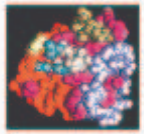
Data integration for the genome sciences - lessons from the FlyMine project

Genomics
Genome annotation

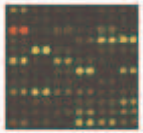
DrosDel
P-element insertions and deletions

RNAi
RNA interference



Proteins
Protein and proteomics data

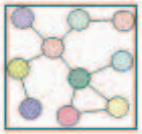
Gene Ontology



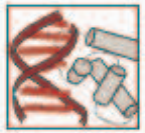
Gene Expression
ArrayExpress

Tiling Path
Microarray Tiling Primers

Disease
Human disease matches from Homophila



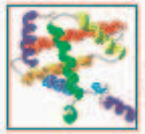
Protein Interactions
IntAct



Transcriptional Regulation
Regulatory regions and transcription factor binding sites

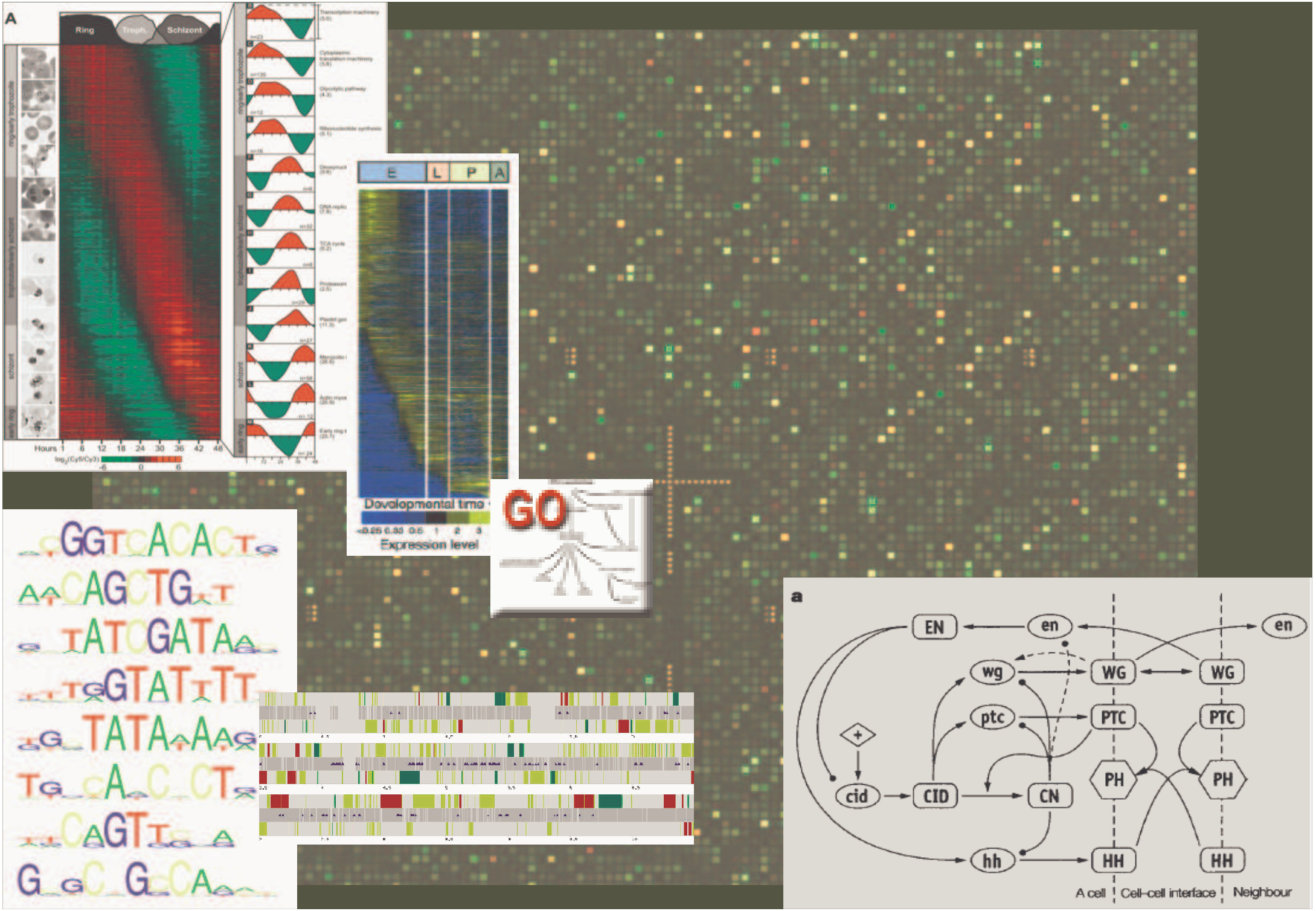
INDAC
Long oligos from the International *Drosophila* Array Consortium

