InterMine

Open source data warehouse and web interface

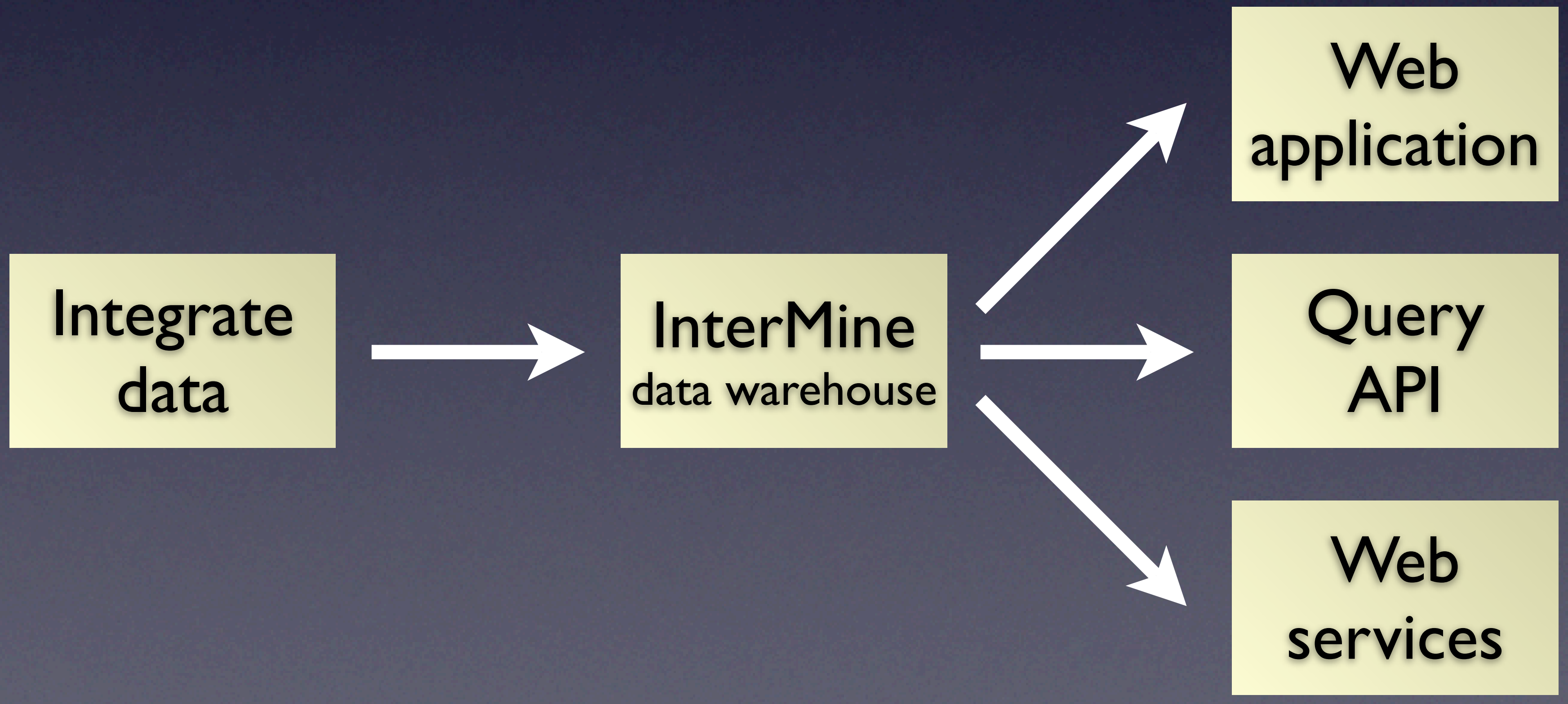
Richard Smith
University of Cambridge

Poster: *E34 (Monday)*

www.intermine.org

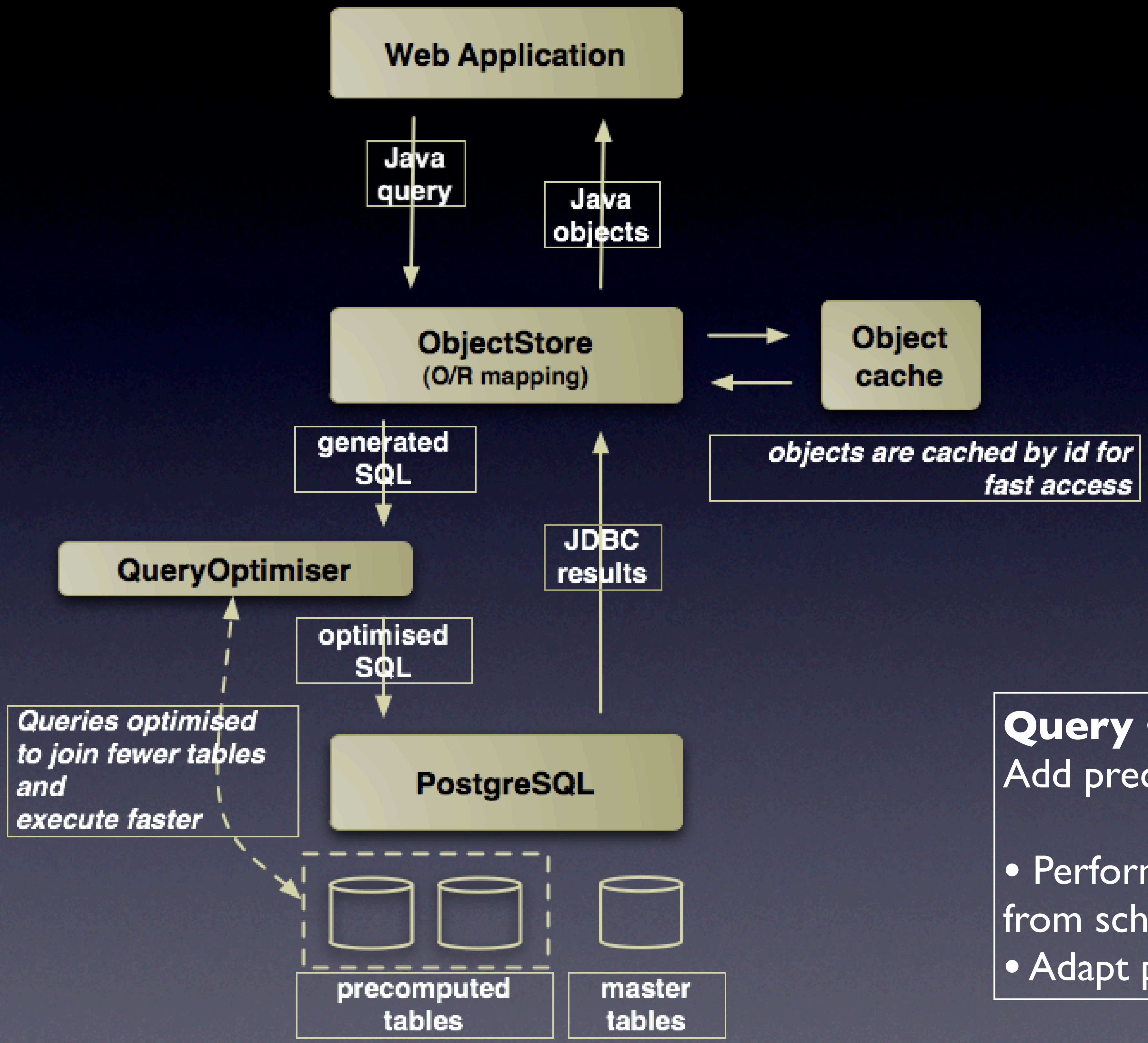
Overview

- Query-optimised data warehouse system
- Java, object-based data model
- Free, open source (LGPL)
- Flexible querying



Projects

- Five developers, since 2002
- FlyMine - www.flymine.org
  - 30+ data sources, *Drosophila* & *Anopheles*
- modENCODE - www.modencode.org
  - *C. elegans/D. melanogaster high throughput*
- BOKU & IMP - Vienna
- MitoMiner - mitochondria
- MilkMine - milk proteins
- *Yeast, Rat, Zebrafish*

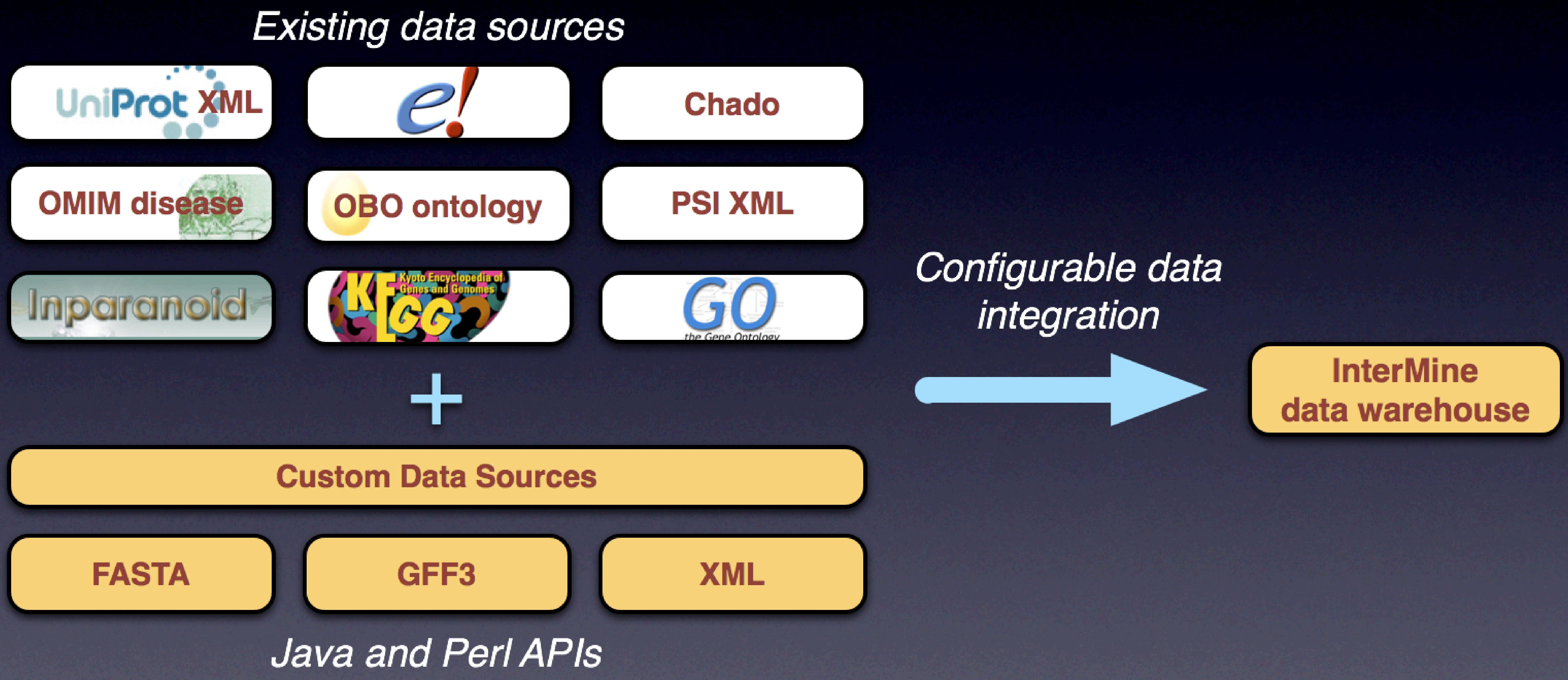


Query Optimisation

Add precomputed tables at any time:

- Performance optimisation separate from schema design
- Adapt performance to actual use

Data Integration



Existing data sources

+

Custom Data Sources

FASTA

GFF3

XML

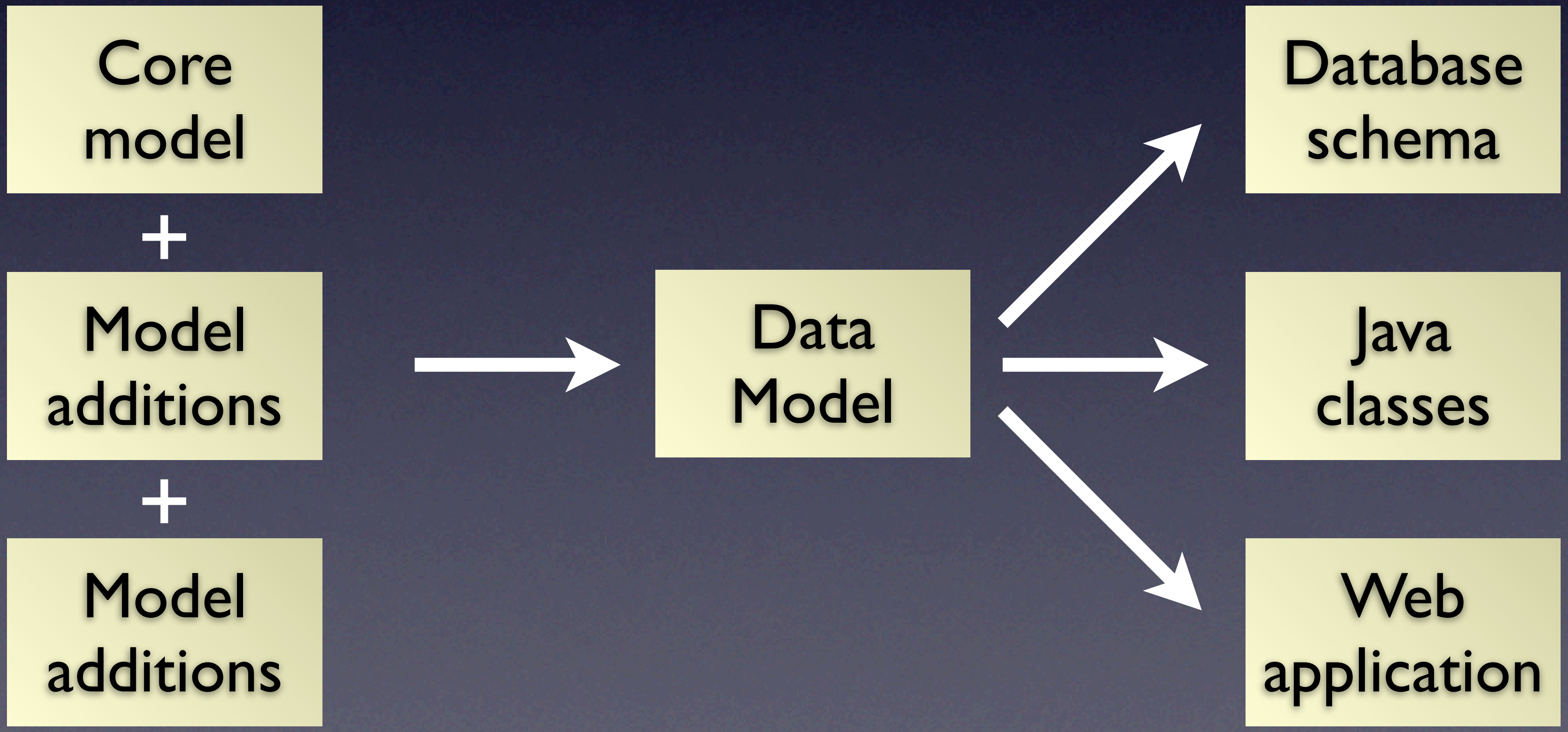
Java and Perl APIs

Configurable data integration

InterMine data warehouse

Auto-generation

- Object model defined by XML file
- Low overhead to extending data model



Custom data

- Any source can add to the data model:

```
<class name="Pathway" is-interface="true" >
  <attribute name="name" type="String"/>
  <collection name="genes" referenced-type="Gene"
              reverse-reference="pathways"/>
</class>
<class name="Gene" is-interface="true">
  <collection name="pathways" referenced-type="Pathway"
              reverse-reference="genes"/>
</class>
```

Configure a new Mine

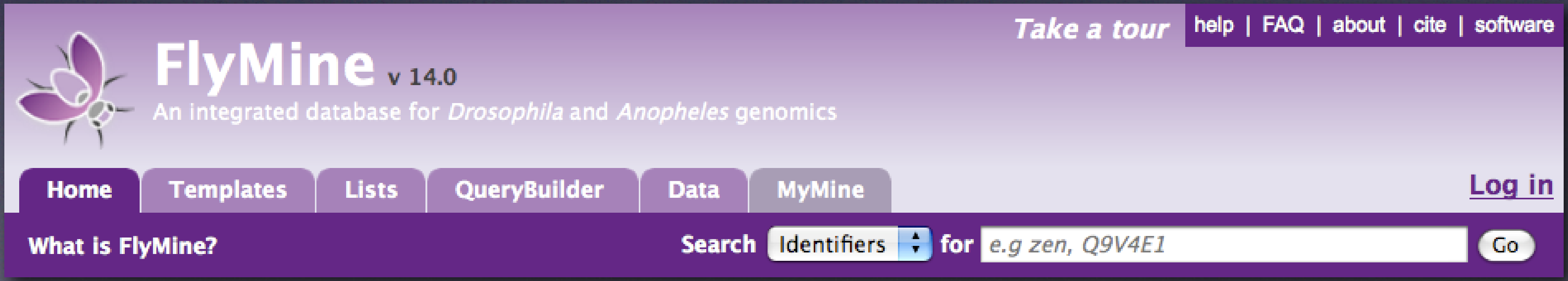
```
<project type="bio">
...
<sources>
  <source name="uniprot" type="uniprot" dump="true">
    <property name="uniprot.organisms"
              value="7227 6239"/>
    <property name="src.data.dir"
              location="/data/uniprot"/>
  </source>
  <source name="my-source" type="pathways">
    <property name="src.data.dir"
              location=”/data/pathways"/>
  </source>
  ...
<sources>
</project>
```

Example Usage

1. Subject specific database:
   - ‘Slices’ of data from repositories
   - Data sets specific to focus
   - e.g. Milk proteins, an organism, a disease
2. Present own data:
   - Web interface for own data
   - Add other sources to provide context

Web Application

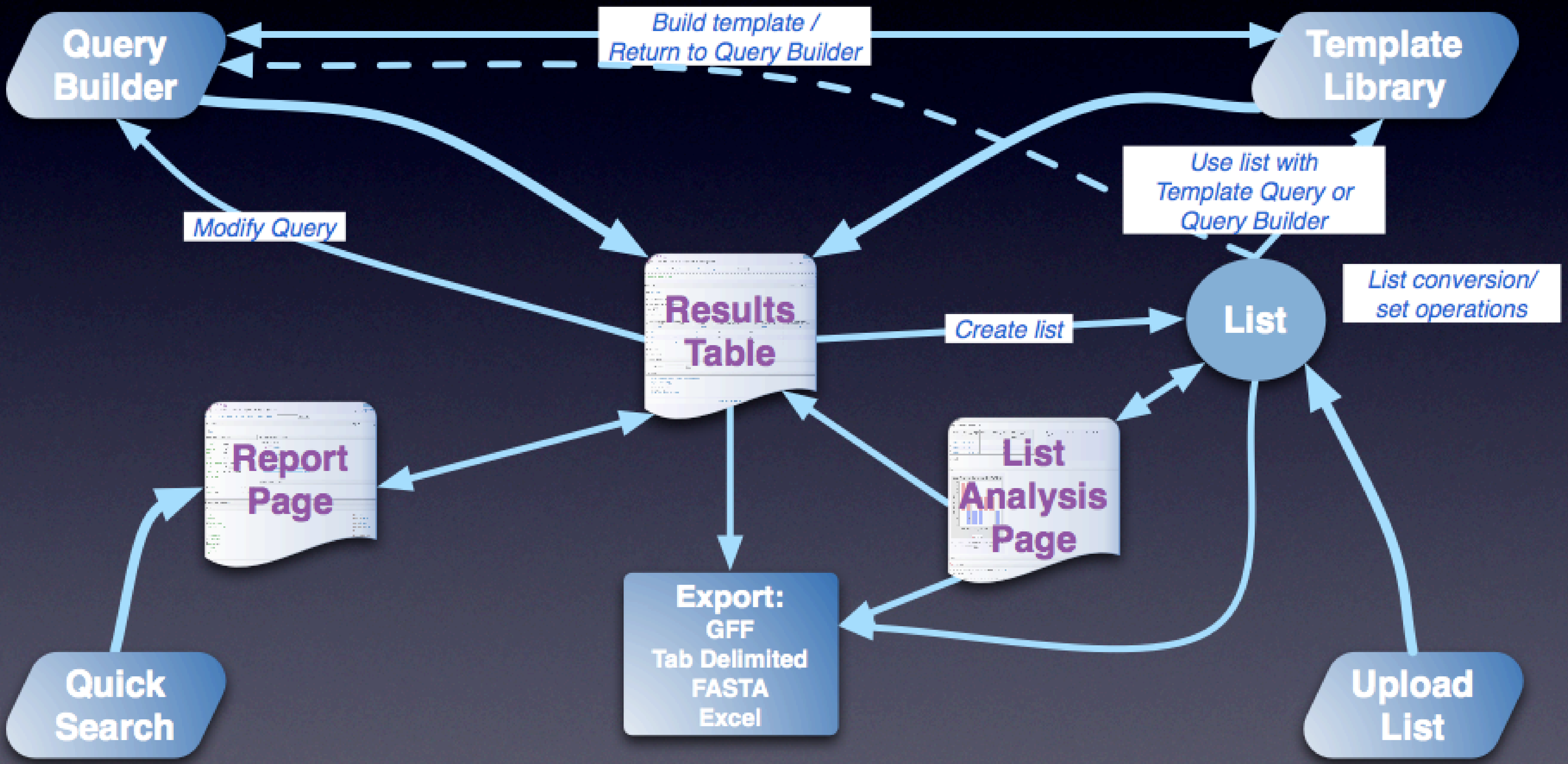
- Works for any data model
- Advanced functionality for bench biologist
- Highly configurable
- Configuration from within web interface


