BIOINFORMATICS ORIGINAL PAPER Vol. 25 no. 22 2009, pages 3001–3004
doi:10.1093/bioinformatics/btp469

Databases and ontologies

Flynet: a genomic resource for *Drosophila melanogaster* transcriptional regulatory networks

Feng Tian[1,2,†], Parantu K. Shah[3,4,†], Xiangjun Liu[1,2], Nicolas Negre[3,4], Jia Chen[2], Oleksiy Karpenko[2], Kevin P. White[3,4,*] and Robert L. Grossman[2,3,*]

[1]School of Medicine, Tsinghua University, Beijing, China 100084, [2]National Center for Data Mining, University of Illinois at Chicago, MC 249, 851 South Morgan Street, Chicago, IL 60607-7045, [3]Institute for Genomics & Systems Biology, The University of Chicago, Cummings Life Sciences Center 431A, 920 East 58th Street, Chicago, IL 60637 and [4]Department of Human Genetics and Department of Ecology and Evolution, Cummings Life Sciences Center 5th Floor, 920 East 58th Street, Chicago, IL 60637, USA

Received on April 30, 2009; revised on July 28, 2009; accepted on July 29, 2009

Advance Access publication August 5, 2009

Associate Editor: Limsoon Wong

ABSTRACT

Motivation: The highly coordinated expression of thousands of genes in an organism is regulated by the concerted action of transcription factors, chromatin proteins and epigenetic mechanisms. High-throughput experimental data for genome wide *in vivo* protein–DNA interactions and epigenetic marks are becoming available from large projects, such as the model organism ENCyclopedia Of DNA Elements (modENCODE) and from individual labs. Dissemination and visualization of these datasets in an explorable form is an important challenge.

Results: To support research on *Drosophila melanogaster* transcription regulation and make the genome wide *in vivo* protein–DNA interactions data available to the scientific community as a whole, we have developed a system called Flynet. Currently, Flynet contains 101 datasets for 38 transcription factors and chromatin regulator proteins in different experimental conditions. These factors exhibit different types of binding profiles ranging from sharp localized peaks to broad binding regions. The protein–DNA interaction data in Flynet was obtained from the analysis of chromatin immunoprecipitation experiments on one color and two color genomic tiling arrays as well as chromatin immunoprecipitation followed by massively parallel sequencing. A web-based interface, integrated with an AJAX based genome browser, has been built for queries and presenting analysis results. Flynet also makes available the *cis*-regulatory modules reported in literature, known and *de novo* identified sequence motifs across the genome, and other resources to study gene regulation.

Contact: grossman@uic.edu

Availability: Flynet is available at https://www.cistrack.org/flynet/.

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Metazoan genomes contain thousands of protein-coding and noncoding RNA genes, whose expression needs to be precisely controlled. Approximately 3–10% of the proteins in the metazoan known proteome are sequence specific transcription factors (TFs) (Kummerfeld and Teichmann, 2006), which bind to specific *cis*-regulatory DNA sequences and modulate the expression of their target genes. These *cis*-regulatory sequences are organized into *cis*-regulatory modules (CRM) containing one or more binding sites for a particular set of TFs. One example of CRMs are enhancers that determine a specific temporal-spatial expression pattern of their target gene (Wang *et al*., 2007).

The various proteins that form the chromatin participate in the regulation of genes (Sims and Reinberg, 2008). For example, the histones forming the nucleosomes can be post-translationally modified to create a chromatin environment that will repress or activate the genes around them. The different associations of TFs with their *cis*-regulatory elements on the DNA can trigger, counteract or modulate these regulatory states of genes. Although detailed studies of individual genes have identified many of the components and basic principles that control transcription, we still lack an understanding of the global architecture of transcription regulatory networks (Babu *et al*., 2004).

Drosophila melanogaster has been used extensively as a model organism to identify components and basic principles of transcription regulation. However, even after decades of research only 661 CRM sequences corresponding to 235 Drosophila genes and 778 transcription factor binding sites (TFBSs) are annotated in the Drosophila Cis-Regulatory Database (http://www.comp.nus.edu.sg/~bioinfo/Drosophila/) that combines information from sources such as RedFly (Halfon *et al*., 2008), DNAse footprint database (Bergman *et al*., 2005) and *Drosophila* Cis-Regulatory Database (Narang *et al*., 2006).

