D594–D598 *Nucleic Acids Research, 2008, Vol. 36, Database issue* *Published online 26 November 2007*
doi:10.1093/nar/gkm876

REDfly 2.0: an integrated database of *cis*-regulatory modules and transcription factor binding sites in *Drosophila*

Marc S. Halfon[1,2,4,5,*], Steven M. Gallo[3,4] and Casey M. Bergman[6]

[1]Department of Biochemistry, [2]Department of Biological Sciences, State University of New York at Buffalo, Buffalo NY 14214, [3]Center for Computational Research, [4]New York State Center of Excellence in Bioinformatics and the Life Sciences, Buffalo NY 14203, [5]Department of Molecular and Cellular Biology, Roswell Park Cancer Institute, Buffalo NY 14263, USA and [6]Faculty of Life Sciences, University of Manchester, Manchester M13 9PT, UK

Received August 13, 2007; Revised September 12, 2007; Accepted October 1, 2007

ABSTRACT

The identification and study of the *cis*-regulatory elements that control gene expression are important areas of biological research, but few resources exist to facilitate large-scale bioinformatics studies of *cis*-regulation in metazoan species. *Drosophila melanogaster*, with its well-annotated genome, exceptional resources for comparative genomics and long history of experimental studies of transcriptional regulation, represents the ideal system for regulatory bioinformatics. We have merged two existing *Drosophila* resources, the REDfly database of *cis*-regulatory modules and the FlyReg database of transcription factor binding sites (TFBSs), into a single integrated database containing extensive annotation of empirically validated *cis*-regulatory modules and their constituent binding sites. With the enhanced functionality made possible through this integration of TFBS data into REDfly, together with additional improvements to the REDfly infrastructure, we have constructed a one-stop portal for *Drosophila cis*-regulatory data that will serve as a powerful resource for both computational and experimental studies of transcriptional regulation. REDfly is freely accessible at http://redfly.ccr.buffalo.edu.

INTRODUCTION

Regulated spatial and temporal control of gene transcription is a fundamental process for all metazoans. Critical to this process is the interaction of transcription factors (TFs) with specific *cis*-regulatory DNA sequences. These regulatory sequences—for instance, enhancers and promoters—are organized in a modular fashion, with each module containing one or more binding sites for a specific combination of TFs (1). We use the term '*cis*-regulatory module' (CRM) as a generic term to refer to all enhancers and similar regulatory elements that are located outside of the core promoter region and which function to regulate transcription in a spatio–temporal-specific manner. We use the more general term '*cis*-regulatory element' to refer to either a CRM or a TF binding site (TFBS).

Despite the clear importance of *cis*-regulatory elements for many areas of biology—for instance, CRMs and TFBSs act as major control nodes in embryonic development, and variation in *cis*-regulatory elements plays an important role in both evolutionary change and normal phenotypic variation (2,3)—our knowledge of these sequences is surprisingly limited. The vast majority of CRMs are not known and, of those that are, relatively few have been characterized in detail. Even *Drosophila melanogaster*, which is a well-studied organism with a richly annotated genome, only has identified CRMs associated with fewer than 2% of its ~14 000 genes (4). Likewise, fewer than 1% of *Drosophila* genes currently have annotated TFBS data (5).

A well-annotated collection of known CRMs and their constituent TFBSs would be of significant use in many important areas of biological research, including studies of transcriptional regulation, genome structure and organization and the evolution of gene regulation. Such a resource would have considerable value in aiding subsequent CRM and TFBS discovery, for example, by providing training data for supervised learning or other bioinformatics approaches. Currently,

