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

Dmitri Papatsenko[1*], Andrey Kislyuk[2], Michael Levine[1] and Inna Dubchak[2]

[1]Department of Molecular and Cell Biology, University of California at Berkeley
[2]Genomics Division, Lawrence Berkeley National Laboratory
*Corresponding author, email: dxp@berkeley.edu

Abstract

We have explored the distributions of fully conserved ungapped blocks in genome-wide pairwise alignments of recently completed species of *Drosophila*: *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 non-functional sequences. At the same time, the blocks in the coding regions carried 3N+2 signature characteristic to synonymic substitutions in the 3rd codon positions. 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 discussed how restrained conservation patterns may help in mapping functional sequence categories and improving genome annotation.

Introduction

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 finding 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 pairwise alignments of *Drosophila*. The distribution of block sizes was originally explored by Bergman and coworkers using pairwise alignments of several genomic intervals of two *Drosophila* species {Bergman, 2001 #38}. 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 {Bergman, 2002 #48; Ashburner, 2005 #47} (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 methods, based on local (PIPMaker for blastz {Schwartz, 2000 #24;Schwartz, 2003 #25;Schwartz, 2003 #26}) or global alignment algorithms (VISTA for AVID and LAGAN) {Mayor, 2000 #19;Bray, 2003 #5;Brudno, 2003 #7;Brudno, 2003 #6}, are very efficient in finding long stretches of conservation, including ultraconserved regions {Pennacchio, 2001 #22;Bejerano, 2004 #1}. However, these methods are not focused, for instance, on efficient finding of transcription regulatory elements on a large scale or binding sites for individual regulatory proteins on a smaller scale {Pollard, 2004 #23;Berman, 2004 #3}. Some programs, however, approach the problem of alignment interpretation in a more accurate way. For instance, 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 one {Siepel, 2004 #27;Siepel, 2005 #28}. While mathematical models based on nucleotide substitution matrices {Bergman, 2001 #38;Siepel, 2005 #28} help in detection of the conserved regions, the role of block size and its relation with sequence function remains relatively unexplored. Strategy of Siepel and coworkers {Siepel, 2005 #28} 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 {Shabalina, 2004 #46}.

Finding functional conservation signatures, such as characteristic block sizes, is especially important for mapping transcription regulatory regions. The comparative analysis of *D. 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 non-coding genome {Emberly, 2003 #9}. 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 {Berman, 2004 #3},

based on the same strategy (window identity scores), showed that *cis*-regulatory elements appear indistinguishable from flanking sequence as there is high amount of non-coding sequence conservation throughout the analyzed gene loci. At the same time, Bergman and coworkers {Bergman, 2001 #38;Dermitzakis, 2003 #39} 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 {Glazov, 2005 #11}, the authors shown 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 towards the interpretation of the alignment patterns based on the block size and explore how sizes are distributed among five different functional sequence categories: coding regions, untranslated regions (UTRs), transcription regulatory regions (promoters and enhancers) and unannotated sequences in the genome of *Drosophila melanogaster*. We also analyzed functional assignment of the longest ungapped blocks (ultraconserved) and conservation of some other functionally important sequences, such as microRNA {Grun, 2005 #42}.

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

Results

1. Current limits in functional interpretation of genome alignments

To demonstrate existing problems with functional alignment interpretation, we explored conservation of some functional regions from *Drosophila* using conventional phylogenetic method, based on window identity scores {Mayor, 2000 #19}. We focused on several of the most annotated developmental gene loci, containing a number of well-known transcription regulatory regions, and fly enhancers {Nazina, 2003 #21}. 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. **Figure 1**, top track shows comparison of VISTA plots, where conserved regions (colored) were calculated with 70% identity in 100 bp window cutoff, and the map of annotated functional regions for the loci of two developmental genes – even-skipped and *fushi-tarazu*. While the coding regions correlated with the conserved regions (peaks) well, the distributions of enhancer regions correlated with the conservational profiles at much lower degree (r = 0.3-0.4). In many cases, the overall conservation level in the enhancer regions was not higher than the conservation level in flanking non-functional genomic intervals {Nazina, 2003 #21}. On the same data set, we also explored correlation between distribution of ungapped conserved blocks with both VISTA profile,

and the functional map (middle track in the **Figure 1**). Surprisingly, we have found that exons do not contain 100% ungapped conserved blocks longer than 40 bases, but such blocks were present in enhancer regions. The distribution of the ungapped blocks was quite different from the VISTA score profile.

This analysis demonstrated that the alignment interpretation based on standard window identity scores (such as VISTA) may be improved further. More information can be extracted from the alignments if they are given consideration of block and gap lengths 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 categories.

