protein by genome-wide rare codons. Furthermore, if the desired functions of a transgene require slow accumulation of its encoded protein in the cellular pool, suboptimal codons may be the better choices. For example, for synthetic genes that form a transcriptional oscillator [245], using suboptimal codons may help maintain the oscillating behavior. This is because it has been shown that the preferential usage of suboptimal codons is necessary for the native proteins underlying circadian rhythm to maintain the concerted fluctuation of concentrations [30,33].

It should be noted that my finding that genome-wide rare codons can be optimal in specific tissues do not reject the assumptions of the translational selection theory in general. This is because my finding also supports that the interaction between anticodons and codons is one of the key ways in which synonymous codon usage affects phenotypes. My findings are more of fine-scale modifications of the translational selection theory, as they show that the codon-anticodon interaction is not constant across time and space. Thus, the codon-anticodon interaction provides more degrees of freedom for synonymous codon usage to regulate gene functions.

4.3 Codon Clusters Represent Nucleic-Acid-Level Domains Affecting Protein Functions

Functional domains of proteins are usually thought to be formed by sequences of amino acid residues with specific biochemical properties. In this regard, researchers have been using multiple experimental and computational tools to detect functional domains from the databases of amino acid sequences. Their efforts have generated fruitful results, including the discovery of transmembrane helices and the development of protein domain databases such as the Pfam database [169].

However, as synonymous codon usage broadly affects protein functions, it is also possible that some functional domains of proteins can be encoded by codon sequences with characteristic codon usage patterns. Indeed, by developing and applying a statistical method to detect putatively functional codon clusters (PFCCs), I have identified in *D. melanogaster* genome over three thousand PFCCs defined by codon usage patterns rather than amino acid sequences. These PFCCs have diverse patterns of synonymous codon usage, and the majority of them do not co-occur with known protein domains that are determined by amino acid sequences. Furthermore, by using voltage-gated sodium channels as examples, I have explained how conserved preference for rare codons in a specific homologous region can be favored by natural selection. Thus, my results suggest that functional domains of proteins are encoded not only by amino acid sequences but also by DNA sequences with characteristic codon usage patterns. To understand how the structures and functions of proteins are regulated *in vivo*, it is necessary to incorporate nucleic-acid-level information in addition to the amino-acid-level one.

4.4 This Thesis Calls for Re-evaluating the Research Paradigms Based on the General Neutrality of Synonymous Mutations

In my thesis, I conclude that the neutrality of synonymous mutations and amino-acid determinism of protein properties only partially capture the critical factors affecting protein functions, associated phenotypes, and molecular evolution of protein-coding sequences. Non-neutral synonymous mutations, context-dependent optimality of rare codons, and codon-usage-encoded functional domains represent prevalent, important, yet likely underappreciated mechanisms regulating organismal functions.

I expect that my thesis will have impacts on evolutionary biology, population genetics, and structural biology. First, the prevalent signatures of natural selection on synonymous mutations indicate that it is likely inappropriate to use synonymous mutations as proxies for neutral mutations without prior evidence. Thus, the methods based on the general neutrality of synonymous mutations are prone to generate biased results in terms of the rates of evolution and the types of natural selection. Second, the broad existence of functional synonymous codon usage suggests that synonymous SNPs could significantly contribute to phenotypic variations, and thus excluding synonymous SNPs – which is a common approach for filtering out “noise” – should not be a standard step in GWAS. Third, the context-dependent optimality of rare codons implies that synonymous codon usage has the potential to play various roles in shaping organismal functions. It also suggests that in genetic engineering practice, the synonymous codon usage of a transgene should be carefully designed so that it is adapted to the desired functions, instead of simply choosing the common, putatively optimal codons. Fourth, the detection of codon-usage-defined functional domains indicates that amino-acid determinism is not enough to explain the structural and functional properties of proteins. Proteins, especially when expressed *in vivo*, should be viewed not only as strings of amino acid residues, but also as highly dynamic, context-dependent entities that incorporate information from multiple levels to ensure normal functions of organisms. In this regard, the best predictor of protein functions may not always be the amino acid sequences.

4.5 Future Directions

As shown by my work, the assumption that synonymous mutations are generally neutral, which underlies multiple statistical methods for analyzing the evolution of protein-coding genes, is not entirely accurate. Therefore, it may be necessary to modify these methods by incorporating the effects of non-neutral synonymous mutations, in order to more accurately assess the impacts of natural selection on protein-coding genes. One possible modification is to estimate the robustness of the results of these methods when a certain proportion of synonymous mutations are assigned as non-neutral mutations. If the results are relatively robust when the proportion of putatively non-neutral synonymous mutations increases, it is likely that the results are reliable; otherwise, researchers may need to admit that the signal-to-noise ratios of the results are not high enough to allow clear inferences about the roles of natural selection.

I have also claimed that simple “codon optimization”, which generally uses common, putatively optimal codons to encode amino acid residues, may not always be the best strategy for choosing codons for engineered genes. Therefore, a more comprehensive algorithm for choosing the most appropriate codons for engineered genes may be necessary. Such an algorithm may be achieved by training an artificial neural network with the known associations between specific codon usage patterns and gene functions.

Although I have computationally identified PFCCs in the *D. melanogaster* genome, I have not succeeded in experimentally investigating whether these PFCCs are truly functional domains and what their specific functions might be. During my graduate study, I tried to test the functional effects of the PFCCs in the type 1 voltage-gated sodium channel (*para*) and the ligand *decapentaplegic* (*dpp*) in the TGF-β signaling pathway. Unfortunately, the manipulation of the *para* codon usage was not technically feasible, because modifying the codon usage pattern of the DIII-IV linker would result in high GC-content that prevented the artificial syntheses of the DNA sequences that would be used for generating transgenic plasmids. For *dpp*, although I successfully synthesized the plasmid carrying the experimental allele, and the two putatively successful CRISPR (http://flycrispr.molbio.wisc.edu/scarless) [246] transformant fruit flies exhibited abnormal wing morphology, these transformants seemed to be dominantly sterile and with low viability, which prevented me from further experimental studies. Nevertheless, for researchers interested in functional codon clusters, it might be worthwhile to experimentally investigate the functional roles of other PFCCs shown in Appendix 3.1. Among these PFCCs, I would recommend starting with the N-terminal rare codon cluster of *lozenge* (*lz*). This is because *lz* plays a key role in eye development so that the phenotypic effects may be easily quantified, and the position and codon usage pattern of its PFCC are similar to those of *dpp*.

As long as researchers could free their mind from the seemingly intrinsic association between synonymous mutations and evolutionary neutrality, they would find diverse interesting directions worthy of investigation.

References

1. Ikemura T. Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: A proposal for a synonymous codon choice that is optimal for the E. coli translational system. J Mol Biol. 1981;151: 389–409. doi:10.1016/0022-2836(81)90003-6

2. Kanaya S, Yamada Y, Kudo Y, Ikemura T. Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of Bacillus subtilis tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis. Gene. 1999;238: 143–155. doi:10.1016/S0378-1119(99)00225-5

3. Sharp PM, Li W-H. The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res. 1987;15: 1281–1295. doi:10.1093/nar/15.3.1281

4. Sharp PM, Cowe E, Higgins DG, Shields DC, Wolfe KH, Wright F. Codon usage patterns in Escherichia coli, Bacillus subtilis, Saccharomyces cerevisiae, Schizosaccharomyces pombe, Drosophila melanogaster and Homo sapiens ; a review of the considerable within-species diversity. Nucleic Acids Res. 1988;16: 8207–8211. doi:10.1093/nar/16.17.8207

5. Angov E. Codon usage: Nature’s roadmap to expression and folding of proteins. Biotechnol J. 2011;6: 650–659. doi:10.1002/biot.201000332

6. Quax TEF, Claassens NJ, Söll D, van der Oost J. Codon Bias as a Means to Fine-Tune Gene Expression. Mol Cell. 2015;59: 149–161. doi:10.1016/j.molcel.2015.05.035

7. Berger EM. Are Synonymous Mutations Adaptively Neutral? Am Nat. 1977;111: 606–607. doi:10.1086/283192

8. Hasegawa M, Yasunaga T, Miyata T. Secondary structure of MS2 phage RNA and bias in code word usage. Nucleic Acids Res. 1979;7: 2073–2079. doi:10.1093/nar/7.7.2073

9. Gouy M, Gautier C. Codon usage in bacteria: correlation with gene expressivity. Nucleic Acids Res. 1982;10: 7055–7074. doi:10.1093/nar/10.22.7055

10. Eyre-Walker AC. An analysis of codon usage in mammals: Selection or mutation bias? J Mol Evol. 1991;33: 442–449. doi:10.1007/BF02103136

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

12. Stenico M, Lloyd AT, Sharp PM. Codon usage in Caenorhabditis elegans : delineation of translational selection and mutational biases. Nucleic Acids Res. 1994;22: 2437–2446. doi:10.1093/nar/22.13.2437

13. Hurst LD, Pál C. Evidence for purifying selection acting on silent sites in BRCA1. Trends Genet. 2001;17: 62–65. doi:10.1016/S0168-9525(00)02173-9

