Haploid selection drives new gene male germline expression

Julia B. Raices,[1,2] Paulo A. Otto,[1] and Maria D. Vibranovski[1]

[1] *Department of Genetics and Evolutionary Biology, Institute of Biosciences, University of São Paulo, São Paulo, Brazil, 05508-090*

New genes are a major source of novelties, and a disproportionate amount of them are known to show testis expression in later phases of male gametogenesis in different groups such as mammals and plants. Here, we propose that this enhanced expression is a consequence of haploid selection during the latter stages of male gametogenesis. Because emerging adaptive mutations will be fixed faster if their phenotypes are expressed by haploid rather than diploid genotypes, new genes with advantageous functions arising during this unique stage of development have a better chance to become fixed. To test this hypothesis, expression levels of genes of differing evolutionary age were examined at various stages of *Drosophila* spermatogenesis. We found, consistent with a model based on haploid selection, that new *Drosophila* genes are both expressed in later haploid phases of spermatogenesis and harbor a significant enrichment of adaptive mutations. Additionally, the observed overexpression of new genes in the latter phases of spermatogenesis was limited to the autosomes. Because all male cells exhibit hemizygous expression for X-linked genes (and therefore effectively haploid), there is no expectation that selection acting on late spermatogenesis will have a different effect on X-linked genes in comparison to initial diploid phases. Together, our proposed hypothesis and the analyzed data suggest that natural selection in haploid cells elucidates several aspects of the origin of new genes by explaining the general prevalence of their testis expression, and a parsimonious solution for new alleles to avoid being lost by genetic drift or pseudogenization.

[Supplemental material is available for this article.]

New gene origination is thought to play a major role in phenotypic evolution (Kaessmann 2010; Chen et al. 2013). New genes may originate from several different mechanisms such as DNA-based duplication, retrotransposition, and de novo origination (for review, see Long et al. 2013). Most emerged genes are lost because of their low fixation probability in large populations or by accumulation of deleterious mutations during the process of becoming pseudogenes (Ohno 1970; Lynch and Conery 2000). Pseudogenization of new genes is especially widespread for duplications, the most common mechanism of gene origination, which often generates an exact same copy of a parental gene and performs a redundant function (Ohno 1970; Lynch and Conery 2000). A model called *neofunctionalization* proposes that new structures can arise after duplication through the accumulation of adaptive mutations (Ohno 1970). Beneficial effects provided by novel functions increase the chance of a new gene to spread in a population and ultimately to be fixed.

Although neofunctionalization of new genes certainly occurs in nature (e.g., Zhang et al. 2004, 2009; Des Marais and Rausher 2008; Rosso et al. 2008; Assis and Bachtrog 2013; Long et al. 2013), there are two poorly explained aspects. On the theoretical side, a new duplicate gene must remain in a population long enough until rare beneficial mutations accumulate (Ohno 1970; Thornton and Long 2002). On the empirical side, there is a strong bias for new gene expression in the male germline in a wide range of organisms such as *Drosophila*, mosquitos, plants, and mammals, including humans (e.g., Betrán et al. 2002; Levine et al. 2006; Vibranovski et al. 2009; Soumillon et al. 2013; Cui et al. 2015). Although competition in male reproduction provides a rich environment for the spread of beneficial novel mutations (Parker 1970; Singh et al. 2002; Manier et al. 2010), recent work argues that new gene testis expression is driven by mechanistic bias (Soumillon et al. 2013). Spermatogenesis usually occurs in three phases along the testis: (1) a diploid mitotic phase that increases cell numbers and size; (2) a meiotic phase characterized by intense transcriptional activity and by the reduction of the DNA amount (*2n* to *n*) in which diploid cells become haploid; (3) a post-meiotic phase in which differentiation and individualization of sperm cells takes place as well as changes in DNA packing proteins (Kimmins and Sassone-Corsi 2005). In mouse, rice, and *Arabidopsis*, new genes tend to be expressed in later phases of male gametogenesis, encompassing meiosis and post-meiosis (Soumillon et al. 2013; Cui et al. 2015). The transcriptionally permissive state of the chromatin during late meiosis and post-meiosis can explain why new genes are more frequently expressed in later phases of spermatogenesis as soon as they originate (Soumillon et al. 2013) but does not necessarily explain how those genes become fixed unless subsequent genetic drift or selection are involved. Although positive selection signature has been found several times for testis-expressed new genes (e.g., Zhang et al. 2009, 2010; Jiang and Assis 2017), the major support for the mechanistic-drift neutral hypothesis comes from the increased expression of both retrogenes and retropseudogenes during mouse late sperm development (Soumillon et al. 2013). Because pseudogenes are nonfunctional, the origin of new genes seems to be a neutral process, shaped by chromatin open state and genetic drift (Soumillon et al. 2013).

[2]Present address: Institute of Science and Technology Austria, 3400 Vienna, Austria
Corresponding author: mdv@ib.usp.br
Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.238824.118.

Drosophila genome composition reflects the milder effects of genetic drift, typically observed for species with population sizes much larger than those of mammals (Kreitman 1983; Lozovskaya et al. 1999). Because natural selection is more relevant to the genomic history of *Drosophila* than that of mammals, the mechanistic biased hypothesis coupled with neutral evolution per se is less likely to completely explain why new genes are expressed in the male germline in a wide range of organisms and therefore other hypotheses should be developed.

Results

Drosophila new genes: spermatogenic expression and positive selection signature

To elucidate the contributions of selection and neutral evolution on the origin of new genes, we analyzed how new *D. melanogaster* genes are expressed along spermatogenesis. In agreement with the data observed in plants and mammals (Fig. 1A), new genes in *Drosophila* are more frequently and highly expressed in later phases of male gamete development, whereas old genes show significantly higher expression and enrichment in mitosis (Fig. 1B). The pattern is consistent for genes originated by the three major mechanisms: DNA-based and RNA-based duplications and de novo origination.