Chromatin Immunoprecipitation (ChIP) followed by microarray hybridization on the whole genome tiling arrays (ChIP-chip; Iyer *et al*., 2001; Ren *et al*., 2000) or followed by massively parallel DNA sequencing (ChIP-seq) (Johnson *et al*., 2007), are now established as powerful methods to identify all of the genomic

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

regions bound by a protein of interest in a given condition (Keles, 2007). Genome wide protein–DNA interaction data and epigenetic marks are now available for many transcription factors and chromatin regulators for *D.melanogaster* as well as other species that are providing details on transcription regulation (Kim and Ren, 2006). Moreover, the National Human Genome Research Institute sponsored model organism ENCyclopedia Of DNA Elements (modENCODE; http://www.modencode.org) Project aims to identify the majority of the sequence-based functional elements in the *Caenorhabditis elegans* and *D.melanogaster* genomes. It is important therefore to develop tools for storing, organizing and analyzing these data sets and to make them available to the scientific community in a usable format.

We have built Flynet as a part of our data management and visualization efforts for the modENCODE project, whose goal is to map the genome wide associations of a large set of the *Drosophila* sequence-specific TFs and chromatin regulator proteins. Flynet is the first public database for *D.melanogaster in vivo* protein–DNA interaction data identified on the whole genome tiling arrays using ChIP-chip as well as ChIP-seq for a variety of transcription factors and chromatin regulator proteins in different experimental conditions. It also makes available known CRMs, well-known and *de novo* identified sequence motifs across the genome, and a list of transcription factors and chromatin regulator proteins in *D.melanogaster* genome, their domain assignments and their orthologs and paralogs across 12 *Drosophila* genomes in the form of multiple sequence alignments. In the following sections we describe the query interface, system architecture, and AJAX based genome browser, as well as tools and resources available as a part of Flynet.

2 METHODS

2.1 Flynet system architecture

The Flynet data system is designed to be a general system for storing, annotating, and visualizing *in vivo* DNA-protein interaction datasets. Flynet includes code for processing, integrating and indexing the data from the several primary data sources. The Flynet database is implemented in MySQL. The user interface is written in Perl and uses Perl's Common Gateway Interface module (CGI.pm) and Cascaded Style Sheets (CSS). Flynet provides two types of front-end tools that let the user interact with the stored data: an AJAX based genome browser (Skinner *et al*., 2009) and a web tool that provides a table-based view of the data and a mechanism for downloading data.

The data in Flynet is stored in a system we call Cistrack (https://www.cistrack.org). Cistrack manages *cis*-regulatory data and was built using an integrated set of tools that we have developed called the Chicago Utilities for Biological Sciences or CUBioS. CUBioS includes: (i) a database; (ii) a light weight database browser; (iii) a cloud for storing auxiliary files, archiving the data and computing over the data; (iv) utilities for uploading and annotating the data and (v) Web 2.0 widgets for accessing, analyzing and visualizing the data.

The *in vivo* protein DNA interaction data are processed using different analysis algorithms (see Methods). The processed data are stored in the Flynet system in three different formats: (i) as ChIP enriched region data for downloading; (ii) as parsed and tiled data for visualization by the genome browser and (iii) as tables in MySQL to support SQL queries. The Flynet data system includes the chromosome, start and end co-ordinates of the ChIP-enriched region, peak location for the binding region, *P*-value, false discovery rate and various other attributes computed by the analysis.

There are seven tables in the Flynet module of Cistrack database that store information about the datasets, binding sites, experimental metadata, motifs, annotations and pre-computed putative target genes (PTGs). The first table contains basic information about the datasets. The TFBS table stores the binding information for each TF. The annotation tables contain annotations from several sources, including Redfly, Flybase and the Gene Ontology. A pre-computed PTG table allows users to search for co-regulators of a gene of interest.

