

# Analysis of Conserved Noncoding DNA in *Drosophila* Reveals Similar Constraints in Intergenic and Intronic Sequences

Casey M. Bergman<sup>1</sup> and Martin Kreitman

Department of Ecology and Evolution, University of Chicago, Chicago, Illinois 60637, USA

Comparative genomic approaches to gene and *cis*-regulatory prediction are based on the principle that differential DNA sequence conservation reflects variation in functional constraint. Using this principle, we analyze noncoding sequence conservation in *Drosophila* for 40 loci with known or suspected *cis*-regulatory function encompassing >100 kb of DNA. We estimate the fraction of noncoding DNA conserved in both intergenic and intronic regions and describe the length distribution of ungapped conserved noncoding blocks. On average, 22%–26% of noncoding sequences surveyed are conserved in *Drosophila*, with median block length ~19 bp. We show that point substitution in conserved noncoding blocks exhibits transition bias as well as lineage effects in base composition, and occurs more than an order of magnitude more frequently than insertion/deletion (indel) substitution. Overall, patterns of noncoding DNA structure and evolution differ remarkably little between intergenic and intronic conserved blocks, suggesting that the effects of transcription per se contribute minimally to the constraints operating on these sequences. The results of this study have implications for the development of alignment and prediction algorithms specific to noncoding DNA, as well as for models of *cis*-regulatory DNA sequence evolution.

The functional annotation of eukaryotic genomic sequences represents one of the greatest challenges in modern biology. Therefore, a diversity of computational approaches have emerged to identify genes and the *cis*-regulatory sequences controlling their expression. A promising class of methods for both gene and *cis*-regulatory prediction is based on comparative sequence analysis (Batzoglou et al. 2000; Loots et al. 2000). These approaches work because functionally constrained sequences are often conserved in evolution, much more so than nonfunctional sequences. Although why comparative sequence analysis enhances functional predictions is widely recognized, the link between molecular evolution and functional constraint is rarely made explicit. This link is most clearly formulated under the neutral theory of molecular evolution, which quantitatively relates functional constraint with the rate and pattern of sequence evolution (Kimura 1983; see also Gillespie 1991). Acknowledging this connection implies that constructing models of molecular evolution should be relevant to the development of methods that predict function from comparative genomic sequence data.

We are specifically interested in modeling the molecular evolution of *cis*-regulatory sequences controlling developmentally regulated gene expression in *Drosophila* (Ludwig et al. 1998, 2000). *Drosophila* is an excellent model system to explore the link between comparative and functional representations of *cis*-regulatory sequences. First, *Drosophila melanogaster* is a complex animal with a compact, completely sequenced genome with excellent physical and genetic maps (Adams et al. 2000; Hoskins et al. 2000). With such rapid progress in the completion of the *D. melanogaster* genome, sequencing of additional *Drosophila* genomes for comparative

analysis is a distinct possibility. Second, this species has a rapid and cost-effective transgenic system that can be adapted for rescue, reporter, misexpression, or knockout studies to test the function of predicted *cis*-regulatory sequences (Ashburner 1989; Rorth et al. 1998; Rong and Golic 2000). Furthermore, recent developments allow the possibility of reciprocal, cross-species transgenic analysis (Horn and Wimmer 2000). Third, the molecular genetics of many developmentally important *cis*-regulatory regions and pathways are well understood, providing the necessary functional context to test predictions based on comparative sequence analysis (Lawrence 1992). Finally, the phylogeny and evolutionary genetics of the genus *Drosophila* present a well-described range of divergence times to calibrate comparative and functional models (Powell 1997). Therefore in *Drosophila*, all of the tools are in place to critically test predictions about *cis*-regulatory structure and function based on comparative sequence data.

For these reasons, the use of comparative sequencing has become a common technique in the analysis of *cis*-regulatory structure/function in *Drosophila*. Unfortunately, the utility of such data in predicting *cis*-regulatory function is limited, as little is known about the expected features of *cis*-regulatory molecular evolution (Stern 2000). A major difficulty impeding the quantitative analysis of *cis*-regulatory sequence evolution is the lack of a framework for the a priori statistical interpretation of noncoding DNA, akin to the genetic code for protein coding sequences. Empirically, however, the pattern of *cis*-regulatory molecular evolution in *Drosophila* and other species can be qualitatively described by highly conserved blocks of DNA separated by unalignable gaps. Conserved blocks of noncoding DNA likely result from the sequence-specific constraints of DNA-protein interactions, although the correspondence between functionally characterized binding sites and conserved sequences is not exact (Dickinson 1991). Remarkably, this mode of molecular evolution does

**<sup>1</sup>Corresponding author.**

**E-MAIL** [cbergma@midway.uchicago.edu](mailto:cbergma@midway.uchicago.edu); **FAX (773) 702-9740**.  
Article and publication are at <http://www.genome.org/cgi/doi/10.1101/gr.178701>.

not lead to drastic changes in the pattern of gene expression across species, as assayed by interspecific transgenic analysis (Tautz 2000). These seemingly paradoxical observations have led to a model in which stabilizing selection acts on the phenotype of gene expression, allowing a flux in the composition of the underlying *cis*-regulatory sequence (Ludwig et al. 2000; Carroll et al. 2001). Despite these insights into the mode of *cis*-regulatory molecular evolution, the comparative analysis of noncoding DNA is only beginning to be placed in a quantitative analytical framework.

In this paper, we study constraints operating on *Drosophila* noncoding sequences to gain insight into the expected features of *cis*-regulatory molecular evolution. Specifically, we use pairwise sequence analysis of noncoding DNA between *D. melanogaster* and *Drosophila virilis* to study molecular evolutionary constraints acting on >100 kb of noncoding DNA sampled from 40 loci scattered throughout the *Drosophila* genome. These two species have been separated for ~40 million years, a divergence that is approximately equal to that between human and mouse, and more than sufficient to discern functional constraint in noncoding sequences (Blackman and Meselson 1986; Hartl and Lozovskaya 1994; Kwiatowski et al. 1994; Russo et al. 1995). We focus our attention primarily on noncoding regions that have been shown functionally to contain *cis*-regulatory activity in at least one of the two species. Both intergenic and intronic DNA is surveyed to analyze the effects of transcription on noncoding molecular evolutionary constraints. This report addresses the following questions about the structure and evolution of conserved noncoding DNA: (1) What is the fraction and density of conserved DNA in noncoding regions? (2) What is the length distribution of ungapped, conserved noncoding blocks? (3) What is the rate and pattern of point substitution in conserved noncoding blocks? and (4) What is the rate and pattern of indel substitution in conserved noncoding blocks? We compare constraints in *Drosophila* noncoding DNA to that of other species, as well as to other types of sequences in the *Drosophila* genome. We evaluate our results using several tools for genome alignment to substantiate our findings and benchmark automated ap-

proaches. Finally, we suggest future prospects for the evolutionary and functional analysis of conserved noncoding DNA.

## RESULTS

The results of our survey for noncoding regions exhibiting primary sequence conservation in *Drosophila* identified the 40 loci in Table 1. The loci are scattered among all five major

**Table 1.** Names, Cytological Positions in *Drosophila melanogaster*, and Accession Nos. of NonCoding Sequences Used in This Study

