

Case Essay

Computational inference of *cis*-regulatory elements involved in transcriptional regulation of sex-biased gene expression in *Drosophila*

Nicolas Jaccard*

Supervisors: Dr. Max Reuter** | Dr. Vincent Plagnol***

Center for Mathematics and Physics in the Life Sciences and Experimental Biology, University College London, London*

Department of Genetics, Evolution and Environment, University College London, London**

UCL Genetics Institute, University College London, London***

Drosophila sex-determination mechanism is a prime example of complex multi-level regulatory network involving transcriptional regulation and alternative splicing of pre-messenger RNA. The different pathways involved have been extensively described, including the role of key molecules such as *doublesex* or *fruitless*, which are thought to regulate the expression of particular genes responsible for sex-specific phenotypes. However, it is yet to be determined if they directly interact with the DNA or if an intermediate layer of regulators comes into play.

The goal of this case essay was to determine if a *cis*-regulatory element present in the 5' flanking region of certain *Drosophila melanogaster* genes could be responsible for their sex-biased expression. Different approaches to computational inference of DNA motifs were integrated in a workflow designed to facilitate the analysis of large amounts of sequences with, as a starting point, sex-biased expression data that was extracted from the literature. A custom R framework was also developed to allow seamless integration of the various tools and to facilitate data visualization.

Motifs with an 'ATCGAT' core were found to be significantly overrepresented upstream of female-biased genes and the proportion of male-biased genes with at least one instance of the sequence in their 5' flanking region was below what was observed for a random control distribution. Similar results were obtained for six related *Drosophila* species. The motif was identified as a binding sequence for BEAF-32 (32 kDa boundary element-association factor) but more interestingly, it also corresponds to the previously reported DNA replicated-related element (DRE), which is the central element of a vast regulatory network. Results seem to indicate that the CRE/DREF system might also have a sex-specific regulatory function, mostly independent of previously described mechanisms.

1. Introduction

1.1 *Drosophila* sex determination

Sexual dimorphism is a trait shared by the greatest part of metazoan organisms and its associated phenotypes have been shown to be the most differentiating characteristics among individuals in a species¹. Sex-related mechanisms also seem to evolve at a fast rate, outpacing what is observed for any other trait, including genes involved in the survival of the organism². It is therefore not surprising that sex differentiation pathways have been found to diverge considerably between the major metazoan model organisms³.

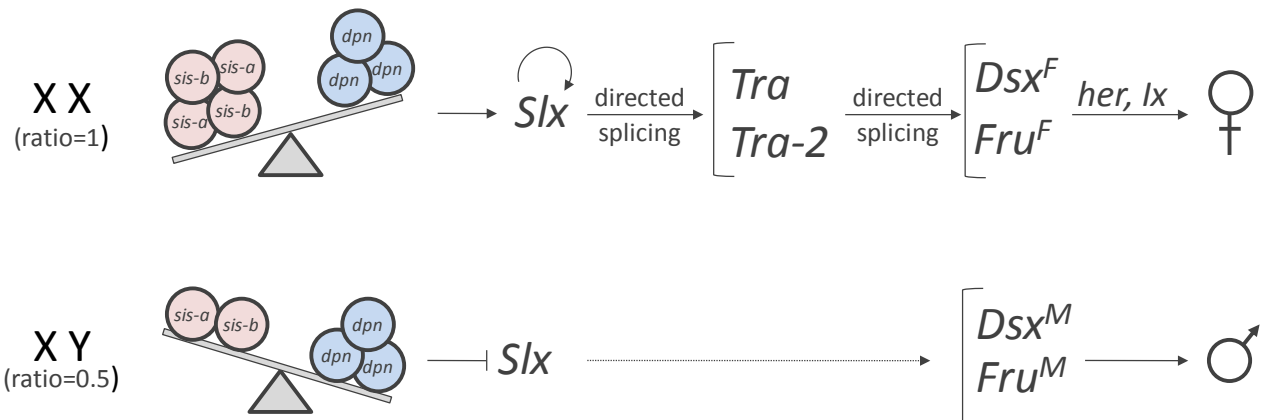


Figure 1 Molecular mechanisms underlying *Drosophila* sex-determination process

Drosophila sex-determination process is a prime example of complex multi-level regulatory network involving transcriptional regulation and alternative splicing of pre-messenger RNA (Figure 1). Unlike in mammals where the sex of an individual is mostly determined by the presence or absence of the Y chromosome, the ratio of X chromosomes to autosomes is thought to constitute the primary signal for sex determination in *Drosophila*. A set of genes located on chromosome X code for 'numerator' proteins such as *sisterless-a* and *sisterless-b* (*sis-a* and *sis-b*) that can activate the regulatory protein *sex-lethal* (*Sxl*)⁴ while autosomes-encoded 'denominator' proteins (e.g. *deadpan*) inhibit its early activation⁵. The ratio of X chromosome to autosomes thus appears to be biologically encoded by the competition between the 'numerator' and 'denominator' proteins. *Sxl* has an affinity for its own pre-messenger RNA and promotes a splicing pattern that results in the active form of the protein. Later in the development, *Sxl* is also transcribed in males but due to the absence of the molecule in its active form, an mRNA with an early stop codon is produced after splicing, leading the translation into an inactive version of the regulator. In females, active *sex-lethal* is already present and can maintain its own production through this auto feedback loop⁶ and direct the splicing of the transformer genes *tra* and *tra2* that will in turn promote female-specific splicing of *doublesex* (*Dsx^F*), which is thought to be one of the main transcriptional regulator of sex-specific genes⁷. In addition, *tra* and *tra2* also affects the splicing of *fruitless* into its female isoform, *Fru^F*. In conjunction with other regulators such as *Her* (hermaphrodite)⁸ and *Ix* (intersex)⁹, *Dsx^F* and *Fru^F* will activate genes related to female traits. In absence of active transformer proteins in males, *doublesex* and *fruitless* will be spliced using the default mechanism, leading to the production of the male-specific *Dsx^M* and *Fru^M* regulators. The latter has been shown to be essential for males to exhibit courtship behavior¹⁰.

Recent studies have examined *Drosophila* transcriptomes and more precisely the expressions of genes that appear to be biased based on the sex of the individual^{1, 11-14}. Most of the studies use mRNA extracted from the whole organism, ruling out the possibility to compare gene expression on a tissue basis, which would allow removing a potential experimental bias due to the high expression of certain genes in sexual tissues¹⁵). Nevertheless, these studies generated large quantities of expression data that can be used as a basis for further investigations.

While regulatory networks involved in sex determination are relatively well understood, it is still not clear how *doublesex* and *fruitless* achieve the regulation of the genes involved in sex-specific phenotypes, more especially it is yet to be determined if they are part of a regulatory network comprising other intermediate transcription factors or if they can modulate the expression of these genes by themselves. This case essay aims at evaluating different computational approaches that might shed light on the regulation mechanisms that lead to gene expression bias in male and female *Drosophila* individuals.

1.2 DNA motifs representation and discovery

In recent years, the amount of data related to biological systems has increased dramatically. New whole-genome sequences are appearing every week, complemented by high throughput analysis of the transcriptome¹⁶, proteome¹⁷ and metabolome¹⁸. All this information can eventually be pooled together to understand fundamental biological mechanisms and decipher complex networks. In the case of transcriptional regulation, it might be interesting to determine where in the genome a known transcription factor can bind or, on the contrary, try to find out which transcription factor might bind to a sequence of interest. Various experimental methods are available to accomplish such tasks, including tiling arrays¹⁹, universal protein-binding microarrays²⁰, ChIP-chip²¹ (chromatin immunoprecipitation selection of bound factors combined with relative enrichment detection using DNA microarrays) or ChIP-seq²² (immunoprecipitation followed by the sequencing of the precipitated fragments). Complementary computational methods have also been developed to analyze the vast amount of data resulting from these high throughput methods or to discover *de-novo* motifs that can then be experimentally characterized.

1.2.1 Regulatory sequence representation

Transcription factors rarely bind to a unique sequence but instead have varying affinity for a set of related oligonucleotides that can be represented as a unique, idealized consensus sequence. In its simplest form, it is made up of the most commonly occurring nucleotide for all positions in the sequences. It can be improved by using a degenerated code (Figure 2.A) that allows for ambiguous positions (for example a D in the sequence can be replaced by anything but C, see IUPAC rules²³). While apparently giving a more realistic description of a binding sequence, this approach has many shortcomings. A degenerated symbol at a certain position implies that all the possible nucleotides it represents are given the same weight irrespective of their actual frequency at this position in the set of binding sequences (positional preference is lost). In addition, it is often not possible to accurately represent all the binding sequences in a set with a single consensus sequence, despite having an extensive degenerated symbols alphabet. Certainly useful to get a qualitative and instinctive appreciation of a sequence, the consensus representation might not be suitable for quantitative analysis of genomic data.

