Available online at www.sciencedirect.com
ScienceDirect
Genomics 88 (2006) 431–442
GENOMICS
www.elsevier.com/locate/ygeno

Conservation patterns in different functional sequence categories of divergent *Drosophila* species

Dmitri Papatsenko [a,*], Andrey Kislyuk [b], Michael Levine [a], Inna Dubchak [b]

[a] *Department of Molecular and Cell Biology, University of California at Berkeley, Berkeley, CA 94720, USA*
[b] *Genomics Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA*

Received 3 February 2006; accepted 21 March 2006
Available online 11 May 2006

Abstract

We have explored the distributions of fully conserved ungapped blocks in genome-wide pair-wise alignments of recently completed species of *Drosophila: D. melanogaster, D. yakuba, D. ananassae, D. pseudoobscura, D. virilis*, and *D. mojavensis*. Based on these distributions we have found that nearly every functional sequence category possesses its own distinctive conservation pattern, sometimes independent of the overall sequence conservation level. In the coding and regulatory regions, the ungapped blocks were longer than in introns, UTRs, and nonfunctional sequences. At the same time, the blocks in the coding regions carried a $3N + 2$ signature characteristic of synonymous substitutions in the third-codon position. Larger block sizes in transcription regulatory regions can be explained by the presence of conserved arrays of binding sites for transcription factors. We also have shown that the longest ungapped blocks, or "ultraconserved" sequences, are associated with specific gene groups, including those encoding ion channels and components of the cytoskeleton. We discuss how restraining conservation patterns may help in mapping functional sequence categories and improve genome annotation.

Keywords: *Drosophila*; Conservation pattern; Vista alignment; Developmental enhancer

There has been a recent explosion in the number of completed animal genomes and a broad sampling of genome alignments is now available for most of the model organisms. Interpretation of genome alignments is a high priority goal, as it will help find new genes, gene control regions, and other functional sequences. Here we attempt to define the sequence conservation patterns in functionally different classes of genomic DNA, including protein coding genes and regulatory DNA sequences. We approach this problem with the help of statistical analysis of ungapped block sizes in genome-wide pair-wise alignments of *Drosophila*. The distribution of block sizes was originally explored by Bergman and coworkers using pair-wise alignments of several genomic intervals of two *Drosophila* species [1]. In the current work, we describe analysis of whole-genome alignments of six *Drosophila* species and compare block size statistics for five functional sequence categories. Details on evolutionary history, biology of the selected species, and impact can be found elsewhere [2,3] (see also http://rana.lbl.gov/drosophila/ for the project status).

Functional differences in the conservation patterns—such as the distribution of ungapped block sizes—are difficult to detect using standard methods, such as the number of matches in a fixed-width window. Most of the methods, based on local (PIPMaker for Blastz [4–6]) or global alignment algorithms (VISTA for AVID and LAGAN) [7–10], are very efficient in finding long stretches of conservation, including ultraconserved regions [11,12]. However, these methods are not focused, for instance, on efficient detection of transcription regulatory elements on a large scale or binding sites for individual regulatory proteins on a smaller scale [13,14]. Some programs, however, approach the problem of alignment interpretation in a more accurate way. For instance, the phastCons program computes conservation scores based on a phylo-HMM, a type of probabilistic model that describes both the process of DNA substitution at each site in a genome and the way this process changes from one site to the next [15,16]. While mathematical models based on nucleotide substitution matrices [1,16] help in detection of the conserved regions, the role of block size and its relation with sequence

* Corresponding author.
E-mail address: dxp@berkeley.edu (D. Papatsenko).

0888-7543/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2006.03.012

function remain relatively unexplored. The strategy of Siepel and coworkers [16] is careful identification of conserved regions and consequent exploration of functional annotations; we attempt to find differences (signatures) between functional sequence categories first. A similar strategy was explored, for instance, in the analysis of orthologous eukaryotic mRNAs [17].

Finding functional conservation signatures, such as characteristic block sizes, is especially important for mapping transcription regulatory regions. The comparative analysis of *Drosophila melanogaster* and *D. pseudoobscura* using conventional window-based features (% of identity) showed that known transcription regulatory regions are only slightly more conserved than the rest of the noncoding genome [18]. The authors of this study found that 50–70% of known binding sites are located in windows with high sequence identity scores, but these percentages are not greatly enriched over what is expected by chance. The study of Berman and coworkers [14], based on the same strategy (window identity scores), showed that *cis*-regulatory elements appear indistinguishable from flanking sequence as there is a high amount of noncoding sequence conservation throughout the analyzed gene loci. At the same time, Bergman and coworkers [1,19] have suggested a connection between the block size and the size of binding site/binding site clusters in regulatory regions of *Drosophila*. In a more recent study by Glazov and coworkers [20], the authors showed that the majority of 100% conserved ungapped blocks are found within intergenic spacers, but not in the coding regions. These results indicated the need for further systematic exploration of the block size phenomenon, especially in transcription regulatory regions.

Here we undertake the next step toward the interpretation of the alignment patterns based on block size and explore how these sizes are distributed among five different functional sequence categories: coding regions, untranslated regions (UTRs), transcription regulatory regions (promoters and enhancers), and unannotated regions in the genome of *D. melanogaster*. We also analyze the functional assignment of the longest ungapped blocks (ultraconserved) and conservation of some other functionally important sequences, such as microRNA [21].

In the case of a pair-wise alignment, the conservation patterns (or signatures) can be described explicitly through an ordered set S of gaps, mismatches, and ungapped conserved blocks with their corresponding lengths. One can see that two different block-gap sets, S_1 ad S_2, may produce the same local sum of matches or the same window identity scores. However, different sizes and arrangements of these blocks and gaps in either of these sets (S_1 and S_2) may be dependent on the biological function of that genomic region. Therefore, a comprehensive exploration of block-mismatch ordered sets S might improve alignment interpretation and lead to a straightforward evaluation of the sequence function.

Results

Current limits on functional interpretation of genome alignments

To demonstrate existing problems with functional alignment interpretation, we explored the conservation of a variety of functional regions from *Drosophila* using a conventional phylogenetic method based on window identity scores [7]. We focused on several of the most annotated developmental gene loci, containing a number of well-known transcription regulatory regions, and fly enhancers [22]. The gene loci were selected on the basis of annotation quality. We compared functional maps for the gene loci (enhancers and coding sequences) with conserved regions, calculated by VISTA. In Fig. 1, the top tracks show a comparison of VISTA plots, in which conserved regions (colored) were calculated with a 70% identity in a 100-bp window cutoff, and the map of annotated functional regions for the loci of two developmental genes—*even-ski pped* and *fushi-tarazu*. While the coding regions correlate with the conserved regions (peaks) well, the distributions of enhancer regions show a low degree of correlation (r = 0.3–0.4) with the conservational profiles. In many cases, the overall conservation level in the enhancer regions is not higher than the conservation level in flanking nonfunctional genomic intervals [22]. On the same dataset, we also explored

