Supplementary Materials and Methods to the manuscript

i-cisTarget: an integrative genomics method for the prediction of regulatory features and *cis*-regulatory modules

Carl Herrmann[1,2,*], Bram Van de Sande[3,*], Delphine Potier[1,3], and Stein Aerts[3]

1 TAGC – Inserm U928 and Aix-Marseille Université, Campus de Luminy, Marseille, France

2 Département de Biologie, Campus de Luminy, Aix-Marseille Université, Marseille, France

3 Laboratory of Computational Biology, Center for Human Genetics, Leuven, Belgium

***** These authors contributed equally to this work

Collections of motifs and iVEs used

The libraries of features that have been used for the analyses in this article are listed in table A. Several more recent libraries are available through the online application and can be used for analysis. This collection is summarized in table B.

Table A: Collection used for analysis in this article (feature database version 1.0)

Source name	**Version/Description**	**Number of features**
Motif Collection		*3731*
JASPAR (1)	3th release, downloaded 11/2010	1179
TRANSFAC professional (2)	01/2010	1300
FlyFactorSurvey (3)	19/11/2010	470
Tiffin (4)	Version 1.2, 01/2007	120
Article O. Elemento (5)	26/01/2005	371
Article A. Stark (6)	04/10/2007	232
SelexConsensus (4)	Version 1.1, 01/2007	59
In-vivo features		*420*
modENCODE features (7)	modENCODE features representing binding densities for transcription factors or chromatin-binding factors, as well as histone modifications, in various cell lines or developmental stages.	300
BDTNP features (8)	Berkeley Drosophila Transcriptional Network Program features, mainly binding locations for 20 sequence specific transcription factor involved in early drosophila development, as well as DNAse I hypersensitive sites (DHS).	45
Furlong features (9)	Features related to the binding locations for five TFs involved in mesoderm development, at various stages.	15
State features (7)	30 “chromatin state” features, based on combinations of chromatin marks in two different cell lines.	60
Total		*4151*

Table B: Collection available via web interface (feature database version 2.0)

Source name	Version	Number of features
Motif Collection		*6383*
JASPAR	3th release, downloaded 09/2011	1316 – 1 = 1315
There is some internal redundany in the jaspar collection, i.e. POL012.1 and MA0108.1 are identical. POL012.1 is therefore removed from the collection.		
TRANSFAC professional	02/01/11	1551
YeTFaSCo (10)	12/30/99	1709
FlyFactorSurvey	29/06/2011	614
hPDI (11)	22/12/2011	437
Tiffin	Version 1.2, 01/2007	120
Article O. Elemento	26/01/2005	371
Article A. Stark	04/10/2007	232 – 4 = 228
There are 4 consensus sequences identical to elements in Elemento (AAATCAAT, TATCGATA, TGACGTCA, TGGCGCC). These are removed from the Stark collection.		
SelexConsensus	Version 1.1, 01/2007	62 – 24 = 38
There are 24 duplicate motifs in this collection that are also present in TRANSFAC and or JASPAR. These are excluded form the library.		
ChIP TF binding sites		*109*
modENCODE		42
BTNP		41
Furlong		26
Non TF binding sites		*216*
modENCODE		208
BDTNP		7
Furlong		1
Histone modifications		*211*
modENCODE		205
BDTNP		0
Furlong		6
Total		*6919*

PWM library and motif clustering

The library of Position Weight Matrices (PWMs) used for the analyses in this article was compiled from JASPAR (1), TRANSFAC (2), FlyFactorSurvey (3) and Tiffin (4). These databases of known binding factor motifs were extended with conserved motifs derived from comparative genomics approaches (5, 6), resulting in a collection of 3731 PWMs. The library of motifs available via the web interface of i-cisTarget utilizes a larger collection of 6383 PWMs assembled from more recent versions of the aforementioned libraries (Table S1 and Supplementary Materials and Methods). To account for the motif redundancy in the output, we use a clustering procedure on the enriched motif features based on the tool STAMP (12), using the sum of squared distances (SSD). The optimal number of clusters is obtained based on the Calinski & Harabasz statistics. PWMs belonging to the same similarity cluster share a common color in the i-cisTarget web output.