2.2 Analysis of *in vivo* protein DNA interaction data

We analyzed peak-based ChIP-chip data generated on Affymetrix genomic tiling arrays as follows. Affymetrix BPMAP files were remapped using xMAN (Li *et al*., 2008) and the latest version of the fly assembly (UCSC dm3/Flybase release 5.8). The remapped BPMAP files are also processed to remove probe redundancy so that each 25-mer probe is mapped no more than once in any 1 kb window along the genome. The UCSC (http://genome.ucsc.edu/) dm3 RepeatMasker and simple repeat files were downloaded and used to create a Repeat Library file for use with MAT (Johnson *et al*., 2006). MAT was run with the remapped BPMAP files, the Repeat Library files generated specifically for the *Drosophila* genome and appropriate parameters for Bandwidth, MaxGap and Minprobe. Bandwidths were taken according to the average DNA fragments lengths from the original publications when reported (Johnson *et al*., 2006). The Repeat Library file is available for download from the Flynet resource page.

Peak-based data from two color arrays were analyzed using the MA2C and CisGenome packages with appropriate parameters for the UCSC dm3 assembly. ChIP-seq data were analyzed with the MACS package using appropriate parameters for the UCSC dm3 assembly. We used the following procedure for identification of binding regions for factors that identify broad regions. The ChIP data were first quantile normalized, replicate information was merged, and fold change for each probe on the array was calculated. The data was smoothed with an appropriate window and given as input for the HMM based segmentation. HMM segmentation using expectation maximization was used to identify the regions (Shah *et al*. in preparation).

2.3 Determination of PTGs

We extracted transcription start site (TSS) information of 15 145 genes from Flybase release 5.8. PTGs for a transcription factor were simply defined as genes whose TSS is closest to the transcription factor binding sites (Zhang, 1998). We also employed insulator information to correctly assign transcription factor binding sites to their target genes (Negre *et al*., in preparation).

2.4 Motif discovery and scanning methods

For each factor position, weight matrices were obtained from known databases or enriched motifs were identified using MEME (Bailey and Elkan, 1994), AlignACE (Hughes *et al*., 2000) and MDscan (Liu *et al*., 2002). All programs were run with default parameters except for MEME, which was restricted to a maximum of 3 iterations and a maximum motif width of 25. The motifs were then evaluated using the motif instance pipeline described in (Kheradpour *et al*., 2007) in order to identify motifs specifically enriched in the insulator regions and to compare the motifs discovered by the different programs. Each motif was scanned at two PWM cutoffs corresponding as determined by TFM (Touzet and Varre, 2007). Motifs were then ranked by their enrichment at several conservation levels (from 0.0 to 1.0 confidence).

3 RESULTS

3.1 Flynet user interface

Flynet is available at https://www.cistrack.org/flynet/. It provides users with the ability to search and browse *in vivo* DNA protein interaction data and related resources.

The Flynet 'search page' allows users to query the database using a transcription factor of interest (TF and DNA associated proteins)

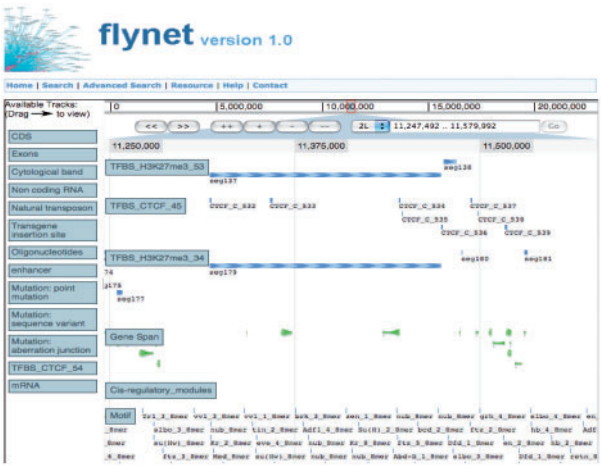

Fig. 1. A snapshot of Flynet JBrowse genome browser showing small peaks for the insulator CTCF (Illumina platform) and regions of histone modification H3K27me3 (Embryo stage E0-4 and S2 cells). JBrowse allows for faster and smoother navigation through the genome without requiring the reloading of the page. The annotation tracks in the left panel could be dynamically added and removed by dragging.

or by a target gene of interest. On the 'Search' page, users can first select single or multiple factors of interest. Next, users can review the associated metadata (e.g. data source, antibody, platform, analysis method, and publication) and use this information to refine the selection, after which it is then possible to browse the selected tracks in the browser.

