BMC Genomics

BioMed Central

Research article

Genomic characterization of a repetitive motif strongly associated with developmental genes in *Drosophila*

Javier Costas*[1,2], Cristina P Vieira[1], Fernando Casares[1] and Jorge Vieira[1]

Address: [1]Instituto de Biologia Molecular e Celular (IBMC), Universidade do Porto, Rua do Campo Alegre 823, 4150 Porto, Portugal and [2]Present address: Unidade de Medicina Molecular, Complexo Hospitalario Universitario de Santiago, rúa Choupana s/n, Edf. Consultas, planta -2, E15706 Santiago de Compostela, Spain

Email: Javier Costas* - bfcostas@usc.es; Cristina P Vieira - cgvieira@ibmc.up.pt; Fernando Casares - fcasares@ibmc.up.pt; Jorge Vieira - jbvieira@ibmc.up.pt

* Corresponding author

Published: 16 December 2003

BMC Genomics 2003, **4**:52

Received: 10 September 2003
Accepted: 16 December 2003

This article is available from: http://www.biomedcentral.com/1471-2164/4/52

Abstract

Background: Non-coding DNA represents a high proportion of all metazoan genomes. Although an undetermined fraction of this DNA may be considered devoid of any function, it also contains important information residing in specific *cis*-regulatory sequences.

Results: We report a 27 bp motif that is overrepresented within the fly genome. This motif does not show any significant similarity with transposon sequences and is strongly associated with genes involved in development and/or signal transduction. The 27 bp motif is preferentially located within introns, and has a tendency to be present in multiple copies around genes. Furthermore, it is often found embedded in known non-coding regulatory regions. The regulatory network defined by this motif is partially shared in *D. pseudoobscura*.

Conclusion: We have identified a 27 bp *cis*-regulatory sequence widely distributed within the *Drosophila* genome in association with developmental genes. This motif may be very useful towards the annotation of functional regulatory regions within the *Drosophila* genome and the construction of regulatory networks of *Drosophila* development.

Background

Coding regions constitute a small portion of metazoan genomes, representing ~24% of the small genome of *Drosophila melanogaster* and less than 2% of the larger human genome [1,2]. Although an unknown proportion of non-coding DNA might be regarded as "junk DNA", non-coding regions also include important information related to essential processes such as transcriptional and post-transcriptional regulation, splicing, higher-order chromatin structure and DNA replication. This information generally lays in specific DNA sequences located both in intergenic regions and introns. Nevertheless, this information remains largely inaccessible to the researchers, due to the reduced knowledge about structure and function of non-coding DNA.

Different approaches have been proposed to infer putative *cis*-regulatory regions. One strategy is based on the identification of overrepresented motifs in sets of coexpressed genes [3-5]. This approach requires prior data on gene expression of large number of genes, generally determined by microarray technology or expressed sequence tags (ESTs).

A second method to locate novel regulatory regions within the genome is the search for statistically

improbable concentration of putative binding sites for a transcription factor or a set of functionally related transcription factors. This method generates testable predictions about the function of the putative regulatory regions. For instance, identification of clusters of binding sites for *dorsal*, *dl*, and *Suppresor of Hairless*, *Su(H)*, have led to the identification of new regulatory regions within the *Drosophila* genome controlled by these genes [6,7]. In other cases, clustering of binding sites for different transcription factors, such as those active in early *Drosophila* development or those determining mesoderm activation also revealed new enhancers [8-10].

A third approach (evolutionary comparative approach) relies on the availability of full genome sequences of several eukaryotes, and is based on the fact that conservation of blocks of non-coding sequence between distantly related species is unlikely and thus implies functional constraint on the conserved blocks (called phylogenetic footprints) [11,12].

All of these approaches represent an essential contribution to one of the major goals in genome research, the construction of regulatory networks, consisting of the linkages between different *cis*-regulatory systems the genes they govern [13].

The *wingless (wg)* gene is a member of the *Wnt* gene family that encode for secreted glycoproteins, which act as key intercellular signaling molecules during animal development [14]. Although the mechanisms of *wg* signaling are beginning to be understood [15], much less is known about how the complex pattern of expression of *wg* is regulated.

While searching the *D. melanogaster wg* intron sequences for putative regulatory regions using an evolutionary comparative approach, we identified a 27 bp long motif that is overrepresented within the *D. melanogaster* genome and that is strongly associated with genes involved in development and/or signal transduction. This motif does not bear any similarity with any of the described *D. melanogaster* transposons. The gene network defined for *D. melanogaster* is partially present in *D. pseudoobscura*. This motif might prove useful in searching for new genes involved in *Drosophila* development, in genome annotation and in the construction of regulatory networks.

Results

Identification of a 27 bp long motif overrepresented in the fly genome

Two transcripts have been found for the *D. melanogaster wg* gene. The longer transcript codes for five exons, while the shorter one codes for only four exons. The 3' end of the first intron of the longer transcript is part of the 5' untranslated region (UTR) of the shorter transcript since the alternative *wg* start codon is located within the second exon of the longer transcript (Release 3.1 of the *D. melanogaster* genome, February 2003). Three out of the four introns of the longer transcript are large (more than 1 Kb long) considering that more than half of *D. melanogaster* introns are less than 80 nucleotides in length [16]. Regulatory sequences are often found within intron sequences (see for instance [17-19]). A detailed analysis of *wg* non-coding regions, including the intron regions, could therefore help understanding how *wg* expression is regulated.

In order to identify putative regulatory regions embedded in the *D. melanogaster wg* introns, we first identified the *D. pseudoobscura* contig that contains the *wg* orthologous intron sequences using BLAST search [20]. In contrast with *wg* intron 2, when the first and third introns were used as a query, many hits of 20 bp or longer were obtained in the *D. pseudoobscura* genome. Visual inspection of these sequences revealed that only hits generated by the first intron are not microsatellites. The conserved signal obtained using the *wg* intron 1 sequence was about 25–30 bp long. This is surprising since the fast turnover rate of *Drosophila* binding sites [21] means that it is unlikely that long motifs are shared between the genomes of species as distant as *D. melanogaster* and *D. pseudoobscura*. In fact, a search for additional dispersed repetitive sequences within the introns of 21 developmental genes (from table 1 of reference [22]) did not detect any sequence as long as this one. Since the *D. pseudoobscura* genome is unannotated and incomplete, this observation motivated us to perform BLAST searches against the *D. melanogaster* genome using as a query the first intron of the *D. melanogaster wg* gene. This led to the identification of the 25–30 bp motif in many regions of the fly genome (more than 300 hits). This motif does not show any significant similarity with the sequences deposited in the transposon database at the Berkeley *Drosophila* Genome Project [23]. Thus, the best hit (element jockey2) presents only 14/16 identities (E-value = 44). We also ruled out the possibility of the motif being a known microRNA, after the search of a database of published microRNAs [24] for sequences homologous to the motif yielded no positive result.

