modMine: flexible access to modENCODE data

Sergio Contrino[1], Richard N. Smith[1], Daniela Butano[1], Adrian Carr[1], Fengyuan Hu[1],
Rachel Lyne[1], Kim Rutherford[1], Alex Kalderimis[1], Julie Sullivan[1], Seth Carbon[2],
Ellen T. Kephart[2], Paul Lloyd[2], E. O. Stinson[2], Nicole L. Washington[2], Marc D. Perry[3],
Peter Ruzanov[3], Zheng Zha[3], Suzanna E. Lewis[2,*], Lincoln D. Stein[3,*] and
Gos Micklem[1,*]

[1]Department of Genetics, University of Cambridge, Downing Street, Cambridge CB2 3EH, UK,
[2]Lawrence Berkeley National Laboratory, Genomics Division, Berkeley, CA 94720, USA and
[3]Ontario Institute for Cancer Research, MaRS Centre, Toronto, ON, Canada M5G 0A3

ABSTRACT

In an effort to comprehensively characterize the functional elements within the genomes of the important model organisms *Drosophila melanogaster* and *Caenorhabditis elegans,* the NHGRI model organism Encyclopaedia of DNA Elements (modENCODE) consortium has generated an enormous library of genomic data along with detailed, structured information on all aspects of the experiments. The modMine database (http://intermine.modencode.org) described here has been built by the modENCODE Data Coordination Center to allow the broader research community to (i) search for and download data sets of interest among the thousands generated by modENCODE; (ii) access the data in an integrated form together with non-modENCODE data sets; and (iii) facilitate fine-grained analysis of the above data. The sophisticated search features are possible because of the collection of extensive experimental metadata by the consortium. Interfaces are provided to allow both biologists and bioinformaticians to exploit these rich modENCODE data sets now available via modMine.

INTRODUCTION

The NHGRI model organism Encyclopaedia of DNA Elements (modENCODE) project (1, http://www.modencode.org) aims to provide the biological research community with a comprehensive encyclopaedia of genomic functional elements for the model organisms *Caenorhabditis elegans* and *Drosophila melanogaster* (2,3). The consortium's research composed of 11 primary projects divided between fly and worm, spans a wide diversity of genomic structures and functions including: identification of novel genes; annotation of gene parts including introns, exons, 5′ and 3′ regulatory elements, alternative splicing and complete gene models; mRNA and ncRNA expression profiles; transcription factor binding sites; profiles of histone modification and chromatin structure; and origins of DNA replication (only *D. melanogaster*).

The project has employed a diverse and constantly improved set of experimental strategies to keep pace with technology. For example, while microarrays were commonly used to acquire data early on, by the end of the project sequencing by synthesis or ligation (or 'next generation sequencing'), platforms were being used for most of the data collection including Chromatin immunopurification (ChIP) studies to map transcription factor binding sites and domains of histone modification, as well as to help determine gene structure and measure gene expression. So that the provenance of all data may be clearly understood, the ordered set of protocols along with key parameters are available for each data set, including the computation methods used to process data and, for example, call peaks within ChIP-seq data.

A particular challenge for this large multifaceted project is helping researchers to find relevant research results among the broad data types and thousands of individual experiments, which would overwhelm typical list-oriented displays. This challenge of providing users with a direct, obvious way to pinpoint relevant data sets can only be met by ensuring the quality and detail of all experimental

metadata. The raw and interpreted data from the consortium as well as the associated experimental metadata are all vetted by the Data Coordination Center (DCC) to ensure consistency and completeness prior to being released to the community (4). Data sets are released to the community immediately after vetting, and after a 9 months embargo, there is no restriction on their use. All data and the publication policy are available at http://www.modencode.org.

