Published in final edited form as:
J Mol Evol. 2012 October ; 75(3-4): 141–150. doi:10.1007/s00239-012-9528-x.

Effects of premature termination codon polymorphisms in the *Drosophila pseudoobscura* subclade

Kenneth B. Hoehn[1], **Suzanne E. McGaugh**[1], and **Mohamed A. F. Noor**[1]
[1]Biology Department, Duke University, Box 90388, Durham, NC 27708 USA

Abstract

Premature termination codon (PTC) mutations can have dramatic effects – both adaptive and deleterious – on gene expression and function. Here, we examine the number and selective effects of PTC mutations within the *Drosophila pseudoobscura* subclade using 18 resequenced genomes aligned to the reference genome. We located and characterized 1679 PTC mutations in 605 genes across each of these genomes relative to the *D. pseudoobscura* reference genome, and use RT-PCR to confirm transcription of a subset of these genes containing PTC mutations. We confirm previous findings that genes containing PTC mutations are less selectively constrained and less broadly expressed than non-PTC containing genes, suggesting that most of these mutations are at least mildly deleterious. Further, we find highly significant codon usage bias in regions downstream of the PTC in 38 of these PTC containing genes, suggesting that some of these PTC mutations – if not alternatively spliced out of the transcript– have neutral effects. Ultimately, these analyses support the view that PTC mutations are mostly detrimental, but are nonetheless common enough in genomes that a subset could be effectively neutral.

Keywords

PTC; nonsense mutation; premature stop codon; null mutation

Introduction

Mutations resulting in premature termination (stop) codon (PTCs) might be presumed to be disadvantageous and may result in the production of truncated, non-functional, proteins (Khajavi et al. 2006; Mort et al. 2008). However, many RNA transcripts harboring PTCs are also subject to nonsense-mediated decay which specifically destroys transcripts with PTCs prior to translation (NMD, Khajavi et al. 2006; Hansen et al. 2009), and thus the predicted truncated protein may be rare or absent (Chang et al. 2007). In either event, the net effects of PTC mutations are generally negative, with pathological effects in humans ranging from Parkinson's disease to myopathy (Lee and Reinhardt 2012; Yamaguchi-Kabata et al. 2008). Other studies have suggested links between PTC mutations and virulence of *Staphylococcus* infections (Young et al. 2011). Nonetheless, PTCs may be an important source of phenotypic variation, as they have been found to be variable across the genome between individuals and across several species (Durbin at al. 2010; Lee and Reinhardt 2012; Jungreis et al. 2011; Yngvadottir et al. 2009). Other studies have shown that some PTCs and other loss-of-function mutations can be adaptive, such as in cases of host-specialization of *Drosophila sechellia* on morinda fruit (Dworkin and Jones 2009) or potentially neutral, such as in the yellow coat color of labrador and golden retrievers (Everts et al. 2000) or the

[2]Corresponding author: Kenneth B. Hoehn, Email of corresponding author: kenneth.hoehn@gmail.com, Telephone: 919-613-8193 (Noor lab), Fax: (919) 660-7293 (Duke Biology Department).

melanic form of deer mice (Kingsley et al. 2009) and cats (Eizirik et al. 2003). This varied pattern of effects, along with their relative ease of detection given the availability of whole genome sequencing and alignment, make PTCs good candidates for studying effects of natural selection across diverse loci within and between closely related species.

Because only a small number of broadly comparative studies on PTCs have been done in model systems, some questions remain about their evolutionary influences. Research in *Drosophila melanogaster* suggests that PTC-containing genes have a narrower range of expression (Lee and Reinhardt 2012), perhaps implying that broadly expressed genes experience stronger purifying selection (Subramanian and Kumar 2004). However, these findings may not be general, and other questions have not yet been addressed. For example, we could not find studies that examined the prevalence of PTCs in relation to recombination rate, which is thought to influence natural selection's effectiveness (Hill and Robertson 1966; Charlesworth et al. 2009). Generally speaking, few studies leveraged multiple genome sequences from within and between numerous closely related species to examine variation in PTCs (but see Lee and Reinhardt 2012; Yngvadottir et al. 2009). Instead, most studies of PTC mutations have leveraged comparisons within one or between two species (e.g. McBride 2007; Hansen et al. 2009), compiled results using single representative genome sequences from disparate clades to study the effects of NMD and fragile-codon robustness (Cusack et al. 2011), or focused primarily on within-species PTC variation (Yngvadottir et al. 2009).

Here, we leverage genomic sequences of the *Drosophila pseudoobscura* subgroup to examine variation within and between species in PTC prevalence and to infer genetic and evolutionary consequences. This system boasts comparisons at multiple levels: two subspecies (*D. pseudoobscura* and *D. ps. bogotana*), a sister species that hybridizes with *D. pseudoobscura* (*D. persimilis*), and a closely related outgroup species that does not hybridize with the other taxa (*D. miranda*). Genome resequence data are available for multiple strains of each of these taxa (McGaugh et al. 2012) as well as precise estimates of local recombination rates at a fine scale (Kulathinal et al. 2008; Stevison and Noor 2010; McGaugh et al. 2012). Further, many genes within the *D. pseudoobscura* group share a 1:1 ortholog relationship with those in *D. melanogaster*, allowing the use of extensive expression data available from this model system.

We explore PTC variation within the *D. pseudoobscura* subclade with 18 resequenced genomes including 10 inbred lines of *D. pseudoobscura,* two *D. ps. bogotana*, three *D. persimilis*, and three *D. miranda*. After detecting and filtering PTC mutations from each of these genomes, we tested for: (i) loss of transcription for genes with PTCs using RT-PCR; (ii) difference of selective constraint of non-synonymous vs. synonymous substitution rate; (iii) broadness of tissue expression between genes with PTC mutations and those without; and (iv) association of PTC prevalence with local recombination rate.

We surveyed four PTC-containing genes, confirmed the PTC mutation through Sanger sequencing of genomic DNA, and demonstrated that the PTC-containing genes were transcribed through RT-PCR and sequencing. We confirm previous findings (Yamaguchi-Kabata et al. 2008; Lee and Reinhardt 2012) that PTC mutations are generally detrimental and that genes with PTC mutations are less broadly expressed than non-PTC genes. We further find no significant relationship between fine-scale recombination rate and PTC presence.

Methods

Genome Sequences

We utilized resequenced genomes from the *Drosophila pseudoobscura* clade and the reference genomes of *D. pseudoobscura* (Richards et al. 2005) v.2.9 and *D. persimilis* (Clark et al. 2007) v.1.3. Illumina reads were obtained from resequenced genomes of *Drosophila pseudoobscura* (N = 10 isofemale lines), *D. ps. bogotana* (N = 2), *D. persimilis* (N = 2), *D. miranda* (N = 3), and *D. lowei* (N = 1). The *D. lowei* sequence was only used for the dN/dS analyses (see below). Sequences and alignment methods have all been published previously (McGaugh et al. 2012; McGaugh and Noor 2012). Briefly, genomes were resequenced from virgin females of inbred lines using various Illumina platforms and aligned to the *D. pseudoobscura* reference genome v2.9 using bwa-0.5.5 (Li and Durbin 2009). Samtools 0.1.6 pileup (Li et al. 2009) consensus calls were used as the base for each resequenced line, and aligned resequenced genomes were filtered for coverage (> 4 reads), quality (≥ 30 Phred score) and bases within 5bp of indels were excluded. FlyBase annotations of the *Drosophila pseudoobscura* genome Release 2.9 were used in demarcating coding, intronic, and intergenic regions.