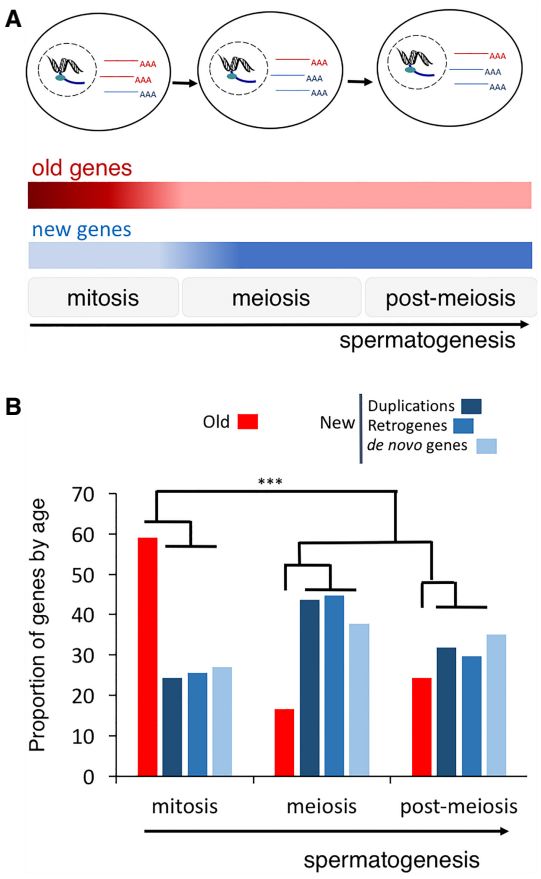

Figure 1. Spermatogenic expression profiles of new and old genes. (*A*) Old genes (red) are expected to be mainly expressed in the diploid phases of spermatogenesis, where mitosis occurs; new genes (blue) are expected to be more highly expressed in later phases of spermatogenesis, meiosis, and post-meiosis. (*B*) Relative proportion calculated separately for old and new *D. melanogaster* genes overexpressed in mitosis, meiosis, or post-meiosis (Methods). Total sample sizes for old and new genes were 5170 and 260, respectively. (***) $P<0.001$, Fisher's exact test. X-linked genes were excluded from these analyses to avoid effects from X-Chromosome inactivation taking place during *Drosophila* male meiosis (Vibranovski et al. 2009).

Although natural selection is important in *Drosophila* evolution in general, late-stage expression of new genes of *D. melanogaster* is also compatible with neutral evolution and permissive transcriptional environment (Soumillon et al. 2013). New genes in general, including those testis-biased expressed, display higher rates of nonsynonymous per synonymous substitutions on their sequences (d_N/d_S) (Supplemental Fig. S1) and significant enrichment with positive selection signature when compared to old genes, together indicating adaptive evolution (Supplemental Table S1; Zhang et al. 2010). By grouping autosomal genes with higher expression in meiosis and post-meiosis as our haploid set, we found an increased ratio of nonsynonymous to synonymous substitutions which at first is consistent with the hypothesis of lack of constraint/transcriptional permissiveness. However, the haploid set of genes is as well enriched with positive selection signature in comparison to genes expressed in diploid cells, the mitotic phase (Fig. 2A; Table 1). Hence, the enrichment of nonsynonymous to synonymous substitutions does not solely result from lack of constraint. Moreover, new haploid genes present a higher proportion of nonsynonymous substitutions over nonsynonymous polymorphisms in comparison to diploid-expressed new genes (α: 0.23 vs. 0.16, $P=0.002$, $n=248$, χ^2 test). Together this indicates that natural selection is involved in the fixation of new genes. If mechanistic bias created by an open chromatin state in later phases of spermatogenesis, followed by genetic drift, has driven the fixation of new genes (Soumillon et al. 2013), we would not observe such elevated rates of positive selection on the sequences of haploid-expressed new genes.

Haploid selection for the evolution of new genes

Alternatively, we propose here that haploid selection, rather than mechanistic bias/genetic drift, taking place in later phases of male gametogenesis would lead to higher rates of fixation of new genes. It is well known that the haploid state will rapidly expose a low frequency adaptive mutation, conferring immediate advantage to the organisms carrying it (Joseph and Kirkpatrick 2004), whereas recessive adaptive mutations would be hidden from natural selection in a heterozygous genotype, preventing or delaying their increase in frequency and their fixation (Fig. 2B; Otto et al. 2015). In other words, adaptive mutations have their frequency increased faster if their phenotypes are produced by a haploid rather than a diploid genotype.

Note that adaptive recessive alleles in any gene would benefit from any haploid selection taking place in the gametogenesis (Joseph and Kirkpatrick 2004). In the case of duplicate genes, adaptive mutations, frequently recessive (Thornton and Long 2002), are known to accumulate during the neofunctionalization process of a duplicated gene (Ohno 1970; Long et al. 2013). We rationalize that haploid selection boosts new gene evolution by further increasing their fixation probability avoiding their loss or pseudogenization. Indeed, if new genes are being driven to fixation primarily owing to selection when expressed in haploid cells, a likely correlate of this is that they are expressed in meiosis and post-meiosis at a high rate. Another correlate is that they present an excess of adaptive sequence signature as observed in

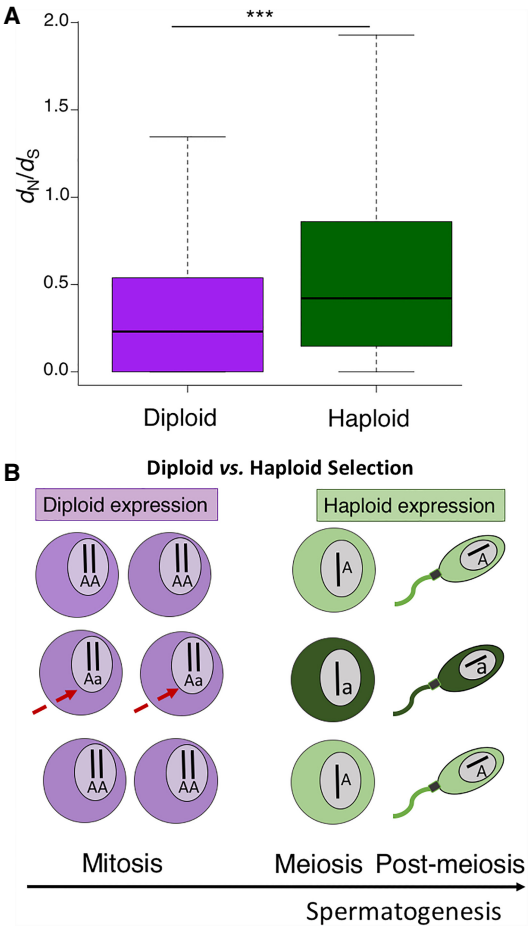

Figure 2. Evolutionary analyses of *D. melanogaster* autosomal genes and systems of selection. (*A*) The ratios of nonsynonymous substitution rate to synonymous substitution rate (d_N/d_S) across different gene expression groups. Diploid (purple) and haploid (green) groups were considered those genes overexpressed in mitosis, and meiosis or post-meiosis, respectively (Methods). Total sample sizes of diploid and haploid genes were 2348 and 2212, respectively, for d_N/d_S. Mann-Whitney-Wilcoxon test was used as the statistical test: (***) $P < 0.001$. (*B*) Haploid selection advantage over diploid selection is shown as haploid cells (green) carrying new allele *a* immediately present a new phenotype (dark green). However, for diploid cells (purple), heterozygote cells do not present a different phenotype (all light purple). The new allele *a* effect is hidden by the dominant allele *A* (dashed red arrow).