14. Duan J, Wainwright MS, Comeron JM, Saitou N, Sanders AR, Gelernter J, et al. Synonymous mutations in the human dopamine receptor D2 (DRD2) affect mRNA stability and synthesis of the receptor. Hum Mol Genet. 2003;12: 205–216. doi:10.1093/hmg/ddg055

15. Herbeck JT, Novembre J. Codon Usage Patterns in Cytochrome Oxidase I Across Multiple Insect Orders. J Mol Evol. 2003;56: 691–701. doi:10.1007/s00239-002-2437-7

16. Ko H-J, Ko S-Y, Kim Y-J, Lee E-G, Cho S-N, Kang C-Y. Optimization of Codon Usage Enhances the Immunogenicity of a DNA Vaccine Encoding Mycobacterial Antigen Ag85B. Infect Immun. 2005;73: 5666–5674. doi:10.1128/IAI.73.9.5666-5674.2005

17. Singh ND, Davis JC, Petrov DA. X-Linked Genes Evolve Higher Codon Bias in Drosophila and Caenorhabditis. Genetics. 2005;171: 145–155. doi:10.1534/genetics.105.043497

18. Morton BR, Wright SI. Selective Constraints on Codon Usage of Nuclear Genes from Arabidopsis thaliana. Mol Biol Evol. 2007;24: 122–129. doi:10.1093/molbev/msl139

19. Wang H-C, Hickey DA. Rapid divergence of codon usage patterns within the rice genome. BMC Evol Biol. 2007;7: S6. doi:10.1186/1471-2148-7-S1-S6

20. Pagani F, Raponi M, Baralle FE. Synonymous mutations in CFTR exon 12 affect splicing and are not neutral in evolution. Proc Natl Acad Sci. 2005;102: 6368–6372. doi:10.1073/pnas.0502288102

21. Kimchi-Sarfaty C, Oh JM, Kim I-W, Sauna ZE, Calcagno AM, Ambudkar SV, et al. A “Silent” Polymorphism in the MDR1 Gene Changes Substrate Specificity. Science. 2007;315: 525–528. doi:10.1126/science.1135308

22. Parmley JL, Hurst LD. Exonic Splicing Regulatory Elements Skew Synonymous Codon Usage near Intron-exon Boundaries in Mammals. Mol Biol Evol. 2007;24: 1600–1603. doi:10.1093/molbev/msm104

23. Takahashi A. Effect of exonic splicing regulation on synonymous codon usage in alternatively spliced exons of Dscam. BMC Evol Biol. 2009;9: 214. doi:10.1186/1471-2148-9-214

24. Zhang G, Hubalewska M, Ignatova Z. Transient ribosomal attenuation coordinates protein synthesis and co-translational folding. Nat Struct Mol Biol. 2009;16: 274–280. doi:10.1038/nsmb.1554

25. Zhou T, Weems M, Wilke CO. Translationally Optimal Codons Associate with Structurally Sensitive Sites in Proteins. Mol Biol Evol. 2009;26: 1571–1580. doi:10.1093/molbev/msp070

26. Agashe D, Martinez-Gomez NC, Drummond DA, Marx CJ. Good Codons, Bad Transcript: Large Reductions in Gene Expression and Fitness Arising from Synonymous Mutations in a Key Enzyme. Mol Biol Evol. 2013;30: 549–560. doi:10.1093/molbev/mss273

27. Gu W, Wang X, Zhai C, Xie X, Zhou T. Selection on Synonymous Sites for Increased Accessibility around miRNA Binding Sites in Plants. Mol Biol Evol. 2012;29: 3037–3044. doi:10.1093/molbev/mss109

28. Qian W, Yang J-R, Pearson NM, Maclean C, Zhang J. Balanced Codon Usage Optimizes Eukaryotic Translational Efficiency. PLOS Genet. 2012;8: e1002603. doi:10.1371/journal.pgen.1002603

29. Lawrie DS, Messer PW, Hershberg R, Petrov DA. Strong Purifying Selection at Synonymous Sites in D. melanogaster. PLOS Genet. 2013;9: e1003527. doi:10.1371/journal.pgen.1003527

30. Zhou M, Guo J, Cha J, Chae M, Chen S, Barral JM, et al. Non-optimal codon usage affects expression, structure and function of clock protein FRQ. Nature. 2013;495: 111–115. doi:10.1038/nature11833

31. Ma L, Cui P, Zhu J, Zhang Z, Zhang Z. Translational selection in human: more pronounced in housekeeping genes. Biol Direct. 2014;9: 17. doi:10.1186/1745-6150-9-17

32. Shin YC, Bischof GF, Lauer WA, Desrosiers RC. Importance of codon usage for the temporal regulation of viral gene expression. Proc Natl Acad Sci. 2015;112: 14030–14035. doi:10.1073/pnas.1515387112

33. Fu J, Murphy KA, Zhou M, Li YH, Lam VH, Tabuloc CA, et al. Codon usage affects the structure and function of the Drosophila circadian clock protein PERIOD. Genes Dev. 2016;30: 1761–1775. doi:10.1101/gad.281030.116

34. Pouyet F, Bailly-Bechet M, Mouchiroud D, Guéguen L. SENCA: A Multilayered Codon Model to Study the Origins and Dynamics of Codon Usage. Genome Biol Evol. 2016;8: 2427–2441. doi:10.1093/gbe/evw165

35. Saikia M, Wang X, Mao Y, Wan J, Pan T, Qian S-B. Codon optimality controls differential mRNA translation during amino acid starvation. RNA. 2016; doi:10.1261/rna.058180.116

36. Machado HE, Lawrie DS, Petrov DA. Strong selection at the level of codon usage bias: evidence against the Li-Bulmer model. bioRxiv. 2017; 106476. doi:10.1101/106476

37. Matsumoto T, John A, Baeza-Centurion P, Li B, Akashi H. Codon Usage Selection Can Bias Estimation of the Fraction of Adaptive Amino Acid Fixations. Mol Biol Evol. 2016;33: 1580–1589. doi:10.1093/molbev/msw027

38. dos Reis M, Savva R, Wernisch L. Solving the riddle of codon usage preferences: a test for translational selection. Nucleic Acids Res. 2004;32: 5036–5044. doi:10.1093/nar/gkh834

39. Smith NGC, Eyre-Walker A. Why Are Translationally Sub-Optimal Synonymous Codons Used in Escherichia coli? J Mol Evol. 2001;53: 225–236. doi:10.1007/s002390010212

40. Fuglsang A. Codon optimizer: a freeware tool for codon optimization. Protein Expr Purif. 2003;31: 247–249. doi:10.1016/S1046-5928(03)00213-4

41. Slimko EM, Lester HA. Codon optimization of Caenorhabditis elegans GluCl ion channel genes for mammalian cells dramatically improves expression levels. J Neurosci Methods. 2003;124: 75–81. doi:10.1016/S0165-0270(02)00362-X

42. Keller EB, Noon WA. Intron splicing: a conserved internal signal in introns of animal pre-mRNAs. Proc Natl Acad Sci. 1984;81: 7417–7420. doi:10.1073/pnas.81.23.7417

43. Keller EB, Noon WA. Intron splicing: a conserved internal signal in introns of Drosophila pre-mRNAs. Nucleic Acids Res. 1985;13: 4971–4981. doi:10.1093/nar/13.13.4971

44. Nakata K, Kanehisa M, DeLisi C. Predictlon of splice junctions in mRNA sequences. Nucleic Acids Res. 1985;13: 5327–5340. doi:10.1093/nar/13.14.5327

45. Branciamore S, Chen Z-X, Riggs AD, Rodin SN. CpG island clusters and pro-epigenetic selection for CpGs in protein-coding exons of HOX and other transcription factors. Proc Natl Acad Sci. 2010;107: 15485–15490. doi:10.1073/pnas.1010506107

46. Matsuo Y. Epigenetics and Codon Usage of the Histone Genes in 12 Drosophila Species. J Phylogenetics Evol Biol. 2017;s5. doi:10.4172/2329-9002.1000178

47. Komar AA, Lesnik T, Reiss C. Synonymous codon substitutions affect ribosome traffic and protein folding during in vitro translation. FEBS Lett. 1999;462: 387–391. doi:10.1016/S0014-5793(99)01566-5

48. Zalucki YM, Gittins KL, Jennings MP. Secretory signal sequence non-optimal codons are required for expression and export of β-lactamase. Biochem Biophys Res Commun. 2008;366: 135–141. doi:10.1016/j.bbrc.2007.11.093

49. Ansari MA, Pedergnana V, Ip CLC, Magri A, Von Delft A, Bonsall D, et al. Genome-to-genome analysis highlights the effect of the human innate and adaptive immune systems on the hepatitis C virus. Nat Genet. 2017;49: 666–673. doi:10.1038/ng.3835

50. Baker Z, Schumer M, Haba Y, Bashkirova L, Holland C, Rosenthal GG, et al. Repeated losses of PRDM9-directed recombination despite the conservation of PRDM9 across vertebrates. de Massy B, editor. eLife. 2017;6: e24133. doi:10.7554/eLife.24133