*To whom correspondence should be addressed: Tel: +716-829-3126; Fax: +716-829-2725; Email: mshalfon@buffalo.edu
Correspondence may also be addressed to Casey M. Bergman. Tel: +44-161-275-1613; Fax: +44-161-275-5082;
Email: casey.bergman@manchester.ac.uk

two databases play an important role in the study of *cis*-regulation in the model organism *D. melanogaster:* the FlyReg DNase I footprint database (5), a database of empirically defined TFBSs; and the REDfly (Regulatory Element Database for Fly) database (4), a highly annotated source of information on experimentally proven CRMs. These resources have served as the basis of a number of large-scale studies of *cis*-regulation and have allowed statistical, computational and comparative genomics methods to be brought to bear on its study (6–20). In this report, we describe the merger of these two databases into REDfly v2.0. With this release, REDfly has become a unified source of *Drosophila cis*-regulatory element annotation with one-stop access to both CRM and TFBS data for one of the best-studied model organisms, and the most comprehensive open-access resource for curated regulatory data in any metazoan species.

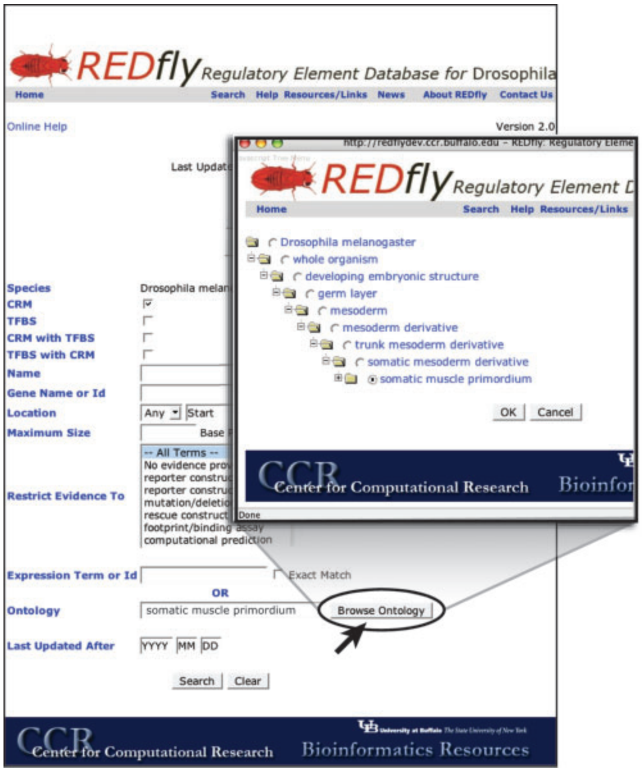
DATABASE CONTENTS

REDfly v2.0 (August 2007) contains records for 665 CRMs (up from 544 in the initial release) and 1341 TFBSs (down from 1367 in the initial FlyReg release due to removal of TFBS not attributed to target genes). Only sequences with empirical support are included in the database. The goal of REDfly is to include all experimentally verified fly CRMs and TFBSs along with their DNA sequence, their associated genes, and the expression patterns they direct. At present, curation has focused on literature reports of sequences that have been unambiguously demonstrated to be sufficient to regulate gene expression, primarily through reporter gene assays in transgenic animals and on TFBSs discovered by DNase I footprinting assays. For the most part, CRMs are included directly as reported in the literature. Where multiple nested sequences with identical activity were reported, the shortest such sequence was selected. Sequences with identical activity that are distinct but minimally overlapping are mostly reported separately, although in some instances of more substantial overlap, one or more sequences were omitted.

TFBS records include primarily DNase I (but not hydroxy-radical or copper nuclease) footprinting experiments that used protein obtained from nuclear extract (either crude or purified) or recombinant expression (either partial or full-length). When a binding factor purified from nuclear extract has been shown to be the derivative of a specific gene, footprints were attributed to the gene encoding that factor; otherwise the binding factor for nuclear extract footprints has been left as ‘unspecified’. Where possible, we followed the rule of precedence in attributing footprint data to a particular reference, unless members of the same research group reported refined coordinates in a subsequent publication. When two or more overlapping motifs for the same TF were reported for a single footprinted region, they were merged and annotated as one footprint.