FlyMine v 14.0

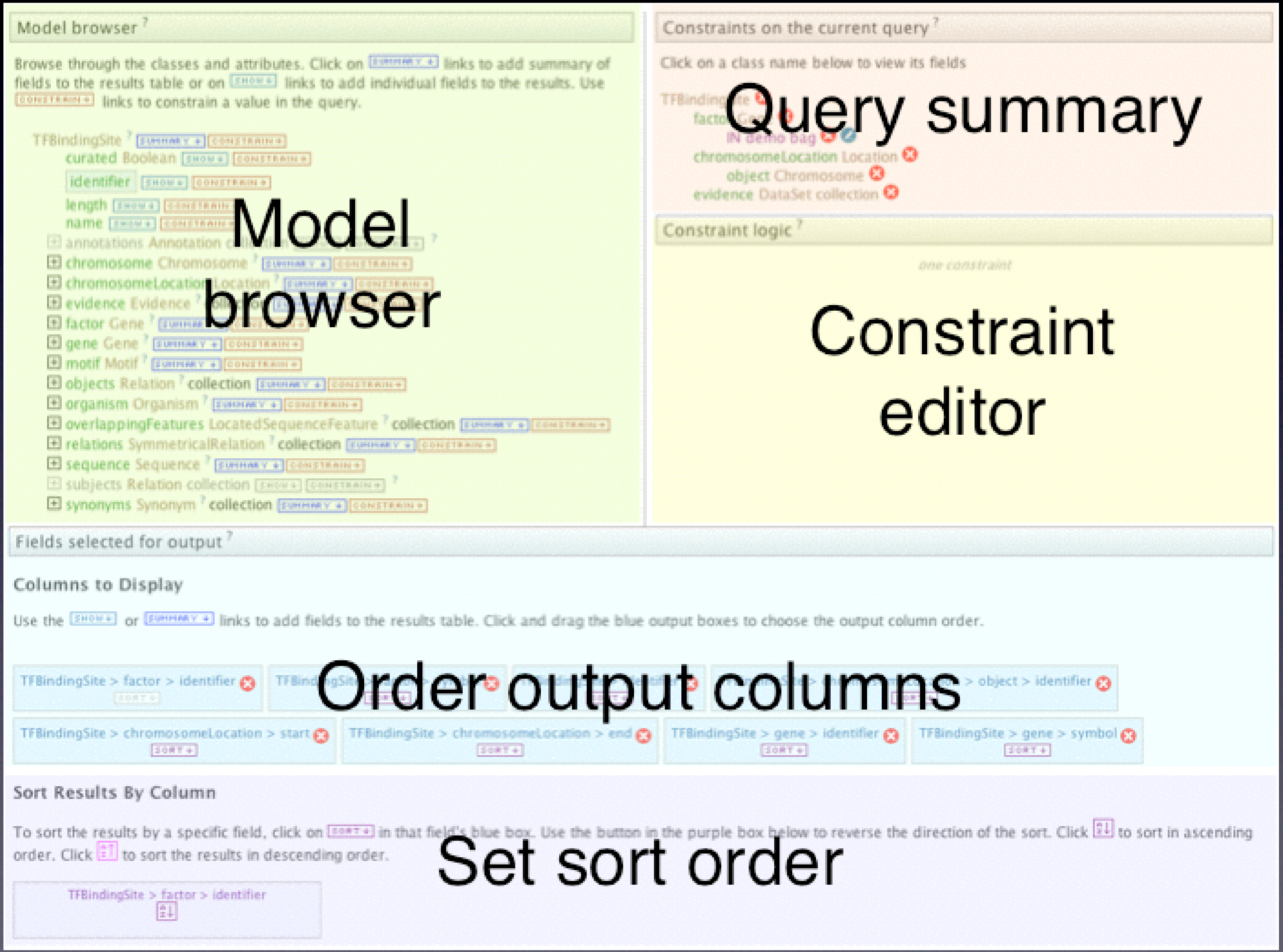
An integrated database for *Drosophila* and *Anopheles* genomics

Search Identifiers for e.g zen, Q9V4E1

Webapp Overview


Query Builder
Build template / Return to Query Builder
Template Library
Use list with Template Query or Query Builder
Modify Query
Results Table
Create list
List
List conversion/ set operations
Report Page
List Analysis Page
Export: GFF Tab Delimited FASTA Excel
Quick Search
Upload List

QueryBuilder



Model browser ?

Browse through the classes and attributes. Click on SUMMARY + links to add summary of fields to the results table or on SHOW + links to add individual fields to the results. Use CONSTRAIN + links to constrain a value in the query.

TFBindingSite ? SUMMARY + CONSTRAIN +
- curated Boolean SHOW + CONSTRAIN +
- identifier SHOW + CONSTRAIN +
- length SHOW + CONSTRAIN +
- name SHOW + CONSTRAIN +
- ⊞ annotations Annotation collection ?
- ⊞ chromosome Chromosome ? SUMMARY + CONSTRAIN +
- ⊞ chromosomeLocation Location ? SUMMARY + CONSTRAIN +
- ⊞ evidence Evidence ? collection
- ⊞ factor Gene ? SUMMARY + CONSTRAIN +
- ⊞ gene Gene ? SUMMARY + CONSTRAIN +
- ⊞ motif Motif ? SUMMARY + CONSTRAIN +
- ⊞ objects Relation ? collection SUMMARY + CONSTRAIN +
- ⊞ organism Organism ? SUMMARY + CONSTRAIN +
- ⊞ overlappingFeatures LocatedSequenceFeature ? collection SUMMARY + CONSTRAIN +
- ⊞ relations SymmetricalRelation ? collection SUMMARY + CONSTRAIN +
- ⊞ sequence Sequence ? SUMMARY + CONSTRAIN +
- ⊞ subjects Relation collection SHOW + CONSTRAIN + ?
- ⊞ synonyms Synonym ? collection SUMMARY + CONSTRAIN +

Model browser

Constraints on the current query ?

Click on a class name below to view its fields

TFBindingSite ⊗
- factor Gene ⊗
  - IN demo bag ⊗
- chromosomeLocation Location ⊗
  - object Chromosome ⊗
- evidence DataSet collection ⊗

Query summary

Constraint logic ?

one constraint

Constraint editor

Fields selected for output ?

Columns to Display

Use the SHOW + or SUMMARY + links to add fields to the results table. Click and drag the blue output boxes to choose the output column order.

- TFBindingSite > factor > identifier ⊗ SORT +
- TFBindingSite > chromosomeLocation > object > identifier ⊗ SORT +
- TFBindingSite > chromosomeLocation > start ⊗ SORT +
- TFBindingSite > chromosomeLocation > end ⊗ SORT +
- TFBindingSite > gene > identifier ⊗ SORT +
- TFBindingSite > gene > symbol ⊗ SORT +

Order output columns

Sort Results By Column

To sort the results by a specific field, click on SORT + in that field's blue box. Use the button in the purple box below to reverse the direction of the sort. Click ⇅ to sort in ascending order. Click ⇅ to sort the results in descending order.

TFBindingSite > factor > identifier ⇅

Set sort order

Query API

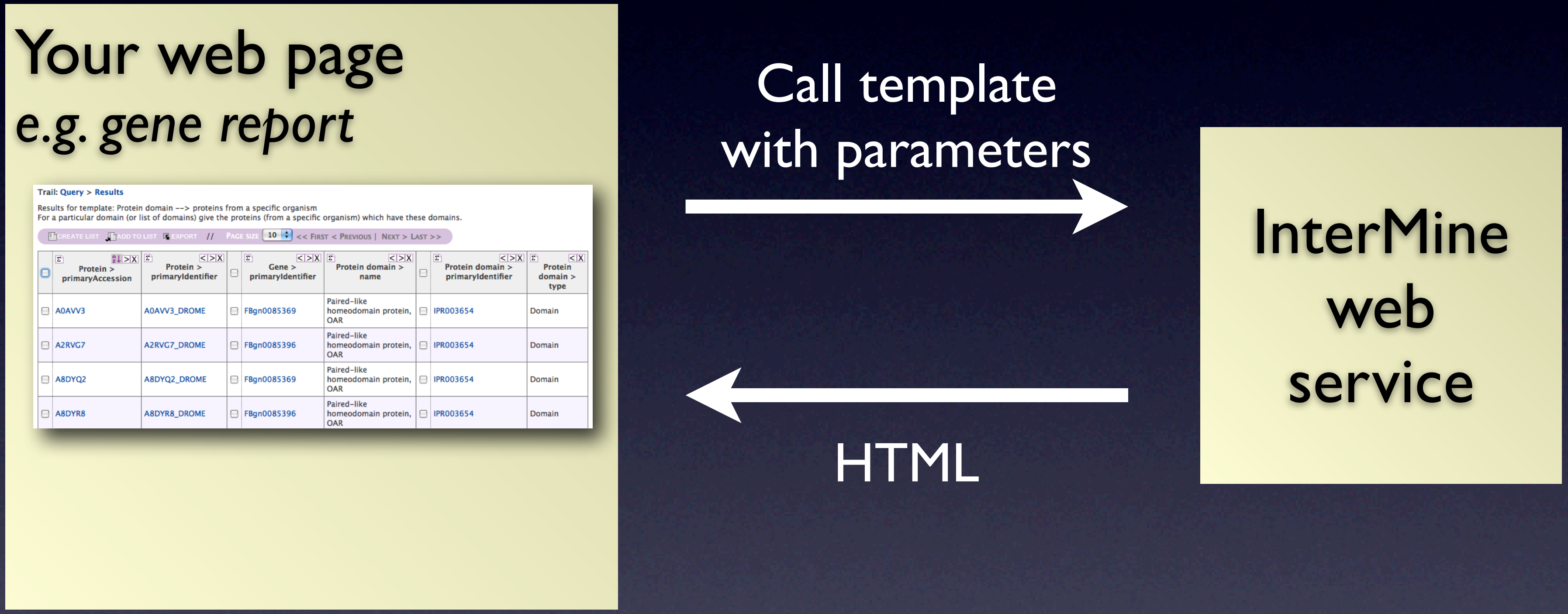
```
PathQuery q = new PathQuery(model);

q.setView("Protein.primaryIdentifier,
          Protein.genes.primaryIdentifier");

q.addConstraint("Protein.proteinDomains.name",
                Constraints.eq("Homeobox"));
q.addConstraint("Protein.organism.shortName",
                Constraints.eq("D. melanogaster"));
```

Web Services

- RESTful web service
- Run queries or templates
  - export XML from web app or use query API
- Java client API
- *Perl client API coming soon*
- *Lists, widgets, logins, tags to be added*

Embedding Templates

- Web service can return HTML



Trail: Query > Results

Results for template: Protein domain --> proteins from a specific organism
For a particular domain (or list of domains) give the proteins (from a specific organism) which have these domains.

