InterMine Webapp Workshop

October 2008

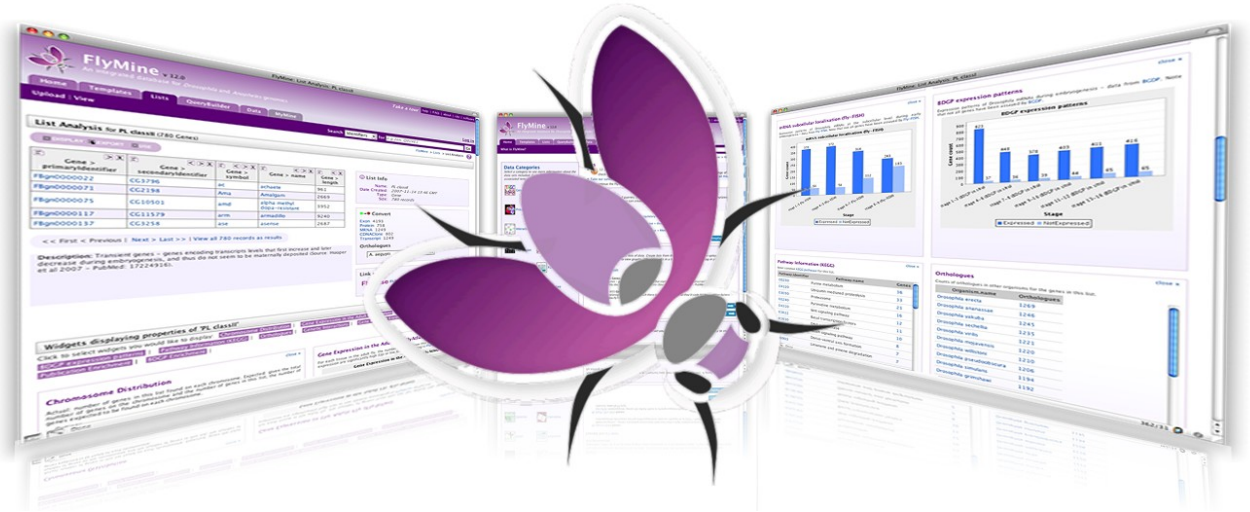


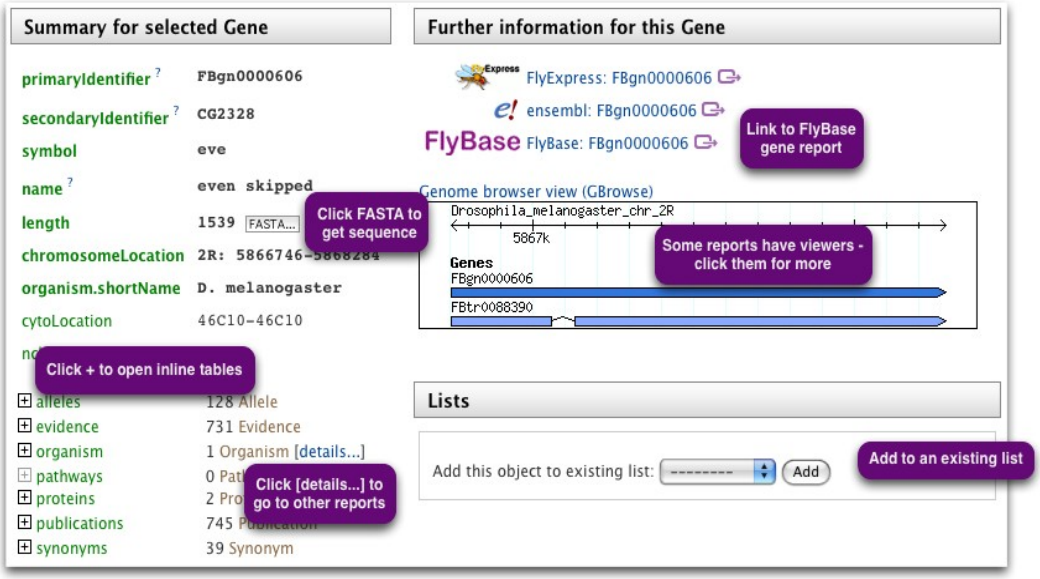
Table of Contents

I. Quick Search and report pages....3
II. Template Queries....5
A. The Templates Page....6
B. Running a template....6
C. Result pages....6
III. Lists and List Analysis pages....8
A. The Lists Page....8
1. - Upload -....8
2. - View -....9
3. List Operations....9
B. List Analysis Page....11
IV. Exporting Data....15
V. Query Builder....17
VI. Super User....22
A. Creating a Template Query....22
B. Tagging....25
1. Template queries and lists....25
2. Fields and collections....25
3. Classes....25
VII. Appendix I: Basic FlyMine - Terminology and concepts....26
Appendix II: Help and Feedback....27

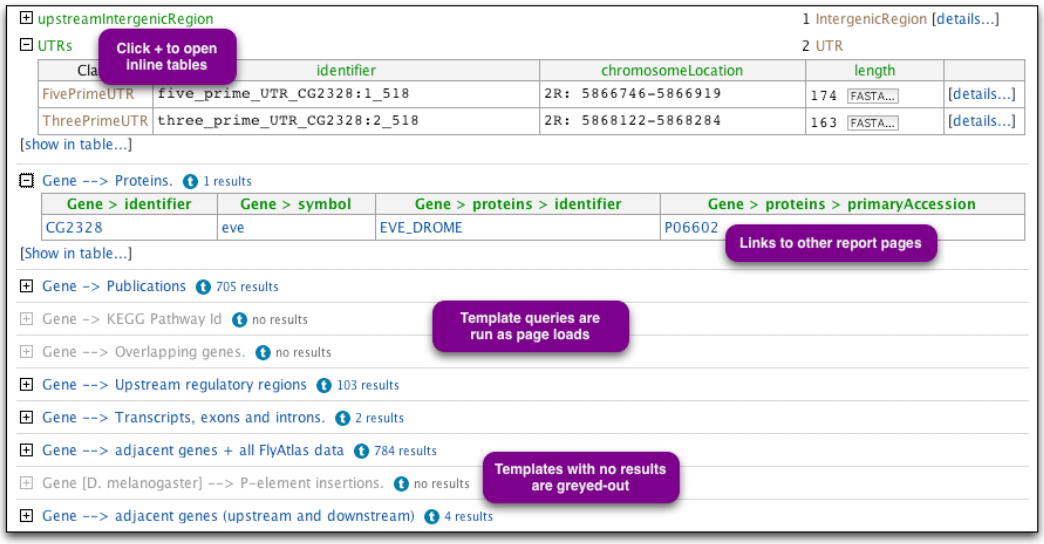
I.Quick Search and report pages

Entering any object identifier (such as a gene name or identifier) in the quick search box will take you straight to the **report page** for that object. For every object in the FlyMine database there is a 'report' page which provides all the information that FlyMine has for that object. All report pages have the same basic structure:

- The top part provides summary information for every object:



- The lower part, 'Further information by category' has more detailed information divided into the data categories. The categories correspond to the categories of data shown on the home page.



In cases where more than one object is returned by your Quick Search, the results will provide

a list of objects which match the query. Clicking on any of them will take you to the corresponding report page for each object. Report pages can also be accessed from result tables by clicking on an object that provides a link.

Some example identifiers: for the *Drosophila* gene zen, type FBgn0004053, CG1046, zen or any of its synonyms.

