EVOLUTIONARY CHANGES IN GENE REGULATION FROM A COMPARATIVE ANALYSIS OF MULTIPLE *DROSOPHILA* SPECIES

LAN HU[1] hulan@bu.edu
DANIEL SEGRÈ[1] dsegre@bu.edu
TEMPLE F. SMITH[1,2] tsmith@darwin.bu.edu

[1]*Graduate Program in Bioinformatics, Boston University, Boston MA 02215, U.S.A.*
[2]*BioMolecular Engineering Research Center, Boston University, Boston MA 02215, U.S.A.*

Exploiting the ortholog/homolog information now available from the complete genomic sequences of twelve species of *Drosophila*, we have investigated the ability of regulatory site recognition methods to find regulatory changes for orthologs linked to chromosomal rearrangements. This has made use of the wealth of synteny information among these species. By comparing orthologs in multiple species, we found that the breakpoint of chromosomal rearrangements could have had an impact on regulatory changes of genes next to it with respect to the gene function and location. Extensions of our approach could be used to shed light on the role of gene regulation in the evolutionary adaptation to different environmental conditions.

Keywords: *Drosophila*; regulatory site; ortholog; chromosomal rearrangement.

1. Introduction

The genomes from twelve species of *Drosophila* (fruit flies) [13] have been recently sequenced, providing a wealth of data for comparative genomics analyses, and in particular for the study of how evolution may have fine-tuned the regulation of specific genes and pathways associated with different lifestyles of these species. These data include species from the two well recognized subgenera of *Drosophila*, *Sophophora* subgenus and *Drosophila* subgenus, that diverged between 40 and 60 million years ago.

While chromosomal rearrangements in *Drosophila* are common, the majority are inversions, which maintain the involved genes within the same Muller Element (ME, a chromosome arm in *D.melanogaster*). Thus few genes appear to have moved between MEs [6]. In a few cases, putative gene homologs are found to change ME, apparently via retrotransposition (i.e. mRNA is retrotranscribed to DNA and reinserted into genome at a new position). In the case of inversion, the regulatory signals may have "traveled" along with the genes, since the range of an inversion is usually large. In the case of retrotransposition, however, the genes normally would not carry along the original regulatory signals. An intriguing question is how different models of DNA rearrangement could have affected the regulatory program of a gene, especially when the upstream region of a gene has been disrupted or left behind.

Upstream small-scale deletions, insertions and point mutations are not the focus of this work. It is not that such events do not play a key role in determining gene regulation and thus expression – as a function of extent, location and/or timing – but these have

been well studied, at least in micro organisms, and are generally assumed to occur more gradually. Rather, we concentrate on the less understood implications of sudden, drastic changes. Potential genetic regulation changes are particularly acute in retrotransposition, because the original regulatory region, lost during a gene transition, is unlikely to be replaced by a compatible and useful transcriptional signal. On the other hand, the fact that these genes survived indicates that, whatever change occurred, it was evolutionarily advantageous, or at least neutral. One possibility is that the moved gene has been fortuitously inserted next to a useful set of regulatory elements, however unlikely that is. Another possibility is that the gene has been inserted in an exon of another gene, regulated in a similar manner. A third possibility is that the gene is placed in a chromosomal region globally maintained at a high transcription level. In such case, given enough time, a minimal upstream region for more specialized regulation could gradually evolve. In general, if a gene is essential, this would require that a second functional copy exists. This could be realized in diploid organisms, in addition to occurring for retrotransposed genes.

Since the fruit fly is a well established model organism, its genetics and development are well studied. The availability of twelve genomes furthermore is expected to provide new comparative genomics insight on the regulation of genes that moved in different ways. In this paper, a method to characterize and compare the potential regulatory sites (PRS) of orthologs across all available species is developed. The method is applied to the central carbon metabolic genes, particularly to those genes that have disrupted upstream region by chromosomal rearrangements along the evolution. The results indicate that comparing common PRS across available species with full synteny breakpoint analysis could help to gain insights of how the breakpoint could affect the regulation of moved genes.

2. Methods

2.1. Synteny analysis

In this study, we are particularly interested in genes that "moved" at the first diverging point in the evolution of *Drosophila*, i.e. about 40~60 million years ago. The expectation would be that these orthologs keep the same neighbor context in one subgenus but not the other. The synteny analysis [5, 6] carried out hence is based on gene neighborhood comparison relative to *D.melanogaster*.

