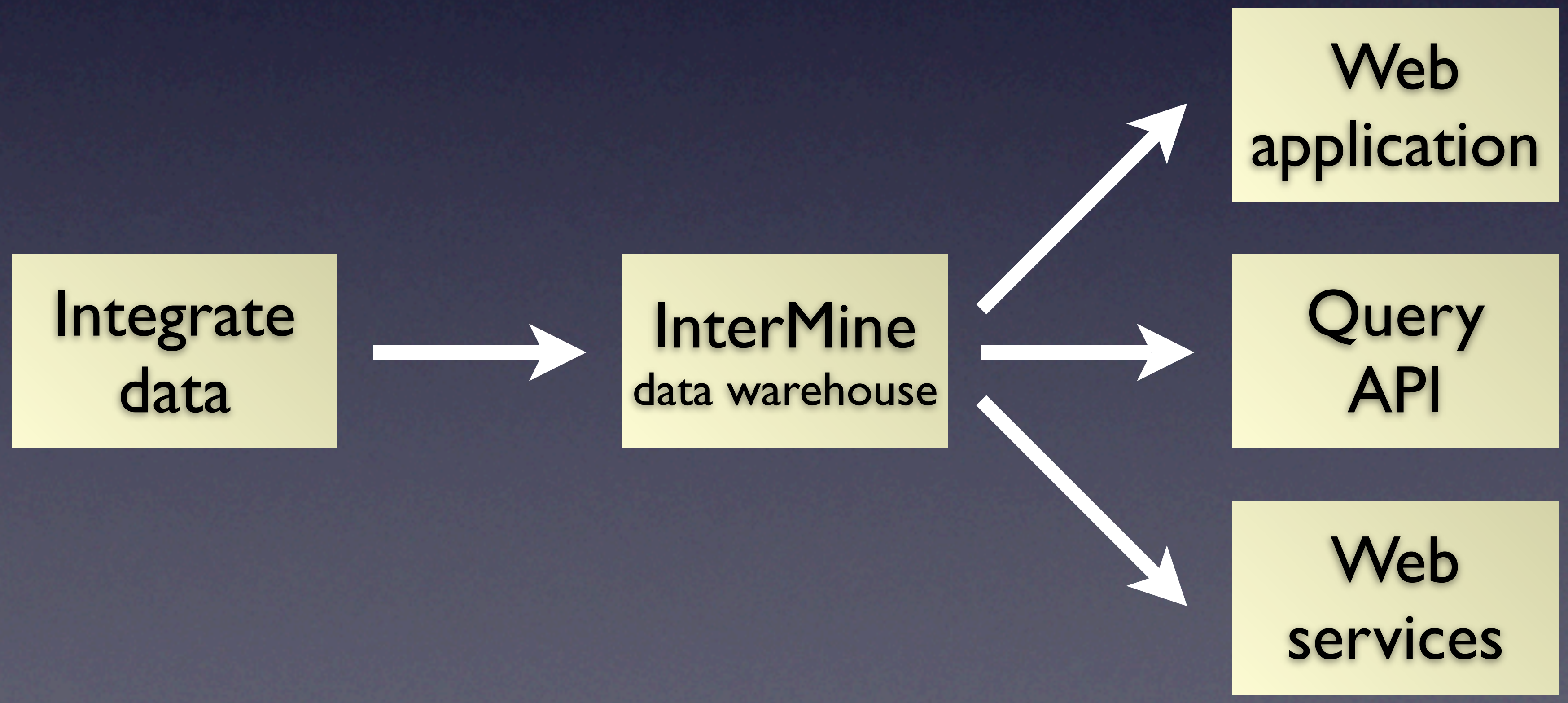
New Mines and New Features

Richard Smith
richard@flymine.org

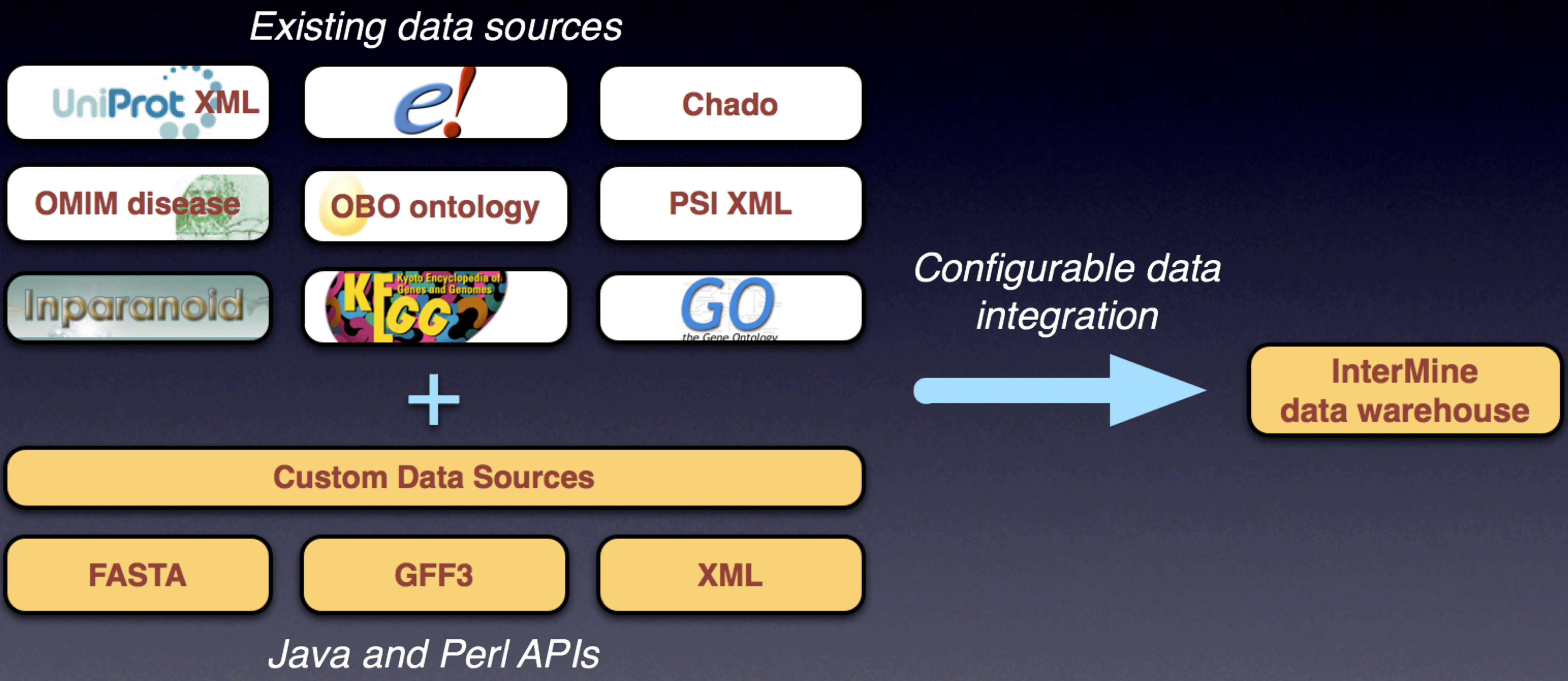
InterMine
www.intermine.org

Overview

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



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

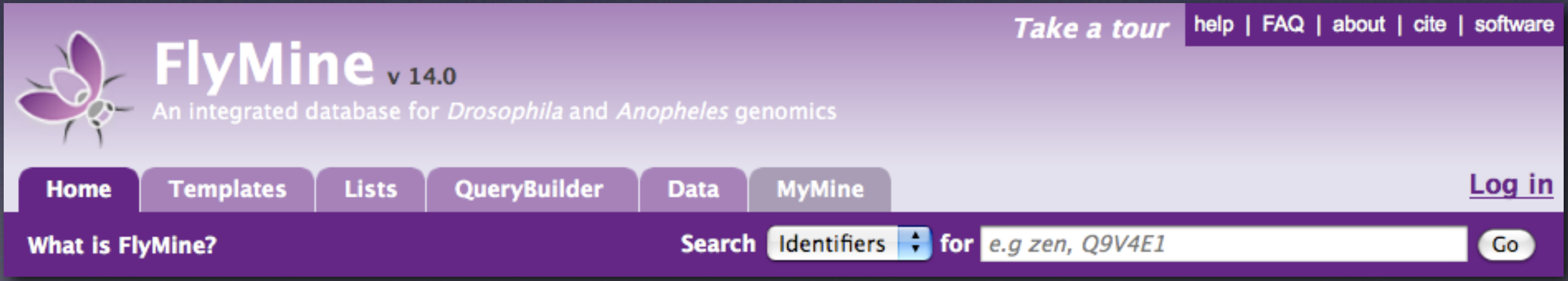
Java and Perl APIs

Configurable data integration

InterMine data warehouse

Web Application

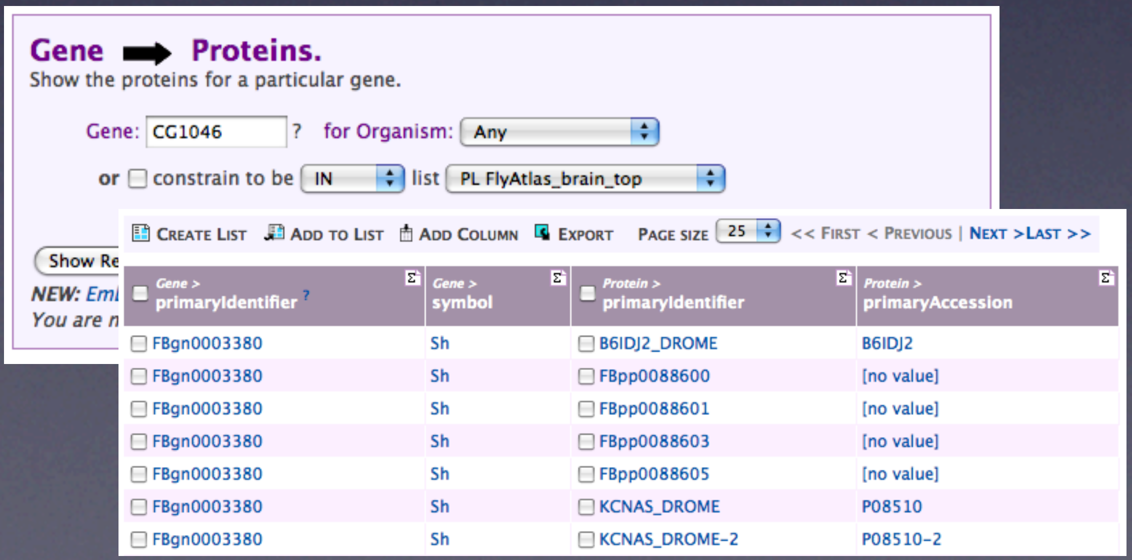
- Works for any data model
- Advanced functionality for bench biologist
- Highly configurable
- Configuration from within web interface



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

Web Interface

- **QueryBuilder** - custom queries, advanced
- **Template queries** - pre-defined queries
- **Report pages** - configurable, templates
- **Lists** - upload, use in queries, analysis widgets
- **Export & API**
- **MyMine**

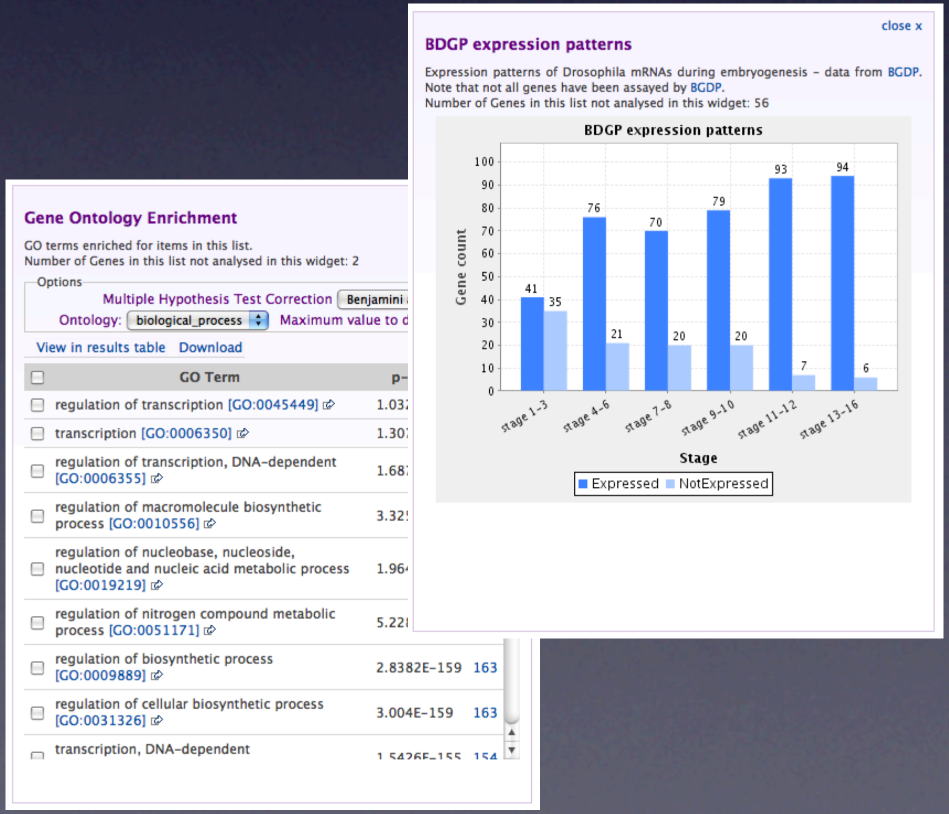


Gene ➡ Proteins.
Show the proteins for a particular gene.

Gene: CG1046 ? for Organism: Any

or □ constrain to be IN list PL FlyAtlas_brain_top

□ Gene > primaryIdentifier ?	Gene > symbol	□ Protein > primaryIdentifier	Protein > primaryAccession
□ FBgn0003380	Sh	□ B6IDJ2_DROME	B6IDJ2
□ FBgn0003380	Sh	□ FBpp0088600	[no value]