This article presents how we have used the InterMine platform (5) to address the above challenge. A prerequisite for providing intuitive, consistent and accurate data mining using modMine is well-annotated data sets, using well-controlled metadata. Experimental metadata is collected using the BIR-TAB format (4), which draws on current MIAME (6) for microarray, MINSEQE (http://www.mged.org/minseqe/) for high-throughput sequencing and other MIBBI (7) experimental metadata specifications. Where available, the project uses standard ontologies such as the Sequence Ontology for genomic features (8) or the MGED Ontology for microarray experiments (9), and controlled vocabularies such gene names from the model databases, strains from the worm and fly stock centers and cell lines from the Drosophila Genomics Resource Center (4). This has allowed us to exercise fine-grained control over presentation and queries in the modMine database (http://intermine.modencode.org), thereby allowing the research community to navigate through modENCODE experiments, to perform sophisticated *ad hoc* queries on the project data and metadata and to select, view, download, integrate and analyse the research results. This article describes features of the modMine database that are useful to biologists, while at the same time highlighting some features helpful to bioinformaticians.

The modMine database and web interface are based on the InterMine data warehousing system (5) to provide researchers with a powerful infrastructure to query the modENCODE data and metadata. Data produced by the modENCODE project are integrated with information from other sources in order to increase their utility. For instance by including mappings to orthologous genes in other organisms (10), the opportunity to carry out comparative studies is provided. Other external data incorporated in modMine include genome annotations from WormBase (11) and FlyBase (12), Gene Ontology annotations (13), physical and genetic interactions (14,15), protein information (16) and protein domains (17). Apart from the ability to integrate data from multiple sources, modMine has other useful features: the ability to work with lists (e.g. of genes or submissions); to access a library of commonly used search tasks available as 'search templates'; and to be able to extract data from a defined list of chromosomal locations and the provision of extensive web services and code generation for bioinformaticians. These features complement other tools used and developed by the modENCODE DCC such as the worm (http://modencode.oicr.on.ca/fgb2/gbrowse/worm/) and fly (http://modencode.oicr.on.ca/fgb2/gbrowse/fly/) GBrowse (18) genome browsers that can be used to view combinations of existing and modENCODE-generated genome annotations and the faceted data sets search tool (http://data.modencode.org/) that makes it easy to download the data from related sets of experiments.

We selected InterMine as a database platform since it provides a wide range of features, is designed to perform efficiently with large data sets and is easily extensible. It supports flexible querying, export of data in a variety of configurable and common formats, analysis of lists of genes with embedded tools, persistence by allowing users private workspaces to store their own queries and lists between sessions and a RESTful API, client libraries and automatic code generation to support programmers. The system is built from an extensible core data model by automatic code generation, which reduces maintenance overheads. We have had considerable experience working with model organism data within InterMine as part of the FlyMine project (5, http://www.flymine.org). FlyMine, as the prototypical InterMine database, has been joined by a series of databases built in collaboration with major model organism databases such as budding yeast (SGD, 19, http://yeastmine.yeastgenome.org/) and rat (RGD, 20, http://ratmine.mcw.edu/), with a zebrafish InterMine database (ZFIN, 21) to be released soon and mouse (MGD, 22) and worm (WormBase, 11) versions under development. InterMine has also been used to present data on mitochondrial proteins (23, http://mitominer.mrc-mbu.cam.ac.uk), *D. melanogaster* transcription factors (24, http://www.flytf.org), drug targets (25, http://targetmine.nibio.go.jp) and to support the metabolic disease research community (http://www..metabolicmine.org).

modMine is the gateway to the very substantial amount of data being generated by the modENCODE project. To provide context, modENCODE data are presented with core data from FlyBase and WormBase. In time, some of the modENCODE data are expected to be integrated in the relevant model organism databases by the respective curation teams. To provide an integrated view of the modENCODE project output in the longer term, modMine will remain available through the Amazon cloud.

DATA SOURCES

The data sources used to build modMine are accessible via the 'Data' tab at the top of the homepage. Within the data table, note in particular the links 'See all fly modENCODE submissions' and the corresponding worm one, which list all the modENCODE data submissions.

The data generated by modENCODE fall into several categories based on the genomic or biological element tested: gene structure, mRNA and ncRNA expression profiling, transcription factor binding sites, histone modification and replacement, chromatin structure, DNA replication and copy number variation.

