Are Nonsense Alleles of *Drosophila melanogaster* Genes under Any Selection?

Nadezhda A. Potapova[1,2], Maria A. Andrianova[1,3], Georgii A. Bazykin[1,3], and Alexey S. Kondrashov[2,4,*]

[1]Institute of Information Transmission Problems (Kharkevich Institute) of the Russian Academy of Sciences, Moscow, Russia

[2]Department of Bioengineering and Bioinformatics, Lomonosov Moscow State University, Moscow, Russia

[3]Skolkovo Institute of Science and Technology, Moscow, Russia

[4]University of Michigan, Ann Arbor, USA

*Corresponding author: E-mail: kondrash@umich.edu.

Accepted: February 6, 2018

Abstract

A gene which carries a bona fide loss-of-function mutation effectively becomes a functionless pseudogene, free from selective constraint. However, there is a number of molecular mechanisms that may lead to at least a partial preservation of the function of genes carrying even drastic alleles. We performed a direct measurement of the strength of negative selection acting on nonsense alleles of protein-coding genes in the Zambian population of *Drosophila melanogaster*. Within those exons that carry nonsense mutations, negative selection, assayed by the ratio of missense over synonymous nucleotide diversity levels, appears to be absent, consistent with total loss of function. In other exons of nonsense alleles, negative selection was deeply relaxed but likely not completely absent, and the per site number of missense alleles declined significantly with the distance from the premature stop codon. This pattern may be due to alternative splicing which preserves function of some isoforms of nonsense alleles of genes.

Key words: drosophila, loss-of-function, nonsense allele, negative selection, neutral evolution.

Introduction

A gene whose function does not contribute to fitness is on its way to becoming a pseudogene. Thus, with few exceptions (Xue et al. 2006; MacArthur et al. 2007), genes must be protected by negative, or purifying, selection which removes their loss-of-function (LoF) alleles. Nevertheless, genotypes of individuals carry substantial numbers of LoF alleles or, more precisely, of alleles that are likely to cause a complete loss of function of a protein-coding gene (MacArthur et al. 2012). Such alleles include nonsense substitutions which produce a premature stop codon as well as frameshift deletions, insertions, and complex mutations. In humans, there are 53–100 LoF alleles per genotype, including 21–27 nonsense alleles (MacArthur et al. 2012; Li et al. 2015). Data on nonsense alleles are more abundant and more reliable than on other LoF alleles, because calling frameshift alleles when genotypes are studied by resequencing is problematic. There are ~30 nonsense alleles per genotype in pig *Sus scrofa* (Groenen et al. 2012), ~100 in an alga *Chlamydomonas reinhardtii* (Flowers et al. 2015), and ~18 in North American or ~35 in Zambian populations of *Drosophila melanogaster* (Lack et al. 2015; Yang et al. 2015).

Large per genotype numbers of nonsense and other LoF alleles may suggest that at least some of them do not, in fact, lead to a complete loss of function. Indeed, there is a number of molecular mechanisms that could ensure at least a partial preservation of function of an allegedly LoF allele, including alternative splicing, stop codon readthrough, and alternative translation initiation (Jagannathan and Bradley 2016). Several cases of functioning nonsense alleles have been described (Prieto-Godino et al. 2016).

Still, there is no doubt that, on an average, a nonsense allele is more deleterious than a missense allele. For instance, a nonsense allele is three times more likely to lead to a disease than a missense allele (Krawczak et al. 1998). Per site prevalence of nonsense alleles in all studied populations is substantially lower than that of missense alleles (Mort et al. 2008; Yamaguchi-Kabata et al. 2008; Kono et al. 2016). Genes that harbor nonsense alleles have narrower expression profiles, are commonly involved in dispensable biological processes, and

have many paralogs, which makes loss of their functions less deleterious (Lee and Reinhardt 2012; MacArthur et al. 2012; Yang et al. 2015).

To investigate the impact of nonsense alleles on the function of affected genes, we performed a direct measurement of the strength of negative selection acting within these alleles in a natural population of *D. melanogaster*.