Intergenic locus	Cytological position	<i>Drosophila melanogaster</i>	<i>Drosophila virilis</i>
<i>achaete-scute</i> †	1B1	AL024453/AL023873	AF060607/AF132809
<i>Antennapedia</i>	84B1	AC001655	M95827
<i>bride of sevenless</i>	96F9	AC019700	L08132
<i>brown</i>	59E2-3	AC005639	L37035
<i>decapentaplegic</i> †	22F1-2	AC004369/AC019923	U95037/X81976
<i>dopa decarboxylase</i> †	37C1	AC007176	X05065/Johnson et al. 1989
<i>D-mef2</i>	46C1-2	AC014124/AC014416	Cripps et al. 1998; Gajewski et al. 1997
<i>e74</i>	74F1	AC019594	X59493
<i>engrailed</i>	48A2	AC020381	Kassis et al. 1989
<i>frmf-amide</i>	46C1-2	AC015179	AH000028
<i>fused</i>	17C5-7	AE003509	U20586
<i>fushi tarazu</i>	84B1	AE001573	Schier and Ghering 1993
<i>glass</i>	91A1	AC014473	U39746
<i>hair</i> †	66D10	AC014797	AF329639/AF329640/Langeland and Carroll 1993
<i>hunchback</i>	85A6-11	U17742	S70575
<i>knirps</i>	77E2	AC020104	L36177
<i>paramyosin</i>	66D14	AE003554	AJ243069
<i>prospero</i>	86E3	AC013194/AC012748	AF190404
<i>runt</i>	19E2	AE003570	Wolff et al. 1999
<i>tailless</i>	100B1	AC014779	AF019361
<i>teashirt</i> †	40A1-4	AC006467	McCormick et al. 1995
<i>tinman</i> †	93E1	AC020256	Lee et al. 1997; Xu et al. 1998
<i>troponin T</i>	12A1-4	AC014434	AJ002263
<i>twist</i>	59C3	AC005975	Pan et al. 1994
<i>wingless</i>	27F1-2	AC017528	AF046865
<i>zerknüllt</i>	84A5	AC002512	L17339
Intronic locus	Cytologic position	<i>Drosophila melanogaster</i>	<i>Drosophila virilis</i>
<i>Antennapedia</i>	84B1	AC001655/AC020267	M95827/M95828
<i>bride of sevenless</i>	96F9	AC019700	L08132
<i>corkscrew</i>	2D3	AC017610	U22356
<i>decapentaplegic</i>	22F1-2	AC019923	U63855
<i>engrailed</i>	48A2	AC020381	Kassis et al. 1986
<i>glass</i>	91A1	AC014473	U39746
<i>Gpdh</i>	26A7-9	AC017294	D10697
<i>hunchback</i>	85A6-11	U17742	X15395
<i>knirps</i>	77E2	AC020104	L36177
<i>miniparamyosin</i>	66D14	AE003554	AJ243070
<i>myosin light chain</i>	98A6	AC013071	L08053
<i>paralytic</i>	14D1-E1	AC014944	U26718
<i>pdm-2</i>	33F1-2	AC006470	U14723
<i>prospero</i>	86E3	AF190403	AF190405
<i>rough</i>	97D5	AC014838	M35372
<i>sevenless</i>	10A2	AC017584	M34544
<i>single minded</i>	87E1	AC020412	AF071932
<i>Staufen</i>	55B4-5	AC004336	AF225924
<i>tinman</i>	93E1	AC020256	Yin et al. 1997
<i>trithorax</i>	88B1-2	AC013943	Z50038
<i>vestigial</i>	49D3-4	AC014851	Williams et al. 1994

If a GenBank accession no. is not indicated, the data has been obtained from the primary reference or by personal communication. A dagger indicates that more than one pair of contigs were surveyed for that locus.

chromosomal arms of the *D. melanogaster* genome, and therefore reflect a sample that is more or less random with respect to positional influences. Some loci have data for both intergenic and intronic regions, therefore the total number of regions surveyed is greater than the total number of loci. Several loci fitting our criteria for inclusion in the data set had no conserved noncoding blocks distinguishable from background similarity using our methods, and were excluded from further analyses (*Adh*, *Amy*, *Gld*, *RP140*, *sisA*, *Sxl*, *Rh1-4*, *elav*, and *su(s)*). For those loci that did show substantial conservation by our methods, initial attempts to find parameters of Wilbur-Lipman, Needleman-Wunsch, and pairwise BLAST algorithms that consistently gave comparable results were unsuccessful, frequently failing to recover easily identifiable dot-matrix homology blocks. This is likely a result of the fact that alignment methods designed for coding sequences perform poorly when stretches of homology are short and gaps are frequent and variable, as is true for noncoding DNA (for further discussion, see Jareborg et al. 1999). Using the combined output of several tools specifically developed for noncoding or genomic alignment, however, we were able to recapitulate results similar to those obtained by filtered dotplot.

Even by our relatively stringent criteria, we find that substantial amounts of intergenic and intronic noncoding DNA in *Drosophila* are subject to primary sequence constraint. When pooled across loci, 29,915 bp (20,501 bp intergenic, 9414 bp intronic) are contained within conserved block sequences, out of 114,015 bp (79,874 bp intergenic, 34,141 bp intronic) and 138,831 bp (95,592 bp intergenic, 43,239 bp intronic) of DNA surveyed in *D. melanogaster* and *D. virilis*, respectively (Table 2). We note that each species has the same amount of DNA in the conserved block component of the data set by definition. Although the fraction of conserved block sequence ranges considerably across regions as a consequence of variation in the length of unalignable DNA, the average fraction of DNA under primary sequence constraint appears to differ little between intergenic and intronic DNA. Globally, there

is a significantly higher fraction of conserved noncoding DNA in the *D. melanogaster* genome relative to *D. virilis* for both intergenic ( $G = 431.65$ , 1 d.f.,  $P < 10^{-2}$ ) and intronic regions ( $G = 347.24$ , 1 d.f.,  $P < 10^{-12}$ ).

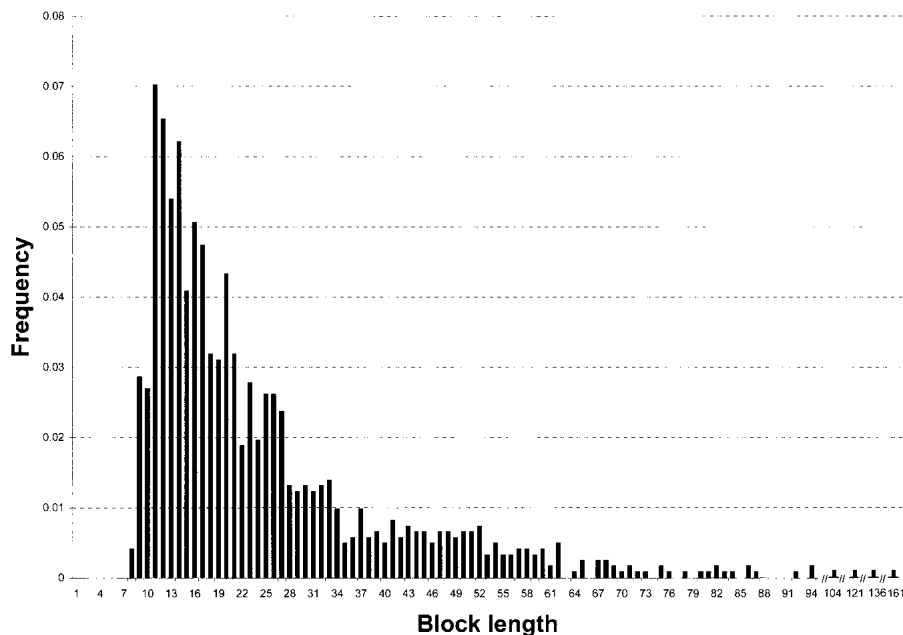
**Table 2.** Number of Nucleotides Surveyed, Conserved, and Percent Sequence Conservation of Intergenic and Intronic Regions Analyzed in This Study

Intergenic region	<i>Drosophila melanogaster</i> bp surveyed	<i>Drosophila virilis</i> bp surveyed	Bp conserved	<i>Drosophila melanogaster</i> % conserved	<i>Drosophila virilis</i> % conserved
<i>achaete-scute</i>	2151	<b>5434</b>	559	26	10
<i>Antennapedia</i>	<b>3894</b>	3624	799	21	22
<i>bride of sevenless</i>	1458	<b>1730</b>	194	13	11
<i>brown</i>	269	<b>286</b>	113	42	40
<i>decapentaplegic</i>	12271	<b>15,562</b>	4173	34	27
<i>dopadecarboxylase</i>	<b>590</b>	523	178	30	34
<i>D-mef2</i>	10117	<b>14,083</b>	1407	14	10
<i>e74</i>	294	<b>331</b>	120	41	36
<i>engrailed</i>	2222	<b>2716</b>	738	33	27
<i>frmf-amide</i>	2072	<b>2623</b>	438	21	17
<i>fused</i>	<b>347</b>	275	179	52	65
<i>fushi tarazu</i>	385	<b>464</b>	135	35	29
<i>glass</i>	<b>1779</b>	1281	488	27	38
<i>hairy</i>	11,942	<b>12,673</b>	3549	30	28
<i>hunchback</i>	3423	<b>3624</b>	808	24	22
<i>knirps</i>	<b>1631</b>	1559	327	20	21
<i>paramyosin</i>	1444	<b>1803</b>	295	20	16
<i>prospero</i>	5525	<b>7714</b>	2053	37	27
<i>runt</i>	5995	<b>6170</b>	498	8	8
<i>tailless</i>	3646	<b>3906</b>	1095	30	28
<i>teashirt</i>	787	<b>846</b>	416	53	49
<i>tinman</i>	<b>1531</b>	1213	445	29	37
<i>troponin T</i>	<b>427</b>	389	64	15	16
<i>twist</i>	<b>1322</b>	1153	301	23	26
<i>wingless</i>	3979	<b>5213</b>	973	24	19
<i>zerknüllt</i>	373	<b>397</b>	156	42	39
<b>Intergenic total</b>	<b>79,874</b>	<b>95,592</b>	<b>20,501</b>	<b>26</b>	<b>21</b>
Intronic region	<i>Drosophila melanogaster</i> bp surveyed	<i>Drosophila virilis</i> bp surveyed	Bp conserved	<i>Drosophila melanogaster</i> % conserved	<i>Drosophila virilis</i> % conserved
<i>Antennapedia</i>	2648	<b>2861</b>	881	33	31
<i>bride of sevenless</i>	1308	<b>1951</b>	399	31	20
<i>corkscrew</i>	521	<b>1102</b>	153	29	14
<i>decapentaplegic</i>	832	<b>929</b>	435	52	47
<i>engrailed</i>	1026	<b>1271</b>	416	41	33
<i>glass</i>	68	<b>111</b>	36	53	32
<i>Gpdh</i>	1448	<b>2346</b>	222	15	9
<i>hunchback</i>	2630	<b>3574</b>	892	34	25
<i>knirps</i>	614	<b>705</b>	273	44	39
<i>miniparamyosin</i>	1002	<b>1563</b>	251	25	16
<i>myosin light chain</i>	<b>1223</b>	1187	400	33	34
<i>paralytic</i>	<b>173</b>	171	80	46	47
<i>pdm-2</i>	<b>708</b>	642	215	30	33
<i>prospero</i>	4603	<b>5511</b>	1327	29	24
<i>rough</i>	2659	<b>4239</b>	627	24	15
<i>sevenless</i>	2266	<b>2606</b>	262	12	10
<i>single minded</i>	3119	<b>4777</b>	851	27	18
<i>Staufen</i>	1170	<b>1628</b>	349	30	21
<i>tinman</i>	309	<b>381</b>	158	51	41
<i>trithorax</i>	<b>5162</b>	4988	871	17	17
<i>vestigial</i>	652	<b>696</b>	316	48	45
<b>Intronic total</b>	<b>34141</b>	<b>43239</b>	<b>9414</b>	<b>28</b>	<b>22</b>
<b>Grand total</b>	<b>114015</b>	<b>138831</b>	<b>29915</b>	<b>26</b>	<b>22</b>