□ FBgn0003380	Sh	□ FBpp0088601	[no value]
□ FBgn0003380	Sh	□ FBpp0088603	[no value]
□ FBgn0003380	Sh	□ FBpp0088605	[no value]
□ FBgn0003380	Sh	□ KCNAS_DROME	P08510
□ FBgn0003380	Sh	□ KCNAS_DROME-2	P08510-2



Gene Ontology Enrichment

GO terms enriched for items in this list.
Number of Genes in this list not analysed in this widget: 2

Options
Multiple Hypothesis Test Correction: Benjamini
Ontology: biological_process
Maximum value to d

□ GO Term	p-	
□ regulation of transcription [GO:0045449]	1.03	
□ transcription [GO:0006350]	1.30	
□ regulation of transcription, DNA-dependent [GO:0006355]	1.68	
□ regulation of macromolecule biosynthetic process [GO:0010556]	3.32	
□ regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process [GO:0019219]	1.96	
□ regulation of nitrogen compound metabolic process [GO:0051171]	5.22	
□ regulation of biosynthetic process [GO:0009889]	2.8382E-159	163
□ regulation of cellular biosynthetic process [GO:0031326]	3.004E-159	163
□ transcription, DNA-dependent	1.5426E-155	154

BDGP expression patterns

Expression patterns of Drosophila mRNAs during embryogenesis - data from BGDP.
Note that not all genes have been assayed by BGDP.
Number of Genes in this list not analysed in this widget: 56

