Sequence Turnover and Tandem Repeats in *cis*-Regulatory Modules in Drosophila

Saurabh Sinha and Eric D. Siggia

Center for Studies in Physics and Biology, The Rockefeller University

The path by which regulatory sequence can change, yet preserve function, is an important open question for both evolution and bioinformatics. The recent sequencing of two additional species of Drosophila plus the wealth of data on gene regulation in the fruit fly provides new means for addressing this question. For regulatory sequences, indels account for more base pairs (bp) of change than substitutions (between *Drosophila melanogaster* and *Drosophila yakuba*), though they are fewer in number. Using *Drosophila pseudoobscura* as an out-group, we can distinguish insertions from deletions (with maximum parsimony criteria), and find a ratio between 1 and 5 (insertions to deletions) that is species dependent and much larger than the ratio of 1/8 for neutral sequences (Petrov and Hartl 1998). Because neutral sequence is rapidly cleared from the genome, most noncoding regions which preserve their length between *D. melanogaster–D. pseudoobscura* and have an excess of insertions over deletions should be functional. A fraction of 15%–18% (i.e., more than 20 standard deviations from random expectation) of the regulatory sequence is covered by low copy number tandem repeats whose repeating unit has an average length of 5–10 bp and which occur preferentially (25%–45% coverage) in indels. All indels may be due to tandem repeats if we extrapolate the detection efficiency of the repeat-finding algorithms using the observed point mutation rate between the species we compare. Sequence creation by local duplication accords with the tendency for multiple copies of transcription factor–binding sites to occur in regulatory modules. Thus, indel events and tandem repeats in particular need to be incorporated into models of regulatory evolution because they can alter the rate at which beneficial variants arise and should also influence bioinformatic algorithms that parse regulatory sequences into binding sites.

Introduction

The continual flood of genome sequences has reinforced multiple functional studies as to the conservative nature of protein evolution. For example, the regulatory genes of the cell cycle are recognizably similar from yeast to mammals (Nasmyth 2001), the Hox genes play similar roles in all metazoans, and intracellular signaling pathways are well conserved (Carroll, Grenier, and Weatherbee 2001; Davidson 2001; Wilkins 2002). Although spontaneous mutations in regulatory DNA can affect dramatic changes in morphology, as attested by homeotic mutations, instances where morphological changes between existing species can be traced to regulatory evolution are rare (Akam 1998; Tautz 2000; Wray et al. 2003). Lacking any quantitative means to infer function from genomic sequence, it is impossible to assess the relative importance of changes in regulatory DNA or proteins, as compared to changes in structural genes, for the generation of new species.

Part of the problem is that in comparison with our ability to locate protein-coding genes (particularly in the compact genomes of model organisms) and decompose them into functional domains (Bateman et al. 2004), regulatory DNA is both hard to locate and assign function to. Even when the relevant *cis*-regulatory modules (CRMs) have been located in the genome and their orthologous sequences identified in related species, it is very unclear which changes in these are neutral and which lead to change in gene expression. Various instances of conserved functionality in the presence of considerable sequence fluidity have been demonstrated (Ludwig, Patel, and Kreitman 1998) but so also have cases where a few base changes give rise to human disease (Dermitzakis and Clark 2002; Rockman et al. 2003). Thus, there is no established test, analogous to the ratio of nonsynonymous/synonymous codon changes, to compare the rate of potentially functional to neutral change. A recently published article (Wong and Nielsen 2004) takes the first steps in this direction.

A better understanding of regulatory evolution has important consequences for bioinformatics. The analysis of regulatory DNA—both parsing it into transcription factor–binding sites and discerning CRMs in the genome—is an important focus of bioinformatics. The techniques developed for such analysis implicitly or otherwise presuppose a model for what constitutes functional change and what changes are neutral. For instance, binding motif finders use an information theoretic measure of sequence difference to detect binding sites and sometimes impose ad hoc constraints on their spacing and order, both in intra- and interspecific analysis. Clearly, a better comprehension of these constraints from an evolutionary perspective will feed into refinement of the algorithms. However, not only the binding sites themselves but also the "background" sequence they reside in are often crucial to the various bioinformatic algorithms. Using the example of the motif finders again, the significance of a motif, in its simplest form, is calculated from the probability that it was obtained by sampling the single-base frequencies of the background noncoding DNA (Lawrence et al. 1993; Bailey and Elkan 1995; Sinha and Tompa 2000). Similarly, a host of programs (Berman et al. 2002; Halfon et al. 2002; Markstein et al. 2002; Rajewsky et al. 2002; Rebeiz, Reeves, and Posakony 2002) have sprung up to locate CRMs (in fruit fly) by counting the number and type of putative binding sites based on a sample of known sites, everything else about the sequence being ignored. However, scattered studies have shown that regulatory DNA is not a series of stereotyped binding sites in a simple random background, as assumed by these methods. Sites can overlap (Papatsenko et al. 2002), and Dermitzakis, Bergman, and Clark (2002) find that there remains a vestige of mutated binding sites

Key words: drosophila, regulation, repeats, insertions, deletions, neutral.

E-mail: saurabh@lonnrot.rockefeller.edu.

Mol. Biol. Evol. 22(4):874–885. 2005
doi:10.1093/molbev/msi090
Advance Access publication January 19, 2005

around functional ones. Also, as we report in this study, CRMs show a substantial sequence turnover through tandem repeats. Such evolutionary features may lead to a remodeling of background sequences for bioinformatics applications. In short, natural sequence has evolved under selection and drift, and knowledge of these processes may enhance the efficacy of search algorithms. Conversely, a variety of bioinformatic tools can be used to test for patterns in regulatory DNA and better study their evolution.

To measure how regulatory change drives evolution, not only do we need an estimate of raw mutation rates but also the fixation rates of various changes. The latter question is still very open because we know almost nothing about the fitness changes associated with sequence change. The most complicated situation is when changes can compensate. If the sequence responsible for some aspect of gene expression is long enough, then multiple mutations in that sequence can be present in one individual and together go to fixation. The calculation of the most rapid path to fixation in this situation is just beginning (Carter and Wagner 2002).

While the processes that lead from a mutation in an individual's regulatory DNA to a significant allele in the population may be subtle and difficult to quantify precisely, it is easy to compare homologous DNA between close enough species and measure the net turnover due to various categories of change in meaningful units. For instance, the synonymous substitution rate is generally the same for all protein-coding genes and indicative of the point mutation rate for neutrally evolving sequence (Li 1997). For regulatory sequence, several authors have documented a pattern of conserved ungapped blocks punctuated by unalignable gaps (Kassis et al. 1989), with a roughly exponential length distribution (Bergman and Kreitman 2001). Nonfunctional ("dead on arrival") transposons have been used to document an excess of deletions over insertions and thus the rate of sequence loss as well as the ratio of indels to point mutations in neutrally evolving sequence (Petrov and Hartl 1998).

In this study, we wish to exploit the recently sequenced genomes of *Drosophila yakuba* and *Drosophila pseudoobscura* to examine a much broader collection of experimentally characterized CRMs and to augment previous two-way comparisons of *Drosophila virilis* and *Drosophila melanogaster* so as to distinguish insertions from deletions. We compute rates for indels and point mutation in the regulatory sequences and contrast them with those in neutral sequence as deduced in Petrov and Hartl (1998). Indels are found to be considerably less frequent than point mutations but account for more base pairs (bp) of change, a fact that needs greater attention in algorithms that exploit cross-species comparison.

Finally, we apply refined algorithms to find all statistically significant tandem repeats and associate them with various mutational processes. Most of the events we find are not perfect repeats, contain only a few copies of the repeat unit, and are not readily discerned by inspection in contrast to typical microsatellites. The density of the tandem repeats we find is high enough to account for all insertion events in regulatory sequence and thus needs to be included in models of molecular evolution as well as in bioinformatic models of regulatory regions. Tandem duplications nicely accord with the general tendency for there to be multiple binding sites of any one factor in a CRM (Carroll, Grenier, and Weatherbee 2001; Davidson 2001).

Materials and Methods

Regulatory Sequences

We worked with release 3 (Celniker et al. 2002) of the *D. melanogaster* genome, the February 2003 release of the *D. pseudoobscura* genome (Human Genome Sequencing Center at Baylor College of Medicine 2003), and the April 2004 assembly of the *D. yakuba* genome (Genome Sequencing Center at Washington University Medical School 2004). We obtained an extensive collection of 76 experimentally validated CRMs that pattern the embryo, from the literature (Schroeder et al. 2004), and mapped them without ambiguity to the other genomes, by homology. Details of the chosen CRMs are given in *Supplementary Material*. The sequences in *D. melanogaster* range in length between 67 and 5,586 bp, with a median of 1,220 bp, and a total length of 101.05 kbp. The data set of these sequences from the three species is henceforth called "REG."