The synteny analysis is schematically illustrated in Fig. 1. Given a gene, A_mel in Fig. 1, from *D. melanogaster*, its adjacent neighbors (X_mel and Y_mel) are extracted, as well as its ortholog and neighbors (if they exist according to annotation) in another *Drosophila* species. Next, for one of the neighbors of the ortholog, N_s in Fig. 1, its ortholog and neighbors are extracted back from *D.melanogaster*. There are two possibilities, as shown in Fig. 1: (1) N_s and Y_mel are orthologs; and (2) N_s and N_mel are orthologs. The first case means that for the neighbor pair of A_mel and Y_mel

in *D.melanogaster*, their orthologs (A_s and N_s respectively) are also neighbors in another species. In the second case, however, the neighbor pair relationship is not consistent, suggesting that there was a breakpoint either between A_mel and Y_mel or between A_s and N_s.

This type of synteny analysis is carried out across seven *Drosophila* species, four from the *Sophophora* subgenus (*D.melanogaster*, *D.yakuba*, *D.erecta*, and *D.ananassae*, diverged about 10~15 million years ago) and three from the *Drosophila* subgenus (*D.mojavensis*, *D.virilis*, and *D.grimshawi*, diverged about 30~35 million years ago). The genes that have possibly "moved" at the first speciation event would keep the same neighbor context in one subgenus but not the other.

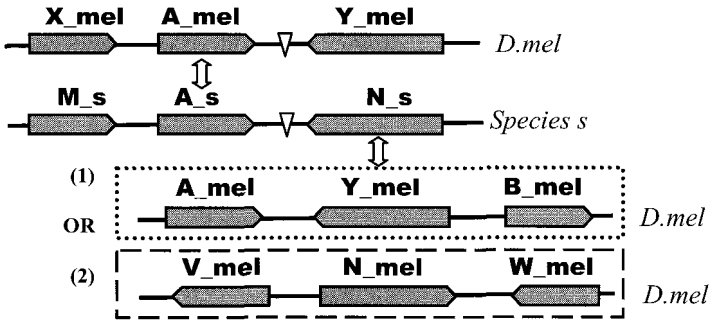

Fig. 1. Schematic illustration of synteny analysis based on gene neighbor context

Double headed arrows indicate that those two genes are orthologs. Pairs A_mel and A_s, N_s and Y_mel, N_s and N_mel are orthologs. Three genes in the same row mean that they are adjacent neighbors. Dotted line box and dashed line box show two possibilities of N's ortholog in *D.mel*: (1) N_s and Y_mel are orthologs. (2) N_s and N_mel are orthologs, showing that there is a breakpoint either between A_mel and Y_mel in *D.mel* or between G and N in species *s* (see main text). White triangles represent the possible breakpoints. *D.mel* = *D.melanogaster*. Species *s* = any other *Drosophila* species. Names with _mel suffix are genes from *D.mel*, _s suffix from Species *s*.

2.2. Regulatory site identification

Many different approaches for potential regulatory sites (motifs) identification have been developed. In general, motif finding falls into two categories: pattern matching to previously identified sites or *de novo* discovery. Pattern matching algorithms (e.g. MotifScanner [10] and Patser [11]) use identified patterns such as position weighted matrices (PWM) or position frequency matrices (PFM) to scan through the sequences and return the segments that have scores over some threshold. The *de novo* discovery approaches use techniques such as Gibbs Sampling (AlignACE [8]) or Expectation Maximization (MEME [3]) to detect the over-represented DNA segments in given sequences. Pattern matching approach largely depends on the patterns which ideally should come from experimentally determined sites. In fruit fly, unfortunately, the number of transcription factors whose binding sites have been characterized is still limited.

The *Drosophila* DNase I Footprint Database (FlyReg 2.0 [4]) has a collection of 1,365 DNase I footprints for *D. melanogaster* from a single experimental data type.

These data have been extracted from 201 primary references and provide a non-redundant set of high quality binding site information for 87 transcription factors. 62 motif models have been curated in the format of PWM's, and 75 in the format of PFM's.

The overall similarity of extracted upstream DNA sequences for orthologs decreases with divergence as expected. To identify the potential regulatory signals, the curated 75 alignment matrices from FlyReg 2.0 are used to scan the upstream regions (both strands) of given genes. Sites that have above threshold scores are returned as putative regulatory sites. Due to the repetitive sequence in a given DNA segment as well as the incompleteness of transcription factor binding sites, overlapping and repetitive sites could be returned by the process. To avoid that, the site with highest in the region is kept and others are discarded.