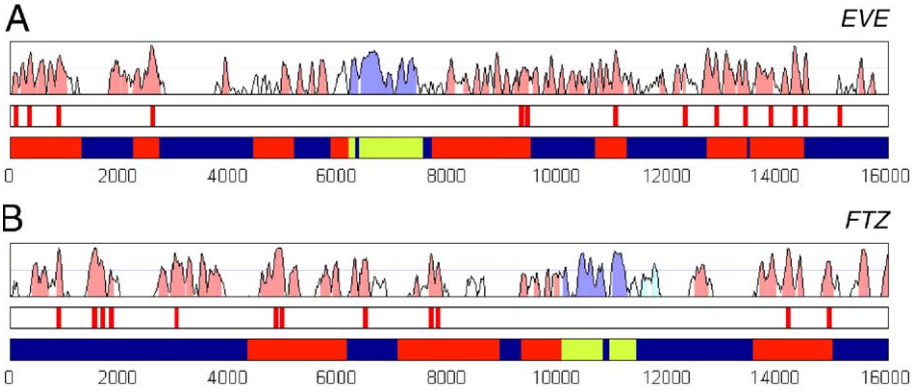

Fig. 1. Patterns of conservation in *eve* and *ftz* gene loci. Conservation profiles of (A) *even-skipped* and (B) *fushi-tarazu* gene loci. The top track in each shows the VISTA plot, the middle track shows the positions of ungapped conserved blocks longer than 40 bases, and the bottom track shows functional maps, in which regulatory regions are in red and exons are in yellow. Without additional treatment (interpretation), the conservation profiles (top) display low correlation with the functional maps. The middle tracks show that blocks longer than 40 are frequently found in enhancers, but not in coding regions.

the correlation between the distribution of ungapped conserved blocks and both the VISTA profile and the functional map (middle track in Fig. 1). Surprisingly, we found that exons do not contain 100% ungapped conserved blocks longer than 40 bases, but such blocks are present in enhancer regions. The distribution of the ungapped blocks is quite different from the VISTA score profile.

This analysis demonstrates that alignment interpretations based on standard window identity scores (such as VISTA) may be improved further. More information can be extracted from the alignments if block and gap lengths are given consideration along with the overall window identity score. For this reason, we decided to focus on statistics of ungapped block lengths and explore whether distribution of some block sizes is related to enhancers, exons, or some other functional sequence category.

Construction and evaluation of pair-wise alignments

To assess the power of the alignment interpretation based on statistics for ungapped blocks, we focused on the genome of *Drosophila*. Our choice of *Drosophila* was dictated by the very rich assortment of recently completed related fly genomes (available from LBNL Web resource: http://rana.lbl.gov/drosophila/) and the outstanding level of genome annotation for *D. melanogaster* [23].

We based our analysis on pair-wise genome alignments between *D. melanogaster* and the most recent genome assemblies of five different *Drosophila* species, *D. yakuba, D. ananassae, D. pseudoobscura, D. virilis*, and *D. mojavensis*. All alignments were obtained and analyzed using VISTA software with the Shuffle-LAGAN alignment module [9,24] (see also Materials and methods). Quality of the alignments was estimated using standard measures, such as coverage of the entire base genome and its functional features (annotated regions) [6].

Table 1 shows summary statistics for the genome-wide pair-wise alignments. These alignments cover different fractions of the *D. melanogaster* genome, depending on the evolutionary distance between compared species and the quality of genome assemblies. The achieved coverage of exons (85.2–97.8%) suggests that a majority of functional sequences are likely to be covered even in distant species, such as *D. mojavensis*. In addition to the standard "coverage" measures, we calculated the total lengths of the ungapped blocks (total number of matches) in the alignments. The ungapped blocks cover 30–70% of the base genome, depending on evolutionary distance.

Table 1
Quality of pair-wise alignments

	Drosophila yakuba	*Drosophila ananassae*	*Drosophila pseudoobscura*	*Drosophila virilis*	*Drosophila mojavensis*
Genome size (Mb)	171.9	167.1	135.8	196.6	189.8
Loose coverage					
Total	0.88	0.82	0.76	0.51	0.45
UTR	0.98	0.92	0.86	0.65	0.59
Exons	0.98	0.96	0.91	0.88	0.85
up100	0.97	0.85	0.79	0.52	0.46
up500	0.96	0.85	0.79	0.44	0.37
Tight coverage					
Total	0.85	0.32	0.22	0.15	0.14
UTR	0.96	0.29	0.15	0.06	0.05
Exons	0.97	0.80	0.70	0.62	0.60
up100	0.95	0.19	0.09	0.04	0.03
up500	0.91	0.15	0.07	0.03	0.03
Ungapped blocks	0.71	0.48	0.54	0.30	0.34

The coverage of genome annotation by pair-wise alignments used in this study is shown. Loose and tight coverage values were calculated according to a previously described method [6]. The bottom row shows the fraction of the base genome covered by ungapped 100% conserved blocks, i.e., the fraction of base pairs of the base genome exactly matching the second genome.

Definition of restrained patterns of conservation in pair-wise alignments

Along with window identity scores and nucleotide substitution matrices, conservation of a DNA sequence can be described by an ordered set of lengths for ungapped conserved blocks, mismatches, and gaps in a pair-wise alignment. The importance of this feature has been demonstrated in several related studies [1,20]. The block-gap sets S (see the introduction) can also be analyzed in multiple alignments; however, in that case there can be many types of gaps and/or ungapped blocks. In addition, construction of multiple alignments is more sensitive to the selected weighting method, so statistical interpretation of multiple alignments is less straightforward. Biological interpretation of multiple alignments is also more difficult due to the presence of repeated signals in functional regions and different ways of evolutionary sequence rearrangement in different species. Defining conserved regions and patterns in alignments of multiple species is a much more complex problem and is described in detail elsewhere [25,26].

Pair-wise alignments are more convenient for building a catalog/statistics for gap, mismatch, and block lengths for the described technical and biological reasons. There can potentially be only one type of ungapped fully conserved blocks and no more than two types of gaps between the blocks. Mismatches in the alignments (when both sequences are present) may be considered as type I gaps. Cases in which either sequence is absent from an alignment are different and may be considered type II gaps. It is unclear how much information can be obtained from the statistics of the lengths of type II gaps (unaligned regions) as they apparently correspond to nonfunctional sequences (insertions), which are not under evolutionary pressure and apparently may vary substantially in size. Similar considerations are applicable, to some extent, to type I gaps (mismatches). In general, the gaps of both types might simply reflect "allowed" distance ranges between some functional elements, residing in blocks. This model may be very simple, but it points out that the size of the ungapped block is more likely to be the functional indicator than the size of the region between two ungapped blocks. Usually, functional regions or sites expose a higher degree of conservation; therefore extended ungapped blocks (ultraconserved regions) may represent higher

