

Identifying Signatures of Selection at the *Enhancer of split* Neurogenic Gene Complex in *Drosophila*

Stuart J. Macdonald and Anthony D. Long

Department of Ecology and Evolutionary Biology, University of California, Irvine

The *Enhancer of split* gene complex (E(spl)-C) is one of the more highly annotated gene regions in *Drosophila*, and the 12 genes within the complex help determine the spacing and patterning of adult bristles. Any E(spl)-C coding, transcribed, or *cis*-regulatory regions experiencing nonneutral evolution are strong candidates to harbor polymorphisms contributing to naturally occurring variation in bristle number. We confirm that the E(spl)-C is strongly conserved and show that 74% of regulatory elements previously identified in *D. melanogaster* are conserved in *D. pseudoobscura*. Regulatory elements in enhancer regions show lower nucleotide diversity and more rare polymorphisms compared with adjacent nonregulatory DNA, suggesting they are under purifying selection, and these effects are particularly pronounced when considering only conserved regulatory elements. The ratio of polymorphism to divergence was significantly different between binding sites and nonbinding sites for transcription factors within enhancer regions, suggesting the action of some form of selection. Too few polymorphisms in regions of the 3' UTR harboring regulatory motifs prevents adequate comparison of diversity and the polymorphism frequency spectrum between 3' UTR motif and nonmotif sequence. We identified at least two broad regions of the gene complex showing strong population subdivision among four populations, which is suggestive of local adaptation or background selection. Finally, two regions of the E(spl)-C exhibit low nucleotide diversity, a high level of rare polymorphisms, and an increase in linkage disequilibrium, which together suggest the action of positive selection. Notably, the gene *m2* shows a significant deviation from neutrality by the McDonald-Kreitman test and resides in one of the two regions putatively experiencing a selective sweep. All sites in regions apparently visible to various selective forces are candidates for future work to determine their phenotypic effects.

Introduction

There is a growing understanding of the importance of *cis*-regulatory sequence variation on the expression of phenotypic traits (e.g., Stern 1998; Ludwig et al. 2000; Robin et al. 2002; Wittkopp, Vaccaro, and Carroll 2002). However, the relationship between molecular variation in regulatory regions, gene expression changes, and phenotypic variation is largely unknown (reviewed by Stern [2000] and Wray et al. [2003]). The number of sternopleural and abdominal bristles of *Drosophila melanogaster* have been employed as model characters for dissecting the genetic basis of variation in complex quantitative traits (reviewed by Mackay [1995]). These external mechanosensory bristles constitute the adult peripheral nervous system in *Drosophila*, and their stereotypical pattern is largely determined by genes of the Notch signaling pathway (reviewed by Jan and Jan [1994]). In this context, the genes of the Notch pathway are an excellent system for assessing the relationship between genotype and phenotype.

Genes of the *Enhancer of split* Complex (E(spl)-C) act at the end of the Notch signaling pathway, and in *D. melanogaster*, the complex harbors 12 transcription units: the E(spl)bHLH genes, *mδ*, *mγ*, *mβ*, *m3*, *m5*, *m7*, and *m8* are seven Notch-responsive basic helix-loop-helix (bHLH) transcription factors that act to repress neural cell fate (Delidakis and Artavanis-Tsakonas 1992; Knust et al. 1992; Jennings et al. 1994), and the E(spl)Brd genes, *mα*, *m2*, *m4*, and *m6* are four Bearded family genes, overexpression of which antagonizes Notch signaling activity (Lai et al. 2000; Lai, Bodner, and Posakony 2000). The single gene *m1* appears unrelated to the Notch pathway

and likely encodes a protease inhibitor of the Kazal family (Wurmbach, Wech, and Preiss 1999; Lai, Bodner, and Posakony 2000). The different transcripts show marked differences in imaginal disc gene expression patterns (de Celis et al. 1996; Singson et al. 1994), implying the genes are functionally separable. This is further evidenced by the preservation of E(spl)-C gene number and organization in *D. hydei* (Maier et al. 1993), a species that diverged from *D. melanogaster* around 60 MYA.

Upstream of each of the E(spl)-C genes, an array of activator/repressor protein-binding sites have been identified (Tietze, Oellers, and Knust 1992; Kramatschek and Campos-Ortega 1994; Eastman et al. 1997; Nellesen, Lai, and Posakony 1999; Lai et al. 2000; Lai, Bodner, and Posakony 2000). The variety of imaginal expression patterns exhibited by the endogenous E(spl)-C genes can be replicated purely by these short enhancer regions *in vivo*; that is, transcription is independently regulated for each gene, and transcriptional control is primarily local (Bailey and Posakony 1995; Nellesen, Lai, and Posakony 1999; Cooper et al. 2000).

In addition to *cis*-regulatory elements, several classes of 3' UTR regulatory motifs are also known from E(spl)-C genes (Leviten, Lai, and Posakony 1997; Lai, Burks, and Posakony 1998; Nellesen, Lai, and Posakony 1999; Lai et al. 2000; Lai, Bodner, and Posakony 2000; Lai 2002). These have been shown to negatively regulate transcript accumulation and elicit phenotypic changes in the number of adult bristles (Lai and Posakony 1997; Lai, Burks, and Posakony 1998). The motifs are perfectly complementary to a subset of microRNAs, and it is postulated that this posttranscriptional regulation may be mediated by the formation of RNA duplexes (Lai and Posakony 1998; Lai 2002).

Many of the identified *cis*-regulatory and 3' UTR motifs have been tested in functional assays, and the E(spl)-C locus is among a handful of loci, such as

Key words: regulatory DNA, enhancer, conservation, selection, population structure.

E-mail: sjm@uci.edu.

Mol. Biol. Evol. 22(3):607–619. 2004

doi:10.1093/molbev/msi046

Advance Access publication November 10, 2004

even-skipped (Ludwig et al. 2000), that are highly characterized with respect to regulatory domains in *Drosophila*. Given the demonstrable effects of E(spl)-C regulatory motifs on gene expression and on adult bristles, effects on natural phenotypic variation attributable to E(spl)-C may be caused by regulatory substitutions rather than changes in coding regions.

If phenotypic variation is largely caused by regulatory variants, then the relatively small proportion of noncoding DNA harboring functional regulatory variants must be distinguished from the much larger amount of non-functional noncoding sequence in any given genome. Using population-genetic and molecular-evolutionary approaches, regions of DNA visible to selection can be identified. Although not all phenotypic variation will be associated with regions showing evidence of past selection, we hypothesize that identified regions will be enriched for functional elements. This is because some fraction of functional regulatory regions may show footprints of past selection, whereas nonfunctional regions can never show such footprints.

The extensive annotation available for E(spl)-C provides an ideal opportunity to examine the pattern of molecular variation across functionally separable domains and determine whether known regulatory regions show evidence of selection. Because members of E(spl)-C likely influence variation in bristle number in adult flies (Long et al. 1995; Norga et al. 2003; Nuzhdin, Dilda, and Mackay 1999; Dilda and Mackay 2002), and because bristle number has been shown to be subject to stabilizing selection (García-Dorado and González 1996), nonneutrally evolving regions in E(spl)-C are more likely to harbor variants affecting bristle number than are neutrally evolving regions. Because any region of E(spl)-C deviating from neutral expectation harbored variants affecting gene function in the past, we speculate that such regions are more likely to harbor segregating functional variants that contribute to current standing variation than are tracts of noncoding DNA showing no evidence for past selection.

We examine the molecular evolution of the E(spl)-C complex in *Drosophila* at several levels to identify regions experiencing nonneutral evolution. First, we assess conservation of the locus, particularly regulatory sequences, between the diverged species *D. melanogaster* and *D. pseudoobscura*. Second, we compare nucleotide diversity within and between the sibling species *D. melanogaster* and *D. simulans*, focusing on differences between regulatory and adjacent nonregulatory sequence. Third, we examine variation in the level of population structure exhibited by different regions of the locus using the F_{ST} statistic. Fourth, we examine E(spl)-C within a single population for evidence of nonneutral evolution, including the action of positive selection. Sites located in regions identified by these population genetics methods are more likely to represent bristle number QTN (quantitative trait nucleotides) than sites in regions showing no evidence for a departure from neutrality, and this hypothesis can be tested in subsequent functional genetics studies of bristle number. Thus, this system holds promise for eventually characterizing the phenotypic effect of interesting DNA variants.

Sequence data from this article have been deposited with the EMBL/GenBank Data Libraries under the accession numbers AY779906 to AY779995.

Materials and Methods

Resequencing

The complete sequence for the third chromosome E(spl)-C locus was obtained for 16 third-chromosome extraction strains of *D. melanogaster*, where the natural chromosome was derived from Napa Valley, Calif., and made homozygous against a balancer chromosome (GenBank accession numbers AY779906 to AY779921) and two inbred strains of *D. simulans* (sim6, AY779922 and sim132, AY779923). The 48,512-bp alignment of these 18 alleles corresponds to positions 21809699 to 21857005 of the 3R *Drosophila melanogaster* Release 3.2.0 genome sequence and begins 3,120 bp 5' of the start codon of gene *mδ* and ends 1,055 bp 3' of the stop codon of gene *m8*.

Fifty-eight overlapping 1-kb PCR amplicons were developed to cover the entire approximately 47-kb region using the software PCR-Overlap (Rieder et al. 1998; all primer sequences available from <http://cstern.bio.uci.edu/pubs.htm>). Addition of 18-nt tails to the 5'-end of each oligo allowed sequencing of all fragments using just two common primers (universal forward: -21M13, 5'-TGTA-AAACGACGGCCAGT-3' and universal reverse: M13reverse, 5'-CAGGAAACAGCTATGACC-3' [Rieder et al. 1998]). All amplicons were directly sequenced using ABI Big Dye terminator chemistry on an ABI377 automated sequencer. Sequence traces for each allele were assembled using the program SeqManII version 5.01 (DNASTAR, Inc.), and the contigs manually aligned using BioEdit (<http://www.mbio.ncsu.edu/BioEdit/bioedit.html>).

