The Evolution of Small Insertions and Deletions in the Coding Genes of *Drosophila melanogaster*

Zechen Chong,[†,1,2] Weiwei Zhai,[†,‡,1,3] Chunyan Li,[1] Min Gao,[1] Qiang Gong,[1] Jue Ruan,[1] Juan Li,[1] Lan Jiang,[1] Xuemei Lv,[1] Eric Hungate,[4] and Chung-I Wu*[,1,4]

[1]Center for Computational Biology and Laboratory of Disease Genomics and Individualized Medicine, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing, China

[2]Graduate University of Chinese Academy of Sciences, Beijing, China

[3]National Center for Mathematics and Interdisciplinary Sciences, Chinese Academy of Sciences, Beijing, China

[4]Department of Ecology and Evolution, University of Chicago

[†]These authors contributed equally to this work.

[‡]Present address: Genome Institute of Singapore, Agency for Science, Technology and Research, 60 Biopolis Street, Genome #02-01, Singapore, Singapore

***Corresponding author:** E-mail: ciwu@uchicago.edu.

The resequencing data for this isogenic ZS30 have been deposited to the NCBI database under the BioProject number PRJNA215656, and the short reads can be accessed through Short Reads Archive with the accession SRX337489.

Associate editor: Doris Bachtrog

Abstract

Studies of protein evolution have focused on amino acid substitutions with much less systematic analysis on insertion and deletions (indels) in protein coding genes. We hence surveyed 7,500 genes between *Drosophila melanogaster* and *D. simulans*, using *D. yakuba* as an outgroup for this purpose. The evolutionary rate of coding indels is indeed low, at only 3% of that of nonsynonymous substitutions. As coding indels follow a geometric distribution in size and tend to fall in low-complexity regions of proteins, it is unclear whether selection or mutation underlies this low rate. To resolve the issue, we collected genomic sequences from an isogenic African line of *D. melanogaster* (ZS30) at a high coverage of 70× and analyzed indel polymorphism between ZS30 and the reference genome. In comparing polymorphism and divergence, we found that the divergence to polymorphism ratio (i.e., fixation index) for smaller indels (size ≤10 bp) is very similar to that for synonymous changes, suggesting that most of the within-species polymorphism and between-species divergence for indels are selectively neutral. Interestingly, deletions of larger sizes (size ≥11 bp and ≤30 bp) have a much higher fixation index than synonymous mutations and 44.4% of fixed middle-sized deletions are estimated to be adaptive. To our surprise, this pattern is not found for insertions. Protein indel evolution appear to be in a dynamic flux of neutrally driven expansion (insertions) together with adaptive-driven contraction (deletions), and these observations provide important insights for understanding the fitness of new mutations as well as the evolutionary driving forces for genomic evolution in *Drosophila* species.

Key words: **population genetics, inferences, coalescent, neolithic transition, expansions.**

Introduction

The origin of and driving force for the evolutionary differences between species have been central topics in evolutionary biology (Kimura 1985; Gillespie 1994). With the availability of many genome sequences, tremendous insights into genome evolution at the level of nucleotide substitutions across the coding genes have gained in recent years (Abecasis et al. 2012). Other than nucleotide substitutions, structural changes, which also constitute a large portion of natural variations, are also of great importance in understanding genome evolution (Mikkelsen et al. 2005; Mills et al. 2011; Abecasis et al. 2012). Partly due to the nature of these changes [e.g., they tend to happen in the context of complex repetitive sequence (Haerty and Golding 2010)] and also due to the lack of efficient methods in characterizing these mutations, the study of structural mutations has been insufficiently undertaken.

Of all types of structural changes, small insertions and deletions (indels) are the most amenable to genomic analysis (Albers et al. 2011). With the recent advances in sequencing technology, especially the developments of longer sequence reads and elaborate algorithms performing realignment procedures after read mapping, indel variant calling has reached an excellent level of accuracy and power in recent years (McKenna et al. 2010; Albers et al. 2011; DePristo et al. 2011; Li 2011; Neuman et al. 2013). Despite these technological advances, the evolutionary driving force acting on small indels has not yet been elucidated.

Many indel mutations, especially those that occur in functionally important domains, will be disruptive to protein

Article

function. These indels are likely to be highly deleterious and will quickly be removed by natural selection. For example, germline indels that are associated with human genetic diseases constitute approximately 23.7% of total mutations in the Human Gene Mutation Database (Stenson et al. 2009). These deleterious indels are likely to contribute very little to within-population polymorphism and even more rarely to species divergence. In this case, much of the observed divergence between species might be due to a small proportion of indel mutations that are mildly deleterious or neutral and get fixed due to genetic drift (Kimura 1968; Ohta 1973). In other words, neutral or mild purifying selection might be driving indel divergence between species.

On the other hand, because of the extensive redundancy in the biological system (Edelman and Gally 2001), removal or rewiring of some genes for a different function through insertion or deletion changes could be highly advantageous (Conant and Wolfe 2008; Innan and Kondrashov 2010). Positive selection might also be the major driving force responsible for indel evolution. Because both hypotheses seem plausible, we examine them and address which forces may be involved in indel evolution.

In this study, we will focus on indel evolution in *D. melanogaster*. A powerful framework for investigating evolutionary forces is the McDonald Kreitman (MK) test, where information about within-population polymorphism is conjugated together with between-species divergence (McDonald and Kreitman 1991). Herein, we look at a subset of indels of a favorable size, where second-generation sequencing data show high sensitivity and specificity (Albers et al. 2011; Bansal and Libiger 2011; Neuman et al. 2013) (see later sections). Because of uncertainties in the sequence alignment between species (Wong et al. 2008), we will restrict ourselves to the coding part of the genome.

Earlier studies investigating nucleotide substitutions indicate widespread positive selection and between 30% and 60% of nucleotide substitutions in *D. melanogaster* coding and noncoding regions are estimated to be adaptive (Fay et al. 2002; Smith and Eyre-Walker 2002; Sawyer et al. 2003; Bierne and Eyre-Walker 2004; Andolfatto 2005, 2007; Sawyer et al. 2007; Hahn 2008; Sella et al. 2009). What will be the driving force for indel evolution in a background of pervasive positive selection on nucleotide changes is the key question we would like to address here.

Results

Indel Data within and between Species

Between-species orthologous coding genes were extracted from the 12 *Drosophila* genome project deposited in the Flybase database (Tweedie et al. 2009). Using a set of stringent criteria, we curated sequence alignments for 7,486 genes between the *Drosophila melanogaster*, *D. simulans* and *D. yakuba*. Because sequence alignment is a critical basis for subsequent analysis (Markova-Raina and Petrov 2011), we adopted a sophisticated alignment procedure that better reflects sequence homology in coding regions (Roshan and Livesay 2006) (see Materials and Methods).

