

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



Genomics

Genome annotation



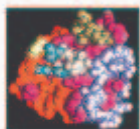
DrosDel

P-element insertions and deletions



RNAi

RNA interference

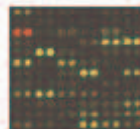


Proteins

Protein and proteomics data



Gene Ontology



Gene Expression

ArrayExpress



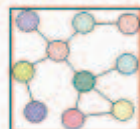
Tiling Path

Microarray Tiling Primers



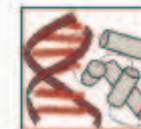
Disease

Human disease matches from Homophila



Protein Interactions

IntAct



Transcriptional Regulation

Regulatory regions and transcription factor binding sites



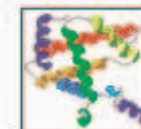
INDAC

Long oligos from the International *Drosophila* Array Consortium



Comparative Genomics

Orthologues and paralogues



Protein Structure

3-D protein structures

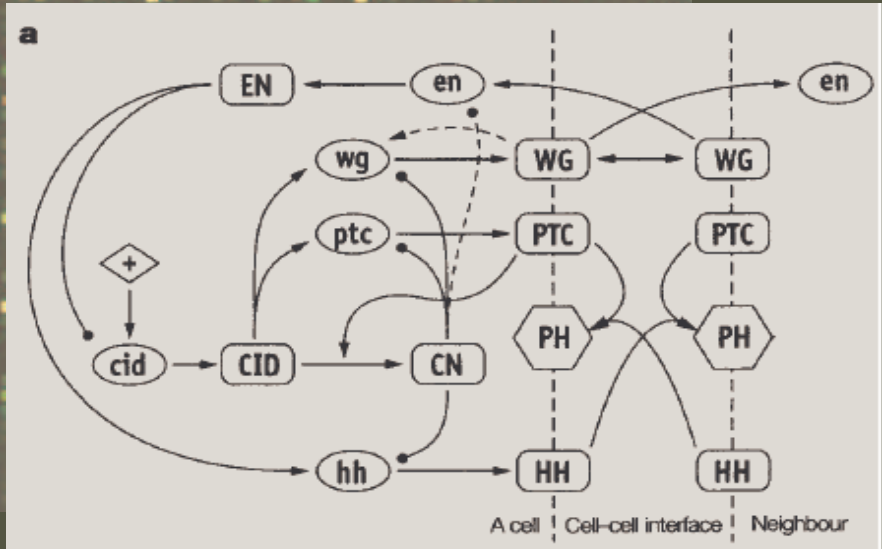
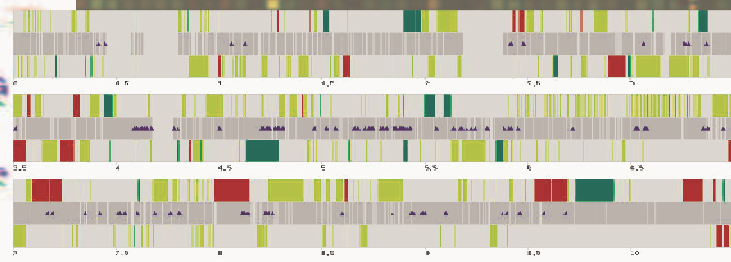
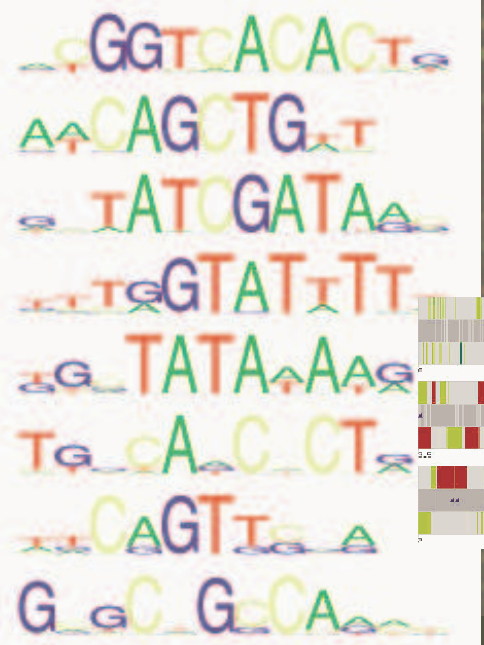
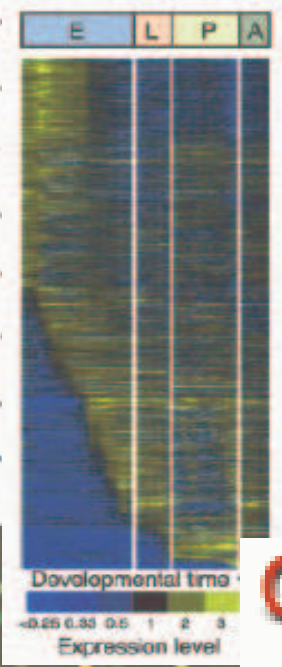
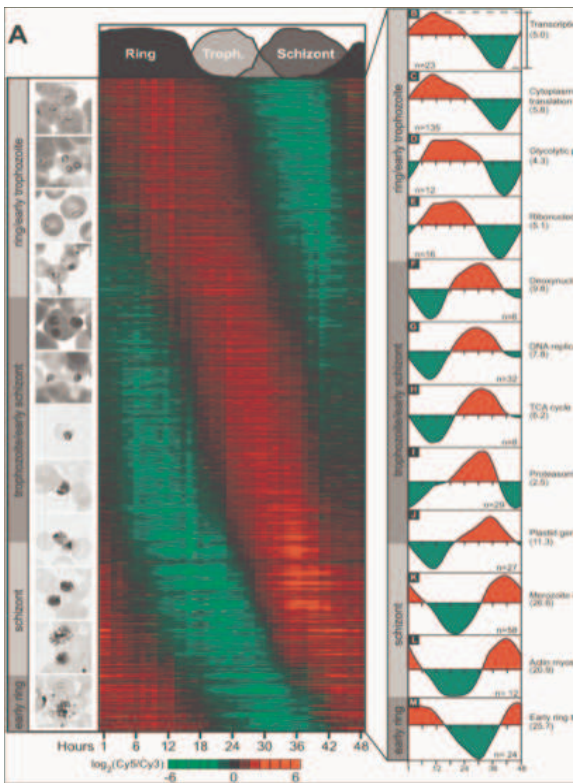
Gos Micklem