Each category contains multiple sets of experiments, with each set varying by the assay method, submitting group and/or experimental condition tested. Results for

each experiment are submitted by the data producing laboratories to the DCC. An individual 'submission' is a single instance of an experiment, which tests varying factors such as life stage, genetic background or antibody, and consists of a package of results, experimental protocols and other metadata. The principal experimental variables include: organism, strain, developmental stage, tissue, cell line, temperature, compound, experimental technique (e.g. ChIP-seq) and the entity assayed (e.g. mRNA and particular histone modification)

For example, the *Drosophila* experiment set 'Chromatin Binding Site Mapping of Transcription Factors in D. melanogaster by ChIP-seq' can be found in the 'TF binding sites' category of experiment (categories are listed under 'Browse all modENCODE data' on the front page), together with two other experiments. This experiment currently has 25 data submissions that cover a range of developmental stages, antibodies and strains. It has identified some 50 000 protein binding sites, and produced 59 submissions to public repositories such as GEO (26), plus approximately two files of evidence per submission: a wiggle format file for the display of dense, continuous data such as probability scores (https://cgwb.nci.nih.gov/goldenPath/help/wiggle.html) and a GFF3 format file for the representation of genomic features on a sequence (http://www.sequenceontology.org/gff3.shtml).

The most recent release of modMine (release 25, September 2011) contains 69 experiments, subdivided in 12 categories and comprising 1664 individual submissions.

In order to coordinate analysis efforts, there have been periodic data freezes: currently modMine is rebuilt quarterly, as well as after every modENCODE data freeze. modMine releases are archived and are available to users (e.g. release 17 of modMine is available at http://intermine.modencode.org/release-17).

Non-modENCODE data sets

The above modENCODE data sets are integrated with data from other sources, which enrich the set of questions that it is possible to ask in modMine. Specifically, it is important to include the reference gene annotations from WormBase (11) and FlyBase (12), and gene function assignments through Gene Ontology annotations (13). Phenotypic data from mutant alleles (11,12) and RNAi screens (27, http://www.flyrnai.org/) can help with assessment of gene function and Reactome pathways (28) provide a broader functional perspective. PubMed gene to publication mappings provide access to the underlying literature but also enable statistical enrichment analysis to help identify papers that are relevant to specified sets of genes. To complement the mapping of transcription factor binding sites to nearby genes, we have included evidence for physical interaction of gene products and of their genetic interactions (14,15) and supplemented this with details on individual proteins from UniProt (16) and their domains (17). Finally, to support comparative studies and linking to related resources, we have included information on gene orthology (10).

Data processing and storage

Data produced by the modENCODE project is vetted and loaded into a Chado database (29), and a specialized XML format (chadoxml) is used to transfer data between different instances of Chado databases. These files are used to generate local Chado database instances of modENCODE data, and modMine loads data from them. Currently, modMine integrates about 70 individual data sets from other sources with the modENCODE data, parsing files in various format (XML, FASTA, GFF3 and key-value pairs) obtained from the producers or using available Chado databases. During this integration phase, conflicts or discrepancies among different sources are resolved by defining the relative priorities of different sources in a field-specific way. The sizes of the different sources vary considerably, from about 1 MB for the PSI (30) ontology file, up to 50 GB for the largest Chado database. Correspondingly, the source processing time also varies greatly from ∼50 s to nearly 18 h. After integration there are a few additional building steps to complete the database, when, for example, sequences are attached to sequence locations and search indexes are created. In total, building modMine takes about a week on a modern high performance Linux server (4x dual core 3.0 GHz AMD Opteron, 32 GB RAM, with a fibre-channel connected SAS 1500 rpm RAID 10 disk array). The size of the postgreSQL database is ∼325 GB before the extensive indexing that optimizes queries to the databases, and around 1 TB after this.

ACCESS, INTERFACE AND UTILITY

The modMine database and web interface provide researchers with a powerful infrastructure to query the modENCODE data and metadata: it adds to the features of the InterMine software platform (5) on which it is based with customized tools and views. Apart from the ability to integrate data from many sources, modMine has many features that make it particularly well suited for accessing modENCODE data. A few examples are described in this article, while a more comprehensive list is available as Supplementary Table S1. In addition, readers may find it useful to refer to the online tutorial http://intermine.modencode.org/help/tour/start.html.