To gather a good sample of within-species indel polymorphism, we conducted a resequencing study to capture natural variations within the *D. melanogaster* population. Because of the nature of structural changes, we wanted to sequence a good nonreference fly genome to high coverage supplying enough information about within-species polymorphism. The comparison between two high-quality genomes should provide a good basis for subsequent evolutionary analysis.

We completed whole-genome sequencing of one *D. melanogaster* African isogenic line (ZS30) collected from Zimbabwe using the Illumina GAIIx platform to a very high coverage of 70× (see supplementary files, Supplementary Material online, for details). Because variant identification is a very important step for our subsequent analysis and indel calling has gained rapid progress in recent years (e.g., realignment procedures), we want to first evaluate the performance of several popular programs and subsequently apply the analysis procedures to our real data. To tailor the pipeline for indel identification, we conducted a simulation study under similar specifics to our real data set (e.g., sequence coverage and reads length) and compared the performances of a set of programs including SAMtools/mpileup (Li et al. 2009; Li 2011), GATK (McKenna et al. 2010; DePristo et al. 2011), Stampy/Dindel (Albers et al. 2011; Lunter and Goodson 2011) as well as a customarized de novo assembly procedure based on Velvet (Zerbino and Birney 2008).

By large, the programs show similar performances in the overall sensitivity (between 90% and 100%) for small indels, but somewhat different in terms of false discovery rate (FDR). For example, SAMtools/mpileup can achieve a very low FDR (1.6%) and high sensitivity (>93.0%) across deletions in the range of 1–30 bp. However, sensitivity for insertions drops sharply around 23–25 bp, possibly due to difficulties in read mapping when there are inserted sequences of unknown content. Other programs such as GATK and Stampy/Dindel show similar performances. Interestingly, de novo assembly can identify longer insertions quite well but show high FDR in small indels, possibly due to the difficulties in mapping large contigs to the reference (see supplementary files, Supplementary Material online, for details).

It is worth emphasizing that similar trends have also been observed in several earlier studies (Albers et al. 2011; Li 2011; Neuman et al. 2013). After balancing performances and the easiness of usage, we used SAMtools/mpileup for subsequent analysis. To focus on a subset of indels of high confidence, we limited ourselves to deletions within the range of 1–30 bp and insertions of 1–20 bp in length.

After mapping the reads to the reference genome with BWA (Li and Durbin 2010), we used SAMtools (Li et al. 2009; Li 2011) to call variants between ZS30 and the reference genome (dm3) (see Materials and Methods). When looking at indels between, as well as within populations, indels show several broad-scale characteristics: 1) a geometric distribution with larger indels being less frequent (figs. 1*A* and 1*B*) (Massouras et al. 2012); 2) coding indels distributed toward both tails of the proteins with higher occurrences at the N and C terminals (figs. 1*C* and 1*D*); 3) when we classify the

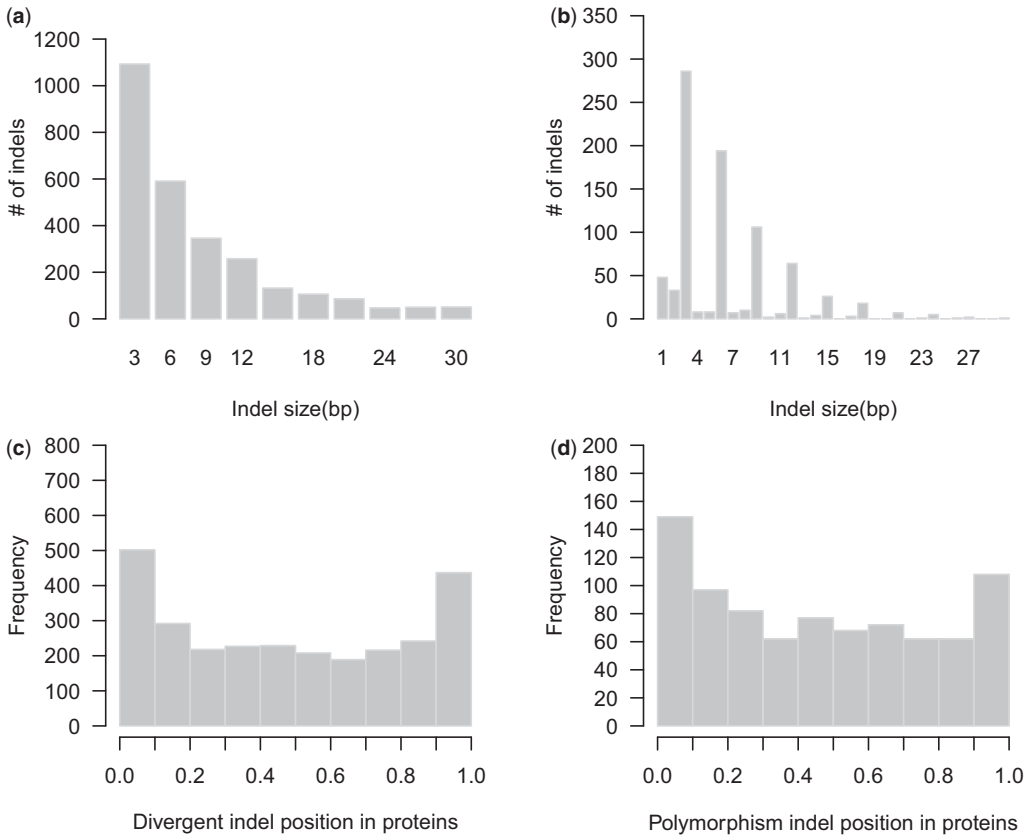

FIG. 1. Coding indel substitutions and polymorphism for the *D. melanogaster* lineage: (*A*) size distribution of fixed indels for the *D. melanogaster* lineage; (*B*) size distribution of polymorphic indels for the *D. melanogaster* population; (*C*) relative positions of fixed indels on their host proteins; (*D*) relative positions of polymorphic indels on their host proteins.

Table 1. Indel Distribution in Domain and Repetitive Regions.