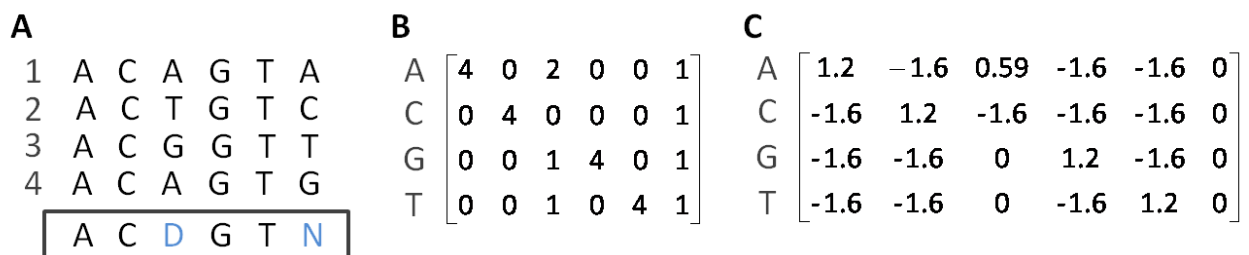


Figure 2 Different representations of a set of sequences. **A** Degenerated consensus sequence **B** Occurrence matrix for the sequences shown in A. **C** Weight matrix for the sequences shown in A (the prior probabilities were chosen to be 0.25 for all nucleotides and the pseudo-factor k was set to 1)

Another way to represent multiple binding sequences consists in building an occurrence matrix (Figure 2.B) where each element contains the frequency at which a nucleotide is observed at a given position. Nucleotide composition varies greatly between species²⁴ but also depending on the type of sequence considered (intergenic, intronic or coding)²⁵. A weight matrix (Figure 2.C) taking into account the nucleotide distribution bias (also called prior residue probability) can be derived from the occurrence matrix by calculating the weight of each of its elements (Equation 1).

$$W_{i,j} = \ln\left(\frac{\text{observed}}{\text{expected}}\right) = \ln\left(\frac{f_{i,j}}{p_j}\right) = \frac{\ln\left(\frac{n_{i,j}}{N}\right)}{p_j} \quad (\text{Equation 1})$$

$W_{i,j}$ the weight of the nucleotide j at position i , $f_{i,j}$ the observed frequency for the nucleotide j at position i , p_j the overall frequency of the nucleotide j (usually estimated from a large number of sequences), N the number of sequences, $n_{i,j}$ the number of observed nucleotide j at position i

If a nucleotide does not appear at a given position in any of the sequence, the frequency is equal to $-\infty$ ²⁶. The introduction of a pseudo-weight has been proposed²⁷ to avoid such a situation (Equation 2).

$$W_{i,j} = \ln\left(\frac{f'_{i,j}}{p_j}\right) = \frac{n_{i,j} + k p_j}{(N + k) p_j} \quad (\text{Equation 2})$$

$W_{i,j}$ the corrected weight of the nucleotide j at position i , $f'_{i,j}$ the corrected observed frequency for the nucleotide j at position i , p_j the overall frequency of the nucleotide j (usually estimated from a large number of sequences), N the number of sequences, $n_{i,j}$ the number of observed nucleotide j at position i , k the pseudo-weight

The weight can be used to score a particular sequence against the matrix. For the matrix shown in Figure 2.C, the sequence ACAGN (where N can be any nucleotide) would have a score of 5.4, which is the best possible score for that matrix. Another metric of interest is the information content of a matrix (also called normalized log-likelihood or relative entropy), a measure of the discrimination between the binding sequence represented by the matrix and a random sequence (given by the background model)²⁷. This notion is interesting as it relates to the thermodynamic aspect of the binding event²⁸.

$$I = \sum_{j=1}^L \sum_{i=1}^4 f'_{i,j} W_{i,j} = \sum_{j=1}^L \sum_{i=1}^4 f'_{i,j} \ln\left(\frac{f'_{i,j}}{p_j}\right) \quad (\text{Equation 3})$$

I the information content of the sequence, $f'_{i,j}$ the corrected observed frequency for the nucleotide j at position i , p_j the overall frequency of the nucleotide j (usually estimated from a large number of sequences), L the number of positions in the sequences (number of columns in the matrix)

Although matrices are a very elegant way to abstract the concept of binding sequence in a form that is easily usable in the context of computational analysis of genomic data, they are difficult to interpret visually. Sequences logos (Figure 3) have been proposed by Schneider and Stephens as a middle-ground between the easily interpretable but flawed consensus sequences and the information rich but cryptic weight matrices²⁹.

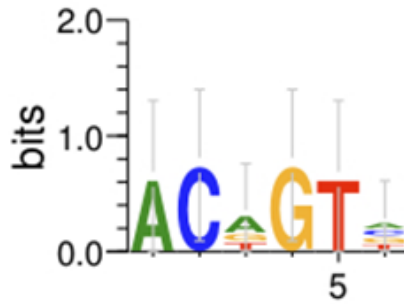


Figure 3 Sequence logo for the set of sequence shown in Figure 2.A (generated using WebLogo 3.0 with *Drosophila* as background model)

They define the sequence conservation of a particular position as the difference between the maximal and observed entropy (Equation 4).

$$R_{seq}(i) = S_{max} - S_{obs} = 2 - \left(-\sum_{n=1}^4 f_{i,j} \log_2 f'_{i,j}\right) \quad (\text{Equation 4})$$

$R_{seq}(i)$ the sequence conservation at position i , $f_{i,j}$ is the observed frequency for the nucleotide j at position i

The height of a nucleotide at a particular position is given by the product of the sequence conservation and its frequency (Equation 4)

$$h_{i,j} = f_{i,j} R_{seq}(i) \quad (\text{Equation 5})$$

$h_{i,j}$ the height of nucleotide j at position i, $f_{i,j}$ is the observed frequency for the nucleotide j at position i, $R_{seq}(i)$ the sequence conservation at position i

1.2.2 De-novo motif discovery

While the link between co-expression and co-regulation is rarely evident³⁰, genes presenting similar expression profiles might share a common *cis*-regulatory element. *De-novo* (unsupervised) motif discovery algorithms have been developed to identify overrepresented patterns in a set of potentially related sequences, with little to no prior information regarding the target sequence.

Table 1 Various scoring metrics for words-based motif discovery as proposed by van Helden³¹. f_{obs} is the observed frequency of a given word, f_{exp} is the frequency of a given word according to the background model, x is the observed occurrences for a given word, p is the probability of occurrence for a given word according to the background model, T is the number of possible positions for a word of length k in a set of n sequences of length L_i , D is the number of tests

Scoring metric	Equation	Description
Log-likelihood ratio	$LLR = f_{obs} \log \left(\frac{f_{obs}}{f_{exp}} \right)$	Probability that the data could have been generated by the model
P-value	$P(X \geq x) = \sum_{i=x}^T \frac{T!}{i! (T-i)!} p^i (1-p)^{T-i}$	Probability to observe at least x occurrence of a word in a succession of trial
Significance	$sig = -\log_{10}(\text{Pvalue} * D)$	Gives a measure of the number of false positives (sig=1 means a false positive should be expected every 10 sets)

The major difficulty is to make the distinction between frequent and over-represented patterns. Indeed, simply trying to identify the most frequent group of nucleotides of a given length in the upstream regions of a set of genes would most likely reveal trivial results such as AT-rich motifs or intrinsic properties of such regions (e.g. TATA box or other promoters-related elements). Instead, a common approach is to look for over-represented motifs, which are DNA patterns whose observed frequencies in the set of analyzed sequences are higher than their expected frequencies according to the background model. The choice of a suitable model is therefore critical to obtain pertinent results. One strategy consists in using input sequences to train a markov chain model³², resulting in good background information if the number of sequences is large enough. An alternative for organisms with fully sequenced and annotated genomes is to calculate frequencies of all possible oligonucleotides in specific regions (e.g. all upstream sequences). Various scoring metrics were proposed by van Helden³¹ to evaluate the over-representation of words (Table 1).