Alternatively, the 'search by genes page' in Flynet allows users to query and retrieve the available data by Flybase identifier, CG identifier or gene symbol, and then to browse the putative regulators and the known *Drosophila* transcriptional CRMs from RedFly. Flynet provides lists of genes that are regulated by a transcription factor by two means: 1) using the closest transcription start site to a TFBS and 2) refining the search criteria using the presence of genome wide binding sites of insulator proteins (Negre *et al*., in preparation).

Flynet also provides an 'advanced search page' that allows users to query by selecting genomic regions and False Discovery Rates (FDR) thresholds for filtering results and identifying results in the genomic repeat regions. The 'download all' option allows users to download the publicly available data matching the specific query and provides quantitative information like enrichment score, enrichment ratios, FDR levels and peak positions.

Flynet utilizes JBrowse (http://www.jbrowse.org/), which allows users to select single or multiple factors from the available datasets and to browse them in a dynamic HTML environment that renders data tracks on the client side (Skinner *et al*., 2009). The use of an JAX-based browser offers several advantages over the existing static HTML based browsers, including a faster and smoother navigation through the genome without requiring the reloading of the page. Moreover, the browser makes it possible to view the portion of the genome and annotations by dragging and double clicking on the annotations. Users can also dynamically add or remove annotation tracks to generate customized views using the browser (for example as shown in Fig. 1).

3.2 Flynet contents

The present version of Flynet contains *in vivo* protein DNA interaction data for 101 datasets for 38 transcription factors (general and sequence specific TFs) and 8 chromatin regulators, including histone modifications in different experimental conditions. The data sources for these are experiments performed on whole genome tiling arrays in one color (Affymetrix) and two colors (Agilent and Nimblegen), as well as massively parallel sequencing using Illumina Genome analyzer. Raw data for *D.melanogaster* TFs and chromatin regulatory proteins include those downloaded from repositories like GEO (Barrett *et al*., 2009) and ArrayExpress (Parkinson *et al*., 2007), as well as those coming from our high-throughput experimental pipeline for identification of transcription factor binding sites as a part of the modENCODE project. These transcription factors show different binding behaviors ranging from sharp peak to broad regions and require different analysis methods and parameter optimizations.

Flynet stores relevant experimental metadata including the gene synonyms, data source (experimental conditions referring to appropriate development stage or cell line), antibody name, array platform, analysis method and literature reference. In addition to the experimental data and metadata, the UCSC dm3 genome assembly, annotations of 15 145 *D.melanogaster* genes from Flybase release 5.8, 162 727 PTG records and 665 CRM records from Redfly 2.0 are also integrated into Flynet.

In the resources section, we make available a list of transcription factors and chromatin regulator proteins in *D.melanogaster* genome, their domain assignments, and their orthologs and paralogs across 12 *Drosophila* genomes in the form of multiple sequence alignments and a collection of position weight matrices.

4 DISCUSSION

Flynet is a web-accessible database of *in vivo* protein DNA interactions integrated with effective searching and advanced browsing capabilities. To our knowledge Flynet is the first public database for *Drosophila* with the explicit goal of making accessible high-throughput data from genome wide studies *on vivo* protein DNA interactions and integrating it with other available data.

Specialized databases providing a list of sequence specific TF (Adryan and Teichmann, 2006), collections of CRM and TFBS information from literature (Halfon *et al*., 2008), *Drosophila cis*-regulatory element database (Narang *et al*., 2006), DNAse I footprint database (Bergman *et al*., 2005) and position weight matrices (Sandelin *et al*., 2004; Wingender *et al*., 1997) are available. Flynet integrates knowledge from some of these sources, includes recent experimental data (Georlette *et al*., 2007; Isogai *et al*., 2007; Kwong *et al*., 2008; Lee *et al*., 2008; Li *et al*., 2008; Matsumoto *et al*., 2007; Misulovin *et al*., 2008; Schwartz *et al*., 2006), and allows users to browse and compare this data easily.