CREATE LIST ADD TO LIST EXPORT // PAGE SIZE 10 << FIRST < PREVIOUS | NEXT > LAST >>

□	Protein > primaryAccession	Protein > primaryIdentifier	□	Gene > primaryIdentifier	Protein domain > name	□	Protein domain > primaryIdentifier	Protein domain > type
□	A0AVV3	A0AVV3_DROME	□	FBgn0085369	Paired-like homeodomain protein, OAR	□	IPR003654	Domain
□	A2RVG7	A2RVG7_DROME	□	FBgn0085396	Paired-like homeodomain protein, OAR	□	IPR003654	Domain
□	A8DYQ2	A8DYQ2_DROME	□	FBgn0085369	Paired-like homeodomain protein, OAR	□	IPR003654	Domain
□	A8DYR8	A8DYR8_DROME	□	FBgn0085396	Paired-like homeodomain protein, OAR	□	IPR003654	Domain

- ‘Embed this template’ link
- Saves remote site from integrating data
- *Widgets coming soon*

Acknowledgments

Biologists Hilde Jannsens, Rachel Lyne

Developers Richard Smith, Jakub Kulaviak, Julie Sullivan, Matthew Wakeling, Xavier Watkins

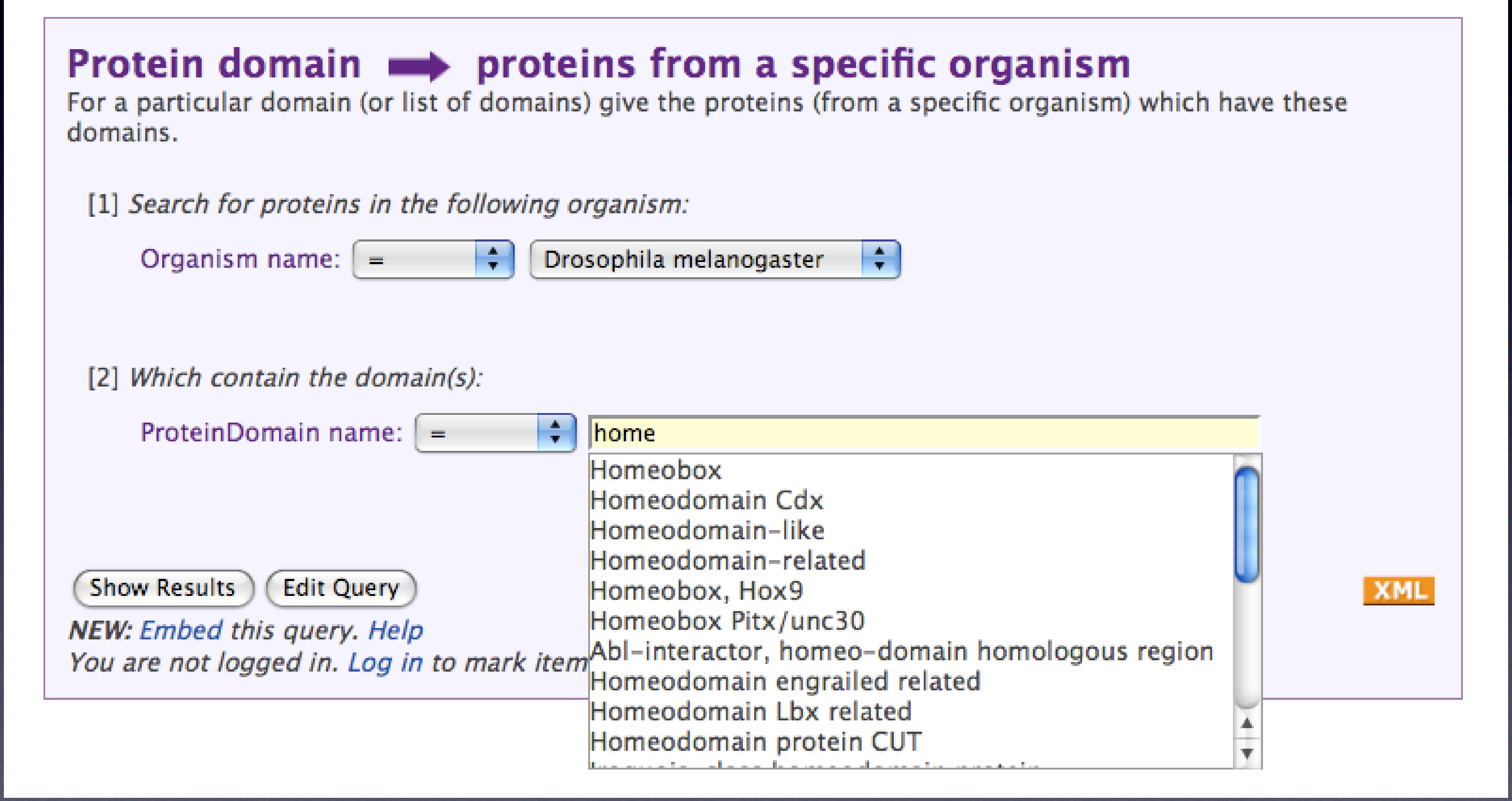
Sys Admin Dan Tomlinson

modENCODE Sergio Contrino, Kim Rutherford

PI Gos Micklem

www.intermine.org

Template Queries



Protein domain → proteins from a specific organism

For a particular domain (or list of domains) give the proteins (from a specific organism) which have these domains.

[1] *Search for proteins in the following organism:*

Organism name: = Drosophila melanogaster

[2] *Which contain the domain(s):*

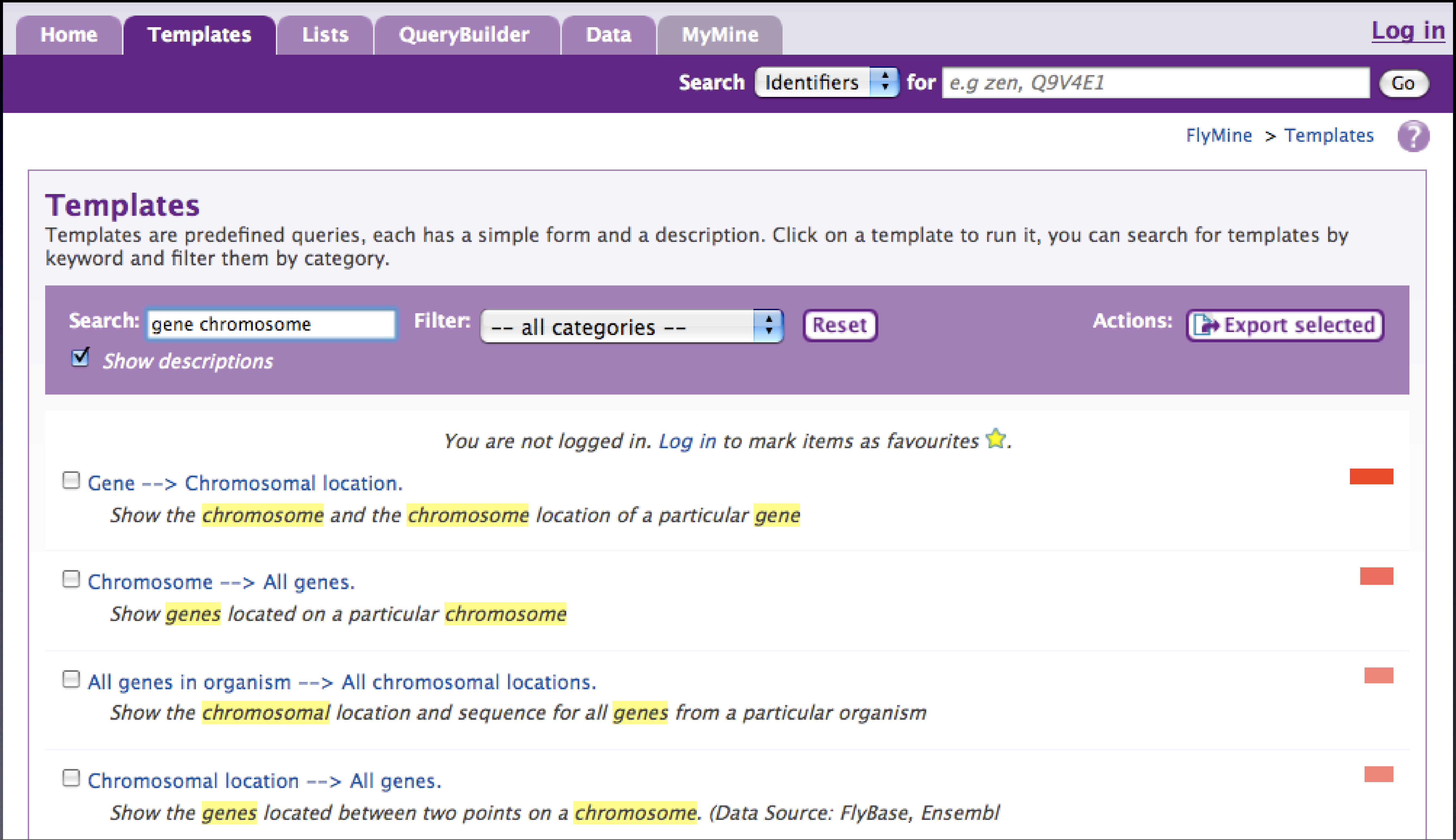
ProteinDomain name: = home

Homeobox
Homeodomain Cdx
Homeodomain-like
Homeodomain-related
Homeobox, Hox9
Homeobox Pitx/unc30
Abl-interactor, homeo-domain homologous region
Homeodomain engrailed related
Homeodomain Lbx related
Homeodomain protein CUT

NEW: Embed this query. Help

You are not logged in. Log in to mark item

Template Library



Templates

Templates are predefined queries, each has a simple form and a description. Click on a template to run it, you can search for templates by keyword and filter them by category.

Search: gene chromosome

Filter: -- all categories --

☑ *Show descriptions*

You are not logged in. Log in to mark items as favourites ☆.

□ Gene --> Chromosomal location.

Show the chromosome and the chromosome location of a particular gene

□ Chromosome --> All genes.

