

Databases and ontologies

## ***Drosophila* DNase I footprint database: a systematic genome annotation of transcription factor binding sites in the fruitfly, *Drosophila melanogaster***

Casey M. Bergman<sup>1,\*</sup>, Joseph W. Carlson<sup>2,3</sup> and Susan E. Celniker<sup>2,3</sup>

<sup>1</sup>Department of Genetics, University of Cambridge, Cambridge CB2 3EH, UK, <sup>2</sup>Department of Genome Sciences, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA and <sup>3</sup>Berkeley *Drosophila* Genome Project, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA

Received on August 5, 2004; revised on November 5, 2004; accepted on November 20, 2004

Advance Access publication November 30, 2004

### ABSTRACT

**Summary:** Despite increasing numbers of computational tools developed to predict *cis*-regulatory sequences, the availability of high-quality datasets of transcription factor binding sites limits advances in the bioinformatics of gene regulation. Here we present such a dataset based on a systematic literature curation and genome annotation of DNase I footprints for the fruitfly, *Drosophila melanogaster*. Using the experimental results of 201 primary references, we annotated 1367 binding sites from 87 transcription factors and 101 target genes in the *D.melanogaster* genome sequence. These data will provide a rich resource for future bioinformatics analyses of transcriptional regulation in *Drosophila* such as constructing motif models, training *cis*-regulatory module detectors, benchmarking alignment tools and continued text mining of the extensive literature on transcriptional regulation in this important model organism.

**Availability:** <http://www.flyreg.org/>

**Contact:** cbergman@gen.cam.ac.uk

The fruitfly *Drosophila melanogaster* has one of the most highly annotated metazoan genome sequences with respect to gene and transposable element content (Misra *et al.*, 2002; Kaminker *et al.*, 2002). In contrast, the *cis*-regulatory sequences that control transcription are only just beginning to be incorporated explicitly into the genome annotation, despite the vast literature of functionally characterized *cis*-regulatory elements that exists for this species (<http://www.flybase.org/>). This lack of a systematic, publicly available compilation of *cis*-regulatory sequences for *D.melanogaster*, such as the SCPD in yeast (Zhu and Zhang, 1999), limits progress in the computational analysis of gene regulation for this important model species. The need for such a resource is clear from the fact that *cis*-regulatory curation efforts of limited scope for genes involved in early development (Ludwig *et al.*, 2000; Spirov *et al.*, 2000; Berman *et al.*, 2002; Papatsenko *et al.*, 2002; Rajewsky *et al.*, 2002; Emberly *et al.*, 2003; Lifanov *et al.*, 2003) have rapidly proven useful for subsequent bioinformatic and comparative studies of gene regulation (e.g. Costas *et al.*, 2003; Grad *et al.*, 2004).

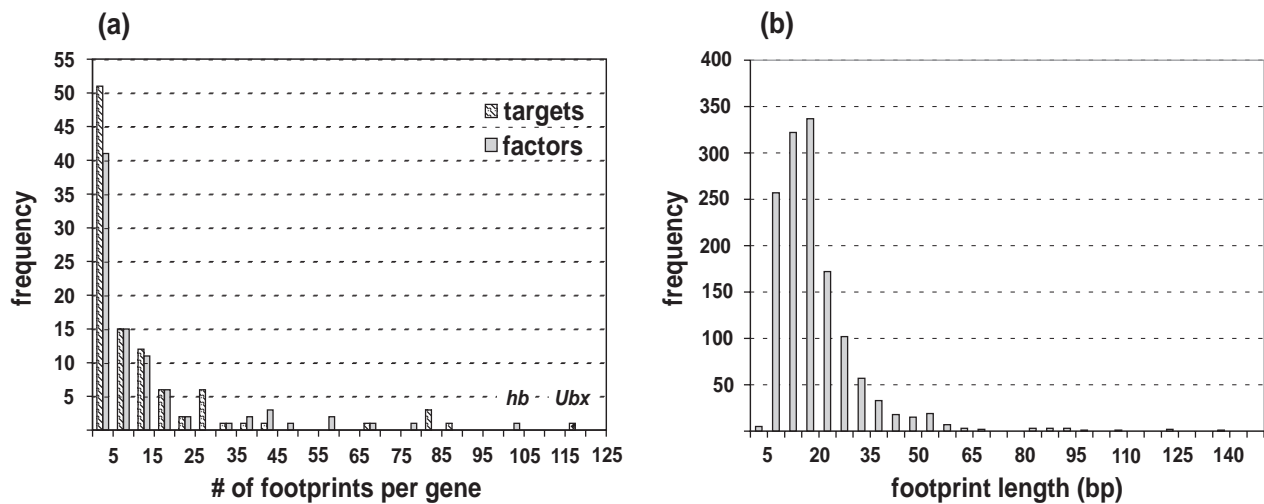
To contribute to a comprehensive annotation of *cis*-regulatory sequences in *D.melanogaster*, we report here a database of DNase

I footprint sequences derived from a systematic literature curation and annotation effort. We have chosen to focus on DNase I footprints data since they are an abundant and high-quality source of data on transcription factor specificity (Galas and Schmitz, 1978). In contrast to previous binding site compilations in *Drosophila*, these data are derived from the same experimental data type, cover all available aspects of development and are explicitly linked to the finished Release 3 genome sequence coordinates (Celniker *et al.*, 2002). The purpose of this note is to present a basic characterization of these data and to make them available in a single database as a resource for computational analyses of transcriptional regulation in one of the most important model organisms.