About 200 sequences, 100 bp long and centered around the motif were collected and aligned using the program diAlign [25]. The software diAlign is especially suitable to perform local multiple alignments to identify homologous stretches of DNA interspersed between sequences of no homology. For a contiguous stretch of 27 bp the most abundant nucleotide is always at a frequency higher than 50%, while elsewhere the frequency of the most frequent nucleotide is always lower than 50%. For 10 out of the 27 positions of the contiguous DNA stretch, the same

Table 1: Distribution of the motif on the different chromosomal arms

Chromosome arm	Observed motifs	Expected[a]
X	48	69.8
2L	71	70.5
2R	64	64.5
3L	85	74.6
3R	100	88.6
4	0	3.7

[a]Based on the total length of each chromosome arm [49].

nucleotide is present in more than 99% of the sequences and is therefore presumed to be critical for the function of the motif. The frequency of the most abundant nucleotide reaches more than 70% in all other positions except nucleotide 23, where C is present in ~55% of the sequences and T in the other 45%. For these positions, the frequency of the 2nd most abundant nucleotide is always lower than 25%. The length of the motif was thus established as 27 bp.

In order to locate all motifs in the *D. melanogaster* genome and to avoid the inclusion of false positives we decided to use an approach similar to the use of PWM to identify binding sites for transcription factors. BLAST E-values are not suitable when analyzing short sequences since they depend on the size of the retrieved sequence. Motifs with mismatches at the end of the sequence relative to the query's sequence will usually be retrieved as shorter sequences than motifs having the same total number of mismatches, but with the mismatches located internally. Therefore different E-values are going to be reported even though the two sequences have the same total number of mismatches relative to the query sequence.

The most important difference relative to the use of conventional PWM that we introduce is that we do not use the actual nucleotide frequencies to weight each position accordingly, since the lack of any functional information prevents us from selecting a specific subset of sequences to construct it. Standard computer software designed to identify putative binding sites for known transcription factor binding sites (such as MatInspector [26]) failed to identify any credible sites embedded in the 27 bp motif sequence (data not shown). The parameters of our *ad hoc* PWM were thus set to identify all the *D. melanogaster* sequences that match the consensus in those positions with the most frequent nucleotide appearing in more than 99% of the sampling sequences (but allowing C or T at position 23, see Material and Methods). We searched for all sequences differing from 0 to 8 nucleotides from the preliminary consensus sequence (based on the sequences retrieved from the BLAST search) at the other positions

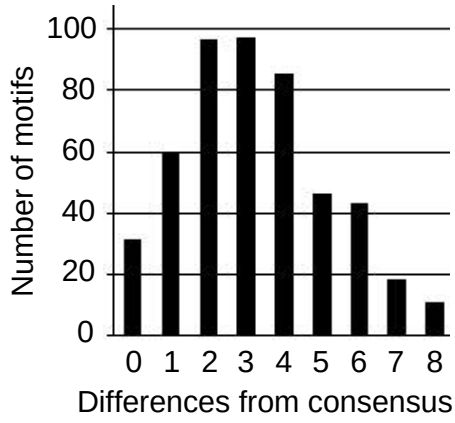

Figure 1
Number of motifs within the *D. melanogaster* genome (Y-axis) as function of the number of differences from consensus (X-axis).

using the server Target Explorer [27,28]. Rather than obtaining a raise in the number of targets as we increase the number of allowed differences (as expected by chance), the distribution shown in Fig. 1 reveals that the majority of target sequences present between 2 and 4 differences from consensus. This distribution strongly suggests a biological function for the sequence. According to this distribution, we set a conservative cut-off value of 4 differences to avoid the inclusion of putative false positives. A total of 368 sequences matched this criterion, representing 75.7% of the identified sequences. All the subsequent analyses were performed based on these 368 sequences, which we, therefore, expect to constitute a representative subset of all relevant sequences. The consensus sequence based on these 368 sequences is shown in Fig. 2, as a pictogram. This consensus sequence is identical to the preliminary consensus sequence (see above), being the relative nucleotide frequencies at each position very similar.

Figure 2
Pictogram of the sequence motif. The height of letters is proportional to their relative frequencies.

Using the same criterion, no matches were found in a set of 20 random sequences of 250000 bp with the same nucleotide composition as the *D. melanogaster* intergenic regions, while ~15 motifs were expected based on the proportion within the *Drosophila* genome (368 repeats / 120 Mb). Thus, the repeat is significantly highly overrepresented within the *Drosophila* genome.

Distribution of the sequence motif

As shown in Table 1, the 27 bp motif is present in all chromosome arms but the small chromosome 4. Nevertheless, this distribution departs from the random expectation based on the total length of each chromosome arm (χ^2 = 13.433, 5df, P < 0.0196), due mainly to an underrepresentation of the motif in the X chromosome, coupled to an overrepresentation on both arms of chromosome 3.

The location of the 27 bp motif relative to *D. melanogaster* genes is shown in Fig. 3. There are 125 motifs within introns and 234 in intergenic regions. Seven motifs are located within the 5' UTR of genes, one within the 3' UTR and one within a coding region of *Spatzle3 (Spz3)*. These numbers should be taken with some caution as some currently annotated intergenic regions may be found to be the introns of larger transcripts. Furthermore, some currently annotated intron regions may contain alternative spliced exons, as well as small nested genes yet to be recognized. At present, there is thus no good approximation for the actual figures on the proportion of intronic and intergenic regions [2]. Therefore, the numbers previously reported (~20 Mb of intron sequences versus ~76 Mb of intergenic sequences [29]) are likely not correct, but can be used as an approximation. The strong deviation from the null hypotheses of identical distribution of sites in intronic versus intergenic regions (χ^2 = 42.557. 1 df, P < 0.0001) suggests, nevertheless, the existence of a significant overrepresentation of the sequence within introns. While 173 of the intergenic motifs are located upstream of the nearest gene, only 61 are downstream (Fig. 3). These values significantly depart from an equal proportion of motifs 5' and 3' of the nearest gene (χ^2 = 43.834, 1 df, P < 0.0001). Interestingly, 20% of the motifs located upstream of genes are within the first 1000 bp from the transcription start.

The distribution of distances between consecutive motifs (Fig. 4) reveals a clear trend for the motif to form clusters. 78.5% of the motifs are included in clusters of at least two motifs within 50 kb, while the proportion detected in the *Drosophila* genome is one motif per ~326 Kb. There are 16 clusters of at least two motifs within 1000 bp. We also analyzed the existence of clusters based on the association with genes, rather than by distance (Fig. 5). Approximately 37% of the genes associated with the repeat present more than one repeat around/within the gene. The proportion of motifs belonging to these clusters around/within genes reaches ~62% of the total number of motifs. We used the Gene Ontology (GO) classification [30] to search for any bias in the molecular function or biological process of genes associated with the motif,

