Accelerated sequence divergence of conserved genomic elements in *Drosophila melanogaster*

Alisha K. Holloway,[1,4] David J. Begun,[1] Adam Siepel,[2] and Katherine S. Pollard[3]

[1]*Department of Evolution and Ecology and Center for Population Biology, University of California, Davis, California 95691, USA;* [2]*Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, New York 14853, USA;* [3]*UC Davis Genome Center and Department of Statistics, University of California, Davis, California 95691, USA*

Recent genomic sequencing of 10 additional *Drosophila* genomes provides a rich resource for comparative genomics analyses aimed at understanding the similarities and differences between species and between *Drosophila* and mammals. Using a phylogenetic approach, we identified 64 genomic elements that have been highly conserved over most of the *Drosophila* tree, but that have experienced a recent burst of evolution along the *Drosophila melanogaster* lineage. Compared to similarly defined elements in humans, these regions of rapid lineage-specific evolution in *Drosophila* differ dramatically in location, mechanism of evolution, and functional properties of associated genes. Notably, the majority reside in protein-coding regions and primarily result from rapid adaptive synonymous site evolution. In fact, adaptive evolution appears to be driving substitutions to unpreferred codons. Our analysis also highlights interesting noncoding genomic regions, such as regulatory regions in the gene *gooseberry–neuro* and a putative novel miRNA.

[Supplemental material is available online at www.genome.org. Sequence data have been submitted to GenBank under accession nos. EU588685–EU588714.]

Comparative genomics approaches have assumed a central role in the identification of functionally important genomic regions (Kellis et al. 2003; Siepel et al. 2005; Xie et al. 2005; Birney et al. 2007). These approaches are based on the neutral theory prediction that sequences that have been highly conserved over tens of millions of years are either functionally important or are mutational cold spots (although no molecular mechanism for generating cold spots has been proposed). Recent population genetic analyses showed that low-frequency alleles are more common in highly conserved sequences, which supports the idea that such sequences, including those that do not encode proteins, are functionally constrained in multiple lineages (Drake et al. 2006; Asthana et al. 2007; Casillas et al. 2007; Katzman et al. 2007). On the other hand, questions remain about the functional importance of conserved sequences. For example, a recent functional analysis provided no evidence for strong viability selection against four conserved noncoding elements in mice (Ahituv et al. 2007).

The conceptual foundation linking conserved function with conserved sequence ignores the biologically interesting question of how biological functions evolve in different lineages. Indeed, from an evolutionary perspective, understanding the causes of rapid sequence evolution may be at least as interesting as understanding the causes of strong sequence conservation. Of particular relevance for identifying potential major functional changes is the identification of genomic regions that are highly conserved over most of a phylogeny, but that evolve very rapidly in at least one lineage. Such phylogenetically restricted rapid evolution could be due to a dramatic change in functional constraint, an increased mutation rate, or a shift in function, which drives large numbers of substitutions through populations under directional selection (Gillespie 1991).

[4]Corresponding author.
E-mail akholloway@ucdavis.edu; fax (530) 752-1449.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.077131.108. Freely available online through the *Genome Research* Open Access option.

Although the statistical analysis of heterogeneous rates of coding sequence evolution among lineages has a long history (Zuckerandl and Pauling 1962; Ohta and Kimura 1971; Langley and Fitch 1973, 1974), only recently have genome assemblies and alignments from multiple species (Blanchette et al. 2004; Clark et al. 2007; Stark et al. 2007) permitted such questions to be pursued in a comprehensive manner that is unbiased with respect to genomic feature. For example, Pollard et al. (2006) used alignments of multiple vertebrate species to identify genomic regions that are highly conserved in most vertebrates, but that have evolved rapidly in humans. These human accelerated regions (HARs) are candidates for contributing to human-specific biology. Interestingly, the majority of these regions were noncoding, and many were located near genes functioning in the nervous system. A more recent genomic analysis (Kim and Pritchard 2007) took a similar approach, but broadly investigated heterogeneous rates of evolution for conserved noncoding sequence across vertebrates. They concluded that short bursts of adaptive evolution drive divergence in conserved noncoding sequences.

The recent availability of multiple genome assemblies (Stark et al. 2007) and alignments (Karolchik et al. 2003, 2004; Blanchette et al. 2004) from *Drosophila* motivates an extension of such approaches to the *Drosophila* model for three main reasons. First, the experimental power of *Drosophila* opens up the possibility of detailed, in vivo functional investigation of candidate regions that are generally highly conserved but evolve rapidly in one lineage. Second, the genome organizations of flies and vertebrates are markedly distinct, with flies having much more compact genomes containing less noncoding DNA. This raises interesting questions as to whether the genomic distribution of lineage-specific increases in substitution rates in flies will also be concentrated in noncoding DNA, or whether differences in the biology and/or population genetics of flies and humans lead to different patterns. Finally, the *Drosophila melanogaster* genome is very well annotated, which facilitates targeted functional studies.

18:1592–1601 ©2008 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/08; www.genome.org

Comparison of functional annotations associated with lineage-specific rate increases in different lineages could provide clues as to potential generalities as well as unique biological functions exhibiting these unusual evolutionary patterns.

Results

Using whole-genome alignments of 10 *Drosophila* species to the *D. melanogaster* reference (Karolchik et al. 2003, 2004; Blanchette et al. 2004), we identified genomic regions that have been highly conserved over tens of millions of years, but show a recent acceleration in the rate of evolution solely along the *D. melanogaster* branch (Fig. 1A). Genomic regions were defined as conserved if they were 96% similar in sequence between *Drosophila simulans*, *Drosophila yakuba*, and *Drosophila erecta* and were at least 100 bp long. We identified 97,901 conserved regions with a mean (and median) length of 140 bp. Next, we assessed acceleration along the *D. melanogaster* branch using a likelihood ratio test (LRT) to compare two models of evolution over the *Drosophila* tree. The three species used to identify conserved regions (*D. simulans*, *D. yakuba*, and *D. erecta*) were excluded from this step in the analysis since, by definition, they were highly conserved. For each candidate region, the LRT compares the likelihood of the multiple alignments under a local null model with no acceleration in *D. melanogaster* to an alternative model with acceleration. There were 400 accelerated regions with an initial, unadjusted *P*-value < 0.05. Sixty-four of the conserved regions were determined to have significant acceleration along the *D. melanogaster* lineage after adjusting for multiple comparisons using the false discovery rate (FDR) (adjusted *P*-value < 0.05; Table 1). Hereafter, we refer to these as *D*rosophila *m*elanogaster *a*ccelerated *r*egions, or DMARs.

Accelerated rates of evolution could result from multiple single substitution events or they could result from microinversions that would cause a short region of sequence to appear to be rapidly diverged. An analysis of possible microinversions showed that only five substitution pairs could have resulted from this process, which only explains ~1% of all substitutions in DMARs. Therefore, the substitution process that leads to DMARs predominantly results from multiple single substitution events.

The 64 DMARs were dispersed fairly evenly throughout the major chromosome arms (Fig. 1B). Relative to the proportion of regions identified on the X chromosome as "conserved" in the first step of the analysis (10.5%), DMARs are significantly over-represented on the X chromosome ($n = 16$, FET two-tailed *P*-value = 0.0151). If DMARs are driven to fixation by directional selection, more efficient selection on the X chromosome could have led to this finding (for review, see Vicoso and Charlesworth 2006).

The majority of DMARs (72%) are found in protein-coding regions (Table 1). There were 46 DMARs in exons, nine in intergenic regions, eight in introns, and a single DMAR in a core promoter/5′ untranslated region (UTR). This distribution of DMARs among genomic features contrasts dramatically with regions in the human genome that show evidence of recent acceleration (HARs), which were found primarily in noncoding regions (Table 2; Pollard et al. 2006). The fact that the majority of HARs were found in noncoding regions may not be surprising considering that only 2% of the human genome is protein-coding. Flies have much more compact genomes, with almost 20% of the genome coding for proteins. However, even after considering genomic content in *Drosophila*, a significant excess of DMARs occur in protein-coding regions (see Table 2).

