D118–D123 *Nucleic Acids Research, 2011, Vol. 39, Database issue* *Published online 21 October 2010*
doi:10.1093/nar/gkq999

REDfly v3.0: toward a comprehensive database of transcriptional regulatory elements in *Drosophila*

Steven M. Gallo[1,2], Dave T. Gerrard[3], David Miner[1,2], Michael Simich[4], Benjamin Des Soye[4], Casey M. Bergman[3] and Marc S. Halfon[2,4,5,6,*]

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

Received September 14, 2010; Accepted October 6, 2010

ABSTRACT

The REDfly database of *Drosophila* transcriptional *cis*-regulatory elements provides the broadest and most comprehensive available resource for experimentally validated *cis*-regulatory modules and transcription factor binding sites among the metazoa. The third major release of the database extends the utility of REDfly as a powerful tool for both computational and experimental studies of transcription regulation. REDfly v3.0 includes the introduction of new data classes to expand the types of regulatory elements annotated in the database along with a roughly 40% increase in the number of records. A completely redesigned interface improves access for casual and power users alike; among other features it now automatically provides graphical views of the genome, displays images of reporter gene expression and implements improved capabilities for database searching and results filtering. REDfly is freely accessible at http://redfly.ccr.buffalo.edu.

INTRODUCTION

With the sequencing of the human genome and those of most of the major model organisms completed—some as long ago as a decade or more—a main priority for genomics has become to achieve a complete annotation of the functional features of the genome including all transcribed regions and regulatory sequences. The National Institutes of Health (NIH)-sponsored ENCODE and modENCODE consortia (1,2), as well as numerous individual efforts, have made great strides in identifying features amenable to discovery by high-throughput genomic methods such as transcripts, transcription start sites (TSSs), transcription factor-binding sites (TFBSs), chromatin conformation and various epigenetic marks. Most of these data can be accessed directly either via the UCSC (3) or ENSEMBL (4) genome browsers, browsers maintained by the project consortia,and/or the major model organism databases, or can be imported with minimal effort as tracks into these browsers.

However, other important classes of functional elements are still poorly represented in the model organism databases and other genomic data clearinghouses. Chief among these are transcriptional *cis*-regulatory modules (CRMs) such as enhancers, silencers and proximal promoter sequences [in keeping with our previous usage (5) we use CRM as a generic term to refer to transcriptional regulatory elements located outside of the core promoter region which regulate gene expression in a spatio-temporal manner]. Despite progress in computational and experimental identification of CRMs by high-throughput methods, reliable CRM discovery and, in particular, accurate determination of CRM structure and function still remains the province of small-scale, low-throughput experimental approaches. Moreover, a wealth of data on CRM structure and function, as well as on the TFBSs that constitute CRM substructure, remains locked in legacy biomedical literature.

We have endeavored to fill this gap in CRM and TFBS annotation for the important model organism *Drosophila melanogaster* with the establishment and curation of the REDfly (6) and FlyReg (7) databases and their subsequent integration in 2008 as REDfly 2.0 (5). Our goal is to include all experimentally verified fly CRMs and TFBSs along with their DNA sequence, their associated genes, and the expression patterns they direct, and, as such, REDfly provides the most comprehensive available resource for curated regulatory data in any metazoan.

*To whom correspondence should be addressed. Tel: +1 716 829 3126; Fax: +1 716 849 6655; Email: mshalfon@buffalo.edu

We report here the release of the third major update to REDfly (v3.0), which introduces the curation of several new types of data, a completely redesigned and enhanced user interface, a new RESTful application programming interface (API) to support the user interface, and an increase of over 800 new records.

NEW DATA CLASSES

Reporter Constructs and Inferred CRMs

With respect to CRMs, the initial releases of REDfly focused on 'minimal' sequences shown to be sufficient to regulate gene expression through reporter gene assays in transgenic animals. Where multiple nested sequences with identical activity were reported, the shortest such sequence was selected and defined as a CRM. REDfly v3.0 maintains this CRM definition, but recognizes that there is value to researchers in knowing all sequences tested in reporter gene assays, regardless of whether they are found to be positive regulators of gene expression or if shorter inclusive sequences with the same regulatory activity exist. We therefore now define CRMs as a subclass of a broader class of reporter constructs (RCs), which includes *all* sequences functionally tested in a transgenic reporter gene assay (Figure 1). REDfly RCs have three associated attributes: 'expression'; 'CRM'; and 'minimization'.