Support for both gene signatures and genomic loci

Two types of data can be analyzed for regulatory feature enrichment via i-cisTarget: sets of related genes, referred to as signatures, and ChIP peak locations (or more general genomic loci). The former type of data can be provided as a simple list of genes separated by newline characters, the latter must be supplied in BED file format, defined and supported by the UCSC Genome Browser to visualize genomic loci/features. A typical BED file specifies each genomic feature in a separated line: the name of the chromosome the feature is located on, the start and end position of the feature on that chromosome and the name of the feature. All these fields must be separated by a whitespace character. The positions must be specified in a 0-based genomic coordinate system, i.e. a chromosome starts at position 0. Furthermore, the end position is not considered part of the interval.

To enable the analysis of multiple gene signatures at once, i.e. batch analysis, support for the GMT format is also provided. In the GMT (Gene Matrix Transposed) file format, introduced by the Broad Institute for their GSEA tool (13), each line corresponds to a different gene signature. The first and second column, separated by a TAB character, specify the unique identifier of the signature and its description. The next columns contain the actual gene identifiers in the signature.

Extending less abundant feature databases for enrichment analysis

Feature databases with a small number of features cannot be used to assess feature enrichment because their small size results in too few AUC values to get a good empirical distribution. Therefore, these feature databases are extended with additional features randomly sampled from the initial features. This sampling is done in the following way: the region at position i of the ranking that defines an additional random feature R is derived using the following formula: $region_i^R = random[U_{f \text{ in } F}(region_i^f)]$ where F is the set of all initial features and $region_i^f$ is the region at position i for feature f. Random sampling is done via a uniform distribution. If a region that is already present in a previous position in the ranking is drawn, a new region is sampled for that position. Sampling effectively produces a new random ranking for these additional features.

Combining gene sets and genomic loci for validation purposes

We used the following gene sets: Sandmann-Mef2 (14) and Liang-zelda (15). The corresponding "control" sets of genomic loci are, respectively, Mef2 ChIP-chip (9), and zelda ChIP-seq (16). These sets of genomic loci were intersected with the 5kb upstream + 5'UTR + first intronic space around the genes of the corresponding gene sets. We also generated sets of random 136K regions that were size matched with the control set. We ran i-cisTarget in parallel on the set of genes, the corresponding control set and the random sets, and compared the ranks and NES of enriched iVEs (NES >= 4 in gene sets) in both sets (Table 2).

Analysis of FlyBase TermLink sets

A collection of 628 co-expressed gene signatures was compiled from FlyBase TermLink. This database contains expression data that is organized in an anatomical ontology (FBbt), which is available in the open source OBO format. Starting from all the genes annotated in the ontology (specified by their FBgn identifier) a collection of FBbt terms was derived, and the annotated genes (FBgn) were collected for each term. Next, only those FBbt terms were retained for which at least one transcription factor was annotated for which a motif is available in the FlyFactorSurvey Database (the version downloaded on 19/11/2010) (3), and the root term (Fbbt:00000001 'organism') was left out. The resulting 628 signatures were extended with the genes associated with all direct and indirect children of the "primary" FBbt term as defined by the FBbt ontological graph. To avoid a too large and thus biological meaningless gene signature expansion, only graph traversal up to a depth of 100 was performed, and only the relationships type 'part_of' and 'has_part' in the ontology were used. On average these signatures have 42

genes. The entire collection of gene sets is analyzed at once with i-cisTarget, using the 'gmt' file input option. We then ask whether i-cisTarget can identify the motif for one or more TFs that are themselves expressed in the same tissue or cell type as its predicted targets. To quantify this statistically, a p-value was calculated based on the hypergeometrical distribution which models the overrepresentation of enriched TFs in the set of TFs that are part of the gene signature. The size of the population for the hypergeometrical distribution is taken as the total number of FlyFactorSurvey TFs used in the analysis; the number of successes in this population is equal to the number of enriched FFS TFs for a given gene signature; the number of TFs part of the gene signature is the sample size and the number of successes in this sample is equated to the number of these TFs that are enriched. For the four chosen sets shown in the text, we re-ran i-cisTarget via the web interface to obtain the HTML report for each set; to include iVEs in the analysis; and to further investigate whether additional motifs are found, besides those from FlyFactorSurvey Database. The i-cisTarget options used are the following: species "dm3", features "all features of version 1.1 databases" (corresponding to 4238 PWMs, 300 modEncode features, 48 BDTNP features, 23 Chip datasets from Furlong lab, 60 chromatin-state features), region mapping "5kb up, 5'UTR and 1st intron", Fraction overlap "0.4", Enrichment-score threshold "2.5", Enrichment analysis "within each database separately", ROC thresold for AUC calculation "0.01". We selected the IVEs and motifs that could be related to TFs present in the input set, and pick up from their respective ranking the optimal subset of predicted targets.