Protein-coding DMARs

DMARs in coding regions can be divided into two groups based on whether substitutions are found primarily at synonymous sites or nonsynonymous sites (Supplemental Table S1). DMARs with primarily synonymous substitutions ($DMAR_{SS}$) were defined as those with fewer than 25% of substitutions at amino acid

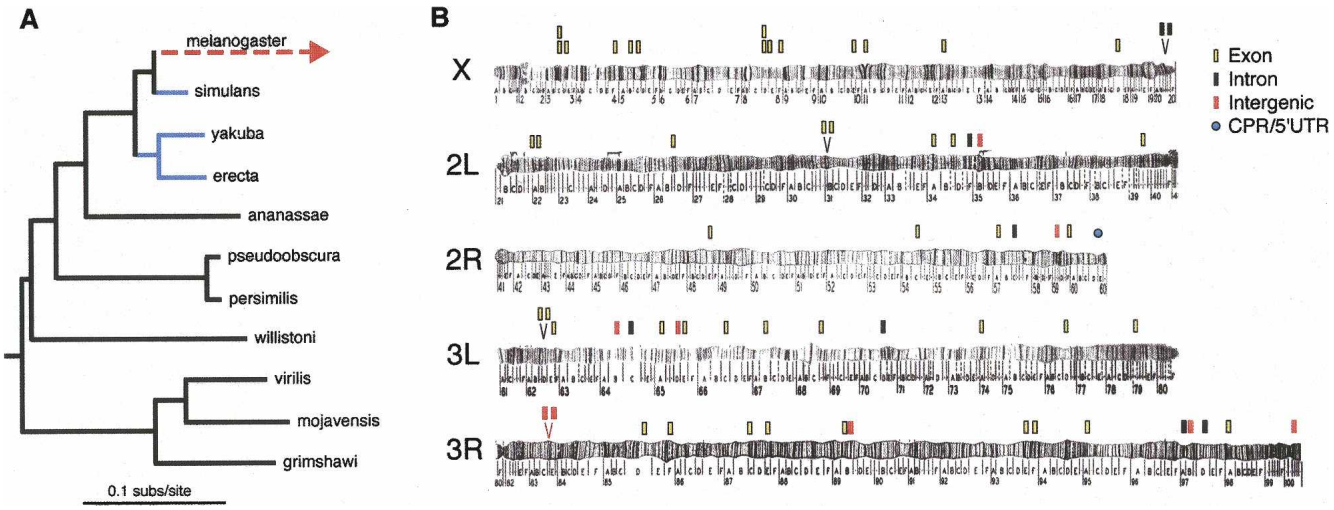

Figure 1. (*A*) Phylogeny of 11 *Drosophila* species with genome sequences. Branch lengths are derived from maximum likelihood analysis of all elements conserved throughout the tree. Branches in blue (*D. simulans*, *D. yakuba*, and *D. erecta*) were used to identify the blocks of at least 100 bp with 96% identity between the three species. All other lineages (and the *D. melanogaster–D. simulans* ancestor) were used to infer whether *D. melanogaster* had an accelerated rate of evolution relative to the expected rate of evolution based on elements conserved throughout the tree. (*B*) Locations of *D. melanogaster* accelerated regions (DMARs). (Stacked bars) Multiple DMARs within a single locus. (Two bars above a "V") Two DMARs that were within the same chromosomal band. DMARs are found predominately in exons (46/64) and are significantly over-represented on the X chromosome (16/64). Chromosome images adapted from Lefevre (1976).

Table 1. Summary of accelerated regions (DMARs)

Feature	Location	Gene name	CG	Rank	LRT	Adjusted *P*-value value	Length
CPR-5′ UTR	2R.20939795	*gsb-n*	*CG2692*	42	8.332	0.0189	157
Exonic-DMAR$_{AA}$	X.5635297	*l(1)G0060*	*CG3125*	27	9.386	0.0150	112
	X.9255013	*CG12139*	*CG12139*	9	11.836	$<1 \times 10^{-6}$	104
	2L.10062752	*CG4804*	*CG4804*	32	9.113	0.0150	122
	2L.10068212	*CG31714*	*CG31714*	58	7.681	0.0400	99
	2R.16552458	*CG9025*	*CG9025*	14	10.959	$<1 \times 10^{-6}$	100
	3L.2189910	*CG5707*	*CG5707*	18	10.355	0.0109	122
	3R.12072722	*Fmr1*	*CG6203*	4	13.094	$<1 \times 10^{-6}$	179
Exonic-DMAR$_{SS}$	X.3060523	*N*	*CG3936*	12	11.405	$<1 \times 10^{-6}$	118
	X.3062953	*N*	*CG3936*	34	8.825	0.0150	106
	X.3386279	*Gas8*	*CG14271*	25	9.394	0.0150	100
	X.5386061	*CG16752*	*CG16752*	24	9.460	0.0150	165
	X.5806638	*CG12236*	*CG12236*	39	8.481	0.0150	106
	X.9100917	*CG12119*	*CG12119*	1	15.191	$<1 \times 10^{-6}$	194
	X.9103460	*CG12119*	*CG12119*	59	7.569	0.0400	268
	X.9514729	*CG3099*	*CG3099*	50	8.052	0.0264	106
	X.11704400	*CG1578*	*CG1578*	23	9.530	0.0109	110
	X.11928697	*Cyp318a1*	*CG1786*	29	9.307	0.0150	116
	X.14887473	*eag*	*CG10952*	21	9.949	0.0109	130
	X.19565890	*Zw*	*CG12529*	22	9.594	0.0109	196
	2L.1181276	*CG4896*	*CG4896*	49	8.059	0.0264	108
	2L.1603227	*CG14351*	*CG14351*	15	10.950	$<1 \times 10^{-6}$	261
	2L.6293584	*Ddr*	*CG33531*	63	7.477	0.0400	100
	2L.13004891	*Tor*	*CG5092*	16	10.482	0.0109	110
	2L.13438707	*Tehao*	*CG7121*	52	8.002	0.0289	163
	2L.21637666	*Ac3*	*CG1506*	55	7.854	0.0323	120
	2R.8043497	*prp8*	*CG8877*	33	9.102	0.0150	168
	2R.13488870	*lack*	*CG4943*	26	9.393	0.0150	107
	2R.19484909	*apt*	*CG5393*	43	8.324	0.0189	220
	3L.2197793	*CG8960*	*CG8960*	19	10.343	0.0109	99
	3L.2666408	*CG16976*	*CG16976*	56	7.804	0.0344	141
	3L.6068855	*CG6610*	*CG6610*	11	11.544	$<1 \times 10^{-6}$	205
	3L.7267671	*unc-13-4A*	*CG32381*	38	8.558	0.0150	107
	3L.8350470	*Idh*	*CG7176*	20	10.195	0.0109	400
	3L.9468982	*Uch-L3*	*CG3431*	7	12.543	$<1 \times 10^{-6}$	99
	3L.11667361	*CG32085*	*CG32085*	40	8.400	0.0150	315
	3L.17338579	*Mip*	*CG6456*	13	11.092	$<1 \times 10^{-6}$	145
	3L.19832444	*kto*	*CG8491*	30	9.194	0.0150	178
	3L.21970046	*CG7470*	*CG7470*	46	8.244	0.0189	202
	3R.5352024	*CG8176*	*CG8176*	41	8.345	0.0189	127
	3R.8454604	*CG6359*	*CG6359*	31	9.146	0.0150	109
	3R.9209200	*CG8863*	*CG8863*	28	9.324	0.0150	104
	3R.11888878	*CG10185*	*CG10185*	44	8.287	0.0189	170
	3R.17348514	*CG7956*	*CG7956*	64	7.471	0.0436	129
	3R.17843301	*CG6439*	*CG6439*	48	8.066	0.0264	100
	3R.19536281	*CG10301*	*CG10301*	60	7.509	0.0400	110
	3R.24870375	*Pkc98E*	*CG1954*	37	8.577	0.0150	114
Intronic DMAR	X.22170917	*CG41476*	*CG41476*	2	14.542	$<1 \times 10^{-6}$	103
	X.22242399	*CG41475;fog*	*CG41475;CG9559*	10	11.725	$<1 \times 10^{-6}$	112
	2L.14011350	*CG33681*	*CG33681*	8	12.218	$<1 \times 10^{-6}$	148
	2R.17064171	*Pu*	*CG9441*	6	12.595	$<1 \times 10^{-6}$	101
	3L.5250477	*alan shepard*	*CG32423*	47	8.139	0.0264	128
	3L.14335683	*fz*	*CG17697*	35	8.698	0.0150	106
	3R.22145321	*CG33970*	*CG33970*	61	7.504	0.0400	116
	3R.22864033	*NepYr*	*CG5811*	3	14.080	$<1 \times 10^{-6}$	129
Intergenic DMAR	2L.14565705	—	—	57	7.775	0.0344	120
	2R.18747326	—	—	53	7.976	0.0295	117
	3L.4633878	—	—	36	8.632	0.0150	101
	3L.6932880	—	—	51	8.016	0.0289	111
	3R.1888158	—	—	5	12.823	$<1 \times 10^{-6}$	133
	3R.1966842	—	—	54	7.902	0.0295	100
	3R.12336598	—	—	62	7.480	0.0400	174
	3R.22247605	—	—	45	8.283	0.0189	102
	3R.27839557	—	—	17	10.398	0.0109	109

changing sites ($n = 39$); the remaining set (DMAR$_{AA}$) have at least 40% of substitutions at amino acid changing sites ($n = 7$). This arbitrary definition marks a natural break in the distribution of nonsynonymous substitution rates; DMARs defined as DMAR$_{AA}$ have high nonsynonymous substitution rates (0.0334–0.0692 substitutions/site) along the *D. melanogaster* lineage, whereas nonsynonymous substitution rates in DMAR$_{SS}$ are 0.0139–0.0200 substitutions/site (Fig. 2; Supplemental Table S1).

