

Complex Interplay of Evolutionary Forces in the *ladybird* Homeobox Genes of *Drosophila melanogaster*

Evgeniy S. Balakirev^{1,2*}, Maria Anisimova^{3,4}, Francisco J. Ayala¹

1 Department of Ecology and Evolutionary Biology, University of California Irvine, Irvine, California, United States of America, **2** A.V. Zhirmunsky Institute of Marine Biology, Far Easter Branch, Russian Academy of Sciences, Vladivostok, Russia, **3** Department of Computer Science, Swiss Federal Institute of Technology ETH Zürich, Zürich, Switzerland, **4** Swiss Institute of Bioinformatics, Lausanne, Switzerland

Abstract

Tandemly arranged paralogous genes *lbe* and *lbl* are members of the *Drosophila* NK homeobox family. We analyzed population samples of *Drosophila melanogaster* from Africa, Europe, North and South America, and single strains of *D. sechellia*, *D. simulans*, and *D. yakuba* within two linked regions encompassing partial sequences of *lbe* and *lbl*. The evolution of *lbe* and *lbl* is highly constrained due to their important regulatory functions. Despite this, a variety of forces have shaped the patterns of variation in *lb* genes: recombination, intragenic gene conversion and natural selection strongly influence background variation created by linkage disequilibrium and dimorphic haplotype structure. The two genes exhibited similar levels of nucleotide diversity and positive selection was detected in the noncoding regions of both genes. However, synonymous variability was significantly higher for *lbe*: no nonsynonymous changes were observed in this gene. We argue that balancing selection impacts some synonymous sites of the *lbe* gene. Stability of mRNA secondary structure was significantly different between the *lbe* (but not *lbl*) haplotype groups and may represent a driving force of balancing selection in epistatically interacting synonymous sites. Balancing selection on synonymous sites may be the first, or one of a few such observations, in *Drosophila*. In contrast, recurrent positive selection on *lbl* at the protein level influenced evolution at three codon sites. Transcription factor binding-site profiles were different for *lbe* and *lbl*, suggesting that their developmental functions are not redundant. Combined with our previous results on nucleotide variation in esterase and other homeobox genes, these results suggest that interplay of balancing and directional selection may be a general feature of molecular evolution in *Drosophila* and other eukaryote genomes.

Citation: Balakirev ES, Anisimova M, Ayala FJ (2011) Complex Interplay of Evolutionary Forces in the *ladybird* Homeobox Genes of *Drosophila melanogaster*. PLoS ONE 6(7): e22613. doi:10.1371/journal.pone.0022613

Editor: William J. Etges, University of Arkansas, United States of America

Received: February 7, 2011; **Accepted:** June 29, 2011; **Published:** July 22, 2011

Balakirev et al. This is an open-access article, free of all copyright, and may be freely reproduced, distributed, transmitted, modified, built upon, or otherwise used by anyone for any lawful purpose. The work is made available under the Creative Commons CC0 public domain dedication.

Funding: This study was supported by Bren Professor Funds at the University of California, Irvine to F. J. Ayala and E. S. Balakirev and Swiss Federal Institute of Technology Zürich to Maria Anisimova. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

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

* E-mail: balakire@uci.edu

Introduction

Genetic changes in the genes that encode transcription factor (TF) proteins can produce fundamental phenotypic differences between species [for review, see 1–4]. Moreover, changes in TF coding sequences can result in profound modifications of the body plan [5,6]. In order to understand how complex phenotypes evolve, we need to understand how genes involved in transcriptional regulation evolve. A global genomic approach has revealed general trends in gene evolution and showed that positive Darwinian selection is an important factor driving molecular evolution [e.g., 7–9]. An important limitation of large-scale genomic studies is that they were unable to identify small-scale, within-gene variation that may directly influence protein function and corresponding phenotypic characteristics. Also, whole genome approaches are unable to reveal population level variation necessary for a better understanding of TF sequence evolution. Quantification of segregating variation within populations at TF loci is necessary to infer selective pressures and to ascertain the functional effects of naturally occurring allelic variation and sequence divergence among orthologs.

The available data on between-species TF variation indicate high rates of sequence evolution among regulatory genes [e.g., 10,11]. Studies of intra-specific nucleotide variation in *Drosophila* have revealed that regulatory genes tend to be less polymorphic than structural genes [12]. In contrast, homeobox genes from the 93DE cluster of *D. melanogaster* exhibit high sequence variation in *bagpipe* (but not in adjacent *tinman*) genes [13]. Also the TFs involved in olfactory pathways in *Caenorhabditis* exhibited more between- and within-species variation than structural chemosensory genes [14]. These data indicate that even adjacent regulatory genes can differ greatly in the level and pattern of sequence variation. This suggests that different members of a regulatory gene cluster may be subject to distinct evolutionary forces [15].

Here we focus on two homeobox genes, *ladybird early* (*lbe*) and *ladybird late* (*lbl*), tandemly-arranged paralogs in *Drosophila*. We analyze the level and pattern of the *lbe* and *lbl* segregating nucleotide variation in natural populations and divergence between close *Drosophila* species in attempt to reveal evolutionary forces governing the evolution of these genes. Both *lbe* and *lbl* genes are members of the NK homeobox gene family that consists of closely linked interacting regulatory genes (in 5' to 3' order: *tinman* (*tin*), *bagpipe* (*bap*), *lbe*, *lbl*, *C15*, and *slouch* (*S59*)), located on

the right arm of *D. melanogaster* chromosome 3 at cytological map position 93DE [16,17]. The coding region of *lbe* is 4,124 bp long and consists of two exons (1,008 and 432 bp) and one intron (2,684 bp). The *lbe* gene encodes a protein of 479 amino acids: the *lbe* homeodomain is located within exon II. The *lbl* coding region is 23,419 bp long and consists of three exons (702, 180, and 237 bp) and two introns (22,012 and 571 bp). The *lbl* gene is alternatively spliced, with three different LBL protein isoforms deduced from the sequence of cDNA clones [18], consisting of 342, 372, and 411 amino acids. The first part of the *lbl* homeodomain is in exon II and the rest in exon III, with a 571-bp intron that interrupts the *lbl* homeodomain. The transcription start site of *lbl* is 8.0 kb downstream of the *lbe* terminal stop codon. No ORF has been recorded in the region between the two genes and both genes are transcribed from the same DNA strand. The two genes show high similarity in the regions extending downstream and immediately upstream from the homeodomain [18,19]. The deduced LBE and LBL amino acid sequences are 97% identical in the homeodomain, 61–81% identical in the upstream conserved region and 77% identical downstream of the homeodomain [17].

The *lbe* and *lbl* genes encode transcription regulators, which play an important role in neurogenesis, myogenesis, and cardiogenesis [17,18,20]. These genes have almost identical expression patterns, although *lbe*, located at the 5' end, is activated slightly earlier than *lbl*; another difference concerns the trunk epidermis, where *lbe* transcripts are much more abundant [18]. Based upon their similar amino acid composition and expression patterns, both genes are often jointly referred to as “ladybird (*lb*)” [17,20]. Analyses of *lb* gain-of-function phenotypes and rescue experiments have led to the conclusion that *lbe* and *lbl* are functionally redundant [17,18]. In addition to *Drosophila*, *lb*-like genes have been detected in the sponge *Sycon raphanus* [21] and the mollusk *Loligo opalescens* [22], and orthologous genes have been found in mouse, chicken, and human genomes [23–25]. It is currently thought that the ladybird genes have an evolutionarily conserved role in development.

We previously investigated nucleotide variability in the *tin* and *bap* homeobox genes, located on the right arm of chromosome 3 of *D. melanogaster* within the 93DE cluster [13]. We now analyze nucleotide variation in *lbe* and *lbl* homeobox genes in 70 strains of *D. melanogaster* in four populations from East Africa (Zimbabwe), Europe (Spain), North (California) and South (Venezuela) America. We sequenced 4,482 bp covering the homologous coding (including the homeodomain) and noncoding (intron and 3'-flanking) regions for both genes (2,044 bp for *lbe* and 2,438 bp for *lbl*). The *lbe* and *lbl* genes display distinctive transcription factor binding-site profiles, suggesting that they are not redundant in developmental function. Negative selection and demography are major factors shaping the pattern of nucleotide polymorphism in the two genes. However there are clear indications of positive selection in the coding and noncoding regions of both genes, as well as balancing selection at synonymous sites in the *lbe* gene.

Results

Nucleotide Polymorphism

We detected similar total nucleotide diversity for *lbe* and *lbl* (Tables 1 and 2) close to the levels observed for *tin* and *bap* from the same 93DE gene cluster [13]. These estimates were within the range found in highly recombining gene regions of *D. melanogaster* [12] and in other regulatory genes [e.g., 15]. Figures S1 and S2 show the polymorphisms observed in 70 lines of *lbe* and *lbl*, including length polymorphisms. No nonsynonymous variability was detected in the *lbe* gene, but three nonsynonymous

polymorphic sites were found in *lbl* (exon III, Fig. S2). While silent nucleotide diversity for *lbe* and *lbl* was similar (Tables 1 and 2), the level of synonymous polymorphism was 4.9 times higher in *lbe* than *lbl* ($P < 0.001$). Synonymous variability of *lbe* was 4.4 times higher than noncoding variation (Table 1), a difference statistically significant in simulations even without recombination ($P = 0.01$), but the observed difference was not significant for *lbl*. There were five polymorphic sites within the *lbe* homeodomain (positions 1045, 1055, 1058, 1088, and 1121, Fig. S1), with $\pi = 0.0086$, slightly higher than for the whole *lbe* coding region ($\pi = 0.0057$). In contrast, there were two polymorphic sites in the *lbl* homeodomain (exon II, positions 837 and 903, Fig. S2), with $\pi = 0.0013$.

Synonymous variation of *lbe* was higher than in *Est-6* and $\psi Est-6$ (for the same population samples), which were among the most polymorphic genes in *D. melanogaster* [26,27]. High synonymous variability within *lbe* exon II was associated with the highest level of pair-wise divergence between *D. melanogaster* and three other *Drosophila* species (Table 1), which was several times higher than the divergence of the noncoding region (0.143–0.179 *vs.* 0.039–0.099). Functional significance could account for the fixation of favored codons, increasing the synonymous divergence in *lbe* exon II. The high variability of *lbe* exon II cannot be accounted for by relaxation of functional constraints, since it contains a functionally important homeodomain (180 bp at position 999–1178, our coordinates), which is conserved within a wide phylogenetic scale [17].