'Expression' has value 'positive' or 'negative' and describes whether or not the sequence was reported to drive gene expression in the reporter gene assay. RCs with positive expression have their expression patterns annotated using the *Drosophila* anatomy ontology, as previously described (6). RCs recorded as 'negative' for regulatory activity should be treated with caution by the user, for as with any negative data, the failure to observe reporter gene activity could simply reflect a failure of the assay rather than a biological result. The sequence might still mediate gene regulation in an unexamined tissue, require a promoter different than the one used in the reporter construct or be a silencer or other form of negative regulatory element not detectable in the assay. Nevertheless, knowledge of which sequences have been tested but failed to show regulatory activity can be instructive in understanding the regulatory landscape of a region and for designing further experiments to explore a locus.

'CRM' is a binary flag indicating whether an RC is the shortest of a set of nested sequences with identical activity (Figure 1). The CRM flag is also set to true if an RC is the only annotated sequence covering a given set of genomic coordinates. In other words, we define an RC as having the property CRM if it is the minimal-length reporter construct in a set of one or more nested reporter constructs that produce the same gene expression pattern.

When an RC is part of a set of nested sequences, rather than a single tested sequence at a particular locus, we say that the set of RCs have undergone 'minimization' (Figure 1). This designation could, for example, aid a

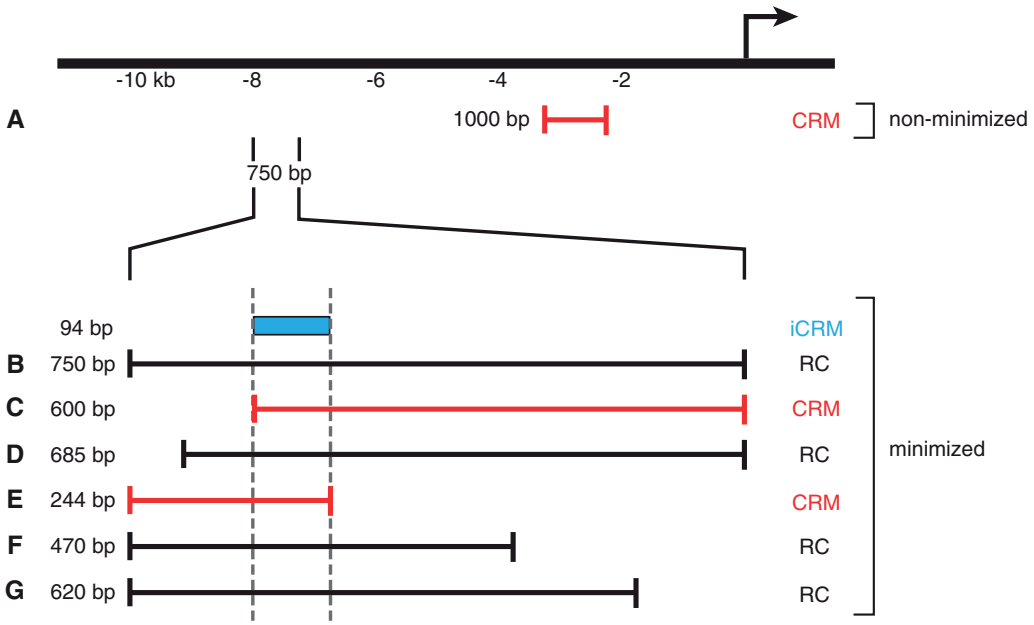

Figure 1. Reporter constructs and their attributes in REDfly. The figure illustrates a hypothetical locus for which seven different reporter constructs (**A–G**) have been tested *in vivo*. Construct A is a 1-kb sequence fragment located roughly 2 kb upstream of the transcription start. Because it is an isolated construct, it is considered to be a CRM that has not been subject to minimization. If this construct showed reporter gene activity, it would be designated as 'expression positive'; otherwise, it would be labeled 'expression negative'. Constructs B–G are part of an overlapping and partially nested series of sequences spanning 750 bp of DNA 7.25 kb upstream of the transcription start. In this example, each drives the identical pattern of reporter gene expression. Because each of these constructs overlaps at least one other, we consider this region and the six constructs to have undergone minimization. Constructs C and E are each the shortest of a respective set of nested sequences and are therefore considered to be CRMs (marked in red). The remaining constructs are designated as RCs (black). A 94-bp sequence marks the minimal region of overlap among all of the constructs and is thus registered in REDfly as an inferred CRM (iCRM, blue).

researcher in deciding whether to undertake experimental analysis of a region; a region that has undergone minimization might have less new information to reveal than one for which only a single construct has been tested *in vivo*. Likewise, knowing whether an RC has been minimized helps with the interpretation of overlaps between *in silico* or chromatin immunoprecipitation (ChIP)-based CRM predictions and annotated CRMs, which if not minimized can be up to several kilobases in length.