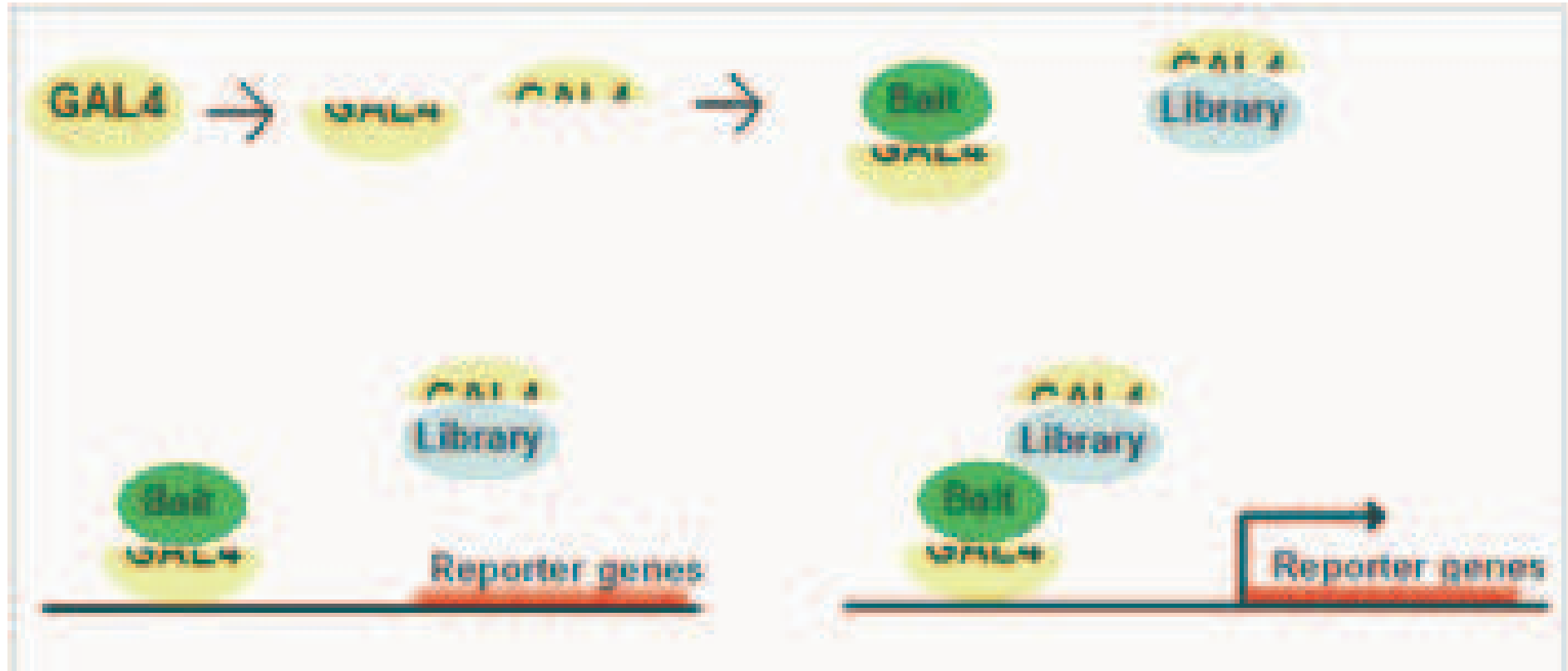
www.flymine.org



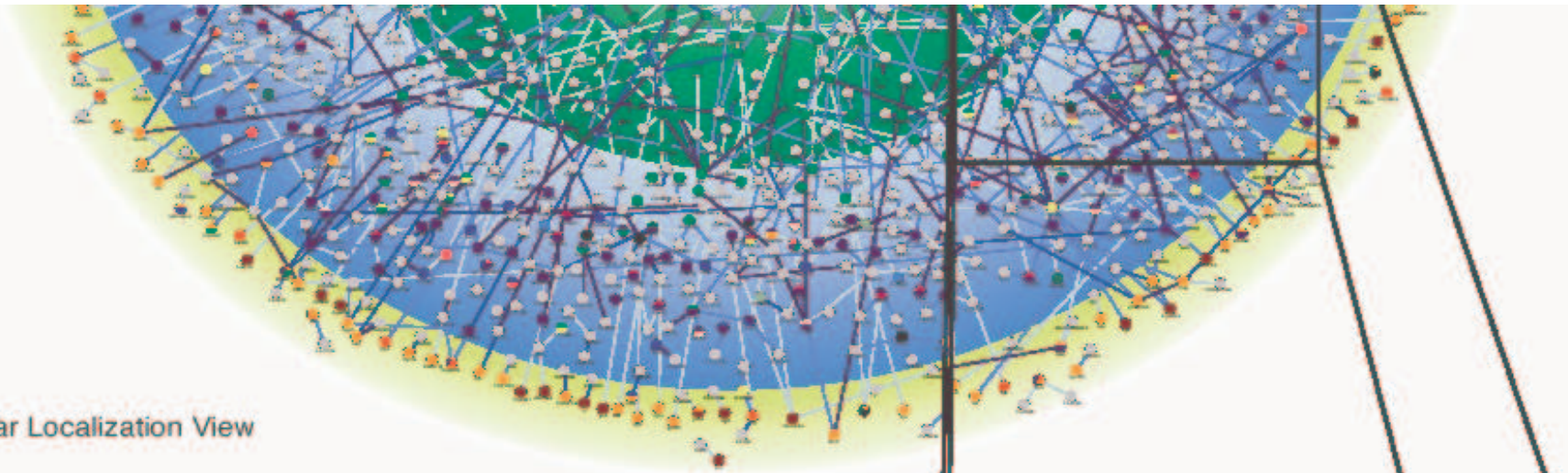
UNIVERSITY OF
CAMBRIDGE



Yeast 2-hybrid screening



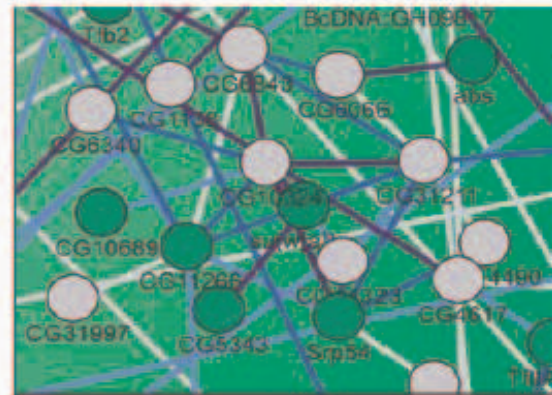
Drosophila



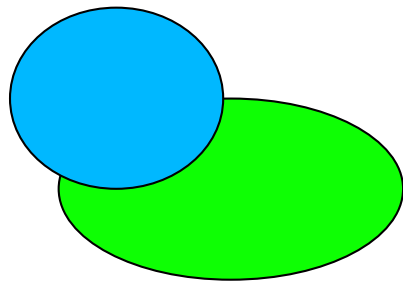
Sub-Cellular Localization View

- 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

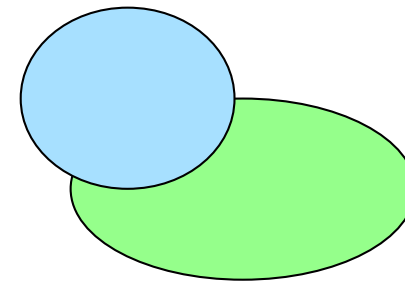
Interaction Ratings	
0.9 - 1.0	Thick dark blue line
0.8 - 0.9	Medium dark blue line
0.65 - 0.8	Medium light blue line
< 0.65	Thin light blue line



Interologs



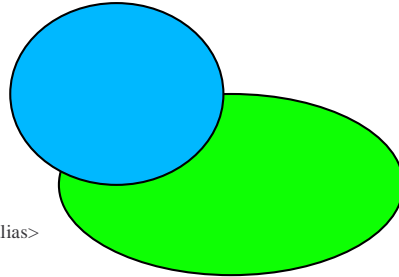
D. melanogaster



C. elegans



PSI for Drosophila

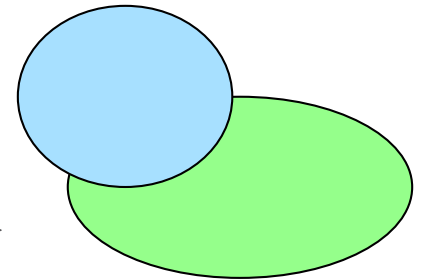


```

<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>

```

PSI data for worm:



```

<interactor id="262">
  <names>
    <shortLabel>q8mxt7_caee1</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_caee1" 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_caee1"/>
  </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>caee1</shortLabel>
      <fullName>Caenorhabditis elegans</fullName>
    </names>
  </organism>

```



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%

Standard data formats?

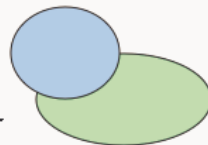


PSI for Drosophila



```
<code><pre></pre></code>
```

PSI data for worm:



```
<code><pre></pre></code>
```

In Paracnid fly/worm orthologues

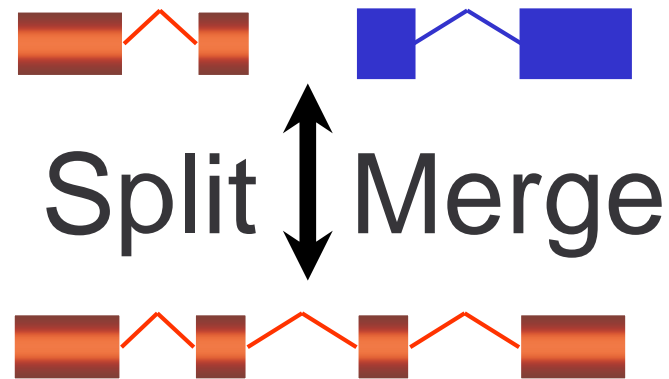
Table with 4 columns: ID, Species, PSI, and Coverage. It lists orthologous genes between Drosophila and worm.

Nothing!