In order to compare the regulatory sites, three kinds of gene sets are constructed. They are a random gene set, a *Sophophora* ortholog set, and a *Drosophila* ortholog set. The random set(s) are generated by randomly choosing 100 genes with at least 2 kb upstream intergenic region from different species independently (which species to choose are based on individual analysis). The reason for having at least 2kb upstream intergenic region is that current annotation of gene span in species other than *D.melanogaster* does not have transcription start site but translation start site estimation, which may introduce non-transcriptional regulation information in regulatory site finding.

The other two sets are ortholog sets. One hundred genes with at least 2kb upstream intergenic region are chosen (with or without functional constraints) from *D.melanogaster* first. For those 100 genes, orthologs from four species in *Sophophora* subgenus (*D.melanogaster*, *D.yakuba*, *D.erecta*, and *D.ananassae*) constitute the *Sophophora* ortholog set; orthologs from three species in *Drosophila* subgenus (*D.mojavensis*, *D.virilis*, and *D.grimshawi*) constitute the *Drosophila* ortholog set.

In each gene set, after site scanning using Patser with p-value 10^{-3} and tiling, common potential regulatory sites (PRS) are obtained for every quadruplet (the *Sophophora* ortholog set) or triplet (the *Drosophila* ortholog set). Common PRSs are more than the intersection of PRSs from given sequences. If a PRS is detected for N (N>1) times in given sequences, it would be counted N times in the final common PRSs. The distribution of the common PRS then is analyzed in the relationship to moved genes.

3. Results

3.1. Functionally independent genes

Using our synteny analysis (see Methods), we identified about 1050 genes likely to have "moved" at the first speciation event relative to *D. melanogaster*. We next set to compare the upstream regulatory regions of such genes, to shed light on the potential implications of such rearrangements on transcriptional patterns.

Our first test was aimed at studying upstream region changes among functionally independent genes. For this purpose, we chose 100 *D. melanogaster* genes without

functional constraints to construct two ortholog sets and applied our regulatory site identification algorithms (See Methods). For each ortholog set, we constructed a corresponding random set from the same species, to use as a baseline for comparison.

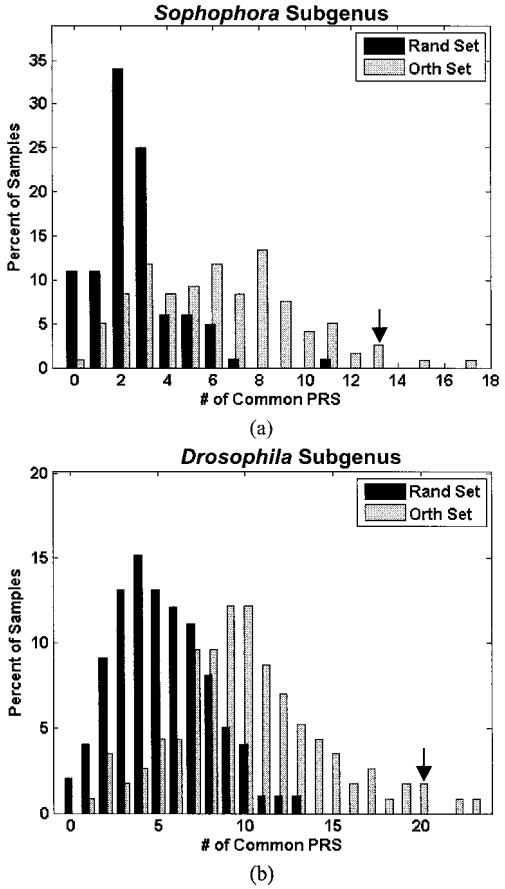

Fig. 2. Distribution of number of common PRS for functionally independent genes

(a) Between the random gene set and the *Sophophora* ortholog set. The random gene set is composed of randomly chosen genes from *D.melanogaster*, *D.yakuba*, *D.erecta*, and *D.ananassae*. (b) Between the random gene set and the *Drosophila* ortholog set. The random gene set is composed of randomly chosen genes from *D.mojavensis*, *D.virilis*, and *D.grimshawi*. In both figures, the black arrow points to where orthologs of *kek1* are binned.

As illustrated in Fig. 2, two ortholog sets have different distribution of common PRS relative to the random gene set, as expected. The orthologs, even functional independent in this case, share more common sites. The next question is whether there is intersection among the genes that have high number of common PRS in both ortholog sets. If there are genes whose orthologs share high number of common PRS in both ortholog sets, this would suggest that the regulation of those genes may not have changed throughout evolution to keep the functions of genes under the selection pressure, and some of the common PRS could be real regulatory sites.

