ReadMe file for ‘DmLc3.flybase.ids.redfly.tfs.xlsx’

The workbook ‘DmLc3.flybase.ids.redfly.tfs.xlsx’ contains the results of two distinct sets of biological significance analyses. They are:

* TRANSFAC Analysis
* DAVID Analyses (the results of these analyses are not presented in the main paper due to space constraint)

How these analyses are performed and how to interpret their results are discussed below.

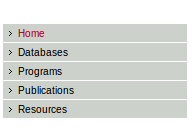
# TRANSFAC Analysis

The goal of this analysis is to compare the predicted behaviours (direct regulatees, downstream regulatees, co-factors etc.) of a given Transcription Factor (TF) with its known behaviours. The predicted behaviours are retrieved from the four predicted Gene Regulatory Networks (GRNs), one for each stage of the Drosophila melanogaster (Dm) lifecycle, that are learnt from the DmLc3 dataset (containing 588 genes). On the other hand, the known behaviours are retrieved from TRANSFAC (http://gene-regulation.com/pub/databases.html) version 7.0 Public database (hereafter, the TRANFAC database).

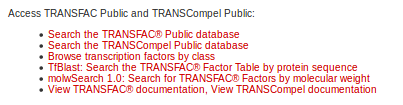
TRANFAC claims to be the ‘The gold standard in the area of transcriptional regulation’ (source: (http://genexplain.com/transfac/ , last accessed on 28th Sep, 2017). There are 233 TFs corresponding to Dm in the TRANSFAC database. One webpage is dedicated for each of the TFs. In each webpage, literary published behaviours (if any) of the corresponding TF are recorded in plain English along with the reference publication(s). For each TF, extracting its known behaviours, that are relevent to the current analysis, requires manually reading the whole webpage and finding out the relevent ones. This process is profoundly time consuming to perform for 233 TFs (i.e. 233 webpages). Instead, a shorter list of only 25 TFs are generated by the following process: The ‘flynet’ article [Ref: Predictive regulatory models in Drosophila melanogaster by integrative inference of transcriptional networks. Marbach D, Roy S, Ay F, Meyer PE, Candeias R, Kahveci T, Bristow CA, Kellis M. Genome Research, 22(7):1334-49, 2012] provides a list of 73 FlyBase (http://flybase.org/) IDs corresponding to 73 known Dm TFs (hereafter, REDfly TFs, because these TFs are extracted from the REDfly (http://redfly.ccr.buffalo.edu/) database by the authors of the flynet article) (file: ‘redfly\_tfs.txt’. It is downloaded from https://storage.googleapis.com/google-code-archive-downloads/v2/code.google.com/flynet/flynet\_v1.1beta.zip , last accessed on 28th Sep, 2017). The next step is to identify the REDfly TFs that are available in the predicted GRNs. But the 588 genes in the predicted GRNs are not identified by their FlyBase IDs; instead they are indentified by their gene symbols. Therefore, the gene symbols are converted to their corresponding FlyBase IDs using the online FlyBase converter (http://flybase.org/static\_pages/downloads/IDConv.html). Only 536 gene symbols out of 588 are converted successfully to their corresponding FlyBase IDs (the steps followd in the conversion process is documented in the worksheet ‘FlyBase\_IDs.xlsx’). Intersection of these 536 FlyBase IDs with the 73 REDfly TF FlyBase IDs produces a list of 25 distinct FlyBase IDs representing 25 known TFs in Dm.

The result of the comparison between the predicted behaviours and the known behaviours of these 25 TFs are documented in the worksheet ‘TRANSFAC Analysis’. Each row represents a TF. To reach the TRANSFAC webpage corresponding to a particular TF, please follow the steps mentioned below:

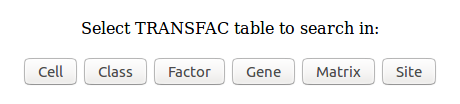
TRANSFAC web interface requires the user to register (create) a free account and log in to access the database. Once logged in, please select the ‘Databases’ menu.



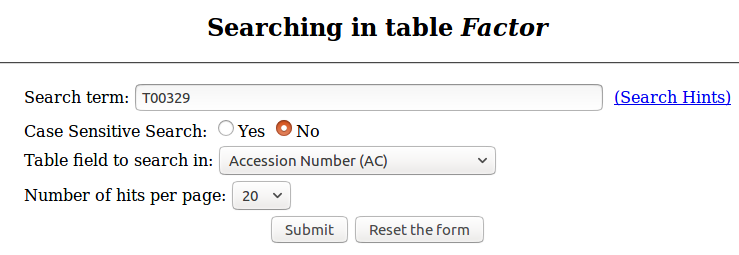
Then click on the link ‘Search the TRANSFAC® Public database’.



Then press the ‘Factor’ button to search in the Transcription Factor (TF) table.