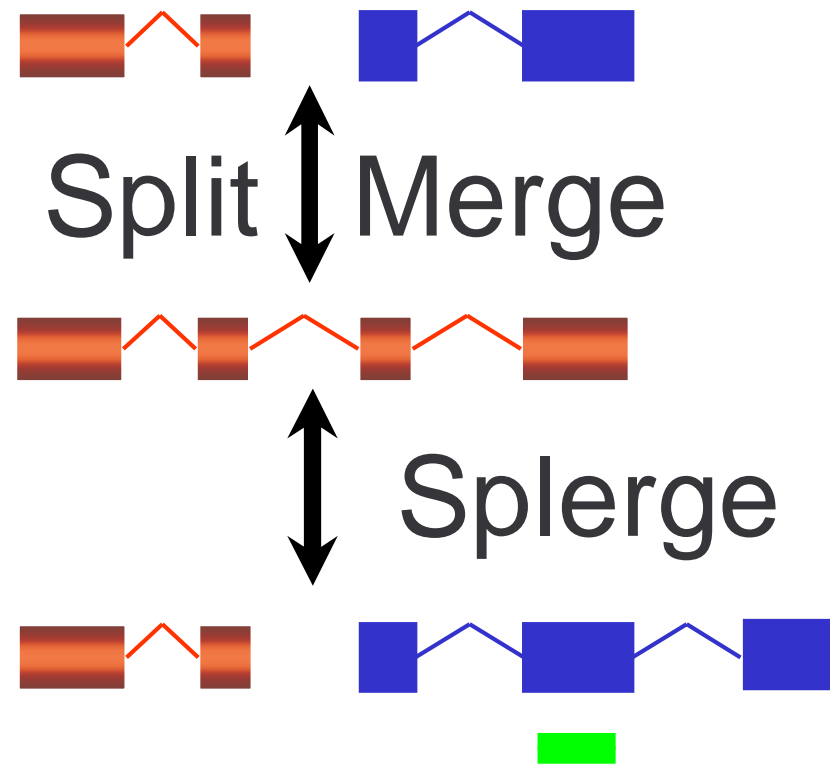
Genomes

Sequence, annotation not stable



Some MODs track annotation history

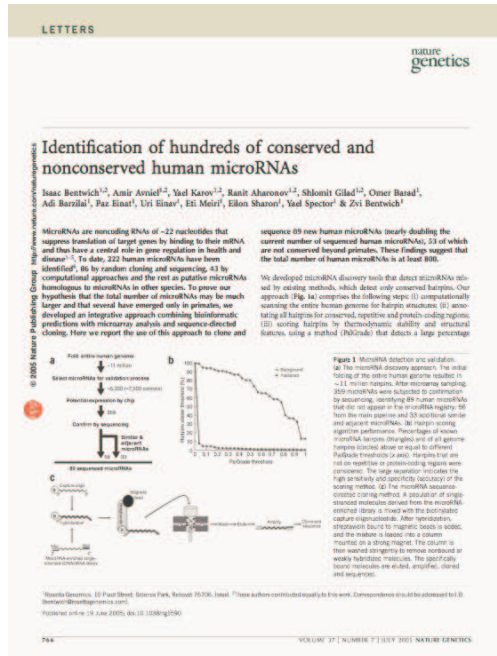




Over time a single microarray **probe** can assay 'different' genes



Fund, publish, freeze



LETTERS

nature genetics

Identification of hundreds of conserved and nonconserved human microRNAs

Isaac Benayahu^{1,2}, Amir Arad^{1,2}, Yael Katzir^{1,2}, Ranit Aharonov^{1,2}, Shoshit Gilad^{1,2}, Ofer Basrai¹, Adi Barzilai¹, Paz Einafi¹, Uri Eliner¹, Eli Meiri¹, Eilon Sharoni^{1,2}, Yael Spector¹ & Zvi Benayahu¹

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 disease¹. To date, 222 human microRNAs have been identified. We used a computational approach to identify conserved and nonconserved human 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, 23 of which are not conserved beyond primates. These findings suggest that the total number of human microRNAs is at least 800.

We developed microRNA discovery tools that detect microRNAs using a variety of methods, which detect only conserved hairpins. Our approach (Fig. 1a) comprises the following steps: (i) computationally scanning the entire human genome for hairpin structures; (ii) annotating all hairpins for conserved, repetitive and possibly coding regions; (iii) sorting hairpins by thermodynamic stability and structural features using a method (DfMicro) that detects a large percentage

FIGURE 1 MicroRNA discovery pipeline. (a) The search pipeline for microRNAs. The initial search of the entire human genome resulted in ~1.1 million hairpins. After removing overlapping hairpins, 200 microRNAs were subjected to confirmation by sequencing, identifying 89 human microRNAs that do not appear in the miRBase registry. 50 hairpins that appear in the miRBase registry and adjacent microRNAs were removed. (b) Heatmap showing sequence conservation of microRNAs across species. The large number of conserved microRNAs indicates the high conservation and evolutionary success of the mature sequence. (c) The microRNA discovery pipeline. A sequence of single-stranded molecules derived from the microRNA precursor is inserted into the 3' UTR of a luciferase reporter construct. After transfection, the reporter construct is transfected into cells and the luciferase activity is measured. The mature microRNA sequence is inserted into a column 'fused' on a string of poly-U. The column is then washed stringently to remove nonbound or weakly hybridized molecules. The specifically bound molecules are eluted, amplified, cloned and sequenced.

© 2005 Nature Publishing Group



Cell, Volume 130

Supplemental Data

The Mitron 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: GGGggggcggGTGAACGACTGCAACAGCA
rev: GgggtagaCTTGCCGTCTCTCCTTTC

miR-1010 for: GGGggggcggAAGGGGACCTTATCGATGT
rev: GgggtagaGGTTGAGAATGCCAGGTAA

miR-1004 for: GGGggggcgggggttccatctgtag
rev: Ggggtagaaggcgagcatctcttaga
```

2. 'Intron' miRtron expression constructs.

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

```
5' ATG GGC GAA AGC GAT TCT TCG AAG AAG GACT TGA TTA AAG GAG GGC AAT TCG AAT GAG 3'
   ATG GGC GAA AGC GAT TCT TCG AAG AAG GACT TGA TTA AAG GAG GGC AAT TCG AAT GAG 3'
```

The top strand sequence of Zmyc is:

```
ATG GAG TTT CTT TCG AAG GAG GACT TGA TTA AAG GAG GGC AAT TCG AAT GAG
```

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