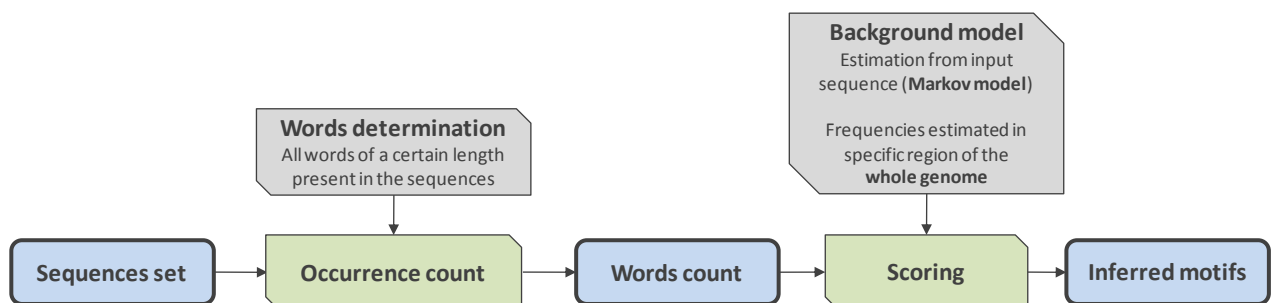


Figure 4 Typical words-based algorithm (adapted from Van Helden²⁶)

Words-based motif discovery (Figure 4) has the advantage to rely on string matching algorithms that have been heavily optimized in other computer sciences fields and in consequence requires less computational power than other strategies such as matrix-based matching. However, if degenerated sequences have to be taken into account, this method is no longer suitable due to the complexity of such an analysis. A solution consists in counting occurrences of individual words before assembling related sequences into a consensus motif²⁶. Many words-based DNA pattern discovery algorithms have been successfully used to identify regulatory motifs in a set of related sequences: RSAT's³³ dna-pattern³¹ and dyad-analysis (spaced words)³⁴, a suffix-tree based algorithm³⁵ and YMF³⁶.

Alternative approaches based on the matrix representation of DNA patterns have recently gained traction. The field was pioneered by Hertz *et al* with their greedy algorithm³⁷ that combines pairs of all possible oligonucleotides of a given length from two related sequences into matrices. Only the highest scoring matrices are kept (e.g. using information content as a scoring metric³⁸) and are in turn combined with all possible oligonucleotides in another related sequence. This process is iterated for all the sequences in the set (Figure 5) and a final sorting of the matrices retain the most relevant motifs.

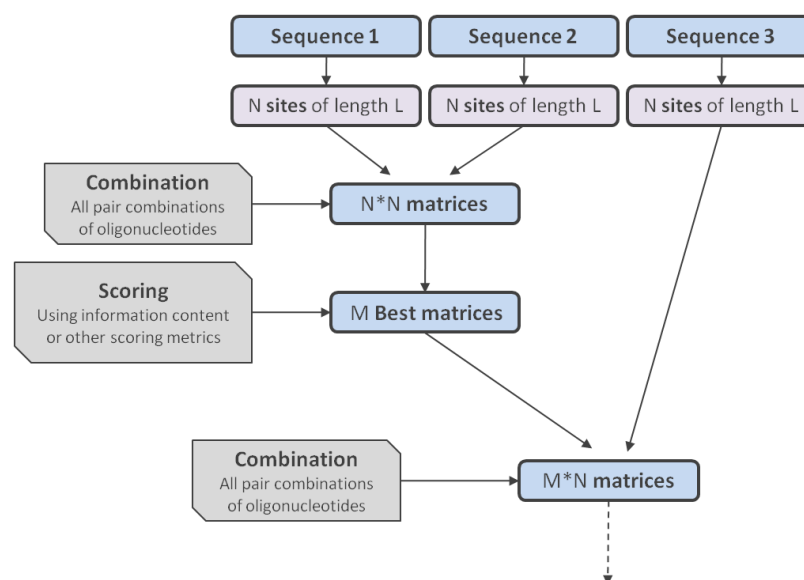


Figure 5 Greed algorithm for matrix-based discovery of DNA patterns (Adapted from Hertz³⁷ *et al* and Van Helden³⁸)

Timothy *et al*³⁹ proposed the use of an expectation maximization (EM) algorithm for finite mixture models that was first introduced by Aitkin *et al*⁴⁰: a weight matrix is built using a word from the target sequences. It then scans all possible words in the set of sequences, calculating at each iteration the probability that the picked word was generated from the initial matrix rather than by the background model and use this information to refine the motif.

Another widely used approach relies on Gibbs sampling, a stochastic equivalent of expectation maximization, first described by Lawrence *et al*⁴¹. Many algorithms adopted this strategy, including MotifSampler³², AlignAce⁴², BioProspector⁴³ and Info-gibbs⁴⁴. In addition, genetic algorithms for motif discovery such as GAME have been developed but the computational power requirements make them unsuitable for large scale motif discovery⁴⁵.

2. Methods

The goal of this study was to identify eventual regulatory sequences that appear upstream of sex-biased genes. The motif discovery workflow (Figure 6) has been designed to integrate heterogeneous genomic analysis tools while being as automated as possible. A custom R framework (ver. 2.10.1) was developed to handle most of the automation tasks and to allow seamless flow of data despite the different input and output formats employed by the various tools. Sequence retrieval was handled by biomaRt⁴⁶ (ver. 2.3.4) and all graphs were generated using the ggplot2 package (ver. 0.8.5). Van Helden's regulatory sequence analysis tool kit⁴⁷ (RSAT) has also been extensively used at various points of the workflow.

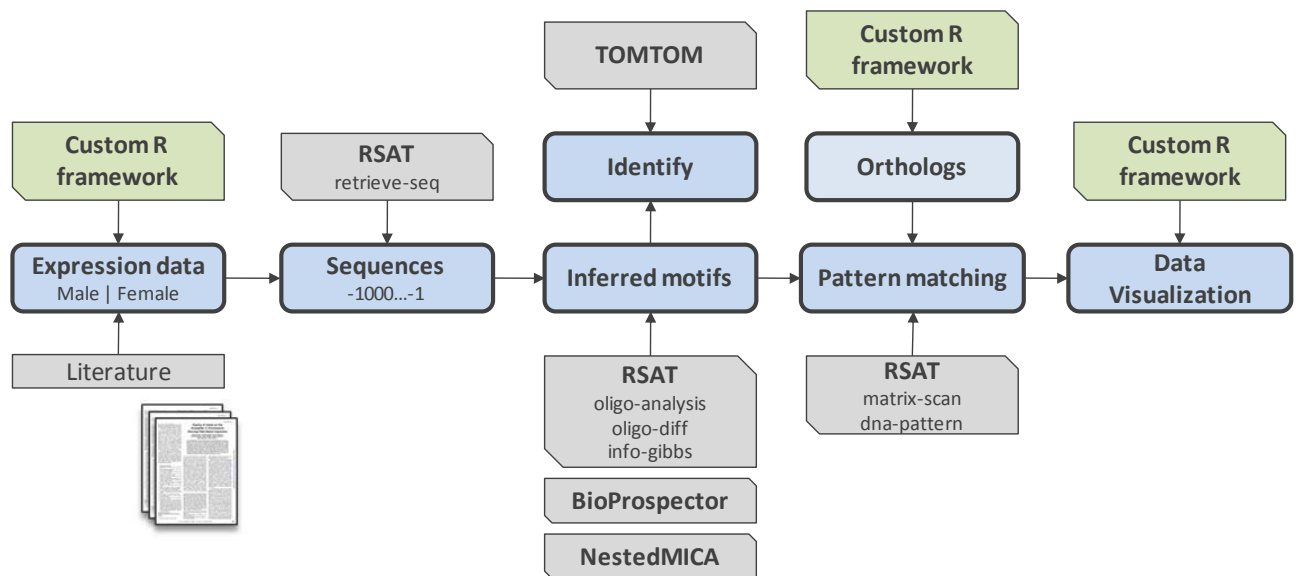


Figure 6 Motif discovery and analysis workflow

Sex-biased expression data for *Drosophila melanogaster* was extracted from Zhang *et al*¹ supplementary material. 1000 bp upstream sequences were then retrieved from Ensembl (Ensembl Genes 57) using RSAT or biomaRt. Unless otherwise specified, masks for low-complexity elements, repeats and coding sequences were applied. Motifs were inferred using RSAT tools (words-based oligo-analysis³¹, oligo-diff and matrice-based info-gibbs⁴⁴), BioProspector⁴³ and Sanger Institute's NestedMICA⁴⁸ (ver. 0.8.0). The background model used was upstream regions of *Drosophila* genome when available or a first order markov chain model if not. The best candidates were compared to known *Drosophila* binding sites using FlyReg⁴⁹ (ver. 2.0) via TOMTOM⁵⁰ (ver. 4.3.0) and were matched (using matrix-scan and dna-pattern³³) against the initial set of sequences but also against those of orthologs from various *Drosophila* species (retrieved from Ensembl Metazoa 3 database). Finally, the R framework was used for data visualization, including spatial location of the expression in the fly body by querying the FlyAtlas database⁵¹.