All REDfly sequence features are mapped to the most current release (release 5; http://www.fruitfly.org/sequence/release5genomic.shtml) of the *D. melanogaster* genome sequence. Coordinates are also provided for the two previous sequence releases for maximum convenience and back-compatibility with other sequence resources. We store the actual DNA sequences as well as the coordinates so that sequences can be downloaded without ambiguity. Because TFBS sequences are often short and therefore cannot be uniquely mapped to the genome, we also include a ‘TFBS with flank’ option that provides ~25 bp of additional sequence both 5′ and 3′ to the TFBS. All records contain hyperlinks to the FlyBase (21) and FlyMine (22) entries for the target gene whose expression is regulated by the CRM or TFBS, and all features can be displayed on Gbrowse or UCSC genome browsers (23,24). For TFBS records, hyperlinks to FlyBase, FlyMine and FlyTF (16) are also available for the TF that binds the site, when known. For CRM records, controlled vocabulary descriptions of the expression pattern mediated by the CRM are provided using the *Drosophila* anatomy ontology (25). This is a key feature of REDfly and allows users to search for expression patterns using a tree-based browser interface (Figure 1). Selecting a term from the tree will query REDfly for any CRMs annotated with that term or any of its descendant terms. Alternatively, users can search for only a single term. Because expression patterns are described using the anatomy ontology, users can link from a CRM record to any other REDfly CRMs that are annotated as mediating the same gene expression pattern, or to records in FlyBase or the Berkeley *Drosophila* Genome Project’s *in situ* expression pattern database (26,27) for genes expressed in that pattern. These features promise to be highly useful for investigating properties of tissue-specific CRMs. For example, we recently made use of the expression pattern annotations to demonstrate that a certain class of CRMs—those that drive gene expression in the *Drosophila* early embryonic blastoderm—have characteristics that distinguish them from other CRMs (6). Detailed instructions on using the ontology to facilitate searching for CRMs that regulate specific expression patterns are provided in REDfly’s online help.

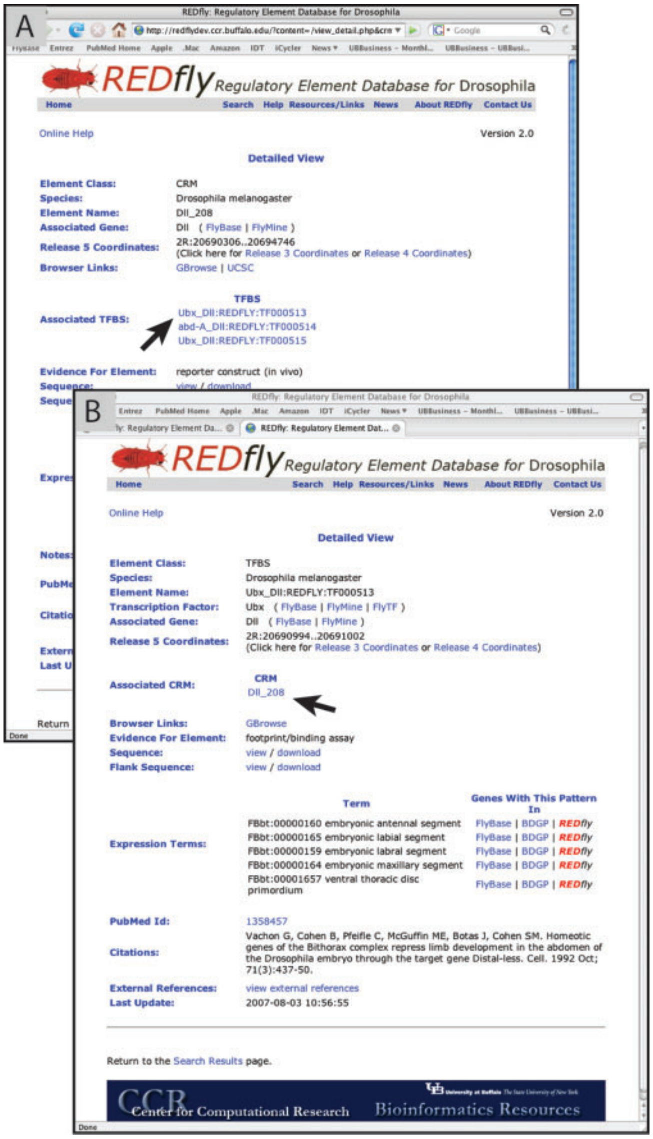
A major advantage of integrating the REDfly and FlyReg databases is the unprecedented level of detailed information that can now be obtained by mapping TFBSs directly to the CRMs of which they are a part. Upon entry of a new CRM or TFBS, the sequence coordinates of the new element are checked against the coordinates of all of the stored TFBSs or CRMs, respectively. If a TFBS falls within a known CRM, the name of the CRM and a link to its REDfly record is provided. Similarly, all CRM records are linked to the REDfly annotations of any TFBSs that fall within them (Figure 2). Searches of REDfly can be restricted to just those TFBSs that map to known CRMs, and vice-versa. Currently, 70% of TFBSs in REDfly map to a known CRM, while 26% of CRMs contain annotated TFBSs. Using these new REDfly features, it is now possible, for example, to investigate the association of TFBS sequences with expression patterns via their corresponding CRMs. REDfly is the only resource for regulatory bioinformatics that provides such a highly