Drosophila data. Old genes, in contrast, have accumulated other functions along evolution, and hence, present a varied expression profile (Zhang et al. 2012) that erases their previous pattern of haploid expression enrichment and signature of positive selection.

Although the haploid selection is presented as a parsimonious explanation for the observed data, there are empirical and theoretical aspects that must be addressed before assuming this hypothesis. On the developmental aspect, later stages of gametogenesis represent the only haploid cells in an extensive list of diploid organisms. However, selection in female haploid stage is not a prominent candidate to boost new gene evolution. In mammalian egg development, the haploid phase is reduced in space and time because meiosis II is only triggered by sperm fertilization (Otto et al. 2015), mostly hampering any role for haploid selection in the evolution of new genes. Likewise, in *D. melanogaster* oogenesis, haploid selection is mostly out of consideration. Only one of the 16 cells from each cyst becomes a viable ovule, whereas the other cells will nourish it by exporting/pumping their entire cytoplasm with organelles, nutrients, and mRNAs to the oocyte (Cáceres and Nilson 2005; He et al. 2011). Through ring canals, the 15 nurse cells supply essential components to the oocyte, which has a nucleus that is mostly transcriptionally quiescent and arrested in meiotic prophase (King and Burnett 1959; Spradling 1993).

In males, all spermatids become individual spermatozoa with independent function and fitness (e.g., Alavioon et al. 2017) and therefore are, in theory, excellent candidates to boost new gene evolution by haploid selection. However, in spermatogenesis of *Drosophila* and of other organisms, RNA products can be shared/leaked among cells across cytoplasmic bridges because cytokinesis within cysts is not complete (Dadoune et al. 2004; Joseph and Kirkpatrick 2004). Therefore, genetically haploid spermatids were for a long time thought to be phenotypically diploid (Braun et al. 1989). However, recent evidence has given support to haploid selection in male gametogenesis bearing cytoplasmic bridges. First, zebrafish harbor differential fitness in a single ejaculation owing to genetic differences among individual sperms from the same individual. Second, compartmentalization of the mouse Sperm Adhesion Molecule I mRNA proved, for at least some transcripts, the lack of, or at least the incomplete sharing of mRNAs through cytoplasmic bridges (Martin-DeLeon et al. 2005).

In agreement with this developmental aspect, new genes are usually not found to be expressed in ovaries in different taxa (e.g., Zhang et al. 2010) reinforcing the higher chances of fixation for new genes bearing expression in meiotic and post-meiotic phases of male gametogenesis (Soumillon et al. 2013; Cui et al. 2015) and not in the female one. Haploid selection in males is not an unlikely phenomenon as in oogenesis where mRNA and oocyte nutrients come from the mixed cytoplasm of neighboring diploid cells, which therefore contribute in the same way to a single ovule fitness and function.

Regarding our haploid selection hypothesis, any RNA leakage occurring between spermatids impacts only the degree of dominance between alleles. Incomplete level of dominance is a proxy of the fitness effects on recessive alleles transcribed in haploid cells with mRNA partially shared through the cytoplasmic bridges. Moreover, one can argue about the recessive nature of adaptive mutations on new genes. It is difficult to imagine gene duplication per se as recessive because one haplotype carries an extra copy. However, the subsequent evolution of a new gene may very likely involve recessive mutations as shown by positive selection signature enriched in X-linked new genes in *Drosophila* (Thornton and Long 2002). Therefore, by varying the degrees of dominance, as done in the population genetic model described in the following section, it is possible to measure how much faster new beneficial alleles could be fixed in the haploid population in general.

Table 1. Positive selection on diploid- and haploid-expressed genes (*D. melanogaster*)

Positive Darwinian selection[a]	Diploid-expressed genes (%)	Haploid-expressed genes (%)
Detected	35 (1.4)	72 (3.5)
Nondetected	2425	2006

[a]Based on flyDIVaS database (Stanley and Kulathinal 2016); Fisher's exact test, $P < 0.001$.

Population genetic model for haploid selection in the evolution of new genes

The population genetic model that follows aims to compare the effects of positive selection in new alleles acting in the haploid phase (gamete) with the corresponding one taking place during the diploid phase. We intend to understand the effect of two major aspects which, as mentioned previously, are crucial for the evolution of new genes: (1) degree of dominance of the adaptive mutation/allele; and (2) genetic drift. Our population genetics model first considers an autosomal locus with alleles A and a that are expressed in diploid phase as well as in haploid phase. In a deterministic model, in which the *Drosophila* population size is considerably large, mating occurs randomly after Hardy-Weinberg ratios and the effects of genetic drift are considered as negligible. The model assumes (1) that s_1 ($0 \le s_1 \le 1$) and hs_1 ($0 \le h \le 1$) are the respective coefficients of selection of genotypes AA and Aa, while 1 is the relative fitness value of genotype aa; that $q = q_1$ and $1-q = 1-q_1$ are the frequencies of alleles a and A; that h is a dominance measure; and that q_1' is the frequency value of the a allele in the next generation; and (2) that s_2 ($0 \le s_2 \le 1$) is the coefficient of selection of A gametes and 1 the relative fitness of gametes carrying the a allele; $q = q_2$ and $1-q = 1-q_2$ are the population frequencies of a and A gametes; and that q_2' is the frequency value of the a allele resulting from gametes that compete among themselves to form the next generation genotypes.