Comparative Genomics
Orthologues and paralogues

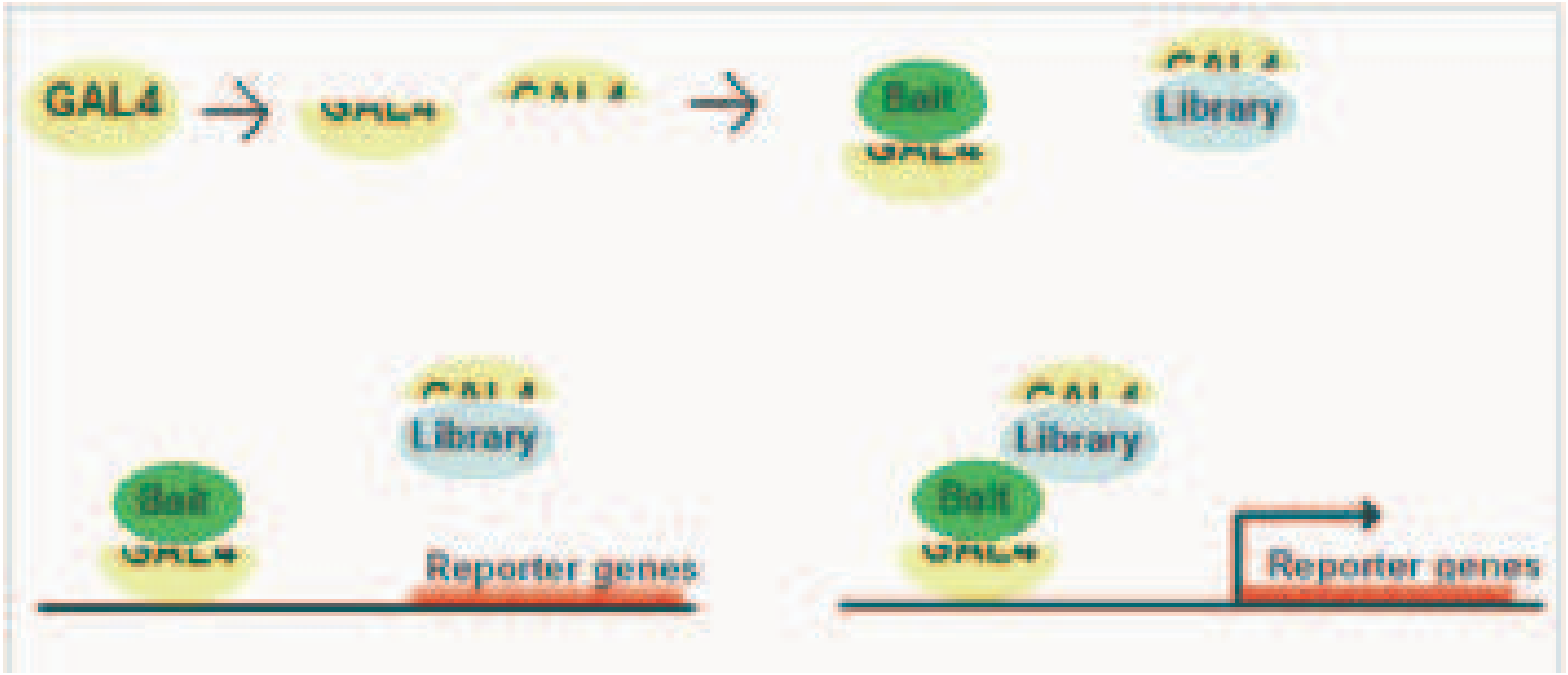


Protein Structure
3-D protein structures

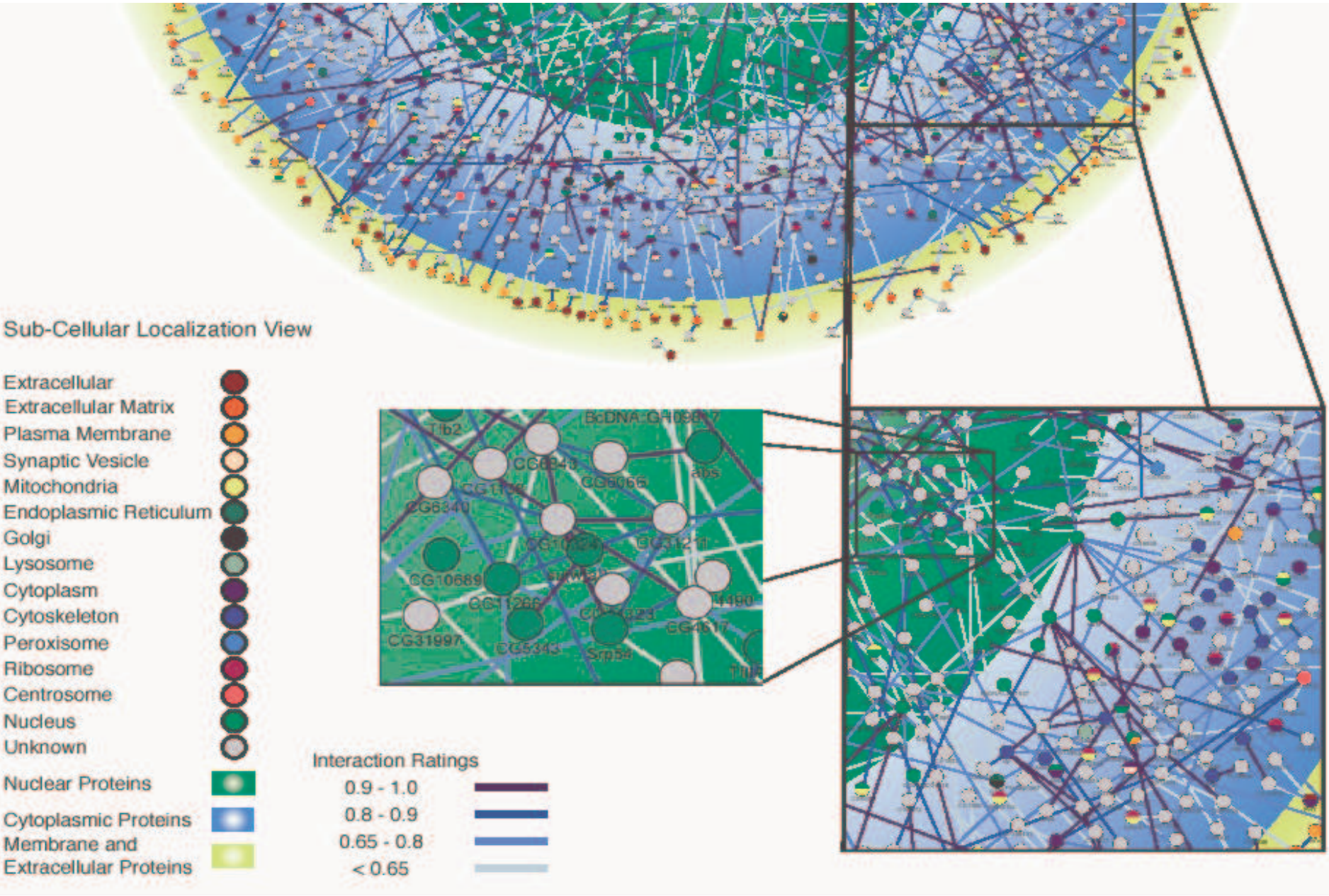
Gos Micklem


A
Ring
Troph.
Schizont
ring/early trophozoite
trophozoite/early schizont
schizont
early ring
Hours 1 6 12 18 24 30 36 42 48
log2(Cy5/Cy3)
-6 0 6
E
L
P
A
Developmental time
<0.25 0.33 0.5 1 2 3
Expression level
GO
a
EN
en
en
wg
WG
WG
ptc
PTC
PTC
+
cid
CID
CN
PH
PH
hh
HH
HH
A cell
Cell–cell interface
Neighbour

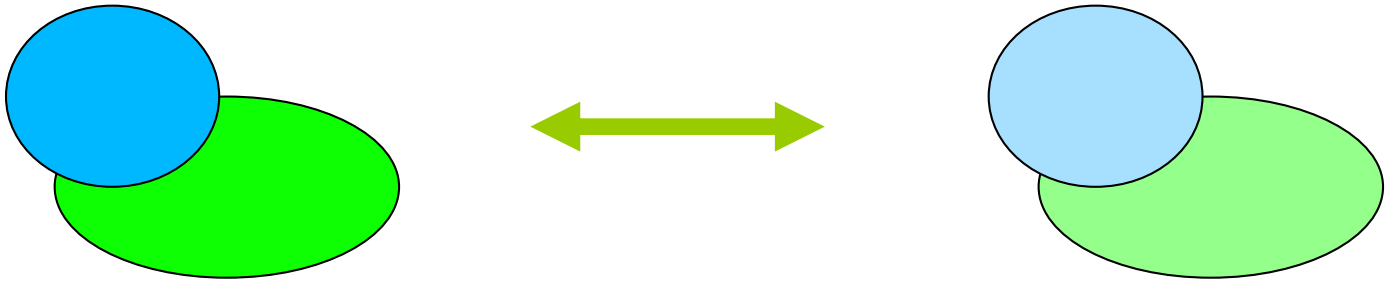
Yeast 2-hybrid screening



Drosophila


Sub-Cellular Localization View
Extracellular
Extracellular Matrix
Plasma Membrane
Synaptic Vesicle
Mitochondria
Endoplasmic Reticulum
Golgi
Lysosome
Cytoplasm
Cytoskeleton
Peroxisome
Ribosome
Centrosome
Nucleus
Unknown
Nuclear Proteins
Cytoplasmic Proteins
Membrane and
Extracellular Proteins
Interaction Ratings
0.9 - 1.0
0.8 - 0.9
0.65 - 0.8
< 0.65
Tlb2
CG6340
CG10689
CG31997
CG5343
Srp54
abs
CG31211
CG4617

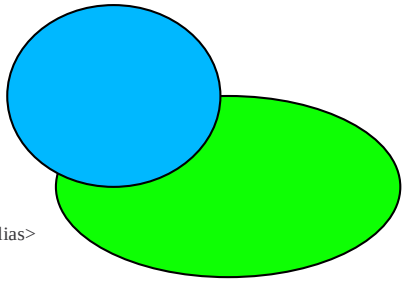
Interologs



D. melanogaster

C. elegans

PSI for Drosophila

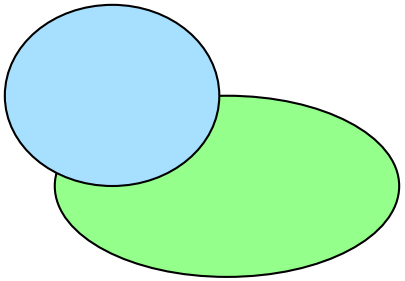


```
<interactor id="6">
 <names>
  <shortLabel>src64_drome</shortLabel>
  <fullName>Tyrosine-protein kinase Src64B</fullName>
  <alias type="gene name" typeAc="MI:0301">Src64B</alias>
 </names>
 <xref>
  <primaryRef db="uniprotkb" dbAc="MI:0486" id="P00528" refType="identity" refTypeAc="MI:0356"
secondary="src64_drome" version="SP_48"/>
  <secondaryRef db="go" dbAc="MI:0448" id="GO:0007391" secondary="P:dorsal closure"/>
 </xref>
 <interactorType>
  <names>
   <shortLabel>protein</shortLabel>
   <fullName>protein</fullName>
  </names>
  <xref>
   <primaryRef db="psi-mi" dbAc="MI:0488" id="MI:0326" refType="identity" refTypeAc="MI:0356"/>
   <secondaryRef db="pubmed" dbAc="MI:0446" id="14755292" refType="primary-reference"
refTypeAc="MI:0358"/>
  </xref>
 </interactorType>
```

InParanoid fly/worm orthologues

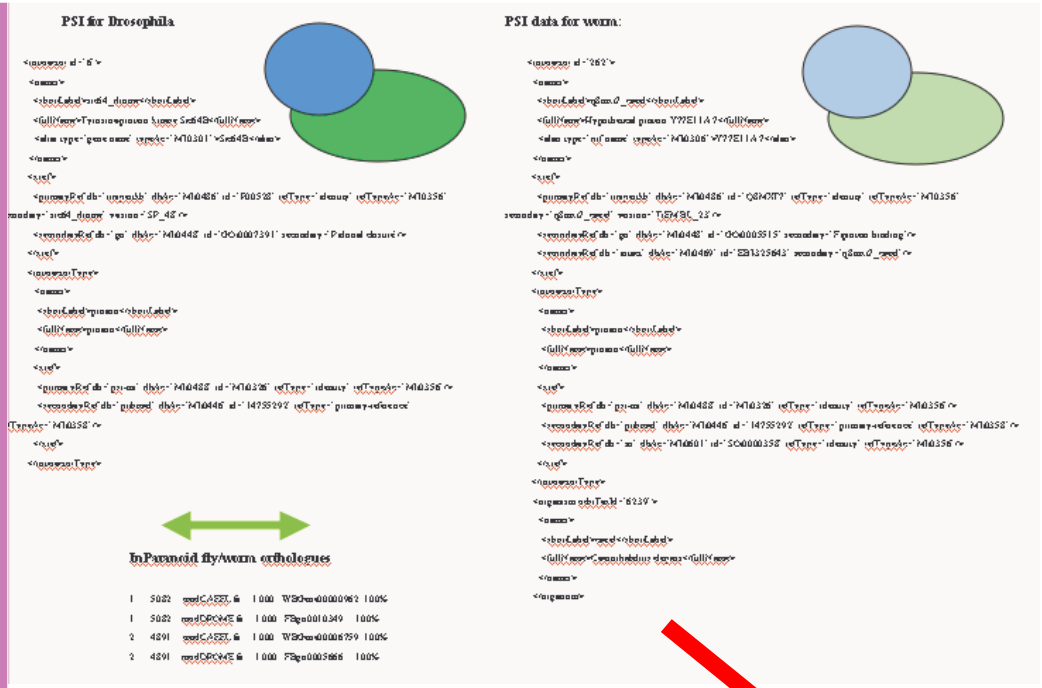
1	5082	modCAEEL.fa	1.000	WBGene00000962	100%
1	5082	modDROME.fa	1.000	FBgn0010349	100%
2	4891	modCAEEL.fa	1.000	WBGene00006759	100%
2	4891	modDROME.fa	1.000	FBgn0005666	100%

