# FlyMine

www.flymine.org

## An integrated database for *Drosophila* and *Anopheles* genomics

## Introduction

FlyMine is an open-source project to build an integrated database of genomic, expression and proteomics data for *Drosophila, Anopheles* and other insects.

With the completion of increasing numbers of genomic sequences has come an explosion in the development of both computational and experimental high-throughput techniques for deciphering molecular functions and their interactions. These techniques include theoretical methods for deducing function such as analysis of protein homologies, structural domain predictions, phylogenetic profiling and protein domain fusions and empirical methods such as microarray gene expression studies, two-hybrid protein-protein interaction experiments, protein co-immunoprecipitation and large-scale RNAi screens.

The result is a huge amount of data, and the challenge is to extract meaniful knowledge and patterns of biological significance that can lead to new experimentally testable hypotheses. Many of the resulting high-throughput data sets, however, are noisy and the data quality can vary significanlty. Stronger inferences can be made if one is able to combine data from different sources. Currently, biological data is stored in a wide variety of formats in numerous different places, making complex queries across these data sources very difficult. One of the current challenges for bioinformatics is therefore to develop resources for the integration and combined analysis of such data.

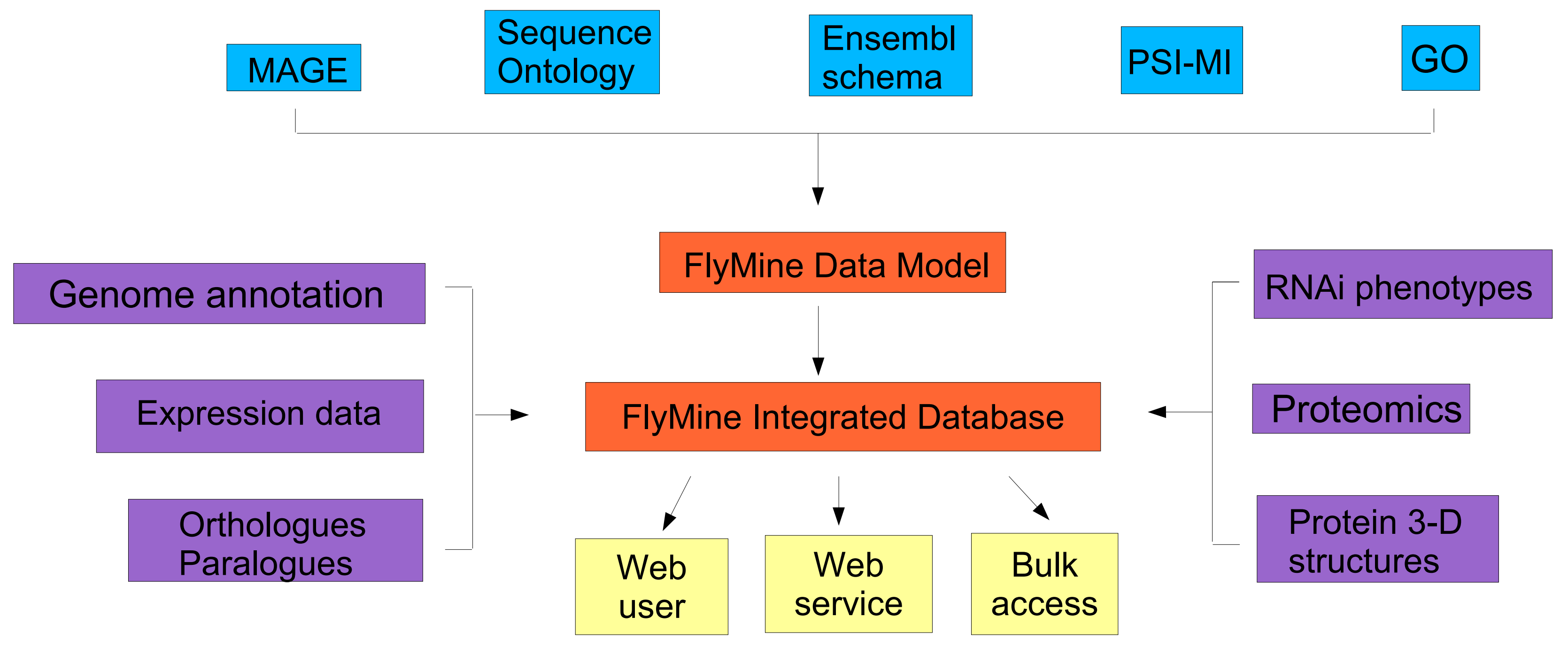
## Technical Description

- FlyMine is built using InterMine (www.intermine.org), an open-source data warehouse system for integrating data from many sources into one object-based database, allowing complex queries to be performed on the integrated data.