Species: Drosophila melan
CRM ☑
TFBS ☐
CRM with TFBS ☐
TFBS with CRM ☐
Name
Gene Name or Id
Location: Any Start
Maximum Size: Base
Restrict Evidence To: -- All Terms --, No evidence pro, reporter constru, reporter constru, mutation/deletio, rescue construct, footprint/binding assay, computational prediction
Expression Term or Id ☐ Exact Match
OR
Ontology: somatic muscle primordium
Last Updated After: YYYY MM DD

○ Drosophila melanogaster
○ whole organism
○ developing embryonic structure
○ germ layer
○ mesoderm
○ mesoderm derivative
○ trunk mesoderm derivative
○ somatic mesoderm derivative
● somatic muscle primordium

Figure 1. Searching for regulatory elements based on the expression patterns they regulate using the 'Ontology' search function. Clicking on the 'Browse Ontology' button (circled) will open a pop-up window with the ontology tree (inset) that can then be navigated until the desired anatomical feature at the preferred level of granularity is reached. All REDfly records annotated with the selected term or any of its descendant terms will be returned in the database search. In the pictured example, the chosen term is 'somatic muscle primordium'; any records containing annotations for 'somatic muscle primordium' or its two descendant terms 'embryonic somatic muscle' and 'larval somatic muscle' will be returned. Alternatively, a user can type a term into the 'Expression Term or Id' field, in which case only records annotated with that term (e.g., 'somatic muscle primordium') will be returned.


A
B
http://redflydev.ccr.buffalo.edu/?content=/view_detail.php&crm
Google
Flybase Entrez PubMed Home Apple .Mac Amazon IDT iCycler News UBBusiness – Monthl... UBBusiness – UBBusi...
REDfly: Regulatory Element Database for Drosophila
REDfly: Regulatory Element Dat...
REDfly Regulatory Element Database for Drosophila
Home
Search
Help
Resources/Links
News
About REDfly
Contact Us
Online Help
Version 2.0
Return
Done
CCR Center for Computational Research
Bioinformatics Resources

Detailed View

Element Class: CRM
Species: Drosophila melanogaster
Element Name: Dll_208
Associated Gene: Dll (FlyBase | FlyMine)
Release 5 Coordinates: 2R:20690306..20694746
(Click here for Release 3 Coordinates or Release 4 Coordinates)
Browser Links: GBrowse | UCSC
Associated TFBS: TFBS
Ubx_Dll:REDFLY:TF000513
abd-A_Dll:REDFLY:TF000514
Ubx_Dll:REDFLY:TF000515
Evidence For Element: reporter construct (in vivo)
Sequence: view / download