Divergence	Domain[a]	Nondomain	Low_Complex[b]	Non-Low_Complex
Total length[a]	3,991,302	3,723,930	381,212	10,828,291
Observed indel number	181 (12.3%)	1,293 (87.7%)	594 (22.7%)	2,027 (77.3%)
Expected indel number	762 (51.7%)	712 (48.3%)	89 (3.4%)	2,532 (96.6%)
P value		$<2.2e^{-16}$		$<2.2e^{-16}$
Polymorphism	**Domain[a]**	**Nondomain**	**Low_Complex[b]**	**Non-Low_Complex**
Total length[a]	3,991,302	3,723,930	381,212	10,828,291
Observed indel number	65 (13.0%)	435 (87.0%)	336 (40.2%)	500 (59.8%)
Expected indel number	259 (51.7%)	241 (48.3%)	28 (3.4%)	808 (96.6%)
P value		$<2.2e^{-16}$		$<2.2e^{-16}$

[a]4,993 proteins have domain information.
[b]Low-complexity region, all 7,486 genes are included.

coding sequences into functional domains according to the Pfam database (Finn et al. 2010), there is a paucity of indels within domain regions (table 1), which is consistent with them being found more in protein tails (domains tend to reside in the middle of the protein); and 4) in contrast to functional domains, coding indels show strong enrichment in regions that has low complexity (i.e., simple sequence grammar, table 1) (Taylor et al. 2004).

When we compared the number of indel substitutions with nucleotide changes along the *D. melanogaster* lineage, we found that the indel substitution rate (2.5×10^{-4}) is estimated to be much lower than synonymous changes (7.13×10^{-2}). The low evolutionary rate observed for small indels is informative about selective forces acting upon indels (Charlesworth B and Charlesworth D 2010). However, the long-term evolutionary rate is affected by both mutation rate and natural selection. The low evolutionary rate of indels can be compatible with a reduced mutation rate or strong purifying selection. Thus, the occurrence of low mutation rate for indels might or might not be related to natural selection, and information gathered from polymorphism will shed light on the underlying mechanisms.

Table 2. The MK Table Including Substitutions and Small Indels.

Mutation Class	*D*/*P*	FI ($FI_{exp} \pm SD$)	*P* Value
Total deletion (1–30 bp)	1,356/426	1.07 (1.09 ± 0.055)	0.6798
Total Insertion (1–30 bp)	1,265/410	1.03 (0.97 ± 0.050)	0.1216
Nonsynonymous changes	56,051/13,074	1.44 (1.20 ± 0.011)	$<10^{-4}$
Synonymous changes	147,242/49,376	–	–
	Smaller size indels		
Deletion (≤10 bp)	954/354	0.90 (1.03 ± 0.059)	0.9923
Insertion (≤10 bp)	1,065/348	1.03 (0.94 ± 0.051)	0.0437
	Middle size indels		
Deletion (11–30 bp)	402/72	1.87 (1.27 ± 0.133) 1.74 (1.27 ± 0.133)[a]	0.0007**
Insertion (11–20 bp)	200/62	1.08 (1.18 ± 0.168) 1.03 (1.18 ± 0.168)[a]	0.7805

NOTE.—*D*/*P*, Divergence to polymorphism; FI, fixation index; Flexp, mean FI across all permuted tables; SD, standard deviation of the permuted FI.
[a]If we explicitly correct for the fact that our sensitivity is 0.93 for deletions and 0.95 for insertions.
**Significant *P* value.

Inferring Natural Selection with MK Test

The following is an analysis to partition natural selection on indels into positive and negative selection, taking into consideration the mutation rate gathered from polymorphism data (table 2). Because synonymous mutations are relatively neutral, the divergence to polymorphism ratio for synonymous changes provides a good null expectation for these two quantities. Other mutation classes (e.g., nonsynonymous changes) can be compared with synonymous changes using a statistic [denoted as fixation index (FI)] defined as a relative ratio between the ratio for that mutation class versus the ratio for the synonymous changes (Shapiro et al. 2007).

In table 2, we can see that the FI for nonsynonymous change is 1.44, indicating that the nonsynonymous substitutions is 1.44 times the expected value when compared with the neutrally evolving synonymous mutations. Previous studies looking at a subset of these genes also found a similar trend for nonsynonymous changes (Fay et al. 2002; Bierne and Eyre-Walker 2004; Andolfatto 2007; Sawyer et al. 2007; Shapiro et al. 2007; Sella et al. 2009). The overrepresentation of fixed differences for nonsynonymous changes is a strong indication of positive selection.

It is interesting to note that the expected value of FI is not 1 when data from multiple loci are pooled (Simpson 1951). As discussed by Shapiro et al. (2007), one of the approaches mitigating this issue is to perform a permutation test by conditioning on the marginal sums. When permuting the individual tables 10,000 times, the expected values for the FI are still much less than the observed values (table 2).

When calculating FI values for insertions or deletions only in the *D. melanogaster* lineage, we found that the observed values for each category are very similar to the permuted results, suggesting that indels are evolving similarly to synonymous changes. Interestingly, when partitioning the indels according to their sizes, deletions of larger sizes show a distinctive pattern from the rest of the categories. The FI for deletions between 11 and 30 bp shows very elevated fixed differences when compared with synonymous changes (observed 1.80, expected 1.27). This value is also much higher compared to nonsynonymous changes (1.80 vs. 1.44), which means that positive selection is playing an even stronger role for the evolution of middle-sized deletions than nonsynonymous changes. By calculating the level of excess in amount of fixed differences when compared with synonymous changes, we found that 44.4% of the fixed deletions in this size group could be driven by positive selection. This proportion stays relatively constant when a variety of other methods are also employed (see Materials and Methods).

When we focus on indels between 11 and 30 bp and separate the MK into genes with normal recombination rates and low recombination rates (see Materials and Methods), the overrepresentation of fixed changes is still prevalent although the value of FI is not significant in genes with lower recombination rates (fig. 2). This is compatible with the expectation of the Hill–Robertson effect, where the effect of natural selection will be greatly reduced in regions of lower recombination (Hill and Robertson 1966).

Among a subset of 4,993 genes with Pfam domain information, no statistically significant differences are observed when breaking down the indels into domain and nondomain regions (fig. 2). However, this is not the case when partitioning them into low-complexity (simple local sequence grammar, which is correlated with repetitive regions, see Materials and Methods) versus non-low-complexity regions. The FI in non-low-complexity regions is a lot larger than in low complexity regions, suggesting that adaptation is mostly driven by deletion mutations in the nonrepetitive part of the coding genes.

