Automated generation of context-specific gene regulatory networks with a weighted approach in *Drosophila melanogaster*

Leandro Murgas\textsuperscript{1,4,†}, Sebastian Contreras-Riquelme\textsuperscript{1,2,†}, J. Eduardo Martínez-Hernandez\textsuperscript{1,3,4}, Camilo Villaman\textsuperscript{1,4}, Rodrigo Santibáñez\textsuperscript{1} and Alberto J. M. Martin\textsuperscript{1}

\textsuperscript{1}Laboratorio de Biología de Redes, Centro de Genómica y Bioinformática, Facultad de Ciencias, Universidad Mayor, Santiago 8580745, Chile
\textsuperscript{2}Facultad de Ciencias de la Vida, Universidad Andres Bello, Santiago 8370146, Chile
\textsuperscript{3}Centro de Modelamiento Molecular, Biofísica y Bioinformática—CM2B2, Facultad de Ciencias Químicas y Farmacéuticas, Universidad de Chile, Santiago 8380492, Chile
\textsuperscript{4}Programa de Doctorado en Genómica Integrativa, Vicerrectoría de Investigación, Universidad Mayor, Santiago, Chile

The regulation of gene expression is a key factor in the development and maintenance of life in all organisms. Even so, little is known at whole genome scale for most genes and contexts. We propose a method, Tool for Weighted Epigenomic Networks in *Drosophila melanogaster* (Fly T-WEoN), to generate context-specific gene regulatory networks starting from a reference network that contains all known gene regulations in the fly. Unlikely regulations are removed by applying a series of knowledge-based filters. Each of these filters is implemented as an independent module that considers a type of experimental evidence, including DNA methylation, chromatin accessibility, histone modifications and gene expression. Fly T-WEoN is based on heuristic rules that reflect current knowledge on gene regulation in *D. melanogaster* obtained from the literature. Experimental data files can be generated with several standard procedures and used solely when and if available. Fly T-WEoN is available as a Cytoscape application that permits integration with other tools and facilitates downstream network analysis. In this work, we first demonstrate the reliability of our method to then provide a relevant application case of our tool: early development of *D. melanogaster*. Fly T-WEoN together with its step-by-step guide is available at https://weon.readthedocs.io.

1. Introduction

The regulation of gene expression is indispensable for adaptation to ever changing contexts and every aspect involved in sustaining life. Gene regulation is mainly carried out by highly specialized proteins, among which transcription factors (TFs) are generally accepted as the key actors [1]. Canonically speaking, the regulation of gene expression works through the binding of TFs to certain sites in the chromatin, TF binding sites (TFBSs), and TFs recognize specific DNA patterns called TF binding motifs. These sites are usually specific for each TF and they are commonly located around the promoter of TF-target genes upstream of their transcription start site. Whereas proximal upstream locations of TFBSs are easily related to the regulation of specific genes [2,3], to determine which genes are controlled by each TF binding to enhancer regions has shown a greater difficulty [4–6]. Moreover, gene expression can be defined as the process by which the final products encoded by genes are
generated, and thus their regulation can also include control of translation and RNA degradation. In this way, several other non-TF regulatory elements are involved in the regulation of gene expression. For example, miRNAs and other ncRNAs are known to act during translation by binding to other RNAs [7,8], while histone modifiers attach or remove post-translational modifications to control the positions of the chromatin that are available to be occupied by TFs.

Several epigenetic marks, including histone modifications [9] and DNA methylation [10], have been related to active and inactive states of chromatin [11,12], therefore influencing the ability of TFs to regulate gene expression. In this way, combinations of epigenetic marks have been related to a specific effect on TF binding and gene expression, coining an epigenetic code that is still not properly understood [9,13]. Even so, there are some generally accepted facts on the relationship between TF binding and epigenetic marks that have made it possible to grasp a general tendency [14]. Nonetheless, chromatin structure and epigenetic marks change dynamically in a context-specific manner, and those changes have been subject to both static and dynamic modelling to predict gene expression [15].

Despite the relationship between epigenetic marks and gene regulation, the determination of the chromatin state for each TFBS remains experimentally difficult and expensive, while computational inference from limited experimental evidence is common in the literature. For instance, CENTIPEDE [16] is probably one of the first computational methods aiming to decipher which TFBS are actually bound at certain experimental condition instead of just defining TFBS from databases such as JASPAR [17]. CENTIPEDE makes use of DNase-seq data in an unsupervised learning algorithm to infer which TFBS are in an open active state and can compare its results with experimental data. Currently, computational analysis has at its disposal several tools to process experimental data related to gene regulation from which choosing is not an easy task. Nonetheless, some collaborative projects employ reliable pipelines, e.g. the TCGA workflow [18] or the ENCODE data processing pipelines (https://github.com/ENCODE-DCC). Often, those computational tools do not provide an intuitive interface, relying entirely on command-line instructions and/or do not report figures to interpret results from such data. For example, CENTIPEDE is a R package and, therefore, requires a minimum coding expertise. Moreover, there are other tools such as Anchor, a Python package [19], Mocap, a Python and R hybrid package [20], and TEPIC, a C++ program [21]. All these methods aim to determine DNA occupancy by TFs, but require expertise from users in compiling, installing dependencies, coding and the use of the command-line interfaces.