Finding data sets of interest

Searching for modENCODE submissions of interest, using the project metadata, is a key activity. A search box specifically for searching modENCODE experiments in this way is available on the front page of the modMine. To find, for instance, ChIP-chip experiments performed with antibodies against the CTCF factor using chromatin from 0 to 12 h embryonic flies, it would be sufficient to enter 'chip–chip AND ctcf and 0–12' (Figure 1). Selection of individual entries brings up submission report pages giving background on the experiment, a section that details the individual steps taken to generate the data, including relevant parameters and links to protocols, as well as allowing data download in various formats. Note that one can use the provided checkboxes to select one or

Search Term: chip–chip AND ctcf and 0–12
Matching submissions: 2

CREATE LIST

☑	DCC id	Organism	Group	Name	Date	Details	Search score
☑	modENCODE_770		PI: White Lab: White	E0–12_dCTCF N–term ChIP–chip	Aug 8, 2008	antibody: No Antibody Control strain: yellow cinnabar brown speck array: Affymetrix Drosophila Tiling Arrays v2.0R developmental stage: Embryo 0–12 h antibody: CTCF–N cell line: not applicable tissue: not applicable	▬▬
☑	modENCODE_769		PI: White Lab: White	E0–12_dCTCF C–term ChIP–chip	Aug 8, 2008	antibody: No Antibody Control strain: yellow cinnabar brown speck array: Affymetrix Drosophila Tiling Arrays v2.0R developmental stage: Embryo 0–12 h cell line: not applicable antibody: CTCF–C tissue: not applicable	▬▬

Figure 1. Experiment search: entering 'chip–chip AND ctcf AND 0–12' into the experiment search box on the front page returns the two CTCF ChIP-chip experiments using material from 0 to12 h embryos. The check boxes have been ticked in preparation for making them into a list of submissions.

more submissions and then create a list of them using the 'create list' button. One of the useful features of the resulting list page is that one can bring up all the selected data tracks in GBrowse in one step. Another feature is that the list of submissions is named and this name can be used in other parts of the modMine user interface, for instance to run a query using a list of submissions. Such a list, of 12 worm chip-seq experiments using transcription factor/GFP fusions and generated using L2 larvae, is found using 'elegans AND chip-seq AND Anti-eGFP AND L2' and is used as an example in the following section.

Template search library

modMine makes it possible to do more than just search for submissions. For instance, we can use the list of 12 chip-seq submissions generated above as input to a 'template search'. A search template is a web form that simplifies a particular task or type of search, and modMine maintains a library of useful template searches (including ones for finding experiment submissions, for instance those that provide information on a particular gene or genes, or those that use a particular protocol). Suggestions for template searches are welcomed from the user community and template searches are often generated in response to particular user needs. Figure 2 illustrates a template search that, given a gene and a ChIP submission, will report ChIP-identified binding sites upstream of the gene.

A powerful feature of templates is that a gene list can be supplied instead of a single gene and likewise a list of ChIP submissions can be supplied instead of just one: for instance the *C. elegans* ChIP-seq submission list described above will automatically appear in the drop-down menu used to select lists of submissions within the template. In addition to a submission or list of submissions, this template search requires a gene or list of genes. For a gene list it is possible to select a self-made list, or one of several preloaded lists, for instance, the set of known *C. elegans* transcription factors (540 genes with Gene

Ontology annotation GO:0003700). Thus it is simple, using this template search, to find the 1666 ChIP-bound regions identified in multiple experiments as being upstream of the 540 *C. elegans* transcription factor genes. Various export options are available for the results: a table can be exported with comma or tab separated fields; the sequence features can be exported in FASTA format with configurable header fields, or in browser extensible data (BED) or GFF3 sequence feature formats. Finally, data can be transported to the Galaxy system for further analysis using generic tab separated value files or BED format files (http://genome .ucsc.edu/FAQ/FAQformat#format).