- Although FlyMine primarily focuses on *Drosophila* and *Anopheles* data, the InterMine software is designed to be completely generic and is not tied to any particular datasource or application.

- The FlyMine model is generated from existing and emerging ontologies and standards (MAGE, SO, PSI and others). This enables data to be loaded from a variety of sources and provides the flexibility to incorporate new data as they emerge. A number of FlyMine specific classes help to link together those generated from the merged ontologies and standards.

- A powerful and flexible query interface enables users to perform arbitrary and complex queries across the data either via a web interface or a programming interface. Users are not confined to 'fill-in-the-blanks' query templates or predefined queries.



MAGE

Sequence Ontology

Ensembl schema

PSI-MI

GO

FlyMine Data Model

Genome annotation

Expression data

Orthologues Paralogues

FlyMine Integrated Database

RNAi phenotypes

Proteomics

Protein 3-D structures

Web user

Web service

Bulk access

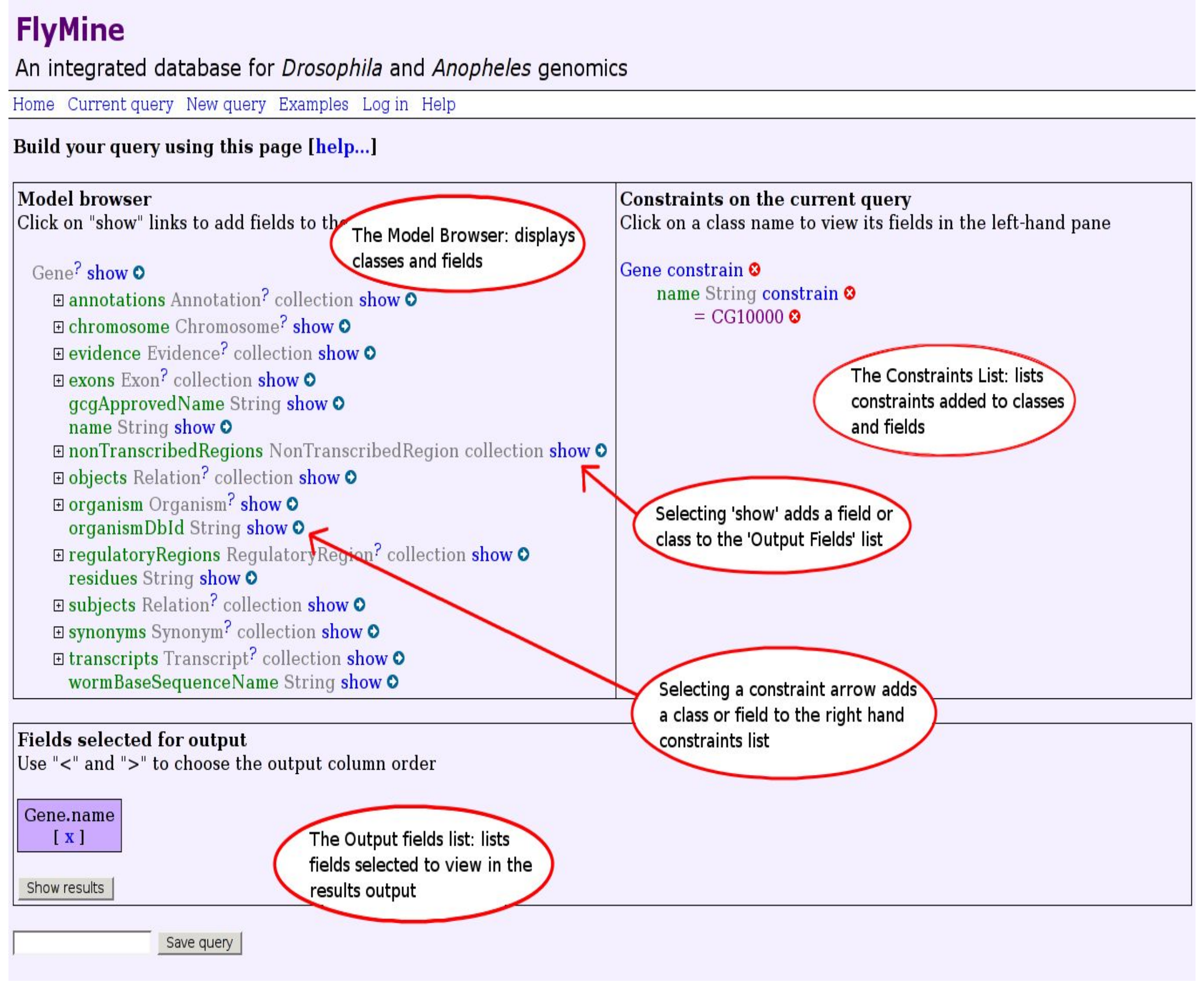
## The FlyMine Query Builder Page

- The query interface consists of three sections: a model browser, a constraints list and an output fields list.

- The model browser allows you to select the classes that contain the objects you require. You can navigate to other classes using the references between them.

- From the model browser classes and fields can be selected to constrain in the constraints list and/or show in the results output.

- Queries can be saved for future reference or futher refinement.



**FlyMine**

An integrated database for *Drosophila* and *Anopheles* genomics

Home Current query New query Examples Log in Help

**Build your query using this page [help...]**

**Model browser**
Click on "show" links to add fields to th

```
Gene? show
  ⊞ annotations Annotation? collection show
  ⊞ chromosome Chromosome? show
  ⊞ evidence Evidence? collection show
  ⊞ exons Exon? collection show
    gcgApprovedName String show
    name String show
  ⊞ nonTranscribedRegions NonTranscribedRegion collection show
  ⊞ objects Relation? collection show
  ⊞ organism Organism? show
    organismDbId String show
  ⊞ regulatoryRegions RegulatoryRegion? collection show
    residues String show
  ⊞ subjects Relation? collection show
  ⊞ synonyms Synonym? collection show
  ⊞ transcripts Transcript? collection show
    wormBaseSequenceName String show
```

**Constraints on the current query**
Click on a class name to view its fields in the left-hand pane

```
Gene constrain
    name String constrain
        = CG10000
```

**Fields selected for output**
Use "<" and ">" to choose the output column order

Gene.name
[ x ]