Our literature curation yielded a total of 201 references with non-redundant experimental data from DNase I footprinting experiments (see Supplemental Files 1 and 2). Our set of references is a superset of all those meeting the same criteria in previous compilations of binding site data for the *Drosophila* early embryo (Ludwig *et al.*, 2000; Spirov *et al.*, 2000; Berman *et al.*, 2002; Papatsenko *et al.*, 2002; Rajewsky *et al.*, 2002; Emberly *et al.*, 2003; Lifanov *et al.*, 2003) and Transfac 5.0 (Wingender *et al.*, 2001). The overlap between the present and previous compilations is detailed in Supplemental File 1. Our current work includes information from 113 primary references not present in any previous compilation, doubling the number of references with curated *Drosophila* DNase I footprint data consolidated in a single public database.

Of the 1367 footprints annotated, 1341 footprints (98%) can be attributed to 101 target genes, with 26 footprints (2%) obtained from chromatin immunoprecipitation experiments having 'unknown' targets (Supplemental File 3). The mean (median) number of footprints annotated per target gene is 13.3 (5), with a skewed distribution (Fig. 1a): the top ten genes (*Ubx*, *Antp*, *h*, *ftz*, *eve*, *dpp*, *kni*, *en*, *Ddc*, *Sgs4*) contribute nearly half (49%) of the footprints mapping to known targets. Likewise, 1164 (85%) of the 1367 footprints annotated can be attributed to 87 purified or recombinant transcription factors, plus an additional 203 footprints (15%) from 'unspecified' factors with unknown identity derived from crude or purified nuclear extract (Supplemental File 3). The distribution of number of footprints per factor is also skewed with a mean (median) number of footprints annotated per factor of 13.4 (6) (Fig. 1). As with the distribution by target, the top ten genes (*hb*, *Trl*, *ftz*, *Ubx*, *en*, *bcd*,

\*To whom correspondence should be addressed.



**Fig. 1.** Distribution of the number of annotated footprints per target gene and transcription factor (a), and distribution of annotated footprint lengths (b). In (a), the target gene with the highest number of footprints annotated is *Ubx*, and the transcription factor with the highest number of footprints annotated is *hb*.

*Kr*, *abd-A*, *z*, *dl*) also contribute nearly half (49%) of the footprints derived from known factors. Although these data represent the most comprehensive collection of binding site data in *Drosophila* to date, it is clear that binding site information is lacking for the majority of factors and genes, a limitation that can hopefully be overcome in the future by high-throughput experimental techniques (e.g. Bulyk et al., 2001).

Individually the 1363 footprints that map to euchromatic arms (four footprints map to heterochromatic scaffolds) comprise a total of 26,983 bp of DNA sequence, but since nearly half (45%,  $n = 613$ ) of the footprints annotated overlap at least one other footprint, these data span only 21,372 bp of genomic DNA, or approximately 0.0183% of the Release 3 euchromatic genome sequence. The footprinted sequences annotated range in length from 5 to 140 bp, and surprisingly have a mean (median) length of 19.8 bp (17 bp) (Fig. 1). In fact, the vast majority (81%,  $n = 1101$ ) of the footprinted sequences annotated are longer than both the 10.5 bp length needed for one turn of the  $\beta$ -form DNA helix (Wolffe, 1998) as well as the core recognition motif length (5–10 bp) typically reported for most transcription factors. The prevalence of long footprinted sequences may simply result from steric hindrance of the transcription factor preventing access to DNase cleavage, but may also suggest an under-appreciated role for non-core motif nucleotides in transcription-factor DNA interactions and/or a high frequency of homo-cooperative binding interactions. Certainly, the magnitude of overlap among footprinted sequences suggests the possibility of extensive hetero-cooperative interactions in these data. With the resource presented here, these and other hypotheses can now be tested using the wide array of experimental and computational methods available for the functional analysis of transcription factor binding sites.