The more elevated FI for deletions compared with insertions is particularly interesting. Because conserving coding sequence is very important for proper function of proteins, adding new elements to the existing functional complex can often be deleterious because it is unlikely that the position, size, and content of the insertion will be in frame and/or inert with respect to the overall protein structure. On the contrary, fine-tuning protein functions through deletions or removing nonessential amino acids can often be highly advantageous

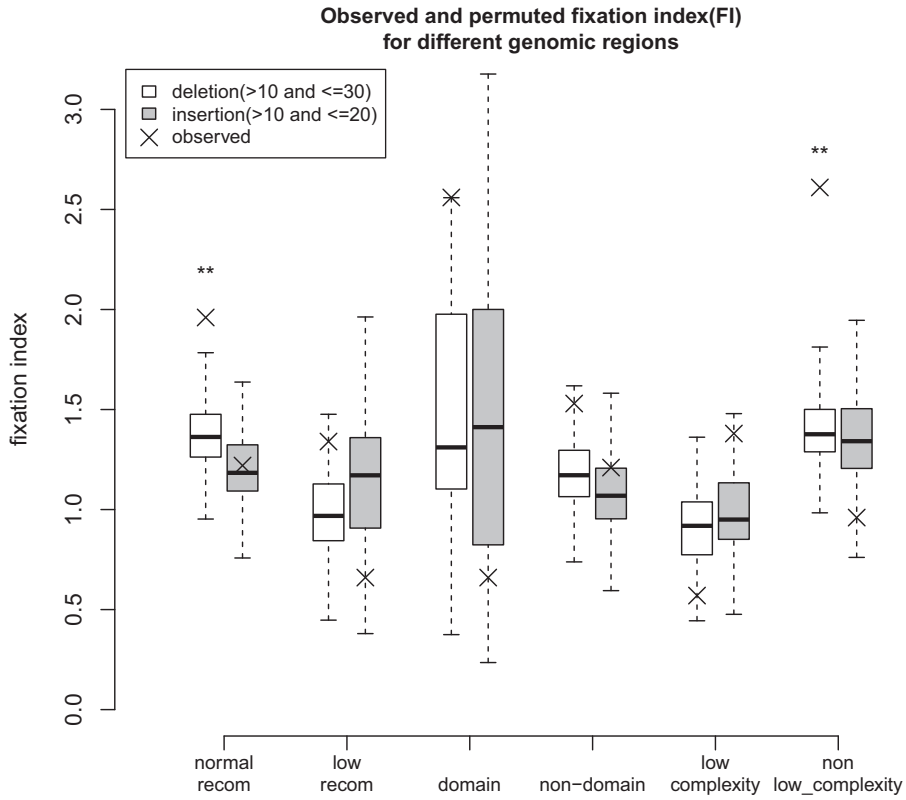

Fig. 2. Observed and permuted FI for different genomic regions. Deletions are restricted to be between 11 and 30 bp and insertions are between 11 and 20 bp. Observed FI is shown as crosses. The significant observed value is marked with **. In the boxplot (with default settings in R), the two edges of the boxes mark the 25% (*L*) and 75% (*U*) of the distribution, the whiskers extended to 1.5 interquantile region (i.e., IQR = $U - L$) in both directions. Between whisker region is roughly 99.3% of the probability density for a standard normal distribution.

(Olson 1999). The higher FI for deletions, as opposed to insertions, is an evidence for this conclusion. Evolutionary adaptation of indels seems to be dominated by increasing compactness by fine-tuning protein elements rather than accumulating complexity through adding new ingredients.

To further explore genes with strong signals of positive selection due to indel evolution, we conducted MK tests across all the surveyed genes. Because of the small number of indels observed for each gene, we combined them in the contingency table. We selected the top 200 most significant genes ($P \leq 0.01$) and carried out gene ontology (GO) analysis using DAVID (Huang et al. 2009). Interestingly, genes related to binding function (e.g., both ion and DNA binding) and transcriptional activities are strongly enriched with high indel divergence (table 3).

A representative example from the list of genes with multiple local sequence alignment surrounding the indels is presented in figure 3. Even though, many indels reside in simple local sequence context with clear evolutionary histories (figs. 3*A* and 3*C*), some of the indels can potentially have multiple possible evolutionary trajectories (figs. 3*B* and 3*E*), which might reflect elevated local mutation rates where indels tend to happen in clusters (figs. 3*B* and 3*D*). These indel changes also seem to correlate with local nucleotide substitutions, where sequence alignment can be quite complex and challenging (fig. 3*D*). This observation has also been pointed out in the context of comparative genomics, where a large proportion of the adaptive evolution detected between species can potentially be due to poor sequence alignment with multiple indels found in local clusters (Fletcher and Yang 2010; Markova-Raina and Petrov 2011). A model-based approach capturing this dynamic history, although it has been attempted (Thorne et al. 1992; Miklos et al. 2009), remains a challenging problem facing the field.

Table 3. GO Analysis for the Genes That Are Significant in the MK Test.

Molecular Function[a]	Fold Enrichment	Gene Number in Our Data Set[b]	*P* Value[c]
Metal ion binding	1.7	33	1.7e-3
Cation binding	1.6	33	2.7e-3
Ion binding	1.6	33	2.9e-3
Transition metal ion binding	1.9	30	3.2e-4
Zinc ion binding	2.4	28	1.6e-5
DNA binding	2.6	25	1.1e-5
Transcription regulator activity	2.2	17	3.9e-3
RNA polymerase II transcription factor activity	3.0	9	8.6e-3

[a]Only molecular function is shown in this table.
[b]This list the number of genes for this category in our significant genes. Different categories are not mutually exclusive and a single gene can appear in multiple categories.
[c]We used cutoff *P* value as 0.01.

```
A yak_nuc: ATGATGTCCTGCCCCCTTCCCATGCAATTACCAATACCAATTCCTTCAATA
   sim_nuc: ATGATGTCCTGCCCCATTCCCATGCAATTGCAAATGCCAATTCCTTCATTA
   mel_nuc: ATGATGTCCTGCCCCATTCCCATG------CCAATGCCAATTTCTTCATTA
   mel_aa :  M  M  S  C  P  I  P  M  -  -  P  M  P  I  S  S  L
   sim_aa :  M  M  S  C  P  I  P  M  Q  L  Q  M  P  I  P  S  L
   yak_aa :  M  M  S  C  P  L  P  M  Q  L  P  I  P  I  P  S  I

B yak_nuc: CAGGTGCACCAGCAACAACACCAACAGCAGCAACAGCAGCAGCAGGCGGCCGTCTTTCACCACCA
   sim_nuc: CAGGTGCACCAGCAGCAACACCAACAGCAG------CAGCAGCAGGCGGCCGTCTTTCACCACCA
   mel_nuc: CAGGTGCACCAGCAGCAACACCAACAG---------CAACAGCAGGCGGCCGTCTTTCACCACCA
   mel_aa :  Q  V  H  Q  Q  Q  H  Q  Q  -  -  -  Q  Q  Q  A  A  V  F  H  H  Q
   sim_aa :  Q  V  H  Q  Q  Q  H  Q  Q  Q  -  -  Q  Q  Q  A  A  V  F  H  H  Q
   yak_aa :  Q  V  H  Q  Q  Q  H  Q  Q  Q  Q  Q  Q  Q  Q  A  A  V  F  H  H  Q
   low_com:  *  *  *  *  *  *              *  *  *  *  *  *  *  *  *  *  *  *  *

C yak_nuc: TCCAGCAACAATGCCAACAGCTCAGGCGACAGCAACAGCAGCGCCGTG
   sim_nuc: ACCAGCAACAATGCCAACAGCTCAAGCGACAGCAACAACAGCGCCATG
   mel_nuc: ACCAGCAACAATGCAAACAGCTCAAGCGACAGCAAC---AGCGCCATG
   mel_aa :  T  S  N  N  A  N  S  S  S  D  S  N  -  S  A  M
   sim_aa :  T  S  N  N  A  N  S  S  S  D  S  N  N  S  A  M
   yak_aa :  S  S  N  N  A  N  S  S  G  D  S  N  S  S  A  V
   low_com:  *  *  *  *  *  *  *  *  *  *  *  *     *  *

D yak_nuc: CAGAACAACACAAGCGCCGTAAGTGCGGCCGGCAACAACACGAACTCCGCCGGAGGACCTAAT
   sim_nuc: CAAAACAACACAAGCGCCGTCAATGCGGGCGGC---------AACGCCGCCGGGGGACCCAAT
   mel_nuc: CAAAAC---ACAAGCGCCGTCAATGCGGGCGGC---AACACCAACGCTGCCGGGGGACCCAAT
   mel_aa :  Q  N  -  T  S  A  V  N  A  G  G  -  N  T  N  A  A  G  G  P  N
   sim_aa :  Q  N  N  T  S  A  V  N  A  G  G  -  -  -  N  A  A  G  G  P  N
   yak_aa :  Q  N  N  T  S  A  V  S  A  A  G  N  N  T  N  S  A  G  G  P  N
   domain :                $  $  $  $  $  $  $     $  $  $  $  $  $  $  $  $
```