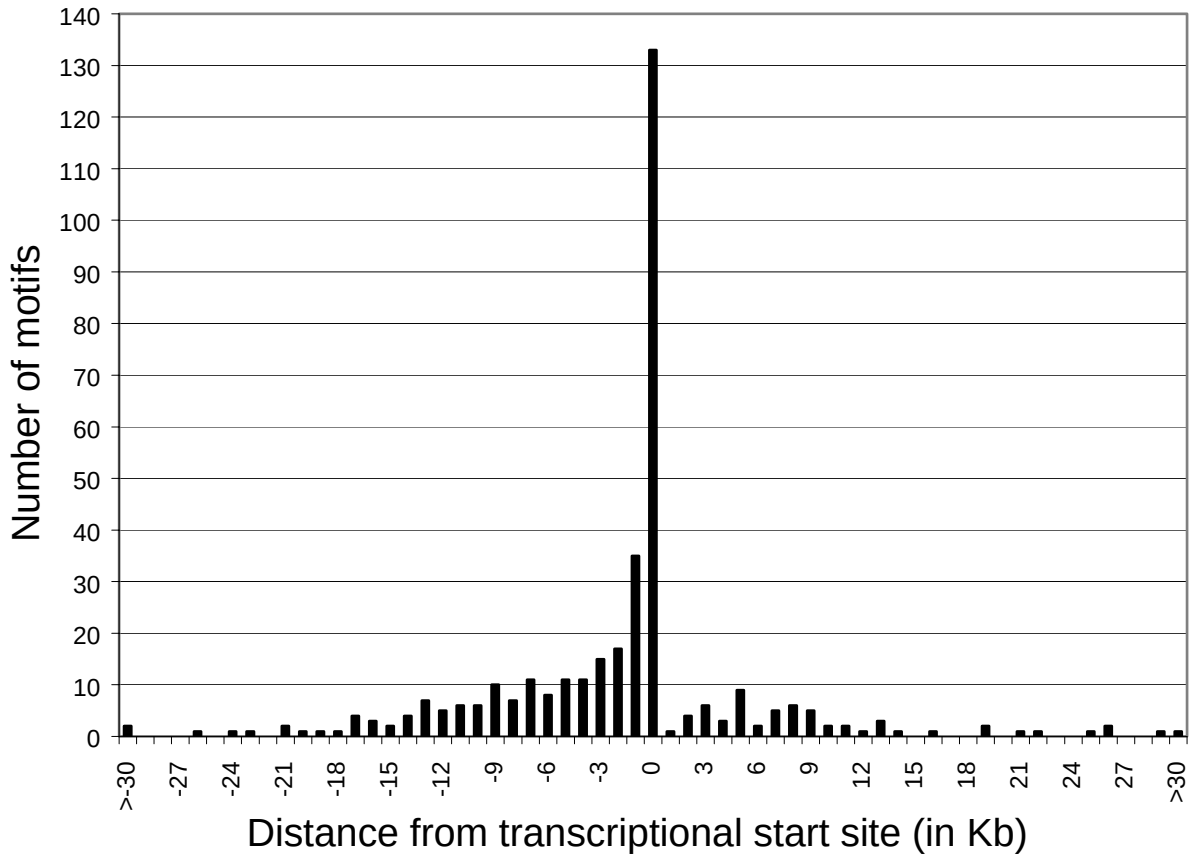

Figure 3
Number of motifs according to their distance from transcriptional start site. Distances in kb. "0" indicates motifs within the gene, including UTRs, introns, and exons (see text).

using the Target Explorer web server. We performed three different analyses. In the first one, all the motifs were included. Bearing in mind that an important proportion of motifs forms clusters, we reanalyzed our data set selecting only one motif per cluster. When a repeat is located in an intergenic region, both genes around the motif are considered in these two analyses, leading to an underestimation of the actual bias towards genes involved in particular processes. Because of that, we performed a third analysis, including those genes with at least one motif within introns, so the association motif-gene is unambiguous. Since each analysis comprises a fraction of the motifs included in the previous one, we expect a concomitant reduction in the power of the test due to the smaller sampling size.

As shown in Table 2, the 27 bp motif is significantly associated with genes whose molecular function is related to signal transduction and/or transcriptional regulation. In regard to the biological process, there is a significant overrepresentation of genes involved in development, and a significant underrepresentation of genes related to cell growth and/or maintenance. For instance, only 6.3% of all genes in the Gene Ontology database are associated with the transcriptional regulation category. Therefore, under the null hypothesis that the 27 bp motif is associated with genes regardless of their molecular function, we expect that approximately 6.3% of the genes in our sample belong to the transcriptional regulation category. If we consider the subset of genes that have the 27 bp motif located in the middle of intron sequence, as much as 14.9% of the genes are associated with the transcriptional

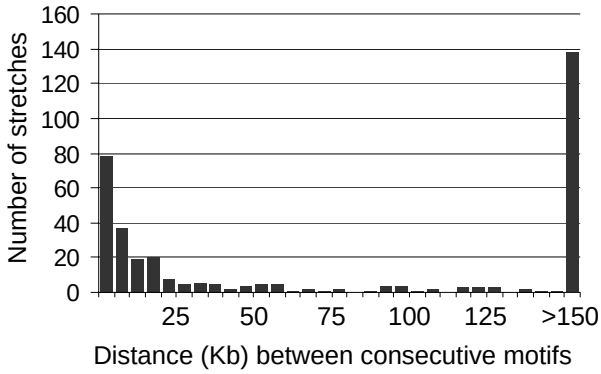

Figure 4
Distribution of distances between consecutive motifs. Range of each distance class, 5000 bp.

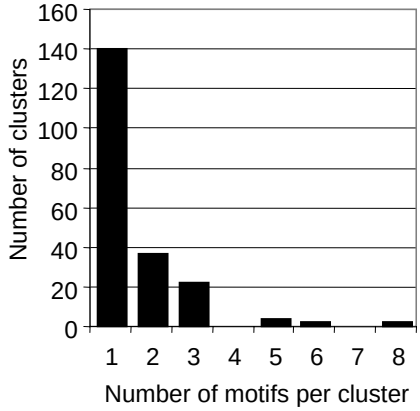

Figure 5
Clusters of sequence motifs. The X-axis represents the different clusters based on the number of motifs they contain, while the Y-axis shows the number of clusters for each class.

regulation category. The difference of the expected and observed proportion is highly significant ($P < 0.005$).

There are 102 well-known genes putatively associated with the repeat, considering both genes around intergenic motifs. Fifty-four of these genes are included in at least one of the overrepresented categories of the Gene Ontology tree. See additional file 1 for the complete list of genes putatively associated with the repeat.

Conservation of the sequence motif in other species

In order to infer general facts about the evolution of the motif, we searched for its presence in the genome of *D. pseudoobscura*, a species with an ongoing genome project [31]. We located the orthologous region of 322 out of the 368 motifs detected in *D. melanogaster*, amounting to 87.5% of the sequences. Using the same criterion as in *D. melanogaster*, we identified 178 conserved motifs within the orthologous region of *D. pseudoobscura*, accounting for 55.3% of the 322 sequences.

If the information associated with each one of the motifs from the same cluster is at least partially redundant, those motifs not belonging to clusters are expected to be more constrained than the clustered ones. Nevertheless, we identified only 62 motifs in *D. pseudoobscura* within the 120 orthologous regions with motif not associated in clusters in *D. melanogaster*. These values do not differ significantly from the null hypothesis of equal degree of conservation in both subsets ($\chi^2 = 1.010$, 1 df, $P < 0.3150$).

We further analyzed the 178 motif pairs to characterize the selective constraint. If the degree of constraint to maintain the similarity to consensus is similar in the two species, we expect a correlation between the number of differences from consensus in each one of the orthologous sequences. If one motif is more constrained in one of the species, this correlation is expected to be lost. The number of pairs with identical number of differences from consensus is significantly higher than expected under the hypothesis of no correlation (57 *vs* 37.76, $\chi^2 = 12.437$, 1 df, $P < 0.0004$), indicative of the action of selection to maintain the degree of divergence (Table 3).