Mines for MODs

FlyMine

Yeast

Zebrafish

Rat

0.5 FTE each
started at different times

Mines for MODs

FlyMine

Yeast

public beta

Zebrafish

Rat

private beta

released

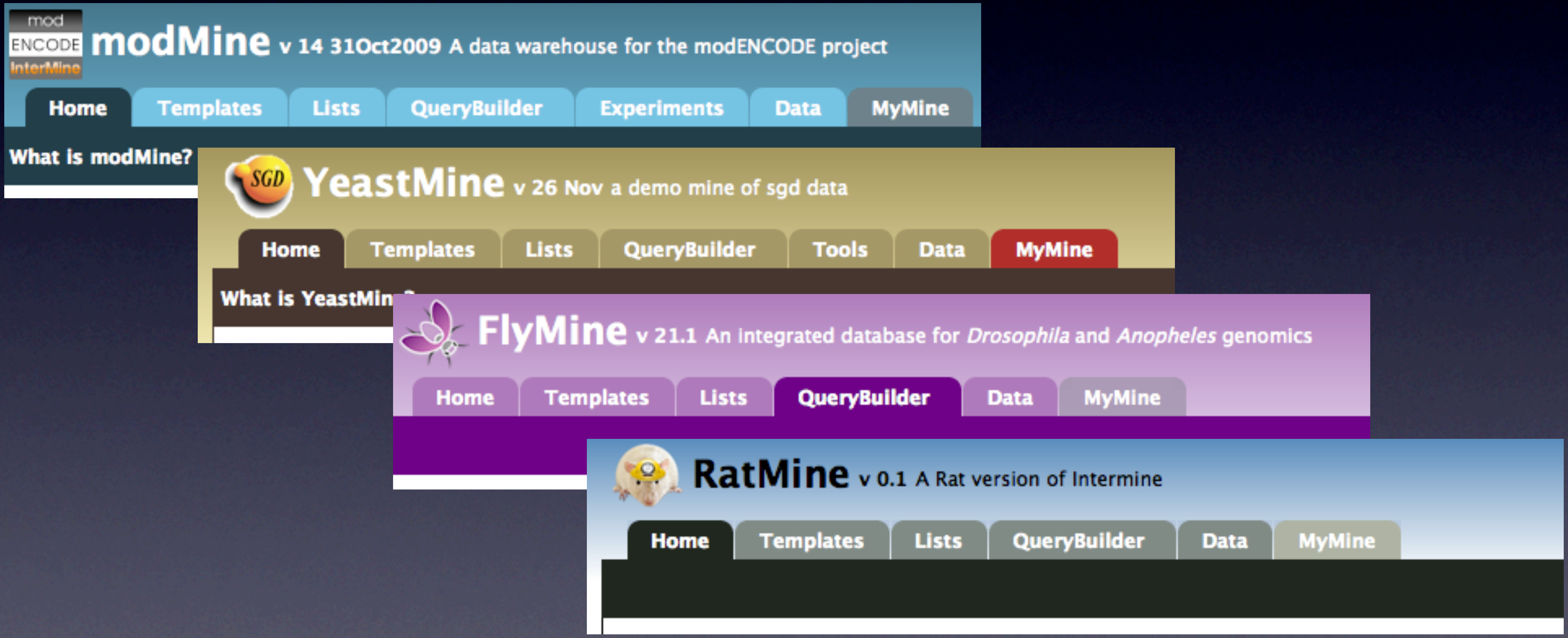
Mouse

Worm

Mines for MODs

- InterMine alongside existing web site
  - contains MOD data + other sources
  - *working on better embedding of InterMine*
- New features for MODs
- MODs replacing some older search tools
- Lots of input for InterMine development!

Common Interface



modMine v 14 31Oct2009 A data warehouse for the modENCODE project

What is modMine?

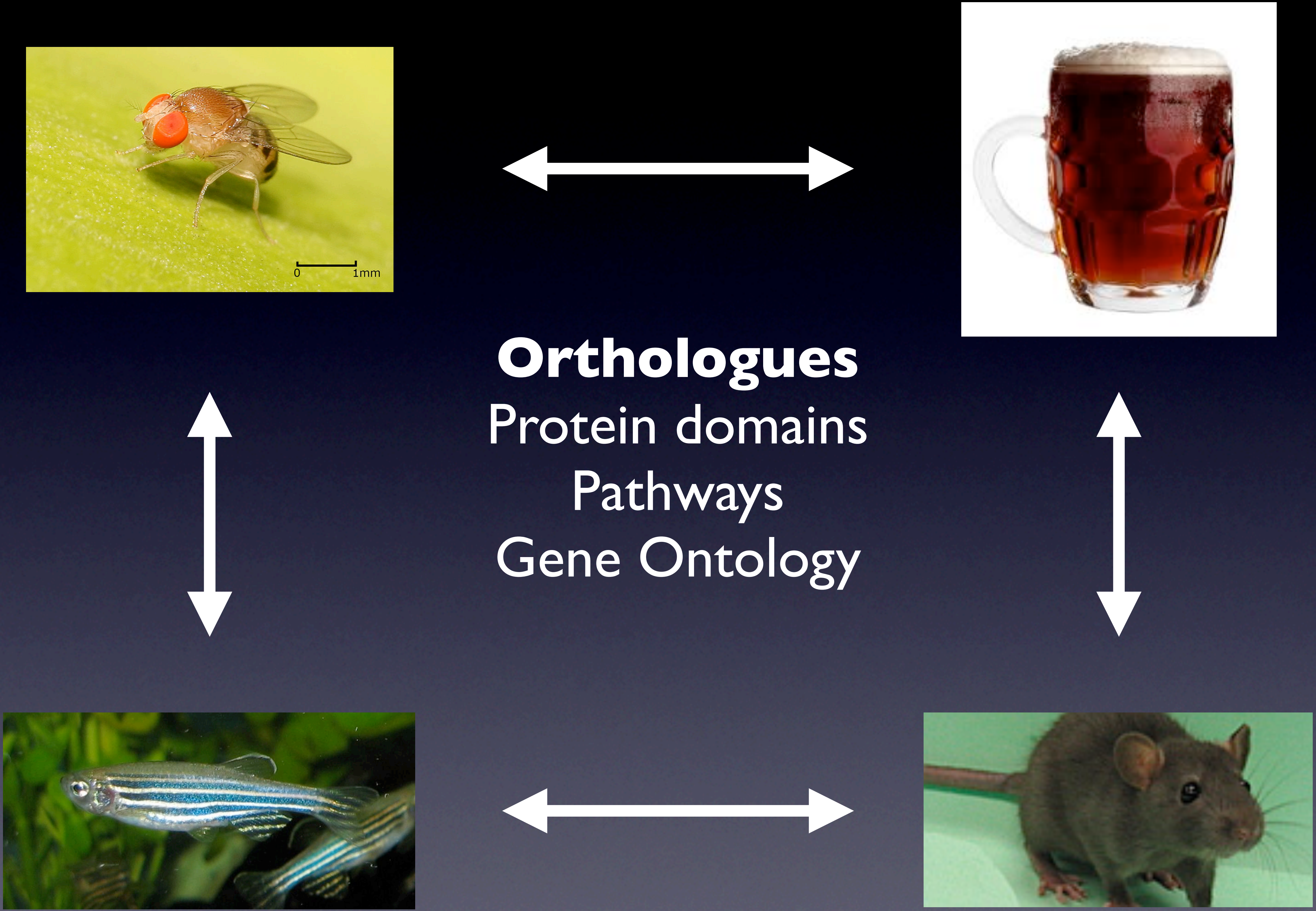
YeastMine v 26 Nov a demo mine of sgd data

What is YeastMin

FlyMine v 21.1 An integrated database for *Drosophila* and *Anopheles* genomics

RatMine v 0.1 A Rat version of Intermine

Your Mine here...