Materials and Methods

Drosophila melanogaster Genome Data Sets

We used the DPGP3 data set of genomes of Zambian *D. melanogaster* haploid embryos *as our main data set* (Lack et al. 2015; http://www.johnpool.net/genomes.html; last accessed February 15, 2018). We used only those 196 genomes for which all the three major chromosomes, 2, 3, and X, were available. We also analyzed two smaller data sets (of ~50 individuals each) from Africa (AGES) and North America (NUZHDIN) (Lack et al. 2015). The annotation file has been downloaded from UCSC Genome Browser for version 3 of *D. melanogaster* genome (http://hgdownload.cse.ucsc.edu/goldenPath/dm3/database/flyBaseGene.txt.gz; last accessed February 15, 2018). Canonical splice variants of genes are from (http://hgdownload.cse.ucsc.edu/goldenPath/dm3/database/flyBaseCanonical.txt.gz; last accessed February 15, 2018). We used the longest isoform for every gene, in order to include as many nonsense mutations as possible in analysis.

Data Filtering

We focused on single-nucleotide nonsense substitutions. A gene was excluded from the analysis if the start codon was not ATG, if the stop codon was not TAA, TAG, or TGA, or if the length of the coding sequence was not in a multiple of 3. About 90 genes that contain at least one nonsense allele with the frequency >0.3 were excluded from estimates of negative selection, because such nonsense alleles are often spurious. Individual nonsense alleles located within the first or the last 5% of the length of the ORF were also excluded (MacArthur et al. 2012) from these estimates. About 73% (1231/1689) of genes and 62% (1726/2786) of nonsense alleles survived this filtering.

pN/pS Estimation

pS was calculated using 4-fold degenerate synonymous sites, and pN was calculated from nondegenerate sites at second positions within each codon only. For each site of the corresponding category, site-specific pN or pS were calculated as

$$1-\sum (n_i/N)^2,$$

where N is the number of genotypes, n_i is the number of genotypes carrying a certain nucleotide, and summation is over all four nucleotides. pN and pS were then obtained by averaging these values over all sites of the corresponding category.

Fig. 1.—Schematic representation of mutation types in nonsense alleles. The presence of a nonsense mutation (shown as a square) subdivides the sample into nonsense and sense alleles. For analysis of pN/pS, we considered only those synonymous or missense mutations that only occurred in nonsense alleles, but did not occur in all nonsense alleles (stars). Such mutations are most likely to have arisen after the nonsense mutation against its background. Mutations in all nonsense alleles (circles), or mutations occurring in some sense alleles (triangles), were not considered.

To calculate pN and pS for nonsense alleles, we used only nonsingleton nonsense alleles (i.e., those observed two or more times in our sample). We analyzed only those synonymous and missense polymorphisms that were nested within the nonsense alleles (fig. 1), because when a polymorphism is present in all nonsense alleles there is a chance that it originated before the nonsense mutation.

Calculation of pN/pS ratios for mutations nested within frequency-matched synonymous alleles with those nested within nonsense alleles was performed in genes with nonsense alleles using the formula described earlier.

The confidence intervals for pN/pS ratios were estimated from 10,000 bootstrap trials resampling case. Bootstrapping was performed by individual genes.

pN and pS estimations for figure 5 were calculated for sliding windows of width 210 (70 codons), with the step 50.

Drosophila melanogaster RNA-Seq Data Sets

We used RNA-seq data set SRR3135045 (von Heckel et al. 2016) for Zambian *D. melanogaster* from the NCBI SRA (https://www.ncbi.nlm.nih.gov/sra; last accessed February 15, 2018). Raw reads were downloaded with SRA Toolkit (v. 2.8.0). Then we trimmed this data using Trimmomatic (v. 0.32) and made quality control using FastQC (v. 0.10.1). Transcriptome was mapped to *D. melanogaster* reference genome (dm3) with TopHat (v. 2.1.0). Coverage for nonsense exons was calculated using BEDTools (v. 2.16.2) with option "coverage -counts –abam."