Fly gene expression heatmaps and their relationship to gene lists

In order to provide an overview of *D. melanogaster* gene expression during development and also across multiple cell lines, modMine generates heatmaps of gene versus conditions on the fly from user-selected gene lists (Figure 3). The heat maps are interactive and options are available to vary the hierarchical clustering parameters. The gene lists can be generated by uploading gene identifiers, or by selecting genes identified during searches. In fact, the heatmaps are embedded in and represent just one of the several tools available in a list analysis page. Other examples include tools that plot observed versus expected chromosome distributions for the list members, and calculate Gene Ontology term enrichment, publication enrichment and protein domain enrichment. These latter tools are an example of the benefit of integrating non-modENCODE data sets in modMine as they help automatically to highlight unexpected properties of the list. For instance, the PubMed gene/publication mappings are used in the publication enrichment tool to highlight any publications that mention an unexpected number of the genes in the list and which may therefore be of particular interest. modMine supports lists of any class of object in the database, for example the list of dataset submissions described above.

Figure 2. A Template Search which, given a gene (or a list of genes) and a ChIP experiment submission (or a list of submissions), returns ChIP positive annotation in the upstream intergenic region. Options have been selected so that the template search is accepting a list of genes as input, and also a list of submissions. Thus, a single template search can retrieve data corresponding to the product of a gene list and a submission list. The example used finds 1666 regions identified in 12 ChIP experiments as being upstream of the 540 *C. elegans* transcription factor genes. Links to additional features: link (**A**) '< embed results/>', generates text that can be pasted into a web page so that the results appear there; (**B**) 'web service URL', provides a URL to fetch results for the template from the command line or a script; (**C**) 'perl | python | java' generates code in the selected language that will run the template query; (**D**) 'export XML' exports an XML version of the template search (Import is available via the QueryBuilder, or in the 'Query History' section of MyMine).
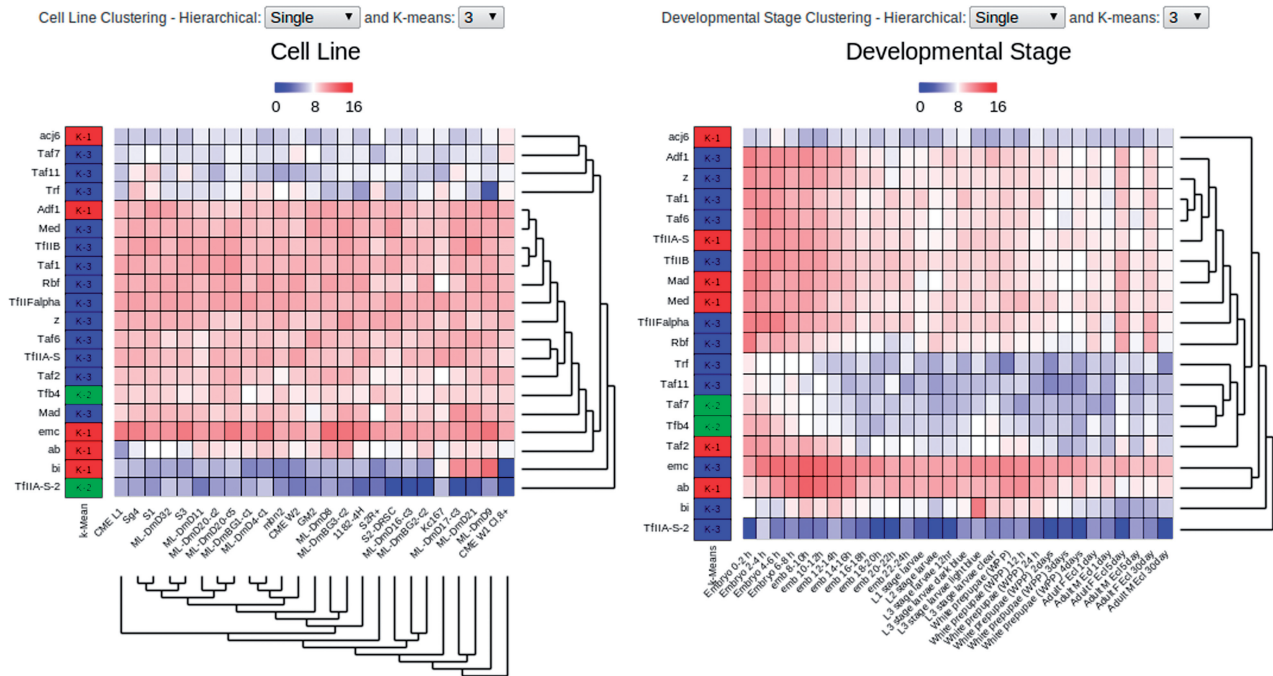