Detailed View

Element Class: TFBS
Species: Drosophila melanogaster
Element Name: Ubx_Dll:REDFLY:TF000513
Transcription Factor: Ubx (FlyBase | FlyMine | FlyTF)
Associated Gene: Dll (FlyBase | FlyMine)
Release 5 Coordinates: 2R:20690994..20691002
(Click here for Release 3 Coordinates or Release 4 Coordinates)
Associated CRM: CRM
Dll_208
Browser Links: GBrowse
Evidence For Element: footprint/binding assay
Sequence: view / download
Flank Sequence: view / download

Expression Terms:

Term	Genes With This Pattern In
FBbt:00000160 embryonic antennal segment	FlyBase | BDGP | REDfly
FBbt:00000165 embryonic labial segment	FlyBase | BDGP | REDfly
FBbt:00000159 embryonic labral segment	FlyBase | BDGP | REDfly
FBbt:00000164 embryonic maxillary segment	FlyBase | BDGP | REDfly
FBbt:00001657 ventral thoracic disc primordium	FlyBase | BDGP | REDfly

PubMed Id: 1358457
Citations: Vachon G, Cohen B, Pfeifle C, McGuffin ME, Botas J, Cohen SM. Homeotic genes of the Bithorax complex repress limb development in the abdomen of the Drosophila embryo through the target gene Distal-less. Cell. 1992 Oct; 71(3):437-50.
External References: view external references
Last Update: 2007-08-03 10:56:55

Return to the Search Results page.

Figure 2. Integration of CRM and TFBS records. An exciting new feature in REDfly is its cross-referencing of CRMs with any TFBSs that fall within them. Panel A shows the detailed view page for the CRM Dll_208, which contains three footprinted TFBSs with records in REDfly. The arrow indicates the link to the record for one of these three TFBSs, Ubx_Dll:REDFLY:TF000513. Panel B shows the record obtained through the link indicated in panel A. Note that the lookup is functional in both directions; from this TFBS record page, there is a link to the CRM in which the binding site is located (arrow).

integrated annotation of CRMs and their constituent TFBSs.

DATABASE SCHEMA

The database schema has been designed to be both fast and extensible so that additional species can be added to the existing database structure at a later date and utilize the same search capabilities that have been developed for REDfly's *Drosophila* data. The tables in the database are grouped into four categories as diagrammed at the REDfly site at http://redfly.ccr.buffalo.edu/?content=/database.php:

(i) Species-specific fixed terms (outlined in red)
(ii) The CRM definition (yellow)
(iii) The binding site definition (blue)
(iv) External reference information (green).

The fixed-term tables—equivalent to the dimension tables of a star database schema—contain information that change infrequently, including anatomy terms (using the controlled vocabulary), evidence terms (also using a controlled vocabulary), chromosome numbers, sequences and gene names and IDs. Fixed-term information can be associated with a particular species, or can be common across all species. By utilizing tables of fixed terms we can load information for multiple species into the database and then reference this information from CRMs or binding sites without duplicating the information. Fixed terms also allow us to reduce query times and prevent the introduction of typographical errors when entering data.

The CRM definition tables consist of the basic information describing each CRM, such as the CRM name, species to which the CRM belongs, free text notes, references to information stored in the fixed-term tables, citation data and references to external websites. The binding site tables describe the TFBSs and provide information similar to that found in the CRM definition tables. A mapping also exists from CRMs to binding sites that are associated with that CRM, and vice versa.

Any reference that a CRM or TFBS record makes to an external site—such as FlyBase (21)—is considered an external reference. The external reference tables contain information on how to construct references to external sites such as a template for the URL, required parameters, etc. Citations are also included in external references.

Snapshots of the MySQL schema (i.e. a database dump) are recorded daily and are available for download. This provides extra backup and versioning protection, as well as an alternative method of access to the data for interested users.