The amount of DNA conserved is by definition the same in each species, thus the size differences are due to length changes in unalignable sequences. The species with the larger regional size is shown in bold.

These results are expected as *D. melanogaster* is known to have a smaller genome size than *D. virilis* (Powell 1997).

We studied the density and length distribution of ungapped conserved blocks, which are important and yet unknown features of noncoding sequences that can be used to increase the sensitivity of genomic alignment and prediction tools. We observed 1225 (825 intergenic, 400 intronic) conserved noncoding blocks that we used to estimate these features empirically. The number of conserved blocks observed in intergenic and intronic DNA fit expected proportions based on the total amount of intergenic and intronic DNA surveyed in *D. melanogaster* ( $\chi^2 = 4.28$ , 1 d.f.,  $P < 0.039$ ) and *D. virilis* ( $\chi^2 = 1.30$ , 1 d.f.,  $P < 0.254$ ). Therefore, from the total data set, the average density of ungapped conserved noncoding blocks is estimated to be 10.7 conserved blocks per kilobase in *D. melanogaster* and 8.9 conserved blocks per kilobase in *D. virilis*. Furthermore, no significant differences were observed between the length distribution of intergenic versus intronic conserved blocks (Kolmogorov-Smirnov test,  $P < 0.10$ ). Because the density and length distribution of blocks does not appear to differ substantially among intergenic and intronic DNA, the data were pooled into one frequency distribution (Fig. 1). The distribution is highly skewed toward conserved blocks shorter than the mean (24.4 bp), with median and modal block lengths of 19 and 11 bp, respectively. Approximately 95% of conserved noncoding blocks are distributed between 10–71 bp, and only one intergenic block and three intronic blocks are >100 bp in length. Using a linear transformation of the data (variate = length – 8), to correct for truncation of the distribution at the lower extreme, we could not reject that our data are obtained from a discrete approximation to a lognormal distribution (mean 2.376, variance 0.926; Kolmogorov-Smirnov test,  $P > 0.10$ ), although other continuous (normal,  $\gamma$ ,  $\chi^2$ ) and discrete (binomial, Poisson, geometric) distributions could be rejected.



**Figure 1** Length distribution of ungapped conserved noncoding blocks in *Drosophila*. As the distributions of intergenic and intronic noncoding block lengths do not differ significantly, they are plotted together. The distribution is truncated at the lower extreme as a result of the criteria used to define conserved blocks (see Methods for details).

For each position in each conserved block, the nucleotide of both species was counted to derive a match–mismatch matrix for conserved noncoding blocks in *Drosophila* (Table 3). This information is critical for understanding the molecular evolutionary dynamics of conserved block substitutions, as well as the statistical evaluation of ungapped local alignments using extreme value (Karlin and Altschul 1990) or Bayesian theory (Zhu et al. 1998). As expected, the majority of sites remain unchanged between *D. melanogaster* and *D. virilis* and have a base composition ([AT]=60%) typical of noncoding regions in *Drosophila* (Moriyama and Hartl 1993). We observed 2157 (1503 intergenic, 654 intronic) nucleotide sites that differ within conserved block sequences. The number of substitutions in intergenic and intronic conserved blocks fit expected proportions ( $\chi^2 = 1.32$ , 1 d.f.,  $P < 0.250$ ), indicating that similar rates of substitution are observed in both types of conserved blocks. In the total data set, ~7.2% of the nucleotide sites in conserved noncoding blocks are substituted between *D. melanogaster* and *D. virilis*.

Not only the rate of substitution per basepair, but the entire structure of the match–mismatch matrix is similar for intergenic and intronic conserved blocks (Table 3). The frequency of a given observation differs between intergenic and intronic blocks by 2.3% or less for identities and 0.9% or less for substitutions. In contrast to similarity of substitution pattern across classes of DNA based on transcriptional state, there appears to be differences in the substitution pattern across species. This is especially apparent for reciprocal intergenic transitions, which show significant differences among observed counts of *mel* A: *vir* G (150) vs. *mel* G: *vir* A (212) ( $\chi^2 = 10.62$ , 1 d.f.,  $P < 1.1 \times 10^{-3}$ ) and *mel* C: *vir* T (241) vs. *mel* T: *vir* C (150) ( $\chi^2 = 21.18$ , 1 d.f.,  $P < 4.0 \times 10^{-6}$ ). Curiously, there is a trend across all reciprocal mismatch cells for substitutions to make *D. virilis* more AT-rich for both intergenic ( $\chi^2 = 20.51$ , 1 d.f.,  $P < 6.0 \times 10^{-6}$ ) and intronic ( $\chi^2 = 9.50$ , 1 d.f.,  $P < 2.1 \times 10^{-3}$ ) conserved blocks.

Unlike reciprocal substitutions, there are no significant differences observed in the pattern of complementary substitutions (i.e., *mel* A: *vir* G vs. *mel* T: *vir* C) across species.

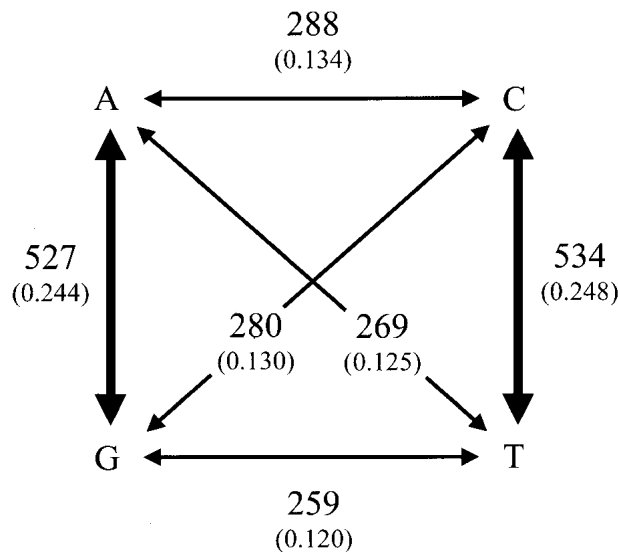
Although we can observe species differences in base usage at variable sites in conserved blocks, we are unable to polarize the direction of substitutions from pairwise sequence data. Considering this and the similarity of the substitution pattern in intergenic and intronic regions, we collapsed the data in Table 3 by summing the total numbers of reciprocal substitutions within the matrix (i.e., *mel* A: *vir* T + *mel* T: *vir* A = A ↔ T) to derive the relative rates of substitution among bases contained in conserved noncoding blocks in *Drosophila* (Fig. 2). This view of the data shows that the relative rates of purine and pyrimidine transition substitutions are equal to each other, as are relative rates for each of the four different transversion substitutions. Relative rates of individual transi-

**Table 3.** Nucleotide Match–Mismatch Table for Conserved Noncoding Blocks between *Drosophila melanogaster* and *Drosophila virilis*