Table 2. Comparison of proportion of accelerated regions in coding and noncoding genomic regions in flies and humans

Species	Genomic region	Percent of genome (%)	Percent of conserved blocks (%)	Percent of accelerated regions (%)	FET *P*-value[a]
Human	Noncoding	98	80	98	5×10^{-5}
	Coding	2	20	2	
Fly	Noncoding	81	75	28	3×10^{-11}
	Coding	19	25	72	

[a]*P*-values from two-tailed tests comparing the percentage of conserved blocks and accelerated regions.

Acceleration of synonymous site divergence

The $DMAR_{SS}$, by definition, are evolving rapidly at synonymous sites in *D. melanogaster*, but slowly at amino acid sites—even in comparison to the gene in which they are found (Fig. 2; Supplemental Table S1). The genes that contain these $DMAR_{SS}$ are evolving slower at amino acid sites than the genomic average (Fig. 2A), while synonymous site evolution of $DMAR_{SS}$-containing genes is comparable to the genomic average (Fig. 2B). These data suggest that evolutionary rates of $DMAR_{SS}$ are not properties of genes, but of small regions within genes.

Rapid synonymous site divergence may indicate a shift in codon usage. Therefore, we examined codon usage in $DMAR_{SS}$, in the genes that contain them, and genome-wide. Our calculation of the number of substitutions to unpreferred codons was based on the mutational opportunity from preferred to unpreferred codons in the inferred ancestor of *D. melanogaster* and *D. simulans* (see Methods; Begun et al. 2007). We counted the number of substitutions from preferred to unpreferred codons and divided by the proportion of preferred codons in the inferred ancestor. Genes containing $DMAR_{SS}$ have more substitutions to unpreferred codons than do a random selection of genes in the genome (0.0565 vs. 0.0456; permutation test *P*-value = 0.002). Even more striking is the dramatic skew toward fixation of unpreferred codons in $DMAR_{SS}$ compared to the remainder of the gene (0.1689 vs. 0.0565; paired *t*-test; *P*-value = 0.0016). Accelerated synonymous site divergence in $DMAR_{SS}$ is attributable to fixation of many unpreferred variants.

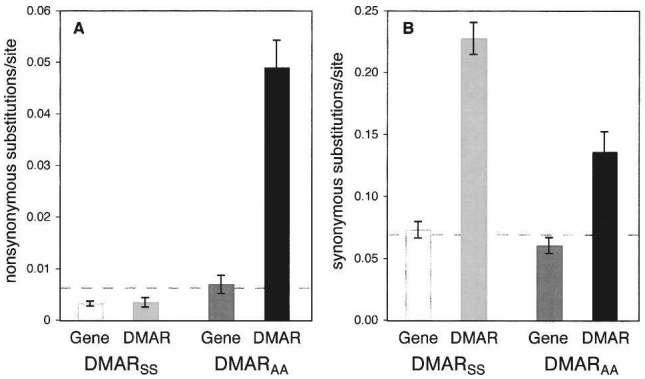

Figure 2. Nonsynonymous (*A*) and synonymous (*B*) substitution rates for DMARs in coding regions and the genes that contain those DMARs. Rates are per nonsynonymous site or synonymous site. Both $DMAR_{SS}$ and the genes that contain them (light gray and white bars) have very low levels of amino acid divergence. (Black and dark gray bars) $DMAR_{AA}$ have high rates of synonymous and nonsynonymous substitution, but the genes that contain them evolve at similar rates to the genomic average (dashed line).

Preferred codons typically end in guanine or cytosine. An overall mutational bias from G|C to A|T could explain increased substitution from preferred to unpreferred codons. Unless the mutational bias was extremely local, it would extend to introns of genes containing $DMAR_{SS}$ since they are intercalated among exons. In fact, several studies have found that G+C content was highly correlated between introns and third positions of codons (Kliman and Hey 1994; Heger and Ponting 2007; Vicario et al. 2007). For $DMAR_{SS}$, introns of $DMAR_{SS}$, and introns of all genes in the genome, we calculated the fraction of G|C to A|T substitutions by counting the number of G|C to A|T substitutions and divided that by the sum of all substitutions from ancestrally G|C nucleotides. The average fraction of G|C to A|T substitutions in introns of genes that contain $DMAR_{SS}$ was similar to the genome average (0.839 vs. 0.851, respectively). The $DMAR_{SS}$, on the other hand, have a significantly higher fraction of G|C to A|T substitutions than do the introns of the $DMAR_{SS}$-containing genes (0.931 vs. 0.839; paired *t*-test, *t*-statistic = 3.00, degrees of freedom [df] = 15, two-tailed *P*-value = 0.0089), which indicates that a gene-sized local mutational bias does not explain the rapid accumulation of unpreferred codons. This finding contrasts sharply with the substitution bias in HARs. In HARs, there was a preponderance of A|T to G|C substitutions, which indicates that biased gene conversion may be driving HAR substitutions.

A second hypothesis for the rapid synonymous site divergence in $DMAR_{SS}$ is that directional selection has fixed these substitutions. Recent work has shown that short introns (<80 bp) have very low levels of constraint (Halligan et al. 2004), which suggests they are composed primarily of neutral sites. In a modified version of the McDonald-Kreitman test (McDonald and Kreitman 1991), we compared ratios of polymorphism and divergence in short introns to synonymous sites in $DMAR_{SS}$. For six of the $DMAR_{SS}$ (two of which are in *Notch*), we have polymorphism data from the DPGP *D. melanogaster* resequencing project (http://www.dpgp.org/melanogaster/). We found that three out of six $DMAR_{SS}$ show a significant excess of synonymous site fixation, which suggests the action of directional selection (Table 3). We also performed this test on the remainder of the gene (without the DMAR) and found that four out of five show evidence of adaptive synonymous site evolution (Table 3). However, as noted previously, codon usage is significantly different between the $DMAR_{SS}$ and the remainder of the gene, with $DMAR_{SS}$ fixing significantly more unpreferred codons (paired *t*-test for six genes with polymorphism data; *P*-value = 0.0078). This difference in substitution pattern may indicate that different mechanisms of evolution are acting on synonymous sites in $DMAR_{SS}$ compared to synonymous sites in regions of the gene that do not have recent accelerations. The identification of $DMAR_{SS}$ may have drawn attention to a class of genes with multiple evolutionary pressures driving synonymous substitution.

In earlier work, one $DMAR_{SS}$-containing gene, *Notch*, was found to harbor a region with rapid synonymous site evolution that overlaps one of the $DMAR_{SS}$ (DuMont et al. 2004). In agreement with our findings for many $DMAR_{SS}$, intensive investigation of the *Notch* region with rapid synonymous site evolution led to the conclusion that directional selection was acting on synonymous sites (DuMont et al. 2004).

Table 3. Counts of polymorphic and fixed sites in $DMAR_{SS}$, $DMAR_{SS}$-containing genes, and introns of $DMAR_{SS}$ genes

			Nonsynonymous		Synonymous		Intron		
Feature	Gene	Codons	Poly	Fix	Poly	Fix	Poly	Fix	FET *P*-value[a]
$DMAR_{SS}$	*Gas8*	33	0	0	2	4	12	7	0.3500
	Notch (1)	38	0	0	1	5	310	197	0.0373
	Notch (2)	34	0	0	0	6	310	197	0.0037
	Tor	36	0	0	0	8	12	9	0.0089
	Tehao	53	0	0	2	4	17	13	0.3911
	CG16752	52	0	0	3	4	97	59	0.4313
Gene	*Gas8*	199	0	0	7	12	12	7	0.1939
	Notch	992	1	3	25	136	310	197	<0.0001
	Tor	1365	0	8	10	49	12	9	0.0011
	Tehao	440	1	4	11	25	17	13	0.0460
	CG16752	225	0	1	5	26	97	59	<0.0001

[a]FET *P*-values from comparisons of polymorphisms (poly) and fixations (fix) in synonymous sites and introns.