Figure 3. Gene expression heatmap tool. Viewing a list of *Drosophila* genes results in a number of interactive analysis tools being run with default settings including the heatmap tool shown here displaying gene expression across different cell lines and across different developmental stages. This tool is implemented with canvasXpress (http://www.canvasxpress.org).

Region search

A useful function is the ability to extract genome features of interest only from those specific regions of the genome that are of interest. This saves the need to download entire data sets and then write and/or run programs to extract the regions. For instance, region search can be used to help compare the chromatin modification profiles around all fly DNA replication origins to a set of randomly selected chromosomal locations of the same size: spans (chromosome, start and end coordinates) can be pasted into or uploaded into the tool, the modENCODE data sets of interest selected and the data downloaded.

Interactive figures from integrative papers

The first two integrative analysis papers from modENCODE presented overviews of the regulatory interactions between transcription factors as well as microRNAs. modMine provides interactive versions of the paper figures [ref. (2), Figure 3; ref. (3), Figure 7] using Cytoscape (31, http://cytoscapeweb.cytoscape.org/). This allows the influences of individual genes to be highlighted, provides links to the corresponding gene report pages and also allows the graph to be downloaded in eXtensible Graph Markup and Modelling Language and scalable vector graphics XML formats, as well as the SIF graphics format. See the 'About the network...' button for help on interacting with the graphs.

Features for programmers

modMine, exploiting its powerful InterMine engine, provides extensive customization and automation facilities for use by programmers. The system allows fine-grained control over the query to be run and the format of the data that is returned.

Web service capabilities

Any query, whether a predefined template (Figure 2B) or constructed with the QueryBuilder, can be run via HTTP web services. This means any user with HTTP web access can directly query the modENCODE data-warehouse from their own programs or embed results in their own pages. InterMine web services are designed using principles of RESTful design, describing their features as stateless resources, defined by parameters, with a variety of representations. Thus, the same set of query results can be retrieved as tab or comma-delimited values, JSON (http://www.json.org) or XML. Various metadata about the InterMine system can also be accessed such as the data-model, and the list and composition of the available Template Searches and Lists. Access to these services can be authenticated, giving a user access to data stored in their personal accounts.

Access from scripts and programs

Web service client libraries have been published in standard repositories such as CPAN (http://www.cpan.org), PyPi (http://pypi.python.org) and Rubygems (http://rubygems.org) for simplifying access in Perl, Python and Ruby. InterMine itself hosts Java and JavaScript libraries for the same purpose. As well as data-retrieval, the JavaScript library allows tables of query results to be embedded in any other web page (Figure 2A). To lower the bar to development using these web services, InterMine can generate code in the above languages for any query that can be run in the web application. (Figure 2C). Documentation for all the resources accessible through the InterMine web service API is available at http://intermine.org/wiki/WebService.

CONCLUSION AND PLANS

We have integrated data sets from the modENCODE project and relevant non-modENCODE data sets to make modMine a database that enables flexible interrogation of these combined resources. Future versions will include further modENCODE data sets as they become available, but will also address limitations in the current system. For instance, soon a new version of the search results pages will be made available, which allows simple rearrangement and filtering of columns of data, as well as sorting by the values in one or more columns. Improvements will also be made to the region download tool, so that, for instance, it can accept a previously generated list of submissions as a way of selecting the types of data to be extracted, or can read GFF3 and other common formats as a way of defining the regions of interest.

modMine will also benefit from work underway as part of the interMOD project, which is building InterMine databases alongside the major model organism databases for budding yeast, mouse, rat, worm and zebrafish. A particular focus of the interMOD project is interoperation between the different databases, in which, when browsing one database, features of interest from the others will be made available automatically.

modMine will be kept available indefinitely using current servers. In addition, a virtual instance of modMine will be implemented on the Amazon EC2 cloud at the end of the project, ensuring that modMine will continue to be available to the research community.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Table 1.