We also sequenced nine of the 12 E(spl)-C coding regions in a single inbred line of *D. yakuba* (yakTAI27, AY779924 to AY779932).

Long Haplotype Analysis

The resequencing data showed two of the 16 alleles to be identical at all but one site across the entire E(spl)-C locus. To localize the ends of this potentially long haplotype, we sequenced approximately 700-bp PCR amplicons approximately 200 kb upstream (esU.200kb.F, 5'-TGGCAGCAATAAATGGATCA-3', esU.200kb.R, 5'-AATCCCAAACAAGCTGGATG-3'; GenBank accession numbers AY779964 to AY779979), 200 kb downstream (esD.200kb.F, 5'-ACTGAGCCAAAGCTGACGTT-3', esD.200kb.R, 5'-AATGTCTTCGGCTGGAATTG-3'; GenBank accession numbers AY779948 to AY779963), and 500 kb downstream (esD.500kb.F, 5'-CGTTGCATTAAGAGCAGCAA-3', esD.500kb.R, 5'-GGACGGAAACGAGAAACAAA-3'; GenBank accession numbers AY779980 to AY779995) of E(spl)-C.

To ascertain whether the long haplotype is associated with the nearby polymorphic cosmopolitan inversion In(3R)Payne (Bridges and Bridges 1938), we crossed each of the 15 extant lines to Canton-S and examined F_1

salivary gland polytene chromosomes using standard protocols.

Finally, to assess whether the presence of a pair of sequences differing by a single site in a set of 16 sequences harboring 1,013 segregating sites deviates from neutral expectation, we employed the haplotype test of Hudson et al. (1994). This procedure involves generating random samples of 16 sequences under a neutral coalescent model, each with 1,013 polymorphic sites, with a specified level of recombination. Each replicate sample is then tested for the presence of a pair of sequences that differ by one site or are identical, and the fraction of samples containing such a pair is an estimate of the P value for the observation.

Assessing E(spl)-C Conservation Using Blast

We used a sliding-window approach to Blast consecutive overlapping small subsections of a 47,677-bp *D. melanogaster* E(spl)-C consensus sequence derived from an alignment of the 16 sequenced alleles, against the homologous region of *D. pseudoobscura* (GenBank accession number AADE01000136, positions 561 to 58129) using the `bl2seq` utility. For each 31-bp *D. melanogaster* query sequence, we recorded the position, orientation, and score of the highest Blast hit in *D. pseudoobscura*, sliding through the region in 15-bp steps. Only Blast hits with scores above 45 were considered in further analyses.

Population Genetic Analyses

Estimates of nucleotide diversity and Tajima's D statistic (Tajima 1989) were performed using DnaSP version 4.0 (Rozas et al. 2003). MK G -tests (McDonald and Kreitman 1991) and tests analogous to the MK and HKA tests (Hudson, Kreitman, and Aguadé 1987) for regulatory sequences were performed on the counts of polymorphisms and fixed differences using a custom routine in the statistical programming language R (www.R-project.org). The probabilities of obtaining the observed values for Tajima's D statistic were determined by simulating neutral genealogies (Hudson 1990) using the program "ms" (Hudson 2002; <http://home.uchicago.edu/~rhudson1/source/mksamples.html>). Simulations were replicated 10,000 times, conditional on the empirical sample size, the observed number of segregating sites, and the alignment length in bp, with the population recombination rate parameter, ρ (or $4N_{or}$) set to the values 0, 1, 10, and 100. For the sliding-window analysis of Tajima's D statistic, we employed the Perlscript SCANMS (Ardell 2004; <http://www.lcb.uu.se/~dave/SCANMS>), using a window of 2 kb with 200-bp steps, which uses the coalescence simulator "ms" to generate probabilities while accounting for multiple testing.

We estimated the value of ρ (denoted by $\hat{\rho}_{w00}$) for the effectively haploid sequence data in a sliding-window framework based on the number of haplotypes and the minimum number of recombination events, as described in Wall et al. (2003). Only those biallelic SNPs and InDels identified by resequencing that had no missing data and showed the minor allele in at least two of the 16 sequenced alleles were used in the analysis (total = 477 sites, window size = 20 sites). We also estimated ρ (denoted by $\hat{\rho}_{H01d}$)

from unphased diploid genotype data (below) using the program RECLIDER (Wall et al. 2003; <http://genapps.uchicago.edu/labweb/index.html>), with a sliding window of 20 segregating sites and an initial estimate of $\rho = 0.01$.

Genotyping

A subset of the polymorphisms identified by resequencing were genotyped in four outbred population samples of *D. melanogaster* using an oligonucleotide ligation assay approach (described in Genissel et al. 2004). Genotyped sites were concentrated in and around the 12 E(spl)-C transcripts (40 in exons, eight in 5' UTRs, 22 in 3' UTRs, 12 in enhancers, and 35 in intergenic regions), and the set was selected such that the members showed minimal linkage disequilibrium (LD) with each other in the resequenced alleles. Population samples were (a) Napa Valley, Calif. ($N = 60$), (b) Southern France ($N = 46$), (c) Madang, Papua New Guinea ($N = 60$), and (d) Benin, West Africa ($N = 60$). The flies from Napa Valley were directly sampled from nature in 2001, whereas the other three samples were harvested from large populations maintained in laboratory cages with overlapping generations since inception from wild-caught individuals in 1999 (Southern France), 1998 (Madang), and 1970 (Benin).

Results

Conservation of the E(spl)-C Region

Using a Blast approach, we assessed E(spl)-C sequence conservation between our 47,677-bp *D. melanogaster* consensus sequence, derived from 16 sequenced alleles, and the homologous region from *D. pseudoobscura*. Previous work by Maier et al. (1993) examined conservation between *D. melanogaster* and *D. hydei* using Southern hybridization, but our methodology should be more sensitive to subtle differences between species.

Figure 1 highlights the similarity at E(spl)-C between *D. melanogaster* and *D. pseudoobscura*, species thought to have diverged 25 MYA (Russo, Takezaki, and Nei 1995). Overall the conservation is strong, although the region around *m1/m2* and the region from *m4* to *m8* both show fewer high Blast hits (score > 45) than the does rest of the locus, suggesting they have undergone greater evolutionary change. The *D. pseudoobscura* locus is slightly expanded relative to *D. melanogaster*, but otherwise, rearrangements appear to be absent, and virtually all the Blast hits are in the same direction (i.e., there are no inversions). Only 13 of the 433 hits with Blast scores above 45 were most similar to reverse complemented *D. pseudoobscura* sequences (nine of these 13 reverse complement hits exist in three coincident triplets, so only seven lines can be easily seen in figure 1). None of these represent identical-by-descent inversions: 12 are hits between sites in the bHLH areas of different, and oppositely transcribed, E(spl)bHLH genes, and one is between positions centered on Suppressor of Hairless – binding sites upstream of different E(spl)Brd genes (these binding sites can exist in either orientation).

Previous work has shown that functionally annotated motifs in *D. melanogaster* are conserved in *D. hydei* for some parts of E(spl)-C (3' UTR of *m4* [Lai and Posakony

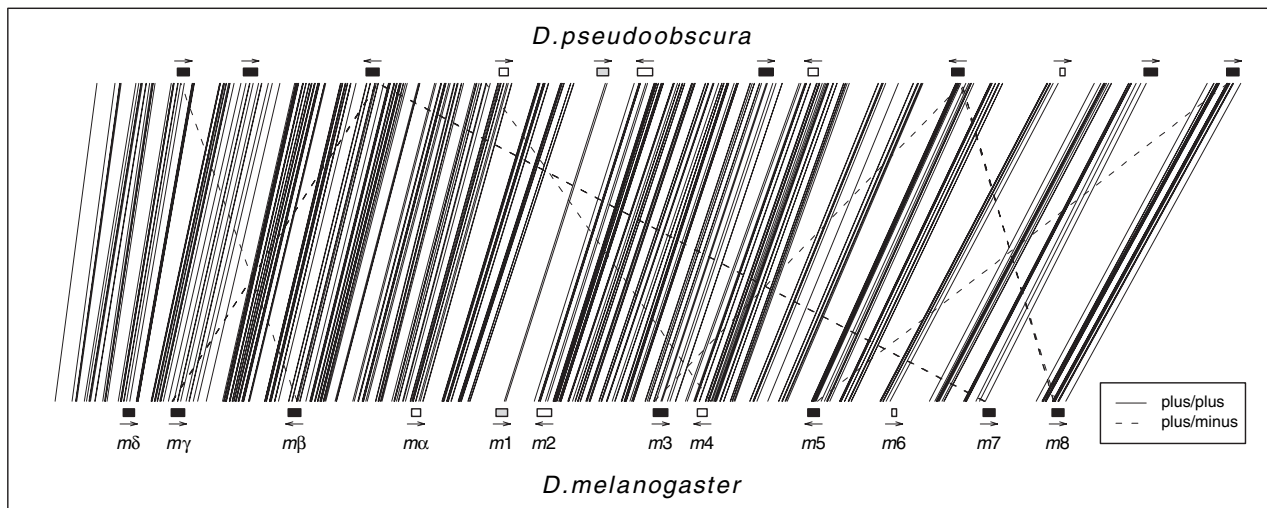


FIG. 1.—Conservation of the *Enhancer of split* complex sequence between *D. melanogaster* and *D. pseudoobscura*. Lines represent Blast hits (score > 45), showing the position in each genome (unbroken lines, hits between sequences in the same orientation; broken lines, hits between a sequence and the reverse complement of that sequence). Boxes are the positions of the translated portions of E(spl)-C (filled boxes, E(spl)bHLH genes; open boxes, E(spl)Brd genes; gray box, Kazal-type protease inhibitor), and arrows show the direction of transcription. See *Materials and Methods* for details of the Blast approach.

