InterMine

Open source data warehouse and web interface

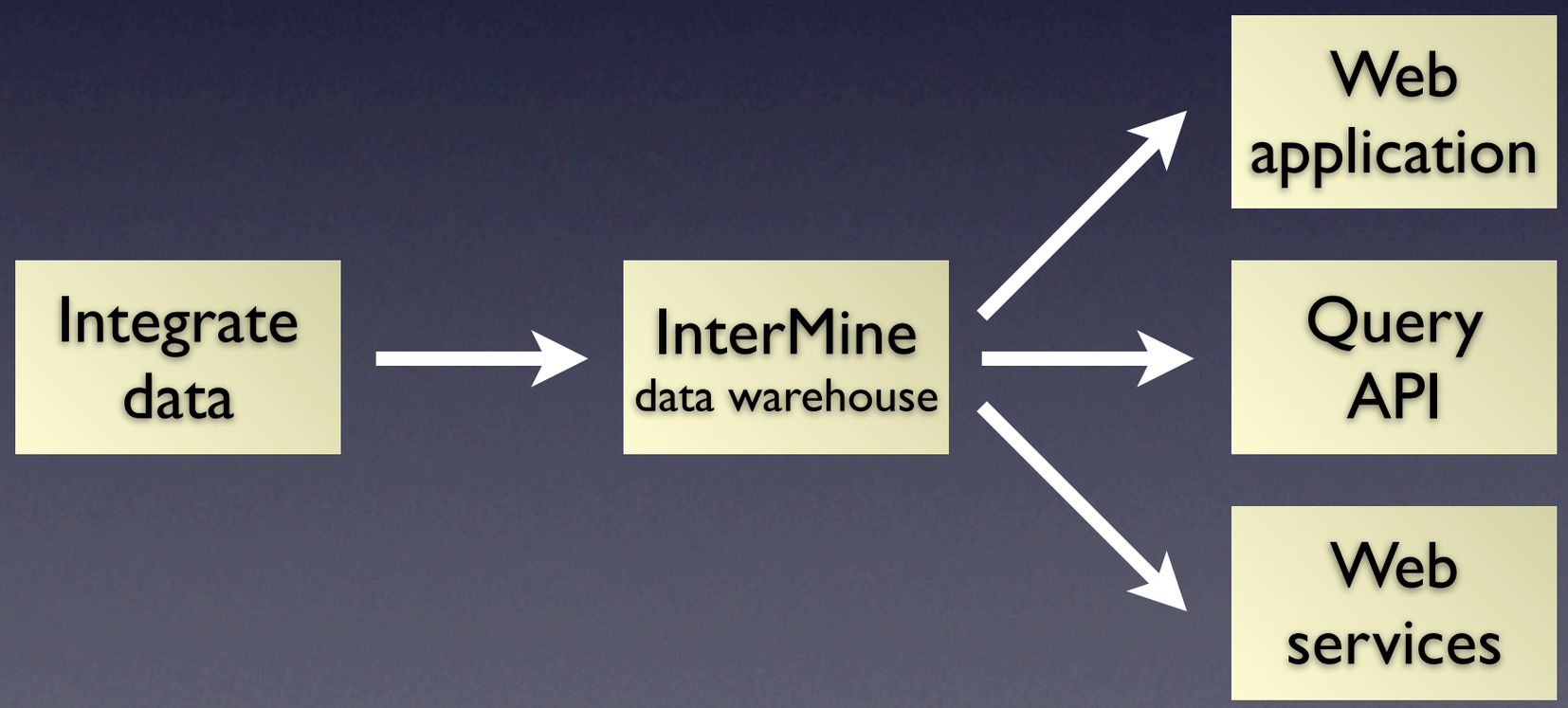
Richard Smith
University of Cambridge

Demo: *Monday 12:15 room 701B*
Poster: *E34 (Monday)*

www.intermine.org

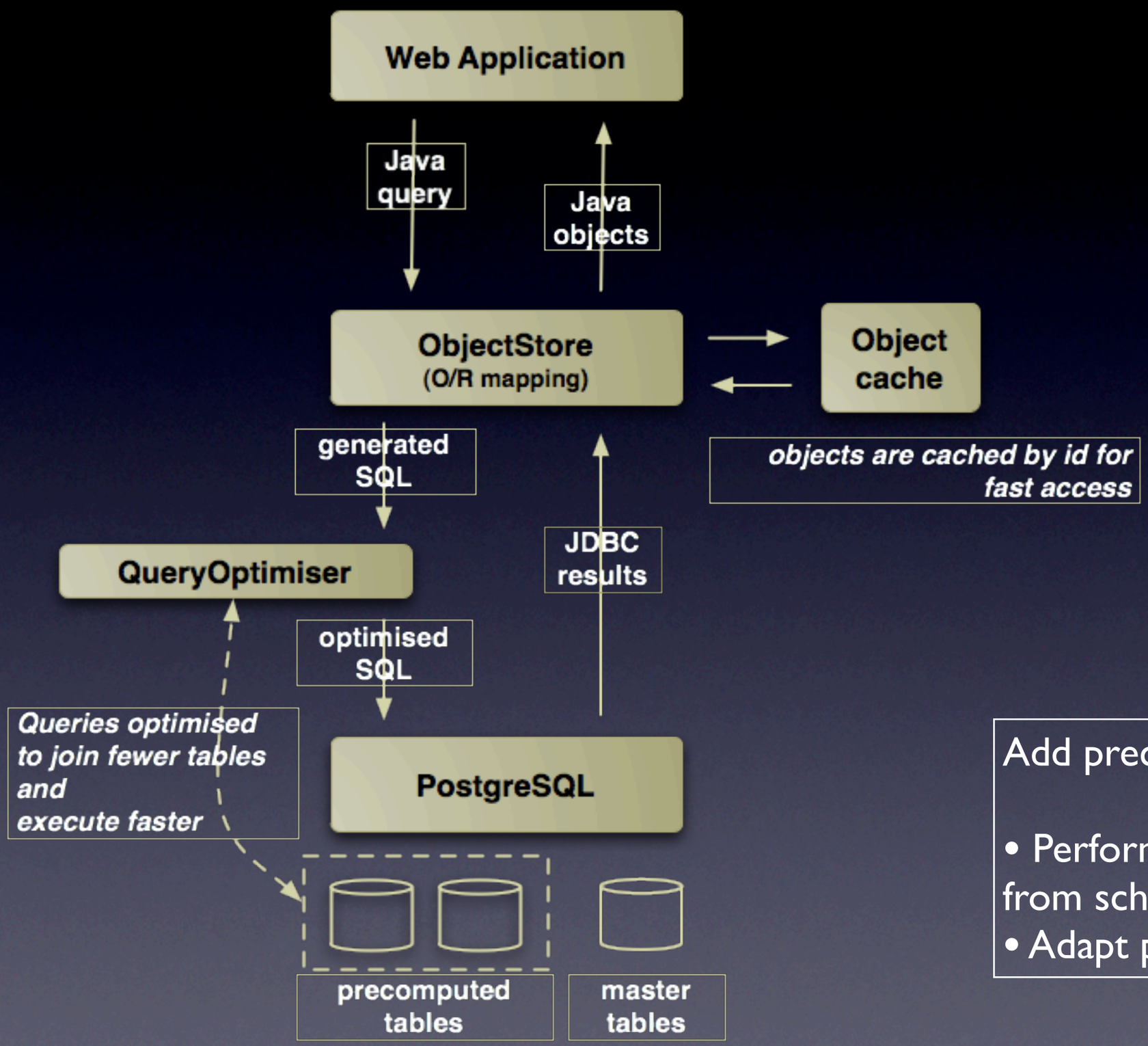
Overview

- Query-optimised data warehouse system
- Java, object-based data model
- 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*



Add precomputed tables at any time:

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

Data Integration

Existing data sources

UniProt XML | e! | Chado

OMIM disease | OBO ontology | PSI XML

Inparanoid | KEGG (Kyoto Encyclopedia of Genes and Genomes) | GO (the Gene Ontology)