II.Template Queries

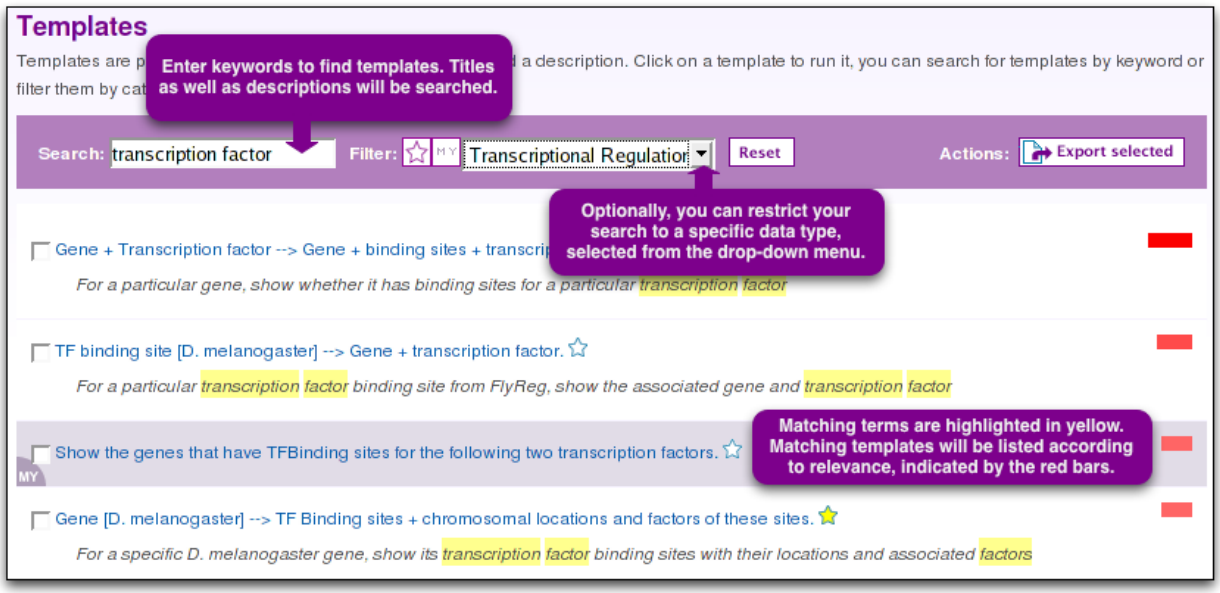
To enable you to quickly and easily carry out a range of queries across the data, FlyMine includes a library of **predefined Template Queries**. A template consists of a query form in which certain variables can be changed. The variables are constraints (filters) on certain types of data, so that a subset of that type is defined.

Template queries can be accessed in four ways:

- Each **data category** page has a set of template queries related to that data type.
- The **templates page,** accessible from the FlyMine home page or via the tabs in the top menu bar available on any FlyMine page, allows you to search or browse for templates.
- Templates can also be searched using the **Quick Search** available on every FlyMine page. Change the drop-down list from 'Identifiers' to 'Templates'.
- Templates also appear on **report pages** and **list analysis pages**.

[It is also possible to make your own template queries using the Query Builder (covered later) and these can be accessed both on the templates page and from your MyMine account]

Each template query has a shorthand title (displayed) which is:
'query starting point(s)' --> 'query output' and a longer description displayed underneath. The 'star' next to each template title enables you to set templates as your 'favourites'. Your favourite templates can be accessed through a filter on the templates page.

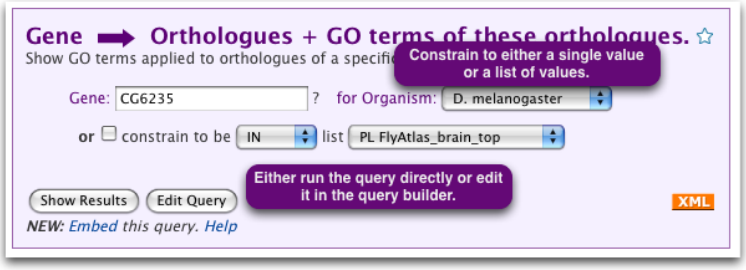


A.The Templates Page

The templates page lists all available templates with their descriptions by default. The descriptions can be hidden by deselecting the check box. The bar above the list of templates allows you to search for a particular template using keywords, show just your favourite templates or just your own templates or filter the templates list according to a data category.

B.Running a template

To access the template form click on the template name. The template form will consist of one or more constraints. You can change the values for each constraint or leave the default values. For most templates it is possible to enter any kind of identifier (e.g. a CGXXXX, FBgn or symbol for *Drosophila* genes). It is also possible to **enter multiple values** (separated by a comma) (e.g. CG10007, CG1046). There is also the option to constrain to a list of items - this will be covered later.

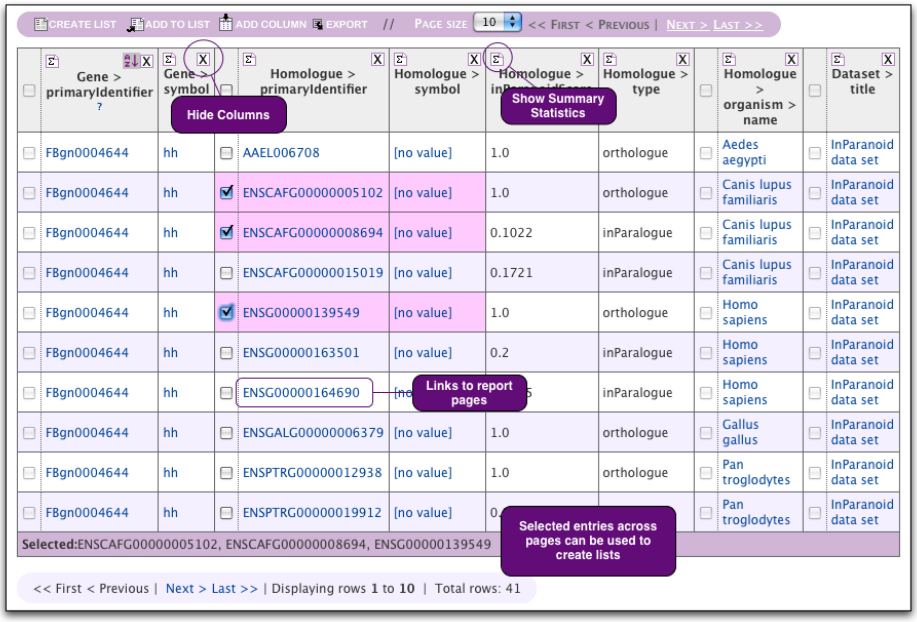


At the bottom of each template form are two options: 'Show Results' and 'Edit Query'.

- The '**Show Results**' button will take you directly to the results page.
- The '**Edit Query**' option allows you to display the template in the FlyMine Query Builder (covered later). This option enables you to use a template query as a starting point for building a modified or more complex query, or to see how a certain query may be constructed using the Query Builder.

C.Result pages

Result pages show the first ten rows of your results by default. The **total number of rows** is displayed underneath the table (this will initially give an approximate number of rows until the total number is calculated). The **∑** in each column header will display **summary statistics** for that column. The type of statistics depends on the type of data in the column - for columns containing text this will display the number of unique values, for numerical data, the min, max, mean and standard deviation will be given (note: at the moment only the first 10 rows of the summary are shown). Above the results table are options for **saving results to lists and for exporting** (covered later).



Exercises:

4.1 Use the templates search on the templates page to find and run the following template query for the default values or for the gene of your choice - and try to understand the results. Look at the summary statistics for some of the columns in your results (accessed in the column header).

Gene --> Transcripts, exons and introns.

Find the answers to the following questions using templates:

4.2 How many D. melanogaster genes that have been annotated with the GO term 'zinc ion binding' have at least one insertion. (Data Source: GO, FlyBase).
[A: GO term name [D. melanogaster] All genes with insertions.]