1997], 3' UTR of *m8* [Lai, Burks, and Posakony 1998], and enhancer regions of *mγ* and *m4* [Nellesen, Lai, and Posakony 1999]. A literature review yielded a large set of regulatory sequences initially identified in the *D. melanogaster* E(spl)-C (i.e., regulatory sequences were identified without recourse to sequence data from species other than *D. melanogaster*). Using positional information taken primarily from Nellesen, Lai, and Posakony's figure 2 (1999) and Lai, Burks, and Posakony's figure 2 (1998), we manually localized 46 3' UTR motifs and 136 upstream regulatory elements in our *D. melanogaster* E(spl)-C consensus sequence and detected 40 (87%) and 94 (69%) of these, respectively, in the *D. pseudoobscura* genome. This shows that regulatory elements in E(spl)-C are generally conserved across species.

Pattern of Nucleotide Diversity Across E(spl)-C

The extensive annotation of E(spl)-C allows us to parse the locus into separate categories and estimate population genetic parameters within each. Table 1 documents nucleotide diversity for various regions of the E(spl)-C locus using a 48,512-bp alignment of 16 *D. melanogaster* and two *D. simulans* alleles. For the coding regions, nucleotide diversity, π , within *D. melanogaster* is 0.0032, a value not inconsistent with those observed for other autosomal loci (Moriyama and Powell's table 1 [1996]). However, $\pi = 0.0047$ for the nonregulatory intergenic portions of the locus, which, although greater than the value for coding regions, is perhaps not as high as expected: Moriyama and Powell (1996) estimate $\pi = 0.0118$ averaged over noncoding regions for autosomal loci. Nucleotide divergence between *D. melanogaster* and *D. simulans* is also much lower for E(spl)-C ($K = 0.0379$) than values observed for X and 3R chromosome loci (Begun and Whitley 2000). A sliding-window analysis of nucleotide diversity shows that polymorphism is not

uniformly low in intergenic regions (fig. 2), and occasional peaks of within-species or between-species diversity exist. For instance, 5' to *mδ*, there is a peak of intraspecific nucleotide diversity, while just 3' to *mδ*, there is a peak of between-species diversity. There are also wide peaks of both within-species and between-species variation 5' to *m3*.

We calculated the value of Tajima's *D* statistic across E(spl)-C, which summarizes the frequency spectrum of observed polymorphisms within species. The complete locus shows $D = -0.688$ ($P > 0.05$ for $\rho = 0, 1$, and 10, $P < 0.01$ for $\rho = 100$), suggesting a slight skew towards rare sites, and this effect is heightened for transcribed regions (exons, $D = -1.155$; 5' UTR, $D = -0.917$; 3' UTR, $D = -0.883$). These values are within the range demonstrated for other loci (Moriyama and Powell 1996).

Under neutrality, the ratio of polymorphisms to fixed differences should be identical for both synonymous and nonsynonymous sites, and deviation from equality can be assessed using the MK test (McDonald and Kreitman 1991). Levels of diversity are generally too low for this test to be applied satisfactorily on a per-gene basis, and grouping the E(spl)bHLH and E(spl)Brd genes shows no significant difference from neutral evolution (table 2). Nonetheless, *m2* seems to exhibit a significant excess of fixed synonymous sites in the *D. melanogaster*–*D. simulans* comparison (MK *G*-test, $P = 0.017$) and approaches significance in the *D. melanogaster*–*D. yakuba* comparison (MK *G*-test, $P = 0.061$), which is suggestive of some form of selection acting on this gene. However, we note that neither test retains significance after correcting for multiple comparisons over genes. Over all the E(spl)-C genes, d_S is highest at *m2* ($m2$ $d_S = 0.145$, whereas for the other 11 genes, $0.035 < d_S < 0.130$), although the d_N/d_S ratio at *m2* is within the range observed for the other genes. Because *m2* does not markedly differ in codon usage from the other genes in any of the tested species (data not shown), the significant MK test at *m2*

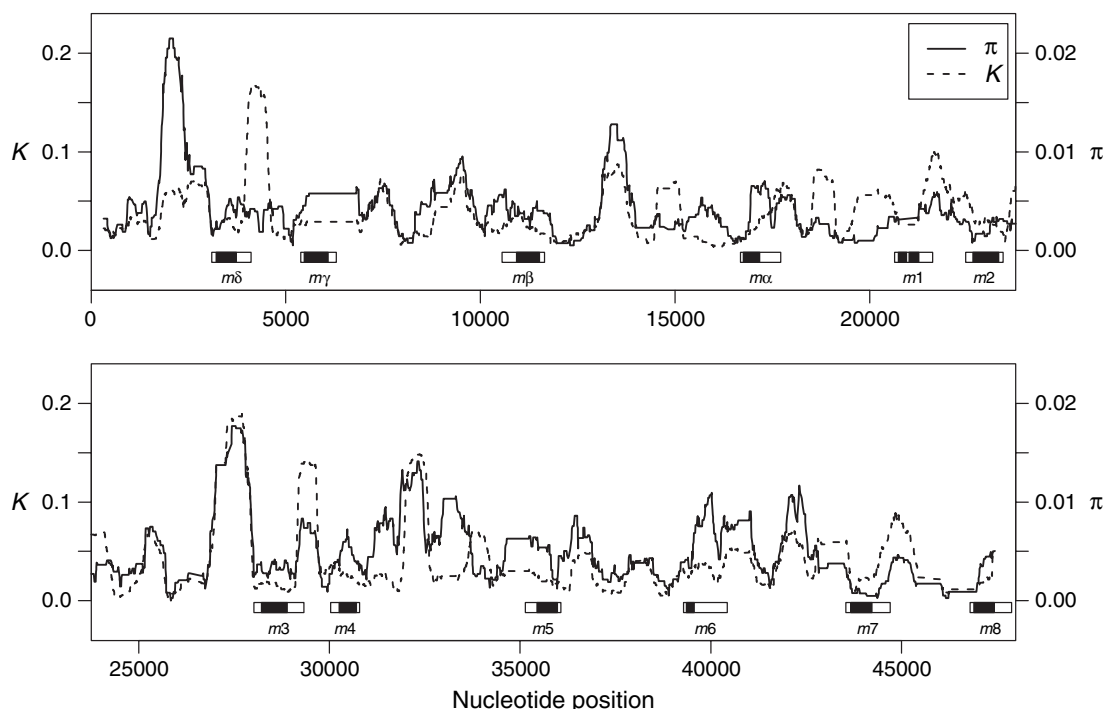


FIG. 2.—Nucleotide diversity across the *Enhancer of split* complex. A sliding-window analysis (500-bp window, 3-bp step) of within-species nucleotide diversity, π (unbroken line), and between-species divergence, K (broken line), was performed on the alignment of 16 *D. melanogaster* and two *D. simulans* alleles, using DnaSP version 4.0 (Rozas et al. 2003). Positions of exons (filled boxes) and UTRs (open boxes) are shown beneath the plots.

does not appear to be related to a change in optimal codon usage at this gene.

We were particularly interested in whether annotated regulatory elements—the 3' UTR motifs or the transcription factor-binding sites in the upstream enhancer modules—could be distinguished from surrounding sequence using statistical tests derived from population genetics theory. We sought to compare adjacent sets of regulatory and non-regulatory sequence, rather than to compare regulatory sequence with intergenic sequence, because the latter comparison is more likely to be confounded by differences in evolutionary history between the two test regions that may mask, or perhaps enhance, any distinguishing properties of regulatory sequence.

Table 1 shows that as a class, the 3' UTR motifs exhibit similar, although marginally lower, intraspecific diversity to the remaining, nonmotif, portions of the 3' UTRs, as well as reduced interspecific nucleotide divergence. This same pattern holds when comparing binding sites and nonbinding sites within the upstream enhancer modules. Considering only those annotated regulatory elements conserved between *D. melanogaster* and *D. pseudoobscura*, the disparity in both intraspecific and interspecific nucleotide diversity between regulatory and nonregulatory sequence increases, particularly for the enhancer module regions.

To test whether nucleotide diversity differs between regulatory and adjacent nonregulatory sequence, we compared the ratio of the number of segregating sites (table 1, S) to nonsegregating sites (table 1, Length - S) between these regions within *D. melanogaster*. The test is not significant for the 3' UTR sequences (all motifs, G -test,

$P = 0.085$; conserved motifs, G -test, $P = 0.083$) but is significant for the conserved enhancer-binding sites (all binding sites, G -test, $P = 0.333$; conserved binding sites, G -test, $P < 10^{-4}$), in the direction of lower diversity in binding sites. This difference either could be explained by selective constraint on enhancer-binding sites or could be generated by a lower mutation rate around these regulatory sites.

We adapted the MK and HKA tests (McDonald and Kreitman 1991; Hudson, Kreitman and Aguadé 1987) to test whether the ratio of fixed changes to polymorphisms is the same within regulatory and nonregulatory sequence sharing similar evolutionary history, as expected under neutrality (table 3). For the 3' UTR, there is no significant difference between regulatory and nonregulatory sequence (G -test, $P = 0.260$), whereas for the enhancer modules, there is a significant difference (G -test, $P = 0.004$), suggesting some form of selection is acting on the enhancer regions of *E(spl)-C* genes. Comparing conserved regulatory and nonregulatory sequence in the upstream enhancer modules eliminates the significant G -test (table 3), perhaps because of the low level of nucleotide diversity within conserved binding site sequences.

Too few polymorphisms and fixed differences may also explain the lack of a significant difference between 3' UTR motifs and nonmotifs. However, because UTRs are transcribed, nonmotif 3' UTR sequences are not necessarily nonfunctional, which may confound the motif versus nonmotif comparison.