In both ortholog sets, we singled out the top 10 genes whose orthologs share most common PRS. The resulting intersection contains CG14220, CG6621, and *kekl* (CG12283, denoted by the black arrow in Fig. 2). The synteny analysis shows that the orthologs of *kekl* in *D.mojavensis*, *D.virilis*, and *D.grimshawi* have been moved. The gene *kekl* has negative regulation of epidermal growth factor receptor activity [1, 2] and is also involved in *Drosophila* oogenesis [7]. Among all the sites detected, there are three sites that are common in both ortholog sets, two of which are binding sites of the transcription factor (TF) *apterous* and one is the binding site of TF *broad.* The TF *apterous* is involved in cell fate commitment and *broad* in cell death and oogenesis, which is consistent with the functions of *kekl*. Hence, the important functions of *kekl* in development should be conserved throughout evolution regardless of the movement of its orthologs in some species. It is likely that those three sites could be real regulatory sites (see Discussion).

3.2. Central carbon metabolic genes

We next sharpened our analysis by testing the extent of upstream region overlap for genes with known and conserved functions. Given our interest in potential correlation with lifestyle and dietary changes among species, we focused on genes coding for metabolic proteins. Metabolic genes in *D.melanogaster* thus are chosen to construct the two ortholog sets. We identified a total of 104 genes involved in *D.melanogaster* central carbon metabolism, i.e. glycolysis, pentose phosphate pathway, and tricarboxylic acid (TCA) cycle. These genes either code for enzymes or have metabolic functional annotations according to GO terms in the three pathways considered (Supplementary material). Similarly to the first test, the orthologs of these 104 genes from *D.melanogaster*, *D.yakuba*, *D.erecta*, and *D.ananassae* constitute the *Sophophora* ortholog set; orthologs from *D.mojavensis*, *D.virilis*, and *D.grimshawi* constitute the *Drosophila* ortholog set. The random gene sets are the same as those in the first test.

As shown in Fig. 3, the distribution of common PRS in both ortholog sets demonstrates a trend similar to the one found in the comparison of genes that do not necessarily share function (Fig. 2). Again, we focused on the top 10 genes that have the most common PRS from both ortholog sets. The resulting intersection includes CG5261, CG5432, and *Hex-A*. Synteny analysis shows that their orthologs in other species keep the same neighbor context. High number of common PRS in these three genes suggests that functional constraint and same gene context could keep similar gene regulation.

In addition, there are eight genes (Table 1) which have upstream region disrupted by the chromosomal rearrangement breakpoints in *Drosophila* subgenus species (*D.mojavensis*, *D.virilis*, and *D.grimshawi*) relative to *D.melanogaster.* As illustrated in Table 1, these eight genes have low number of common PRS across both subgenera, yet have more common PRS's if they are compared in either subgenus, raising the possibility that those genes may have undergone different regulation mechanisms after the first

speciation event. Using the corresponding random set as background, we quantified the significance of these findings through a Z-test (Supplementary material).

Pyruvate dehydrogenate kinase (*Pdk*) has significant p-values (<0.01) in both ortholog sets, suggesting that the importance of this enzyme has kept the gene under strong selection despite the movement of its orthologs and disruption of the upstream region. Low number of common PRS in all species and high number of common PRS in either subgenus could imply that the movement of this gene may have introduced new regulatory signal to keep, and possibly fine-tune the gene's function.