RECENT IMPROVEMENTS TO REDFLY

REDfly XML

In addition to the inclusion of TFBSs and their association with corresponding CRMs, we have implemented a number of other key improvements to REDfly. In particular, we have developed extensible markup language (XML) representations for both CRM and TFBS records and have enabled XML formatted data as one of the download options. The XML format is the most comprehensive available for REDfly and allows for a complete dump of the database contents. Development of the XML format has also helped us to automate our input procedures and should help to increase the pace of updates and additions to the database.

Data sharing with ORegAnno

We have also established data-sharing standards with the ORegAnno community regulatory annotation database (28) and implemented a two-way exchange of data. All REDfly data are automatically shared with ORegAnno, where they represent 33 and 27% of the total number of curated CRM and TFBS records in ORegAnno, respectively. Although ORegAnno lacks *Drosophila*-specific functionality and several of the detailed annotation fields contained in REDfly, the inclusion of our data within ORegAnno allows for alternative access through the ORegAnno web-services and the newly implemented ORegAnno tracks in the UCSC browser (23). REDfly data that do not correspond to core ORegAnno fields are stored as ‘metadata’ within the ORegAnno record. Importantly, ORegAnno is an open community-based annotation platform. Therefore, community users can annotate fly CRMs and TFBSs via the easy-to-use ORegAnno interface that includes automatic mappings to the current genome build. REDfly can then automatically retrieve these data from ORegAnno and map them to the appropriate REDfly fields using the XML representations. During the synchronization process, REDfly also performs real-time queries to NCBI and UCSC to augment the ORegAnno data with related information such as literature citations and mapping of feature locations to multiple genome sequence releases. The records are then passed to the REDfly curators for validation, for the addition of any further annotation not provided in the ORegAnno metadata, and for connection to external *Drosophila*-specific resources not supported by ORegAnno. Over time, we anticipate that this will be the primary route of entry for REDfly data, either from the community or by our own curators. In this way, we are able to take advantage of ORegAnno’s general database cross-referencing functions, community-based annotation model and UCSC Browser tracks while still maintaining REDfly’s ability to provide *Drosophila*-specific information such as expression pattern data, links to external *Drosophila* resources and cross-references between CRMs and TFBSs.

PLANNED DEVELOPMENTS

In addition to continued curation of REDfly, we have targeted several areas for development within the near future. Important among these is to expand our curation to include TFBS data from sources other than footprint experiments, e.g from electrophoretic mobility shift assay (EMSA) and chromatin immunoprecipitation experiments. We also plan to annotate a broader range of CRMs, including negative regulatory elements (silencers), and to increase the amount of annotation for each CRM to include features such as the FlyBase transgenic transposon ID for the reporter gene construct used in the assay that defined the CRM, and the position of the CRM with respect to the organization of the gene it regulates (e.g. transcription start site, exon boundaries). These additions will increase the comprehensiveness of REDfly as a source of *Drosophila cis*-regulatory data and facilitate the mining of these data in more diverse and sophisticated ways. A further planned development is the addition of images of reporter gene expression driven by each CRM in order to associate *cis*-regulatory sequences directly with embryonic expression patterns; this work is being conducted in collaboration with the FlyExpress project (29).

Over the longer term, we plan to incorporate a more extensive use of formal ontologies to describe not only expression pattern data but also experimental evidence, assay types and sequence features in order to maximize opportunities for data mining and for interoperability with other databases. Toward this goal, we have been working with the Sequence Ontology (SO) developers (30) to expand and refine SO’s treatment of *cis*-regulatory sequences. We note that REDfly is easily adaptable to curation of *cis*-regulatory elements from species other than *D. melanogaster* with only minor modifications to the current schema that raises the possibility of incorporating multi-species regulatory data—either by direct curation or via our links with ORegAnno—into the database. This, along with our use of ontologies to allow interspecies mapping of genes and tissues, has the potential to make REDfly an unparalleled platform for comparing regulatory strategies and studying the organization of regulatory elements throughout evolution.