Our model is based on the analytical and numerical analyses of the expression

$$\frac{\Delta q_2}{\Delta q_1} = \frac{\{1-(1-q)sx[1-q(1-2h)]\}}{\{x[1-(1-q)s][1-h-q(1-2h)]\}},$$

where $\Delta q_2 = q_2' - q$; $\Delta q_1 = q_1' - q$; $s_2 = s$; and $s_1 = sx$, so that $s_1/s_2 = x$.

The expression $\Delta q_2/\Delta q_1$ is the incremental rate, a pertinent variable for comparing the evolutionary gain of frequency (fixation rate) of the allele a under the alternative hypotheses of positive selection acting during the haploid and diploid phases, respectively.

The analysis of the incremental rate shows that, if $0 < x \le 1$, $\Delta q_2/\Delta q_1$ is always larger than 1, irrespective of the value h (the dominance factor) can take. For small values of q (which is usually the case for new mutations in large populations), the gain in gene frequency $\Delta q_2/\Delta q_1$ (fixation rate) of the allele a in the haploid phase is much larger when $h = 1$ (aa completely recessive) than when $h = 0$ (dominant case). Extensive computer-assisted numerical analysis showed that this is also true when $h < 1/2$ for all possible combinations of q, x, and s values and that for small or very small (<0.01 or 0.001) frequency values of q the gain in gene frequency $\Delta q_2/\Delta q_1$ when $h = 1$ is about $(1-q)/q$ times larger than when $h = 0$. For example, when $q = 0.001$, the gain in gene frequency of the allele a is about 999 times larger in the case $h = 1$ than in the case $h = 0$.

When $x > 1$, the fixation rate of the a allele depends on some specific combinations of q, s, x, and h values that can be expressed through specific mathematical formulas, fully detailed in Supplemental Methods. Example of one of the results obtained from the analysis of the model is shown in Figure 3A. Therefore, new beneficial alleles could be fixed faster in the haploid population than in the diploid, even when there is incomplete dominance of the alleles. Such a result expands the conditions in which our model works. The numerical analysis of the model shows clearly that what is important to the fixation rate of the a allele is the ratio x between the selective values $s_1 = sx$ and $s_2 = s$ (of homozygous AA and gametes A, respectively) and not the value of the coefficient of selection s considered separately. Hence, the patterns observed (Fig. 3A–E) should be robust for other values of s as well.

The analysis of a large number of cases (of the order of 10^6), generated by computer-assisted random combinations of q, $s_1 = sx$, $s_2 = s$, and h, showed that, even in the nonadvantageous situation when $x > 1$, in about 38.5% of cases the rate of frequency gain (fixation rate) of the a allele is larger under the system of positive selection during the haploid phase than during the diploid phase ($\Delta q_2/\Delta q_1 > 1$). Considering that this is exactly what always takes place when $x \le 1$, we have just evidenced the importance of the mechanism of positive selection acting during the haploid phase of male gametogenesis in the process of fixation of new genes.

To consider the effects of random genetic drift, millions of diploid populations with distinct sizes were computer-simulated (Supplemental Methods). From each one out of 50 genotypic compositions obtained for each population of size N with a particular combination of the four parameters {q, s, h, x}, the frequency $q_1' = [2N(aa) + N(Aa)]/2N$ was directly estimated and used to calculate the value of $\Delta q_1 = q_1' - q$, which was then compared to $\Delta q_2 = q(1\text{-}q)s/[1-(1-q)s]$ to compute the number of times in which $\Delta q_1 > \Delta q_2$. The value of Δq_2, unlike what happened to the value of Δq_1, was directly estimated from the formula derived in the deterministic model, because the selection in the haploid model obviously results from a practically infinite number of gametes that compete among themselves to form the genotypes of the next *Drosophila* generation. Example results obtained from analyses of the model and simulations are shown in Figure 3B–E.

The percentage figures obtained in the case $x \le 1$ correspond to the function $y = e^{3.84}/N^{0.30}$ [$F_{(1,10)} = 1404.30$, $P = 0.00001$, $r^2 = 0.993$], a formula that can be directly used to obtain the percentage of cases in which the gain (entirely because of random genetic drift) in the diploid phase is larger than in the haploid phase (Fig. 3F). Even with very large population numbers, however, on average in ~5% of cases $x \le 1$ the selective gain (fixation rate of the a allele) will be larger in the diploid than in the haploid phase. In any case, and for any population number, the number of cases in which the haploid gain predominates is overwhelming despite drift, whose effects can be assumed to be negligible in the case $x \le 1$ for population sizes of the order of 400 or more. The percentage figures obtained for the case $x > 1$, on the other hand, indicate that they do not differ significantly (and independently from the population size N) from the overall value obtained in the deterministic model (around 61.5%) (Fig. 3F). We conclude therefore that random genetic drift does not interfere significantly with the dynamics of the deterministic model we described for this specific case. The complete description of all methods and procedures, the detailed derivation of all formulas we used, and extensive mathematical analyses (analytical and numerical) of the model are contained in Supplemental Methods.

Autosomes versus X Chromosomes

For organisms with male as the heterogametic sex, the X Chromosome is always hemizygous and therefore X-linked genes have only one type of allele expressed in all male cells, including the entire spermatogenesis (Fig. 4A). The Faster-X effect proposes that X-linked genes will evolve more rapidly than autosomal loci as adaptive alleles are sheltered from positive selection on autosomes but are fully expressed in males when X-linked (Charlesworth et al. 1987), an analogous effect of the haploid selection (Joseph and Kirkpatrick 2004). Therefore, X-linked mutations are expected to