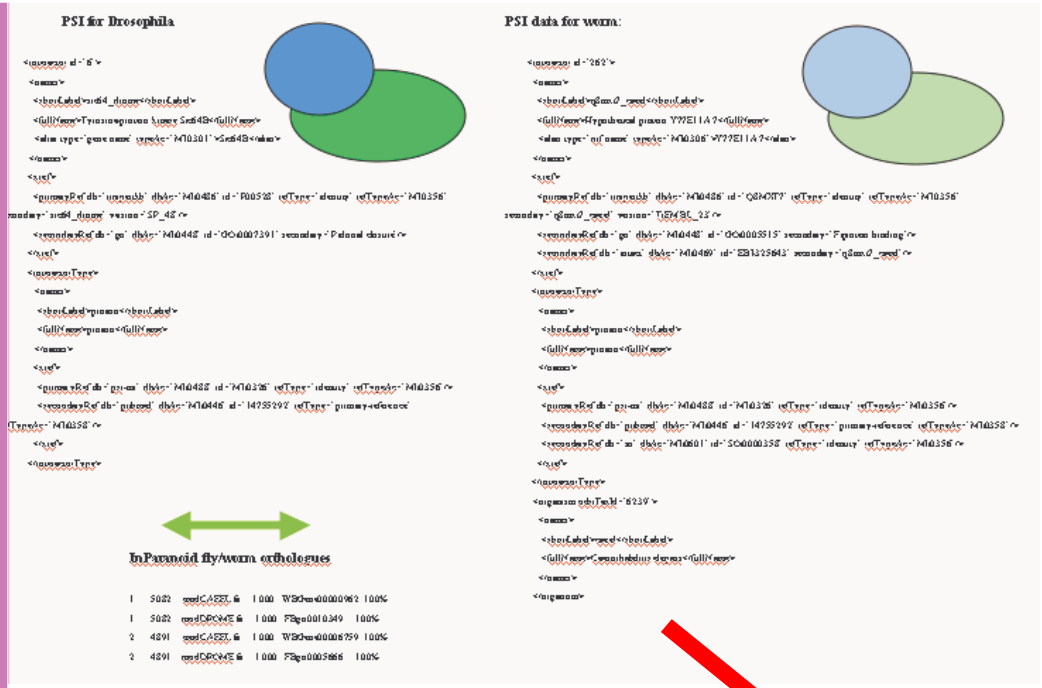
PSI data for worm:



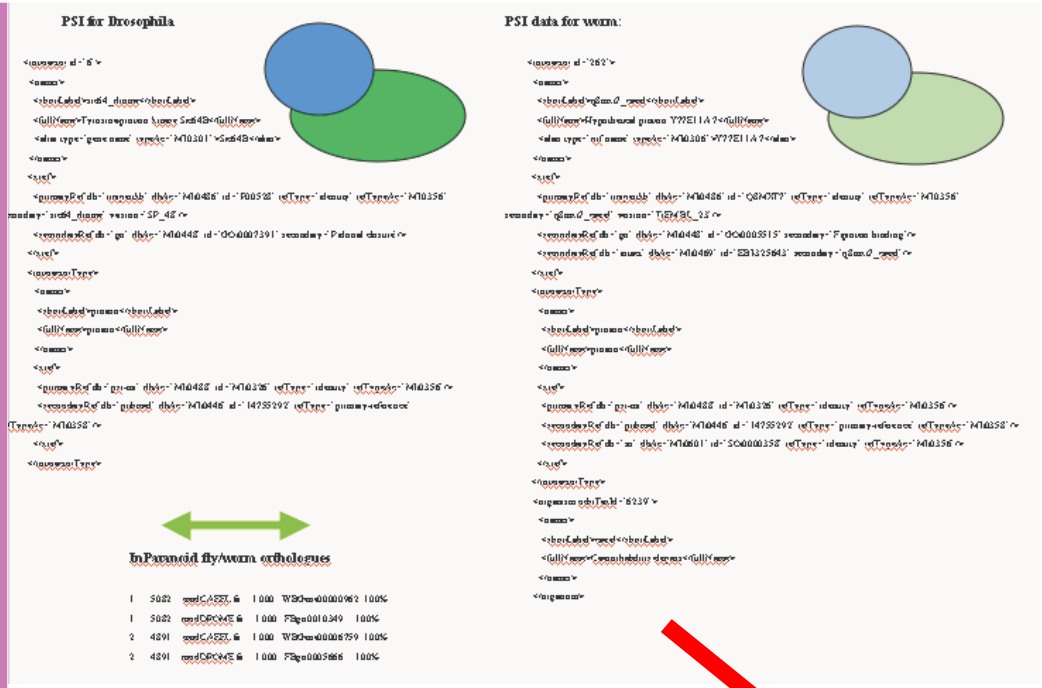
```
<interactor id="262">
 <names>
  <shortLabel>q8mxt7_caeel</shortLabel>
  <fullName>Hypothetical protein Y77E11A.7</fullName>
  <alias type="orf name" typeAc="MI:0306">Y77E11A.7</alias>
 </names>
 <xref>
  <primaryRef db="uniprotkb" dbAc="MI:0486" id="Q8MXT7" refType="identity" refTypeAc="MI:0356"
secondary="q8mxt7_caeel" version="TrEMBL_23"/>
  <secondaryRef db="go" dbAc="MI:0448" id="GO:0005515" secondary="F:protein binding"/>
  <secondaryRef db="intact" dbAc="MI:0469" id="EBI-325643" secondary="q8mxt7_caeel"/>
 </xref>
 <interactorType>
  <names>
   <shortLabel>protein</shortLabel>
   <fullName>protein</fullName>
  </names>
  <xref>
   <primaryRef db="psi-mi" dbAc="MI:0488" id="MI:0326" refType="identity" refTypeAc="MI:0356"/>
   <secondaryRef db="pubmed" dbAc="MI:0446" id="14755292" refType="primary-reference" refTypeAc="MI:0358"/>
   <secondaryRef db="so" dbAc="MI:0601" id="SO:0000358" refType="identity" refTypeAc="MI:0356"/>
  </xref>
 </interactorType>
 <organism ncbiTaxId="6239">
  <names>
   <shortLabel>caeel</shortLabel>
   <fullName>Caenorhabditis elegans</fullName>
  </names>
 </organism>
```

Standard data formats?


PSI for Drosophila
PSI data for worm:
InParanoid fly/worm orthologues


PSI for Drosophila
PSI data for worm:
InParanoid fly/worm orthologues

Nothing!



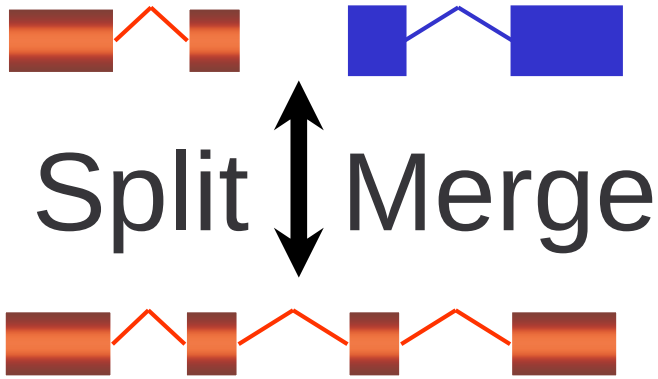
None?

Naming?