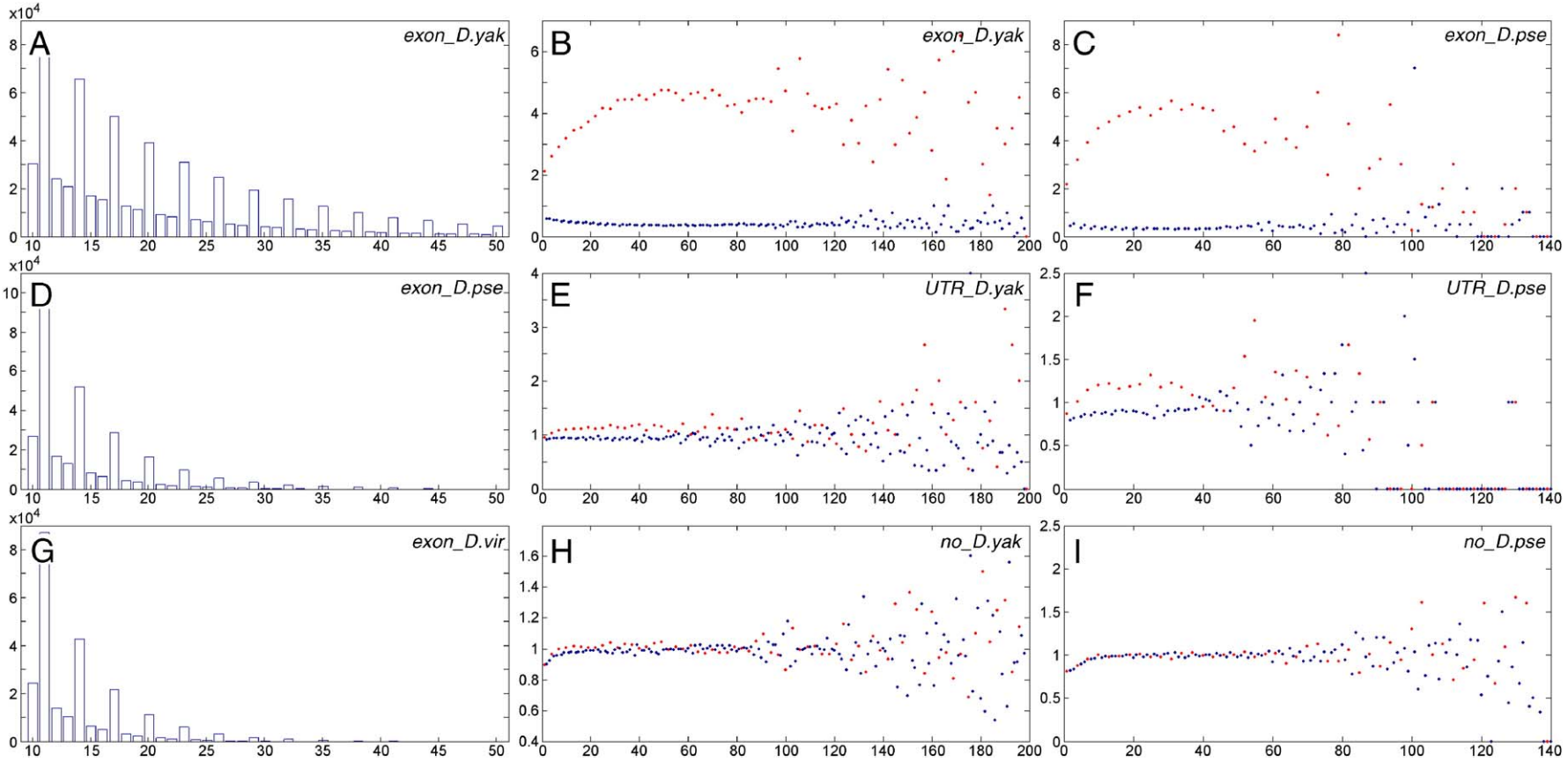

Fig. 2. The power of the $3N+2$ signal in exons and other sequences. Frequency histograms (A, D, G) show the presence of the $3N+2$ signal in exons. Results of filtering (see Eq. (2)) show that even very long ungapped blocks (>100 bases) in exons still fit to the $3N+2$ size (B, C, see data series in red). The signal is also present in untranslated regions (UTRs, E, F) and even in some sequences without any functional annotation (H), but to a much lower degree.

biological interest than very long gaps—the simple absence of alignments.

Here, we begin with defining alignment patterns through a size of ungapped 100% conserved blocks leaving the incorporation of type I and type II gaps as well as exploration of the multiple alignments among our prospective goals. Our systematic study was performed on a series of whole-genome pair-wise alignments recently obtained with the LAGAN global alignment algorithm. According to a detailed study by Pollard and colleagues [13], LAGAN yields rather accurate and specific alignments of functionally constrained coding and noncoding sequences in *Drosophila*. Along with other global alignment techniques, it has high sensitivity not only over functional maps (annotated functional features), but over entire populations of noncoding sequences as well.

Distribution of exon-specific block sizes across the genome of Drosophila

While peculiarities of sequence conservation in regulatory and other noncoding regions are quite obscure, the coding regions (CDS) represent an ideal model for exploring restrained alignment patterns or signatures. It is well known that the third position of amino acid codons can be subject to synonymous substitutions. In the example of human–mouse partial genome alignments, Dermitzakis and coworkers have shown that the direct consequence of synonymous substitutions is overrepresentation of ungapped blocks with the size $3N + 2$ in the coding regions [27].

To explore the distribution of $3N + 2$ blocks in genome-wide pair-wise alignments of *Drosophila* we generated frequency histograms for the ungapped block sizes for each considered pair-wise alignment between the *Drosophila* species. Fig. 2 shows that in all cases exons are highly enriched by the ungapped blocks with size $3N + 2$ (up to five or six times, see Figs. 2B and 2C). To provide a more sensitive method than frequency histograms, we performed signal filtering. We calculated the excess E of the $3N + 2$ fraction as the difference between the frequency F of the $3N + 2$ fraction and the expectation, approximated by the average frequency between the two neighboring bins:

$$E = F(3N + 2) - (F(3N + 1) + F(3N + 3))/2. \quad (1)$$

The signal filtering allowed the detection of some prevalence of the $3N + 2$ fraction in functional sequence categories other than the CDS category. We have found that this signal is still present in UTRs and in introns, but it is much weaker than in exons (up to 1.25 times enrichment of the $3N + 2$ fraction, see Figs. 2E and 2F). Some traces of the signal were even found in sequences without any functional annotation (see Fig. 2H), but the signal (see Fig. 2E) was relatively weak. No $3N + 2$ signal was detected in enhancer regions (data not shown). Overall, the prevalence of the $3N + 2$ fraction was distributed among functional categories as follows: exons > UTRs > introns > unknown. Possible reasons of this effect are given under Discussion.

The presence and distribution of the $3N + 2$ signal in *Drosophila* support previous findings by Dermitzakis and coworkers [27] obtained from human chromosome 21. Our signal filtering shows that even blocks in the range of 60–100 bases in exons (*D. melanogaster–D. pseudoobscura* alignments) carry the $3N + 2$ signature and the traces of the signal are present in untranslated regions and in some unannotated sequences as well (see Fig. 2). The test also shows that the restrained functional patterns are not lost in our most recent LAGAN/VISTA pair-wise alignments and that these signatures are specific to functional sequence categories.

Regulatory regions and UTRs possess their own signatures

To detect the possible presence of functional signatures in categories other than CDS functional sequence categories, we analyzed differences in the block frequency histograms built for seven functional sequence classes: enhancers, promoters, 5′ UTRs, exons, introns, 3′ UTRs, and "unknown" (sequences without annotation). The large enhancer and promoter datasets have not been previously subjected to this type of analysis. To suppress the effect of $3N + 2$ bias and possible small sample errors we considered wider block size ranges: 1–10, 11–20, 21–30, 31–40, 41–60, 61–80, 81–100, 100–265. The histograms are available in supplementary Table S1.