Acceleration of amino acid divergence

In genes that contain $DMAR_{AA}$, the rate of amino acid and synonymous site divergence is similar to the genomic average (Fig. 2). In contrast, the $DMAR_{AA}$ are evolving rapidly not only at amino acid changing sites (Fig. 2A), but also at synonymous sites (twofold higher than the genomic average) (Fig. 2B). The genes containing $DMAR_{AA}$ do not differ significantly from the genomic average with respect to substitutions to unpreferred codons (0.0588 vs. 0.0456; permutation test *P*-value = 0.067). The small sample size ($n = 7$) may increase variance in the permutation test and make rejecting the null hypothesis of no difference between $DMAR_{AA}$ genes and the genomic average difficult. Regardless, like $DMAR_{SS}$, the proportion of substitutions to unpreferred codons in $DMAR_{AA}$ is significantly higher than in the remainder of the gene (0.1384 vs. 0.0588; paired *t*-test, df = 6, *t*-statistic = 6.338, *P*-value = 7.2×10^{-4}).

In order to address whether directional selection may have acted to fix amino acid substitutions of $DMAR_{AA}$, we collected sequence data from *D. melanogaster* inbred lines for three genes [*Fmr1*, *l(1)G0060*, and *CG12139*] (Table 4). The $DMAR_{AA}$ and surrounding sequence for *Fmr1* and *CG12139* have very little polymorphism, which could indicate the action of recent directional selection. In fact, in comparison to the levels of synonymous polymorphism and divergence at the *Adh* locus (polymorphism data from the DPGP *D. melanogaster* resequencing project; http://www.dpgp.org/melanogaster/), there are fewer polymorphic synonymous sites than would be expected under a neutral model for both *Fmr1* and *CG12139* (Table 4; Hudson et al. 1987). For *l(1)G0060*, polymorphism relative to divergence was not significantly different from the neutral expectation.

Table 4. HKA tests for recent adaptive evolution near $DMAR_{AA}$

		Nonsynonymous		Synonymous		
Gene	Alleles	Poly	Fix	Poly	Fix	FET *P*-value[a]
CG12139	11	0	6	1	7	0.0421
Fmr1	10	0	7	1	6	0.0457
l(1)G0060	12	4	7	6	4	0.5966
Adh	34	2	1	10	2	

[a]FET *P*-values were from comparisons with *Adh*.

Ontology

Two biological processes (cell–cell signaling and cell communication) and two molecular functions (signal transducer activity and receptor activity) are over-represented among protein-coding genes containing DMARs (permutation test *P*-value < 0.01). The biological process signal transduction was also slightly over-represented (permutation test *P*-value = 0.038). There is notable overlap of genes among these terms. In fact, six genes are associated with at least four of these ontology terms (Supplemental Table S3). One other biological process, catabolism, is also significantly over-represented among DMARs in coding regions, but this ontology category does not overlap extensively with the aforementioned. Interestingly, catabolism and several specific types of receptor activity also appear to be enriched in the set of protein-coding genes with significantly accelerated amino acid evolution in *D. melanogaster* (see Table S21 in Begun et al. 2007). In comparison, in HARs, DNA binding and transcriptional regulation of genes near HARs were over-represented, which, once again, highlights the different biological processes and mechanisms that drive recent accelerations in the human and fly lineages. For DMARs, the biological significance of accelerated evolution in cell signaling genes is an interesting topic for future investigations.

DMARs in noncoding DNA

Intergenic and intron accelerated regions

Annotation of the *D. melanogaster* genome was used to determine the location of DMARs. Therefore, it is possible that the intergenic DMARs are actually protein-coding regions in other species and that *D. melanogaster* has lost one or more genes (or exons). The accelerated rate of evolution in a putatively intergenic region would then be due to relaxation of purifying selection in *D. melanogaster*. We investigated whether DMARs in intergenic regions were predicted to be protein-coding genes in *D. simulans*, *D. yakuba*, or *D. erecta* (Stark et al. 2007). In fact, none of the intergenic sequences were parts of predicted proteins in any of those three species. Additionally, we found that none of the DMARs fall within noncoding RNAs included in release 5.2 of the *D. melanogaster* annotation. However, two intergenic DMARs are near genes and may serve some *cis*-regulatory function. DMAR 2R.18747326 is 1009 bp from the 5′-UTR of *inaD*, and DMAR 3R.4633878 is 559 bp from the 3′-end of *CG13716*. There is no

annotated 3′-UTR for *CG13716*. It is possible that DMAR 3R.4633878 is part of the *CG13716* 3′-UTR given that the average length of 3′-UTRs in *Drosophila* is 318 bp and 3′-UTRs > 500 bp are not uncommon.

Intronic DMARs are found primarily in first introns (five of eight), and the remaining DMARs are in the largest introns of the gene. Introns often harbor regulatory elements, and it is possible that these DMARs serve some regulatory function. However, there are no known regulatory elements in intronic DMARs (FlyReg 2.0 [Bergman et al. 2005]; REDFly [Gallo et al. 2006]).

Intergenic and intronic DMARs could be unannotated noncoding RNAs. We took two approaches to address this question. First, we investigated whether whole-genome tiling-array experiments on total RNA (Stolc et al. 2004) revealed expression in the regions of any intronic or intergenic DMARs. In fact, two are expressed (DMAR X.22170917 and DMAR 3R.22145321). These expression profiles are based on total RNA; therefore, it is possible that unprocessed RNA was detected (Stolc et al. 2004). Second, we examined the predicted secondary structure of intergenic and intronic DMARs using EvoFold (Pedersen et al. 2006). All species for which there was available sequence for each DMAR were used in analyses. We also used RNAfold from the Vienna RNA package v1.6.4 (Hofacker et al. 1994) to compare the optimal secondary structures of *D. melanogaster* and *D. simulans*. Supplemental Figures S1–S17 show the optimal secondary structures as well as plots of base-pairing for the minimum free energy structure (lower left) and the probability of base-pairing (upper right) for each DMAR. Supplemental Table S2 shows numerical results from both EvoFold and RNAfold. There were four DMARs—X.22170917 (which also shows evidence of transcription), 3L.6932880, 3R.1888158, and 3R.1966842 (Supplemental Table S2; Supplemental Figs. S1, S9, S11, S12, respectively)—with high folding potential scores from the EvoFold analysis, which could indicate secondary structure. Predictions from RNAfold do not show any convincing secondary structure for three of these DMARs. However, in *D. melanogaster*, intergenic DMAR 3R.1966842 folds into a single hairpin (Supplemental Fig. S18), much like an miRNA, whereas *D. simulans* has a Y-shaped optimal structure (Supplemental Fig. S12). *Drosophila simulans* has a much weaker hairpin structure when forced onto the *D. melanogaster* optimal structure. Three substitutions along the *D. melanogaster* lineage increase complementary base-pairing in the hairpin (Supplemental Fig. S18).

Using miRScan (Lim et al. 2003a,b), we found that 3R.1966842 has significant potential for being an miRNA, with a total score (11.48) similar to the scores from known miRNAs in vertebrates and *Caenorhabditis elegans* and substantially higher than those of most non-miRNAs (Lim et al. 2003a,b). The Heidelberg RNA study (Hild et al. 2003) shows expression in the region of DMAR 3R.1966842, with the expressed probe residing within the DMAR. Population data from lines that are isogenic for chromosome 3 ($n = 75$) show that the DMAR sequence is fixed in *D. melanogaster*, except for three singleton polymorphisms that do not influence secondary structure (see Supplemental Fig. S18). Expression data are needed to validate whether the mature RNA is of the appropriate size to be considered a miRNA.

EvoFold (Pedersen et al. 2006) has been used to detect secondary structure throughout the genomes of flies (Stark et al. 2007), and these predictions are available as tracks on the UCSC Genome Browser (Karolchik et al. 2003, 2004). We found that there are nine EvoFold predictions that overlap with DMARs (Supplemental Table S4). However, only five of these have substitutions within the DMAR, and of these only one DMAR, X.22170886, which is contained within the intron of *CG41476*, has convincing secondary structure. Given this analysis and the analysis of entire DMAR sequences, it seems unlikely that substitutions within DMARs for changes in secondary structure would be a general driving force in the evolution of DMARs.