Then for each gene, we calculated the relative density of nonsense exon reads as the following ratio:

$$(N\_\text{nons}/L\_\text{nons})/\sum_i (N\_\text{free}_i/L\_\text{free}_i),$$

where N_nons is the number of reads mapped onto an exon carrying the nonsense mutation, L_nons is the length of this exon, N_free$_i$ is the number of reads mapped onto the ith

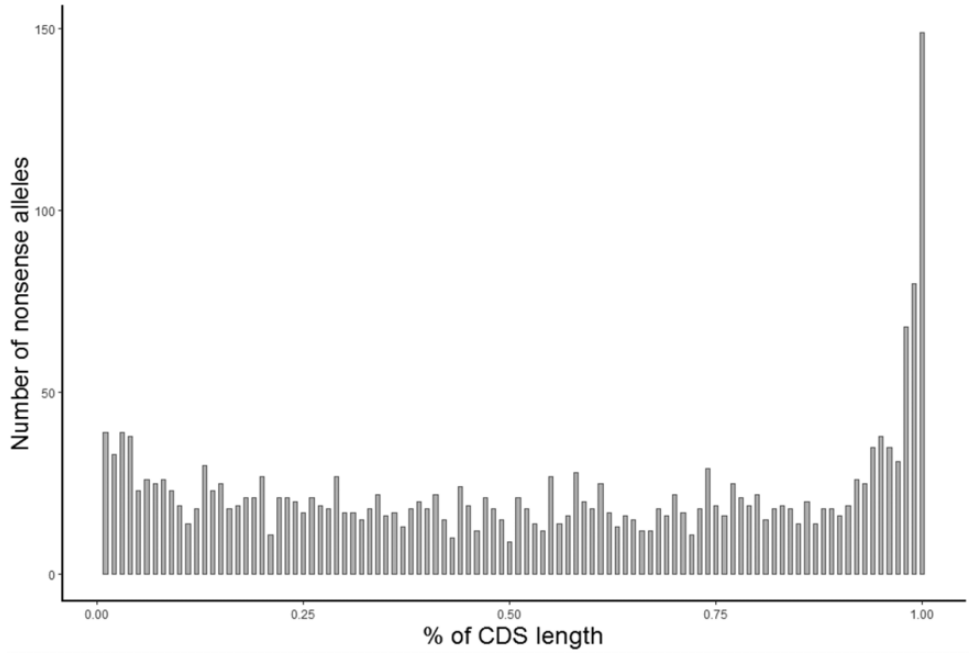

Fig. 2.—Relative positions of nonsense mutations within coding portions of genes which harbor them.

nonsense-free exon, L_free_i is the length of the ith exon. The final value was obtained by averaging over all genes.

Results

Prevalence of Nonsense Alleles

We investigated only nonsense alleles that resulted from a single nucleotide substitution. Below, the term "nonsense allele" refers to both a nonsense mutation and a haplotype which carries it.

In 196 haploid genotypes of Zambian *D. melanogaster* we detected, within canonical isoforms of 13,300 protein-coding genes, 1,726 nonsense alleles within 1,231 genes. Among these genes, 767 carried only singleton nonsense alleles, that is, nonsense alleles that were observed in just a single genotype. The remaining 464 genes carried both singleton and nonsingleton or only nonsingleton nonsense alleles. The total number of singleton nonsense alleles was 1,236. On an average, each genotype contained 35 genes with 36 nonsense alleles (including singletons), or 30 genes with 31 nonsense alleles (excluded singletons). The proportions of genes that harbor nonsense alleles were similar between chromosomes 2 and 3, but twice as low for X chromosome (supplementary table 1, Supplementary Material online), in line with higher efficiency of negative selection against nonsense alleles in hemizygous state (Mackay et al. 2012).

Figure 2 demonstrates a significant excess of nonsense alleles near the 3′-end of genes (χ^2 test, P value = 8.518e-13), where they may not always destroy the function completely (Wetterbom et al. 2009; Lee and Reinhardt 2012). In contrast, there is no significant excess of nonsense alleles near the 5′-end of genes (χ^2 test, P value = 0.6818).

Figure 3 presents the distribution of frequency x of nonsense alleles. In agreement with the data obtained previously (Li and Stephan 2006), we see a substantial excess of very rare nonsense alleles (i.e., of singletons and of those that appeared twice in our sample of 196 genotypes) over the neutral expectation of $\sim 1/x$ (Mann–Whitney test, P value $= 3.6 \times 10^{-11}$) (Wright 1931; Kimura 1983), which must be at least partially due to negative selection against them.