+

Custom Data Sources

FASTA | GFF3 | XML

Java and Perl APIs

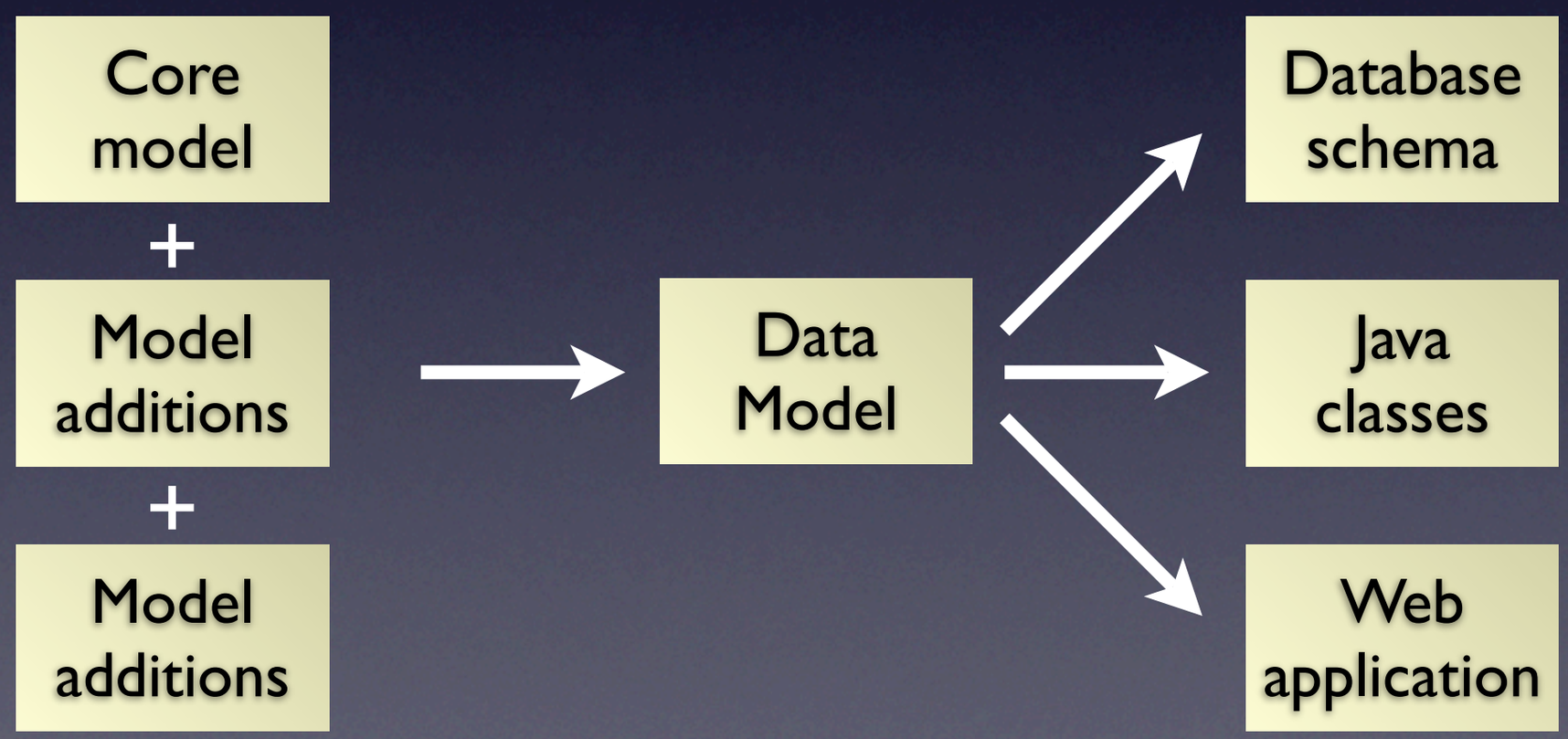
Configurable data integration

→

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:

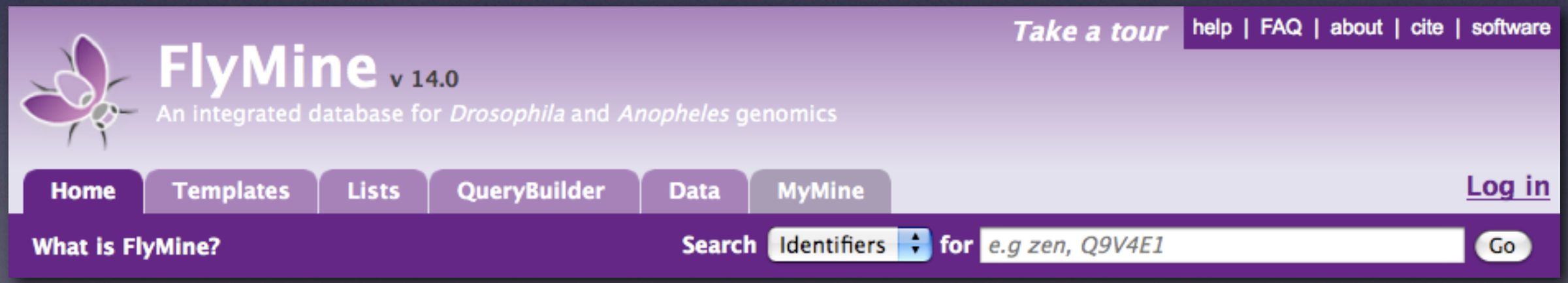
```
<class name="Pathway" is-interface="true" >
  <attribute name="name" type="String"/>
  <collection name="genes" referenced-type="Gene"
              reverse-reference="pathways"/>
</class>
<class name="Gene" is-interface="true">
  <collection name="pathways" referenced-type="Pathway"
           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

Configure a new Mine

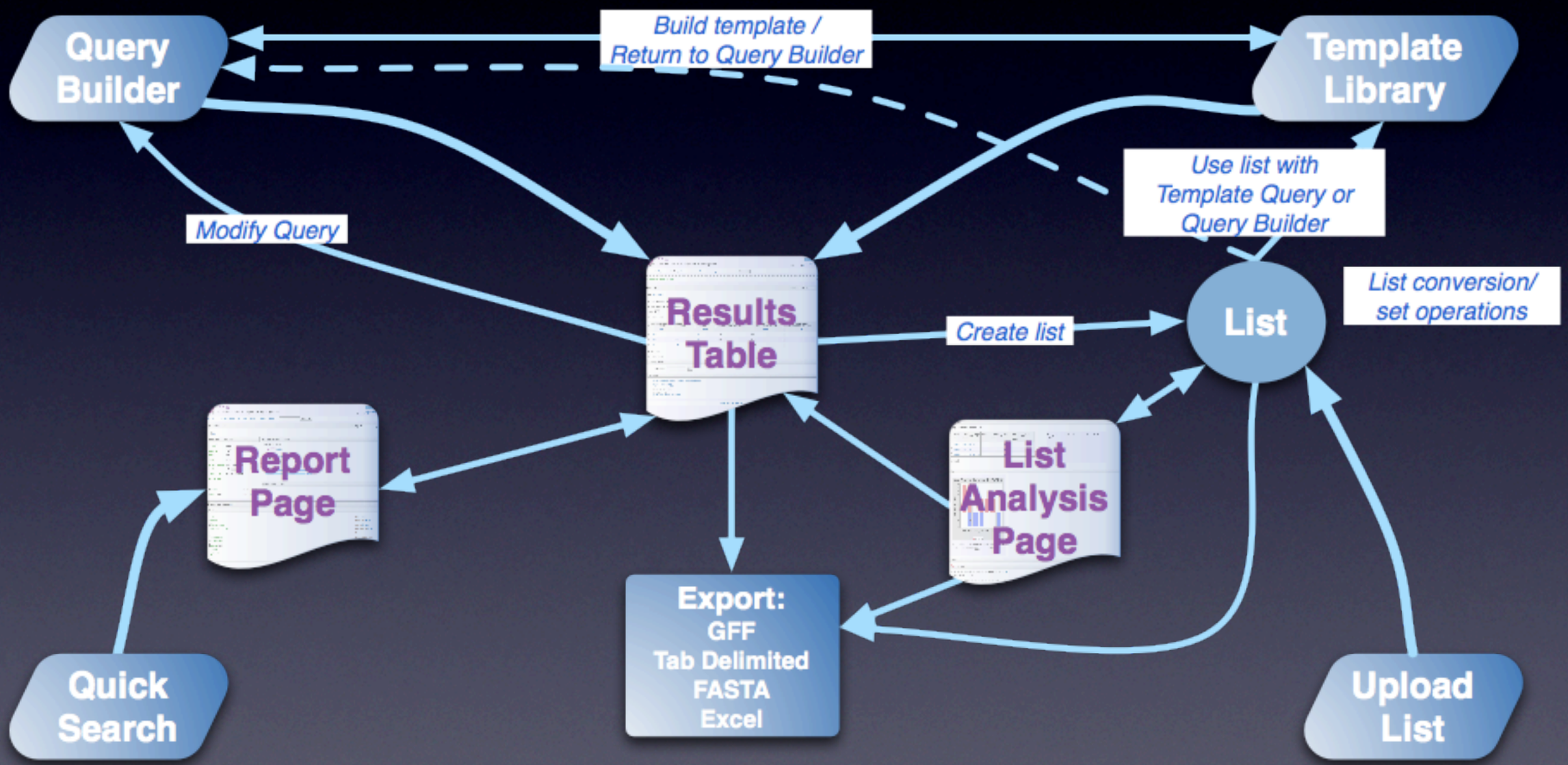
```
<project type="bio">
...
<sources>
  <source name="uniprot" type="uniprot" dump="true">
    <property name="uniprot.organisms"
              value="7227 6239"/>
    <property name="src.data.dir"
              location="/data/uniprot"/>
  </source>
  <source name="my-source" type="pathways">
    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