We constructed a second, smaller data set of 17 CRMs, with orthologous sequences from *D. melanogaster*, *D. pseudoobscura*, and *D. virilis*, the latter being obtained from the noncoding sequence collected by Bergman and Kreitman (2001). We call this the REG2 data set, and the *D. melanogaster* CRMs in this set have a total length of 29.74 kbp. (See *Supplementary Material* for details.)

Pseudogenes

We obtained a list of 105 gene-pseudogene pairs in *D. melanogaster*, annotated for Release 3 of the genome, from Harrison et al. (2003). We ran Blast (TBlastN) to obtain preliminary alignment of the gene to the pseudogene, extracted alignment boundaries, and truncated both sequences at these boundaries. The truncated gene-pseudogene pair was then aligned by a variant of the Needleman-Wunsch algorithm that treats the gene as a sequence of codons and the pseudogene as a sequence of nucleotides, allowing indels in codon units in the former and in single-base units in the latter, using the BLOSSUM matrix for match/mismatch scores. We partitioned the alignment at exon boundaries in the gene and narrowed down each partition to regions anchored on both ends by ungapped blocks of 10 bp with at least 70% identity, thereby obtaining a set of well-aligned exonic regions. We used the codeml program from the PAML package on the aligned codons and computed *dN* and *dS* (Nei and Gojobori 1986). We then extracted only those pseudogenes with $dS \leq 1.25$ and dN/dS ratio ≥ 0.3, so that the period of neutral evolution (time since pseudogene formation) is comparable to the time since gene duplication. (For a more sophisticated discussion see Bustamante, Nielsen, and Hartl [2002].) The set of such pseudogenes is henceforth called the "PGENES" data set and consists of 6–9 pseudogenes, with a total length of $\sim$2 kbp. We shall use the pseudogenes data only to infer indels and point mutations (from two-way comparison) because we are not certain if these events occurred in the pseudogene or the functional gene. The above restrictions (on *dN*, *dS* values) do not specify whether the gene

Table 1
Nucleotide Substitution Matrix Used in Alignment

	A	C	G	T	N
A	91	−114	−31	−123	−43
C	−114	100	−125	−31	−43
G	−31	−125	100	−114	−43
T	−123	−31	−114	91	−43
N	−43	−43	−43	−43	−43

duplication from which the pseudogene arose occurred pre- or postspeciation. This latter aspect of the pseudogene's history is not crucial to our analysis.

Sequence Alignment

All sequence alignments for the REG (and REG2) data sets were done using the MLagan program of Brudno et al. (2003). MLagan is a progressive multiple alignment program that uses the pairwise alignment tool Lagan. The latter first performs local alignment based on multiple exactly conserved words and uses the local alignments as anchors to perform limited-area Needleman-Wunsch alignment between successive anchors. The scoring matrix used was the default nucleotide substitution matrix of the program (see table 1). The gap extension penalty was set to 10, and the gap initiation penalty (γ) was varied uniformly between 500 and 1,200 in eight different runs of the program on each data set. We identified the first and last contiguous stretch (block) of at least five nongap columns in the alignment and focused all analysis on the portion of the alignment bordered by these two stretches. This was done to address the combined effects of (1) the incompleteness of some orthologous sequences and (2) incorrectly aligned columns at the termini, a common problem in alignment programs.

Our ability to correctly infer insertion and deletion events (*indels*) depends on the quality of the ungapped "anchor" regions that delimit them. The anchor is terminated by the next indel in any of the species being aligned. Because the anchors are of variable length and mismatch density, and can involve two or three species, we put a uniform probability measure on all events by simulating random sequence, aligning as for the genomic data, and then marking the 1% contour line in the length-mismatch parameter space. Anchors with a probability of 1% or less of occurring by chance for either two or three sequences are deemed acceptable. A table of the marginal number of mutations as a function of length is available in *Supplementary Material*.

For two-way alignments between *D. melanogaster* and *D. yakuba*, 90%–99% of indels (for γ = 600…1,200) have good anchors on both sides. In two-way comparisons, all the sequences are scored as either indels or ungapped blocks, and the dependence of statistics on the gap initiation penalty (γ) defines one's confidence in the result. Three-way alignments are needed only to distinguish insertions and deletions, and two other filters are imposed as described in the next section, before we make this assignment.

Insertion and Deletion Statistics

Three-way alignments were done for the REG data set. *Drosophila pseudoobscura* was treated as an out-group, and hence we were able to call insertions-deletions only when either *D. melanogaster* or *D. yakuba*, but not both, had a gap. The obscura group, of which *D. pseudoobscura* is a member, diverged from the melanogaster group between 25 (Russo, Takezaki, and Nei 1995) and 30 (Schlotterer et al. 1994) Myr ago, while *D. melanogaster* and *D. yakuba* diverged roughly 10 Myr ago (Powell 1997). We did not infer any insertions/deletions in *D. pseudoobscura* or in the common ancestor of the two in-group species.

The alignment between the two in-group species is of excellent quality as already noted. Two-way alignments between *D. melanogaster* and *D. pseudoobscura* have good anchors bracketing 81%–88% of indels (for γ = 800…1,200). Thus, the ambiguities where they exist are local and typically involve overlapping gaps in two species, with their relative positions sensitive to the alignment parameters. To filter out such unstable indels, we proceed as follows: the gap initiation penalty γ is varied in the broad range $\{500, 600, \ldots, 1{,}200\}$, alignments are done, and insertions/deletions are detected for each value of γ. We then intersect the sets of indels corresponding to $(\gamma - 200)$, γ, $(\gamma + 200)$, keeping only those indels that have the same type (insertion or deletion) in all three sets. This procedure screens for indel events that are robust to local changes of γ, while still allowing a useful plot of events versus γ. This filter eliminates 40% of indel events for γ = 500 and 22% for γ = 1,200.

Given a stable alignment, there is still ambiguity in cleanly distinguishing insertions from deletions by maximum parsimony, as shown in figure 1. Scenarios (A) (or (B)) can be unambiguously called as a single insertion (or deletion) event in an in-group species because the alternative explanation would involve two events. Scenario (C) is called a deletion in the in-group (IG2), overlapping a deletion in the out-group (OG), and scenario (D) is called as overlapping deletion events in the two in-group species (IG1 and IG2). Alternative explanations in either case would require at least three events. Scenarios (E) and (F) have ambiguous explanations, and we therefore use the lengths of the indels (in bp) to choose the most parsimonious one. In either case, there are two competing explanations, each with two events—an in-group event and an out-group event. (In scenario (F), there is a third possibility with two events on the two in-group species, but this is ignored because the likelihood of an in-group event is much lower than the likelihood of an out-group event, given the phylogenetic distances between species.) The total length of the events (in bp) under each explanation is computed, and their ratio taken. If this ratio (or its inverse) is less than a threshold of 0.8, the parsimonious event is inferred; otherwise, the indel is labeled as being ambiguous, and dropped.

This second filter eliminates a further 39% of the insertion/deletion events for γ = 500 (57% for γ = 1,200). For those that remain, we examine the anchors and find that 90%–95% of the insertions have good anchors both sides (for γ = 800…1,200). This level of uncertainty will not

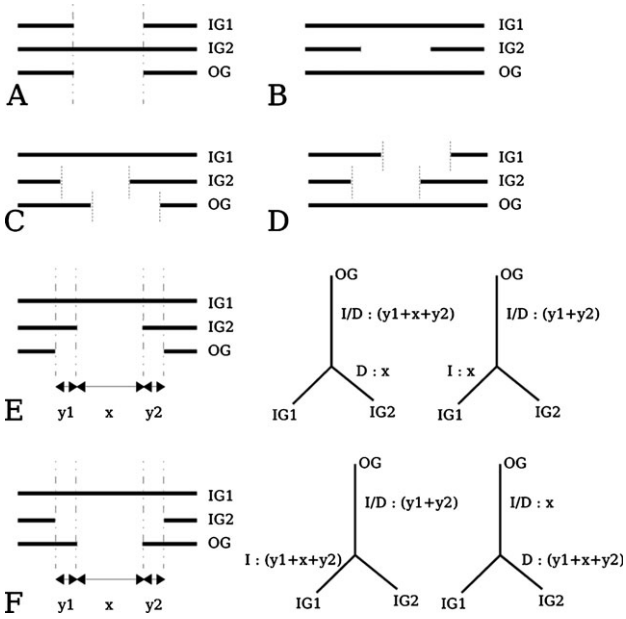

FIG. 1.—Maximum parsimony of insertions and deletions. All possible ways in which gaps may occur in three-way alignments of the in-group species IG1, IG2, and the out-group species OG. Bold lines indicate sequence, and "holes" in the lines indicate gaps in the alignment. (*A*)–(*D*): Unambiguous call of insertion (*A*) or deletion (*B–D*) in the in-group species because alternative explanations would require more number of events. (*E*), (*F*): There exist two alternative explanations, shown by two different trees next to each alignment. A label on a branch in the tree describes what the event is (I: insertion, D: deletion) along that edge and how many base pairs are involved. An event on the out-group branch is always ambiguous because an insertion in the out-group can also be interpreted as a deletion in the common ancestor of the in-group species, and vice versa. To choose one of the two trees, the total length (in base pairs) of events is calculated and if one is some threshold factor smaller than the other, it is chosen; otherwise the indel is declared ambiguous.

