Published in final edited form as:

Genesis. 2015 August ; 53(8): 547–560. doi:10.1002/dvg.22869.

Cross-Organism Analysis Using InterMine

Rachel Lyne[1,2], **Julie Sullivan**[1,2], **Daniela Butano**[1,2], **Sergio Contrino**[1,2], **Josh Heimbach**[1,2], **Fengyuan Hu**[1,2], **Alex Kalderimis**[1,2], **Mike Lyne**[1,2], **Richard N. Smith**[1,2], **Radek Štěpán**[1,2], **Rama Balakrishnan**[3], **Gail Binkley**[3], **Todd Harris**[4], **Kalpana Karra**[3], **Sierra A. T. Moxon**[5], **Howie Motenko**[6], **Steven Neuhauser**[6], **Leyla Ruzicka**[5], **Mike Cherry**[3], **Joel Richardson**[6], **Lincoln Stein**[4], **Monte Westerfield**[5,7], **Elizabeth Worthey**[8], and **Gos Micklem**[1,2]

[1]Cambridge Systems Biology Centre, University of Cambridge, Cambridge CB2 1QR, United Kingdom

[2]Department of Genetics, University of Cambridge, Cambridge CB2 3EH, United Kingdom

[3]Department of Genetics, Stanford University, Stanford, CA 94305-5120, USA

[4]Ontario Institute for Cancer Research, Toronto, ON, M5G0A3, Canada

[5]ZFIN, University of Oregon, Eugene, OR, 97403, USA

[6]The Jackson Laboratory, Bar Harbor, Maine, 04609, USA

[7]Institute of Neuroscience, University of Oregon, Eugene, OR, 97403, USA

[8]Human and Molecular Genetics Center, Medical College of Wisconsin, Milwaukee, WI, 53226, USA

Abstract

InterMine is a data integration warehouse and analysis software system developed for large and complex biological datasets. Designed for integrative analysis it can be accessed through a user-friendly web interface. For bioinformaticians, extensive web services as well as programming interfaces for most common scripting languages support access to all features.

The web interface includes a useful identifier look-up system, and both simple and sophisticated search options. Interactive results tables enable exploration, and data can be filtered, summarised and browsed. A set of graphical analysis tools provide a rich environment for data exploration including statistical enrichment of sets of genes or other entities.

InterMine databases have been developed for the major model organisms, budding yeast, nematode worm, fruit fly, zebrafish, mouse and rat together with a newly developed human database. Here we describe how this has facilitated interoperation and development of cross-organism analysis tools and reports. InterMine as a data exploration and analysis tool is also described. All the InterMine based systems described in the paper are resources freely available to the scientific community.

Corresponding Author: Gos Micklem: Cambridge Systems Biology Centre, University of Cambridge, Cambridge CB2 1QR, United Kingdom; g.micklem@gen.cam.ac.uk, +44 (1223) 760240.

Keywords

Data integration; Data analysis; Comparative analysis; Cross-organism analysis; Integrative analysis; Genomics; Proteomics

Introduction

With the advent of widespread whole-genome sequencing and high-throughput genome-wide experiments, managing the accelerating growth in biological data production remains an ongoing challenge.

The data being produced are not just large in volume but also diverse and often challenging to analyse. The regulation of biological systems occurs at many levels and can only be fully understood by analysing many types of data together to maximise the knowledge extracted from high-volume datasets. The searches required are often complex and the data may be contained in several databases or flat data files, or have been studied in several organisms. Bringing the data together in a suitable form and analysing them can be a particularly formidable task for a biologist. InterMine is a data warehousing system for the integration of such diverse biological data, together with an intuitive web-based user-interface as well as web services and application programming interfaces (APIs) for bioinformaticians.

The InterMine platform was originally developed over 10 years ago for *Drosophila melanogaster* data (FlyMine, www.flymine.org, Lyne et al., 2007) and has since grown to cover many organisms and data types, including annotated genome features from the modENCODE project (modMine, www.intermine.modencode.org, Contrino et al., 2012), *Toxoplasma gondii* data (toxoMine www.toxomine.org), drug discovery data (targetMine www.targetmine.nibio.go.jp, Chen et al., 2011), plant genomics data (phytoMine, http://phytozome.jgi.doe.gov/phytomine), mitochondrial proteomics (mitoMiner, http://mitominer.mrc-mbu.cam.ac.uk, Smith AC et al., 2012), Drosophila transcription factors (flyTF, http://www.flytf.org, Pfreundt et al., 2010) and microbial genomics data (INDIGOmine (http://www.cbrc.kaust.edu.sa/indigo, Alam et al., 2013). Although many biological data management systems have been established, and in-particular we note BioMart (Smedley et al., 2015), The Eukaryotic Pathogen Databases (EuPathDB, Harb et al., 2015) and BioCyc (Caspi et al., 2014), each is appropriate in different scenarios and the InterMine system provides several unique features. In addition, the fact that InterMine has been adopted by the major model organisms to provide an advanced interface to MOD data gives it a unique position in cross-organism analysis and translational research.

Model organism databases (MODs) curate and collate genomic data for a specific organism, or a range of related organisms. Such databases exist for the major model organisms, mouse (MGI, Eppig et al., 2015), rat (RGD, Shimoyama et al., 2015), zebrafish (ZFIN, Howe et al., 2013), fly (FlyBase, dos Santos et al., 2015), nematode (wormbase, Yook et al., 2011) and budding yeast (SGD, Cherry et al., 2012). However, each of these databases are run independently and with different underlying infrastructure, thus providing a barrier to comparative analysis. The launch of the InterMOD project in 2009 extended the range of organisms available through an InterMine database. The InterMOD project funded five of the major model organisms, mouse (MouseMine, www.mousemine.org, Eppig et al., 2015), rat (RatMine, http://ratmine.mcw.edu/), zebrafish (ZebrafishMine, http://www.zebrafishmine.org, Ruzicka et al., 2015), nematode (WormMine, http://www.wormbase.org/tools/wormmine) and budding yeast (YeastMine, www.yeastmine.yeastgenome.org, Balakrishnan et al., 2012) to build data platforms using the InterMine system. This has not only provided each of these MODs with a powerful query system for their data, but also unites each MOD with a common platform - thus facilitating uniform and consistent cross-organism analysis. Throughout this paper these databases, together with the analogous FlyMine database, will collectively be referred to as the MOD-InterMine databases. To complement this project the InterMine team have also created a HumanMine database (www.humanmine.org), which generalises the earlier metabolicMine database (Lyne et al., 2013) and is focussed on human genomics and proteomics datasets, thus helping to allow the interpretation of model organism data in a biomedical context.

In this paper we provide an overview of the InterMine system as an inter- and intra-organism analysis platform, describing use of the InterMine search and analysis tools and some of the challenges in cross-organism analysis.

The InterMine System

InterMine has been described in detail elsewhere (see Smith RN et al., 2012 for a more technical overview), but we briefly describe here the main features that make InterMine a useful system. At its core InterMine consists of the ObjectStore, a custom object/relational mapping system written in Java and optimized for read-only database performance. Object queries from the web application or web services are sent to the ObjectStore which generates SQL to execute in the underlying PostgreSQL database and materializes objects from the results. InterMine is able to integrate data from a wide variety of sources in many formats commonly used with biological data, including GFF3, FASTA, OBO, BioPAX, GAF, PSI and Chado and includes a powerful identifier resolution system such that any outdated identifiers from a dataset can be replaced with the current ones. The integrated data can be accessed through a sophisticated web interface described in more detail below. In addition, the range of analysis tools provided by the MOD-InterMine databases is extended through interoperation with both Galaxy (Goecks et al., 2010) and Genome Space (www.genomespace.org). Galaxy is a bioinformatics web-based platform particularly suited for analysis of biological sequence data while Genome Space provides an interoperability framework to a diverse range of bioinformatics tools allowing easier transmission of data between tools. Data from InterMine searches, including sequence data, can be seamlessly uploaded into both Galaxy and Genome Space for further analysis.

For bioinformaticians, the InterMine databases can also be accessed programmatically through the same RESTful web services that are used to create the web interface. Client libraries are available in a number of programming languages including Perl, Python, Ruby, Java and Javascript. The web services allow the creation of powerful automated analysis workflows, their use being aided by automatic code generation from searches within the web interface and extensive documentation (http://intermine.readthedocs/en/latest/web-services). The web services and their use have been described in detail in Kalderimis et al., 2014.