Then copy the TRANSFAC Accession Number of the TF from the ‘TF’s TRANSFAC Accession No.’ column and paste in the ‘Search term’ textbox. Select ‘Accession Number (AC)’ from the ‘Table field to search in:’ drowdown box. After that, press the ‘submit’ button. It will take you to the TRANSFAC webpage corresponding to the given TF.



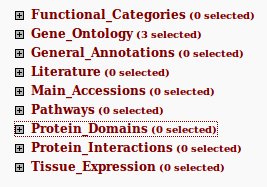
# Four DAVID (https://david.ncifcrf.gov/summary.jsp) Functional Annotation Clustering Analyses:

## Analysis 1: (for the E stage)

‘output20170908045031/di.subnet.adj.matrix.flybase.20170908045031.RData’ is produced by subsetting the E stage predicted rolled GRN ‘output/di.net.adj.matrix.20170908045031.RData’. It is done by retaining only the 536 genes for which FlyBase IDs can be found among the total 588 genes. The gene symbol to FlyBase ID conversion process is described in Section ‘TRANSFAC Analysis’.

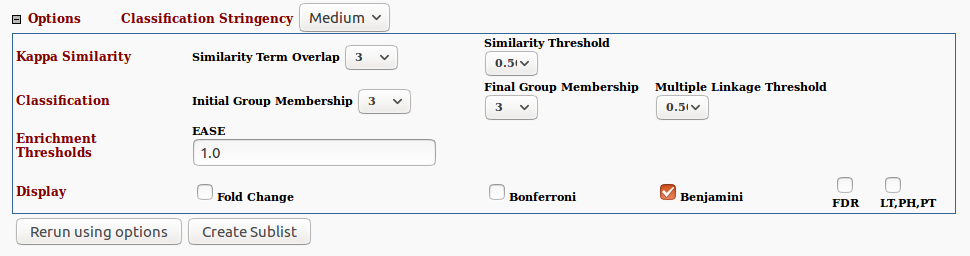
‘output20170908045031/di.subnet.adj.matrix.flybase.20170908045031.RData’ contains 536 genes and 1051 edges. The genes with top three out-degrees are FBgn0000351 (out-degree = 19), FBgn0003525 (out-degree = 18) and FBgn0003719 (out-degree = 17). For each of these three genes, a gene list is prepared that includes the gene itself and its predicted targets. Then these three separate lists are combined through union operation to produce a single gene list. This final list contains 43 gtenes (worksheet ‘david.genelist.E’). In DAVID (version 6.8), ‘Functional Annotation’ tool is selected and this list is uploaded as the ‘Gene List’. The list of all 536 genes (worksheet ‘david.bg.genelist’) is uploaded as the ‘Background’. For uploading of both the ‘Gene List’ and the ‘Background’, ‘FLYBASE\_GENE\_ID’ is chosen from the ‘Select Identifier’ dropdown box. It results in display of ‘Annotation Summary Results’. From the results, all already checked checkboxes are unchecked. Subsequently, three checkboxes titled ‘GOTERM\_BP\_DIRECT’, ‘GOTERM\_CC\_DIRECT’ and ‘GOTERM\_MF\_DIRECT’ under the heading ‘Gene\_Ontology’ are checked. Thus, the results section looks like Figure 1:

**Figure 1**



After that, the ‘Functional Annotation Clustering’ button is clicked to produce the results as recorded in the worksheet ‘david.results.E’. The ‘Options’ and the ‘Classification Stringency’ are kept unchanged i.e. at their default values as shown in Figure 2.

**Figure 2**



## Analysis 2: (for the L stage)

‘output20170913180657/di.subnet.adj.matrix.flybase.20170913180657.RData’ is produced by subsetting the L stage predicted rolled GRN ‘output20170913180657/di.net.adj.matrix.20170913180657.RData’. It is done by retaining only the 536 genes for which FlyBase IDs can be found among the total 588 genes. The gene symbol to FlyBase ID conversion process is described in Section ‘TRANSFAC Analysis’.

‘output20170913180657/di.subnet.adj.matrix.flybase.20170913180657.RData’ contains 536 genes and 133 edges. The genes with top three out-degrees are FBgn0266098 (out-degree = 12), FBgn0004921 (out-degree = 11), FBgn0003502 (out-degree = 10) and FBgn0259984 (out-degree = 10). For each of these four genes, a gene list is prepared that includes the gene itself and its predicted targets. Then these four separate lists are combined through union operation to produce a single gene list. This final list contains 47 gtenes (worksheet ‘david.genelist.L’). In DAVID (version 6.8), ‘Functional Annotation’ tool is selected and this list is uploaded as the ‘Gene List’. The list of all 536 genes (worksheet ‘david.bg.genelist’) is uploaded as the ‘Background’. For uploading of both the ‘Gene List’ and the ‘Background’, ‘FLYBASE\_GENE\_ID’ is chosen from the ‘Select Identifier’ dropdown box. It results in display of ‘Annotation Summary Results’. From the results, all already checked checkboxes are unchecked. Subsequently, three checkboxes titled ‘GOTERM\_BP\_DIRECT’, ‘GOTERM\_CC\_DIRECT’ and ‘GOTERM\_MF\_DIRECT’ under the heading ‘Gene\_Ontology’ are checked. Thus, the results section looks like Figure 1.