Using the two tests described above, we also examined whether the enhancer regions of the *E(spl)-C*

Table 1
Nucleotide Diversity Throughout the *Enhancer of split* Complex Locus

| Region ^a | Length ^b | S^c | π^d | θ^d | K^e | D^f |
|---|---------------------|-------|---------|------------|--------|--------|
| Full locus | 38,920 | 690 | 0.0045 | 0.0054 | 0.0379 | -0.688 |
| Intergenic ^g | 15,567 | 288 | 0.0047 | 0.0056 | 0.0496 | -0.611 |
| Exons | 5,110 | 76 | 0.0032 | 0.0045 | 0.0218 | -1.155 |
| 5' UTRs | 1,134 | 22 | 0.0046 | 0.0059 | 0.0277 | -0.917 |
| 3' UTRs | 4,208 | 77 | 0.0044 | 0.0056 | 0.0356 | -0.883 |
| Motifs ^h | 727 | 8 | 0.0039 | 0.0033 | 0.0160 | 0.607 |
| Nonmotifs ^h | 2,792 | 56 | 0.0045 | 0.0060 | 0.0380 | -1.024 |
| Motifs (conserved) ^h | 661 | 7 | 0.0039 | 0.0032 | 0.0139 | 0.791 |
| Nonmotifs (conserved) ^h | 2,853 | 57 | 0.0045 | 0.0060 | 0.0380 | -1.023 |
| Enhancer modules ⁱ | 10,604 | 183 | 0.0046 | 0.0053 | 0.0264 | -0.470 |
| Binding sites ⁱ | 2,494 | 37 | 0.0033 | 0.0045 | 0.0169 | -1.029 |
| Nonbinding sites ⁱ | 8,330 | 147 | 0.0049 | 0.0054 | 0.0288 | -0.345 |
| Binding sites (conserved) ⁱ | 1,826 | 12 | 0.0012 | 0.0020 | 0.0080 | -1.602 |
| Nonbinding sites (conserved) ⁱ | 8,928 | 171 | 0.0053 | 0.0058 | 0.0298 | -0.385 |

^a Data generated using DnaSP version 4.0 (Rozas et al. 2003) based on the 48,512-bp alignment of 16 *D. melanogaster* and two *D. simulans* alleles. A set of binary vectors describing the parsing of the E(spl)-C region is available from <http://cstern.bio.uci.edu/pubs.htm>.

^b Length of sequence in base pairs, excluding alignment gaps, and regions showing any missing data.

^c The number of segregating sites within *D. melanogaster*.

^d Estimates of population-level heterozygosity.

^e Average proportion of nucleotide differences between *D. melanogaster* and *D. simulans*, corrected according to Jukes and Cantor (1969).

^f Tajima's (1989) *D* statistic.

^g The complete locus, excluding any sequence between the start of the most distal upstream regulatory element and the end of the 3' UTR for all genes.

^h Motifs = the sequence of each annotated 3' UTR motif including the 10 bp both upstream and downstream; nonmotifs = the remainder of the 3' UTR sequences, excluding those 3' UTRs that do not harbor annotated motifs; motifs (conserved) = only those annotated 3' UTR motifs (and 10-bp flanking sequences) that could be identified in *D. pseudoobscura*; nonmotifs (conserved) = the remainder of the 3' UTR sequences, including those 3' UTR motifs not conserved between *D. melanogaster* and *D. pseudoobscura*, but excluding those 3' UTRs that do not harbor annotated conserved motifs.

ⁱ Enhancer modules = the region between the start of the most distal annotated upstream regulatory element and the end of the most proximal element for each gene; binding sites = the sequence of each annotated element including 10 bp both upstream and downstream; nonbinding sites = the remainder of the enhancer module sequences; binding sites (conserved) = the sequence of those binding sites (and 10-bp flanking sequences) that could be detected in *D. pseudoobscura*; nonbinding sites (conserved) = the remainder of the enhancer module sequence, including those elements not conserved in *D. pseudoobscura*.

genes could be distinguished from intergenic sequence. Nucleotide diversity did not differ significantly between the enhancers and intergenic regions, but there was a significant difference in the ratio of fixed changes to polymorphisms (*G*-test, $P = 0.007$). This difference appears to be almost entirely related to the binding sites

in the enhancers, as $P = 0.098$ for intergenic versus enhancer-nonbinding regions, and $P = 0.0002$ for intergenic versus enhancer-binding regions.

The polymorphism spectrum, as summarized by Tajima's *D* statistic, also appears to differ between regulatory and nonregulatory DNA: *D* is positive for 3' UTR

Table 2
MK Tests for *Enhancer of split* Coding Regions

| Gene | P_S^a | P_R^a | <i>D. simulans</i> (sim6) | | | | <i>D. yakuba</i> (yakTAI27) | | | |
|-------------------|----------------|---------|---------------------------|---------|----------------------------|---|-----------------------------|---------|----------------------------|--------------------------------|
| | | | F_S^a | F_R^a | FET P value ^b | MK <i>G</i> -test ^c (P value) | F_S^a | F_R^a | FET P value ^b | MK <i>G</i> -test ^c |
| bHLH ^d | 42 | 12 | 42 | 14 | 0.824 | 0.118 (0.732) | 85 | 14 | 1.000 | 0.188 (0.665) |
| Brd ^d | 14 | 8 | 31 | 6 | 0.114 | 3.015 (0.082) | 50 | 15 | 0.256 | 1.743 (0.187) |
| <i>mδ</i> | 8 | 2 | 8 | 2 | 1.000 | 0.000 (1.000) | 27 | 5 | 1.000 | 0.102 (0.750) |
| <i>mγ</i> | 3 ^e | 0 | 4 | 1 | 1.000 | NA | 16 | 2 | 1.000 | NA |
| <i>mβ</i> | 6 ^e | 0 | 9 | 1 | 1.000 | NA | 19 | 4 | 1.000 | NA |
| <i>mα</i> | 5 | 0 | 3 | 3 | 0.182 | NA | 7 | 5 | 0.245 | NA |
| <i>m1</i> | 1 | 3 | 5 | 2 | 0.242 | 2.284 (0.131) | 12 | 4 | 0.101 | 3.404 (0.065) |
| <i>m2</i> | 3 ^e | 4 | 17 | 2 | 0.028 | 5.743 (0.017) | 25 | 9 | 0.075 | 3.510 (0.061) |
| <i>m3</i> | 11 | 2 | 6 | 0 | 1.000 | NA | — | — | — | — |
| <i>m4</i> | 5 | 2 | 9 | 0 | 0.175 | NA | 14 | 1 | 0.227 | 1.802 (0.179) |
| <i>m5</i> | 6 | 2 | 7 | 4 | 1.000 | 0.281 (0.596) | — | — | — | — |
| <i>m6</i> | 1 | 2 | 2 | 1 | 1.000 | 0.680 (0.410) | 4 | 0 | 0.143 | NA |
| <i>m7</i> | 2 | 0 | 5 | 4 | 0.491 | NA | 23 | 3 | 1.000 | NA |
| <i>m8</i> | 6 | 6 | 3 | 2 | 1.000 | 0.142 (0.706) | — | — | — | — |

^a Manual counts of silent and replacement polymorphisms and fixed differences in each gene. Sites within gaps and nonbiallelic sites were excluded. Dash (—) indicates those genes not sequenced in the *D. yakuba* strain, yakTAI27.

^b P value from a Fisher's exact test.

^c McDonald-Kreitman *G*-test (McDonald and Kreitman 1991). Test is undefined for genes marked NA.

^d Tests conducted using sites from all seven E(spl)bHLH genes (four E(spl)bHLH genes for *D. melanogaster*-*D. yakuba* comparison) or all four E(spl)Brd genes. For the *D. melanogaster*-*D. yakuba* tests, *D. melanogaster* polymorphism counts are reduced from the presented values: E(spl)bHLH, $P_S = 17$ and $P_R = 2$; E(spl)Brd, $P_S = 13$ and $P_R = 8$.

^e For the *D. melanogaster*-*D. yakuba* comparisons, these counts were reduced by a single site, as the *D. yakuba* sequence had a different allele than the two segregating in *D. melanogaster*.

Table 3
Polymorphism and Divergence in *Enhancer of split* Regulatory Regions

| Regions Tested ^a | P _{NR} ^b | P _R ^b | F _{NR} ^b | F _R ^b | FET <i>P</i> value ^c | <i>G</i> -test Statistic (<i>P</i> value) |
|-------------------------------------|------------------------------|-----------------------------|------------------------------|-----------------------------|---------------------------------|--|
| 3' UTRs (all motifs) | 65 | 9 | 114 | 9 | 0.309 | 1.269 (0.260) |
| 3' UTRs (conserved motifs) | 66 | 8 | 115 | 8 | 0.294 | 1.115 (0.291) |
| Enhancers (all binding sites) | 168 | 40 | 316 | 37 | 0.005 | 8.200 (0.004) |
| Enhancers (conserved binding sites) | 194 | 12 | 338 | 13 | 0.290 | 1.323 (0.250) |

^a Test 3' UTR motif versus nonmotif sequence, and enhancer-binding site versus nonbinding site sequence, both for all motifs/binding sites, and only those conserved between *D. melanogaster* and *D. pseudoobscura*. These regions are described in the footnote to table 1.

^b Manual counts of regulatory and nonregulatory *D. melanogaster* polymorphisms and fixed differences between *D. melanogaster* and the *D. simulans* strain sim6. Sites within gaps and nonballelic sites were excluded. Polymorphism counts presented may not exactly correspond to the number of segregating sites given in table 1, as, here, polymorphisms were counted irrespective of missing sequence data, whereas in table 1, regions of the alignment harboring any missing sequence data were eliminated.

^c *P* value from a Fisher's exact test.

motifs, whereas for nonmotifs *D* is negative, although this comparison may not be reliable because of the very low polymorphism exhibited by 3' UTR motif regions (table 1). The enhancer-module sequences show a negative value of *D*, with binding-site regions showing considerably lower *D* than nonbinding-site regions (binding sites, *D* = -1.029; nonbinding sites, *D* = -0.345). This difference is increased when comparing conserved binding sites to nonbinding sites. These data suggest that polymorphisms present in, or close to, binding sites for regulatory transcription factors are more rare than are those polymorphisms present in the surrounding nonbinding-site regions of the enhancers.