Selection may be acting mainly to maintain the relative affinity (roughly approximated as the number of differences from the consensus) or/and to maintain the exact sequence of the motif in any particular context. Fig. 6 shows the observed and expected number of differences at each position of the motif between the 178 orthologous pairs between species, based on the nucleotide frequencies at each position within each one of the species. In all positions, the number of motif pairs with nucleotide difference is lower than expected. This difference is significant in most cases (Fig. 6), indicating that selection acts to maintain the appropriate motif variant at each particular location.

We also searched for the presence of the motif in other species using BLAST search. We identified one sequence within the first intron of the gene *Om(1D)* (Accession No. X56682) from *D. ananassae*, species that belongs also to the subgenus Sophophora (as *D. melanogaster* and *D. pseudoobscura*). This gene is the orthologous to the *D. mela-*

Table 2: Gene Ontology classification of genes associated with the motifs

GO classification[a]	*Drosophila* genome	All motifs	One repeat per cluster	Introns
Molecular function				
binding	2301 (27.2%)	95 (30.8%)	56 (28.4%)	19 (28.4%)
enzyme	3006 (35.5%)	90* (29.2%)	58 (29.4%)	17 (25.4%)
signal transducer	730 (8.6%)	44*** (14.3%)	24 (12.2%)	12* (17.9%)
transcriptional regulation	532 (6.3%)	38*** (12.3%)	28*** (14.2%)	10** (14.9%)
transporter	763 (9.0%)	16* (5.2%)	15 (7.6%)	4 (6.0%)
Total[b]	8463	308	197	67
Biological process				
cell communication	854 (20.2%)	58** (29.4%)	32 (27.1%)	12 (26.1%)
cell growth/maintenance	2439 (57.8%)	59*** (29.9%)	39*** (33.1%)	13*** (28.3%)
development	734 (17.4%)	69*** (35.0%)	41*** (34.7%)	19*** (41.3%)
Total[b]	4220	197	118	46

χ^2 test of independence: * P < 0.05, ** P < 0.005, *** P < 0.0005. [a]Only those categories containing more than 5% of the annotated genes are shown. [b]The total number of classifications is greater than the total number of genes since each gene is usually classified under different categories. This number is used to calculate the proportions showed between parentheses.

Table 3: Correlation between values for differences from consensus between *D. melanogaster* and *D. pseudoobscura*

D. pseu\D. mel[a]	0	1	2	3	4	total
0	6 (1,41)	1 (2,70)	3 (3,57)	1 (2,29)	1 (2,02)	12
1	4 (3,77)	12 (7,19)	9 (9,53)	4 (6,11)	3 (5,39)	32
2	5 (4,95)	13 (9,44)	11 (12,50)	8 (8,02)	5 (7,08)	42
3	5 (6,02)	10 (11,46)	21 (15,18)	11 (9,74)	4 (8,59)	51
4	1 (4,84)	4 (9,21)	9 (12,21)	10 (7,83)	17 (6,91)	41
total	21	40	53	34	30	178

[a]Differences from consensus for each sequence pair: row, *D. melanogaster*; column, *D. pseudoobscura*. Expected values (between parentheses) are based on the proportion of each category in each species.

nogaster Bar-H1 (*B-H1*) that also presents the motif in orthologous location. We also identified the motif in two genes of *D. virilis*, from the Drosophila subgenus. One copy is located 5' from the *actin E2* gene (Accession No. AF358263). The orthologous sequence in *D. melanogaster* also contains a sequence equivalent to those of the motif, but presenting 5 differences from the consensus. The other copy located in *D. virilis* is present within the enhancer region of the *achaete-scute* (*ac-sc*) complex (Accession No. AF132809). The orthologous sequences in *D. melanogaster* and *D. pseudoobscura* lacked the motif. The 27 bp motif seems not to be present in the *Anopheles gambiae* genome. BLAST searches against the *A. gambiae* genome using as a query the first intron of the *A. gambiae wg* gene does not retrieve any sequence with similar characteristics to the *Drosophila* 27 bp motif (data not shown).

Discussion

Most essential processes, such as transcriptional and post-transcriptional regulation, DNA replication, or higher-order chromatin structure, are under the control of *cis*-acting elements located within non-coding DNA (see for instance, [32-34]). Here, we describe a 27 bp long repetitive DNA sequence within the *Drosophila* genome that, based on its characteristics, may be considered one of these *cis*-acting elements.

Mainly, there is a significant association of the 27 bp long motif to genes involved in development, whose molecular function is related to signal transduction and/or transcriptional regulation (Table 2). This association may be indeed stronger than shown in Table 2, taking into account that any given gene is usually classified under several different categories of the Gene Ontology classification. For instance, although only 41.3% of the biological process classifications of the 22 genes with the motif present within an intron are annotated as involved in development (Table 2), 19 of them (~86%) are indeed known to be involved in development.

Several components of main signaling pathways are associated with the motif, such as: *Delta* (*Dl*) and *Serrate* (*Ser*)

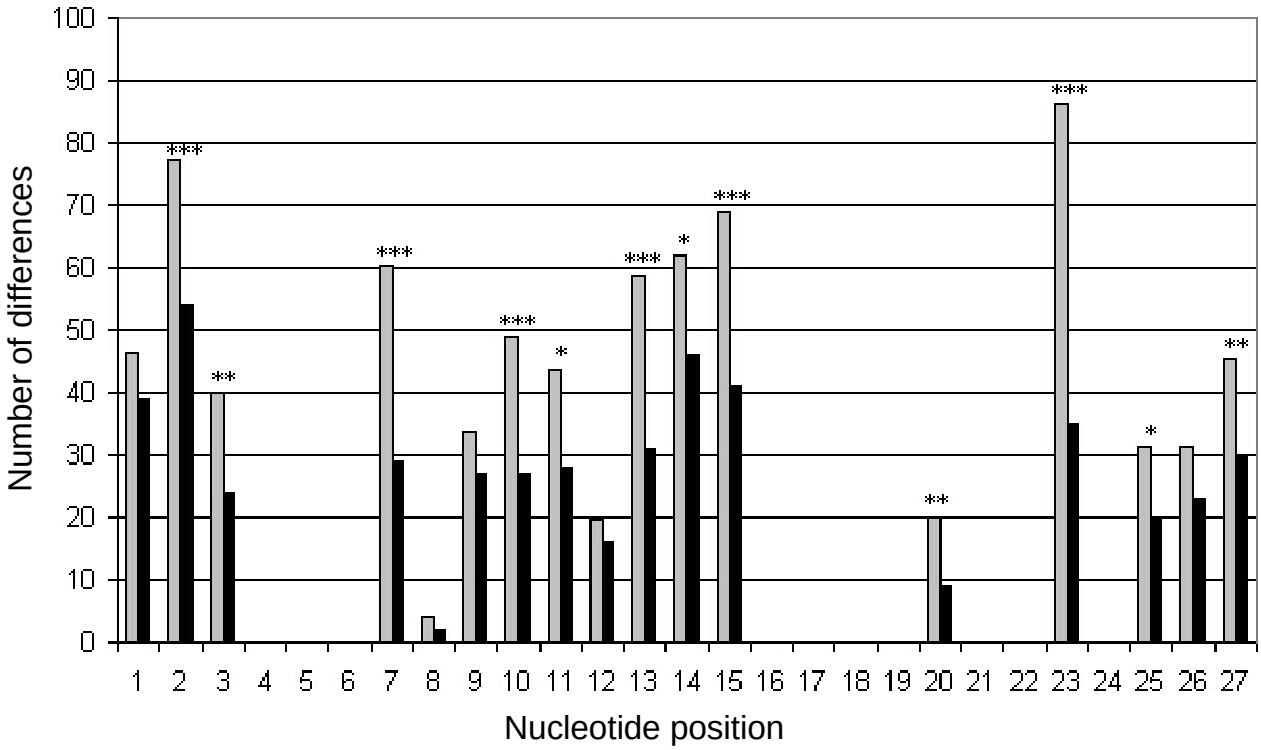

Figure 6
Number of expected (gray) and observed (black) differences at each nucleotide position for the 178 orthologous motif-pairs between *D. melanogaster* and *D. pseudoobscura*. Statistically significant differences are shown (* P < 0.05, ** P < 0.005, *** P < 0.0005).