Negative Selection in Nonsense Alleles

Next, we asked whether selection affects polymorphisms that segregate at the genetic background of a nonsense allele. First, we compared the numbers of missense and synonymous SNPs in genes with and without nonsense alleles. After that, we compared these numbers in nonsense-carrying versus nonsense-free alleles of genes that possess nonsingleton nonsense alleles (fig. 4).

Table 1 presents data on the strengths of negative selection, characterized by pN/pS ratios, in classes of genes defined by the presence, within our sample of genotypes, of nonsense alleles in them. Not surprisingly, genes that carry nonsense alleles, and especially nonsingleton nonsense alleles, are, on an average, under weaker selection. Values for pN/pS calculated separately only for those alleles of genes that do not carry nonsense mutations are only slightly lower than the

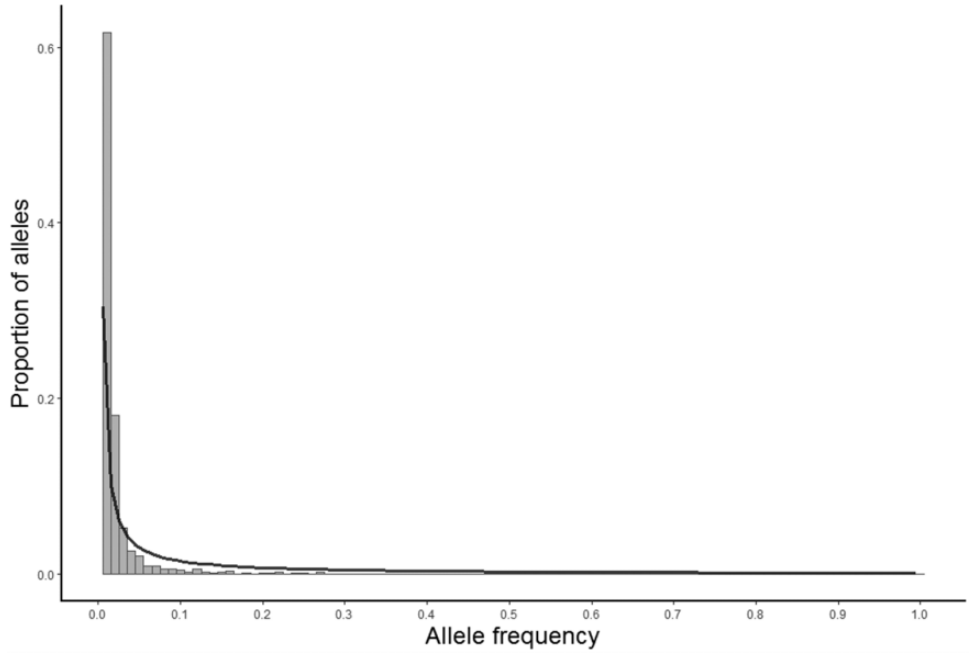

Fig. 3.—The observed distribution of frequencies of nonsense (gray bars) alleles and the $\sim 1/x$ expected distribution (black line).

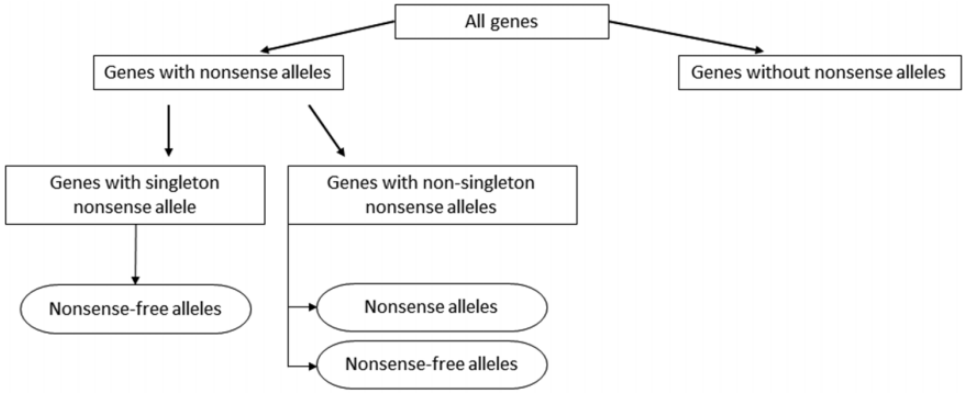

Fig. 4.—Workflow of data analysis. Classes of genes are shown in rectangles and classes of alleles are shown in ovals.