First, we estimated whether the histograms obtained for the enhancer regions are significantly different from the other datasets. We have found that block distributions in enhancer regions are strikingly different for most of the cases (see Table 2). While this standard statistical test showed an example of overall differences between sequence categories (frequency histograms), we were also interested in identifying fine differences/similarities between the functional classes in each block size range (bin). We compared a fraction of the blocks in each bin of each of the functional categories with the total

Table 2
Differences between enhancers and other sequence categories

	Prm	UTR	Exon	Intron	Unknown	All blocks
D.m.–D.yak.	7.72×10^{-44}	2.12×10^{-29}	4.52×10^{-115}	2.46×10^{-29}	7.09×10^{-8}	8.12×10^{-24}
D.m–D.ana.	2.77×10^{-47}	4.70×10^{-23}	3.29×10^{-179}	**2.22×10^{-1}**	3.15×10^{-2}	6.99×10^{-7}
D.m.–D.pse.	7.69×10^{-61}	2.21×10^{-35}	1.73×10^{-161}	1.96×10^{-3}	**4.57×10^{-1}**	5.12×10^{-8}
D.m–D.vir.	6.50×10^{-5}	8.97×10^{-19}	3.13×10^{-88}	5.15×10^{-2}	6.05×10^{-2}	2.90×10^{-6}
D.m–D.moj.	6.86×10^{-3}	3.02×10^{-14}	6.06×10^{-64}	6.48×10^{-2}	1.65×10^{-2}	2.51×10^{-4}

The p values obtained from χ^2 test are shown. Block frequency histograms for enhancers were compared with frequency histograms of all other sequence categories for blocks longer than 10 bases (see exact bin ranges under Results). While the distribution of block sizes in enhancers is close to that of introns and unannotated sequences (see numbers in bold), these three categories are still distinguishable, especially in *D. melanogaster–D. yakuba* alignments ($p = 7.09 \times 10^{-8}$).

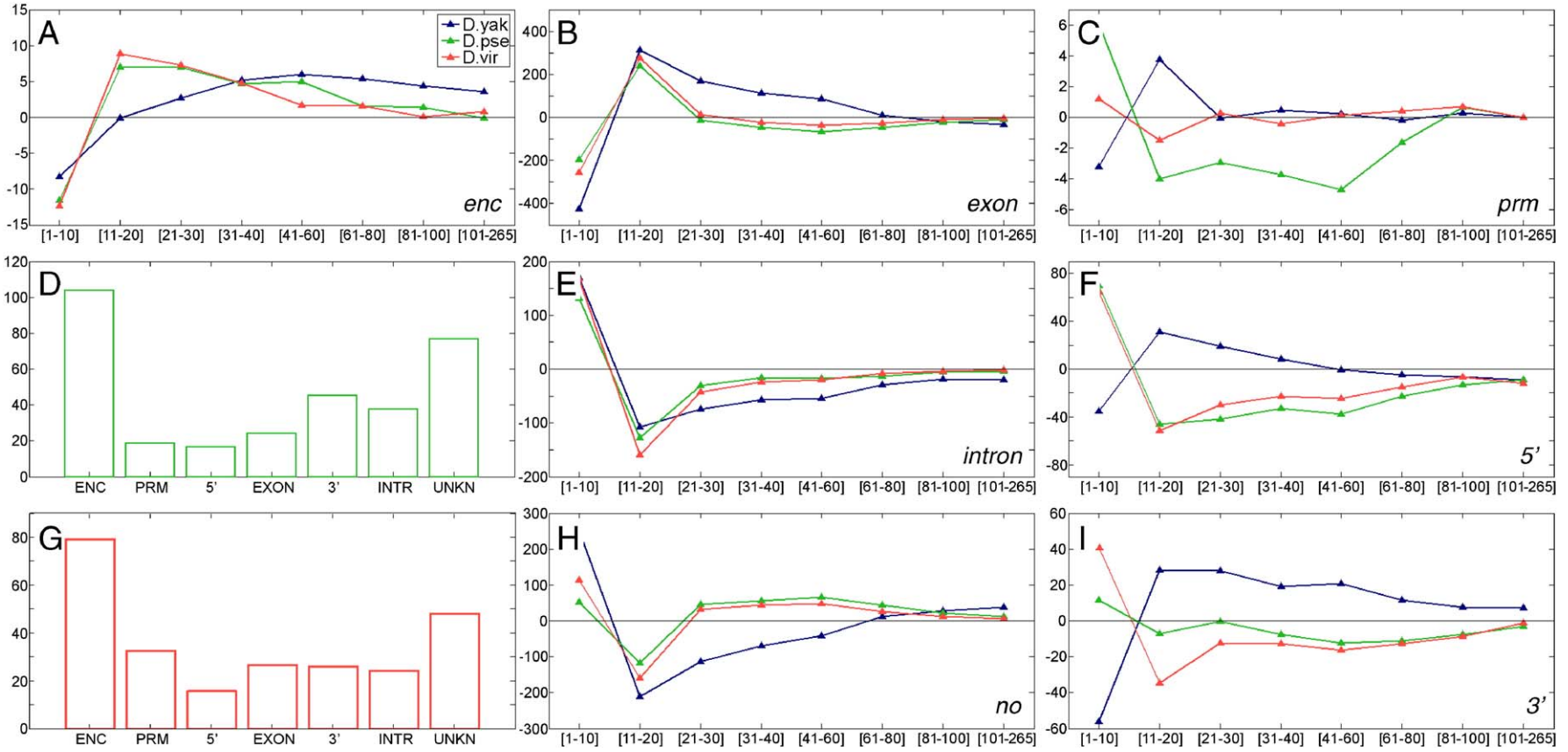

Fig. 3. Unequal distribution of block sizes among different sequence categories. (A–C, E, F, H, I) The *z* score profiles for fractional abundance of block in different block size ranges. (D, G) The relative amounts of ungapped blocks for all sequence categories in the range 31–40. Data series in blue correspond to *D. melanogaster–D. yakuba* alignments, data series in green are based on *D. melanogaster–D. pseudoobscura*, and those in red on *D. melanogaster–D. virilis* alignments. While shorter blocks (11–20) are more abundant in exons and enhancers (A, B), the enhancers also contain substantial fraction of longer blocks (>30 bases). In introns (E) and sequences without annotation (F) very small blocks (<11 bases) are more abundant. However, unannotated regions are also enriched by the longer blocks, suggesting the presence of unknown enhancers or other functional regions. In promoter (C) and untranslated regions (F, I) the longer blocks are not frequent or quickly disrupted in evolution.

fraction of blocks in the same bin obtained from the entire genome alignment (all categories). The z score was calculated for each bin as follows [28]:

$$z = \frac{p_1 - p_2}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \quad (2)$$

In this formula, p_1 is the fraction of blocks observed in a block size range (i.e., 1–10) for the analyzed sequence category, p_2 is the fraction of blocks observed in the same size range for all other sequence categories, n_1 is the total number of blocks for the analyzed category, n_2 is the total number of blocks for all other sequence categories. This statistic clearly shows that the distribution of block sizes is unequal among functionally different sequence classes (see Fig. 3).

We observed that both the enhancer regions and the exons contain a larger amount of 20–30 blocks, but the enhancers are also enriched in ungapped blocks longer than 20 bases, which are present in introns and unknown fractions (compare Figs. 3A and 3B). This effect is more striking in the case of the pair-wise alignment between *D. melanogaster* and *D. pseudoobscura*. Clearly, many blocks containing transcription regulatory signals survive "longer" in evolution than blocks in exons, which are broken due to synonymous substitutions in the third position of codons.

In some cases we detected up to 50–80% prevalence of ungapped blocks in enhancers in the range 21–30 (Figs. 3D and 3G) or one additional block (with respect to noise) of that length in nearly every enhancer (124 sequences in the enhancer dataset total). For longer blocks (>30 bases) we also detected some overrepresentation of the ungapped blocks in enhancers; however, in this case it was more difficult to judge due to the small size of the enhancer dataset. Nevertheless, in *D. melanogaster–D. pseudoobscura* alignments, ~40% of enhancers contain 100% conserved blocks longer than 35–40 bases and few contain very long blocks exceeding 60 or more bases. In the case of enhancers and exons, the z-score profiles across the size ranges are in agreement for all considered combinations of species (see lines of different color in Fig. 3).