2. Construction and evaluation of pairwise alignments

In order 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 very rich assortment of recently completed related fly genomes (Available from LBNL web resource: http://rana.lbl.gov/drosophila/), and outstanding level of genome annotation for *D.melanogaster* {Misra, 2002 #20}.

We based our analysis on pairwise 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 Shuffle-LAGAN alignment module {Brudno, 2003 #7;Frazer, 2004 #10} (see also "Methods" section). Quality of the alignments was estimated using standard measures, such as coverage of the entire base genome and its functional features (annotated regions) {Schwartz, 2003 #26}.

Table 1 shows summary statistics for the genome-wide pairwise alignments. These alignments cover different fractions of the *D.melanogaster* genome, depending on the evolutionary distance between compared species and quality of genome assemblies. The achieved coverage of exons (85.2 – 97.8%) suggested that 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 also calculated the total lengths of the ungapped blocks (total number of matches) in the alignments. The ungapped blocks covered 30-70% of the base genome, depending on evolutionary distance, so at least this or (highly likely) much higher fraction of the base genome (*D.melanogaster*) can be annotated based on the ungapped block statistics.

3. Definition of restrained patterns of conservation in pairwise alignments

Along with window identity scores and nucleotide substitution matrices, conservation of a DNA sequence can be described by a sequence of lengths for ungapped conserved blocks, mismatches, and gaps in a pair-wise alignment. The importance of this feature has been demonstrated earlier in several related studies {Bergman, 2001 #38;Glazov, 2005 #11}. The block-gap sequences *S* (see 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 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, described in details elsewhere {Margulies, 2003 #18;Fitch, 1971 #29}.

For the described technical and biological reasons, pairwise alignments are more convenient for building a catalog/statistics for gap, mismatch and block lengths; there can be potentially 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, when either sequence is absent from an alignment are different, so they may be considered as type II gaps. It is unclear how much information can be obtained from the statistics of lengths of the type II gaps (unaligned regions) as they apparently correspond to non-functional sequences (insertions), which are not under evolutionary pressure and apparently may substantially vary in the size as well. Similar considerations are applicable, to some extent, to the type I gaps (or 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 that the size of the ungapped block is more likely to be the functional indicator than the size of the gap between two ungapped blocks. Usually, functional regions or sites expose higher degree of conservation, therefore extended ungapped blocks (ultraconserved regions) may represent higher biological interest than very long gaps, - simply absence of alignments.

Here, we decided to begin with defining alignment patterns through a size of ungapped 100% conserved blocks leaving incorporation of the 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 pairwise alignments recently obtained with LAGAN global alignment algorithm. According to a detailed study by Pollard and colleagues {Pollard, 2004 #23}, 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 population of noncoding sequences as well.

4. Distribution of exon-specific block sizes across 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 3rd position of amino acid codons can be a subject to synonymic substitutions. On the example of human-mouse partial genome alignments, Dermitzakis and coworkers have shown that the direct consequence of synonymic substitutions is overrepresentation of ungapped blocks with the size 3N+2 in the coding regions {Dermitzakis, 2002 #41}.

To explore distribution of 3N+2 blocks in genome-wide pairwise alignments of *Drosophila* we generated frequency histograms for the ungapped block sizes for each considered pairwise alignment between the *Drosophila* species. **Figure 2** shows that in all cases exons are highly enriched by the ungapped blocks with the size 3N+2 (up to 5-6 times, see **Figure 2 B**, **C**). In order to provide more sensitive method than frequency histogram, we performed signal filtering. We calculated excess *E* of 3N+2 fraction as a

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

Presence and distribution of 3N+2 signal in *Drosophila* supported previous finding by Dermitzakis and coworkers {Dermitzakis, 2002 #41}, obtained for human chromosome 21. Our signal filtering has shown that even blocks in the range 60-100 bases in exons (*D.melanogaster-D.pseudoobscura* alignments) carry 3N+2 signature and the traces of the signal are present in untranslated regions and in some unannotated sequences (see **Figure 2**). The test has also shown that the restrained functional patterns are not lost in our most recent LAGAN/VISTA pairwise alignments and these signatures are specific to functional sequence categories.

5. Regulatory regions and UTRs possess their own signatures

In order to detect possible presence of the functional signatures in other than CDS functional sequence categories we analyzed differences in the block frequency histograms built for seven functional sequence classes: enhancers, promoters, 5' UTR's, exons, introns, 3' UTR's, and "unknown" (sequences without annotation). The large enhancer and promoter datasets have not been subjected to this type of analysis before. 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 is 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 to identify fine differences/similarities between the functional classes in the each block size range (bin). We compared fraction of blocks in each bin of each functional category with the total fraction of blocks in the same bin obtained from entire genome alignment (all categories). We calculated *z*-scores for each bin as follows {Glantz, 2005 #30}:

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

In this formula p_1 is fraction of blocks observed in a block size range (i.e. [1-10]) for the analyzed sequence category; p_2 is 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 **Figure 3**).