Detection, Characterization, and Expression of PTC mutations

We wrote custom Perl scripts to scan each of the aligned, resequenced genomes for PTCs. In this study, PTCs were defined as stop codons within a CDS occurring before the identified stop codon position in the annotated reference *D. pseudoobscura* genome. For each gene, we recorded the percentage of the coding sequence before the PTC, in the strain being analyzed, that fit our filtering criteria and was not within 5 bp +/− an indel relative to the reference genome. We then scored each PTC gene based on predicted percent truncation of CDS, number of strains containing a PTC in that gene, and number of Gene Ontology (GO) notes – which give categorizations of the gene's functionality - available from FlyBase (www.flybase.org; McQuilton et al. 2011). Predicted percent coding sequence truncation, as defined here, refers to the percent of a coding sequence downstream of a PTC (e.g. 90% truncation of a 100 codon gene has a PTC at codon 10). Because NMD in *Drosophila* is not dependent on exon-junction complexes (Cusack et al. 2011), predicted sequence truncation was used as a proxy for the potential effect a PTC mutation might have on a gene. We then used the Batch Download tool from FlyBase (McQuilton et al. 2011) to infer orthologs in *D. melanogaster* for each gene when available.

To confirm that a subset of predicted PTC-containing genes were transcribed, we selected four genes – GA21687, GA22859, GA19967, and GA15892 from the above list to assay for transcription using RT-PCR. All four genes had *D. melanogaster* homologues with known mutant phenotypes or electronic annotation through InterPro (via FlyBase; Apweiler et al. 2000), and were predicted to suffer at least 50% coding sequence truncation. We selected genes with a broad range of occurrence in the genomes surveyed – two genes (GA21687; chromosome 2, GA22859; XL chromosome arm) contained a PTC in only one of the 18 genomes, one gene (GA15892; XR chromosome arm) contained a PTC in two genomes, and one gene (GA19967; XR group 8) contained a PTC in 14 genomes.

We first designed primers flanking the PTC mutations of each gene, and confirmed the presence of each PTC mutation by Sanger sequencing of genomic DNA. Having confirmed that these mutations were not next-generation sequencing errors, we then began testing for transcription by designing primers flanking an intron in each of these four genes of interest, as mature mRNA transcripts should be missing the introns. The introns for each gene were >50 bp, so that the size difference between mature mRNA transcripts and genomic DNA

could be easily visualized on an agarose gel. As a positive control, we developed primers for a non-PTC containing gene (GA15693; chromosome 3).

The Qiagen RNeasy and QiaShredder kit was used to isolate RNA from 20–30 male and female flies. We used RT-PCR protocol of Michalak and Noor (2003) to amplify each of these segments. PCR products were visualized on a 2% agarose gel, and we compared the presence of bands in the PTC gene lane to those in the control non-PTC gene to ensure the reaction worked. For each of the four genes, if a band of the proper size was found in the experimental lane, we interpreted this as evidence for the presence of an mRNA transcript. Because genes transcribed into mRNA lack introns, we also confirmed that the gene was transcribed by using Sanger sequencing to verify the absence of an introns in the cDNA.

To test the range of expression of genes with PTC mutations, we first surveyed FlyBase (www.flybase.org; McQuilton et al. 2011) to retrieve *D. melanogaster* orthologs of each gene with PTC mutations expected to cause at least 50% predicted sequence truncation. We used these orthologs to access multiple tissue microarray data from FlyAtlas (www.flyatlas.org; Chintapalli et al. 2007). FlyAtlas “present” call data was obtained from four duplicate arrays run for each tissue tested; thus, each gene has a score ranging from 0/4 – 4/4 for each of 20 tissues. This corresponds to the number of duplicate arrays in which the gene was called as “present” in a given tissue. As in Lee and Reinhardt (2012), we excluded non-protein coding genes and called each gene as expressed if it was detected in either ¾ or 4/4 microarray replicates. Next, we determined the number of tissues in which a gene was expressed. This ranged in value from 0, for no repeatably detectable expression in any tissue, to 20, for repeatably detectable expression in all tissues. A Fisher's exact test was used to determine whether PTCs were enriched in the category of no repeatably detectable expression in any tissues (expressed in 0 of 20 tissues), versus any tissues (expressed in 1–20 of 20 tissues). Like Lee and Reinhardt (2012), we used raw FlyAtlas expression data to determine the tissue in which a gene was expressed most highly, and then used Fisher's exact test with a Bonferroni-adjusted p value to determine if, for each tissue type tested, there was enrichment of PTCs or non-PTCs reaching their highest level of expression in that tissue.

Recombination Analysis

We used fine scale recombination maps (McGaugh et al. 2012) to determine whether there was an association between the number of PTC-containing genes in a recombination window and recombination rate in that window. Briefly, three recombination maps were generated for chromosome 2 and the XR and XL chromosome arms. Two maps were made for *D. pseudoobscura* (Flagstaff and Pikes Peak), and one was made for *D. miranda* through different backcrosses of inbred lines. In all maps, recombination was measured over intervals that were, in most cases, < 200kb (mean interval lengths were 195kb, 208kb, and 187kb, for *D. pseudoobscura Pikes Peak*, *D. pseudoobscura Flagstaff*, and *D. miranda*, respectively)

We filtered out any PTC mutation that did not truncate at least 50% of the coding sequence to be more confident that the PTCs tested could have a selective effect. We then performed a linear least squares regression between the percentage of genes in a recombination window containing a PTC to the rate of recombination in that window. These analyses were controlled by species and strain, such that only PTC mutations in *D. miranda* were used with *D. miranda* recombination maps, only PTCs in the *D. pseudoobscura* lines used to make the Flagstaff maps (Flagstaff16 and Flagstaff18) were used with *D. pseudoobscura* Flagstaff recombination maps, and only PTCs in the *D. pseudoobscura* lines used to make the Pikes Peak maps (Pikes Peak 1134 and 1137) were used with *D. pseudoobscura* Pikes Peak recombination maps.

Codon Bias Analysis

Variation in stop codon location could reflect loss of a stop codon or addition of a premature termination codon. Functionally important genes tend to evolve biased codon usage, often preferentially using G- or C-ending codons in this species group (Akashi and Schaeffer 1997). To provide evidence that the PTC occurs prior to the end of the functionally important region of a gene, and the PTCs we observe are not simply a result of misannotation of the reference genome, we measured codon usage bias within the reference sequence of each PTC gene, and sequences of each derived PTC gene downstream of the PTC. We assigned a codon as preferred in *D. pseudoobscura* if it was designated preferred in all three methods used inVicario et al. (2007). We used a binomial test to examine significant deviations from 1 of the preferred/unpreferred ratio. This was done on a per gene per resequenced genome basis. To simplify the assessment of statistical significance, we only included 2-fold and 4-fold codons which had 1/2 preferred and 1/2 unpreferred codons. Significance was assessed with a Bonferroni adjusted p-value. Using only codons 3' to the most downstream PTC mutation in the resequenced genome containing the most downstream PTC per gene, we further measured the average ratio of preferred to unpreferred codons in all genes pooled together, and used a binomial test to assess the significance of this preferred/unpreferred ratio in deviating from 1.

PAML and dN/dS Analysis

All of the genomic data described above (one *D. lowei*, three *D. miranda*, three *D. persimilis*, two *D. pseudoobscura bogotana* and 11 *D. pseudoobscura* genomes) was used in codeml in PAML v4.4d (Yang 2007) to calculate maximum likelihood estimates of dN and dS. Genes that did not have sufficient coverage in at least eight *D. pseudoobscura* genomes and one genome from each: *D. lowei*, *D. pseudoobscura bogotana*, *D. persimilis,* and *D. miranda* were excluded from this analysis.