To overcome these difficulties, we created an efficient and easy to use method, Tool Weighted EpigenOmic Network (Fly T-WEoN), that is able to generate Drosophila melanogaster context-specific gene regulatory networks (GRNs). This method employs a series of filters, that once applied to a reference network, remove TF–gene regulations that are unlikely taking place according to current knowledge on the relationship between epigenetic and TFBS activation. Specificity on resulting networks is provided by the time and context for which the omic data employed by each filter were generated. Our tool is available as a Cytoscape application that provides a user-friendly and intuitive interface where researchers easily introduce their data processed with standard protocols to generate context-specific GRNs.

Figure 1. Flowchart describing Fly T-WEoN. The TF–gene reference network is filtered by DNA methylation, then by chromatin accessibility of regulatory sites and third by histone marks. TF–gene and miRNA–gene networks are then filtered according to RNA-seq expression of the regulators and edges in the resulting networks are then scored according to the number of filters passed and provided as edge weights in the context-specific GRN.

2. Methods

2.1. Construction of a reference gene regulatory network

A reference GRN is a network that contains all known regulatory interactions between gene products and genes, regardless of developmental stage, environment or cell type in an organism. To create a reference network for D. melanogaster, we combined TFBS information from the ENCODE data repository [22] and FlyBase [23] to then infer regulatory relationships based on distance of TFBSs to the transcription start site (TSS) of each gene in the genome of the fly version 6.32 (see electronic supplementary material, NetsInfo for details). To determine whether a TF regulates a gene, we chose distance thresholds between TFBSs and the TSS of each gene, so if the TFBS falls within this distance, we assumed it regulates the respective gene. We created three reference networks with different distance thresholds, 1500, 2000 and 5000 nucleotides inspired by other approaches [24]. In the case of miRNA, genetic relationships based on experimentally determined targets from miRecords [25] and miRTarBase [26] were also retrieved and incorporated into the reference networks.

2.2. Filtering the reference network

In order to determine which regulatory relationships are taking place in any experimental context of interest, we defined several filters, each relying on a different type of experimental data as input. The filtering process was implemented in PERL and is the backend software of the Cytoscape [27] application developed to provide a tool with a user-friendly interface. The filtering procedure generates a time- and tissue-specific GRN depending on the experimental condition in which experimental data used were generated.

Our method considers experimental information following this order for each TFBS: chromatin accessibility (DNase-seq), methylation of the DNA, histone modifications around the TFBS, the expression of each TF with known TFBSs in the reference network and miRNA quantification (figure 1). First, if there is a positive signal in the TFBSs for DNA methylation, Fly T-WEoN assumes that TF cannot bind its TFBS and the filter removes the regulation accordingly. Second, if chromatin
accessibility data, e.g. DNase-seq, show a positive signal within the chosen distance threshold used to assign a TF to the regulation of a gene, this indicates that a TF can bind the corresponding region and therefore the edge is not removed. The next filter considers if the chromatin is in open or closed state based on histone marks experimentally associated with this process. For example, trimethylation of the Histone H3 Lys27 [28,29] or trimethylation of the Histone H4 Lys20 [9,30,31] are marks associated with inactive chromatin. The effects of the histone marks considered by default in the histone marks filter are described in Table 1, and sequencing reports in BED format were used as provided in ENCODE and FlyBase (see electronic supplementary material, Data processing for a brief explanation of the protocols followed). Each of these filters takes in consideration if the epigenetic mark can be associated with one of the TFBS of each TF associated with the regulation of each gene. Finally, the last filter considers if the gene coding a regulator (TF or miRNA) is expressed; regulations emerging with one of the TFBS of each TF associated with the regulation takes in consideration if the epigenetic mark can be associated with the chosen distance threshold used to assign a TF to the regulatory region and therefore the edge is not removed. For example, trimethylation of the Histone H3 Lys27 [28,29] or trimethylation of the Histone H4 Lys20 [9,30,31] is marks associated with inactive chromatin. The effects of the histone marks considered by default in the histone marks filter are described in Table 1, and sequencing reports in BED format were used as provided in ENCODE and FlyBase (see electronic supplementary material, Data processing for a brief explanation of the protocols followed). Each of these filters takes in consideration if the epigenetic mark can be associated with the chosen distance threshold used to assign a TF to the regulatory region and therefore the edge is not removed.

2.2.1. Scoring edges
Fly T-WEoN assigns weights to edges in the resulting network. The weight of each edge is calculated by adding a score of one for each filter that the edge passes. By default, edges have no weight, so a weight of one means the edge passed only the expression filter, a weight of two means it passed an additional filter such as a histone mark, and a weight of three indicates that the edge passed the expression filter, and for example, two different histone modifications indicated its binding site was active.

2.3. Validation
To assess the reliability of GRNs generated with Fly T-WEoN, we used as gold standard a network created with all TF ChIP-seq experiments available in the ENCODE repository for the third instar larval stage or L3 of D. melanogaster. We chose this stage because there are experimental ChIP-seq data for 32 different TF (all already included in the list of ChIP-seq experiments employed to generate the reference networks) and for 10 histone marks as well as RNA-seq data. All experiments considered were carried out in equivalent conditions (see electronic supplementary material, NetsInfo for the list and IDs of experiments used). The gold-standard network was created by first removing edges from the reference network arising from genes coding for any regulator that is not among those 32 TFs, and second, by removing those edges whose TFBS was not occupied by its respective TF.