In contrast to enhancers, the promoter regions (198 sequences) display no preference for the long ungapped blocks. Instead, these regions appear to be highly flexible in evolution as their block sizes are, in general, smaller than in other sequence categories (see Fig. 3F). Surprisingly in *D. melanogaster–D. yakuba* alignments (blue line), there is some prevalence of blocks in the range 11–20, while in the *D. melanogaster–D. pseudoobscura* alignments and other species combinations, this signal disappears.

Somewhat similar conservation signatures are found between 3′ UTRs and 5′ UTRs (Figs. 3F and 3I). In all these regions, blocks in the range 11–20 are overrepresented at short evolutionary distances (alignments with *D. yakuba*) and are completely disrupted at longer evolutionary distances. Results in Fig. 3 also show that the conservation signatures between 5′ UTRs and promoter regions are quite similar in some cases. This is to be expected given the fact that most *Drosophila* promoters are close to the 5′ ends of genes. Table 3 shows similarities in the z-score profiles (correlation matrices) for all considered sequence classes in three species combinations. One can see that the signatures identified in promoters and 5′ UTRs

Table 3
Similarities in signatures of conservation between sequence categories

		Enc	Prm	5′ UTR	Exon	Intron	3′ UTR	Unknown
D.mel.–D.yak.	Enc	1	**0.41**	0.43	0.62	–0.7	*0.8*	–0.5
	Prm	**0.41**	1	*0.89*	*0.91*	–0.9	*0.8*	–0.9
	5′ UTR	**0.43**	*0.89*	1	*0.97*	–0.9	*0.88*	–1
	Exon	**0.62**	*0.91*	*0.97*	1	–1	*0.96*	–1
	Intron	–0.7	–0.9	–0.9	–1	1	–1	*0.95*
	3′ UTR	*0.8*	*0.8*	*0.88*	*0.96*	–1	1	–0.9
	Unknown	–0.5	–0.9	–1	–1	*0.95*	–0.9	1
D.mel.–D.pse.	Enc	1	–0.2	–1	**0.69**	–0.9	–0.7	–0.3
	Prm	–0.2	1	0.29	–0.3	0.29	0.22	*0.15*
	5′ UTR	–1	0.29	1	–0.7	*0.9*	*0.81*	*0.28*
	Exon	**0.69**	–0.3	–0.7	1	–0.9	–0.4	–0.9
	Intron	–0.9	0.29	*0.9*	–0.9	1	**0.69**	0.67
	3′ UTR	–0.7	0.22	*0.81*	–0.4	**0.69**	1	0.11
	Unknown	–0.3	0.15	0.28	–0.9	**0.67**	0.11	1
D.mel.–D.vir.	Enc	1	–0.1	–1	*0.86*	–0.9	–0.9	–0.7
	Prm	–0.1	1	0.14	–0.2	0.12	0.03	0.21
	5′ UTR	–1	0.14	1	–0.9	*0.97*	*0.98*	**0.73**
	Exon	*0.86*	–0.2	–0.9	1	–1	–0.9	–1
	Intron	–0.9	0.12	*0.97*	–1	1	*0.97*	*0.87*
	3′ UTR	–0.9	0.03	*0.98*	–0.9	*0.97*	1	**0.75**
	Unknown	–0.7	0.21	**0.73**	–1	*0.87*	**0.75**	1

A similarity matrix (Pearson correlation values) for the z-score profiles shown in Fig. 3 is shown. Underscoring indicates low correlation ($r < 0.3$); bold, moderate correlation ($0.3 < r < 0.8$); italic, high correlation. In the *D. melanogaster–D. pseudoobscura* and *D. melanogaster–D. virilis* alignments the distributions of block sizes in enhancers are similar to those of exons. At the same time, blocks in exons conform to the $3N + 2$ rule, while blocks in enhancers do not. Note that the "unknown" and exon datasets are dependent to a certain degree, as they contribute the largest number of blocks to the total amount. Instead, the enhancer and exon fractions are nearly independent due to the small contribution (small sample size) of the enhancer fraction.

produce high correlation (r = 0.89) in the case of the *D. melanogaster–D. yakuba* alignment and moderate to low correlation in the case of more distant species (r = 0.29, 0.14).

Finally, one of the most interesting observations was that sequences with no annotation or introns produce signatures opposite to those of exons (Fig. 3C; Table 3, negative correlations). However, in contrast to introns, unannotated sequences contain a moderately abundant fraction of long blocks in the range >20 bases, which may suggest the presence of some yet unannotated enhancers and other functional elements in the fly genome. The presence of this fraction also explains some similarity between the unknown sequences and the enhancers detected in the χ^2 test (see Table 2). Similarity between unannotated sequences and introns is also rather expected as some introns are very long and may contain other genes and regulatory sequences and, in this sense, are not quite different from the intergenic regions without functional annotation.

In general, the analysis of fractional differences between block size distributions clearly demonstrates the presence of signatures inherent to different functional sequence categories.

Ultraconserved Drosophila sequences

Along with rather short conserved blocks, eukaryotic genomes also contain much more extended regions of high identity, sometimes called ultraconserved sequences [12,20]. In this study, we extracted ultraconserved ungapped blocks longer than 59 bases (2303 blocks, 167,778 bases total length) from *D. melanogaster–D. virilis* pair-wise alignments and browsed genome annotations for the extracted sequences.

In the case of regulatory sequences, we found ultraconserved blocks in the following enhancers: Bicoid-dependent enhancer of *giant* (112 bases long), late enhancer of *forkhead* (85 bases), Dorsal-dependent enhancers of *m7* and *snail* (77, 71 bases, correspondingly), stripe 4 + 6 enhancer of *even-skipped* (65 bases), late *even-skipped* enhancer (64 bases), and Bicoid-dependent enhancer of *sloppy-paired* (61 bases). In fact, a number of Bicoid- and Dorsal-dependent enhancers also contain ultraconserved sequences just below the cutoff size (i.e., ~50 or so bases). The frequency of the longest blocks (>50 bases) in enhancers is 7.6×10^{-4}, while this value for the entire genome (all datasets taken together) is 3.5×10^{-4}. Analyses of promoter regions have shown a lower abundance of the ultraconserved regions (as well as other blocks, see Fig. 3). We have found only two blocks longer than 60 bases in the proximal promoter of *mhc* (81 bases) and *tml* (64 bases), while the promoter dataset is comparable in size with the enhancer set. The full list of blocks >30 bases, identical between *D. melanogaster* and *D. virilis*, is available in Supplementary Table S3.

Similarly, we identified all genes containing ultraconserved exons (>59 bases) in the *D. virilis–D. melanogaster* alignments. A total of 240 protein coding genes were found. Fig. 4 summarizes their encoded functions. Nearly a fourth of these genes encode proteins that participate in membrane transport and encode ion channels (see gene names, etc., in Supplementary Table S2). Most of them contain related protein domains, so conservation in this group is likely caused by a specific protein domain structure. The second largest group of genes with ultraconserved sequences encode proteins engaged in cytoskeleton functions. These genes contain a variety of diverse protein domains, so it is likely that the conservation has a functional basis. Glaszov and coworkers obtained similar results in a recent study [20]. Nearly 12% of the long ungapped sequence blocks are associated with genes encoding transcription factors, which is higher than expected by chance (5%). The distribution of the remaining ultraconserved sequences is more or less proportional to the group fraction among all *Drosophila* genes. We have also collected from the *D. melanogaster–D. virilis* alignments all ungapped blocks longer than 30 bases from regions without functional annotation (Supplementary Table S3). These may be helpful as a cross-reference in future analyses, such as finding new enhancers [29] or other functional sequences. For instance, we have found that 19 of 78 *Drosophila* microRNA-encoding regions [37] contain ungapped blocks longer than 30 bases (see Supplementary Table S4).