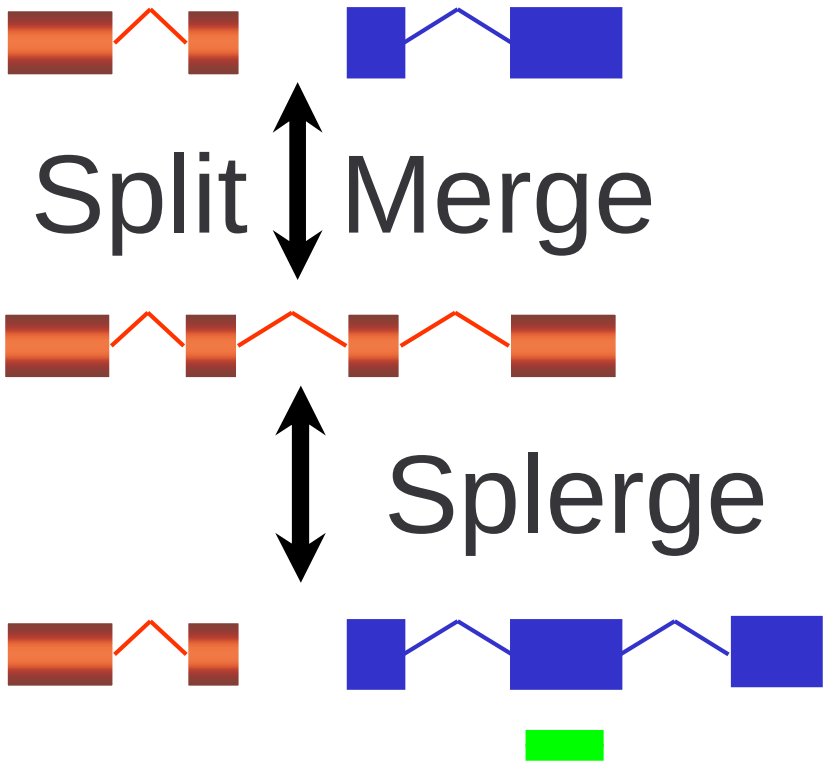
Timing?

Genomes

Sequence, annotation not stable

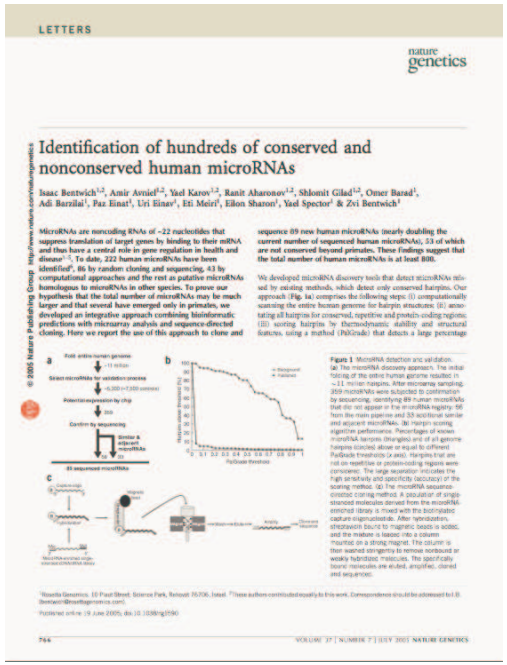


Some MODs track annotation history



Over time a single microarray probe can assay ‘different’ genes

Fund, publish, freeze



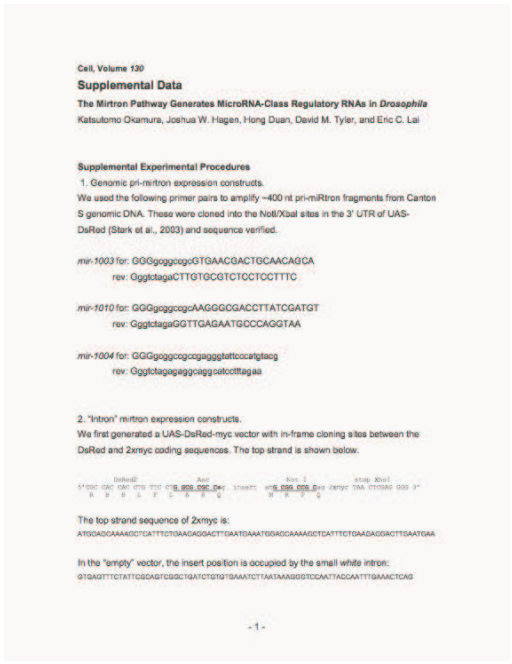
LETTERS

nature genetics

Identification of hundreds of conserved and nonconserved human microRNAs

Isaac Bentwich[1,2], Amir Avniel[1,2], Yael Karov[1,2], Ranit Aharonov[1,2], Shlomit Gilad[1,2], Omer Barad[1], Adi Barzilai[1], Paz Einat[1], Uri Einav[1], Eti Meiri[3], Eilon Sharon[3], Yael Spector[3] & Zvi Bentwich[1]

MicroRNAs are noncoding RNAs of ~22 nucleotides that suppress translation of target genes by binding to their mRNA and thus have a central role in gene regulation in health and disease[1–3]. To date, 222 human microRNAs have been identified[4], 86 by random cloning and sequencing, 43 by computational approaches and the rest as putative microRNAs homologous to microRNAs in other species. To prove our hypothesis that the total number of microRNAs may be much larger and that several have emerged only in primates, we developed an integrative approach combining bioinformatic predictions with microarray analysis and sequence-directed cloning. Here we report the use of this approach to clone and sequence 89 new human microRNAs (nearly doubling the current number of sequenced human microRNAs), 33 of which are not conserved beyond primates. These findings suggest that the total number of human microRNAs is at least 800.



Cell, Volume 120

Supplemental Data

The Mirtron Pathway Generates MicroRNA-Class Regulatory RNAs in *Drosophila*

Katsutomo Okamura, Joshua W. Hagen, Hong Duan, David M. Tyler, and Eric C. Lai

Supplemental Experimental Procedures

1. Genomic pri-mirtron expression constructs.

We used the following primer pairs to amplify ~400 nt pri-miRtron fragments from Canton S genomic DNA. These were cloned into the NotI/XbaI sites in the 3' UTR of UAS-DsRed (Stark et al., 2003) and sequence verified.

mir-1003 for: GGGgcggcgcGTGAACGACTGCAACAGCA
rev: GggtctagaCTTGTGCGTCTCCTCCTTTC

mir-1010 for: GGGgcggcgcAAGGGCGACCTTATCGATGT
rev: GggtctagaGGTTGAGAATGCCCAGGTAA

mir-1004 for: GGGgcggcgcgcgagggtattcccatgtacg
rev: Gggtctagagaggcaggcatcctttagaa

2. "Intron" mirtron expression constructs.

We first generated a UAS-DsRed-myc vector with in-frame cloning sites between the DsRed and 2xmyc coding sequences. The top strand is shown below.

The top strand sequence of 2xmyc is:

ATGGACGCAAAAGCTCATTTCTGAACAGGACTTCAATGAAATGGACGCAAAAGCTCATTTCTGAACAGGACTTGAATGAA

In the "empty" vector, the insert position is occupied by the small *white* intron:

GTGAGTTTCTATTCGCAGTCGGCTGATCTGTGTGAAATCTTAATAAAGGGTCCAATTACCAATTTGAAACTCAG

- 1 -

Supplementary data/
Database online but not maintained

Synchronisation

Secondary Data

IntAct PSI for Drosophila

1) has UniProt ID and a gene symbol

2) contains secondary data - includes GO and InterPro data

3) has a sequence which may not match UniProt

IntAct updates every two weeks so they may keep up to date. But GO terms often don't match GO terms in the UniProt record.

IntAct has trEmbl sequences, but trEmbl records disappear over time…

Synonyms/ multiple identifiers

Lab independently discover and name genes (Collected by Model Organism Databases)

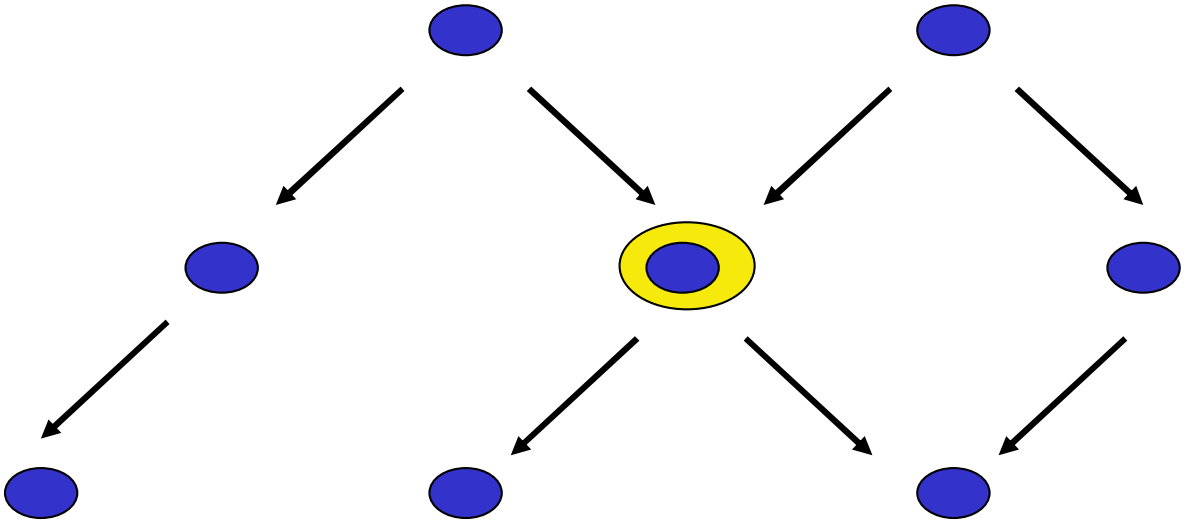
Data sources use different identifiers to refer to the same thing:
e.g. Zen, CG…., FBgn…

Need authoritative source to merge data based on different identifiers

The Three Gene Ontologies

- *Molecular Function* — elemental activity or task
  nuclease, DNA binding, transcription factor
- *Biological Process* — broad objective or goal
  mitosis, signal transduction, metabolism
- *Cellular Component* — location or complex
  nucleus, ribosome, origin recognition complex

DAG Structure



- is-a
  subclass; *a* is a type of *b*
- part-of
  physical part of (component)
  subprocess of (process)

Directed acyclic graph: each child may have one or more parents

Sequence Ontology

Naming of sequence features and their relationships:

Gene --> transcripts --> polypeptides

Well defined and uniform meaning across databases

Rules for assignment?