Population Subdivision at E(spl)-C

To examine the pattern of population subdivision at E(spl)-C, we genotyped 117 polymorphisms in samples of *D. melanogaster* from four continents: Napa Valley (North America), southern France (Europe), Madang (Australia), and Benin (Africa). Ancestral African populations of *D. melanogaster* are thought to have colonized Europe after the last ice age and more recently been introduced by man to North America from Europe and to Australia from African and European populations (David and Capi 1988).

Figure 3 (lower panel) shows the frequency of polymorphisms in the samples from Southern France, Madang, and Benin, as a deviation from the frequency in Napa Valley. The majority of sites show similar frequencies in the southern France and Napa Valley samples, likely reflecting their recent shared ancestry, whereas a number of sites show large frequency differences in the samples from Benin, and particularly, Madang.

The degree of population differentiation at a polymorphic site can be summarized using the F_{ST} statistic (fig. 3, upper panel), and values are mostly within the range of previous data (Begun and Aquadro 1993). However, a few zones, notably around *mδ* and *m6*, show a generally elevated level of population structure, caused primarily by frequency differences in the Madang and Benin samples. There are also three individual sites showing F_{ST} above 0.5: an intergenic C/T polymorphism at position 4752 in the *D. melanogaster* alignment, a G/A polymorphism in the 3' UTR of *mβ* at position 10601, and a synonymous A/G variant in *m2* at position 22644. The latter variant shows

dramatically higher F_{ST} than the surrounding sites, and has a much lower allele frequency in the Madang sample than in the other three samples (pA = 0.96, 1.00, 1.00, and 0.38, in Napa Valley, Benin, Southern France, and Madang, respectively). Genotyping accuracy for this single-nucleotide polymorphism (SNP) in the Madang population was confirmed by resequencing 15 diploid individuals (GenBank accession numbers AY779933 to AY779947): in all cases genotype calls from the SNP assay matched those obtained by sequencing.

Although the genotyped sites can be separated by the region in which they reside (e.g., exon, UTR, enhancer, and so on), and it is possible to look at differentiation across the different functional regions, the variance in F_{ST} within each category is too high for any meaningful interpretation (data not shown).

Haplotype Structure Around E(spl)-C

Our resequencing effort demonstrated the presence of a long haplotype across E(spl)-C: two of the 16 lines were identical at all but one of 1,013 biallelic SNPs and simple InDels. The discrepant site, an A/G polymorphism at position 19662 in the *D. melanogaster* alignment, exhibits the minor G-allele in one of the two lines showing the long haplotype and the major A-allele in all other lines. Sequencing approximately 700 bp from regions 200 kb upstream, 200 kb downstream, and 500 kb downstream of E(spl)-C, showed that 12/16, 33/33, and 21/33 biallelic polymorphisms, respectively, had the same allele in the two lines with the long haplotype. Hence, the haplotype breaks down 0 to 200 kb upstream, and 200 to 500 kb downstream of E(spl)-C. Because E(spl)-C is present at cytological position 96F9-10, and the breakpoint of the common inversion polymorphism In(3R)Payne is thought to be 96A18-19 (Bridges and Bridges 1938), a distance of approximately 1,300 kb, we were concerned that this inversion may be present in the pair of lines showing the long haplotype. However, cytological analysis of the 15 extant sequenced strains, including the pair showing the long haplotype, revealed no inversions close to the cytological position of E(spl)-C.

We used the haplotype test put forward by Hudson et al. (1994) to determine whether the presence of a pair of sequences differing at just one of the 1,013 polymorphic sites across E(spl)-C deviates from neutral expectation.

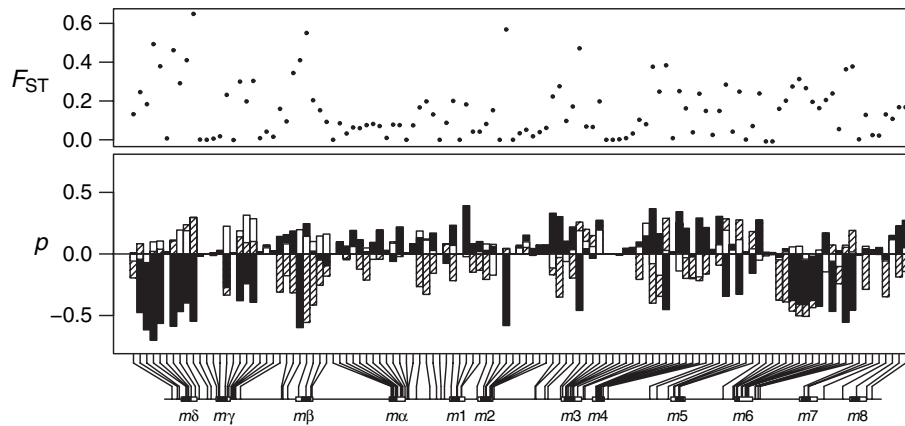


FIG. 3.—Population structure at the *Enhancer of split* complex. The positions of 117 genotyped polymorphisms are shown relative to the transcribed portions of the locus (exons, filled boxes; UTRs, open boxes). The bottom panel shows the frequency of each site in each of the three samples, southern France (open bars), Benin (hatched bars), and Madang (filled bars), as a deviation from the frequency of the major allele in the Napa Valley sample. Thus, positive values show that the major allele is more frequent than in Napa Valley, and negative values show that the site is less frequent than in Napa Valley. The top panel shows the estimated population subdivision across all four samples, as measured by the F_{ST} statistic (Weir and Cockerham 1996).

Using values of the population recombination rate from $\rho = 0$ to 2000, the P value for the test was $0.528 > P > 0.118$, suggesting the observation is consistent with neutrality. Without polymorphism data from the remainder of the region covered by the long haplotype (greater than 200 kb of sequence), we are unable to assess whether the existence of the full haplotype is also consistent with neutral expectation.

Finally, we note that six SNPs defining the sequenced haplotype were genotyped in the larger outbred population samples, yet none of the 226 individuals show genotypes consistent with the existence of this haplotype, and we conclude that the haplotype must be rare.

Indications of Positive Selection at E(spl)-C

We compared the within-species polymorphism (π), the frequency spectrum of observed polymorphisms as summarized by Tajima's D statistic, and the population recombination rate (ρ) per base pair across E(spl)-C. Because the locus is almost 50 kb, encompassing various coding and regulatory regions, a sliding-window approach is likely to be more informative than simply examining the summary statistics for the entire region. Figure 4 shows the pattern of π , D , and ρ across the 47,677-bp *D. melanogaster* E(spl)-C alignment. Spatial variation in π and D is correlated: π and D were calculated over all 86 independent 500-bp windows of E(spl)-C, showing a Pearson correlation coefficient of $r = 0.527$ ($P < 10^{-6}$) and a 95% confidence interval of 0.355 to 0.665. Recombination across E(spl)-C varies threefold to fourfold, showing three pronounced peaks of recombination at approximately 8.9 kb, approximately 14.2 kb, and approximately 19.1 kb and two marked dips. These two zones of very low recombination—about $m1/m2$ and $m7/m8$ —correspond to regions of reduced nucleotide diversity, and strongly negative D (fig. 4).

It is worthwhile to point out that the genotyped diploid individuals yield a different pattern of recombination than do the resequenced alleles (fig. 5), predominantly

showing a flat profile with no pronounced dips or peaks, aside from a single area in the Madang sample showing a high level of LD. This difference stems from our SNP selection process, whereby SNPs were picked to minimize the level of LD among members of the final genotyped set. Variation in the magnitude of recombination shown by the four samples, which maximally shows an almost fourfold difference between samples, can be explained by ascertainment bias—initially identifying polymorphisms within the Napa Valley population, and genotyping those sites in different population samples.

The observed pattern of increased LD, reduced nucleotide diversity, and an excess of rare polymorphisms in the sequenced alleles matches the predictions of the “hitchhiking” model of positive selection (Kaplan, Hudson, and Langley 1989; Braverman et al. 1995; Fay and Wu 2000; Kim and Stephan 2000; Andolfatto and Przeworski 2001). Under this model, neutral variants linked to an advantageous mutation are swept to fixation, reducing haplotype and nucleotide diversity, and because most observed polymorphisms will have arisen postsweep, they generally will be rare. To test whether the value of D is significantly negative in the candidate sweep areas, we employed a coalescence approach, conditioning on testing multiple windows, and found that all values of D estimated from the sequence data are within simulated 95% confidence limits. However, we note that if we had a priori reasons to expect positive selection around $m1/m2$ and $m7/m8$ and had merely tested these regions, we would have found significant evidence for negative Tajima's D statistic ($m1/m2$: $D = -1.65$; $m7/m8$: $D = -1.79$; $P < 0.05$ for $\rho = 0, 1, 10, \text{ and } 100$ in each case). The aforementioned significant MK test at the gene $m2$ supports a hypothesis of positive selection in this section of E(spl)-C.