Acceleration in a regulatory region

The *gooseberry-neuro* (*gsb-n*) gene contains a DMAR in the core promoter (102 bp) that extends into the 5′ UTR (55 bp). This gene is a tandem duplicate and is transcribed in the opposite direction from its partner, *gooseberry*; both are transcription factors that are expressed during early development (Baumgartner et al. 1987; Gutjahr et al. 1993). The two genes have nonoverlapping regulatory modules (Li et al. 1993; Li and Noll 1994a,b), but do have partially redundant function; *gooseberry* regulates *gsb-n* and is able to perform the functions of *gsb-n* (Gutjahr et al. 1993). Both of these genes have well-characterized regulatory regions. Unfortunately, comparative expression data from *D. melanogaster* and *D. simulans* are not available for the appropriate developmental stage. Functional investigation of changes in the timing, levels, and spatial patterns of expression are warranted and will be a target of future studies.

Discussion

We identified 64 genomic regions that have been highly conserved over many millions of years, but that have recently experienced a burst of evolution along the *D. melanogaster* lineage. Protein-coding regions harbor the majority of DMARs, and rapid synonymous site evolution was the most common source of divergence. Synonymous site substitutions were overwhelmingly skewed toward unpreferred codons. We ruled out the possibility of a local mutation bias by comparing the substitution bias in $DMAR_{SS}$ and their associated introns. Comparisons of polymorphism and divergence in $DMAR_{SS}$ and nearby introns suggest that directional selection may be the driving force behind these rapid bursts of evolution at synonymous sites. An alternative hypothesis is that rapidly evolving mutation rates can explain these highly unusual genomic regions. In this scenario, $DMAR_{SS}$ and the population genetic evidence for their adaptive divergence could be explained by a recent increase in mutation rate and bias along the branch leading to *D. melanogaster*, followed by a second, more recent change back to ancestor-like mutation rates and patterns. The finely tuned requirements for the timing of these changes make this hypothesis less parsimonious, but given that these are some of the most unusual genomic regions in *D. melanogaster*, the possibility cannot be ruled out.

Rapidly evolving *D. melanogaster* genes often have lower levels of codon bias (Akashi 1994, 1995; Akashi et al. 2007), but, in general, this is not associated with adaptive evolution (Akashi 1995, 1996; Singh et al. 2007). In fact, fixation of unpreferred codons is attributed to the reduced efficacy of selection in *D. melanogaster* due to smaller population sizes (Akashi 1995, 1996; Vicario et al. 2007). However, a genome-wide computational analysis of unpreferred codon usage of mRNAs in flies, yeast, and bacteria showed that some unpreferred codons are fixed by directional selection in both bacteria and flies (Neafsey and Galagan 2007). Interestingly, in that study, none of the DMARs genes were identified as having evidence of directional selection acting on unpreferred codon usage. In a second genomic study (Singh et

al. 2007), only the *Notch* gene showed evidence of selection for unpreferred codon usage. Most likely, these analyses identify a different set of loci from our study because analysis of the entire gene would miss DMARs-like short stretches of unpreferred codon usage.

Prior intensive investigation at the *Notch* locus has identified regions with patterns of substitution similar to our findings for DMARs with rapid synonymous site evolution (DuMont et al. 2004; Nielsen et al. 2007). In fact, the *Notch* locus contained two DMARs with rapid synonymous site evolution, and one DMAR (X.3062953) is located in the region noted as the "3′ region" in DuMont et al. (2004). That study found an excess of unpreferred codon fixation and ruled out the possibility that changes in mutation rate and/or low levels of recombination could explain the pattern completely (DuMont et al. 2004). They concluded that directional selection on synonymous sites has driven the fixation of these unpreferred codons. Our results for *Notch* DMARs and synonymous site DMARs are in agreement with their findings.

Preferred codons are thought to be favored by selection on translational accuracy (e.g., fidelity of translation) (Akashi 1994), efficiency (e.g., tRNA abundance) (Akashi 2001, 2003), and/or robustness (e.g., proper folding despite mistranslation) (Drummond et al. 2005, 2006). In the case of translational efficiency, experimental work has shown that the use of unpreferred codons reduced the rate of translation in yeast (Purvis et al. 1987), *Drosophila* (Carlini and Stephan 2003), *Escherichia coli* (Parker 1989; Andersson and Kurland 1990; Komar et al. 1999), and humans (Kimchi-Sarfaty et al. 2007).

While unpreferred codons that reduce translational efficiency would typically be selected against, in some cases selection may act to reduce rates of protein translation (Konigsberg and Godson 1983; Purvis et al. 1987; Andersson and Kurland 1990; Thanaraj and Argos 1996; Komar et al. 1999). For example, protein folding often occurs before the completion of protein synthesis; pausing caused by the use of rare codons can allow for proper protein folding (Purvis et al. 1987; Thanaraj and Argos 1996; Komar et al. 1999). Directional selection may also act to fix unpreferred (or rare) codons to reduce translational efficiency and therefore protein levels (Konigsberg and Godson 1983; Andersson and Kurland 1990). The two hypotheses for how selection favors unpreferred codons make different predictions for the distribution of unpreferred codon usage. In cases of selection for reduced translational efficiency acting on overall protein abundance, we might expect unpreferred substitutions to be distributed throughout the mRNA. On the other hand, short segments of a coding region that use a high proportion of unpreferred codons may be more effective in causing sufficient ribosomal pausing at a particular position to induce proper folding (Purvis et al. 1987; Thanaraj and Argos 1996; Komar et al. 1999). That is, the physical proximity of unpreferred codons may have a multiplicative effect on translation rates (Purvis et al. 1987). One example of this phenomenon is in the pyruvate kinase gene in yeast. Five rare codons are used just before a predicted fold and are hypothesized to cause a pause in protein synthesis specifically at this location (Purvis et al. 1987). We hypothesize that ribosomal pausing for proper protein folding is a more tenable mechanism for explaining the abundance of DMARs with fixations of unpreferred codons than the alternative of reducing translation efficiency. Why the demands of protein folding would change between two closely related *Drosophila* species remains an open question.

Adaptive protein evolution is pervasive in *Drosophila* (Smith and Eyre-Walker 2002; Eyre-Walker 2006; Begun et al. 2007); thus, it is not surprising that two of three DMARs with multiple nonsynonymous substitutions showed evidence of directional selection in our population data. These $DMAR_{AA}$ are also evolving more rapidly at synonymous sites than the genome-wide average, which could be due to Hill-Robertson interference (Hill and Robertson 1966). Interference would reduce the efficacy of selection against unpreferred codons. However, this phenomenon is more commonly observed in regions of reduced recombination, and DMARs are not restricted to regions of reduced recombination.

Although the majority of DMARs are located in coding regions, this genomic study is unbiased with respect to genomic location and was not restricted to highlighting unusual patterns of evolution solely in known protein-coding regions. In fact, one interesting finding from this study has been the identification of a putative novel miRNA in *D. melanogaster*. Our study also identified several other genomic regions that will be the focus of future investigations, such as the core promoter and 5′ UTR of the *gooseberry-neuro* gene.

Conclusions

This comprehensive investigation of genomic elements that have been conserved over long periods of evolutionary time but that have had a recent burst of evolution in the *D. melanogaster* lineage suggests that DMARs may result from adaptive evolution. Intriguingly, many DMARs are attributable to recent accelerated synonymous site divergence and the accumulation of unpreferred codons. Population genetic evidence suggests that directional selection on synonymous sites plays a role in this phenomenon; though unusual, nonequilibrium mutational variation is not ruled out. Our findings reveal that DMARs contrast sharply in location, mechanism, and functional properties compared to HARs, which indicates that the biological and ecological differences between humans and flies are important factors in driving the evolutionary properties of genomes. Functional characterization of DMARs is now necessary to determine how radical changes in genotype are reflected in phenotype.