We used a tree rooted with *D. lowei* and considered the branches leading to (*D. persimilis* (*D. pseudoobscura*, *D. pseudoobscura bogotana*)) to be the foreground branches, though we used a free-ratio model which does not utilize that particular information. The following specifications were used in codeml: i) Maximum likelihood estimates of dN and dS were calculated with pairwise comparisons of sequences (runmode = −2); ii) Codon frequencies were estimated from the nucleotide frequencies observed in the data (F3X4 method); iii) ω (d_N/d_S) was not constrained to be equal on all branches of the tree (so-called ‘free-ratio’ model; model = 1); iv) No sites model was specified (NSsites = 0); v) Transition-transversion ratio was estimated for each gene; vi) A single α for all sites was assumed; vii) All ambiguous characters or alignment gaps were removed from all sequences (cleandata = 1); and viii) Branch lengths were estimated with maximum likelihood from starting values (fix_blength = 1).

Because dN/dS can be highly variable in short gene sequences, we controlled these measurements (similar to Lee and Reinhardt 2012) by excluding any genes with a coding sequence shorter than 100 amino acids, and removing any genes with a dS measurement of < 0.001, which is indicative of an extreme poverty of synonymous substitutions and would otherwise result in highly exaggerated dN/dS ratios. We used Mann-Whitney tests to assess significance of difference between dN/dS ratios of PTC and non-PTC containing genes.

Minor Allele Frequency Analysis

Demonstrating that the minor allele frequency (MAF) of PTC alleles is significantly lower than the MAF of neutral fourfold degenerate codons would suggest that premature stop codon alleles tend to be deleterious. We first determined the average MAF for each fourfold degenerate codon in PTC containing genes. Because it was possible for each gene to have

more than one site containing a PTC, we characterized each PTC allele by its 5'-most PTC mutation. We then determined the MAF of PTC alleles in these same genes by discarding all genes that contain more than one PTC allele, and then determining the frequency of the PTC allele in the remaining genes. Because we expect spurious PTC mutations resulting from misannotation of the reference genome to be at abnormally high frequency, and because we only examine synonymous alleles at frequency < 0.5, we only considered PTC-containing genes in which the PTC mutation was the minor allele (frequency < 0.5). We took two statistical approaches to comparing MAF: 1) using a paired-Wilcoxon test, and 2) tallying the instances in which the PTC MAF was lower than the average fourfold degenerate codon MAF in the same gene, and using a binomial test to assess the significance of this sign difference. Because *D. persimilis*, *D. pseudoobscura bogotana*, and *D. miranda* only had 2–3 intraspecific genomes each, we only included 10 genomes from *D. pseudoobscura* in this analysis.

Results and Discussion

In total, we found hundreds of PTC mutations in the *D. pseudoobscura* subclade, and while our results are consistent with the hypothesis that PTC mutations are generally deleterious, further analyses suggest that some of these genes are still transcribed, and that a subset may be neutral.

PTC mutations are widespread in the *Drosophila pseudoobscura* subclade

Within the 18 Drosophila genomes searched, we detected 1679 PTCs in 605 genes, within a total set of 10,468 genes. Each gene in this set contained a putative PTC in an average of 2.78 of the 18 genomes (range = 1–18, median = 2, mode = 1). Similarly in a survey of 44 *D. melanogaster* genomes 438 genes were found to contain PTC polymorphisms (Lee and Reinhardt 2012). We expect our dataset to be biased towards less deleterious PTC mutations because these genomes were obtained from homozygous inbred lines which would have eliminated the most strongly deleterious alleles (Lee and Reinhardt 2012). Though the average predicted sequence truncation was 43%, we found the distribution of predicted sequence truncation to be highly skewed towards 0–10% sequence truncation, meaning that much of the coding sequence was left intact, with a slight increase in 80–100% predicted truncation (Figure 1). A similar distribution is found in *D. melanogaster* (Lee and Reinhardt 2012), and this suggests that PTC mutations are generally deleterious, and that the observable PTC mutations in inbred lines are skewed towards leaving most of the coding region intact. Importantly, some putative PTC mutations may be not deleterious because of alternate or incorrectly inferred start sites, or splicing that removes the PTC mutation.

Four genes molecularly surveyed are transcribed despite PTC mutations

We selected four genes for RT-PCR and Sanger sequencing of the cDNA product-- these mutations were predicted to result in a high degree of truncation, and were found in a range of numbers of genome sequences (Table 1). All four PTC mutations were confirmed with Sanger sequencing of genomic DNA to not be next-generation sequencing errors or pseudogenes. For GA19967, we amplified regions with RT-PCR upstream and downstream of the PTC. For GA21687, transcript was only amplified downstream of the PTC. For GA22859 and GA15892, we only amplified a transcript with RT-PCR upstream of the PTC, though, we did not attempt to amplify a region downstream.

If these genes were pseudogenes or misannotations of noncoding sequence, we would likely find no transcription or intron splicing. In contrast, we found evidence for transcription and intron splicing in all four of these genes. These PTCs may persist because of alternative spicing around the PTC mutation, though in the case of GA19967, we actually amplified the

PTC in the transcript, indicating that the PTC was not spliced out of an existing mRNA transcript. However, this approach does not test for translation of truncated proteins – only transcription of mRNA. Many mRNA transcripts which contain a PTC are degraded through nonsense-mediated decay (NMD), which would prevent translation though not prevent transcription (Chang et al. 2007).

PTC containing genes are less selectively constrained than non-PTC containing genes

We used dN/dS estimates between PTC containing genes and non-PTC containing genes to test the hypothesis that genes harboring PTC mutations are, as a group, under less selective constraint than other genes. We found that PTC containing genes have significantly higher dN/dS ratios than non-PTC containing genes (Wilcoxon rank sum, W = 1047585, 671314.5, 722402.5, and $P < 0.001$ for *D. pseudoobscura, D. persimilis*, and *D. miranda*, respectively). In order to create a more conservative data set, we filtered out genes with a dN/dS ratio greater than or equal to 0.8, and repeated our analysis. Results were consistent for all three species (W = 836836, 477253, and 546452.5, and $P < 0.001$ for *D. pseudoobscura, D. persimilis,* and *D. miranda*, respectively). Higher dN/dS ratios provide evidence for either increased positive selection or relaxed purifying selection (Lee and Reinhardt 2012). However, because previous work (Lee and Reinhardt 2012) has shown that patterns of PTC mutations in *D. melanogaster* are consistent with reduced purifying selection on PTC containing genes, we interpret these results as evidence for reduced selective constraint on PTC containing genes. This interpretation is consistent with the hypothesis that PTC mutations are generally deleterious.

Orthologs of PTC-containing genes are more narrowly expressed than non-PTC orthologs

Another means for inferring selective constraint is to analyze a gene's expression profile. We expect genes expressed in many tissues to be under stronger purifying selection than genes expressed in fewer tissues (Subramanian and Kumar 2004; Begun and Lindfors 2005; Van Dyken and Wade 2010). As with Lee and Reinhardt (2012), we used microarray expression data from FlyAtlas (Chintapalli et al. 2007) to assess these differences in expression. However, because FlyAtlas was created using genes in *D. melanogaster*, our sample size and statistical power was reduced to genes in *D. pseudoobscura* with known orthologs in *D. melanogaster*. In total, 259 of the original 605 PTC genes in *D. pseudoobscura* had orthologs in *D. melanogaster* – 150 of these were found in FlyAtlas, and only 54 had a predicted truncation of > 50%. In total, 5020 non-PTC containing genes in our dataset had orthologs in *D. melanogaster* that were also used in FlyAtlas. Despite this reduction in sample size, we found that PTC containing genes were marginally more likely than non-PTC containing genes to have no persistently detectable expression in any tissue both when compared to non-PTC genes detectably expressed in any tissue (Fisher's exact test, $P = 0.05$) and non-PTC genes expressed in all 20 tissues (Fisher's exact test, $P = 0.061$; Figure 2a). We also found that PTC-containing genes were qualitatively less likely to be expressed in all 20 tissues, though this relationship is not significant (Fisher's exact test, $P = 0.85$). We also tested for enrichment of PTC mutations in particular tissue types; unlike Lee and Reinhardt (2012) we found no significant enrichment of either PTC or non-PTC containing genes in any particular tissue type using a Bonferroni adjusted p-value (Figure 2b), though this may likely be due to reduced sample size of our dataset since we were using only those genes which had orthologs in *D. melanogaster*. Thus, while we were unable to confirm Lee and Reinhardt's (2012) findings that PTC mutations are enriched in specific tissues such as the larval fat body, we were able to show that PTC mutations were significantly more likely to be not detectably expressed in any tissue. This latter finding is further consistent with the hypothesis that PTC mutations are, as a group, detrimental.