Show genes located on a particular chromosome

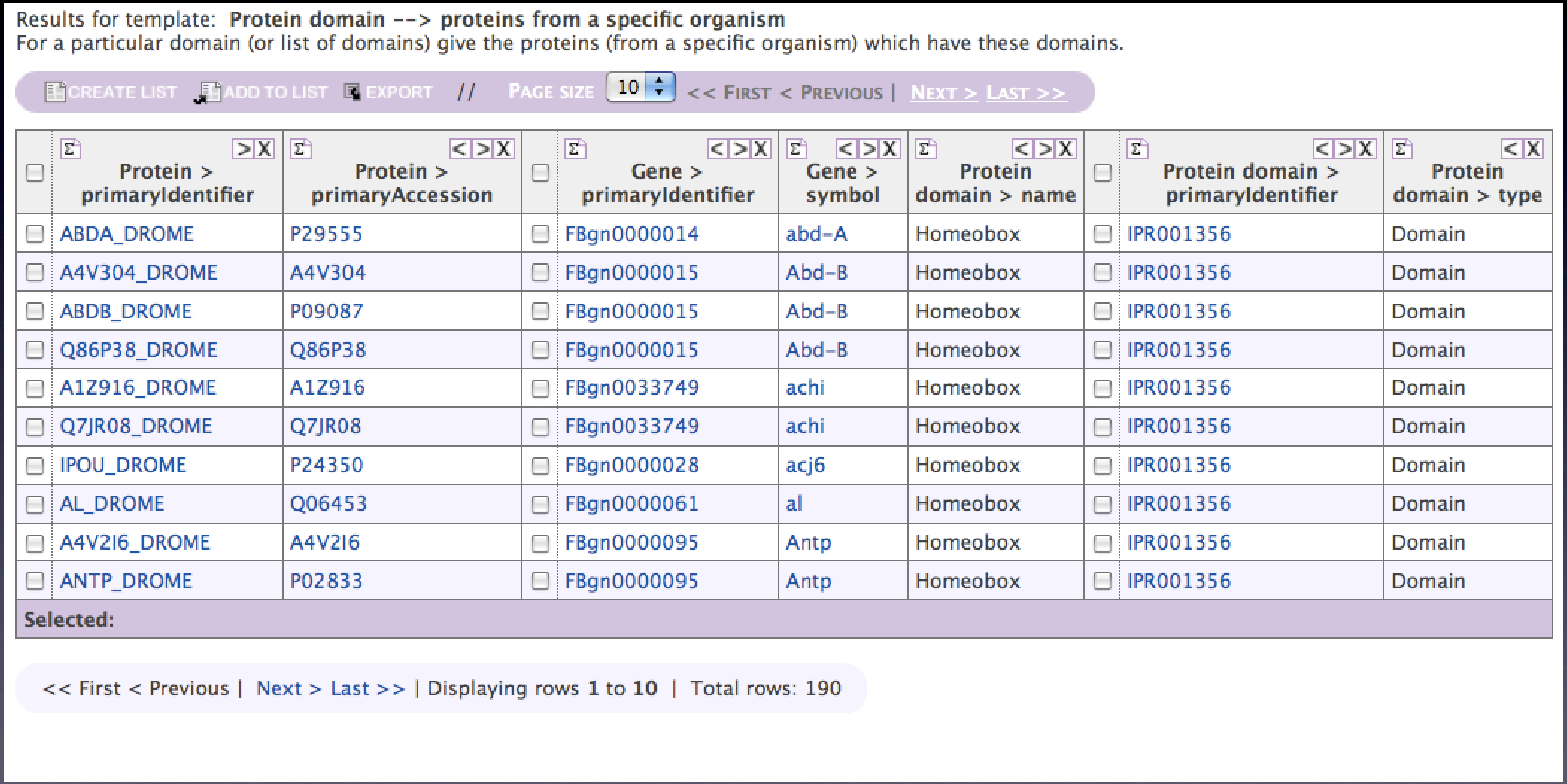
□ All genes in organism --> All chromosomal locations.

Show the chromosomal location and sequence for all genes from a particular organism

□ Chromosomal location --> All genes.

Show the genes located between two points on a chromosome. (Data Source: FlyBase, Ensembl

Results



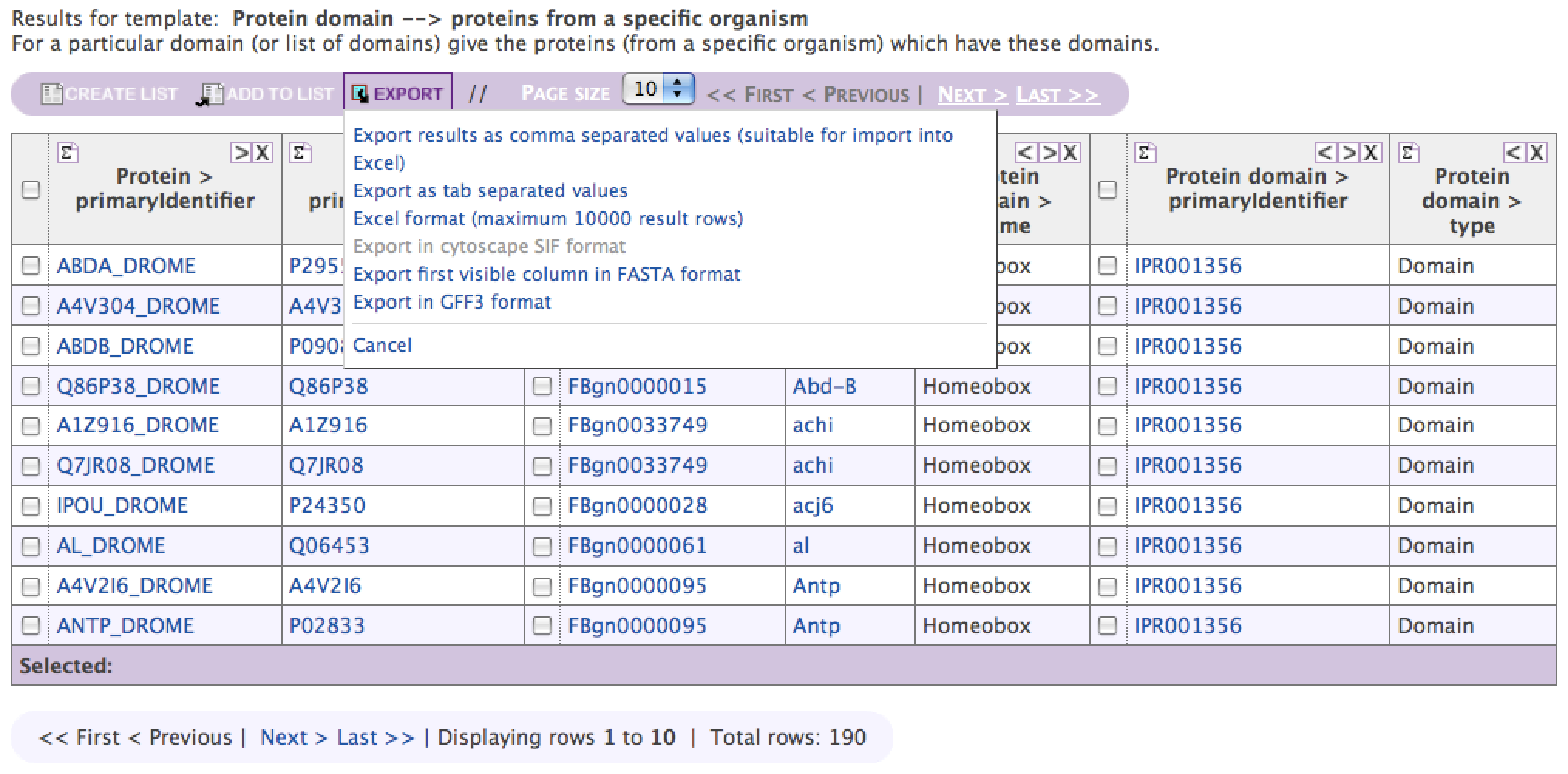
Results for template: **Protein domain --> proteins from a specific organism**
For a particular domain (or list of domains) give the proteins (from a specific organism) which have these domains.

□	Protein > primaryIdentifier	Protein > primaryAccession	□	Gene > primaryIdentifier	Gene > symbol	Protein domain > name	□	Protein domain > primaryIdentifier	Protein domain > type
□	ABDA_DROME	P29555	□	FBgn0000014	abd-A	Homeobox	□	IPR001356	Domain
□	A4V304_DROME	A4V304	□	FBgn0000015	Abd-B	Homeobox	□	IPR001356	Domain
□	ABDB_DROME	P09087	□	FBgn0000015	Abd-B	Homeobox	□	IPR001356	Domain
□	Q86P38_DROME	Q86P38	□	FBgn0000015	Abd-B	Homeobox	□	IPR001356	Domain
□	A1Z916_DROME	A1Z916	□	FBgn0033749	achi	Homeobox	□	IPR001356	Domain
□	Q7JR08_DROME	Q7JR08	□	FBgn0033749	achi	Homeobox	□	IPR001356	Domain
□	IPOU_DROME	P24350	□	FBgn0000028	acj6	Homeobox	□	IPR001356	Domain
□	AL_DROME	Q06453	□	FBgn0000061	al	Homeobox	□	IPR001356	Domain
□	A4V2I6_DROME	A4V2I6	□	FBgn0000095	Antp	Homeobox	□	IPR001356	Domain
□	ANTP_DROME	P02833	□	FBgn0000095	Antp	Homeobox	□	IPR001356	Domain

Selected:

<< First < Previous | Next > Last >> | Displaying rows **1** to **10** | Total rows: 190

Export



Results for template: **Protein domain --> proteins from a specific organism**
For a particular domain (or list of domains) give the proteins (from a specific organism) which have these domains.

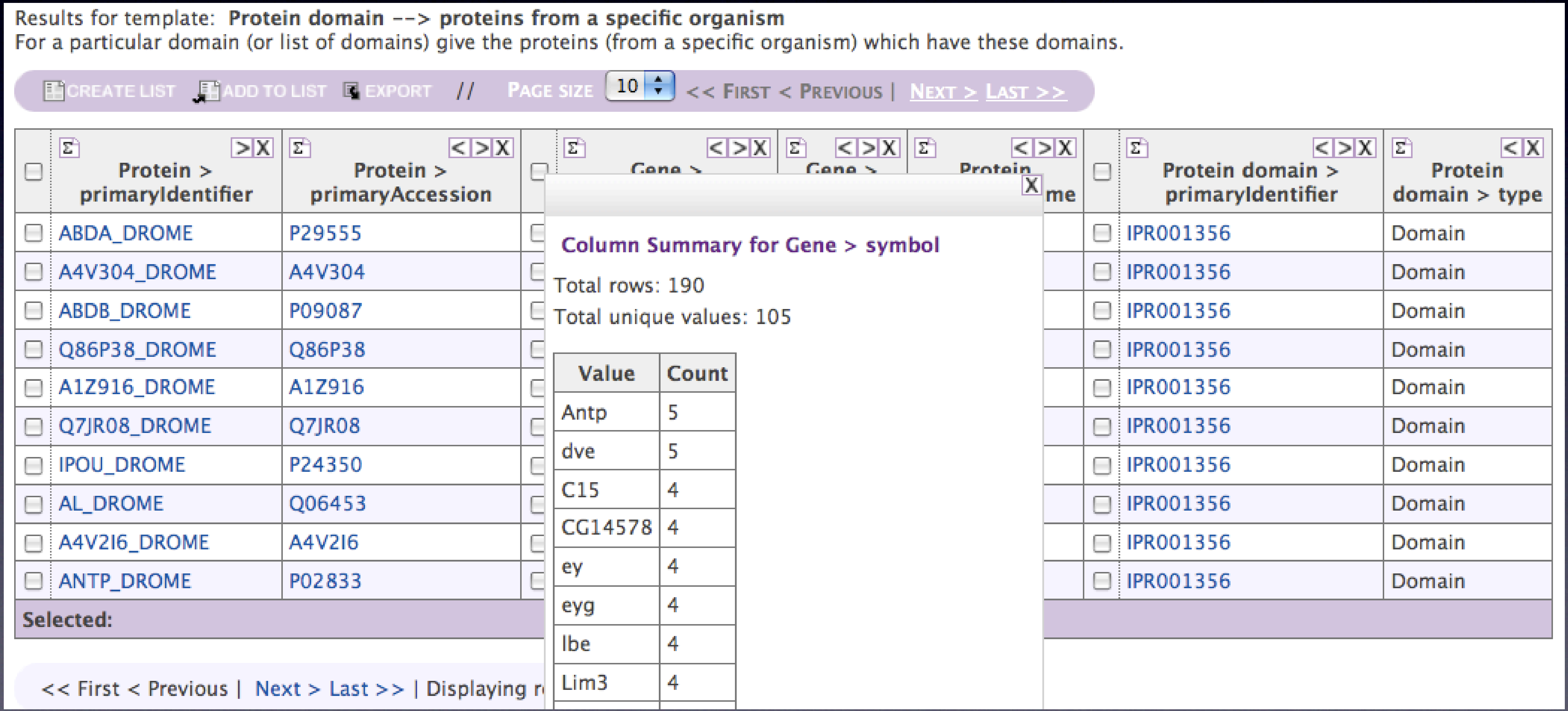
Export results as comma separated values (suitable for import into Excel)
Export as tab separated values
Excel format (maximum 10000 result rows)
Export in cytoscape SIF format
Export first visible column in FASTA format
Export in GFF3 format
Cancel