<i>Drosophila melanogaster</i>	<i>Drosophila virilis</i>			
	A	C	G	T
<b>A</b>				
Intergenic	5501 (0.268)	86 (0.004)	150 <sup>a</sup> (0.007)	82 (0.004)
Intronic	2738 (0.291)	39 (0.004)	68 (0.007)	31 (0.003)
Total	8239 (0.275)	125 (0.004)	218 <sup>a</sup> (0.007)	113 (0.004)
<b>C</b>				
Intergenic	108 (0.005)	3803 (0.186)	92 (0.004)	241 <sup>a</sup> (0.012)
Intronic	55 (0.006)	1676 (0.178)	45 (0.005)	72 (0.008)
Total	163 (0.005)	5479 (0.183)	137 (0.005)	313 <sup>a</sup> (0.010)
<b>G</b>				
Intergenic	212 <sup>a</sup> (0.010)	107 (0.005)	3937 (0.192)	78 (0.004)
Intronic	97 (0.010)	36 (0.004)	1613 (0.171)	51 (0.005)
Total	309 <sup>a</sup> (0.010)	143 (0.005)	5550 (0.186)	129 (0.004)
<b>T</b>				
Intergenic	96 (0.005)	150 <sup>a</sup> (0.007)	101 (0.005)	5755 (0.281)
Intronic	60 (0.006)	71 (0.008)	29 (0.003)	2733 (0.290)
Total	156 (0.005)	221 <sup>a</sup> (0.007)	130 (0.004)	8488 (0.284)

Each cell represents the observed numbers of matched and mismatched bases for nucleotide positions aligned in intergenic, intronic, and total conserved blocks. Rounded frequencies of observations are in parentheses and are scaled relative to intergenic (20,499), intronic (9,414), or grand (29,913) totals, which do not include ambiguous nucleotides.  
<sup>a</sup>Significant differences between reciprocal substitutions at  $P < 0.05/18 = 0.00278$ .

tions are twofold greater than relative rates of individual transversions, indicating transition bias in these sequences. This twofold bias toward transitions leads to equal numbers of observed transition and transversion substitutions, as there are twice as many possible transversions as transitions. We caution that symmetry in the relative rate matrix should not be interpreted as stationarity in the substitution process in light of the lineage effects detected above.



**Figure 2** Relative rates of point substitution in *Drosophila* conserved noncoding blocks. Frequencies of observations are in parentheses below total observed counts. Reciprocal substitutions were pooled (i.e., *mel* A:*vir* T+*mel* T:*vir* A = A↔T) because our divergence data is pairwise, and therefore the polarity of the change is ambiguous. Bold arrows indicate transition substitutions.

In addition to the rate and pattern of point substitution, we can study the properties of conserved noncoding block indel substitution when two conserved blocks are contiguous in one species but interrupted by an insertion in the other species. There are 96 observations of this kind that can be ascribed to de novo indel events, although these events cannot be distinguished as insertions or deletions from pairwise data alone. *D. melanogaster* has insertion of this kind relative to *D. virilis* 49 times (32 intergenic, 17 intronic), and *D. virilis* has an insertion relative to *D. melanogaster* 47 times (31 intergenic, 16 intronic). Moreover, the length distribution of inserted sequences does not differ significantly between *D. melanogaster* and *D. virilis* (Kolmogorov-Smirnov test,  $P > 0.10$ ). From these observations, we can infer that the apparent rate and pattern of indel substitution in conserved noncoding blocks do not show lineage effects. The number of

indel events in intergenic versus intronic conserved blocks fits the expected proportions based on the total amount of conserved block DNA in each class ( $\chi^2 = 0.376$ , 1 d.f.,  $P < 0.539$ ). Additionally, no significant differences in the indel length distribution were observed between intergenic versus intronic regions (Kolmogorov-Smirnov test,  $P > 0.10$ ). Therefore, both the rate and pattern of indel substitution in *Drosophila* conserved noncoding blocks are similar in intergenic and intronic sequences.

Because the indel length distribution does not differ significantly across species or transcriptional state, and as we cannot discriminate insertions from deletions, we pooled indel substitutions for further analysis. The total rate of indel substitution in conserved block DNA is estimated to be 0.32%, a rate more than 20-fold less than point substitution. The length distribution of indel substitutions is skewed toward small (1–5 bp) sequences with a long tail of larger indels (Fig. 3A). The mean and median indel lengths are 7.73 bp and 2 bp, respectively. The relationship between the natural log of indel size and the natural log of indel frequency for the combined data set is linear and highly correlated (Spearman's coefficient of rank correlation;  $R = -0.740$ ,  $P < 2.4 \times 10^{-5}$ ) (Fig. 3B). This pattern has been found for an analysis of indel substitutions in a variety of data sets and suggests that the frequency distribution of indels follows a  $\zeta$  distribution (Johnson and Kotz 1969; Gu and Li 1995). We attempted to fit our data to this function by using the maximum likelihood estimate,  $\rho$ , and approximate variance of the  $\zeta$  distribution (Johnson and Kotz 1969). We estimate  $\rho$  ( $\pm$  one standard deviation) for indels in *Drosophila* conserved noncoding blocks to be  $0.6 \pm 0.06$ , a parameter estimate similar to that found for indels in organellar and mammalian nuclear DNA [note that  $1 + \rho = b$  of Gu and Li (1995)].

The results of all molecular evolutionary analyses are dependent on the underlying sequence alignments used, and our study is no exception. To substantiate the validity of our

alignments and to benchmark tools designed for the alignment of noncoding genomic regions, we performed a compatibility analysis of our method with several recently developed alignment platforms: DiAlign, DNA Block Aligner (DBA), VISTA, and Lamark (Morgenstern 2000; Dubchak et al. 2000; Jareborg et al. 1999; S. Shabalina and A. Kondrashov, pers. comm.). We considered a block to be compatible if conserved block sequences defined by our method were also contained within an alignment block produced by an automated method. Because alignment is expected to differ among methods, particularly around block edges, we did not require exact correspondence of alignments. Of the 1225 blocks identified by our filtered dotplot analysis, 888 (72.4%), 1030 (84.0%), 1074 (87.6%), and 1122 (91.5%) blocks were identified by DBA, VISTA, Lamark, and DiAlign, respectively. One thousand two hundred and nine (98.6%) conserved blocks identified by our method were identified by at least one of the four automated tools, 1158 (94.5%) by at least two, 1018 (83.0%) by at least three, and 745 (60.3%) were identified by all four.

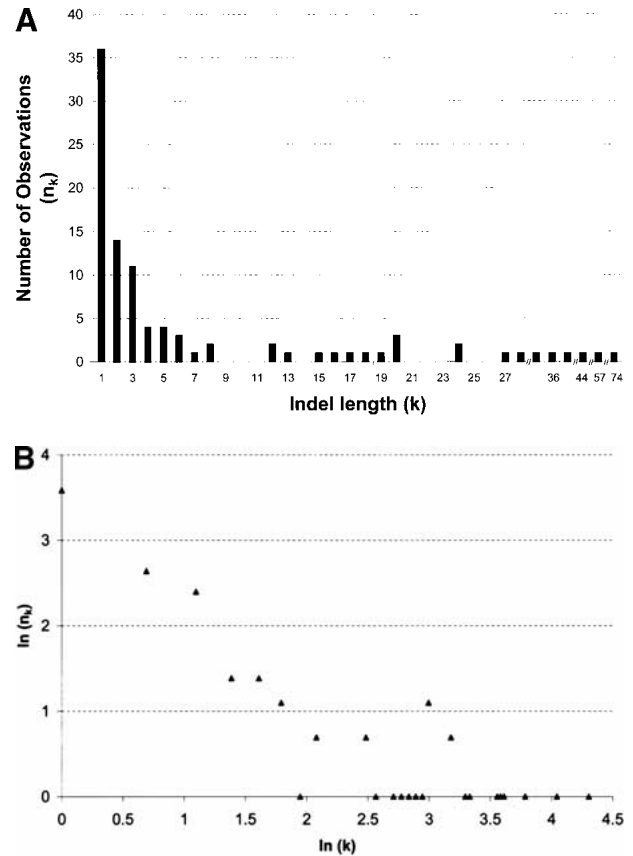
## DISCUSSION

### Estimates of Noncoding Sequence Conservation in Eukaryotic Genomes

Our finding that 22%–26% of *Drosophila* noncoding sequences are highly constrained is similar to published estimates of the fraction sequence conservation in noncoding regions of other complex eukaryotes. The fraction of conserved sequences in upstream and intron regions in the mouse genome have been estimated to be 36% and 23%, respectively, using a Hidden Markov Model (HMM) approach to noncoding alignment (Jareborg et al. 1999). Similarly, a comparison of both intergenic and intronic muscle-specific regulatory regions between human and mouse using a Bayesian alignment procedure reveals that 19% of human sequences are highly conserved (Wasserman et al. 2000). These two estimates may in fact be compatible given that the fraction of conserved sequences can vary across species because of changes in genome size, as we have observed in *Drosophila*. An analysis of conservation between *Caenorhabditis elegans* and *Caenorhabditis briggsae* using dynamic programming alignment methods reveals that for both intergenic and intronic regions the fraction of noncoding DNA under constraint is at least 18% for both species (Shabalina and Kondrashov 1999). Despite differences in organismal complexity, sampling, and alignment methods, these different analyses give remarkably similar values. In summary, preliminary estimates suggest that ~20%–30% of nucleotide sites may be expected to be conserved in functionally constrained noncoding regions of eukaryotic genomes.