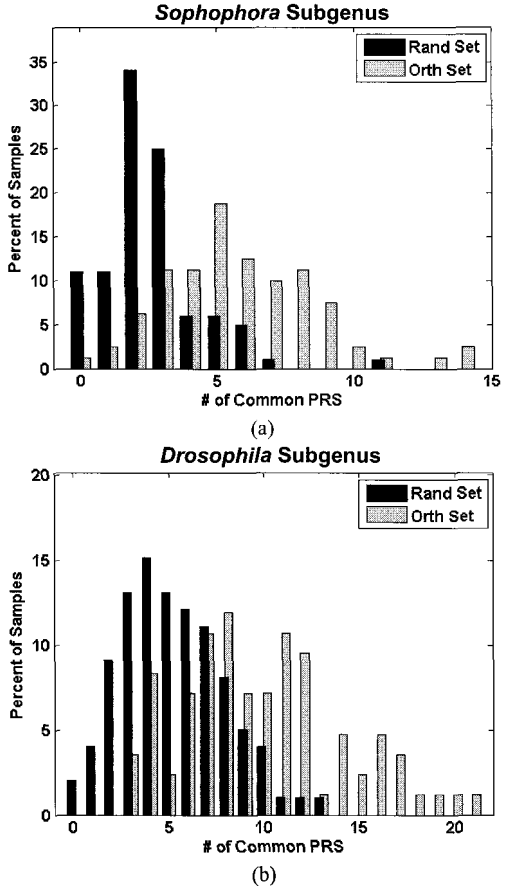

Fig. 3. Distribution of number of common PRS for central carbon metabolic genes

(a) Between the random gene set and the *Sophophora* ortholog set. The random gene set is the same as that in Fig. 2(a). (b) Between the random gene set and the *Drosophila* ortholog set. The random gene set is the same as that in Fig. 2(b).

Genes CG2964 and CG13369 have large p-values in both orthologs sets (>0.4 and >0.5 respectively), which indicates that the two genes cannot be discriminated from the random gene set. While this may potentially mean that the level of expression could have changed significantly after the rearrangement, additional evidence points to a different

possible explanation. There has been evidence of expression similarity within large chromosomal domains in *D. melanogaster*: genes with unrelated function may have similar transcriptional levels merely due to their common chromosomal location [9]. The two genes under discussion, CG2964 and CG13369, turn out to fall into chromosomal domains that share similar expression profile (data not shown). This supports the possibility that "moved" genes could be fortuitously inserted next to useful regulatory elements, such as enhancers which could be far away from the gene itself. Therefore regardless of the small number of common PRS in front of these two genes, "remote" regulatory signal would still guarantee appropriate transcription level of the genes.

Table 1. Common PRS of genes with disrupted upstream region in central carbon metabolism. Numbers in parenthesis are p-values from Z-test.

Gene Name	# of Common PRS (p-value)			Molecular Function in Central Carbon Metabolism
	Sophophora ortholog set	*Drosophila* ortholog set	Both	
Pdk	8 (0.0022)	16 (0.0001)	4	pyruvate dehydrogenase kinase activity
CG2964	4 (0.4178)	7 (0.5170)	1	pyruvate kinase activity; carbohydrate kinase activity
CG13369	3 (0.8044)	7 (0.5170)	0	pentose-phosphate shunt
CG5362	9 (0.0003)	10 (0.0783)	3	L-lactate dehydrogenase activity; L-malate dehydrogenase activity
Mdh	8 (0.0022)	7 (0.5170)	1	NAD binding; malate dehydrogenase activity
CG9467	6 (0.0529)	8 (0.3083)	1	NAD binding; glyceraldehyde-3-phosphate dehydrogenase (phosphorylating) activity
CG7349	4 (0.4178)	12 (0.0123)	0	succinate dehydrogenase activity
CG6666	5 (0.1697)	8 (0.3083)	0	succinate dehydrogenase (ubiquinone) activity

Genes CG5362 and *Mdh* have significant p-value in one ortholog set but not the other (0.0003 and 0.0783 for CG5326, 0.0022 and 0.5170 for *Mdh*). This may suggest that the species from different subgenera have different metabolic requirements as a result of adaptation to different environments. The characterization of TF binding sites used in this particular measurement is solely based on *D.melanogaster*, which could introduce a bias towards species from *Sophophora* subgenus that are close to *D.melanogaster* and may not be sufficient to detect regulatory signal in more divergent species from the *Drosophila* subgenus.

4. Discussion

While our analysis shows potentially significant patterns that link genome rearrangements to regulatory adaptation of metabolic genes, one should keep in mind that

the currently available set of characterized TF binding sites is rather limited, and mainly specific to *D.melanogaster* development. This inevitably introduces a bias in recognizing PRS in orthologs. The results, however, still demonstrate that comparing common PRS across available species with full synteny breakpoint analysis could help gain valuable information, for solving the puzzle of the regulation of "moved" genes.

The broad question of how DNA breakpoints affect the regulation of moved genes in *Drosophila* still remains largely unanswered. The method presented in this study takes the perspective of identifying common potential regulatory sites for sets of orthologs across different species, and attempts to elucidate how breakpoints within the upstream region of a gene could affect its regulation in light of its function.

In both tests we presented, the method shows that orthologs tend to have more common potential regulatory sites regardless of functional dependence, which is expected because the selection of function of orthologs may select the regulation as well.

In the test of functional independent gene sets (Fig. 2), the detection of gene *kek1* shows that this method could capture the common PRS in orthologs over ~50 million year span even when a gene has been "moved". This could be due to the constraint from conservation of a gene with important developmental function.