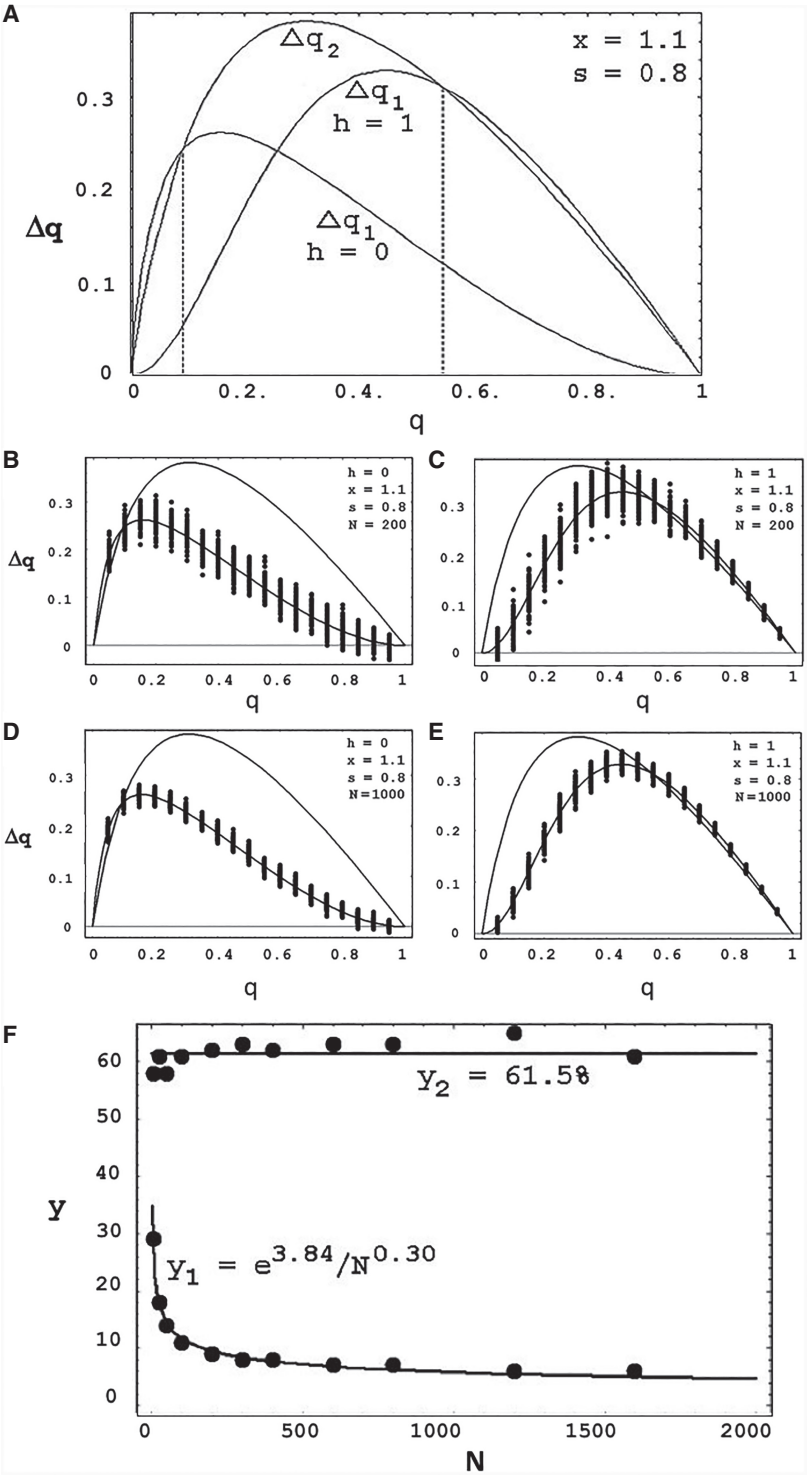

Figure 3. Simulations from the developed mathematical model. (*A*) Comparison between the values of Δq_2 and Δq_1 obtained in the deterministic model for the cases {$h=0$, $x=1.1$, $s=0.8$} and {$h=1$, $x=1.1$, $s=0.8$}. Results (case $x>1$) obtained from drift/selection simulations for cases $h=1$ (*B*,*D*) and $h=0$ (*C*,*E*) for population sizes (*N*) 200 (*B*,*C*) and 1000 (*D*,*E*), keeping the parameters as prescribed in *A* (case $x=1.1$). (*F*) The Δq_1 simulated populations are shown as black dots (around the curves representing Δq_1 in *A*). The black dots represent the percentage of cases in which the selective gain in the fixation process is larger in the diploid than in the haploid phase, because of random genetic drift depending on selection and population size *N*. The *upper* set and the line y_2 correspond to the case $x>1$, and the *lower* set and the function y_1 correspond to the alternative case $x\leq 1$. The percentage figures obtained in the case $x\leq 1$ correspond with negligible statistical error to the function $y=e^{3.84}/N^{0.30}$ [$F_{(1,10)}=1404.30$, $P=0.00001$, $r^2=0.993$].

have high fixation probabilities, especially if beneficial to males (Charlesworth et al. 1987). In support to the Faster-X hypothesis, the observation of rapid divergence of gene duplicates on the *D. melanogaster* X Chromosome (Thornton and Long 2002) indicates that adaptations at duplicate loci are recessive on average.

The haploid selection hypothesis predicts that higher expression in later phases of male germline is a frequent feature of new autosomal genes rather than of X-linked new genes. A correlate of this prediction is that new X-linked genes are expected to show elevated signature of positive selection as well as high expression in all male germline (including mitosis) and not only on later phases. Therefore, if our haploid hypothesis is correct, we then expect the relative proportion of autosomal and X-linked expressed new genes to be different among *Drosophila* spermatogenic phases (Fig. 4B). Because haploid phases are comparatively more advantageous for autosomal genes than mitotic ones and there is no differential fitness effect for X-linked genes along spermatogenesis, we observed a significantly higher proportion of autosomal new genes expressed in post-meiosis opposed to lower proportion of those expressed in mitosis (Fig. 4C). This result supports the idea that haploid expression in later phases of male germline is beneficial only for those genes located in autosomal and not in the X Chromosome.

Note that such pattern is absent for the analyzed old genes (Fig. 4B). Instead, there is a meiotic depletion of old X-linked genes, which has been previously pointed out as an effect of meiotic sex chromosome inactivation (MSCI) in *Drosophila* (Vibranovski et al. 2009). Curiously, meiotic depletion is not observed for new genes suggesting that they are not immediately susceptible to the inactivation process as they emerge.