51. Barnes KG, Weedall GD, Ndula M, Irving H, Mzihalowa T, Hemingway J, et al. Genomic Footprints of Selective Sweeps from Metabolic Resistance to Pyrethroids in African

Malaria Vectors Are Driven by Scale up of Insecticide-Based Vector Control. PLOS Genet. 2017;13: e1006539. doi:10.1371/journal.pgen.1006539

52. Bellott DW, Skaletsky H, Cho T-J, Brown L, Locke D, Chen N, et al. Avian W and mammalian Y chromosomes convergently retained dosage-sensitive regulators. Nat Genet. 2017;49: 387–394. doi:10.1038/ng.3778

53. Blanc-Mathieu R, Perfus-Barbeoch L, Aury J-M, Rocha MD, Gouzy J, Sallet E, et al. Hybridization and polyploidy enable genomic plasticity without sex in the most devastating plant-parasitic nematodes. PLOS Genet. 2017;13: e1006777. doi:10.1371/journal.pgen.1006777

54. Blanco-Melo D, Gifford RJ, Bieniasz PD. Co-option of an endogenous retrovirus envelope for host defense in hominid ancestors. Coffin J, editor. eLife. 2017;6: e22519. doi:10.7554/eLife.22519

55. Bobay L-M, Ochman H. Impact of Recombination on the Base Composition of Bacteria and Archaea. Mol Biol Evol. 2017;34: 2627–2636. doi:10.1093/molbev/msx189

56. Boscaro V, Kolisko M, Felletti M, Vannini C, Lynn DH, Keeling PJ. Parallel genome reduction in symbionts descended from closely related free-living bacteria. Nat Ecol Evol. 2017;1: 1160. doi:10.1038/s41559-017-0237-0

57. Botero-Castro F, Figuet E, Tilak M-K, Nabholz B, Galtier N. Avian Genomes Revisited: Hidden Genes Uncovered and the Rates versus Traits Paradox in Birds. Mol Biol Evol. 2017;34: 3123–3131. doi:10.1093/molbev/msx236

58. Campos JL, Zhao L, Charlesworth B. Estimating the parameters of background selection and selective sweeps in Drosophila in the presence of gene conversion. Proc Natl Acad Sci. 2017;114: E4762–E4771. doi:10.1073/pnas.1619434114

59. Campos JL, Johnston KJA, Charlesworth B. The Effects of Sex-Biased Gene Expression and X-Linkage on Rates of Sequence Evolution in Drosophila. Mol Biol Evol. 2018;35: 655–665. doi:10.1093/molbev/msx317

60. Chen B, Zhang B, Xu L, Li Q, Jiang F, Yang P, et al. Transposable Element-Mediated Balancing Selection at Hsp90 Underlies Embryo Developmental Variation. Mol Biol Evol. 2017;34: 1127–1139. doi:10.1093/molbev/msx062

61. Clark EL, Bush SJ, McCulloch MEB, Farquhar IL, Young R, Lefevre L, et al. A high resolution atlas of gene expression in the domestic sheep (Ovis aries). PLOS Genet. 2017;13: e1006997. doi:10.1371/journal.pgen.1006997

62. Couce A, Caudwell LV, Feinauer C, Hindré T, Feugeas J-P, Weigt M, et al. Mutator genomes decay, despite sustained fitness gains, in a long-term experiment with bacteria. Proc Natl Acad Sci. 2017;114: E9026–E9035. doi:10.1073/pnas.1705887114

63. Croucher NJ, Campo JJ, Le TQ, Liang X, Bentley SD, Hanage WP, et al. Diverse evolutionary patterns of pneumococcal antigens identified by pangenome-wide immunological screening. Proc Natl Acad Sci. 2017;114: E357–E366. doi:10.1073/pnas.1613937114

64. Cubillos-Ruiz A, Berta-Thompson JW, Becker JW, Donk WA van der, Chisholm SW. Evolutionary radiation of lanthipeptides in marine cyanobacteria. Proc Natl Acad Sci. 2017;114: E5424–E5433. doi:10.1073/pnas.1700990114

65. Daub JT, Moretti S, Davydov II, Excoffier L, Robinson-Rechavi M. Detection of Pathways Affected by Positive Selection in Primate Lineages Ancestral to Humans. Mol Biol Evol. 2017;34: 1391–1402. doi:10.1093/molbev/msx083

66. De La Torre AR, Li Z, Van de Peer Y, Ingvarsson PK. Contrasting Rates of Molecular Evolution and Patterns of Selection among Gymnosperms and Flowering Plants. Mol Biol Evol. 2017;34: 1363–1377. doi:10.1093/molbev/msx069

67. Dietschi Q, Tuberosa J, Rösingh L, Loichot G, Ruedi M, Carleton A, et al. Evolution of immune chemoreceptors into sensors of the outside world. Proc Natl Acad Sci. 2017;114: 7397–7402. doi:10.1073/pnas.1704009114

68. Ebel ER, Telis N, Venkataram S, Petrov DA, Enard D. High rate of adaptation of mammalian proteins that interact with Plasmodium and related parasites. PLOS Genet. 2017;13: e1007023. doi:10.1371/journal.pgen.1007023

69. Eberlein C, Nielly-Thibault L, Maaroufi H, Dubé AK, Leducq J-B, Charron G, et al. The Rapid Evolution of an Ohnolog Contributes to the Ecological Specialization of Incipient Yeast Species. Mol Biol Evol. 2017;34: 2173–2186. doi:10.1093/molbev/msx153

70. Echave J, Spielman SJ, Wilke CO. Causes of evolutionary rate variation among protein sites. Nat Rev Genet. 2016;17: 109–121. doi:10.1038/nrg.2015.18

71. Elgvin TO, Trier CN, Tørresen OK, Hagen IJ, Lien S, Nederbragt AJ, et al. The genomic mosaicism of hybrid speciation. Sci Adv. 2017;3: e1602996. doi:10.1126/sciadv.1602996

72. Escalona T, Weadick CJ, Antunes A. Adaptive Patterns of Mitogenome Evolution Are Associated with the Loss of Shell Scutes in Turtles. Mol Biol Evol. 2017;34: 2522–2536. doi:10.1093/molbev/msx167

73. Evans AL, Blackburn JWD, Taruc K, Kipp A, Dirk BS, Hunt NR, et al. Antagonistic Coevolution of MER Tyrosine Kinase Expression and Function. Mol Biol Evol. 2017;34: 1613–1628. doi:10.1093/molbev/msx102

74. Faria NR, Quick J, Claro IM, Thézé J, de Jesus JG, Giovanetti M, et al. Establishment and cryptic transmission of Zika virus in Brazil and the Americas. Nature. 2017;546: 406–410. doi:10.1038/nature22401

75. Feyertag F, Alvarez-Ponce D. Disulfide Bonds Enable Accelerated Protein Evolution. Mol Biol Evol. 2017;34: 1833–1837. doi:10.1093/molbev/msx135

76. Figueiró HV, Li G, Trindade FJ, Assis J, Pais F, Fernandes G, et al. Genome-wide signatures of complex introgression and adaptive evolution in the big cats. Sci Adv. 2017;3: e1700299. doi:10.1126/sciadv.1700299

77. Fulgione A, Koornneef M, Roux F, Hermisson J, Hancock AM. Madeiran Arabidopsis thaliana Reveals Ancient Long-Range Colonization and Clarifies Demography in Eurasia. Mol Biol Evol. 2018;35: 564–574. doi:10.1093/molbev/msx300

78. Gao D, Chu Y, Xia H, Xu C, Heyduk K, Abernathy B, et al. Horizontal Transfer of Non-LTR Retrotransposons from Arthropods to Flowering Plants. Mol Biol Evol. 2018;35: 354–364. doi:10.1093/molbev/msx275

79. Good BH, McDonald MJ, Barrick JE, Lenski RE, Desai MM. The dynamics of molecular evolution over 60,000 generations. Nature. 2017;551: 45–50. doi:10.1038/nature24287

80. Guéguen L, Duret L. Unbiased Estimate of Synonymous and Nonsynonymous Substitution Rates with Nonstationary Base Composition. Mol Biol Evol. 2018;35: 734–742. doi:10.1093/molbev/msx308

81. Hargreaves AD, Zhou L, Christensen J, Marlétaz F, Liu S, Li F, et al. Genome sequence of a diabetes-prone rodent reveals a mutation hotspot around the ParaHox gene cluster. Proc Natl Acad Sci. 2017;114: 7677–7682. doi:10.1073/pnas.1702930114

82. Heidmann O, Béguin A, Paternina J, Berthier R, Deloger M, Bawa O, et al. HEMO, an ancestral endogenous retroviral envelope protein shed in the blood of pregnant women and expressed in pluripotent stem cells and tumors. Proc Natl Acad Sci. 2017;114: E6642–E6651. doi:10.1073/pnas.1702204114

83. Hu H, Uesaka M, Guo S, Shimai K, Lu T-M, Li F, et al. Constrained vertebrate evolution by pleiotropic genes. Nat Ecol Evol. 2017;1: 1722. doi:10.1038/s41559-017-0318-0

84. Huber CD, Kim BY, Marsden CD, Lohmueller KE. Determining the factors driving selective effects of new nonsynonymous mutations. Proc Natl Acad Sci. 2017;114: 4465–4470. doi:10.1073/pnas.1619508114