After that, the ‘Functional Annotation Clustering’ button is clicked to produce the results as recorded in the worksheet ‘david.results.L’. The ‘Options’ and the ‘Classification Stringency’ are kept unchanged i.e. at their default values as shown in Figure 2.

## Analysis 3: (for the P stage)

‘output20170913200300/di.subnet.adj.matrix.flybase.20170913200300.RData’ is produced by subsetting the P stage predicted rolled GRN ‘output20170913200300/di.net.adj.matrix.20170913200300.RData’. It is done by retaining only the 536 genes for which FlyBase IDs can be found among the total 588 genes. The gene symbol to FlyBase ID conversion process is described in Section ‘TRANSFAC Analysis’.

‘output20170913200300/di.subnet.adj.matrix.flybase.20170913200300.RData’ contains 536 genes and 649 edges. The genes with top three out-degrees are FBgn0002441 (out-degree = 20), FBgn0016794 (out-degree = 17), FBgn0021895 (out-degree = 15), FBgn0004110 (out-degree = 15) and FBgn0003149 (out-degree = 15). For each of these five genes, a gene list is prepared that includes the gene itself and its predicted targets. Then these five separate lists are combined through union operation to produce a single gene list. This final list contains 74 gtenes (worksheet ‘david.genelist.P’). In DAVID (version 6.8), ‘Functional Annotation’ tool is selected and this list is uploaded as the ‘Gene List’. The list of all 536 genes (worksheet ‘david.bg.genelist’) is uploaded as the ‘Background’. For uploading of both the ‘Gene List’ and the ‘Background’, ‘FLYBASE\_GENE\_ID’ is chosen from the ‘Select Identifier’ dropdown box. It results in display of ‘Annotation Summary Results’. From the results, all already checked checkboxes are unchecked. Subsequently, three checkboxes titled ‘GOTERM\_BP\_DIRECT’, ‘GOTERM\_CC\_DIRECT’ and ‘GOTERM\_MF\_DIRECT’ under the heading ‘Gene\_Ontology’ are checked. Thus, the results section looks like Figure 1.

After that, the ‘Functional Annotation Clustering’ button is clicked to produce the results as recorded in the worksheet ‘david.results.P’. The ‘Options’ and the ‘Classification Stringency’ are kept unchanged i.e. at their default values as shown in Figure 2.

## Analysis 4: (for the A stage)

‘output20170913203923/di.subnet.adj.matrix.flybase.20170913203923.RData’ is produced by subsetting the A stage predicted rolled GRN ‘output20170913203923/di.net.adj.matrix.20170913203923.RData’. It is done by retaining only the 536 genes for which FlyBase IDs can be found among the total 588 genes. The gene symbol to FlyBase ID conversion process is described in Section ‘TRANSFAC Analysis’.

‘output20170913203923/di.subnet.adj.matrix.flybase.20170913203923.RData’ contains 536 genes and 79 edges. The genes with top three out-degrees are FBgn0038755 (out-degree = 44), FBgn0004921 (out-degree = 23) and FBgn0003882 (out-degree = 6). For each of these three genes, a gene list is prepared that includes the gene itself and its targets. Then these three separate lists are combined through union operation to produce a single gene list. This final list contains 75 gtenes (worksheet ‘david.genelist.A’). In DAVID (version 6.8), ‘Functional Annotation’ tool is selected and this list is uploaded as the ‘Gene List’. The list of all 536 genes (worksheet ‘david.bg.genelist’) is uploaded as the ‘Background’. For uploading of both the ‘Gene List’ and the ‘Background’, ‘FLYBASE\_GENE\_ID’ is chosen from the ‘Select Identifier’ dropdown box. It results in display of ‘Annotation Summary Results’. From the results, all already checked checkboxes are unchecked. Subsequently, three checkboxes titled ‘GOTERM\_BP\_DIRECT’, ‘GOTERM\_CC\_DIRECT’ and ‘GOTERM\_MF\_DIRECT’ under the heading ‘Gene\_Ontology’ are checked. Thus, the results section looks like Figure 1.

After that, the ‘Functional Annotation Clustering’ button is clicked to produce the results as recorded in the worksheet ‘david.results.A’. The ‘Options’ and the ‘Classification Stringency’ are kept unchanged i.e. at their default values as shown in Figure 2.