In terms of positive selection sequence signature, we also observed a difference between genes located on the X Chromosome and autosomes. On one hand, overall new X-linked genes are enriched with positive selection signature when compared with old ones (13% vs. 3%, $P<0.02$, Fisher's exact test), a pattern not observed for autosomal new genes in general ($P=0.33$, Fisher's exact test). On the other hand, X-linked genes do not show significantly more signature of positive selection on genes expressed in the

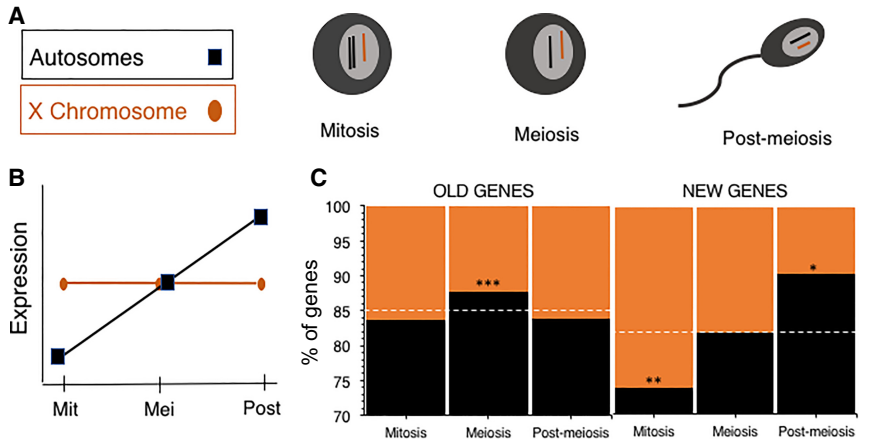

Figure 4. Expression profiles of autosomal and X-linked genes according to the haploid selection hypothesis. (*A*) Diploid and haploid states of autosomal (black) and X-linked (orange) genes along the three phases of spermatogenesis. (*B*) Relative expression prediction of autosomal and X-linked genes along male mitosis (Mit), meiosis (Mei), and post-meiosis (Post). (*C*) Relative proportion of autosomal and X-linked *D. melanogaster* genes calculated for each spermatogenic phase showed separately for old and new genes (Methods). Dashed white lines represent the average autosomal proportion among ages. Total sample sizes for autosomal and X-linked genes, respectively: 5170, 956 (old); 260, 57 (new). (***) $P < 0.001$; (**) $P < 0.01$; (*) $P < 0.05$; Fisher's exact test: one phase versus other phases grouped.

later phases of spermatogenesis in comparison to the mitotic ones ($P = 0.40$, Fisher's exact test), as observed for genes located on autosomes (Fig. 2A).

The chromosomal discrepancy on both the proportions of genes expressed in different phases and on the enrichment of positive selection corroborates that both faster evolution for the X Chromosome and haploid selection for autosomes are occurring in males and contributing to the evolution of new genes.

Discussion

We showed that *Drosophila* new genes, as it happens in mammals and plants, are more expressed in later phases of male gametogenesis. Contrary to expectations under fixation by genetic drift, new genes expressed in male meiosis and post-meiosis are enriched with positive selection signature. Additionally, only new genes located in the autosomes are found to be preferentially expressed in later phases of spermatogenesis. All three lines of evidence could be explained by competing hypotheses. On one hand, the expression bias of new genes in late stages (Fig. 1) is compatible with the neutral hypothesis (permissive transcriptional environment). On the other hand, excessive adaptive signature (Fig. 2; Table 1) on genes expressed in the late stages could be explained by sexual selection (e.g., sperm competition). Finally, the observation that autosomal new genes are underrepresented in mitosis but overrepresented in post-meiosis (Fig. 4) could be somehow a partial consequence of the gene movement out the X Chromosome because of MSCI acting on the X Chromosome in meiosis and in post-meiosis.

Here, we proposed only one model to explain all presented observations rather than recruiting three different hypotheses. Although all other models likely coexist and by acting in the spermatogenesis may also contribute to the evolution of new genes, the haploid selection model states that natural selection taking place in haploid cells elucidates several aspects of their origin. The model explains the prevalence of new gene testis expression and gives a parsimonious solution for new alleles, particularly recessive ones, to avoid being lost by genetic drifts or pseudogenization.

In addition, by comparing population genetics equations under haploid and diploid selection systems, we established the robustness of our model in the presence of genetic drift and with different degrees of dominance. Those parameters are important to consider during the evolution of new genes for the following reasons. First, new gene testis expression is a widespread phenomenon occurring in species with a large range of population size and therefore is subject to different levels of genetic drift. Second, in different spermatogenesis, RNA products can be shared among cells across cytoplasmic bridges. Therefore, if leak of expression from the dominant allele follows, new adaptive recessive mutations might be sheltered, or at least partially hidden, from positive selection regardless of the system. Third, we do not know whether the nature of new genes is likely to be recessive or whether subsequent evolution always involves recessive mutations. Varying those parameters in our simulations revealed that even if the selection coefficient is higher in the diploid cell population, the fixation can still be faster in the haploid population in cases in which dominance is not complete and in the presence of genetic drift, therefore further strengthening the theoretical side of our hypothesis.

To sum it up, our study shed light on the importance of haploid phase, which has been neglected for many years as a potential contributor to the causes of male fertility, because it can now be directly related to novelties in morphogenesis, motility, and sperm fertility. Moreover, haploid selection as a main player on the origin of new genes places them as prominent candidates to provide valuable information on new biological pathways and functions that may affect the sperm.

Nevertheless, it is important to acknowledge that a quarter of new genes are also overexpressed in male mitosis and therefore might have been fixed by positive selection in diploid phases. Further studies might help to understand other processes responsible for the fixation of new genes that are expressed in diploid cells and involve completely dominant alleles.

Methods

Spermatogenesis stage-specific expression analyses

For expression analyses, we used microarray data for each spermatogenesis phase (Vibranovski et al. 2009) from dissected parts of *D. melanogaster* testis enriched with cells from mitosis, meiosis, and post-meiosis. Each gene had the expression compared between a pair of phases (mitotic and meiotic phase, meiotic and post-meiotic, mitotic and post-meiotic) and categorized as underexpressed, overexpressed, as well as equally expressed in each pairwise comparison, according to a Bayesian statistical model (Vibranovski et al. 2009). Using these gene categories, it was possible to create expression classes representing all the genes and how their expression

occurs throughout the three spermatogenic phases (Supplemental Fig. S2). To evaluate the haploid fitness, it is important to investigate genes that have been transcribed by haploid cells rather than analyze mRNAs transcribed in previous diploid phases and stored in the cytoplasm (Schäfer et al. 1995). Because spermatogenesis is a developmental and temporal biological process, using genes more expressed in later phases than early ones guarantees haploid transcription and therefore association with haploid selection (Vibranovski et al. 2009, 2010). Because crossing over does not occur in *Drosophila* males, we considered meiosis to be in haploid state as homologous chromosomes are separated in its first cell division (Meiosis I), so a recessive allele is no longer masked by a dominant one for the sake of fitness.