85. Huylmans AK, Macon A, Vicoso B. Global Dosage Compensation Is Ubiquitous in Lepidoptera, but Counteracted by the Masculinization of the Z Chromosome. Mol Biol Evol. 2017;34: 2637–2649. doi:10.1093/molbev/msx190

86. Iranzo J, Cuesta JA, Manrubia S, Katsnelson MI, Koonin EV. Disentangling the effects of selection and loss bias on gene dynamics. Proc Natl Acad Sci. 2017;114: E5616–E5624. doi:10.1073/pnas.1704925114

87. Janiak MC, Chaney ME, Tosi AJ. Evolution of Acidic Mammalian Chitinase Genes (CHIA) Is Related to Body Mass and Insectivory in Primates. Mol Biol Evol. 2018;35: 607–622. doi:10.1093/molbev/msx312

88. Kimura M. Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. Nature. 1977;267: 275–276. doi:10.1038/267275a0

89. Kita R, Venkataram S, Zhou Y, Fraser HB. High-resolution mapping of cis-regulatory variation in budding yeast. Proc Natl Acad Sci. 2017;114: E10736–E10744. doi:10.1073/pnas.1717421114

90. Koch AS, Brites D, Stucki D, Evans JC, Seldon R, Heekes A, et al. The Influence of HIV on the Evolution of Mycobacterium tuberculosis. Mol Biol Evol. 2017;34: 1654–1668. doi:10.1093/molbev/msx107

91. Kursel LE, Malik HS. Recurrent Gene Duplication Leads to Diverse Repertoires of Centromeric Histones in Drosophila Species. Mol Biol Evol. 2017;34: 1445–1462. doi:10.1093/molbev/msx091

92. Lee KB, Wang J, Palme J, Escalante-Chong R, Hua B, Springer M. Polymorphisms in the yeast galactose sensor underlie a natural continuum of nutrient-decision phenotypes. PLOS Genet. 2017;13: e1006766. doi:10.1371/journal.pgen.1006766

93. Levin TC, Malik HS. Rapidly Evolving Toll-3/4 Genes Encode Male-Specific Toll-Like Receptors in Drosophila. Mol Biol Evol. 2017;34: 2307–2323. doi:10.1093/molbev/msx168

94. Lin Y-C, Wang J, Delhomme N, Schiffthaler B, Sundström G, Zuccolo A, et al. Functional and evolutionary genomic inferences in Populus through genome and population sequencing of American and European aspen. Proc Natl Acad Sci. 2018;115: E10970–E10978. doi:10.1073/pnas.1801437115

95. Liu H, Li Y, Chen D, Qi Z, Wang Q, Wang J, et al. A-to-I RNA editing is developmentally regulated and generally adaptive for sexual reproduction in Neurospora crassa. Proc Natl Acad Sci. 2017;114: E7756–E7765. doi:10.1073/pnas.1702591114

96. Liu Q, Zhou Y, Morrell PL, Gaut BS. Deleterious Variants in Asian Rice and the Potential Cost of Domestication. Mol Biol Evol. 2017;34: 908–924. doi:10.1093/molbev/msw296

97. Lu T-C, Leu J-Y, Lin W-C. A Comprehensive Analysis of Transcript-Supported De Novo Genes in Saccharomyces sensu stricto Yeasts. Mol Biol Evol. 2017;34: 2823–2838. doi:10.1093/molbev/msx210

98. Maclean CJ, Metzger BPH, Yang J-R, Ho W-C, Moyers B, Zhang J. Deciphering the Genic Basis of Yeast Fitness Variation by Simultaneous Forward and Reverse Genetics. Mol Biol Evol. 2017;34: 2486–2502. doi:10.1093/molbev/msx151

99. Mähler N, Wang J, Terebieniec BK, Ingvarsson PK, Street NR, Hvidsten TR. Gene co-expression network connectivity is an important determinant of selective constraint. PLOS Genet. 2017;13: e1006402. doi:10.1371/journal.pgen.1006402

100. Marques DA, Taylor JS, Jones FC, Palma FD, Kingsley DM, Reimchen TE. Convergent evolution of SWS2 opsin facilitates adaptive radiation of threespine stickleback into different light environments. PLOS Biol. 2017;15: e2001627. doi:10.1371/journal.pbio.2001627

101. McDonald JH, Kreitman M. Adaptive protein evolution at the Adh locus in Drosophila. Nature. 1991;351: 652–654. doi:10.1038/351652a0

102. Meslin C, Cherwin TS, Plakke MS, Hill J, Small BS, Goetz BJ, et al. Structural complexity and molecular heterogeneity of a butterfly ejaculate reflect a complex history of selection. Proc Natl Acad Sci. 2017;114: E5406–E5413. doi:10.1073/pnas.1707680114

103. Miyata T, Yasunaga T. Molecular evolution of mRNA: A method for estimating evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences and its application. J Mol Evol. 1980;16: 23–36. doi:10.1007/BF01732067

104. Mock T, Otillar RP, Strauss J, McMullan M, Paajanen P, Schmutz J, et al. Evolutionary genomics of the cold-adapted diatom *Fragilariopsis cylindrus*. Nature. 2017;541: 536–540. doi:10.1038/nature20803

105. Morgan AP, Pardo-Manuel de Villena F. Sequence and Structural Diversity of Mouse Y Chromosomes. Mol Biol Evol. 2017;34: 3186–3204. doi:10.1093/molbev/msx250

106. Mostowy RJ, Croucher NJ, De Maio N, Chewapreecha C, Salter SJ, Turner P, et al. Pneumococcal Capsule Synthesis Locus cps as Evolutionary Hotspot with Potential to Generate Novel Serotypes by Recombination. Mol Biol Evol. 2017;34: 2537–2554. doi:10.1093/molbev/msx173

107. Nielsen R. Statistical tests of selective neutrality in the age of genomics. Heredity. 2001;86: 641–647. doi:10.1046/j.1365-2540.2001.00895.x

108. O’Toole ÁN, Hurst LD, McLysaght A. Faster Evolving Primate Genes Are More Likely to Duplicate. Mol Biol Evol. 2018;35: 107–118. doi:10.1093/molbev/msx270

109. Pacheco MA, Matta NE, Valkiūnas G, Parker PG, Mello B, Stanley CE, et al. Mode and Rate of Evolution of Haemosporidian Mitochondrial Genomes: Timing the Radiation of Avian Parasites. Mol Biol Evol. 2018;35: 383–403. doi:10.1093/molbev/msx285

110. Partha R, Chauhan BK, Ferreira Z, Robinson JD, Lathrop K, Nischal KK, et al. Subterranean mammals show convergent regression in ocular genes and enhancers, along with adaptation to tunneling. Odom DT, editor. eLife. 2017;6: e25884. doi:10.7554/eLife.25884

111. Payne S, McCarthy S, Johnson T, North E, Blum P. Nonmutational mechanism of inheritance in the Archaeon Sulfolobus solfataricus. Proc Natl Acad Sci. 2018;115: 12271–12276. doi:10.1073/pnas.1808221115

112. Ponte I, Romero D, Yero D, Suau P, Roque A. Complex Evolutionary History of the Mammalian Histone H1.1–H1.5 Gene Family. Mol Biol Evol. 2017;34: 545–558. doi:10.1093/molbev/msw241

113. Pucholt P, Wright AE, Conze LL, Mank JE, Berlin S. Recent Sex Chromosome Divergence despite Ancient Dioecy in the Willow Salix viminalis. Mol Biol Evol. 2017;34: 1991–2001. doi:10.1093/molbev/msx144

114. Robinson J, Guethlein LA, Cereb N, Yang SY, Norman PJ, Marsh SGE, et al. Distinguishing functional polymorphism from random variation in the sequences of >10,000 HLA-A, -B and -C alleles. PLOS Genet. 2017;13: e1006862. doi:10.1371/journal.pgen.1006862

115. Rogers RL, Slatkin M. Excess of genomic defects in a woolly mammoth on Wrangel island. PLOS Genet. 2017;13: e1006601. doi:10.1371/journal.pgen.1006601

116. Roggiani M, Yadavalli SS, Goulian M. Natural variation of a sensor kinase controlling a conserved stress response pathway in Escherichia coli. PLOS Genet. 2017;13: e1007101. doi:10.1371/journal.pgen.1007101

117. Rolo J, Worning P, Nielsen JB, Sobral R, Bowden R, Bouchami O, et al. Evidence for the evolutionary steps leading to mecA-mediated β-lactam resistance in staphylococci. PLOS Genet. 2017;13: e1006674. doi:10.1371/journal.pgen.1006674

118. Rutledge GG, Böhme U, Sanders M, Reid AJ, Cotton JA, Maiga-Ascofare O, et al. *Plasmodium malariae* and *P. ovale* genomes provide insights into malaria parasite evolution. Nature. 2017;542: 101–104. doi:10.1038/nature21038

119. Salojärvi J, Smolander O-P, Nieminen K, Rajaraman S, Safronov O, Safdari P, et al. Genome sequencing and population genomic analyses provide insights into the adaptive landscape of silver birch. Nat Genet. 2017;49: 904–912. doi:10.1038/ng.3862