Orthologues

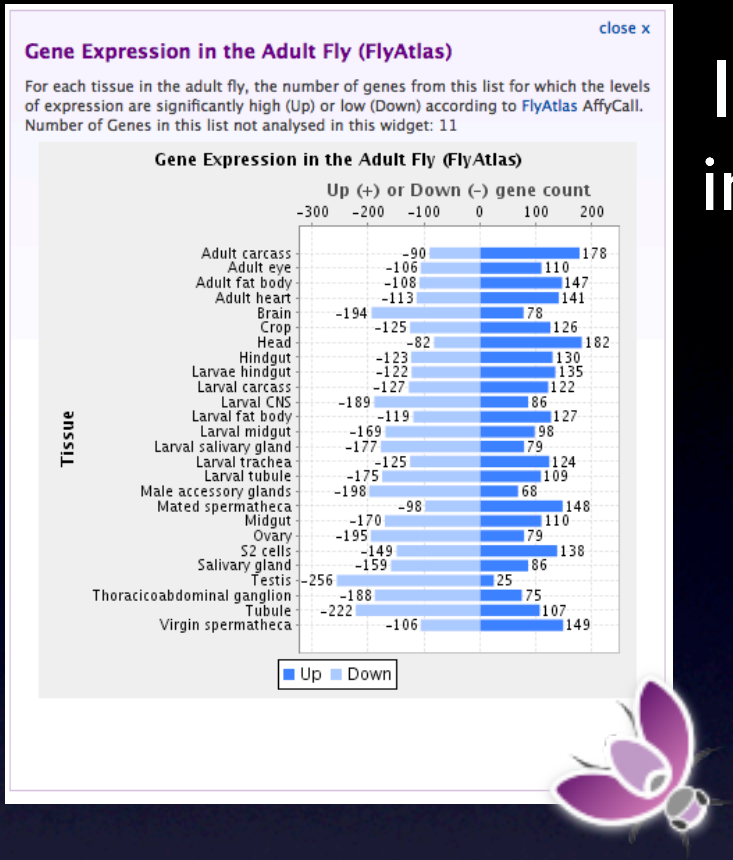
Protein domains

Pathways

Gene Ontology

Images from Wikipedia and RGD

gene list in FlyMine



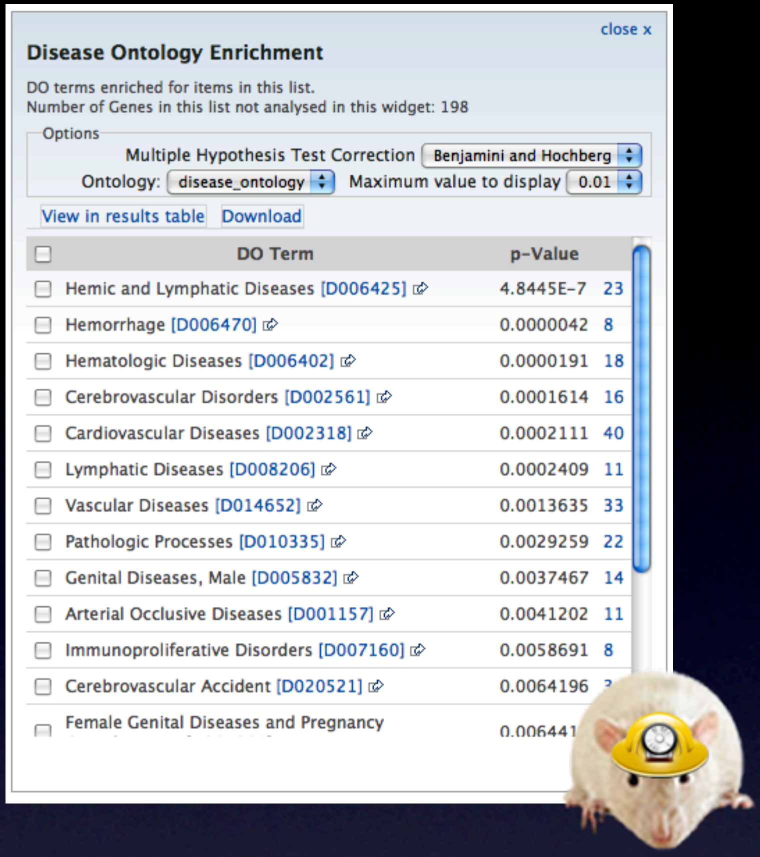
close x

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: 11

141 up-regulated
in heart

export to RGD
(orthologues)



close x

Disease Ontology Enrichment

DO terms enriched for items in this list.
Number of Genes in this list not analysed in this widget: 198

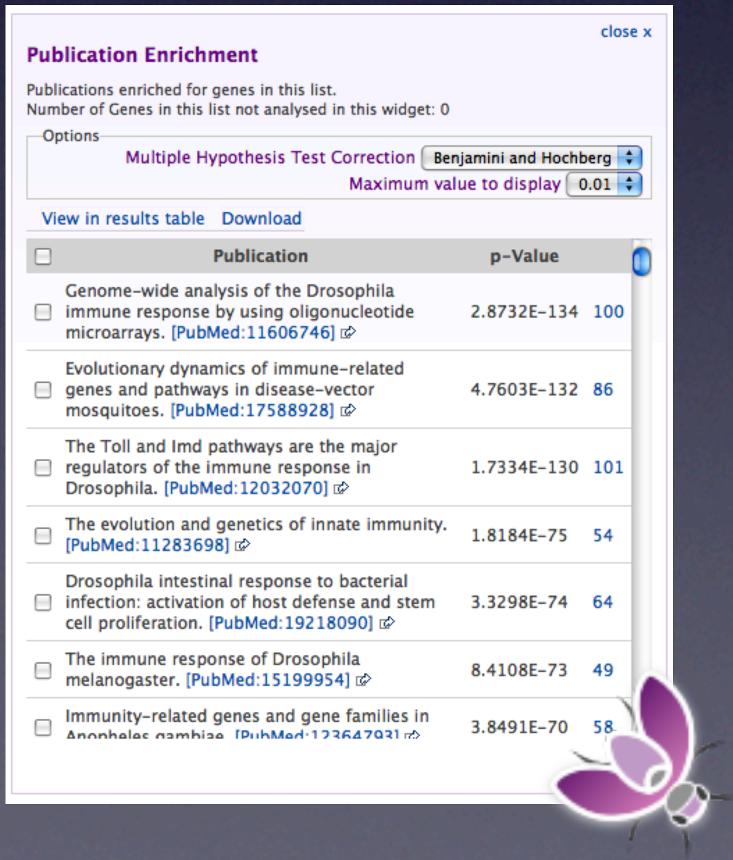
Options
Multiple Hypothesis Test Correction: Benjamini and Hochberg
Ontology: disease_ontology
Maximum value to display: 0.01

View in results table | Download

□	DO Term	p-Value	
□	Hemic and Lymphatic Diseases [D006425]	4.8445E-7	23
□	Hemorrhage [D006470]	0.0000042	8
□	Hematologic Diseases [D006402]	0.0000191	18
□	Cerebrovascular Disorders [D002561]	0.0001614	16
□	Cardiovascular Diseases [D002318]	0.0002111	40
□	Lymphatic Diseases [D008206]	0.0002409	11
□	Vascular Diseases [D014652]	0.0013635	33
□	Pathologic Processes [D010335]	0.0029259	22
□	Genital Diseases, Male [D005832]	0.0037467	14
□	Arterial Occlusive Diseases [D001157]	0.0041202	11
□	Immunoproliferative Disorders [D007160]	0.0058691	8
□	Cerebrovascular Accident [D020521]	0.0064196	3
□	Female Genital Diseases and Pregnancy	0.006441	