```
GTAGT TTT CTT TCG AAG GAG GACT TGA TTA AAG GAG GGC AAT TCG AAT GAG
```

- 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





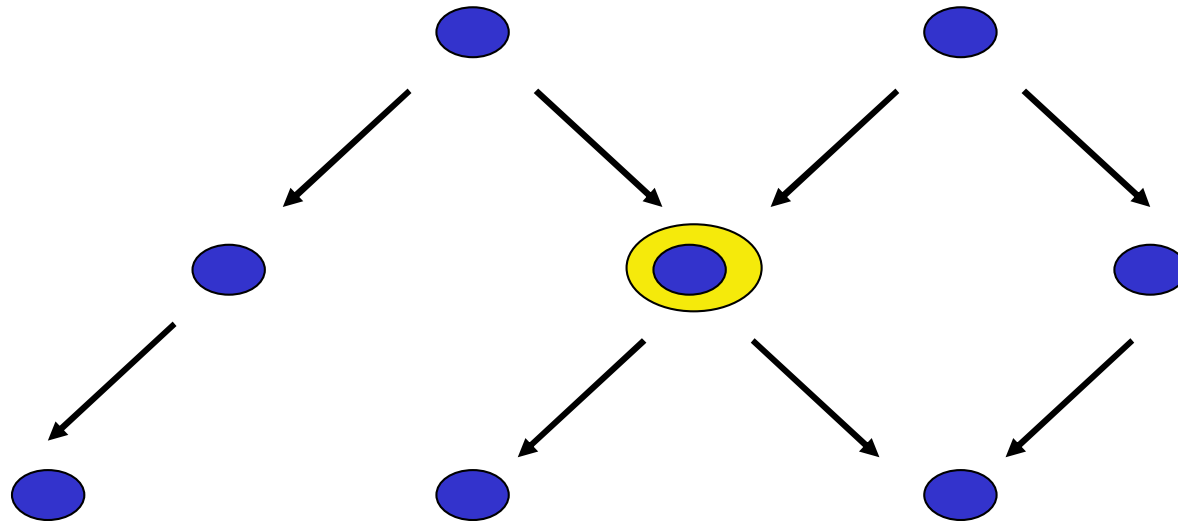
The Three Gene Ontologies

- *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

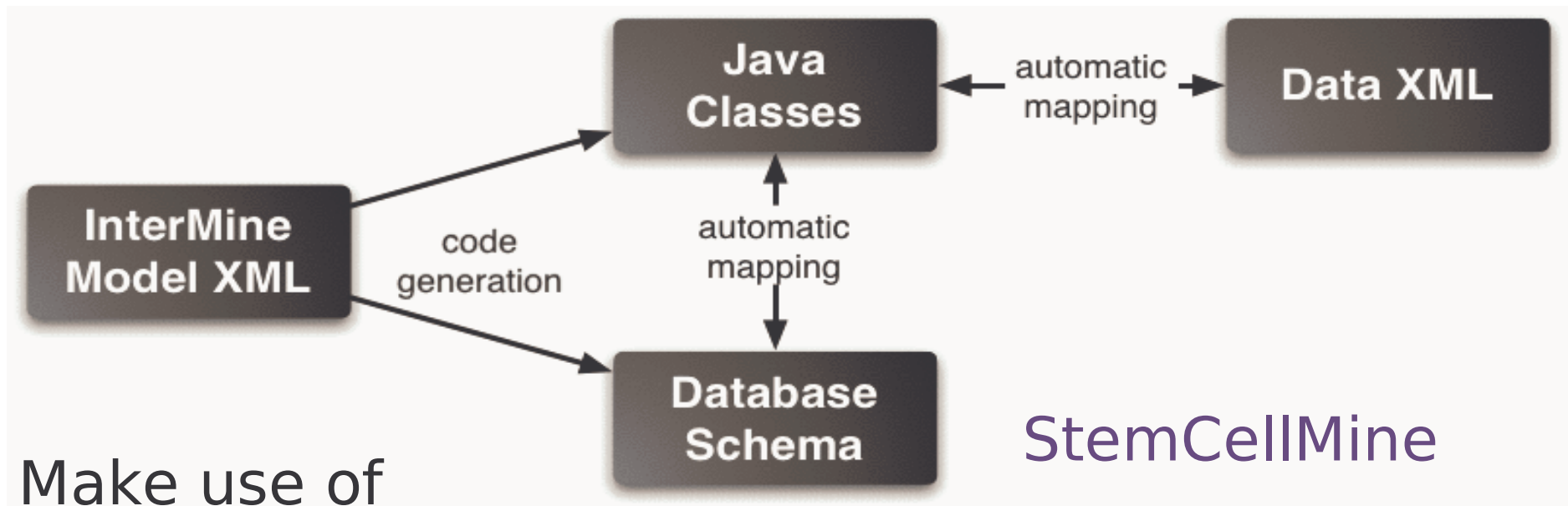
Operate on lists as easily as single entities

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

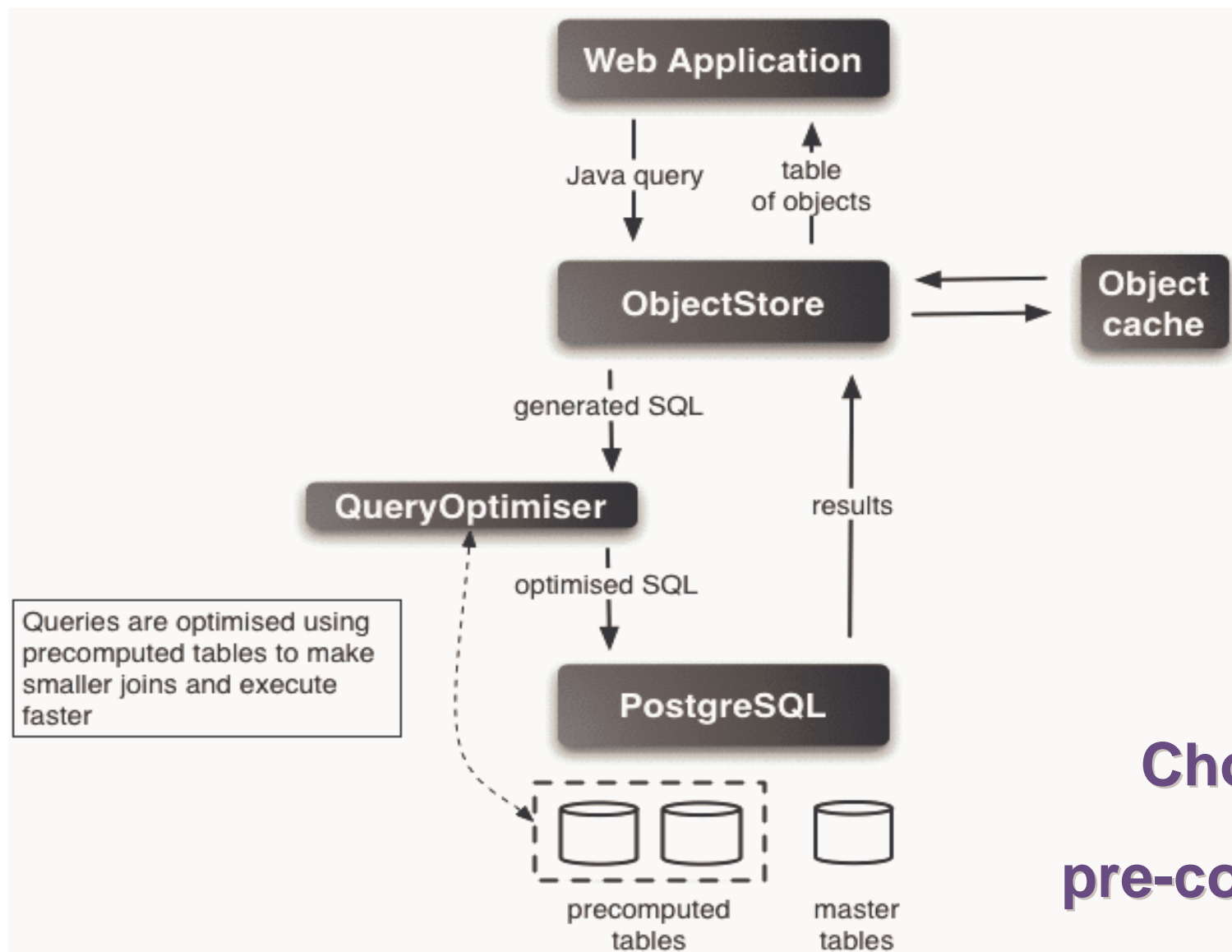


Project Stats

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



InterMine Query Optimisation

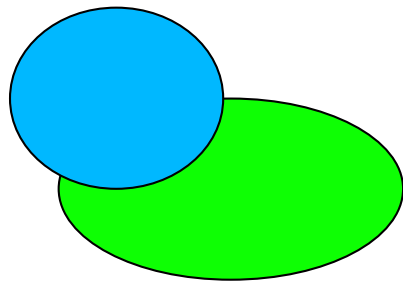


**Choice of
pre-computes?**

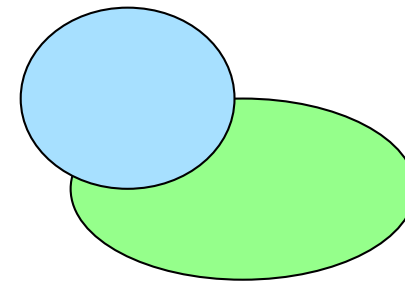


Query Complexity

- Interologs**



D. melanogaster



C. elegans

- Encapsulation**

Query templates



Complex Query: Search for Interologs

Model browser ?

Browse through the classes and attributes. Click on **SHOW** links to add fields to the results table. Use **CONSTRAIN** links to constrain a value in the query.

Gene **CONSTRAIN**+

accession **SHOW** **CONSTRAIN**+

curated Boolean **SHOW** **CONSTRAIN**+

identifier **SHOW** **CONSTRAIN**+

length **SHOW** **CONSTRAIN**+

name **SHOW** **CONSTRAIN**+

organismDbId **SHOW** **CONSTRAIN**+

symbol **SHOW** **CONSTRAIN**+

wildTypeFunction **SHOW** **CONSTRAIN**+

allGoAnnotation GOAnnotation collection **CONSTRAIN**+

annotations Annotation collection **CONSTRAIN**+

CDSs CDS collection **CONSTRAIN**+

chromosome Chromosome **CONSTRAIN**+

chromosomeLocation Location **CONSTRAIN**+

clones CDNAClone collection **CONSTRAIN**+

comment Comment **CONSTRAIN**+

downstreamIntergenicRegion IntergenicRegion **CONSTRAIN**+

evidence Evidence collection **CONSTRAIN**+

