OPEN ACCESS Freely available online

PLOS | GENETICS

Strong Purifying Selection at Synonymous Sites in *D. melanogaster*

David S. Lawrie[1,2]*, Philipp W. Messer[2], Ruth Hershberg[2,3], Dmitri A. Petrov[2]

1 Department of Genetics, Stanford University, Stanford, California, United States of America, **2** Department of Biology, Stanford University, Stanford, California, United States of America, **3** Rachel and Menachem Mendelovitch Evolutionary Processes of Mutation and Natural Selection Research Laboratory, Department of Genetics, The Ruth and Bruce Rappaport Faculty of Medicine, Technion–Israel Institute of Technology, Haifa, Israel

Abstract

Synonymous sites are generally assumed to be subject to weak selective constraint. For this reason, they are often neglected as a possible source of important functional variation. We use site frequency spectra from deep population sequencing data to show that, contrary to this expectation, 22% of four-fold synonymous (4D) sites in *Drosophila melanogaster* evolve under very strong selective constraint while few, if any, appear to be under weak constraint. Linking polymorphism with divergence data, we further find that the fraction of synonymous sites exposed to strong purifying selection is higher for those positions that show slower evolution on the Drosophila phylogeny. The function underlying the inferred strong constraint appears to be separate from splicing enhancers, nucleosome positioning, and the translational optimization generating canonical codon bias. The fraction of synonymous sites under strong constraint within a gene correlates well with gene expression, particularly in the mid-late embryo, pupae, and adult developmental stages. Genes enriched in strongly constrained synonymous sites tend to be particularly functionally important and are often involved in key developmental pathways. Given that the observed widespread constraint acting on synonymous sites is likely not limited to Drosophila, the role of synonymous sites in genetic disease and adaptation should be reevaluated.

Citation: Lawrie DS, Messer PW, Hershberg R, Petrov DA (2013) Strong Purifying Selection at Synonymous Sites in *D. melanogaster*. PLoS Genet 9(5): e1003527. doi:10.1371/journal.pgen.1003527

Editor: Joshua B. Plotkin, University of Pennsylvania, United States of America

Received January 24, 2013; **Accepted** April 8, 2013; **Published** May 30, 2013

Funding: The work was supported by the NIH grants RO1GM100366 and RO1GM097415 to DAP. RH is supported by an ERC FP7 CIG grant, by a Yigal Allon Fellowship awarded by the Israeli Council for Higher Education, and by the Robert J. Shillman Career Advancement Chair. PWM was supported by NIH grants RO1GM100366 and RO1GM097415. DSL is supported by the Stanford Genome Training Program (SGTP; NIH/NHGRI) and by NIH grants RO1GM100366 and RO1GM097415. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: dlawrie@stanford.edu

Introduction

As there are 64 codons and only 20 amino acids, most amino acids can be encoded by more than a single codon. Mutations that alter coding sequences (CDS), but do not alter amino acid sequences are referred to as synonymous mutations. Synonymous sites are then the collection of potential synonymous mutations present in a gene. Predicated on the assumption that the CDS of a gene is simply the recipe for making the protein, synonymous mutations were long thought to have no functional effect, in other words to be "silent" and thus selectively neutral [1,2]. As a result, synonymous variation is often used as the neutral reference when measuring selection at functionally important, non-synonymous sites [3–7].

The observation of codon usage bias in many organisms was the first indication of possible functionality encoded by synonymous sites [8,9]. Different codons for the same amino acid are often utilized at unequal frequencies across the genome. Highly expressed genes and codons encoding functionally important amino acids generally display particularly biased patterns of codon usage [9–11]. This observation led to the theory that selection for translation optimization generates higher levels of codon bias [12–15]. In other words, it is thought that the speed and accuracy of mRNA translation is higher for a subset of codons, referred to as "optimal" ("preferred") codons [14–19]. Such codons are translated more accurately and more efficiently because they are recognized by more abundant tRNA molecules with more specific anti-codon binding [14,20,21]. While this kind of selection acting on synonymous mutations is widely accepted, it is generally estimated to be weak - nearly, if not quite, neutral [22–31]. Synonymous variation is therefore still often thought to lack any major functional or evolutionary importance. In this paper, we further investigate the functionality of synonymous sites through detecting the action of purifying selection. If synonymous sites harbor highly deleterious variants under strong purifying selection, then that must change our view of the functional importance of synonymous sites and their potential role in genetic disease, as a possible source for adaptation, and as the neutral foil in tests for selection.

Previous tests for selection on synonymous sites have often been consistent with the presence of weak purifying selection operating on synonymous variation. Using the rate of divergence between species, the signal of purifying selection comes from a lower number of inferred substitutions on a phylogenetic tree at sites allowing for synonymous mutations, compared to the expectation provided by a neutral reference. Simply comparing the rate of evolution between a test and a neutral reference set can be problematic when weak purifying selection and mutational biases interact [32]. Nonetheless, synonymous sites do indeed appear to evolve slower than expected under neutrality for many organisms in a manner seemingly consistent with weak selection [22,29–31,33–40]. More evidence for

Author Summary

Synonymous mutations do not alter the sequence of amino acids encoded by the gene in which they occur. These synonymous mutations were thus long thought to have no effect on the function of the ensuing protein or the fitness of the organism. At four-fold degenerate sites, every possible mutation is synonymous. For this reason, they are often neglected as a possible source of important functional changes. Using a deep sampling of the variation within a population of the fruit fly *Drosophila melanogaster*, we show that, contrary to this expectation, 22% of synonymous mutations at four-fold degenerate sites are strongly deleterious to the point of absence in the Drosophila population. The underlying biological function disrupted by these mutations is unknown, but is not related to the forces generally believed to be the principal actors shaping the evolution of synonymous sites. Genes with many such possible deleterious synonymous mutations tend to be particularly functionally important, highly expressed, and often involved in key developmental pathways. Given that the observed functional importance of synonymous sites is likely not limited to Drosophila, the role of synonymous sites in genetic disease and adaptation should be reevaluated.

weak selection acting at synonymous sites comes from the study of polymorphism within species. Purifying selection reduces the frequency of deleterious alleles in the population. To measure its effect, the site frequency spectrum (SFS) tabulates the fraction of observed SNPs in all frequency classes across the sites of interest. The overabundance of low frequency SNPs relative to the neutral expectation is the signal of purifying selection operating on the test sites. From this signal, one can calculate the strength of the selective force and the proportion of the test sites it affects [41,42]. Such methods have been applied to studying the effects of selection on synonymous sites in a variety of Drosophila species, and have found evidence of weak selection - often favoring optimal codons [24–28].

Studies using divergence and polymorphism to infer selection as described above are, however, unable to detect the action of strong purifying selection. Tests that rely on divergence are limited in power to distinguish strong purifying selection from weak or moderate purifying selection. The problem lies in the efficacy of purifying selection (constraint) over a tree: a small, linear increase in the strength of constant purifying selection causes a large, exponential drop in the rate of evolution [43–45]. Weak to moderate constraint is thus capable of conserving sites over even large phylogenetic distances and increasing the number of species/tree length results in only a limited increase in power to distinguish strong from moderate or weak purifying selection [32]. Unlike tests on divergence that have difficulty distinguishing between strong and weak constraint, tests using the SFS of observed polymorphism can miss strong purifying selection entirely. While both weak and strong constraint eliminate variation from the population, strong selection does so far more efficiently. Therefore, at sites of strong constraint there are few SNPs and only at very low frequency in the population. Without sequencing enough members of a population to attain a deep sample, such SNPs will not be in the SFS of observed polymorphism. With no signal in the shape of the SFS from shallow population sequencing data, any strong selection acting on synonymous sites could not be detected via these methods.

While strong selection does not significantly affect the shape of a shallow-sample SFS, the lack of polymorphism can itself be a powerful signal of the action of selection [46,47]. Knowing how many mutations should be present in the population sample, as compared to the amount actually present, can allow the estimation of the fraction of sites under strong selection. To do this, one needs a large sample of sites as the density of polymorphism is always low - on the order of a few percent. Differentiating between low densities in the test set and the neutral reference thus requires a large number of sites from each.

Note that both weak purifying selection and lower rates of mutation can likewise cause a paucity of SNPs. Ultra-low frequency variants can distinguish the signal of strong constraint from that of a variation in the mutation rate between the neutral reference and the set of sites being tested. While mutational cold spots only lead to a lower number of SNPs, under strong purifying selection some mutations should still be observable at very low frequencies in a deep enough sample of the population. Weak selection, meanwhile, will affect the shape of the spectrum beyond the rare alleles and can be estimated from that. Combined, the lack of polymorphism and the excess of rare variants from a genome-wide, deep sample, could give the necessary power to quantify the intensity of the strong constraint and the fraction of sites it affects. Thus, our dataset needs to include both a wide sample of sites from the genome, as well as a deep sample from the population.

The Drosophila Genetic Reference Panel (DGRP) for *D. melanogaster* provides such a dataset [48]. With 168 sequenced-inbred lines, this data set represents the whole genome (thus providing us with the widest possible sample of sites from the genome). The data also provides a deep sample of the variation within the *D. melanogaster* population of North Carolina. Using DGRP polymorphism, we estimate that, contrary to long held expectations, a substantial fraction of the synonymous sites in *D. melanogaster* is evolving under strong selective constraint. The discovery of strong selection on codon usage in Drosophila should dramatically change our collective perspective on the functional and evolutionary significance of synonymous sites.

Results

To detect the action of selection on DGRP variation in synonymous sites, we need a neutral reference against which to compare the site-frequency spectrum and SNP density of synonymous sites. Short introns in Drosophila have been shown to be evolving neutrally or nearly so [49–52]. We therefore use sites from introns shorter than 86 bp as the neutral reference, also removing the edges of these introns, 16 bp away from the intron start and 6 bp away from the intron end, as they may contain splicing elements [52].

For our collection of synonymous sites, to prevent any confusion of synonymous vs. non-synonymous selection acting on a given codon position, we focused on the third codon positions of the four-fold degenerate amino acids (Proline, Alanine, Threonine, Glycine, and Valine). All possible mutations in the third codon position are synonymous for these five amino acids. The third codon positions of these amino acids will from hereon be referred to as 4D sites. So that we could later relate our results from this analysis on polymorphism within *D. melanogaster* to divergence across Drosophila, we used only those 4D sites from genes with 1–1 orthologs across the twelve sequenced Drosophila species as our test set [53].

To normalize the number of *D. melanogaster* strains sequenced at each position and any sequencing differences between short introns and 4D sites, we took only those positions which had their base pair called in at least 130 out of 168 strains and further

resampled all SNPs to a depth of 130 homozygous strains (see Materials and Methods). The resulting data set consists of 863,972 4D sites, 5.58% of them containing a SNP, and 870,364 sites in short introns with 6.0% of these being polymorphic. By comparing the density and SFS of polymorphism between 4D and short intron sites, we can quantify the strength of selective forces operating on 4D sites and the fraction of such sites they affect.

Before doing so, several potential confounding factors to such an analysis need to be removed. The greatest of these is the difference in GC content between short introns and 4D synonymous sites. The GC content of 4D sites in *D. melanogaster* is 64%, compared to only 31% for short introns. Mutation is known to be generally biased towards A/T with particularly high rates of mutation from C:G to T:A [30,51,54–57]. With a higher GC content, 4D sites are thus expected to be subject to a higher mutation rate on average compared to short introns increasing their relative density of polymorphism. This mutational-GC effect could mask any effects of selection on 4D sites, which if present, would reduce the density of polymorphism in 4D sites compared to short introns.

A further complication is that there are spatial variations in the rates of mutation and recombination and in the amount and severity of linked selection across the genome [51,58–61]. Sweeps, strong background selection, and variation in mutation rates may all influence the density of polymorphism in short intron sites relative to 4D sites [62,63].

Outlined in Figure 1A is our bootstrap procedure to control for GC content and spatial variation in levels of polymorphism. We first pair 4D sites with short intron positions, requiring a short intron and 4D pair to have identical major alleles and be within 1 KB of each other. Such pairs are then sampled with replacement, first drawing a 4D site and then picking at random one of its possible short intron partners, until the number of random pairs drawn equals the population of all 4D sites with short intron pairs. This process matches the GC content of the neutral reference to the test set, ensures the same spatial sampling of the intronic and 4D sites, and as a side bonus, normalizes the total number from each.

Strong purifying selection on synonymous sites

Figure 1B shows the SFS for the SNPs in short introns and 4D sites from one bootstrap run. The shapes of the short intron and 4D spectra appear nearly identical. However, this similarity in the shapes of the spectra for 4D and short intron sites belies a large disparity in the density of polymorphism between the two sets. We measured the density of polymorphism in short intron and 4D sites and calculated the standard error of our measurement over 10 bootstrap runs. We find that 4D sites have approximately 22.1% (+/−0.6%) fewer segregating sites as compared to short introns (Figure 1C).

To account for the relative paucity of polymorphism in 4D sites when the spectra of 4D and short introns SNPs are so similar, we combined both facets of information in a maximum-likelihood method allowing for the effects of multiple selective forces and demography on polymorphism (see Materials and Methods). We extended the SFS to include the number of non-polymorphic sites, the "zero"-frequency class, in our 4D and short intron bootstrap samples. Using such "amplitude" information along with the distribution of alleles over the observed frequency classes enables better maximum-likelihood estimation for parameters of strong selection. In this model, selection is parameterized by the effective selection coefficient $4N_e s$: where s is the selection coefficient and N_e is the effective population size of the organism. In our maximum-likelihood model, we used three categories of selection, neutral: $4N_e s = 0$, weak purifying: $|4N_e s| < 5$, strong purifying: $|4N_e s| > 100$. The point estimates for the fraction of sites and the strength of constraint in each selection category can be seen in Table 1. While there is no evidence of extant weak selection acting differentially on 4D sites and short introns, ~22% of 4D sites are estimated to be under very strong constraint, $4N_e s \sim -283 +/- 28.3$ (standard error estimate by bootstrap). When a coarse-grained demographic correction was applied to the SFS we obtained results that, though quantitatively are somewhat different ($4N_e s \sim -370.1 +/- 105$), are qualitatively similar in that for both cases $100 << |4N_e s| << 700$ – the calculable limit of our program (see Text S1).