4.3 How many CRMs (=Cis-Regulatory Modules, i.e. regulatory regions) are know for the D. melanogaster gene fork head?
[A: Gene [D. melanogaster] --> CRMs.]

4.4 How many D. melanogaster genes code for a protein with a basic leucine zipper domain?
[A: Protein Domain --> Proteins + Genes coding for these proteins.]

III.Lists and List Analysis pages

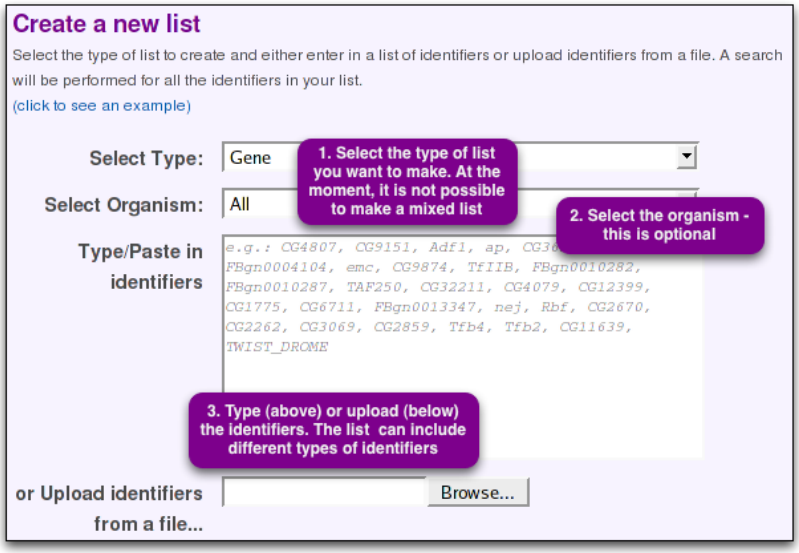
All queries in FlyMine can be performed on lists (e.g. a list of genes). You can create lists yourself (either by uploading or from the results of queries) or you can use a set of '**public**' lists already available in FlyMine. Lists can be used to **constrain queries** (either a template query or in the Query Builder), to **perform logical operations** (union, intersect, subtract) and can be exported. You can **search** lists and **copy** lists.
The '**Lists page**' can be accessed from the FlyMine home page or via the tabs in the top menu bar available on any FlyMine page. Lists that you have created yourself can also be accessed through your '**My Mine**' page. To save your lists permanently, you need to be logged-in to FlyMine.

A.The Lists Page

The lists page is divided into two tabs - '**Upload**' and '**View**':

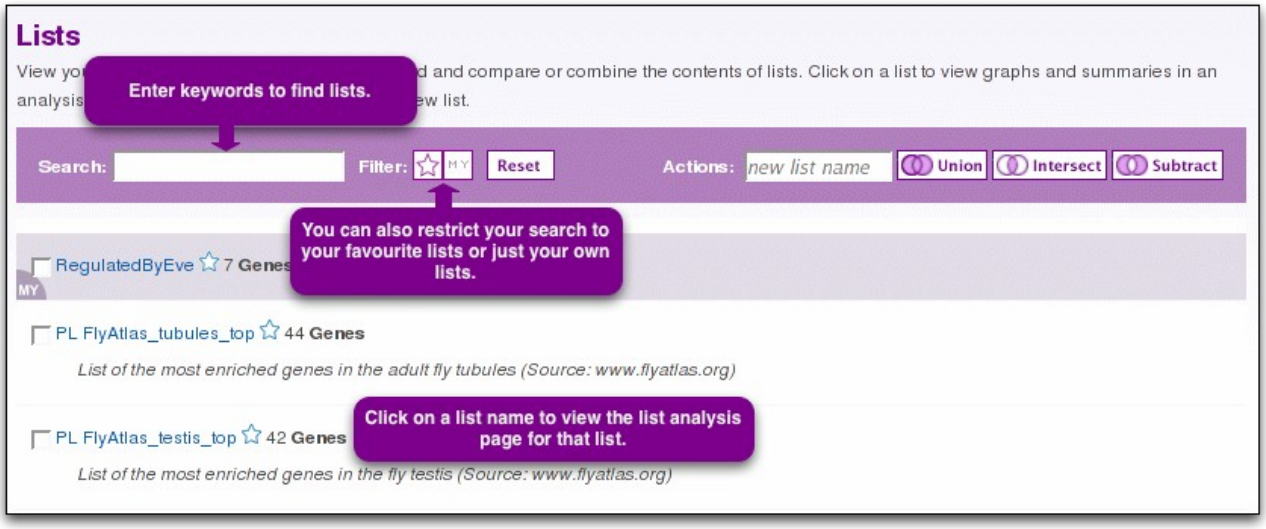
1. - Upload -

Use the list upload tab to create a list from an external source. This page allows you to either type/paste-in a list of identifiers or upload a file containing the list. Note that you can only create a list of one 'type' of thing (e.g. genes), so you must set the 'type' of identifier you are uploading first from the drop-down list. It is possible, however, to have a mix of the type of identifier (e.g. for a list of *Drosophila* genes your list may contain both gene symbols, CG identifiers and FBgns).
Optionally you can also set the organism for your list of identifiers - this is useful if you have, for example, gene symbols in your list which may appear in more than one organism. When you create your list all your identifiers will be compared against the FlyMine database and any discrepancies will be reported to you on the '**list confirmation page**'. This is described in more detail in the exercises.



2.- View -

The 'View' tab shows all your lists and all public lists with their descriptions by default. The descriptions can be hidden by deselecting the check box. All public list names begin with 'PL', while your own lists are labelled 'My' and appear in a coloured bar. Like templates, you can set lists as favourites if you want to be able to access them quickly. The bar above the lists allows you to search for a particular list using keywords, show just your favourite lists or show just your own lists.



3.List Operations

If you have more than one list, it is possible to carry out list operations - currently union, intersect and subtract are available. These operations are available both on the Lists-View page and on the lists tab of your MyMine account. Simply select the check boxes for the lists you wish to use, enter a name for the new list and select the type of operation you wish to perform. You can copy lists in a similar way. Without a new name, FlyMine will name the list original_list_copyx. You can copy multiple lists all at once.

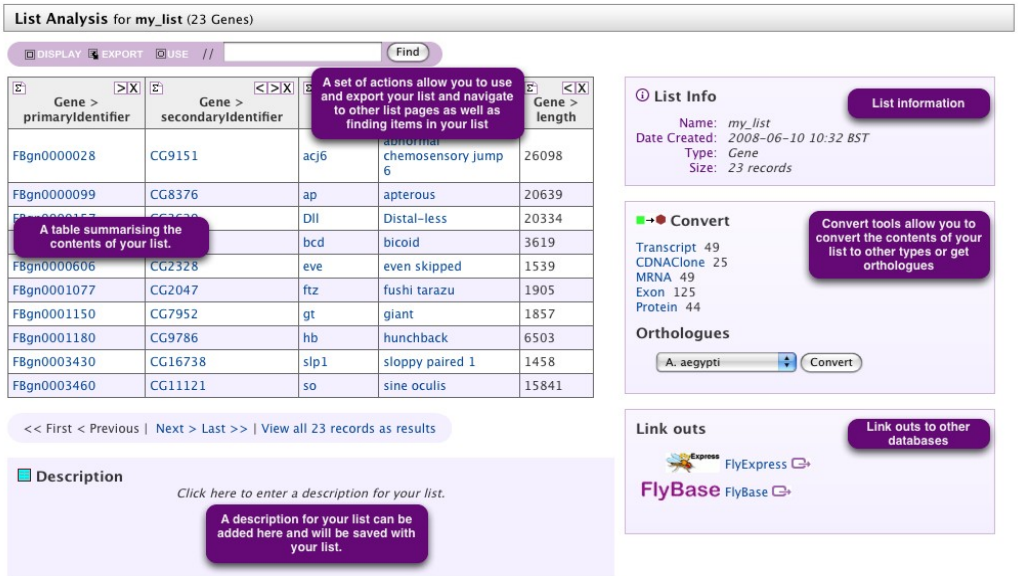
Lists
View your own and public lists, search by keyword and compare or combine the contents of lists
an analysis page, select lists
using checkboxes to perform set operations. Click 'Upload' above to import a new list.
2. Enter a name for your new list
Search:
Reset
Actions:
New list name
Union
Intersect
Subtract
Copy
1. Select the lists you wish to subtract using the check boxes.
You are not logged in. Log in to save lists permanently and to mark item
3. Select the logical operation you wish to perform.
PL FlyAtlas_brain_top 44 Genes
List of the most enriched genes in the adult fly brain (Source: www.flyatlas.org)
PL FlyAtlas_head_top 49 Genes
List of the most enriched genes in the adult fly head (Source: www.flyatlas.org)