Discussion

There is considerable interest in understanding regulatory sequence evolution and its impact on phenotypic

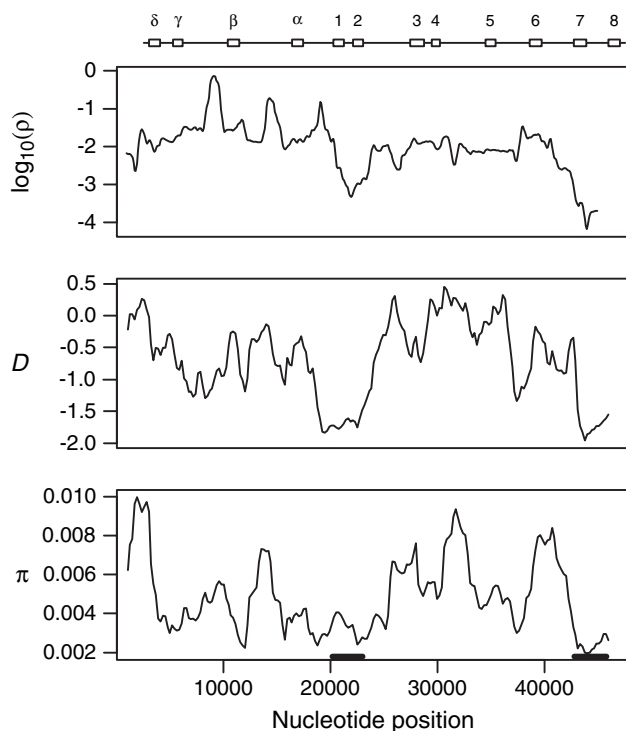


FIG. 4.—Rate of recombination, Tajima's D statistic, and nucleotide diversity across the *Enhancer of split* complex. All statistics are calculated from an alignment of 16 resequenced *D. melanogaster* E(spl)-C alleles. The rate of recombination, ρ , per base pair was calculated from 477 biallelic sites with a window size of 20 sites (see *Materials and Methods*), and Tajima's D statistic and nucleotide diversity, π , were calculated using DnaSP version 4.0 (Rozas et al. 2003), with window size = 2 kb, and step size = 200 bp. For each statistic, the plot represents a smoothed curve through the data using the *ksmooth* function in the statistical programming language R (www.R-project.org). Positions of transcripts are shown above the plots. Thick lines, of arbitrary length, on the x -axis represent regions of E(spl)-C that appear to show signatures of past positive selection (see *Results* for details).

change, and this has proved challenging because of the relatively fluid nature of promoter organization, the complexity of the *cis*-regulatory and *trans*-regulatory factors likely to be responsible for conferring specific spatial and temporal patterns of transcript expression, and the relative lack of data compared with coding sequences (see Wray et al. [2003] for a review). All of this contributes to the difficulty identifying sites in noncoding DNA that affect expression of any given phenotype. In this regard, the *Enhancer of split* locus in *Drosophila* is valuable, as it has been extensively studied in the context of neuronal cell fate determination and encompasses large areas of well-annotated regulatory sequences. We sought to examine whether we could distinguish these known regulatory regions on the basis of primary sequence data, as well as identify other sites or regions within the E(spl)-C under selection using a number of complementary approaches.

Deep Phylogenetic Divergence

Numerous regulatory motifs have been localized to the 3' UTRs of E(spl)-C genes, and many transcription factor-binding sites identified in enhancer regions. We

show that 134 of the 182 functional elements initially identified in *D. melanogaster* are conserved in *D. pseudoobscura*, suggesting that these 134 elements maintain a similar function in the two species. We also observed that flanking sequences were often conserved along with the regulatory element, and this raises two possibilities: genomic context may be an important determinant of the binding efficiency at a given site, and/or the species may be insufficiently distant for all nonfunctional sequences to have diverged. This issue may be resolved using phylogenetic shadowing (Boffelli et al. 2003) at E(spl)-C across several species of *Drosophila*. Because mutations should accumulate randomly in nonfunctional DNA within each of the species, only truly important regions will be conserved among all of the tested species. This method has been successful in identifying functional regions of the yeast genome (Kellis et al. 2003).

Overall, we were unable to detect 26% of the regulatory motifs identified in *D. melanogaster* in *D. pseudoobscura*, and given that many of these elements have not been functionally assayed, the simplest explanation is that these nonconserved elements do not have regulatory function and are, therefore, unconstrained and free to evolve. However, some or all may represent real species differences.

It seems unlikely that the species differ in transcript expression because of differences in the binding-site complement of the upstream enhancer modules, as the *mγ* gene enhancer modules from the diverged species *D. melanogaster* and *D. hydei* drive nearly identical patterns of gene expression in *D. melanogaster* larval wing discs (Nellesen, Lai, and Posakony 1999). Another possibility might be that expression patterns across species are maintained despite changes in the DNA sequence of enhancer regions. This hypothesis predicts that known transcription factor-binding sites in *D. melanogaster* that are not conserved in *D. pseudoobscura* should be functionally substituted with other (perhaps unrecognized) binding sites to achieve identical transcript expression. This model of functional compensation during enhancer evolution has been used to explain the maintenance of *even-skipped* stripe 2 embryonic expression in *D. melanogaster* and *D. pseudoobscura*, despite few of the *D. melanogaster* binding sites being conserved in *D. pseudoobscura* at this locus (Ludwig et al. 2000).

Shallow Divergence and Polymorphism

The E(spl)-C shows a lower level of nucleotide diversity than the average observed for other autosomal loci (see Moriyama and Powell [1996]). It is possible that the entire region is particularly constrained, although this seems unlikely, given that UTR and intergenic sequences—regions thought to experience different levels of evolutionary functional constraint—exhibit similar diversity (table 1). Andolfatto, Depaulis, and Navarro (2001) showed that loci up to 1,000 kb away from an inversion breakpoint can be subject to a reduction in diversity, and because E(spl)-C is positioned approximately 1300 kb downstream of the breakpoint of the common inversion polymorphism In(3R)Payne, it appears unlikely it has had

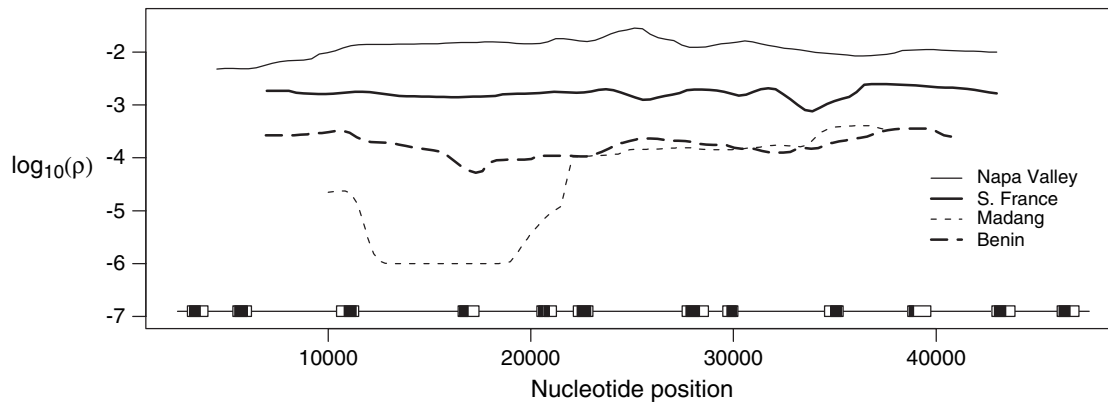


FIG. 5.—Recombination rate across the *Enhancer of split* complex in samples of diploid individuals. The population recombination rate, ρ , per base pair was estimated for each of the four samples using the RECSLIDER program with a window size of 20 sites and an initial estimate of $\rho = 0.01$ (see *Materials and Methods*). The number of biallelic sites used in the analyses were 106 (Napa Valley), 89 (southern France), 73 (Benin), and 58 (Madang) and differed because of fixation of some SNPs (actual sites used available from <http://cstern.bio.uci.edu/pubs.htm>). This explains why not all curves extend to the same position on the x -axis. Each plot represents a smoothed curve through the data using the `ksmooth` function in the statistical programming language R (www.R-project.org). Positions of exons (filled boxes) and UTRs (open boxes) are shown below the curves.

a major role in shaping variability at E(spl)-C. Nevertheless, data from Andolfatto, Depaulis, and Navarro (2001) are based only on 10 genes, and the breakpoints of In(3R)Payne have not been precisely mapped. Systematic sequencing of regions various distances from the breakpoint would allow the extent of diversity suppression to be measured more precisely.

We note that any effect of In(3R)Payne on the diversity at E(spl)-C would be limited to *D. melanogaster*—between-species diversity would be unaffected. Thus, the reduction in divergence between *D. melanogaster* and *D. simulans* at E(spl)-C can be used to reject the hypothesis that an inversion influences E(spl)-C diversity within *D. melanogaster*.

Our observed conservation of regulatory sequences across species suggests that these elements are largely under purifying selection. Using our polymorphism data, we show that in the E(spl)-C enhancer modules, binding sites show much lower levels of diversity than do non-binding sites, and more of the polymorphisms are rare, suggesting a similar process of purifying selection. This effect is particularly apparent when we consider only those binding sites conserved between *D. melanogaster* and *D. pseudoobscura*, suggesting that conserved regulatory binding sites are less likely than nonconserved binding sites to contribute to standing phenotypic variation or microevolution. Therefore, strategies to identify regulatory regions based on sequence conservation across two or more evolutionarily diverged species (Boffelli et al. 2003; Kellis et al. 2003) may in fact be less likely to detect elements influencing complex trait variation within a species.

We also demonstrated, using a test similar to the MK and HKA tests (McDonald and Kreitman 1991; Hudson, Kreitman, Aguadé 1987), that the ratio of fixed changes to polymorphisms differs between binding sites and non-binding sites in upstream enhancer regions of E(spl)-C. Phinchongsakuldit, MacArthur, and Brookfield (2003) have previously reported a similar result for the *bx-32.8* enhancer of *Ubx*. Unfortunately, the tests are difficult to

interpret, and it is unclear how to polarize any deviation from neutrality. The data from E(spl)-C enhancer regions and from Phinchongsakuldit, MacArthur, and Brookfield (2003) are compatible with too few fixed regulatory sites, too many polymorphic regulatory sites, too few polymorphic nonregulatory sites, or too many fixed nonregulatory sites. None of these are mutually exclusive.