Accessibility

REDfly is freely available to all users without restriction at http://redfly.ccr.buffalo.edu. A snapshot of the current MySQL schema is posted daily on the REDfly server. Source code and other detailed information is available upon request.

ACKNOWLEDGEMENTS

We thank Guruprasad Sarvothaman for programming assistance and Stephen Montogomery, Obi Griffith, Bryan Chu, Erin Pleasance for assistance with the REDfly-ORegAnno data exchange. Grant support is from National Institutes of Health (HG002489 to M.S.H.). Funding to pay the Open Access publication charges for this article was provided by the University at Buffalo.

Conflict of interest statement. None declared.

REFERENCES

1. Davidson,E.H. (2006) *The Regulatory Genome: Gene Regulatory Networks in Development and Evolution, 1st edn*. Academic Press, Burlington, MA.
2. Carroll,S.B., Grenier,J.K. and Weatherbee,S.D. (2001) *From DNA to Diversity. Molecular Genetics and the Evolution of Animal Design*. Blackwell Science, MA.
3. Wray,G.A. (2007) The evolutionary significance of cis-regulatory mutations. *Nat. Rev. Genet*., **8**, 206–216.
4. Gallo,S.M., Li,L., Hu,Z. and Halfon,M.S. (2006) REDfly: a Regulatory Element Database for Drosophila. *Bioinformatics*, **22**, 381–383.
5. Bergman,C.M., Carlson,J.W. and Celniker,S.E. (2005) Drosophila DNase I footprint database: a systematic genome annotation of transcription factor binding sites in the fruitfly, *Drosophila melanogaster*. *Bioinformatics*, **21**, 1747–1749.
6. Li,L., Zhu,Q., He,X., Sinha,S. and Halfon,M.S. (2007) Large-scale analysis of transcriptional cis-regulatory modules reveals both common features and distinct subclasses. *Genome Biol*., **8**, R101.
7. Brody,T., Rasband,W., Baler,K., Kuzin,A., Kundu,M. and Odenwald,W.F. (2007) cis-Decoder discovers constellations of conserved DNA sequences shared among tissue-specific enhancers. *Genome Biol*., **8**, R75.
8. De Renzis,S., Elemento,O., Tavazoie,S. and Wieschaus,E.F. (2007) Unmasking activation of the zygotic genome using chromosomal deletions in the Drosophila embryo. *PLoS Biol*., **5**, e117.
9. Kim,J. and Sinha,S. (2007) Indelign: a probabilistic framework for annotation of insertions and deletions in a multiple alignment. *Bioinformatics*, **23**, 289–297.
10. Sandmann,T., Girardot,C., Brehme,M., Tongprasit,W., Stolc,V. and Furlong,E.E.M. (2007) A core transcriptional network for early mesoderm development in *Drosophila melanogaster*. Genes Dev., **21**, 436–449.
11. Sandmann,T., Jensen,L.J., Jakobsen,J.S., Karzynski,M.M., Eichenlaub,M.P., Bork,P. and Furlong,E.E. (2006) A temporal map of transcription factor activity: mef2 directly regulates target genes at all stages of muscle development. *Dev. Cell*, **10**, 797–807.
12. Macdonald,S.J. and Long,A.D. (2006) Fine scale structural variants distinguish the genomes of *Drosophila melanogaster* and *D. pseudoobscura*. *Genome Biol*., **7**, R67.
13. Papatsenko,D., Kislyuk,A., Levine,M. and Dubchak,I. (2006) Conservation patterns in different functional sequence categories of divergent Drosophila species. *Genomics*, **88**, 431–442.
14. Down,T.A., Bergman,C.M., Su,J. and Hubbard,T.J. (2007) Large-scale discovery of promoter motifs in *Drosophila melanogaster*. *PLoS Comput. Biol*., **3**, e7.