Minor allele frequency of PTC mutations is significantly lower than minor allele frequency of neutral sites

We compared the minor allele frequencies of PTC alleles and fourfold degenerate codon sites to assess the deleterious nature of PTC mutations. Using a paired Wilcoxon test, we found that PTC minor alleles had a significantly lower frequency than fourfold degenerate codon minor alleles in the same genes ($V = 26364$, $P = 0.0303$). We also tallied the number of instances in which the PTC minor allele had a lower frequency than the average fourfold degenerate codon MAF in the same gene. We found that PTC minor alleles were marginally significantly more likely to be at a lower frequency than fourfold degenerate minor alleles on a gene-by-gene basis (Binomial test, $P = 0.0587$). Because minor alleles PTC mutations tend to occur at a lower frequency than neutral sites, this is a relatively strong indication that they are as a group under negative selective forces.

No significant relationship found between PTC content and recombination rate

We used fine-scale recombination maps to directly analyze the relationship between PTC mutations and local recombination rate. Recombination is expected to increase the efficacy of natural selection in removing deleterious mutations (Hill and Robertson 1966; Charlesworth et al. 1993), and our results above suggest that PTC mutations are generally selected against; thus, we expect to find a negative relationship between the percentage of genes containing a PTC in a recombination window and the recombination rate of that window. Using linear least squares regression, however, we found no significant relationship between of the percentage of PTC containing genes and recombination rate ($R^2 < 0.001$ and $P > 0.6$ in all comparisons; Figure 3a). We performed the same analysis but comparing the relationship between predicted coding sequence disruption and recombination rate, hypothesizing that PTC mutations would be skewed towards reduced sequence disruption in regions of high recombination. We found similar patterns of non-significance of relationship ($R^2 < 0.002$ and $P > 0.5$ in all comparisons; Figure 3b). Because weakly deleterious mutations are preferentially retained in regions of low recombination (reviewed in Charlesworth 2012), this lack of a relationship between recombination rate and PTC abundance may suggest that either a nontrivial subset of the PTC mutations we surveyed are nearly neutral, or that our approach simply lacked the power to detect a relationship between PTC density and recombination rate variation in ~200kb intervals. The latter option seems particularly likely because the overwhelming majority of recombination intervals had zero, one, or two, PTC mutations.

Biased codon usage suggests neutrality or positive selection on some PTC containing genes

We further examined the evolutionary forces affecting PTC-containing genes by examining the codon bias 3' downstream of the putative PTC mutations. If the original *D. pseudoobscura* reference genome sequence has either a sequencing error or the *D. pseudoobscura* reference genome is misannotated, the true stop codon for a transcript may appear as a PTC. In these cases, if the reference is misannotated, we do not expect to observe codon usage bias downstream of the putative PTC. On the other hand, if we observe codon usage bias downstream of the putative PTC in *multiple* strains, then the spread of the PTC could be neutral or perhaps even adaptive (Vicario et al. 2007).

We used a binomial test to assess whether the preferred to unpreferred codon usage ratio deviated significantly from one in the sequence downstream of the most downstream PTC for each gene in each genome, and a Bonferroni-adjusted p value as a cutoff to identify significantly biased genes. The average preferred/unpreferred codon ratio in these downstream regions was 1.506 counting GA28861 (a far outlier which had a bias ratio of 75), and 1.36 excluding GA28861. Using only the genome containing the most downstream

PTC for each gene, we also pooled all codons surveyed, yielding a highly significant preferred/unpreferred codon ratio of 1.32 (binomial test, $P < 0.001$), indicating a weak preferred codon bias in the full set. Further, for 85 PTC mutations in 38 genes across our 18 genomes, we found significant codon usage bias downstream of the PTC even after Bonferroni correction. In these 38 genes, the average preferred/unpreferred ratio was 4.08 (2.08 without GA28861). These genes had a much higher average predicted truncation of 82% instead of 43% found in all PTC genes. That these genes exhibited statistically significant binomial tests may actually reflect the number of codons available for the test, as statistical significance for codon usage bias will be detected for genes with more codons (i.e. more percent truncation). This may account for the difference in predicted truncation between these two groups. The high preferred codon usage in these genes strongly suggests that the regions downstream of the PTC in many of these genes were at least historically under selection for translational efficiency. Many of the genes exhibiting preferred codon usage downstream of their respective PTC are found in multiple genomes and have interesting orthologs in *D. melanogaster.* Of particular interest are genes which contain PTC mutations in almost all genomes, such as GA24678, whose *D. melanogaster* ortholog is involved in neurotransmitter transport, or genes that contain PTC synapomorphies, such as GA15481, which was found exclusively in all three *D. miranda* genomes and has orthologs involved in dorsal/ventral axis specification. In the case of PTCs found in multiple lineages, the PTC mutation may be old and possibly neutral. We cannot exclude, however, that the PTC is spliced out during expression of many of these genes or that the PTC mutations became common variants despite being detrimental because of linkage with adaptive alleles.

Conclusion

Premature termination codons represent an important class of mutations, in part because they have the potential to cause dramatic change in gene function and expression that can have both adaptive and deleterious phenotypic effects (Dworkin and Jones 2009; Yamaguchi-Kabata et al. 2008). While previous work (Lee and Reinhardt 2012) has addressed the effects of primarily intraspecific PTC mutations in *D. melanogaster*, the effects of PTC mutations in other species has not been investigated as carefully.

Consistent with Lee and Reinhardt's (2012) findings, genes in our data set which contain PTC mutations exhibit narrower expression and more relaxed purifying selection than genes in our dataset without PTC mutations, and are highly skewed towards reduced predicted sequence truncations. We see no correlation of PTC mutations with recombination rate, implying that these mutations are not consistently "very weakly deleterious" and preferentially retained in regions of low recombination. Our further analyses also suggest that these mutations tend to be detrimental, but that a subset may be neutral because they exhibited at least historical CUB and were found in multiple lineages. While suggestive, we cannot completely exclude the alternative hypothesis that all PTC mutations we've studied are deleterious but that they spread due to strong genetic drift or linkage to sweeping advantageous variants. Nonetheless, this research has confirmed patterns observed in another *Drosophila* species and adds a list of candidate genes for further study of the evolutionary fate of PTC mutations.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

D. Bachtrog and S. Nuzhdin provided some sequence data, and C. Machado provided flies for sequencing. Thanks to J. Merrill and two anonymous reviewers for comments on the manuscript. This research was funded by NIH grants GM092501 and GM086445. Data was deposited in the Short Read Archive (accession numbers SRA044960.1, SRA044955.2, SRA044956.1); see also http://pseudobase.biology.duke.edu/.