exons Exon collection **CONSTRAIN**+

goAnnotation GOAnnotation collection **CONSTRAIN**+

microArrayResults MicroArrayResult collection **CONSTRAIN**+

objects Relation collection **CONSTRAIN**+

omimDiseases Disease collection **CONSTRAIN**+

organism Organism **CONSTRAIN**+

orthologues Orthologue collection **CONSTRAIN**+

overlappingFeatures LocatedSequenceFeature collection **CONSTRAIN**+

probeSets ProbeSet collection **CONSTRAIN**+

proteins Protein collection **CONSTRAIN**+

regulatoryRegions RegulatoryRegion collection **CONSTRAIN**+

relations SymmetricalRelation collection **CONSTRAIN**+

rnaIResults RNAIResult collection **CONSTRAIN**+

sequence Sequence **CONSTRAIN**+

subjects Relation collection **CONSTRAIN**+

synonyms Synonym collection **CONSTRAIN**+

transcripts Transcript collection **CONSTRAIN**+

upstreamIntergenicRegion IntergenicRegion **CONSTRAIN**+

UTRs UTR collection **CONSTRAIN**+

Fields selected for output ?

Click and drag the blue output boxes to choose the output column order

Gene > identifier Gene > symbol Gene > proteins > identifier Gene > orthologues > subject > proteins > interactionRoles > interaction > interactors > protein > genes > orthologues > subject > proteins > identifier

Constraints on the current query ?

Click on a class name to view its fields

Gene

organism Organism

name

= Drosophila melanogaster (C)

orthologues Orthologue collection

subject Gene

proteins Protein collection

interactionRoles ProteinInteractor collection

interaction ProteinInteraction

interactors ProteinInteractor collection

protein Protein

= Gene > orthologues > subject > proteins (A)

genes Gene collection

orthologues Orthologue collection

subject Gene

proteins Protein collection