3. Results

A few examples of discovered motifs are presented in the following results. The 1000 bp 5' flanking sequences of the genes presenting a high bias in term of expression were used to build two sex-specific inferences sets (containing 18 and 40 sequences for females and males respectively). Both matrix and words-based motif discovery algorithms were ran independently for each set of sequences (Table 2 and Table 3)

Table 2 Examples of overrepresented motifs resulting from matrix-based pattern discovery using Matrix-scan on sex-specific sequence sets

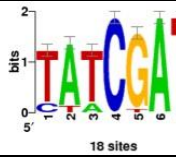
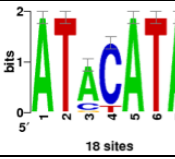
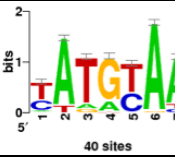
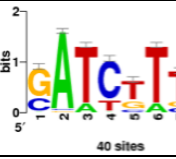
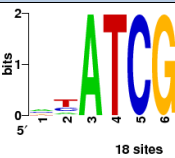
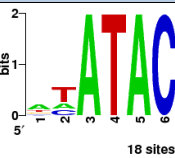
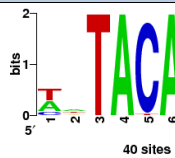
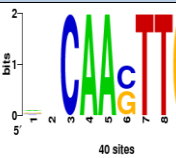
	Female-biased set		Male-biased set	
	Motif F.1	Motif F.2	Motif M.1	Motif M.2
				
Log likelihood ratio	153.027	155.264	249.36	295.648
Information content	7.459	7.556	5.794	6.934

Table 3 Examples of overrepresented motifs resulting from words-based pattern discovery using dna-pattern on sex-specific sequence sets. O. occurrence is the observed number of occurrences and E. occurrence is the expected occurrence

	Female-biased set		Male-biased set	
	Motif F.3	Motif F.4	Motif M.3	Motif M.4
				
O. Occurrence	16	14	49	20
E. occurrence	4.77	2.79	22.51	6.29
Significance	0.80	1.78	2.43	1.08

BioProspector's best motif for the set of 18 female genes is shown in Figure 7.B, no relevant overrepresented motif was found for the male-biased genes. Additional analyses were also performed on larger sets of sequences including all biased genes (470 and 1779 genes respectively for female and male respectively). The most overrepresented motif in the female-biased set according to NestedMICA is shown in Figure 7.A with again no relevant results for the male set. The two sets were then compared using oligo-diff in order to determine which oligonucleotides were overrepresented in a set respectively to the other. The core of the motif shown in Figure 7.C (TATCGA) has been identified as being enriched in the set of female-biased genes with an observed frequency of 1.14E-03 compared to 5.44E-04 in the male-biased set.



Figure 7 Various motif identification results. **A** NestedMICA best match for all female biased genes **B** BioProspector match for the 18 most biased female genes **C** Oligo-diff result for the comparison between all the biased genes for both genders

Further analyzing all these motifs would be a daunting task and well out of the scope of this study. Instead, it was decided to focus on patterns emerging when comparing the results from the different approaches, more especially those containing the 'TATCGAT' core motif (motifs F.1, F.3 and Figure 6), as they were the only inferred motifs that were identified as being solely overrepresented in female-biased sequences sets. The rest of this document will therefore focus on this core motif (Figure 8) and try to determine whether it can potentially

be linked with an eventual regulatory mechanism responsible for the sex-biased expression of genes in *Drosophila*.

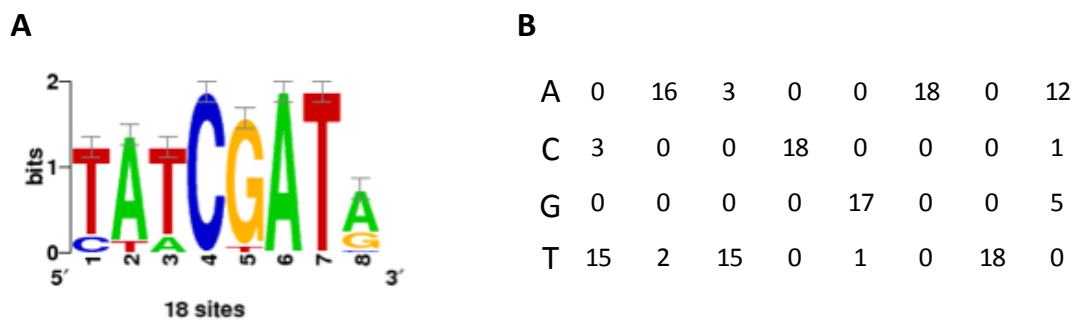


Figure 8 Motif of interest **A** sequence logo **B** Occurrence matrix

The experimental dataset contained 2249 genes (470 and 1779 were female and male-biased respectively). For statistical validation of the results, 100'000 sets of 470 and 1779 genes were randomly sampled in the *Drosophila melanogaster* genome. 1000 bp upstream sequences for the control sets and sex-biased genes were retrieved before being scanned with the matrix representing the motif. Figure 9.A shows the proportions of sex-biased genes with at least one hit for the target sequence alongside the control distributions.

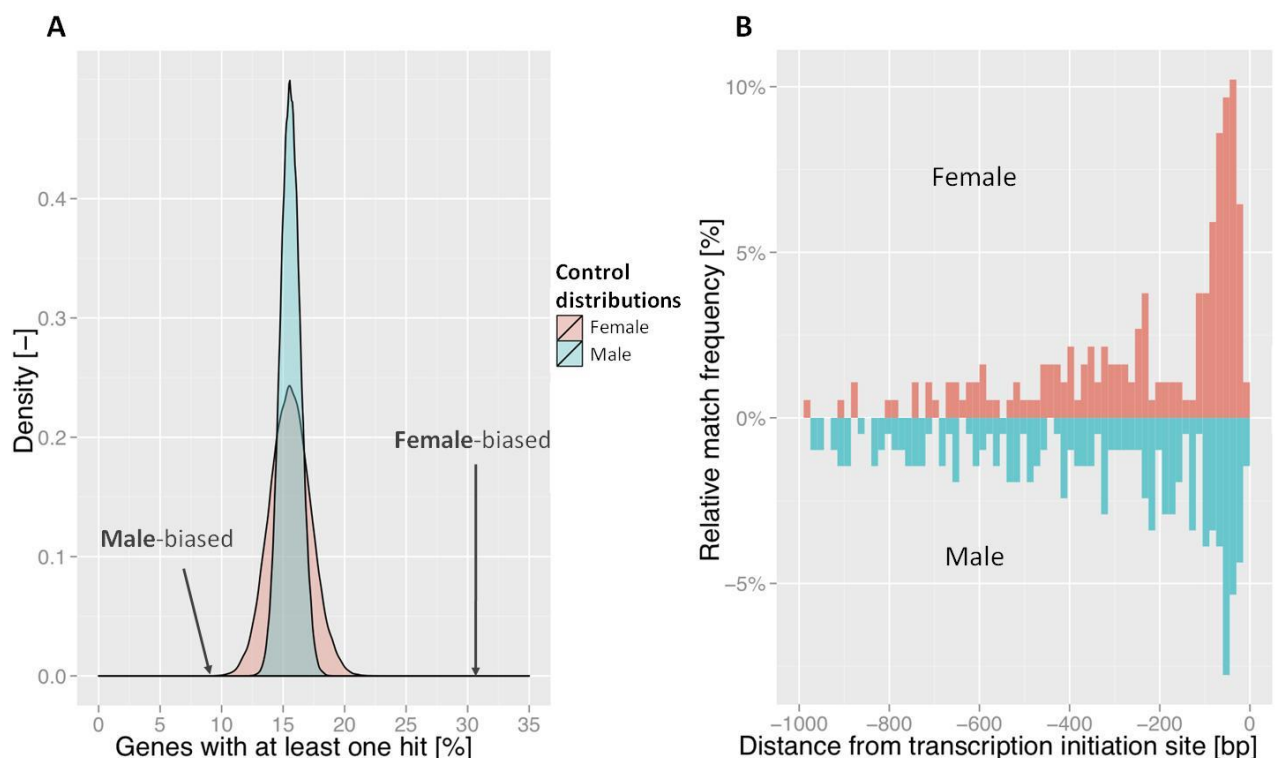


Figure 9 **A** Proportions of sex-biased genes with at least one hit for the target motif in their 1000 bp upstream region alongside kernel density estimations for control distributions **B** Relative frequency distribution of the motif distance from the transcription initiation site (only one transcript considered per gene, one transcript can have multiple hits)

Among the genes that showed a female-biased expression, 30.6% had at least one instance of the motif in their upstream region, which was significantly higher than what was obtained for the female control set whereas only 9.2% of the male-biased genes were found to have an instance of the sequence in their 5' flanking region, a figure markedly lower than that observed in the corresponding control distribution (Table 4).

Table 4 Proportions of sex-biased genes having at least one hit for the target motif and properties of the control distributions generated from randomly picking sets of genes in *Drosophila melanogaster* genome (Figure 9)

Bias	Genes scanned	At least one hit [%]	Controls sets		
			Mean	0.025 quantile	0.975 quantile
Female	470	30.6	15.575	12.34	18.94
Male	1779	9.2	15.81	13.99	17.20

The target sequence was predominantly located within 100 bp of the transcription initiation site (Figure 10.B). In addition, the frequency of matches directly upstream of the transcript was significantly higher for female-biased genes compared to their male counterpart. The sequence preference of the motif (Figure 8) differed according to the gender the genes were biased towards and the distance from the transcription initiation site (Figure 10) but the core 'ATCGAT' was present in all cases.