GO terms often inherited through sequence similarity during genome annotation

Evidence and provenance important…

Objects aren’t named consistently

Identifiers can change with time

Standard data formats are good

Evidence/Provenance are important

FlyMine/InterMine Aims

Generic, extensible data integration platform

Flexible querying (no SQL, schema knowledge)

High performance even though flexible

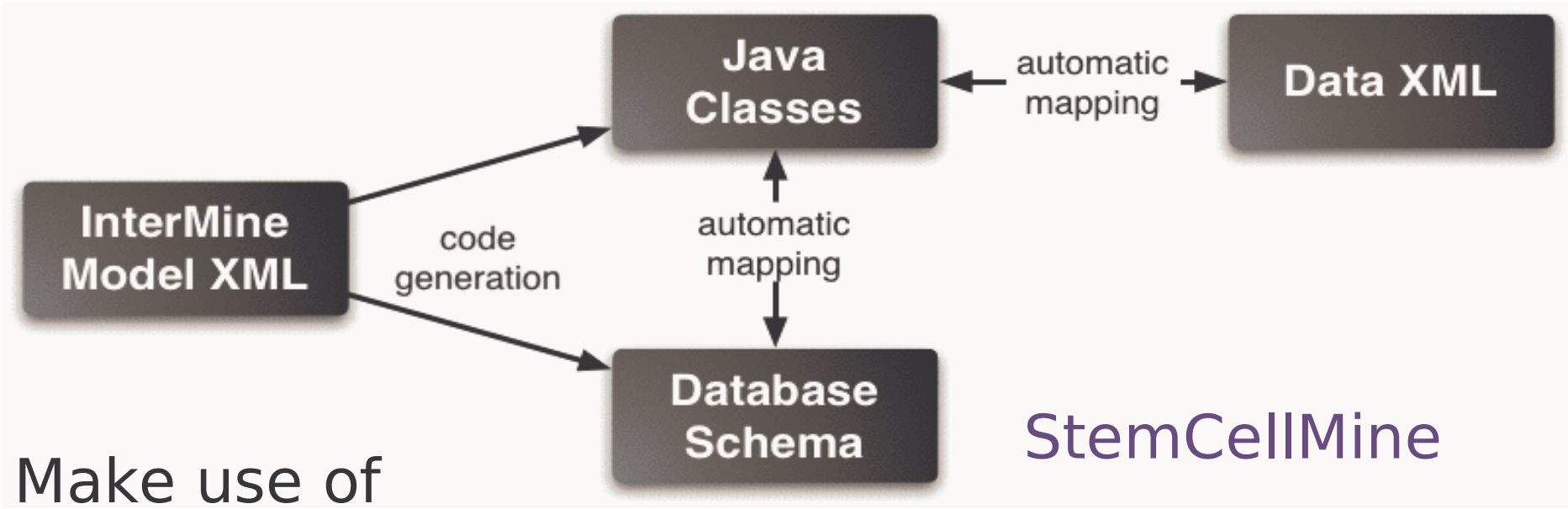
Encapsulation of complex queries for easy sharing and re-use

Operate on lists as easily as single entities

FlyMine:

(*Drosophila/ Anopheles* genomics/ proteomics)

InterMine Maximum Laziness Principle



Make use of

Standards for data

Model e.g. Sequence Ontology

StemCellMine

mitoMine

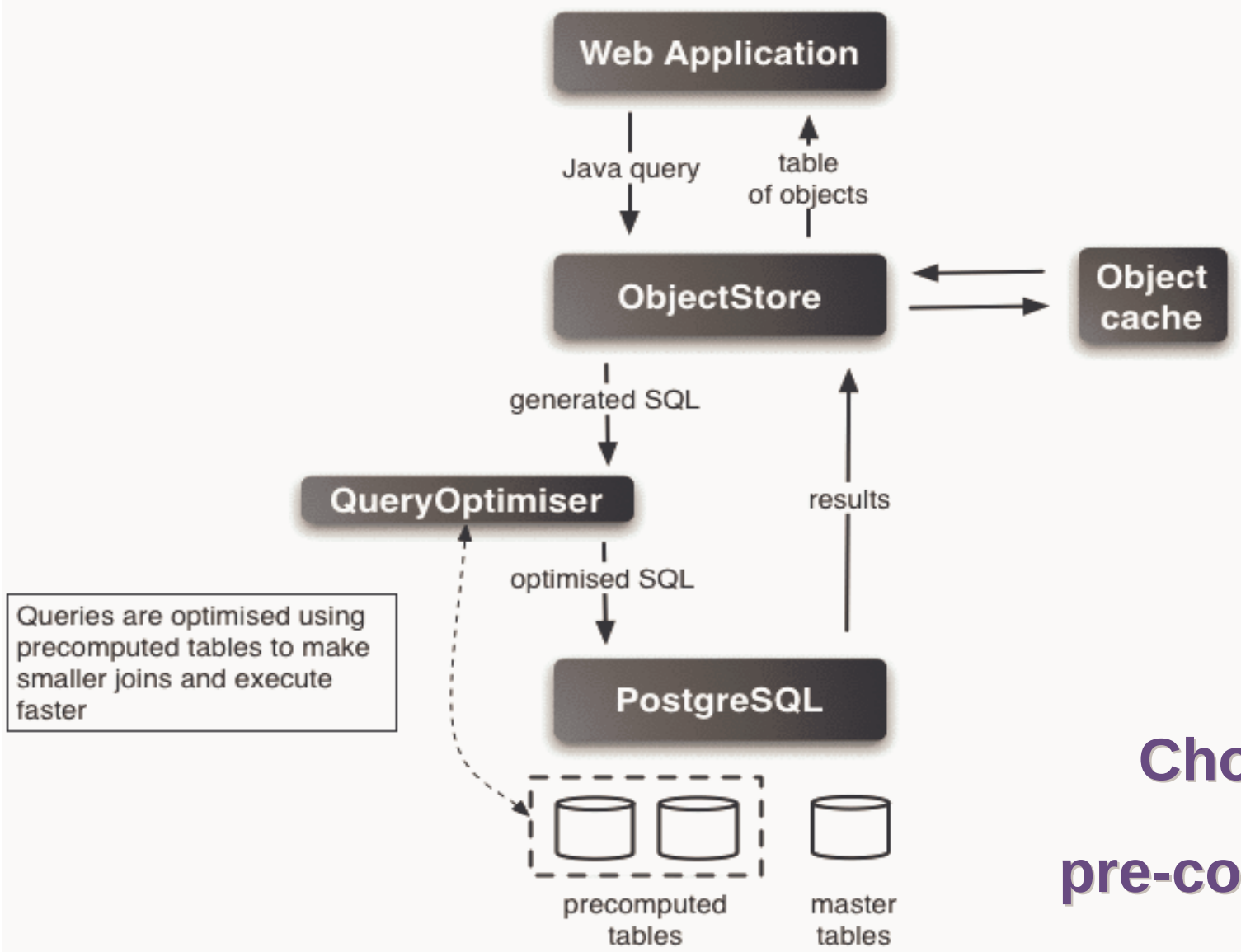
milkMine

modENCODE DCC

Project Stats

- Team of 7 FTE
  - 5 developers, one sys admin,
  - 1 biologist/ bioinformatician
- Java/ postgreSQL
- Struts/JSP/Ajax for webapp)
- Open Source
- SVN: 125,000 lines of code
- 57,000 lines of tests

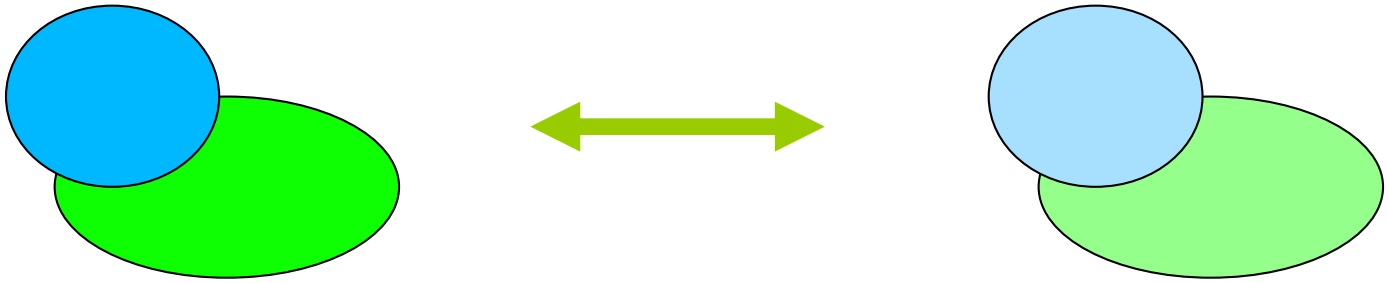
InterMine Query Optimisation



Choice of pre-computes?