2.3.1. Network reliability: edges
We estimated the performance of Fly T-WEoN by considering the presence/absence of edges in the final network as a binary classification problem. In this set-up, a true positive (TP) is defined as an edge present in the context-specific network generated after applying the filters and in the gold-standard network. Similarly, a false negative (FN) edge is absent in the network generated by Fly T-WEoN but it is present in the gold standard, while a false positive (FP) edge is present in the network and absent in the gold standard. Importantly, true negatives (TNs) indicate edges absent in both the gold standard and in the network created by Fly T-WEoN. Finally, once all edges are assigned to either of the three types TF, FP or FN, they were used to calculate precision (P, equation (2.1)), recall (R, equation (2.2)) and F1 (equation 2.3), metrics that serve as indicators of the reliability of the context-specific networks. Each of these metrics has a value in the [0,1] range, with greater values indicating a better classification. To evaluate the effect of distance threshold, we also calculated the performance metrics using the reference networks generated using the three distance thresholds 1.5, 2 and 5 kb (see electronic supplementary material, NetsInfo).

2.3.2. Network reliability: local topology
GRNs are formed by combinations of graphlets, induced subgraphs that have been associated to specific functions [34]. Graphlets can be used to describe local topology of nodes in GRNs, and the presence or the absence of the graphlets in which a node participates indicate functional variation for that gene in two realizations of the same network [35]. In addition, the presence or absence of graphlets in two versions of the same network can be considered as a binary classification problem, and thus the same metrics calculated for edges indicate how similar is the local topology of each gene in the gold-standard network and in the predicted GRNs, or their overall topological similarity. We employed LoTo [36] to calculate precision, recall and the F1 metrics calculated for the presence/absence of graphlets in every pairwise network comparison. If these metrics only consider graphlets in which the same gene participates, they serve to indicate variations in the local topology of that node. Whereas, if the metrics are calculated for all graphlets in the networks, they serve to indicate global topological similarity between the two networks.

2.4. Fruit fly early embryo development
To demonstrate the utility of Fly T-WEoN, we generated networks for six different stages of early embryo development in fruit fly (D. melanogaster). We employed RNA-seq experiments and histone marks data downloaded from different databases such as modENCODE and modMine projects [22,37], and the FlyBase database [23] (see electronic supplementary material,
Table 2. Description of the reference networks employed in Fly T-WEoN. Reference networks were created by assigning TFs to the regulation of specific genes based on a distance threshold between the TFBS and the gene. All three networks described in the table include the same 350 TFs.

| threshold (kb) | genes   | edges   |
|---------------|---------|---------|
| 1.5           | 15 576  | 1 094 130 |
| 2             | 15 899  | 1 190 168 |
| 5             | 16 665  | 1 679 173 |

Table 3. Gold standard networks used to validate Fly T-WEoN. Networks made with the 32 TFs at different distance thresholds between TFBSs and the TSS of each gene. Number of different genes and edges present in each of the networks made by assigning a TF to the regulation of a gene if the TF is bound within the distance and the gene TSS. Percentages indicate the ratio of edges and genes present in these networks compared to the subnetworks made with all TFBS for the same 32 TFs.

| threshold (kb) | genes percent | edges percent |
|---------------|---------------|---------------|
| 1.5           | 83.96%        | 80.30%        |
| 2             | 84.75%        | 80.56%        |
| 5             | 89.19%        | 82.19%        |

Generated by our approach. The gold-standard network was made with 32 different TFs and their binding sites determined by ChIP-seq experiments in equivalent experimental conditions. The reference networks made at 1.5, 2 and 5 kb thresholds are described in table 3.

Not surprisingly, larger distance thresholds include more TF–gene interactions for genes to which we cannot assign regulators otherwise, and thus networks built using greater thresholds contain more nodes.

3.2.1. Network similarity: edges

Using the L3 example, the lowest score of edges in the predicted networks is two and the highest eleven. This is due to the number of Fly T-WEoN filtering steps applied, so a score of two implies that the TF from which an edge is originated is expressed and there is at least a single histone modification supporting its existence. Scores of three, and above, mean that there are at least two types of histone modification indicating that the link exists.

As shown in table 4 for a threshold of 1.5 kb, Fly T-WEoN generates networks with very high similarity to the gold-standard network in our benchmarking. Starting with edges of score two or greater, the network generated by Fly T-WEoN contains 97.8% of the edges of the gold-standard network ($R = 0.978$), decreasing the recall as the edges score increases. Also, the $F1$ value follows the same trend: it displays its highest value using this score ($F1 = 0.884$) and decreases as the minimum score for the edges increases. Moreover, the precision follows a different tendency, with its highest value with score $\geq 6$ ($P = 0.810$). The worst performance is obtained with a score of eleven, the maximum, with which Fly T-WEoN recovers 0.1% of the edges of the gold-standard network ($R = 0.001$, $P = 0.646$ and $F1 = 0.001$), indicating low similarity between edges present in the predicted networks and the gold standard.

3.2.2. Global topological similarity calculated with graphlets

The trend for graphlet based results is similar to that based on single edges, shown in table 5.

Using a minimum score of two, Fly T-WEoN is able to recover 95.7% of the graphlets found in the gold-standard network ($R = 0.957$), but it tends to overpredict graphlets as indicated by the much lower precision ($P = 0.662$). Also, the $F1$ value had its greatest value with a score of at least two ($F1 = 0.782$), indicating again high similarity between the predicted and gold-standard networks.