B.List Analysis Page

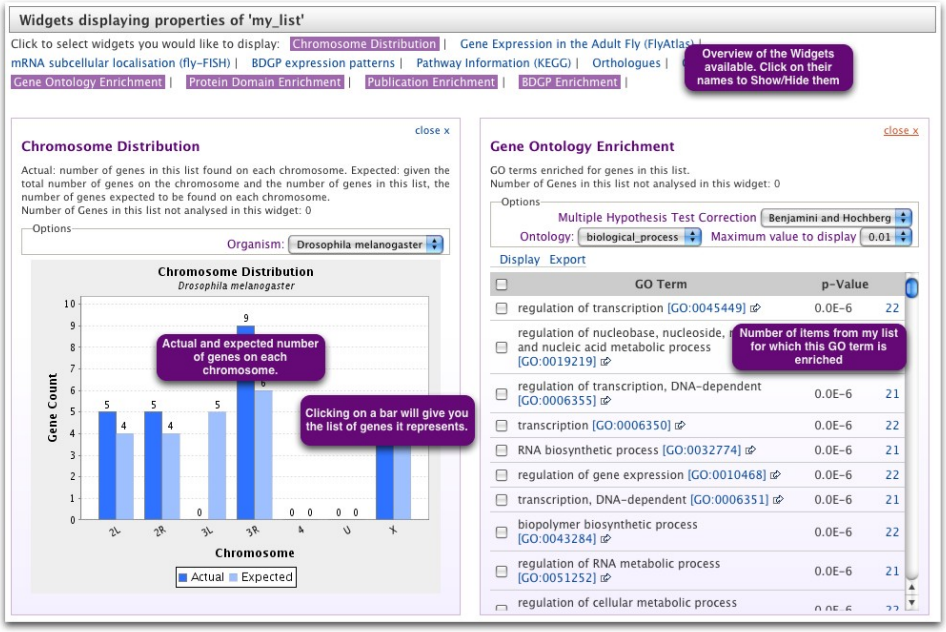
Details of the contents of a list together with additional information about a list can be viewed through the '**List Analysis**' page. The analysis page for a list is automatically displayed when a list is created. List analysis pages can be accessed at other times by clicking on the list name on the 'list page'.

All list analysis pages have the same structure and consist of three sections:

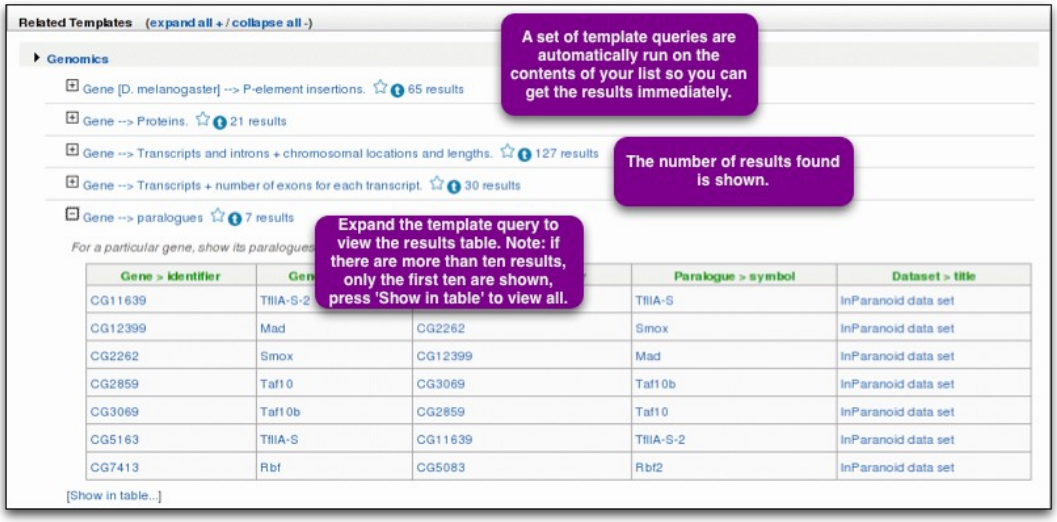
- a **summary** of the contents of the list includes a table with the first 10 obects of the lists, a list description, List info, Convert tools and Link outs (if available). Above the table, there are Display / Export / Use options as well as a Find box. Note that searching a list is case-sensitive and wildcards cannot be used.



- a series of '**widgets**' giving more information on the list, for example, for a list of genes, a graph of the chromosome distribution, a statistical analysis of the GO term distribution (GO term enrichment), KEGG pathway data, protein domain enrichment and publication enrichment.

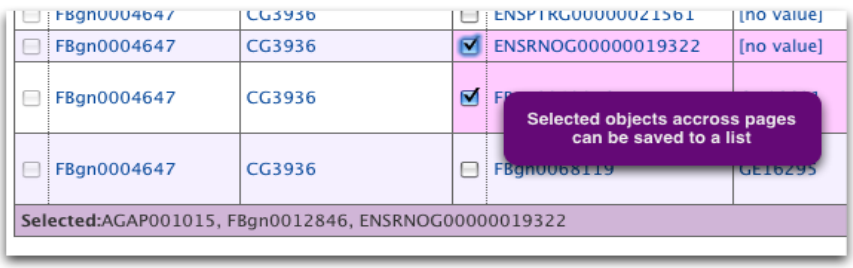


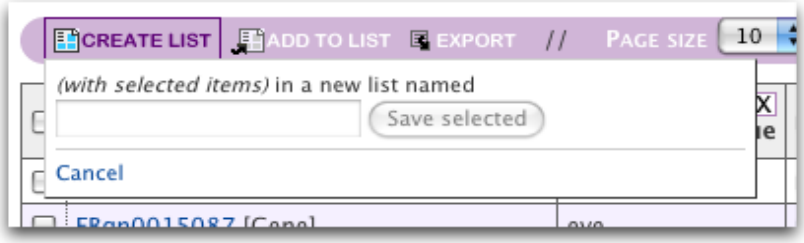
- a set of **template queries** that have been pre-run on the set of objects in the list:



Lists can be created from the results of queries:

Data from queries can also be saved as a list. To save your data as a list select the required data on a results page and either add it to an existing list or save it as a new list, using the





boxes displayed above the results. Selecting the check boxes next to the column headings will select the entire column for saving. Alternatively, individual items can be saved by selecting check boxes next to the items. Selected items will be listed in the box below the results table.

Exercise:

In this tutorial we will upload an example list of genes. We will then use this list to look at a list analysis page, run a template query and save a further list from the results.