(ligands), *Notch* (*N*) (receptor) and *Su(H)* (nuclear transducer) of the *N* signaling pathway; *wg* (ligand) and *frizzled3* (*fz3*) (receptor) of the *Wnt* signaling pathway; *Epidermal growth factor receptor* (*Egfr*) (receptor) and *vein* (*vn*) (ligand) of the *Egfr* signaling pathway; *hedgehog* (*hh*) (ligand) of the *hh* signaling pathway; or *decapentaplegic (dpp)* (ligand) of the *TGF*-β receptor signaling pathway. There are also several selector or selector-like genes [35], conserved transcription factors that act controlling the development of morphogenetic fields giving rise to specific adult structures, such as *twist* (*twi*) in mesoderm tissues, *Distal-less* (*Dll*) in the ventral appendages, *pannier* (*pnr*) in dorso-medial domains of trunk and head, *brachyenteron* (*byn*) in posterior terminal structures, *engrailed* (*en*) in posterior compartments, and *B-H1* and *Bar-H2* (*B-H2*) in neural tissues (see FlyBase [36] for a description of these genes' function, plus references therein). Thus, this motif may define an important regulatory network, linking together several fundamental genes active during *Drosophila* development (Table 2 and Additional file 1).

Second, our strategy to search for the conservation of the motif in *D. pseudoobscura* (see Methods) revealed that 70% of the regions around the motif in *D. melanogaster* present at least 70 identical nucleotides out of 100 bp of sequence in *D. pseudoobscura*. A recent comparison of non-coding regions between *D. melanogaster* and *D. pseudoobscura* [12] revealed that only 28% of the non-coding sequences are conserved between these two species. The conserved non-coding sequences (defined as windows of at least 10 bp with at least 90% of nucleotide identity) tend to be spatially clustered. Therefore, these data strongly indicate that

the motif described in this paper is generally located within regulatory regions of genes.

In agreement with this prediction, several copies of the motif are located within known regulatory regions. Thus, the six motifs associated to *dpp* (Additional file 1) are located within the 3' "disk region" of the gene, an enhancer controlling the expression of *dpp* in imaginal discs [37]. Two motifs located between 10 and 15 Kb upstream of *Ser* are included within the "dorsal wing regulator" enhancer (DWR), which directs the expression of *Ser* in the wing disc [38]. The motif associated to *Su(H)* is located within the autoregulatory socket enhancer (ASE), a discrete cell specific transcriptional enhancer active only in the socket cells of external sensory organs [39]. This enhancer is regulated by the *Su(H)*'s own protein product, containing eight predicted high-affinity binding sites for the Su(H) protein. The motif is embedded within these binding sites. In contrast to the previous examples, the regulatory sequences of *Dl* are dispersed over a large stretch of DNA rather than being concentrated in discrete zones. The first intron of *Dl*, which presents one motif, contains a quantitative enhancer of transcription acting on several different organs [40].

Several characteristics of the motif, such as its trend to form clusters within/around genes (Fig. 5) or its biased location in regard to the transcription units (Fig. 3), might be a consequence of its association with regulatory regions of genes associated with signal transduction pathways, and transcription factors involved in several developmental processes. In general, these genes present several independent enhancers located not only upstream, but also downstream or in intronic regions. The modularity of the enhancer architecture [13] is in agreement with our observation of similar constraints acting on those motifs belonging to clusters and the remaining, non-clustered, motifs.

In a similar way, one could image that the underrepresentation of the motif in the X chromosome might be due to a biased distribution of developmental genes on this chromosome. Nevertheless, this does not seem to be the case, since137 of the 681 fly genes associated with the GO term "development", and whose chromosomal locations are known, lay on the X chromosome, slightly over the expected value of 128, estimated based on the sequence length of each chromosomal arm (χ^2 test; $P > 0.05$). Alternatively the explanation for the underrepresentation of the motif in the X chromosome might be related to the lower effective population size of this chromosome when compared to autosomes (3/4 that of autosomes). Because of that, according to the nearly neutral theory of molecular evolution (see [41] for a recent review), natural selection is expected to be less efficient to create and/or maintain this sequence motif in the X chromosome.

It should be noted that although this 27 bp motif does not show any significant similarity with any of the transposable elements listed in the transposon database at the Berkeley *Drosophila* Genome Project [23], we cannot rule out a transposable element origin for this motif. All transposons are known to be underrepresented on the X chromosome relative to the autosomes [42]. The underrepresentation of the 27 bp motif in the X chromosome could thus simply reflect its origin. It has been suggested that in an unknown proportion of cases transposons might be a source of "ready-to-use" regulatory motifs [43,44].

The comparative genomic approach revealed that the regulatory network defined by this motif in *D. melanogaster* is partially shared with *D. pseudoobscura*. Furthermore, conserved motifs seem to be constrained to maintain not only location but also the exact sequence variant at each particular position (Table 3 and Fig. 6), as described previously in the case of binding sites for transcription factors envolved in early Drosophila development [21]. Although the early stage of the *D. pseudoobscura* genome project precludes a full comparison, our results indicate that more than half of the motifs are conserved between the two species. Two facts suggest that this figure might underestimate the actual number of conserved sequences. First, while the consensus sequence for the motif is identical in both species, the inferred PWM might be different, as we used the PWM derived from *D. melanogaster* to classify a sequence from *D. pseudoobscura* as matching the motif. In fact, the average number of differences between the 178 motifs of *D. melanogaster* whose orthologous has been identified in *D. pseudoobscura* and the consensus sequence is 2.07, while this figure reaches 2.43 in the case of the *D. pseudoobscura* motifs. Second, the cut-off score (allowing for a maximum of four differences from consensus in the changing positions) might be too stringent, leading to the detection of only high affinity binding sites-containing motifs conserved in both species. In fact, we found several cases of orthologous regions in *D. pseudoobscura* containing a sequence that differs from the consensus in only 5 or 6 differences, but that were, nevertheless, excluded from our data set for further analysis. A similar situation occurs in the case of the *actin E2* from *D. virilis*, whose orthologous sequence in *D. melanogaster* presents 5 differences from consensus.

The detection of the motif in other *Drosophila* species from the *Drosophila* subgenus shows that this motif arose within the genome before the radiation of the genus. Its absence in *Anopheles* is expected, taking into account that

only a very small proportion of regulatory regions are conserved between these two genera of Diptera [12,45].

Finally, it is interesting to remark that one concern of *cis*-regulatory prediction algorithms is the rate of false positives [9,10]. This problem is not present in the case of the motif described here due to its unusually long length compared to other regulatory motifs, which makes its appearance by chance highly improbable. This characteristic and the others discussed previously, makes this motif very useful towards the annotation of functional regulatory regions within the *Drosophila* genome and the construction of regulatory networks of *Drosophila* development. It may also be useful for inferring the function of a number of genes that show no similarity with other known genes. Functional tests will be required to characterize the function of this motif.