### Structural and Evolutionary Properties of Conserved Noncoding Blocks in *Drosophila*

We found that average density and length distributions of conserved blocks are statistically indistinguishable between intergenic and intronic regions in our sample. The distribution of conserved block lengths pooled over intergenic and intronic regions reveals that the majority of noncoding sequence constraints act continuously over only short stretches of DNA. The median block length (19 bp) is small and the mass of the distribution lies between 8 and 75 bp (Fig. 1). We



**Figure 3** Length distribution of indel substitutions in *Drosophila* conserved noncoding blocks. Because the length distributions of indels for intergenic and intronic conserved blocks do not differ significantly, they are plotted together. (A) Number of observations ( $n_k$ ) plotted by length ( $k$ ). (B) Correlation of  $\ln(n_k)$  and  $\ln(k)$ . The Spearman coefficient of rank correlation between  $\ln(n_k)$  and  $\ln(k)$  is  $R = -0.740$  ( $P < 2.4 \times 10^{-5}$ ).

were not able to reject the hypothesis that the pooled length distribution of conserved blocks is generated by a lognormal function using the best-fit parameter estimates from the data, although we could reject other continuous and discrete distributions. Although the blocks of conservation we detected are in general quite small, they are on average longer than the length of a single transcription factor binding site, and therefore likely correspond to the module level of *cis*-regulatory structure (Arnone and Davidson 1997). It is also interesting to note that the distribution of conserved block lengths is quite different from the distribution of exon lengths in *Drosophila*; conserved noncoding blocks are much shorter on average than the expected length of *Drosophila* exons (141.1 bp) (Deutsch and Long 1999). These data should help the construction or parameterization of alignment and prediction algorithms that discriminate noncoding from coding DNA.

In addition to block length, the rate and pattern of point substitution also did not differ statistically between intergenic and intronic blocks. We estimate that ~7% of nucleotide sites are substituted in conserved noncoding blocks between *D. melanogaster* and *D. virilis*, a value similar to one obtained for a sample of loci between mouse and human (Wasserman et al. 2000). Substituted sites within highly constrained noncoding

sequences showed two noteworthy features in the relative rates of point substitution — transition bias and lineage effects in base composition (Table 2; Fig. 2). We observed a 2:1 bias toward transition substitutions in our data, which is similar to estimates based on the divergence of coding sequences and nonfunctional dead-on-arrival retro-elements (Moriyama and Powell 1997; Petrov and Hartl 1999). Polymorphisms in *Drosophila* noncoding and fourfold degenerate coding sites also show a 2:1 transition rate bias (Moriyama and Powell 1996). Therefore a 2:1 bias toward transitions may be a general feature of molecular evolution throughout the *Drosophila* genome. The relative contributions of mutation, selection, and other evolutionary forces to generating this pattern, however, remain unclear. Evidence for the second notable feature, lineage-specific changes in base composition, has also been observed for synonymous substitutions in *Drosophila*-coding sequences (Moriyama and Hartl 1993; Rodríguez-Trelles et al. 2000). Therefore, changes in base composition among different lineages may ramify throughout both noncoding and coding regions of the *Drosophila* genome. Transition bias operating in conjunction with changes in base composition indicate that a nonstationary, nonhomogeneous model is necessary to adequately describe the subtleties of conserved noncoding block point substitution in *Drosophila* (Galtier and Gouy 1998).

In contrast to point substitution, rates of indel substitution are at least 20-fold less frequent in conserved noncoding blocks and appear to show no lineage effect. This decrease in the rate of indel substitution in noncoding blocks may reflect differential mutation rates or more severe selective constraints on indel substitution relative to point substitution. Order-of-magnitude differences in the rates of point and indel substitution have been observed previously in comparative analyses of mammalian noncoding DNA (Saitou and Ueda 1994). Like point substitution, however, the rate and pattern of indel substitution is similar for intergenic and intronic sequences. The pooled length distribution of indels in conserved noncoding blocks is skewed toward short sequences (Fig. 3A), as has been noted for *de novo* indels in inactive retro-element sequences in both *D. melanogaster* and *D. virilis* and in analyses of polymorphism and divergence in the *D. melanogaster* species group (Petrov and Hartl 1998; Comeron and Kreitman 2000). This skew is sufficient to produce a negative correlation between frequency and length for indel substitutions in *Drosophila* conserved noncoding blocks (Fig. 3B). This result adds to a variety of different data sets that suggest that the  $\zeta$  distribution can describe indel substitution, and that a logarithmic gap penalty is appropriate for the alignment of neighboring conserved noncoding blocks in *Drosophila* (Gu and Li 1995).

A major conclusion of our findings is that most features of noncoding DNA conservation are indistinguishable between intergenic and intronic regions. This is true for the average density and length distribution of conserved blocks, the rate and pattern of point substitution, as well as for the rate and pattern of indel substitution. We suggest that the similar properties of intergenic and intronic conserved blocks reflect similar mechanistic constraints operating on these sequences, and that transcription per se does not substantially influence major features of noncoding sequence evolution in *Drosophila*. This finding has an important implication that can substantially reduce the complexity of large-scale comparative sequence analyses in *Drosophila*. Namely, our results would indicate that a single model for the identification of

conserved noncoding DNA is sufficient for both intergenic and intronic compartments of the *Drosophila* genome.

### Methodological Considerations for the Interpretation of Noncoding Conservation

Because there is no reading frame to constrain alignments, results based on pairwise sequence comparisons of noncoding DNA are critically dependent on alignment methods and parameters. We have attempted to be relatively conservative in our criteria for including sequences in the conserved block component of our data set, to ensure that the majority of substituted sites analyzed are contained within sequences under purifying selection. In our opinion, it is first necessary to understand the pattern and relative rates of substitution in conserved noncoding blocks to statistically identify the boundaries of conserved blocks for subsequent analyses (Karlin and Altschul 1990; Zhu et al. 1998). We suspect, however, that there are sequences that have functional constraint that are not conserved at the level >70% nucleotide identity, especially nucleotides flanking the edges of conserved blocks. For this reason, the fraction of constrained sequences in noncoding regions with known or suspected *cis*-regulatory function based on our analysis is likely underestimated. Conversely, it is clear that some noncoding regions exhibit little if any primary sequence constraint, and therefore genomic averages of noncoding constraint may be lower than what we report for functional regions. Moreover, our block definition tends to bias the length distribution of conserved blocks toward lower values and increase the number of independent blocks relative to true values. Using stringent block criteria also leads to underestimating the total rates, but not the relative rates, of point and indel substitutions relative to true values. In the absence of a good substitution model to assess the significance of conserved blocks, problems such as these can only be ameliorated empirically by multiple-species sequence comparisons.

To evaluate potential biases in our alignment methods, we compared our results with those derived from four independent automated genomic alignment tools. Such a comparison is helpful for substantiating the results of our filtered dotplot approach, as well as calibrating automated tools for large-scale noncoding sequence analyses in the future. From the combined output of DiAlign, DBA, VISTA, and Lamark, >98% of blocks identified by our method can be automatically identified, although only 60% are identified by all four methods. These results indicate that on the order of only 2% or less of blocks in our data set have no evidence of being conserved using an automated method, even though they contain matches that meet our criteria. We analyzed the 60% of blocks that were compatible across all methods (the compatible set) for properties of noncoding conservation reported above. As expected, the estimated fraction of DNA conserved in noncoding regions is lower in the compatible set (*D. melanogaster*: 19.4%; *D. virilis*: 16.0%). Also, the length distribution of conserved blocks differs between the total data set and compatible set (Kolmogorov-Smirnov test,  $P < 0.001$ ). Fewer short blocks are included in the compatible set, as reflected in an increase in the location of the length distribution of the compatible set — the mean and median block lengths are 29.8 bp and 25 bp, respectively. Importantly, however, our conclusions about the rate and pattern of point substitution are not dependent on alignment method. In the compatible set, 7.0% of conserved block nucleotide positions are substituted,

compared with 7.2% of sites in the total data set. Moreover, the overall structure of the match–mismatch matrix does not differ between the total and compatible data sets ( $\chi^2=7.85$ , 15 d.f.,  $P<0.93$ ), nor does the distribution of indel sizes (Kolmogorov–Smirnov test,  $P<0.10$ ). Despite similarity in the pattern of indel substitution in the total and compatible data sets, however, the estimated rate of indel substitution in the compatible data set (0.15%) is twofold lower than the total data set (0.32%). Therefore, differences in alignment procedures may affect inferences about the fraction of conservation and length distribution of conserved blocks, however, inferences concerning substitution properties are not substantially affected.