organism Organism

interactingProteins Protein collection

interactionRoles ProteinInteractor collection

interaction ProteinInteraction

interactors ProteinInteractor collection

protein Protein

= Gene > proteins (D)

organism Organism

name

= Caenorhabditis elegans (E)

proteins Protein collection

organism Organism

= Gene > orthologues > subject > proteins > interactionRoles > interaction > interactors > protein > genes > orthologues > subject > proteins > identifier

Constraint logic:

A and B and D and C and E edit..



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:
 or constrain to be bag

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

Organism name:
 or constrain to be bag

Show Results

Edit Query



Search Template Library

Search predefined template queries

Search: Choose aspect --

Enter a keyword to find template queries relating to a certain type of data. Select 'Public templates' to search pre-defined templates, 'My templates' to search templates you have created yourself or 'Everything' to search all templates

91 results for **gene transcript**. (0.0050 seconds)

- Gene --> Transcripts + number of exons for each transcript ☆ t
- Gene --> Transcript identifiers + transcript identifiers from orthologues ☆ t
- Gene --> Transcripts and exons + chromosomal locations and lengths ☆ t
- TF binding site [D. melanogaster] --> Gene + transcription factor ☆ t
- Chromosomal location --> Genes, transcripts, INDAC long oligos ☆ t
- Gene --> Transcripts, exons and introns ☆ t
- Gene --> Transcripts and introns + chromosomal locations and lengths ☆ t
- TranscriptionFactor --> Genes regulated ☆ t
- Chromosomal location --> All genes, transcripts, exons ☆ t