120. Salvador-Martínez I, Coronado-Zamora M, Castellano D, Barbadilla A, Salazar-Ciudad I. Mapping Selection within Drosophila melanogaster Embryo’s Anatomy. Mol Biol Evol. 2018;35: 66–79. doi:10.1093/molbev/msx266

121. Sharir-Ivry A, Xia Y. The Impact of Native State Switching on Protein Sequence Evolution. Mol Biol Evol. 2017;34: 1378–1390. doi:10.1093/molbev/msx071

122. Shaw AE, Hughes J, Gu Q, Behdenna A, Singer JB, Dennis T, et al. Fundamental properties of the mammalian innate immune system revealed by multispecies comparison of type I interferon responses. PLOS Biol. 2017;15: e2004086. doi:10.1371/journal.pbio.2004086

123. Shewaramani S, Finn TJ, Leahy SC, Kassen R, Rainey PB, Moon CD. Anaerobically Grown Escherichia coli Has an Enhanced Mutation Rate and Distinct Mutational Spectra. PLOS Genet. 2017;13: e1006570. doi:10.1371/journal.pgen.1006570

124. Shropshire JD, On J, Layton EM, Zhou H, Bordenstein SR. One prophage WO gene rescues cytoplasmic incompatibility in Drosophila melanogaster. Proc Natl Acad Sci. 2018;115: 4987–4991. doi:10.1073/pnas.1800650115

125. Steige KA, Laenen B, Reimegård J, Scofield DG, Slotte T. Genomic analysis reveals major determinants of cis-regulatory variation in Capsella grandiflora. Proc Natl Acad Sci. 2017;114: 1087–1092. doi:10.1073/pnas.1612561114

126. Sweeney CG, Rando JM, Panas HN, Miller GM, Platt DM, Vallender EJ. Convergent Balancing Selection on the Mu-Opioid Receptor in Primates. Mol Biol Evol. 2017;34: 1629–1643. doi:10.1093/molbev/msx105

127. Tamarit D, Neuvonen M-M, Engel P, Guy L, Andersson SGE. Origin and Evolution of the Bartonella Gene Transfer Agent. Mol Biol Evol. 2018;35: 451–464. doi:10.1093/molbev/msx299

128. Tian X, Ruan J-X, Huang J-Q, Yang C-Q, Fang X, Chen Z-W, et al. Characterization of gossypol biosynthetic pathway. Proc Natl Acad Sci. 2018;115: E5410–E5418. doi:10.1073/pnas.1805085115

129. Tobler R, Nolte V, Schlötterer C. High rate of translocation-based gene birth on the Drosophila Y chromosome. Proc Natl Acad Sci. 2017;114: 11721–11726. doi:10.1073/pnas.1706502114

130. Turissini DA, McGirr JA, Patel SS, David JR, Matute DR. The Rate of Evolution of Postmating-Prezygotic Reproductive Isolation in Drosophila. Mol Biol Evol. 2018;35: 312–334. doi:10.1093/molbev/msx271

131. Vakirlis N, Hebert AS, Opulente DA, Achaz G, Hittinger CT, Fischer G, et al. A Molecular Portrait of De Novo Genes in Yeasts. Mol Biol Evol. 2018;35: 631–645. doi:10.1093/molbev/msx315

132. Vicens A, Borziak K, Karr TL, Roldan ERS, Dorus S. Comparative Sperm Proteomics in Mouse Species with Divergent Mating Systems. Mol Biol Evol. 2017;34: 1403–1416. doi:10.1093/molbev/msx084

133. Wang X, Xu Y, Zhang S, Cao L, Huang Y, Cheng J, et al. Genomic analyses of primitive, wild and cultivated citrus provide insights into asexual reproduction. Nat Genet. 2017;49: 765–772. doi:10.1038/ng.3839

134. Warner MR, Mikheyev AS, Linksvayer TA. Genomic Signature of Kin Selection in an Ant with Obligately Sterile Workers. Mol Biol Evol. 2017;34: 1780–1787. doi:10.1093/molbev/msx123

135. Whittle CA, Extavour CG. Causes and evolutionary consequences of primordial germ-cell specification mode in metazoans. Proc Natl Acad Sci. 2017;114: 5784–5791. doi:10.1073/pnas.1610600114

136. Wilson BA, Foy SG, Neme R, Masel J. Young genes are highly disordered as predicted by the preadaptation hypothesis of *de novo* gene birth. Nat Ecol Evol. 2017;1: 0146. doi:10.1038/s41559-017-0146

137. Yang Z, Bielawski JP. Statistical methods for detecting molecular adaptation. Trends Ecol Evol. 2000;15: 496–503. doi:10.1016/S0169-5347(00)01994-7

138. Yang K, Huang L-Q, Ning C, Wang C-Z. Two single-point mutations shift the ligand selectivity of a pheromone receptor between two closely related moth species. Dicke M, editor. eLife. 2017;6: e29100. doi:10.7554/eLife.29100

139. Young BC, Wu C-H, Gordon NC, Cole K, Price JR, Liu E, et al. Severe infections emerge from commensal bacteria by adaptive evolution. Holden MT, editor. eLife. 2017;6: e30637. doi:10.7554/eLife.30637

140. Yu FB, Blainey PC, Schulz F, Woyke T, Horowitz MA, Quake SR. Microfluidic-based mini-metagenomics enables discovery of novel microbial lineages from complex environmental samples. Thrash C, editor. eLife. 2017;6: e26580. doi:10.7554/eLife.26580

141. Yue J-X, Li J, Aigrain L, Hallin J, Persson K, Oliver K, et al. Contrasting evolutionary genome dynamics between domesticated and wild yeasts. Nat Genet. 2017;49: 913–924. doi:10.1038/ng.3847

142. Zarin T, Tsai CN, Ba ANN, Moses AM. Selection maintains signaling function of a highly diverged intrinsically disordered region. Proc Natl Acad Sci. 2017;114: E1450–E1459. doi:10.1073/pnas.1614787114

143. Zhang R, Deng P, Jacobson D, Li JB. Evolutionary analysis reveals regulatory and functional landscape of coding and non-coding RNA editing. PLOS Genet. 2017;13: e1006563. doi:10.1371/journal.pgen.1006563

144. Zhang S-J, Wang C, Yan S, Fu A, Luan X, Li Y, et al. Isoform Evolution in Primates through Independent Combination of Alternative RNA Processing Events. Mol Biol Evol. 2017;34: 2453–2468. doi:10.1093/molbev/msx212

145. Zhang W, Chen S, Abate Z, Nirmala J, Rouse MN, Dubcovsky J. Identification and characterization of Sr13, a tetraploid wheat gene that confers resistance to the Ug99 stem rust race group. Proc Natl Acad Sci. 2017;114: E9483–E9492. doi:10.1073/pnas.1706277114

146. Zhao Z-M, Campbell MC, Li N, Lee DSW, Zhang Z, Townsend JP. Detection of Regional Variation in Selection Intensity within Protein-Coding Genes Using DNA Sequence

Polymorphism and Divergence. Mol Biol Evol. 2017;34: 3006–3022. doi:10.1093/molbev/msx213

147. Kimura M, Ohta T. On Some Principles Governing Molecular Evolution. Proc Natl Acad Sci. 1974;71: 2848–2852. doi:10.1073/pnas.71.7.2848

148. Kimura M. Evolutionary Rate at the Molecular Level. Nature. 1968;217: 624–626. doi:10.1038/217624a0

149. King JL, Jukes TH. Non-Darwinian Evolution. Science. 1969;164: 788–798. doi:10.1126/science.164.3881.788

150. Cooper GM, Shendure J. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. Nat Rev Genet. 2011;12: 628–640. doi:10.1038/nrg3046

151. Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C, et al. Targeted capture and massively parallel sequencing of 12 human exomes. Nature. 2009;461: 272–276. doi:10.1038/nature08250

152. Hampe J, Franke A, Rosenstiel P, Till A, Teuber M, Huse K, et al. A genome-wide association scan of nonsynonymous SNPs identifies a susceptibility variant for Crohn disease in *ATG16L1*. Nat Genet. 2007;39: 207–211. doi:10.1038/ng1954

153. Burton PR, Clayton DG, Cardon LR, Craddock N, Deloukas P, Duncanson A, et al. Association scan of 14,500 nonsynonymous SNPs in four diseases identifies autoimmunity variants. Nat Genet. 2007;39: 1329–1337. doi:10.1038/ng.2007.17

154. Xu Z, Taylor JA. SNPinfo: integrating GWAS and candidate gene information into functional SNP selection for genetic association studies. Nucleic Acids Res. 2009;37: W600–W605. doi:10.1093/nar/gkp290

155. Fong C, Ko DC, Wasnick M, Radey M, Miller SI, Brittnacher M. GWAS Analyzer: integrating genotype, phenotype and public annotation data for genome-wide association study analysis. Bioinformatics. 2010;26: 560–564. doi:10.1093/bioinformatics/btp714