Signal of strong selection not due to mutation. One concern is the existence of some mutational difference between short introns and 4D sites beyond regional effects accounted for in the bootstrap, lowering the density of polymorphism in 4D sites relative to short introns. Such a difference is unlikely. As short introns are also transcribed, any transcription-coupled repair should affect both short introns and 4D sites equally, and its effects should be controlled for in our analysis. Furthermore, there have been no reports of such transcription-coupled repair lowering the mutation rate of transcribed regions in *D. melanogaster* [64]. The rate of mutation at a site may also be affected by that site's immediate neighbors, a phenomenon known as context-dependence [65–67]. However, when we used matching triplets in the bootstrap (i.e. the 4D site plus its immediate 3′ and 5′ neighbors paired against a similar trio of short intron sites), to account for any possible dinucleotide biases, we found no evidence of a mutational difference between 4D and short intron sites affecting our signal of strong constraint (see Text S2A).

Our maximum-likelihood estimation of the intensity of strong selection is itself evidence against a mutational force underlying the disparity in polymorphism between intronic and 4D sites. Simply lowering the mutation rate of 4D sites while maintaining their neutrality with respect to short introns would affect only the relative density of polymorphism and not the shape of the SFS. Therefore, a lower mutational rate at 4D positions acts equivalently to infinitely strong purifying selection (i.e. presence of lethals) in the above maximum-likelihood estimation. With our sample depth, the maximum-likelihood estimation has the power to detect the difference between a lower rate of mutation at 4D sites and strong constraint operating at $4N_e s \sim -300$ (see Text S2B). Infinite selection or mutation on 4D sites would yield an estimate of $4N_e s \sim -700$ (see Text S2B), far outside the range of any estimate of the strength of selection we made. The finite estimate of constraint that we obtained with and without demographic correction argues against the possibility that a mutational difference between 4D sites and short introns explains our results.

With no significant involvement of other forces eliminating synonymous polymorphism, the percent of missing variation in 4D sites is therefore a reasonable proxy for the fraction of sites under strong constraint. Thus ~22% of synonymous sites appear to be under very strong purifying selection in *D. melanogaster*.

Evolutionary history of synonymous sites across Drosophila

Exposing the action of the strong constraint on divergence between Drosophila species affirms the functional importance of these 4D sites across evolutionary history and reveals how these constrained synonymous sites evolve. If the strong constraint at 4D sites we identified within *D. melanogaster* has been constant across Drosophila, we would expect it to result in the complete conservation of the constrained 4D sites. If, on the other hand, the strong constraint is not constant and there is functional turnover at these sites, then we would expect to see substitutions

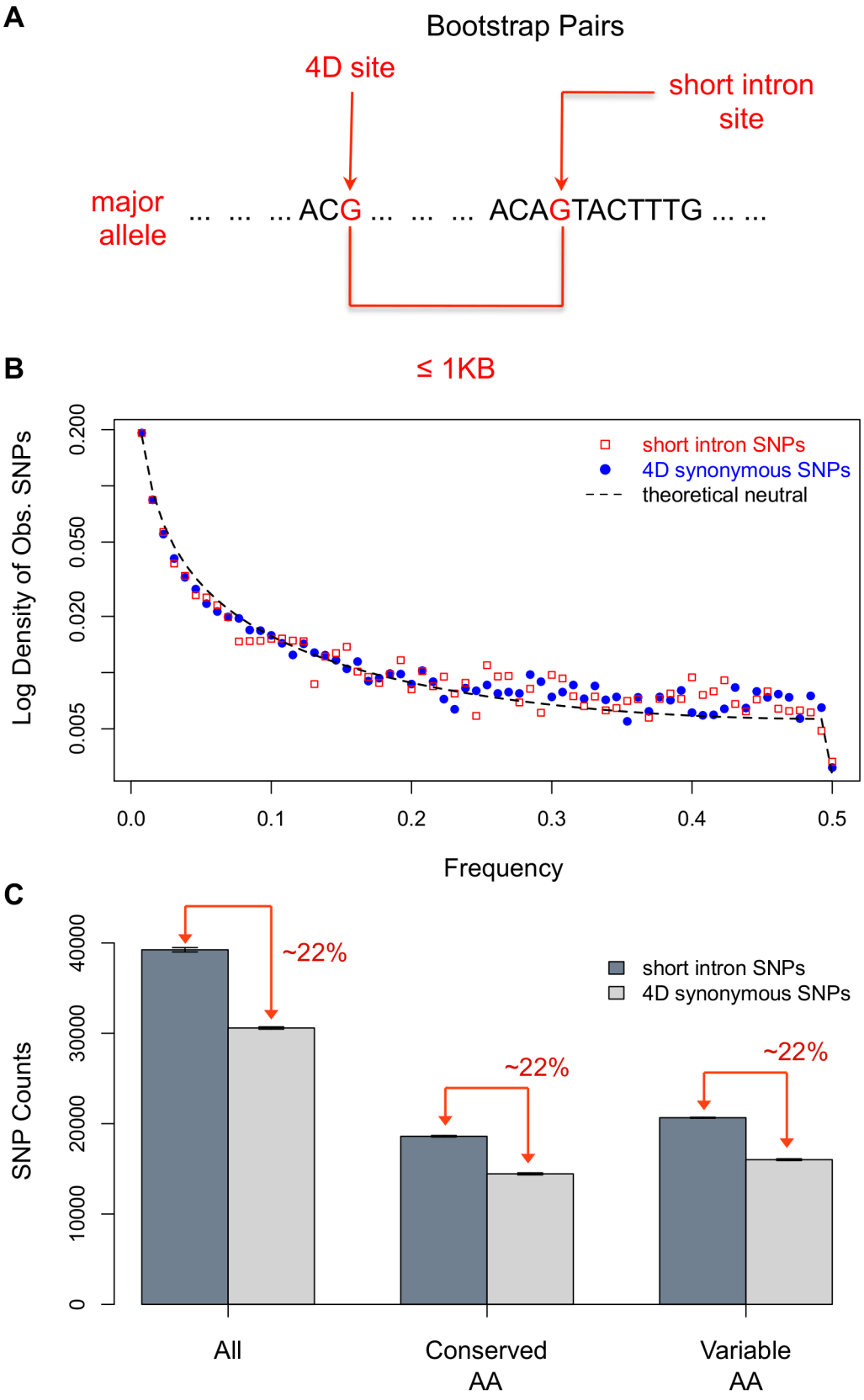

Figure 1. The signal of strong selection acting on 4D sites. (A) Overview of the bootstrap method. We sample 4D sites and their nearby (<1 KB apart) short intron pairs with replacement in order to control for linked selection and variation in GC content and mutation/recombination rates between the neutral reference (short introns) and the test set (4D sites). The short intron, 4D pair must have the same nucleotide as their major allele. (B) The folded Site Frequency Spectra (SFS) of observed SNPs from short introns, 4D sites, and the theoretical neutral distribution in a

population with constant size. The SNPs were resampled to 130 strains and folded using the minor allele frequency. (C) The ratio of the amount of polymorphism in short introns versus 4D sites in all, conserved, and variable amino acids with standard error bars. Conserved amino acids are those present and identical in the 12 sequenced Drosophila genomes. Variable amino acids are defined as being not conserved according to the above definition. Ten bootstraps were done for each category (all, conserved, and variable) of 4D site. Lifting the restriction on distance and only controlling for GC content in the bootstrap produces identical results as above (not shown). To be conservative, we continued to use the distance restriction in the bootstrap. Note, had we simply taken the density of polymorphism as is without correction of GC content, we would've only seen a 7% drop in the density of polymorphism from short introns to 4D sites (5.58% vs 6.0% segregating in 4D versus short intron sites).
doi:10.1371/journal.pgen.1003527.g001

occurring even at constrained sites along the Drosophila species tree. In order to compare the divergence between species to the constraint within a species, we considered only those 4D sites in amino acids conserved across the twelve Drosophila species from *D. melanogaster* to *D. grimshawi*. This simplifies the analysis as only the synonymous third position of the codon has been allowed to change over time. Thus, we can focus solely on the evolution of the synonymous site itself rather than consider the evolution of the entire codon. Figure 1C shows that the conservation of the amino acid has no bearing on the fraction of missing polymorphism in 4D sites. As such, the 4D sites of conserved amino acids provide a representative sample with which to study the strong constraint over the evolution of all 4D sites.

The gene orthologs in the other species were obtained from the 12 Drosophila Genome Consortium data realigned by PRANK [53,68,69]. We used the established 12 Drosophila species tree and re-estimated the branch lengths on the aligned 4D sites with PhyML (see Materials and Methods) [70]. From these alignments we removed the sequences belonging to *D. melanogaster* and *D. willistoni*. The *D. melanogaster* sequences were removed because the polymorphism data was extracted from this species and we wished to avoid a false concordance between the results from polymorphism and divergence. The *D. willistoni* sequences were removed, because the branch length leading to *D. willistoni* is long and the codon bias of *D. willistoni* is significantly different than from the rest of the twelve Drosophila species [71]. Having removed these species, the expected number of substitutions over the now ten Drosophila species tree for synonymous positions in otherwise conserved four-fold amino acids is estimated by PhyML at 3.1 subs/site [70]. To obtain site-wise estimates of conservation, we then inferred the number of substitutions along this tree for each 4D site independently using GERP (see Materials and Methods) [72,73].

Figure 2 shows that the percentage of sites under strong constraint declines monotonically as the rate of evolution increases. 40.8% (+/−1.9%) of completely conserved sites (0 substitution class), and only 7.1% (+/−3.0%) of the fastest evolving sites (≥9.3 substitution class) are predicted to be under strong constraint. This difference in the amount of constraint between fast and slow-evolving sites allowed us to carry out a further control for any variation in mutation rate between short introns and 4D sites. We carried out an identical bootstrap procedure but pairing slow-evolving 4D sites with neighboring fast-evolving 4D sites instead of short introns as a neutral reference. We recapitulated our result of strong constraint at 4D sites by using slow- versus fast-evolving 4D sites (see Text S2C).

This correlation between a 4D site's conservation across species and strong constraint within a species underscores the functional importance of these synonymous positions over the evolutionary history of the Drosophila clade. However, over 80% of the sites currently under strong constraint in *D. melanogaster* fall outside the 0 substitution class, i.e. are not conserved across the ten Drosophila species. Indeed, over 11% of 4D sites under strong constraint in *D. melanogaster* have each acquired 6.2 or more substitutions over the tree, evolving quickly at more than twice the average rate. As even a moderate amount of selection results in complete conservation if it has been consistent over the tree, this suggests there has been functional turnover at these functionally important synonymous sites.

Table 1. Estimated proportion of 4D sites and $4N_es$ for each selection class.

Selection Category[a]	Fraction of Sites[b]	Strength[c]
Neutral	77.4% (+/−0.6%)	0
Weak Constraint	0	N/A
Strong Constraint	22.6% (+/−0.6%)	−283 (+/−28.3)

[a]Selection categories are defined as follows => Neutral: $4N_es = 0$, Weak Constraint: $|4N_es| < 5$, and Strong Constraint: $|4N_es| > 100$ - defining Strong Constraint: $|4N_es| > 5$ gives exactly the same Maximum Likelihood Estimator (MLE) for the fraction/strength of the strong category;
[b]mean of the MLEs for the fraction of sites in each category over the ten bootstrap runs (+/− s.e.);
[c]mean of the MLEs for the strength of strong selection over the ten bootstrap runs (+/− s.e.); $4N_e\mu$ (θ) = 0.0132.
doi:10.1371/journal.pgen.1003527.t001

Codon bias

Codon bias is generally thought to be the product of background substitution biases combined with a weak selective force within genes skewing codon usage towards optimal (preferred) codons to increase translation efficiency and accuracy [19]. In Drosophila, translationally preferred codons are always G- or C-ending (except for in *D. willistoni*) [71]. The five four-fold degenerate amino acids have the following preferred codons: Alanine - GCC; Glycine - GGC; Proline - CCC; Threonine - ACC; and Valine - GTG [71]. Selection for codon bias is thus likely responsible for driving the GC content of 4D synonymous sites in *D. melanogaster* to 64% and to over 67% in the 4D sites of amino acids conserved over the 12 Drosophila species. While codon bias increases in conserved amino acids [17], as stated above, the strong selection at synonymous sites inferred in this paper does not (Figure 1C). To explore the relationship between codon bias and the strong constraint, we measured the fraction of sites under strong constraint within each codon, in unpreferred versus preferred codons conserved from *D. sechellia-D.grimshawi*, and across genes ranked by codon bias.