We observed that both the enhancer regions and exons contain larger amount of [20-30] blocks, but the enhancers are also enriched by ungapped blocks longer than 20 bases, present in introns and "unknown" fractions (compare **Figure 3**, **A**, **B**). This effect is more striking in the case of pairwise 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 synonymic 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] (**Figure 3**, **D**, **G**), or one additional (to the noise) block 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 that case it was more difficult to judge due to the smaller size of the enhancer dataset. Nevertheless in *D.melanogaster* - *D.pseudoobscura* alignments ~ 40 % of enhancers contained 100% conserved blocks longer than 35-40 bases and few contained very long blocks exceeding 60 or more bases. In the case of enhancers and exons, the *z*-score profiles across the size ranges were in agreement for all considered combinations of species (see lines of different color in **Figure 3**).

In difference from enhancers, the promoter regions (198 sequences) displayed 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 (**Figure 3**, **F**). 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 combination this signal disappears.

Somewhat similar conservation signatures were found between 3' UTRs and 5' UTRs (**Figure 3 F**, **I**). 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 **Figure 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 produce high correlation (r=0.89) in the case of *D.melanogaster*-*D.yakuba* alignments 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 and introns produced opposite signatures to that of exons (**Figure 3, C, Table 3**, negative correlation). However, in contrast to introns, unannotated sequences contained moderately abundant fraction of long blocks in the range > 20 bases, which may suggest presence of some yet unannotated enhancers and other functional elements

in fly genome. Presence of this fraction also explains some similarity between the "unknown" sequences and enhancers detected in the chi-square test (see **Table 2**). Similarity between unannotated sequences and introns is also rather expected as some introns are very long, 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 difference between block size distributions has clearly demonstrated presence of signatures, inherent to different functional sequence categories.

6. Ultraconserved Drosophila sequences

Along with rather short conserved blocks, eukaryotic genomes also contain much more extended regions of high identity, sometimes called ultraconserved sequences {Bejerano, 2004 #1;Glazov, 2005 #11}. In this study, we extracted ultraconserved ungapped blocks longer than 59 bases (2303 blocks, 167,778 bases total length) from *D.melanogaster-D.virilis* pairwise alignments and browsed genome annotations for the extracted sequences.

In the case of regulatory sequences, we have 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 contained ultraconserved sequences just below the cutoff size (i.e. ~ 50 or so bases). The frequency of the longest blocks (>50 bases) in enhancers is 7.6E-04, while this value for the entire genome (all data sets, taken together) is 3.5E-04. Analysis of promoter regions has shown lower abundance of the ultraconserved regions (as well as other blocks, see **Figure 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 data set is comparable by its size with enhancers. 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 D.virilis-D.melanogaster alignments. A total of 240 protein coding genes were found, **Figure 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 **Table S2**, supplement). Most of them contain related protein domains, so conservation in this group is likely caused by 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 co-workers obtained similar results in a recent study {Glazov, 2005 #11}. 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 (**Table S3**, supplement). These may be helpful as a cross-reference in future analyses, such as

finding new enhancers {Papatsenko, 2005 #31} or other functional sequences. For instance, we have found that 19 out of 78 Drosophila micro RNA encoding regions contain ungapped blocks longer than 30 bases (see **Table S4**, supplement).

Exploration of the ultraconserved fraction demonstrated that the restrained signatures, such as the block lengths, may be helpful not only in discrimination between different functional categories (i.e. enhancers vs. exons), but may also 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 Drosophilids where there is a substantial conservation (identity score) of intergenic regions, even among distant species. **Figure 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 {Bejerano, 2004 #1}.

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. In order 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 identified 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 pairwise 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 {Dermitzakis, 2002 #41}. While the ungapped blocks in exons are expected to be "broken" in approximately every 3rd position, 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 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 3N+2 bias in UTRs (see **Figure 2 E**, **F**). 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 **Figure 2 H**).