values for all alleles of the same genes, because nonsense alleles are rare. By contrast, the value of pN/pS obtained for nonsingleton nonsense alleles on the basis of missense and synonymous polymorphisms nested within them is much higher than that for alleles that do not carry a nonsense mutation, and is not significantly different from 1, indicating that selection against missense mutations is reduced or absent. Such polymorphisms are present in only 140 out of the 464 genes that harbor nonsingleton nonsense alleles, leading to wide confidence interval for this value (pN and pS values separately are shown in supplementary tables 2 and 3, Supplementary Material online). Still, the same analysis using smaller data sets of other *D. melanogaster* populations shows similar patterns (supplementary tables 4 and 5, Supplementary Material online).

We analyzed Zambian *D. melanogaster* transcriptome data and found that the ratio of the densities of reads for nonsense-carrying over nonsense-free exons is 0.39 [95% CI: 0.32–0.46], indicating that a substantial proportion of nonsense-carrying exons are incorporated only in rare splice isoforms. However, when we considered 66 genes (out of 464 nonsense-carrying genes in our data set) with two or more annotated splice isoforms and subdivided coding sites within each of these genes into two categories, those incorporated into all isoforms and only into a subset of isoforms, we did not observe any difference between per site prevalences of nonsense alleles in these

Table 1
Average pN/pS in Zambian Population of *Drosophila melanogaster**

	All 12,842 Genes	11,611 Genes Without Nonsense Alleles	767 Genes With Only Singleton Nonsense Alleles	464 Genes With Only Nonsingleton Nonsense Alleles
All alleles	0.106 [0.103–0.109]	0.093 [0.091–0.095]	0.186 [0.171–0.203]	0.262 [0.238–0.290]
Nonnonsense alleles	0.106 [0.103–0.109]	0.093 [0.091–0.095]	0.186 [0.171–0.202]	0.256 [0.231–0.283]
Polymorphisms nested within 567 nonsingleton nonsense alleles	–	–	–	0.803 [0.584–1.121]

*95% CIs are shown in square brackets.

Table 2
Average pN/pS for Polymorphisms Nested within Nonsense Alleles*

pN/pS	121 Nonsense Alleles of 99 One-Exon Genes	446 Nonsense Alleles of 335 Multiexon Genes	
		Exon with Nonsense Mutation	Exons without Nonsense Mutation
	1.020 [0.770–1.530]	0.906 [0.556–1.500]	0.787 [0.459–1.359]

*95% CIs are shown in square brackets.