The results of our compatibility analysis also indicate that the majority of discordant blocks are missed uniquely by only one of the three methods. For instance, DBA systematically neglects many of the shortest blocks in our data set, which represent the mass of the conserved block distribution in our analysis (Fig. 1). DBA also has the tendency to insert bases and gaps in the output local alignments that are not present in the input sequences; this is likely a consequence of finite symbol emission probabilities of the HMM architecture underlying the alignment algorithm. VISTA, on the other hand, tends to omit small blocks that flank longer, strongly conserved blocks, or omit small blocks that lie between two larger blocks that have strongly conserved spacing. These effects are likely attributable to the global alignment nature of the algorithm that might compromise small local alignments at the expense of aligning larger regions. Because LAMARK was designed for a hierarchical search strategy, a single parameter search such as ours leads to many local alignments off the main diagonal, which necessitated imposing colinearity on the output to filter real from additional alignments. DiAlign identified the highest proportion of blocks in our data set with the minimum number of spurious alignments using a single set of parameters. DiAlign, however, occasionally excludes flanking nucleotides from “regions of similarity” (i.e., conserved blocks) that are clearly aligned in the output. Because each method has characteristic difficulties, we conclude that the use of multiple noncoding alignment tools is currently advisable to identify conserved sequences in noncoding regions.

Finally, our discussion of noncoding constraints must consider the methodological limitations imposed by pairwise sequence analysis. In addition to making the definition of block edges problematic, pairwise data limits our understanding of the constraints on noncoding sequences in a number of important ways. As noted previously, pairwise data cannot distinguish the polarity of evolutionary changes, and therefore the relative rates of reciprocal substitutions, or insertions versus deletions, cannot be estimated individually. Moreover, pairwise data does not allow the observation of multiple substitutions at the same nucleotide position, for which we have made no corrections in our analyses. Multiple substitution in conserved noncoding blocks may not be a serious concern, however, since variant sites in the *Drosophila even-skipped* stripe two enhancer are generally substituted only once on the phylogeny (Ludwig et al. 1998). Not detecting multiple hits, however, reveals a more general limitation of pairwise sequence analysis — the inability to assess heterogeneity in the rate and pattern of substitution across sites. It is well established in coding sequences that both the rate and pattern of substitution vary across sites and lineages (Yang 1996a,b). Furthermore, variation in the rate or pattern of substitution

can influence estimation of other evolutionary parameters such as transition bias (Wakeley 1994; Huelsenbeck and Nielsen 1999). Therefore, proper estimation of a substitution model for conserved noncoding blocks will require multiple species sequence comparisons to address these limitations of pairwise data.

### Future Prospects for the Comparative Analysis of Noncoding DNA

Our results provide insight into the mode of molecular evolution for a subset of highly conserved noncoding sequences. Future analyses of multiple species comparisons will be necessary to evaluate the generality of our results and construct more realistic substitution models for highly conserved noncoding DNA. Multiple species comparisons within *Drosophila* are especially important as our results suggest that the pattern of substitution in conserved noncoding DNA may fluctuate across lineages. For the same reason, it is also worth investigating conserved noncoding block substitution models in other complex eukaryotic (e.g., mammalian, rhabdoid) lineages. Multiple species comparisons may also allow for null models of sequence evolution to be applied to data in hopes of potentially identifying Darwinian selection operating on noncoding sequences. Finally, comparing sequences from species more closely related than *D. melanogaster* and *D. virilis* will allow for a more thorough understanding of the evolutionary dynamics of weakly constrained noncoding sequences, as their rate of evolution should be higher than for sequences studied here. Modeling the molecular evolution of noncoding sequences in general will require much additional research, as our results and others show that the majority of noncoding nucleotides are not under strong primary sequence constraint.

Just as more widespread taxonomic sampling aids molecular evolutionary modeling, whole genome comparative analyses should offer a wealth of information relevant to modeling the individual units of noncoding structure and function. For instance, a genomic database of conserved noncoding blocks will be particularly useful for modeling structural properties of individual DNA–protein interactions like transcription factor-binding site specificity. With such a resource, models of binding-site specificity can be inferred from similarities intrinsic to a database of conserved noncoding blocks and be confirmed, rather than developed, experimentally. As proof of this principle, Wasserman et al. (2000) were able to reconstruct the usage matrices for three myogenic transcription factors using conserved blocks from a sample of skeletal muscle regulatory regions. It is generally appreciated that noncoding conservation can be used to locate regulatory sequences, and that conservation can be used in combination with binding-site prediction to identify potential upstream regulators of these sequences (Duret and Bucher 1997; Fickett and Wasserman 2000). Using binding-site prediction in conjunction with conservation in this manner is “top-down” (sensu Bucher 1999) and requires detailed a priori knowledge about which sequences a particular factor binds, a step that precludes efficient whole genome analysis. A “bottom-up” approach like clustering conserved blocks should rapidly provide many models of transcription factor specificity that can be used to make functional predictions.

Finally, future comparative genomic analyses will also help our understanding of the higher order structural organization present in noncoding regions of eukaryotic genomes.

Advances in this direction will require linking the pattern of noncoding primary sequence conservation to higher order functional units through specific structural models. For example, functional analysis of enhancer structure points to the importance of hierarchical spatial constraints operating between sequence-specific elements (Ondek et al. 1988). Under such a framework, modeling the spatial organization of conserved noncoding blocks will potentially aid the functional annotation of enhancer sequences. Other functional noncoding sequences will certainly benefit from the reciprocal development of higher-order models of structure and molecular evolution as well. In our view, a hallmark of success for models of noncoding molecular evolution will be their ability to assist functional predictions in comparative genomic data. Free from the constraints of the genetic code, the analysis of noncoding DNA present a unique opportunity to develop and test models of molecular evolution that interface with those that predict structure and function.

## METHODS

### Data Collection

A clone of the *D. virilis hairy* 5' region contained in a *P*-element vector was obtained from J. Langeland (Langeland and Carroll 1993). The insert was digested using *NotI* and *Asp718*, and shotgun sequenced as described previously (Andolfatto et al. 1999). This fragment has been submitted to GenBank under the accession number AF329639. A PCR product derived from *D. virilis* genomic DNA (Pasadena, CA strain 1052) was amplified using the Expand system (Roche Molecular) using primers designed from GenBank accession M87885 and the homologous region to GenBank accession S78746 (Langeland and Carroll 1993). The PCR primers for this reaction are *vir\_h\_2322U24*: 5'-CCATCTCGCGAGCGTGTC CAAAGC-3' and *vir\_h\_6547L24*: 5'-GTATTGGGCACCGCT GTCGTCTCC-3'. The reaction used 1.5  $\mu$ L of genomic DNA (protocol 48, Ashburner 1989) and the cycling conditions on an MJ Research PTC-200 for this reaction were: initial denaturation at 92°C for 2 min, 10 rounds of denaturation at 92°C for 10 sec, annealing at 65°C for 30 sec, and extension at 68°C for 3 min 45 sec, followed by 19 rounds of the same conditions adding 20 sec per round to the extension time, terminated by a 68°C incubation for 7 min. This long-distance PCR product has been submitted to GenBank under the accession number AF329640. The three *D. virilis hairy* 5' fragments were joined into one contig for the final comparative analysis with *D. melanogaster*.

We attempted to generalize the pattern of sequence conservation derived from preliminary analysis of the *hairy* region by searching PubMed and GenBank for entries that contained *D. virilis* homologs of sequences with known or suspected *cis*-regulatory function in *D. melanogaster*. Where possible, sequences were downloaded from GenBank; additional sequences were obtained by personal communication or transcribed from figures in the primary reference (Table 1). *D. melanogaster* sequences were obtained and oriented from the BDGP database via preliminary BLAST analysis with the *D. virilis* homolog. Sequences were edited so that the beginning and end of each region would correspond to conserved blocks.