For the fly genome, gene centric databases like Flybase and Flymine (Lyne *et al*., 2007) provide genomic and protein sequences, annotations, GO terms, protein structure, protein–protein interactions and pathway information, as well as various types of functional genomics data, including gene expression profiling and phenotypic information from various screens. However, these databases do not contain the large amount of publicly available ChIP-chip and ChIP-Seq data in an analyzed format.

Flynet, on the other hand, takes a TF-centric approach and specializes in transcriptional regulation information. It includes data being produced as the part of modENCODE project, as well as data deposited in GEO and ArrayExpress. In fact, 41 of the 101 datasets (∼40%) currently in Flynet are non-modENCODE data sets. In this way, Flynet provides users with ability to view and analyze modENCODE data, other published data, *de novo* computed motifs and CRM information along the genome.

Flynet data is analyzed by experts in a uniform manner using state-of-the art methods, undergoes a manual check for quality, and is based on the latest version of the fly genome. Methods for analyzing ChIP-chip and ChIP-seq data are still evolving. In fact, we have developed an HMM based segmentation algorithm for handling 'region' based data for K9 and K27 tri-methylation of Histone H3. We plan to reanalyze the available data with better methods as they are developed.

The data residing in Flynet includes data generated using different experimental platforms and under various experimental conditions, including different developmental stages, different cell lines and different antibodies. Because of the variety of data included in Flynet, similarities and differences in genome wide occupancy of these factors can be analyzed, as well as the strengths and weaknesses of different data generation platforms. For example, information on insulator binding regions in Flynet can be used for determining likely target genes for transcription factor binding sites. Flynet will be updated with new data, as it becomes available. We will also plan to update Flynet with every major update of *D.melanogaster* genome assembly and genome annotations.

The data residing in Flynet will be helpful in identification of new CRMs and the comparative analysis of similarities and differences in genome wide *in vivo* protein–DNA interactions and epigenetic marks.

ACKNOWLEDGEMENTS

The authors would like to thank Ian Holmes' laboratory for developing the AJAX based genome browser used in this project and making it available as an open source project. They also like to thank Manolis Kellis' laboratory for providing sequence motifs to display as an annotation track.

Funding: Chicago Biomedical Consortium with support from The Searle Funds at The Chicago Community Trust, NIH through grants P50 GM081892 and U01 HG004264, and the Chinese Student-Exchange Program.

Conflict of Interest: none declared.

REFERENCES

Adryan,B. and Teichmann,S.A. (2006) FlyTF: a systematic review of site-specific transcription factors in the fruit fly Drosophila melanogaster. *Bioinformatics*, **22**, 1532–1533.

Babu,M.M. *et al.* (2004) Structure and evolution of transcriptional regulatory networks. *Curr. Opin. Struct. Biol.*, **14**, 283–291.

Bailey,T.L. and Elkan,C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **2**, 28–36.

Barrett,T. *et al.* (2009) NCBI GEO: archive for high-throughput functional genomic data. *Nucleic Acids Res.*, **37**, D885–D890.

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.

Georlette,D. *et al.* (2007) Genomic profiling and expression studies reveal both positive and negative activities for the Drosophila Myb MuvB/dREAM complex in proliferating cells. *Genes Dev.*, **21**, 2880–2896.

Halfon,M.S. *et al.* (2008) REDfly 2.0: an integrated database of cis-regulatory modules and transcription factor binding sites in Drosophila. *Nucleic Acids Res.*, **36**, D594–D598.

Hughes,J.D. *et al.* (2000) Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. *J. Mol. Biol.*, **296**, 1205–1214.

Isogai,Y. *et al.* (2007) Novel TRF1/BRF target genes revealed by genome-wide analysis of Drosophila Pol III transcription. *EMBO J.*, **26**, 79–89.