impact any of our conclusions. All three-way alignments for $\gamma = 800$ are available, with an indication of which events are retained after the filtering steps, at http://uqbar.rockefeller.edu/~saurabh/turnover/alignments.html. An unbiased sample of the alignments (for the first five modules in lexicographic order) is presented in *Supplementary Material*.

Tandem Repeat Statistics

Two different programs were used to detect tandem repeats because our experience with the two programs shows some exclusivity in the repeats they are able to detect. The first of them, called "Tandem Repeat Finder" (TRF) is from Benson (1999) and was run with parameter settings (match = 2, mismatch = 3, indel = 5, match probability = 0.8, indel probability = 0.1, minimum score = 25, maximum period = 500). This program detects repeats by first searching for exactly repeated short words and extending such "seeds" to longer approximate repeats. It scores the repeats using a stochastic model specified by percent identity and frequency of insertions and deletions. The second, called "Mreps" is described in Kolpakov, Bana, and Kucherov (2003) and was executed with parameters (resolution = 3, minimum period = 3). Mreps finds approximate repeats without relying on exactly repeated seeds, and its statistical criteria for reporting repeats are also different from that of TRF. The two programs were run separately, and we also computed the union of their results by overlaying the tandem repeats found by each on the same sequence. Each program reports the length (periodicity) and copy number (two or more) of each detected repeat. These statistics were also noted. The stringency of tandem repeat detection by each program was evaluated by running it on a randomly generated sequence of length 20 kbp and computing the masked bases as a fraction of the total length, and doing 100 replicates of this experiment. TRF masks 1.5% (±0.5) of random sequence on an average, Mreps masks 1.48% (±0.47), and their union masks 2.55% (±0.63) on average.

Transcription Factor–Binding Sites

A set of 32 CRMs involved in A-P axis patterning in Drosophila was selected. Position weight matrices (PWMs) of the transcription factors Bicoid, Hunchback, Caudal, Kruppel, Knirps, Tailless, Giant, Dstat, and the torRE binding factor were obtained from Rajewsky et al. (2002). The Stubb program (Sinha, van Nimwegen, and Siggia 2003) was run on each module separately with all the above PWMs and made to predict binding site occurrences that have posterior probability above 0.3. The same exercise was repeated for each *D. melanogaster* enhancer, as well as its orthologous sequence in *D. pseudoobscura*. (In a separate project we have shown by protein homology modeling that there are no residue changes in positions where the abovementioned protein factors contact the DNA and correspondingly, no systematic changes in the inferred binding sites when experimental sites are mapped between the species.)

Results

Substitutions and Indels

The pairwise alignments between *D. melanogaster* and *D. yakuba* depend weakly on the scoring parameters; hence, we plot their statistics against the gap initiation penalty γ (fig. 2). A larger γ penalizes new gaps more, thus the relative number of mismatched columns goes up as γ increases. (The histogram of lengths of ungapped aligned blocks in the REG data set, for $\gamma = 800$, is presented in *Supplementary Material*.)

We define the substitution rate in a two-way alignment as the fraction of aligned columns that have mismatched bases. Regulatory sequences (REG) have a substitution rate in the range 0.1–0.12 for $\gamma = 500$–1,200. Jukes-Cantor correction for multiple hits yields a corresponding range of 0.11–0.13. (When alignments are ambiguous, there will be a much greater variation of the inferred point mutation rate with γ because the most mutated regions will be classified as gaps for small γ.) For comparison, two random sequences with a 40%/60% GC/AT bias, aligned without gaps, would have a substitution rate of 0.74, while the maximum substitution rates when gaps are allowed ($\gamma = 1{,}200$) is ~0.5, as estimated from simulations. To compare with

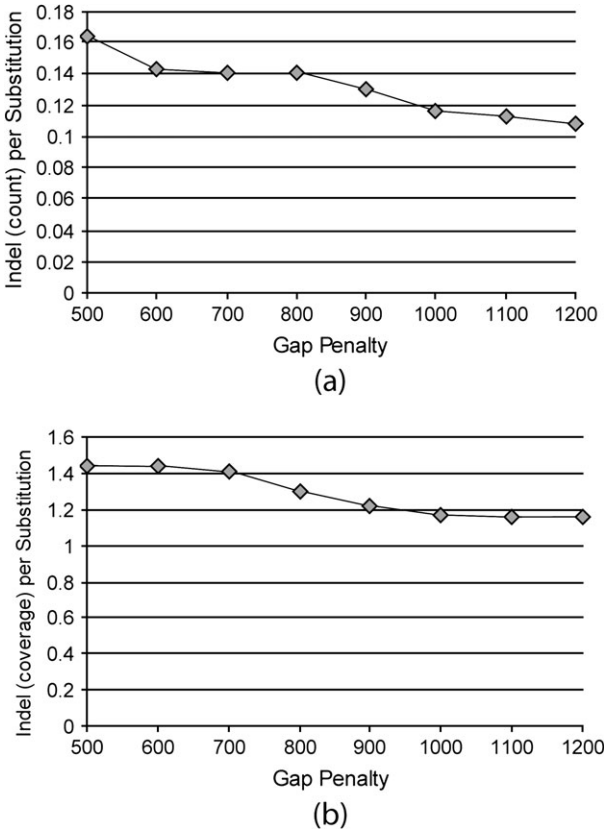

FIG. 2.—Indels per substitution in regulatory (REG) sequence based on number of events (*a*), and total length of indels (*b*). Plotted values are medians over all sequences in the corresponding data set.

the substitution rate in the REG set, we computed the synonymous substitution rate in functional genes (from the PGENES data set) between *D. melanogaster* and *D. yakuba*. The proportion of synonymous sites that are changed between the two species, i.e., the p_S value according to Nei and Gojobori (1986) had an average of 0.23 ($\pm$0.06). (Jukes-Cantor correction yields an average *dS* of 0.28 $\pm$ 0.08.) We thus observe greater conservation in regulatory regions, an evidence of functional constraints.

An *indel* is a contiguous stretch of gaps (in the same species) in an alignment. It could be the result of an insertion or a deletion. Figure 2*a* counts the *number* of indels per substitution, while figure 2*b* counts the sum of their *lengths* (per substitution). Thus, indels occur at roughly 10% of the point mutation rate but account for slightly more base pairs of change. Similar values were found for the REG2 data set. For the PGENES data set, the indel to substitution rate is in the range of 0.02–0.05 (resp. 0.2–0.6) based on counts (resp. coverage) over a range of γ. (The large uncertainties reflect the limited amount of reliable data we have on pseudogenes.) The median length of an indel is about 5–6 bp in regulatory regions (REG).

In summary, we have found (1) a point substitution rate of 0.1–0.12 in REG versus a synonymous substitution rate of 0.23 $\pm$ 0.06 in *D. melanogaster–D. yakuba* genes and (2) the number of indels is an order of magnitude less than the number of substitutions but accounts for more base pairs of change.