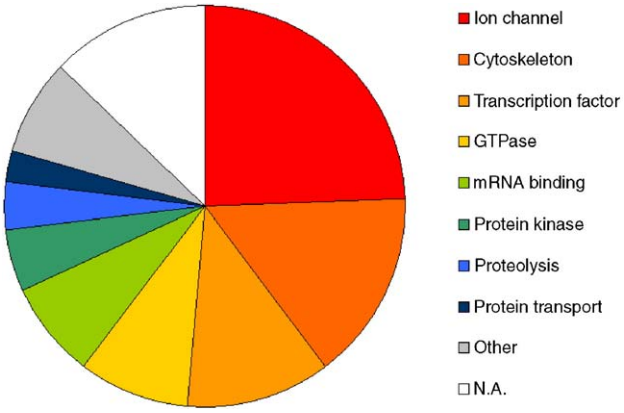

Fig. 4. Distribution of longest blocks among functional gene categories. Most of the ungapped blocks longer than 60 bases were found in exons of ion channel proteins (24%), in genes encoding proteins related to the cytoskeleton (14%), and in genes encoding transcription factors (12%). Exons of other gene categories are not significantly enriched by blocks longer than 60 bases (see also Supplementary Table S1).

Exploration of the ultraconserved fraction demonstrates that the restrained signatures, such as the block lengths, not only may be helpful in discrimination between different functional categories (i.e., enhancers vs exons), but also may provide information on some function-related differences within a category, as we demonstrated in the example with exons.

Discussion

While construction of genome-wide alignments has become a routine procedure, biological interpretation of the information contained in these alignments (patterns of conservation or signatures) is still at the inception stage. Here we have demonstrated that assessment of block lengths brings information that may be helpful in the interpretation of genome alignment data, particularly among species of *Drosophila* in

which there is a substantial conservation (identity score) of intergenic regions, even among distant species. Fig. 1 demonstrates some problems connected with the interpretation based on the window identity scores. Previous studies also dealt with difficulties in the detection of certain functional categories, such as regulatory DNAs (e.g., enhancers), based on standard "window identity score" methods [12].

The key assumption of the present analysis is that function may be reflected in restrained conservation patterns, which do not necessarily depend on total window identity scores. To reveal restrained patterns we conducted a statistical analysis of ungapped conserved block size distributions among different functional sequence categories. Based on statistical analysis we identify specific signatures for the following functional sequence categories: enhancers, promoters, 5′ UTRs, 3′ UTRs, introns, and unknown or unannotated sequences. We have found, for instance, that ungapped blocks with lengths of 21–30 bases (*D. melanogaster–D. virilis* alignments) are overrepresented in enhancers, but not in any of the other sequence categories.

Our findings strongly confirm that specific signatures of conservation are present in functional sequence classes and they can be detected in the pair-wise alignments based on block size statistics.

Signature of exons

The fraction of ungapped conserved blocks with the length $3N + 2$ is highly enriched in exons [27]. While the ungapped blocks in exons are expected to be "broken" in approximately every third position, the prevalence of the $3N + 2$ fraction in some other functional sequence categories was rather unexpected. There are several possible reasons for this. The first is the precision of the genome annotations. It is known that gene-finding algorithms are imprecise and the positions of exon borders contain errors. Clearly, these mapping errors contribute to the presence of a $3N + 2$ bias in UTRs (see Figs. 2E and 2F). In principle, the $3N + 2$ signature can be used as an independent benchmarking test for gene-mapping programs. Along with exon-mapping errors, pseudo genes and "pseudo exons" (changed translation start site) may also contribute to the $3N + 2$ bias (see Fig. 2H).

Fractional differences of block size ranges (see Eq. (2)) also distinguish exons from other sequence categories (see Fig. 3 and Table 3). This type of analysis has revealed strong prevalence of 11- to 20-bp blocks in exons; moreover, this prevalence was quite independent of evolutionary distances between selected species. Apparently, in evolution, exons swiftly break into $3N + 2$ fragments 11, 14, 17, and 20 ($N = 3–6$) but further disruption is under heavy evolutionary pressure. We also found that exons comprise the vast majority of the ultraconserved fraction (longest ungapped blocks). This might also be considered as a signature, but its analysis is less proficient in the alignment interpretation as they are rare by definition. Higher interest represents analysis of the block size distribution among exons of genes with different functional assignment (see Fig. 4). Strength of the $3N + 2$ signal may be increased by a parallel assessment of several pair-wise alignments or even multiple alignments.

Signature of regulatory DNAs

There is currently no code that links primary DNA sequence to enhancer function, as seen for protein coding regions [30,31]. Phylogenetic methods are also inefficient in mapping regulatory sequences (see Fig. 1). Therefore, the identification of alignment signatures is of particular interest in the case of transcription regulatory regions. Here we considered two major types of transcription regulatory regions, proximal promoters [32] and enhancer regions (124 sequences, available at https://webfiles.berkeley.edu/dap5/public_html/index.html).

Statistical analysis of block frequency histograms has demonstrated that in enhancer and promoter regions the block size distributions are different from the other functional sequence categories (see Figs. 3A, 3D, and 3G, Supplementary Fig. S1). Correlation values in Table 3 show that there is a certain level of similarity between enhancers and exons (prevalence of blocks in the 11–20 range), but enhancers contain no traces of $3N + 2$ signal (data not shown). In addition, enhancers contain a larger proportion of extended sequence blocks, 21–40 and 61–100 bp, than exons. The basis for such extensive DNA conservation in enhancers is not known. Most functional signals in enhancers correspond to binding sites for individual sequence-specific transcription factors. Perhaps the larger blocks of conservation correspond to composite elements containing two or more tightly linked binding sites [33]. The conservation of such elements could explain ungapped blocks of 11–30 bp. In principle, enhancers can be identified by the prevalence of 11- to 30-bp blocks lacking the $3N + 2$ signal seen for exons. Earlier, Bergman and coworkers [1,19] observed that the block length in noncoding DNA, on average, is larger than the length of a single binding site. They also attributed this phenomenon to the module level of enhancer structure [33,34], i.e., to the presence of the linked binding sites or binding site clusters.

In contrast to enhancers, clear specific signature of conservation was not detected in promoter regions. In addition to core elements, such as TATA, CAAT, and DPE [32], promoters might also contain composite elements or linked binding sites, such as those in enhancers. However, in general, signatures detected in promoter regions were more similar to those seen in UTRs (see Figs. 3F and 3I and Table 3). These results may suggest that commonly accepted automatic partition of promoter regions (−200, +50, relative to transcription start site) may not be optimal for this sort of analysis. The identification of unique promoter signatures must await the compilation of a more reliable dataset.

Interpretation of signatures in unknown fraction

Sequences without any functional annotation have shown some prevalence of long ungapped blocks (see Fig. 3H). This finding, at first glance, is surprising. However, it is possible that at least some of the long blocks in the unknown fraction also

belong to enhancers or other transcription regulatory regions. One has also to take into consideration that most of exons, UTRs, and introns are already known, but a large fraction of regulatory regions, especially these that are far from the transcription start, is still "hidden" among the unannotated sequences. In fact, precision of the current promoter- and enhancer-finding algorithms is not even close to the precision of gene-finding algorithms.