FUNDING

Conflict of interest statement. None declared.

REFERENCES

1. 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.

2. Gerstein,M.B., Lu,Z.J., Van Nostrand,E.L., Cheng,C., Arshinoff,B.I., Liu,T., Yip,K.Y., Robilotto,R., Rechtsteiner,A., Ikegami,K. *et al.* (2010) Integrative analysis of the Caenorhabditis elegans genome by the modENCODE Project. *Science*, **30**, 1775–1787.

3. The modENCODE Consortium *et al.* (2010) Identification of functional elements and regulatory circuits by Drosophila modENCODE. *Science*, **330**, 1787–1797.

4. Washington,N.L., Stinson,E.O., Perry,M.D., Ruzanov,P., Contrino,S., Smith,R., Zha,Z., Lyne,R., Carr,A., Kephart,E. *et al.* (2011) The modENCODE Data Coordination Center: lessons in harvesting comprehensive experimental Details. *Database*, doi: 10.1093/database/bar023.

5. Lyne,R., Smith,R., Rutherford,K., Wakeling,M., Varley,A., Guillier,F., Janssens,H., Ji,W., Mclaren,P., North,P. *et al.* (2007) FlyMine: an integrated database for Drosophila and Anopheles genomics. *Genome Biol.*, **8**, Article RI29.

6. Brazma,A., Hingamp,P., Quackenbush,J., Sherlock,G., Spellman,P., Stoeckert,C., Aach,J., Ansorge,W., Ball,C.A., Causton,H.C. *et al.* (2001) Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat. Genet.*, **29**, 365–371.

7. Taylor,C.F., Field,D., Sansone,S.A., Aerts,J., Apweiler,R., Ashburner,M., Ball,C.A., Binz,P.A., Bogue,M., Booth,T. *et al.* (2008) Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. *Nat. Biotechnol.*, **26**, 889–896.

8. Eilbeck,K., Lewis,S., 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.

9. Whetzel,P.L., Parkinson,H., Causton,H.C., Fan,L., Fostel,J., Fragoso,G., Game,L., Heiskanen,M., Morrison,N., Rocca-Serra,P. *et al.* (2006) The MGED Ontology; a resource for semantics-based description of microarray experiments. *Bioinformatics*, **22**, 866–873.

10. Ruan,J., Li,H., Chen,Z., Coghlan,A., Coin,L.J., Guo,Y., Hériché,J.K., Hu,Y., Kristiansen,K., Li,R. *et al.* (2008) TreeFam: 2008 update. *Nucleic Acids Res.*, **36**, D735–D740.

11. Harris,T.W., Antoshechkin,I., Bieri,T., Blasiar,D., Chen,J., Chen,W.J., De La Cruz,N., Davis,P., Duesbury,M., Fang,R. *et al.* (2010) WormBase: a comprehensive resource for nematode research. *Nucleic Acids Res.*, **38**, D463–D467.

12. 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.

13. GO ontology consortium. (2006) The Gene Ontology (GO) project in 2006. *Nucleic Acids Res.*, **34**, D322–D326.

14. Aranda,B., Achuthan,P., Alam-Faruque,Y., Armean,I., Bridge,A., Derow,C., Feuermann,M., Ghanbarian,A.T., Kerrien,S., Khadake,J. *et al.* (2010) The IntAct molecular interaction database in 2010. *Nucleic Acids Res.*, **38**, D525–D531.

15. Stark,C., Breitkreutz,B.-J., Chatr-aryamontri,A., Boucher,L., Oughtred,R., Livstone,M.S., Nixon,J., Van Auken,K., Wang,X., Shi,X. *et al.* (2011) The BioGRID Interaction Database: 2011 update. *Nucleic Acids Res.*, **39**, D698–D704.