Fractional differences of block size ranges (see formula 2) also distinguish exons from other sequence categories (see **Figure 3**, and **Table 3**). This type of analysis has revealed strong prevalence of 11-20 bp blocks in exons; moreover, this prevalence was quite independent from evolutionary distances between selected species. Apparently, in evolution, exons swiftly break into 3N+2 fragments 11, 14, 17, 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 **Figure 4**). Strength of 3N+2 signal may be increased by a parallel assessment of several pairwise 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 {Berman, 2002 #2;Lifanov, 2003 #16}. Phylogenetic methods are also inefficient in mapping regulatory sequences (see **Figure 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 {Kutach, 2000 #15} 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 **Figure 3 A**, **D**, **G**). Correlation values in **Table 3** show that there is a certain level of similarity between enhancers and exons (prevalence of blocks in [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 2 or more tightly linked binding sites {Makeev, 2003 #17}. The conservation of such elements could explain ungapped blocks of 11-30 bp. In principle, enhancers can be identified by the prevalence of 11-30 bp blocks lacking the 3N+2 signal seen for exons. Earlier, Bergman and coworkers {Bergman, 2001 #38;Dermitzakis, 2003 #39} observed that the block length in non-coding 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 {Arnone, 1997 #43;Makeev, 2003 #17}, i.e. to the presence of the linked binding sites or binding site clusters.

In difference from enhancers, clear specific signature of conservation was not detected in promoter regions. In addition to core elements, such as TATA, CAAT, DPE etc {Kutach, 2000 #15}, promoters might also contain composite elements or linked binding sites, such as these in enhancers. However, in general, signatures detected in promoter regions were more similar to those seen in UTRs (see **Figure 3 F**, **I** and **Table 3**). These results may suggest that commonly accepted automatic partition of promoter regions (-200, +50, relatively 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 **Figure 3 H**). This finding, on the 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 take into consideration that most of exons, UTRs and introns are already known, but the large fraction of regulatory regions, especially these that are far from 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 connection between the block length and sequence function, so there is even a chance that some structural regions, "parasitic" or other repetitive DNA is 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 **Table S2** supplement) that may help in future analysis of the 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 miRNA genes, (see **Table S4**, supplement) since there is extensive conservation of the 80-100 bp stem-loop structure, the pre-miRNA, that is processed into the mature 21-24 nt miRNA. 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 {Glazov, 2005 #11}.

Prospective directions in alignment interpretation

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

We approached the interpretation problems from considering ungapped block lengths and their statistics present in different functional sequence categories (signatures). Current study can be extended into 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 (3^{rd} 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 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 pairwise 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 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 only in 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 Jan. 2003; *D. pseudoobscura* July 2003 (Baylor College of Medicine); *D. virilis* Jul. 2004 (Agencourt Bioscience Corporation); *D. ananassae* Jul. 2004 (TIGR); *D. yakuba* Apr. 2004 (Release 1.0) (Washington University School of Medicine in St. Louis); *D. mojavensis* Aug. 2004 (Agencourt Bioscience Corporation).

Alignment methods

We used the Berkeley Genome Pipeline infrastructure for the construction of genome-wide pairwise 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 {Brudno, 2003 #6}. 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 Shuffle-LAGAN glocal chaining algorithm to local alignments produced by translated BLAT {Kent, 2002 #14}. After that, we applied Super map, the fully symmetric whole-genome extension to the Shuffle-LAGAN algorithm {Brudno, 2003 #7}. To ensure that only non-duplicate, 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 {Schwartz, 2003 #26}.

The constructed genome-wide pairwise alignments of different species of Drosophila are available at the following 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" – fraction of sequences without any available annotation.

Exons, introns, UTRs and "unknown" data sets were based on standard Drosophila genome annotations (release 3.1) and were obtained as RefSeq data set for D. melanogaster from the UCSC genome browser {Browser #36}.

198 promoter regions were downloaded from *Drosophila* Core Promoter Database (DCPD, by A. Kutach, S. Iyama, J. Kadonaga) {Kutach, 2000 #15}. The selected promoter segments were adjusted to cover region -250 - +50 relatively to transcription start sites of the corresponding genes. 124 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 {Lifanov, 2003 #16} and from recently introduced REDfly database available from M. Halfon web resource {Gallo, 2006 #44}.

Acknowledgements

The authors are grateful to Michael Brudno and Alexander Poliakov for their extensive work on *Drosophila* alignments analyzed in the paper, Casey Bergman for careful manuscript reading, critical remarks, and suggested changes. This work was supported by National Heart, Lung, and Blood Institute, National Institutes of Health Grant U1HL66681B; U.S. Department of Energy's Office of Science, Biological, and Environmental Research Program, Lawrence Berkeley National Laboratory Contract DE-AC02-05CH11231 to I.D. and the National Institutes of Health Grant GM 46638 to M.L.