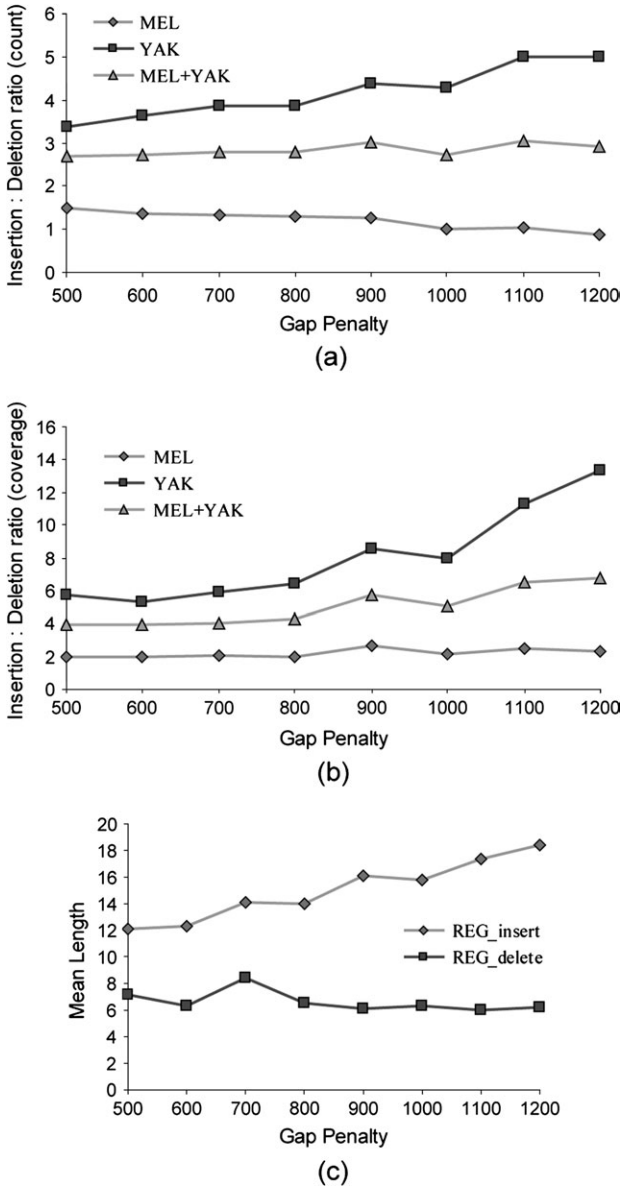

FIG. 3.—Insertion and deletion events in regulatory sequences. (*a*) Ratio of raw numbers of insertions and deletions. (*b*) Ratio of total lengths of insertions and deletions. The insertions and deletions are counted for *Drosophila melanogaster* alone (MEL), for *Drosophila yakuba* alone (YAK), and for both species together (MEL + YAK). (*c*) Mean lengths of insertions and deletions. REG_insert and REG_delete: insertions and deletions (respectively) in regulatory sequences.

Insertions and Deletions

We can distinguish insertions from deletions in the REG data set by using *D. pseudoobscura* as an out-group, and figure 3*a* plots the ratio of their number and figure 3*b* the ratio of their coverages. Insertions outweigh deletions in regulatory regions while in neutral regions Petrov and Hartl (1998) find a ratio of 1:8, consistent with the rapid loss of neutral sequence. (We have too few events in our PGENES data set to make an accurate statement.) There are over 200 insertion/deletion events in the REG data for all values of γ, all robust to local changes in parameters.

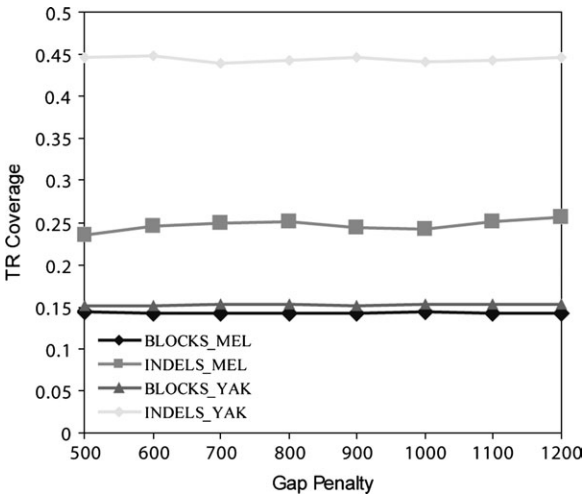

FIG. 4.—Tandem repeat coverage. The fraction of lengths of sequences that are marked as tandem repeats. All numbers are for regulatory sequences (REG). BLOCKS_MEL and BLOCKS_YAK: TR coverage in *aligned* columns, for *Drosophila melanogaster* and *Drosophila yakuba* separately. INDELS_MEL and INDELS_YAK: TR coverage in unaligned positions (gaps), for *D. melanogaster* and *D. yakuba*, respectively.

Our results largely hold true even if we include all detected indels, instead of restricting to the unambiguous ones. The enrichment for insertions over deletions was also significant in the REG2 data set, where the events were between *D. melanogaster* and *D. pseudoobscura*, with *D. virilis* serving as out-group (data not shown). Figure 3(*a* and *b*) also shows that the insertion to deletion ratio is higher in *D. yakuba* than in *D. melanogaster*.

The trends evident in figure 3(*a* and *b*) suggest that insertions are longer than deletions for regulatory sequence. Figure 3*c* makes this point quantitatively. The mean length of the insertion/deletion events in the regulatory regions is insensitive to the gap penalty, reflecting the unambiguous alignments. For neutral sequences, deletions were found to be longer than insertions (Petrov and Hartl 1998).

We repeated our analysis on noncoding sequence from *Drosophila erecta*, obtained from Bergman et al. (2002), with their orthologs from *D. melanogaster* and *D. yakuba*, the latter two being the in-group species, with *D. erecta* used as out-group. The total length of this sequence was over 130 kbp in each species. The three-way alignments were of excellent quality, and ambiguities due to the parsimony criteria were also rare. We again found in this data set an excess of insertions over deletions.

Net Sequence Retention

Two-way comparisons can reveal both the net change in length of functional sequence between the two species and the shrinkage of neutral sequence. Petrov and Hartl (1998) found that due to a considerable excess of deletions over insertions, a pseudogene in Drosophila has a "half-life" (average time for sequence length to halve) of about 14.3 Myr. Thus, assuming a divergence time of 25–30 Myr between *D. melanogaster* and *D. pseudoobscura*, we expect a pseudogene to shrink to about a quarter of its length if it became redundant at the time of speciation.

In contrast, for the sequences in REG, the length ratio between the two species has an average of 1.2 ± 0.26, i.e., the lengths are maintained to within 20% of each other on an average and furthermore as shown in figure 3*b*, both *D. melanogaster* and *D. yakuba* regulatory regions are increasing relative to their common ancestor. We therefore inquired as to the conservation in length of all the noncoding sequences and chose *D. melanogaster* and *D. pseudoobscura* for comparison so as to allow time for more change. We aligned large portions (>50 Mbp) of the two genomes, extracted ungapped blocks of high sequence conservation (more than 10 bp long and more than 70% identical), and considered the spacing between such blocks. We restricted attention to those pairs of blocks that are separated by between 500 and 1,000 bp (typical size of regulatory modules) in *D. melanogaster* and observed the corresponding spacing in *D. pseudoobscura*. We found that 81% of the time, the spacing did not change by more than 50% of its length in *D. melanogaster*.

To exclude the possibility that most of the noncoding sequence is simply shrinking at the neutral rate in both genomes, we repeated the analysis that leads to figure 3*b* for 100 randomly chosen noncoding regions of length 1 kbp from *D. melanogaster*, along with their orthologs from *D. yakuba* and *D. pseudoobscura*. We found an insertion to deletion (coverage) ratio between 3 and 6 (for $\gamma = 500 \ldots 1{,}200$), which is slightly below those reported in figure 3*b* (MEL + YAK), but clearly shows a predominance of insertions. This predominance is also evident when counting the *numbers* of events, similar to figure 3*a*. Furthermore, 79%–89% of the random noncoding regions had an excess of insertions over deletions and 78%–82% had more base pair coverage of insertions. Considering the rate of sequence loss from neutral regions, this is evidence that most of the noncoding sequence is of comparable functionality between the two species.

Tandem Repeats

We have seen above that insertions play an important role in the sequence turnover in CRMs. A question that then begs itself is—"how is new regulatory sequence created?" Here, we examine the possible role of tandem repeats in the creation of new regulatory sequence. A tandem repeat is a sequence that is repeated at least twice, in tandem. However, in reality, the tandem repeats we detect need not be exact copies nor do they have to occur strictly in tandem. Mutations and small indels accumulated during evolution may induce inexactness on these tandem repeats. This renders their detection difficult, and sophisticated programs such as TRF (Benson 1999) and Mreps (Kolpakov, Bana, and Kucherov 2003) have been developed to solve this problem computationally. We applied each of these programs on the regulatory sequences and computed a union of the detected repeats.

We see, in figure 4, that the fraction of sequence (in REG) that is covered by tandem repeats is significantly higher in indels than in aligned regions ($P < 10^{-100}$ for *D. melanogaster*, and $P = 0$ for *D. yakuba*, for $\gamma = 800$, binomial proportions test) but not significantly different between insertions and deletions (data not shown).