40 related to
Cardiovascular
disease

export to FlyMine
(orthologues)



close x

Publication Enrichment

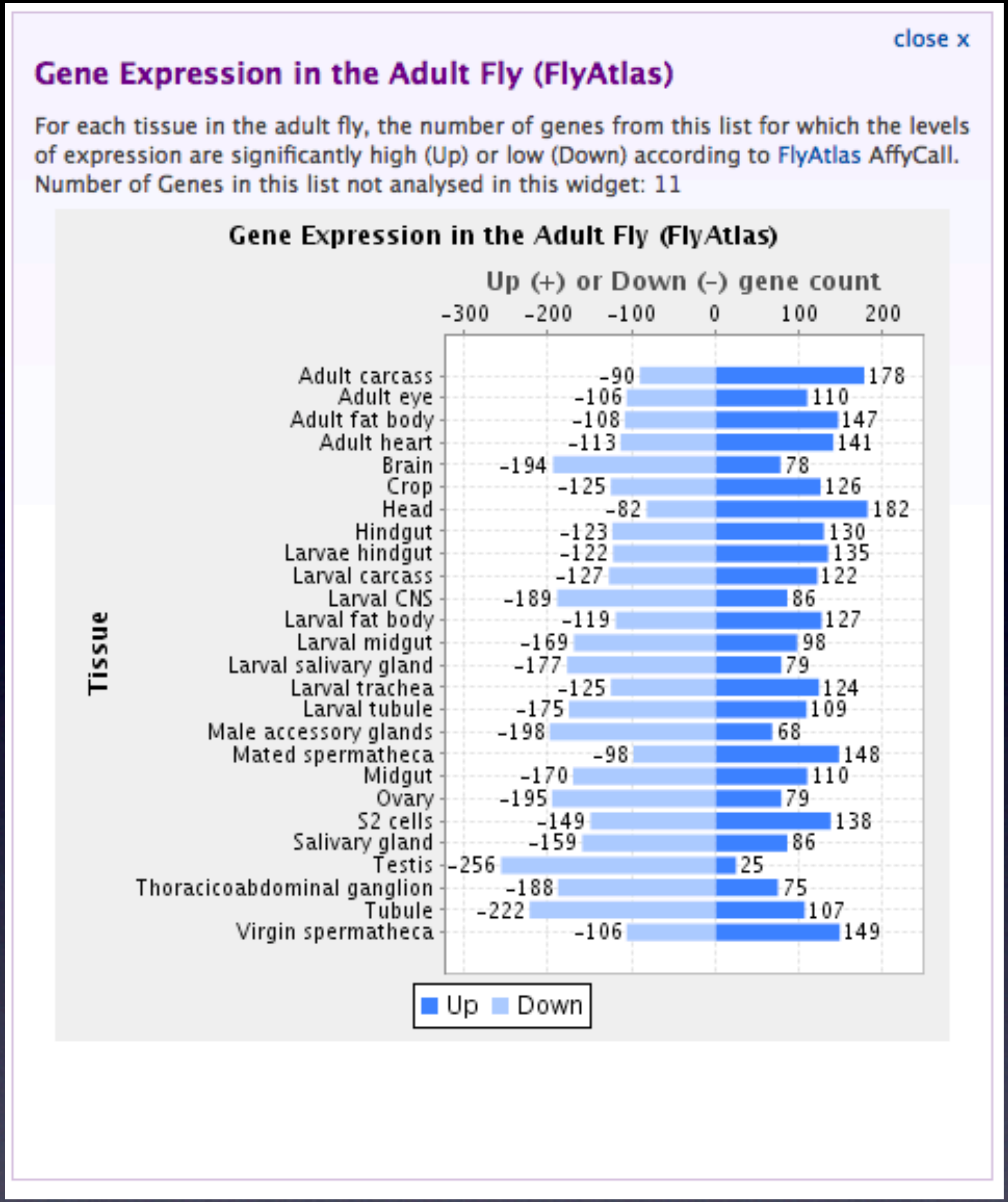
Publications enriched for genes in this list.
Number of Genes in this list not analysed in this widget: 0

Options
Multiple Hypothesis Test Correction: Benjamini and Hochberg
Maximum value to display: 0.01

View in results table | Download

□	Publication	p-Value	
□	Genome-wide analysis of the Drosophila immune response by using oligonucleotide microarrays. [PubMed:11606746]	2.8732E-134	100
□	Evolutionary dynamics of immune-related genes and pathways in disease-vector mosquitoes. [PubMed:17588928]	4.7603E-132	86
□	The Toll and Imd pathways are the major regulators of the immune response in Drosophila. [PubMed:12032070]	1.7334E-130	101
□	The evolution and genetics of innate immunity. [PubMed:11283698]	1.8184E-75	54
□	Drosophila intestinal response to bacterial infection: activation of host defense and stem cell proliferation. [PubMed:19218090]	3.3298E-74	64
□	The immune response of Drosophila melanogaster. [PubMed:15199954]	8.4108E-73	49
□	Immunity-related genes and gene families in Anopheles gambiae. [PubMed:12364793]	3.8491E-70	58

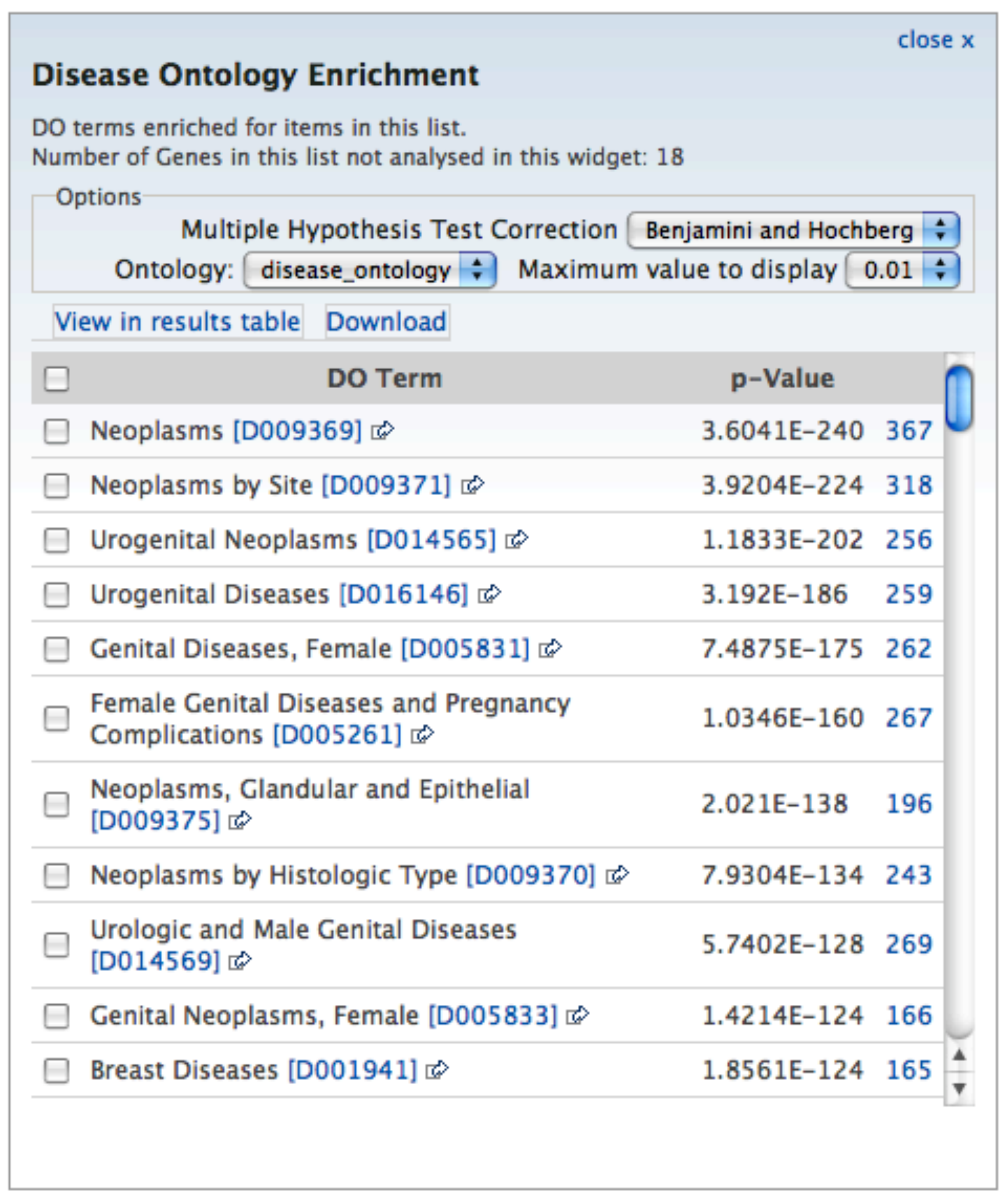
- look for relevant publications
- export sequences