3. Results

3.1. Reference networks

The three reference networks provided as default in our tool are described in table 2.

As expected, increasing the cut-off employed to assign TFs based on the distance TFBS–TSS, the number of genes and edges in each reference GRN increases.

3.2. Method validation

We employed the L3 context-specific GRN described in the Methods section to estimate the reliability of the networks generated by our approach. The gold-standard network was made with 32 different TFs and their binding sites determined by ChIP-seq experiments in equivalent experimental conditions. The reference networks made at 1.5, 2 and 5 kb thresholds are described in table 3.

Not surprisingly, larger distance thresholds include more TF–gene interactions for genes to which we cannot assign regulators otherwise, and thus networks built using greater thresholds contain more nodes.
precision was obtained using a minimum score of five ($P = 0.665$), which also supports that networks obtained by Fly T-WEoN contain more graphlets than gold-standard networks, even at the maximum precision. The lowest values for the performance metrics were obtained using weights greater than or equal to 10, with predicted networks recovering 0% of the graphlets present in the gold-standard network.

### 3.3. An example case: fruit fly early development

#### 3.3.1. Network sizes

Six time-specific networks were created with Fly T-WEoN encompassing the fly embryonic development (0–24 h) in consecutive time ranges of 4 h (0–4 h, 4–8 h, 8–12 h, 12–16 h, 16–20 h and 20–24 h). These networks were made using available data of histone modifications and RNA-seq. These networks have different numbers of edges, graphlets, regulatory nodes (TFs) and total number of genes, as shown in table 6.

The largest network belongs to the 16–20 h time range, with the largest numbers for nodes, total connections and regulatory nodes (10 993, 928 599 and 345, respectively). The smallest network is the network for the 0–4 h time range, which has the lowest number of total connections, and regulatory nodes (718 583 and 305, respectively), while the network for time range 4–8 h has the lowest number of nodes (7886).

#### 3.3.2. Network comparisons

We compared each network with the network representing the next time interval obtaining $F_1$ values greater than 0.85 (table 7). These results indicate that despite changes, most of the regulatory network remains unaltered between time lapses.
Thus, indicating that relatively small changes in the network account for all stages of early embryo development.

We also analysed the $F_1$ values by types of genes, TF and non-TF coding and non-coding genes (table 8).

Thus, indicating that relatively small changes in the network account for all stages of early embryo development.

We also analysed the $F_1$ values by types of genes, TF and non-TF coding and non-coding genes (table 8).

Table 6. Characterization of embryo development networks. The table shows the number of edges and regulatory nodes for each of the networks created for the six time intervals during early development of the fruit fly. Regulatory nodes indicate the number of TFs in each network and the total number of genes and edges in the networks are also displayed. These networks were obtained by removing unlikely edges from a reference network were TFBSs located at most at 1.5 kb upstream the TSS are used to assign the TFs that bind to that TFBS to the regulation of each gene.

| node type       | 0–4 h | 4–8 h | 8–12 h | 12–16 h | 16–20 h | 20–24 h |
|-----------------|-------|-------|--------|---------|---------|---------|
| total edges     | 718 583 | 733 863 | 803 613 | 888 537 | 928 599 | 840 567 |
| TF nodes        | 305   | 324   | 340    | 335     | 345     | 339     |
| total nodes     | 8811  | 7886  | 8554   | 10 528  | 10 993  | 11 146  |

Table 7. Comparisons of embryo development networks using graphlets. The table displays the number of true positive graphlets (TP), graphlets in the first network (belonging to the earlier time interval) that are present in the later network; false positive graphlets (FP), those present in the later network but absent in the earlier one; and false negative graphlets (FN), those that are only present in the earlier network and not in the later network. TP, FP and FN graphlets were used to calculate precision ($P$), recall ($R$) and $F_1$ metrics.

| comparison   | 0–4 to 4–8 h | 4–8 to 8–12 h | 8–12 to 12–16 h | 12–16 to 16–20 h | 16–20 to 20–24 h |
|--------------|--------------|---------------|-----------------|------------------|------------------|
| TP           | 960 856 575  | 154 840 150   | 170 821 683     | 0.849            | 0.861            | 0.855 |
| FP           | 1 015 103 337 | 237 272 519   | 100 593 375     | 0.910            | 0.811            | 0.857 |
| FN           | 1 162 138 649 | 358 797 508   | 90 237 194      | 0.928            | 0.764            | 0.838 |
| FP           | 1 347 047 992 | 264 431 780   | 173 888 152     | 0.886            | 0.836            | 0.860 |
| FN           | 1 297 280 051 | 74 185 782    | 314 199 708     | 0.805            | 0.946            | 0.870 |

Table 8. Total number of genes by type and $F_1$ interval in each of the comparisons of embryo development consecutive networks using graphlets. The table displays the number of genes in each of the four $F_1$ intervals $[0.0, 0.5)$, $[0.5, 0.7)$, $[0.7, 0.9)$ and $[0.9, 1.0]$ in each of the five comparisons performed between GRNs depicting gene regulation at each time lapse. $F_1$ values closer to 0 indicate larger local topological variation, while closer to 1 indicate fewer variations in the graphlets in which a gene participates.

| comparison       | 0–4 to 4–8 h | 4–8 to 8–12 h | 8–12 to 12–16 h | 12–16 to 16–20 h | 16–20 to 20–24 h |
|------------------|--------------|---------------|-----------------|------------------|------------------|
| all genes        | [0.0, 0.5)   | [0.5, 0.7)    | [0.7, 0.9)      | [0.9, 1.0)       | [0.0, 0.5)       |
| TFs              | [0.0, 0.5)   | [0.5, 0.7)    | [0.7, 0.9)      | [0.9, 1.0)       | [0.0, 0.5)       |
| coding genes     | [0.0, 0.5)   | [0.5, 0.7)    | [0.7, 0.9)      | [0.9, 1.0)       | [0.0, 0.5)       |
| non-coding genes | [0.0, 0.5)   | [0.5, 0.7)    | [0.7, 0.9)      | [0.9, 1.0)       | [0.0, 0.5)       |

Thus, indicating that relatively small changes in the network account for all stages of early embryo development.