On the other hand, little is known about the connection between block length and sequence function, so there is even a chance that some structural regions or "parasitic" or other repetitive DNAs are responsible for the presence of the long blocks among the unknown fraction. Solving problems related to interpretation of the alignments found in the unannotated regions will require further analysis and better genome annotation using independent techniques. Therefore we have collected long ungapped blocks from unannotated regions (>30 bases) and generated a database (see Supplementary Table S2) that may help in future analysis of sequences with no functional annotation.

A number of ultraconserved sequences were found in regions of unknown function. It is conceivable that some of these are associated with unknown regulatory DNAs since just a small fraction of such DNAs are known. Others are associated with microRNA genes (see Supplementary Table S4) since there is extensive conservation of the 80- to 100-bp stem–loop structure, the pre-microRNA, which is processed into the mature 21- to 24-nt microRNA. In addition, some ultraconserved blocks may be associated with sequences involved in chromosome integrity and condensation of heterochromatin. More details on functional assignment of ultraconserved sequences from *Drosophila* can be found in the recent dedicated study [20].

Prospective directions in alignment interpretation

As we discussed, construction of genome-wide alignments is only a first step in the phylogenetic analysis of genome information; undoubtedly, it will require an interpretation step to achieve efficient mapping of biologically significant features.

We approached the interpretation problems by considering ungapped block lengths and their statistics present in different functional sequence categories (signatures). Current study can be extended in several directions. First, it will be very helpful to include consideration of the type I gaps (mismatches) between the blocks. Small gaps (i.e., 1–2 bases) might be especially important as they often correspond to breaks within functional patterns, as in the case with exons (third position of codons). Thus, we have already observed that masking of the short type I gaps (mismatches) will dramatically change statistics for the conserved blocks. Second, the consideration of type I gaps and blocks can simply be extended to block-gap Markov models that can be trained using the same functional sequence classes. We expect these models to be more informative and selective than our current signatures, based exclusively on the ungapped blocks. Supposedly, statistical interpretation of a sliding window containing only a few blocks and gaps may appear to be inefficient due to the lack of the information. However, for most basic model organisms, there is typically more than one related genome, so several pair-wise alignments can simultaneously be assessed using a mapping algorithm.

In their turn, multiple alignments will also require more efficient methods of interpretation. To some extent, they can be analyzed using a very similar approach accounting for blocks and gaps between them; however, this consideration will require more parameters, as the same blocks and gaps may be present in only some of the aligned sequences. As we discussed above, multiple alignments are also more ambiguous, so their interpretation using statistical approaches is expected to be more complicated. Finally, the statistical alignment interpretations can be combined with existing methods of gene mapping, promoter finding, and binding site/binding site cluster recognition.

Perhaps, the conserved signatures reported in this study for *Drosophila* may be identified in other organisms as well. We expect, however, significant signature variations between densely packed fly genomes and, for instance, much more "sparse" (i.e., containing more "background") vertebrate genomes.

Materials and methods

Drosophila genome assemblies

The following assemblies were used in the analysis: *D. melanogaster* Genome Assembly, BDGP release 3.1 January 2003; *D. pseudoobscura* July 2003 (Baylor College of Medicine); *D. virilis* July 2004 (Agencourt Bioscience Corp.); *D. ananassae* July 2004 (TIGR); *D. yakuba* April 2004 (release 1.0) (Washington University School of Medicine in St. Louis); *D. mojavensis* August 2004 (Agencourt Bioscience Corp.).

Alignment methods

We used the Berkeley Genome Pipeline infrastructure for the construction of genome-wide pair-wise alignments of *D. melanogaster* with *D. pseudoobscura, D. virilis, D. ananassae, D. yakuba*, and *D. mojavensis*.

To align genomes we have implemented new algorithms that used an efficient combination of both global and local alignment methods [10]. The sequences of each species were mapped to the *D. melanogaster* genome as follows. First, we obtained a map of large blocks of conserved synteny between the two species by applying the Shuffle-LAGAN glocal chaining algorithm to local alignments produced by translated BLAT [35]. After that, we applied Super Map, the fully symmetric whole-genome extension to the Shuffle-LAGAN algorithm [9]. To ensure that only nonduplicate, unique homology regions were selected for pattern analysis, only dual-monotonic alignment regions as produced by Super Map were used. Then, in each syntenic block, we applied Shuffle-LAGAN a second time to obtain a more fine-grained map of small-scale rearrangements such as inversions. The sensitivity of alignments was measured by fractions of sequence features covered by alignments (see Table 1) using the techniques first applied to the human–mouse alignment [6].

The constructed genome-wide pair-wise alignments of different species of *Drosophila* are available at the URL http://pipeline.lbl.gov/downloads.shtml and can be accessed for browsing and various types of analysis through the VISTA browser at http://pipeline.lbl.gov.

Construction of functional datasets

In the current work, we explored the following seven functional sequence categories: enhancers, proximal promoters, 5′ UTRs, exons, introns, 3′ UTRs, and unknown—the fraction of sequences without any available annotation.

Exons, introns, UTRs, and unknown datasets were based on standard *Drosophila* genome annotations (release 3.1) and were obtained as a RefSeq dataset for *D. melanogaster* from the UCSC genome browser [12].

One hundred ninety-eight promoter regions were downloaded from the *Drosophila* Core Promoter Database (by A. Kutach, S. Iyama, J. Kadonaga) [32]. The selected promoter segments were adjusted to cover region −250 to +50 relative to transcription start sites of the corresponding genes. One hundred twenty-four experimentally validated enhancer regions were compiled from available databases and relevant literature, including most recent publications. Enhancer sequences are available for download from the enhancer collection by D. Papatsenko [31] and from the recently introduced REDfly database available from the M. Halfon Web resource [36].

Acknowledgments

The authors are grateful to Michael Brudno and Alexander Poliakov for their extensive work on *Drosophila* alignments analyzed in the paper and Michael Cipriano for help with the manuscript. This work was supported by National Heart, Lung, and Blood Institute, National Institutes of Health, Grant U1HL66681B; the U.S. Department of Energy's Office of Science, Biological, and Environmental Research Program Lawrence Berkeley National Laboratory Contract DE-AC03-76SF00098 to I.D.; and National Institutes of Health Grant GM 46638 to M.L.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ygeno.2006.03.012.