6.1. **Upload the example list.**

- Navigate to the 'upload' tab of the lists page.
- Above the box for typing/pasting in identifiers, click on '(click to see an example)' to add the example to the box.
- Set the '**type**' of list to Gene. Leave the organism constraint blank. Press '**Create List**' to upload the list. This will take you to the '**List confirmation**' page:

6.2. **List Confirmation Page**.

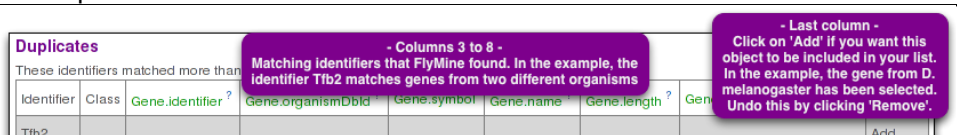
- Before a list is created it is checked against the FlyMine database to ensure that everything in the list can be found. Any discrepancies are reported on the '**list confirmation**' page. The demo list matches 26 genes successfully but contains two discrepancies - a low quality match and a converted type. These are explained in more detail below:

Low quality matches: A low quality match is one that does not match the main identifiers in the database but which was found by searching the synonyms. In this case FBgn0001251 is an old (or secondary) identifier for the gene. However, it has found the gene we want (with current identifier FBgn0001325) so we can add this to our list by clicking on 'add' in the last column.

Duplicate genes: You may also see discrepancies called 'Duplicate genes'. This can occur if, for example, you have gene symbols in your list and the symbol is found in more than one organism. In such a case, all occurrences of the gene will be listed and you can select the one you intended to be in your list. Such discrepancies can be avoided by using the optional organism constraint on the list upload page.

Converting types: When we uploaded our list we specified that we wanted to make a list of genes. However one item in our list is actually a protein identifier (TWIST_DROME). FlyMine has found this protein in the database along with the gene that encodes it. It is not possible to make a list of mixed 'types' but we can now add the gene for this protein to our list.

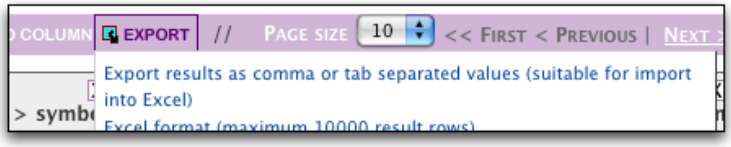
An example of 'duplicate' items is shown below:



IV.Exporting Data

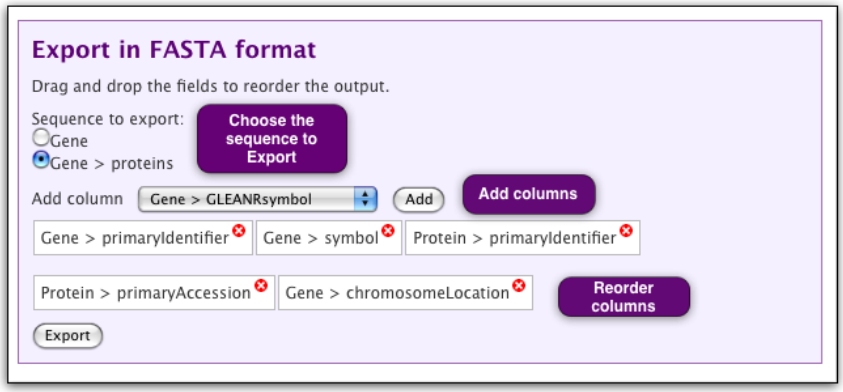
Exporting lists and result tables

Data can be exported from lists and results pages in various formats - excel (or open office) format, comma separated values, tab separated values, or for items with sequences, FASTA format (see below). To export the contents of a list or a results page, use the button in the bar above the table, and select the export format.



Currently all the results are exported, regardless of which items are selected in the table. You are then presented with a range of options depending on which format you selected.

Add column: Lets you add a column to the exported data even if it was not displayed before.
Reorder fields: It is possible to change the order of the fields displayed before exporting them by using drag and drop. This change is not reflected in the results table.
FASTA format: If more than one columns contains objects that have a nucleotide or amino acid sequence, it is possible to export all sequences in FASTA format. You can select which object you want to export the sequence of in the export options. If objects appear more than once in the column, the sequence will only be exported once.



Exercises:

7.1 **Export** the sequences of the regulated genes from exercise 6.4.

- load the list analysis page for the 'regulated genes' list you created in exercise 6.4.
- Click on EXPORT in the toolbar above the table.
- Choose ‘Export in FASTA format’.
- Select the option ‘TFBindingSite > gene’ to export the sequences of the regulated genes.
- Click on ‘Export’.

7.2 **Export** the corresponding protein sequences of regulated genes from exercise 6.4.

- Find the template: Gene --> Proteins.
- Run it for the list of 'regulated genes' from exercise 6.4.
- Click on EXPORT in the toolbar above the table, and select ‘Export in FASTA format’.
- Select the option ‘Gene > proteins’ to select the Protein sequence. Click on ‘Export’.

V.Query Builder

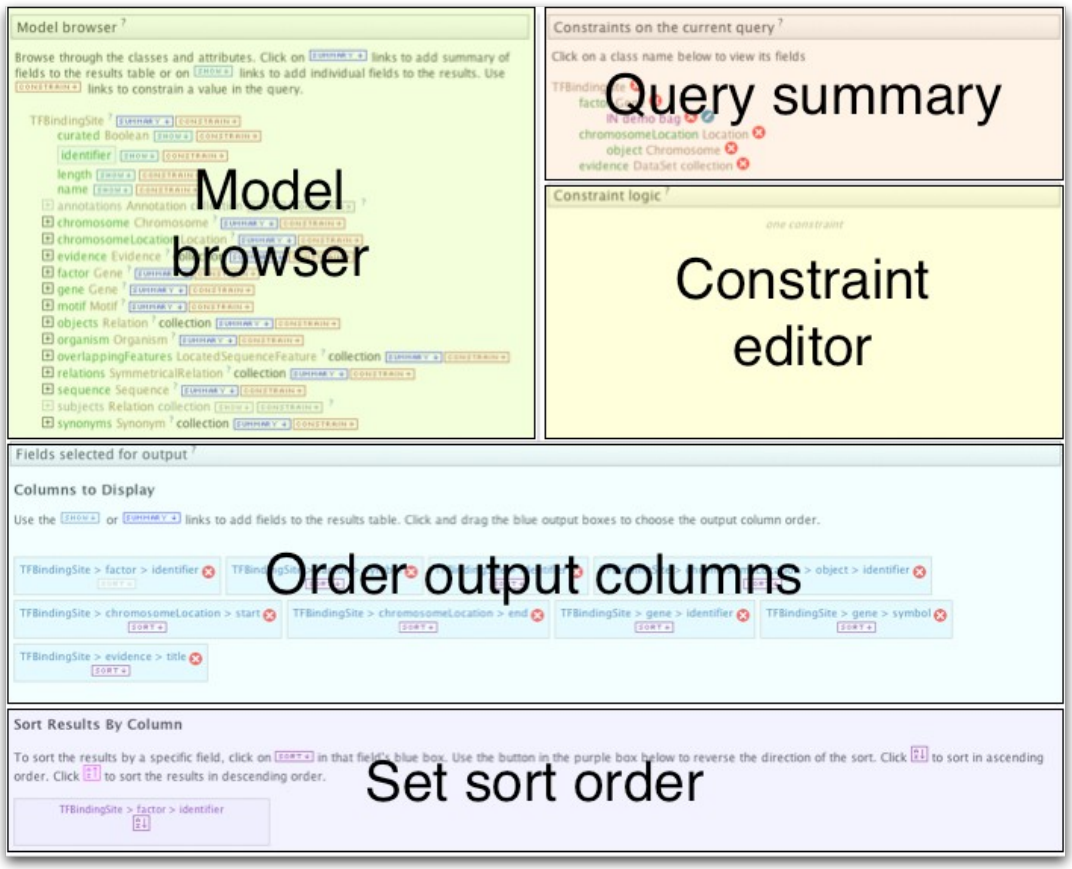
Advanced users can use a **flexible query interface**

- to construct their own data mining queries across the multiple integrated data sources,
- to modify existing template queries or
- to create their own template queries.