16. The UniProt Consortium. (2010) The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res.*, **38**, D142–D148.

17. Hunter,S., Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Binns,D., Bork,P., Das,U., Daugherty,L., Duquenne,L. *et al.* (2009) InterPro: the integrative protein signature database. *Nucleic Acids Res.*, **37**, D211–D215.

18. Stein,L.D., Mungall,C., Shu,S., Caudy,M., Mangone,M., Day,A., Nickerson,E., Stajich,J.E., Harris,T.W., Arva,A. *et al.* (2002) The generic genome browser: a building block for a model organism system database. *Genome Res.*, **12**, 1599–1610.

19. Hong,E.L., Balakrishnan,R., Dong,Q., Christie,K.R., Park,J., Binkley,G., Costanzo,M.C., Dwight,S.S., Engel,S.R., Fisk,D.G. *et al.* (2008) Gene Ontology annotations at SGD: new data sources and annotation methods. *Nucleic Acids Res.*, **36**, D577–D581.

20. Twigger,S.N., Shimoyama,M., Bromberg,S., Kwitek,A.E. and Jacob,H.J. (2007). RGD Team. (2007) The Rat Genome Database, update 2007 – easing the path from disease to data and back again. *Nucleic Acids Res.*, **35**, D658–D662.

21. Bradford,Y., Conlin,T., Dunn,N., Fashena,D., Frazer,K., Howe,D.G., Knight,J., Mani,P., Martin,R., Moxon,S.A. *et al.* (2011) ZFIN: enhancements and updates to the zebrafish model organism database. *Nucleic Acids Res.*, **39**, D822–D829.

22. Blake,J.A., Bult,C.J., Kadin,J.A., Richardson,J.E. and Eppig,J.T. (2011). the Mouse Genome Database Group. (2011) The Mouse Genome Database (MGD): premier model organism resource for mammalian genomics and genetics. *Nucleic Acids Res.*, **39**, D842–D848.

23. Smith,A.C. and Robinson,A.J. (2009) MitoMiner, an integrated database for the storage and analysis of mitochondrial proteomics data. *Mol. Cell Proteomics*, **8**, 1324–37.

24. 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.

25. Chen,Y.-A., Tripathi,L.P. and Mizuguchi,K. (2011) TargetMine, an integrated data warehouse for candidate gene prioritisation and target discovery. *PLoS One*, **6**, e17844.

26. Barrett,T., Troup,D.B., Wilhite,S.E., Ledoux,P., Evangelista,C., Kim,I.F., Tomashevsky,M., Marshall,K.A., Phillippy,K.H., Sherman,P.M. *et al.* (2011) NCBI GEO: archive for functional genomics data sets-10 years on. *Nucleic Acids Res.*, **39**, D1005–D1010.

27. Flockhart,I., Booker,M., Kiger,A., Boutros,M., Armknecht,S., Ramadan,N., Richardson,K., Xu,A., Perrimon,N. and Mathey-Prevot,B. (2006) FlyRNAi: the Drosophila RNAi screening center database. *Nucleic Acids Res.*, **34**, D489–D494.

28. Croft,D., O'Kelly,G., Wu,G., Haw,R., Gillespie,M., Matthews,L., Caudy,M., Garapati,P., Gopinath,G., Jassal,B. *et al.* (2011) Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res.*, **39**, D691–D697.

29. Mungall,C.J. and Emmert,D.B. (2007). FlyBase Consortium. (2007) A Chado case study: an ontology-based modular schema for representing genome-associated biological information. *Bioinformatics*, **23**, i337–i346.

30. Kerrien,S., Orchard,S., Montecchi-Palazzi,L., Aranda,B., Quinn,A.F., Vinod,N., Bader,G.D., Xenarios,I., Wojcik,J., Sherman,D. *et al.* (2007) Broadening the horizon–level 2.5 of the HUPO-PSI format for molecular interactions. *BMC Biol.*, **5**, 44.

31. Lopes,C.T., Franz,M., Kazi,F., Donaldson,S.L., Morris,Q. and Bader,G.D. (2010) Cytoscape Web: an interactive web-based network browser. *Bioinformatics*, **26**, 2347–2348.