References

Akashi H, Schaeffer SW. Natural Selection and the Frequency Distributions of ``silent'' DNA Polymorphism in Drosophila. Genetics. 1997; 146:295–307. [PubMed: 9136019]

Apweiler R, Attwood TK, Bairoch A, Bateman A, Birney E, Biswas M, Bucher P, Cerutti L, Corpet F, Croning MDR, Durbin R, Falquet L, Fleischmann W, Gouzy J, Hermjakob H, Hulo N, Jonassen I, Kahn D, Kanapin A, Karavidopoulou Y, et al. InterPro—an integrated documentation resource for protein families, domains and functional sites. Bioinformatics. 2000; 16:1145–1150. [PubMed: 11159333]

Begun DJ, Lindfors HA. Rapid Evolution of Genomic Acp Complement in the melanogaster Subgroup of Drosophila. Molecular Biology and Evolution. 2005; 22:2010–2021. [PubMed: 15987879]

Chang Y-F, Imam JS, Wilkinson MF. The Nonsense-Mediated Decay RNA Surveillance Pathway. Annual Review of Biochemistry. 2007; 76:51–74.

Charlesworth B. The Effects of Deleterious Mutations on Evolution at Linked Sites. Genetics. 2012; 190:5–22. [PubMed: 22219506]

Charlesworth B, Betancourt AJ, Kaiser VB, Gordo I. Genetic Recombination and Molecular Evolution. Cold Spring Harbor Symposia on Quantitative Biology. 2009; 74:177–186.

Charlesworth B, Morgan MT, Charlesworth D. The effect of deleterious mutations on neutral molecular variation. Genetics. 1993; 134:1289–1303. [PubMed: 8375663]

Chintapalli VR, Wang J, Dow JAT. Using FlyAtlas to identify better Drosophila melanogaster models of human disease. Nature Genetics. 2007; 39:715–720. [PubMed: 17534367]

Clark AG, Eisen MB, Smith DR, Bergman CM, Oliver B, Markow TA, Kaufman TC, Kellis M, Gelbart W, Iyer VN, Pollard DA, Sackton TB, Larracuente AM, Singh ND, Abad JP, Abt DN, Adryan B, Aguade M, Akashi H, Anderson WW, et al. Evolution of genes and genomes on the Drosophila phylogeny. Nature. 2007; 450:203–218. [PubMed: 17994087]

Consortium T1000 GP. A map of human genome variation from population-scale sequencing. Nature. 2010; 467:1061–1073. [PubMed: 20981092]

Cusack BP, Arndt PF, Duret L, Crollius HR. Preventing Dangerous Nonsense: Selection for Robustness to Transcriptional Error in Human Genes. PLoS Genetics. 2011; 7:e1002276. [PubMed: 22022272]

Dworkin I, Jones CD. Genetic Changes Accompanying the Evolution of Host Specialization in Drosophila sechellia. Genetics. 2009; 181:721–736. [PubMed: 19033155]

Dyken JDV, Wade MJ. The Genetic Signature of Conditional Expression. Genetics. 2010; 184:557–570. [PubMed: 19966065]

Eizirik E, Yuhki N, Johnson WE, Menotti-Raymond M, Hannah SS, O'Brien SJ. Molecular Genetics and Evolution of Melanism in the Cat Family. Current Biology. 2003; 13:448–453. [PubMed: 12620197]

Everts RE, Rothuizen J, van Oost BA. Identification of a premature stop codon in the melanocyte-stimulating hormone receptor gene (MC1R) in Labrador and Golden retrievers with yellow coat colour. Animal Genetics. 2000; 31:194–199. [PubMed: 10895310]

Hansen KD, Lareau LF, Blanchette M, Green RE, Meng Q, Rehwinkel J, Gallusser FL, Izaurralde E, Rio DC, Dudoit S, Brenner SE. Genome-Wide Identification of Alternative Splice Forms Down-Regulated by Nonsense-Mediated mRNA Decay in Drosophila. PLoS Genetics. 2009; 5:e1000525. [PubMed: 19543372]

Hill WG, Robertson A. The effect of linkage on limits to artificial selection. Genetics Research. 1966; 8:269–294.

Jungreis I, Lin MF, Spokony R, Chan CS, Negre N, Victorsen A, White KP, Kellis M. Evidence of abundant stop codon readthrough in Drosophila and other metazoa. Genome Research. 2011; 21:2096–2113. [PubMed: 21994247]

Khajavi M, Inoue K, Lupski JR. Nonsense-mediated mRNA decay modulates clinical outcome of genetic disease. European Journal of Human Genetics. 2006; 14:1074–1081. [PubMed: 16757948]

Kingsley EP, Manceau M, Wiley CD, Hoekstra HE. Melanism in Peromyscus Is Caused by Independent Mutations in Agouti. PLoS ONE. 2009; 4:e6435. [PubMed: 19649329]

Kulathinal RJ, Bennett SM, Fitzpatrick CL, Noor MAF. Fine-scale mapping of recombination rate in Drosophila refines its correlation to diversity and divergence. Proceedings of the National Academy of Sciences USA . 2008; 105:10051–10056.

Lee YCG, Reinhardt JA. Widespread Polymorphism in the Positions of Stop Codons in Drosophila melanogaster. Genome Biology and Evolution. 2012; 4:533–549. [PubMed: 22051795]

Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics. 2009; 25:1754–1760. [PubMed: 19451168]

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009; 25:2078–2079. [PubMed: 19505943]

McBride CS. Rapid evolution of smell and taste receptor genes during host specialization in Drosophila sechellia. Proceedings of the National Academy of Sciences USA. 2007; 104:4996–5001.

McGaugh SE, Noor MAF. Genomic impacts of chromosomal inversions in parapatric Drosophila species. Philosophical Transactions of the Royal Society B. 2012; 367:422–429.

McGaugh SE, Heil CSS, Manzano-Winkler B, Loewe L, Goldstein S, Himmel T, Noor MAF. Recombination modulates how selection affects linked sites in Drosophila. PLoS Biology. 2012 in press.

McQuilton P, St. Pierre SE, Thurmond J. the FlyBase Consortium. FlyBase 101 - the basics of navigating FlyBase. Nucleic Acids Research. 2011; 40:D706–D714. [PubMed: 22127867]

Michalak P, Noor MAF. Genome-Wide Patterns of Expression in Drosophila Pure Species and Hybrid Males. Molecular Biology and Evolution. 2003; 20:1070–1076. [PubMed: 12777520]

Mort M, Ivanov D, Cooper DN, Chuzhanova NA. A meta-analysis of nonsense mutations causing human genetic disease. Human Mutation. 2008; 29:1037–1047. [PubMed: 18454449]

Richards S, Liu Y, Bettencourt BR, Hradecky P, Letovsky S, Nielsen R, Thornton K, Hubisz MJ, Chen R, Meisel RP, Couronne O, Hua S, Smith MA, Zhang P, Liu J, Bussemaker HJ, van Batenburg MF, Howells SL, Scherer SE, Sodergren E, et al. Comparative genome sequencing of Drosophila pseudoobscura: Chromosomal, gene, and cis-element evolution. Genome Research. 2005; 15:1–18. [PubMed: 15632085]

Stevison LS, Noor MAF. Genetic and Evolutionary Correlates of Fine-Scale Recombination Rate Variation in *Drosophila persimilis*. Journal of Molecular Evolution. 2010; 71:332–345. [PubMed: 20890595]

Subramanian S, Kumar S. Gene Expression Intensity Shapes Evolutionary Rates of the Proteins Encoded by the Vertebrate Genome. Genetics. 2004; 168:373–381. [PubMed: 15454550]

Vicario S, Moriyama EN, Powell JR. Codon usage in twelve species of Drosophila. BMC Evolutionary Biology. 2007; 7:226. [PubMed: 18005411]

Yamaguchi-Kabata Y, Shimada MK, Hayakawa Y, Minoshima S, Chakraborty R, Gojobori T, Imanishi T. Distribution and Effects of Nonsense Polymorphisms in Human Genes. PLoS ONE. 2008; 3:e3393. [PubMed: 18852891]

Yang Z. PAML 4: Phylogenetic Analysis by Maximum Likelihood. Molecular Biology and Evolution. 2007; 24:1586–1591. [PubMed: 17483113]

Yngvadottir B, Xue Y, Searle S, Hunt S, Delgado M, Morrison J, Whittaker P, Deloukas P, Tyler-Smith C. A Genome-wide Survey of the Prevalence and Evolutionary Forces Acting on Human Nonsense SNPs. The American Journal of Human Genetics. 2009; 84:224–234.