- Chromosomal location [D. melanogaster] --> TF binding sites + chromosomal locations, genes and transcription factors ☆ t
- DrosDel deletion [D. melanogaster] --> Chromosomal location + genes + transcripts + overlapping INDAC oligos ☆ t
- Gene [D. melanogaster] --> TF Binding sites + chromosomal locations and factors of these sites ☆ t
- Gene [D. melanogaster] --> TF Binding sites and regulatory regions + chromosomal locations of these elements ☆ t
- Protein1 + Protein2 [D. melanogaster] --> Genes + TF binding sites where both bind ☆ t
- Protein [D. melanogaster] --> TF binding sites + chromosomal locations and genes ☆ t
- Organism --> All transcripts ☆ t
- Transcript --> INDAC long oligo ☆ t
- Organism [D. melanogaster] --> TF Binding sites + chromosomal location and factors of the sites. ☆ t
- Intergenic region [D. melanogaster] --> Regulatory regions + chromosomal position and length of each region. ☆ t
- Gene --> Overlapping genes ☆ t
- Gene --> Orthologues ☆ t
- Gene --> Proteins ☆ t
- Gene symbol --> Gene identifier ☆ t

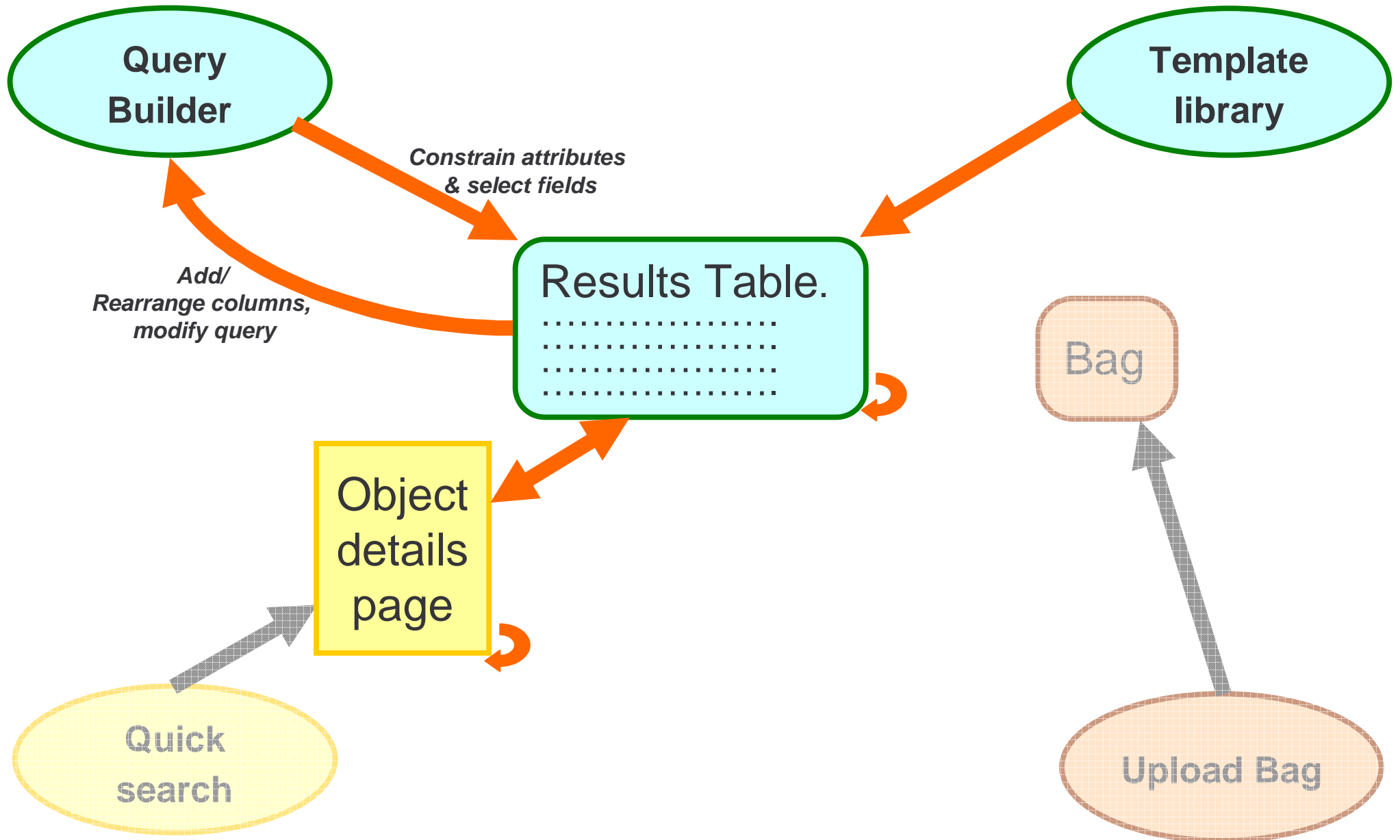
Search using Key words

Results graded according to similarity to key words

Click on 't' to access template form

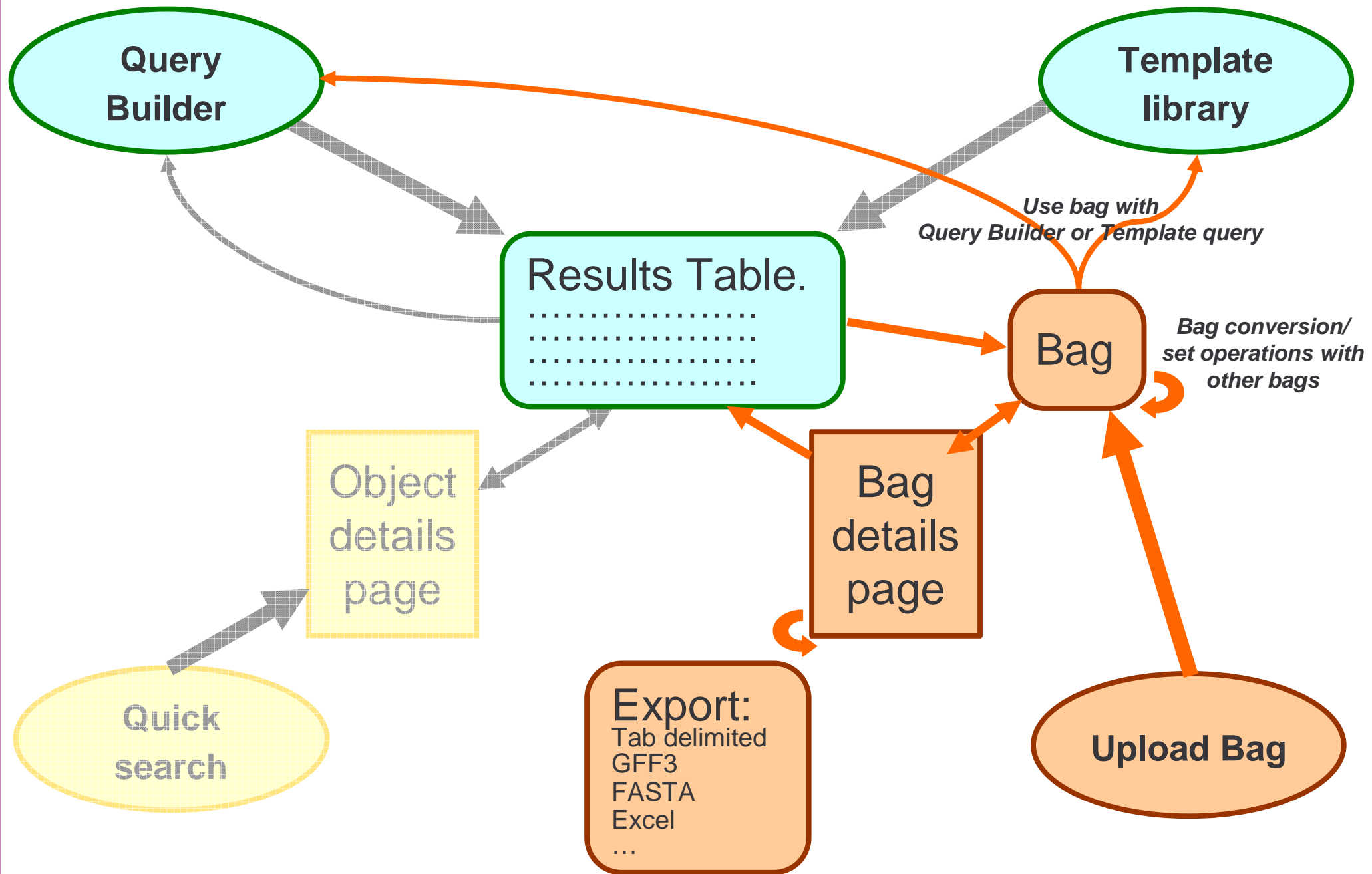
Pre-Compute templates





Results Table





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)

<input type="checkbox"/>	Bag name		Number of objects
<input type="checkbox"/>	test_genes1	<input checked="" type="checkbox"/>	2 values
<input type="checkbox"/>	test_genes2	<input checked="" type="checkbox"/>	4 values

New bag name:

Synonyms
Multiple/old identifiers
Duplicates
Wrong class (e.g.
proteins not genes)

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
```