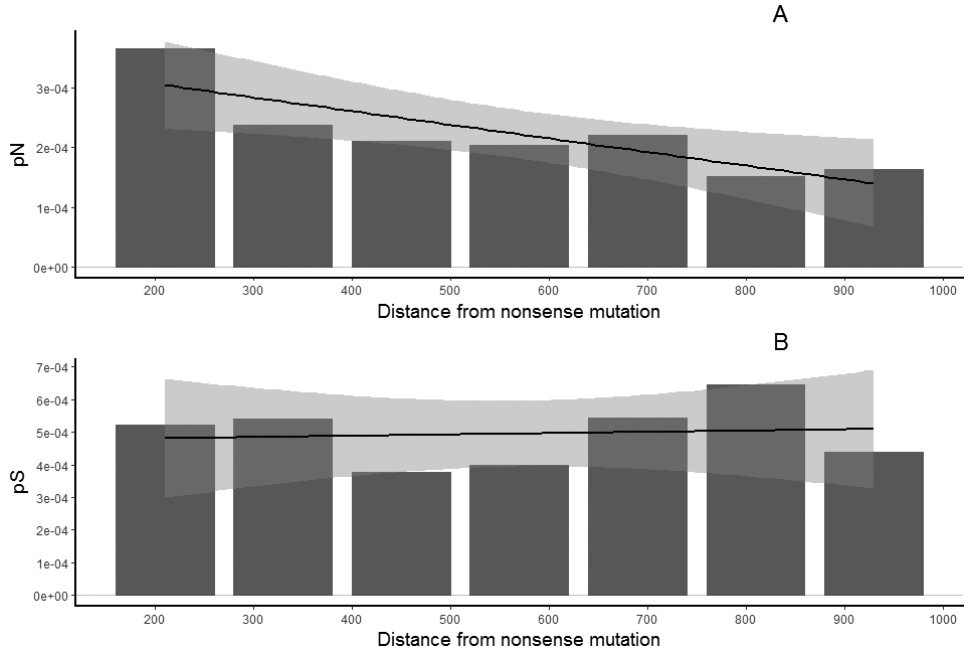

Fig. 5.—Dependencies of pN (*A*) and pS (*B*) in nonsense alleles on the distance along the CDS (introns excluded) from the nonsense mutation. Black lines show the linear trend with 95% CI indicated by shading.

two categories of sites (0.0016 and 0.0019, respectively, χ^2 test, P value = 0.9973).

Table 2 presents data on pN/pS ratios for polymorphisms nested within nonsingleton nonsense alleles separately for one-exon genes, for exons of multiexon genes that do not carry a nonsense mutation, and for exons of multiexon genes that carry a nonsense mutation.

Figure 5 shows how the values of pN and pS, calculated for nested polymorphisms only, depend on the distance from the premature stop codon along the coding sequence of the gene. pN decreases with distance from the stop codon (the slope of the linear trend: -2.3×10^{-7} [95% CI: -3.9×10^{-7} to -5.9×10^{-8}], $P = 0.01787$), while pS does not change (the slope is 3.9×10^{-8}, with 95% CI: -3.8×10^{-7} to 4.6×10^{-7}, $P = 0.8186$). We also investigated the dependence of pN and pS of a gene on the rate of recombination in it (Fiston-Lavier et al. 2010), and did not find any statistically significant relationships (data not reported).

Discussion

Our main goal was to determine whether nonsense mutation-carrying alleles of genes retain some residual function and, thus, remain under some negative selection. We have found that within the exon of a nonsense allele which carries a premature stop codon, the pN/pS ratio, calculated on the basis of missense and synonymous mutations that are nested within the nonsense allele and therefore likely appeared after the nonsense mutation, is not significantly different from 1, indicating total relaxation of selection (table 2). For other exons of nonsense alleles, we obtained a slightly lower value of pN/pS, which, however, is not significantly different from the first one or from 1.

pN, but not pS, declines with the distance from a premature stop codon (fig. 5). This contrast suggests that selection plays a role in the decline of pN. There could be two not mutually exclusive causes for this pattern. First, residual negative selection probably operates on exons that do not carry a premature stop codon, because some of them are not incorporated into all isoforms produced by alternative splicing. Second, even if negative selection is absent throughout the whole nonsense mutation-carrying allele, the observed effect could be due to its recombination with functional alleles depleted of missense substitutions. Unfortunately, our data are insufficient to discriminate between these two possibilities, although the lack of dependency of pN of a gene on its rate of recombination may be interpreted as favoring the alternative splicing explanation.

Nonsense alleles are mostly rare, so that missense and synonymous mutations nested within them are even rarer. In fact, all such mutations that are present in the data we used are singletons. Because negative selection leads to an excess of rare alleles, pN/pS ratios calculated on the basis of such mutations must be inflated, even without any relaxation of selection. We investigated this by calculating the pN/pS ratios for mutations nested within frequency-matched synonymous alleles with those nested within nonsense alleles. As expected, the average of these ratios (0.611 [0.556–0.770]) is much higher than for all mutations; however, it was still significantly <1. Thus, relaxation of negative selection acting on nonsense alleles of genes appears to be real.

Analyses of small samples of genotypes from two other populations produced results similar to those reported above, but with even wider confidence intervals (supplementary tables 4 and 5, Supplementary Material online). Obviously, it would be very desirable to analyze a much larger sample of genotypes. A data set of $>1{,}000$ Drosophila genotypes is available (Lack et al. 2015), but they are of multiple geographical origins, so that using the values of pN/pS from this data set as proxies for negative selection is problematic. Unfortunately, the already available massive data on human diploid genotypes are not easy to use for our purposes, because it is not always possible to distinguish maternal and paternal sequences. Hopefully, larger sets of haploid genotypes from the same population, which in the case of *Drosophila* can be obtained either from haploid embryos or from inbred lines, will soon become available.