The codon targets of strong selection. Despite the fact that the conservation, and thus presumably the functional importance, of the amino acids appears not to matter, the fraction of 4D sites under strong constraint does fluctuate across the different amino acids: Alanine −22.3% (+/−0.9%); Glycine −15.0% (+/−1.8%); Proline −18.0% (+/−1.7%); Threonine −24.8% (+/−1.0%); Valine −28.5% (+/−1.2%). In order to identify the fraction of synonymous sites under constraint for an individual codon within each amino acid, we first assigned 4D sites to codons by their ancestral state, which we determined by parsimony using *D. sechellia* as the outgroup. We determine ancestral/derived alleles, known as "polarizing" a site, based on this principle. However, a

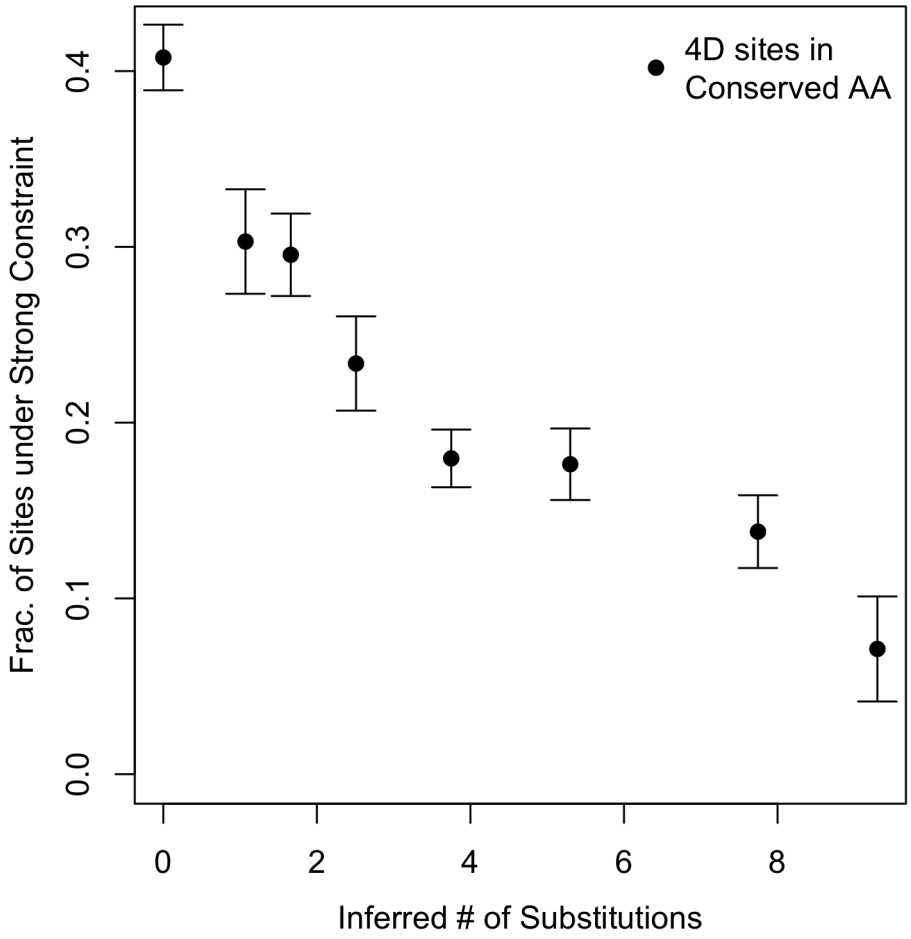

Figure 2. Conservation versus constraint at 4D sites in conserved amino acids. For each 4D site in a conserved amino acid, we use GERP to infer the number of substitutions that have occurred at that site across the Drosophila tree (removing *D. melanogaster* and *D. willistoni* from the analysis). We define eight rate classes defined by the number of inferred substitutions across the tree - a proxy for the rate of evolution at the site - and bin the 4D sites accordingly. The class of the slowest evolving sites consists of those codons completely conserved across the ten Drosophila species (0 inferred substitutions along the tree at the 4D site). The fastest evolving class meanwhile has sites with greater than or equal to 9.3 substitutions per site. The remaining substitution classes are spread at intermediate values with a view to roughly equilibrate the number of sites in each class. The substitution bins (b) are as follows: ($b_1 = 0$, $0 < b_2 \leq 1.4$, $1.4 < b_3 \leq 1.92$, $1.92 < b_4 \leq 3.10$, $3.10 < b_5 \leq 4.40$, $4.40 < b_6 \leq 6.20$, $6.20 < b_7 < 9.30$, $b_8 \geq 9.30$). 10 bootstraps were done for the 4D sites within each bin and their short introns partners. Error bars represent the s.e. of the estimates.
doi:10.1371/journal.pgen.1003527.g002

substitution between *D. sechellia* and *D. melanogaster* will cause a site to be unpolarizable. Because more monomorphic 4D sites than polymorphic ones are unpolarizable, simply removing unpolarizable sites would cause a shift in the density of SNPs in 4D sites and alter our signal of constraint. Therefore, these ancestrally ambiguous sites (<8%) were assigned to their respective codons by their major allele so as not to remove sites during polarization.

Figure 3 shows the relationship between the amount of constraint in all the 4D sites for all codons grouped by their optimality and amino acid as well as the amount of each codon in the bootstrap analysis, reflecting the abundance of that codon in the genome. Two striking observations result from this analysis. First, while optimal codons are more frequently constrained than non-optimal codons, 25% (P) versus 18% (U) over all, for any individual amino acid the optimal codon may not have the highest fraction of sites under constraint. For Proline and Valine, the 4D sites of the unpreferred codons CCA/CCG (Proline) and GTT (Valine) are the most frequently strongly constrained. Second, there are some codons that have no apparent strong constraint on their 4D site - i.e. their 4D SNP density matches or exceeds the SNP density of short introns. These codons with seemingly neutral 4D sites are also used rarely in the genome relative to the other codons for that amino acid. These results are qualitatively similar when restricting the analysis to conserved amino acids (not shown).

There would thus appear to be strong selection on codon usage beyond the canonical selection for optimal codons. Figure 3 defines which codons are "favored" by strong constraint for each of the five four-fold amino acids. Sometimes these are also the previously defined optimal codons, but sometimes they are not. Even though there is propensity of strong constraint to affect particular codons, each four-fold amino acid has more than one codon with some fraction of its synonymous positions across the genome under strong constraint.

Our procedure polarizing sites by parsimony to a single species outgroup and then by major allele can misidentify the ancestral allele. Thus SNPs can be grouped with the wrong set of monomorphic sites, subtly changing the SNP densities across the codons. For instance, the negative fraction of sites under constraint - indicative of an excess of 4D polymorphism relative to short introns - for Proline's codon CCT is likely a product of this

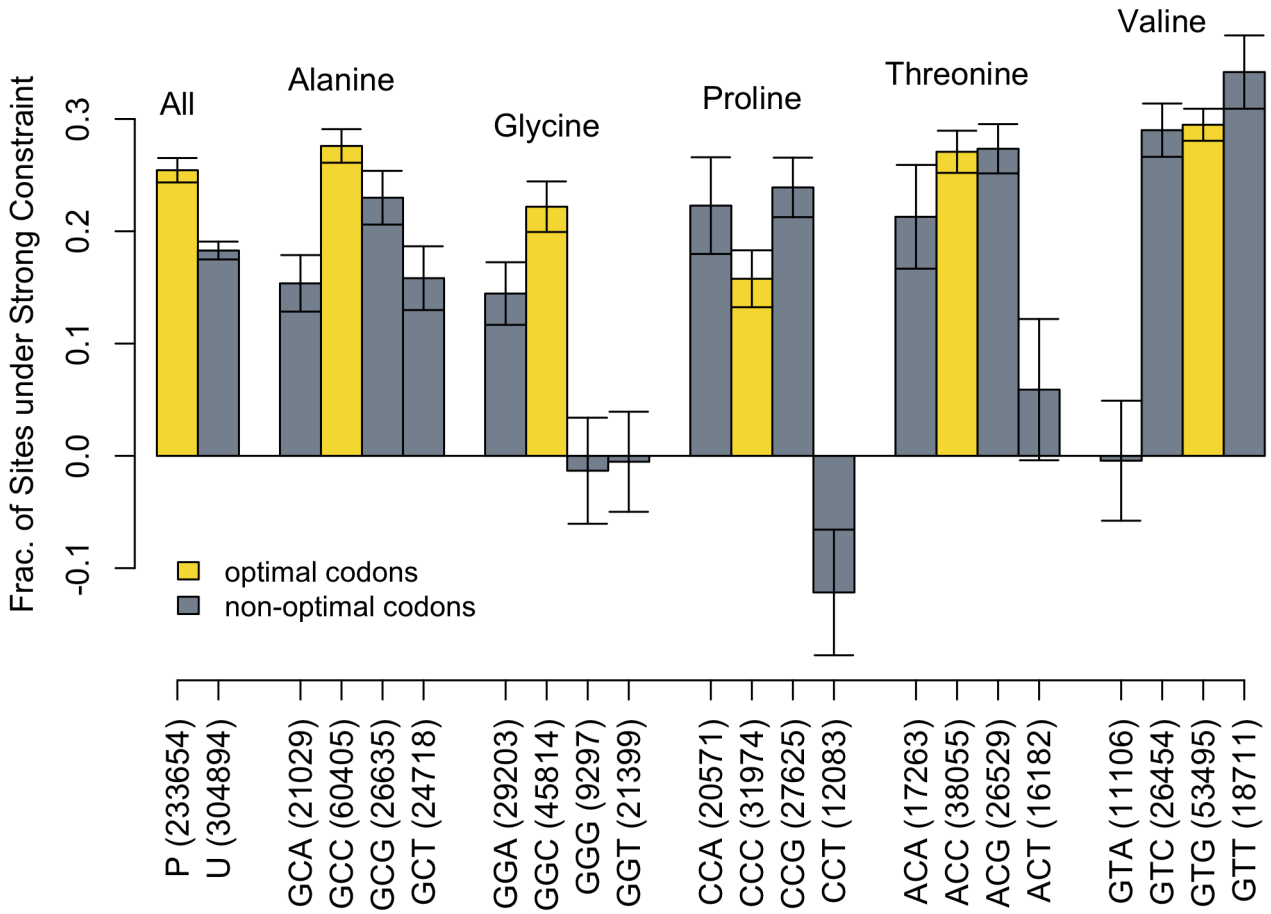

Figure 3. Constraint across codons. For each amino acid, we list the codons and, in parentheses, the number of 4D sites from each codon used in the bootstrap analysis – representing, in relative terms, the abundance of each codon in the genome. P-codons are all 4D sites from optimal codons grouped together, while U-codons are all 4D sites from non-optimal codons. 4D sites were binned into codons either by their ancestral allele as determined by parsimony to *D. sechellia* or by major allele if there is a substitution at that site between *D. sechellia* and *D. melanogaster*. Gold bars are the optimal codons for each amino acid, while dark grey bars are the non-optimal codons. 10 bootstraps determine the fraction of sites under constraint for each codon-type. Error bars represent the s.e. of the estimates. A negative value indicates an excess of polymorphism at 4D sites compared to short introns and is likely due to mispolarization assigning SNPs to the wrong codon.
doi:10.1371/journal.pgen.1003527.g003

mispolarization. It is more likely that codon CCT is similar to GGG, GGT, ACT, and GTA and has a neutral or nearly neutral level of polymorphism. Thus while the relationship between codons is worth noting, the exact numerical fraction of sites under constraint for each individual codon are all slightly biased beyond the nominal standard error. This bias is difficult to quantify but is not expected to be strong for most codon categories as *D. sechellia* and *D. melanogaster* are closely related species with few substitutions to throw off polarization. To eliminate any such biases from mispolarization and concurrently study the long-term signals of selection on 4D sites with respect to codon optimality and strong constraint, we refocused our analysis on only conserved codons.

Conserved codons: The dueling signals of codon bias and strong constraint. To more accurately quantify the relationship between selection for optimal codons and strong constraint, we restricted our bootstrap to include only those 4D sites from conserved amino acids in the 0 substitution class - i.e. those 4D sites conserved across the ten Drosophila species from *D. sechellia* to *D. grimshawi* (excluding *D. melanogaster* and *D. willistoni*). In such conserved codons, there are only a few substitutions along the *D. melanogaster* lineage at the 4D sites. In over 98% of these conserved 4D sites, both segregating and monomorphic, *D. melanogaster* shares an allele with the ten Drosophila species outgroup - the ancestral allele by parsimony. Such support as ten species sharing the same allele provides more confidence in the polarization of the SNPs at these sites. Also, roughly the same percent of monomorphic as polymorphic sites are removed as unpolarizable (less than 2% each) so that the act of polarization itself does not affect the density of 4D SNPs. This restriction allows for confident polarization by parsimony without changing the relative density of SNPs between 4D sites and short introns. As there are too few such conserved 4D sites to analyze each codon individually as above, we only consider the broad classes of preferred and unpreferred 4D sites. Note, however, these fully conserved codons are precisely where the action of selective forces has been most efficacious over evolutionary history.

In contrast to the results from all codons, when limiting the analysis to conserved codons (Figure 4), a higher fraction of unpreferred than preferred 4D sites are under strong constraint −53% (U) to 38% (P). However, 4D sites in the optimal state in *D. melanogaster* have been conserved across the ten Drosophila species to a greater extent, almost three times as often, than their non-optimal counterparts (Figure 4). This is expected because weak selection for codon bias in other Drosophila species on the

Drosophila tree is expected to generate conservation at optimal codons over and above the strong constraint we identify in this paper. Therefore, although fewer non-optimal codons are conserved in total, more of the conserved non-optimal codons have been so conserved because of the strong constraint.

Strong constraint across genes ranked by codon bias. To compare the gene targets of selection for codon bias and the strong constraint, we ranked genes by their Effective Number of Codons (ENC) [11] and Frequency of Optimal Codons (FOP) [20], each obtained from the database SEBIDA [74]. While neither metric accounts for local GC content, we used them to broadly classify genes by the extent of their codon bias (high, medium, low). We then performed 10 bootstrap runs on all the 4D sites within each gene-class. From Table 2, we can see that highly biased genes (having a high FOP and low-med ENC) have a slightly lower fraction of sites under strong constraint than genes with lower codon bias. Thus strong constraint acting on synonymous sites in *D. melanogaster* operates largely independently from canonical codon bias.