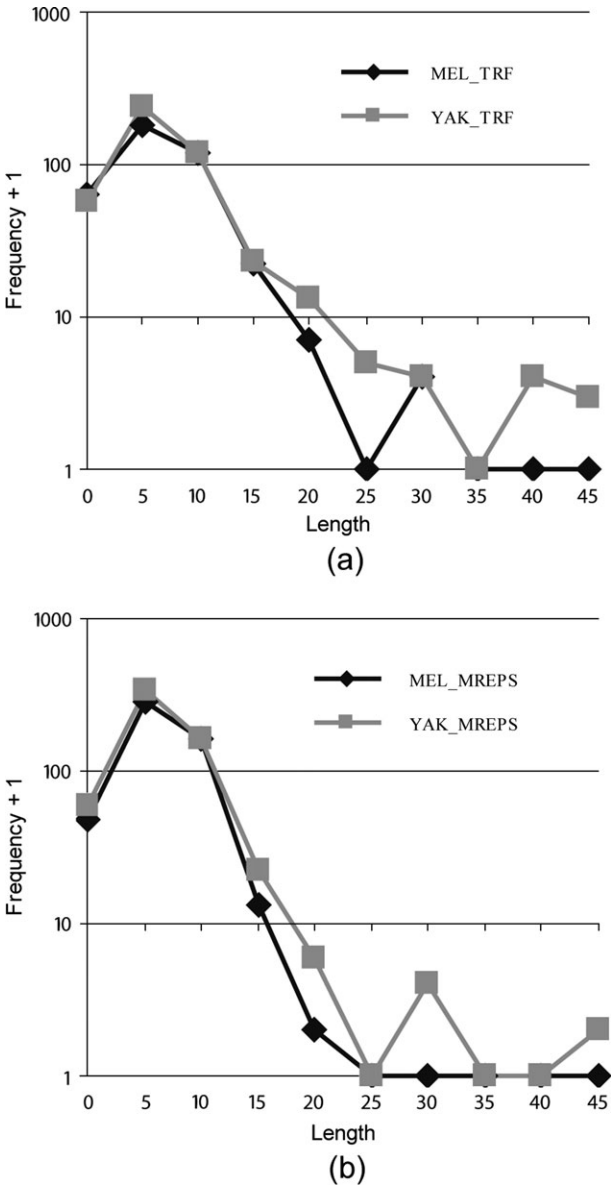

FIG. 5.—Length distribution of tandem repeats as reported by each of the two programs (*a*) TRF and (*b*) MREPS on regulatory sequences from *Drosophila melanogaster* (MEL) and *Drosophila yakuba* (YAK). The *y* axis has logarithmic scale. On the *x* axis, the point "*x*" represents repeats of length *x* to *x* + 4.

Overall, approximately 18% of *D. yakuba* sequence and 15% of *D. melanogaster* sequence is covered by tandem repeats, considerably more than expected by chance (2.55% ± 0.63%, based on random simulations; see *Materials and Methods*). Moreover, the repeat coverage in indels is 45% in *D. yakuba* and 25% in *D. melanogaster*. Figure 5*a* and *b* shows the distributions of the length of (each repeating unit of) tandem repeats as detected by the two programs separately. The typical length of the repeating unit is between 5 and 10 bp, with repeats longer than 20 bp being rare. The typical copy number of repeats is 2 or 3 (data not shown). We do not report the above statistics for the PGENES data set because most of these sequences are vestiges of coding sequence. (The tandem repeat coverage of the functional coding sequences is about 10%.)

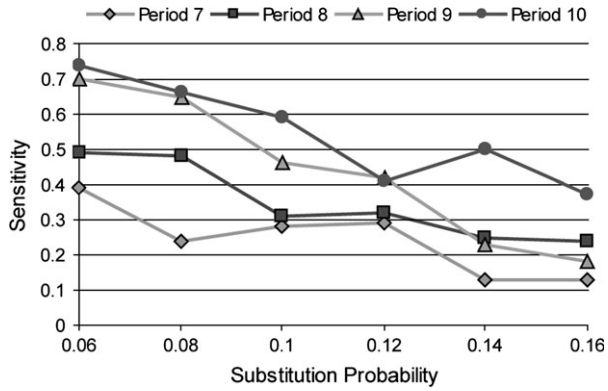

FIG. 6.—Sensitivity of tandem repeat detection. The *x* axis is the substitution probability between initial synthetic sequence (where the repeat was planted) and the final sequence (where the repeat detection was done). Two programs TRF and Mreps were used to detect repeats. The *y* axis is the fraction of times that the planted repeat was detected in the final sequence. The different tracks represent different repeat lengths. ("Period" refers to the length of each repeating unit.) In all experiments, exactly two copies of the repeat were planted in tandem.

As mentioned above, the sensitivity of tandem repeat detection is reduced due to mutations in the repeats. To quantify this effect, we performed simulation studies where we artificially created a tandem repeated sequence, subjected it to a specific substitution probability, and observed how often the combination of the repeat-finding programs is able to detect the repeat. (No indels were included in the simulations.) This exercise was done for various values of the repeat length and substitution probability, with a copy number of 2. Figure 6 reports the results for repeat length of 7–10 and substitution probability in the interesting range of values (0.06–0.16). For each combination of values, 100 independent simulations were done, and we report the fraction of times that the repeat was detected. Note that at a substitution rate of 0.1–0.12, which is what we find in regulatory sequences (REG) and for repeats of length 7–10, the programs retrieve about 25%–60% of the ancestral repeats. Thus, the 45% and 25% tandem repeat coverage seen in indels, which considers only detectable repeats, is consistent with all of the regulatory sequence insertions being generated via tandem repeats. Interestingly, the same conclusion can be reached using the REG2 data set and *D. melanogaster–D. pseudoobscura* comparisons. Here, the substitution probability is in the range 0.25–0.30, and the repeat detection efficiency in this range (for repeats of length 10 and copy number 2) is about 5%–15%. The tandem repeat coverage of indels is about 20% in *D. melanogaster* and about 30% in *D. pseudoobscura*, which is again consistent with all insertions arising from repeats.

Transcription Factor–Binding Sites in Tandem Repeats

Because CRMs are rich in tandem repeats and because they harbor several transcription factor–binding sites, we sought to determine the overlap between repeats and binding sites, having hypothesized that tandem repeats are an important mechanism for introducing new copies of sites. The simplest question to ask here is: "what fraction of binding sites are in the form of tandem repeats?" We ran both

Table 2
Examples of Predicted Binding Sites Masked as Tandem Repeats

Species	Factor	Site Count	Masked Count	Masked Fraction	Site Strength[a]	CRM[b]
Drosophila pseudoobscura	Bicoid	27	9	0.33	11.72	btd_head[c]
Drosophila pseudoobscura	Kruppel	23	9	0.39	10.09	h_stripe7[d]
Drosophila melanogaster	Hunchback	22	8	0.36	7.69	h_stripe7[d]
Drosophila pseudoobscura	Knirps	20	8	0.4	4.66	h_stripes_3+4[d]
Drosophila pseudoobscura	Hunchback	20	7	0.35	4.8	h_stripe7[d]
Drosophila pseudoobscura	Kruppel	27	7	0.25	11.09	h_stripes_1+5[d]
Drosophila melanogaster	Hunchback	44	7	0.15	14.83	run_stripe3[d]
Drosophila melanogaster	Hunchback	31	6	0.19	10.98	Kr_CD2_AD1[d]
Drosophila melanogaster	Knirps	19	6	0.31	4.55	h_stripe6[d]
Drosophila melanogaster	Hunchback	29	6	0.2	10.07	kni_kd[e]

[a] Sum of strengths of prediction (on a scale of 0 to 1) of all sites of the factor.
[b] *cis*-Regulatory module.
[c] Wimmer et al. (1995).
[d] Emberly et al. (2003).
[e] Pankratz et al. (1992).

TRF and Mreps programs on known CRMs involved in patterning of the early Drosophila embryo, on *D. melanogaster* sequences and their *D. pseudoobscura* orthologs separately. On the same sequences, we also detected putative binding sites of various maternal and gap transcription factors using the Stubb program (Sinha, van Nimwegen, and Siggia 2003) and computed what fraction of these predicted binding sites is repeat masked. (A separate study showed that a program similar to Stubb successfully retrieves over 50% of known sites [Schroeder et al. 2004].) We found that about 20% of predicted binding sites in *D. pseudoobscura* and about 15% of those in *D. melanogaster* overlap (>50% of site length) with tandem repeats.