Our interest lies in the molecular evolutionary analysis of *cis*-regulatory sequences involved in transcriptional regulation, so we focused whenever possible on 5' and 3' nontranscribed, noncoding sequences with experimentally verified *cis*-regulatory transcriptional function. Transcriptional regulatory elements are found in the introns of genes involved in many developmentally regulated genes in *Drosophila* (e.g., *Ubx*, *eyeless*, *B-tubulin*), so we also chose to include introns

with known or suspected regulatory function and long introns (>1 kb) in the dataset. Coding and noncoding exons were excluded from our analysis to the limits of resolution of the annotated transcript structure in GenBank or the primary reference. In general, a given pair of sequences is terminated by either the transcription initiation site for 5' intergenic sequences, or the first block downstream of reported polyadenylation site for 3' intergenic sequences. For intronic sequences, the first and last blocks are contained entirely within the intron.

### Data Analysis

We performed a manual analysis of homologous pairs of sequences using the Filtered DotPlot implementation in the MegAlign program (DNASTar) (Maizel and Lenk 1981). This type of pairwise sequence analysis affords an interactive and exhaustive search for conservation that can be directly visualized. The parameters used in the initial search were percent match: 70%; minimum window: 1; window size: 10. We filtered top-scoring segments using a locus-to-locus heuristic threshold based on the shape of the tail of the distribution of segment scores. We then chose a colinear path of blocks among the top-scoring segments to generate a set of local alignments spanning the entire region of homology (Supplemental Table 1, available on-line at <http://www.genome.org>). To ensure that the substitutions observed in our data are truly within conserved sequences, we trimmed the blocks in our data set so that at least three nucleotides of identity flanked each block to avoid spurious alignment of nucleotides around the core of conservation. In general, off-main-diagonal high-scoring segments were omitted because they were caused by simple sequence repeats in which one species had a higher-scoring match elsewhere closer to the main diagonal. For all loci the counter-diagonal was also analyzed, but high-scoring segments in the opposite orientation were also generally restricted to simple repeats. We justify the use of this method based on previous estimates of what would be statistically significant alignment blocks between *D. melanogaster* and *D. virilis*, in conjunction with the established pattern of colinearity of conservation in *cis*-regulatory sequences (Hartl and Lozovskaya 1994; Jareborg et al. 1999).

During the course of this study, we were provided with an automated alignment tool called *Lamarck* based on a dot-plot algorithm developed by S. Shabalina and A. Kondarashov (NCBI, pers. comm.), which we used to evaluate our alignment procedure. The parameters of the *Lamarck* search were six matches in a window size of seven, with each significant block requiring a segment score of 10 (i.e., 10 contiguous windows of six/seven matches offset by one base). We imposed colinearity of the blocks from the output of *Lamarck* and compared the results of the automatic and manual analyses. Additionally, we employed default parameters of the DNA Block Aligner (DBA), a finite state/hidden Markov algorithm available in the Wise package, to evaluate our choice of blocks (Jareborg et al. 1999). Both of these platforms generate a set of local alignments rather than a true global alignment. We also used default parameters of the DiAlign v2.1 (T=0, with regions of maximum similarity denoted by five "\*" alignment method (Morgenstern 2000). Finally, we submitted our dataset to the global VISTA genomic alignment tool using a window size of 10 and identity of 70% (Dubchak et al. 2000). We attempted to choose parameters for *Lamarck* and VISTA that were comparable with those used in the filtered dotplot analysis.

Conserved blocks in both *D. melanogaster* and *D. virilis* sequences were parsed using helper applications written in the C programming language. For each conserved block, identical and variant nucleotides were counted relative to the plus strand of the local transcription unit in *D. melanogaster*. Insertions causing conserved blocks that are contiguous in one

species to be separated in the other species were treated as insertion/deletion (indel) events. G-tests were used to evaluate the difference in the percent of conserved sequences among transcription classes caused by changes in genome size. Nonparametric tests were used to evaluate differences in length distributions of blocks and indel substitutions by transcriptional class and species as the data are not distributed normally. Goodness of fit to various expectations was evaluated using  $\chi^2$  tests. Tests were considered significant if the statistic had probability less than (0.05/number of tests) to correct for multiple testing.

## ACKNOWLEDGMENTS

We thank Jim Langeland for providing the *D. virilis hairy* clone; Misha Ludwig for providing *D. virilis* flies; Etsuko Moriyama for communicating a list of introns in *D. virilis*; and the authors listed in Table 1 who supplied sequence data. C.M.B. particularly thanks Josep Comeron for many insightful discussions and advice on statistical analyses, as well as Mark Biggin, Steve Dorus, Carrie Grimsley, and members of the laboratory of Harinder Singh for helpful comments on an earlier version of this manuscript. C.M.B. is supported by a National Science Foundation (NSF) pre-doctoral fellowship and an NSF G.A.A.N. training grant in Evolutionary Genomics. This work was supported by a grant from the University of Chicago Hinds Fund to C.M.B. and NSF grant MCB-9604477 to M.K. and Michael Ludwig.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

## REFERENCES

- Adams, M.D., Celniker, S.E., Holt, R.A., Evans, C.A., Gocayne, J.D., Amanatides, P.G., Scherer, S.E., Li, P.W., Hoskins, R.A., Galle, R.F., et al. 2000. The genome sequence of *Drosophila melanogaster*. *Science* **287**: 2185–2195.
- Andolfatto, P., Wall, J.D., and Kreitman, M. 1999. Unusual haplotype structure at the proximal breakpoint of In(2L)t in a natural population of *Drosophila melanogaster*. *Genetics* **153**: 1297–1311.
- Arnone, M.I. and Davidson, E.H. 1997. The hardwiring of development: Organization and function of genomic regulatory systems. *Development* **124**: 1851–1864.
- Ashburner, M. 1989. *Drosophila: A laboratory manual*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.
- Batzoglou, S., Pachter, L., Mesirov, J.P., Berger, B., and Lander, E.S. 2000. Human and mouse gene structure: Comparative analysis and application to exon prediction. *Genome Res.* **10**: 950–958.
- Blackman, R.K. and Meselson, M. 1986. Interspecific nucleotide sequence comparisons used to identify regulatory and structural features of the *Drosophila hsp82* gene. *J. Mol. Biol.* **188**: 499–515.
- Bucher, P. 1999. Regulatory elements and expression profiles. *Curr. Opin. Struct. Biol.* **9**: 400–407.
- Carroll, S.B., Grenier, J.K., and Weatherbee, S.D. 2001. *From DNA to diversity: Molecular genetics and the evolution of animal design*. Blackwell Science, Inc., Oxford, UK.
- Comeron, J.M. and Kreitman, M. 2000. The correlation between intron length and recombination in *Drosophila*: Dynamic equilibrium between mutational and selective forces. *Genetics* **156**: 1175–1190.
- Cripps, R.M., Black, B.L., Zhao, B., Lien, C.L., Schulz, R.A., and Olson, E.N. 1998. The myogenic regulatory gene Mef2 is a direct target for transcriptional activation by Twist during *Drosophila* myogenesis. *Genes & Dev.* **12**: 422–434.
- Deutsch, M. and Long, M. 1999. Intron-exon structures of eukaryotic model organisms. *Nucleic Acids Res.* **27**: 3219–3228.
- Dickinson, W.J. 1991. The evolution of regulatory genes and patterns in *Drosophila*. In *Evolutionary biology* (ed. M.K. Hecht, B.W., R.J. MacIntyre), pp. 127–173. Plenum Publishing Corp., New York.
- Dubchak, I., Brudno, M., Loots, G.G., Pachter, L., Mayor, C., Rubin, E.M., and Frazer, K.A. 2000. Active conservation of noncoding sequences revealed by three-way species comparisons. *Genome Res.* **10**: 1304–1306.
- Duret, L. and Bucher, P. 1997. Searching for regulatory elements in human noncoding sequences. *Curr. Opin. Struct. Biol.* **7**: 399–406.
- Fickett, J.W. and Wasserman, W.W. 2000. Discovery and modeling of transcriptional regulatory regions. *Curr. Opin. Biotechnol.* **11**: 19–24.
- Gajewski, K., Kim, Y., Lee, Y.M., Olson, E.N., and Schulz, R.A. 1997. D-mef2 is a target for Tinman activation during *Drosophila* heart development. *EMBO J.* **16**: 515–522.
- Galtier, N. and Gouy, M. 1998. Inferring pattern and process: Maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis. *Mol. Biol. Evol.* **15**: 871–879.
- Gillespie, J.H. 1991. *The causes of molecular evolution*. Oxford University Press, New York.
- Gu, X. and Li, W.H. 1995. The size distribution of insertions and deletions in human and rodent pseudogenes suggests the logarithmic gap penalty for sequence alignment. *J. Mol. Evol.* **40**: 464–473.
- Hartl, D.L. and Lozovskaya, E.R. 1994. Genome evolution: Between the nucleosome and the chromosome. *EXS* **69**: 579–592.
- Horn, C. and Wimmer, E.A. 2000. A versatile vector set for animal transgenesis. *Dev. Genes Evol.* **210**: 630–637.
- Hoskins, R.A., Nelson, C.R., Berman, B.P., Laverty, T.R., George, R.A., Ciesiolka, L., Naeemuddin, M., Arenson, A.D., Durbin, J., David, R.G., et al. 2000. A BAC-based physical map of the major autosomes of *Drosophila melanogaster*. *Science* **287**: 2271–2274.
- Huelsenbeck, J.P. and Nielsen, R. 1999. Variation in the pattern of nucleotide substitution across sites. *J. Mol. Evol.* **49**: 86–93.
- Jareborg, N., Birney, E., and Durbin, R. 1999. Comparative analysis of noncoding regions of 77 orthologous mouse and human gene pairs. *Genome Res.* **9**: 815–824.
- Johnson, N.L. and Kotz, S. 1969. *Distributions in statistics: Discrete distributions*. Houghton Mifflin Company, Boston, MA.
- Johnson, W.A., McCormick, C.A., Bray, S.J., and Hirsh, J. 1989. A neuron-specific enhancer of the *Drosophila* dopa decarboxylase gene. *Genes & Dev.* **3**: 676–686.
- Karlin, S. and Altschul, S.F. 1990. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl. Acad. Sci.* **87**: 2264–2268.
- Kassis, J.A., Poole, S.J., Wright, D.K., and O'Farrell, P.H. 1986. Sequence conservation in the protein coding and intron regions of the engrailed transcription unit. *EMBO J.* **5**: 3583–3589.
- Kassis, J.A., Desplan, C., Wright, D.K., and O'Farrell, P.H. 1989. Evolutionary 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.
- Kimura, M. 1983. *The neutral theory of molecular evolution*. Cambridge University Press, Cambridge, UK.
- Kwiatowski, J., Skarecky, D., Bailey, K., and Ayala, F.J. 1994. Phylogeny of *Drosophila* and related genera inferred from the nucleotide sequence of the *Cu,Zn Sod* gene. *J. Mol. Evol.* **38**: 443–454.
- Langeland, J.A. and Carroll, S.B. 1993. Conservation of regulatory elements controlling hairy pair-rule stripe formation. *Development* **117**: 585–596.
- Lawrence, P. 1992. *The making of a fly: The genetics of animal design*. Blackwell Scientific Publications, Oxford, UK.
- Lee, Y.M., Park, T., Schulz, R.A., and Kim, Y. 1997. Twist-mediated activation of the NK-4 homeobox gene in the visceral mesoderm of *Drosophila* requires two distinct clusters of E-box regulatory elements. *J. Biol. Chem.* **272**: 17531–17541.
- Loots, G.G., Locksley, R.M., Blankespoor, C.M., Wang, Z.E., Miller, W., Rubin, E.M., and Frazer, K.A. 2000. Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. *Science* **288**: 136–140.
- Ludwig, M., Patel, N., and Kreitman, M. 1998. Functional analysis of eve stripe 2 enhancer evolution in *Drosophila*: Rules governing conservation and change. *Development* **125**: 949–958.
- Ludwig, M.Z., Bergman, C., Patel, N., and Kreitman, M. 2000. Evidence for stabilizing selection in a eukaryotic cis-regulatory element. *Nature* **403**: 564–567.
- Maizel, J.V. and Lenk, R.P. 1981. Enhanced graphic matrix analysis of nucleic acid and protein sequences. *Proc. Natl. Acad. Sci.* **78**: 7665–7669.
- McCormick, A., Core, N., Kerridge, S., and Scott, M.P. 1995. Homeotic response elements are tightly linked to tissue-specific elements in a transcriptional enhancer of the *teashirt* gene. *Development* **121**: 2799–2812.