Variability of *lbl* exon III ($\pi = 0.0045$) was 3.5 times higher than exon II ($\pi = 0.0013$); a significant difference ($P < 0.01$) in coalescent simulations. Increased total variability within *lbl* exon III was accompanied by decreased silent divergence between *D. melanogaster* and *D. simulans*, which was 4.3 times lower in exon III than in exon II (0.0237 *vs.* 0.1020). Relaxation of functional constraints is one possible explanation for the prevalence of replacement substitutions in *lbl* exon III. However, these patterns indicated that the *lbl* coding region was under strong negative selection (see below), possibly imposed by alternative splicing [28] described for this gene [17]. Elevated replacement substitutions may indicate a functional shift of the *lbl* coding region evolving under positive selection. Below, we used neutrality tests and codon models to test this hypothesis. There was also a significant difference in population variability between *lbl* introns, 4.2 times higher in intron II than in intron I ($\pi = 0.0220$ *vs.* 0.0053, Table 2). A parallel difference in species divergence was observed between these two introns (see $K_{met-sim}$ and $K_{met-sec}$ in Table 2). The elevated divergence in the *lbl* intron II is puzzling. The region is rich with transcription factor binding sites (see below the section “Binding Site Profile”) and its complex architecture might be related to the specific evolutionary dynamic of intron enhancers that can evolve beyond recognizable sequence similarity while retaining function [e.g., 29]. For more details concerning nucleotide polymorphism, see Text S1 and Tables S1, S2, S3.

Recombination and Gene Conversion

The method of Hudson and Kaplan [30] revealed a minimum of 10 recombination events for *lbe*, 14 for *lbl*, and one between them. Estimates of the recombination rate ρ and the ρ/θ ratio were higher for *lbl* than for *lbe* ($\rho = 0.016$ and 0.006, respectively; Table S4). Previously we found a large difference (~ 33 times) in recombination rate between *tin* ($\rho = 0.001$) and *bap* ($\rho = 0.026$), within the 93DE cluster [13]. For *esterase* genes on the left arm of *D. melanogaster* chromosome 3 at cytological map position 68F7-F8, the rate was nearly three times higher for *Est-6* than for $\psi Est-6$ (0.021 *vs.* 0.008) [26,27]. This suggests that noticeably different recombination rates are common in tandemly associated paralogs.

Table 1. Nucleotide diversity and divergence in the *lbe* gene region of *D. melanogaster*.

	<i>lbe</i> exon II				Full sequence			
	Intron I	Syn	Nsyn	Total	3'-fl. region	Ncod	Silent	All sites
N	948	100	329	429	568	1516	1616	1945
S	21 (7)	8 (0)	0 (0)	8 (0)	23 (5)	44 (12)	52 (12)	52 (12)
π	0.0035	0.0244	0	0.0057	0.0090	0.0055	0.0067	0.0056
θ	0.0046	0.0166	0	0.0039	0.0084	0.0060	0.0068	0.0056
$K_{mel-sim}$	0.0257	0.1425	0	0.0309	0.0632	0.0393	0.0454	0.0374
$K_{mel-sec}$	0.0330	0.1425	0	0.0309	0.0705	0.0469	0.0526	0.0433
$K_{mel-yak}$	0.0792	0.1791	0	0.0382	0.1333	0.0986	0.1035	0.0845

Note. — Calculations based on 70 *D. melanogaster* lines from three populations: Barcelona, El Rio (California) and Venezuela, plus three lines from Zimbabwe and one *lbe* sequence from GenBank (accession number NT_033777.2). N, number of sites (indels excluded); S, polymorphic sites (number of singletons in parentheses); π , average number of nucleotide differences per site among all pairs of sequences [104, p. 256]; θ , average number of segregating nucleotide sites among all sequences, based on the expected distribution of neutral variants in a panmictic population at equilibrium [105]; $K_{mel-sim}$, $K_{mel-sec}$ and $K_{mel-yak}$ refer to nucleotide differences between *D. melanogaster* and *D. simulans*, *D. sechellia* or *D. yakuba*, respectively; Syn, synonymous sites; Nsyn, nonsynonymous sites; Ncod, noncoding (intronic and flanking regions); Silent, silent sites (synonymous and noncoding sites).

doi:10.1371/journal.pone.0022613.t001

The recombination rate was similar for European and North American samples, but much lower for South America (Table S4). A similar trend was observed for *Est-6* and ψ *Est-6* [26,27].

Sawyer's method [31] detected gene conversion events within *lbe* (except Venezuela) and *lbl*, but the number of significant events was considerably higher for *lbl* (Table S5). The average fragment length was 1,202 bp for *lbe* but 703 bp for *lbl*. Previously, we observed similar differences in average fragment length for *tin* (1,396 bp) and *bap* (665 bp) [13]. There was no evidence of intergenic gene conversion, likely due to low nucleotide similarity between *lbe* and *lbl* (53.5%), insufficient for efficient intergenic conversion.

Haplotype Structure

The *lbe* haplotype structure (excluding recombinants, see Fig. S1) is shown in Fig. 1, left. Due to recombination and gene conversion, this tree is not good reflection of the genealogical process, but serves to show the genetic structure of the data. There were two main *lbe* haplotype groups (1 and 2 in Fig. 1) and two sub-groups (1a and 2a). The main haplotype groups exhibited 19 nucleotide differences: 14 fixed within noncoding regions and five almost fixed within coding regions (excepting two recombinant variants detected for Ven-S-21F and Zim-S-44F; Fig. S1). The groups were differentially associated with indels. Group 2 was

completely associated with three deletions (1-, 4-, and 30-bp, within intron I and the 3'-flanking region; \blacktriangle 1, \blacktriangle 4, and \blacktriangle 6, Fig. S1). Group 1 was associated with two polymorphic indels (22-bp insertion and 8-bp deletion within the 3'-flanking region; \blacktriangledown 1 and \blacktriangle 7; Fig. S1). These five indels were not found in the *lbe* sequences of *D. simulans* and *D. sechellia*, suggesting that they may represent a derived condition after the split of *D. melanogaster* from the other two species. The difference between the two *lbe* haplotype groups was highly significant ($F_{st} = 0.84$; $K_{st} = 0.75$; $P < 0.001$).

There were two *lbe* haplotype sub-groups (1a and 2a); related respectively to group 1 and to group 2 (Fig. 1, left). Sub-group 1a had seven nearly fixed differences from group 1 (excepting one recombinant variant, ER-S-26F), all within noncoding regions (Fig. S1). The coding regions were identical for all *lbe* group 1. Sub-group 2a differed from group 2 by nine fixed nucleotide differences (excepting one recombinant variant, Zim-S-44F), all within the 3'-flanking region. There were six polymorphic sites within the coding region with different frequencies for group 2 and sub-group 2a; however, none of these differences were fixed. Sub-group 2a was associated with two indels (2-bp deletion and 10-bp insertion within the 3'-flanking region; \blacktriangle 8 and \blacktriangledown 2; Fig. S1).

Strong haplotype structure was also observed for *lbl*, with two main haplotype groups (1 and 2), and two sub-groups, 1a and 1b

Table 2. Nucleotide diversity and divergence in the *lbl* gene region of *D. melanogaster*.

	<i>lbl</i> exon II + exon III					Full sequence			
	Intron I	Syn	Nsyn	Total	Intron II	3'-fl. region	Ncod	Silent	All sites
N	725	94	311	405	264	643	1632	1726	2037
S	16 (3)	3 (1)	3 (0)	6 (1)	24 (6)	25 (11)	65 (20)	68 (21)	71 (21)
π	0.0053	0.0049	0.0025	0.0031	0.0220	0.0069	0.0086	0.0084	0.0075
θ	0.0046	0.0066	0.0020	0.0031	0.0189	0.0081	0.0083	0.0082	0.0072
$K_{mel-sim}$	0.0357	0.0585	0.0099	0.0209	0.1200	0.0674	0.0599	0.0598	0.0517
$K_{mel-sec}$	0.0299	0.0760	0.0112	0.0257	0.1223	0.0781	0.0623	0.0631	0.0547
$K_{mel-yak}$	0.0736	0.1496	0.0140	0.0452	0.2772	0.1885	0.1411	0.1416	0.1186

See Table 1, Note.