Show results

Save query

The Model Browser: displays classes and fields

The Constraints List: lists constraints added to classes and fields

Selecting 'show' adds a field or class to the 'Output Fields' list

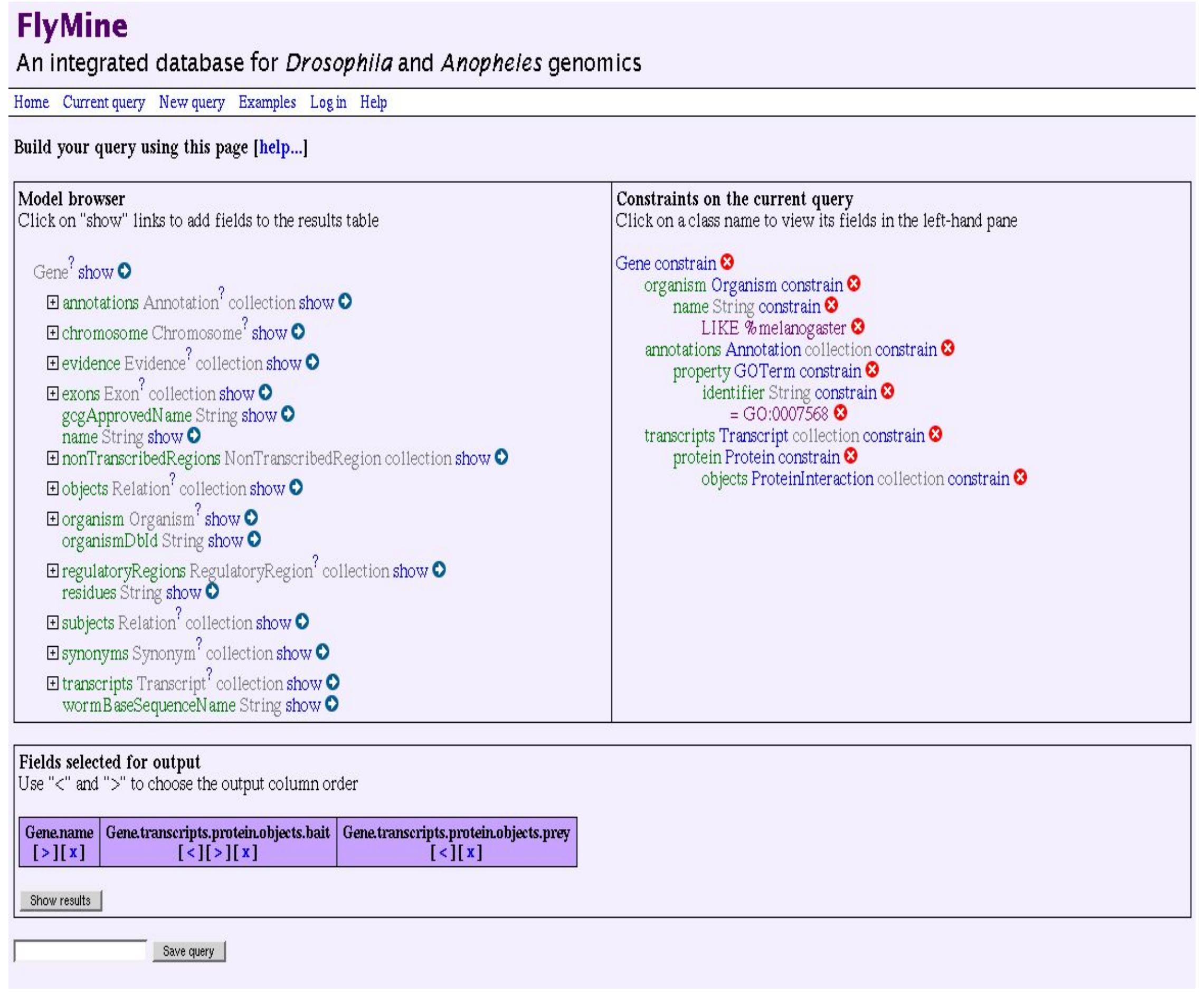
Selecting a constraint arrow adds a class or field to the right hand constraints list

The Output fields list: lists fields selected to view in the results output

## A Query Example: building a query across integrated data sources

The screenshot below illustrates how to construct the query:

**"For *Drosophila* genes that have been annotated with GO term GO:0007568, and whose proteins have protein interaction data, show the gene name and the proteins involved in the interactions".**



**FlyMine**

An integrated database for *Drosophila* and *Anopheles* genomics

Home Current query New query Examples Log in Help

**Build your query using this page [help...]**

**Model browser**
Click on "show" links to add fields to the results table

```
Gene? show
  ⊞ annotations Annotation? collection show
  ⊞ chromosome Chromosome? show
  ⊞ evidence Evidence? collection show
  ⊞ exons Exon? collection show
    gcgApprovedName String show
    name String show
  ⊞ nonTranscribedRegions NonTranscribedRegion collection show
  ⊞ objects Relation? collection show
  ⊞ organism Organism? show
    organismDbId String show
  ⊞ regulatoryRegions RegulatoryRegion? collection show
    residues String show
  ⊞ subjects Relation? collection show
  ⊞ synonyms Synonym? collection show
  ⊞ transcripts Transcript? collection show
    wormBaseSequenceName String show
```

**Constraints on the current query**
Click on a class name to view its fields in the left-hand pane

```
Gene constrain
    organism Organism constrain
        name String constrain
            LIKE %melanogaster
    annotations Annotation collection constrain
        property GOTerm constrain
            identifier String constrain
                = GO:0007568
    transcripts Transcript collection constrain
        protein Protein constrain
            objects ProteinInteraction collection constrain
```

**Fields selected for output**
Use "<" and ">" to choose the output column order

| Gene.name<br>[ > ][ x ] | Gene.transcripts.protein.objects.bait<br>[ < ][ > ][ x ] | Gene.transcripts.protein.objects.prey<br>[ < ][ x ] |
|---|---|---|