References

1. Portales-Casamar,E., Thongjuea,S., Kwon,A., Arenillas,D., Zhao,X., Valen,E., Yusuf,D., Lenhard,B., Wasserman,W. and Sandelin,A. (2009) JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. *Nucleic Acids Res*.

2. Matys,V., Kel-Margoulis,O.V., Fricke,E., Liebich,I., Land,S., Barre-Dirrie,A., Reuter,I., Chekmenev,D., Krull,M., Hornischer,K., et al. (2006) TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res*, **34**, D108–10.

3. Zhu,L.J., Christensen,R.G., Kazemian,M., Hull,C.J., Enuameh,M.S., Basciotta,M.D., Brasefield,J.A., Zhu,C., Asriyan,Y., Lapointe,D.S., et al. (2011) FlyFactorSurvey: a database of Drosophila transcription factor binding specificities determined using the bacterial one-hybrid system. *Nucleic Acids Res*, **39**, D111–7.

4. Down,T.A., Bergman,C.M., Su,J. and Hubbard,T.J.P. (2007) Large-scale discovery of promoter motifs in Drosophila melanogaster. *PLoS Comput Biol*, **3**, e7.

5. Elemento,O. and Tavazoie,S. (2005) Fast and systematic genome-wide discovery of conserved regulatory elements using a non-alignment based approach. *Genome Biol*, **6**, R18.

6. Stark,A., Lin,M.F., Kheradpour,P., Pedersen,J.S., Parts,L., Carlson,J.W., Crosby,M.A., Rasmussen,M.D., Roy,S., Deoras,A.N., et al. (2007) Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures. *Nature*, **450**, 219–232.

7. modENCODE Consortium, Roy,S., Ernst,J., Kharchenko,P.V., Kheradpour,P., Negre,N., Eaton,M.L., Landolin,J.M., Bristow,C.A., Ma,L., et al. (2010) Identification of functional elements and regulatory circuits by Drosophila modENCODE. *Science*, **330**, 1787–1797.

8. Li,X.-Y., MacArthur,S., Bourgon,R., Nix,D., Pollard,D.A., Iyer,V.N., Hechmer,A., Simirenko,L., Stapleton,M., Luengo Hendriks,C.L., et al. (2008) Transcription factors bind thousands of active and inactive regions in the Drosophila blastoderm. *PLoS Biol*, **6**, e27.

9. Zinzen,R.P., Girardot,C., Gagneur,J., Braun,M. and Furlong,E.E.M. (2009) Combinatorial binding predicts spatio-temporal cis-regulatory activity. *Nature*, **462**, 65–70.

10. de Boer,C.G. and Hughes,T.R. (2011) YeTFaSCo: a database of evaluated yeast transcription factor sequence specificities. *Nucleic Acids Res*, 10.1093/nar/gkr993.

11. Xie,Z., Hu,S., Blackshaw,S., Zhu,H. and Qian,J. (2010) hPDI: a database of experimental human protein-DNA interactions. *Bioinformatics*, **26**, 287–289.

12. Mahony,S. and Benos,P.V. (2007) STAMP: a web tool for exploring DNA-binding motif similarities. *Nucleic Acids Res*, **35**, W253–8.

13. Subramanian,A., Tamayo,P., Mootha,V.K., Mukherjee,S., Ebert,B.L., Gillette,M.A.,

Paulovich,A., Pomeroy,S.L., Golub,T.R., Lander,E.S., et al. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA*, **102**, 15545–15550.

14. Sandmann,T., Jensen,L.J., Jakobsen,J.S., Karzynski,M.M., Eichenlaub,M.P., Bork,P. and Furlong,E.E.M. (2006) A Temporal Map of Transcription Factor Activity: Mef2 Directly Regulates Target Genes at All Stages of Muscle Development. *Dev Cell*, **10**, 797–807.

15. Liang,H.-L., Nien,C.-Y., Liu,H.-Y., Metzstein,M.M., Kirov,N. and Rushlow,C. (2008) The zinc-finger protein Zelda is a key activator of the early zygotic genome in Drosophila. *Nature*, **456**, 400–403.

16. Harrison,M.M., Li,X.-Y., Kaplan,T., Botchan,M.R. and Eisen,M.B. (2011) Zelda Binding in the Early Drosophila melanogaster Embryo Marks Regions Subsequently Activated at the Maternal-to-Zygotic Transition. *PLoS Genet*, **7**, e1002266.