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: 11



close x

Disease Ontology Enrichment

DO terms enriched for items in this list.
Number of Genes in this list not analysed in this widget: 18

Options
Multiple Hypothesis Test Correction: Benjamini and Hochberg
Ontology: disease_ontology
Maximum value to display: 0.01

□	DO Term	p-Value	
□	Neoplasms [D009369]	3.6041E-240	367
□	Neoplasms by Site [D009371]	3.9204E-224	318
□	Urogenital Neoplasms [D014565]	1.1833E-202	256
□	Urogenital Diseases [D016146]	3.192E-186	259
□	Genital Diseases, Female [D005831]	7.4875E-175	262
□	Female Genital Diseases and Pregnancy Complications [D005261]	1.0346E-160	267
□	Neoplasms, Glandular and Epithelial [D009375]	2.021E-138	196
□	Neoplasms by Histologic Type [D009370]	7.9304E-134	243
□	Urologic and Male Genital Diseases [D014569]	5.7402E-128	269
□	Genital Neoplasms, Female [D005833]	1.4214E-124	166
□	Breast Diseases [D001941]	1.8561E-124	165

- fetch results dynamically between Mines
- other developers can create mash-ups

metabolicMine

- Metabolic diseases: *diabetes, obesity*
- Close collaboration with major labs
- Human, mouse, rat
- Lots of data!
  - SNPs, HapMap, expression, ENCODE...
  - Already have parsers for many sources

metabolicMine

- We’ve identified a gene, what does it do?
  - currently looking at many sites per gene
- Prioritise candidate genes
- Compare lists, find common attributes
- Upload genome coordinates -> features

InterMine 0.94

new features

Search

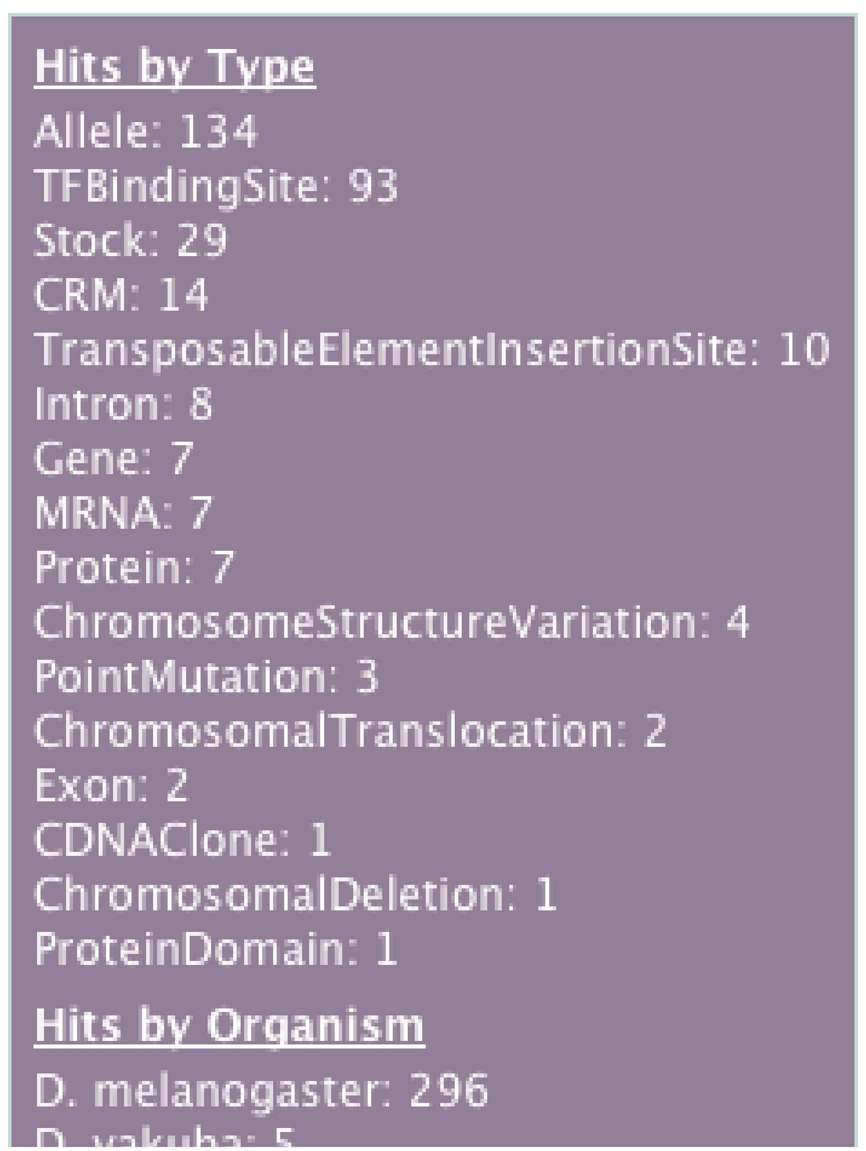
`Templates and QueryBuilder are structured searches`

- Lucene search over whole database
- FlyMine: 45 mins, 2.5 GB index file
  - parallel fetching, precomputes
- Each object is a document
  - attributes are fields
  - also related data e.g. GO, pathways

Faceted Search

- BOBO - created by LinkedIn
- Group results by 'facets'
- Filter by multiple facets
- Facet on any property

  e.g. pathway, expression
- *Display is hard*



Hits by Type

Allele: 134
TFBindingSite: 93
Stock: 29
CRM: 14
TransposableElementInsertionSite: 10
Intron: 8
Gene: 7
MRNA: 7
Protein: 7
ChromosomeStructureVariation: 4
PointMutation: 3
ChromosomalTranslocation: 2
Exon: 2
CDNAClone: 1
ChromosomalDeletion: 1
ProteinDomain: 1