As a further aid to researchers, we introduce a new class of data, the 'inferred CRM' (iCRM). Often, two overlapping sequences have the same regulatory activity *in vivo*, which suggests (but does not prove) that the overlapping region may contain the minimal CRM (Figure 1). These overlaps can arise from RCs that were assayed in different publications and therefore are only discovered through integrated curation in REDfly. Note that because iCRMs have no direct empirical evidence supporting their functionality, they are not considered RCs by REDfly.

FlyBase (8) curates transgenic constructs, including reporter gene constructs, reported in publications (using the identifier FBtpxxxxxxx). Although these database entries contain a variety of useful meta-data (such as molecular data about the construct, available alleles and synonyms or secondary IDs), few of the FBtp constructs curated in FlyBase have associated genomic coordinates or gene expression data. REDfly now includes the FlyBase identifier for each RC that has also been curated by FlyBase, with a link to the FlyBase report for that FBtp construct. In the future, this REDfly ↔ FlyBase mapping will be used to supplement the core REDfly annotations and to enhance interoperability between REDfly and other *Drosophila* databases.

Electrophoretic mobility shift assay-based TFBSs

We have expanded the annotation of TFBSs by curating those confirmed using electrophoretic mobility shift assays (EMSAs, gel shift) in addition to DNase I footprinted sites. The EMSA data are of necessity not as precise as those from DNase I footprinting; whereas footprinting provides an exact sequence for the protected regions, TFBSs obtained from EMSA experiments formally can be said only to bind somewhere within the sequence of the probe used in the assay (typically 20–50 bp in length). However, in most cases, the authors have provided a presumed binding sequence within the probe, and we have used this to represent the binding site.

Regulatory element-target gene positional data

As a new feature in REDfly, we now report the position of each regulatory element with respect to each of the transcripts of its associated gene. The coordinates of each element are evaluated against the most current genome annotation to determine whether the element lies 5′ to, 3′ to, within an intron of, or within (or overlaps) the transcribed region of each associated transcript. These fields are searchable within the database, so that a user can search only for regulatory elements found within introns, for example, or can exclude anything located 3′ to the gene.

Conservation of regulatory element–target gene syntenic relationships

REDfly 3.0 also contains information on the conservation of local synteny between CRMs and the TSSs of their respective target genes. Because large-scale genomic rearrangements that disrupt local synteny have the potential to disrupt the spacing and orientation of CRMs relative to their target genes, conservation of synteny provides support for the annotated CRM–target gene relationship and indicates that this functional association is conserved over evolutionary time. Syntenic regions were identified by modifying the method of Engstrom *et al.* (9) applied to *D. melanogaster*, *D. ananassae*, *D. yakuba*, *D. erecta*, *D. pseudoobscura*, *D. willistoni*, *D. virilis*, *D. mojavensis* and *D. grimshawi* using a 10-kb cutoff (DTG and CMB, unpublished data). Of an initial 737 CRMs surveyed, 78% of individual CRM-to-target TSS relationships are found within the same syntenic block across these nine species, providing evolutionary evidence that the CRM–target gene interaction is real and conserved. An additional 13% of CRM-to-target TSS pairings fall in regions between adjacent syntenic blocks, and 9% overlap breaks in synteny. These may represent misannotated CRM-target gene relations, gain/loss of regulatory interactions or CRMs that have not been sufficiently minimized.

NEW RECORDS

REDfly v3.0 increases the total number of REDfly records by 877 from 2079 (737 CRMs + 1342 TFBSs) to 2956 (1334 RCs + 1662 TFBSs). The number of curated publications has increased by 91 from 392 to 483 and the number of genes for which there is at least one CRM has risen by over 40% from 240 to 342. TFBS records include sites for 30 new transcription factors associated with 45 additional target genes (113 total, up from 83, and 142 total, up from 97, respectively). Annotation of additional RCs from publications for which REDfly already records one or more CRMs is in progress and will be completed in the near future. Curation of the literature is ongoing and new records will continue to be added on a rolling basis following the complete curation of each individual publication.