The observation of too few fixed changes within transcription factor-binding sites suggests conservation of the binding sites across species. In contrast, a greater number of polymorphisms within binding sites could indicate the maintenance of balanced polymorphisms but could also be explained if mutations in binding sites are slightly deleterious, as they will then contribute to within-species heterozygosity, but are unlikely to become fixed (Nachman et al. 1996). The greater number of rare polymorphisms observed in enhancer-binding sites also provides support for the idea that mutations here are mildly deleterious, as such polymorphisms are less likely to attain high frequency compared with neutral substitutions. An excess of fixed nonregulatory changes is possible if there is selection on nonbinding sites, which could imply that some of the “nonregulatory” portions of the enhancer modules are actually functional.

Local Adaptation

Another way to detect selection is to examine the level of population differentiation. It is known that *D. melanogaster* is not a panmictic population, as extensive among-population variation has been demonstrated for several loci, including the mitochondrial DNA (Hale and Singh 1991; Begun and Aquadro 1993), but under neutral evolution, all loci are expected to show the same level of differentiation among subpopulations. Natural selection can alter the apparent level of subdivision at variants that are favored in some populations; thus, by examining geographic variation in allele frequency, one can identify targets of local adaptation (e.g., the *Duffy* blood group locus in humans [Hamblin and Di

Rienzo 2000], clinal variation at the *Adh* locus in *D. melanogaster* [Berry and Kreitman 1993]). We have shown that across the E(spl)-C locus, there is variation in the degree of observed population structure (fig. 3). In particular, regions of elevated F_{ST} around the genes *mδ* and *m6* are possibly indicative of functional population differentiation, and because some of the sites are in noncoding regions, may show differential regulatory activity across populations. However, in general, sites in regulatory regions did not show a level of population differentiation different from sites in other nonregulatory regions.

We note that the level of population differentiation can also be elevated by background selection against deleterious alleles, as demonstrated by Nordborg (1997) using coalescent simulation. However, in terms of our goal of identifying regions under selection, detecting either background selection or diversifying selection implicates a region as having function.

Within-Population Selection

Two regions (around *m1/m2* and *m7/m8* [fig. 4]) exhibit patterns—high LD, low diversity, and skewed polymorphism-frequency distribution—indicative of past positive selection (Kim and Stephan 2000; Andolfatto and Przeworski 2001). The observation of a significant excess of fixed synonymous mutations at *m2* is also consistent with a scenario of past selection in/around this gene. However, the evidence is insufficient to lend strong support to a hypothesis of positive selection. The two putatively selected regions have negative values of D but are not significant after multiple testing, and although the nucleotide diversity is reduced in these zones, other regions also exhibit low levels of heterozygosity. These difficulties highlight a concern that as we collect population-genetic data sets encompassing very large genomic regions, effects will need to be much more pronounced to be found significant using sliding-window-type approaches.

There are several other reasons why selective sweeps are difficult to detect. The power of Tajima's D statistic to detect a selective sweep is strongly dependent not only on the number of sequenced alleles but also on the selective strength of the event and on the number of generations since it occurred (Simonsen, Churchill, and Aquadro 1995). Old sweeps will be obscured by the accumulation of neutral mutations, whereas weak sweeps will reduce heterozygosity less efficiently. Thus, it is possible that the two putative cases of hitchhiking we outline represent selection of weakly advantageous mutations or are perhaps very distant events and, hence, do not achieve significance. Also, as demonstrated by Kim and Stephan (2002), individual realizations of a simulated selective sweep vary in the size of the area of reduced heterozygosity, the extent of the reduction, and the position of the valley relative to the selected polymorphism.

However, other models can account for aspects of the data we observe without implicating a selective sweep. Background, or purifying selection also eliminates variation (Charlesworth, Morgan, and Charlesworth 1993), although this model does not predict a skew in the poly-

morphism frequency spectrum (Kim and Stephan 2000; Andolfatto and Przeworski 2001).

Conclusion

Ultimately, we wish to understand how coding and regulatory sequence variation influences the expression of phenotypic traits, including molecular phenotypes such as transcript abundance and alternative splicing. The well-characterized *Drosophila* bristle number model system, with its attendant list of viable neurogenic candidate genes (Mackay 1995), represents perhaps the best route to achieving these aims. Even so, screening through every noncoding polymorphism and testing for an association with bristle number is a daunting task, and methods are required to aid an informed choice of those sites to genotype.

We have described nucleotide diversity across different functional regions of the bristle number candidate locus E(spl)-C, showing that previously identified regulatory elements are visible to selection. We also highlight other regions exhibiting signatures of nonneutral evolution, implying they are also of functional importance in regulating E(spl)-C genes. Because E(spl)-C is likely to have a role in the genetic control of natural variation in bristle number (Dilda and Mackay 1995; Long et al. 1995; Norga et al. 2003; Nuzhdin, Dilda, and Mackay 1999), a trait under stabilizing selection (García-Dorado and González 1996), sites within these candidate functional regions are more likely to be QTN for bristle number than are sites in regions showing neutral evolution. Regions showing no departure from neutrality may still harbor functional sites, but natural selection has not acted in a detectable manner on these regions in the recent past.

In general, enriching association-mapping studies for sites more likely to contribute to phenotypic variation will streamline the process of detecting genetic variants underlying natural variation in complex traits. We suggest that together with other motivating factors, selecting sites for genotyping in association studies should be informed by the results of sequence analysis methods that detect the action of natural selection.

Acknowledgments

We thank L. Partridge and W. J. Kennington for the non-American population samples of flies, D. J. Begun for the *sim6* strain, M. Long and K. Thornton for the *yakTAI27* strain, J. D. Wall for the sliding-window analysis of recombination rate software, and B. S. Gaut and two anonymous reviewers for valuable comments on the manuscript. All data are available at <http://cstern.bio.uci.edu/pubs.htm>. This work was supported by National Institutes of Health grant GM 58564 to A.D.L.

Literature Cited

- Andolfatto, P., F. Depaulis, and A. Navarro. 2001. Inversion polymorphisms and nucleotide variability in *Drosophila*. *Genet. Res.* 77:1–8.

- Andolfatto, P., and M. Przeworski. 2001. Regions of lower crossing over harbor more rare variants in African populations of *Drosophila melanogaster*. *Genetics* **158**:657–665.
- Ardell, D. H. 2004. SCANMS: adjusting for multiple comparisons in sliding window neutrality tests. *Bioinformatics* **20**:1986–1988.
- Bailey, A. M., and J. W. Posakony. 1995. Suppressor of Hairless directly activates transcription of *Enhancer of split* complex genes in response to Notch receptor activity. *Genes Dev.* **9**:2609–2622.
- Begun, D. J., and C. F. Aquadro. 1993. African and North American populations of *Drosophila melanogaster* are very different at the DNA level. *Nature* **365**:548–550.
- Begun, D. J., and P. Whitley. 2000. Reduced X-linked nucleotide polymorphism in *Drosophila simulans*. *Proc. Natl. Acad. Sci. USA* **97**:5960–5965.
- Berry, A., and M. Kreitman. 1993. Molecular analysis of an allozyme cline: alcohol dehydrogenase in *Drosophila melanogaster* on the East Coast of North America. *Genetics* **134**:869–893.
- Boffelli, D., J. McAuliffe, D. Ovcharenko, K. D. Lewis, I. Ovcharenko, L. Pachter, and E. M. Rubin. 2003. Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science* **299**:1391–1394.
- Braverman, J. M., R. R. Hudson, N. L. Kaplan, C. H. Langley, and W. Stephan. 1995. The hitchhiking effect on the site frequency spectrum of DNA polymorphism. *Genetics* **140**:783–796.
- Bridges, C. B., and P. N. Bridges. 1938. Salivary analysis of inversion-3R-Payne in the “venation” stock of *Drosophila melanogaster*. *Genetics* **23**:111–114.
- Charlesworth, B., M. T. Morgan, and D. Charlesworth. 1993. The effect of deleterious mutations on neutral molecular variation. *Genetics* **134**:1289–1303.
- Cooper, M. T. D., D. M. Tyler, M. Furriols, A. Chalkiadaki, C. Delidakis, and S. Bray. 2000. Spatially restricted factors cooperate with Notch in the regulation of *Enhancer of split* genes. *Dev. Biol.* **221**:390–403.
- David, J. R., and P. Cappy. 1988. Genetic variation of *Drosophila melanogaster* natural populations. *Trends Genet.* **4**:106–111.
- de Celis, J. F., J. de Celis, P. Ligoxygakis, A. Preiss, C. Delidakis, and S. Bray. 1996. Functional relationships between *Notch*, *Su(H)* and the bHLH genes of the *E(spl)* complex: the *E(spl)* genes mediate only a subset of *Notch* activities during imaginal development. *Development* **122**:2719–2728.
- Delidakis, C., and S. Artavanis-Tsakonas. 1992. The *Enhancer of split [E(spl)]* locus of *Drosophila* encodes seven independent helix-loop-helix proteins. *Proc. Natl. Acad. Sci. USA* **89**:8731–8735.
- Dilda, C. L., and T. F. C. Mackay. 2002. The genetic architecture of *Drosophila* sensory bristle number. *Genetics* **162**:1655–1674.
- Eastman, D. S., R. Slee, E. Skoufos, L. Bangalore, S. Bray, and C. Delidakis. 1997. Synergy between Suppressor of Hairless and Notch in regulation of *Enhancer of split mg* and *md* expression. *Mol. Cell. Biol.* **17**:5620–5628.
- Fay, J. C., and C.-I. Wu. 2000. Hitchhiking under positive Darwinian selection. *Genetics* **155**:1405–1413.
- García-Dorado, A., and J. A. González. 1996. Stabilizing selection detected for bristle number in *Drosophila melanogaster*. *Evolution* **50**:1573–1578.
- Genissel, A., T. Pastinen, A. Dowell, T. F. C. Mackay, and A. D. Long. 2004. No evidence for an association between common nonsynonymous polymorphisms in *Delta* and bristle number variation in natural and laboratory populations of *Drosophila melanogaster*. *Genetics* **166**:291–306.
- Hale, L. R., and R. S. Singh. 1991. A comprehensive study of genic variation in natural populations of *Drosophila melanogaster*. IV. Mitochondrial DNA variation and the role of history vs. selection in the genetic structure of geographic populations. *Genetics* **129**:103–117.
- Hamblin, M. T., and A. Di Rienzo. 2000. Detection of the signature of natural selection in humans: evidence from the *Duffy* blood group locus. *Am. J. Hum. Genet.* **66**:1669–1679.
- Hudson, R. R. 1990. Gene genealogies and the coalescent process. Pp. 1–44 in D. Futuyma and J. Antonovics, eds. *Oxford surveys in evolutionary biology*. Oxford University Press, Oxford.
- . 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* **18**:337–338.
- Hudson, R. R., K. Bailey, D. Skarecky, J. Kwiatowski, and F. J. Ayala. 1994. Evidence for positive selection in the superoxide dismutase (*Sod*) region of *Drosophila melanogaster*. *Genetics* **136**:1329–1340.
- Hudson, R. R., M. Kreitman, and M. Aguadé. 1987. A test of neutral molecular evolution based on nucleotide data. *Genetics* **116**:153–159.
- Jan, Y. N., and L. Y. Jan. 1994. Genetic control of cell fate specification in *Drosophila* peripheral nervous system. *Annu. Rev. Genet.* **28**:373–393.
- Jennings, B., A. Preiss, C. Delidakis, and S. Bray. 1994. The Notch signaling pathway is required for *Enhancer of split* bHLH protein expression during neurogenesis in the *Drosophila* embryo. *Development* **120**:3537–3548.
- Jukes, T. H., and C. R. Cantor. 1969. Evolution of protein molecules. Pp. 21–120 in H. M. Munro, ed. *Mammalian protein metabolism*. Academic Press, New York.
- Kaplan, N. L., R. R. Hudson, and C. H. Langley. 1989. The “hitchhiking effect” revisited. *Genetics* **123**:887–899.
- Kellis, M., N. Patterson, M. Endrizzi, B. Birren, and E. S. Lander. 2003. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423**:241–254.
- Kim, Y., and W. Stephan. 2000. Joint effects of genetic hitchhiking and background selection on neutral variation. *Genetics* **155**:1415–1427.
- . 2002. Detecting a local signature of genetic hitchhiking along a recombining chromosome. *Genetics* **160**:765–777.
- Knust, E., H. Schrons, F. Grawe, and J. A. Campos-Ortega. 1992. Seven genes of the *Enhancer of split* complex of *Drosophila melanogaster* encode helix-loop-helix proteins. *Genetics* **132**:505–518.
- Kramatschek, B., and J. A. Campos-Ortega. 1994. Neuroectodermal transcription of the *Drosophila* neurogenic genes *E(spl)* and *HLH-m5* is regulated by proneural genes. *Development* **120**:815–826.
- Lai, E. C. 2002. Micro RNAs are complementary to 3′ UTR sequence motifs that mediate negative post-transcriptional regulation. *Nat. Genet.* **30**:363–364.
- Lai, E. C., R. Bodner, J. Kavalier, G. Freschi, and J. W. Posakony. 2000. Antagonism of Notch signaling activity by members of a novel protein family encoded by the *Bearded* and *Enhancer of split* gene complexes. *Development* **127**:291–306.
- Lai, E. C., R. Bodner, and J. W. Posakony. 2000. The *Enhancer of split* complex of *Drosophila* includes four Notch-regulated members of the Bearded gene family. *Development* **127**:3441–3455.
- Lai, E. C., C. Burks, and J. W. Posakony. 1998. The K box, a conserved 3′ UTR sequence motif, negatively regulates accumulation of *Enhancer of split* complex transcripts. *Development* **125**:4077–4088.
- Lai, E. C., and J. W. Posakony. 1997. The Bearded box, a novel 3′ UTR sequence motif, mediates negative post-transcriptional