Without considering gene type (all genes), most of them are in the $F_1$ ranges with less topological variation ($[0.7, 0.9]$ and $[0.9, 1.0]$), evidencing that, as happened with global topology, the local topology of a majority of genes...
remains unaltered between consecutive time lapses. The same trend is displayed by the TF-coding genes, with most of them in the range [0.7, 0.9). With respect to non-TF coding genes, again most of them fall into F1 interval ranges with less topological variation ([0.7, 0.9) and [0.9, 1.0]). Notably, there are large proportions of ncRNA

| comparison         | GO term                                           | GO ID           | fold enrichment | p-value    |
|--------------------|---------------------------------------------------|-----------------|-----------------|------------|
| 0–4 to 4–8 h       | cell differentiation                              | GO0030154       | 2.29            | 1.14 × 10⁻⁴|
|                    | developmental process                             | GO0032502       | 2.02            | 1.49 × 10⁻²|
|                    | cellular developmental process                    | GO0048869       | 2.15            | 3.12 × 10⁻⁴|
|                    | sulfur compound metabolic process                 | GO0006790       | 2.53            | 1.06 × 10⁻³|
|                    | anatomical structure development                  | GO0048856       | 1.92            | 1.41 × 10⁻³|
|                    | cellular modified amino acid metabolic process    | GO0006575       | 2.94            | 1.85 × 10⁻³|
|                    | glutathione metabolic process                     | GO0006749       | 3.52            | 2.40 × 10⁻³|
|                    | cell fate commitment                              | GO0045165       | 4.82            | 3.54 × 10⁻³|
|                    | neurogenesis                                      | GO0022008       | 2.38            | 4.14 × 10⁻³|
|                    | generation of neurons                             | GO0048699       | 2.38            | 5.81 × 10⁻³|
|                    | multicellular organisal process                   | GO0032501       | 1.58            | 9.06 × 10⁻³|
| 4–8 to 8–12 h      | developmental process                             | GO0032502       | 1.87            | 1.15 × 10⁻³|
|                    | cell differentiation                              | GO0030154       | 2.04            | 2.07 × 10⁻³|
|                    | chaperone-mediated protein folding                | GO0061077       | 4.18            | 3.16 × 10⁻³|
|                    | cellular developmental process                    | GO0048869       | 1.91            | 4.46 × 10⁻³|
|                    | anatomical structure development                  | GO0048856       | 1.79            | 6.33 × 10⁻³|
| 8–12 to 12–16 h    | cellular modified amino acid metabolic process    | GO0006575       | 4.18            | 2.61 × 10⁻⁹|
|                    | glutathione metabolic process                     | GO0006749       | 5.00            | 2.04 × 10⁻⁸|
|                    | cofactor metabolic process                        | GO0051186       | 2.24            | 3.25 × 10⁻⁵|
|                    | sulfur compound metabolic process                 | GO0006790       | 2.32            | 1.10 × 10⁻⁴|
|                    | response to drug                                  | GO0042493       | 2.92            | 1.39 × 10⁻³|
|                    | organic acid metabolic process                    | GO0006082       | 1.66            | 2.81 × 10⁻³|
|                    | small molecule metabolic process                  | GO0044281       | 1.46            | 3.85 × 10⁻³|
|                    | transmembrane transport                           | GO0055085       | 1.55            | 6.24 × 10⁻³|
|                    | carboxylic acid metabolic process                 | GO0019752       | 1.62            | 6.36 × 10⁻³|
|                    | organic anion transport                           | GO0015711       | 2.05            | 6.50 × 10⁻³|
|                    | oxoacid metabolic process                         | GO0043436       | 1.60            | 6.65 × 10⁻³|
|                    | aminoglycan metabolic process                     | GO0006022       | 3.43            | 6.94 × 10⁻³|
|                    | anion transport                                   | GO0006820       | 1.84            | 8.37 × 10⁻³|
|                    | defence response                                  | GO0006952       | 3.25            | 8.89 × 10⁻³|
| 12–16 to 16–20 h   | aminoglycan metabolic process                     | GO0006022       | 3.98            | 7.78 × 10⁻⁴|
|                    | amino sugar catalytic process                     | GO0046348       | 3.62            | 2.28 × 10⁻³|
|                    | chemical synaptic transmission                    | GO0007268       | 2.09            | 2.73 × 10⁻³|
|                    | anterograde trans-synaptic signalling             | GO0098916       | 2.09            | 2.73 × 10⁻³|
|                    | response to drug                                  | GO0042493       | 2.64            | 3.05 × 10⁻³|
|                    | synaptic signalling                               | GO0099536       | 2.06            | 4.36 × 10⁻³|
|                    | trans-synaptic signalling                         | GO0099537       | 2.06            | 4.36 × 10⁻³|
|                    | multicellular organisal process                   | GO0032501       | 1.41            | 7.76 × 10⁻³|
| 16–20 to 20–24 h   | aminoglycan metabolic process                     | GO0006022       | 4.25            | 7.65 × 10⁻⁴|
|                    | amino sugar catalytic process                     | GO0046348       | 4.25            | 7.65 × 10⁻⁴|
|                    | drug metabolic process                            | GO0017144       | 1.85            | 5.51 × 10⁻³|
coding genes in the range that displays larger topological variations, hinting they play a relevant role in the developmental stages depicted by the networks. Detailed information on which genes show greater variation on their local topology and the GRN for each time point can be found in the electronic supplementary material (file LoTo_Embryo and EmbryoNetworks, respectively).