InterMine is open source software with an active developer community, communicating extensively through a developer mailing list. The InterMod project was specifically funded to provide an InterMine database for the organisms specified in the paper. However, we welcome interest from other groups and are happy to provide technical support. We also welcome contributions in terms of new tools, code improvements and suggestions for new features from the biological community.

Facilitating cross-organism analysis

Ultimately the aim of many studies in model organisms is to further understanding of human biology and disease and eventually facilitate translation of research into clinical practice. However, the transfer of biological knowledge across organisms can be time-consuming and requires specialised knowledge. The InterMine system brings data from the participating model organisms together under a common platform, thus providing a unique opportunity to provide interoperation between related data types and providing researchers with a single common platform through which to access and analyse the data. In creating such a unified common platform across multiple organisms we have faced many challenges, ranging from the use of diverse terms for the same entity and varying conventions for data representation, to the challenge of relating ontologies across species and the multifarious sets of homologue mappings between species.

A number of links between cross-species data can potentially be used for comparative analysis. One of the most widely used is orthology. Orthologues are genes in different species that evolved from a common ancestral gene by speciation. Often, orthologues retain the same function in the course of evolution and thus provide the most common route to knowledge transfer between organisms. Indeed, many annotations are based solely on transfer of data between orthologous genes (e.g. Reactome pathways, Croft et al., 2011, Gene Ontology annotations, Ashburner et al., 2000). Such orthologous relationships provided the most obvious initial choice of data for cross-InterMine links and from any gene or list of genes one can “jump” between MOD-InterMine databases through orthologues (Figure 1, Sullivan et al., 2013). However, this method is not without its challenges. Numerous algorithms have been developed to identify all orthologous genes between organisms, with over 30 phylogenomic databases providing orthologue sets (Sonnhammer et al., 2014). These databases differ not only in their algorithms but also in the underlying datasets used for analysis and the range of organisms analysed. Thus, comparison between sets is difficult and one set of predictions may be better for one organism than another. Consequently, annotations inferred through orthologues may also be erroneous and this should be considered when analysing such data. However, orthologues currently provide the best relationships between genes in different organisms and with collection of more data and refinements in methods their overall quality is likely to improve over time. Currently each MOD-InterMine database provides it’s own preferred set of orthologues for interoperation, thus when moving between databases, the orthologue set of the database being moved to is used. This has the additional advantage that curated orthologues provided by some MODs can be used (Table 1). In addition to "jumping" between MOD-InterMine databases, orthology relationships allow the collation of annotations from multiple MOD-InterMine databases and their display in tabular form. For instance, tables showing pathway annotations for a gene, from *D. melanogaster, M. musculus* and *H. sapiens* and rat (*R. norvegicus)* disease annotations, are available on FlyMine gene report pages (Sullivan et al., 2013).

One way to facilitate consistent queries across databases is to create a shared data model, where data is represented in a uniform way across all instances. The Sequence Ontology (SO, Eilbeck et al., 2005), is used to provide the basis for the InterMine core model. This core is expanded to include types of data that are not covered by the SO, such as protein interaction data, publications and species-specific features. Each MOD-InterMine is therefore based on the same core model thus enabling some consistency in searches against all MOD-InterMine databases and this has made it possible to create a number of pre-defined searches that return equivalent data from each database. In addition, efforts have been made to unify additional aspects of the data model (Sullivan et al., 2013), with some success, including interaction data and protein domain data.

Problems unifying the entire data models, however, have been encountered (Sullivan et al., 2013), particularly where species-specific ontologies have been used to annotate experimental data, for instance phenotypes. For example, the model organism databases have each evolved their own species-specific ontologies for the annotation of anatomy and phenotype. In the case of mouse phenotypes, data stored in MGI is annotated with the Mammalian Phenotype Ontology (Smith and Eppig., 2012), while fly experiments in FlyBase are annotated with the Drosophila Phenotype Ontology (Osumi-Sutherland et al., 2013). Such differences are further complicated by different phenotype representations. ZFIN, for example, describe phenotypes using an Entity-Quality (EQ) methodology, with E being the affected entity and Q how it is affected (Sprague et al., 2008), while other ontologies, such as the Mammalian Phenotype Ontology and Human Phenotype Ontology (Kohler et al., 2014), are pre-constructed, and therefore use only one term per annotation. Such differences lead to diversity in both underlying data and the data model, and make it difficult both to create a unified data model and to create systematic links between data from different organisms. Sophisticated algorithms are required to create maps between comparable phenotype terms and to create species-agnostic ontologies. The development and application of such algorithms is in progress (e.g. Kohler et al., 2013, Mungall et al., 2012) and the use of such mappings should become an integral part of cross-organism interoperation through InterMine in the future.

InterMine is continually reviewing the data and data models and is developing tools to overcome disparities in underlying data and data structure to help improve cross-organism analysis and provide new connections between equivalent data, for instance through synteny and expression data, in order to promote powerful comparative analysis (Sullivan et al., 2013).

Data Content

The core of the MOD-InterMine databases consists of biological domain knowledge - data extracted from functional databases that provide links between a gene or protein and its function, often through a controlled vocabulary or ontology, such as the Gene Ontology (Ashburner et al., 2000), KEGG (Kanehisa et al., 2012) and Reactome pathways (Croft et al., 2011) (Table 1). Such sources provide a background of interpretable biological knowledge which researchers can apply to their own data. In addition, the InterMine databases load experiment data such as expression data from ArrayExpress and GEO, and analysed high-throughput data such as those produced by RNA-seq and CHIP-seq, as well as RNAi phenotype screens (Table 1). The range and depth of data loaded is being expanded for most InterMine databases. Integration of such data provides a path to inference of causal relationships between biological entities, for instance, between gene expression and diseases, genes and phenotypes and for the dissection of disease traits.

All the InterMine databases provide a data page - usually accessible through the "data sources" tab on any page. This page lists all the data loaded into the particular InterMine instance along with links to the original source, publications and the version or date the data was loaded. This is a good starting point for users to familiarise themselves with the range of data available. The content of each MOD-InterMine database varies, but some key datasets are available in all (Table 1).

The InterMine web interface

The InterMine interface is designed to facilitate both data discovery and data exploitation. The data discovery interfaces include a keyword search and detailed interactive report pages collating all integrated data for each object or list of objects in the databases. For data exploitation, several query interfaces of varying sophistication are provided to allow the researcher to interrogate the data in a number of ways. The interfaces have been designed to overcome some of the obstacles to large scale data analysis and this will be demonstrated through two use-cases, one involving data exploration and one illustrating an approach to data exploitation through candidate gene filtering.

Knowledge discovery

Keyword search and report pages

The simplest way to search InterMine databases, the keyword search, allows all types of data within a database to be interrogated, thus facilitating data exploration and discovery. The search is executed by an Apache Lucene search and index system (www.lucene.apache.org) and the faceted results allow drilling down through the returned categories to show data of interest.

One advantage of data integration is that all the information about a particular biological entity can be surveyed together without the need to visit several databases. InterMine collates data for each biological entity into comprehensive report pages allowing in-depth exploration of the entity and related entities. Each item returned by the keyword search links to such a report page and every object within an InterMine database has a report page, whether it is a gene, a binding site or a publication. Report pages collate the data into interactive tables and graphical displays. Related entities are interconnected (for example a gene report page will link to its associated protein report page), thus allowing navigation through the relationships that exist at different biological levels, for example from gene expression to protein interactions.

We will illustrate knowledge discovery through the InterMine interface by examining the relationships between the functional levels of biological data for the *D. melanogaster eyeless* (*ey*) gene in FlyMine, and cross-examining these data for a mouse orthologue, *Pax6* (Figure 1). The *eyeless* gene is a homeobox transcription factor known to be essential for eye development in Drosophila. A keyword search for "*ey"* in FlyMine returns several categories of data including genes, publications, interaction experiments and an RNAi screen result. For our example, we are interested in analysing the *D. melanogaster ey* gene and so we navigate to the report page for this gene. First we examine the knowledge domain data to identify key annotations for this gene - the Gene Ontology annotations indicate that this gene is indeed involved in compound eye development with a molecular function indicating transcription factor binding activity.

The pathway data, however, only indicate involvement in "Developmental Biology", a high-level annotation term, but also interestingly point to an involvement in regulation of pancreatic cell development. Disease annotations of the orthologous rat gene (available directly on the FlyMine gene report page) indicate several eye-related diseases including aniridia, coloboma of optic nerve, and eye abnormalities as well as an involvement in diabetes-related illnesses.