USER INTERFACE

To accommodate our new data classes and to improve the user experience for both casual and power users of REDfly, we have wholly redesigned the REDfly user interface (Figure 2). Search, results and detailed record views are now all contained within separate panes of the same browser window. Automatic term suggestion and auto-completion has been enabled for many search fields, and context-specific help links and mouse-over help displays are available for most fields and features.

Search

Search functionality has been improved so that essentially all data classes, attributes and associated fields are available for searching. The search pane is divided into 'basic search' and 'advanced search' sections. Basic search (Figure 2A) allows for searching by gene name, FlyBase ID (both gene IDs and transgenic construct IDs), element name, PubMed ID or recent updates; the latter will return all records entered on the most recent date of data entry/update. Options to 'browse all' records and to 'download all' RCs, all CRMs or all TFBSs are also available. The advanced search pane (Figure 2B) is divided into two tabs, one for RC/CRM options and one for TFBS options. RC/CRM options include searching for all records, for CRM records only or for CRMs with associated TFBS data only. These can be further filtered for positive versus negative expression and for whether or not a construct is part of a minimized set.

TFBS-specific options allow for searching all TFBSs or only those with associated RC/CRM data. Gene names can be used to search all TFBS records or only those where the named gene either is the target or encodes the transcription factor, respectively.

Results

Search results appear as a list in a pane located below the search pane, with results for RCs/CRMs, TFBSs and iCRMs listed in separate tabs (Figure 2C). The number