Web Application

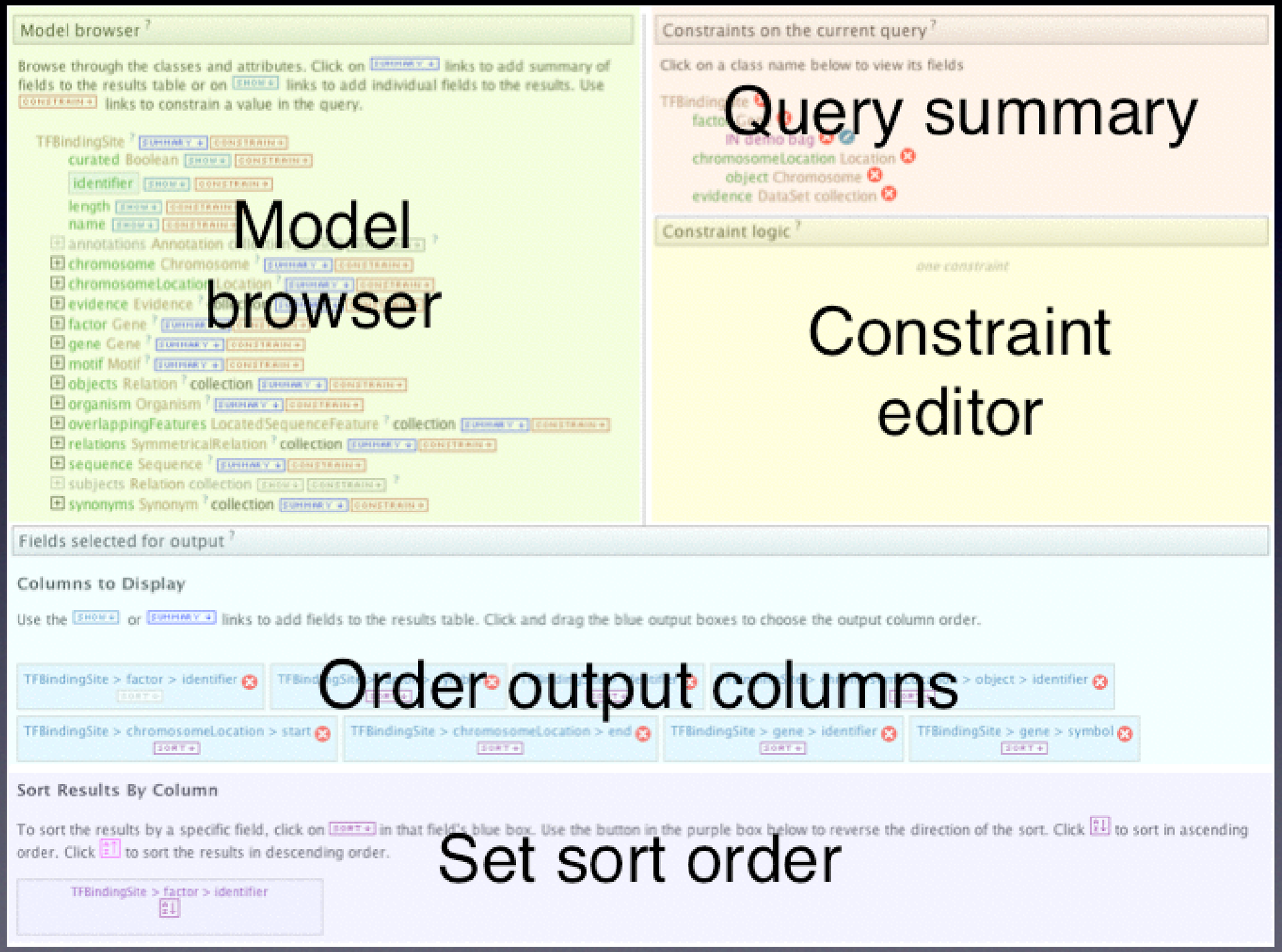
- Works for any data model
- Advanced functionality for bench biologist
- Highly configurable
- Configuration from within web interface