doi:10.1371/journal.pone.0022613.t002

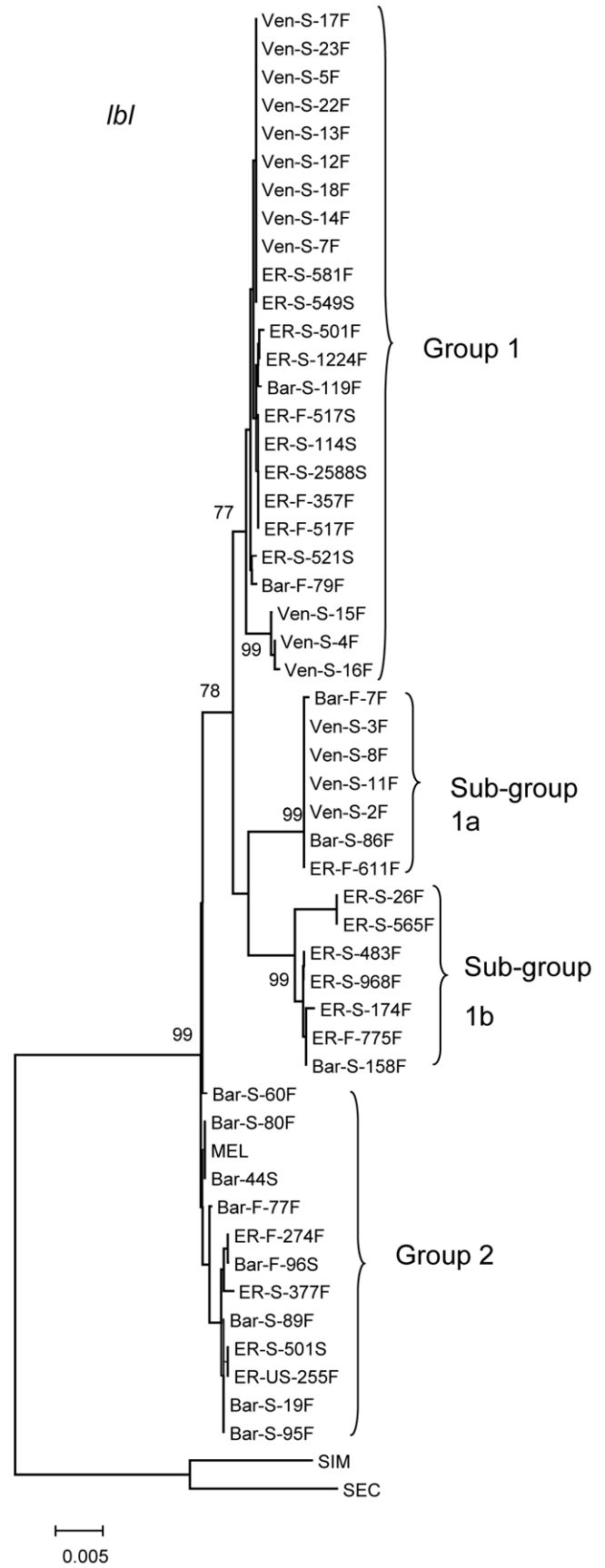
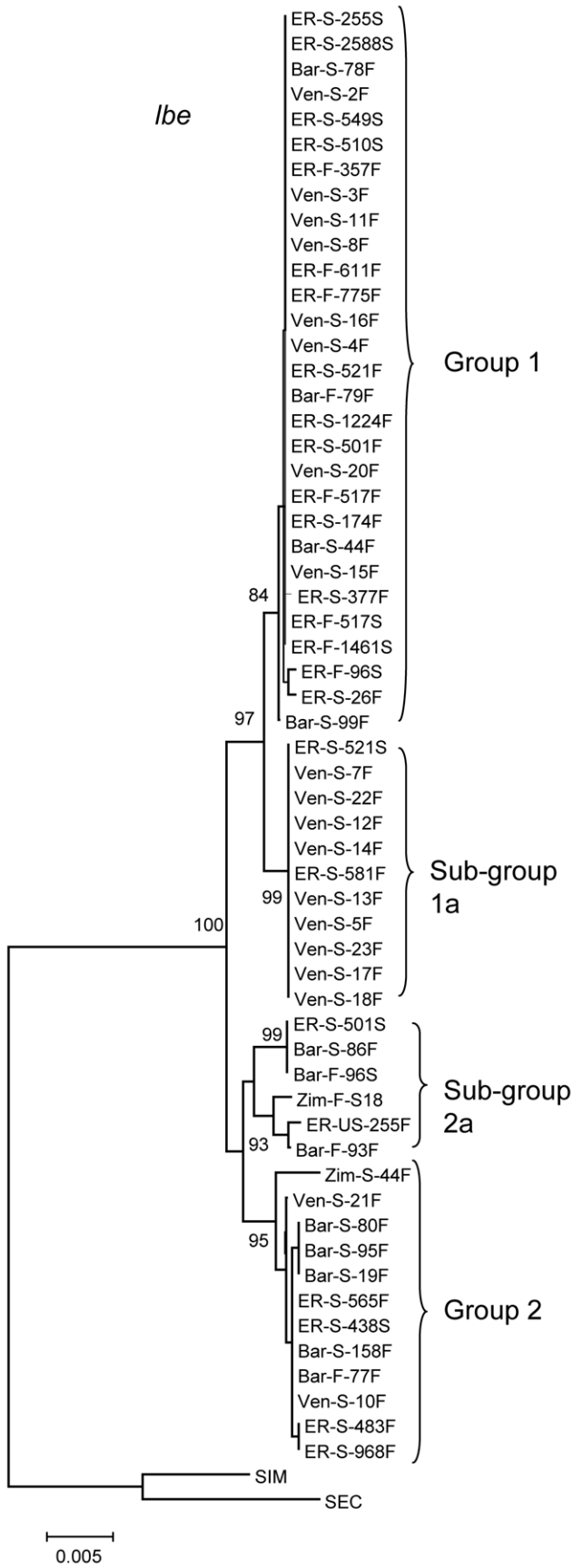


Figure 1. Neighbor-joining tree of *lbe* (left) and *lbl* (right) haplotypes of *D. melanogaster*, excluding recombinant sequences. The tree is based on Kimura 2-parameter distance. Numbers at the nodes are bootstrap percent support values based on 10,000 replications. Strains as in Fig. S1. Haplotype groups are encompassed in brackets. SEC, *D. sechellia*; SIM, *D. simulans*. doi:10.1371/journal.pone.0022613.g001

(Fig. 1, right). The *lbl* haplotype groups were differentially associated with indels: sub-group 1a was fully associated with a 12-bp insertion and a 26-bp deletion within intron II (▼2 and ▲2, Fig. S2); sub-group 1b was associated with a 9-bp deletion within exon III (▲6); group 2 was associated with two deletions (78- and 237-bp long) and a 24-bp insertion (▲1, ▼3, and ▲3); there were no indels within group 1. The difference between the *lbl* haplotype groups was highly significant ($F_{st} = 0.86$; $K_{st} = 0.54$; $P < 0.001$). Total sequence divergence (D_{xy}) among the *lbl* haplotype lineages was 0.0106 (ignoring indels), similar to *lbe* (0.0081).

Group 1 *lbe* haplotypes were most frequent in our data set, but variability was low (π , total = 0.0002), 5.5 times lower than in group 2. Group 2 haplotypes were less frequent and more variable (π , total = 0.0011). Group 2 is likely the ancestral state, consistent with the higher polymorphism of group 2 haplotypes, and supported by the comparison with *D. simulans* and *D. sechellia*. Group 1 may have evolved under directional selection (high frequency and low variability haplotype profile). For the North American sample (excluding recombinants), there were 37 polymorphic sites, and there was a subset of 17 sequences (haplotype group 1) with only four polymorphic sites (Fig. S1). The haplotype test [32] was significant ($P = 0.02$) with recombination rate $\rho = 0.015$. For the South American sample, the test was also significant ($P < 0.01$), as it was significant for the total dataset, contrasting group 1 with all available sequences ($P = 0.02$). The haplotype test was not significant for the full length of the *lbl* gene region, which could be explained by elevated number of polymorphic sites within intron I and 3'-flanking region (Fig. S2). However the test was significant ($P = 0.025$) for the *lbl* region including exon II, intron II, and exon III for the total dataset. Thus, the main haplotype group 1 might evolve under the influence of directional selection in both the *lbe* and *lbl* genes. For more details concerning haplotype structure, see Text S2.

Sliding Window Analysis

There were noticeable peaks of nucleotide variability along the *lbe* and *lbl* genes (Fig. 2). For *lbe*, bursts of variability were observed for intron I, exon II, and the 3'-flanking region with the most pronounced peak of silent polymorphism in exon II (midpoint coordinates 1073–1254; Fig. 2A). For *lbl*, there were three significant peaks within noncoding regions with the most pronounced peak observed for intron II, 3–4 times more polymorphic than other noncoding regions of *lbl*. The coding region of *lbl* also showed a significant peak of nucleotide variability in exon III accompanied by decrease of divergence that strongly contrasted with the high divergence in the adjacent regions (Fig. 2B). Peaks of nucleotide diversity were accompanied by increased levels of linkage disequilibrium (Fig. 2C) and significant values of Kelly's [33], Wall's [34], and Tajima's [35] neutrality test statistics (Fig. 2D, E, and F), suggesting that positive (balancing or diversifying) selection may overcome the dominant effect of negative selection in this highly constrained functional region.

Patterns of polymorphism and divergence were different in *lbe* and *lbl*. For *lbe* the highest silent polymorphism and divergence were observed in exon II, but for *lbl* in a noncoding region, intron II. We suggest that there may be multiple targets of positive selection within the noncoding and coding regions of both genes. Despite the absence of replacement substitutions in *lbe*, the most pronounced *lbe* peak of variability was located within exon II,

seven and three times more polymorphic than intron I and the 3'-flanking region, respectively. One explanation is that balancing selection acts on synonymous sites of *lbe* (additional support for this premise below).

Two types of selection might be involved in the evolution of the *lb* genes: balancing selection that creates elevated nucleotide variation around target polymorphic sites, and directional selection that creates significant excesses of very similar sequences exhibiting very low levels of variation. A similar scenario was inferred to explain patterns of nucleotide polymorphism for *Est-6* and *bagpipe* in *D. melanogaster* [13,26,36].

Tests of Neutrality and Maximum Likelihood Analysis of Selective Pressures

Neutrality tests detected significant deviations from neutrality in the *lb* region (Table S7), based on linkage disequilibrium (LD) between segregating sites with recombination ($\rho = 0.015$, see above). When applied to the full *ladybird* gene region, Kelly's [33] Z_{ns} and Wall's [34] B and Q tests were highly significant for the whole dataset, and for each population separately (Table S7). The tests were also significant separately for the *lbe* and *lbl* genes (except the European sample) as well as for the coding and flanking regions of both genes (Table S7). Significant values of Kelly's and Wall's statistics were grouped around the peaks of linkage disequilibrium (Fig. 2), and centered around the coding and noncoding regions of *lbe* and *lbl*, which supporting the hypothesis that these sites were targets of balancing selection. For more detailed information concerning LD, see Text S3, Fig. S3, and Table S6.

However, neutrality tests are typically affected by demography and so may be difficult to interpret [37,38]. We applied model-based maximum likelihood (ML) methods to confirm the observations made above (results summarized in Table 3). All results from the ML analyses revealed below held for the full sample and for the *D. melanogaster* strains separately, whether or not recombinant sequences were removed.

The test for variability of diversifying selection on the protein was not significant for *lbe* (M0 vs. M3) [39] as stringent purifying selection prohibited nonsynonymous changes. In contrast, this test was highly significant for *lbl* (Table 3). The test for positive selection (M7 vs. M8, Table 3) was significant for *lbl*, suggesting that 2% of sites evolved under strong diversifying selection ($\omega \approx 24.5$), while the remaining sites were very conserved ($\omega \approx 0.1$). Sites 118Q, 122S and 131L (in bold in Fig. S2) in *lbl* were subject to positive selection at the protein level (posterior probability > 0.99). Lower, but significant, levels of positive selection were detected in other genes (Table 3), with 7, 3, and 1 sites under positive selection, respectively for *Est-6*, $\psi Est-6$, and *bap* (posterior probability > 0.99), consistent with previous results [13,40].