Gene expression data reveal that the *eyeless* gene is expressed in the adult eye, head, brain, thoracicoabdominal ganglion and the larval CNS. Navigating to the report page for the eyeless protein (Pax6, Uniprot Identifier PAX6_DROME), one learns that the protein has both a homeobox domain and a paired domain, often found in transcriptional regulators that control aspects of development. The protein interaction viewer reveals several physical interactions for the eyeless protein. Analysis of a list created from the interaction network identifies a number of genes enriched for eye development and morphogenesis (see Data Exploitation for further details on list analysis and enrichment). Further exploration of the gene report page reveals that the *eyeless* gene is highly conserved across species, with links to orthologous genes available to RatMine, MouseMine, ZebrafishMine and HumanMine. Linking to the orthologous mouse *Pax6* gene and again viewing the Gene Ontology annotations we learn that this gene has transcriptional regulatory activity and is involved in eye and brain development. Expression data indicate that it is expressed in various brain structures, the eye and the pancreas (among other tissues) and annotations to the Mammalian Phenotype Ontology show mutant phenotypes including abnormal eye and brain morphology and abnormal pancreatic function. Disease annotations, via orthologous human genes, include eye conditions such as aniridia, anterior segment mesenchymal dysgenesis and keratitis. Examination of the human *Pax6* gene provides similar data, thus suggesting that eye morphogenesis is under similar genetic control in both insects and vertebrates and that the Drosophila *ey* gene and mouse and human *Pax6* genes have an additional role in pancreatic function. Although eye development and morphogenesis have been well studied, this example illustrates how the connections between the data for multiple organisms allows a researcher to build up a detailed picture of what is already known about a particular set of genes and their biological function.

Template searches

Data exploration in a more structured fashion is available through *template searches* and the *query builder*. Template searches are predefined database searches provided via a simple interface with customisation available through one or more configurable filters. Template searches range from simple searches covering one data type (e.g. Genes -> Pathways) to more complex searches spanning several data types (e.g. Expression + Interactions -> Genes) (Figure 2).

InterMine databases provide a library of these pre-defined templates providing users with immediate and easy access to the integrated data. For more advanced searching, researchers can use the Query Builder to either modify existing template searches or to build their own queries from scratch. The Query Builder provides an intuitive interface for browsing the underlying InterMine data model, selecting and applying constraints (filters) to the data types of interest and configuring the columns of data required in the output (Figure 3).

Importantly, both template searches and the query builder can be configured to operate with lists of entities allowing, for instance, all the genes identified in a screen to be examined at once. Likewise, both template searches and the query builder deliver results in an interactive table and this provides additional spreadsheet-like analysis tools such as filtering, sorting, addition and removal of columns and the creation of lists for further analysis. A particular feature of the results table are the column summary buttons at the top of each column: these summarise the complexity of data present in the column and are especially useful when examining large result sets (Figure 4). The column summaries provide a way to interactively filter the data in results tables, as they provide a mechanism to select only those rows that carry particular ranges of values in the specified column.

Data can also be exported from results tables in a number of formats, notably tab-separated, comma-separated, XML, GFF3, JSON, BED and FASTA (for the corresponding protein or nucleic acid sequences). Template searches, the query builder and the results tables allow researchers to define multiple constraints (filters) on the integrated data in order to examine whether a particular relationship exists. For example, we could run the template “Gene —> GO terms” with our *eyeless* gene. Filtering through the results tables would allow us to identify only those annotations that refer to the eye and that have an experimental evidence code (in this case IMP, inferred from mutant phenotype) (Figure 4). Secondly, we could run the template “Expression + Interactions —> Genes” to show genes that interact with *eyeless* that are also expressed in the adult eye. We could then create a list of these genes for further analysis (see below for analysis of lists). This latter template illustrates how different data types can be combined to both answer a particular question, identify genes for further analysis and add weight to an assertion implied by particular data.

Although template searches are configured independently for each MOD-InterMine, a number of core templates enabling consistent cross-organism analysis are available. The use of standard template searches across MOD-InterMine databases can overcome some of the problems of data heterogeneity between organisms and most of the common data types can be accessed through template searches in each MOD-InterMine (Table 1). Even if the underlying data model differs, templates can hide some of this complexity from the user by providing a simplified interface to the data.

Data exploitation

The InterMine databases are especially suited to the analysis of lists of data. A 'list' is a working unit, allowing researchers to pass data from one analysis step to another, thus creating iterative analyses. A list is a set of one or more single entities - e.g a list of genes or a list of proteins. Such lists can be uploaded into an InterMine database or created within one from results tables and analysis tools. Lists can be used in searches, which, together with the filtering feature of the results table, provides the means to create lists with defined properties - for example, all genes expressed in a certain tissue, all genes with a particular GO annotation. Additionally, a set of list operations (union, intersect and subtractions) allows further manipulation of lists.

We illustrate the use of lists in InterMine, particularly for comparative analysis, to analyse a set of candidate genes. Many high-throughput experimental methods ultimately result in a list of candidate genes which the researcher must investigate further - often to reduce the list to those genes or proteins that look most interesting to the trait or process being studied, and to formulate a hypothesis for further study. Our example will take a hypothetical set of human genes identified in a screen for lipid and cholesterol markers as part of a study on atherosclerosis. As our results are from a human study, we are interested in finding model systems in other organisms that may help further our study. A typical analysis workflow is now described in a series of steps (Figure 5):

1. Upload candidate gene set—A common problem in the biological sciences is keeping track of up-to-date identifiers for entities like genes and proteins. Identifiers often change between genome annotation releases and so researchers can find that they are working with sets of outdated identifiers, and that identifiers from publications have become outdated. InterMine provides an *Identifier Resolution System* to help researchers resolve identifiers into the most up to date set and is automatically instigated whenever a user uploads a list. For more details on the identifier resolution system see Smith RN et al., 2012. Thus the first step in our analysis workflow is to upload our candidate set and check we have an up-to-date and correct set of identifiers. Since we have a list of human genes we will upload them into the HumanMine database.

2. Identify genes in the list already associated with the disease—The InterMine databases provide disease and phenotype data which can be interrogated through template searches. For this example, we will use a template from HumanMine, "Gene + HPO parent term → Genes", to identify genes from our list annotated with the HPO (Human Phenotype Ontology) term "Atherosclerosis". This is a high-level ontology term and the template search has been constructed so that it also returns genes annotated to any child-terms such as coronary atherosclerosis, precocious atherosclerosis and myocardial infarction. All

InterMine searches can be executed against either a single entity or a list of entities and so we are able to generate results for all members of our list of genes in one step. A second list is created from the results table, containing the genes returned by the search - i.e. those genes with a atherosclerosis-related annotation.

3. Create a list of the genes not associated with the disease—Further creation of sets of interesting entities is possible through the list operations, union, intersect, subtract and asymmetric difference. In addition it is possible to add entities to an already existing list from results tables. In our example we use an asymmetric difference to create a new list which contains only the genes which do not have an atherosclerosis-related HPO annotation. This is achieved by selecting the lists you wish to perform the operation on, in this case, our original uploaded list and the list created from the template search in the last step. The asymmetric difference will find those entities in the first list which are not present in the second list and create a new list containing just these entities. This is our set of "non-atherosclerosis" genes that we will analyse further.

4. Functional interpretation of the non-atherosclerosis list in other organisms—So far we have carried out our example analysis within the HumanMine database. However, we are interested in investigating the model organisms to find any suitable for further studies. From our set of "non-atherosclerosis" genes we see that orthologous genes exist in budding yeast, fruit fly, rat, mouse and zebrafish. We choose to investigate those in mouse and zebrafish further as we expect these to most closely resemble the human system. We can navigate straight to both the MouseMine and ZebrafishMine databases where the set of genes orthologous to our "non-atherosclerosis" set will be displayed.

We start our analysis of the orthologous mouse and zebrafish genes by examining some enrichment statistics. Set enrichment analysis is a popular way to systematically analyse lists of e.g. genes (Huang et al., 2009) by comparison with already available knowledge. Such overrepresentation analysis is provided by the InterMine databases automatically and for a number of types of annotations, including Gene Ontology terms, protein domains, pathways, publications, disease ontology terms (MouseMine), Mammalian Phenotype Ontology terms (MouseMine) and Anatomy Ontology terms (MouseMine). Enrichment analysis compares the frequency of individual annotations within the list to the annotation frequency in a reference list. In InterMine this is calculated using the hypergeometric distribution with a default Holm-Bonferonni multiple testing correction applied and p-value cut-off of 0.05. More stringent methods of multiple test correction, Benjamini Hochberg or Bonferonni (Huang et al., 2009), can also be applied and the p-value cut-off can be changed. The default reference population is all the genes that have the particular annotation within the relevant genome, but this reference set can also be defined by the user. Such set enrichment analysis allows one to summarise the functional properties of a list as the set of overrepresented functions. The tables of enriched annotations provided as output by InterMine enable the creation of sub-lists of entities for one or more enriched terms, thus facilitating the further analysis of such sets. FlyMine, for example, has been used for enrichment analysis in several studies (e.g. Favrin et al., 2013, Aleksic et al., 2013, Lowe et al., 2014, Seridi et al., 2014).