- Morgenstern, B. 2000. A space-efficient algorithm for aligning large genomic sequences. *Bioinformatics* **16**: 948–949.
- Moriyama, E.N. and Hartl, D.L. 1993. Codon usage bias and base composition of nuclear genes in *Drosophila*. *Genetics* **134**: 847–858.
- Moriyama, E.N. and Powell, J.R. 1996. Intraspecific nuclear DNA variation in *Drosophila*. *Mol. Biol. Evol.* **13**: 261–277.
- . 1997. Synonymous substitution rates in *Drosophila*: Mitochondrial versus nuclear genes. *J. Mol. Evol.* **45**: 378–391.
- Ondek, B., Gloss, L., and Herr, W. 1988. The SV40 enhancer contains two distinct levels of organization. *Nature* **333**: 40–45.
- Pan, D., Valentine, S.A., and Courey, A.J. 1994. The bipartite *D. melanogaster* twist promoter is reorganized in *D. virilis*. *Mech. Dev.* **46**: 41–53.
- Petrov, D.A. and Hartl, D.L. 1998. High rate of DNA loss in the *Drosophila melanogaster* and *Drosophila virilis* species groups. *Mol. Biol. Evol.* **15**: 293–302.
- . 1999. Patterns of nucleotide substitution in *Drosophila* and mammalian genomes. *Proc. Natl. Acad. Sci.* **96**: 1475–1479.
- Powell, J.R. 1997. Progress and prospects in evolutionary biology: The *Drosophila* model. Oxford University Press, Oxford, UK.
- Rodríguez-Trelles, F., Tarrío, R., and Ayala, F.J. 2000. Evidence for a high ancestral GC content in *Drosophila*. *Mol. Biol. Evol.* **17**: 1710–1717.
- Rong, Y.S. and Golic, K.G. 2000. Gene targeting by homologous recombination in *Drosophila*. *Science* **288**: 2013–2018.
- Rorth, P., Szabo, K., Bailey, A., Laverty, T., Rehm, J., Rubin, G.M., Weigmann, K., Milan, M., Benes, V., Ansorge, W., et al. 1998. Systematic gain-of-function genetics in *Drosophila*. *Development* **125**: 1049–1057.
- Russo, C.A., Takezaki, N., and Nei, M. 1995. Molecular phylogeny and divergence times of drosophilid species. *Mol. Biol. Evol.* **12**: 391–404.
- Saitou, N. and Ueda, S. 1994. Evolutionary rates of insertion and deletion in noncoding nucleotide sequences of primates. *Mol. Biol. Evol.* **11**: 504–512.
- Schier, A.F. and Gehring, W.J. 1993. Analysis of a fushi tarazu autoregulatory element: Multiple sequence elements contribute to enhancer activity. *EMBO J.* **12**: 1111–1119.
- Shabalina, S.A. and Kondrashov, A.S. 1999. Pattern of selective constraint in *C. elegans* and *C. briggsae* genomes. *Genet. Res.* **74**: 23–30.
- Stern, D.L. 2000. Perspective: Evolutionary developmental biology and the problem of variation. *Evolution* **54**: 1079–1091.
- Tautz, D. 2000. Evolution of transcriptional regulation. *Curr. Opin. Genet. Dev.* **10**: 575–579.
- Wakeley, J. 1994. Substitution-rate variation among sites and the estimation of transition bias. *Mol. Biol. Evol.* **11**: 436–442.
- Wasserman, W.W., Palumbo, M., Thompson, W., Fickett, J.W., and Lawrence, C.E. 2000. Human-mouse genome comparisons to locate regulatory sites. *Nat. Genet.* **26**: 225–228.
- Williams, J.A., Paddock, S.W., Vorwerk, K., and Carroll, S.B. 1994. Organization of wing formation and induction of a wing-patterning gene at the dorsal/ventral compartment boundary. *Nature* **368**: 299–305.
- Wolff, C., Pepling, M., Gergen, P., and Klingler, M. 1999. Structure and evolution of a pair-rule interaction element: runt regulatory sequences in *D. melanogaster* and *D. virilis*. *Mech. Dev.* **80**: 87–99.
- Xu, X., Yin, Z., Hudson, J.B., Ferguson, E.L., and Frasch, M. 1998. Smad proteins act in combination with synergistic and antagonistic regulators to target Dpp responses to the *Drosophila* mesoderm. *Genes & Dev.* **12**: 2354–2370.
- Yang, Z. 1996a. Among-site rate variation and its impact on phylogenetic analyses. *Trends Ecol. & Evol.* **11**: 367–372.
- . 1996b. Maximum-likelihood models for combined analyses of multiple sequence data. *J. Mol. Evol.* **42**: 587–596.
- Yin, Z., Xu, X.L., and Frasch, M. 1997. Regulation of the twist target gene tinman by modular cis-regulatory elements during early mesoderm development. *Development* **124**: 4971–4982.
- Zhu, J., Liu, J.S., and Lawrence, C.E. 1998. Bayesian adaptive sequence alignment algorithms. *Bioinformatics* **14**: 25–39.

Received January 5, 2001; accepted in revised form May 14, 2001.