156. Huang X, Kurata N, Wei X, Wang Z-X, Wang A, Zhao Q, et al. A map of rice genome variation reveals the origin of cultivated rice. Nature. 2012;490: 497–501. doi:10.1038/nature11532

157. Wong K-C, Zhang Z. SNPdryad: predicting deleterious non-synonymous human SNPs using only orthologous protein sequences. Bioinformatics. 2014;30: 1112–1119. doi:10.1093/bioinformatics/btt769

158. Shoemaker SC, Ando N. X-rays in the Cryo-Electron Microscopy Era: Structural Biology’s Dynamic Future. Biochemistry. 2018;57: 277–285. doi:10.1021/acs.biochem.7b01031

159. Wang H-W, Wang J-W. How cryo-electron microscopy and X-ray crystallography complement each other. Protein Sci. 2017;26: 32–39. doi:10.1002/pro.3022

160. Beveridge R, Migas L, Das R, Pappu R, Kriwacki R, Barran PE. Ion Mobility Mass Spectrometry Uncovers the Impact of the Patterning of Oppositely Charged Residues on the Conformational Distributions of Intrinsically Disordered Proteins. 2018; doi:10.26434/chemrxiv.7312277.v1

161. Cohan MC, Posey AE, Grigsby SJ, Mittal A, Holehouse AS, Buske PJ, et al. Evolved sequence features within the intrinsically disordered tail influence FtsZ assembly and bacterial cell division. bioRxiv. 2018; 301622. doi:10.1101/301622

162. Chintapalli VR, Wang J, Dow JAT. Using FlyAtlas to identify better *Drosophila melanogaster* models of human disease. Nat Genet. 2007;39: 715–720. doi:10.1038/ng2049

163. Gelbart WM, Emmert DB. Flybase high throughput expression pattern data. FlyBase Anal FlybaseorgreportsFBrf0221009html 29 Oct 2013 Date Last Accessed. 2013;

164. Graveley BR, May G, Brooks AN, Carlson JW, Cherbas L, Davis CA, et al. The D. melanogaster transcriptome: modENCODE RNA-Seq data for dissected tissues. Pers Commun FlyBase 20114 13. 2011;

165. The Gene Ontology Consortium. Expansion of the Gene Ontology knowledgebase and resources. Nucleic Acids Res. 2017;45: D331–D338. doi:10.1093/nar/gkw1108

166. Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, Bhai J, et al. Ensembl 2018. Nucleic Acids Res. 2018;46: D754–D761. doi:10.1093/nar/gkx1098

167. Bischof J, Maeda RK, Hediger M, Karch F, Basler K. An optimized transgenesis system for Drosophila using germ-line-specific φC31 integrases. Proc Natl Acad Sci. 2007;104: 3312–3317. doi:10.1073/pnas.0611511104

168. Brand AH, Perrimon N. Targeted gene expression as a means of altering cell fates and generating dominant phenotypes. Development. 1993;118: 401–415.

169. El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, et al. The Pfam protein families database in 2019. Nucleic Acids Res. 2019;47: D427–D432. doi:10.1093/nar/gky995

170. Burga A, Wang W, Ben-David E, Wolf PC, Ramey AM, Verdugo C, et al. A genetic signature of the evolution of loss of flight in the Galapagos cormorant. Science. 2017;356: eaal3345. doi:10.1126/science.aal3345

171. Chan PP, Lowe TM. GtRNAdb: a database of transfer RNA genes detected in genomic sequence. Nucleic Acids Res. 2009;37: D93–D97. doi:10.1093/nar/gkn787

172. Chaney JL, Clark PL. Roles for Synonymous Codon Usage in Protein Biogenesis. Annu Rev Biophys. 2015;44: 143–166. doi:10.1146/annurev-biophys-060414-034333

173. Yannai A, Katz S, Hershberg R, Dagan T. The Codon Usage of Lowly Expressed Genes Is Subject to Natural Selection. Genome Biol Evol. 2018;10: 1237–1246. doi:10.1093/gbe/evy084

174. Bulmer M. The selection-mutation-drift theory of synonymous codon usage. Genetics. 1991;129: 897–907.

175. Yang Z, Nielsen R. Mutation-Selection Models of Codon Substitution and Their Use to Estimate Selective Strengths on Codon Usage. Mol Biol Evol. 2008;25: 568–579. doi:10.1093/molbev/msm284

176. Rodrigue N, Philippe H, Lartillot N. Mutation-selection models of coding sequence evolution with site-heterogeneous amino acid fitness profiles. Proc Natl Acad Sci. 2010; 200910915. doi:10.1073/pnas.0910915107

177. Tamuri AU, dos Reis M, Goldstein RA. Estimating the Distribution of Selection Coefficients from Phylogenetic Data Using Sitewise Mutation-Selection Models. Genetics. 2012;190: 1101–1115. doi:10.1534/genetics.111.136432

178. Gilchrist MA, Chen W-C, Shah P, Landerer CL, Zaretzki R. Estimating Gene Expression and Codon-Specific Translational Efficiencies, Mutation Biases, and Selection Coefficients from Genomic Data Alone. Genome Biol Evol. 2015;7: 1559–1579. doi:10.1093/gbe/evv087

179. Kubatko L, Shah P, Herbei R, Gilchrist MA. A codon model of nucleotide substitution with selection on synonymous codon usage. Mol Phylogenet Evol. 2016;94: 290–297. doi:10.1016/j.ympev.2015.08.026

180. Galtier N, Roux C, Rousselle M, Romiguier J, Figuet E, Glémin S, et al. Codon Usage Bias in Animals: Disentangling the Effects of Natural Selection, Effective Population Size, and GC-Biased Gene Conversion. Mol Biol Evol. 2018;35: 1092–1103. doi:10.1093/molbev/msy015

181. Galtier N, Piganeau G, Mouchiroud D, Duret L. GC-Content Evolution in Mammalian Genomes: The Biased Gene Conversion Hypothesis. Genetics. 2001;159: 907–911.

182. Rodríguez F, Oliver JL, Marín A, Medina JR. The general stochastic model of nucleotide substitution. J Theor Biol. 1990;142: 485–501. doi:10.1016/S0022-5193(05)80104-3

183. Bérard J, Guéguen L. Accurate Estimation of Substitution Rates with Neighbor-Dependent Models in a Phylogenetic Context. Syst Biol. 2012;61: 510–521. doi:10.1093/sysbio/sys024

184. Drake JW, Charlesworth B, Charlesworth D, Crow JF. Rates of Spontaneous Mutation. Genetics. 1998;148: 1667–1686.

185. Stoltzfus A, McCandlish DM. Mutational Biases Influence Parallel Adaptation. Mol Biol Evol. 2017;34: 2163–2172. doi:10.1093/molbev/msx180

186. Davis JJ, Olsen GJ. Modal Codon Usage: Assessing the Typical Codon Usage of a Genome. Mol Biol Evol. 2010;27: 800–810. doi:10.1093/molbev/msp281

187. Xiang H, Zhang R, Iii RRB, Liu T, Zhang L, Pombert J-F, et al. Comparative Analysis of Codon Usage Bias Patterns in Microsporidian Genomes. PLOS ONE. 2015;10: e0129223. doi:10.1371/journal.pone.0129223

188. Bera BC, Virmani N, Kumar N, Anand T, Pavulraj S, Rash A, et al. Genetic and codon usage bias analyses of polymerase genes of equine influenza virus and its relation to evolution. BMC Genomics. 2017;18: 652. doi:10.1186/s12864-017-4063-1

189. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J R Stat Soc Ser B Methodol. 1995;57: 289–300.

190. Oliver JL, Bernaola-Galván P, Carpena P, Román-Roldán R. Isochore chromosome maps of eukaryotic genomes. Gene. 2001;276: 47–56. doi:10.1016/S0378-1119(01)00641-2

191. Hambuch TM, Parsch J. Patterns of Synonymous Codon Usage in Drosophila melanogaster Genes With Sex-Biased Expression. Genetics. 2005;170: 1691–1700. doi:10.1534/genetics.104.038109

192. Gene Ontology Consortium. Gene Ontology Consortium: going forward. Nucleic Acids Res. 2015;43: D1049–D1056. doi:10.1093/nar/gku1179

193. Pechmann S, Chartron JW, Frydman J. Local slowdown of translation by nonoptimal codons promotes nascent-chain recognition by SRP *in vivo*. Nat Struct Mol Biol. 2014;21: 1100–1105. doi:10.1038/nsmb.2919

194. Liu H, Rahman SU, Mao Y, Xu X, Tao S. Codon usage bias in 5′ terminal coding sequences reveals distinct enrichment of gene functions. Genomics. 2017;109: 506–513. doi:10.1016/j.ygeno.2017.07.008

195. Plotkin JB, Robins H, Levine AJ. Tissue-specific codon usage and the expression of human genes. Proc Natl Acad Sci. 2004;101: 12588–12591. doi:10.1073/pnas.0404957101

196. Dittmar KA, Goodenbour JM, Pan T. Tissue-Specific Differences in Human Transfer RNA Expression. PLOS Genet. 2006;2: e221. doi:10.1371/journal.pgen.0020221

197. Crick FHC. Codon—anticodon pairing: The wobble hypothesis. J Mol Biol. 1966;19: 548–555. doi:10.1016/S0022-2836(66)80022-0

198. de Muro MA. Probe Design, Production, and Applications. In: Walker JM, Rapley R, editors. Molecular Biomethods Handbook. Totowa, NJ: Humana Press; 2008. pp. 41–53. doi:10.1007/978-1-60327-375-6_4

199. Kopp A, Blackman RK, Duncan I. Wingless, decapentaplegic and EGF receptor signaling pathways interact to specify dorso-ventral pattern in the adult abdomen of Drosophila. Development. 1999;126: 3495–3507.