Young, Bernadette C.; Golubchik, Tanya; Batty, Elizabeth M.; Fung, Rowena; Larner-Svensson, Hanna; Votintseva, Antonina A.; Miller, Ruth R., et al. Evolutionary Dynamics of Staphylococcus

Aureus During Progression from Carriage to Disease. Proceedings of the National Academy of Sciences USA. 2012; 109:4550–4555.

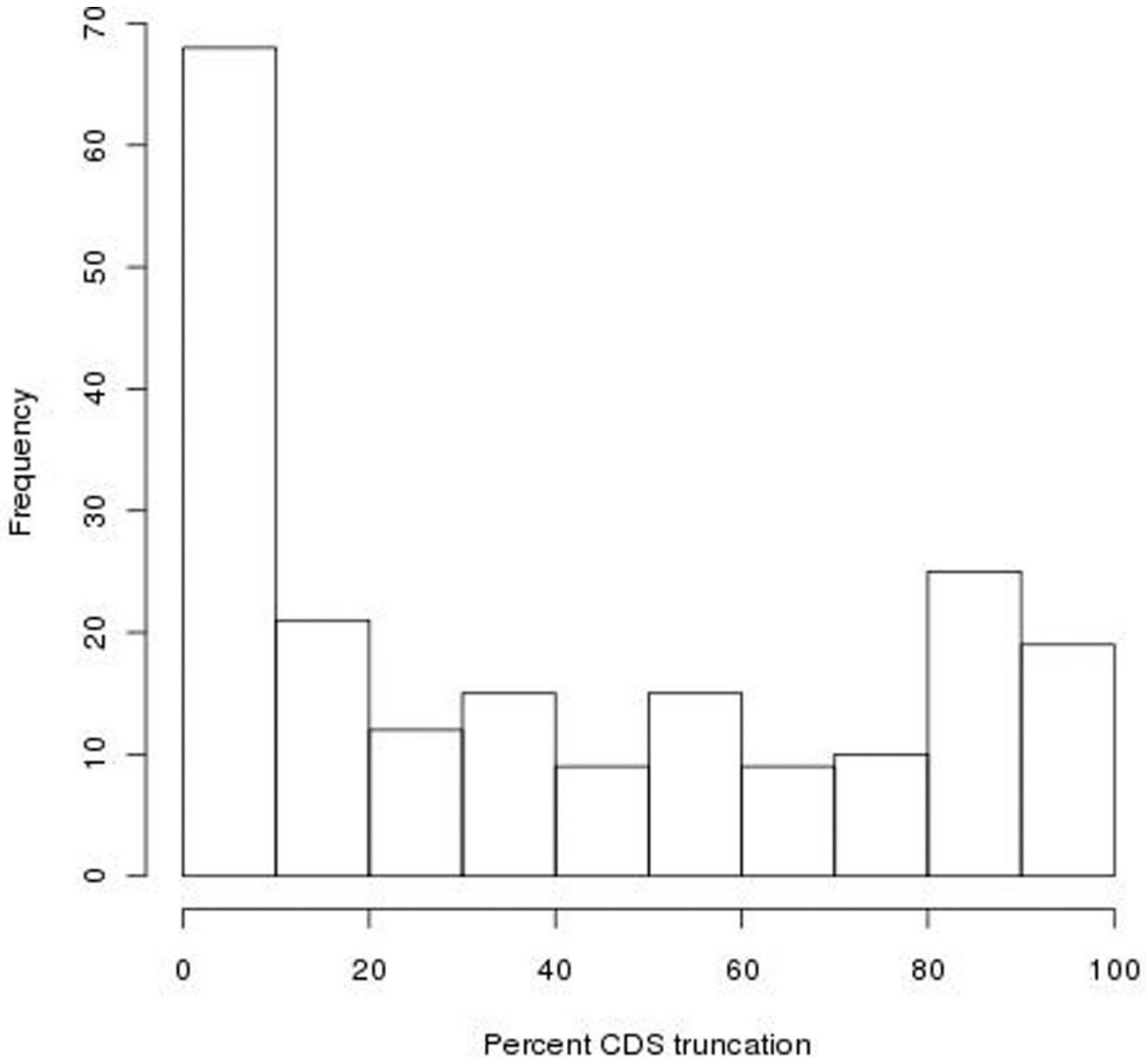
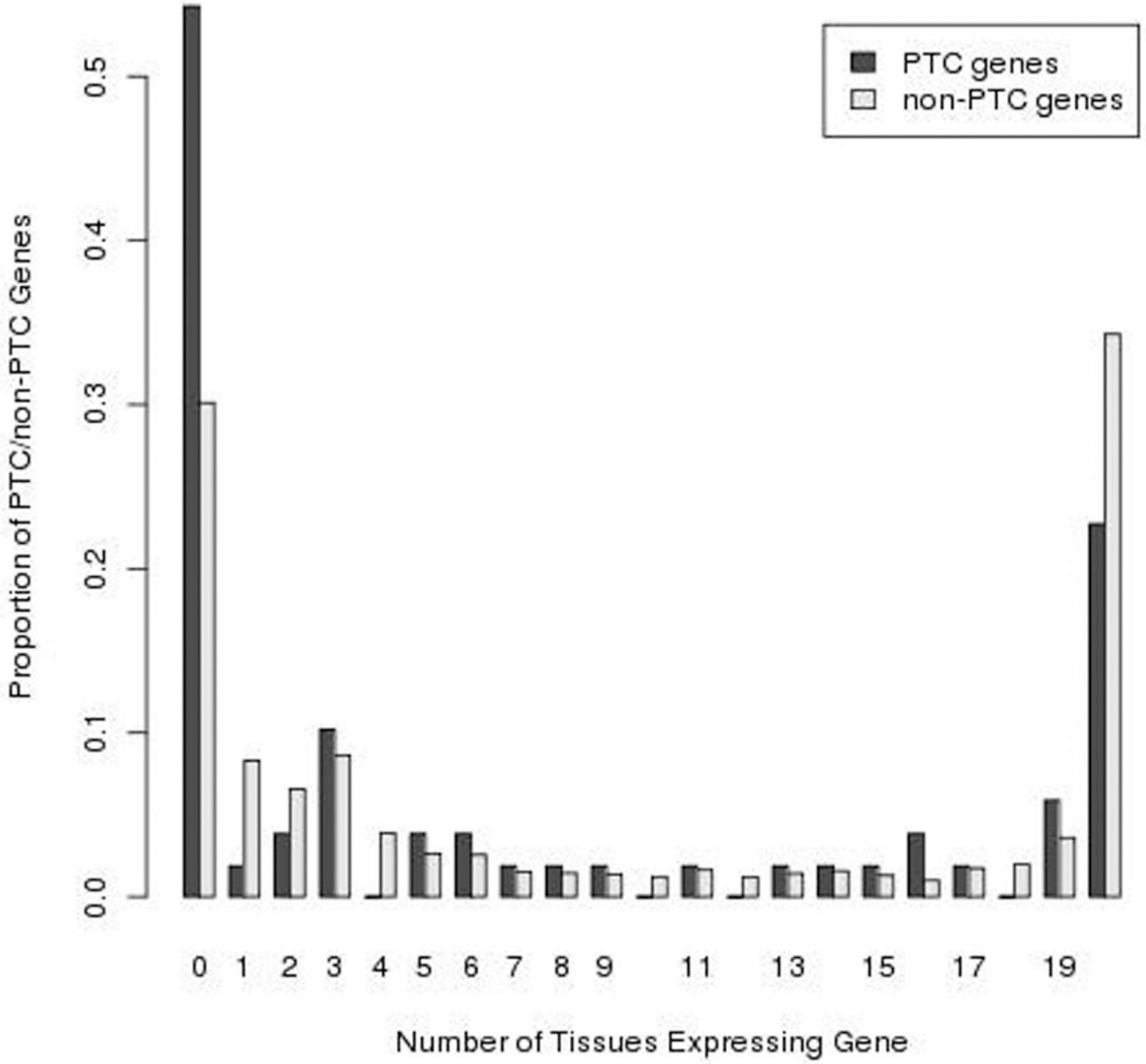


Figure 1. Distribution of sequence truncation in *D. pseudoobscura*

Distribution of predicted coding sequence disruption for genes in *D. pseudoobscura*.