References

[1] C.M. Bergman, M. Kreitman, Analysis of conserved noncoding DNA in Drosophila reveals similar constraints in intergenic and intronic sequences, Genome Res. 11 (2001) 1335–1345.
[2] C.M. Bergman, B.D. Pfeiffer, D.E. Rincon-Limas, R.A. Hoskins, A. Gnirke, C.J. Mungall, A.M. Wang, B. Kronmiller, J. Pacleb, S. Park, M. Stapleton, K. Wan, R.A. George, P.J. de Jong, J. Botas, G.M. Rubin, S.E. Celniker, Assessing the impact of comparative genomic sequence data on the functional annotation of the Drosophila genome, Genome Biol. 3 (2002) (RESEARCH0086).
[3] M. Ashburner, C.M. Bergman, Drosophila melanogaster: a case study of a model genomic sequence and its consequences, Genome Res. 15 (2005) 1661–1667.
[4] S. Schwartz, Z. Zhang, K.A. Frazer, A. Smit, C. Riemer, J. Bouck, R. Gibbs, R. Hardison, W. Miller, PipMaker—A web server for aligning two genomic DNA sequences, Genome Res. 10 (2000) 577–586.
[5] S. Schwartz, L. Elnitski, M. Li, M. Weirauch, C. Riemer, A. Smit, E.D. Green, R.C. Hardison, W. Miller, MultiPipMaker and supporting tools: alignments and analysis of multiple genomic DNA sequences, Nucleic Acids Res. 31 (2003) 3518–3524.
[6] S. Schwartz, W.J. Kent, A. Smit, Z. Zhang, R. Baertsch, R.C. Hardison, D. Haussler, W. Miller, Human–mouse alignments with BLASTZ, Genome Res. 13 (2003) 103–107.
[7] C. Mayor, M. Brudno, J.R. Schwartz, A. Poliakov, E.M. Rubin, K.A. Frazer, L.S. Pachter, I. Dubchak, VISTA: visualizing global DNA sequence alignments of arbitrary length, Bioinformatics 16 (2000) 1046–1047.
[8] N. Bray, I. Dubchak, L. Pachter, AVID: a global alignment program, Genome Res. 13 (2003) 97–102.
[9] M. Brudno, S. Malde, A. Poliakov, C.B. Do, O. Couronne, I. Dubchak, S. Batzoglou, Glocal alignment: finding rearrangements during alignment, Bioinformatics 19 (2003) i54–i62.
[10] M. Brudno, C.B. Do, G.M. Cooper, M.F. Kim, E. Davydov, E.D. Green, A. Sidow, S. Batzoglou, LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA, Genome Res. 13 (2003) 721–731.
[11] L.A. Pennacchio, E.M. Rubin, Genomic strategies to identify mammalian regulatory sequences, Nat. Rev. Genet. 2 (2001) 100–109.
[12] G. Bejerano, M. Pheasant, I. Makunin, S. Stephen, W.J. Kent, J.S. Mattick, D. Haussler, Ultraconserved elements in the human genome, Science 304 (2004) 1321–1325.
[13] D.A. Pollard, C.M. Bergman, J. Stoye, S.E. Celniker, M.B. Eisen, Benchmarking tools for the alignment of functional noncoding DNA, BMC Bioinform. 5 (2004) 6.
[14] B.P. Berman, B.D. Pfeiffer, T.R. Laverty, S.L. Salzberg, G.M. Rubin, M.B. Eisen, S.E. Celniker, Computational identification of developmental enhancers: conservation and function of transcription factor binding-site clusters in Drosophila melanogaster and Drosophila pseudoobscura, Genome Biol. 5 (2004) R61.
[15] A. Siepel, D. Haussler, Combining phylogenetic and hidden Markov models in biosequence analysis, J. Comput. Biol. 11 (2004) 413–428.
[16] A. Siepel, D. Haussler, Statistical Methods in Molecular Evolution, Springer-Verlag, Berlin/New York, 2005.
[17] S.A. Shabalina, A.Y. Ogurtsov, I.B. Rogozin, E.V. Koonin, D.J. Lipman, Comparative analysis of orthologous eukaryotic mRNAs: potential hidden functional signals, Nucleic Acids Res. 32 (2004) 1774–1782.
[18] E. Emberly, N. Rajewsky, E.D. Siggia, Conservation of regulatory elements between two species of Drosophila, BMC Bioinform. 4 (2003) 57.
[19] E.T. Dermitzakis, C.M. Bergman, A.G. Clark, Tracing the evolutionary history of Drosophila regulatory regions with models that identify transcription factor binding sites, Mol. Biol. Evol. 20 (2003) 703–714.
[20] E.A. Glazov, M. Pheasant, E.A. McGraw, G. Bejerano, J.S. Mattick, Ultraconserved elements in insect genomes: a highly conserved intronic sequence implicated in the control of homothorax mRNA splicing, Genome Res. 15 (2005) 800–808.
[21] D. Grun, Y.L. Wang, D. Langenberger, K.C. Gunsalus, N. Rajewsky, microRNA target predictions across seven Drosophila species and comparison to mammalian targets, PLoS Comput. Biol. 1 (2005) e13.
[22] A.G. Nazina, D.A. Papatsenko, Statistical extraction of Drosophila cis-regulatory modules using exhaustive assessment of local word frequency, BMC Bioinform. 4 (2003) 65.
[23] S. Misra, M.A. Crosby, C.J. Mungall, B.B. Matthews, K.S. Campbell, P. Hradecky, Y. Huang, J.S. Kaminker, G.H. Millburn, S.E. Prochnik, C.D. Smith, J.L. Tupy, E.J. Whitfied, L. Bayraktaroglu, B.P. Berman, B.R. Bettencourt, S.E. Celniker, A.D. de Grey, R.A. Drysdale, N.L. Harris, J. Richter, S. Russo, A.J. Schroeder, S.Q. Shu, M. Stapleton, C. Yamada, M. Ashburner, W.M. Gelbart, G.M. Rubin, S.E. Lewis, Annotation of the Drosophila melanogaster euchromatic genome: a systematic review, Genome Biol. 3 (2002) (RESEARCH0083).
[24] K.A. Frazer, L. Pachter, A. Poliakov, E.M. Rubin, I. Dubchak, VISTA: computational tools for comparative genomics, Nucleic Acids Res. 32 (2004) W273–W279.
[25] E.H. Margulies, M. Blanchette, D. Haussler, E.D. Green, Identification and characterization of multi-species conserved sequences, Genome Res. 13 (2003) 2507–2518.
[26] W. Fitch, Toward defining the course of evolution: minimum change for a specified tree topology, Syst. Zool. 20 (1971) 406–416.
[27] E.T. Dermitzakis, A. Reymond, R. Lyle, N. Scamuffa, C. Ucla, S. Deutsch, B.J. Stevenson, V. Flegel, P. Bucher, C.V. Jongeneel, S.E. Antonarakis, Numerous potentially functional but non-genic conserved sequences on human chromosome 21, Nature 420 (2002) 578–582.
[28] S.A. Glantz, Primer on Biostatistics, McGraw–Hill, New York, 2005.
[29] D. Papatsenko, M. Levine, Computational identification of regulatory DNAs underlying animal development, Nat. Methods 2 (2005) 529–534.
[30] B.P. Berman, Y. Nibu, B.D. Pfeiffer, P. Tomancak, S.E. Celniker, M. Levine, G.M. Rubin, M.B. Eisen, 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 (2002) 757–762.
[31] A.P. Lifanov, V.J. Makeev, A.G. Nazina, D.A. Papatsenko, Homotypic regulatory clusters in Drosophila, Genome Res. 13 (2003) 579–588.

[32] A.K. Kutach, J.T. Kadonaga, The downstream promoter element DPE appears to be as widely used as the TATA box in Drosophila core promoters, Mol. Cell. Biol. 20 (2000) 4754–4764.
[33] V.J. Makeev, A.P. Lifanov, A.G. Nazina, D.A. Papatsenko, Distance preferences in the arrangement of binding motifs and hierarchical levels in organization of transcription regulatory information, Nucleic Acids Res. 31 (2003) 6016–6026.
[34] M.I. Arnone, E.H. Davidson, The hardwiring of development: organization and function of genomic regulatory systems, Development 124 (1997) 1851–1864.
[35] W.J. Kent, BLAT—The BLAST-like alignment tool, Genome Res. 12 (2002) 656–664.
[36] S.M. Gallo, L. Li, Z. Hu, M.S. Halfon, REDfly: a regulatory element database for Drosophila, Bioinformatics 22 (2006) 381–383.
[37] S. Griffiths-Jones, The microRNA Registry, Nucleic Acids Res. 32 (2004) D109–D111.