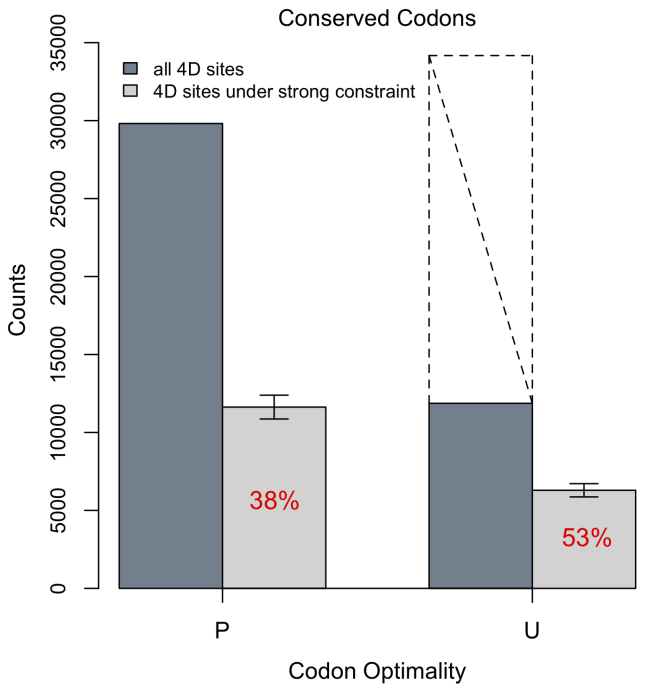

Figure 4. Codon optimality versus constraint in conserved codons. Codons are conserved from *D. sechellia-D. grimshawi* (excluding *D. willistoni*). The conserved codons were separated into those that were ancestrally preferred (P) and those that were ancestrally unpreferred (U) using polarization with the *D. sechellia-D. grimshawi* (excluding *D. willistoni*) outgroup. 10 bootstraps were done within each class. Error bars represent the s.e. of the estimates. The dark bars represent the counts of all sites that fall into each class while the light bars represent the number of sites estimated to be under strong constraint via the bootstrap procedure. The dashed line indicates what the count of total unpreferred conserved codons would have been had unpreferred 4D sites been conserved to the same extent as preferred 4D sites in otherwise conserved amino acids, i.e. the dashed line represents the proportion of U:P in all conserved amino acids. More than half (53%) of those unpreferred codons that are conserved across the ten Drosophila species are under strong purifying selection in *D. melanogaster*; 38% of preferred conserved codons are under strong selection.
doi:10.1371/journal.pgen.1003527.g004

Strong constraint as a function of different genic features

Table 3 summarizes our analyses of how the extent of strong constraint is influenced by different genic features such as gene length, the location of the synonymous site along the gene, the chromosome on which the gene is located, whether or not the synonymous site falls within splice junctions, and nucleosome binding. Many of the associations below, while suggestive, are marginal in effect. The dominant pattern is that strong constraint at synonymous sites appears to be ubiquitous across different gene classes and functional elements within genes.

Spatial distribution of strong constraint within genes. Looking at the distribution of constrained sites within genes, we focused on those sites that are within 75 bp from the start or stop codon and compared them to the 4D sites that lie in the middle of the gene. For this comparison, we took only those genes with a CDS longer than 150 bp. ~31% of 4D sites near the translation start and stop are under strong constraint. This is nearly a 50% increase in the fraction of sites under strong constraint as compared to the middle of the gene where only ~21% of 4D sites are under strong constraint on average. Breaking the spatial distribution of 4D sites across the middle of the gene into finer segments, we find no other peaks of strong constraint beyond those at the 5′ and 3′ edges of the genes (see Figure S1).

Bulk nucleosomes. Bulk nucleosomes wind themselves over ~146 bp of DNA, attaching at semi-regular intervals. Nucleosome-bound regions are associated with both the presence of purifying selection and lower rates of mutation [75–79]. Canvassing all 4D sites in the 146 bp regions around known bulk nucleosomal binding sites [80], we find a small increase in the fraction of missing polymorphism in these 4D sites bound by nucleosomes (Table 3). However, we have reason to believe that this slight increase above 22% is due to weak selection acting on nucleosome-bound sites and is likely not related to the strong constraint we measure in this paper (see Text S3). This potential weak-selective force does not impact our other results as it affects both short introns and 4D sites and we only measure selective differences between short introns and 4D sites.

Splice junctions. To investigate whether strong constraint can be explained by the need to maintain splice junctions, we tested 4D sites near intron-exon splice junctions - i.e. within 48 bp of a splice site. Around 26.0% of such 4D sites are under strong constraint (Table 3). This might indicate a role for splicing enhancers in the strong constraint, but Table 3 also shows that multi-exon genes and single-exon genes have similar amounts of strong constraint. The inference on the single-exon genes is particularly noisy, especially so given that our bootstrap method controls for distance to short introns. However, only about one-fifth of our 4D sites fall near splice sites and the modest enrichment of constraint near splice sites is not enough to explain the ubiquitous constraint at 4D sites across the genome or especially in single-exon genes.

Gene length. Longer genes tend to have slightly more sites under strong constraint than shorter genes (Table 3). Interestingly this correlation is stronger when taking intron and UTR length into account than when considering the CDS sequence alone. This pattern is the opposite of what is seen for codon bias in Drosophila [15,81].

X-linked versus autosomal genes. In Table 3, we show that X-linked genes have a slightly lower fraction of sites under strong constraint than autosomal genes. This pattern is again the opposite of what is seen for codon bias [28,82]. As selection is more efficient on the X chromosome [83], the cause for this

Table 2. Strong constraint in genes grouped by codon bias.

FOP[a]	Fraction of Sites[c]	ENC[b]	Fraction of Sites[c]
high FOP	18.7% (+/−1.7%)	low ENC	21.8% (+/−1.5%)
medium FOP	23.1% (+/−1.0%)	medium ENC	21.8% (+/−0.8%)
low FOP	23.2% (+/−0.9%)	high ENC	23.0% (+/−1.0%)

[a]Genes are ranked in descending order by their Frequency of Optimal Codons (FOP) with the top, middle, and bottom third forming the high, medium, and low FOP classes respectively;
[b]genes are ranked in ascending order by their Effective Number of Codons (ENC) with the top, middle, and bottom third forming the low, medium, and high ENC classes respectively;
[c]mean fraction of 4D sites under strong constraint in each category over 10 bootstrap runs (+/− s.e.).
doi:10.1371/journal.pgen.1003527.t002

difference is not clear and might reflect some difference in the types of genes located on the X as opposed to the autosomes.

Strong constraint correlates with gene expression level over development

To map how strong constraint at synonymous sites varies with gene expression over development, we ranked genes by their expression levels at each developmental time point in the ModEncode data set [84]. We split the genes evenly into three categories of expression - highly, moderately, and lowly expressed - within each developmental stage and ran 10 bootstraps for the 4D sites of the genes within each expression category in each developmental time point. The results are shown in Figure 5.

The overall gene expression level across development correlates well with the fraction of sites under strong constraint with lowly expressed genes tending to have fewer sites under strong constraint and highly expressed genes tending to have more sites under strong constraint. This pattern is strongest for the genes expressed highly in mid-late embryos, pupae, and adult males. The association of strong constraint with these developmental stages is further enhanced when the "high" expression group has been split in half into "high" and "very high" expression level categories (see Figure S2). In contrast to this preference of strong selection for genes highly expressed in embryo, pupal, and adult stages, codon bias is highest for genes whose expression peaks in larval stages [85].

Strong constraint over gene ontology

The difference in density of polymorphism between 4D sites and short introns does not allow for precise measurements of constraint on the synonymous sites of single genes. To identify a set of genes that are under particularly strong constraint at synonymous sites, we ranked genes by the fraction of their conserved amino acids that are unpreferred and conserved from *D. sechellia to D. grimshawi*, in the 0-substitution class (see Materials and Methods). Our method left 4,877 genes capable of being ranked of which we took the top sixth (812 genes, see Dataset S1) as our gene set enriched for strong constraint.

To validate our method of selecting genes under strong constraint, we checked that our 812-gene set is indeed enriched for strong constraint at 4D sites. We performed a bootstrap analysis on the 4D sites of variable amino acids in the genes in and out of this top set. Estimating constraint using 4D sites from variable amino acids provides a measure of the fraction of synonymous sites under constraint independent from our surrogate using conserved amino acids. In the top 812 genes, we find a ~30% reduction in polymorphism at 4D sites in variable amino acids; in all 4,065 genes not in the top 812 set, we find an average of ~21% of 4D sites in variable amino acids under strong constraint. As such, our top 812 genes are enriched for almost 50% more 4D sites under strong constraint than the average gene. Note that any individual gene in the 812-set does not necessarily have elevated levels of strong constraint at its synonymous sites, nor does any individual gene of the 4,065 necessarily have a lower fraction of 4D sites under strong constraint.

In order to examine whether genes under strong constraint at synonymous sites tend to be enriched for certain functions, we

Table 3. Strong constraint over different genic features.

Category	Fraction of Sites[a]	Category	Fraction of Sites[a]
5′ 75 bp of CDS[b]	30.7% (+/−3.0%)	3′ 75 bp of CDS[c]	31.5% (+/−2.5%)
Bulk Nucleosomes[d]	24.2% (+/−0.7%)	splice junctions[e]	26.0% (+/−1.0%)
multi-exon genes[f]	22.0% (+/−0.6%)	single-exon genes[g]	21.8% (+/−4.4%)
long genes[h]	25.8% (+/−0.9%)	long CDSs[i]	24.1% (+/−0.8%)
medium genes[h]	19.3% (+/−0.6%)	medium CDSs[i]	19.2% (+/−1.0%)
short genes[h]	17.3% (+/−1.3%)	short CDSs[i]	20.3% (+/−1.8%)
autosomal genes[j]	22.6% (+/−0.5%)	X-linked genes[k]	19.2% (+/−1.3%)

[a]Mean fraction of 4D sites under strong constraint in each category over 10 bootstrap runs (+/− s.e.);
[b]4D sites within 75 bp of the translation start site (longest transcript);
[c]4D sites within 75 bp of stop codon (longest transcript);
[d]4D sites in bulk nucleosome footprints;
[e]4D sites within 48 bp of a splice junction;
[f]4D sites from multi-exon genes;
[g]4D sites from single-exon genes;
[h]genes are ranked in descending order by their gene length (UTR + all exons + all introns) with the top, middle, and bottom third forming the long, medium, and short gene classes respectively;
[i]genes are ranked in descending order by their CDS length (longest transcript) with the top, middle, and bottom third forming the long, medium, and short CDS classes respectively;
[j]4D sites from Autosomal genes;
[k]4D sites from X-linked genes.
doi:10.1371/journal.pgen.1003527.t003

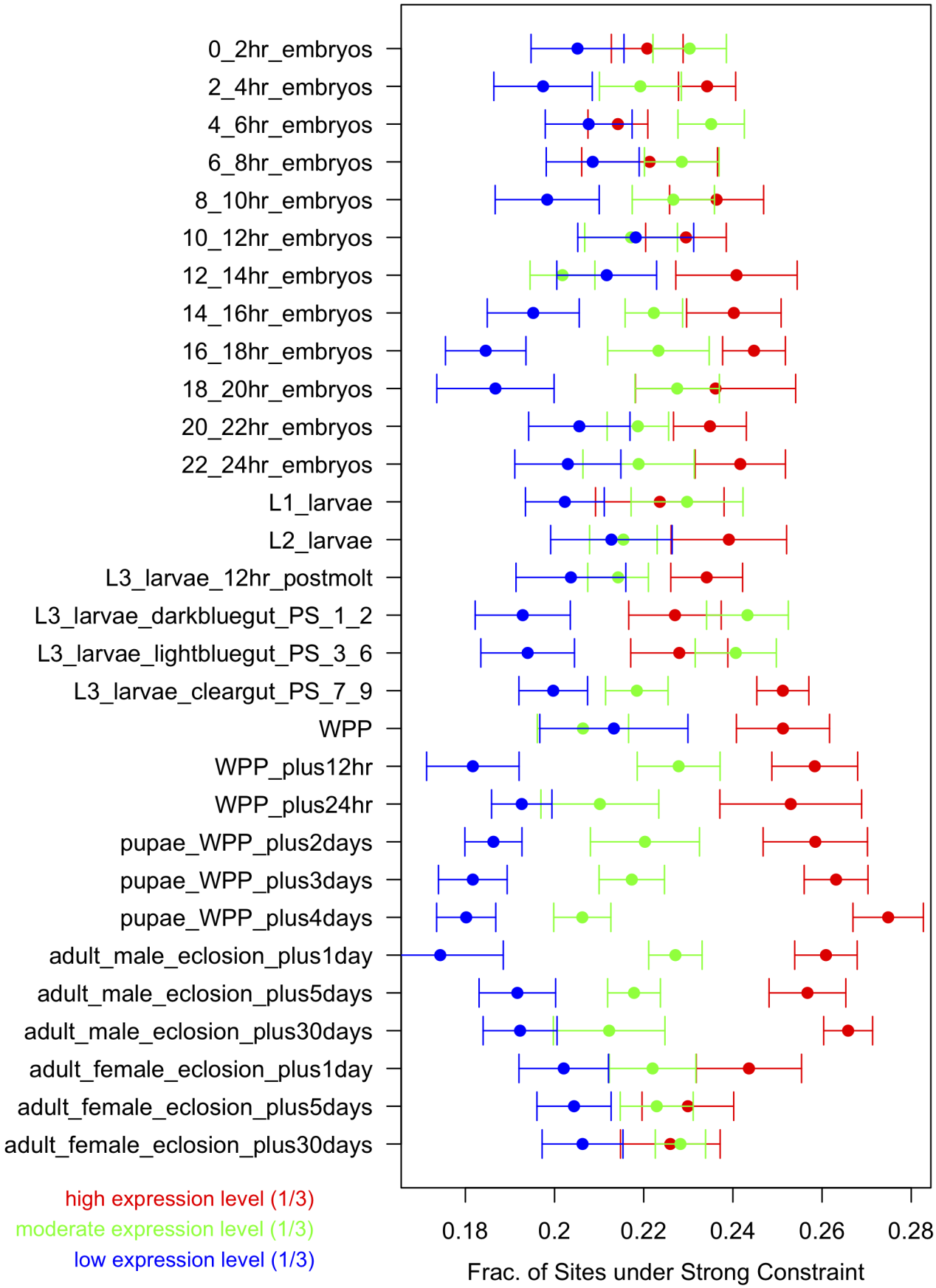
0_2hr_embryos
2_4hr_embryos
4_6hr_embryos
6_8hr_embryos
8_10hr_embryos
10_12hr_embryos
12_14hr_embryos
14_16hr_embryos
16_18hr_embryos
18_20hr_embryos
20_22hr_embryos
22_24hr_embryos
L1_larvae
L2_larvae
L3_larvae_12hr_postmolt
L3_larvae_darkbluegut_PS_1_2
L3_larvae_lightbluegut_PS_3_6
L3_larvae_cleargut_PS_7_9
WPP
WPP_plus12hr
WPP_plus24hr
pupae_WPP_plus2days
pupae_WPP_plus3days
pupae_WPP_plus4days
adult_male_eclosion_plus1day
adult_male_eclosion_plus5days
adult_male_eclosion_plus30days
adult_female_eclosion_plus1day
adult_female_eclosion_plus5days
adult_female_eclosion_plus30days
high expression level (1/3)
moderate expression level (1/3)
low expression level (1/3)
0.18
0.2
0.22
0.24
0.26
0.28
Frac. of Sites under Strong Constraint