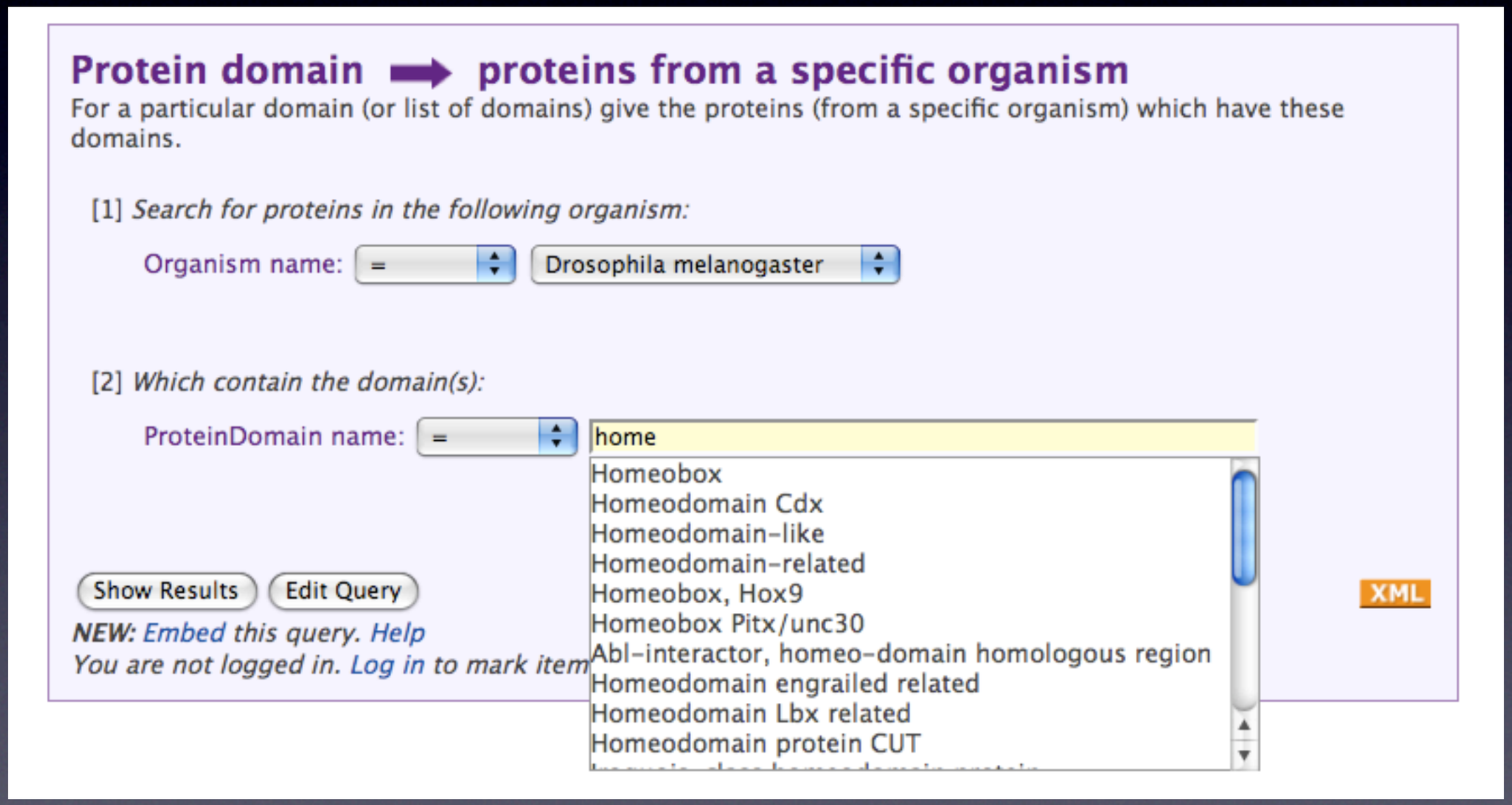
Webapp Overview



QueryBuilder



Template Queries



Template Library

Home | Templates | Lists | QueryBuilder | Data | MyMine | Log in

Search Identifiers for e.g zen, Q9V4E1 Go

FlyMine > Templates

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 -- Reset Actions: Export selected

☑ 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.

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

Protein > primaryIdentifier	Protein > primaryAccession	Gene > primaryIdentifier	Gene > symbol	Protein domain > name	Protein domain > primaryIdentifier	Protein domain > type
Q9Y1P6_DROME	Q9Y1P6	FBgn0027364	Six4	Homeobox	IPR001356	Domain
Q9Y0Z9_DROME	Q9Y0Z9	FBgn0003896	tup	Homeobox	IPR001356	Domain
Q9XZU0_DROME	Q9XZU0	FBgn0023489	Pph13	Homeobox	IPR001356	Domain
Q9XZC2_DROME	Q9XZC2	FBgn0019650	toy	Homeobox	IPR001356	Domain
Q9W4B3_DROME	Q9W4B3	FBgn0053980	CG33980	Homeobox	IPR001356	Domain
Q9W4B2_DROME	Q9W4B2	FBgn0029775	CG4136	Homeobox	IPR001356	Domain
Q9W423_DROME	Q9W423	FBgn0040918	Lag1	Homeobox	IPR001356	Domain
Q9W3C6_DROME	Q9W3C6	FBgn0030058	CG11294	Homeobox	IPR001356	Domain
RX_DROME	Q9W2Q1	FBgn0020617	Rx	Homeobox	IPR001356	Domain
Q9W2P8_DROME	Q9W2P8	FBgn0008636	hbn	Homeobox	IPR001356	Domain

Selected:

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

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

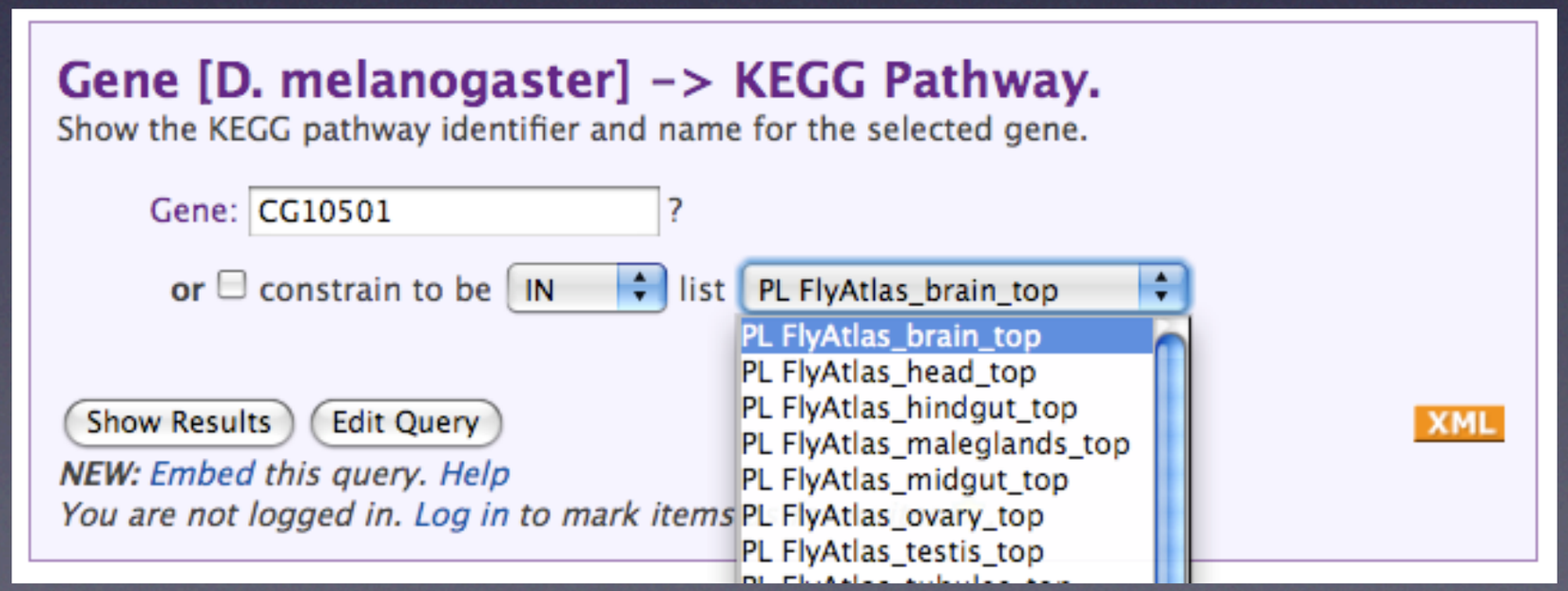
Gene [D. melanogaster] --> CRMs + TF binding sites overlapping these CRMs. 114 results

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

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

Lists

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



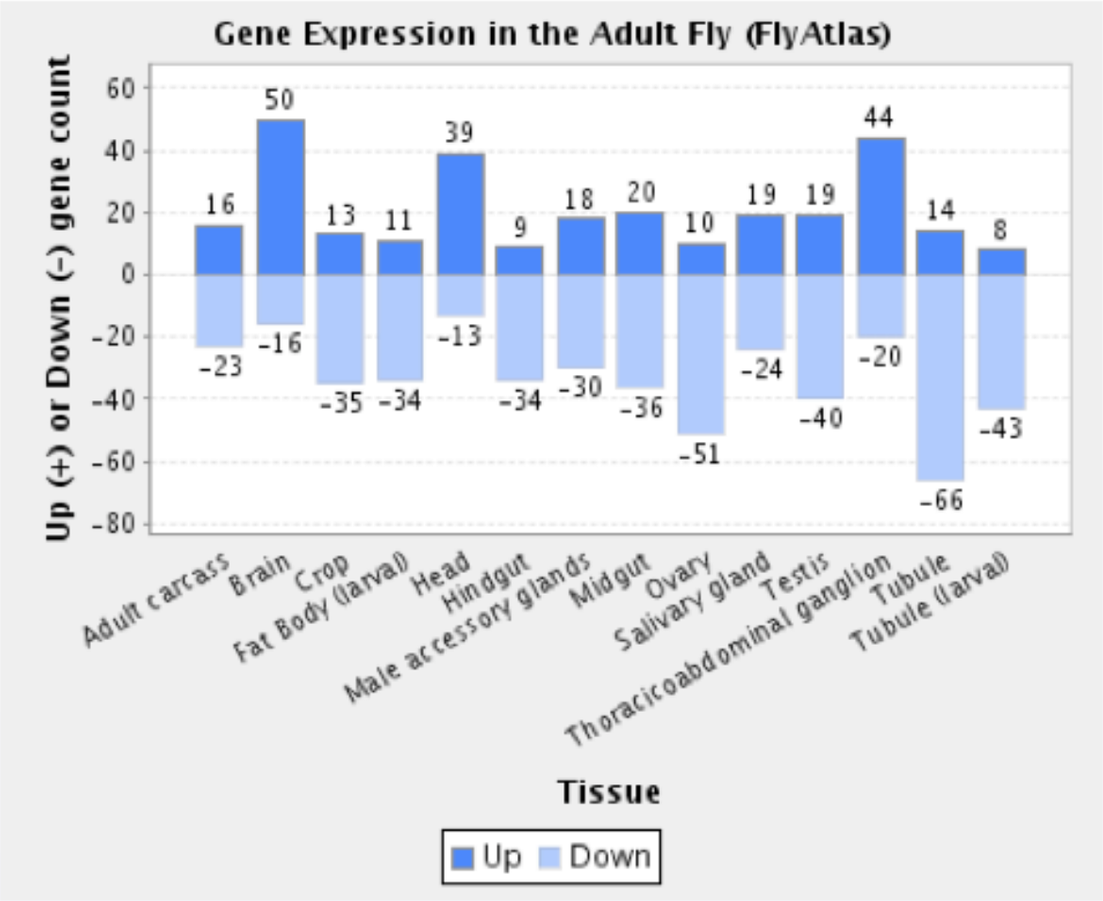
Widgets displaying properties of 'homeobox genes'