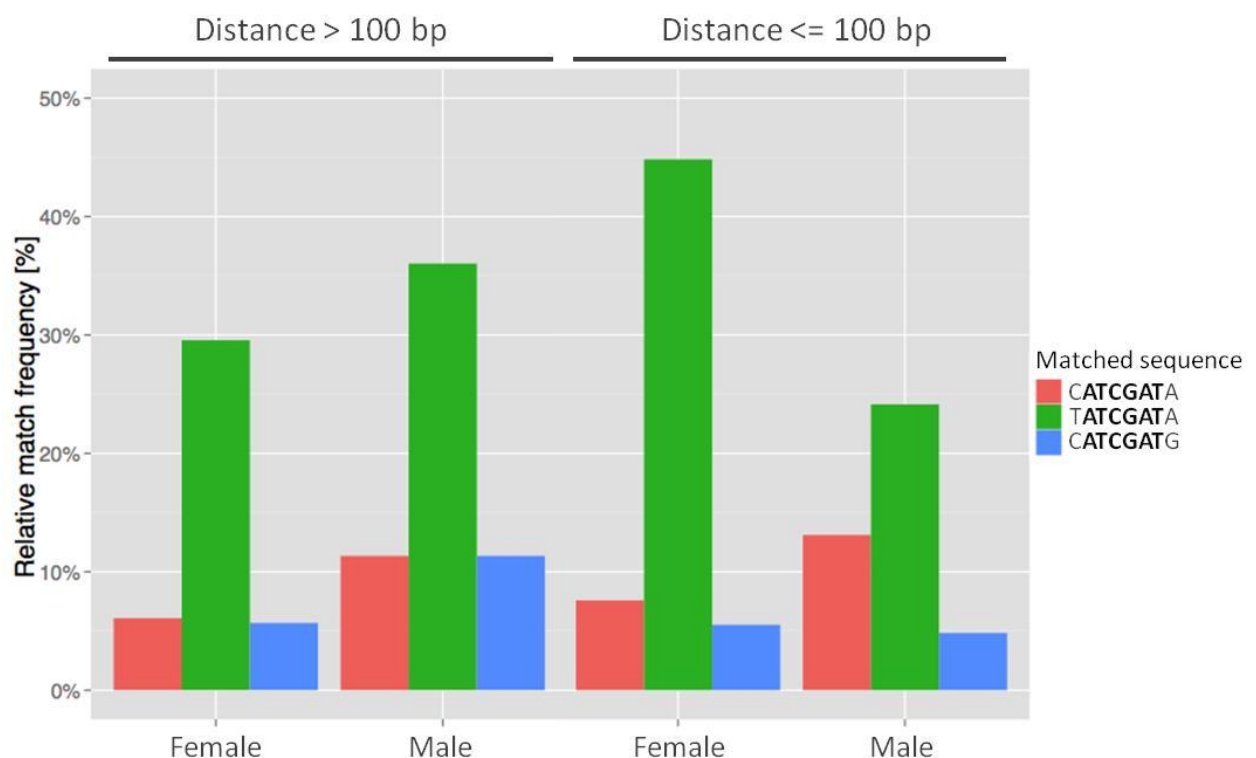


Figure 10 Sequences matched by the matrix describing the motif of interested according to its distance from the transcription initiation site and the sex-bias of the genes (Figure 8)

The frequency distribution of the female to male expression ratio for the genes whose upstream sequence contains at least one instance of the motif also varied according to the distance from transcription start site (Figure 11). The degree of bias (frequencies of extreme ratio) did not appear to be related the location of the target sequence.

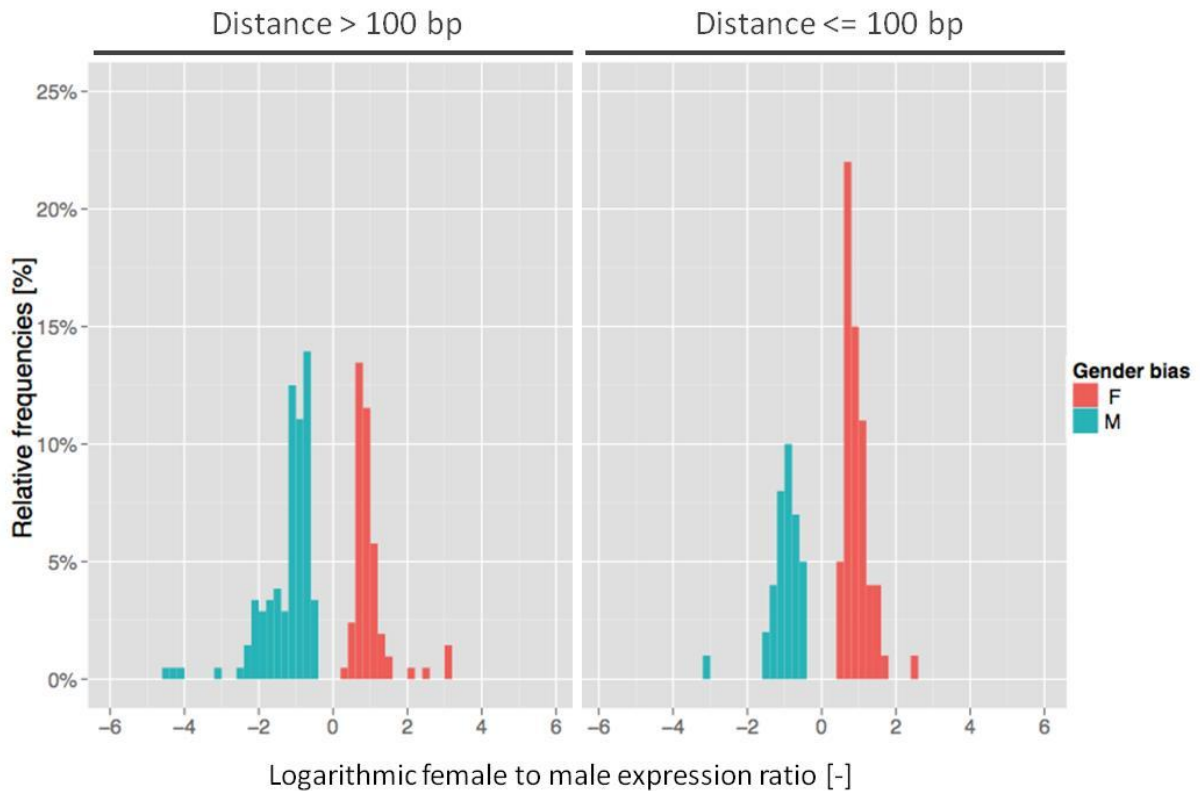


Figure 11 Relative frequencies distributions according to the distance of the motif from transcription initiation site

1336 of the sex-biased genes had orthologs in 11 of the *Drosophila* species whose genome had been sequenced. Among those, 793 had at least one ortholog with the motif present in their 1000 bp upstream region (Figure 12).

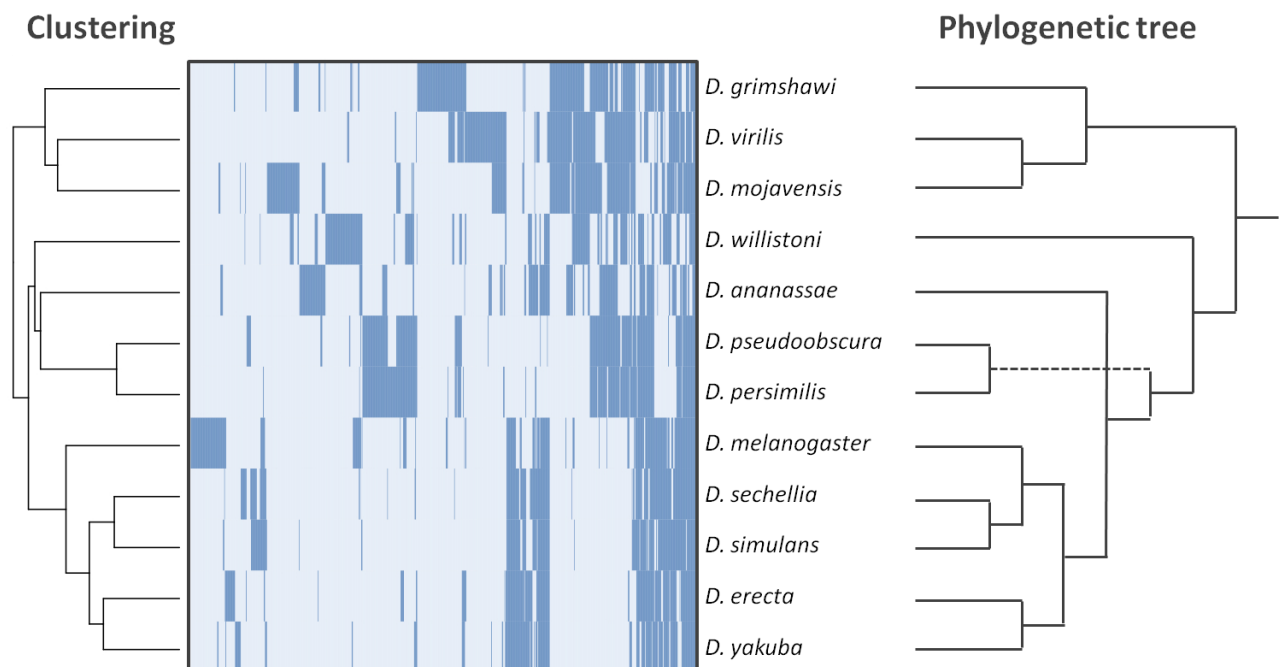


Figure 12 Matrix showing whether genes (x axis) have an ortholog from a *Drosophila* species (y axis) with the motif in its 1000bp upstream region (dark blue=at least a match, light blue=no match). The dendrogram on the left is an automatic clustering, the dendrogram on the right is the phylogenetic tree of the sequenced *Drosophila* species⁵²