15. Dewey,C.N., Huggins,P.M., Woods,K., Sturmfels,B. and Pachter,L. (2006) Parametric alignment of Drosophila genomes. *PLoS Comput. Biol*., **2**, e73.
16. 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.
17. Glazov,E.A., Pheasant,M., McGraw,E.A., Bejerano,G. and Mattick,J.S. (2005) Ultraconserved elements in insect genomes: a highly conserved intronic sequence implicated in the control of homothorax mRNA splicing. *Genome Res*., **15**, 800–808.
18. Moses,A.M., Pollard,D.A., Nix,D.A., Iyer,V.N., Li,X.-Y., Biggin,M.D. and Eisen,M.B. (2006) Large-scale turnover of functional transcription factor binding sites in Drosophila. *PLoS Comp. Biol*., **2**, e130.
19. Pierstorff,N., Bergman,C.M. and Wiehe,T. (2006) Identifying cis-regulatory modules by combining comparative and compositional analysis of DNA. *Bioinformatics*, **22**, 2858–2864.
20. Pollard,D.A., Moses,A.M., Iyer,V.N. and Eisen,M.B. (2006) Detecting the limits of regulatory element conservation and divergence estimation using pairwise and multiple alignments. *BMC Bioinformatics*, **7**, 376.
21. Crosby,M.A., Goodman,J.L., Strelets,V.B., Zhang,P. and Gelbart,W.M. (2007) FlyBase: genomes by the dozen. *Nucleic Acids Res*., **35**, D486–D491.
22. Lyne,R., Smith,R., Rutherford,K., Wakeling,M., Varley,A., Guillier,F., Janssens,H., Ji,W., McLaren,P. *et al*. (2007) FlyMine: an integrated database for Drosophila and Anopheles genomics. *Genome Biol*., **8**, R129.
23. Kent,W.J., Sugnet,C.W., Furey,T.S., Roskin,K.M., Pringle,T.H., Zahler,A.M. and Haussler,D. (2002) The human genome browser at UCSC. *Genome Res*., **12**, 996–1006.
24. Stein,L.D., Mungall,C., Shu,S., Caudy,M., Mangone,M., Day,A., Nickerson,E., Stajich,J.E., Harris,T.W. *et al*. (2002) The generic genome browser: a building block for a model organism system database. *Genome Res*., **12**, 1599–1610.
25. Drysdale,R. (2001) Phenotypic data in FlyBase. *Brief Bioinform*., **2**, 68–80.
26. Tomancak,P., Berman,B.P., Beaton,A., Weiszmann,R., Kwan,E., Hartenstein,V., Celniker,S.E. and Rubin,G.M. (2007) Global analysis of patterns of gene expression during Drosophila embryogenesis. *Genome Biol*., **8**, R145.
27. Tomancak,P., Beaton,A., Weiszmann,R., Kwan,E., Shu,S., Lewis,S.E., Richards,S., Ashburner,M., Hartenstein,V. *et al*. (2002) Systematic determination of patterns of gene expression during Drosophila embryogenesis. *Genome Biol*., **3**, RESEARCH0088.
28. Montgomery,S.B., Griffith,O.L., Sleumer,M.C., Bergman,C.M., Bilenky,M., Pleasance,E.D., Prychyna,Y., Zhang,X. and Jones,S.J.M. (2006) ORegAnno: an open access database and curation system for literature-derived promoters, transcription factor binding sites and regulatory variation. *Bioinformatics*, **22**, 637–640.
29. Van Emden,B., Ramos,H., Panchanathan,S., Newfeld,S.J. and Kumar,S. (2006) FlyExpress: an image-matching web-tool for finding genes with overlapping patterns of expression in Drosophila embryos. Arizona State University, Tempe, AZ, www.flyexpress.net.
30. Eilbeck,K., Lewis,S.E., Mungall,C.J., Yandell,M., Stein,L., Durbin,R. and Ashburner,M. (2005) The Sequence Ontology: a tool for the unification of genome annotations. *Genome Biol*., **6**, R44.