Methods

Genome alignments

MULTIZ alignments that were made in December 2006 using the *D. melanogaster* Release 5 assembly as the reference sequence (http://www.fruitfly.org/sequence/README.RELEASE5) were downloaded from the UCSC Genome Browser (Karolchik et al. 2003, 2004; Blanchette et al. 2004; http://genome.ucsc.edu/admin/cvs.html). The multiple alignments were generated from high-quality pairwise alignments produced by UCSC's chaining and netting pipeline (Kent et al. 2003), which uses conserved synteny to ensure orthology of aligned regions. Repeat regions and regions of low complexity were masked prior to alignment. The resulting 15-way alignments included *D. melanogaster*, *D. simulans*, *Drosophila sechellia*, *D. yakuba*, *D. erecta*, *Drosophila ananassae*, *Drosophila pseudoobscura*, *Drosophila persimilis*, *Drosophila willistoni*, *Drosophila mojavensis*, *Drosophila virilis*, and *Drosophila grimshawi*, as well as sequences of *Anopheles gambiae*, *Apis mellifera*, and *Tribolium castaneum*. We removed the *D. sechellia* sequence before analysis because of the low coverage of the genome. The mosquito, bee, and beetle sequences were also removed before analysis. We also deleted gaps that were inserted due to the non-fly and *D. sechellia* genome sequences and

removed any blocks that overlapped a known transposable element annotated in Release 5.1 of the *D. melanogaster* genome.

Identification of conserved regions

Conserved blocks were defined as those that were at least 100 bp long and had at least 96% sequence similarity between *D. simulans*, *D. yakuba*, and *D. erecta*. We used mafBlocker (Pollard et al. 2006) to identify conserved blocks. Conserved blocks that included sequence data from at least two additional species outside of the *melanogaster* subgroup were retained.

Assessment of significant acceleration

For all conserved blocks, we used likelihood ratio tests (LRTs) to determine whether the *D. melanogaster* branch had a significantly faster rate of evolution than expected. We excluded *D. simulans*, *D. yakuba*, and *D. erecta* from the LRT so that results were independent of the initial identification of the conserved regions. For the LRT, we used *D. melanogaster*, the inferred *D. melanogaster–D. simulans* ancestor, and at least two species from the following: *D. ananassae*, *D. pseudoobscura*, *D. persimilis*, *D. willistoni*, *D. mojavensis*, *D. virilis*, and *D. grimshawi*. The *D. melanogaster–D. simulans* ancestor was used as a node on the tree (with 0 branch length) so that we could ascribe evolutionary changes specifically to the *D. melanogaster* branch. The ancestral state was derived by a majority rule parsimony analysis of the *D. melanogaster*, *D. simulans*, and *D. yakuba* trio; instances of no majority were called "N."

For all conserved blocks, we used phyloFit to estimate two models of evolution (Siepel and Haussler 2004). The null model was derived by rescaling branch lengths from 15-species whole-genome MULTIZ alignments so that relative substitution rates remain constant across branches, but each conserved region has its own rate (branch lengths represented in Fig. 1A). Estimates of base frequencies and the substitution matrix were also taken from the combined 15-species whole-genome MULTIZ alignments. The alternative model included the same rescaling plus allowed the *D. melanogaster* branch to have an accelerated rate of evolution.

We assessed the statistical significance of regions identified as accelerated along the *D. melanogaster* branch by simulation using parametric bootstrapping. First, we generated 1 million alignments based on parameters from the 15-species whole-genome MULTIZ alignments. The simulated null alignments were 140 bp, which was the mean (and median) of the conserved regions we identified. The LRT statistic was then computed for each alignment. For each of the conserved elements we identified in step 1, the empirical *P*-value is equal to the proportion of simulated data sets with a larger LRT statistic. Based on the number of simulated data sets, the smallest *P*-value that can be estimated is $P = 1 \times 10^{-6}$. Empirical *P*-values were adjusted for multiple comparisons by the method of Benjamini and Hochberg (Benjamini and Hochberg 1995), which controls the false discovery rate (FDR). Any region with FDR adjusted *P*-value ≤ 0.05 was taken as having a significant acceleration along the *D. melanogaster* branch. We, therefore, expect that the proportion of false positives in this set is <5%. For each DMAR with an FDR adjusted *P*-value ≤ 0.05, we ensured that each DMAR was the reciprocal best BLAST hit with *D. simulans*.

Release 5.1 of the *D. melanogaster* annotation was used to determine whether DMARs were located in intergenic, coding, intron, or UTR sequence. Initial identification of conserved regions required that they were at least 100 bp long in *D. simulans*, *D. yakuba*, and *D. erecta*. Some DMARs may be shorter than 100 nt for two reasons. First, there may have been deletions along the *D. melanogaster* lineage. Second, DMARs were placed in categories (e.g., coding, intron, UTR) based on the location of the majority of nucleotides, which, in a small number of cases, resulted in a few conserved nucleotides being trimmed from one end. This only occurred when the DMAR was located primarily in a coding region, and estimates of polymorphism and divergence of synonymous and nonsynonymous sites would have been compromised by including noncoding nucleotides.

Molecular methods

D. melanogaster population data for *Fmr1*, *l(1)G0060*, and *CG12139* (n = 9–11 alleles) were from isofemale lines from Malawi, Africa. For population sampling of DMAR 3R.1966842 (n = 75), we used *D. melanogaster* lines that were collected by A.G. Clark (Maryland); these lines are isogenic for chromosome 3. DNA was PCR-amplified using Promega GoTaq Flexi DNA polymerase (Promega) for *Fmr1* and *CG12139* and AmpliTaq (Applied Biosystems) for *l(1)G0060* and DMAR 3R.1966842. PCR products were ligated into a PCR4 TOPO vector (Invitrogen). Ligations were transformed and plated, with the resulting colonies subjected to PCR using vector primers with AmpliTaq (Applied Biosystems). One clone was randomly selected from each line for sequencing. Colony PCR products were purified and sequenced at the University of California, Davis, College of Biological Sciences DNA Sequencing Facility. Sequences were submitted to GenBank under accession nos. EU588685–EU588714. Information for substitutions in the population sample of DMAR 3R.1966842 is in Supplemental Figure S18.

D. melanogaster population data for *Gas8*, *Notch*, *Tor*, *Tehao*, and *CG16752* were obtained from the *Drosophila* Population Genomics Project (http://www.dpgp.org/). DPGP data serve as a community resource and consist of 7 Mb of population data for 40 U.S. strains and 10 African strains that were resequenced using array-based sequencing technology (Affymetrix GeneChip CustomSeq Resequencing Arrays). Singleton single-nucleotide polymorphisms were eliminated before analysis. Data are available at http://www.dpgp.org/melanogaster.

Sequence analysis

D. melanogaster divergence from the inferred *D. melanogaster/D. simulans* ancestor was estimated using gestimator from the libsequence C++ library (Thornton 2003). The expected nucleotide heterozygosity (π) was estimated as the average pairwise difference between *D. melanogaster* alleles (Nei 1987; Weir 1990). For coding regions, the numbers of synonymous and nonsynonymous sites were counted using the method of Nei and Gojobori (1986). The pathway between two codons was calculated as the average number of synonymous and nonsynonymous changes from all possible paths between the pair. Substitutions to/from G|C from/to A|T were counted using the inferred *D. melanogaster*/*D. simulans* sequence described above. Substitutions to/from preferred and unpreferred codons in *D. melanogaster* were also estimated from the inferred *D. melanogaster*/*D. simulans* ancestor (Begun et al. 2007).

Polarized McDonald-Kreitman tests (McDonald and Kreitman 1991) used *D. melanogaster* polymorphism data and *D. simulans* and *D. yakuba* reference sequences to infer the *D. simulans*/*D. melanogaster* ancestral state. We took the conservative approach of using the pathway between codons that minimized the number of nonsynonymous substitutions along the *D. melanogaster* lineage. A Perl script for McDonald-Kreitman tests is available from the Corresponding Author. Hudson-Kreitman-Aguade tests (Hudson et al. 1987) were carried out using DnaSP version 4 (Rozas et al. 2003).

Gene Ontology

We used Gene Ontology terms from the Flybase Gene Ontology terms (http://flybase.org/genes/lk/function) in combination with the generic Gene Ontology Slim set of ontology terms (http://geneontology.org/GO.slims.shtml#avail). The proportion of genes containing a DMAR was calculated for each ontology term. We determined whether each ontology term had a higher proportion of genes with DMARs than would be expected from the empirical distribution. We derived the empirical distribution for each ontology term by drawing the same number of genes that were annotated with each term from all genes that were present in conserved blocks. We used only genes contained in blocks previously identified as conserved in case there was some bias present in the set of genes contained within conserved regions. We then calculated the proportion in the resampled data set that contained DMARs. We used 10,000 resampled data sets to derive the empirical distribution for each term.