200. Takashima S, Paul M, Aghajanian P, Younossi-Hartenstein A, Hartenstein V. Migration of Drosophila intestinal stem cells across organ boundaries. Development. 2013;140: 1903–1911. doi:10.1242/dev.082933

201. Clément Y, Sarah G, Holtz Y, Homa F, Pointet S, Contreras S, et al. Evolutionary forces affecting synonymous variations in plant genomes. PLOS Genet. 2017;13: e1006799. doi:10.1371/journal.pgen.1006799

202. Chen R, Davydov EV, Sirota M, Butte AJ. Non-Synonymous and Synonymous Coding SNPs Show Similar Likelihood and Effect Size of Human Disease Association. PLOS ONE. 2010;5: e13574. doi:10.1371/journal.pone.0013574

203. Campos JL, Zeng K, Parker DJ, Charlesworth B, Haddrill PR. Codon Usage Bias and Effective Population Sizes on the X Chromosome versus the Autosomes in Drosophila melanogaster. Mol Biol Evol. 2013;30: 811–823. doi:10.1093/molbev/mss222

204. Gohli J, Anmarkrud JA, Johnsen A, Kleven O, Borge T, Lifjeld JT. Female Promiscuity Is Positively Associated with Neutral and Selected Genetic Diversity in Passerine Birds. Evolution. 2013;67: 1406–1419. doi:10.1111/evo.12045

205. Martincorena I, Raine KM, Gerstung M, Dawson KJ, Haase K, Van Loo P, et al. Universal Patterns of Selection in Cancer and Somatic Tissues. Cell. 2017;171: 1029-1041.e21. doi:10.1016/j.cell.2017.09.042

206. Hodgkinson A, Eyre-Walker A. Variation in the mutation rate across mammalian genomes. Nat Rev Genet. 2011;12: 756–766. doi:10.1038/nrg3098

207. Singh ND, Arndt PF, Clark AG, Aquadro CF. Strong Evidence for Lineage and Sequence Specificity of Substitution Rates and Patterns in Drosophila. Mol Biol Evol. 2009;26: 1591–1605. doi:10.1093/molbev/msp071

208. Gingold H, Tehler D, Christoffersen NR, Nielsen MM, Asmar F, Kooistra SM, et al. A Dual Program for Translation Regulation in Cellular Proliferation and Differentiation. Cell. 2014;158: 1281–1292. doi:10.1016/j.cell.2014.08.011

209. Yoshihisa T, Ohshima C, Yunoki-Esaki K, Endo T. Cytoplasmic splicing of tRNA in Saccharomyces cerevisiae. Genes Cells. 2007;12: 285–297. doi:10.1111/j.1365-2443.2007.01056.x

210. Wichtowska D, Turowski TW, Boguta M. An interplay between transcription, processing, and degradation determines tRNA levels in yeast. Wiley Interdiscip Rev RNA. 2013;4: 709–722. doi:10.1002/wrna.1190

211. Turowski TW, Tollervey D. Transcription by RNA polymerase III: insights into mechanism and regulation. Biochem Soc Trans. 2016;44: 1367–1375. doi:10.1042/BST20160062

212. Shukla A, Bhargava P. Regulation of tRNA gene transcription by the chromatin structure and nucleosome dynamics. Biochim Biophys Acta BBA - Gene Regul Mech. 2018;1861: 295–309. doi:10.1016/j.bbagrm.2017.11.008

213. Wen P, Xiao P, Xia J. dbDSM: a manually curated database for deleterious synonymous mutations. Bioinformatics. 2016;32: 1914–1916. doi:10.1093/bioinformatics/btw086

214. Meurer A, Smith CP, Paprocki M, Čertík O, Kirpichev SB, Rocklin M, et al. SymPy: symbolic computing in Python. PeerJ Comput Sci. 2017;3: e103. doi:10.7717/peerj-cs.103

215. Kraft D. A software package for sequential quadratic programming. Forschungsbericht Dtsch Forsch- Vers Für Luft- Raumfahrt DFVLR. 1988;88–28.

216. Brown A, Shao S, Murray J, Hegde RS, Ramakrishnan V. Structural basis for stop codon recognition in eukaryotes. Nature. 2015;524: 493–496. doi:10.1038/nature14896

217. Warnes GR, Bolker B, Bonebakker L, Gentleman R, Liaw WHA, Lumley T, et al. gplots: Various R Programming Tools for Plotting Data. R package version 3.0.1. Httpscranr-Proj. 2015;

218. Lu B, LaMora A, Sun Y, Welsh MJ, Ben-Shahar Y. ppk23-Dependent Chemosensory Functions Contribute to Courtship Behavior in Drosophila melanogaster. PLOS Genet. 2012;8: e1002587. doi:10.1371/journal.pgen.1002587

219. Hill A, Zheng X, Li X, McKinney R, Dickman D, Ben-Shahar Y. The Drosophila postsynaptic DEG/ENaC channel ppk29 contributes to excitatory neurotransmission. J Neurosci. 2017; 3850–16. doi:10.1523/JNEUROSCI.3850-16.2017

220. Schindelin J, Arganda-Carreras I, Frise E, Kaynig V, Longair M, Pietzsch T, et al. Fiji: an open-source platform for biological-image analysis. Nat Methods. 2012;9: 676–682. doi:10.1038/nmeth.2019

221. Chartier M, Gaudreault F, Najmanovich R. Large-scale analysis of conserved rare codon clusters suggests an involvement in co-translational molecular recognition events. Bioinformatics. 2012;28: 1438–1445. doi:10.1093/bioinformatics/bts149

222. Clarke TF, Clark PL. Increased incidence of rare codon clusters at 5’ and 3’ gene termini: implications for function. BMC Genomics. 2010;11: 118. doi:10.1186/1471-2164-11-118

223. Clarke IV TF, Clark PL. Rare Codons Cluster. PLOS ONE. 2008;3: e3412. doi:10.1371/journal.pone.0003412

224. Chaney JL, Steele A, Carmichael R, Rodriguez A, Specht AT, Ngo K, et al. Widespread position-specific conservation of synonymous rare codons within coding sequences. PLOS Comput Biol. 2017;13: e1005531. doi:10.1371/journal.pcbi.1005531

225. Mita K, Ichimura S, Zama M, James TC. Specifie codon usage pattern and its implications on the secondary structure of silk fibroin mRNA. J Mol Biol. 1988;203: 917–925. doi:10.1016/0022-2836(88)90117-9

226. Alvarez-Valin F, Lamolle G, Bernardi G. Isochores, GC 3 and mutation biases in the human genome. Gene. 2002;300: 161–168. doi:10.1016/S0378-1119(02)01043-0

227. Peng Z, Zaher H, Ben-Shahar Y. Natural selection on gene-specific codon usage bias is common across eukaryotes. bioRxiv. 2018; 292938. doi:10.1101/292938

228. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: tool for the unification of biology. Nat Genet. 2000;25: 25–29. doi:10.1038/75556

229. Krogh A, Larsson B, von Heijne G, Sonnhammer ELL. Predicting transmembrane protein topology with a hidden markov model: application to complete genomes11Edited by F. Cohen. J Mol Biol. 2001;305: 567–580. doi:10.1006/jmbi.2000.4315

230. von Heijne G. Membrane protein structure prediction: Hydrophobicity analysis and the positive-inside rule. J Mol Biol. 1992;225: 487–494. doi:10.1016/0022-2836(92)90934-C

231. Potter SC, Luciani A, Eddy SR, Park Y, Lopez R, Finn RD. HMMER web server: 2018 update. Nucleic Acids Res. 2018;46: W200–W204. doi:10.1093/nar/gky448

232. Cestèle S, Catterall WA. Molecular mechanisms of neurotoxin action on voltage-gated sodium channels. Biochimie. 2000;82: 883–892. doi:10.1016/S0300-9084(00)01174-3

233. Yu FH, Catterall WA. Overview of the voltage-gated sodium channel family. Genome Biol. 2003;4: 207. doi:10.1186/gb-2003-4-3-207

234. Liebeskind BJ, Hillis DM, Zakon HH. Evolution of sodium channels predates the origin of nervous systems in animals. Proc Natl Acad Sci. 2011;108: 9154–9159. doi:10.1073/pnas.1106363108

235. Ramaswami M, Tanouye MA. Two sodium-channel genes in Drosophila: implications for channel diversity. Proc Natl Acad Sci. 1989;86: 2079–2082. doi:10.1073/pnas.86.6.2079

236. Hong CS, Ganetzky B. Spatial and temporal expression patterns of two sodium channel genes in Drosophila. J Neurosci. 1994;14: 5160–5169. doi:10.1523/JNEUROSCI.14-09-05160.1994