The synonymous rate d_S in the coding sequence varied significantly for both *lbe* and *lbl* (M3 vs. M3-Dual in Table 3). Coefficient of variation (CV) was used to compare the variability of nonsynonymous and synonymous rates estimated under the M3-Dual model. Synonymous variation was higher in *lbe* with $CV_S = 3.59$, compared to 2.27 in *lbl*. Interestingly, the synonymous variation in *lbl* was similar to that in other genes, such as *Est-6*, $\psi Est-6$ and *bap* with $CV_S = 2.29$, 2.05, and 2.33, respectively. In these genes, the nonsynonymous rate variation (CV_N) was much

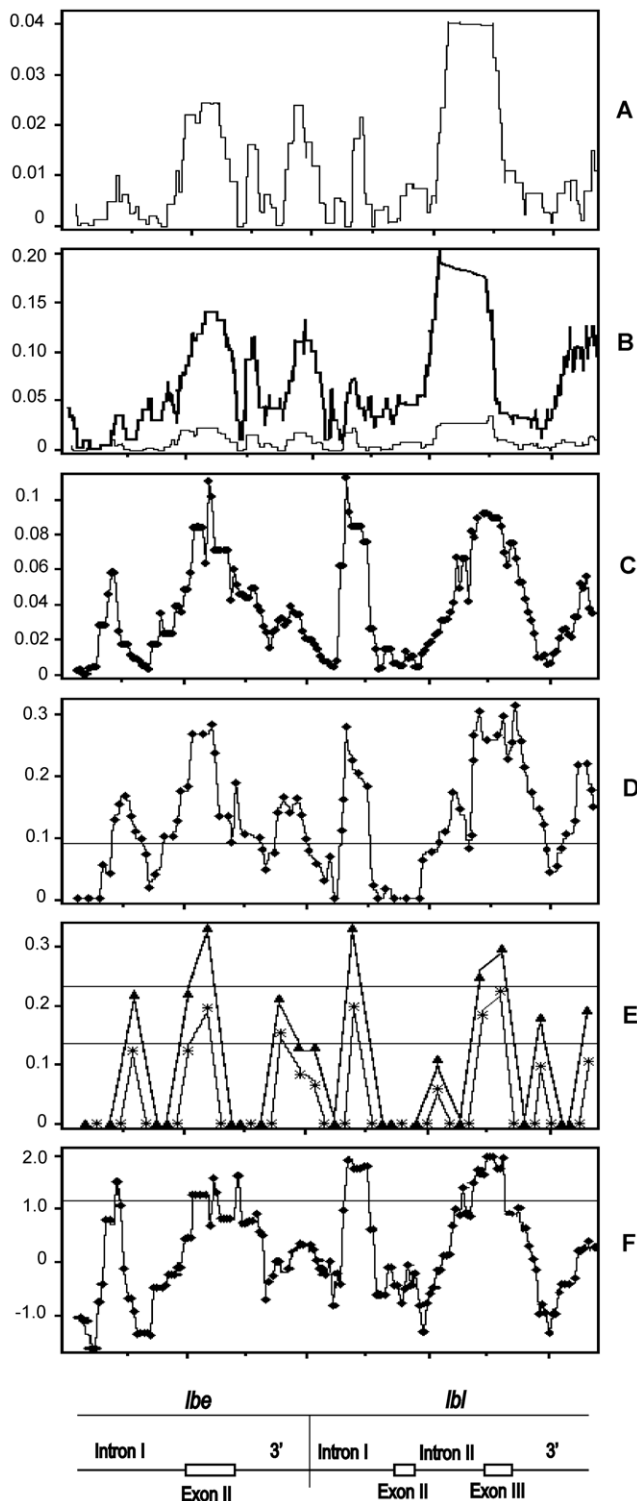


Figure 2. Sliding window plots along the *lbe* and *lbl* genes of *D. melanogaster*. A, silent nucleotide diversity; B, silent nucleotide diversity (thin line) and divergence (thick line); C, linkage disequilibrium measured by D ; D, E, and F, neutrality test statistics of Kelly's Z_{ns} [33], Wall's B and Q [34], and Tajima's D [35], respectively. Window sizes are 100 nucleotides with one-nucleotide increments for A and B; 250 nucleotides with 25-nucleotide increments for C; 250 nucleotides with 30-nucleotide increments for D and F; and 250 nucleotides with 150-nucleotide increments for E. A schematic representation of the region investigated is displayed at bottom. 99% confident intervals for

neutrality test statistics (D, E, and F plots) are marked by thin horizontal lines (there are two horizontal lines in E: Top is for the B statistic and bottom is for the Q statistic of Wall's [34] test) obtained by coalescent simulations conditioned on the number of polymorphic sites with the recombination rate equal to 0.015.

doi:10.1371/journal.pone.0022613.g002

higher than for the synonymous rate, with CV_N 4.4–6 and CV_S 2–2.3. For ψ_{Est-6} and *tin* the variation in d_N and d_S was similar. In contrast, synonymous variation was not only higher in *lbe* comparatively to other genes, but was accompanied by a total absence of nonsynonymous substitutions ($CV_N = 0$). Seven percent of sites in *lbe* evolved with unusually high $d_S = 14.3$, whereas at the remaining sites the rate was constrained to d_S 0.01–0.08.

Widespread positive selection in synonymous sites was previously reported for mammalian genes [41] where selection may act through mRNA destabilization affecting mRNA levels and translation [42,43]. Since protein folding is thought to occur simultaneously with protein translation from mRNA, the use of preferred and unpreferred codons may affect protein translation rates [44]. All polymorphic sites in the *lbe* coding region (Fig. S1) showed anomalously high rates of synonymous substitutions (posterior probabilities >0.99). These sites may be responding to diversifying selection on synonymous codons, perhaps affecting the speed of translation, with possible implications for protein folding (see below “mRNA Secondary Structure Stability” for details).

Synonymous codon usage is typically high in *Drosophila* [e.g., 45,46]. In our data, the test comparing selection-mutation models FMutSel0 vs. FMutSel was highly significant for both *lb* genes and all other members of the 93DE gene cluster (Table 3), suggesting that natural selection was a driving force in the evolution of synonymous codons.

Finally, the LRT for positive selection in noncoding regions (see “Materials and Methods”) was significant for both *lb* genes (Table 3). For *lbe*, 5% sites in the noncoding region were estimated to evolve by positive selection, with a substitution rate more than six times higher than the average d_S in the coding region of *lbe* ($\zeta_1 = 6.55$). For *lbl*, a larger proportion of sites (14%) was estimated to be under positive selection, although the estimated selection pressure was lower, with $\zeta_1 = 2.81$. Such differences in estimated positive selection pressure on noncoding regions are especially striking considering that the average d_S is much higher for *lbe* than for *lbl* (estimates from M0, Table 3). For more detailed information concerning ML analyses, see Text S4.

mRNA Secondary Structure Stability

We calculated RNA secondary structure free energy for the representative sequences of the main *lbe* and *lbl* haplotype groups (Table 4) using the program RNAstructure [47]. The major *lbe* haplotype groups (I and II, Fig. 1) significantly differed ($P = 0.0001$) with respect to mRNA secondary structure while the *lbe* group I haplotypes were less stable than group II. In contrast to *lbe*, the difference in mRNA stability between *lbl* haplotypes was small and not statistically significant ($P > 0.5$). Predicted mRNA stability was much higher for *lbe* than for *lbl* (Mann-Whitney test $P < 0.0001$) for both haplotype groups (Table 4).

Thus, *lbe* haplotypes divergent in only synonymous changes exhibited significant differences in mRNA stability. This observation highlights the potential significance of *lbe* synonymous variation, providing indirect evidence for the functional basis of balancing selection maintaining synonymous variation in this gene. Given the evidence from other studies that differences in mRNA secondary structure stability can affect mRNA decay [48], gene expression [49,50], and level of protein translation [51,52],

Table 3. Likelihood ratio tests LRTs based on codon models of evolution.

LRTs null vs. alternative	Genes ^a	LRT statistic ^b	P-value ^b	Estimates of interest ^c
M0 vs. M3	<i>lbe</i>	0	1	$\omega = 0, \kappa = 2.09, t = 0.37, d_S = 0.80, d_N = 0$
	<i>lbl</i>	45.37	<0.001	$\omega = 0.22, \kappa = 3.72, t = 0.27, d_S = 0.22, d_N = 0.05$
	<i>Est-6</i>	164.50	<0.001	$\omega = 0.23, \kappa = 3.14, t = 0.44, d_S = 0.33, d_N = 0.08$
	ψ <i>Est-6</i>	184.62	<0.001	$\omega = 0.29, \kappa = 1.64, t = 0.61, d_S = 0.44, d_N = 0.13$
	<i>bap</i>	50.43	<0.001	$\omega = 0.10, \kappa = 4.77, t = 0.22, d_S = 0.24, d_N = 0.02$
	<i>tin</i>	0	1	$\omega = 0.13, \kappa = 3.23, t = 0.12, d_S = 0.13, d_N = 0.02$
M3* vs. Dual	<i>lbe</i>	54.64	<0.001	$CV_S = 3.59, CV_N = 0$
	<i>lbl</i>	19.47	<0.001	$CV_S = 2.27, CV_N = 6.00$
	<i>Est-6</i>	130.38	<0.001	$CV_S = 2.29, CV_N = 4.42$
	ψ <i>Est-6</i>	132.06	<0.001	$CV_S = 2.05, CV_N = 2.13$
	<i>bap</i>	28.08	<0.001	$CV_S = 2.33, CV_N = 5.77$
FMutSel0 vs. FMutSel	<i>lbe</i>	101.37	<0.001	$P_+ = 0.22, S_+ = 0.73, S_- = -3.53$
	<i>lbl</i>	75.21	<0.001	$P_+ = 0.04, S_+ = 2.27, S_- = -4.60$
	<i>Est-6</i>	128.93	<0.001	$P_+ = 0.25, S_+ = 0.83, S_- = -1.47$
	ψ <i>Est-6</i>	109.21	<0.001	$P_+ = 0.31, S_+ = 0.64, S_- = -0.96$
M7 vs. M8	<i>lbl</i>	21.14	<0.001	$p = 8.84, q = 99, p_0 = 0.978, \omega = 24.52, p_1 = 0.022$
	<i>Est-6</i>	63.15	<0.001	$p = 0.33, q = 2.24, p_0 = 0.986, \omega = 9.20, p_1 = 0.014$
	ψ <i>Est-6</i>	27.21	<0.001	$p = 0.01, q = 0.03, p_0 = 0.994, \omega = 9.05, p_1 = 0.006$
	<i>bap</i>	11.47	0.003	$p = 0.01, q = 0.13, p_0 = 0.992, \omega = 9.90, p_1 = 0.008$
Selection on noncoding regions $\zeta = 1$ vs. $\zeta \geq 1$	<i>lbe</i>	178.86	<0.001	$\zeta_0 = 0.20, p_0 = 0.95, \zeta_1 = 6.55, p_1 = 0.05$
	<i>lbl</i>	25.11	<0.001	$\zeta_0 = 0.08, p_0 = 0.86, \zeta_1 = 2.81, p_1 = 0.14$

M3* is the HYPHY implementation of model M3 [98]. Tests comparing M3* vs. Dual and M7 vs. M8 were performed only for genes where M0 vs. M3 was significant. The test FMutSel0 vs. FMutSel was performed only for genes with large samples.