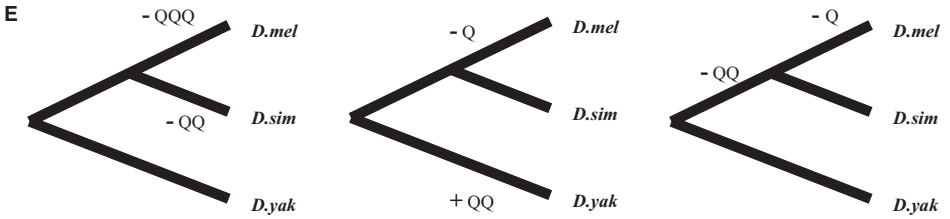

FIG. 3. Multiple sequence alignment for an example gene. Four segments for the gene hephaestus (FBgn0011224) is shown in panels *A*–*D*. Low-com stands for low complexity regions (see Materials and Methods). Domain stands for a Pfam domain (PF00076, RNA recognition motif). (*E*) Three possible evolutionary history for segment B.

Despite the fact that further explorations inspecting these genes in the protein structure databank failed to find a good hit with structural folding (possibly due to the limited information available so far), an enrichment test looking at Pfam domains within these 200 genes do find several interesting protein domains that are strongly enriched (supplementary table S2, Supplementary Material online). Some of the protein domains, including the zinc finger domains, are known to be playing important roles and are found to be evolving under positive selection (Oliver et al. 2009). Functional fine-tuning of binding affinities or selection for efficient protein folding or translational efficiency may be driving the positive selection in these genes.

Discussion

In this study, we completed a whole genome comparison of small indels along the *D. melanogaster* lineage using its sister and outgroup species. We found that indel evolution, both within populations and between species, is much lower than that of nucleotide substitutions. Combining polymorphism and divergence, we found that the FI for smaller indels (size $\leq$ 10 bp) is very similar to that of synonymous changes and is largely consistent with the neutral expectation. This observation is compatible with either neutral evolution for small indels or a mixed of positive and negative selection acting jointly creating a pattern that is in line with neutral prediction. Interestingly, middle-sized deletions (between 11 and 30 bp) are found to be under positive selection and 44.4% of fixed middle-sized deletions are estimated to be driven by adaptive evolution.

Even though most of the within-species polymorphism and between-species divergence for small indels are estimated to be neutral, this does not mean that raw indel mutations are all selectively neutral. If we assume the middle sections in short introns (i.e., bases 8–30 of introns shorter than 100 bp, designated as intron FEI site) (Halligan and Keightley 2006) are neutrally evolving, the observed indel polymorphism rate in these putatively neutral regions is about 3.3×10^{-3} for indels of size of multiples of three (denoted as 3N indels) and 1.6×10^{-3} for nonthree indels (denoted as non-3N indels), while the corresponding number in coding regions is 7.61×10^{-5} and 1.25×10^{-5}. This means that about 77.1% and 99.2% of raw indel mutations are strongly deleterious (Zichner et al. 2013). These deleterious mutations will contribute very little to both polymorphism and divergence.

Mutation Effect, Population Size, and Adaptive Evolution

The contrast observed between indels, as well as between indels of different sizes, indicates an interesting landscape

about the fitness distribution of these mutations. For example, as the deletion size increases, the fitness consequences will also become larger. Deletions will either become highly deleterious or strongly advantageous, and the proportion of neutral deletions will become very small. In this scenario, the percentage of deletion mutations that are fixed between species will be mostly due to those that are selectively advantageous. Thus, a very high FI is observed for larger deletions. On the contrary, the probability of insertions being advantageous is much lower, because the size, position, and content have to be correct simultaneously, adaptive evolution seems to be playing a negligible role in the evolution of insertions.

Previous results using nucleotide changes (Fay et al. 2002; Smith and Eyre-Walker 2002; Sawyer et al. 2003; Bierne and Eyre-Walker 2004; Andolfatto 2005; Begun et al. 2007; Sawyer et al. 2007; Shapiro et al. 2007; Sella et al. 2009) and copy number variations (Emerson et al. 2008) in *Drosophila* species have revealed extensive positive selection in *Drosophila* species. Positive selection in *Drosophila* leads us to ask the following: Why are there so many adaptive changes in these fly species (Hahn 2008; Sella et al. 2009)? The most intuitive explanation is that large effective population size is the major factor accounting for the evolution of these mutations. When population size is very large, the fitness effect of new mutations ($S = 4Ns$) will be greatly elevated. Mutations will either become highly deleterious or strongly advantageous and lead to the conclusion that most of the fixed differences are due to adaptive changes.

It is interesting to note that positive selection in middle-sized deletions, rather than insertions, indicates the existence of a phenomenon that might go beyond the genic regions. In the coding part of the genome, acceptable insertions have to take not only the correct size and position but also the right form and content. This might preclude positive selection in insertions. However, the whole *Drosophila* genome has evolved into a very compact form without much redundancy (Petrov et al. 2000; Petrov 2002; Peterson et al. 2009). This genomewide pattern suggests that the trend observed in coding regions of genes can also potentially be true for the noncoding part of the genome (Andolfatto 2005). Positive selection in species of large effective population size might be a more general occurrence. Of course, other processes including the nonallelic gene conversion might also be contributing to the overall genome size evolution (Assis and Kondrashov 2012). Nevertheless, natural selection could potentially be heavily involved in the evolution of genomes, with different combinations of positive and negative selection acting in diverse species (Lynch 2007).