Table 2 presents some examples of binding sites occurring in repeated sequence. For instance, the hairy stripe7 enhancer in *D. pseudoobscura* has 23 strong and weak (predicted) binding sites of the transcription factor Kruppel. (If each site is scored between 0 and 1 based on its "strength" [Sinha, van Nimwegen, and Siggia 2003], the sum of scores for the 23 sites comes out to 10.09, indicating that the sites are medium strength on average.) Of the 23 predicted sites, 9 (39%) are repeat masked. Similarly, we find (predicted) binding sites of various other factors overlapping with tandem repeats in this and other enhancers.

Discussion

Methodology

Harrison et al. (2003) identified over 100 (gene-pseudogene) pairs in *D. melanogaster*. We then did a custom protein to nucleotide alignment between them and imposed the conditions: $dS \leq 1.25$ to ensure good alignments, $dN/dS \geq 0.3$ so that the time between gene duplication and degeneration of one copy is small, and scored only regions interior to exons. Very little sequence remained. Evidently the formation of pseudogenes has been a rare phenomena since the *D. melanogaster–D. pseudoobscura* split. (The synonymous codons are almost randomized between these two species.)

Our count of indels assumes that each contiguous insertion or deletion is a single event. If two insertions (or deletions, or an insertion and a deletion) occur at the same locus (the "problem of multiple hits"), this assumption is no longer valid. We therefore present statistics on the indel *coverage*, with the goal of summarizing the net change through indels, rather than making claims about their *rates* over short evolutionary periods.

Our measurements of insertions and deletions are based on gapped regions detected by the alignment program, and the traditional interpretation of alignment gaps being indels. It is possible that occasionally, adjacent copies of an insertion and a deletion (of similar lengths) are reported where in reality they are orthologous sequences that are unalignable due to their high divergence. In such cases, our procedure would overcount indels. The proximity of our two in-group species (*D. melanogaster* and *D. yakuba*) minimizes the problem and in addition we consider a large range of gap penalties, and only report insertions and deletions that are invariant under change of gap penalty. As the gap penalty increases, such regions should become aligned, with high substitution rate, and will therefore not be counted as indels. The enrichment for insertions in regulatory regions is observed even at high gap penalties, therefore arguing for the validity of our conclusion. Also, the observation that the reported indels contain a significantly higher proportion of tandem repeats implicates them as being insertion or deletion events rather than unalignable orthologous sequences.

Regulatory Sequences Are Rich in Insertions

One important and unexpected finding was the preponderance of insertions over deletions in regulatory sequence, as measured by coverage of events of each type. We wish to emphasize that this finding is based on the assumption that the maximum parsimony principle is applicable for the phylogeny considered here. The overall length of the REG set (after removing end-gaps in the alignments) is 99.5 kbp in *D. melanogaster* and 103.3 kbp *D. yakuba*. A rough calculation shows that their last common ancestor had approximately the same length (of sequences in REG) as *D. melanogaster*, which is about 4% smaller than *D. yakuba*, arguing that the regulatory regions are undergoing

rapid turnover. The rapid elimination of neutral sequence by contrast shows that the regulatory regions are as small as possible consistent with function. The absence of large deletions in the regulatory regions (in comparison with the neutral rates) is an example of Fisher's geometric model of adaptation (Fisher 1930), which predicts that mutations of large effect (e.g., large deletions) are more likely to be deleterious. That more sequence change arises from indels rather than point substitutions necessitates their inclusion in models of regulatory evolution.

Bergman et al. (2002) have studied the conservation patterns in noncoding regions of *D. melanogaster* and four other species (*D. erecta*, *D. pseudoobscura*, *Drosophila willistoni*, and *Drosophila littoralis*), but they did only pairwise alignments and searched for regions with a high density of conserved ungapped bases in order to screen for functional regulatory sequence. They did not compute indel statistics or contrast the changes in neutral versus functional sequence. Kim (2001) also studies conservation patterns in a regulatory sequence (the hairy enhancer) in *D. melanogaster* and six other Drosophila species. This study finds highly conserved blocks interspersed with highly variable regions but does not attempt to discriminate between insertions and deletions.

Currently, the best evidence of sequence loss in neutral regions comes from Petrov and Hartl (1998) who observed a 1:8 ratio of insertions to deletions, with substantially longer deletions than insertions on average. They used multiple, closely related species, so their ability to score events is better than ours. The average size of their insertions was ~3 bp and they had nine events, while our PGENES set gave an average indel size of 4–6 bp. Both numbers are substantially smaller than the typical size of insertions to regulatory sequence. Zhang and Gerstein (2003) found a 1:3 ratio of insertions to deletions for mammalian pseudogenes.

Bergman and Kreitman (2001) studied the presence of indels in noncoding sequences implicated in *cis*-regulation in Drosophila. They performed two-way comparison of 100 kbp of noncoding sequence (at more than 40 loci) between *D. melanogaster* and *D. virilis*. While they did not count insertions and deletions separately, they found an overall indel per substitution rate of less than 0.05, which is somewhat lower than what we observe. This is because their indel definition was more conservative than ours, requiring conserved sequence blocks (minimum length 10 bp, minimum similarity 70%) on either side. In doing so, they retained only a small subset of the indels in their data. They found a point substitution rate of 7.2% and an indel rate of 0.32% derived from 96 indels with a median length of 2 bp. On the other hand, we classify all sequence (in two-way alignments) as indels or as aligned blocks, partly because we work with closer species than they did and partly because our data set comprises experimentally verified CRMs only. We therefore impose no explicit constraints on what delineates an indel, although the average length of an ungapped block is ~15–25 bp, similar to that in Bergman and Kreitman (2001). For instance, at a gap penalty value of $\gamma = 800$, the median length of ungapped blocks is 28; 91% of the ungapped blocks are over 10 bp long and 89% have a percent identity of more than 70%.

Regulatory Sequences Are Rich in Tandem Repeats

The second important focus of this study is the role of tandem repeats in sequence turnover, particularly for regulatory sequences. We find short tandem repeats ("mini"-satellites) with two to three copies of a repeat unit of 3–20 bp that cover 15%–18% of regulatory modules overall and a far higher fraction, 25%–45%, of the indels. These repeats are quite different from conventional microsatellites (Ellegren 2004), which have more copies of a shorter repeat unit, and have been used as genetic markers for a long time. By extrapolation of our ability to detect tandem repeats, based on the observed point mutation rate between *D. melanogaster* and *D. yakuba* regulatory regions, all indels could be due to tandem repeats. We intentionally chose fairly tolerant scoring parameters in our TRFs, resulting in $2.55 \pm 0.63\%$ of random sequence being masked. The observed fraction is still very significant, and more stringent parameters would make the correction for our detection efficiency a bigger extrapolation.

It is an almost universal rule that the sites for the factors that regulate a CRM in fly are present in multiple copies. We do not posit that tandem duplications occur preferentially at functional protein-binding sites; however, in comparison with point mutations, random duplication of sequence could certainly shorten the time necessary to create a sequence variant with enough fitness advantage to rapidly fix in the population.

An unexpected finding for the regulatory data is that deletion events are as repeat rich as insertions. It is generally thought that tandem duplications arise by the copying of contiguous sequence through polymerase slippage or unequal crossing over (Achaz 2002; Ellegren 2004). These same mechanisms can operate in reverse and remove a duplicated sequence once it has been created, provided there have not been point mutations which destroy the periodicity and act as a ratchet to favor the retention of tandem insertions. To distinguish insertions from deletions, we required comparison with the *D. pseudoobscura* out-group. Hence, what we score as a deletion in *D. melanogaster* or *D. yakuba* must have been present in *D. pseudoobscura*, whereas insertions are newer sequences, having arisen since the split between the first two species. Because our ability to detect tandem repeats is limited by substitutions to relatively recent events (fig. 6), we again expected to find more repeats in insertions than in deletions. However, the molecular mechanisms of indels are varied, for instance nontandem repeated sequence can also mediate the addition and removal of sequence at rates that depend very much on whether the sequence between the repeats contains palindromes (Rosche, Ripley, and Sinden 1998). Selection further complicates reconstructing the fate of duplicated sequence. These problems do not however affect the observed repeat coverage of indels as a group.

Several previous studies have touched upon the issue of tandem repeats and their function. Their connection to insertions and deletions in neutral sequence was noted in Petrov and Hartl (1998), who studied the evolution of the non-LTR retrotransposable element Helena in the *D. melanogaster* subgroup. They observed that about half of the deletions were flanked by tandem repeats of size

1–7 bp, and six of the nine insertions (with average size 2.9 ± 3.5 bp) were also tandem repeats in the same size range. The insertions and tandem repeats found in that study are typically smaller than those in our data set and also fewer in number.