^aResults for *lb* genes are shown for full samples. However, all LRT results remain the same when only *D. melanogaster* was analyzed (with or without the recombinants).

^bLRT statistic is double the difference between the likelihood values optimized under the alternative and null models: $2(\ln L_{alt} - \ln L_{null})$. P-value is computed using the χ^2 -distribution with d.f. = 4 for LRTs of M0 vs. M3, and M3* vs. Dual. For LRT of M7 vs. M8 d.f. = 2, and for FMutSel0 vs. FMutSel d.f. = 41. For the LRT of positive selection on noncoding regions d.f. = 1. Significant P-values in bold.

^cEstimates under M0: $\omega = d_N/d_S$; tree length t is measured by number of expected nucleotide substitutions per codon over the tree; transition/transversion ratio κ (estimated under M3, or under M0 if the test M0 vs. M3 is not significant); d_S , tree length for synonymous substitutions, d_N , tree length for nonsynonymous sites. Estimates under Dual model: coefficients of variation (=standard deviation/mean) for distributions of synonymous and nonsynonymous rates, CV_S and CV_N respectively. Estimates under mutation-selection model FMutSel: P_+ is the proportion of mutations with advantageous effect ($S = 2N_S > 0$); S_+ is mean selection coefficient of all advantageous mutations; S_- is mean selection coefficient of all deleterious mutations. Estimates under M8: p, q are parameters controlling the shape of the Beta-distribution; p_0 is proportion of sites in the sequence with ω -values from beta-distribution (between 0 and 1); $\omega = d_N/d_S$, with discrete class allowed to be > 1 (under positive selection); p_1 is proportion of sites in the discrete class. In the combined coding and noncoding model [102], sites in the noncoding region come from two categories: proportion p_0 are under negative selection with $\zeta_0 < 1$, and p_1 are under positive selection or evolving neutrally with $\zeta_1 \geq 1$; ζ_0 and ζ_1 are estimates of the ratio of substitution rate in the noncoding region to average synonymous rate in the coding region for the two categories.

doi:10.1371/journal.pone.0022613.t003

we propose that *lbe* synonymous polymorphisms may be important contributors to adaptive variation in *D. melanogaster*. Because there was no difference in secondary structure stability between *lbl* mRNA transcripts representing two main haplotype groups, mRNA secondary structure for this gene may not be a target for

Table 4. Free energy (ΔG , kcal/mol) of *lbe* and *lbl* mRNA base pairing.

	Group I (ER-S-2588S)		Group II (ER-S-565F)		P
	Mean	St. dev	Mean	St. dev	
<i>lbe</i>	-144.639	1.6323	-146.734	2.0640	0.0001
<i>lbl</i>	-90.061	2.4110	-89.676	2.3594	N.S.

St. dev.: standard deviation.

doi:10.1371/journal.pone.0022613.t004

positive selection. We showed above (section “Tests of Neutrality and Maximum Likelihood Analysis of Selective Pressures”) that for the *lbl* gene, positive selection operated on the protein level.

Predicted mRNA stability was much higher for *lbe* than for *lbl* (Table 4; Mann-Whitney test $P < 0.0001$), consistent with the GC difference between the genes. Total GC content was significantly higher in *lbe* than *lbl* (59.5% vs. 51.3%; Wilcoxon test $P = 0.0001$; Mann-Whitney test $P = 0.0001$). It was shown previously that increased levels of GC in coding sequences have a stabilizing effect on mRNA secondary structure [e.g., 53]. Accordingly, *lbe* evolution was associated with increased stability and balancing selection on mRNA secondary structure whereas *lbl* evolution was accompanied by lower mRNA stability and positive selection at the protein level.

Binding Site Profile

Drosophila transcription factor motifs were obtained from FlyReg database curated motifs [54], and we used ClusterDraw2 [55] to scan for binding site matches and binding site clusters. The

program uses a polynomial model to determine statistical significance of binding site clusters. The cluster significance cutoff was set to 3, corresponding to a significance level of $P=0.001$. We analyzed 44 transcription factor motifs and detected significant clusters for 30 motifs (Table S8). The spatial distribution of clusters was not uniform along *lb* genes because the vast majority of significant clusters were located within non-coding sequences. The deviation from equal proportion of significant clusters was highly significant for both *lbe* ($\chi^2=29.13$, d.f. = 1, $P<0.001$) and *lbl* ($\chi^2=32.30$, d.f. = 1, $P<0.001$). Twelve of 30 motifs were significant for both genes. Ten clusters were significant only for *lbe* (part of a specific component of the *lbe* regulatory profile). Eight clusters were significant only for *lbl* (specific component of the *lbl* regulatory profile). In *lbe* and *lbl*, these clusters coincided for 40% of the binding sites, whereas the remaining clusters were gene-specific. The distribution of significant binding-site clusters for *lbe* and *lbl* was highly asymmetric, with a proportion of specific clusters $>50\%$. The difference in binding-site profiles suggests that the genes are not redundant in developmental function.

Discussion

Our analyses revealed a dimorphic haplotype structure for both *lb* genes. Despite similar levels of total nucleotide diversity in *lb* genes, synonymous nucleotide variability and the variation in the synonymous rate of change were much higher in *lbe* than in *lbl* and other genes from the 93DE cluster (*tin* and *bap*) as well as *Est-6* and ψ *Est-6* that are among the most polymorphic genes of *D. melanogaster* [13,26,27]. We attribute this high synonymous variation to balancing selection on *lbe* synonymous sites. Resch et al. [41] showed for mammalian genes that positive selection at synonymous sites may act through mRNA destabilization affecting mRNA levels and translation. This observation is in accordance with widespread compensatory evolution at the molecular level, caused by epistatic selection maintaining mRNA secondary structures [56]. The mechanism underlying epistatic selection is based on a model of compensatory fitness interactions [57], which suggests that mutations in RNA helices are individually deleterious but become neutral in appropriate combinations. The presence of significant excess of synonymous variation and clear influence of this variation on mRNA secondary structure stability suggests adaptive compensatory evolution in the *lbe* gene.

There are numerous studies devoted to synonymous site evolution in *Drosophila* [recent reviews in 43,58–60]. The main focus of these investigations is codon usage bias where different synonymous codons are used with different frequencies. It was shown that codon usage is tuned to optimize for expression and is adapted to tRNA pools of the organism. This is a type of purifying selection on synonymous sites preserving the usage of optimal codons [e.g., 42,45,61].

Analysis of the *Notch* locus in *D. melanogaster* has identified a region with accelerated synonymous site divergence [62–64]. The authors found an excess of fixed unpreferred codons and concluded that directional selection on synonymous sites had driven the fixation of these unpreferred codons. Later Holloway et al. [65] detected similar patterns for 64 genomic elements, a majority of which reside in protein-coding regions in the *D. melanogaster* genome. A genome-wide computational analysis showed that some unpreferred codons were fixed by directional selection in both bacteria and flies [66,67]. Thus selection on synonymous sites is not limited to the preferential fixation of mutations that enhance the speed or accuracy of translation because in some situations selection for unpreferred codons can impede translation efficiency. Neafsey and Galagan [66] found

that regulatory genes are particularly likely to be subject to selection for unpreferred codon usage. They suggested that low translational efficiency can be favored by reducing expressional noise through regulatory cascades [66]. Holloway et al. [65] hypothesized that ribosomal pausing for proper protein folding is a more tenable mechanism for explaining the preferable fixation of unpreferred codons than the alternative of reducing translation efficiency. However it was demonstrated that translational initiation of the ribosome can locally destabilize secondary structures and move along the mRNA without any significant delays [68] suggesting that the protein conformation alone cannot explain non-uniformity in translation elongation.

On the inter-specific level, we also found prevalent fixation of unpreferred codons in the *lbe* exon II (in *D. melanogaster* – *D. simulans* or *D. sechellia* comparisons, six out of eight synonymous changes lead to unpreferred codons). This pattern could not be explained by changes in mutation rates and/or low levels of recombination (see the section "Tests of Neutrality and Maximum Likelihood Analysis of Selective Pressures"). It is reasonable to suggest that directional selection on synonymous sites has driven the fixation of these unpreferred codons as in case with *Notch* and some other loci in *Drosophila* [62–67]. However intra-specific patterns of synonymous site variability in the *lbe* exon II suggest the involvement of balancing selection maintaining two different forms of mRNA molecules. Using site-specific codon models [69,70], specifically developed to analyze site-specific variation, we detected a few sites in the *lbe* exon II where, despite strong codon bias due to negative selection on the whole gene, we observed very high rates of synonymous change consistent with balancing selection on those precise sites. Thus we found a site-specific phenomenon that cannot be explained by the influence of codon usage bias alone because divergent *lbe* coding haplotypes could not be a by-product of the selection on codon bias. Our data indicate that different mechanisms are involved in evolution of synonymous sites in the *lbe* gene compared to loci that have recent acceleration synonymous site divergence like *Notch* and some others (see references above). We argue that *lbe* intra-specific synonymous polymorphism is due to balancing selection maintaining two mRNA forms that can provide necessary functional flexibility [51,58].

We detected contrasting patterns in nonsynonymous variation and rates in the *lbe* and *lbl* genes. While nonsynonymous mutations in *lbe* are prohibited, supposedly due to their strong detrimental effects on the LBE protein, three nonsynonymous mutations observed in *lbl* are predicted to have occurred under the influence of recurrent diversifying selection on the LBL protein. The level of the *lbl* nonsynonymous variability and d_N rate variation was much higher than its level of synonymous variability. The excess of replacement substitutions cannot be due to relaxation of functional constraints, because this region contains the homeobox region, which is highly conserved on a wide phylogenetic scale [17]. Moreover, strong selective constraints on the *lbl* coding region may be imposed by the alternative splicing described for this gene [17]. We suggest that the *lbl* gene may pass through evolutionary periods of functional transformation marked by the prevalence of replacement substitutions within *lbl* exon III, against a background of intensive negative selection.