Gene age classification

To analyze the gene ages, we used data from Zhang et al. (2010) in which they dated the origination of *D. melanogaster* genes by accessing ortholog genes presence and absence in close *Drosophila* species. Accordingly, genes that originated more than 63 million years ago were considered old genes, like the genes that are present in both representatives of the subgenus *Sophophora* and the subgenus *Drosophila* (Russo et al. 1995). Genes younger than 63 million years were considered new genes.

Evolutionary analyses

Data of genes d_N (nonsynonymous substitutions), d_S (synonymous substitutions), p_N (nonsynonymous polymorphisms), and p_S (synonymous polymorphisms) were obtained from Zhang et al (2010) by inferring those parameters from sequence comparisons between *D. melanogaster* and *Drosophila simulans*. Ratio distributions of d_N/d_S were compared using the Mann-Whitney-Wilcoxon test.

We estimated α, the proportion of substitutions fixed by adaptive mutation, in new genes by implementing a multilocus McDonald–Kreitman test using Distribution of Fitness Effects (DoFE) (Bierne and Eyre-Walker 2004). As in Zhang et al. (2010), we used the LikeLihood-Ratio test (LLR) to measure whether haploid genes have different α compared to diploid ones. We performed a χ^2 test in which the null hypothesis was that α was the same in the two group of genes as described in Zhang et al. (2010).

We used a comparative genomics resource for *Drosophila* divergence and selection, the flyDIVaS database (Stanley and Kulathinal 2016) to retrieve codon-based tests of positive Darwinian selection for our group of genes. Using orthologous comparisons between sequence from the *D. melanogaster* subgroup, the database test positive selection for each gene using three different selection models and correcting for multiple tests with False Discovery Rates (FDR). We considered a gene as evolving under positive selection if significant difference from the neutral hypothesis was found for at least one model tested. Most of our results maintained the same applying only flyDIVaS's first model (Stanley and Kulathinal 2016). New genes originated in *D. melanogaster* or in *D. melanogaster* subgroup ($n = 126$; branches 5 and 6 from Zhang et al. 2010) were not included for flyDIVaS's analyses because they have no orthologous counterparts. However, they correspond to <15% of the total number of new genes analyzed in the rest of the study.

Chromosomal location analyses

Autosomal and X Chromosome gene location was obtained from tables available in Vibranovski et al. (2009) and confirmed in Zhang et al. (2010). Relative proportions were compared using Fisher's exact test.

Acknowledgments

This work was supported by grants from Fundação de Amparo à Pesquisa do Estado de São Paulo (IC 2013/06117-7; MS 2014/16899-5; JP 2015/20844-4 e CEPID 2013/08028-1). M.D.V. was the recipient of a research productivity fellowship from Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq: 307863/2014-0). We thank Renan B. Lemes (IB USP) for his help in preparing the computer programs we used, Ms. Lilian Dluhosch for reviewing the English usage, and Steve Dorus for helping to organize gene classes in expression groups. We also thank Manyuan Long, Timothy L. Karr, Margarida Cardoso-Moreira, Bernardo de Carvalho, Gustavo França, and Angela Vianna-Morgante for comments and input. The useful comments, suggestions, and corrections from three referees are also gratefully acknowledged.

Author contributions: M.D.V. conceived and supervised the study. J.B.R. and M.D.V. formulated the hypothesis and performed the analyses. P.A.O. developed the mathematical model. J.B.R., P.A.O., and M.D.V. wrote the manuscript.

References

Alavioon G, Hotzy C, Nakhro K, Rudolf S, Scofield DG, Zajitschek S, Maklakov AA, Immler S. 2017. Haploid selection within a single ejaculate increases offspring fitness. *Proc Natl Acad Sci* **114:** 8053–8058. doi:10.1073/pnas.1705601114

Assis R, Bachtrog D. 2013. Neofunctionalization of young duplicate genes in *Drosophila*. *Proc Natl Acad Sci* **110:** 17409–17414. doi:10.1073/pnas.1313759110

Betrán E, Thornton K, Long M. 2002. Retroposed new genes out of the X in *Drosophila*. *Genome Res* **12:** 1854–1859. doi:10.1101/gr.6049

Bierne N, Eyre-Walker A. 2004. Genomic rate of adaptive amino acid substitution in *Drosophila*. *Mol Biol Evol* **21:** 1350–1360. doi:10.1093/molbev/msh134

Braun RE, Behringer RR, Peschon JJ, Brinster RL, Palmiter RD. 1989. Genetically haploid spermatids are phenotypically diploid. *Nature* **337:** 373–376. doi:10.1038/337373a0

Cáceres L, Nilson LA. 2005. Production of *gurken* in the nurse cells is sufficient for axis determination in the *Drosophila* oocyte. *Development* **132:** 2345–2353. doi:10.1242/dev.01820

Charlesworth B, Coyne JA, Barton NH. 1987. The relative rates of evolution of sex chromosomes and autosomes. *Am Nat* **130:** 113–146. doi:10.1086/284701

Chen S, Krinsky BH, Long L. 2013. New genes as drivers of phenotypic evolution. *Nat Rev Genet* **14:** 645–660. doi:10.1038/nrg3521

Cui X, Lv Y, Chen M, Nikoloski Z, Twell D, Zhang D. 2015. Young genes out of the male: an insight from evolutionary age analysis of the pollen transcriptome. *Mol Plant* **8:** 935–945. doi:10.1016/j.molp.2014.12.008

Dadoune JP, Siffroi JP, Alfonsi MF. 2004. Transcription in haploid male germ cells. *Int Rev Cytol* **237:** 1–56. doi:10.1016/S0074-7696(04)37001-4

Des Marais DL, Rausher MD. 2008. Escape from adaptive conflict after duplication in an anthocyanin pathway gene. *Nature* **454:** 762–765. doi:10.1038/nature07092

He L, Wang X, Montell DJ. 2011. Shining light on *Drosophila* oogenesis: live imaging of egg development. *Curr Opin Genet Dev* **21:** 612–619. doi:10.1016/j.gde.2011.08.011

Jiang X, Assis R. 2017. Natural selection drives rapid functional evolution of young *Drosophila* duplicate genes. *Mol Biol Evol* **34:** 3089–3098. doi:10.1093/molbev/msx230