Figure 5. Strong constraint versus gene expression across development. Within each developmental time point, genes were ranked by their level of expression and then grouped into high, moderate, and low expression levels - each group comprising of one-third of all genes. Within each gene set within each time point, the fraction of 4D synonymous sites under strong constraint was calculated using the bootstrap. 10 bootstraps were done within each such class. Error bars represent the s.e. of the estimates.
doi:10.1371/journal.pgen.1003527.g005

used DAVID 6.7 [86,87]. DAVID takes all the genes in the background data set (4,877 genes) and all genes in the test data set (812 genes) and looks for the enrichment of biological terms and gene families in the test set relative to the background. In Table 4, we list a subset of those biological terms found by DAVID's functional annotation clustering run on high stringency (for full information on the top 13 clusters, see Table S3). We find that in genes enriched for strong constraint, we co-enrich for many important functional gene sets. In particular, we co-enrich for genes critical in pupae-to-adult morphogenesis and in late embryogenesis. This finding is consistent with the result that genes expressed highly in late embryos, pupae, and adults have elevated levels of strong constraint at 4D sites. Many other functional classes important to the basic development and functioning of *D. melanogaster* appear to have a higher fraction of synonymous sites under strong constraint including: transcription factors, ribosomal genes, immunoglobulin genes, genes regulating gamete production – particularly oogenesis, cell-signaling genes – particularly synaptic transmission, and more.

Discussion

The strong constraint at synonymous sites in *D. melanogaster* measured in this paper represents a powerful force. We estimate that ~22% of synonymous sites are experiencing, on average, a selective pressure between $4N_e s \sim -250$–-500 against deleterious mutations. This strength of selection is as strong or stronger as has been measured via population genetic techniques at any class of sites, including non-synonymous ones [46,47]. Mutations at strongly constrained synonymous sites should never rise above low frequency in the population and certainly will never fix, barring tight linkage to a very advantageous allele or a shift in the functional properties of the site. While detectable within a population, these mutations are effectively lethal over evolutionary time.

We tested a number of controls to rule out the possibility that our observation of strong purifying selection results from other forces with possibly similar signals: A lower mutation rate, for example, can cause a signal indistinguishable from strong selection in polymorphism if the sample depth of the population is too shallow. To account for this, and at the same time account for any variation in the amount of linked selection between 4D sites and short intron sites, we used a bootstrap to control for GC content and distance between the 4D and short intron sites. We also performed bootstraps controlling for dinucleotide content between 4D and short intron sites and performed bootstraps pairing slow-evolving 4D sites against fast-evolving 4D sites as the neutral reference. Neither revealed a mutational force underlying the ~22% drop in 4D polymorphism compared to short introns. As revealed by simulations, the finite estimate we obtained of the strength of strong selection is itself evidence against a mutational force being responsible for our signal, as mutational variation would behave like infinite selection on 4D sites. While we do not have the frequency depth from the population necessary to estimate a full distribution of selection coefficients for the strong constraint force, our point estimate of $4N_e s \sim -283$ for these 22% of sites is statistically significantly different from the value of $4N_e s \sim -700$ (the computational limit of our program) expected if the signal was due to variation in mutation rate.

We also controlled for deviations from mutation-selection equilibrium affecting both the 4D and short intron site frequency spectra using a frequency-dependent correction. Such deviations include demography, shared (linked) selection between 4D sites and short introns, and our own approximations to the SFS. Controlling for these deviations resulted in a higher estimate of the strength of selection ($4N_e s \sim -370$) with larger error bars, but still significantly far from the boundary of $4N_e s \sim -700$.

A constant influx of weakly advantageous alleles in coding sequences, as is expected to occur in *D. melanogaster* [60], could affect variation at nearby 4D sites more than at short introns. The resulting genetic draft generated by adaptive substitutions in coding sequences would weaken the apparent intensity of purifying selection on 4D sites by bringing strongly deleterious alleles to higher frequencies, making our above estimates of selection intensity conservative [88]. Even so, strong selection, rather than a mutational difference, would still underlie our signal, as genetic draft cannot alter the frequency of synonymous mutations that are simply absent from the population. On the other hand, sweeps of weakly advantageous alleles in coding sequences could eliminate polymorphism in 4D sites more so than in short introns. Narrow selective sweeps in coding sequences reducing variation at

Table 4. Functional clusters in genes enriched for strong constraint.

Cluster #[a]	Overall Functional Annotation[b]	Enrichment[c]
1	transcriptional regulation	9.69
2	imaginal disc development	9.28
3	homeobox protein domain	7.57
4	eye morphogenesis	7.49
6	epithelium development	6.07
8	immunoglobulin domain	5.93
9	ribosomal proteins	5.36
10	cell signaling	4.59
12	gamete generation	4.33
13	neuron development	3.51

[a]Functional annotation clusters ranked by significance by DAVID [86,87]. These clusters are groups of similar or related biological annotation terms, with similarity determined by a simple stringency setting - in the above, a high stringency setting was used. The significance of the overall cluster reflects the combined enrichment in the test gene set of the individual annotation terms within a cluster (see c). Clusters 5, 7, and 11 are not reported here as their biological terms were similar to clusters 4, 1, and 4 & 13 respectively, so provided no new information. The full information for the top 13 clusters is reported in Table S3;
[b]Summary description of the type of annotation terms within each cluster. The specific annotation terms for each cluster are in the supplement;
[c]The enrichment score of the overall cluster as calculated by DAVID in the test gene set with respect to the background gene set. According to the description of enrichment scores by DAVID, each individual annotation term within a cluster has a p-value, or significance, for the enrichment of that term in the test gene set. The enrichment score of the overall cluster is then the geometric mean of these p-values. Thus the higher the enrichment score, the lower the p-values are for all terms in the overall annotation cluster and the more significantly enriched the overall cluster is in the test gene set. The p-values for the enrichment of the annotation terms in each cluster are reported in Table S3.
doi:10.1371/journal.pgen.1003527.t004

otherwise neutral 4D sites is, however, an unlikely explanation for our observations. When comparing 4D sites from different substitution rate classes against each other, we found a signal of strong constraint at conserved 4D sites relative to fast-evolving 4D sites. As sweeps should not affect the overall substitution rate of linked sites, strong purifying selection on synonymous sites is the best explanation for the lack of polymorphism at 4D sites relative to short introns.

Our ability to detect strong selection and differentiate it from other forces critically depends on the availability of deep and genome-wide population data. Previous data sets could only find weak or no constraint, thus always confirming our collective biological intuition that synonymous sites had little functional or evolutionary importance. In a shallower sample of even genome-wide data, the highly deleterious variants would be simply missing from the sample and there would be no power to distinguish strong selection from a variation in the rate of mutation. As an example, we simulated 4D sites evolving under the selective regime inferred from the real data (22% of sites at $4N_es = -283$) but with only 60 instead of 130 homozygous strains. Attempting to re-estimate the strength of selection from such a shallow sample results in the observation of seemingly infinite selection operating on 22% of 4D sites. Simulating 60 strains under a scenario where neutral 4D sites have a 22% lower mutation rate than do short introns results in the same inference of infinite selection. Genome-wide, deep population data sets were not available before recently and thus strong constraint could never before be unambiguously detected at synonymous sites.

Interestingly, the strong constraint in *D. melanogaster* appears to be a largely orthogonal force to canonical codon usage bias, favoring an overlapping, but different set of codons with subtly different gene targets. Codon bias increases as the conservation of amino acids increases, while the strong constraint targets the 4D sites of both conserved and variable amino acids equally. We further identified the codons under strong constraint and, for any given amino acid, the codon(s) with the highest fraction of sites under constraint were not necessarily the optimal codon. Other studies have likewise noted signals of selection favoring non-optimal codons in Drosophila [25,30,33,89,90]. Overall, preferred 4D sites do have greater amounts of strong constraint acting on them, but the strong selective force targets a substantial fraction of the unpreferred 4D sites as well. There is also a weak anti-correlation between genes with a high fraction of constraint and genes with high codon bias, which extends to various gene features. Long genes are associated with higher levels of strong constraint at 4D sites as opposed to shorter genes, in opposition to codon bias in Drosophila [15,81]. X-linked genes have a lower fraction of 4D sites under constraint than autosomal genes, wheras codon usage bias is stronger on the X [28,82]. While both codon bias and the fraction of 4D sites under strong constraint are correlated with highly expressed genes, codon usage bias is strongest in genes with their highest expression in larval stages [85] as opposed to the strong constraint seen most often in genes expressed highly in mid-late embryo, pupal, and male adult stages.

The pattern of conservation over 4D sites supports the existence of weak selection in Drosophila favoring the conservation of preferred 4D sites across the twelve species, but it appears to have been relaxed in *D. melanogaster*. In our SFS analysis, we were not only able to gauge the intensity of strong selection, but also show a lack of contribution from weak purifying selection to our signal. If any weak selection is still acting differentially on synonymous sites relative to short introns, then it is not powerful enough to be detected by our SFS model or contribute much to our signal of lost polymorphism. These results recapitulate some earlier results on *D. melanogaster* [24], although see [25–28]. While weak selection on 4D sites in *D. melanogaster* may not have vanished completely, the large influx of mutations and substitutions away from optimal codons corroborates some relaxation of constraint for codon bias in *D. melanogaster* [25,30,31,33,38]. Overall, weak selection for codon bias would seem to be less of a force on synonymous sites in *D. melanogaster* than in its sister species where weak selection for codon bias can be detected with far less ambiguity [24,30,31,33,34,40]. Thus, evidence suggests that there are at least two major, orthogonal forces affecting the evolution of 4D sites in Drosophila: the weak selective force driving codon bias that favors optimal codons, present in other Drosophila species, but relaxed in *D. melanogaster*; and an extant strong selective force targeting both optimal and non-optimal codons in *D. melanogaster* and across the Drosophila phylogeny. The function engendering the strong constraint appears to be independent of the translation optimization for efficiency and accuracy governing canonical codon usage bias.

The presence of splicing enhancers and nucleosomes do not explain the pattern of strong purifying selection either. However, the function underlying the strong constraint of synonymous sites may yet prove to be acting at the level of gene regulation. Those genes where strong selection on synonymous sites acts most frequently are often highly expressed regulatory proteins, operating in essential, tightly controlled developmental pathways. These are genes where the regulation of gene expression will matter most. Regulation of gene expression may be acting at the level of mRNA structures, mRNA stability, miRNA binding sites, and the modulation of translation rate [91–104]. Choice of synonymous codons might affect all of these levels of gene regulation. It should be noted that these various hypotheses are not mutually exclusive and may be intertwined. mRNA structures - as well as their avoidance, especially near the start of ORFs - may be involved in translation initiation/elongation, modulation of mRNA half-life, and accessibility of the mRNA to proteins and miRNAs [98,99,103–105]. Indeed, signatures of selection have been associated both with mRNA accessibility and mRNA structures and overall folding energy [97,99,105,106]. Our initial analysis found no enrichment of conserved unpreferred codons, a first-pass marker of the action of the strong constraint described in this paper, in either structured or unstructured mRNA as determined by ds/ssRNA sequencing [107] (not shown). This analysis, however, is at best preliminary and a strong possibility remains that the function underlying the strong constraint at synonymous sites is related to mRNA structure. miRNAs also have a host of different functional effects in different species and different genes within a species but are well known in their role of mRNA degradation [108,109]. The dynamics of translation not only affect the overall rate at which proteins are created, but also affect how these proteins fold and even the mRNA half-life [91–94,110–112]. The possibility that strong selection acts at the level of modulating translation rate through the presence of slow/fast sites is interesting as the translation speed of a codon is not necessarily related to codon optimality and tight control has been inferred at the beginning and end of ORFs in some species [96,100–102,113,114]. Given the pattern of the strong constraint across the different codons both optimal and non-optimal, the strong selective force may be due to the abundance of wobble vs. Watson-Crick tRNAs available for that codon. Ascertaining the functional mechanism underlying the observed strong constraint acting on synonymous sites could reveal deep insights into the regulation of gene expression.