Indel Evolution in Low-Complexity Regions

Several earlier studies looking at indel evolution in repetitive sequences revealed evidence of positive selection based on the observation of higher evolutionary rate. As we found in this study, a higher evolutionary rate might be confounded by the fact that repetitive regions have a more elevated indel mutation rate, which may not be correlated with natural selection (Schully and Hellberg 2006). Nevertheless, this overall pattern does not rule out the possibility that sequence changes in individual genes can still be positively selected (Ometto et al. 2005; Schully and Hellberg 2006; Parsch et al. 2010). The overall dynamics of protein length evolution might be taking a balance between neutrally evolving insertions (e.g., melting-down process, Lynch 2007) and adaptive-driven deletions shown here.

Future Directions

In this work, we only sequenced one isogenic line for polymorphism analysis. In principle, larger samples (e.g., 8–11) for polymorphism could be used to conduct a more sophisticated population genetics study. For example, populationwide allele frequency distribution might provide a better means for looking at the fitness effect of polymorphic mutations and provide a more elaborate picture about indel evolution (Eyre-Walker and Keightley 2007). However, considering the fact that the genetic variation is captured in the first few samples and the total number of variants increases very slowly with the sample size (on the order of $\ln(n)$) (Watterson 1975), using only two high-quality lines to gather polymorphism, information will only affect the power of the current approach. In other words, further sampling will tend to strengthen the conclusions drawn here. In addition, as the reference genomes improve for many sister species, for example *simulans* (Hu et al. 2013), further analysis might be endowed with an opportunity to push the study of this type to the whole genome level. Our study might be one of the many forthcoming studies looking at this problem using genomic approaches. With many individuals being sequenced across the tree of life, including the *Drosophila* groups (e.g., DPGP: Langley et al. 2012; DGRP: Mackay et al. 2012; Pool et al. 2012; Singh et al. 2013), we might be able to draw a more elaborate picture about the mode of evolution of indels at a much larger scale.

Materials and Methods

Sequence Alignments between Species

The coding sequences and amino acid sequences of 7,486 (about half of the total *D .melanogaster* genes) orthologous genes among *D. melanogaster*, *D. simulans*, and *D. yakuba* were downloaded from Flybase (Tweedie et al. 2009). Only one-to-one orthologous genes were retained. We required all the genes to have high-quality sequence, without a single "N" (assembled with unknown sequence or gap) in any of the sequences for each set of orthologous proteins. Multiple sequence alignment was then carried out using Probalign Version 1.4 (Roshan and Livesay 2006) for each orthologous gene set. Probalign outperforms other multiple sequence alignment programs in that 1) it estimates amino acid posterior probabilities from a partition function of alignments and 2) it computes the alignment of maximal expected accuracy (Roshan and Livesay 2006). After obtaining protein alignments, the aligned amino acid sequences were then backtranslated to nucleotide sequence alignment.

Delineating Lineage-Specific Indel and Nucleotide Substitutions

Using the *D. yakuba* sequence as an outgroup, we polarized indels on the *D. melanogaster* and *D. simulans* branches separately. Nucleotide substitutions were estimated using the codeml program from PAML (Yang 2007). All annotations are based on gene or genome positions of *D. melanogaster* (Flybase version r5.24). The evolutionary rate for small indels on the *D. melanogaster* and *D. simulans* lineage was estimated as the number of counted indels divided by the total aligned gene length.

Annotating Low-Complexity Regions across the Genome

Low-complexity regions in the protein coding sequences were queried using the SEG program (Wootton and Federhen 1996). Locally optimized low-complexity segments were produced at defined levels of stringency. The parameter settings were adopted from previous definitions of local compositional complexity (Wootton and Federhen 1993). The segment lengths and the number of segments per sequence were determined automatically by the algorithm. Here, to detect longer and more repetitive regions, we used the parameter settings suggested by Huntley and Clark (2007) with a window length of 15 and a complexity cutoff of 1.9.

Annotating Recombination Rates for the Genes

We extracted the recombination rates for the genes through the web-based tool *D. melanogaster* recombination rate calculator (RRC) version 2.2 (Fiston-Lavier et al. 2010). The low recombination rate was defined as those with recombination rate less than 0.002 cM/kB (Shapiro et al. 2007).

Genome Sequencing and Variant Calling

The isogenic line (ZS30) was constructed using flies collected from Zimbabwe and subsequently extracted with chromosomes 2 + 3 balancer (CyO + TM3) followed by inbreeding to get rid of X linked polymorphisms (Hollocher et al. 1997). A paired-end library with an insert size of 700 bp and a mate-pair library with an insert size of 1,500 bp were prepared from the genomic DNA of ~50 flies extracted using standard protocols. Illumina Genome Analyser II platform was employed to generate reads with length of 80 bp. The throughput of the data is presented in supplementary table S1, Supplementary Material online. All the sequenced reads were subsequently mapped to the *D. melanogaster* genome reference release 3 using BWA Version 0.5.9-r16 (Li and Durbin 2010). The mapped reads cover about 70× of the *Drosophila* genome.

Variants were called with SAMtools Version 0.1.16 (Li et al. 2009; Li 2011). First, the command "samtools mpileup–C50–E–uf $REF $BAM" was executed to generate a file of raw variants, $VAR. Then, the final set of the filtered variants was generated using the command "vcfutils.pl varFilter–D250 $VAR."

Data Simulation and Indel Calling Evaluation

We simulated short reads matching our real data set with wgsim (Li et al. 2009; Li 2011). In total, we simulated 27,500 indels whose size ranged from 1 to 80 bp with a geometric-like distribution. The simulated data set subsequently went through the same pipeline as our real data. Sensitivity is defined as the proportion of the indels that can be detected using the current approach. In other words, it is true positive/(true positive + false negative). FDR is defined as the false positive/(true positive + false positive). The details of the simulations are presented in detail in the supplementary file, Supplementary Material online

The MK Test

A contingency table with both point mutations (nonsynonymous and synonymous) and insertion and deletion polymorphisms was first compiled. The synonymous mutations were used as the baseline category representing mutations evolving under neutral evolution. Contingency tests were constructed by comparing variants of interest against synonymous mutations. The FI for a mutation category (e.g., indels) was calculated as (indel substitution/indel polymorphism)/(synonymous substitution/synonymous polymorphism) with the counts from the MK table.