Show results

Save query

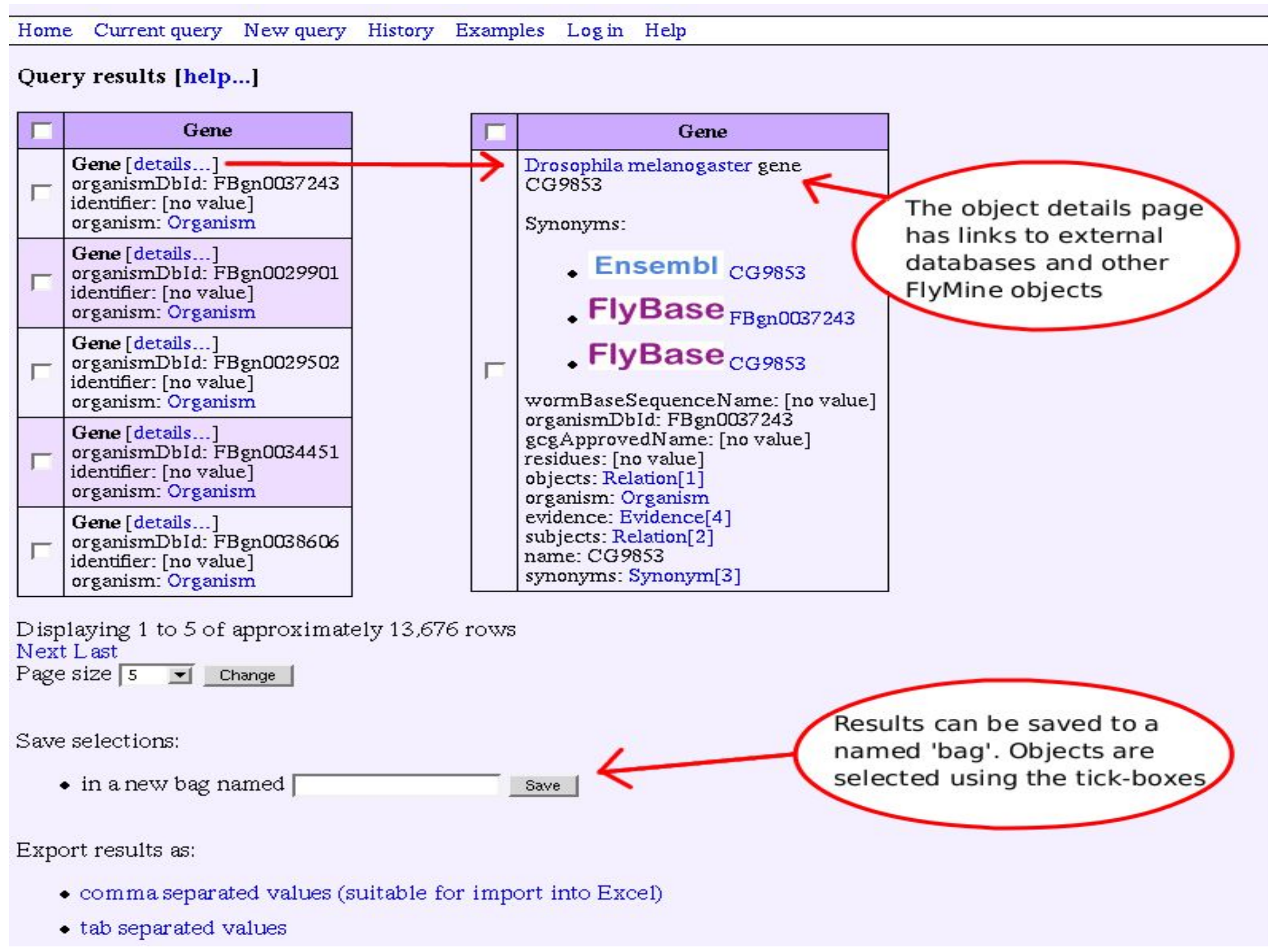
## The Result Display

- The result display includes links to external databases and class and attribute browsing capability. Thus the interface allows for both well-defined hypothesis driven analysis as well as more exploratory analysis.

- Integration of display and analysis tools is in progress, and will allow users to upload their results to a suitable viewing or analysis tool.