Query Complexity

- **Interologs**

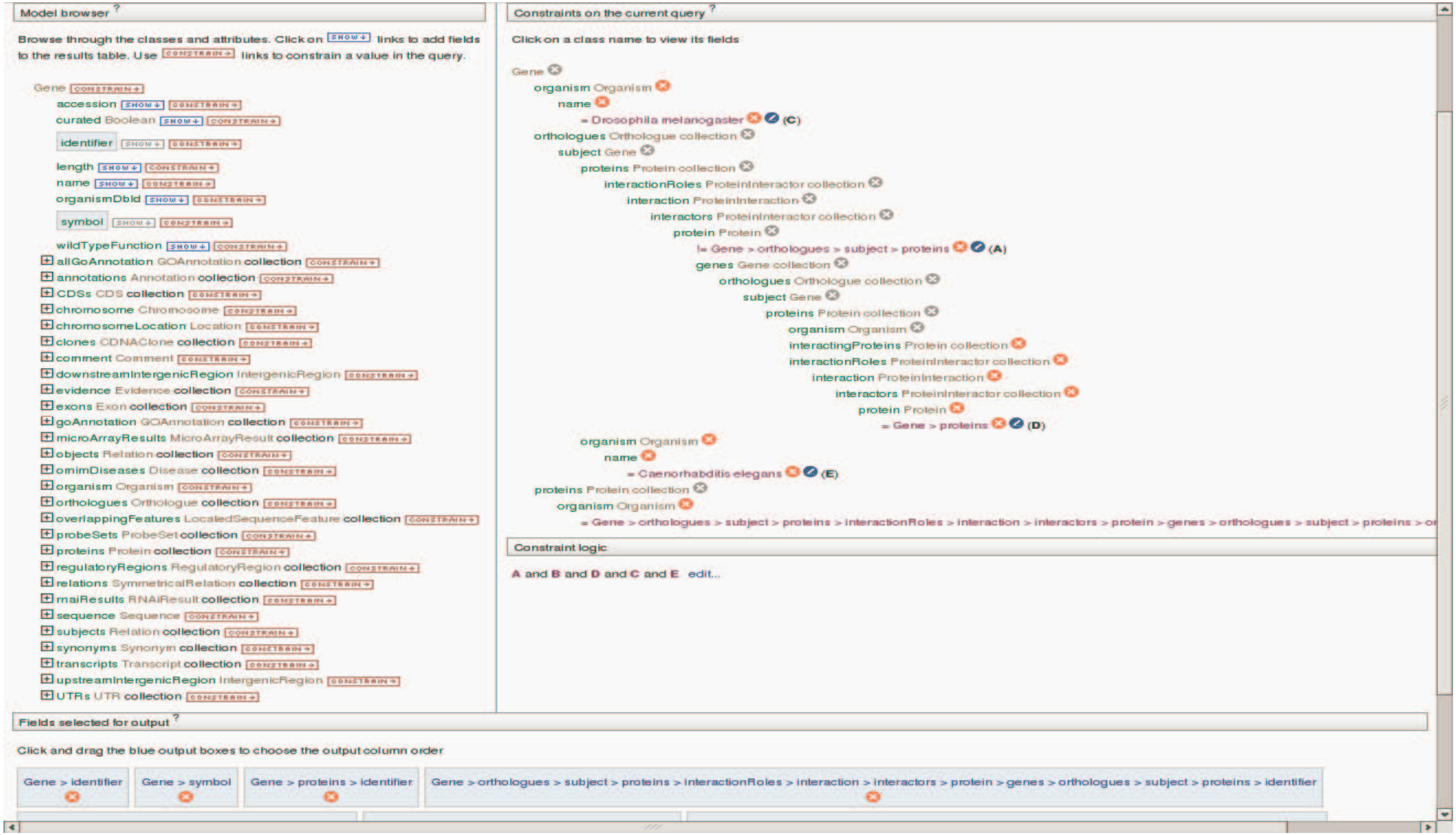


D. melanogaster *C. elegans*

- **Encapsulation**

  Query templates

Complex Query: Search for Interologs



Complex Query simplified as a template

This is a template query - edit the values below

All pairs of interacting proteins in organism1 --> All pairs of interacting proteins in organism2.

[1] *For proteins that interact in this organism:*

Organism name: = Drosophila melanogaster

or constrain to be IN bag test_genes1

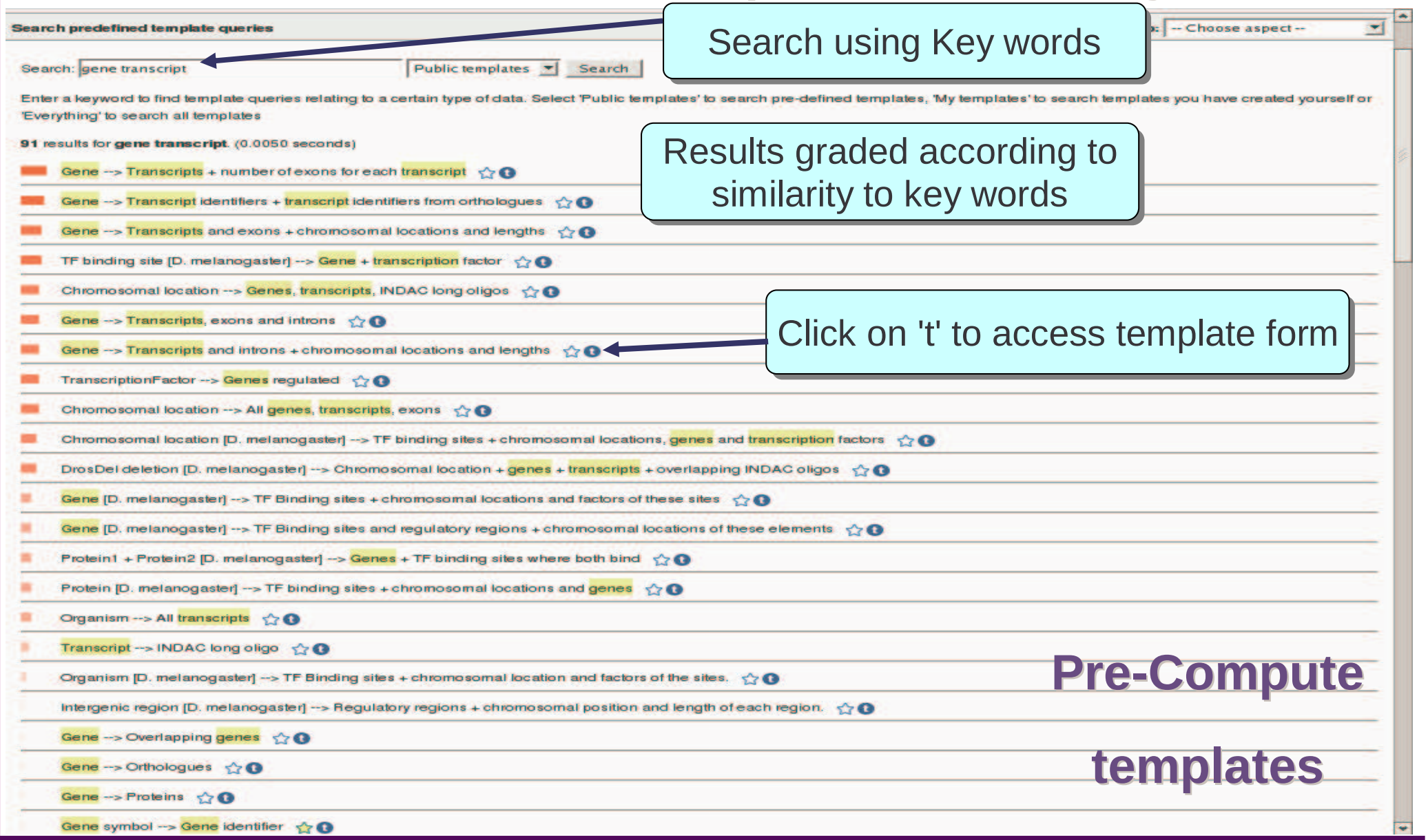
[2] *search for interologues in the following organism:*