The analysis of eight genes and their orthologs that have upstream region disrupted by chromosomal rearrangement breakpoints in central carbon metabolism (Table 1) exemplifies the possible effect that the breakpoint could have on gene regulation. The importance of gene functions requires continuation of compatible regulation, despite nearby breakpoints. In the case of *Pdk*, it may also have evolved new regulatory signal to keep the gene function after the first speciation event. Some genes that "moved" without their own regulatory signal, such as CG2964 and CG13369, could have been fortuitously inserted into a highly expressed region, hence guaranteeing continuity of their functionality. For some other genes, such as CG5362 and *Mdh*, the difference in common PRS could be a sign of actual difference in gene regulation caused by the specific metabolic requirement due to adaptation to different environments. For example, *D.melanogaster* is a sympatric cosmopolitan species, feeding on necrotic fruit. *D.mojavensis* is a cactophilic species that is specifically found on the rotten arms of cacti in deserts.

Exploiting the ortholog/homolog information now available from the complete genomic sequences of twelve species of *Drosophila*, we have investigated the ability of regulatory site recognition method to find regulatory changes linked to chromosomal rearrangements, particularly the genes next to the breakpoints. We have shown that breakpoints could have multi-level effects on the regulatory changes of those genes in two subgenera of *Drosophila* with respect to the gene function and gene location. With the availability of whole genome expression data, it will give better understanding of how breakpoints change gene regulation along the evolution.

Acknowledgments

We thank AAA site [12] for genome assemblies. We also thank Douglas Smith for permission of using sequence data and Venky Iyer at Eisen Lab UC Berkeley for ortholog model. *D.erecta*, *D.ananassae*, *D.mojavensis*, *D.virilis* and *D.grimshawi* were sequenced by Agencourt Biosystems. *D.yakuba* was sequenced by Washington University. This work is sponsored by NSF grant DBI-0516000.

References

[1] Alvarado, D., Rice, A.H., and Duffy, J.B., Knockouts of *Kekkon1* define sequence elements essential for *Drosophila* epidermal growth factor receptor inhibition, *Genetics*, 166(1):201-211, 2004.

[2] Alvarado, D., Rice, A.H., and Duffy, J.B., Bipartite inhibition of *Drosophila* epidermal growth factor receptor by the extracellular and transmembrane domains of *Kekkon1*, *Genetics*, 167(1):187-202, 2004.

[3] Bailey, T.L. and Elkan, C., Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, 28-36, 1994.

[4] Bergman, C.M., Carlson, J.W., and Celniker, S.E., *Drosophila* DNase I footprint database: a systematic genome annotation of transcription factor binding sites in the fruitfly, Drosophila melanogaster, *Bioinformatics*, 21(8):1747-1749, 2005.

[5] Bhutkar A., Russo, S., Smith, T. F., and Gelbart, W.M., Techniques for Multi-Genome Synteny Analysis to Overcome Assembly Limitations, *Genome Inform.*, 17(2): 152-161, 2006.

[6] Drosophila Comparative Genome Sequencing and Analysis Consortium, Genomics on a phylogeny: Evolution of Genes and Genomes in the Genus Drosophila, Submitted

[7] Ghiglione, C., Carraway, K.L., Amundadottir, L.T., Boswell, R.E., Perrimon, N., and Duffy, J.B., The transmembrane molecule *kekkon1* acts in a feedback loop to negatively regulate the activity of the Drosophila EGF receptor during oogenesis, *Cell*, 96(6):847-856, 1999.

[8] Hughes, J.D., Estep, P.W., Tavazoie, S., and Church, G.M., Computational identification of cis-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*, *J. Mol. Biol.*, 296(5):1205-1214, 2000.

[9] Spellman, P.T. and Rubin, G.M., Evidence for large domains of similarly expressed genes in the *Drosophila* genome, *J. Biol.*, 1(1):5, 2002.

[10] Thijs, G., Marchal, K., Lescot, M., Rombauts, S., Moor, B.D., Rouzé, P., and Moreau, Y., A Gibbs sampling method to detect overrepresented motifs in the upstream regions of coexpressed genes, *J.Comput. Biol.*, 9(2):447-464, 2002.

[11] van Helden, J., Regulatory sequence analysis tools, *Nucleic Acids Res.*, 31(13):3593-3596, 2003.

[12] http://rana.lbl.gov/drosophila/

[13] http://www.flybase.org/