In our example, we will first look at the Mammalian Phenotype Ontology (MPO), and Gene Ontology enrichments for our set of “non-atherosclerosis” mouse orthologues. This information is provided automatically when we open the list analysis page. We notice that there are a number of genes enriched for lipid and cholesterol phenotypes and processes, including abnormal lipid level, abnormal homeostasis (MPO) and lipid and cholesterol transport (Gene Ontology). For our “non-atherosclerosis” zebrafish genes the Gene Ontology term enrichments point to a number of the genes being involved in lipid and cholesterol localisation and transport (Figure 6). We decide to investigate those genes involved in cholesterol transport further. Lists containing only those genes with this Gene Ontology annotation are created from both the mouse and zebrafish Gene Ontology enrichment tables. The zebrafish set contains six genes. A simple search for orthologues of these six genes within ZebrafishMine reveals that all but one of these genes, *Cetp*, has a mouse orthologue while all six have an orthologue in human. The *Cetp* gene is thus targeted for further analysis. The report page in HumanMine reveals that the human *Cetp* gene may affect susceptibility to atherosclerosis, is expressed in the liver and is secreted into the plasma. A literature search through ZebrafishMine reveals that the *Cetp* gene is indeed expressed in zebrafish but not mice and this difference makes zebrafish an important model for the study of atherosclerosis (Fang et al., 2014).

Our analysis example illustrates how lists can form the basic unit of iterative analysis both within a single InterMine and across multiple MOD-InterMine databases. Lists can be passed from one InterMOD database to another and can be successively filtered through a range of analysis tools, with a series of intermediate sub-lists being created. In our hypothetical use-case we showed how analysis mediated by lists used across multiple MOD-InterMine databases highlighted fundamental similarities and differences between organisms and led to the identification of the most suitable model for further study. Searches, list operations and graphical and statistical displays enable the creation of lists of entities with a defined set of properties. Lists are not limited to biological entities such as genes or proteins - lists of, for example, related phenotypes or regulatory regions can also be made. Creating a user account with the relevant InterMine database allows lists and searches to be stored between sessions.

In addition to tables of enriched functional annotation, InterMine provides further information on the properties of lists by means of graphical displays and charts that summarise, e.g. gene expression and chromosomal distribution. For instance, in the FlyMine database, bar graphs show expression over development and in adult fly tissues. Like the enrichment tables, such graphs are interactive, allowing the creation of sub-lists of genes from a particular category. Thus, list analysis pages are analogous to the report pages for single entities and collect together information and tools relevant to a list of entities, including as illustrated above, enrichment statistics, graphical displays as well as tables displaying the results of template searches using the list.

Conclusions

InterMine, through the InterMOD project, has strengthened facilities for translational research by providing a platform for interoperation and comparative interpretation of model organism data. Together with HumanMine, the MOD-InterMine databases form a platform to support translational biomedical research by facilitating analysis of the underlying processes that produce complex disease.

Above, we described the data and use of the web interface in InterMine to carry out analysis of integrated biological data and how consistent cross-organism analysis is possible. InterMine aids serendipitous biological discovery through the keyword search, the report and list analysis pages and has been discussed together with its use as a data analysis platform. InterMine allows researchers to exploit the wealth and diversity of data for analysis and examine the associations between data types in order to help make assertions about biological processes. InterMine is particularly powerful for the analysis of lists and an example use-case showed how a list can be gradually filtered through several analysis tools and across several organisms.

Cross-organism analysis has many challenges and some of these have been discussed here. InterMine will continue to establish links between the MOD-InterMine databases and standardised comparative analysis tools are being developed. In addition, the range of cross-species information displayed within one MOD-InterMine database can be expanded. Such collaboration between the major model organisms will help to provide reliable resources with common tools, consistent data and standardised analysis procedures.

Acknowledgments

Grant Support: This work was supported by the Wellcome Trust [099133] and National Human Genome Research Institute [R01HG004834]

The authors would like to thank our users and collaborators for their continuing interest and support. The content is solely the responsibility of the authors and does not necessarily represent the official views of any of the funding bodies.

This work was supported by The Wellcome Trust [099133] and the National Human Genome Research Institute [R01HG004834].

References

Alam I, Antunes A, Kamau AA, Ba Alawi W, Kalkatawi M, Stingl U, Bajic VB. INDIGO - INtegrated data warehouse of microbial genomes with examples from the red sea extremophiles. PLoS One. 2013 Dec 6.8(12):e82210. [PubMed: 24324765]

Aleksic J, Ferrero E, Fischer B, Shen SP, Russell S. The role of Dichaete in transcriptional regulation during Drosophila embryonic development. BMC Genomics. 2013 Dec 8.14:861. [PubMed: 24314314]

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000 May; 25(1):25–9. [PubMed: 10802651]

Balakrishnan R, Park J, Karra K, Hitz BC, Binkley G, Hong EL, Sullivan J, Micklem G, Cherry JM. YeastMine-an integrated data warehouse for Saccharomyces cerevisiae data as a multipurpose tool-kit. Database (Oxford). 2012:bar062. [PubMed: 22434830]

Caspi R, Altman T, Billington R, Dreher K, Foerster H, Fulcher CA, Holland TA, Keseler IM, Kothari A, Kubo A, Krummenacker M, Latendresse M, Mueller LA, Ong Q, Paley S, Subhraveti P, Weaver DS, Weerasinghe D, Zhang P, Karp PD. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. Nucleic Acids Res. 2014 Jan; 42(Database issue):D459–71. [PubMed: 24225315]

Chen YA, Tripathi LP, Mizuguchi K. TargetMine, an integrated data warehouse for candidate gene prioritisation and target discovery. PLoS One. 2011; 6(3):e17844. [PubMed: 21408081]

Cherry JM, Hong EL, Amundsen C, Balakrishnan R, Binkley G, Chan ET, Christie KR, Costanzo MC, Dwight SS, Engel SR, Fisk DG, Hirschman JE, Hitz BC, Karra K, Krieger CJ, Miyasato SR, Nash RS, Park J, Skrzypek MS, Simison M, Weng S, Wong ED. Saccharomyces Genome Database: the genomics resource of budding yeast. Nucleic Acids Res. 2012 Jan; 40(Database issue):D700–5. [PubMed: 22110037]

Contrino S, Smith RN, Butano D, Carr A, Hu F, Lyne R, Rutherford K, Kalderimis A, Sullivan J, Carbon S, Kephart ET, Lloyd P, Stinson EO, Washington NL, Perry MD, Ruzanov P, Zha Z, Lewis SE, Stein LD, Micklem G. modMine: flexible access to modENCODE data. Nucleic Acids Res. 2012 Jan; 40(Database issue):D1082–8. [PubMed: 22080565]

Croft D, O'Kelly G, Wu G, Haw R, Gillespie M, Matthews L, Caudy M, Garapati P, Gopinath G, Jassal B, Jupe S, Kalatskaya I, Mahajan S, May B, Ndegwa N, Schmidt E, Shamovsky V, Yung C, Birney E, Hermjakob H, D'Eustachio P, Stein L. Reactome: a database of reactions, pathways and biological processes. Nucleic Acids Res. 2011 Jan; 39(Database issue):D691–7. [PubMed: 21067998]

dos Santos G, Schroeder AJ, Goodman JL, Strelets VB, Crosby MA, Thurmond J, Emmert DB, Gelbart WM. FlyBase Consortium. FlyBase: introduction of the Drosophila melanogaster Release 6 reference genome assembly and large-scale migration of genome annotations. Nucleic Acids Res. 2015 Jan; 43(Database issue):D690–7. [PubMed: 25398896]

Eilbeck K, Lewis SE, Mungall CJ, Yandell M, Stein L, Durbin R, Ashburner M. The Sequence Ontology: a tool for the unification of genome annotations. Genome Biol. 2005; 6(5):R44. [PubMed: 15892872]

Eppig JT, Blake JA, Bult CJ, Kadin JA, Richardson JE. The Mouse Genome Database Group. The Mouse Genome Database (MGD): facilitating mouse as a model for human biology and disease. Nucleic Acids Res. 2015 Jan 28.43:D726–36. (Database issue). [PubMed: 25348401]

Epping JT, Richardson JE, Kadin JA, Blake JA, Bult CJ. the MGD Team . Mouse Genome Database: from sequence to phenotypes and disease models. Genesis. 2015 this issue.