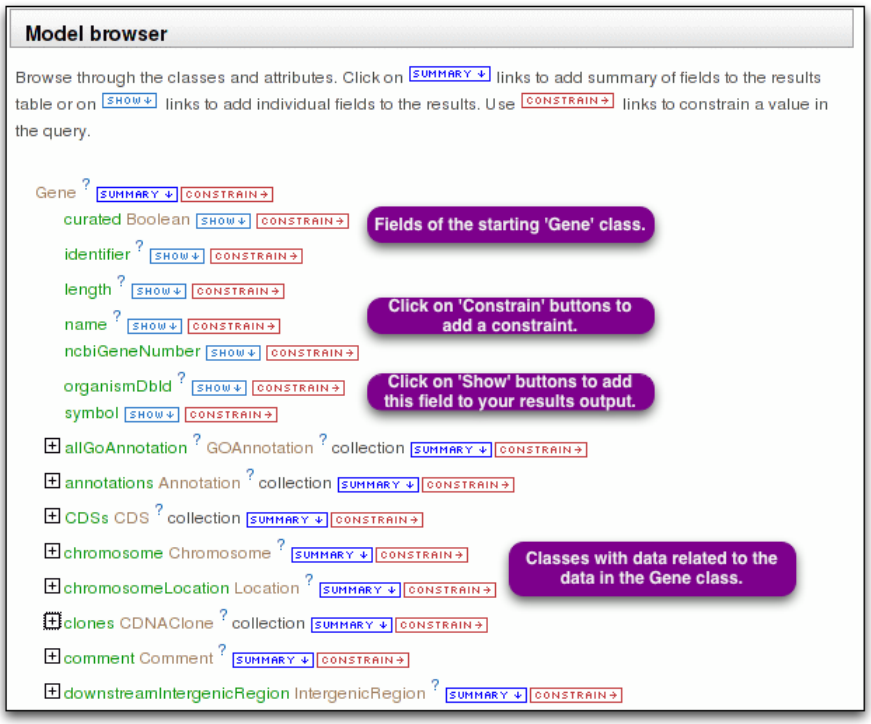
Access the FlyMine Query Builder at the bottom of the FlyMine home page, from the tabs in the main menu bar on any FlyMine page or choose a query starting point from a data category page. The Query Builder interface enables you to browse through classes and fields of the data model, selecting those that you wish to constrain and those that you wish to include in your results (see appendix I for definition of some terms). Building a query involves three main steps:

1. Browsing the model browser to find the classes of data you wish to query.
2. Applying constraints to this data to define the subset of that type of data required and
3. Defining your results table by selecting the fields you want to include.

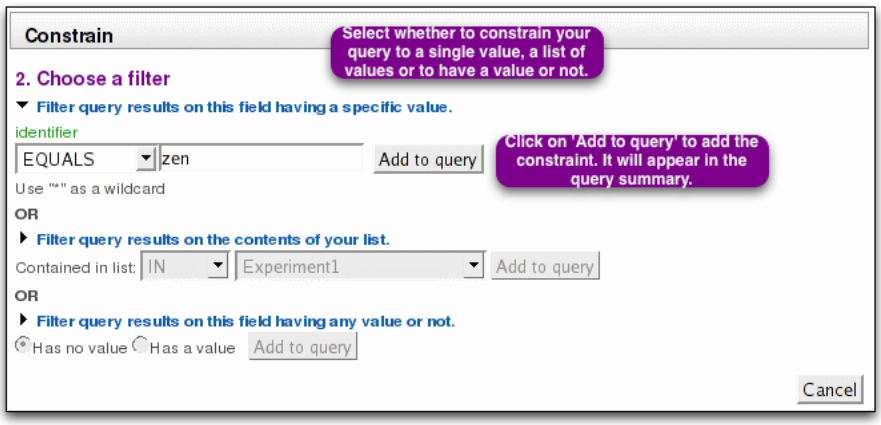
The main **FlyMine Query Builder page** is composed of several sections:



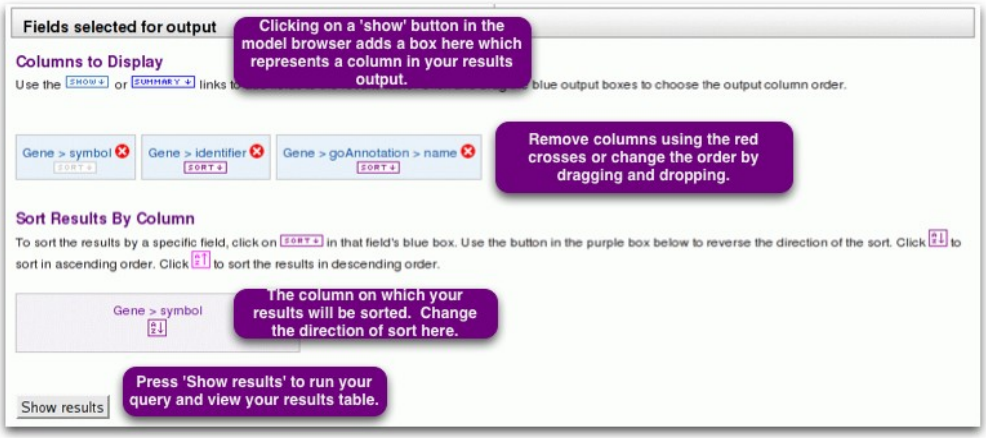
A. The **model browser** (left) allows you to browse through the FlyMine data model. This displays an initial starting class and its fields with classes containing related information listed below. Each class can be expanded to show the fields for that class and in turn, any classes related to that class. For example, if the starting class is 'Gene', the fields shown will include 'name', 'identifier' and 'symbol' and the classes of related information include 'Proteins', 'Transcripts' and 'Microarray results'. The model browser thus allows you to browse through the types of data that are available and select those that you wish to query. For each field within the model browser there is a 'SHOW' and a 'CONSTRAIN' option. These are used to either select a field to show in the results output or to constrain in the constraints list (see below).



B. The **constraints list** (right) allows you to specify the constraints you wish to apply to particular classes and fields. A constraint is like a filter - you filter out a subset of objects in a class which have certain characteristics (defined by the constraint). To apply a constraint to a field, navigate to the field in the model browser and click on 'constrain' next to it. This will add the field to the constraints list and bring up a box (lower right) where you specify your constraint. You can change the relationship (*and, or*) between the different constraints using the 'constraint logic'.