## ACKNOWLEDGEMENTS

We thank Nicholas Blanchard for assistance with literature curation; FlyBase Cambridge for access to the *Drosophila* offprint collection; Michael Ashburner, Douda Bensasson, Thomas Down and Rachel Drysdale for suggestions on data format and representation;

and three anonymous reviewers and Nikolaus Rajewsky for helpful comments on the manuscript. This work was supported in part by NIH grants HG00750 and GH002673 to G.Rubin and SEC, respectively. CMB is supported by NIH training grant T32 HL07279 to E.Rubin and by a Royal Society USA Research Fellowship.

## SUPPLEMENTARY DATA

Supplementary data for this paper are available on *Bioinformatics* online.

## REFERENCES

- Berman, B.P., Nibu, Y., Pfeiffer, B.D., Tomancak, P., Celniker, S.E., Levine, M., Rubin, G.M. and Eisen, M.B. (2002) Exploiting transcription factor binding site clustering to identify *cis*-regulatory modules involved in pattern formation in the *Drosophila* genome. *Proc. Natl Acad. Sci. USA*, **99**, 757–762.
- Bulyk, M.L., Huang, X., Choo, Y. and Church, G.M. (2001) Exploring the DNA-binding specificities of zinc fingers with DNA microarrays. *Proc. Natl Acad. Sci. USA*, **98**, 7158–7163.
- Celniker, S.E., Wheeler, D.A., Kronmiller, B., Carlson, J.W., Halpern, A., Patel, S., Adams, M., Champe, M., Dugan, S.P., Frise, E. et al. (2002) Finishing a whole genome shotgun sequence assembly: release 3 of the *Drosophila* euchromatic genome sequence. *Genome Biol.*, **3**.
- Costas, J., Casares, F. and Vieira, J. (2003) Turnover of binding sites for transcription factors involved in early *Drosophila* development. *Gene*, **310**, 215–220.
- Emberly, E., Rajewsky, N. and Siggia, E.D. (2003) Conservation of regulatory elements between two species of *Drosophila*. *BMC Bioinformatics*, **4**, 57.
- Galas, D.J. and Schmitz, A. (1978) DNase footprinting: a simple method for the detection of protein-DNA binding specificity. *Nucleic Acids Res.*, **5**, 3157–3170.
- Grad, Y.H., Roth, F.P., Halfon, M.S. and Church, G.M. (2004) Prediction of similarly-acting *cis*-regulatory modules by subsequence profiling and comparative genomics in *D.melanogaster* and *D.pseudoobscura*. *Bioinformatics*, **20**, 2738–2750.
- Kaminker, J.S., Bergman, C.M., Kronmiller, B., Carlson, J., Svirskas, R., Patel, S., Frise, E., Wheeler, D.A., Lewis, S.E., Rubin, G.M., Ashburner, M. and Celniker, S.E. (2002) The transposable elements of the *Drosophila melanogaster* euchromatin: a genomics perspective. *Genome Biol.*, **3**.
- Lifanov, A.P., Makeev, V.J., Nazina, A.G. and Papatsenko, D.A. (2003) Homotypic regulatory clusters in *Drosophila*. *Genome Res.*, **13**, 579–588.
- Ludwig, M.Z., Bergman, C., Patel, N. and Kreitman, M. (2000) Evidence for stabilizing selection in a eukaryotic *cis*-regulatory element. *Nature*, **403**, 564–567.
- Misra, S., Crosby, M.A., Mungall, C.J., Matthews, B.B., Campbell, K.S., Hradecky, P., Huang, Y., Kaminker, J.S., Millburn, G.H., Prochnik, S.E. et al. (2002) Annotation

- of the *Drosophila melanogaster* euchromatic genome: a systematic review. *Genome Biol.*, **3**.
- Papatsenko,D.A., Makeev,V.J., Lifanov,A.P., Regnier,M., Nazina,A.G. and Desplan,C. (2002) Extraction of functional binding sites from unique regulatory regions: the *Drosophila* early developmental enhancers. *Genome Res.*, **12**, 470–481.
- Rajewsky,N., Vergassola,M., Gaul,U. and Siggia,E.D. (2002) Computational detection of genomic *cis*-regulatory modules applied to body patterning in the early *Drosophila* embryo. *BMC Bioinformatics*, **3**, 30.
- Spirov,A.V., Bowler,T. and Reinitz,J. (2000) HOX Pro: a specialized database for clusters and networks of homeobox genes. *Nucleic Acids Res.*, **28**, 337–340.
- Wingender,E., Chen,X., Fricke,E., Geffers,R., Hehl,R., Liebich,I., Krull,M., Matys,V., Michael,H., Ohnhauser,R. *et al.* (2001) The TRANSFAC system on gene expression regulation. *Nucleic Acids Res.*, **29**, 281–283.
- Wolffe,A. (1998) *Chromatin*. Academic Press, San Diego, p. 8.
- Zhu,J. and Zhang,M.Q. (1999) SCPD: a promoter database of the yeast *Saccharomyces cerevisiae*. *Bioinformatics*, **15**, 607–611.