It is interesting to note that while not identical, the dendrogram resulting from automatic clustering of species according to whether genes have a hit or not was similar to the phylogenetic tree of *Drosophila* with most of the groups remaining intact. A more in-depth analysis of that conservation was possible using the gender-bias experimental data¹ for six of the related *Drosophila* species (Figure 13). As it was observed for *Drosophila melanogaster*, the proportion of female-biased genes with at least one hit for the target sequence was remarkably higher than that of genes presenting a male-biased expression.

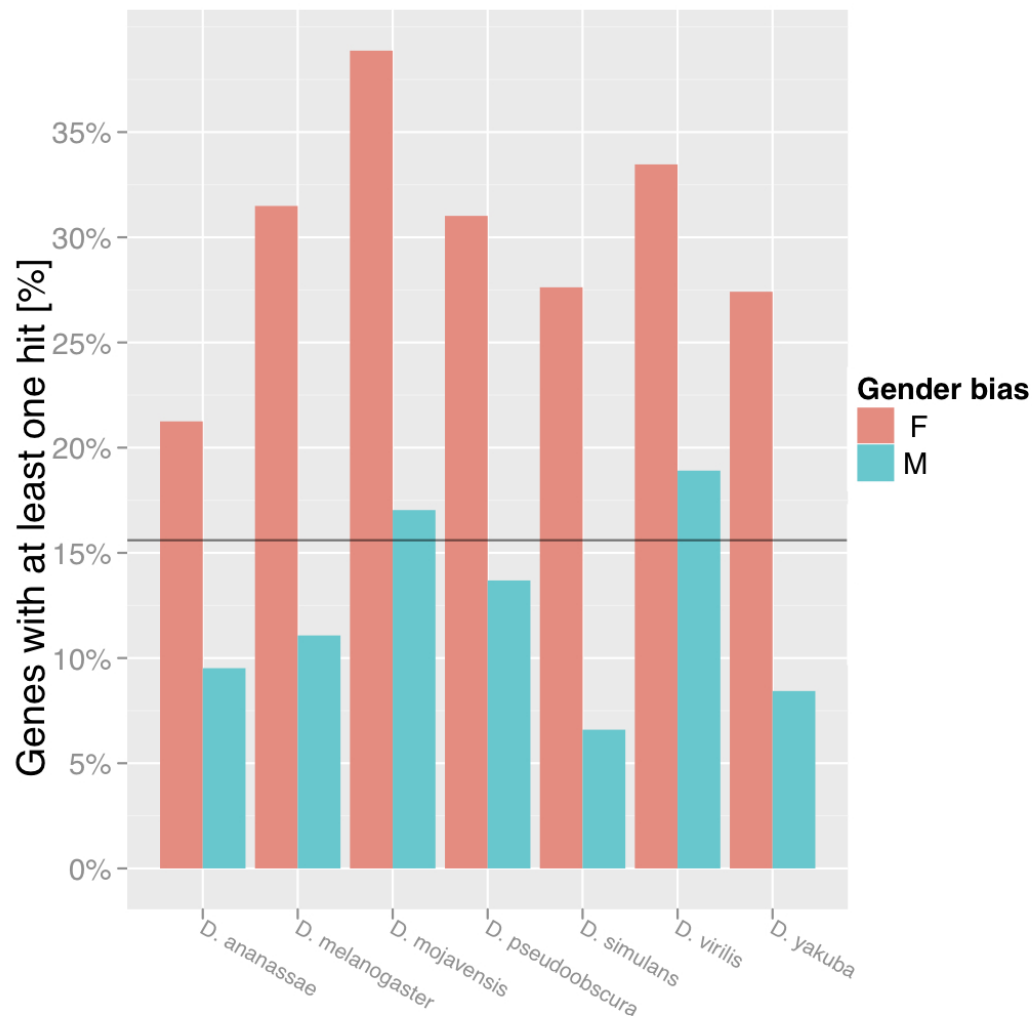


Figure 13 Proportions of sex-biased genes with at least one hit in their 1000 bp upstream region. The horizontal line is the mean of the random distribution generated by picking random sets of 1000 genes in the *Drosophila melanogaster* genome

Using the pattern matching results together with the FlyAtlas database⁵¹, it was possible to qualitatively determine whether parts of the fly's body where genes are expressed could be related with the presence of the motif in their upstream region (Figure 14). When only genes with a positive match for the motif in their upstream region are taken into account, an overall shift can be observed compared to the situation where all genes are considered: fewer genes were favorably expressed in testis and the number of instances where genes were mainly expressed in ovaries increased.

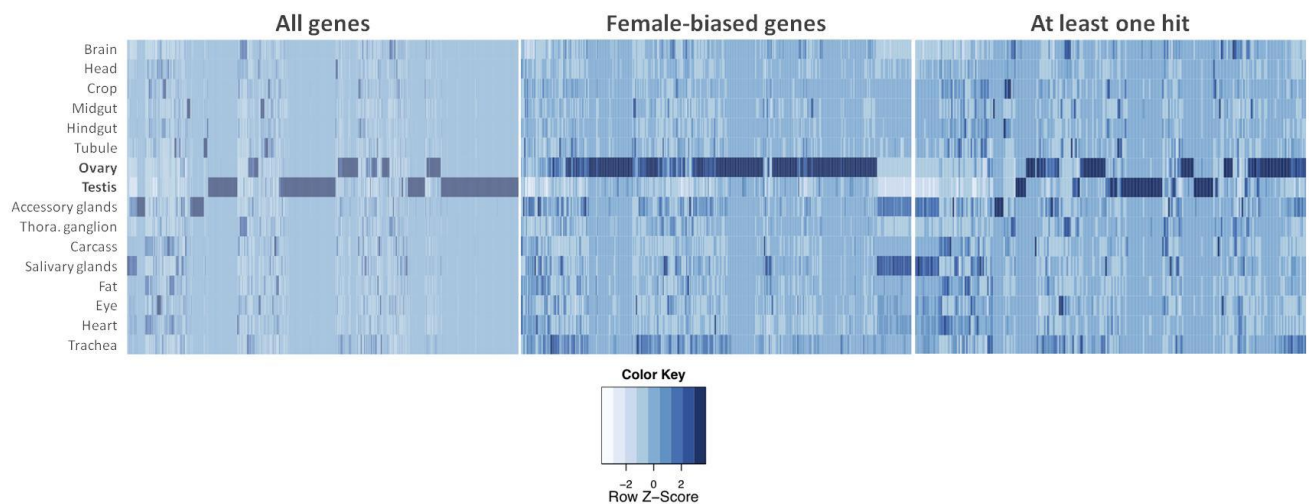


Figure 14 Heat map showing the body part of the fly (y axis) where genes (x axis) are favorably expressed. The intensity of color represents the z-score (distance from the mean in term of standard deviation) normalized per row. (expression data extracted from FlyAtlas⁵¹)

TOMTOM was used to compare the motif with known *Drosophila* regulatory elements (Figure 15.A). The matches with the lowest p-values were the transcription factors BEAF-32 (32 kDa boundary element-association factor) and Dref (replication-related element binding factor). The motif actually corresponds to the previously reported DNA replication-related element (DRE) sequence (5'-TATCGATA)⁵³, a promoter-activating element that was found to regulate many mechanisms in *Drosophila*⁵⁴, such as the cell cycle⁵⁵, differentiation⁵⁶, catalase expression⁵⁷, wing vein development⁵⁸ or proteasome degradation activity⁵⁹.