Because the MK test requires all sites to come from the same genealogy, when information from multiple genes are aggregated into one table, the potential false correlation due to heterogeneities in these individual tables needs to be corrected. We conducted a permutation test similar to that of Shapiro et al (2007). The FIs from multiple permuted tables were then collected, and the empirical *P* value was evaluated as the percentage of the replicates where the permuted FI was greater or equal to the observed value.

Proportion of Substitutions Fixed by Adaptive Evolution

A simple method for calculating the proportion of substitutions fixed by positive selection is $\alpha = 1 - (D_s \times P_X)/(D_X \times P_s)$, where *D* and *P* represent fixed differences and polymorphism, respectively. Subscript s represents synonymous changes and x represents mutations of interest (e.g., nonsynonymous or indels). More elaborate methods (Smith and Eyre-Walker 2002; Bierne and Eyre-Walker 2004) from several recent studies are also implemented using in the DoFE package available at Dr. Adam Eyre-Walker's webpage.

GO Annotation and Functional Analysis

Candidate genes were first selected by conducting a MK test for each surveyed gene (Fisher-exact test *P* value < 0.01). GO analysis was then carried out using the DAVID web server version 6.7 (Huang et al. 2009). Only GO terms with *P* values less than 0.01 were retained in the final results.

Supplementary Material

Supplementary file and tables S1 and S2 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

The authors thank Chau-Ti Ting for helpful discussions. This work was supported by National Natural Science Foundation of China (31000957) and Ministry of Science and Technology (2012CB316505), as well as National Natural Science Foundation of China (91131011, 31071914, and 31061160189). C-I.W., W.Z., and Z.C. conceived the study. C.L, M.G, Q.G, J.R., J.L., L.J., and X.L helped with the experiments, Z.C., W.Z., and C.I.W analyzed the data. Z.C., W.Z., E.H, and C.I.W. wrote the manuscript.

References

Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* 491:56–65.

Albers CA, Lunter G, MacArthur DG, McVean G, Ouwehand WH, Durbin R. 2011. Dindel: accurate indel calls from short-read data. *Genome Res.* 21:961–973.

Andolfatto P. 2005. Adaptive evolution of non-coding DNA in *Drosophila*. *Nature* 437:1149–1152.

Andolfatto P. 2007. Hitchhiking effects of recurrent beneficial amino acid substitutions in the *Drosophila melanogaster* genome. *Genome Res.* 17:1755–1762.

Assis R, Kondrashov AS. 2012. A strong deletion bias in nonallelic gene conversion. *PLoS Genet.* 8:e1002508.

Bansal V, Libiger O. 2011. A probabilistic method for the detection and genotyping of small indels from population-scale sequence data. *Bioinformatics* 27:2047–2053.

Begun DJ, Holloway AK, Stevens K, et al. (13 co-authors). 2007. Population genomics: whole-genome analysis of polymorphism and divergence in *Drosophila simulans*. *PLoS Biol.* 5:2534–2559.

Bierne N, Eyre-Walker A. 2004. The genomic rate of adaptive amino acid substitution in *Drosophila*. *Mol Biol Evol.* 21:1350–1360.

Charlesworth B, Charlesworth D. 2010. Elements of evolutionary genetics. Greenwood Village (Colorado): Roberts & Company Publishers.

Conant GC, Wolfe KH. 2008. Turning a hobby into a job: how duplicated genes find new functions. *Nat Rev Genet.* 9:938–950.

DePristo MA, Banks E, Poplin R, et al. (18 co-authors). 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet.* 43:491–498.

Edelman GM, Gally JA. 2001. Degeneracy and complexity in biological systems. *Proc Natl Acad Sci U S A.* 98:13763–13768.

Emerson JJ, Cardoso-Moreira M, Borevitz JO, Long M. 2008. Natural selection shapes genome-wide patterns of copy-number polymorphism in *Drosophila melanogaster*. *Science* 320:1629–1631.

Eyre-Walker A, Keightley PD. 2007. The distribution of fitness effects of new mutations. *Nat Rev Genet.* 8:610–618.

Fay JC, Wyckoff GJ, Wu CI. 2002. Testing the neutral theory of molecular evolution with genomic data from *Drosophila*. *Nature* 415: 1024–1026.

Finn RD, Mistry J, Tate J, et al. (14 co-authors). 2010. The Pfam protein families database. *Nucleic Acids Res.* 38:D211–222.

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

Fletcher W, Yang Z. 2010. The effect of insertions, deletions, and alignment errors on the branch-site test of positive selection. *Mol Biol Evol.* 27:2257–2267.

Gillespie J. 1994. The causes of molecular evolution. New York: Oxford University Press.

Haerty W, Golding GB. 2010. Genome-wide evidence for selection acting on single amino acid repeats. *Genome Res.* 20:755–760.

Hahn MW. 2008. Toward a selection theory of molecular evolution. *Evolution* 62:255–265.

Halligan DL, Keightley PD. 2006. Ubiquitous selective constraints in the *Drosophila* genome revealed by a genome-wide interspecies comparison. *Genome Res.* 16:875–884.

Hill WG, Robertson A. 1966. The effect of linkage on limits to artificial selection. *Genet Res.* 8:269–294.

Hollocher H, Ting C-T, Pollack F, Wu CI. 1997. Incipient speciation by sexual isolation in *Drosophila melanogaster*: variation in mating preference and correlation between sexes. *Evolution* 51:1175–1181.

Hu TT, Eisen MB, Thornton KR, Andolfatto P. 2013. A second-generation assembly of the *Drosophila simulans* genome provides new insights into patterns of lineage-specific divergence. *Genome Res.* 23:89–98.

Huang DW, Sherman BT, Lempicki RA. 2009. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc.* 4:44–57.

Huntley MA, Clark AG. 2007. Evolutionary analysis of amino acid repeats across the genomes of 12 *Drosophila* species. *Mol Biol Evol.* 24: 2598–2609.

Innan H, Kondrashov F. 2010. The evolution of gene duplications: classifying and distinguishing between models. *Nat Rev Genet.* 11: 97–108.

Kimura M. 1968. Evolutionary rate at the molecular level. *Nature* 217: 624–626.

Kimura M. 1985. The neutral theory of molecular evolution. New York: Cambridge University Press.

Langley CH, Stevens K, Cardeno C, et al. (18 co-authors). 2012. Genomic variation in natural populations of *Drosophila melanogaster*. *Genetics* 192:533–598.

Li H. 2011. Improving SNP discovery by base alignment quality. *Bioinformatics* 27:1157–1158.

Li H, Durbin R. 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* 26:589–595.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Proc GPD. 2009. The sequence alignment/map format and SAMtools. *Bioinformatics* 25:2078–2079.

Lunter G, Goodson M. 2011. Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Res.* 21: 936–939.