Fang L, Liu C, Miller YI. Zebrafish models of dyslipidemia: relevance to atherosclerosis and angiogenesis. Transl Res. 2014 Feb; 163(2):99–108. [PubMed: 24095954]

Favrin G, Bean DM, Bilsland E, Boyer H, Fischer BE, Russell S, Crowther DC, Baylis HA, Oliver SG, Giannakou ME. Identification of novel modifiers of Aβ toxicity by transcriptomic analysis in the fruitfly. Sci Rep. 2013 Dec 16.3:3512. [PubMed: 24336499]

Howe DG, Bradford YM, Conlin T, Eagle AE, Fashena D, Frazer K, Knight J, et al. ZFIN, the Zebrafish Model Organism Database: increased support for mutants and transgenics. Nucleic Acids Res. 2013; 41(D1):D854–60. [PubMed: 23074187]

Goecks J, Nekrutenko A, Taylor J. Galaxy Team . Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol. 2010; 11(8):R86. [PubMed: 20738864]

Harb OS, Roos DS. The Eukaryotic Pathogen Databases: a functional genomic resource integrating data from human and veterinary parasites. Methods Mol Biol. 2015; 1201:1–18. [PubMed: 25388105]

Huang da W, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 2009 Jan; 37(1):1–13. [PubMed: 19033363]

Kalderimis A, Lyne R, Butano D, Contrino S, Lyne M, Heimbach J, Hu F, Smith R, Stěpán R, Sullivan J, Micklem G. InterMine: extensive web services for modern biology. Nucleic Acids Res. 2014 Jul; 42(Web Server issue):W468–72. [PubMed: 24753429]

Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M. KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res. 2012 Jan; 40(Database issue):D109–14. [PubMed: 22080510]

Köhler S, Doelken SC, Ruef BJ, Bauer S, Washington N, Westerfield M, Gkoutos G, Schofield P, Smedley D, Lewis SE, Robinson PN, Mungall CJ. Construction and accessibility of a cross-species phenotype ontology along with gene annotations for biomedical research. Version 2. F1000Res. 2013 Feb 1.2:30. revised 2014 Jan 21. [PubMed: 24358873]

Köhler S, Doelken SC, Mungall CJ, Bauer S, Firth HV, Bailleul-Forestier I, Black GC, Brown DL, Brudno M, Campbell J, FitzPatrick DR, Eppig JT, Jackson AP, Freson K, Girdea M, Helbig I, Hurst JA, Jähn J, Jackson LG, Kelly AM, Ledbetter DH, Mansour S, Martin CL, Moss C, Mumford A, Ouwehand WH, Park SM, Riggs ER, Scott RH, Sisodiya S, Van Vooren S, Wapner RJ, Wilkie AO, Wright CF, Vulto-van Silfhout AT, de Leeuw N, de Vries BB, Washingthon NL, Smith CL, Westerfield M, Schofield P, Ruef BJ, Gkoutos GV, Haendel M, Smedley D, Lewis SE, Robinson PN. The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. Nucleic Acids Res. 2014 Jan; 42(Database issue):D966–74. [PubMed: 24217912]

Lowe N, Rees JS, Roote J, Ryder E, Armean IM, Johnson G, Drummond E, Spriggs H, Drummond J, Magbanua JP, Naylor H, Sanson B, Bastock R, Huelsmann S, Trovisco V, Landgraf M, Knowles-Barley S, Armstrong JD, White-Cooper H, Hansen C, Phillips RG, Lilley KS, Russell S, St Johnston D. Analysis of the expression patterns, subcellular localisations and interaction partners of Drosophila proteins using a pigP protein trap library. Development. 2014 Oct; 141(20):3994–4005. [PubMed: 25294943]

Lyne R, Smith R, Rutherford K, Wakeling M, Varley A, Guillier F, Janssens H, Ji W, Mclaren P, North P, Rana D, Riley T, Sullivan J, Watkins X, Woodbridge M, Lilley K, Russell S, Ashburner M, Mizuguchi K, Micklem G. FlyMine: an integrated database for Drosophila and Anopheles genomics. Genome Biol. 2007; 8(7):R129. [PubMed: 17615057]

Lyne M, Smith RN, Lyne R, Aleksic J, Hu F, Kalderimis A, Stepan R, Micklem G. metabolicMine: an integrated genomics, genetics and proteomics data warehouse for common metabolic disease research. Database (Oxford). 2013 Aug.9:bat060. [PubMed: 23935057]

Mungall CJ, Torniai C, Gkoutos GV, Lewis SE, Haendel MA. UK Drosophila Protein Trap Screening Consortium. Uberon, an integrative multi-species anatomy ontology. Genome Biol. 2012 Jan 31.13(1):R5. [PubMed: 22293552]

Osumi-Sutherland D, Marygold SJ, Millburn GH, McQuilton PA, Ponting L, Stefancsik R, Falls K, Brown NH, Gkoutos GV. The Drosophila phenotype ontology. J Biomed Semantics. 2013 Oct 18.4(1):30. [PubMed: 24138933]

Pfreundt U, James DP, Tweedie S, Wilson D, Teichmann SA, Adryan B. FlyTF: improved annotation and enhanced functionality of the Drosophila transcription factor database. Nucleic Acids Res. 2010 Jan; 38(Database issue):D443–7. [PubMed: 19884132]

Ruzicka L, Bradford YM, Frazer K, Howe DG, Paddock H, Ramachandran S, Singer A, Toro S, Van Slyke CE, Eagle AE, Fashena D, Kalita P, Knight J, Mani P, Martin R, Moxon SAT, Pich C, Schaper K, Shao X, Westerfield M. ZFIN, the Zebrafish Model Organism Database: updates and new directions. Genesis. 2015 This Issue.

Seridi L, Ryu T, Ravasi T. Dynamic epigenetic control of highly conserved noncoding elements. PLoS One. 2014 Oct 7.9(10):e109326. [PubMed: 25289637]

Shimoyama M, De Pons J, Hayman GT, Laulederkind SJ, Liu W, Nigam R, Petri V, Smith JR, Tutaj M, Wang SJ, Worthey E, Dwinell M, Jacob H. The Rat Genome Database 2015: genomic, phenotypic and environmental variations and disease. Nucleic Acids Res. 2015 Jan; 43(Database issue):D743–50. [PubMed: 25355511]

Smedley D, Haider S, Durinck S, Pandini L, Provero P, Allen J, Arnaiz O, Awedh MH, Baldock R, Barbiera G, Bardou P, Beck T, Blake A, Bonierbale M, Brookes AJ, Bucci G, Buetti I, Burge S, Cabau C, Carlson JW, Chelala C, Chrysostomou C, Cittaro D, Collin O, Cordova R, Cutts RJ, Dassi E, Genova AD, Djari A, Esposito A, Estrella H, Eyras E, Fernandez-Banet J, Forbes S, Free RC, Fujisawa T, Gadaleta E, Garcia-Manteiga JM, Goodstein D, Gray K, Guerra-Assunção JA, Haggarty B, Han DJ, Han BW, Harris T, Harshbarger J, Hastings RK, Hayes RD, Hoede C, Hu S, Hu ZL, Hutchins L, Kan Z, Kawaji H, Keliet A, Kerhornou A, Kim S, Kinsella R, Klopp C, Kong L, Lawson D, Lazarevic D, Lee JH, Letellier T, Li CY, Lio P, Liu CJ, Luo J, Maass A, Mariette J, Maurel T, Merella S, Mohamed AM, Moreews F, Nabihoudine I, Ndegwa N, Noirot C, Perez-Llamas C, Primig M, Quattrone A, Quesneville H, Rambaldi D, Reecy J, Riba M, Rosanoff S, Saddiq AA, Salas E, Sallou O, Shepherd R, Simon R, Sperling L, Spooner W, Staines DM, Steinbach D, Stone K, Stupka E, Teague JW, Dayem Ullah AZ, Wang J, Ware D, Wong-Erasmus M, Youens-Clark K, Zadissa A, Zhang SJ, Kasprzyk A. The BioMart community portal: an innovative alternative to large, centralized data repositories. Nucleic Acids Res. 2015 Apr 20. pii: gkv350.