Hits by Organism

D. melanogaster: 296
D. yakuba: 5

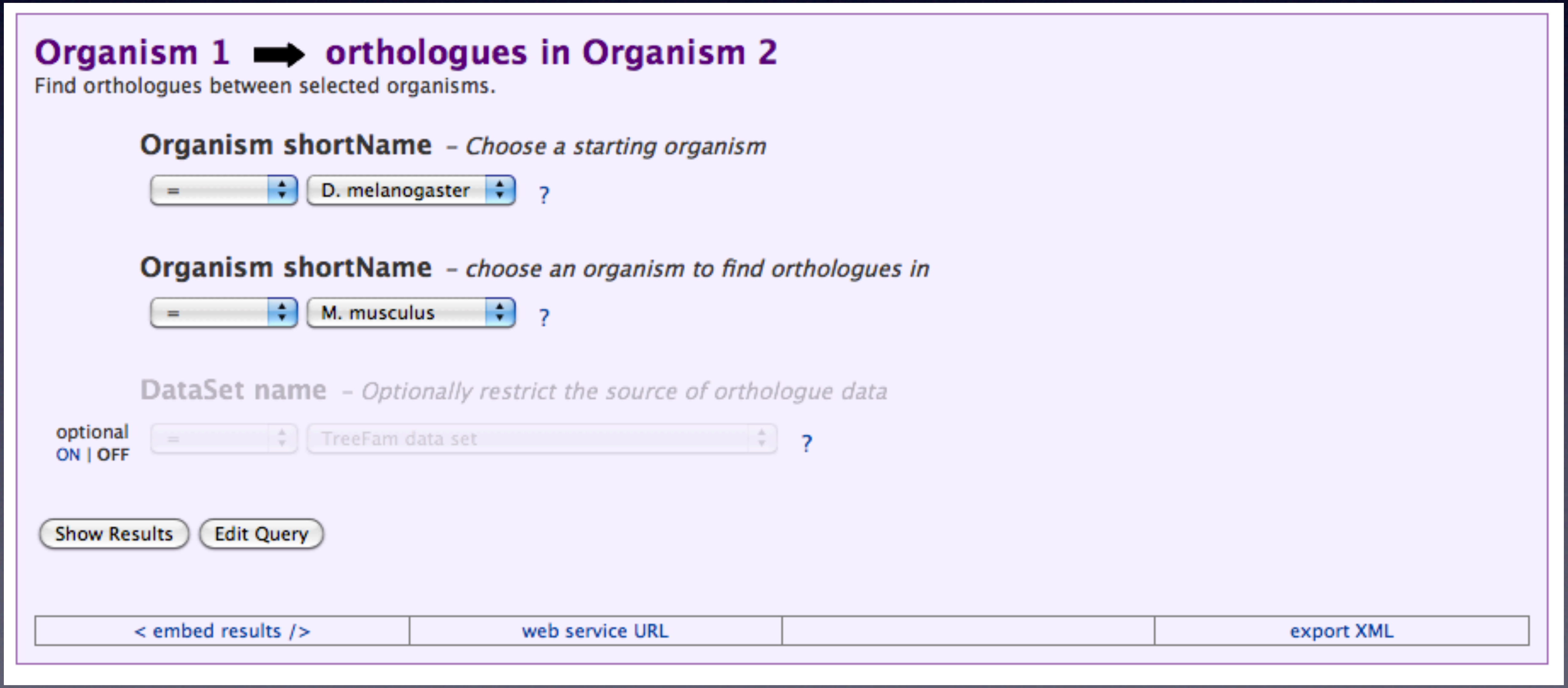
`Integrates with lists and templates`

Configurable

```
index.references.BioEntity = synonyms organism
index.references.OntologyTerm = synonyms
index.references.Gene = pathways omimDiseases
proteins.proteinDomains goAnnotation.ontologyTerm
index.references.Protein = proteinDomains
index.ignore = Annotation AnalysisResult Comment Homologue
Interaction Image OntologyAnnotation OntologyRelation
OntologyAnnotation ProteinStructure Reporter Sequence
SymmetricalRelation Synonym
index.facet.Category = Category
index.facet.Organism = organism.shortName
index.boost.Gene = 2.0
index.boost.Protein = 1.5
```

Templates

- Problem: too many similar templates



Organism 1 ➡ orthologues in Organism 2

Find orthologues between selected organisms.

Organism shortName – *Choose a starting organism*

= | D. melanogaster | ?

Organism shortName – *choose an organism to find orthologues in*

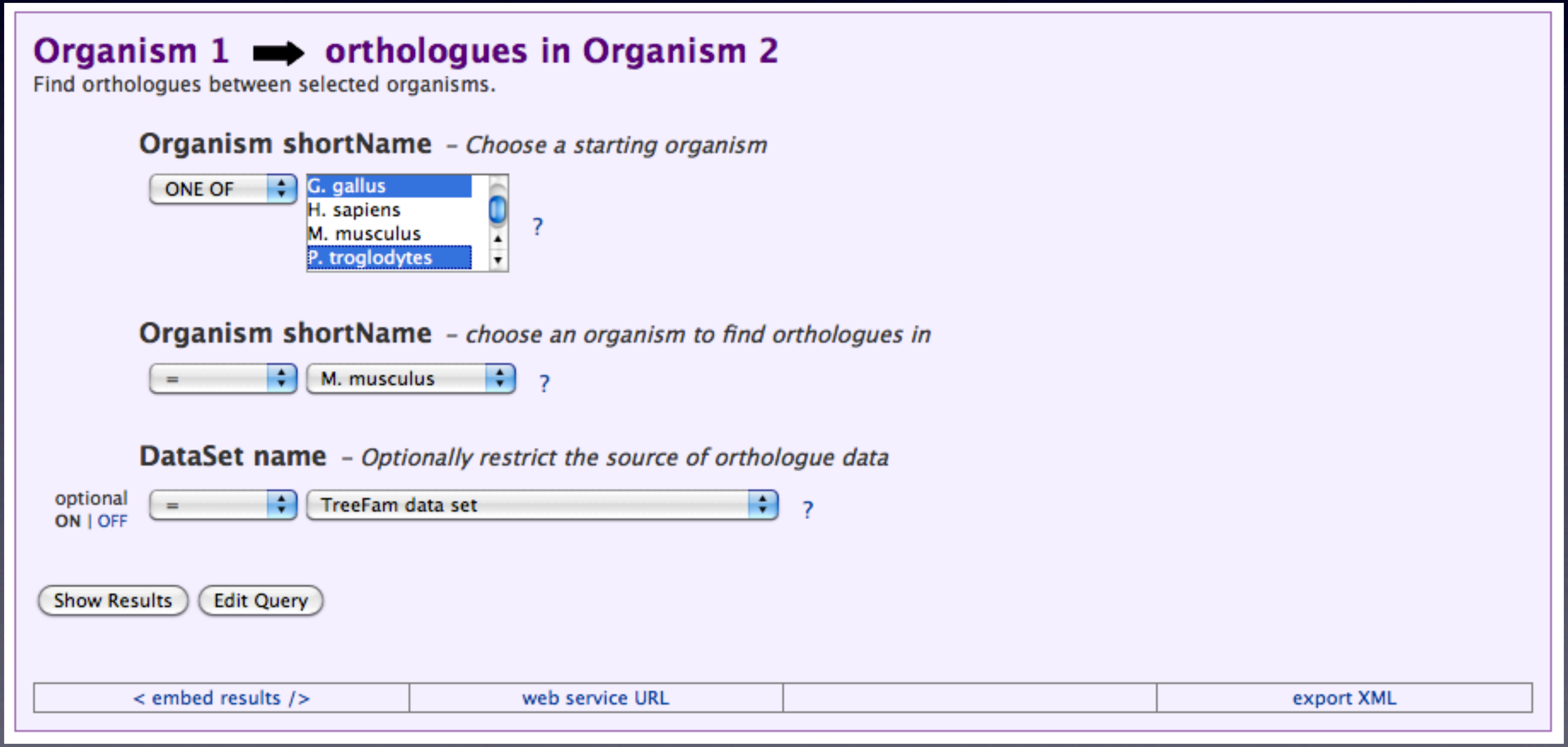
= | M. musculus | ?

DataSet name – *Optionally restrict the source of orthologue data*

optional ON | OFF — = | TreeFam data set | ?

Templates

- Select multiple values



Organism 1 ➡ orthologues in Organism 2

Find orthologues between selected organisms.

Organism shortName – *Choose a starting organism*

ONE OF: G. gallus, P. troglodytes (selected from G. gallus, H. sapiens, M. musculus, P. troglodytes)