3.3.3. Functional analysis of genes with altered local topology

After performing comparisons of networks representing consecutive developmental stages, we analysed the function of genes with altered local topology. To do so, we employed the statistical enrichment of GO-Slim Biological Process terms with PANTHER [42] for genes in each of F1 ranges previously defined.

Focusing on the analysis of genes with F1 in the range [0–0.5), the enrichment test denoted several GO terms that are known to be involved in embryonic development (table 9).

For example, we found enriched GO terms ‘developmental process’ and ‘anatomical structure development’ in genes in the lowest F1 range in the comparisons spanning the first 12 h (0–4 h to 4–8 h and 4–8 to 8–12 h). In the comparisons spanning the last 12 h, we found enriched functional terms related to metabolism and metabolite transport processes such as ‘glutathione metabolic process’, ‘transmembrane transport’ and ‘aminoglycan metabolic process’. Genes in intervals with moderate topological variation (F1 range [0.5–0.7); see electronic supplementary material, GO file) showed enrichment in GO terms related to defence response, metabolic, and developmental process. For the comparison of 4–8 h and 8–12 h networks, genes in this F1 range, enriched terms were ‘animal organ development’, ‘cytoplasmic translation’, and ‘cell development’. In the case of the comparison 8–12 to 12–16 h, enriched terms associated with cell signalling, GO terms ‘signalling’ and ‘cell communication’. Finally, for the comparison of 12–16 to 16–20 h, overrepresented terms were related with cell structure and

Figure 2. Comparison of subnetworks composed of all those genes showing larger local variation in the 0–4 h to 4–8 h comparison. The network shown is formed by 594 nodes (44 TFs) and 3107 edges coloured according to their existence in the earlier network, in the later network or in both.
cell cycle, GO Slim terms such as ‘establishment of spindle orientation’ and ‘cell cycle’. In the case of the comparison 16–20 to 20–24 h no GO term was significantly enriched.

3.3.4. Subnetworks of nodes showing largest topological variations at early stages

To further investigate the application of our approach to the early embryo development example, we created subnetworks made of only those nodes that have $F_1$ in the $[0.0,0.5)$ range for each comparison. We then compared subnetworks depicting consecutive stages using LoTo. As an example we show the comparison of the two earlier stages (0–4 to 4–8 h) in figures 2 and 3, the results of the comparison showing only TFs. All these subnetworks can be found as a Cytoscape session in the electronic supplementary material.

3.3.5. Dorsal–ventral patterning in the 0–4 h network

We first looked in the three reference networks for the cascade governed by dorsal, finding that 42 of its 61 known edges were present in all three reference networks, and that three edges were only present in the 5 kb network (figure 4a). When analysing the 16 absent regulations, we saw that there are three causes: missing TFBS or in other words in the set of experimentally determined TFBSs there are no sites near certain genes as happened with $brk \rightarrow tld$ and $mad \rightarrow tsg$; TFBS that are further away than the distance cut-offs we employed to assign TF to the regulation of genes but yet can be found at 8–20 kb from the gene TSS, as happened for regulations $brk \rightarrow zen$, $brk \rightarrow pnr$, $ind \rightarrow msh$, $med \rightarrow shn$, $sna \rightarrow sim$, $sna \rightarrow ths$, $tin \rightarrow eve$ and $zen \rightarrow tup$; and TFBS that are close to the TSS but downstream, as happened for regulations $brk \rightarrow twi$, $sna \rightarrow vnd$, $sna \rightarrow vn$, $sna \rightarrow ind$, $sna \rightarrow sog$ and $twi \rightarrow sim$. Next, we determined if this subnetwork was present in the filtered networks, identifying 37 of the 45 regulations that would be happening in the 0–4 h period in the 5 kb network (37 out of 42 for the 1.5 and 2 kb networks; figure 4b). After examination of available epigenomic data, we saw that missing edges were not related to any epigenetic mark indicating active TFBS or were solely linked to a single peak belonging to mark related to inactive TFBS.

4. Discussion

Inference of gene regulation relationships from genomic data is a particularly hard and costly task. This is due to the use of high quality antibodies to determine the bound state of TFs to the open chromatin. Even with aid of computational tools, the determination of gene regulations is an open problem contributed by a gap knowledge of how TFs and other regulators of gene expression work, and by a general lack of genomic data.
suitable for the prediction of such regulations. The inexpensive RNA-seq and chromatin accessibility through footprint sequencing are commonly used to infer condition-specific networks, but these still require corroboration that again, is usually made with comparisons to ChIP-seq experiments of each TF. However, TF ChIP-seq experiments are unavailable for most conditions of model organisms, including even those that have been deeply studied. Importantly, the numbers of ChIP-seq used to determine histone modifications are increasing in data repositories, and given the relationship between histone marks and TF binding in chromatin [9,13,14], we created FlyT-WEoN to generate context-specific GRNs in D. melanogaster.