Smith AC, Blackshaw JA, Robinson AJ. MitoMiner: a data warehouse for mitochondrial proteomics data. Nucleic Acids Res. 2012 Jan; 40(Database issue):D1160–7. [PubMed: 22121219]

Smith CL, Eppig JT. The Mammalian Phenotype Ontology as a unifying standard for experimental and high-throughput phenotyping data. Mamm Genome. 2012 Oct; 23(9–10):653–68. [PubMed: 22961259]

Smith RN, Aleksic J, Butano D, Carr A, Contrino S, Hu F, Lyne M, Lyne R, Kalderimis A, Rutherford K, Stepan R, Sullivan J, Wakeling M, Watkins X, Micklem G. InterMine: a flexible data warehouse system for the integration and analysis of heterogeneous biological data. Bioinformatics. 2012 Dec 1; 28(23):3163–5. [PubMed: 23023984]

Sonnhammer EL, Gabaldón T, Sousa da Silva AW, Martin M, Robinson-Rechavi M, Boeckmann B, Thomas PD, Dessimoz C. Quest for Orthologs consortium. Big data and other challenges in the quest for orthologs. Bioinformatics. 2014 Nov 1; 30(21):2993–8. [PubMed: 25064571]

Sprague J, Bayraktaroglu L, Bradford Y, Conlin T, Dunn N, Fashena D, Frazer K, Haendel M, Howe DG, Knight J, Mani P, Moxon SA, Pich C, Ramachandran S, Schaper K, Segerdell E, Shao X, Singer A, Song P, Sprunger B, Van Slyke CE, Westerfield M. The Zebrafish Information Network: the zebrafish model organism database provides expanded support for genotypes and phenotypes. Nucleic Acids Res. 2008 Jan.36(Database)

Sullivan J, Karra K, Moxon SA, Vallejos A, Motenko H, Wong JD, Aleksic J, Balakrishnan R, Binkley G, Harris T, Hitz B, Jayaraman P, Lyne R, Neuhauser S, Pich C, Smith RN, Trinh Q, Cherry JM, Richardson J, Stein L, Twigger S, Westerfield M, Worthey E, Micklem G. InterMOD: integrated data and tools for the unification of model organism research. Sci Rep. 2013; 3:1802. [PubMed: 23652793]

Yook K, Harris TW, Bieri T, Cabunoc A, Chan J, Chen WJ, Davis P, de la Cruz N, Duong A, Fang R, Ganesan U, Grove C, Howe K, Kadam S, Kishore R, Lee R, Li Y, Muller HM, Nakamura C, Nash B, Ozersky P, Paulini M, Raciti D, Rangarajan A, Schindelman G, Shi X, Schwarz EM, Ann Tuli M, Van Auken K, Wang D, Wang X, Williams G, Hodgkin J, Berriman M, Durbin R, Kersey P, Spieth J, Stein L, Sternberg PW. WormBase 2012: more genomes, more data, new website. Nucleic Acids Res. 2012 Jan; 40(Database issue):D735–41. [PubMed: 22067452]

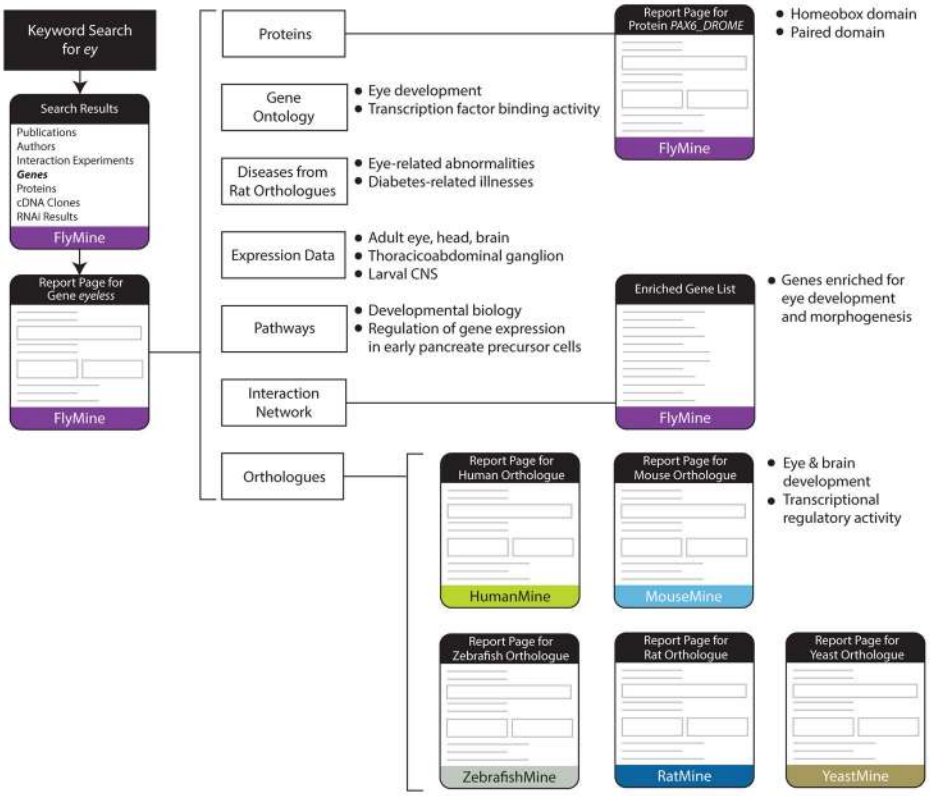
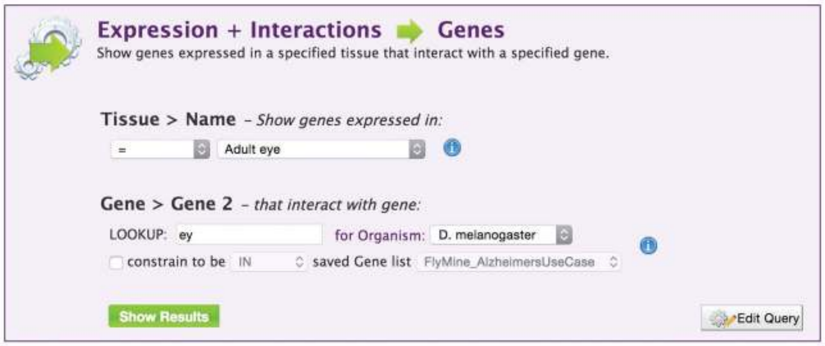

Keyword Search for *ey*
Search Results
Publications
Authors
Interaction Experiments
Genes
Proteins
cDNA Clones
RNAi Results
FlyMine
Report Page for Gene *eyeless*
FlyMine
Proteins
Gene Ontology
• Eye development
• Transcription factor binding activity
Diseases from Rat Orthologues
• Eye-related abnormalities
• Diabetes-related illnesses
Expression Data
• Adult eye, head, brain
• Thoracicoabdominal ganglion
• Larval CNS
Pathways
• Developmental biology
• Regulation of gene expression in early pancreate precursor cells
Interaction Network
Orthologues
Report Page for Protein *PAX6_DROME*
FlyMine
• Homeobox domain
• Paired domain
Enriched Gene List
FlyMine
• Genes enriched for eye development and morphogenesis
Report Page for Human Orthologue
HumanMine
Report Page for Mouse Orthologue
MouseMine
• Eye & brain development
• Transcriptional regulatory activity
Report Page for Zebrafish Orthologue
ZebrafishMine
Report Page for Rat Orthologue
RatMine
Report Page for Yeast Orthologue
YeastMine

Figure 1. Data exploration through the InterMine web interface

Data exploration through the InterMine web interface, illustrating navigation between data types within a report page, between report pages and between report pages for orthologous genes in different organisms. The workflow begins with a keyword search for *ey* in FlyMine followed by navigation to the *D. melanogaster ey* gene report page, where several data types are examined. Navigation to the corresponding protein report page, PAX6_DROME, allows protein domain data to be viewed. Links to report pages for orthologous genes allow data for equivalent genes in human, mouse, rat, zebrafish and yeast to be examined.



Expression + Interactions ➡ Genes

Show genes expressed in a specified tissue that interact with a specified gene.