Thus the two *lb* homeobox genes show contrasting patterns of nucleotide and codon evolution. Moreover, we have found a highly asymmetric distribution of significant binding-site clusters, with $>50\%$ of binding-site clusters specific for either *lbe* and *lbl*. The distinct binding-site profiles of *lbe* and *lbl* suggest that the genes are not redundant in developmental function. On the

contrary, Jagla et al. [17,18,20], assert that *lbe* and *lbl* are functionally redundant.

We have previously suggested that the pattern of nucleotide variability of the *Est-6* and *bap* coding regions in *D. melanogaster* are shaped by the influence of both directional and balancing selection [13,26,36]. Here, directional selection accounts for the excess of nearly identical sequences, and balancing selection prevents the complete fixation of haplotypes and increases the level of nucleotide variation. The present data show that both type of selection are involved in *lbe* and *lbl* evolution of *D. melanogaster*. A similar account has been proposed for the *Adh* and *TFL1* loci of *Arabidopsis thaliana* [71,72], the *Acp29AB* locus of *D. melanogaster* [73], the *Pan I* locus of Atlantic cod *Gadus morhua* [74], the MHC *DQB1* locus of marine mammals *Orcinus orca*, *Tursiops truncatus*, and *T. aduncus* [75], and the human *AVPR1B* gene [76]. Therefore, the operation and interaction of balancing and directional selection appears to be a general feature of molecular evolution in *Drosophila* and other eukaryote genomes.

The interaction between selective and neutral processes, nevertheless, should be cautiously interpreted given the modest sample size of sequences and the relatively short sequence lengths [e.g., 77]. Moreover, non-selective factors such as demography could partly account for the patterns of the polymorphisms. Demographic and selective forces shaping nucleotide polymorphism patterns in a species like *D. melanogaster* are difficult to disentangle because of its complicated history, including both recent worldwide migration and adaptation to drastically new environments [see, e.g., 78–80]. Patterns of polymorphism should be influenced by both of these evolutionary forces and is apparent in our data obtained for *Sod*, *Est-6*, ψ *Est-6*, *tin*, *bap*, *lbe*, and *lbl* located on the third chromosome from four natural *D. melanogaster* populations (Africa, Europe, North and South America). Comparative analysis showed significant peaks of variability in the *Est-6* region observed both in African and non-African samples, but dimorphic structure was detected only in non-African samples [26]. This observation supports the hypothesis that dimorphic haplotype structure is generated by demographic process during the recent history of *D. melanogaster* caused by admixture of differentiated populations. Significant peaks of increased nucleotide variability accompanied by peaks of LD and centered on the functionally important sites may reflect the effects of balancing selection [13,26,36,40,81] – this hypothesis was predicted by theoretical analysis [82–86].

Each gene family has its own evolutionary history. Consequently, a full understanding of their evolution may require comprehensive data be obtained for all multigene families in the genome. Distinguishing between demography and selection, or establishing the relative importance of these evolutionary factors in the patterning of molecular variation may not be sufficient to achieve a deeper understanding of the whole nature of molecular variation evolving under multidirectional evolutionary forces. Consequently, future investigations are needed in other species and genes of *Drosophila* in order to resolve these problems.

Materials and Methods

The *D. melanogaster* strains derive from wild flies collected in Europe (Barcelona, Spain; 19 strains), North America (California; 28 strains), and South America (Caracas, Venezuela; 19 strains). The strains were made fully homozygous for the third chromosome by crosses with balancer stocks, as described by Seager and Ayala [87]. Chung-I Wu kindly provided the isofemale *D. melanogaster* strains from East Africa (Sengwa, Zimbabwe). The three African strains included in our analysis were homozygous for

the *lbe* and *lbl* gene regions. The strains used in the present study were previously investigated for the β -*esterase* gene cluster [26,27,36] and the homeobox genes *tin* and *bap* [13]. The *lb* sequences of *D. melanogaster*, as well as those of *D. sechellia*, *D. simulans*, and *D. yakuba* were obtained from GenBank (accession numbers: NT_033777.2, AAKO01001614, AAEU02001386, AAEU02001382, AAGH01024581, and AAKO01001614).

DNA Extraction, Amplification, and Sequencing

Total genomic DNA was extracted using the tissue protocol of the QIAamp Tissue Kit (QIAGEN®). The primers used for the *lbe* PCR amplification reactions were: 5'-aacgtgctcgagata-acaatgacc-3' (forward primer) and 5'-agaagaaccatcgattgtaagaag-3' (reverse primer). The primers used for the *lbl* PCR amplification reactions were: 5'-atttcctgtacttggctgag-3' (forward primer) and 5'-tgttggcgaaatagtgaatatctg-3' (reverse primer). Methods are as previously described [36]. The sequences of both strands were determined for each line, using 12 overlapping internal primers spaced, on average, 500 nucleotides. At least two independent PCR amplifications were sequenced for each polymorphic site in all *D. melanogaster* strains to prevent possible PCR and sequencing errors. The GenBank accession numbers for the sequences are FJ754496 – FJ754564 and FJ754565 – FJ754633.

DNA Sequence Analysis

The *lbe* and *lbl* sequences were assembled using the program SeqMan (Lasergene, DNASTAR, Inc.). Multiple alignment was carried out manually and using the program CLUSTAL W [88]. The "sliding window" method of Hudson and Kaplan [83] and most intra-specific analyses were performed using DnaSP v. 4.10.9 [89] and PROSEQ v. 2.9 [90]. Departures from neutral expectations were investigated using HKA [91], Tajima's [35], McDonald and Kreitman's [92], Kelly's [33], and Wall's [34] tests. Simulations based on the coalescent process with or without recombination [93–95] were performed with DnaSP and PROSEQ to estimate the probabilities of the observed values of Tajima's *D*, Kelly's ζ_{NS} and Wall's *B* and *Q* statistics and confidence intervals of the nucleotide diversity values. Simulations with 10,000 replicates were conditional on the sample size, the observed number of segregating sites, and the DNA alignment length, with the population recombination rate parameter ($\rho = 4N_e\mu$) set to the gene estimates. The permutation approach of Hudson et al. [96] was used to estimate the significance of sequence differences between populations and haplotype families and the method of Sawyer [31] to detect gene conversion events. The population recombination rate was analyzed with a permutation-based approach [97].

Codon-Based Sequence Analyses

Probabilistic Markov codon-substitution models were fitted to coding alignments. Model parameters were estimated using maximum likelihood. These models measure selective pressure using the ratio of nonsynonymous to synonymous substitution rates $\omega = d_N/d_S$, which may vary among sites. Positive or negative selection is evidenced by significant deviations of the ω -ratio from 1. We used models that assume constant synonymous rates M0, M3, M7, M8 [98] and FMutSel0, FMutSel [99] as implemented in PAML v. 4 [100], and a model accounting for variability of synonymous rate over sites GYxHKY Dual GDD 3×3 of [70], later referred as M3-Dual and implemented in HYPHY [69]. Hypotheses concerning selection, codon bias, and rate variability were tested using likelihood ratio tests (LRTs). For a review on the application of codon models see [101]. Models combining coding and noncoding sequences were used to test for positive

selection on noncoding regions, as implemented in EvoNC [102]. The strength of selection on noncoding regions was measured by ζ , the ratio of the substitution rate in noncoding regions relative to the synonymous rate in coding regions. Under neutrality, these rates are expected to be similar ($\zeta \approx 1$). Significant deviations from 1 may be considered as evidence of positive ($\zeta > 1$) or negative ($\zeta < 1$) selection on noncoding regions. Consequently, the null model allowed two classes of sites in noncoding regions: a neutral class with $\zeta = 1$ and a class of sites evolving under negative selection where the average exonic synonymous rate was higher than the substitution rate in the noncoding regions ($\zeta < 1$). The alternative model also allowed two classes of sites, but the rate ratio was estimated for both classes under constraints: $\zeta \geq 1$ for positive and neutral selection class, and $\zeta < 1$ for the negatively selected class. A Bayesian approach was used to predict sites affected by positive selection in both coding and noncoding regions [70,98,102,103].

Supporting Information

Figure S1 DNA polymorphism in the *lbe* gene of 70 strains of *Drosophila melanogaster*. Symbols for strains: ER, El Rio; Ven, Venezuela; Bar, Barcelona; letters before and after the number refer to the electrophoretic allele observed in earlier studies at two loci: *esterase-6*, before the hyphen, and *superoxide dismutase*, after the hyphen (S, Slow; F, Fast; US, Ultra Slow). MEL, the *lbe* sequence obtained from GenBank (accession number, NT 033777.2). Lines are arranged successively according to genetic similarity. Numbers on top represent the position of segregating sites and the start of a deletion or insertion. Nucleotides are numbered from the beginning of our sequence. Coding regions of the genes are underlined below the top, reference sequence. Dots indicate same nucleotide as reference sequence. A hyphen represents deleted nucleotides. ▲ denotes a deletion; † absence of a deletion; ▼ insertion; ‡ absence of an insertion. Numbers after symbols for the deletions and insertions refer to the particular deletions and insertions. ▲1, a single nucleotide deletion of G (position 676); ▲2, a single nucleotide deletion of T (position 812); ▲3, a 5-bp deletion of TGGAA (position 1710–1714); ▲4, a 4-bp deletion of TAAA (position 1829–1832); ▲5, a 29-bp deletion of TTCAAATGAAG-GTGTTCGTATAATATCA (position 1876–1904); ▲6, a 30-bp deletion of TCGTATAATATCAATATCCAACTACA-A (position 1892–1921); ▼1, a 22-bp insertion of TAGTTGCTC-CATGTAACCATGT (position 1953–1974); ▲7, a 8-bp deletion of AGCAACTA (position 1975–1982); ▲8, a 2-bp deletion of AA (position 2006–2007); ▼2, a 10-bp insertion of TGATTTTTTT (position 2008–2017). Coordinates for functional regions of genes are: 1–950 (*lbe*, intron I), 951–1382 (*lbe*, exon II), 1383–2044 (*lbe*, 3'-flanking region). (DOC)

Figure S2 DNA polymorphism in the *lbl* gene of 70 strains of *Drosophila melanogaster*. Amino acid replacement polymorphisms are marked with asterisks (nucleotide under selection is in boldface). Coordinates for the functional regions of the gene are: 1–738 (*lbl*, intron I), 738–918 (*lbl*, exon II), 919–1523 (*lbl*, intron II), 1524–1766 (*lbl*, exon III), 1767–2438 (*lbl*, 3'-flanking region). ▼1, a single nucleotide insertion of A (position 973); ▲1, a 78-bp deletion of AATATATTTTTTTGCTGCAA-ATCTGCTGTTTTTCGCTTTTCTCAGCGAAAT-ATGTA-CTATTTTCAGTTAAAATATAAT (position 1059–1136); ▼2, a 12-bp insertion of ATAATAAAATAT (position 1132–1143); ▼3, a 24-bp insertion of AAAATATTAATTAATT-TATTATTA (position 1137–1160); ▲2, a 26-bp deletion of CACAAAAGA-