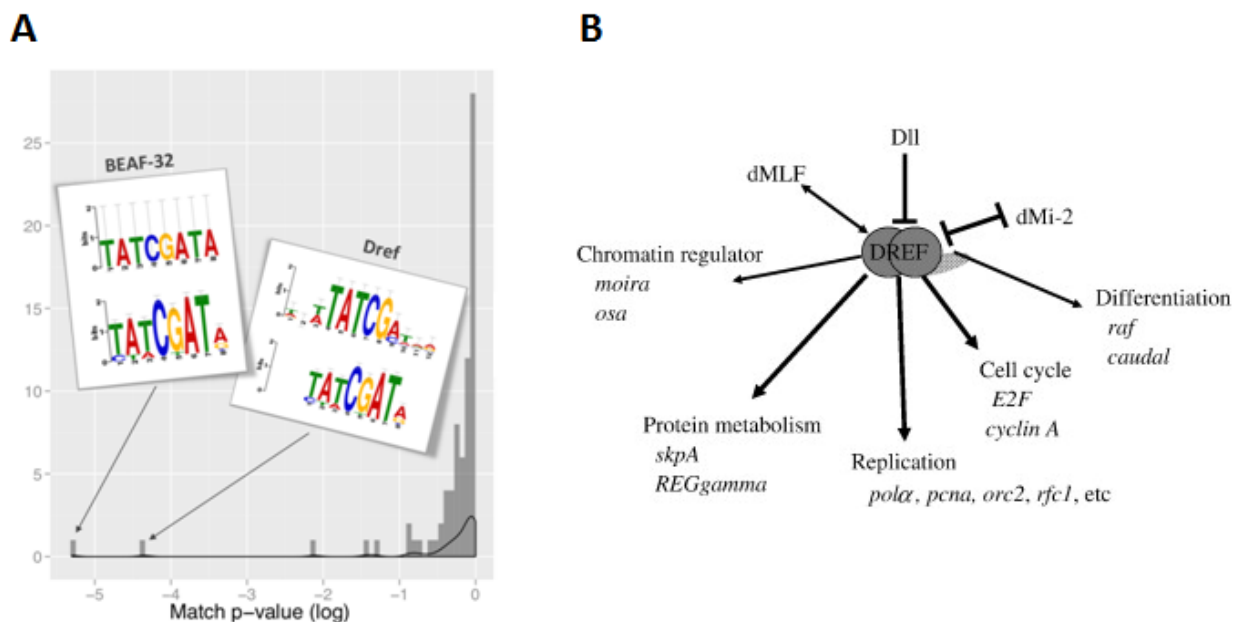


Figure 15 A Distribution of the match p-value when scanning the FlyReg database using the inferred motif. **B** DREF regulatory network (source: Matsukage *et al*⁵⁵)

4. Discussion

The motif was shown to be significantly overrepresented in the 5' flanking regions of female-biased genes both in *Drosophila melanogaster* (Figure 9.A) and other related species (Figure 13). In addition, the distribution of the distance between the motif and the transcription initiation site seemed to hint at a potential proximal regulatory role (Figure 9.B), which would imply that hits farther than 100 bp away from the start site would represent

background noise generated by the stochastic nature of the analysis. Moreover, the sequence preference in the region close to the promoter of female-biased genes (Figure 10) showed a core motif with a small degeneracy, which is characteristic of what is commonly observed in the case of transcription factor binding sequences⁶⁰. The motif was identified as a binding sequence shared by BEAF-32 and Dref (Figure 15.A). The latter seemed the best candidate as its target sequence, the DNA replication-related element, was extensively studied and characterized, in opposition to BEAF-32 for which little data is available, especially concerning its binding sequence.

It is yet still to be determined how DRE, a single regulatory element, could be responsible for the regulation of such a large number of seemingly unrelated functions (Figure 15.B). It has been proposed that instead of acting alone, Dref could interact with auxiliary factors to specifically target certain group of genes involved in a particular cellular function. Another explanation would be that Dref competes for binding with inhibitory chromatin factors⁵⁵. No sex-specific regulatory behavior has been reported in the literature yet although new targets are identified on a regular basis. Goldman and Arbeitman recently studied genes that are regulated by *transformer* and *doublesex* in adult head tissues⁶¹. While a direct comparison is difficult, as it would require more tissue-specific expression data, little to no overlaps were found with the results presented in this document. This might indicate that the sex-specific regulatory mechanism described here might be a completely separate molecular process.

It has been shown that regulatory sequences tend to be conserved among closely related species^{52, 62}. If the motif is indeed a binding sequence for a regulatory molecule, it might also be found in the upstream region of orthologous genes from other *Drosophila* species. The degree of conservation of the motif seemed to be similar to what should have been expected, given the phylogenetic tree of *Drosophila* (Figure 12).

However, some results seemed contradictory. Even though the proportion of male-biased genes with the motif in their upstream region was lower when compared to that of genes with a female-biased expression, the distributions of the distance from the transcription initiation site were strikingly similar for both genders (Figure 9.B). In addition, some genes with a positive match were found to be expressed in testis and male accessory glands (Figure 14). This could be explained by some genes being incorrectly labeled as sex-biased.

Looking at the effect of the distance from the initiation site on the distribution of the logarithmic expression ratios (Figure 11), the results seemed to confirm that hits within 100 bp of the genes promoters corresponded to actual regulatory sites, promoting the expression of female-biased genes, whereas distant sites do not have a noticeable effect on the distributions and could be seen as background noise. Again, the residual male-biased gene expression in the case of proximal hits might be due to the way the bias was experimentally determined.

The sequence preference of distal hits for male-biased genes (Figure 10) was markedly different when compared to that of female-biased genes, which could signify that the motifs weren't identical, similar to the small variations between BEAF-32 and Dref binding sequences⁶³. This might explain some of the inconsistency in the results, both sequences could be detected by the matrix (Figure 8) but only one would have a sex-specific regulatory role.

5. Conclusion

Common motif discovery pitfalls⁶⁴ seem to have been avoided (rigorous control sets, links with known biological entities). Further investigations would be necessary to establish the biological relevance of these results. For instance, tissue-specific gene expression data for both sexes would allow determining if these findings remain true when the genital organs experimental bias is eliminated. Repeating the analysis using other sex-biased datasets might also help explain some of the contradictory results. It is also yet to be established in what measure the hypothetical sex-specific regulatory function of Dref/CRE interacts with known sex-determination mechanisms.

6. References

1. Zhang, Y., et al., *Constraint and turnover in sex-biased gene expression in the genus Drosophila*. Nature, 2007. **450**(7167): p. 233-7.
2. Wyckoff, G.J., W. Wang, and C.I. Wu, *Rapid evolution of male reproductive genes in the descent of man*. Nature, 2000. **403**(6767): p. 304-9.
3. Koopman, P. and K.A. Loffler, *Sex determination: the fishy tale of Dmrt1*. Curr Biol, 2003. **13**(5): p. R177-9.
4. Cline, T.W., *Evidence that sisterless-a and sisterless-b are two of several discrete "numerator elements" of the X/A sex determination signal in Drosophila that switch Sxl between two alternative stable expression states*. Genetics, 1988. **119**(4): p. 829-62.
5. Younger-Shepherd, S., et al., *deadpan, an essential pan-neural gene encoding an HLH protein, acts as a denominator in Drosophila sex determination*. Cell, 1992. **70**(6): p. 911-22.
6. Bell, L.R., et al., *Sex-lethal, a Drosophila sex determination switch gene, exhibits sex-specific RNA splicing and sequence similarity to RNA binding proteins*. Cell, 1988. **55**(6): p. 1037-46.
7. Ryner, L.C. and B.S. Baker, *Regulation of doublesex pre-mRNA processing occurs by 3'-splice site activation*. Genes Dev, 1991. **5**(11): p. 2071-85.
8. Li, H. and B.S. Baker, *Her, a gene required for sexual differentiation in Drosophila, encodes a zinc finger protein with characteristics of ZFY-like proteins and is expressed independently of the sex determination hierarchy*. Development, 1998. **125**(2): p. 225-35.
9. Garrett-Engle, C.M., et al., *intersex, a gene required for female sexual development in Drosophila, is expressed in both sexes and functions together with doublesex to regulate terminal differentiation*. Development, 2002. **129**(20): p. 4661-75.
10. Demir, E. and B.J. Dickson, *fruitless splicing specifies male courtship behavior in Drosophila*. Cell, 2005. **121**(5): p. 785-94.
11. Jin, W., et al., *The contributions of sex, genotype and age to transcriptional variance in Drosophila melanogaster*. Nat Genet, 2001. **29**(4): p. 389-95.
12. Ranz, J.M., et al., *Sex-dependent gene expression and evolution of the Drosophila transcriptome*. Science, 2003. **300**(5626): p. 1742-5.
13. Arbeitman, M.N., et al., *Gene expression during the life cycle of Drosophila melanogaster*. Science, 2002. **297**(5590): p. 2270-5.
14. Parisi, M., et al., *Paucity of genes on the Drosophila X chromosome showing male-biased expression*. Science, 2003. **299**(5607): p. 697-700.
15. Ellegren, H. and J. Parsch, *The evolution of sex-biased genes and sex-biased gene expression*. Nat Rev Genet, 2007. **8**(9): p. 689-98.
16. Auer, H., D.L. Newsom, and K. Kornacker, *Expression Profiling Using Affymetrix GeneChip Microarrays*. Methods Mol Biol, 2009. **509**: p. 35-46.
17. Wingren, C. and C.A. Borrebaeck, *High-throughput proteomics using antibody microarrays*. Expert Rev Proteomics, 2004. **1**(3): p. 355-64.
18. Han, J., et al., *Towards high-throughput metabolomics using ultrahigh-field Fourier transform ion cyclotron resonance mass spectrometry*. Metabolomics, 2008. **4**(2): p. 128-140.
19. Mockler, T.C., et al., *Applications of DNA tiling arrays for whole-genome analysis*. Genomics, 2005. **85**(1): p. 1-15.
20. Berger, M.F. and M.L. Bulyk, *Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors*. Nat Protoc, 2009. **4**(3): p. 393-411.
21. Buck, M.J. and J.D. Lieb, *ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments*. Genomics, 2004. **83**(3): p. 349-60.
22. Park, P.J., *ChIP-seq: advantages and challenges of a maturing technology*. Nat Rev Genet, 2009. **10**(10): p. 669-80.
23. IUPAC-IUB commission on biochemical nomenclature (CBN). *Abbreviations and symbols for nucleic acids, polynucleotides and their constituents*. J Mol Biol, 1971. **55**(3): p. 299-310.
24. Bosco, G., et al., *Analysis of Drosophila species genome size and satellite DNA content reveals significant differences among strains as well as between species*. Genetics, 2007. **177**(3): p. 1277-90.
25. Ometto, L., D. De Lorenzo, and W. Stephan, *Contrasting patterns of sequence divergence and base composition between Drosophila introns and intergenic regions*. Biol Lett, 2006. **2**(4): p. 604-7.
26. Van Helden, J., *Pattern discovery String-based approaches*. 2009, Bruxelles: Pattern discovery String-based approaches.
27. Hertz, G.Z. and G.D. Stormo, *Identifying DNA and protein patterns with statistically significant alignments of multiple sequences*. Bioinformatics, 1999. **15**(7-8): p. 563-77.