Joseph SB, Kirkpatrick M. 2004. Haploid selection in animals. *Trends Ecol Evol* **19:** 592–597. doi:10.1016/j.tree.2004.08.004

Kaessmann H. 2010. Origins, evolution, and phenotypic impact of new genes. *Genome Res* **20:** 1313–1326. doi:10.1101/gr.101386.109

Kimmins S, Sassone-Corsi P. 2005. Chromatin remodelling and epigenetic features of germ cells. *Nature* **434:** 583–589. doi:10.1038/nature03368

King RC, Burnett RG. 1959. Autoradiographic study of uptake of tritiated glycine, thymidine, and uridine by fruit fly ovaries. *Science* **129:** 1674–1675. doi:10.1126/science.129.3364.1674

Kreitman M. 1983. Nucleotide polymorphism at the alcohol dehydrogenase locus of *Drosophila melanogaster*. *Nature* **304:** 412–417. doi:10.1038/304412a0

Levine MT, Jones CD, Kern AD, Lindfors HA, Begun DJ. 2006. Novel genes derived from noncoding DNA in *Drosophila melanogaster* are frequently X-linked and exhibit testis-biased expression. *Proc Natl Acad Sci* **103:** 9935–9939. doi:10.1073/pnas.0509809103

Long M, VanKuren NW, Chen S, Vibranovski MD. 2013. New gene evolution: Little did we know. *Annu Rev Genet* **47:** 307–333. doi:10.1146/annurev-genet-111212-133301

Lozovskaya ER, Nurminsky DI, Petrov DA, Hartl DL. 1999. Genome size as a mutation-selection-drift process. *Genes Genet Syst* **74:** 201–207. doi:10.1266/ggs.74.201

Lynch M, Conery JS. 2000. The evolutionary fate and consequences of duplicate genes. *Science* **290:** 1151–1155. doi:10.1126/science.290.5494.1151

Manier MK, Belote JM, Berben KS, Novikov D, Stuart WT, Pitnick S. 2010. Resolving mechanisms of competitive fertilization success in *Drosophila melanogaster*. *Science* **328:** 354–357. doi:10.1126/science.1187096

Martin-DeLeon PA, Zhang H, Morales CR, Zhao Y, Rulon M, Barnoski BL, Chen H, Galileo DS. 2005. Spam1-associated transmission ratio distortion in mice: elucidating the mechanism. *Reprod Biol Endocrinol* **3:** 32. doi:10.1186/1477-7827-3-32

Ohno S. 1970. *Evolution by gene duplication*. Springer, New York.

Otto SP, Scott MF, Immler S. 2015. Evolution of haploid selection in predominantly diploid organisms. *Proc Natl Acad Sci* **112:** 15952–15957. doi:10.1073/pnas.1512004112

Parker GA. 1970. Sperm competition and its evolutionary consequences in the insects. *Biol Rev Cambridge Philosophic Soc* **45:** 525–567. doi:10.1111/j.1469-185X.1970.tb01176.x

Rosso L, Marques AC, Reichert AS, Kaessmann K. 2008. Mitochondrial targeting adaptation of the hominoid-specific glutamate dehydrogenase driven by positive Darwinian selection. *PLoS Genet* **4:** e1000150. doi:10.1371/journal.pgen.1000150

Russo CA, Takezaki N, Nei M. 1995. Molecular phylogeny and divergence times of drosophilid species. *Mol Biol Evol* **12:** 391–404. doi:10.1093/oxfordjournals.molbev.a040214

Schäfer M, Nayernia K, Engel W, Schäfer U. 1995. Translational control in spermatogenesis. *Dev Biol* **172:** 344–352. doi:10.1006/dbio.1995.8049

Singh SR, Singh BN, Hoenigsberg HF. 2002. Female remating, sperm competition and sexual selection in *Drosophila*. *Genet Mol Res* **1:** 178–215.

Soumillon M, Necsulea A, Weier M, Brawand D, Zhang X, Gu H, Barthès P, Kokkinaki M, Nef S, Gnirke A, et al. 2013. Cellular source and mechanisms of high transcriptome complexity in the mammalian testis. *Cell Rep* **3:** 2179–2190. doi:10.1016/j.celrep.2013.05.031

Spradling AC. 1993. Developmental genetics of oogenesis. In *The development of* Drosophila melanogaster (ed. Bate M, Martinez Arias A), Vol. 1, pp. 1–70. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Stanley CE Jr, Kulathinal RJ. 2016. *flyDIVaS*: a comparative genomics resource for *Drosophila* divergence and selection. *G3 (Bethesda)* **6:** 2355–2363. doi:10.1534/g3.116.031138

Thornton K, Long M. 2002. Rapid divergence of gene duplicates on the *Drosophila melanogaster* X chromosome. *Mol Biol Evol* **19:** 918–925. doi:10.1093/oxfordjournals.molbev.a004149

Vibranovski MD, Lopes HL, Karr TL, Long M. 2009. Stage-specific expression profiling of *Drosophila* spermatogenesis suggests that meiotic sex chromosome inactivation drives genomic relocation of testis-expressed genes. *PLoS Genet* **5:** e1000731. doi:10.1371/journal.pgen.1000731

Vibranovski MD, Chalopin DS, Lopes HL, Long M, Karr TL. 2010. Direct evidence for postmeiotic transcription during *Drosophila melanogaster* spermatogenesis. *Genetics* **186:** 431–433. doi:10.1534/genetics.110.118919

Zhang J, Dean AM, Brunet F, Long M. 2004. Evolving protein functional diversity in new genes of *Drosophila*. *Proc Natl Acad Sci* **101:** 16246–16250. doi:10.1073/pnas.0407066101

Zhang Y, Lu S, Zhao S, Zheng X, Long M, Wei L. 2009. Positive selection for the male functionality of a co-retroposed gene in the hominoids. *BMC Evol Biol* **9:** 252. doi:10.1186/1471-2148-9-252

Zhang YE, Vibranovski MD, Krinsky BH, Long M. 2010. Age-dependent chromosomal distribution of male-biased genes in *Drosophila*. *Genome Res* **20:** 1526–1533. doi:10.1101/gr.107334.110

Zhang YE, Landback P, Vibranovski M, Long M. 2012. New genes expressed in human brains: implications for annotating evolving genomes. *Bioessays* **34:** 982–991. doi:10.1002/bies.201200008

Received June 8, 2018; accepted in revised form May 31, 2019.