Regardless of the specific functional mechanism underlying the strong constraint, experimental evidence from a wide range of

species substantiates an important functional role for synonymous sites. Directed mutagenesis studies targeting synonymous sites as well as studies of natural polymorphism have found consequential changes in protein levels and functionality due to natural synonymous variation and induced mutations [111–113,115–127]. In an experiment done on the Alcohol dehydrogenase (Adh) gene in *D. melanogaster*, changing 10 wild-type preferred Leucine alleles to unpreferred alleles in the 5′ region of the gene lowers the enzymatic activity of collected Adh by 25% [119]. The authors proposed that disruption of the sites' translational efficiency and accuracy caused the drop in activity, but also noted that the functional effect was far larger than expected given the assumption of only weak selection on synonymous sites [119]. 'Humanized' versions of protein coding sequences, with codons replaced with synonymous, putatively optimal codons in humans, show much greater protein expression and function when transfected into mammalian cells than the originals or synthetic versions using a non-mammalian species' set of optimal codons [115–118]. Human gene *Multidrug Resistance* 1 (*MDR*1) contributes to the drug resistance of cancer cells [122]. Both naturally occurring alleles as well as induced novel mutations at synonymous sites in *MDR*1 affect the resulting protein's conformation, altering its substrate specificity in human cell lines [122]. In the *E. coli* gene *ompA*, exchanging eight frequently-used codons for synonymous infrequently-used codons near the gene start results in a 3-fold reduction in mRNA levels and a 10-fold reduction in synthesis of protein OmpA [112]. Meanwhile exchanging codons with low-abundance tRNAs to synonymous codons with high-abundance tRNAs in *E. coli* gene *suf*I - or increasing the abundance of those tRNAs - results in misfolding of the protein *in vitro* and *in vivo* [110].

What about the presence of strong constraint in the synonymous sites of other species? In addition to the above functional assays, there are reported to be a significant fraction of synonymous sites under an unknown intensity of constraint in many species [22,29,35–37,39,97,128–130] and there is evidence for strong selection in humans [47]. For example, when compared to "neutral" controls, there is a reduction in polymorphism density and/or a lower rate of divergence at synonymous sites for many tetrapods including chicken, hominids, murids, and mammals in general [22,29,35–37,128]. Further, some of these species have undetectable or weak levels of codon bias, presumably commensurate with their small effective population sizes and thus the weakness of selection in favor of optimal codons [36,131]. Using a similar model to the one described in this paper, Keightley and Halligan (2011) found evidence to support that weak selection alone is unable to explain the pattern of diversity at 4D synonymous sites in humans [47]. While that study lacked the sample depth of polymorphism to be able to gauge the intensity of the strong selection, they estimated that 11% of 4D sites are evolving under a strong selection regime of $|4N_e s|>40$ [47]. Our results from Drosophila with a deeper population sample lend credence to the hypothesis that, in humans too, a force of strong constraint is responsible for the lack of polymorphism at 4D sites rather than a mutational force or other confounding factors. For many species, there has been no conclusion that the constraint on their respective synonymous sites is strong, but many of the signals are consistent with what we find in Drosophila with the fraction of sites under constraint, the amount of missing polymorphism, and the lack of relationship to codon bias. Thus with genome-wide, deep population SNP data becoming available for many of these other species, we may well find strong selection on synonymous sites to be ubiquitous.

As synonymous sites have often been used as the neutral reference in tests for purifying and adaptive selection, many estimates of the fraction of sites under constraint in other classes, such as non-synonymous sites, UTRs, and many others, are likely to be conservative. This result from population genetics supports findings that synonymous sites may harbor many, important causal variants and that studies ignoring the potential contribution of synonymous mutations may be likewise unnecessarily conservative [91]. Turnover at these strongly constrained synonymous sites could also represent a significant source of interspecies functional divergence and adaptation. The potential of synonymous sites to be sources of adaptation and genetic disease merits further investigation. Although the functionality underlying this strong constraint remains unknown, recent studies have uncovered a myriad of different types of functional information encoded into the CDS of genes beyond the protein recipe, including controls for translational efficiency and accuracy, splicing enhancers, microRNA binding, nucleosome positioning, and more. With the discovery of a significant fraction of sites under strong constraint in Drosophila, two things become clear: the role of synonymous sites in the biology of genomes is far greater than the neutral, "silent" part they were once assumed to play; and we still have much to learn about the functionality encoded in genes.

Materials and Methods

Data

The SNP data set from DGRP (http://dgrp.gnets.ncsu.edu/data/) consists of 168 inbred lines from a population of North Carolina *D. melanogaster* [48]. The SNPs were annotated as synonymous, non-synonymous, and intronic using Flybase release 5.33 (ftp://ftp.flybase.net/genomes/Drosophila_melanogaster/dmel_r5.33_FB2011_01/) [132]. If a position was found in multiple gene annotations, only those sites where the SNP was synonymous in all sites was called synonymous. Short intron sites are defined as those sites falling in introns of less than length 86 bp, 16 bp away from the intron start and 6 bp away from the intron end in order to eliminate any functional sequences at the edges of the introns [52]. Eliminating 16 bp from each side did not change SNP density (not shown). Any remaining purifying selection, especially strong purifying selection, in short introns makes our results more conservative. Four-fold (4D) sites are the collection of 3rd codon positions for the following amino acids: Proline, Alanine, Threonine, Glycine, and Valine.

All sites were resampled to a depth of 130 strains. All sites with sequence information for fewer than 130 strains were excluded. For SNPs at sites with more than 130 strains or which contained heterozygous lines at that position, a 130 allele subset was chosen randomly. If the SNP was no longer polymorphic after this random resampling, that position was moved into the non-polymorphic site class. We also removed any position with more than 2 alleles present.

We restricted our analysis to genes with 1–1 orthologs across the 12 Drosophila species tree [53] and where the longest transcript annotation had remained intact in release 5.33 - even if it is no longer the longest transcript in release 5.33. We used the remaining 5,709 coding sequences aligned with PRANK from Markova-Raina and Petrov (2011) [68,69].

SFS: Maximum likelihood and simulation

To determine the distribution of selective effects on a group of sites based on the shape and the amplitude of the SFS, we assume a two-state framework where sites are either monomorphic in the wild-type state or polymorphic with a neutral or deleterious mutation at some observed frequency in the population. Using short introns as a neutral reference, our model aims to capture the

fraction of synonymous sites falling into three broad selection categories – those with neutral, weakly deleterious, or strongly deleterious mutations – and estimate the effective selection coefficients acting on those mutations.

Strong constraint can be difficult to capture as strong selection has a greater effect on the amplitude of the SFS, the total number of observed mutations, than on its shape, the frequency distribution of observed mutations. Using a similar expansion to the standard SFS to Keightly and Eyre-Walker (2007) [46], we add the zero-frequency class, the fraction of monomorphic sites, to the SFS. The SNP density provides the additional information necessary to infer the action of strong constraint.

Equal to $4N_e\mu$, θ is mutation rate scaled by the effective population size and determines the neutral SNP density. The short intron SFS, used as neutral reference, anchors our estimate of θ which in turn allows us to estimate the amount of missing synonymous polymorphism in each selection category, c. As purifying selection increases, the overall density of observed polymorphism is reduced in the fraction of 4D sites in that selection class and the expected distribution of mutation is further skewed towards rare frequencies in the population. Each category has a single selection parameter, γ_c, a point estimate of the effective strength of selection, $4N_e s$, operating on the 4D sites in that class. For those 4D sites in the neutral category, $\gamma_c = 0$. For those in the weakly deleterious category, $0<|\gamma_c|<5$. For those in the strongly deleterious category, $|\gamma_c|>5$ or 100 – the choice of boundary did not affect results.

For our sample of n chromosomes from the population, assuming mutation-selection balance, we have the following analytical prediction for the SFS, $g(x)$ – the expected fraction of 4D sites with SNPs at frequency x in the sample [43]:

$$g(x,c) = \theta \cdot f_c \cdot L \frac{\left(1 - e^{-\gamma_c \cdot (1-x)}\right)}{(1 - e^{-\gamma_c}) x (1-x)} \tag{1}$$

$$\text{if } \gamma_c = 0 \text{ then } g(x,c) = \theta \cdot f_c \cdot L / x \tag{2}$$

$g(x,c)$ is the contribution of each selection category to the overall SFS. L is the total number of 4D sites while f_c is the fraction of 4D sites in each selection category c.

$$g(x) = \sum_c g(x,c) \tag{3}$$

$$\text{if } x = 0 \text{ then } g(0) = L - m \text{ where } m = \sum_{x=1/n}^{x<1} g(x) \tag{4}$$

$g(0)$ are the zero-frequency class, monomorphic, sites and are what gives the SFS "amplitude" information – the density, rather than just the shape, of the spectrum. While m is the total number observed SNPs in the sample.

The theoretical SFS for intronic sites is the same as above, only all sites are assumed to be neutral. However, any real SFS does not reflect the true frequency distribution of the SNPs in the population, but rather a binomial sampling of those SNPs and frequencies. The above is thus an approximation, as the probability of a site with a SNP at a given frequency in the sample from the population is not quite the same as the probability of a site with a SNP at a given frequency in the population as a whole. However, it is much more computational efficient for both speed and memory to use the approximation.

With this theoretical prediction of the distribution of sites over each frequency class in both the neutral reference (short intron SFS) and test set of sites (4D SFS), we can use maximum-likelihood to fit the parameters of our model to real data sets. Our model has 5 free parameters: θ, $(\gamma_{weak}, \gamma_{strong})$, and $(f_{neutral}, f_{weak}, f_{strong})$ where $f_{neutral} = 1 - f_{weak} - f_{strong}$. The total likelihood, λ, of the model's fit to the data, D, is equal to product of the fit the short intron and 4D sites spectra:

$$\lambda_{full}\left(D|\theta,\overline{\gamma},\overline{f}\right) = \lambda_{4D}\left(D|\theta,\overline{\gamma},\overline{f}\right) \times \lambda_{SI}(D|\theta) \tag{5}$$

λ_{4D} and λ_{SI} are the likelihood of the observed SFS given the expected SFS as determined by the free parameters and equations (1)–(4). These likelihoods are the multinomial probability of observing a certain number of sites, k, with SNPs in frequency class x in the sample given theoretical expectations. Taking short intron sites as an example (same for both):

$$\lambda_{SI}(D|\theta) = \prod_{x=0}^{x=1/2} (p(x|\theta))^{k_x} \text{ where}$$
$$p(x|\theta) = \begin{cases} g(0)/L & \text{if } x = 0 \\ g(1/2)/L & \text{if } x = 1/2 \\ (g(x) + g(1-x))/L & o.w. \end{cases} \tag{6}$$

Equation (6) is thus the probability that the folded theoretical SFS, $g(x)$, matches the empirical folded SFS, k_x. We folded the spectrum to avoid any problems with inferring the ancestral state.

We then maximized the parameters θ, $(f_{neutral}, f_{weak}, f_{strong})$, and $(\gamma_{weak}, \gamma_{strong})$ in Matlab using fminsearch, an implementation of the Nelder-Mead simplex method [133], on the negative log-likelihood of λ_{full}. The observed spectra were obtained from the bootstrapped 4D and short intron pairs. Where simulations were needed in this study, theoretical spectra were calculated using the above equations (1)–(4) and then the parameters were re-estimated by the outlined maximum-likelihood procedure on those theoretical spectra acting in place of the empirical data.

Frequency-dependent correction of SFS. We also employed a frequency-correction developed in Eyre-Walker et al (2006) [41] to control for demography or any weak and linked selection affecting both the short introns and 4D sites and to also correct for the approximation to the true SFS mentioned above. This allows the short intron SFS to not only act as neutral reference for the amplitude of the 4D SFS, but also its shape. With the correction, each frequency class now has a modifier, α_x, which adjusts the probability of seeing a site with a SNP at frequency of x in the sample. As the α's are shared between the short intron and 4D SFS, they control for confounding factors affecting both spectra. This frequency-correction modifies equations (5) and (6) like so:

$$\lambda_{full}\left(D|\theta,\overline{\gamma},\overline{f},\overline{\alpha}\right) = \lambda_{4D}\left(D|\theta,\overline{\gamma},\overline{f},\overline{\alpha}\right) \times \lambda_{SI}(D|\theta,\overline{\alpha}) \tag{7}$$

$$\lambda_{SI}(D|\theta,\overline{\alpha}) = \prod_{x=0}^{x=1/2} (\alpha_x p(x|\theta))^{k_x} \text{where } \alpha_0 = 1 \tag{8}$$

While this correction is robust for many confounding factors [41], it adds a free parameter for every frequency-class except the first one. The parameter for the zero-frequency class, α_0, is set to 1

to anchor the maximum-likelihood estimation of the α's. With 65 frequency classes, this adds 64 free parameters to the basic model of 5 free parameters.

Phylogenetic tree and conservation

We used the determined 15 species Insect tree topology from the UCSC genome browser (http://hgdownload.cse.ucsc.edu/goldenPath/dm3/phastCons15way/) and paired it down to the 12 Drosophila species [134]. We then input that tree topology into PhyML v3.0 (http://www.atgc-montpellier.fr/phyml) [70] and allowed it to re-estimate the branch lengths on all 4D sites in conserved amino acids using the HKY85 model [135] without a discrete gamma model and without invariant sites. The nucleotide frequencies and transition-transversion rate ratio were inferred by maximum-likelihood. The resulting tree can be found in Text S4.

GERPcol from GERP++ (http://mendel.stanford.edu/SidowLab/downloads/gerp/) [73] was run on the collection of all 4D sites from all 12 Drosophila species excluding *D. melanogaster* and *D. willistoni*, estimating the Rscore (tree length - inferred # of substitutions) for each site independently. We input into GERP the tree and transition-transversion ratio from the PhyML results. As these two programs use different parameterizations of the transition-transversion ratio, we translated one to the other (see Text S4).

GO category enrichment

Our signal from polymorphism does not afford us a precise measurement of constraint on the 4D sites of a single gene (not enough information). Therefore, we use a surrogate to infer the amount of strong constraint at the 4D sites of individual genes. Looking only at sites without SNPs, we use the percentage of 4D sites in conserved amino acids that are unpreferred and themselves conserved from *D. sechellia to D. grimshawi* (i.e. in the 0-substitution class) as our measure of how extensive the strong constraint has been on the 4D sites of the gene in question. As unpreferred 4D sites in the 0-substitution class have the highest fraction of sites under strong constraint (53%), the reasoning is that the more such sites exist in a gene, the more likely there has been extensive constraint acting on all 4D sites. Since not all genes have enough conserved amino acids to allow a reasonable calculation of the above surrogate, we used only those genes where at least 20% of the four-fold amino acids were conserved along the tree, leaving 4,877 genes in the analysis. We ranked genes by this surrogate and took the top 812 genes ($\sim$ top sixth of genes). We then used the functional annotation clustering tool from DAVID 6.7 (http://david.abcc.ncifcrf.gov/home.jsp) set on high stringency to look for enrichment of GO category terms in this gene set [86,87].