Conclusions

We have identified a *cis*-regulatory sequence motif widely distributed within the *Drosophila* genome in association with genes involved in development and/or signal transduction. Due to the unusual long size of this motif (27 bp) in comparison with other regulatory motifs, its appearance by chance is highly improbable. Because of that, this motif may be very useful towards the annotation of functional regulatory regions within the *Drosophila* genome as well as the construction of regulatory networks of *Drosophila* development.

Methods

BLAST searches and sequence alignments

BLAST searches were conducted using the BLAST server from NCBI [46] and the BLAST server from the *D. pseudoobscura* Genome Project [20]. The program diAlign [25] was used to perform local multiple alignments to identify homologous stretches of DNA interspersed between sequences of no homology.

Identification and location of a 27 bp motif in the D. melanogaster *genome*

A strategy similar to the use of position weigth matrices (PWMs) for the identification of binding sites for transcription factors was used to search for the presence of a 27 bp motif in the *D. melanogaster* genome previously identified by BLAST searches. It should be noted, however, that in our case we do not use binding affinity/functional information on the observed nucleotide frequencies to weight each position accordingly. We use the web server Target Explorer [27,28]; that easily allows for the editing of the PWM using the following general rules. In positions where the most frequent nucleotide appears in => 90% of the sampling sequences, any non-matching nucleotide was weighted very negatively (-20), while a matching nucleotide is given a +1 weight. In the remaining positions, a weight of +1 was given to the nucleotide present in the preliminary consensus and a weight of -1 was given to the other nucleotides. In the case of nucleotide position 23, where C and T seem to be equally used, the presence of either nucleotide was weighted as +1. To identify those sequences differing from consensus in less than 5 differences, for instance, the cut-off score was set as +18. This approach completely excludes any sequence that differs at any one of the almost invariant positions. The identification of the motifs was performed using the release 3 of the *D. melanogaster* genome and the program Patser [47] as implemented by Target Explorer. Both the location of the repeats in regard to the nearest transcription units and the classification of associated genes according to Gene Ontology categories were also performed by Target Explorer.

Generation of random sequences was done by the random generator tool available at the Regulatory Sequence Analysis Tools web page [48]. This program generates random sequences with the same oligonucleotide composition as observed in the intergenic regions of the selected organism (*D. melanogaster*) by a Markov chain probabilistic model.

Identification of D. pseudoobscura *genomic regions orthologous to those of* D. melanogaster *containing the 27 bp motif*

In order to identify the *D. pseudoobscura* genomic regions orthologous to those of *D. melanogaster* presenting the motif, we employed two different approaches: (1) BLAST searches against the *D. pseudoobscura* sequencing reads from the *D. pseudoobscura* Genome Project web page [20] using as a query each one of the motifs identified in *D. melanogaster* surrounded by 300 bp flanking sequences upstream and downstream from the motif. We considered a *D. pseudoobscura* sequencing read as the orthologous one if there was at least a stretch of 70/100 identical nucleotides. We used a word size of 7 nucleotides and a percent identity of 70%. If the orthologous region is identified but the sequencing read does not contain the motif, we search the *D. pseudoobscura* genomic region in between conserved orthologous blocks flanking the motif. To do so, we search for the contig containing this sequence and align this region with the *D. melanogaster* region using diAlign [25]. (2) If no orthologous region is identified according to the previous criterion and the 27 bp motif is known to be within intron sequence in *D. melanogaster*, we searched for the *D. pseudoobscura* orthologous region using the corresponding *D. melanogaster* whole transcript. In the case of intergenic regions, we considered only those sequence contigs that include both genes around the motif, with one exception; if the motif is present close to the transcript (<1000 bp) we analyzed 10000 bp of the corresponding *D. pseudoobscura* orthologous region.

Authors' contributions

JC conceived this study, participated in its design, carried out most analyses and drafted the manuscript. CPV and JV participated in the identification of the motif in *D. pseudoobscura*. CPV, FC and JV participated in the design of the study and contributed in the final stages of manuscript preparation. All authors read and approved the final manuscript.

Additional material

Additional File 1

List of genes putatively associated with the repeat. [a]Position relative to the transcriptional start site. Positions in italics indicate motifs that are located within the primary transcript. Genes included in at least one of the overrepresented categories (signal transducer, transcriptional regulation) of the Gene Ontology tree are marked (*).

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2164-4-52-S1.xls]

Acknowledgements

This work was supported by grant POCTI/37402 from Fundação para a Ciência e a Tecnologia (FCT), Portugal to FC; JC is a recipient of the postdoctoral fellowship SFRH/BPD/7094/2001 from FCT. CPV is a recipient of the postdoctoral fellowship SFRH/BPD/5592/2001 from FCT. FC is an EMBO Young Investigator and Leukemia and Lymphoma Society Special Fellow.