Figure and Table legends

Figure 1. Patterns of conservation in *eve* and *ftz* gene loci

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

Figure 2. Power of 3N+2 signal in exons and other sequences

Frequency histograms (A, D, G) show presence of the 3N+2 signal in exons. Results of filtering (see equation (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.

Figure 3. Unequal distribution of block sizes among different sequence categories
Panels (A-C, E, F, H, I) show *z*-score profiles for fractional abundance of block in different block size ranges. Panels (D, G) show relative amount 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 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 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.

Figure 4. Distribution of longest blocks among functional gene categories
Most of ungapped blocks longer than 60 bases were found in exons of ion channels proteins (24%), in genes encoding proteins related to cytoskeleton (14%) and in genes encoding transcription factors (12%). Exons of other gene categories are not significantly enriched by block longer than 60 bases (see also **Table 1** supplement).

Table 1. Quality of pairwise alignments
Table shows coverage of genome annotation by pairwise alignments used in this study. Loose and tight coverage values were calculated according to previously described method {Schwartz, 2003 #26}. Bottom row shows fraction of the base genome covered by ungapped 100% conserved blocks, i.e. fraction of base pairs of the base genome exactly matching the second genome.

Table 2. Differences between enhancers and other sequence categories
Table shows *p*-values obtained from chi-square test. 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 in the "Results" section). While the distribution of block sizes in enhancers is close to that of intros and unannotated sequences (see red numbers), these three categories are still distinguishable, especially in *D.melanogaster-D.yakuba* alignments (p=7.09E-08).

Table 3. Similarities in signatures of conservation between sequence categories
Table shows similarity matrix (Pearson correlation values) for the *z*-score profiles shown in the **Figure 3**. Blue color indicates low correlation, (r <0.3), green color – moderate correlation (0.3<r<0.8), red color – high correlation. In *D.melanogaster-D.pseudoobscura* and in *D.melanogaster-D.virilis* alignments distribution of block sizes in enhancers is similar to that of exons. In the same time, blocks in exons confer to 3N+2 rule, while blocks in enhancers are not. Note that the "unannotated" and the exon datasets are dependent to a certain degree, as they contribute largest number of blocks to the total amount. Instead, enhancer and exon fractions are nearly independent due to the small contribution (small sample size) of the enhancer fraction.

Supplementary Figure 1. Effect of block length range selection (bin selection)

Figure compares the relative amount of ungapped blocks for all sequence categories, as in the Figure 3, D, G, but for various block size ranges (shown on the right). Enhancers prevail among the blocks in the size range [16-25] and longer, regardless on the size range (bin) selection.

Supplementary Table S1. Block frequency histograms
Table shows frequency histograms for different combinations of species and functional datasets.

Supplementary Table S2. Ultraconserved ungapped blocks in coding sequences
Drosophila genome release 3.2 coordinates, gene names, and 100% conserved block sequences for all blocks longer than 60 bases identified in coding regions based on *D.melanogaster-D.virilis* alignments.

Supplementary Table S3. Ultraconserved ungapped blocks in unannotated sequences
Drosophila genome release 3.2 coordinates, block sequences for all blocks longer than 60 bases identified in unannotated regions based on *D.melanogaster-D.virilis* alignments.

Supplementary Table S4. Ultraconserved regions encoding micro RNA
Table shows names, coordinates (genome 4.0) and block sizes for micro RNAs overlapping ungapped blocks longer than 30 bases. Last column shows fraction of the ungapped blocks with the corresponding length and longer in base genome. *Drosophila* micro RNA were downloaded from miRBase {Griffiths-Jones, 2004 #12}. The blocks were extracted from *D.melanogaster-D.virilis* alignments (see supplementary **Table 2**).

References

Tables:

Table 1.

	D.yakuba	*D. ananassae*	*D.pseudoobscura*	*D. virilis*	*D. 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

Table 2.

	Prm	UTR	Exon	Intron	Unknown	All blocks
D.m.-D.yak.	7.72E-44	2.12E-29	4.52E-115	2.46E-29	7.09E-08	8.12E-24
D.m-D.ana	2.77E-47	4.70E-23	3.29E-179	2.22E-01	3.15E-02	6.99E-07
D.m.-D.pse.	7.69E-61	2.21E-35	1.73E-161	1.96E-03	4.57E-01	5.12E-08
D.m-D.vir	6.50E-05	8.97E-19	3.13E-88	5.15E-02	6.05E-02	2.90E-06
D.m-D.moj.	6.86E-03	3.02E-14	6.06E-64	6.48E-02	1.65E-02	2.51E-04

Table 3.

		Enc	Prm	5'UTR	Exon	Intron	3'UTR	Unknown
D.mel-D.yakuba	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.pseudo.	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.virilis								
	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