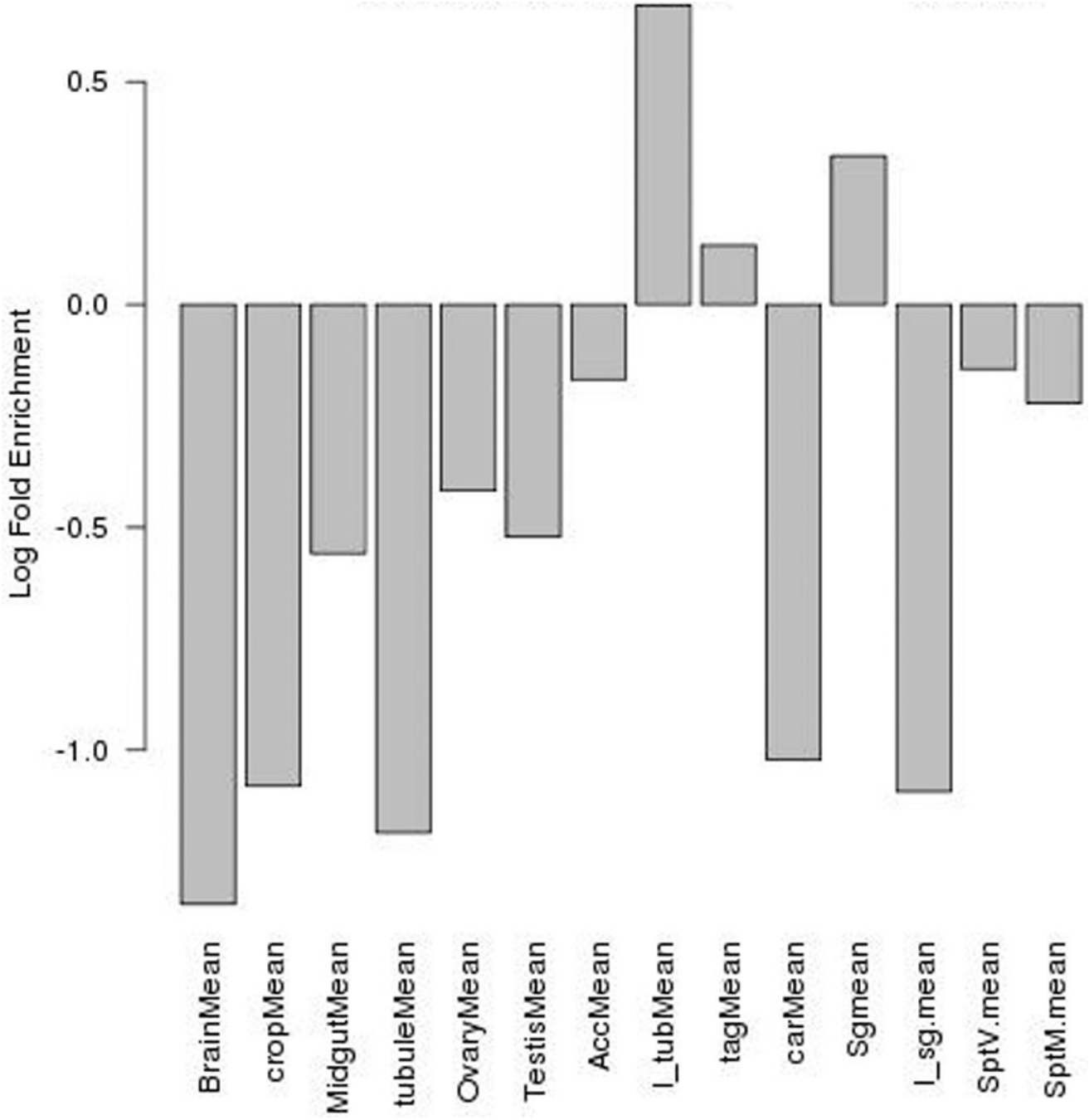


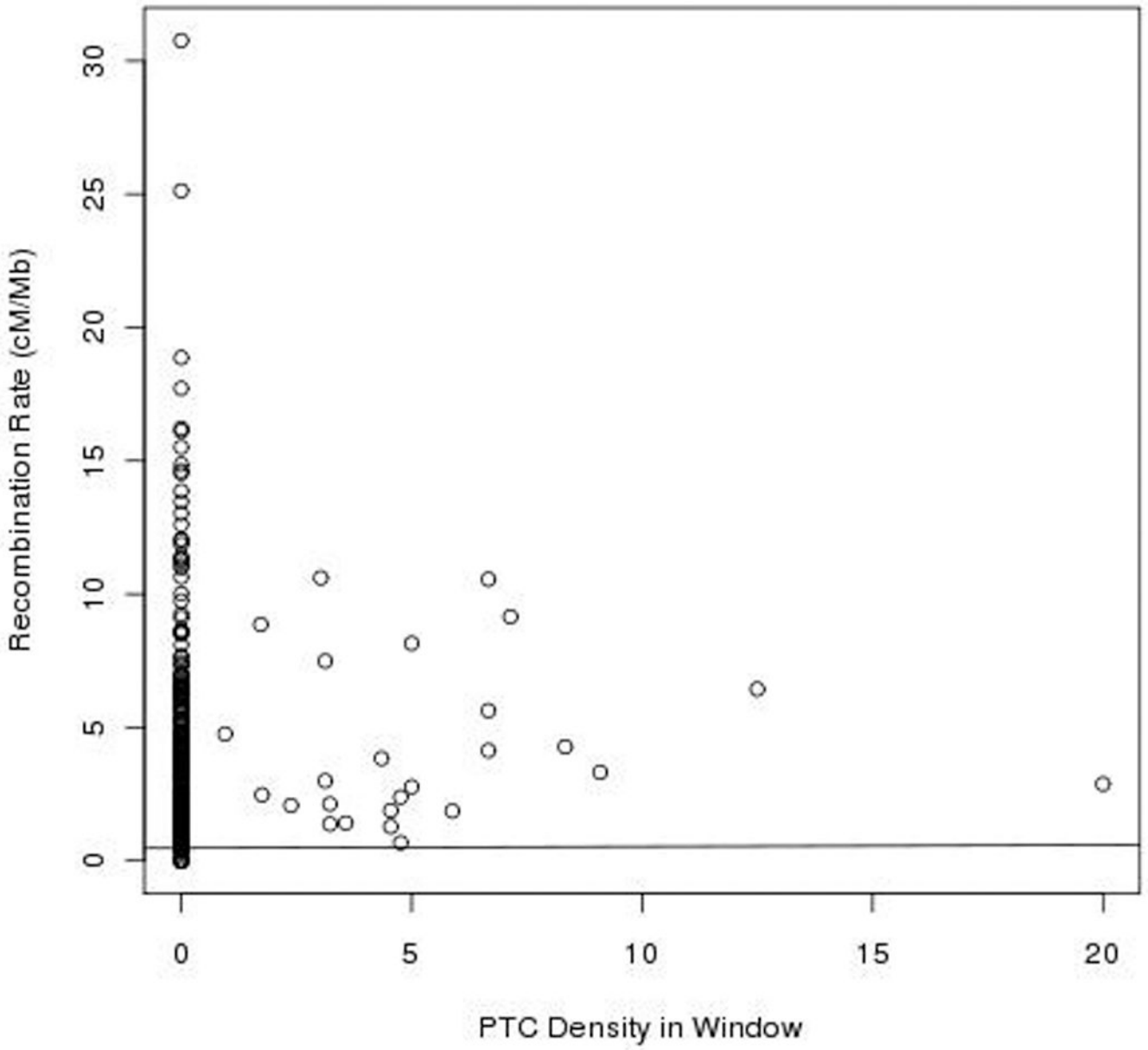
Figure 2.

a: Distribution of Tissue Expression

Distribution of tissue expression count for *D. melanogaster* orthologs of premature termination codon (PTC) containing and non-PTC containing genes.

b: Log Fold Enrichment of PTC Genes by Tissues

Log fold enrichment of PTC genes with highest expression in different tissues.