☐	Σ >X Protein > primaryIdentifier	Σ pri				<>X tein ain > me	☐	Σ <>X Protein domain > primaryIdentifier	Σ <X Protein domain > type
☐	ABDA_DROME	P295				box	☐	IPR001356	Domain
☐	A4V304_DROME	A4V3				box	☐	IPR001356	Domain
☐	ABDB_DROME	P090				box	☐	IPR001356	Domain
☐	Q86P38_DROME	Q86P38	☐	FBgn0000015	Abd-B	Homeobox	☐	IPR001356	Domain
☐	A1Z916_DROME	A1Z916	☐	FBgn0033749	achi	Homeobox	☐	IPR001356	Domain
☐	Q7JR08_DROME	Q7JR08	☐	FBgn0033749	achi	Homeobox	☐	IPR001356	Domain
☐	IPOU_DROME	P24350	☐	FBgn0000028	acj6	Homeobox	☐	IPR001356	Domain
☐	AL_DROME	Q06453	☐	FBgn0000061	al	Homeobox	☐	IPR001356	Domain
☐	A4V2I6_DROME	A4V2I6	☐	FBgn0000095	Antp	Homeobox	☐	IPR001356	Domain
☐	ANTP_DROME	P02833	☐	FBgn0000095	Antp	Homeobox	☐	IPR001356	Domain

Selected:

<< First < Previous | Next > Last >> | Displaying rows **1** to **10** | Total rows: 190

Column Summary



Results for template: **Protein domain --> proteins from a specific organism**
For a particular domain (or list of domains) give the proteins (from a specific organism) which have these domains.

	Protein > primaryIdentifier	Protein > primaryAccession		Protein domain > primaryIdentifier	Protein domain > type
□	ABDA_DROME	P29555	□	IPR001356	Domain
□	A4V304_DROME	A4V304	□	IPR001356	Domain
□	ABDB_DROME	P09087	□	IPR001356	Domain
□	Q86P38_DROME	Q86P38	□	IPR001356	Domain
□	A1Z916_DROME	A1Z916	□	IPR001356	Domain
□	Q7JR08_DROME	Q7JR08	□	IPR001356	Domain
□	IPOU_DROME	P24350	□	IPR001356	Domain
□	AL_DROME	Q06453	□	IPR001356	Domain
□	A4V2I6_DROME	A4V2I6	□	IPR001356	Domain
□	ANTP_DROME	P02833	□	IPR001356	Domain

Selected:

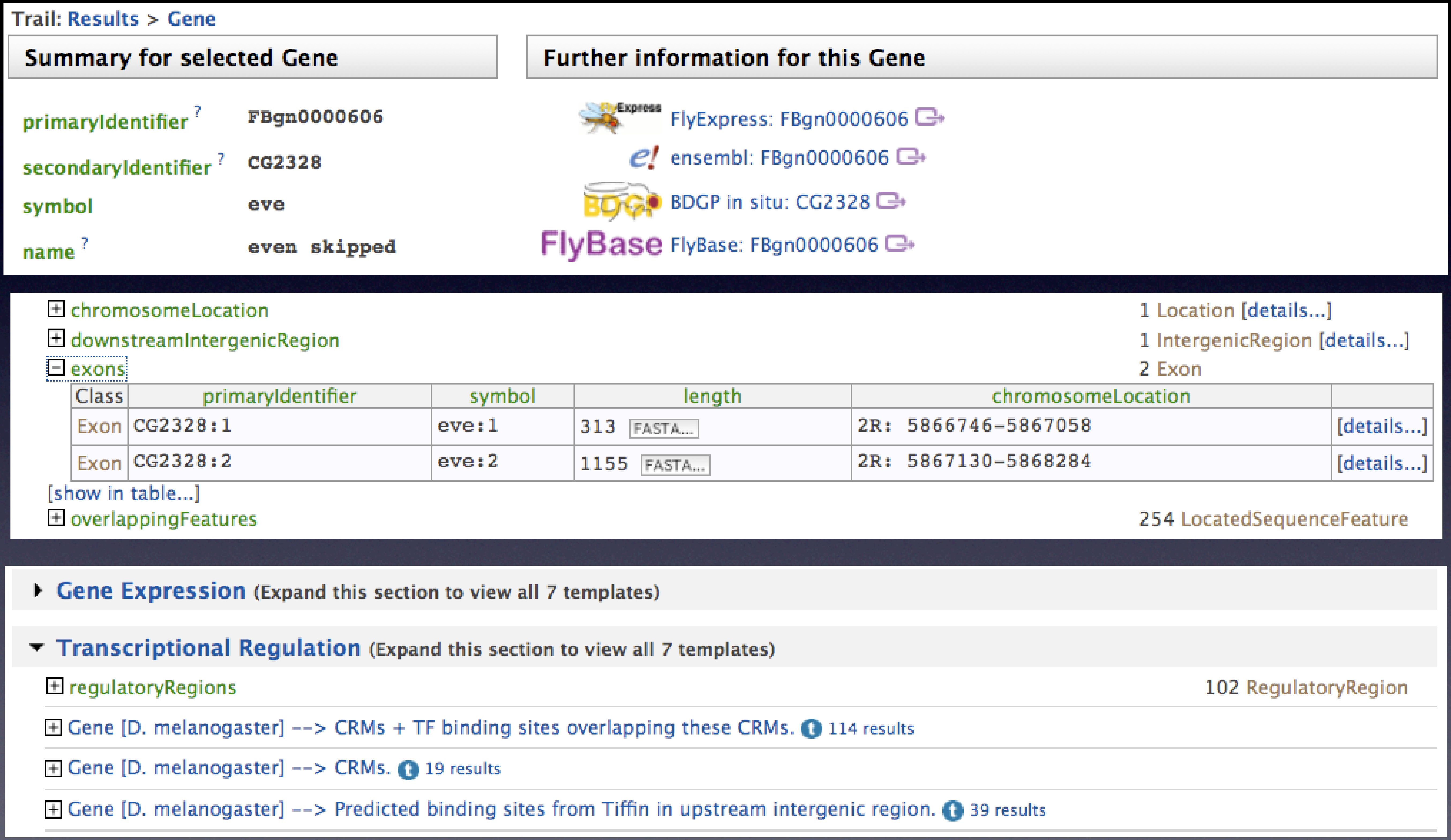
Column Summary for Gene > symbol

Total rows: 190

Total unique values: 105

Value	Count
Antp	5
dve	5
C15	4
CG14578	4
ey	4
eyg	4
lbe	4
Lim3	4

Reports



Trail: Results > Gene

Summary for selected Gene

primaryIdentifier ?	FBgn0000606
secondaryIdentifier ?	CG2328
symbol	eve
name ?	even skipped

Further information for this Gene

FlyExpress: FBgn0000606

ensembl: FBgn0000606

BDGP in situ: CG2328

FlyBase: FBgn0000606

⊞ chromosomeLocation — 1 Location [details...]

⊞ downstreamIntergenicRegion — 1 IntergenicRegion [details...]

⊟ exons — 2 Exon

Class	primaryIdentifier	symbol	length	chromosomeLocation	
Exon	CG2328:1	eve:1	313 FASTA...	2R: 5866746-5867058	[details...]
Exon	CG2328:2	eve:2	1155 FASTA...	2R: 5867130-5868284	[details...]

[show in table...]

⊞ overlappingFeatures — 254 LocatedSequenceFeature

▸ **Gene Expression** (Expand this section to view all 7 templates)

▾ **Transcriptional Regulation** (Expand this section to view all 7 templates)

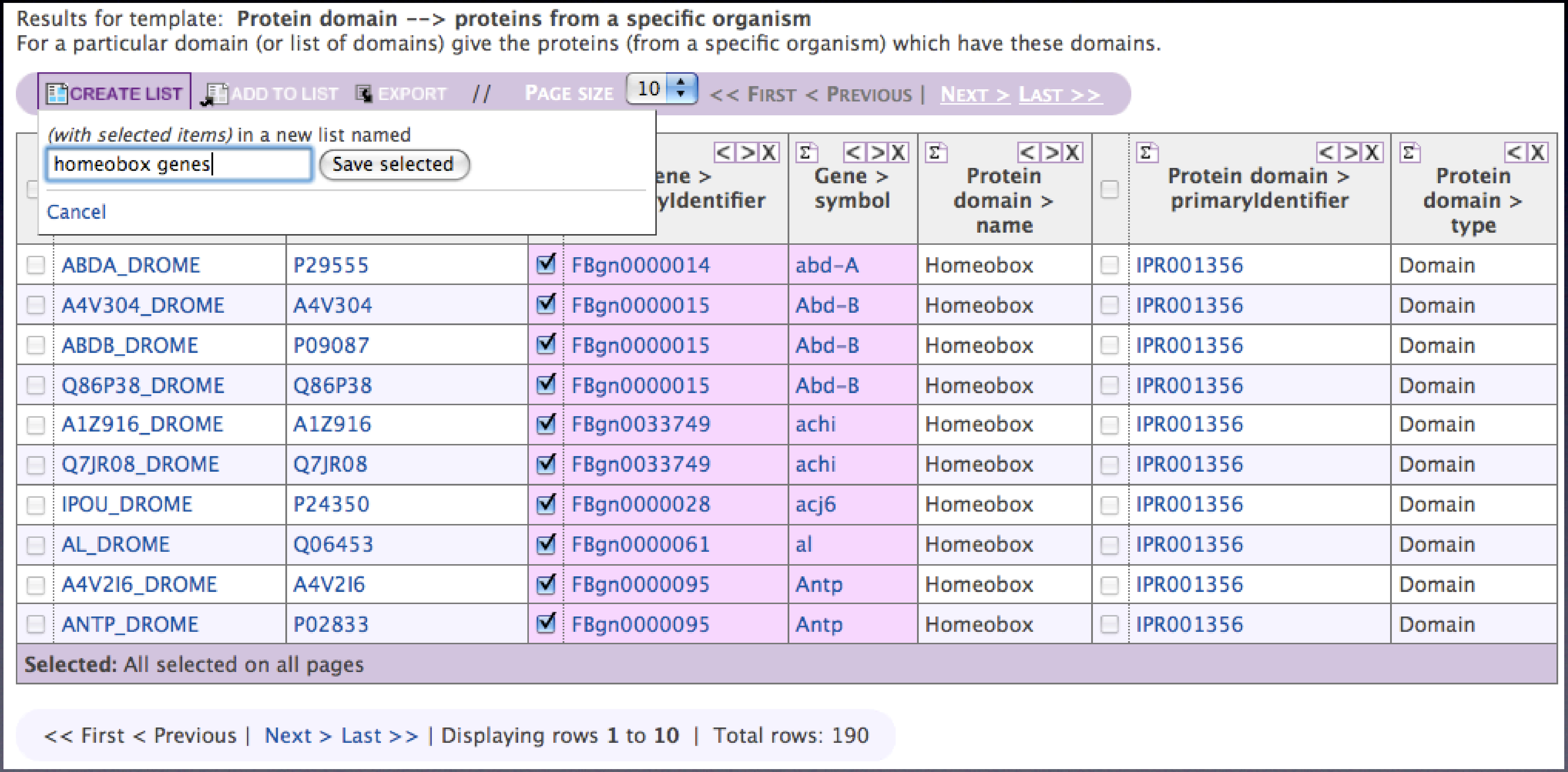
⊞ regulatoryRegions — 102 RegulatoryRegion

⊞ Gene [D. melanogaster] --> CRMs + TF binding sites overlapping these CRMs. 114 results

⊞ Gene [D. melanogaster] --> CRMs. 19 results

⊞ Gene [D. melanogaster] --> Predicted binding sites from Tiffin in upstream intergenic region. 39 results

Create list



Results for template: **Protein domain --> proteins from a specific organism**
For a particular domain (or list of domains) give the proteins (from a specific organism) which have these domains.