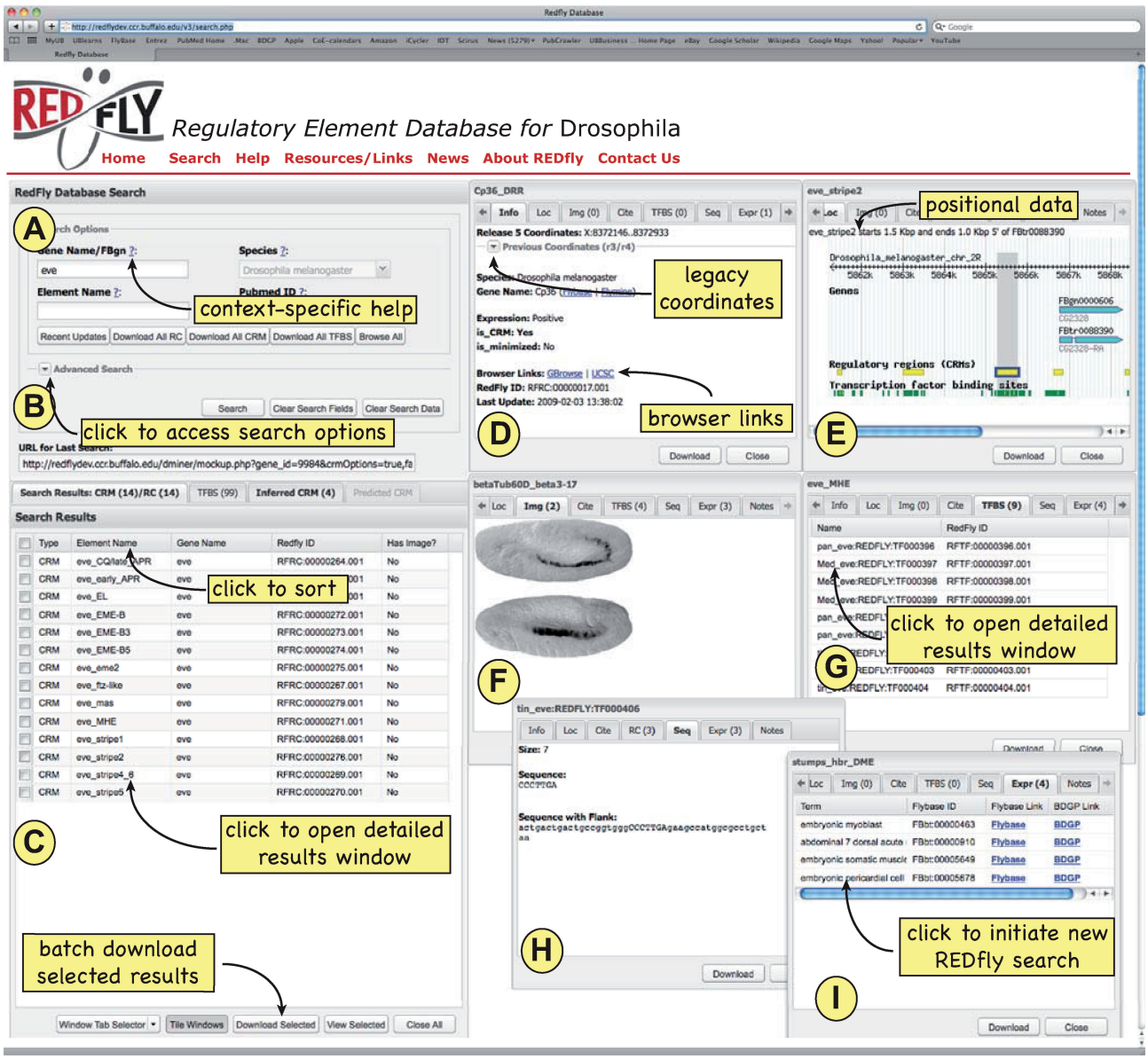

Figure 2. The new REDfly user interface. See text for details. Search options (**A**,**B**), results overview (**C**) and detailed results (**D–I**) are all displayed within a single web browser window. (D–I) Detailed results are displayed as individual floating windows that can be stacked or tiled as desired; on a large monitor, a dozen or more individual records can be fully tiled for simultaneous viewing.

of results returned for each data class is indicated in the tab header, and results may be sorted by clicking on the column header. Users can choose to download one or more records directly from this pane in a variety of formats using the 'download' button at the bottom. Clicking on a row will open a detailed view window for the record to the right of the search pane. Alternatively, multiple records can be selected using the check boxes along the left-hand side of the pane and then clicking on the 'view selected' button at the bottom. Multiple detailed view windows open in a stack; the 'tile windows' button will tile these in the browser window.

Detailed record view

Results for each selected record are presented in a detailed view window composed of multiple tabs displaying different sections of the information for each entry (Figure 2D–I). The 'window tab selector' at the bottom of the search results pane can be used to bring the selected tab to the foreground in each open detailed view window. Each detailed view window contains the following tabs: 'basic information', 'location', 'images', 'citation and evidence', 'associated TFBSs' (for RC/CRM records), 'associated RCs' (for TFBS records), 'sequence', 'expression' and 'notes'. Individual records can be downloaded using the 'download' button at the bottom of each window.

The basic information tab (Figure 2D) contains overview data for the element such as its genomic coordinates, the name of its associated gene(s) and links to the FlyBase and UCSC genome browsers. The location tab (Figure 2E) presents a graphical overview of the locus showing all annotated transcripts and regulatory elements, without requiring navigation to a separate genome browser. The images tab (Figure 2F; RCs/CRMs only) shows the expression pattern of the reporter gene. These images are provided courtesy of FlyExpress (http://www.flyexpress.net), and clicking on the image will bring the user to the FlyExpress website, from which a search can be initiated for other genes with a similar expression pattern. Images are currently available for only a subset of REDfly records, based on availability from the FlyExpress curators. The citation tab provides, and links to, the published reference for the element and also describes the evidence supporting the REDfly annotation. All RC and CRM records are linked to the REDfly annotations of any TFBSs that fall within them. These are listed in the TFBS tab (for RC/CRM records; Figure 2G)—clicking within a row will open a window with detailed results for that record. Likewise, if a TFBS falls within a known RC/CRM, the name of the RC/CRM and a link to its REDfly record is provided in the RC tab. The sequence tab (Figure 2H) displays the size (in base pairs) and sequence of the current feature, while the expression tab lists the gene expression pattern the feature regulates (Figure 2I). Clicking on a row in the expression terms list will initiate a REDfly search in a new browser window for all records containing the specified term. The notes tab contains any free-text notes elaborating on the expression pattern or other aspects of the annotation.