TATCCATTTCTG-GATAT (position 1161–1186); ▲3, a 237-bp deletion of CACAAAAGATATCCATTTCTGG-ATAT-GAATGAAGTGCATCTTATTCGGACTGACAATTTTGTAG-GGAAATGGTAGAGTCCGCTGGCAGTGACTATTTAT-TTTTCAGTCAACATGTATATAATGTGCATTGTTTTTC-TTTTGGCTGAGTAAATGTATCTCGAACTCGACTACAA-CTCTCTGTTGTTTTTTCTAACCATTTTGTCTAATGT-CAGCAAATTAATAAATATATGCGTCCT (position 1161–1398); ▲4, a single nucleotide deletion of T (position 1263); ▲5, a single nucleotide deletion of A (position 1349); ▲6, a 9-bp deletion of ACACCAGCA (position 1719–1727); ▲7, a 6-bp deletion of GCACCA (1728–1733); ▲8, a 32-bp deletion of CGTTCGCCGTTGAGAATAA-TCGTAAACCATTC (position 2347–2378). Other comments: see Figure S1. (DOC)

Figure S3 Fisher exact test of nonrandom associations between pairs of *lbe* and *lbl* polymorphisms. Singleton mutations are excluded from the analysis. Each box in the matrix represents the comparison of two polymorphic sites. Location of the segregating sites on *lbe* and *lbl* genes is shown on the diagonal, which indicates the position of the 5'-flanking, coding, and 3'-flanking regions. $0.01 < P < 0.05$ (grey); $0.001 < P < 0.01$ (dark grey); $P < 0.001$ (black). Intergenic associations are boxed. (TIF)

Table S1 Nucleotide diversity and divergence in the *lbe* gene region of *D. melanogaster*. (DOC)

Table S2 Nucleotide diversity and divergence in the *lbl* gene region of *D. melanogaster*. (DOC)

Table S3 Nucleotide diversity and divergence in the ladybird gene region of *D. melanogaster*. (DOC)

Table S4 Recombination estimates (ρ). (DOC)

Table S5 Gene conversion events in the *lbe* and *lbl* genes of *D. melanogaster*. (DOC)

Table S6 Linkage disequilibrium between functional regions of the *lbe* and *lbl* genes. (DOC)

Table S7 Kelly's (Kelly 1997) and Wall's (Wall 1999) tests of neutrality for the *lbe* and *lbl* gene regions. (DOC)

Table S8 Statistical significance of the binding site cluster for the *lbe* and *lbl* genes. (DOC)

Text S1 Nucleotide polymorphism. (DOC)

Text S2 Haplotype structure. (DOC)

Text S3 Linkage disequilibrium. (DOC)

Text S4 Neutrality and maximum likelihood (ML) analysis of selective pressures. (DOC)

Acknowledgments

We thank Elena Balakireva and Andrei Tatarenkov for encouragement and help; and three anonymous reviewers for valuable comments.

References

- Prud'homme B, Gompel N, Carroll SB (2007) Emerging principles of regulatory evolution. *Proc Natl Acad Sci U S A* 104(Suppl 1): 8605–8612.
- Wray GA (2007) The evolutionary significance of cis-regulatory mutations. *Nat Rev Genet* 8: 206–216.
- Carroll SB (2008) Evo-devo and an expanding evolutionary synthesis: A genetic theory of morphological evolution. *Cell* 134: 25–36.
- Lynch VJ, Wagner GP (2008) Resurrecting the role of transcription factor change in developmental evolution. *Evolution* 62: 2131–2154.
- Galant R, Carroll SB (2002) Evolution of a transcriptional repression domain in an insect Hox protein. *Nature* 415: 910–913.
- Ronshaugen M, McGinnis N, McGinnis W (2002) Hox protein mutation and macroevolution of the insect body plan. *Nature* 415: 914–917.
- Huff CD, Harpending HC, Rogers AR (2010) Detecting positive selection from genome scans of linkage disequilibrium. *BMC Genomics* 11: 8.
- Pavlidis P, Jensen JD, Stephan W (2010) Searching for footprints of positive selection in whole-genome SNP data from nonequilibrium populations. *Genetics* 185: 907–922.
- Zhong M, Lange K, Papp JC, Fan R (2010) A powerful score test to detect positive selection in genome-wide scans. *Europ J Human Genet* 18: 1148–1159.
- Clark AG, Eisen MB, Smith DR, Bergman CM, Oliver B, et al. (2007) Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature* 450: 203–218.
- Haerty W, Artieri C, Khezri N, Singh RS, Gupta BP (2008) Comparative analysis of function and interaction of transcription factors in nematodes: Extensive conservation of orthology coupled to rapid sequence evolution. *BMC Genomics* 9: 399.
- Moriyama EN, Powell JR (1996) Intraspecific nuclear DNA variation in *Drosophila*. *Mol Biol Evol* 13: 261–277.
- Balakirev ES, Ayala FJ (2004a) Nucleotide variation in the *tinman* and *bagpipe* homeobox genes of *Drosophila melanogaster*. *Genetics* 166: 1845–1856.
- Jovelin R, Dunham JP, Sung FS, Phillips PC (2009) High nucleotide divergence in developmental regulatory genes contrasts with the structural elements of olfactory pathways in *Caenorhabditis*. *Genetics* 181: 1387–1397.
- Purugganan MD (2000) The molecular population genetics of regulatory genes. *Mol Ecol* 9: 1451–1461.
- Kim Y, Nirenberg N (1989) *Drosophila* NK-homeobox genes. *Proc Natl Acad Sci USA* 86: 7716–7720.
- Jagla K, Bellard M, Frasch M (2001) A cluster of *Drosophila* homeobox genes involved in mesoderm differentiation program. *BioEssays* 23: 125–133.
- Jagla K, Jagla T, Heitzler P, Dretzen G, Bellard F, et al. (1997) *ladybird*, a tandem of homeobox genes that maintain late wingless expression in terminal and dorsal epidermis of the *Drosophila* embryo. *Development* 124: 91–100.
- Jagla K, Stanceva I, Dretzen G, Bellard F, Bellard M (1994) A distinct class of homeodomain proteins is encoded by two sequentially expressed *Drosophila* genes from the 93D/E cluster. *Nucl Acids Res* 22: 1202–1207.
- De Graeve F, Jagla T, Daponte J-P, Rickert C, Dastugue B, et al. (2004) The *ladybird* homeobox genes are essential for the specification of a subpopulation of neural cells. *Dev Biol* 270: 122–134.
- Manuel M, Le Parco Y (2000) Homeobox gene diversification in the calcareous sponge, *Sycon raphanus*. *Mol Phylogenet Evol* 17: 97–107.
- Lee SE, Gates RD, Jacobs DK (2003) Gene fishing: the use of a simple protocol to isolate multiple homeodomain classes from diverse invertebrate taxa. *J Mol Evol* 56: 509–516.
- Jagla K, Dollé P, Mattei M-G, Jagla T, Schuhbaur B, et al. (1995) Mouse *Lbx1* and human *LBX1* define a novel mammalian homeobox gene family related to the *Drosophila lady bird* genes. *Mech Dev* 53: 345–356.
- Dietrich S, Schubert FR, Healy C, Sharpe PT, Lumsden A (1998) Specification of the hypaxial musculature. *Development* 125: 2235–2249.
- Kanamoto T, Terada K, Yoshikawa H, Furukawa T (2006) Cloning and expression pattern of *lhx3*, a novel chick homeobox gene. *Gene Expr Patterns* 6: 241–246.
- Balakirev ES, Ayala FJ (2003) Nucleotide variation of the *Est-6* gene region in natural populations of *Drosophila melanogaster*. *Genetics* 165: 1901–1914.
- Balakirev ES, Ayala FJ (2004b) The β -*esterase* gene cluster of *Drosophila melanogaster*: Is ψ *Est-6* a pseudogene, a functional gene, or both? *Genetica* 121: 165–179.
- Xing Y, Lee C (2005) Evidence of functional selection pressure for alternative splicing events that accelerate evolution of protein subsequences. *Proc Natl Acad Sci U S A* 102: 13526–13531.
- Meireles-Filho ACA, Stark A (2009) Comparative genomics of gene regulation-conservation and divergence of cis-regulatory information. *Curr Opin Genet Dev* 19: 565–570.
- Hudson RR, Kaplan N (1985) Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics* 111: 147–164.
- Sawyer SA (1989) Statistical tests for detecting gene conversion. *Mol Biol Evol* 6: 526–538.
- Hudson RR, Bailey K, Skarecky D, Kwiatowski J, Ayala FJ (1994) Evidence for positive selection in the superoxide dismutase *Sod* region of *Drosophila melanogaster*. *Genetics* 136: 1329–1340.
- Kelly JK (1997) A test of neutrality based on interlocus associations. *Genetics* 146: 1197–1206.
- Wall JD (1999) Recombination and the power of statistical tests of neutrality. *Genet Res* 74: 65–79.
- Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123: 585–595.
- Balakirev ES, Balakirev EI, Ayala FJ (2002) Molecular evolution of the *Est-6* gene in *Drosophila melanogaster*: Contrasting patterns of DNA variability in adjacent functional regions. *Gene* 288: 167–177.
- Wayne ML, Simonsen K (1998) Statistical tests of neutrality in the age of weak selection. *Trends Ecol Evol* 13: 1292–1299.
- Nielsen R (2001) Statistical tests of selective neutrality in the age of genomics. *Heredity* 86: 641–647.
- Anisimova M, Bielawski JP, Yang Z (2001) Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. *Mol Biol Evol* 18: 1585–1592.
- Balakirev ES, Anisimova M, Ayala FJ (2006) Positive and negative selection in the β -*esterase* gene cluster of the *Drosophila melanogaster* subgroup. *J Mol Evol* 62: 496–510.
- Resch AM, Carmel L, Mariño-Ramírez L, Ogurtsov AY, Shabalina SA, et al. (2007) Widespread positive selection in synonymous sites of mammalian genes. *Mol Biol Evol* 24: 1821–1831.
- Parmley JL, Chamary JV, Hurst LD (2006) Evidence for purifying selection against synonymous mutations in mammalian exonic splicing enhancers. *Mol Biol Evol* 23: 301–309.
- Chamary JV, Parmley JL, Hurst LD (2006) Hearing silence: non-neutral evolution at synonymous sites in mammals. *Nat Rev Genet* 7: 98–108.
- Frydman J (2001) Folding of newly translated proteins in vivo: the role of molecular chaperones. *Annu Rev Biochem* 70: 603–647.
- Akashi H (1994) Synonymous codon usage in *Drosophila melanogaster*: natural selection and translational accuracy. *Genetics* 136: 927–935.
- Duret L (2002) Evolution of synonymous codon usage in metazoans. *Curr Opin Genet Dev* 12: 640–649.
- Reuter JS, Mathews DH (2010) RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinformatics* 11: 129.
- Ling SH, Cheng Z, Song H (2009) Structural aspects of RNA helicases in eukaryotic mRNA decay. *Biosci Rep* 29: 339–349.
- Klaff P, Riesner D, Steger G (1996) RNA structure and the regulation of gene expression. *Plant Mol Biol* 32: 89–106.
- Floris M, Mahgoub H, Lanet E, Robaglia C, Menand B (2009) Post-transcriptional regulation of gene expression in plants during abiotic stress. *Int J Mol Sci* 10: 3168–3185.
- Nackley AG, Shabalina SA, Tchivileva IE, Satterfield K, Korchynskiy O, et al. (2006) Human catechol-O-methyltransferase haplotypes modulate protein expression by altering mRNA secondary structure. *Science* 314: 1930–1933.
- Kudla G, Murray AW, Tollervey D, Plotkin JB (2009) Coding-sequence determinants of gene expression in *Escherichia coli*. *Science* 324: 255–258.
- Basak S, Mukhopadhyay P, Gupta SK, Ghosh TC (2010) Genomic adaptation of prokaryotic organisms at high temperature. *Bioinformation* 4: 352–356.
- Down TA, Bergman CM, Hubbard TJP (2007) Large-scale discovery of promoter motifs in *Drosophila melanogaster*. *PLoS Comput Biol* 3: e7. doi:10.1371/journal.pcbi.0030007.
- Papatsenko D (2007) ClusterDraw web server: a tool to identify and visualize clusters of binding motifs for transcription factors. *Bioinformatics* 23: 1032–1034.
- Kirby DA, Muse SV, Stephan W (1995) Maintenance of pre-mRNA secondary structure by epistatic selection. *Proc Natl Acad Sci U S A* 92: 9047–9051.
- Kimura M (1985) The role of compensatory neutral mutations in molecular evolution. *J Genet* 64: 7–19.
- Parmley JL, Hurst LD (2007) How do synonymous mutations affect fitness? *BioEssays* 29: 515–519.
- Hershberg R, Petrov DA (2008) Selection on codon bias. *Annu Rev Genet* 42: 287–299.
- Plotkin JB, Kudla G (2011) Synonymous but not the same: the causes and consequences of codon bias. *Nat Rev Genet* 12: 32–42.
- Akashi H (1995) Inferring weak selection from patterns of polymorphism and divergence at “silent” sites in *Drosophila* DNA. *Genetics* 139: 1067–1076.