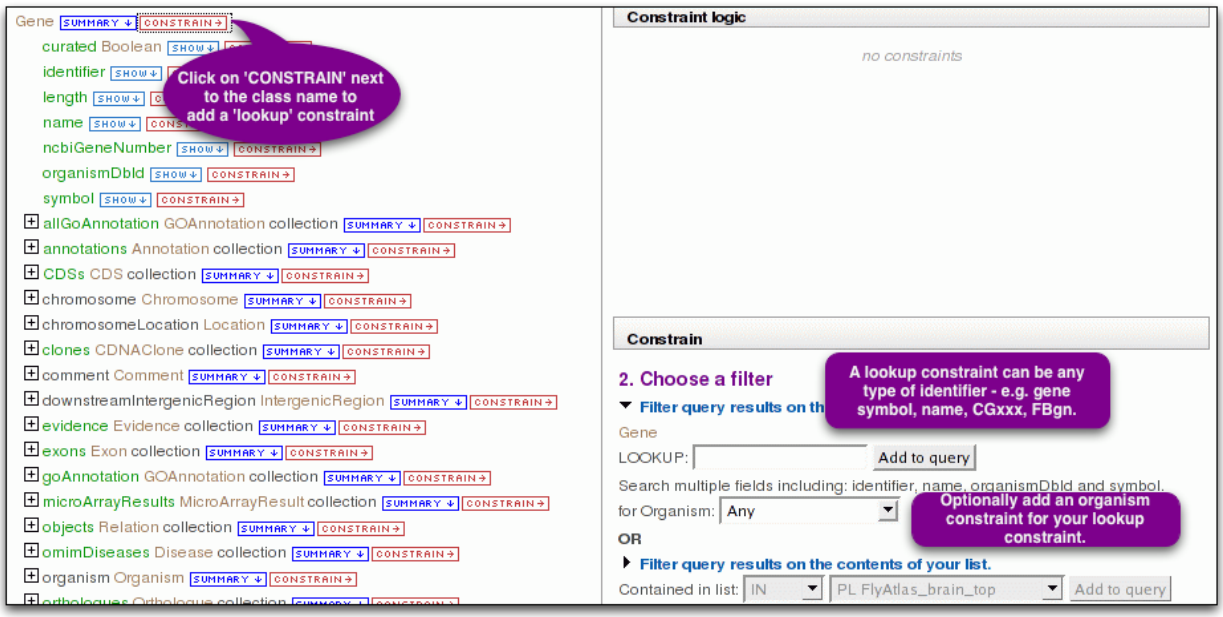
C. Underneath the model browser, the '**fields selected for output**' displays the fields you have selected to show in your results. (Each field is represented by a box which will become a column in the results table, Initially, the output fields list will say 'No fields selected for output'). To select a field to show in the results output you press 'SHOW' next to the field. (Alternatively selecting 'SUMMARY' next to the class name will add the main fields for that class to your results output). The order of fields in the output fields list can be changed by dragging and dropping them. Fields can be deleted from the output fields list using the 'x'. Fields that have been selected for output will appear in a blue box in the model browser and constraint list. By default results will be sorted alphabetically, in ascending order, based on the first column. The column to sort on can be changed by clicking on the '**sort**' box under the column name - note this changes the name of the column displayed under '**Sort Results By Column**' . The **direction of sorting** (ascending/descending) can be changed by clicking on the box under the field name in this section.



Press '**Show results**' to view your results in the results page. You can go back to edit your query by clicking on '**Query**' in the **trail**. The results page will by default display just the first ten rows of your results (this can be changed using the drop-down box above the results table). Below the results table the total number of rows returned by your query is displayed (this will initially say 'Approximate rows' as the total number is calculated, changing to 'Total rows' once the exact number is known.

LOOKUP constraints

There are two ways in which you can add a constraint for the identifier for a particular object. For example, for a particular *Drosophila melanogaster* gene you may want to constrain either the CG identifier, the FBgn identifier, the gene symbol or the gene name. Each of these options is available as a field under the gene class (identifier holds the CG ids and organismDbId holds the FBgn ids) and so depending on the type of identifier you have you can select which field you need to constrain. Alternatively, however, FlyMine includes a feature called 'LOOKUP constraints', which allows you to add any type of identifier to the constraint. To add a LOOKUP constraint, click on the 'CONSTRAIN' button next to the actual class name. Such LOOKUP constraints also allow you to specify the organism in the same constraint (useful, for example, if you are using gene symbols which may occur in more than one organism).



Exercises:

This example introduces the use of the Query Builder to collect all gene identifiers from genes on the 4th chromosome of *D. melanogaster*.

8.1. Go to the 'QueryBuilder' tab, select **Gene** from the list of Data Types and click on 'Select'.

8.2 Select the fields for output:

- Click on 'Summary' next to Gene in the model browser to add predefined fields to the output. Scroll down to the bottom of the page and notice that fields have been added to the output (5 in total).
- Scroll back up and add the Gene length by clicking on 'Show' next to length. You can re-order output fields (columns) by using drag and drop, so go ahead and drag the **Gene > length** between **Gene > symbol** and **Gene > name**.

8.3 Only retrieve all genes on the 4th chromosome of D. melanogaster.

- To constrain the current query to D. melanogaster, click on the '+' sign next to organism in the column on the left. Click on 'CONSTRAIN' next to 'name' (under 'organism). This brings up a 'Constrain' section to the right (you might have to scroll down to view it). Select 'Filter query results on this field having a specific value' and choose Drosophila melanogaster from the drop-down menu. Click 'Add to query' next to the drop-down menu. Notice how the constraint is now displayed in the **Constraints in the current query** section on the top right.
- Next , constrain the current query for genes on the 4th chromosome, by clicking on 'CONSTRAIN' next to 'primaryIdentifier' under 'chromosome'. In the 'Constrain' section to the right, enter a value for the chromosomal identifier: the identifier for the 4th chromosome is '4'. Click 'Add to query' next to the drop-down menu.

8.4 Scroll down and click 'Show results' to run the query. How many genes did you find?

8.5 Re-run the query with the results sorted Z-A on the Gene symbol.

VI.Super User

The SuperUser is the administrator of your InterMine webapp. The SuperUser can use tagging to configure the appearance and functionality of the webapp.

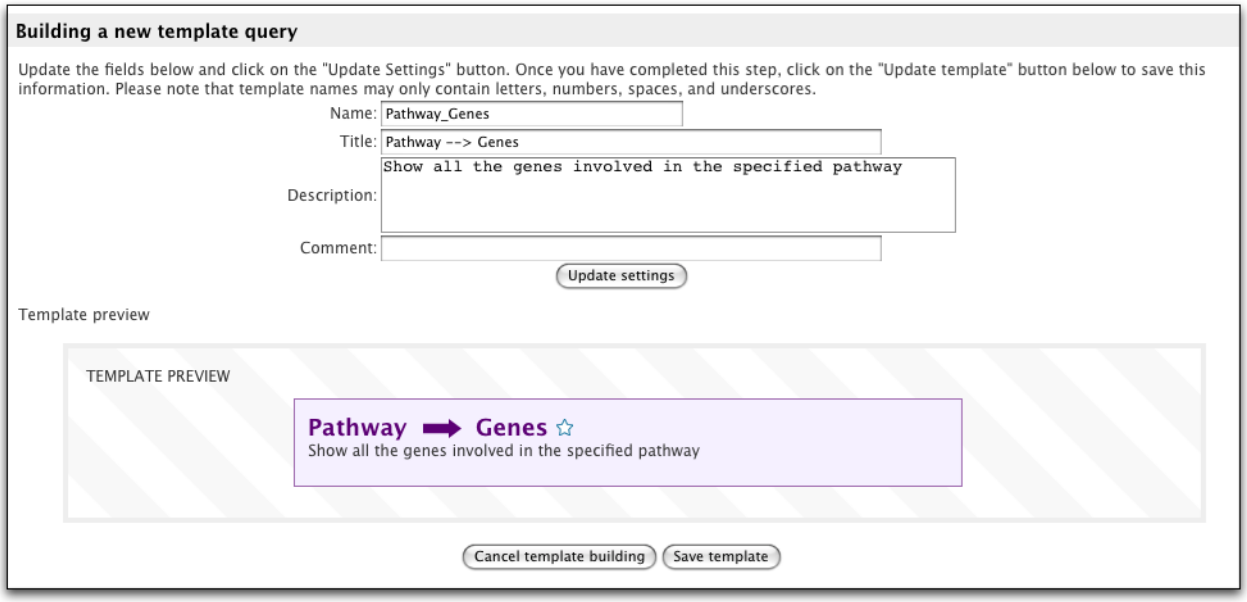
The SuperUser account is created when the UserProfile database is built using the email address specified in the malariamine.properties file and a randomly generated password. Refer to WebappRelease for instructions on retrieving your password, if you don't already have it.