IMPLEMENTATION

In REDfly 3.0, we have adopted the Model View Controller (MVC) architecture. This paradigm keeps the domain logic and the user interface separate, isolating the effect of future design changes and facilitating the sharing of REDfly data with collaborators. The user interface has been redesigned using the ExtJS JavaScript application framework while all domain logic is provided by a RESTful API.

User interface

The REDfly user interface has been completely redesigned to provide the user with the look and feel of a desktop application rather than a traditional web page. This was possible with the maturation of JavaScript user interface toolkits such as ExtJS (http://www.sencha.com/products/js/), which allow us to create complete user interfaces with functionality that was previously available only to users running a locally installed application. This includes expandable widgets, search term suggestion, sortable results tables and tabbed windows, all with the ability to interact with multiple data sources simultaneously while providing a seamless experience for the user. The user interface is responsible only for the collection of user input and the display of information provided by the RESTful API. The user interface additionally accepts parameters via a URL for direct searching. This feature facilitates linking to REDfly from external sites and also allows users to bookmark and cite specific results.

RESTful API

We have designed REDfly 3.0 to expose domain logic (i.e. search results, CRM inference, data curation, etc.) through a set of URLs that implement our RESTful API. This web-based API is responsible for performing core logic operations on behalf of the application and returning the results in a number of formats facilitating access by collaborators. The API is organized into entities, actions to be performed on those entities and the format of the return data. For example, for an RC, actions are available for searching the set of RCs, retrieving a single RC using its REDfly identifier, listing any images associated with an RC and downloading the results of a search in a number of common file formats. While our own user interface makes extensive use of this API, it is also available to collaborators and allows for direct access to REDfly data for inclusion or ingestion by other tools. API documentation is provided at http://redfly.ccr.buffalo.edu/api/explorer.php.

ACCESSIBILITY

REDfly is freely available to all users without restriction at http://redfly.ccr.buffalo.edu. Source code and other detailed information are available upon request.

ACKNOWLEDGEMENTS

We thank the students from the 2008 and 2010 'Biochemistry 502: Genome Annotation' classes at the University at Buffalo for help with curation, Kyle Marcus for assistance with the REDfly user interface and Sudhir Kumar and Michael McCutchan of Arizona State University for making available images from FlyExpress.

FUNDING

The National Science Foundation (grant EF0843229 to M.S.H.); Human Frontier Science Program (grant RGY0084/2008-C to C.M.B.). Funding for open access charge: The National Science Foundation (grant EF0843229 to M.S.H.).

Conflict of interest statement. None declared.

REFERENCES

1. The ENCODE Project Consortium. (2004) The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science*, **306**, 636–640.
2. Celniker,S.E., Dillon,L.A., Gerstein,M.B., Gunsalus,K.C., Henikoff,S., Karpen,G.H., Kellis,M., Lai,E.C., Lieb,J.D., MacAlpine,D.M. *et al*. (2009) Unlocking the secrets of the genome. *Nature*, **459**, 927–930.
3. 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.
4. Hubbard,T.J., Aken,B.L., Ayling,S., Ballester,B., Beal,K., Bragin,E., Brent,S., Chen,Y., Clapham,P., Clarke,L. *et al*. (2009) Ensembl 2009. *Nucleic Acids Res*., **37**, D690–D697.
5. Halfon,M.S., Gallo,S.M. and Bergman,C.M. (2008) REDfly 2.0: an integrated database of *cis*-regulatory modules and transcription factor binding sites in Drosophila. *Nucleic Acids Res*., **36**, D594–D598.
6. Gallo,S.M., Li,L., Hu,Z. and Halfon,M.S. (2006) REDfly: a regulatory element database for *Drosophila*. 10.1093/bioinformatics/bti794. *Bioinformatics*, **22**, 381–383.
7. 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 fruit fly, *Drosophila melanogaster*. *Bioinformatics*, **21**, 1747–1749.
8. Tweedie,S., Ashburner,M., Falls,K., Leyland,P., McQuilton,P., Marygold,S., Millburn,G., Osumi-Sutherland,D., Schroeder,A., Seal,R. *et al*. (2009) FlyBase: enhancing Drosophila Gene Ontology annotations. *Nucleic Acids Res*., **37**, D555–D559.
9. Engstrom,P.G., Ho Sui,S.J., Drivenes,O., Becker,T.S. and Lenhard,B. (2007) Genomic regulatory blocks underlie extensive microsynteny conservation in insects. *Genome Res*., **17**, 1898–1908.