BIOINFORMATICS APPLICATIONS NOTE Vol. 22 no. 12 2006, pages 1532–1533
doi:10.1093/bioinformatics/btl143

Genome analysis

FlyTF: a systematic review of site-specific transcription factors in the fruit fly *Drosophila melanogaster*

Boris Adryan* and Sarah A. Teichmann

MRC Laboratory of Molecular Biology, Cambridge CB2 2QH, UK

Received on March 6, 2006; revised on March 28, 2006; accepted on April 10, 2006
Advance Access publication April 13, 2006
Associate Editor: Alex Bateman

ABSTRACT

Summary: We present a manually annotated catalogue of site-specific transcription factors (TFs) in the fruit fly *Drosophila melanogaster*. These were identified from a list of candidate proteins with transcription-related Gene Ontology (Go) annotation as well as structural DNA-binding domain assignments. For all 1052 candidate proteins, a defined set of rules was applied to classify information from the literature and computational data sources with respect to both DNA-binding and transcriptional regulatory properties. We propose a set of 753 TFs in the fruit fly, of which 23 are confident novel predictions of this function for previously uncharacterized proteins.

Availability: http://www.flytf.org/

Contact: boris@mrc-lmb.cam.ac.uk

Supplementary information: Supplementary data are available at http://www.flytf.org/

The genome sequence of the fruit fly *Drosophila melanogaster* was published almost 6 years ago (Adams *et al*., 2000). Despite progress in the functional annotation of the genes (Misra *et al*., 2002), roughly one-third of the predicted genes in *D.melanogaster* are still of unconfirmed existence (Ashburner and Bergman, 2005) and about one-fifth have no functional annotation according to the Gene Ontology database (Ashburner *et al*., 2000). This level of annotation holds for transcription factors (TFs), too. Regulation of gene expression by TFs is crucial to development and differentiation as well as the physiology of the fly. Networks of TF inter-regulation are known to drive development, for instance the segmentation network (Schroeder *et al*., 2004). TFs that recognize specific DNA sequences are arguably the core information-carrying molecules in gene regulatory networks, and therefore of particular interest for functional characterization.

Computational database support for site-specific transcriptional regulatory interactions has focused on the *cis*-regulatory sequences that are bound by TFs, rather than on the proteins themselves. For instance, the FlyReg database (Bergman *et al*., 2005, http://www.flyreg.org/) documents DNase I sensitive footprints while other databases focus on *cis*-regulatory modules (Gallo *et al*., 2006) or on specific developmental enhancers, for instance the Hox gene promoter regions (Spirov *et al*., 2000). However, there is no database for the complementary proteins that bind these sites, even though there is a wealth of literature on *D.melanogaster* sequence-specific TFs, and there is at least one systematic computational approach for TF prediction (Kummerfeld and Teichmann, 2006). In order to make such a resource available to the scientific community, we have developed a database of characterized and putative site-specific TFs in *D.melanogaster* called FlyTF, available at http://www.flytf.org/.

Our comprehensive database of site-specific TFs in the fruit fly *D.melanogaster* is based on Release 4 of the genome sequence. It is derived from a systematic literature curation of 1052 candidate TFs, which were extracted from a combination of GO annotation queries (see Supplementary Material on current GO annotations, September 2005) and the DBD TF Prediction Database (Kummerfeld and Teichmann, 2006, http://www.transcriptionfactor.org/). The GO queries yielded 1005 candidate proteins, 592 candidate TFs were retrieved from DBD. These DBD predictions are based on DNA-binding domain assignments and are benchmarked as having high accuracy (∼97%) and coverage of ∼65%. There were 47 candidate TFs from DBD that were not previously identified by the GO searches.

This set of 1052 candidate TFs was then subject to careful literature curation. This curation was focused on two separate aspects of the molecular function of TFs: DNA-binding on the one hand and transcription regulatory properties on the other. We assessed the evidence for these two properties of each TF using FlyBase (Drysdale & Crosby, 2005), in particular the Gene ontology and References sections. For instance, annotation based on automated electronic annotation would only be accepted if we could find further experimental evidence in the literature. Assignments of references to GO annotation were used as pointers to the literature, and further literature references were extracted from PubMed and the iHOP search tool (Hoffmann and Valencia, 2005). The most important references for each protein, as well as key sentences for the references, are included in the database entry for each TF, as explained below. Evidence from the carefully benchmarked DBD predictions, as well as annotation of a candidate TF with target genes in the *Drosophila* DNase I Footprint Database (Bergman *et al*., 2005), and the data-mining project FlyMine (http://www.

*To whom correspondence should be addressed.