Papatsenko et al. (2002) also explored the link between tandem repeats and Drosophila enhancers and found a few examples where weak or medium-strength predicted binding sites form tandem clusters in fly enhancers. However, unlike a claim made in that work, we found no striking conservation of tandem repeats in general. We find a greater repeat coverage in indels than in aligned regions. Our study aims at performing a comprehensive assessment of repeat coverage of regulatory sequences, the focus being on associating these repeats with sequence turnover.

Thomas et al. (2004) searched for exact repeats 25 bp or greater within several kilobase pairs of each other, for several mammalian genomes. They used multiple species to distinguish insertions from deletions and found a preponderance of the former. Because of their very stringent filter they only found several thousand such doublets in the human genome and not a meaningful number in the fly. Achaz, Netter, and Coissac (2001) in an earlier study explored a broader class of repeats in a variety of genomes but did not focus so specifically on the rates of various processes (see also Achaz 2002).

Hancock et al. (1999) observed a spatial clustering of specific short words in 5′ and 3′ regions of the hunchback gene and found binding sites (for Bicoid in this case) in such "repeat" regions. However, these repeats are not necessarily in tandem and are merely more clustered than expected by chance. Several instances of tandem repeats (minisatellites) being associated with regulatory function are available from the literature (Trepicchio and Krontiris 1992; Shinder et al. 1994; Shi et al. 2000; Carroll, Grenier, and Weatherbee 2001; Andrioli et al. 2002; Lovejoy et al. 2003). In some cases, direct repeats of binding sites may lead to cooperative binding of a transcription factor. For example, Bicoid bound to a strong site helps Bicoid bind to a nearby weak site (Burz et al. 1998).

Mobile elements have for some time been postulated to be an important catalyst for regulatory change (see Britten 1996 and references therein.) These repeat elements are much larger than the events we have categorized.

Conclusion

We have contrasted the evolution of known regulatory modules with neutral sequence. The regulatory sequence roughly maintains its length as expected (because most modules retain their function between species), with insertions and deletions playing significant but balancing roles. The tendency of regulatory regions in both *D. melanogaster* and *D. yakuba* to increase relative to their common ancestor was unexpected, especially when contrasted with the magnitude of the sequence loss from the neutral regions. The former observation extended to generic noncoding sequence plus the consistency in the length of homologous noncoding regions argues that most of the sequenced euchromatin regions in fly are functional. The rates of sequence loss in mammals are far slower (Petrov and Hartl 1998), so a similar argument cannot be made.

Our detection of repeats is limited to tandem ones, though limited indels between and within the repeating units are allowed. We are not sensitive to repeats spaced by the size of the repeat unit or more. On the scale of a regulatory module, one can look for such repetitions with standard motif finders. Limited computational experiments of this sort exist, e.g., Rajewsky et al. (2002), and when done on well-characterized CRMs, suggest that there are high-quality repeats which do not easily match the factors known to regulate the module. The functionality of such sites has not been investigated experimentally. With more extensive data on closer species, it would be interesting to use point mutations that disrupt strict periodicity to date the repeats and to more accurately correlate them with insertion and deletion events. Also the repeat size deserves to be better quantified—it appears to be comparable to a protein-binding site.

Merely creating a mechanism to copy sequence does not mean that protein-binding sites will be copied more often than nonsites. However, if multiple sites entail some cooperativity in protein binding, then copying will certainly increase the variance in fitness of the module and thus decrease the time necessary to find a variant with a selective advantage. The tails of the fitness distribution can matter a great deal for the rate of evolutionary innovation. It is also not excluded that active binding sites *are* preferentially copied, e.g., an enzyme could mark the unprotected sites with a "do not copy" signal, or the sequence bias of slippage could amplify certain binding sites preferentially. Thus, the biochemical machinery required to copy short sequence elements and place them within the confines of a module could itself be actively selected.

Supplementary Material

1. Genomic coordinates of CRMs used in the analysis (REG and REG2 data sets).
2. Table used in deciding if an aligned block is a good anchor. For a specific gap initiation penalty (γ), for any given length l, this table lists the maximum number of columns with mismatches that may be present in a block of length l for it to be considered a good anchor. (There are separate tables for two-way and three-way alignments.)
3. Three-way alignments for five modules in the REG data set. These are the first five modules in lexicographic order of their names.
4. Histogram of block lengths from two-way alignments of *D. melanogaster* and *D. yakuba* (REG) for $\gamma = 800$.

Acknowledgments

Support was provided by the NSF under grant DMR0129848, the NIH under grant GM66434, and the Keck foundation (to S.S.). We thank Guillaume Achaz and Colin Meiklejohn for discussions and comments on a preliminary version of the manuscript.

Literature Cited

Achaz, G. 2002. Etude de la dynamique des génomes: les répétitions intrachromosomiques. Doctoral dissertation, L'Universite Pierre and Marie Curie, Paris, France. (http://www.oeb.harvard.edu/faculty/wakeley/guillaume/index.html).

Achaz, G., P. Netter, and E. Coissac. 2001. Study of intrachromosomal duplications among the eukaryote genomes. Mol. Biol. Evol. **18**(12):2280–2288.

Akam, M. 1998. Hox genes, homeosis and the evolution of segment identity: no need for hopeless monsters. Int. J. Dev. Biol. **42**(3):445–451.

Andrioli, L., V. Vasisht, E. Theodosopoulou, A. Oberstein, and S. Small. 2002. Anterior repression of a Drosophila stripe enhancer requires three position-specific mechanisms. Development **129**(21):4931–4940.

Bailey, T. L., and C. Elkan. 1995. Unsupervised learning of multiple motifs in biopolymers using expectation maximization. Mach. Learn. **21**(1–2):51–80.

Bateman, A., L. Coin, R. Durbin et al. (13 co-authors). 2004. The Pfam protein families database. Nucleic Acids Res. **32**: D138–D141.

Benson, G. 1999. Tandem repeats finder—a program to analyze DNA sequences. Nucleic Acids Res. **27**(2):573–580.

Bergman, C., and M. Kreitman. 2001. Analysis of conserved noncoding DNA in Drosophila reveals similar constraints in intergenic and intronic sequences. Genome Res. **11**: 1335–1345.

Bergman, C., B. Pfeiffer, D. Rincon-Limas et al. (17 co-authors). 2002. Assessing the impact of comparative genomic sequence data on the functional annotation of the Drosophila genome. Genome Biol. **3**(12).

Berman, B. P., Y. Nibu, B. D. Pfeiffer, P. Tomancak, S. E. Celniker, M. Levine, G. M. Rubin, and M. B. Eisen. 2002. Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome. Proc. Natl. Acad. Sci. USA **99**: 757–762.

Britten, R. 1996. DNA sequence insertion and evolutionary variation in gene regulation. Proc. Natl. Acad. Sci. USA **93**(18):9374–9377.

Brudno, M., C. Do, G. Cooper, M. Kim, E. Davydov, E. Green, A. Sidow, S. Batzoglou, and NISC Comparative Sequencing Program. 2003. LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. Genome Res. **13**(4):721–731.

Burz, D., R. Rivera-Pomar, H. Jackle, and S. Hanes. 1998. Cooperative DNA-binding by Bicoid provides a mechanism for threshold-dependent gene activation in the Drosophila embryo. EMBO J. **17**(20):5998–6009.

Bustamante, C., R. Nielsen, and D. Hartl. 2002. A maximum likelihood method for analyzing pseudogene evolution: implications for silent site evolution in humans and rodents. Mol. Bio. Evol. **19**:110–117.

Carroll, S., J. Grenier, and S. Weatherbee. 2001. From DNA to diversity: molecular genetics and the evolution of animal design. Blackwell Scientific, Malden, Mass.

Carter, A., and G. Wagner. 2002. Evolution of functionally conserved enhancers can be accelerated in large populations: a population-genetic model. Proc. R. Soc. Lond. B. Biol. Sci. **269**(1494):953–960.

Celniker, S., D. Wheeler, B. Kronmiller et al. (32 co-authors). 2002. Finishing a whole genome shotgun: release 3 of the Drosophila melanogaster euchromatic genome sequence. Genome Biol. **3**(12).

Davidson, E. 2001. Genomic regulatory systems. Academic Press, San Diego, Calif.