Tissue > Name – *Show genes expressed in:*

= Adult eye

Gene > Gene 2 – *that interact with gene:*

LOOKUP: ey for Organism: D. melanogaster

□ constrain to be IN saved Gene list FlyMine_AlzheimersUseCase

Figure 2. A template search from FlyMine

A template search, Expression + Interactions → Genes, from the FlyMine database, showing two constraints (filters), one for a tissue (in this case adult eye) and one for the interacting gene (in this case *ey*). This template will return any genes expressed in the adult eye that also interact (physically or genetically) with *ey*.

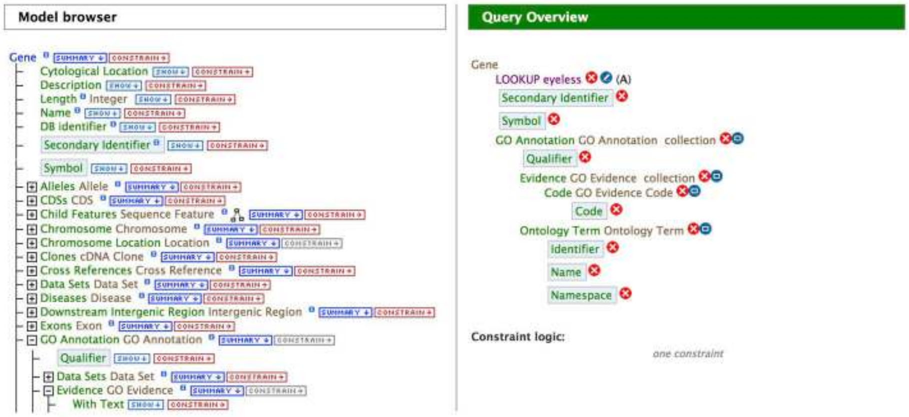
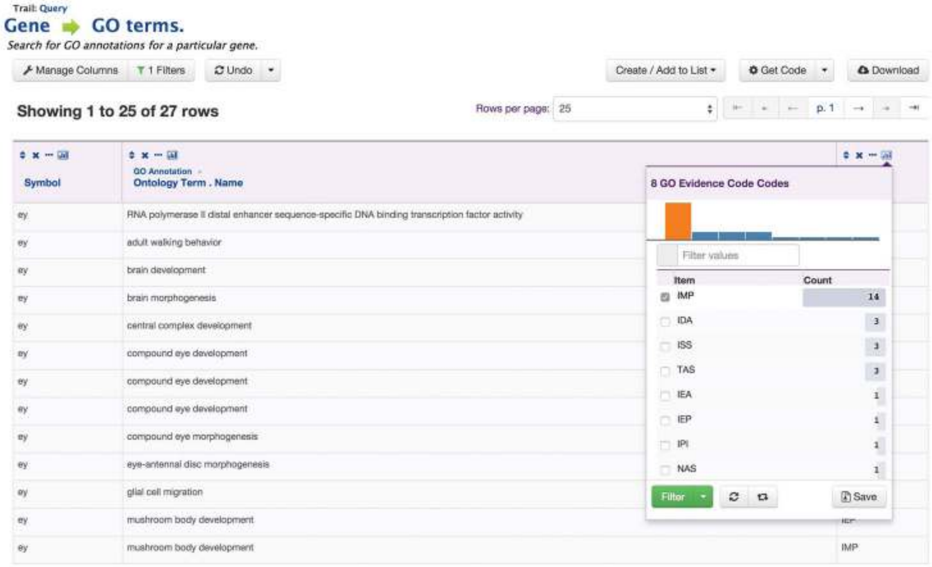

Model browser
Gene SUMMARY CONSTRAIN
Cytological Location SHOW CONSTRAIN
Description SHOW CONSTRAIN
Length Integer SHOW CONSTRAIN
Name SHOW CONSTRAIN
DB identifier SHOW CONSTRAIN
Secondary Identifier SHOW CONSTRAIN
Symbol SHOW CONSTRAIN
Alleles Allele SUMMARY CONSTRAIN
CDSs CDS SUMMARY CONSTRAIN
Child Features Sequence Feature SUMMARY CONSTRAIN
Chromosome Chromosome SUMMARY CONSTRAIN
Chromosome Location Location SUMMARY CONSTRAIN
Clones cDNA Clone SUMMARY CONSTRAIN
Cross References Cross Reference SUMMARY CONSTRAIN
Data Sets Data Set SUMMARY CONSTRAIN
Diseases Disease SUMMARY CONSTRAIN
Downstream Intergenic Region Intergenic Region SUMMARY CONSTRAIN
Exons Exon SUMMARY CONSTRAIN
GO Annotation GO Annotation SUMMARY CONSTRAIN
Qualifier SHOW CONSTRAIN
Data Sets Data Set SUMMARY CONSTRAIN
Evidence GO Evidence SUMMARY CONSTRAIN
With Text SHOW CONSTRAIN
Query Overview
Gene
LOOKUP eyeless (A)
Secondary Identifier
Symbol
GO Annotation GO Annotation collection
Qualifier
Evidence GO Evidence collection
Code GO Evidence Code
Code
Ontology Term Ontology Term
Identifier
Name
Namespace
Constraint logic:
one constraint

Figure 3. The InterMine query builder

The query builder allows navigation of the data model (left pane), where “Constrain” buttons allow the configuration of constraints (filters) on the attribute or class and “Show” buttons add an attribute to the results output. The right pane shows a summary of the query as it is built. A query that will return all Gene Ontology annotations for the *D. melanogaster ey* gene, together with the associated evidence code, is shown.



Search for GO annotations for a particular gene.

Showing 1 to 25 of 27 rows

Rows per page: 25

Symbol	GO Annotation . Ontology Term . Name	
ey	RNA polymerase II distal enhancer sequence-specific DNA binding transcription factor activity	
ey	adult walking behavior	
ey	brain development	
ey	brain morphogenesis	
ey	central complex development	
ey	compound eye development	
ey	compound eye development	
ey	compound eye development	
ey	compound eye morphogenesis	
ey	eye-antennal disc morphogenesis	
ey	glial cell migration	
ey	mushroom body development	IEP
ey	mushroom body development	IMP

8 GO Evidence Code Codes

Filter values

	Item	Count
☑	IMP	14
□	IDA	3
□	ISS	3
□	TAS	3
□	IEA	1
□	IEP	1
□	IPI	1
□	NAS	1

Figure 4. An InterMine results table

An InterMine results table, generated by running the template search "Gene -> GO terms" with the *ey* gene in the FlyMine database. A "column summary" for the Gene Ontology evidence code column is shown, allowing filtering of the results table to show only terms annotated through specific evidence codes. Note that some columns have been removed from the original results for illustration purposes.