28. Stormo, G.D. and D.S. Fields, *Specificity, free energy and information content in protein-DNA interactions*. Trends Biochem Sci, 1998. **23**(3): p. 109-13.
29. Schneider, T.D. and R.M. Stephens, *Sequence logos: a new way to display consensus sequences*. Nucleic Acids Res, 1990. **18**(20): p. 6097-100.
30. Yeung, K.Y., M. Medvedovic, and R.E. Bumgarner, *From co-expression to co-regulation: how many microarray experiments do we need?* Genome Biol, 2004. **5**(7): p. R48.
31. van Helden, J., B. Andre, and J. Collado-Vides, *Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies*. J Mol Biol, 1998. **281**(5): p. 827-42.
32. Thijs, G., et al., *A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling*. Bioinformatics, 2001. **17**(12): p. 1113-22.
33. Turatsinze, J.V., et al., *Using RSAT to scan genome sequences for transcription factor binding sites and cis-regulatory modules*. Nat Protoc, 2008. **3**(10): p. 1578-88.
34. van Helden, J., A.F. Rios, and J. Collado-Vides, *Discovering regulatory elements in non-coding sequences by analysis of spaced dyads*. Nucleic Acids Res, 2000. **28**(8): p. 1808-18.
35. Pavesi, G., G. Mauri, and G. Pesole, *An algorithm for finding signals of unknown length in DNA sequences*. Bioinformatics, 2001. **17** **Suppl 1**: p. S207-14.
36. Sinha, S. and M. Tompa, *YMF: A program for discovery of novel transcription factor binding sites by statistical overrepresentation*. Nucleic Acids Res, 2003. **31**(13): p. 3586-8.
37. Hertz, G.Z., G.W. Hartzell, 3rd, and G.D. Stormo, *Identification of consensus patterns in unaligned DNA sequences known to be functionally related*. Comput Appl Biosci, 1990. **6**(2): p. 81-92.
38. Van Helden, J., *Matrix-based pattern discovery algorithms*. 2009, Bruxelles: Université Libre de Bruxelles.
39. Bailey, T.L. and C. Elkan, *Fitting a mixture model by expectation maximization to discover motifs in biopolymers*. Proc Int Conf Intell Syst Mol Biol, 1994. **2**: p. 28-36.
40. Aitkin, M. and D. Rubin, *Estimation and Hypothesis Testing in Finite Mixture Models*.
41. Lawrence, C.E., et al., *Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment*. Science, 1993. **262**(5131): p. 208-14.
42. Roth, F.P., et al., *Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation*. Nat Biotechnol, 1998. **16**(10): p. 939-45.
43. Liu, X., D.L. Brutlag, and J.S. Liu, *BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes*. Pac Symp Biocomput, 2001: p. 127-38.
44. Defrance, M. and J. van Helden, *info-gibbs: a motif discovery algorithm that directly optimizes information content during sampling*. Bioinformatics, 2009. **25**(20): p. 2715-22.
45. Wei, Z. and S.T. Jensen, *GAME: detecting cis-regulatory elements using a genetic algorithm*. Bioinformatics, 2006. **22**(13): p. 1577-84.
46. Gentleman, R.C., et al., *Bioconductor: open software development for computational biology and bioinformatics*. Genome Biol, 2004. **5**(10): p. R80.
47. van Helden, J., *Regulatory sequence analysis tools*. Nucleic Acids Res, 2003. **31**(13): p. 3593-6.
48. Down, T.A. and T.J. Hubbard, *NestedMICA: sensitive inference of over-represented motifs in nucleic acid sequence*. Nucleic Acids Res, 2005. **33**(5): p. 1445-53.
49. Bergman, C.M., J.W. Carlson, and S.E. Celniker, *Drosophila DNase I footprint database: a systematic genome annotation of transcription factor binding sites in the fruitfly, Drosophila melanogaster*. Bioinformatics, 2005. **21**(8): p. 1747-9.
50. Gupta, S., et al., *Quantifying similarity between motifs*. Genome Biol, 2007. **8**(2): p. R24.
51. Chintapalli, V.R., J. Wang, and J.A. Dow, *Using FlyAtlas to identify better Drosophila melanogaster models of human disease*. Nat Genet, 2007. **39**(6): p. 715-20.
52. Stark, A., et al., *Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures*. Nature, 2007. **450**(7167): p. 219-32.
53. Hirose, F., et al., *Isolation and characterization of cDNA for DREF, a promoter-activating factor for Drosophila DNA replication-related genes*. J Biol Chem, 1996. **271**(7): p. 3930-7.
54. Hirose, F., et al., *Drosophila Mi-2 negatively regulates dDREF by inhibiting its DNA-binding activity*. Mol Cell Biol, 2002. **22**(14): p. 5182-93.
55. Matsukage, A., et al., *The DRE/DREF transcriptional regulatory system: a master key for cell proliferation*. Biochim Biophys Acta, 2008. **1779**(2): p. 81-9.
56. Choi, Y.J., et al., *Transcriptional regulation of the Drosophila caudal homeobox gene by DRE/DREF*. Nucleic Acids Res, 2004. **32**(12): p. 3734-42.
57. Park, S.Y., et al., *Transcriptional regulation of the Drosophila catalase gene by the DRE/DREF system*. Nucleic Acids Res, 2004. **32**(4): p. 1318-24.
58. Yoshida, H., et al., *DREF is required for EGFR signalling during Drosophila wing vein development*. Genes Cells, 2004. **9**(10): p. 935-44.

59. Masson, P., J. Lundgren, and P. Young, *Drosophila* proteasome regulator REGgamma: transcriptional activation by DNA replication-related factor DREF and evidence for a role in cell cycle progression. J Mol Biol, 2003. **327**(5): p. 1001-12.
60. Zhang, C., et al., A clustering property of highly-degenerate transcription factor binding sites in the mammalian genome. Nucleic Acids Res, 2006. **34**(8): p. 2238-46.
61. Goldman, T.D. and M.N. Arbeitman, Genomic and functional studies of *Drosophila* sex hierarchy regulated gene expression in adult head and nervous system tissues. PLoS Genet, 2007. **3**(11): p. e216.
62. Sosinsky, A., et al., Discovering transcriptional regulatory regions in *Drosophila* by a nonalignment method for phylogenetic footprinting. Proc Natl Acad Sci U S A, 2007. **104**(15): p. 6305-10.
63. Yoshida, H., et al., Over-expression of DREF in the *Drosophila* wing imaginal disc induces apoptosis and a notching wing phenotype. Genes Cells, 2001. **6**(10): p. 877-86.
64. Lusk, R.W. and M.B. Eisen, Evolutionary mirages: selection on binding site composition creates the illusion of conserved grammars in *Drosophila* enhancers. PLoS Genet, 2010. **6**(1): p. e1000829.