With the proposed methodology, we first built reference networks based on three distance thresholds of 1.5, 2 and 5 kb between TFBSs and TSS of genes (described in table 2). These networks were then used as the starting point to build condition-specific GRNs for the L3 developmental stage and used them to validate our methodology based on the concatenation of simple filters. We employed ChIP-seq for ten different histone marks and 32 different TFs to build gold-standard networks to then compare them with FlyT-WEoN networks. Each of the filters in our tool uses current knowledge on the known relationship that exists between epigenetic marks and TF activity. We observed that even if the filtering approach may seem to be too simple it still recovers correctly most of the edges found in the gold-standard network we made for that stage. Furthermore, FlyT-WEoN applies a weight system on edges, increasing these weights according to how many filters did each edge pass. Our results show that, at least in our test, using a weight of greater than or equal to 2 produces the most reliable GRNs. This weight means that at least one histone mark and the expression of the TF agrees with the existence of each edge. The worse performance shown with greater weights can be explained due to that by increasing the weight value, the number of edges and graphlets in the networks decrease. However, using only edges with greater weights decreases the reliability of the edges (tables 4 and 5). Which suggests that the known effect of different epigenetic marks is contradictory, and thus our simple filtering approach fails to gather the complexity of the epigenetic code.

To highlight the differences and similarities between FlyT-WEoN and other approaches, we report a brief comparison between FlyT-WEoN and four other methods in table 10. The other methods used for the comparison were CENTIPEDE [16], Anchor [19], TEPIC [21] and Mocap [20]. It is

Figure 4. Dorsal–ventral patterning in the 0–4 h network. (a) shows the conservation of the dorsal cascade in the reference networks, while (b) displays the same subnetwork in the 0–4 h time interval using the 5 kb distance threshold. TFs are shown in orange, non-TF nodes are depicted in blue, dark green edges indicate gene regulations present in all three reference networks, light green edges are regulations present only in the reference network of 5 kb, and red edges are gene regulations that were not detected by the procedure employed to generate all reference networks (a) or did not pass all filters in the 0–4 h network.
important to stress that none of these methods was designed or even tested for *D. melanogaster*, and thus a quantitative comparison is not straightforward. Given the heterogeneous data employed by these methods, the absence of actual context-specific GRNs, and the lack of specific tools for *D. melanogaster*, it is not possible to perform quantitative comparisons between them, and thus only qualitative comparisons are possible. Our comparison (table 10) highlights the main characteristics of Fly T-WEoN, i.e. the intuitive way to use Fly T-WEoN and the integration of its results in Cytoscape, when compared with the other four approaches. It is very important to highlight that these tools use different types of data (table 10) in dissimilar context to those used by Fly T-WEoN. This makes it even more difficult to make a quantitative performance comparison between them. Also, there are no context-specific data available for all data types used by Fly T-WEoN (DNase, RNAseq, DNA methylation and TFs ChIP-seq), which does not allow for a full comparison.

Regarding the example of embryonic development of *D. melanogaster*, we created 6 different networks, each depicting transcriptional control by TFs for each of the four-hour intervals of the first 24 h of a fly embryo. We opted for this condition and time intervals because these were the conditions for which there are more epigenetic and transcriptional data at modENCODE and GEO datasets. Importantly, the stages represented by our networks are when cells and tissues in *D. melanogaster* are more homogeneous, and thus all omic data employed are deemed to be more significant. When comparing these GRNs with LoTo [36], we observed that the networks increase the number of nodes and connections as development progresses. This may indicate that in later stages of development transcriptional regulation becomes a more complex process that involves a greater number of TFs in greater number of cell fates and tissues. Comparisons of overall similarity between networks representing consecutive time intervals showed that the largest variation takes place

---

**Table 10. Qualitative comparison of different methods and Fly T-WEoN.** The table indicates for each tool the language used in its implementation, its purpose, its advantages and disadvantages and general user-friendliness.

| tool        | language | purpose                                      | input data                                                                 | (dis) advantages                                                                 | GUI |
|-------------|----------|----------------------------------------------|-----------------------------------------------------------------------------|---------------------------------------------------------------------------------|-----|
| CENTIPEDE   | R        | infers bound TFBS from Chip-seq of histone modifications and DNase-seq | matrix of read counts around motif matches based on DNAseq or ChIP reads and the following prior information: PWM score for motif matches represented in the genome, conservation score based on evolutionary information of motif and motif distance to TSS | easy to run and very intuitive to generate output data; however it needs many previous steps of data preprocessing to generate the correct input file | no  |
| Anchor [19] | Python   | predicts in vivo TF bindings profiles across cell types | genomic coordinates (BED file), DNase-seq data (BAM file and BigWig file), DNA sequence (genome fasta file), TFs motifs and Gencode GFF file | needs various preprocessing steps of all data (long times, computing intensive), then, it is easy to run | no  |
| Tepic [21]  | C++, R, Python | prediction and analysis of TFBS from epigenetic data, supporting more than 30 species | genome sequence in fasta file, genome annotation file (GTF) | easy to run; however, the output is not friendly for posterior analysis and requires post-processing | no  |
| Mocap [20]  | R, Python | classification of TFBSs from integration of chromatin accessibility, motif scores, TF footprints, CpG/GC content, evolutionary conservation | DNase-Seq or ATAC-Seq counts, BigWig, motif matrix | low time consuming, but it requires high computing performance. It is easy to run, but it is only available for mouse and human and the output requires post-processing | no  |
| Fly T-WEoN  | Perl, Java | apply filters from different genomic and epigenomic experiments to a reference network in order to generate context-specific GRNs | BED files from histone PTMs, methylation sequencing, DNA accessibility sequencing and RNA-seq file of counts, RPKM, or FPKM | the major advantage is the possibility to generate a context-specific GRN without further preprocessing of data in a friendly way. However it is only implemented for fly | yes |
between 8–12 h and 12–16 h networks and that the overall topology of the networks changes less in the last transition between developmental stages included, i.e. in the comparison of 16–20 h and 20–24 h networks.