237. Rohl CA, Boeckman FA, Baker C, Scheuer T, Catterall WA, Klevit RE. Solution Structure of the Sodium Channel Inactivation Gate,. Biochemistry. 1999;38: 855–861. doi:10.1021/bi9823380

238. Qu Y, Rogers JC, Tanada TN, Catterall WA, Scheuer T. Phosphorylation of S1505 in the cardiac Na+ channel inactivation gate is required for modulation by protein kinase C. J Gen Physiol. 1996;108: 375–379. doi:10.1085/jgp.108.5.375

239. Numann R, Catterall WA, Scheuer T. Functional modulation of brain sodium channels by protein kinase C phosphorylation. Science. 1991;254: 115–118. doi:10.1126/science.1656525

240. Scheuer T. Regulation of sodium channel activity by phosphorylation. Semin Cell Dev Biol. 2011;22: 160–165. doi:10.1016/j.semcdb.2010.10.002

241. Lindsay HA, Baines R, ffrench-Constant R, Lilley K, Jacobs HT, O'Dell KMC. The Dominant Cold-Sensitive Out-Cold Mutants of Drosophila melanogaster Have Novel Missense Mutations in the Voltage-Gated Sodium Channel Gene paralytic. Genetics. 2008;180: 873–884. doi:10.1534/genetics.108.090951

242. Fagerberg L, Hallström BM, Oksvold P, Kampf C, Djureinovic D, Odeberg J, et al. Analysis of the Human Tissue-specific Expression by Genome-wide Integration of Transcriptomics and Antibody-based Proteomics. Mol Cell Proteomics. 2014;13: 397–406. doi:10.1074/mcp.M113.035600

243. Katoh K, Standley DM. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Mol Biol Evol. 2013;30: 772–780. doi:10.1093/molbev/mst010

244. Li W, Cowley A, Uludag M, Gur T, McWilliam H, Squizzato S, et al. The EMBL-EBI bioinformatics web and programmatic tools framework. Nucleic Acids Res. 2015;43: W580–W584. doi:10.1093/nar/gkv279

245. Schwarz-Schilling M, Kim J, Cuba C, Weitz M, Franco E, Simmel FC. Building a Synthetic Transcriptional Oscillator. In: Coutts AS, Weston L, editors. Cell Cycle Oscillators: Methods and Protocols. New York, NY: Springer New York; 2016. pp. 185–199. doi:10.1007/978-1-4939-2957-3_10

246. Xi L, Schmidt JC, Zaug AJ, Ascarrunz DR, Cech TR. A novel two-step genome editing strategy with CRISPR-Cas9 provides new insights into telomerase action and TERT gene expression. Genome Biol. 2015;16: 231. doi:10.1186/s13059-015-0791-1

Appendices

Appendix 2.1 Sequences of Fluorescent Reporter cDNAs

>mCherry_Common

ATGGTGAGCAAGGGCGAGGAGGATAACATGGCCATCATCAAGGAGTTCATGCGCTT
CAAGGTGCACATGGAGGGCAGCGTGAACGGCCACGAGTTCGAGATCGAGGGCGAG
GGCGAGGGCCGCCCCTACGAGGGCACCCAGACCGCCAAGCTGAAGGTGACCAAGG
GCGGCCCCCTGCCCTTCGCCTGGGATATCCTGAGCCCCCAGTTCATGTACGGCAGCA
AGGCCTACGTGAAGCACCCCGCCGATATCCCCGATTACCTGAAGCTGAGCTTCCCCG
AGGGCTTCAAGTGGGAGCGCGTGATGAACTTCGAGGATGGCGGCGTGGTGACCGTG
ACCCAGGATAGCAGCCTGCAGGATGGCGAGTTCATCTACAAGGTGAAGCTGCGCGG
CACCAACTTCCCCAGCGATGGCCCCGTGATGCAGAAGAAGACCATGGGCTGGGAGG
CCAGCAGCGAGCGCATGTACCCCGAGGATGGCGCCCTGAAGGGCGAGATCAAGCAG
CGCCTGAAGCTGAAGGATGGCGGCCACTACGATGCCGAGGTGAAGACCACCTACAA
GGCCAAGAAGCCCGTGCAGCTGCCCGGCGCCTACAACGTGAACATCAAGCTGGATA
TCACCAGCCACAACGAGGATTACACCATCGTGGAGCAGTACGAGCGCGCCGAGGGC
CGCCACAGCACCGGCGGCATGGATGAGCTGTACAAGAGCCGCTAG

>EGFP_Common

ATGAGCCGCGTGAGCAAGGGCGAGGAGCTGTTCACCGGCGTGGTGCCCATCCTGGT
GGAGCTGGATGGCGATGTGAACGGCCACAAGTTCAGCGTGAGCGGCGAGGGCGAG
GGCGATGCCACCTACGGCAAGCTGACCCTGAAGTTCATCTGCACCACCGGCAAGCT
GCCCGTGCCCTGGCCCACCCTGGTGACCACCCTGACCTACGGCGTGCAGTGCTTCAG
CCGCTACCCCGATCACATGAAGCAGCACGATTTCTTCAAGAGCGCCATGCCCGAGG
GCTACGTGCAGGAGCGCACCATCTTCTTCAAGGATGATGGCAACTACAAGACCCGC
GCCGAGGTGAAGTTCGAGGGCGATACCCTGGTGAACCGCATCGAGCTGAAGGGCAT
CGATTTCAAGGAGGATGGCAACATCCTGGGCCACAAGCTGGAGTACAACTACAACA
GCCACAACGTGTACATCATGGCCGATAAGCAGAAGAACGGCATCAAGGTGAACTTC
AAGATCCGCCACAACATCGAGGATGGCAGCGTGCAGCTGGCCGATCACTACCAGCA
GAACACCCCCATCGGCGATGGCCCCGTGCTGCTGCCCGATAACCACTACCTGAGCAC
CCAGAGCGCCCTGAGCAAGGATCCCAACGAGAAGCGCGATCACATGGTGCTGCTGG
AGTTCGTGACCGCCGCCGGCATCACCCTGGGCATGGATGAGCTGTACAAGTAG

>EGFP_RareKEQ

ATGAGCCGCGTGAGCAAAGGCGAAGAACTGTTCACCGGCGTGGTGCCCATCCTGGT
GGAACTGGATGGCGATGTGAACGGCCACAAATTCAGCGTGAGCGGCGAAGGCGAA
GGCGATGCCACCTACGGCAAACTGACCCTGAAATTCATCTGCACCACCGGCAAACT
GCCCGTGCCCTGGCCCACCCTGGTGACCACCCTGACCTACGGCGTGCAATGCTTCAG
CCGCTACCCCGATCACATGAAACAACACGATTTCTTCAAAAGCGCCATGCCCGAAG
GCTACGTGCAAGAACGCACCATCTTCTTCAAAGATGATGGCAACTACAAAACCCGC
GCCGAAGTGAAATTCGAAGGCGATACCCTGGTGAACCGCATCGAACTGAAAGGCAT
CGATTTCAAAGAAGATGGCAACATCCTGGGCCACAAACTGGAATACAACTACAACA
GCCACAACGTGTACATCATGGCCGATAAACAAAAAAACGGCATCAAAGTGAACTTC
AAAATCCGCCACAACATCGAAGATGGCAGCGTGCAACTGGCCGATCACTACCAACA
AAACACCCCCATCGGCGATGGCCCCGTGCTGCTGCCCGATAACCACTACCTGAGCAC
CCAAAGCGCCCTGAGCAAAGATCCCAACGAAAAACGCGATCACATGGTGCTGCTGG
AATTCGTGACCGCCGCCGGCATCACCCTGGGCATGGATGAACTGTACAAATAG

Appendix 2.2 Analytical Solutions of Equation Systems Used to Estimate μ Values

Please refer to Appendix 2.2.docx.

Appendix 2.3 tRNA Northern Blot Images

Please refer to Appendix 2.3.zip.

Appendix 2.4 Fluorescent Reporter Expression Images

Please refer to Appendix 2.4.zip.

Appendix 2.5 Real-time qRT-PCR Data

Please refer to Appendix 2.5.xlsx.

Appendix 2.6 Computer Code Used in Chapter 2

Please refer to Appendix 2.6.zip.

Appendix 3.1 Detected PFCCs

Please refer to Appendix 3.1.xlsx.

Appendix 3.2 Association Between Genes with N-terminal PFCCs (RLI<0.1) and GO Terms

Please refer to Appendix 3.2.xlsx.

Appendix 3.3 Association Between Genes with PFCCs and GO Terms

Please refer to Appendix 3.3.xlsx.

Appendix 3.4 Association Between PFCCs and Transmembrane Helices

Please refer to Appendix 3.4.xlsx.

Appendix 3.5 Association Between PFCCs and Pfam Protein Domains

Please refer to Appendix 3.5.xlsx.

Appendix 3.6 Nav Homologs

Please refer to Appendix 3.6.xlsx.

Appendix 3.7 Computer Code Used in Chapter 3

Please refer to Appendix 3.7.zip.