Secondary structure analysis

We estimated the secondary structure of DMARs using EvoFold (Pedersen et al. 2006) and RNAFold (Hofacker et al. 1994). Additionally, we uploaded the coordinates of DMARs as a custom track on the UCSC Genome Browser to determine whether there were any predicted smaller regions of secondary structure that would not have been identified in examination of the secondary structure of the entire DMAR sequences.

EvoFold identifies functional RNA structures in multiple sequence alignments using a probabilistic model that takes into account evolutionary relationships between species in the alignment (Pedersen et al. 2006). RNAfold uses a dynamic programming algorithm to predict structures with minimum free energies and computes the equilibrium partition functions and base-pairing probabilities (Zuker and Stiegler 1981; McCaskill 1990; Hofacker et al. 1994).

Acknowledgments

We thank Angie Hinrichs at UCSC for providing the 15-way whole-genome MULTIZ alignments, Ryan Bickel and Mia Levine at UC Davis for valuable comments and suggestions, Melissa Eckert at UC Davis for laboratory advice and assistance, Elizabeth Milano and Umbreen Arshad at UC Davis for laboratory assistance, and Hiram Clawson at UCSC for help with installation of the Kent library. We also thank four anonymous reviewers for comments that improved this work. A.K.H. and D.J.B. were funded by NIH Grant R01-GM071926 to D.J.B.; A.S. was funded by NSF Faculty Early Career Development grant DBI-0644111.

References

Ahituv, N., Zhu, Y., Visel, A., Holt, A., Afzal, V., Pennacchio, L.A., and Rubin, E.M. 2007. Deletion of ultraconserved elements yields viable mice. *PLoS Biol.* **5:** e234. doi: 10.1371/journal.pbio.0050234.

Akashi, H. 1994. Synonymous codon usage in *Drosophila melanogaster*: Natural selection and translational accuracy. *Genetics* **136:** 927–935.

Akashi, H. 1995. Inferring weak selection from patterns of polymorphism and divergence at "silent" sites in *Drosophila* DNA. *Genetics* **139:** 1067–1076.

Akashi, H. 1996. Molecular evolution between *Drosophila melanogaster* and *D. simulans*: Reduced codon bias, faster rates of amino acid substitution, and larger proteins in *D. melanogaster*. *Genetics* **144:** 1297–1307.

Akashi, H. 2001. Gene expression and molecular evolution. *Curr. Opin. Genet. Dev.* **11:** 660–666.

Akashi, H. 2003. Translational selection and yeast proteome evolution. *Genetics* **164:** 1291–1303.

Akashi, H., Goel, P., and John, A. 2007. Ancestral inference and the study of codon bias evolution: Implications for molecular evolutionary analyses of the *Drosophila melanogaster* subgroup. *PLoS One* **2:** e1065. doi: 10.1371/journal.pone.0001065.

Andersson, S.G. and Kurland, C.G. 1990. Codon preferences in free-living microorganisms. *Microbiol. Rev.* **54:** 198–210.

Asthana, S., Noble, W.S., Kryukov, G., Grant, C.E., Sunyaev, S., and Stamatoyannopoulos, J.A. 2007. Widely distributed noncoding purifying selection in the human genome. *Proc. Natl. Acad. Sci.* **104:** 12410–12415.

Baumgartner, S., Bopp, D., Burri, M., and Noll, M. 1987. Structure of two genes at the gooseberry locus related to the paired gene and their spatial expression during *Drosophila* embryogenesis. *Genes & Dev.* **1:** 1247–1267.

Begun, D.J., Holloway, A.K., Stevens, K.S., Hillier, L.W., Poh, Y., Hahn, M.W., Nista, P.M., Jones, C.D., Kern, A.D., Dewey, C., et al. 2007. Population genomics: Whole-genome analysis of polymorphism and divergence in *Drosophila simulans*. *PLoS Biol.* **5:** e310. doi: 10.1371/journal.pbio.0050310.

Benjamini, Y. and Hochberg, Y. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Statist. Soc. Ser. B* **57:** 289–300.

Bergman, C.M., Carlson, J.W., and Celniker, S.E. 2005. *Drosophila* DNase I footprint database: A systematic genome annotation of transcription factor binding sites in the fruitfly, *Drosophila melanogaster*. *Bioinformatics* **21:** 1747–1749.

Birney, E., Stamatoyannopoulos, J.A., Dutta, A., Guigo, R., Gingeras, T.R., Margulies, E.H., Weng, Z., Snyder, M., Dermitzakis, E.T., Thurman, R.E., et al. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447:** 799–816.

Blanchette, M., Kent, W.J., Riemer, C., Elnitski, L., Smit, A.F., Roskin, K.M., Baertsch, R., Rosenbloom, K., Clawson, H., Green, E.D., et al. 2004. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.* **14:** 708–715.

Carlini, D.B. and Stephan, W. 2003. In vivo introduction of unpreferred synonymous codons into the *Drosophila Adh* gene results in reduced levels of ADH protein. *Genetics* **163:** 239–243.

Casillas, S., Barbadilla, A., and Bergman, C.M. 2007. Purifying selection maintains highly conserved noncoding sequences in *Drosophila*. *Mol. Biol. Evol.* **24:** 2222–2234.

Clark, A.G., Eisen, M.B., Smith, D.R., Bergman, C.M., Oliver, B., Markow, T.A., Kaufman, T.C., Kellis, M., Gelbart, W., Iyer, V.N., et al. 2007. Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature* **450:** 203–218.

Drake, J.A., Bird, C., Nemesh, J., Thomas, D.J., Newton-Cheh, C., Reymond, A., Excoffier, L., Attar, H., Antonarakis, S.E., Dermitzakis, E.T., et al. 2006. Conserved noncoding sequences are selectively constrained and not mutation cold spots. *Nat. Genet.* **38:** 223–227.

Drummond, D.A., Bloom, J.D., Adami, C., Wilke, C.O., and Arnold, F.H. 2005. Why highly expressed proteins evolve slowly. *Proc. Natl. Acad. Sci.* **102:** 14338–14343.

Drummond, D.A., Raval, A., and Wilke, C.O. 2006. A single determinant dominates the rate of yeast protein evolution. *Mol. Biol. Evol.* **23:** 327–337.

DuMont, V.B., Fay, J.C., Calabrese, P.P., and Aquadro, C.F. 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.

Eyre-Walker, A. 2006. The genomic rate of adaptive evolution. *Trends Ecol. Evol.* **21:** 569–575.

Gallo, S.M., Li, L., Hu, Z., and Halfon, M.S. 2006. REDfly: A regulatory element database for *Drosophila*. *Bioinformatics* **22:** 381–383.

Gillespie, J.H. 1991. *The causes of molecular evolution*. Oxford University Press, New York.

Gutjahr, T., Patel, N.H., Li, X., Goodman, C.S., and Noll, M. 1993. Analysis of the *gooseberry* locus in *Drosophila* embryos: *gooseberry* determines the cuticular pattern and activates *gooseberry neuro*. *Development* **118:** 21–31.

Halligan, D.L., Eyre-Walker, A., Andolfatto, P., and Keightley, P.D. 2004. Patterns of evolutionary constraints in intronic and intergenic DNA of *Drosophila*. *Genome Res.* **14:** 273–279.

Heger, A. and Ponting, C.P. 2007. Variable strength of translational selection among 12 *Drosophila* species. *Genetics* **177:** 1337–1348.

Hild, M., Beckmann, B., Haas, S.A., Koch, B., Solovyev, V., Busold, C., Fellenberg, K., Boutros, M., Vingron, M., Sauer, F., et al. 2003. An integrated gene annotation and transcriptional profiling approach towards the full gene content of the *Drosophila* genome. *Genome Biol.* **5:** R3. doi: 10.1186/gb-2003-5-1-r3.