62. Bauer DuMont VL, Fay JC, Calabrese PP, Aquadro CF (2004) DNA variability and divergence at the *Notch* locus in *Drosophila melanogaster* and *D. simulans*: A case of accelerated synonymous site divergence. *Genetics* 167: 171–185.
63. Bauer DuMont VL, Singh ND, Wright MH, Aquadro CF (2009) Locus-specific decoupling of base composition evolution at synonymous sites and introns along the *Drosophila melanogaster* and *Drosophila sechellia* lineages. *Genome Biol Evol* 2009: 67–74.
64. Nielsen R, Bauer DuMont VL, Hubisz MJ, Aquadro CF (2007) Maximum likelihood estimation of ancestral codon usage bias parameters in *Drosophila*. *Mol Biol Evol* 24: 228–235.
65. Holloway AK, Begun DJ, Siepel A, Pollard KS (2008) Accelerated sequence divergence of conserved genomic elements in *Drosophila melanogaster*. *Genome Res* 18: 1592–1601.
66. Neafsey DE, Galagan JE (2007) Positive selection for unpreferred codon usage in eukaryotic genomes. *BMC Evol Biol* 7: 119. DOI: 10.1186/1471-2148-7-119.
67. Singh ND, Bauer DuMont VL, Hubisz MJ, Nielsen R, Aquadro CF (2007) Patterns of mutation and selection at synonymous sites in *Drosophila*. *Mol Biol Evol* 24: 2687–2697.
68. Leibhaber SA, Cash FE, Shakin SH (1984) Translationally associated helix-destabilizing activity in rabbit reticulocyte lysate. *J Biol Chem* 259: 15597–15602.
69. Kosakovsky Pond SL, Frost SD, Muse SV (2005) HyPhy: hypothesis testing using phylogenies. *Bioinformatics* 21: 676–679.
70. Kosakovsky Pond SL, Muse SV (2005) Site-to-site variation of synonymous substitution rates. *Mol Biol Evol* 22: 2375–2385.
71. HanfStingl U, Berry A, Kellogg EA, Costa JT 3rd, W. Rüdiger W, et al. (1994) Haplotype divergence coupled with lack of diversity at the *Arabidopsis thaliana* alcohol dehydrogenase locus: roles for both balancing and directional selection? *Genetics* 138: 811–828.
72. Olsen KM, Womack A, Garrett AR, Suddith JI, Purugganan MD (2002) Contrasting evolutionary forces in the *Arabidopsis thaliana* floral developmental pathway. *Genetics* 160: 1641–1650.
73. Aguadé M (1999) Positive selection drives the evolution of the Acp29AB accessory gland protein in *Drosophila*. *Genetics* 152: 543–551.
74. Pogson GH (2001) Nucleotide polymorphism and natural selection at the Pantophysin *Pan I* locus in the Atlantic cod, *Gadus morhua* L. *Genetics* 157: 317–330.
75. Vassilakos D, Natoli A, Dahlheim M, Hoelzel R (2009) Balancing and directional selection at exon-2 of the MHC *DQB1* locus among populations of odontocete cetaceans. *Mol Biol Evol* 26: 681–689.
76. Cagliani R, Fumagalli M, Pozzoli U, Riva S, Cereda M, et al. (2009) A complex selection signature at the human *AVPR1B* gene. *BMC Evol Biol* 9: 123.
77. Jensen JD, Thornton KR, Aquadro CF (2008) Inferring selection in partially sequenced regions. *Mol Biol Evol* 25: 438–446.
78. Lachaise D, Silvain JF (2004) How two Afrotropical endemics made two cosmopolitan human commensals: the *Drosophila melanogaster*-*D. simulans* palaeogeographic riddle. *Genetica* 120: 17–39.
79. Haddrill PR, Thornton KR, Charlesworth B, Andolfatto P (2005) Multilocus patterns of nucleotide variability and the demographic and selection history of *Drosophila melanogaster* populations. *Genome Res* 15: 790–799.
80. Stephan W, Li H (2007) The recent demographic and adaptive history of *Drosophila melanogaster*. *Heredity* 98: 65–68.
81. Balakirev ES, Balakirev EI, Rodriguez-Trelles F, Ayala FJ (1999) Molecular evolution of two linked genes, *Est-6* and *Sod*, in *Drosophila melanogaster*. *Genetics* 153: 1357–1369.
82. Strobeck C (1983) Expected linkage disequilibrium for a neutral locus linked to a chromosomal arrangement. *Genetics* 103: 545–555.
83. Hudson RR, Kaplan N (1988) The coalescent process in models with selection and recombination. *Genetics* 120: 831–840.
84. Kaplan NL, Darden T, Hudson RR (1988) The coalescent process in models with selection. *Genetics* 120: 819–29.
85. Charlesworth D (2006) Balancing selection and its effects on sequences in nearby genome regions. *PLoS Genet* 2(4): e64. DOI: 10.1371/journal.pgen.0020064.
86. Charlesworth B, Nordborg M, Charlesworth D (1997) The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations. *Genet Res* 70: 155–174.
87. Seager RD, Ayala FJ (1982) Chromosome interactions in *Drosophila melanogaster*. I. Viability studies. *Genetics* 102: 467–483.
88. Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 22: 4673–4680.
89. Rozas J, Sánchez-DelBarrio JC, Messeguer X, Rozas R (2003) DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics* 19: 2496–2497.
90. Filatov DA (2002) PROSEQ: a software for preparation and evolutionary analysis of DNA sequence data sets. *Mol Ecol Notes* 2: 621–624.
91. Hudson RR, Kreitman M, Aguadé M (1987) A test of neutral molecular evolution based on nucleotide data. *Genetics* 116: 153–159.
92. McDonald JH, Kreitman M (1991) Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature* 351: 652–654.
93. Hudson RR (1983) Properties of a neutral allele model with intragenic recombination. *Theor Popul Biol* 23: 183–201.
94. Hudson RR (1990) Gene genealogies and the coalescent process. *Oxf Surv Biol* 7: 1–44.
95. Hudson RR (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* 18: 337–338.
96. Hudson RR, Boos D, Kaplan NL (1992) A statistical test for detecting geographic subdivision. *Mol Biol Evol* 9: 138–151.
97. McVean G, Awadalla P, Fearnhead P (2002) A coalescent-based method for detecting and estimating recombination from gene sequences. *Genetics* 160: 1231–1241.
98. Yang Z, Nielsen R, Goldman N, Pedersen AM (2000) Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* 155: 431–449.
99. Yang Z, Nielsen R (2008) Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. *Mol Biol Evol* 25: 568–579.
100. Yang Z (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* 24: 1586–1591.
101. Anisimova M, Kosiol C (2009) Investigating protein-coding sequence evolution with probabilistic codon substitution models. *Mol Biol Evol* 26: 255–271.
102. Wong WS, Nielsen R (2004) Detecting selection in noncoding regions of nucleotide sequences. *Genetics* 167: 949–958.
103. Yang Z, Wong WS, Nielsen R (2005) Bayes empirical bayes inference of amino acid sites under positive selection. *Mol Biol Evol* 22: 1107–1118.
104. Nei M (1987) *Molecular Evolutionary Genetics*. New York: Columbia University Press. 512 p.
105. Watterson GA (1975) On the number of segregating sites in genetical models without recombination. *Theor Popul Biol* 10: 256–276.