References

1. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor Miklos GL, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, Levine AJ, Roberts RJ, Simon M, Slayman C, Hunkapiller M, Bolanos R, Delcher A, Dew I, Fasulo D, Flanigan M, Florea L, Halpern A, Hannenhalli S, Kravitz S, Levy S, Mobarry C, Reinert K, Remington K, Abu-Threideh J, Beasley E, Biddick K, Bonazzi V, Brandon R, Cargill M, Chandramouliswaran I, Charlab R, Chaturvedi K, Deng Z, Di Francesco V, Dunn P, Eilbeck K, Evangelista C, Gabrielian AE, Gan W, Ge W, Gong F, Gu Z, Guan P, Heiman TJ, Higgins ME, Ji RR, Ke Z, Ketchum KA, Lai Z, Lei Y, Li Z, Li J, Liang Y, Lin X, Lu F, Merkulov GV, Milshina N, Moore HM, Naik AK, Narayan VA, Neelam B, Nusskern D, Rusch DB, Salzberg S, Shao W, Shue B, Sun J, Wang Z, Wang A, Wang X, Wang J, Wei M, Wides R, Xiao C, Yan C, Yao A, Ye J, Zhan M, Zhang W, Zhang H, Zhao Q, Zheng L, Zhong F, Zhong W, Zhu S, Zhao S, Gilbert D, Baumhueter S, Spier G, Carter C, Cravchik A, Woodage T, Ali F, An H, Awe A, Baldwin D, Baden H, Barnstead M, Barrow I, Beeson K, Busam D, Carver A, Center A, Cheng ML, Curry L, Danaher S, Davenport L, Desilets R, Dietz S, Dodson K, Doup L, Ferriera S, Garg N, Gluecksmann A, Hart B, Haynes J, Haynes C, Heiner C, Hladun S, Hostin D, Houck J, Howland T, Ibegwam C, Johnson J, Kalush F, Kline L, Koduru S, Love A, Mann F, May D, McCawley S, McIntosh T, McMullen I, Moy M, Moy L, Murphy B, Nelson K, Pfannkoch C, Pratts E, Puri V, Qureshi H, Reardon M, Rodriguez R, Rogers YH, Romblad D, Ruhfel B, Scott R, Sitter C, Smallwood M, Stewart E, Strong R, Suh E, Thomas R, Tint NN, Tse S, Vech C, Wang G, Wetter J, Williams S, Williams M, Windsor S, Winn-Deen E, Wolfe K, Zaveri J, Zaveri K, Abril JF, Guigo R, Campbell MJ, Sjolander KV, Karlak B, Kejariwal A, Mi H, Lazareva B, Hatton T, Narechania A, Diemer K, Muruganujan A, Guo N, Sato S, Bafna V, Istrail S, Lippert R, Schwartz R, Walenz B, Yooseph S, Allen D, Basu A, Baxendale J, Blick L, Caminha M, Carnes-Stine J, Caulk P, Chiang YH, Coyne M, Dahlke C, Mays A, Dombroski M, Donnelly M, Ely D, Esparham S, Fosler C, Gire H, Glanowski S, Glasser K, Glodek A, Gorokhov M, Graham K, Gropman B, Harris M, Heil J, Henderson S, Hoover J, Jennings D, Jordan C, Jordan J, Kasha J, Kagan L, Kraft C, Levitsky A, Lewis M, Liu X, Lopez J, Ma D, Majoros W, McDaniel J, Murphy S, Newman M, Nguyen T, Nguyen N, Nodell M, Pan S, Peck J, Peterson M, Rowe W, Sanders R, Scott J, Simpson M, Smith T, Sprague A, Stockwell T, Turner R, Venter E, Wang M, Wen M, Wu D, Wu M, Xia A, Zandieh A, Zhu X: **The sequence of the human genome.** *Science* 2001, **291:**1304-1351.
2. Misra S, Crosby MA, Mungall CJ, Matthews BB, Campbell KS, Hradecky P, Huang Y, Kaminker JS, Millburn GH, Prochnik SE, Smith CD, Tupy JL, Whitfied EJ, Bayraktaroglu L, Berman BP, Bettencourt BR, Celniker SE, de Grey AD, Drysdale RA, Harris NL, Richter J, Russo S, Schroeder AJ, Shu SQ, Stapleton M, Yamada C, Ashburner M, Gelbart WM, Rubin GM, Lewis SE: **Annotation of the *Drosophila melanogaster* euchromatic genome: a systematic review.** *Genome Biol* 2002, **3(12):**RESEARCH0083.
3. Roth FP, Hughes JD, Estep PW, Church GM: **Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation.** *Nat Biotechnol* 1998, **16:**939-945.
4. Stathopoulos A, Van Drenth M, Erives A, Markstein M, Levine M: **Whole-genome analysis of dorsal-ventral patterning in the Drosophila embryo.** *Cell* 2002, **111:**687-701.
5. Zhang Y, Ma C, Delohery T, Nasipak B, Foat BC, Bounoutas A, Bussemaker HJ, Kim SK, Chalfie M: **Identification of genes expressed in *C. elegans* touch receptor neurons.** *Nature* 2002, **418:**331-335.
6. Markstein M, Markstein P, Markstein V, Levine MS: **Genome-wide analysis of clustered Dorsal binding sites identifies putative target genes in the Drosophila embryo.** *Proc Natl Acad Sci USA* 2002, **99:**763-768.
7. Rebeiz M, Reeves NL, Posakony JW: **SCORE: a computational approach to the identification of cis-regulatory modules and target genes in whole-genome sequence data Site clustering over random expectation.** *Proc Natl Acad Sci USA* 2002, **99:**9888-9893.
8. Berman BP, Nibu Y, Pfeiffer BD, Tomancak P, Celniker SE, Levine M, Rubin GM, Eisen MB: **Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome.** *Proc Natl Acad Sci USA* 2002, **99:**757-762.
9. Halfon MS, Grad Y, Church GM, Michelson AM: **Computation-based discovery of related transcriptional regulatory modules and motifs using an experimentally validated combinatorial model.** *Genome Res* 2002, **12:**1019-1028.
10. Rajewsky N, Vergassola M, Gaul U, Siggia ED: **Computational detection of genomic cis-regulatory modules applied to body patterning in the early Drosophila embryo.** *BMC Bioinformatics* 2002, **3:**30.
11. International Human Genome Sequence Consortium: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409:**860-921.
12. Bergman CM, Pfeiffer BD, Rincon-Limas DE, Hoskins RA, Gnirke A, Mungall CJ, Wang AM, Kronmiller B, Pacleb J, Park S, Stapleton M, Wan K, George RA, de Jong PJ, Botas J, Rubin GM, Celniker SE: **Assessing the impact of comparative genomic sequence data on the functional annotation of the Drosophila genome.** *Genome Biol* 2002, **3:**RESEARCH0086.
13. Arnone MI, Davidson EH: **The hardwiring of development: organization and function of genomic regulatory systems.** *Development* 1997, **124:**1851-1864.
14. Wodarz A, Nusse R: **Mechanisms of Wnt signaling in development.** *Annu Rev Cell Dev Biol* 1998, **14:**59-88.
15. Hlsken J, Behrens J: **The Wnt signalling pathway.** *J Cell Sci* 2000, **113:**3545-3546.
16. Mount SM, Burks C, Hertz G, Stormo GD, White O, Fields C: **Splicing signals in Drosophila: intron size, information content, and consensus sequences.** *Nucleic Acids Res* 1992, **20:**4255-4262.
17. Stanewsky R, Lynch KS, Brandes C, Hall JC: **Mapping of elements involved in regulating normal temporal period and timeless RNA expression patterns in *Drosophila melanogaster*.** *J Biol Rhythms* 2002, **17:**293-306.
18. Friggi-Grelin F, Coulom H, Meller M, Gomez D, Hirsh J, Birman S: **Targeted gene expression in Drosophila dopaminergic cells**