Supporting Information

Dataset S1 Genes enriched for high constraint. Table of the top 812 genes enriched for high constraint at 4D sites.
(DOC)

Figure S1 Spatial distribution of strong constraint within coding sequences. 4D sites were binned by their distance to the translation start site in the longest transcript for each gene. Each bin represents 5% of transcript length to control for different transcript lengths. 10 bootstraps to determine the fraction of sites under constraint were done within each bin. Error bars represent the s.e. of the estimates.
(TIF)

Figure S2 Strong constraint versus gene expression across development. Genes are grouped and analyzed as in Figure 5. Here, the Figure 5 "high expression level" gene set has been halved, creating a "very high expression level" and "high expression level" group, each containing one-sixth of all genes. Error bars represent the s.e. of the estimates.
(TIF)

Table S1 Conservation versus constraint at 4D sites in conserved amino acids. The proportion of constrained 4D sites in each substitution class using fast-evolving 4D sites as the "neutral" reference and the relationship of those results to the proportion of constrained 4D sites in each substitution class using short introns as the reference.
(DOC)

Table S2 Estimated proportion of slow-evolving 4D sites and $4N_e s$ for each selection class. Maximum likelihood results using fast-evolving 4D sites as the "neutral" reference.
(DOC)

Table S3 Gene ontology clusters. Table of the full information for the top 13 GO clusters as reported by DAVID 6.7 from the 812 genes most enriched for strong constraint at 4D sites [86,87].
(DOC)

Text S1 Demographic correction of SFS. Maximum likelihood results when correcting for demography and other non-mutation-selection balance forces in the SFS.
(DOC)

Text S2 Mutation versus strong constraint. Discusses extra controls for distinguishing the strong constraint from possible confounding mutational effects: *S2.A – Tri-nucleotide bootstrap results, S2.B – Power to detect low mutation versus strong purifying selection on 4D sites, S2.C – Slow- versus fast-evolving 4D site bootstrap results.*
(DOC)

Text S3 Weak selection on bulk nucleosomes. Evidence suggests a selective force operating in regions bound by bulk nucleosomes, but that the force appears to be different from the strong purifying selection inferred in the main result of this paper.
(DOC)

Text S4 Phylogenetic tree and parameters of 4D sites. The 12 Drosophila species tree and nucleotide substitution parameters inferred on 4D synonymous sites.
(DOC)

Acknowledgments

The authors would like to thank David Enard, Anna-Sophie Fiston-Lavier, Penka Markova-Raina, Sandeep Venkataram, and all members of the Petrov lab for helpful feedback and support of this project. We would also like to thank the editors and three anonymous reviewers for their helpful comments and suggestions.

Author Contributions

Analyzed the data: DSL PWM RH DAP. Wrote the paper: DSL PWM RH DAP. Conceived and designed methodologies: DSL PWM RH DAP. Designed and wrote software used in analysis: DSL.

References

1. Kimura M. (1968) Evolutionary rate at the molecular level. Nature 217(5129): 624.
2. King JL, Jukes TH. (1969) Non-darwinian evolution. Science 164(881): 788–798.
3. McDonald JH, Kreitman M. (1991) Adaptive protein evolution at the adh locus in drosophila. Nature 351(6328): 652–654.
4. Sawyer SA, Hartl DL. (1992) Population genetics of polymorphism and divergence. Genetics 132(4): 1161–1176.

5. Bustamante CD, Fledel-Alon A, Williamson S, Nielsen R, Hubisz MT, et al. (2005) Natural selection on protein-coding genes in the human genome. Nature 437(7062): 1153–1157.
6. Eyre-Walker A, Keightley PD. (1999) High genomic deleterious mutation rates in hominids. Nature 397(6717): 344–347.
7. Yang Z. (2007) PAML 4: Phylogenetic analysis by maximum likelihood. Mol Biol Evol 24(8): 1586–1591.
8. Ikemura T. (1981) Correlation between the abundance of escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: A proposal for a synonymous codon choice that is optimal for the E. coli translational system. J Mol Biol 151(3): 389–409.
9. Grantham R, Gautier C, Gouy M, Jacobzone M, Mercier R. (1981) Codon catalog usage is a genome strategy modulated for gene expressivity. Nucleic Acids Res 9(1): 213–213.
10. Gouy M, Gautier C. (1982) Codon usage in bacteria: Correlation with gene expressivity. Nucleic Acids Res 10(22): 7055–7074.
11. Sharp PM, Devine KM. (1989) Codon usage and gene expression level in dictyosteiium discoidtum: Highly expressed genes do [prefer [optimal codons. Nucleic Acids Res 17(13): 5029–5040.
12. Akashi H, Eyre-Walker A. (1998) Translational selection and molecular evolution. Curr Opin Genet Dev 8(6): 688–693.
13. Plotkin JB, Kudla G. (2010) Synonymous but not the same: The causes and consequences of codon bias. Nature Reviews Genetics 12(1): 32–42.
14. Hershberg R, Petrov DA. (2009) General rules for optimal codon choice. PLoS Genet 5: e1000556. doi:10.1371/journal.pgen.1000556
15. Akashi H. (2001) Gene expression and molecular evolution. Curr Opin Genet Dev 11(6): 660–666.
16. Drummond DA, Wilke CO. (2008) Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. Cell 134(2): 341–352.
17. Akashi H. (1994) Synonymous codon usage in drosophila melanogaster: Natural selection and translational accuracy. Genetics 136(3): 927–935.
18. Stoletzki N, Eyre-Walker A. (2007) Synonymous codon usage in escherichia coli: Selection for translational accuracy. Mol Biol Evol 24(2): 374–381.
19. Hershberg R, Petrov DA. (2008) Selection on codon bias. Annu Rev Genet 42: 287–299.
20. Ikemura T. (1985) Codon usage and tRNA content in unicellular and multicellular organisms. Mol Biol Evol 2(1): 13–34.
21. Ikemura T. (1982) Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes. differences in synonymous codon choice patterns of yeast and escherichia coli with reference to the abundance of isoaccepting transfer RNAs. J Mol Biol 158(4): 573–597.
22. Yang Z, Nielsen R. (2008) Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. Mol Biol Evol 25(3): 568–579.
23. Hartl DL, Moriyama EN, Sawyer SA. (1994) Selection intensity for codon bias. Genetics 138(1): 227–234.
24. Andolfatto P, Wong KM, Bachtrog D. (2011) Effective population size and the efficacy of selection on the X chromosomes of two closely related drosophila species. Genome Biology and Evolution 3: 114.
25. Clemente F, Vogl C. (2012) Evidence for complex selection on four-fold degenerate sites in drosophila melanogaster. J Evol Biol 25: 2582–2595.
26. Vogl C, Clemente F. (2012) The allele-frequency spectrum in a decoupled moran model with mutation, drift, and directional selection, assuming small mutation rates. Theor Popul Biol 81(3): 197–209.
27. Zeng K, Charlesworth B. (2009) Estimating selection intensity on synonymous codon usage in a nonequilibrium population. Genetics 183(2): 651–662.
28. Zeng K, Charlesworth B. (2010) Studying patterns of recent evolution at synonymous sites and intronic sites in drosophila melanogaster. J Mol Evol 70(1): 116–128.
29. Chamary J, Parmley JL, Hurst LD. (2006) Hearing silence: Non-neutral evolution at synonymous sites in mammals. Nature Reviews Genetics 7(2): 98–108.
30. Nielsen R, DuMont VLB, Hubisz MJ, Aquadro CF. (2007) Maximum likelihood estimation of ancestral codon usage bias parameters in drosophila. Mol Biol Evol 24(1): 228–235.
31. Akashi H. (1995) Inferring weak selection from patterns of polymorphism and divergence at "silent" sites in drosophila DNA. Genetics 139(2): 1067–1076.
32. Lawrie DS, Petrov DA, Messer PW. (2011) Faster than neutral evolution of constrained sequences: The complex interplay of mutational biases and weak selection. Genome Biology and Evolution 3: 383.
33. Singh ND, DuMont VLB, Hubisz MJ, Nielsen R, Aquadro CF. (2007) Patterns of mutation and selection at synonymous sites in drosophila. Mol Biol Evol 24(12): 2687–2697.
34. Akashi H. (1996) Molecular evolution between drosophila melanogaster and D. simulans: Reduced codon bias, faster rates of amino acid substitution, and larger proteins in D. melanogaster. Genetics 144(3): 1297–1307.
35. Eőry L, Halligan DL, Keightley PD. (2010) Distributions of selectively constrained sites and deleterious mutation rates in the hominid and murid genomes. Mol Biol Evol 27(1): 177–192.
36. Künstner A, Nabholz B, Ellegren H. (2011) Significant selective constraint at 4-fold degenerate sites in the avian genome and its consequence for detection of positive selection. Genome Biology and Evolution 3: 1381.
37. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. (2010) Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res 20(1): 110–121.
38. McVean GAT, Vieira J. (2001) Inferring parameters of mutation, selection and demography from patterns of synonymous site evolution in drosophila. Genetics 157(1): 245–257.
39. Zhou T, Gu W, Wilke CO. (2010) Detecting positive and purifying selection at synonymous sites in yeast and worm. Mol Biol Evol 27(8): 1912–1922.
40. Haddrill PR, Zeng K, Charlesworth B. (2011) Determinants of synonymous and nonsynonymous variability in three species of drosophila. Mol Biol Evol 28(5): 1731–1743.
41. Eyre-Walker A, Woolfit M, Phelps T. (2006) The distribution of fitness effects of new deleterious amino acid mutations in humans. Genetics 173(2): 891–900.
42. Boyko AR, Williamson SH, Indap AR, Degenhardt JD, Hernandez RD, et al. (2008) Assessing the evolutionary impact of amino acid mutations in the human genome. PLoS Genet 4: e1000083. doi:10.1371/journal.pgen.1000083
43. Wright S. (1938) The distribution of gene frequencies under irreversible mutation. Proc Natl Acad Sci U S A 24(7): 253.
44. Fisher RA. (1930) The distribution of gene ratios for rare mutations. Proceedings of the Royal Society of Edinburgh 50: 205–220.
45. Kimura M. (1962) On the probability of fixation of mutant genes in a population. Genetics 47(6): 713.
46. Keightley PD, Eyre-Walker A. (2007) Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies. Genetics 177(4): 2251–2261.
47. Keightley PD, Halligan DL. (2011) Inference of site frequency spectra from high-throughput sequence data: Quantification of selection on nonsynonymous and synonymous sites in humans. Genetics 188(4): 931–940.
48. Mackay TFC, Richards S, Stone EA, Barbadilla A, Ayroles JF, et al. (2012) The drosophila melanogaster genetic reference panel. Nature 482(7384): 173–178.
49. Clemente F, Vogl C. (2012) Unconstrained evolution in short introns?–An analysis of genome-wide polymorphism and divergence data from drosophila. J Evol Biol 25: 1975–1990.
50. Parsch J, Novozhilov S, Saminadin-Peter SS, Wong KM, Andolfatto P. (2010) On the utility of short intron sequences as a reference for the detection of positive and negative selection in drosophila. Mol Biol Evol 27(6): 1226–1234.
51. Singh ND, Arndt PF, Clark AG, Aquadro CF. (2009) Strong evidence for lineage and sequence specificity of substitution rates and patterns in drosophila. Mol Biol Evol 26(7): 1591–1605.
52. Haddrill PR, Charlesworth B, Halligan DL, Andolfatto P. (2005) Patterns of intron sequence evolution in drosophila are dependent upon length and GC content. Genome Biol 6(8): R67.
53. Clark AG, Eisen MB, Smith DR, Bergman CM, Oliver B, et al. (2007) Evolution of genes and genomes on the drosophila phylogeny. Nature 450(7167): 203–218.
54. Hershberg R, Petrov DA. (2010) Evidence that mutation is universally biased towards AT in bacteria. PLoS Genet 6: e1001115. doi:10.1371/journal.pgen.1001115
55. Lynch M, Sung W, Morris K, Coffey N, Landry CR, et al. (2008) A genome-wide view of the spectrum of spontaneous mutations in yeast. Proceedings of the National Academy of Sciences 105(27): 9272–9277.
56. Hildebrand F, Meyer A, Eyre-Walker A. (2010) Evidence of selection upon genomic GC-content in bacteria. PLoS Genet 6: e1001107. doi:10.1371/journal.pgen.1001107
57. Petrov DA, Hartl DL. (1999) Patterns of nucleotide substitution in drosophila and mammalian genomes. Proceedings of the National Academy of Sciences 96(4): 1475–1479.
58. Sella G, Petrov DA, Przeworski M, Andolfatto P. (2009) Pervasive natural selection in the drosophila genome? PLoS Genet 5: e1000495. doi:10.1371/journal.pgen.1000495
59. Haddrill PR, Halligan DL, Tomaras D, Charlesworth B. (2007) Reduced efficacy of selection in regions of the drosophila genome that lack crossing over. Genome Biol 8(2): R18.
60. Andolfatto P. (2007) Hitchhiking effects of recurrent beneficial amino acid substitutions in the drosophila melanogaster genome. Genome Res 17(12): 1755–1762.
61. Macpherson JM, Sella G, Davis JC, Petrov DA. (2007) Genomewide spatial correspondence between nonsynonymous divergence and neutral polymorphism reveals extensive adaptation in drosophila. Genetics 177(4): 2083–2099.
62. Smith JM, Haigh J. (1974) The hitch-hiking effect of a favourable gene. Genet Res 23(1): 23–35.
63. Charlesworth B, Morgan M, Charlesworth D. (1993) The effect of deleterious mutations on neutral molecular variation. Genetics 134(4): 1289–1303.
64. Sekelsky JJ, Brodsky MH, Burtis KC. (2000) DNA repair in drosophila insights from the drosophila genome sequence. J Cell Biol 150(2): F31–F36.
65. Hernandez RD, Williamson SH, Bustamante CD. (2007) Context dependence, ancestral misidentification, and spurious signatures of natural selection. Mol Biol Evol 24(8): 1792–1800.
66. Arndt PF, Burge CB, Hwa T. (2003) DNA sequence evolution with neighbor-dependent mutation. Journal of Computational Biology 10(3–4): 313–322.
67. Siepel A, Haussler D. (2004) Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. Mol Biol Evol 21(3): 468–488.