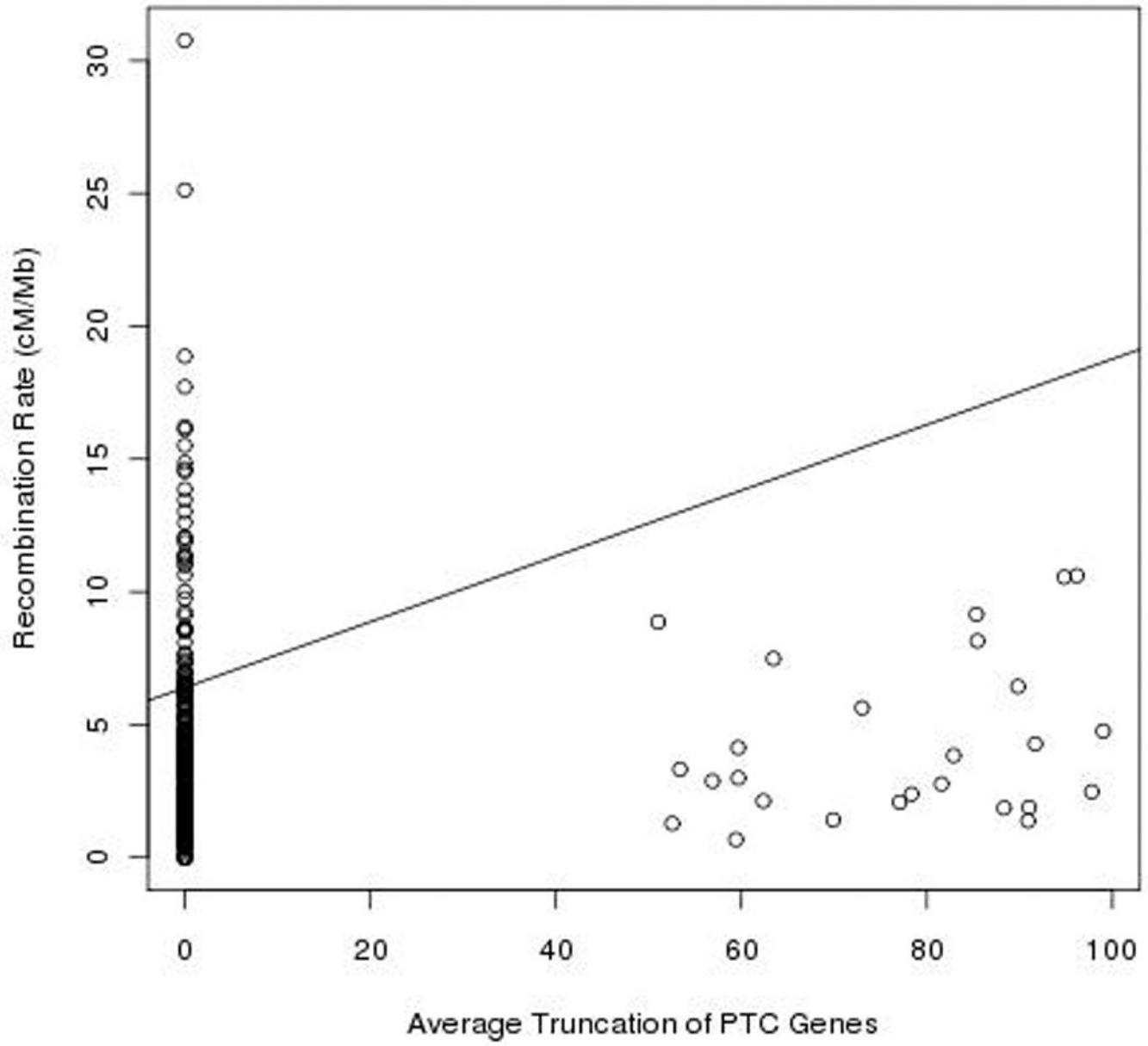


Figure 3.

a: PTC Density and Recombination in *D. pseudoobscura*

Linear least squares regression of PTC density and recombination rate in *D. pseudoobscura* Pikes Peak. Similar results were found for *D. miranda* and *D. pseudoobscura* Flagstaff (Supplementary figures S1, S2).

b: % Truncation and Recombination in *D. pseudoobscura*

Linear least squares regression of average predicted sequence truncation of PTC mutations in a recombination window, and recombination rate in that window. *D. pseudoobscura* Pikes Peak shown. Similar results were found for *D. miranda* and *D. pseudoobscura* Flagstaff (Supplementary figures S3, S4)

Table 1

PTC containing genes selected for RT-PCR and Sanger sequencing of cDNA. All four were Sanger sequenced to confirm that the PTC mutation observed was not a sequencing error in the Illumina data. Sanger sequencing of RT-PCR products showed that each of these genes are transcribed despite the presence of a PTC mutation.

***D. pseudoobscura* ID**	***D mel* Ortholog**	**Genomes with PTC mutation**	**Predicted % truncation**	**Prominent GO notes for *D. mel* ortholog**
GA15892	*SNF1A*	2 – Flagstaff14, Mather32	51.87	establishment or maintenance of epithelial cell apical/basal polarity
GA19967	*Mth*	14 – Found in all species except *D. miranda*	89.86	determination of adult lifespan
GA21687	*Cht5*	1 - Flagstaff14	96.61	chitin catabolic process
GA22859	*Cyp4d2*	1 – Pikes Peak 1134	87.05	oxidation-reduction process

Table 2

Table of PTC containing genes showing significantly high CUB and orthology GO information in *D. melanogaster*. Each of these are either found in multiple lineages of the same species (synpamorphies in the case of GA15481), or are found in all species surveyed.

***D. pseudoobscura* ID**	***D. melanogaster* Ortholog**	**Genomes with PTC mutation**	**Preferred/unpreferred codon ratio**	**Predicted % truncation**	**Prominent GO notes for *D. mel* ortholog**
GA24678	*List*	13 – Found in all species	2.–2.12	95.94	neurotransmitter transport, response to lithium ion
GA15892	*SNF1A*	2 – Unique to *D. pseudoobscura*	2.6–2.8	51.87	establishment or maintenance of epithelial cell apical/basal polarity
GA24335	*lectin-46C*	7 – Unique to *D. pseudoobscura*	3–3.5	24.29	regulation of female receptivity, post-mating, sperm competition
GA15481	*phl*	3 – Unique to *D. miranda*	1.7	89.88	border follicle cell migration, dorsal/ventral axis specification, ovarian follicular epithelium,
GA16951	*CG32506*	3 – Unique to *D. persimilis*	1.93	97.22	regulation of Rab GTPase activity