using regulatory sequences from tyrosine hydroxylase. *J Neurobiol* 2003, **54:**618-627.
19. van Steensel B, Delrow J, Bussemaker HJ: **Genomewide analysis of Drosophila GAGA factor target genes reveals context-dependent DNA binding.** *Proc Natl Acad Sci USA* 2003, **100:**2580-2585.
20. **BLAST against Baylor *D. pseudoobscura* data** [http://www.hgsc.bcm.tmc.edu/blast/?organism=Dpseudoobscura]
21. Costas J, Casares F, Vieira J: **Turnover of binding sites for transcription factors involved in early Drosophila development.** *Gene* 2003, **310:**215-220.
22. Bergman C, Kreitman M: **Analysis of conserved noncoding DNA in *Drosophila* reveals similar constraints in intergenic and intronic sequences.** *Genome Res* 2001, **11:**1335-1345.
23. **Berkeley Drosophila Genome Project** [http://www.fruitfly.org]
24. **The miRNA Registry** [http://www.sanger.ac.uk/Software/Rfam/mirna/index.shtml]
25. Morgenstern B: **DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment.** *Bioinformatics* 1999, **15:**211-218.
26. **Genomatix: MatInspector** [http://www.genomatix.de/software_services/software/MatInspector/MatInspector_stb.html]
27. **Target Explorer** [http://trantor.bioc.columbia.edu/Target_Explorer]
28. Sosinsky A, Bonin CP, Mann RS, Honig B: **Target Explorer: an automated tool for the identification of new target genes for a specified set of transcription factors.** *Nucleic Acids Res* 2003, **31:**3589-3592.
29. Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF, George RA, Lewis SE, Richards S, Ashburner M, Henderson SN, Sutton GG, Wortman JR, Yandell MD, Zhang Q, Chen LX, Brandon RC, Rogers YH, Blazej RG, Champe M, Pfeiffer BD, Wan KH, Doyle C, Baxter EG, Helt G, Nelson CR, Gabor GL, Abril JF, Agbayani A, An HJ, Andrews-Pfannkoch C, Baldwin D, Ballew RM, Basu A, Baxendale J, Bayraktaroglu L, Beasley EM, Beeson KY, Benos PV, Berman BP, Bhandari D, Bolshakov S, Borkova D, Botchan MR, Bouck J, Brokstein P, Brottier P, Burtis KC, Busam DA, Butler H, Cadieu E, Center A, Chandra I, Cherry JM, Cawley S, Dahlke C, Davenport LB, Davies P, de Pablos B, Delcher A, Deng Z, Mays AD, Dew I, Dietz SM, Dodson K, Doup LE, Downes M, Dugan-Rocha S, Dunkov BC, Dunn P, Durbin KJ, Evangelista CC, Ferraz C, Ferreira S, Fleischmann W, Fosler C, Gabrielian AE, Garg NS, Gelbart WM, Glasser K, Glodek A, Gong F, Gorrell JH, Gu Z, Guan P, Harris M, Harris NL, Harvey D, Heiman TJ, Hernandez JR, Houck J, Hostin D, Houston KA, Howland TJ, Wei MH, Ibegwam C, Jalali M, Kalush F, Karpen GH, Ke Z, Kennison JA, Ketchum KA, Kimmel BE, Kodira CD, Kraft C, Kravitz S, Kulp D, Lai Z, Lasko P, Lei Y, Levitsky AA, Li J, Li Z, Liang Y, Lin X, Liu X, Mattei B, McIntosh TC, McLeod MP, McPherson D, Merkulov G, Milshina NV, Mobarry C, Morris J, Moshrefi A, Mount SM, Moy M, Murphy B, Murphy L, Muzny DM, Nelson DL, Nelson DR, Nelson KA, Nixon K, Nusskern DR, Pacleb JM, Palazzolo M, Pittman GS, Pan S, Pollard J, Puri V, Reese MG, Reinert K, Remington K, Saunders RD, Scheeler F, Shen H, Shue BC, Siden-Kiamos I, Simpson M, Skupski MP, Smith T, Spier E, Spradling AC, Stapleton M, Strong R, Sun E, Svirskas R, Tector C, Turner R, Venter E, Wang AH, Wang X, Wang ZY, Wassarman DA, Weinstock GM, Weissenbach J, Williams SM, Woodage T, Worley KC, Wu D, Yang S, Yao QA, Ye J, Yeh RF, Zaveri JS, Zhan M, Zhang G, Zhao Q, Zheng L, Zheng XH, Zhong FN, Zhong W, Zhou X, Zhu S, Zhu X, Smith HO, Gibbs RA, Myers EW, Rubin GM, Venter JC: **The genome sequence of *Drosophila melanogaster*.** *Science* 2000, **287:**2185-2195.
30. The Gene Ontology Consortium: **Gene Ontology: tool for the unification of biology.** *Nature Genet* 2000, **25:**25-29.
31. **Human Genome Sequencing Center at Baylor College of Medicine: Drosophila Genome Project** [http://www.hgsc.bcm.tmc.edu/projects/drosophila]
32. Spradling AC: **ORC binding, gene amplification, and the nature of metazoan replication origins.** *Genes Dev* 1999, **13:**2619-2623.
33. Labrador M, Corces VG: **Setting the boundaries of chromatin domains and nuclear organization.** *Cell* 2002, **111:**151-154.
34. Arnosti DN: **Analysis and function of transcriptional regulatory elements: Insights from Drosophila.** *Annu Rev Entomol* 2003, **48:**579-602.
35. Mann RS, Morata G: **The developmental and molecular biology of genes that subdivide the body of Drosophila.** *Annu Rev Cell Dev Biol* 2000, **16:**243-271.
36. **Flybase: a database of the Drosophila genome** [http://flybase.bio.indiana.edu]
37. Blackman RK, Sanicola M, Raftery LA, Gillevet T, Gelbart WM: **An extensive 3' cis-regulatory region directs the imaginal disk expression of *decapentaplegic*, a member of the TGF-beta family in Drosophila.** *Development* 1991, **111:**657-666.
38. Bachmann A, Knust E: **Positive and negative control of Serrate expression during early development of the Drosophila wing.** *Mech Dev* 1998, **76:**67-78.
39. Barolo S, Walker RG, Polyanovsky AD, Freschi G, Keil T, Posakony JW: **A notch-independent activity of suppressor of hairless is required for normal mechanoreceptor physiology.** *Cell* 2000, **103:**957-969.
40. Haenlin M, Kunisch M, Kramatschek B, Campos-Ortega JA: **Genomic regions regulating early embryonic expression of the Drosophila neurogenic gene *Delta*.** *Mech Dev* 1994, **47:**99-110.
41. Ohta T: **Near-neutrality in evolution of genes and gene regulation.** *Proc Natl Acad Sci USA* 2002, **99:**16134-16137.
42. Bartolomé C, Maside X, Charlesworth B: **On the abundance and distribution of transposable elements in the genome of Drosophila melanogaster.** *Mol Biol Evol* 2002, **19:**926-937.
43. Jordan IK, Rogozin IB, Glazko GV, Koonin EV: **Origin of a substantial fraction of human regulatory sequences from transposable elements.** *Trends Genet* 2003, **19:**68-72.
44. Makalowski W: **Not junk after all.** *Science* 2003, **300:**1246-1247.
45. Zdobnov EM, von Mering C, Letunic I, Torrents D, Suyama M, Copley RR, Christophides GK, Thomasova D, Holt RA, Subramanian GM, Mueller HM, Dimopoulos G, Law JH, Wells MA, Birney E, Charlab R, Halpern AL, Kokoza E, Kraft CL, Lai Z, Lewis S, Louis C, Barillas-Mury C, Nusskern D, Rubin GM, Salzberg SL, Sutton GG, Topalis P, Wides R, Wincker P, Yandell M, Collins FH, Ribeiro J, Gelbart WM, Kafatos FC, Bork P: **Comparative genome and proteome analysis of *Anopheles gambiae* and *Drosophila melanogaster*.** *Science* 2002, **298:**149-159.
46. **NCBI BLAST Home Page** [http://www.ncbi.nlm.nih.gov/BLAST]
47. Hertz GZ, Stormo GD: **Identifying DNA and protein patterns with statistically significant alignments of multiple sequences.** *Bioinformatics* 1999, **15:**563-577.
48. **Regulatory Sequence Analysis Tools** [http://rsat.ulb.ac.be/rsat/]
49. Celniker SE, Wheeler DA, Kronmiller B, Carlson JW, Halpern A, Patel S, Adams M, Champe M, Dugan SP, Frise E, Hodgson A, George RA, Hoskins RA, Laverty T, Muzny DM, Nelson CR, Pacleb JM, Park S, Pfeiffer BD, Richards S, Sodergren EJ, Svirskas R, Tabor PE, Wan K, Stapleton M, Sutton GG, Venter C, Weinstock G, Scherer SE, Myers EW, Gibbs RA, Rubin GM: **Finishing a whole-genome shotgun: Release 3 of the *Drosophila melanogaster* euchromatic genome sequence.** *Genome Biol* 2002, **3:**RESEARCH0079.