Overall, we investigated the possibility of some residual negative selection acting on nonsense alleles of protein-coding genes of *D. melanogaster*. Our results are consistent with complete relaxation of selection within those exons that carry premature stop codons. However, there may be some weak negative selection within other exons, possible due to alternative splicing of the nonsense-containing exon.

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

We thank Anna Klepikova and Michail Schelkunov and all laboratory team for helpful comments and useful discussions for this article. This work was supported by the Russian Science Foundation (grant number 14-50-00150).

Literature Cited

Fiston-Lavier AS, Singh ND, Lipatov M, Petrov DA. 2010. *Drosophila melanogaster* recombination rate calculator. Gene 463(1–2):18–20.

Flowers JM, et al. 2015. Whole-genome resequencing reveals extensive natural variation in the model green alga *Chlamydomonas reinhardtii*. Plant Cell 27(9):2353–2369.

Groenen MAM, et al. 2012. Analyses of pig genomes provide insight into porcine demography and evolution. Nature 491(7424):393–398.

Jagannathan S, Bradley RK. 2016. Translational plasticity facilitates the accumulation of nonsense genetic variants in the human population. Genome Res. 26(12):1639.

Kimura M. 1983. The neutral theory of molecular evolution. Cambridge: Cambridge University Press.

Kono TJY, et al. 2016. The role of deleterious substitutions in crop genomes. Mol Biol Evol. 33(9):2307–2317.

Krawczak M, Ball EV, Cooper DN. 1998. Neighboring-nucleotide effects on the rates of germ-line single-base-pair substitution in human genes. Am J Hum Genet. 63(2):474–488.

Lack JB, et al. 2015. The drosophila genome nexus: a population genomic resource of 623 *Drosophila melanogaster* genomes, including 197 from a single ancestral range population. Genetics 199(4):1229–1241.

Lee YCG, Reinhardt JA. 2012. Widespread polymorphism in the positions of stop codons in *Drosophila melanogaster*. Genome Biol Evol. 4(4):533–549.

Li AH, et al. 2015. Analysis of loss-of-function variants and 20 risk factor phenotypes in 8, 554 individuals identifies loci influencing chronic disease. Nat Genet. 47:1–5.

Li H, Stephan W. 2006. Inferring the demographic history and rate of adaptive substitution in Drosophila. PLoS Genet. 2(10):e166.

MacArthur DG, et al. 2007. Loss of ACTN3 gene function alters mouse muscle metabolism and shows evidence of positive selection in humans. Nat Genet. 39(10):1261–1265.

MacArthur DG, et al. 2012. A systematic survey of loss-of-function variants in human protein-coding genes. Science 335(6070):823–828.

Mackay TF, et al. 2012. The Drosophila melanogaster Genetic Reference Panel. Nature 482(7384):173–178.
Mort M, Ivanov D, Cooper DN, Chuzhanova NA. 2008. A meta-analysis of nonsense mutations causing human genetic disease. Hum Mutat. 29(8):1037–1047.
Prieto-Godino LL, et al. 2016. Olfactory receptor pseudo-pseudogenes. Nature 539(7627):93–97.
von Heckel K, Stephan W, Hutter S. 2016. Canalization of gene expression is a major signature of regulatory cold adaptation in temperate *Drosophila melanogaster*. BMC Genomics 17:574.
Wetterbom A, Gyllensten U, Cavelier L, Bergström TF. 2009. Genome-wide analysis of chimpanzee genes with premature termination codons. BMC Genomics 10:56.
Wright S. 1931. Evolution in mendelian populations. Genetics 16(2):97–159.
Xue Y, et al. 2006. Spread of an inactive form of caspase-12 in humans is due to recent positive selection. Am J Hum Genet. 78(4):659–670.
Yamaguchi-Kabata Y, et al. 2008. Distribution and effects of nonsense polymorphisms in human genes. PLoS One 3(10):e3393.
Yang H, et al. 2015. Expression profile and gene age jointly shaped the genome-wide distribution of premature termination codons in a *Drosophila melanogaster* population. Mol Biol Evol. 32(1):216–228.

Associate editor: Paul Sharp