PAGE SIZE: 10

(with selected items) in a new list named

homeobox genes

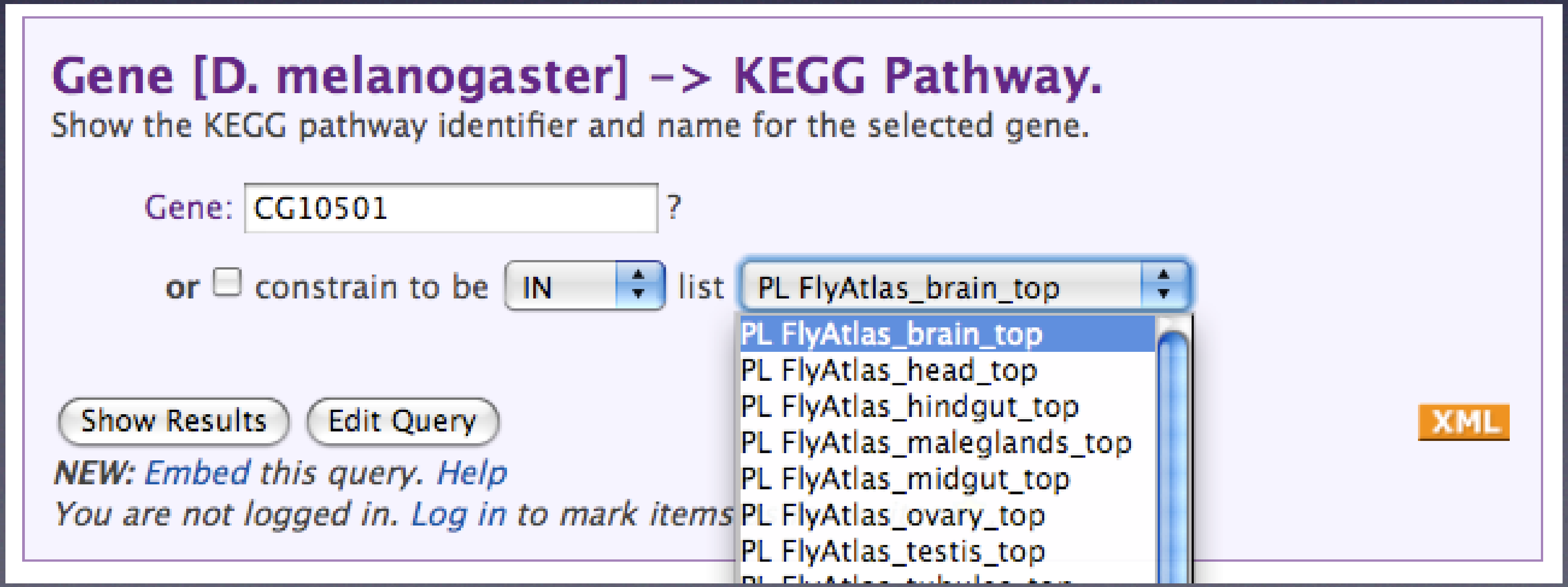
				ene > yIdentifier	Gene > symbol	Protein domain > name	□	Protein domain > primaryIdentifier	Protein domain > type
□	ABDA_DROME	P29555	☑	FBgn0000014	abd-A	Homeobox	□	IPR001356	Domain
□	A4V304_DROME	A4V304	☑	FBgn0000015	Abd-B	Homeobox	□	IPR001356	Domain
□	ABDB_DROME	P09087	☑	FBgn0000015	Abd-B	Homeobox	□	IPR001356	Domain
□	Q86P38_DROME	Q86P38	☑	FBgn0000015	Abd-B	Homeobox	□	IPR001356	Domain
□	A1Z916_DROME	A1Z916	☑	FBgn0033749	achi	Homeobox	□	IPR001356	Domain
□	Q7JR08_DROME	Q7JR08	☑	FBgn0033749	achi	Homeobox	□	IPR001356	Domain
□	IPOU_DROME	P24350	☑	FBgn0000028	acj6	Homeobox	□	IPR001356	Domain
□	AL_DROME	Q06453	☑	FBgn0000061	al	Homeobox	□	IPR001356	Domain
□	A4V2I6_DROME	A4V2I6	☑	FBgn0000095	Antp	Homeobox	□	IPR001356	Domain
□	ANTP_DROME	P02833	☑	FBgn0000095	Antp	Homeobox	□	IPR001356	Domain

Selected: All selected on all pages

<< First < Previous | Next > Last >> | Displaying rows **1** to **10** | Total rows: 190

Lists

- Lists of any type - e.g. *genes, protein domains, organisms*
- Use a list in any query
- Save lists from results pages
- Upload identifiers



Gene [D. melanogaster] -> KEGG Pathway.

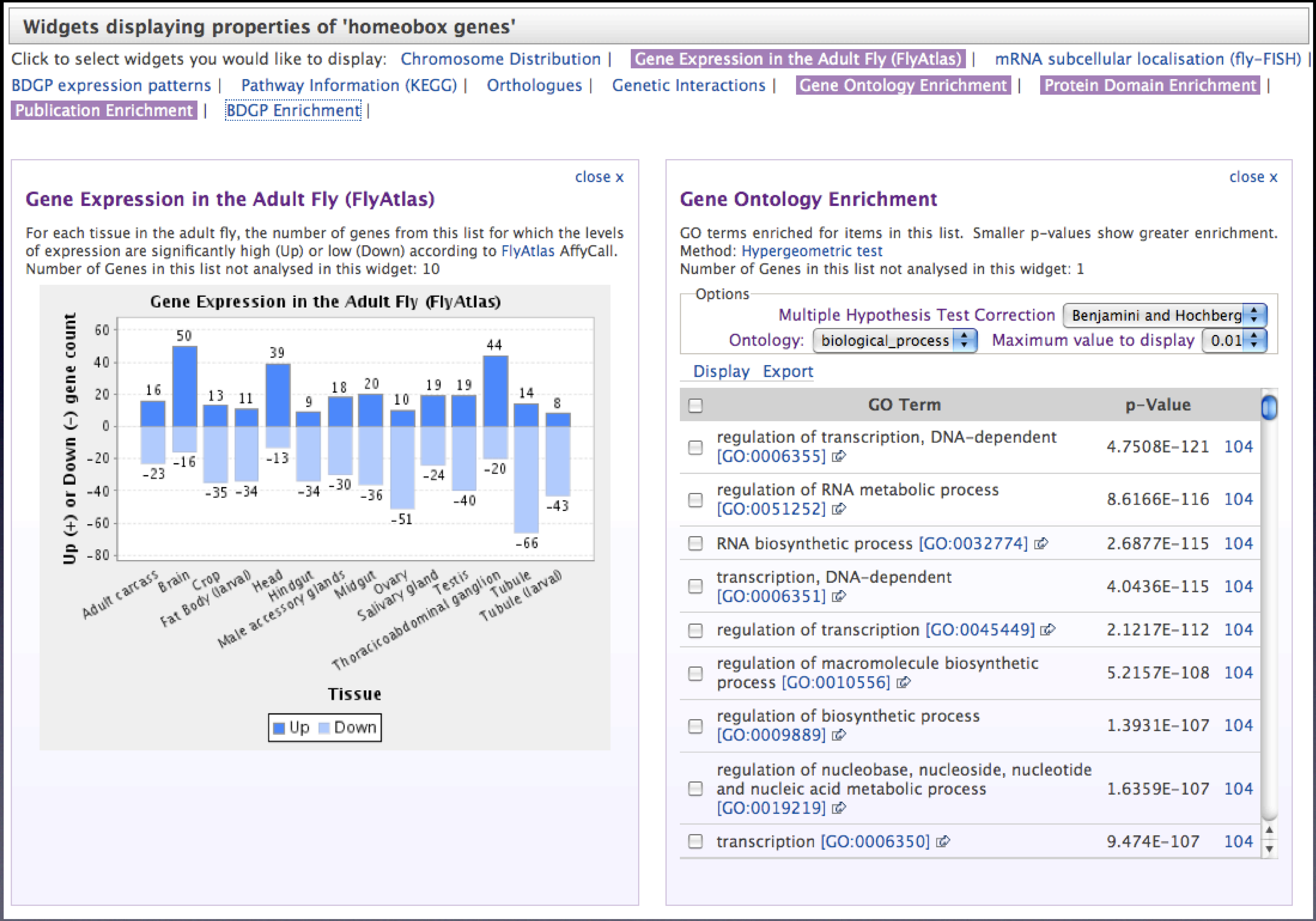
Show the KEGG pathway identifier and name for the selected gene.