68. Löytynoja A, Goldman N. (2008) Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. Science 320(5883): 1632–1635.
69. Markova-Raina P, Petrov D. (2011) High sensitivity to aligner and high rate of false positives in the estimates of positive selection in the 12 drosophila genomes. Genome Res 21(6): 863–874.
70. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, et al. (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. Syst Biol 59(3): 307–321.
71. Vicario S, Moriyama EN, Powell JR. (2007) Codon usage in twelve species of drosophila. BMC Evolutionary Biology 7(1): 226.
72. Cooper GM, Stone EA, Asimenos G, Green ED, Batzoglou S, et al. (2005) Distribution and intensity of constraint in mammalian genomic sequence. Genome Res 15(7): 901–913.
73. Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, et al. (2010) Identifying a high fraction of the human genome to be under selective constraint using GERP. PLoS Comput Biol 6: e1001025. doi:10.1371/journal.pcbi.1001025
74. Gnad F, Parsch J. (2006) Sebida: A database for the functional and evolutionary analysis of genes with sex-biased expression. Bioinformatics 22(20): 2577–2579.
75. Chen X, Chen Z, Chen H, Su Z, Yang J, et al. (2012) Nucleosomes suppress spontaneous mutations base-specifically in eukaryotes. Science 335(6073): 1235–1238.
76. Dai Z, Dai X, Xiang Q. (2011) Genome-wide DNA sequence polymorphisms facilitate nucleosome positioning in yeast. Bioinformatics 27(13): 1758–1764.
77. Prendergast JGD, Semple CAM. (2011) Widespread signatures of recent selection linked to nucleosome positioning in the human lineage. Genome Res 21(11): 1777–1787.
78. Warnecke T, Batada NN, Hurst LD. (2008) The impact of the nucleosome code on protein-coding sequence evolution in yeast. PLoS Genet 4: e1000250. doi:10.1371/journal.pgen.1000250
79. Warnecke T, Supek F, Lehner B. (2012) Nucleoid-associated proteins affect mutation dynamics in E. coli in a growth phase-specific manner. PLoS Comput Biol 8: e1002846. doi:10.1371/journal.pcbi.1002846
80. Mavrich TN, Jiang C, Ioshikhes IP, Li X, Venters BJ, et al. (2008) Nucleosome organization in the drosophila genome. Nature 453(7193): 358–362.
81. Moriyama EN, Powell JR. (1998) Gene length and codon usage bias in drosophila melanogaster, saccharomyces cerevisiae and escherichia coli. Nucleic Acids Res 26(13): 3188–3193.
82. Singh ND, Davis JC, Petrov DA. (2005) X-linked genes evolve higher codon bias in drosophila and caenorhabditis. Genetics 171(1): 145–155.
83. Singh ND, Larracuente AM, Clark AG. (2008) Contrasting the efficacy of selection on the X and autosomes in drosophila. Mol Biol Evol 25(2): 454–467.
84. Graveley BR, Brooks AN, Carlson JW, Duff MO, Landolin JM, et al. (2010) The developmental transcriptome of drosophila melanogaster. Nature 471(7339): 473–479.
85. Vicario S, Mason CE, White KP, Powell JR. (2008) Developmental stage and level of codon usage bias in drosophila. Mol Biol Evol 25(11): 2269–2277.
86. Da Wei Huang BTS, Lempicki RA. (2008) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nature Protocols 4(1): 44–57.
87. Sherman BT, Lempicki RA. (2009) Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res 37(1): 1–13.
88. Messer PW, Petrov DA. (2012) The McDonald-kreitman test and its extensions under frequent adaptation: Problems and solutions. arXiv Preprint arXiv:1211.0060 [Q-Bio.PE].
89. DuMont VB, Fay JC, Calabrese PP, Aquadro CF. (2004) DNA variability and divergence at the notch locus in drosophila melanogaster and D. simulans: A case of accelerated synonymous site divergence. Genetics 167(1): 171–185.
90. DuMont VLB, Singh ND, Wright MH, Aquadro CF. (2009) Locus-specific decoupling of base composition evolution at synonymous sites and introns along the drosophila melanogaster and drosophila sechellia lineages. Genome Biology and Evolution 1: 67.
91. Sauna ZE, Kimchi-Sarfaty C. (2011) Understanding the contribution of synonymous mutations to human disease. Nature Reviews Genetics 12(10): 683–691.
92. Deana A, Belasco JG. (2005) Lost in translation: The influence of ribosomes on bacterial mRNA decay. Genes Dev 19(21): 2526–2533.
93. Purvis IJ, Bettany AJE, Santiago TC, Coggins JR, Duncan K, et al. (1987) The efficiency of folding of some proteins is increased by controlled rates of translation *in vivo*: A hypothesis. J Mol Biol 193(2): 413–417.
94. Zhang G, Ignatova Z. (2011) Folding at the birth of the nascent chain: Coordinating translation with co-translational folding. Curr Opin Struct Biol 21(1): 25–31.
95. Ingolia NT, Lareau LF, Weissman JS. (2011) Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. Cell 147(4): 789–802.
96. Novoa EM, Ribas de Pouplana L. (2012) Speeding with control: Codon usage, tRNAs, and ribosomes. Trends in Genetics 28(11): 574–581.
97. Stoletzki N. (2008) Conflicting selection pressures on synonymous codon use in yeast suggest selection on mRNA secondary structures. BMC Evolutionary Biology 8(1): 224.
98. Gu W, Zhou T, Wilke CO. (2010) A universal trend of reduced mRNA stability near the translation-initiation site in prokaryotes and eukaryotes. PLoS Comput Biol 6: e1000664. doi:10.1371/journal.pcbi.1000664
99. Gu W, Wang X, Zhai C, Xie X, Zhou T. (2012) Selection on synonymous sites for increased accessibility around miRNA binding sites in plants. Mol Biol Evol 29(10): 3037–3044.
100. Tuller T, Veksler-Lublinsky I, Gazit N, Kupiec M, Ruppin E, et al. (2011) Composite effects of gene determinants on the translation speed and density of ribosomes. Genome Biol 12(11): R110.
101. Tuller T, Carmi A, Vestsigian K, Navon S, Dorfan Y, et al. (2010) An evolutionarily conserved mechanism for controlling the efficiency of protein translation. Cell 141(2): 344–354.
102. Pechmann S, Frydman J. (2013) Evolutionary conservation of codon optimality reveals hidden signatures of cotranslational folding. Nature Structural & Molecular Biology 20: 237–243.
103. Zur H, Tuller T. (2012) Strong association between mRNA folding strength and protein abundance in S. cerevisiae. EMBO Rep 13: 272–277.
104. Tuller T, Waldman YY, Kupiec M, Ruppin E. (2010) Translation efficiency is determined by both codon bias and folding energy. Proceedings of the National Academy of Sciences 107(8): 3645–3650.
105. Katz L, Burge CB. (2003) Widespread selection for local RNA secondary structure in coding regions of bacterial genes. Genome Res 13(9): 2042–2051.
106. Park C, Chen X, Yang J, Zhang J. (2013) Differential requirements for mRNA folding partially explain why highly expressed proteins evolve slowly. Proceedings of the National Academy of Sciences 110(8): E678–E686.
107. Li F, Zheng Q, Ryvkin P, Dragomir I, Desai Y, et al. (2012) Global analysis of RNA secondary structure in two metazoans. Cell Rep 1: 69–82.
108. Bartel DP. (2009) MicroRNAs: Target recognition and regulatory functions. Cell 136(2): 215–233.
109. Kusenda B, Mraz M, Mayer J, Pospisilova S. (2009) MicroRNA biogenesis, functionality and cancer relevance. Biomedical Papers 150(2): 205–215.
110. Zhang G, Hubalewska M, Ignatova Z. (2009) Transient ribosomal attenuation coordinates protein synthesis and co-translational folding. Nature Structural & Molecular Biology 16(3): 274–280.
111. Deana Massaferno AE, Ehrlich Szalmian RM, Reiss C. (1996) Synonymous codon selection controls in vivo turnover and amount of mRNA in escherichia coli bla and ompA genes. J Bacteriol 2718: 2720.
112. Deana A, Ehrlich R, Reiss C. (1998) Silent mutations in the escherichia coli ompA leader peptide region strongly affect transcription and translation in vivo. Nucleic Acids Res 26(20): 4778–4782.
113. Qian W, Yang JR, Pearson NM, Maclean C, Zhang J. (2012) Balanced codon usage optimizes eukaryotic translational efficiency. PLoS Genet 8: e1002603. doi:10.1371/journal.pgen.1002603
114. Stadler M, Fire A. (2011) Wobble base-pairing slows in vivo translation elongation in metazoans. RNA 17(12): 2063–2073.
115. André S, Seed B, Eberle J, Schraut W, Bültmann A, et al. (1998) Increased immune response elicited by DNA vaccination with a synthetic gp120 sequence with optimized codon usage. J Virol 72(2): 1497–1503.
116. Kim CH, Oh Y, Lee TH. (1997) Codon optimization for high-level expression of human erythropoietin (EPO) in mammalian cells. Gene 199(1–2): 293–301.
117. Zolotukhin S, Potter M, Hauswirth WW, Guy J, Muzyczka N. (1996) A "humanized" green fluorescent protein cDNA adapted for high-level expression in mammalian cells. J Virol 70(7): 4646–4654.
118. Nagata T, Uchijima M, Yoshida A, Kawashima M, Koide Y. (1999) Codon optimization effect on translational efficiency of DNA vaccine in mammalian cells: Analysis of plasmid DNA encoding a CTL epitope derived from microorganisms. Biochem Biophys Res Commun 261(2): 445–451.
119. Carlini DB, Stephan W. (2003) In vivo introduction of unpreferred synonymous codons into the drosophila adh gene results in reduced levels of ADH protein. Genetics 163(1): 239–243.
120. Duan J, Wainwright MS, Comeron JM, Saitou N, Sanders AR, et al. (2003) Synonymous mutations in the human dopamine receptor D2 (DRD2) affect mRNA stability and synthesis of the receptor. Hum Mol Genet 12(3): 205–216.
121. Griseri P, Bourcier C, Hieblot C, Essafi-Benkhadir K, Chamorey E, et al. (2011) A synonymous polymorphism of the tristetraprolin (TTP) gene, an AU-rich mRNA-binding protein, affects translation efficiency and response to herceptin treatment in breast cancer patients. Hum Mol Genet 20(23): 4556–4568.
122. Kimchi-Sarfaty C, Oh JM, Kim IW, Sauna ZE, Calcagno AM, et al. (2007) A "silent" polymorphism in the MDR1 gene changes substrate specificity. Science 315(5811): 525–528.
123. Komar AA, Lesnik T, Reiss C. (1999) Synonymous codon substitutions affect ribosome traffic and protein folding during in vitro translation. FEBS Lett 462(3): 387–391.
124. Letzring DP, Dean KM, Grayhack EJ. (2010) Control of translation efficiency in yeast by codon–anticodon interactions. RNA 16(12): 2516–2528.
125. Pagani F, Raponi M, Baralle FE. (2005) Synonymous mutations in CFTR exon 12 affect splicing and are not neutral in evolution. Proc Natl Acad Sci U S A 102(18): 6368–6372.
126. Kudla G, Murray AW, Tollervey D, Plotkin JB. (2009) Coding-sequence determinants of gene expression in escherichia coli. Science 324(5924): 255–258.

127. Agashe D, Martinez-Gomez NC, Drummond DA, Marx CJ. (2013) Good codons, bad transcript: Large reductions in gene expression and fitness arising from synonymous mutations in a key enzyme. Mol Biol Evol 30(3): 549–560.
128. Waldman YY, Tuller T, Keinan A, Ruppin E. (2011) Selection for translation efficiency on synonymous polymorphisms in recent human evolution. Genome Biology and Evolution 3: 749.
129. Ingvarsson PK. (2010) Natural selection on synonymous and nonsynonymous mutations shapes patterns of polymorphism in populus tremula. Mol Biol Evol 27(3): 650–660.
130. Vishnoi A, Sethupathy P, Simola D, Plotkin JB, Hannenhalli S. (2011) Genome-wide survey of natural selection on functional, structural, and network properties of polymorphic sites in saccharomyces paradoxus. Mol Biol Evol 28(9): 2615–2627.
131. dos Reis M, Savva R, Wernisch L. (2004) Solving the riddle of codon usage preferences: A test for translational selection. Nucleic Acids Res 32(17): 5036–5044.
132. McQuilton P, Pierre SES, Thurmond J. (2012) FlyBase 101–the basics of navigating FlyBase. Nucleic Acids Res 40(D1): D706–D714.
133. Nelder JA, Mead R. (1965) A simplex method for function minimization. The Computer Journal 7(4): 308–313.
134. Meyer LR, Zweig AS, Hinrichs AS, Karolchik D, Kuhn RM, et al. (2013) The UCSC genome browser database: Extensions and updates 2013. Nucleic Acids Res 41(D1): D64–D69.
135. Hasegawa M, Kishino H, Yano T. (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol 22(2): 160–174.