Iyer,V.R. *et al.* (2001) Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature*, **409**, 533–538.

Johnson,D.S. *et al.* (2007) Genome-wide mapping of *in vivo* protein-DNA interactions. *Science*, **316**, 1497–1502.

Johnson,W.E. *et al.* (2006) Model-based analysis of tiling-arrays for ChIP-chip. *Proc. Natl Acad. Sci. USA*, **103**, 12457–12462.

Keles,S. (2007) Mixture modeling for genome-wide localization of transcription factors. *Biometrics*, **63**, 10–21.

Kheradpour,P. *et al.* (2007) Reliable prediction of regulator targets using 12 Drosophila genomes. *Genome Res.*, **17**, 1919–1931.

Kim,T.H. and Ren,B. (2006) Genome-wide analysis of protein-DNA interactions. *Annu. Rev. Genomics Hum. Genet.*, **7**, 81–102.

Kummerfeld,S.K. and Teichmann,S.A. (2006) DBD: a transcription factor prediction database. *Nucleic Acids Research*, **34**, D74–D81.

Kwong,C. *et al.* (2008) Stability and dynamics of polycomb target sites in Drosophila development. *PLoS Genet.*, **4**, e1000178.

Lee,C. *et al.* (2008) NELF and GAGA factor are linked to promoter-proximal pausing at many genes in Drosophila. *Mol. Cell Biol.*, **28**, 3290–3300.

Li,W. *et al.* (2008) xMAN: extreme MApping of OligoNucleotides. *BMC Genomics*, **9**(Suppl 1), S20.

Li,X.Y. *et al.* (2008) Transcription factors bind thousands of active and inactive regions in the Drosophila blastoderm. *PLoS Biol*, **6**, e27.

Liu,X.S. *et al.* (2002) An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. *Nat. Biotechnol.*, **20**, 835–839.

Lyne,R. *et al.* (2007) FlyMine: an integrated database for Drosophila and Anopheles genomics. *Genome Biol.*, **8**, R129.

Matsumoto,A. *et al.* (2007) A functional genomics strategy reveals clockwork orange as a transcriptional regulator in the Drosophila circadian clock. *Genes Dev*, **21**, 1687–1700.

Misulovin,Z. *et al.* (2008) Association of cohesin and Nipped-B with transcriptionally active regions of the Drosophila melanogaster genome. *Chromosoma*, **117**, 89–102.

Narang,V. *et al.* (2006) Computational annotation of transcription factor binding sites in D. Melanogaster developmental genes. *Genome Inform*, **17**, 14–24.

Parkinson,H. *et al.* (2007) ArrayExpress—a public database of microarray experiments and gene expression profiles. *Nucleic Acids Res.*, **35**, D747–D750.

Ren,B. *et al.* (2000) Genome-wide location and function of DNA binding proteins. *Science*, **290**, 2306–2309.

Sandelin,A. *et al.* (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.*, **32**, D91–D94.

Schwartz,Y.B. *et al.* (2006) Genome-wide analysis of Polycomb targets in Drosophila melanogaster. *Nat Genet*, **38**, 700–705.

Sims,R.J.,3rd and Reinberg,D. (2008) Is there a code embedded in proteins that is based on post-translational modifications?. *Nat. Rev. Mol. Cell. Biol.*, **9**, 815–820.

Skinner,M.E. *et al.* (2009) JBrowse: A next-generation genome browser. *Genome Res*.

Touzet,H. and Varre,J.S. (2007) Efficient and accurate *P*-value computation for Position Weight Matrices. *Algorithms Mo.l Biol.*, **2**, 15.

Wang,Z. *et al*. (2007) Unravelling the world of cis-regulatory elements. *Med. Biol. Eng. Comput.*, **45**, 709–718.

Wingender,E. *et al.* (1997) TRANSFAC database as a bridge between sequence data libraries and biological function. *Pac. Symp. Biocomput.*, 477–485.

Zhang,M.Q. (1998) Identification of human gene core promoters in silico. *Genome Res.*, **8**, 319–326.