Dermitzakis, E., C. Bergman, and A. Clark. 2002. Tracing the evolutionary history of Drosophila regulatory regions with models that identify transcription factor binding sites. Mol. Biol. Evol. **20**(5):703–714.

Dermitzakis, E., and A. Clark. 2002. Evolution of transcription factor binding sites in Mammalian gene regulatory regions: conservation and turnover. Mol. Biol. Evol. **19**:1114–1121.

Ellegren, H. 2004. Microsatellites: simple sequences with complex evolution. Nature Rev. Genet. **5**(6):435–445.

Emberly, E., N. Rajewsky, and E. Siggia. 2003. Conservation of regulatory elements between two species of Drosophila. BMC Bioinformatics **4**(57).

Fisher, R. 1930. The genetical theory of natural selection. Oxford University Press, Oxford.

Genome Sequencing Center at Washington University Medical School. 2004. Drosophila yakuba genome. (http://www.genome.wustl.edu/projects/yakuba/).

Halfon, M., Y. Grad, G. Church, and A. Michelson. 2002. Computation-based discovery of related transcriptional regulatory modules and motifs using an experimentally validated combinatorial model. Genome Res. **12**:1019–1028.

Hancock, J., P. Shaw, F. Bonneton, and G. Dover. 1999. High sequence turnover in the regulatory regions of the developmental gene hunchback in insects. Mol. Biol. Evol. **16**: 253–265.

Harrison, P., D. Milburn, Z. Zhang, P. Bertone, and M. Gerstein. 2003. Identification of pseudogenes in the Drosophila melanogaster genome. Nucleic Acids Res. **31**(3):1033–1037.

Human Genome Sequencing Center at Baylor College of Medicine. 2003. Drosophila genome project. (http://www.hgsc.bcm.tmc.edu/projects/drosophila/).

Kassis, J., C. Desplan, D. Wright, and P. O'Farrell. 1989. Volutionary conservation of homeodomain-binding sites and other sequences upstream and within the major transcription unit of the Drosophila segmentation gene engrailed. Mol. Cell. Biol. **9**:4304–4311.

Kim, J. 2001. Macro-evolution of the hairy enhancer in Drosophila species. J. Exp. Zool. **291**(2):175–185.

Kolpakov, R., G. Bana, and G. Kucherov. 2003. Mreps: efficient and flexible detection of tandem repeats in DNA. Nucleic Acids Res. **31**(13):3672–3678.

Lawrence, C. E., S. F. Altschul, M. S. Boguski, J. S. Liu, A. F. Neuwald, and J. C. Wootton. 1993. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. Science **262**:208–214.

Li, W. 1997. Molecular evolution. Sinauer Associates, Sunderland, Mass.

Lovejoy, E., A. Scott, C. Fiskerstrand, V. Bubb, and J. Quinn. 2003. The serotonin transporter intronic VNTR enhancer correlated with a predisposition to affective disorders has distinct regulatory elements within the domain based on the primary DNA sequence of the repeat unit. Eur. J. Neurosci. **17**(2):417–420.

Ludwig, M., N. Patel, and M. Kreitman. 1998. Functional analysis of eve stripe 2 enhancer evolution in Drosophila: rules governing conservation and change. Development **125**(5):949–958.

Markstein, M., P. Markstein, V. Markstein, and M. S. Levine. 2002. Genome-wide analysis of clustered Dorsal binding sites identifies putative target genes in the Drosophila embryo. Proc. Natl. Acad. Sci. USA **99**:763–768.

Nasmyth, K. 2001. A prize for proliferation. Cell **107**(6):689–701.

Nei, M., and T. Gojobori. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol. **3**(5):418–426.

Pankratz, M., M. Busch, M. Hoch, E. Seifert, and H. Jackle. 1992. Spatial control of the gap gene knirps in the Drosophila

embryo by posterior morphogen system. Science **255**(5047): 986–989.

Papatsenko, D. A., V. J. Makeev, A. P. Lifanov, M. Regnier, A. G. Nazina, and C. Desplan. 2002. Extraction of functional binding sites from unique regulatory regions: the Drosophila early developmental enhancers. Genome Res. **12**:470–481.

Petrov, D., and D. Hartl. 1998. High rate of DNA loss in the Drosophila melanogaster and Drosophila virilis species groups. Mol. Biol. Evol. **15**(3):293–302.

Powell, J. 1997. Progress and prospects in evolutionary biology: the Drosophila model. Oxford University Press, New York.

Rajewsky, N., M. Vergassola, U. Gaul, and E. Siggia. 2002. Computational detection of genomic cis-regulatory modules applied to body patterning in the early Drosophila embryo. BMC Bioinformatics **3**(30).

Rebeiz, M., N. Reeves, and J. Posakony. 2002. SCORE: a computational approach to the identification of cis-regulatory modules and target genes in whole-genome sequence data. Proc. Natl. Acad. Sci. USA **99**:9888–9893.

Rockman, M., M. Hahn, N. Soranzo, D. Goldstein, and G. Wray. 2003. Positive selection on a human-specific transcription factor binding site regulating IL4 expression. Curr. Biol. **13**(23):2118–2123.

Rosche, W., L. Ripley, and R. Sinden. 1998. Primer-template misalignments during leading strand DNA synthesis account for the most frequent spontaneous mutations in a quasi-palindromic region in Escherichia coli. J. Mol. Biol. **284**: 633–646.

Russo, C., N. Takezaki, and M. Nei. 1995. Molecular phylogeny and divergence times of Drosophilid species. Mol. Biol. Evol. **12**:391–404.

Schlotterer, C., M. Hauser, A. von Haeseler, and D. Tautz. 1994. Comparative evolutionary analysis of rDNA ITS regions in Drosophila. Mol. Biol. Evol. **11**:513–522.

Schroeder, M., M. Pearce, J. Fak, H. Fan, U. Unnerstall, E. Emberly, N. Rajewsky, E. Siggia, and U. Gaul. 2004. Transcriptional control in the segmentation gene network of Drosophila. PLoS Biol. **2**(9).

Shi, X., H. Blair, X. Yang, J. McDonald, and X. Cao. 2000. Tandem repeat of C/EBP binding sites mediates PPARgamma2 gene transcription in glucocorticoid-induced adipocyte differentiation. J. Cell Biochem. **76**(3):518–527.

Shinder, G., S. Manam, B. Ledwith, and W. Nichols. 1994. Minisatellite DNA-binding proteins in mouse brain, liver, and kidney. Exp. Cell Res. **213**(1):107–112.

Sinha, S., and M. Tompa. 2000. A statistical method for finding transcription factor binding sites. Pp. 344–354 *in* Bourne, P., M. Gribskov, R. Altman, N, Jensen, D. Hope, T. Lenauer, J. Mitchell, E. Scheef, C. Smith, S. Strande, H. Weissig, eds. Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology. AAAAI Press, Menlo Park, Calif.

Sinha, S., E. van Nimwegen, and E. Siggia. 2003. A probabilistic method to detect regulatory modules. Bioinform **19**(Suppl. 1): 292–301.

Tautz, D. 2000. Evolution of transcriptional regulation. Curr. Opin. Genet. Dev. **10**:575–579.

Thomas, E., N. Srebro, J. Sebat, N. Navin, J. Healy, B. Mishra, and M. Wigler. 2004. Distribution of short paired duplications in mammalian genomes. Proc. Natl. Acad. Sci. USA **101**(28):10349–10354.

Trepicchio, W., and T. Krontiris. 1992. Members of the rel/NF-kappa B family of transcriptional regulatory proteins bind the HRAS1 minisatellite DNA sequence. Nucleic Acids Res. **20**(10):2427–2434.

Wilkins, A. 2002. The evolution of developmental pathways. Sinauer Associates, Sunderland, Mass.

Wimmer, E., M. Simpson-Brose, S. Cohen, C. Desplan, and H. Jackle. 1995. Trans- and cis-acting requirements for blastodermal expression of the head gap gene buttonhead. Mech. Dev. **53**(2):235–245.

Wong, W., and R. Nielsen. 2004. Detecting selection in noncoding regions of nucleotide sequences. Genetics **167**:949–958.

Wray, G., M. Hahn, E. Abouheif, J. Balhoff, M. Pizer, M. Rockman, and L. Romano. 2003. The evolution of transcriptional regulation in eukaryotes. Mol. Biol. Evol. **20**(9):1377–1419.

Zhang, Z., and M. Gerstein. 2003. Patterns of nucleotide substitution, insertion and deletion in the human genome inferred from pseudogenes. Nucleic Acids Res. **31**(18):5338–5348.

Edward Holmes, Associate Editor

Accepted December 14, 2004