Organism name: = Caenorhabditis elegans

or constrain to be IN bag test_genes1

Show Results | Edit Query

Search Template Library

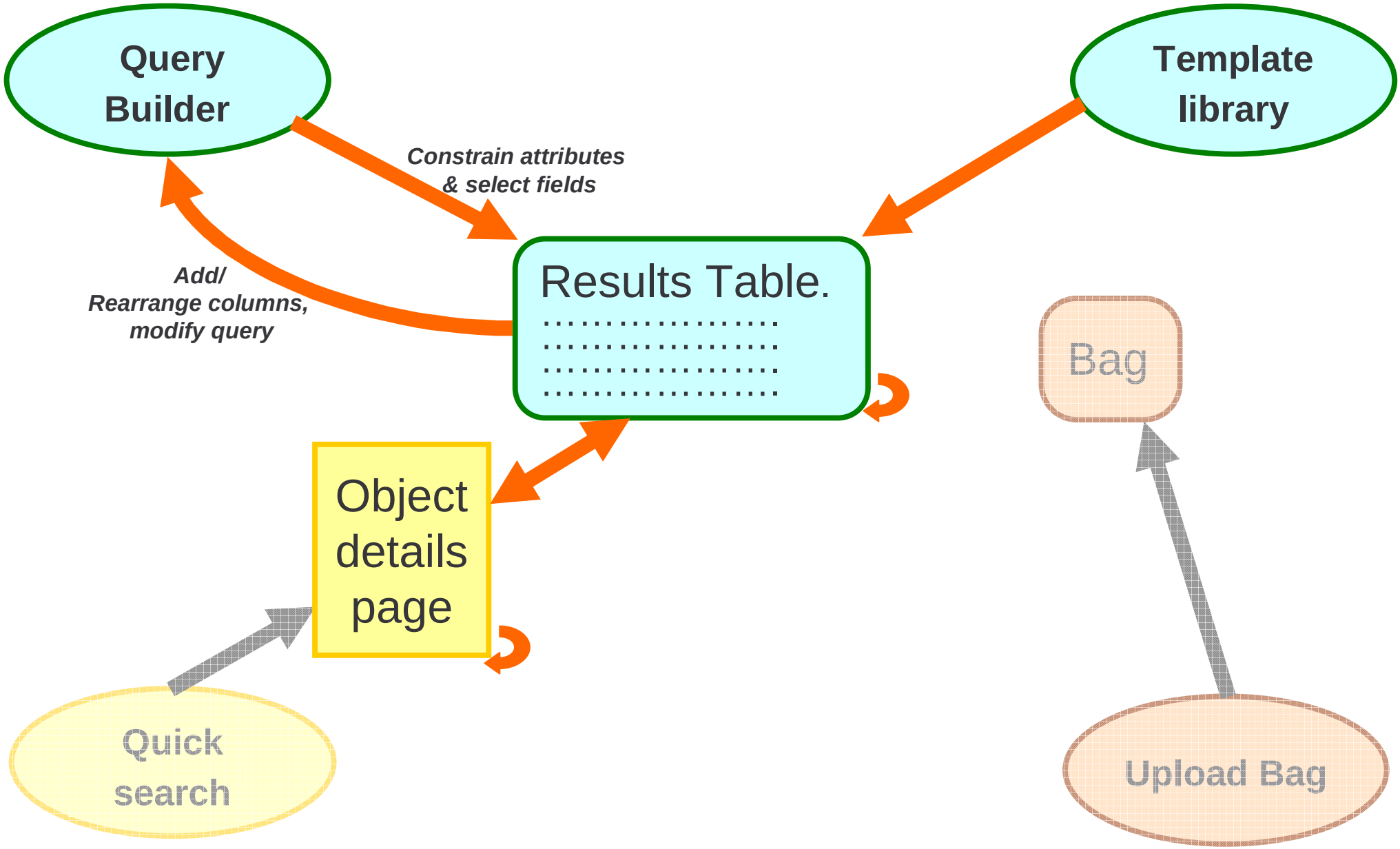


Search using Key words

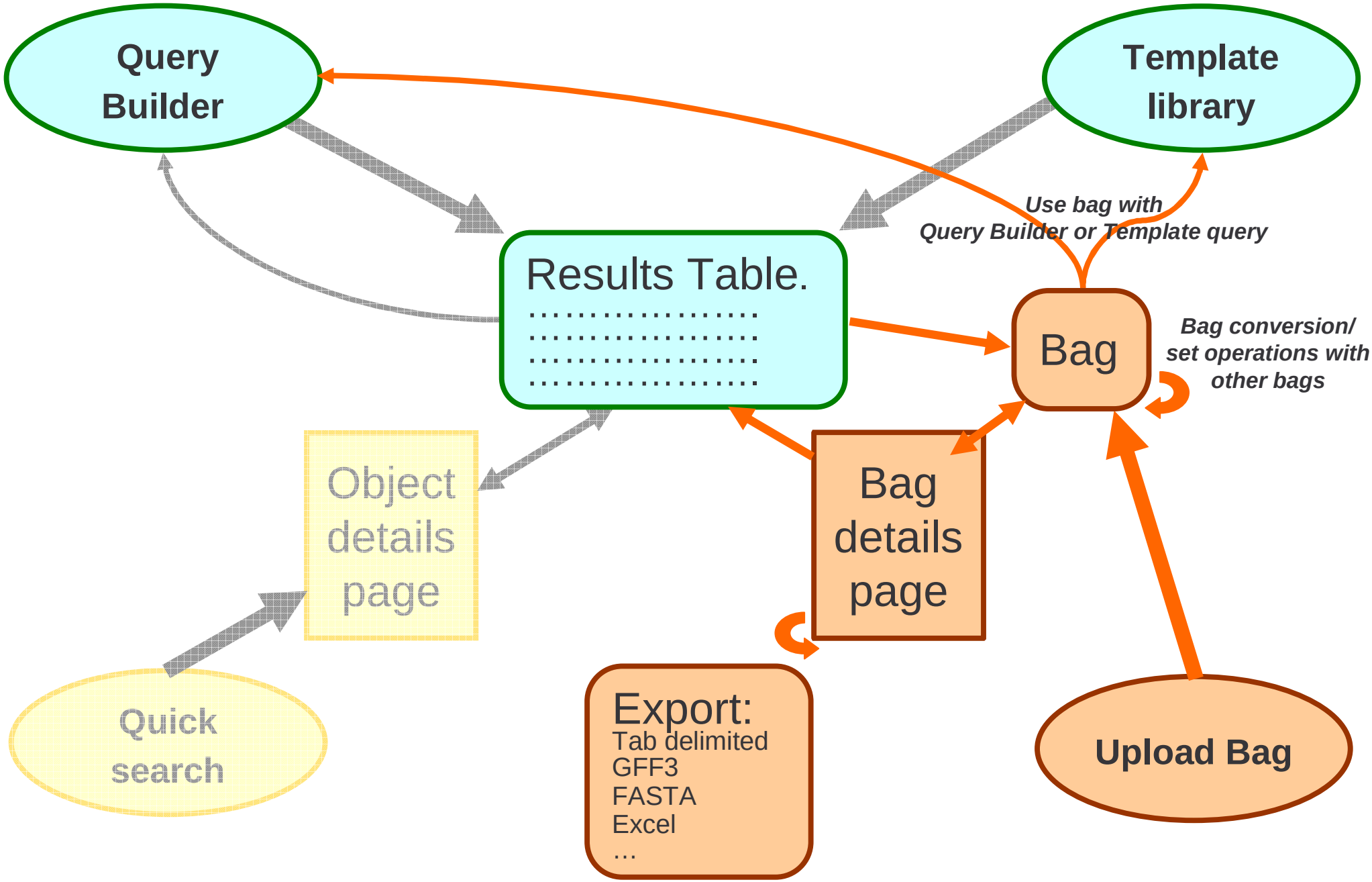
Results graded according to similarity to key words