With respect to variations in the local topology of single nodes determined by F1 calculated for the presence/absence of graphlets, most genes had small variations in all comparisons, a trend observed for all gene types in the networks (TFs, protein coding and ncRNAs). The only exception is ncRNA coding genes, mainly IncRNAs, which are almost as numerous in the F1 range that indicates largest topological variation as in the range depicting the lowest variation. These findings agree with previous knowledge on the role played by IncRNA in D. melanogaster development [43]. Regarding our observation of relatively few TFs displaying large variations in their local topology, and that those with larger changes (lowest F1) are densely linked between them, these findings agree with the concept of clusters of master regulators [44]. In this concept, a small cluster of highly interconnected TFs are the master regulators controlling the other regulators whose function is to act as effectors or ‘fine-tuners’ of the orders given by the master regulators.

In our example, regarding the master regulator concept, the ‘fine-tuners’ would be regulatory nodes found in the F1 ranges with higher values that are linked to the master regulators and to many other genes that do not code for regulators. Nonetheless, it should also be considered that especially at the earlier stages of the embryo, there are many TFs that are inherited from the mother [45], and given that our approach uses as approximation for TF activity the expression of their coding genes, maternal TFs are disregarded. The fact of observing an increasing number of nodes as the networks depict later stages also agrees with known facts regarding developments, as tissues and specialized cells appear, both regulators and non-regulator genes tend to perform more specialized functions [45]. Our functional analysis also corroborates this (table 9), more general functions related at the earlier stages and more specialized functions as development progresses, validating again the networks generated with Fly T-WEoN.

Considering the subnetwork guiding dorsal–ventral patterning, we have shown how our approach is able to recover most known regulatory events that are involved in this process. For example, regulatory interactions arising from snai or by brk were almost all missed by our approach to construct the reference network. Regarding the effect of all filters employed on this example, in the same way as the overall performance estimation made with the L3 network, we also looked at how well inferred is this network in the filtered 0-4h interval (figure 4b), showing how from those edges found in the regulatory network, only those involving TFBSs related to no marks or to a single negative mark were missing in the contextualized subnetwork.

5. Conclusion

Here, we demonstrated the reliability of our tool, Fly T-WEoN, with results indicating that most of the regulatory events depicted by edges in its resulting networks are likely taken place. In addition to this validation, and given the current lack of tools that integrate epigenetic data for the construction of GRNs in D. melanogaster, we also provided a qualitative comparison with other approaches, helping in this way to stress the usability of our method. The minimum input required by Fly T-WEoN is a quantification of the expression of genes, but the results we show here prove how the quality of the network improves by using other epigenetic data or quantification of miRNAs.

We finally demonstrated through a case study the usefulness of genomic data to filter out known regulations from a reference network and make context-specific gene regulatory networks where functions of genes with varying regulation correlate with the development stage. Moreover, we developed a Cytoscape app for Fly T-WEoN that serves as a tool for the presented method, allowing users to create and visualize context-specific GRNs from their processed RNA-seq, DNase-seq, bisulphite-seq, and ChIP-seq datasets or data obtained from public databases. We expect to further develop a backend software harnessing machine learning algorithms that would allow final users to predict gene expression from minimal and cheap genomic data, and extend the current method from fruit fly to other model organisms, specially human.

References

1. Lambert SA et al. 2018 The human transcription factors. Cell 172, 650–665. (doi:10.1016/j.cell.2018.07.029)
2. García-Alonso L, Holland CH, Ibrahim MM, Turei D, Saez-Rodriguez J. 2019 Benchmark and integration of resources for the estimation of human transcription factor activities. Genome Res. 29, 1363–1375. (doi:10.1101/gr.240663.118)
3. Liu Z-P, Wu C, Miao H, Wu H. 2015 RegNetwork: an integrated database of transcriptional and post-transcriptional regulatory networks in human and mouse. Database 2015, 2015, ba095. (doi:10.1093/database/bav095)
4. Fishilevich S et al. 2017 GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. Database 2017, ba028. (doi:10.1093/database/bax028)
5. Krivega I, Dean A. 2012 Enhancer and promoter interactions-long distance calls. Curr. Opin. Genet. Dev. 22, 79–85. (doi:10.1016/j.gde.2011.11.001)
6. Yang J, Corces VG. 2012 Insulators, long-range interactions, and genome function. Curr. Opin.