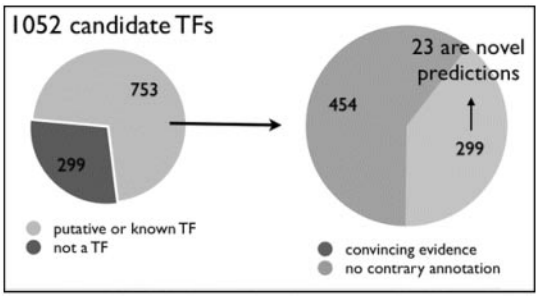

Fig. 1. We curated a total of 1052 candidate proteins. For 753 of them, there is evidence for DNA-binding and transcriptional regulatory activity. (The remainder are either not DNA-binding or not a TF.) Of the 753 candidate TFs, we find convincing evidence in the literature for 454, while there is no annotation to the contrary for 299 putative TFs. Of this latter group 23 represent novel predictions from DBD based on DNA-binding domain assignment without any other functional annotation.

flymine.org/) are also documented in the entry for each protein in FlyTF.

The database entry for each TF includes the literature references and key sentences as mentioned above, cross-links to relevant databases such as DBD, FlyReg and FlyMine, the GO annotation for the TF and four different properties of each protein for which we provide our curator's verdict. The four properties are 'DNA-binding', 'Site-specific TF', 'putative short-range TF' and 'known or putative long-range TF'. They cover the DNA-binding function of the protein and three aspects of the transcriptional regulatory activity. The categorization is detailed in depth in the Supplementary Material.

The entries for each individual protein out of the 1052 candidate proteins can be queried at http://ww.flytf.org/ by gene name, FlyBase identifier, protein family, evidence for DNA-binding, etc. The lists of TFs with different levels of evidence are available for download.

The curation procedure described above yielded 753 site-specific TFs as shown in Figure 1. This means that at least 5% of the fly genome encodes TFs, which agrees with previous estimates (van Nimwegen, 2003; Ashburner and Bergman *et al*., 2005). Approximately 450 of these are experimentally characterized with literature evidence, a further 270 had some previous transcription-related annotation in GO and 23 are entirely novel predictions from DBD (see Supplementary Material). We anticipate that this dataset will provide a framework for future computational and experimental research on the *Drosophila* transcriptional regulatory network from the perspective of the TFs acting on *cis*-regulatory elements. We plan to update the database in the future as new annotation becomes available.

ACKNOWLEDGEMENTS

The authors thank Sarah Kummerfeld for help with the DBD TF Database, Michael Bremang for helpful comments on the manuscript and Michael Ashburner for fruitful discussions on the annotation of TFs. B.A. is supported by an EMBO Longterm Fellowship and SAT is an EMBO Young Investigator. Funding to pay the Open Access publication charges for this article was provided by the Medical Research Council.

Conflict of Interest: none declared.

REFERENCES

Adams,M.D. *et al*. (2000) The genome sequence of *Drosophila melanogaster*. *Science*, **287**, 2185–2195.

Ashburner,M. *et al*. (2000) Gene ontology: tool for the unification of biology. *Nat. Genet*., **25**, 25–29.

Ashburner,M. and Bergman,C.M. (2005) *Drosophila melanogaster*: a case study of a model genomic sequence and its consequences. *Genome Res*., **15**, 1661–1667.

Bergman,C.M. *et al*. (2005) *Drosophila* DNase I footprint database: a systematic genome annotation of transcription factor binding sites in the fruitfly, *Drosophila melanogaster*. *Bioinformatics*, **21**, 1747–1749.

Drysdale,R.A. and Crosby,M.A. (2005) FlyBase: genes and gene models. *Nucleic Acids Res*., **33**, 390–395.

Gallo,S.M. *et al*. (2006) REDfly: a regulatory element database for *Drosophila*. *Bioinformatics*, **21**, 381–383.

Hoffmann,R. and Valencia,A. (2005) Implementing the iHOP concept for navigation of biomedical literature. *Bioinformatics*, **21**, 252–258.

Kummerfeld,S.K. and Teichmann,S.A. (2006) DBD: a transcription factor prediction database. *Nucleic Acids Res*., **34**, 74–81.

Misra,S. *et al*. (2002) Annotation of the *Drosophila melanogaster* euchromatic genome: a systematic review. *Genome Biol*., **3**, RESEARCH0083.

Schroeder,M.D. *et al*. (2004) Transcriptional control in the segmentation gene network of *Drosophila*. *PLoS Biol*., **2**, E271.

Spirov,A.V. *et al*. (2000) HOX Pro: a specialized database for clusters and networks of homeobox genes. *Nucleic Acids Res*., **28**, 337–340.

van Nimwegen,E. (2003) Scaling laws in the functional content of genomes. *Trends Genet*., **19**, 479–484.