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

Hofacker, I.L., Fontana, W., Stadler, P.F., Bonhoeffer, S., Tacker, M., and

Schuster, P. 1994. Fast folding and comparison of RNA secondary structures. *Monatsh. Chem.* **125:** 167–188.
Hudson, R.R., Kreitman, M., and Aguade, M. 1987. A test of neutral molecular evolution based on nucleotide data. *Genetics* **116:** 153–159.
Karolchik, D., Baertsch, R., Diekhans, M., Furey, T.S., Hinrichs, A., Lu, Y.T., Roskin, K.M., Schwartz, M., Sugnet, C.W., Thomas, D.J., et al. 2003. The UCSC Genome Browser Database. *Nucleic Acids Res.* **31:** 51–54.
Karolchik, D., Hinrichs, A.S., Furey, T.S., Roskin, K.M., Sugnet, C.W., Haussler, D., and Kent, W.J. 2004. The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.* **32:** D493–D496.
Katzman, S., Kern, A.D., Bejerano, G., Fewell, G., Fulton, L., Wilson, R.K., Salama, S.R., and Haussler, D. 2007. Human genome ultraconserved elements are ultraselected. *Science* **317:** 915.
Kellis, M., Patterson, N., Endrizzi, M., Birren, B., and Lander, E.S. 2003. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423:** 241–254.
Kent, W.J., Baertsch, R., Hinrichs, A., Miller, W., and Haussler, D. 2003. Evolution's cauldron: Duplication, deletion, and rearrangement in the mouse and human genomes. *Proc. Natl. Acad. Sci.* **100:** 11484–11489.
Kim, S.Y. and Pritchard, J.K. 2007. Adaptive evolution of conserved noncoding elements in mammals. *PLoS Genet.* **3:** 1572–1586.
Kimchi-Sarfaty, C., Oh, J.M., Kim, I.W., Sauna, Z.E., Calcagno, A.M., Ambudkar, S.V., and Gottesman, M.M. 2007. A "silent" polymorphism in the MDR1 gene changes substrate specificity. *Science* **315:** 525–528.
Kliman, R.M. and Hey, J. 1994. The effects of mutation and natural selection on codon bias in the genes of *Drosophila*. *Genetics* **137:** 1049–1056.
Komar, A.A., Lesnik, T., and Reiss, C. 1999. Synonymous codon substitutions affect ribosome traffic and protein folding during in vitro translation. *FEBS Lett.* **462:** 387–391.
Konigsberg, W. and Godson, G.N. 1983. Evidence for use of rare codons in the *dnaG* gene and other regulatory genes of *Escherichia coli*. *Proc. Natl. Acad. Sci.* **80:** 687–691.
Langley, C.H. and Fitch, W.M. 1973. The constancy of evolution: A statistical analysis of the alpha and beta haemoglobins, cytochrome *c*, and fibrinopeptide A. In *Genetic structure of populations* (ed. N.E. Morton), pp. 246–262. University of Hawaii Press, Honolulu.
Langley, C.H. and Fitch, W.M. 1974. An estimation of the constancy of the rate of molecular evolution. *J. Mol. Evol.* **3:** 161–177.
Lefevre, G. 1976. A photographic representation and interpretation of the polytene chromosomes of *Drosophila melanogaster* salivary glands. In *The genetics and biology of* Drosophila (eds. M. Ashburner and E. Novitski) pp. 31–66. Academic Press, London.
Li, X. and Noll, M. 1994a. Compatibility between enhancers and promoters determines the transcriptional specificity of *gooseberry* and *gooseberry neuro* in the *Drosophila* embryo. *EMBO J.* **13:** 400–406.
Li, X. and Noll, M. 1994b. Evolution of distinct developmental functions of three *Drosophila* genes by acquisition of different *cis*-regulatory regions. *Nature* **367:** 83–87.
Li, X., Gutjahr, T., and Noll, M. 1993. Separable regulatory elements mediate the establishment and maintenance of cell states by the *Drosophila* segment-polarity gene *gooseberry*. *EMBO J.* **12:** 1427–1436.
Lim, L.P., Glasner, M.E., Yekta, S., Burge, C.B., and Bartel, D.P. 2003a. Vertebrate microRNA genes. *Science* **299:** 1540.
Lim, L.P., Lau, N.C., Weinstein, E.G., Abdelhakim, A., Yekta, S., Rhoades, M.W., Burge, C.B., and Bartel, D.P. 2003b. The microRNAs of *Caenorhabditis elegans*. *Genes & Dev.* **17:** 991–1008.
McCaskill, J.S. 1990. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers* **29:** 1105–1119.
McDonald, J.H. and Kreitman, M. 1991. Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature* **351:** 652–654.
Neafsey, D.E. and Galagan, J.E. 2007. Positive selection for unpreferred codon usage in eukaryotic genomes. *BMC Evol. Biol.* **7:** 119. doi: 10.1186/1471-2148-7-119.
Nei, M. 1987. *Molecular evolutionary genetics*. Columbia University Press, New York.
Nei, M. and Gojobori, T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* **3:** 418–426.
Nielsen, R., Bauer DuMont, V.L., Hubisz, M.J., and Aquadro, C.F. 2007. Maximum likelihood estimation of ancestral codon usage bias parameters in *Drosophila*. *Mol. Biol. Evol.* **24:** 228–235.
Ohta, T. and Kimura, M. 1971. On the constancy of the evolutionary rate of cistrons. *J. Mol. Evol.* **1:** 18–25.
Parker, J. 1989. Errors and alternatives in reading the universal genetic code. *Microbiol. Rev.* **53:** 273–298.
Pedersen, J.S., Bejerano, G., Siepel, A., Rosenbloom, K., Lindblad-Toh, K., Lander, E.S., Kent, J., Miller, W., and Haussler, D. 2006. Identification and classification of conserved RNA secondary structures in the human genome. *PLoS Comput. Biol.* **2:** e33. doi: 10.1371/journal.pcbi.0020033.
Pollard, K.S., Salama, S.R., Lambert, N., Lambot, M.A., Coppens, S., Pedersen, J.S., Katzman, S., King, B., Onodera, C., Siepel, A., et al. 2006. An RNA gene expressed during cortical development evolved rapidly in humans. *Nature* **443:** 167–172.
Purvis, I.J., Bettany, A.J., Santiago, T.C., Coggins, J.R., Duncan, K., Eason, R., and Brown, A.J. 1987. The efficiency of folding of some proteins is increased by controlled rates of translation in vivo. A hypothesis. *J. Mol. Biol.* **193:** 413–417.
Rozas, J., Sanchez-DelBarrio, J.C., Messeguer, X., and Rozas, R. 2003. DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics* **19:** 2496–2497.
Siepel, A. and Haussler, D. 2004. Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. *Mol. Biol. Evol.* **21:** 468–488.
Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S., et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15:** 1034–1050.
Singh, N.D., Bauer Dumont, V.L., Hubisz, M.J., Nielsen, R., and Aquadro, C.F. 2007. Patterns of mutation and selection at synonymous sites in *Drosophila*. *Mol. Biol. Evol.* **24:** 2687–2697.
Smith, N.G. and Eyre-Walker, A. 2002. Adaptive protein evolution in *Drosophila*. *Nature* **415:** 1022–1024.
Stark, A., Lin, M.F., Kheradpour, P., Pedersen, J.S., Parts, L., Carlson, J.W., Crosby, M.A., Rasmussen, M.D., and Roy, S. 2007. Discovery of functional elements in 12 *Drosophila* genomes using evolutionary signatures. *Nature* **450:** 219–232.
Stolc, V., Gauhar, Z., Mason, C., Halasz, G., van Batenburg, M.F., Rifkin, S.A., Hua, S., Herreman, T., Tongprasit, W., Barbano, P.E., et al. 2004. A gene expression map for the euchromatic genome of *Drosophila melanogaster*. *Science* **306:** 655–660.
Thanaraj, T.A. and Argos, P. 1996. Ribosome-mediated translational pause and protein domain organization. *Protein Sci.* **5:** 1594–1612.
Thornton, K. 2003. Libsequence: A C++ class library for evolutionary genetic analysis. *Bioinformatics* **19:** 2325–2327.
Vicario, S., Moriyama, E.N., and Powell, J.R. 2007. Codon usage in twelve species of *Drosophila*. *BMC Evol. Biol.* **7:** 226. doi: 10.1186/1471-2148-7-226.
Vicoso, B. and Charlesworth, B. 2006. Evolution on the *X* chromosome: Unusual patterns and processes. *Nat. Rev. Genet.* **7:** 645–653.
Weir, B.S. 1990. *Genetic data analysis*. Sinauer, Sunderland, MA.
Xie, X., Lu, J., Kulbokas, E.J., Golub, T.R., Mootha, V., Lindblad-Toh, K., Lander, E.S., and Kellis, M. 2005. Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals. *Nature* **434:** 338–345.
Zuker, M. and Stiegler, P. 1981. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res.* **9:** 133–148.
Zuckerandl, E. and Pauling, L. 1962. Molecular disease, evolution, and genetic heterogeneity. In *Horizons in biochemistry* (eds. M. Kasha and B. Pullman), pp. 189–225. Academic Press, New York.

Received February 6, 2008; accepted in revised form June 19, 2008.