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

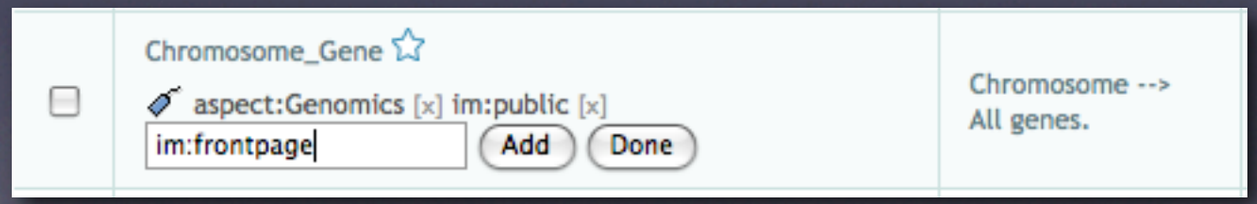
Ontology: biological_process Maximum value to display: 0.01

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

Superuser

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



Query API

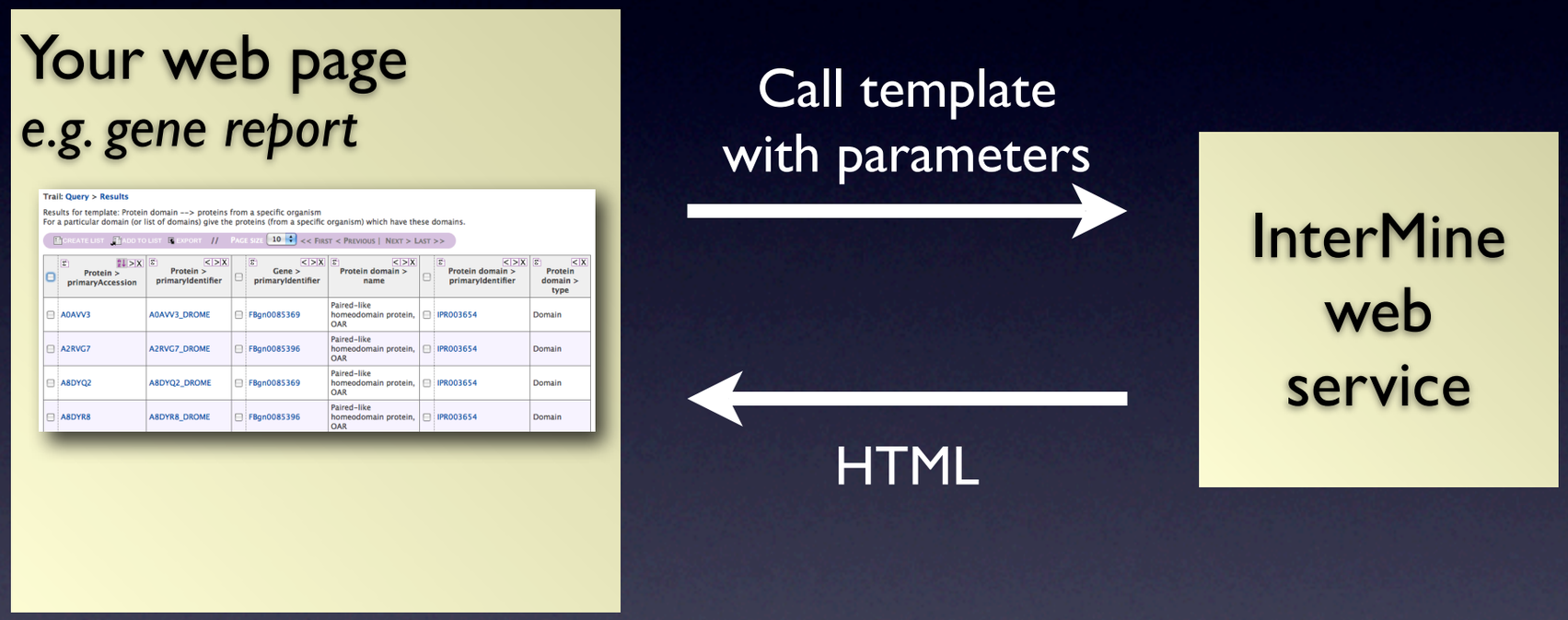
```
PathQuery q = new PathQuery(model);

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



- '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