Organism shortName – *choose an organism to find orthologues in*

= M. musculus

DataSet name – *Optionally restrict the source of orthologue data*

optional ON | OFF

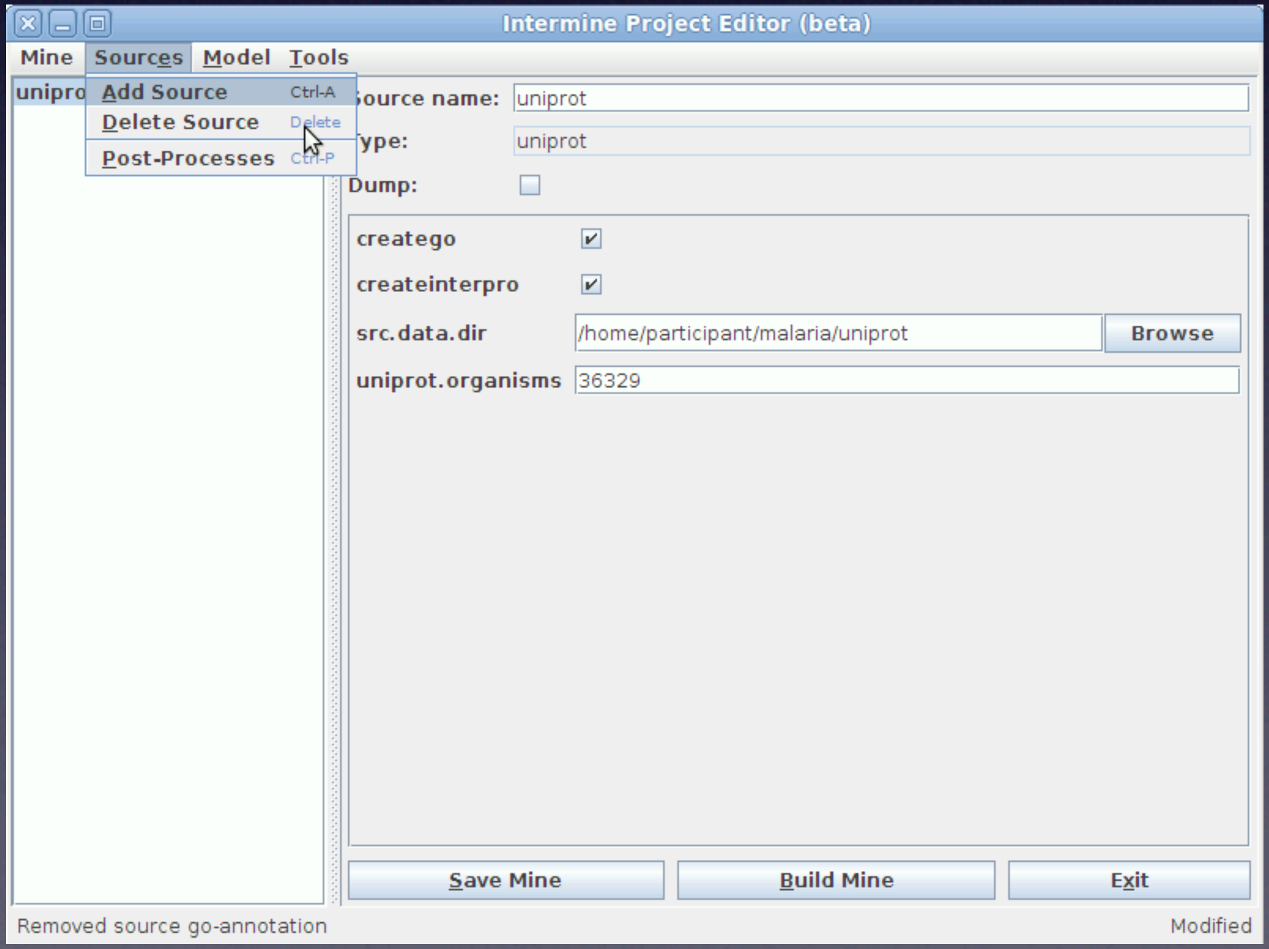
= TreeFam data set

Performance

- Faster builds:
  - FlyMine: 18 hours to 10 hours (75GB)
- Better results cache
- More caches!
- Faster export

MineManager

- GUI to simplify installation
- Simple selection of data sources



Source name: uniprot

Type: uniprot

Dump: □

creatego ☑

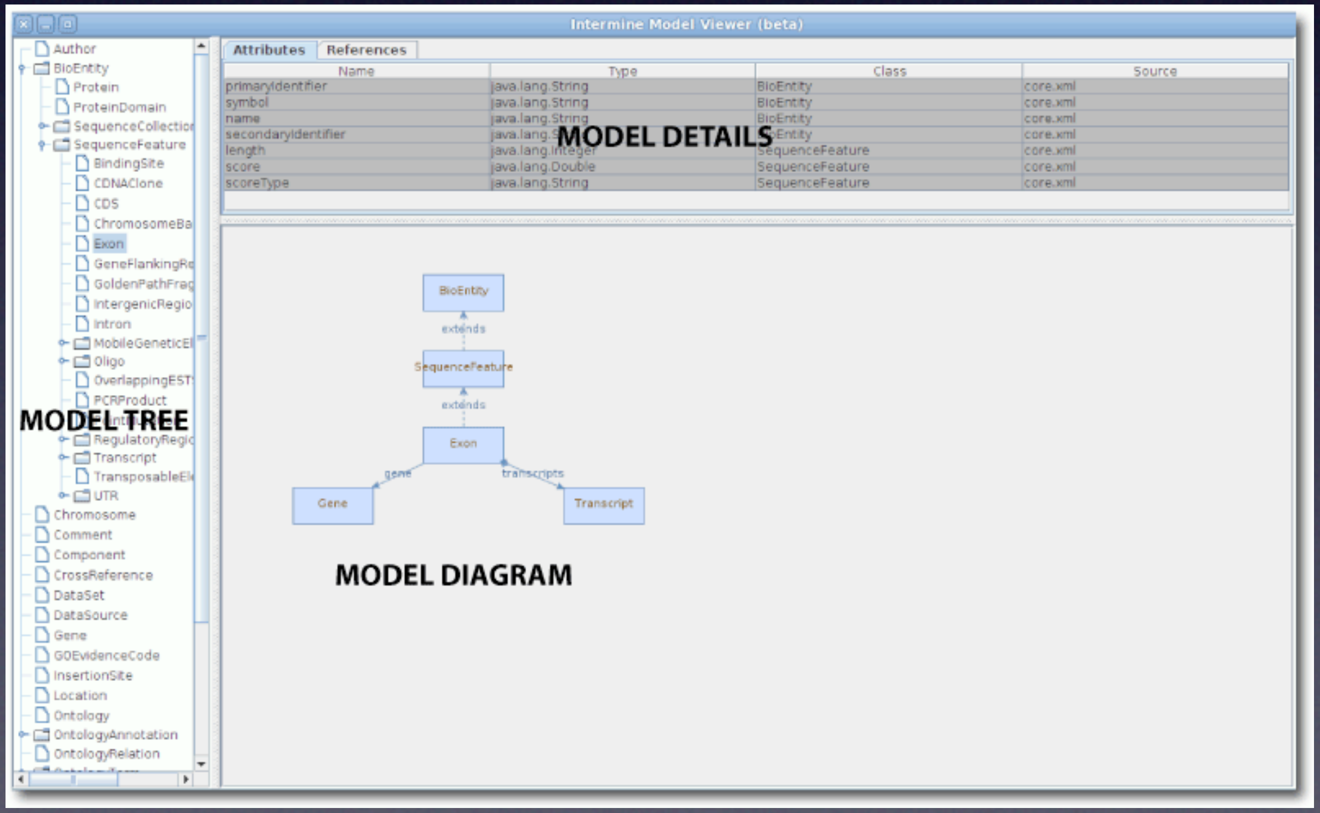
createinterpro ☑

src.data.dir /home/participant/malaria/uniprot

uniprot.organisms 36329

Removed source go-annotation

Modified



Name	Type	Class	Source
primaryIdentifier	java.lang.String	BioEntity	core.xml
symbol	java.lang.String	BioEntity	core.xml
name	java.lang.String	BioEntity	core.xml
secondaryIdentifier	java.lang.S	Entity	core.xml
length	java.lang.Integer	SequenceFeature	core.xml
score	java.lang.Double	SequenceFeature	core.xml
scoreType	java.lang.String	SequenceFeature	core.xml

And theres more...

- Galaxy integration
- Improved Perl web service API
- Automatic SO -> model generation
- Upload list of genome regions
- CytoscapeWeb plugin

Workshop

Come and build your own InterMine

Tomorrow!

We're hiring: calling all Java/Database/performance gurus.

InterMine Team

Biologists	Rachel Lyne, Adrian Carr, Mike Lyne
Developers	Richard Smith, Sergio Contrino, Julie Sullivan, Matthew Wakeling, Alex Kalderimis, Fengyuan Hu, Daniela Butano
Students	Nils Kolling, Richard B.
Sys Admin	Dan Tomlinson
PI	Gos Micklem

www.intermine.org

richard@flymine.org