- Bags of data can be saved and used to constrain further queries and perform logical operations. A user log-in system enables queries and bags of data to be saved between sessions.

- Data can be downloaded in a variety of formats, for example tab-delimited text for use in spreadsheets.



Home Current query New query History Examples Log in Help

**Query results [help...]**

| □ | Gene |
|---|---|
| □ | **Gene** [details...]<br>organismDbId: FBgn0037243<br>identifier: [no value]<br>organism: Organism |
| □ | **Gene** [details...]<br>organismDbId: FBgn0029901<br>identifier: [no value]<br>organism: Organism |
| □ | **Gene** [details...]<br>organismDbId: FBgn0029502<br>identifier: [no value]<br>organism: Organism |
| □ | **Gene** [details...]<br>organismDbId: FBgn0034451<br>identifier: [no value]<br>organism: Organism |
| □ | **Gene** [details...]<br>organismDbId: FBgn0038606<br>identifier: [no value]<br>organism: Organism |

| □ | Gene |
|---|---|
| □ | Drosophila melanogaster gene<br>CG9853<br>Synonyms:<br>• Ensembl CG9853<br>• FlyBase FBgn0037243<br>• FlyBase CG9853<br>wormBaseSequenceName: [no value]<br>organismDbId: FBgn0037243<br>gcgApprovedName: [no value]<br>residues: [no value]<br>objects: Relation[1]<br>organism: Organism<br>evidence: Evidence[4]<br>subjects: Relation[2]<br>name: CG9853<br>synonyms: Synonym[3] |

Displaying 1 to 5 of approximately 13,676 rows
Next Last
Page size 5 Change

Save selections:

- in a new bag named Save

Export results as:

- comma separated values (suitable for import into Excel)
- tab separated values

The object details page has links to external databases and other FlyMine objects

Results can be saved to a named 'bag'. Objects are selected using the tick-boxes

## Data Sources First Release

| Data | Organism | Source | Publication | Website |
|---|---|---|---|---|
| | | | | |
| Genome Annotation | *D. melanogaster* | Ensembl | | http://www.ensembl.org |
| | *A. gambiae* | Ensembl | | http://www.ensembl.org |
| | | | | |
| 2-hybrid protein interactions | *D. melanogaster* | intAct | Giot et al (2003) Science 302 1727-1736 | http://inparanoid.cgb.ki.se/index.html |
| | *D. melanogaster* | HybriGenics | | http://pim.hybrigenics.com/pimriderext/droso/index.html |
| | *C. elegans* | intAct | Li et al (2004) Science 303 540-543 | http://inparanoid.cgb.ki.se/index.html |
| | | | | |
| Microarray expression data | *D. melanogaster* | ArrayExpress | | http://www.ebi.ac.uk/arrayexpress |
| | | | | |
| RNAi phenotypes | *C. elegans* | WormBase | Kamath et al (2003) Nature 421 231-237 | http://www.wormbase.org |
| | | | Fraser et al (2000) Nature 408 325-330 | |
| | | | Simmer et al (2003) PloS Biol 1 (1) E12 | |
| | | | | |
| Orthologues/ Paralogues | *D. melanogaster:A. gambiae* | Ensembl | | http://www.ensembl.org |
| | *D. melanogaster:C. elegans* | InParanoid | | http://inparanoid.cgb.ki.se/index.html |
| | | | | |

## Further Information and Download

- Further information and documentation can be found on the FlyMine website, www.flymine.org, by joining one of the FlyMine electronic mailing lists (details on the website), or by email to info@flymine.org. The intermine code is available for download under the open-source LGPL licence from www.intermine.org. We welcome co-developers.

- Help is available in the form of a quick start guide, query examples, query tutorials and a user manual. There is also an analysis support help desk. The help pages are under continued development.

- The FlyMine team: Rachel Lyne, Richard Smith, Kim Rutherford, Matthew Wakeling, Mark Woodbridge, Wenyan Ji, François Guillier, and Gos Micklem.

Department of Genetics, University of Cambridge,
Downing Street, Cambridge CB2 3EH, UK
Tel: +44 1223 333377 Email: info@flymine.org

FlyMine is funded by the Wellcome Trust (grant no. 067205), awarded to M. Ashburner, G. Micklem, S. Russell, K. Lilley, and K. Mizuguchi.