Gene: CG10501 ?

or ☐ constrain to be IN list PL FlyAtlas_brain_top

NEW: Embed this query. Help

You are not logged in. Log in to mark items



Widgets displaying properties of 'homeobox genes'

Click to select widgets you would like to display: Chromosome Distribution | Gene Expression in the Adult Fly (FlyAtlas) | mRNA subcellular localisation (fly-FISH) | BDGP expression patterns | Pathway Information (KEGG) | Orthologues | Genetic Interactions | Gene Ontology Enrichment | Protein Domain Enrichment | Publication Enrichment | BDGP Enrichment |

close x

Gene Expression in the Adult Fly (FlyAtlas)

For each tissue in the adult fly, the number of genes from this list for which the levels of expression are significantly high (Up) or low (Down) according to FlyAtlas AffyCall.
Number of Genes in this list not analysed in this widget: 10

close x

Gene Ontology Enrichment

GO terms enriched for items in this list. Smaller p-values show greater enrichment.
Method: Hypergeometric test
Number of Genes in this list not analysed in this widget: 1

Options

Multiple Hypothesis Test Correction: Benjamini and Hochberg

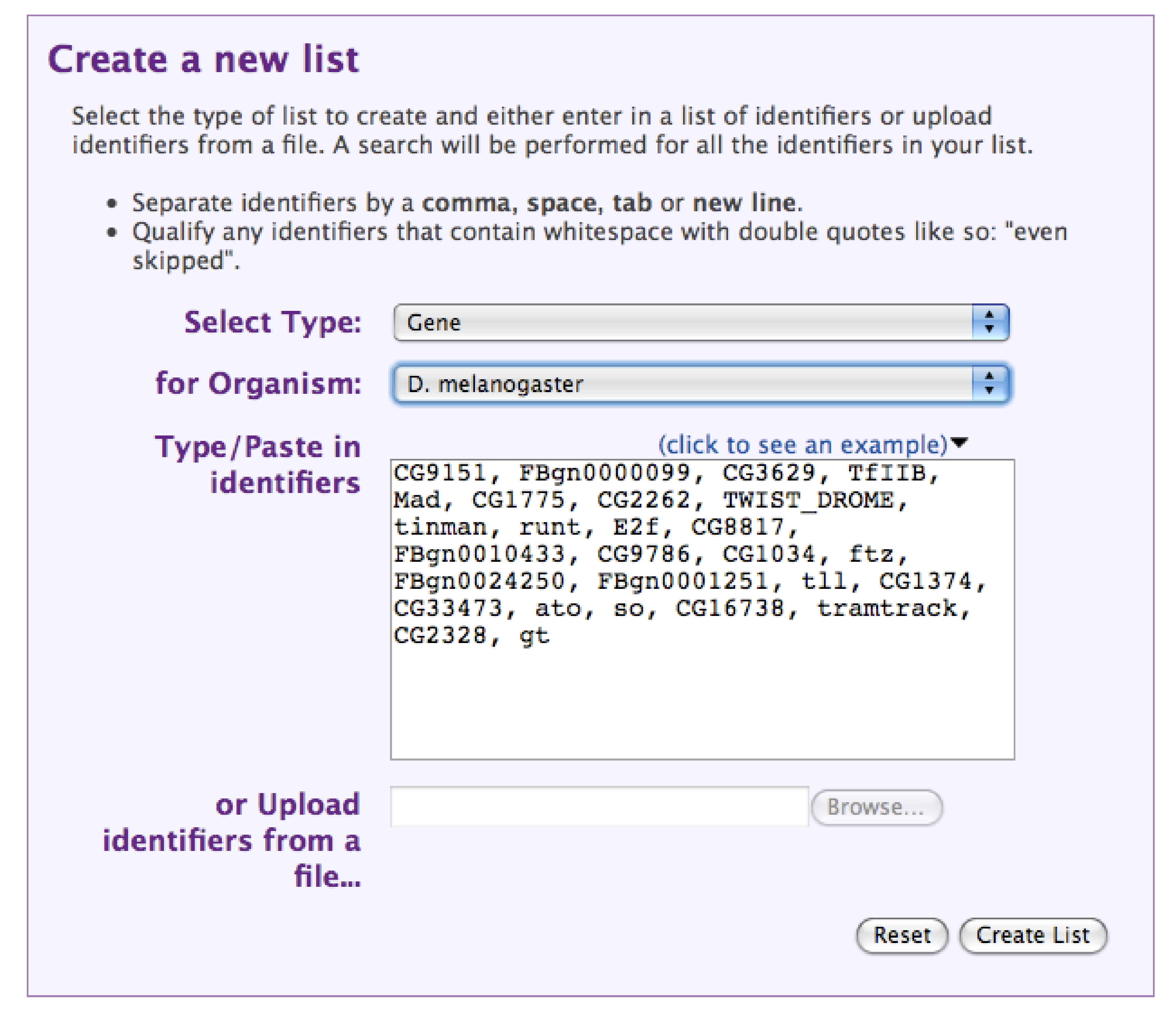
Ontology: biological_process

Maximum value to display: 0.01

Display Export

□	GO Term	p-Value	
□	regulation of transcription, DNA-dependent [GO:0006355]	4.7508E-121	104
□	regulation of RNA metabolic process [GO:0051252]	8.6166E-116	104
□	RNA biosynthetic process [GO:0032774]	2.6877E-115	104
□	transcription, DNA-dependent [GO:0006351]	4.0436E-115	104
□	regulation of transcription [GO:0045449]	2.1217E-112	104
□	regulation of macromolecule biosynthetic process [GO:0010556]	5.2157E-108	104
□	regulation of biosynthetic process [GO:0009889]	1.3931E-107	104
□	regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process [GO:0019219]	1.6359E-107	104
□	transcription [GO:0006350]	9.474E-107	104

List Upload



Create a new list

Select the type of list to create and either enter in a list of identifiers or upload identifiers from a file. A search will be performed for all the identifiers in your list.

- Separate identifiers by a **comma**, **space**, **tab** or **new line**.
- Qualify any identifiers that contain whitespace with double quotes like so: "even skipped".

Select Type: Gene

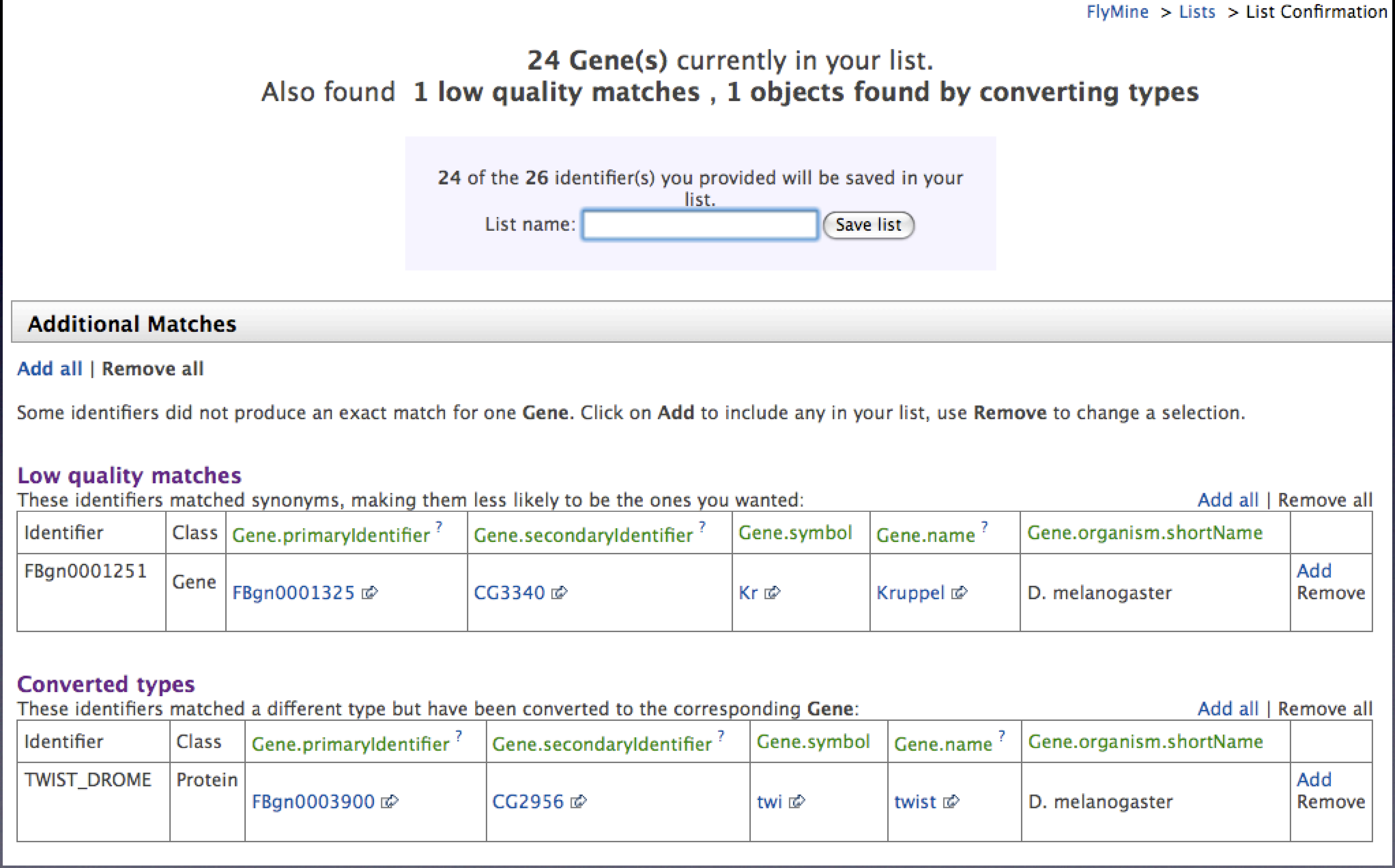
for Organism: D. melanogaster

Type/Paste in identifiers

```
CG9151, FBgn0000099, CG3629, TfIIB,
Mad, CG1775, CG2262, TWIST_DROME,
tinman, runt, E2f, CG8817,
FBgn0010433, CG9786, CG1034, ftz,
FBgn0024250, FBgn0001251, tll, CG1374,
CG33473, ato, so, CG16738, tramtrack,
CG2328, gt
```

or Upload identifiers from a file...

List Upload



FlyMine > Lists > List Confirmation

24 Gene(s) currently in your list.
Also found **1 low quality matches , 1 objects found by converting types**

24 of the **26** identifier(s) you provided will be saved in your list.

List name: ______ Save list

Additional Matches

Add all | **Remove all**

Some identifiers did not produce an exact match for one **Gene**. Click on **Add** to include any in your list, use **Remove** to change a selection.

Low quality matches

These identifiers matched synonyms, making them less likely to be the ones you wanted:

Add all | Remove all

Identifier	Class	Gene.primaryIdentifier ?	Gene.secondaryIdentifier ?	Gene.symbol	Gene.name ?	Gene.organism.shortName	
FBgn0001251	Gene	FBgn0001325	CG3340	Kr	Kruppel	D. melanogaster	Add Remove

Converted types

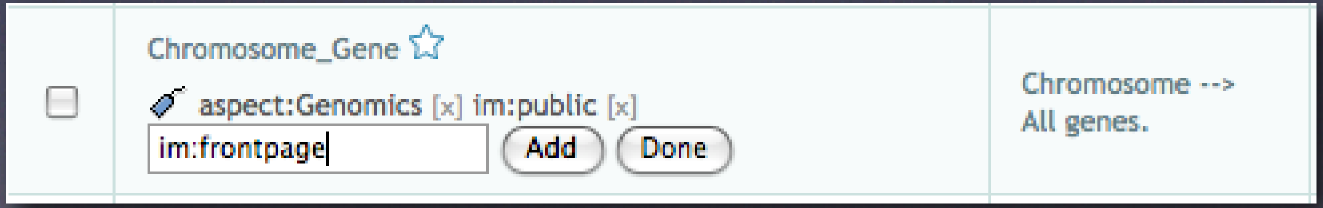
These identifiers matched a different type but have been converted to the corresponding **Gene**:

Add all | Remove all

Identifier	Class	Gene.primaryIdentifier ?	Gene.secondaryIdentifier ?	Gene.symbol	Gene.name ?	Gene.organism.shortName	
TWIST_DROME	Protein	FBgn0003900	CG2956	twi	twist	D. melanogaster	Add Remove

Superuser

- Non-programmer can configure
- Public template queries
- Public lists
- Templates on report pages
- Tagging



□ Chromosome_Gene ☆

aspect:Genomics [x] im:public [x]

im:frontpage

Chromosome -->
All genes.