- regulation of *Bearded* and *Enhancer of split* complex gene expression. *Development* **124**:4847–4856.
- . 1998. Regulation of *Drosophila* neurogenesis by RNA:RNA duplexes? *Cell* **93**:1103–1104.
- Leviten, M. W., E. C. Lai, and J. W. Posakony. 1997. The *Drosophila* gene *Bearded* encodes a novel small protein and shares 3' UTR sequence motifs with multiple *Enhancer of split* complex genes. *Development* **124**:4039–4051.
- Long, A. D., S. L. Mullaney, L. A. Reid, J. D. Fry, C. H. Langley, and T. F. C. Mackay. 1995. High resolution mapping of genetic factors affecting bristle number in *Drosophila melanogaster*. *Genetics* **139**:1273–1291.
- Ludwig, M. Z., C. Bergman, N. H. Patel, and M. Kreitman. 2000. Evidence for stabilizing selection in a eukaryotic enhancer element. *Nature* **403**:564–567.
- Mackay, T. F. C. 1995. The genetic basis of quantitative variation: numbers of sensory bristles of *Drosophila melanogaster* as a model system. *Trends Genet.* **11**:464–470.
- Maier, D., B. M. Marte, W. Schäfer, Y. Yu, and A. Preiss. 1993. *Drosophila* evolution challenges postulated redundancy in the *E(spl)* gene complex. *Proc. Natl. Acad. Sci. USA* **90**:5464–5468.
- McDonald, J. H., and M. Kreitman. 1991. Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature* **351**:652–654.
- Moriyama, E. N., and J. R. Powell. 1996. Intraspecific nuclear DNA variation in *Drosophila*. *Mol. Biol. Evol.* **13**:261–277.
- Nachman, M. W., W. M. Brown, M. Stoneking, and C. F. Aquadro. 1996. Nonneutral mitochondrial DNA variation in humans and chimpanzees. *Genetics* **142**:953–963.
- Nellesen, D. T., E. C. Lai, and J. W. Posakony. 1999. Discrete enhancer elements mediate selective responsiveness of *Enhancer of split* complex genes to common transcriptional activators. *Dev. Biol.* **213**:33–53.
- Nuzhdin, S. V., C. L. Dilda, and T. F. C. Mackay. 1999. The genetic architecture of selection response: inferences from fine-scale mapping of bristle number quantitative trait loci in *Drosophila melanogaster*. *Genetics* **153**:1317–1331.
- Nordborg, M. 1997. Structured coalescent processes on different time scales. *Genetics* **146**:1501–1514.
- Norga, K. K., M. C. Gurganus, C. L. Dilda, A. Yamamoto, R. F. Lyman, P. H. Patel, G. M. Rubin, R. A. Hoskins, T. F. C. Mackay, and H. J. Bellen. 2003. Quantitative analysis of bristle number in *Drosophila* mutants identifies genes involved in neural development. *Curr. Biol.* **13**:1388–1397.
- Phinchongsakuldit, J., S. MacArthur, and J. F. Y. Brookfield. 2004. Evolution of developmental genes: molecular micro-evolution of enhancer sequences at the *Ubx* locus in *Drosophila* and its impact on developmental phenotypes. *Mol. Biol. Evol.* **21**:348–363.
- Rieder, M. J., S. L. Taylor, V. O. Tobe, and D. A. Nickerson. 1998. Automating the identification of DNA variations using quality-based fluorescence re-sequencing: analysis of the human mitochondrial genome. *Nucleic Acids Res.* **26**:967–973.
- Robin, C., R. F. Lyman, A. D. Long, C. H. Langley, and T. F. C. Mackay. 2002. *hairy*: a quantitative trait locus for *Drosophila* sensory bristle number. *Genetics* **162**:155–164.
- Rozas, J., J. C. Sánchez-DelBarrio, X. Messeguer, and R. Rozas. 2003. DnaSP: DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics* **19**:2496–2497.
- Russo, C. A., N. Takezaki, and M. Nei. 1995. Molecular phylogeny and divergence times of drosophilid species. *Mol. Biol. Evol.* **12**:391–404.
- Simonsen, K. L., G. A. Churchill, and C. F. Aquadro. 1995. Properties of statistical tests of neutrality for DNA polymorphism data. *Genetics* **141**:413–429.
- Singson, A., M. W. Leviten, A. G. Bang, X. H. Hua, and J. W. Posakony. 1994. Direct downstream targets of proneural factors in the imaginal disc include genes involved in lateral inhibitory signaling. *Genes Dev.* **8**:2058–2071.
- Stern, D. L. 1998. A role of *Ultrabithorax* in morphological differences between *Drosophila* species. *Nature* **396**:463–466.
- Stern, D. L. 2000. Evolutionary developmental biology and the problem of variation. *Evolution* **54**:1079–1091.
- Tajima, F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**:585–595.
- Tietze, K., N. Oellers, and E. Knust. 1992. Enhancer of split^D, a dominant mutation of *Drosophila*, and its use in the study of functional domains of a helix-loop-helix protein. *Proc. Natl. Acad. Sci. USA* **89**:6152–6156.
- Wall, J. D., L. A. Frisse, R. R. Hudson, and A. Di Rienzo. 2003. Comparative linkage-disequilibrium analysis of the β -globin hotspot in primates. *Am. J. Hum. Genet.* **73**:1330–1340.
- Weir, B. S., and C. C. Cockerham. 1984. Estimating *F*-statistics for the analysis of population structure. *Evolution* **38**:1358–1370.
- Wittkopp, P. J., K. Vaccaro, and S. B. Carroll. 2002. Evolution of *yellow* gene regulation and pigmentation in *Drosophila*. *Curr. Biol.* **12**:1547–1556.
- Wray, G. A., M. W. Hahn, E. Abouheif, J. P. Balhoff, M. Pizer, M. V. Rockman, and L. A. Romano. 2003. The evolution of transcriptional regulation in eukaryotes. *Mol. Biol. Evol.* **20**:1377–1419.
- Wurmbach, E., I. Wech, and A. Preiss. 1999. The *Enhancer of split* complex of *Drosophila melanogaster* harbors three classes of Notch responsive genes. *Mech. Dev.* **80**:171–180.

Diethard Tautz, Associate Editor

Accepted November 2, 2004