We recommend that you login and change your password immediately to something easier to remember.

A.Creating a Template Query

All logged in users can create template queries, but the **SuperUser** can make them available to all users by **tagging** them as public templates. Making a template query is an easy way to get users of your webapp to the data they want very quickly.

Template building is completely integrated into the **query builder**. We are going to build a simple template that returns all genes related to a pathway (*Pathway --> Genes*).



Exercices:

1. **Build a query**

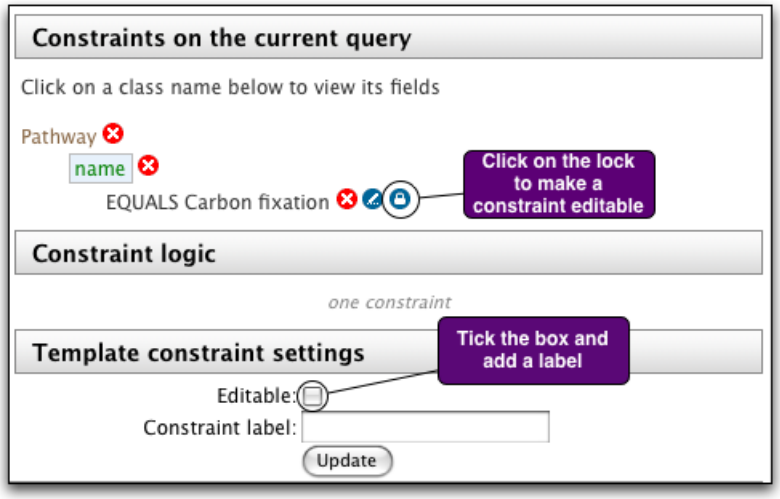
- Click on the QueryBuilder tab on the top menu bar
- Select the pathway class and click the select button.
- Add columns to the query output (the blue boxes)
    - In the Model Browser section (top left), click on the blue SUMMARY button next to Pathway.
        - These are the columns that will appear in the results
        - You can change the order in which the columns appear by moving the blue boxes in the 'Fields selected for output' section (bottom section)
        - Which fields are added to the view list when you click on summary is configured in the webconfig-model.xml file.
    - In the Model Browser section (top left), click on the blue SUMMARY button next to the genes collection.
- Add a filter to the query
    - In the Model Browser section (top left), click on the brown CONSTRAIN button next to the Pathway.name field.
    - This will open the constraint form in a panel on the right of the screen.
    - Enter a Pathway (eg. Carbon fixation) and click on the 'Add to query' button.
    - On FlyMine, typing in the constraint box here would suggest pathways as you type.

2. **Change to template-building mode**

- Verify that you are logged in as the SuperUser (admin mode)
- Click the button "Start building a template query" that appears below the output section
- Update the template's name, title and description and then click the 'Update settings' button.
    - name
        - Pathway_Genes
        - Used by javascript, webservices, etc to identify this template, must be unique
    - title and description
        - Pathway --> Genes
        - Title and description should be human readable
        - Both are displayed on the templates page and the template form

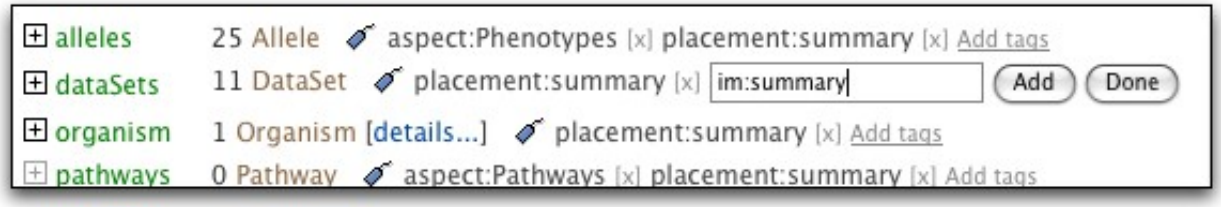
        -
        - Both are used in the template search, so they should be as descriptive as possible
    - comment is used and only viewable by the SuperUser

- Make the constraint editable
    - When in template building mode, constraints listed at the top-right gain an extra icon - a little padlock. The closed padlock means that the constraint is locked and not editable in a template query.
    - Click the padlock, which will bring up the settings for that constraint.
    - Check the editable checkbox and click the 'Add to query' button



- Save the template
    - Scroll down to the bottom of the page to the 'Building a new template query' section
    - You'll see a preview of your template
    - Click on the 'Save template' button below the preview to save your template.

B.Tagging



1.Template queries and lists

The SuperUser can change where templates and lists appear by adding tags via the templates and lists pages in the MyMine section of the webapp.

Tag	Purpose
im:public	make list/template viewable by all users
im:frontpage	put list/template on home page
im:converter	template used in generating links in the 'Convert' section on the list analysis page

2.Fields and collections

The SuperUser can change how fields are displayed by adding tags via the report page.

Tag	Purpose
im:hidden	hides the field/collection
im:summary	add collection to 'Summary' section of report page

3.Classes

The SuperUser can change how classes are displayed by adding tags via the model browser.

Tag	Purpose
im:aspect	class appears on aspect page
im:preferredBagTyp e	class appears first in the class selection

VII.Appendix I: Basic FlyMine - Terminology and concepts

Classes, objects and fields

The FlyMine database is built on top of an object-based data model. This data model defines the way in which the different types of data are related to each other and is made up of classes, which describe a particular type of data, and the relationships between classes.

- A **class** can be thought of as a holder to describe information about a specific type of thing. Things which have similar properties belong to the same class. Example: in FlyMine there is a 'Gene' class. The Gene class describes properties of genes.

- Every database entry is called an **object** and is an actual member of a class. Example: the *Notch* gene is a member of the Gene class.

- Each class has a set of **fields** (attributes) that provide information about the type of item being described. Example: the Gene class includes the fields 'name', 'symbol' and identifier' among others.

- Every object of a class has the set of properties (fields) that describe that class, although some of these fields may be empty (i.e. there may not be any data available for some fields).

- A class can be a parent class for more specialised classes (sub-classes). Example: the transcript class is a parent class for classes describing more specialised transcripts such as tRNA, snRNAs, etc.

References

- Classes are linked together through '**References**'. A reference is a pointer from one class to another class. Example: the Gene class has a reference to the Transcript class.

- In some cases, a reference will link an object in one class to just one object in another class. Example: the *Drosophila zen* gene, has a link to the Organism object '*Drosophila*' of the Organism class.

- In other cases, a reference will link one object in one class to more than one object in another class. Example: if a gene is made up of more than one exon, this gene object will link to several exons in the Transcript class. This type of reference is called a '**Collection**' (list).

The FlyMine Data Model

- The set of classes and their references is known as the **FlyMine Data Model**.

- Where ever possible, any of the classes and their relationships have been determined by pre-existing ontologies and models. For example, a large part of the data model is defined by the Sequence Ontology (http://song.sourceforge.net), which describes the relationships between features annotated to a genomic sequence.

Appendix II: Help and Feedback

FlyMine Website www.flymine.org:

- On-line documentation: Help links for each page (? in top right); Tour available on home page.
- Mailing list for announcements from the FlyMine website

Contact:

- Email: support@flymine.org
- Use the FEEDBACK FORM accessible on every FlyMine page
- Call us: 01223 760262

We welcome any type of Feedback!