Click on 't' to access template form

Pre-Compute templates

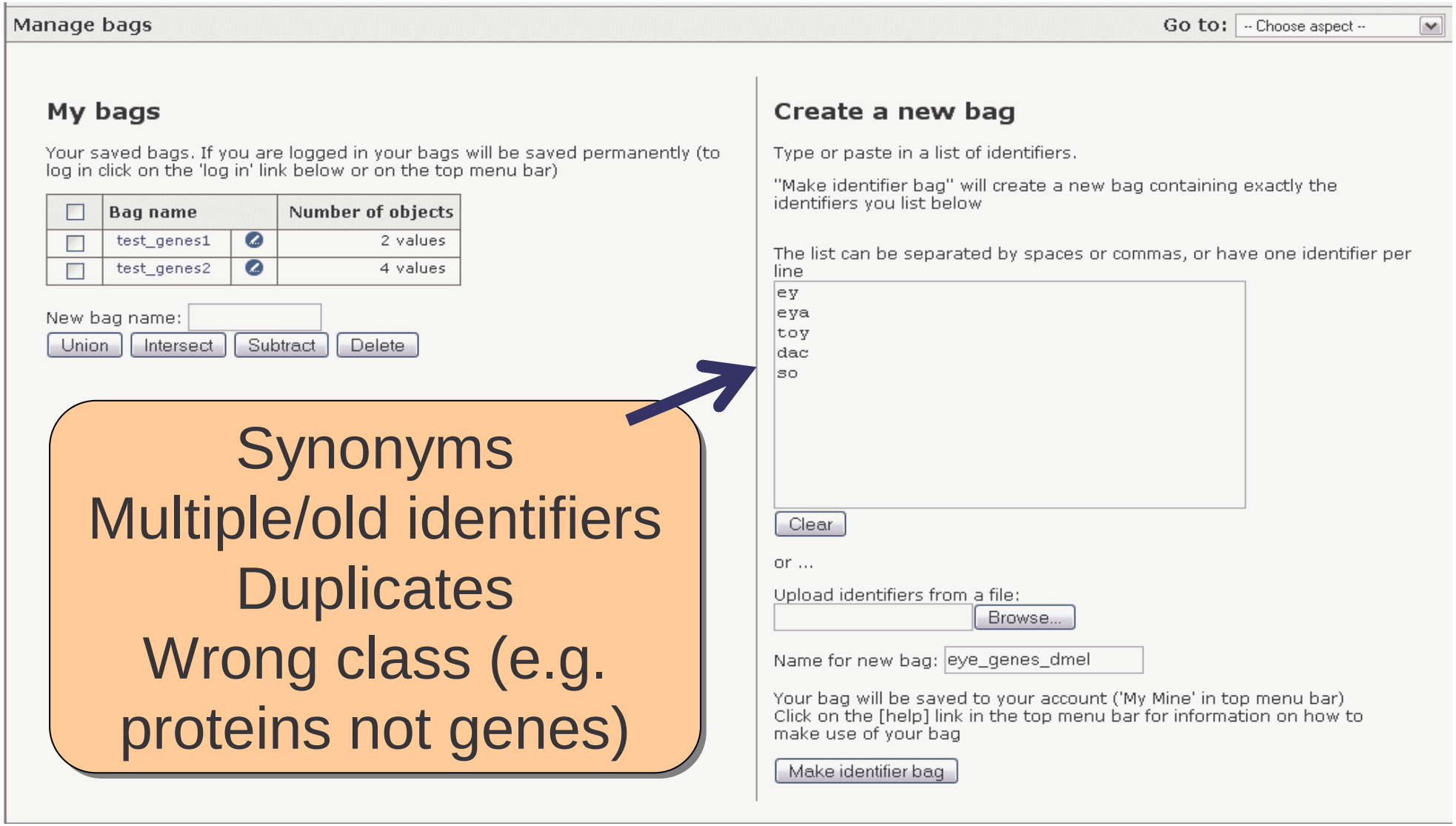


Results Table

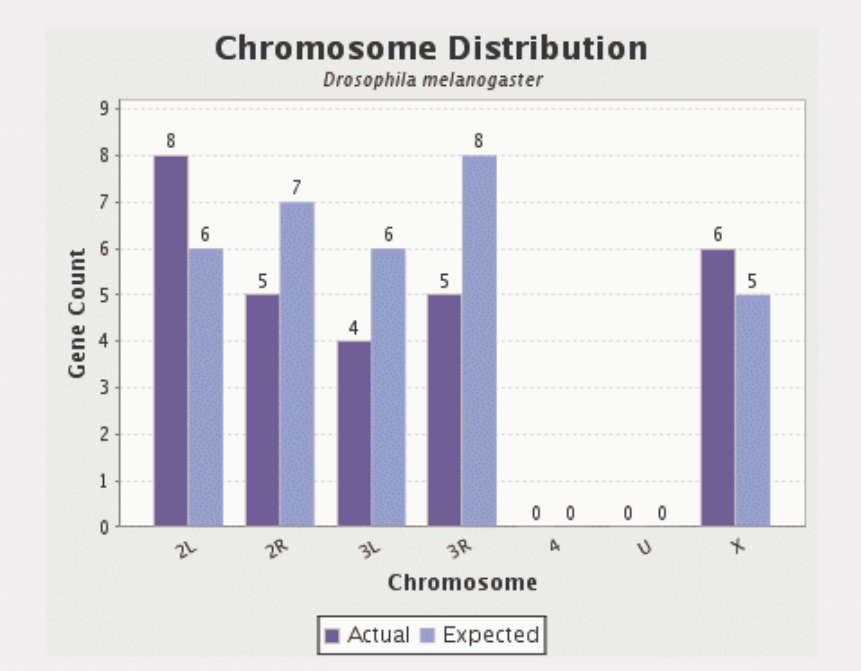

Query Builder
Template library
Use bag with
Query Builder or Template query
Results Table.
Bag
Bag conversion/
set operations with
other bags
Object details page
Bag details page
Quick search
Export:
Tab delimited
GFF3
FASTA
Excel
...
Upload Bag

Bags

Bag upload


Manage bags
Go to: -- Choose aspect --
My bags
Your saved bags. If you are logged in your bags will be saved permanently (to log in click on the 'log in' link below or on the top menu bar)
Bag name
Number of objects
test_genes1
2 values
test_genes2
4 values
New bag name:
Union
Intersect
Subtract
Delete
Create a new bag
Type or paste in a list of identifiers.
"Make identifier bag" will create a new bag containing exactly the identifiers you list below
The list can be separated by spaces or commas, or have one identifier per line
ey
eya
toy
dac
so
Clear
or ...
Upload identifiers from a file:
Browse...
Name for new bag: eye_genes_dmel
Your bag will be saved to your account ('My Mine' in top menu bar)
Click on the [help] link in the top menu bar for information on how to make use of your bag
Make identifier bag
Synonyms
Multiple/old identifiers
Duplicates
Wrong class (e.g. proteins not genes)

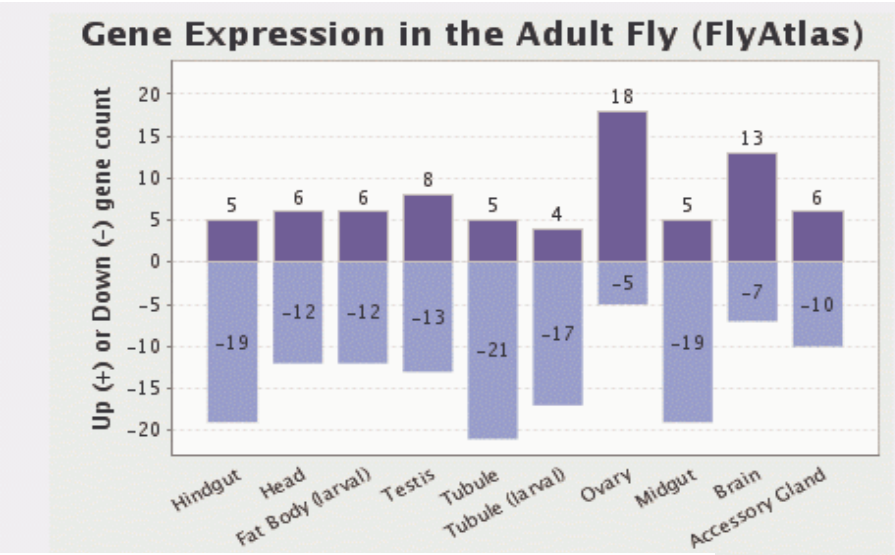
Bag Details Page



Pathway Information (KEGG)

Pathway > identifier	Pathway > name	Genes
dme03022	Basal transcription factors	15
dme04350	TGF-beta signaling pathway	6
dme04310	Wnt signaling pathway	3
dme04630	Jak-STAT signaling pathway	1
dme04330	Notch signaling pathway	1

Most common KEGG pathways for this bag (see here).



Gene Ontology Enrichment

GO terms that are enriched for genes in this bag compared to the reference population. Smaller p-values show greater enrichment. Method: Hypergeometric test with Bonferroni error correction (using a significance value of 0.05).

Reference population: All genes from [Drosophila melanogaster, Homo sapiens, Saccharomyces cerevisiae].

Ontology: Biological process — Update results

GO Term	p-value	
regulation of transcription [GO:0045449]	5.9246496E-44	[31 genes]
regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolism [GO:0019219]	6.3077770E-43	[31 genes]
regulation of transcription, DNA-dependent [GO:0006355]	1.5726087E-42	[30 genes]
regulation of cellular metabolism [GO:0031323]	9.2044437E-42	[31 genes]
transcription initiation from RNA polymerase II promoter [GO:0006367]	2.3488679E-41	[20 genes]
regulation of metabolism [GO:0019222]	4.4976530E-41	[31 genes]

ag for which the
cording to FlyAtlas

Discretisation? Up/down, p(up), p(down)

Acknowledgements

Richard Smith
Kim Rutherford
Matthew Wakeling
Xavier Watkins
Julie Sullivan
Rachel Lyne
Hilde Janssens
François Guillier
Philip North

Andrew Varley, Mark Woodbridge, Tom Riley, Peter Mclaren, Debashis Rana, Wenyan Ji, Markus Brosch, Florian Reisinger

www.flymine.org www.intermine.org

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