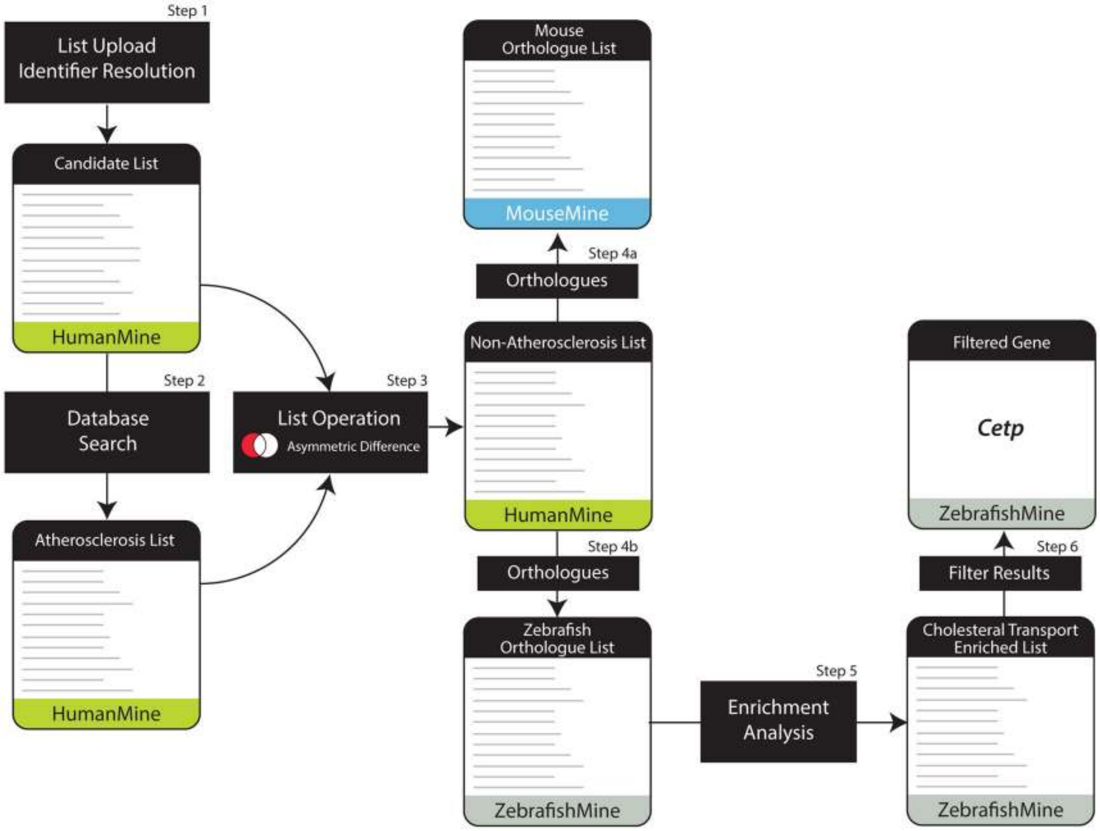
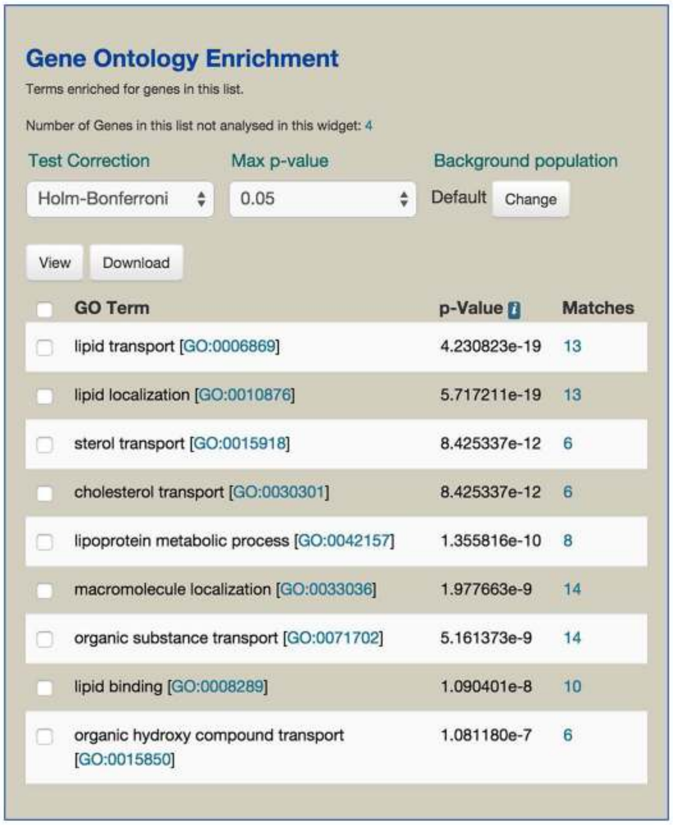


Figure 5. Data Exploitation through the InterMine web interface

A hypothetical workflow in which a candidate gene list is filtered through several consecutive analysis tools. Step1: a candidate gene list, identified through a screen for lipid and cholesterol markers as part of a study on atherosclerosis, is uploaded to the HumanMine database. Step 2: A search of the database identifies those genes from the candidate list that are already associated with the disease atherosclerosis. A list is made of these genes. Step3: Using the list operation, asymmetric distribution, a new list is created which does not contain the genes identified as already being associated with atherosclerosis. This list is called the non-atherosclerosis set. Step 4: Links to MouseMine and ZebrafishMine directly from HumanMine allow lists of mouse (Step 4a) and zebrafish (Step 4b) genes orthologous to the non-atherosclerosis list to be analysed in the respective databases. Enrichment statistics for various annotations can be viewed, and in particular an enrichment for the Gene Ontology term “Cholesterol transport” is noted. Step 5: The zebrafish genes from the list annotated with the Gene Ontology term “Cholesterol transport” are saved as a list. Step 6: A database search and filtering for homologues of these genes reveals a gene, *Cetp*, present in Human and Zebrafish but not in mouse.



Gene Ontology Enrichment

Terms enriched for genes in this list.

Number of Genes in this list not analysed in this widget: 4

Test Correction: Holm-Bonferroni

Max p-value: 0.05

Background population: Default

□	GO Term	p-Value	Matches
□	lipid transport [GO:0006869]	4.230823e-19	13
□	lipid localization [GO:0010876]	5.717211e-19	13
□	sterol transport [GO:0015918]	8.425337e-12	6
□	cholesterol transport [GO:0030301]	8.425337e-12	6
□	lipoprotein metabolic process [GO:0042157]	1.355816e-10	8
□	macromolecule localization [GO:0033036]	1.977663e-9	14
□	organic substance transport [GO:0071702]	5.161373e-9	14
□	lipid binding [GO:0008289]	1.090401e-8	10
□	organic hydroxy compound transport [GO:0015850]	1.081180e-7	6

Figure 6.

Gene Ontology enrichment analysis of a list of genes in ZebrafishMine. A Gene Ontology enrichment table showing terms from the Gene Ontology biological process ontology enriched in a set of zebrafish genes. A hypergeometric distribution is used to calculate the p-value, which is shown in the table, after a Holm-Bonferonni test correction has been applied. The number of genes with each annotation are shown. Lists of genes with each Gene Ontology annotation can be created directly from the table.

Table 1

Many common data types are available in the MOD-InterMine databases. This table indicates those that are available in each database (dark green), those data types that are planned for a future release (light green) and those that are currently either not available or planned for inclusion for that particular organism (white). Many of the data types are derived from different experimental procedures. For example gene expression is provided by microarray data, RNA_seq and in-situ hybridisations while phenotype data includes RNAi phenotypes and interaction data both high-throughput data such as yeast two-hybrid and smaller scale experiments. The experiment used in each case can be accessed through database searches and a detailed breakdown of data types is provided on the data page of each MOD-InterMine database.

	MouseMine	**WormMine**	**FlyMine**	**RatMine**	**YeastMine**	**ZebrafishMine**	**HumanMine**
GENOME ANNOTATION							
sequence	🟨	🟩	🟩	🟨	🟩	🟨	🟩
genes	🟩	🟩	🟩	🟩	🟩	🟩	🟩
transcripts	🟨	🟩	🟩	🟨	🟩	🟨	🟩
exons	🟨	🟩	🟩	🟨	🟩	🟨	🟩
regulatory data		🟩	🟩	🟨	🟩	🟨	🟨
PROTEINS							
sequence	🟨	🟩	🟩	🟩	🟩	🟨	🟩
feature locations	🟨	🟩	🟩	🟩	🟩	🟨	🟩
domain content	🟨	🟩	🟩	🟩	🟩	🟨	🟩
protein-protein interactions	🟨	🟨	🟩	🟨	🟩	🟨	🟩
protein expression							🟩
COMPARATIVE							
curated orthologs	🟨	🟩	🟩	🟨		🟩	
curated paralogs	🟨	🟩	🟩	🟨		🟩	
computed orthologs	🟩		🟩	🟨	🟩	🟨	🟩
computed paralogs	🟩		🟩	🟨	🟩	🟨	🟩
GENETICS							
alleles	🟩	🟩	🟩	🟨		🟩	🟩
SNPs		🟩		🟩		🟩	🟩

	MouseMine	WormMine	FlyMine	RatMine	YeastMine	ZebrafishMine	HumanMine
genotypes	🟩	🟩	🟩	▨		🟩	🟩
linkage		🟩		▨			
QTLs	🟩			🟩			
genetic interactions	🟩	▨	🟩	▨	🟩	▨	🟩
gene expression	🟩	🟩	🟩	▨	🟩	🟩	🟩
STOCKS							
mutant lines	🟩	🟩	🟩	🟩		🟩	
transgenic lines	🟩	🟩	🟩	🟩		🟩	
strains	🟩	🟩	🟩	🟩		🟩	
ANNOTATIONS							
gene ontology	🟩	🟩	🟩	🟩	🟩	🟩	🟩
disease	🟩	▨	🟩	🟩	🟩	🟩	🟩
phenotypes	🟩	🟩	🟩	🟩	🟩	🟩	🟩
anatomy	🟩	🟩	🟩	🟩		🟩	
pathways	▨	🟩	🟩	🟩	🟩	🟩	🟩
publications	🟩	🟩	🟩	🟩	🟩	🟩	🟩