Lynch M. 2007. The origins of genome architecture. Sunderland (MA): Sinauer Associates.

Mackay TFC, Richards S, Stone EA, et al. (52 co-authors). 2012. The *Drosophila melanogaster* genetic reference panel. *Nature* 482: 173–178.

Markova-Raina P, Petrov D. 2011. High sensitivity to aligner and high rate of false positives in the estimates of positive selection in the 12 *Drosophila* genomes. *Genome Res.* 21:863–874.

Massouras A, Waszak SM, Albarca-Aguilera M, et al. (11 co-authors). 2012. Genomic variation and its impact on gene expression in *Drosophila melanogaster*. *PLoS Genet.* 8:e1003055.

McDonald JH, Kreitman M. 1991. Adaptive protein evolution at the Adh locus in *Drosophila*. *Nature* 351:652–654.

McKenna A, Hanna M, Banks E, et al. (11 co-authors). 2010. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20:1297–1303.

Mikkelsen TS, Hillier LW, Eichler EE, et al. (68 co-authors). 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* 437:69–87.

Miklos I, Novak A, Satija R, Lyngso R, Hein J. 2009. Stochastic models of sequence evolution including insertion-deletion events. *Stat Methods Med Res.* 18:453–485.

Mills RE, Pittard WS, Mullaney JM, et al. (11 co-authors). 2011. Natural genetic variation caused by small insertions and deletions in the human genome. *Genome Res.* 21:830–839.

Neuman JA, Isakov O, Shomron N. 2013. Analysis of insertion-deletion from deep-sequencing data: software evaluation for optimal detection. *Brief Bioinform.* 14:46–55.

Ohta T. 1973. Slightly deleterious mutant substitutions in evolution. *Nature* 246:96–98.

Oliver PL, Goodstadt L, Bayes JJ, Birtle Z, Roach KC, Phadnis N, Beatson SA, Lunter G, Malik HS, Ponting CP. 2009. Accelerated evolution of the Prdm9 speciation gene across diverse metazoan taxa. *PLoS Genet.* 5:e1000753.

Olson MV. 1999. When less is more: gene loss as an engine of evolutionary change. *Am J Hum Genet.* 64:18–23.
Ometto L, Stephan W, De Lorenzo D. 2005. Insertion/deletion and nucleotide polymorphism data reveal constraints in *Drosophila melanogaster* introns and intergenic regions. *Genetics* 169:1521–1527.
Parsch J, Novozhilov S, Saminadin-Peter SS, Wong KM, Andolfatto P. 2010. On the utility of short intron sequences as a reference for the detection of positive and negative selection in *Drosophila*. *Mol Biol Evol.* 27:1226–1234.
Peterson BK, Hare EE, Iyer VN, Storage S, Conner L, Papaj DR, Kurashima R, Jang E, Eisen MB. 2009. Big genomes facilitate the comparative identification of regulatory elements. *PloS One.* 4:e4688.
Petrov DA. 2002. DNA loss and evolution of genome size in *Drosophila*. *Genetica* 115:81–91.
Petrov DA, Sangster TA, Johnston JS, Hartl DL, Shaw KL. 2000. Evidence for DNA loss as a determinant of genome size. *Science* 287: 1060–1062.
Pool JE, Corbett-Detig RB, Sugino RP, et al. (11 co-authors). 2012. Population genomics of sub-Saharan *Drosophila melanogaster*: African diversity and non-African admixture. *PLoS Genet.* 8: e1003080.
Roshan U, Livesay DR. 2006. Probalign: multiple sequence alignment using partition function posterior probabilities. *Bioinformatics* 22: 2715–2721.
Sawyer SA, Kulathinal RJ, Bustamante CD, Hartl DL. 2003. Bayesian analysis suggests that most amino acid replacements in *Drosophila* are driven by positive selection. *J Mol Evol.* 57:S154–S164.
Sawyer SA, Parsch J, Zhang Z, Hartl DL. 2007. Prevalence of positive selection among nearly neutral amino acid replacements in *Drosophila*. *Proc Natl Acad Sci U S A.* 104:6504–6510.
Schully SD, Hellberg ME. 2006. Positive selection on nucleotide substitutions and indels in accessory gland proteins of the *Drosophila pseudoobscura* subgroup. *J Mol Evol.* 62:793–802.
Sella G, Petrov DA, Przeworski M, Andolfatto P. 2009. Pervasive natural selection in the drosophila genome? *PLoS Genet.* 5:e1000495.
Shapiro JA, Huang W, Zhang CH, et al. (12 co-authors). 2007. Adaptive genic evolution in the *Drosophila* genomes. *Proc Natl Acad Sci U S A.* 104:2271–2276.
Simpson EH. 1951. The interpretation of interaction in contingency tables. *J R Stat Soc B.* 13:238–241.
Singh ND, Jensen JD, Clark AG, Aquadro CF. 2013. Inferences of demography and selection in an African population of *Drosophila melanogaster*. *Genetics* 193:215–228.
Smith NGC, Eyre-Walker A. 2002. Adaptive protein evolution in *Drosophila*. *Nature* 415:1022–1024.
Stenson PD, Mort M, Ball EV, Howells K, Phillips AD, Thomas NS, Cooper DN. 2009. The human gene mutation database: 2008 update. *Genome Med.* 1:13.
Taylor MS, Ponting CP, Copley RR. 2004. Occurrence and consequences of coding sequence insertions and deletions in mammalian genomes. *Genome Res.* 14:555–566.
Thorne JL, Kishino H, Felsenstein J. 1992. Inching toward reality: an improved likelihood model of sequence evolution. *J Mol Evol.* 34: 3–16.
Tweedie S, Ashburner M, Falls K, et al. (11 co-authors). 2009. FlyBase: enhancing *Drosophila* gene ontology annotations. *Nucleic Acids Res.* 37:D555–D559.
Watterson GA. 1975. On the number of segregating sites in genetical models without recombination. *Theor Popul Biol.* 7: 256–276.
Wong KM, Suchard MA, Huelsenbeck JP. 2008. Alignment uncertainty and genomic analysis. *Science* 319:473–476.
Wootton JC, Federhen S. 1993. Statistics of local complexity in amino-acid-sequences and sequence databases. *Comput Chem.* 17: 149–163.
Wootton JC, Federhen S. 1996. Analysis of compositionally biased regions in sequence databases. *Methods Enzymol.* 266: 554–571.
Yang ZH. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol.* 24:1586–1591.
Zerbino DR, Birney E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* 18:821–829.
Zichner T, Garfield DA, Rausch T, Stutz AM, Cannavo E, Braun M, Furlong EE, Korbel JO. 2013. Impact of genomic structural variation in *Drosophila melanogaster* based on population-scale sequencing. *Genome Res.* 23:568–579.