or ...

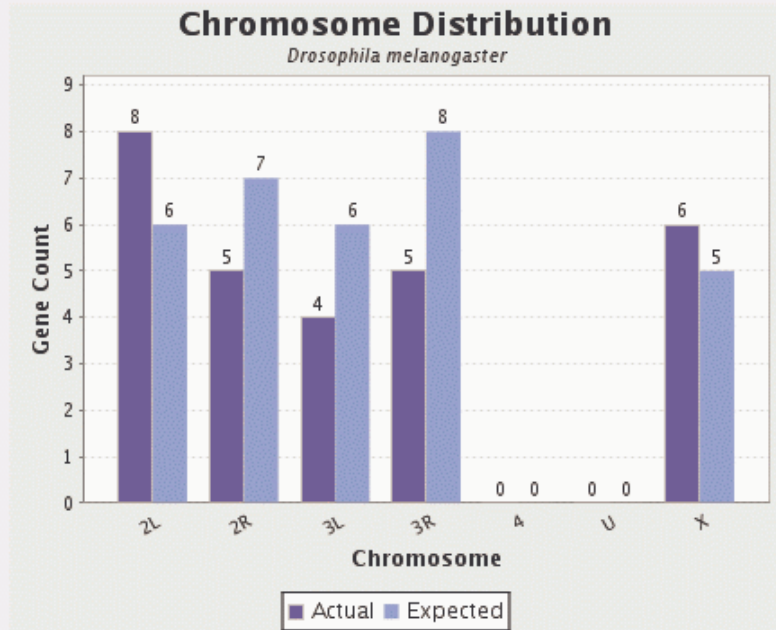
Upload identifiers from a file:

Name for new bag:

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



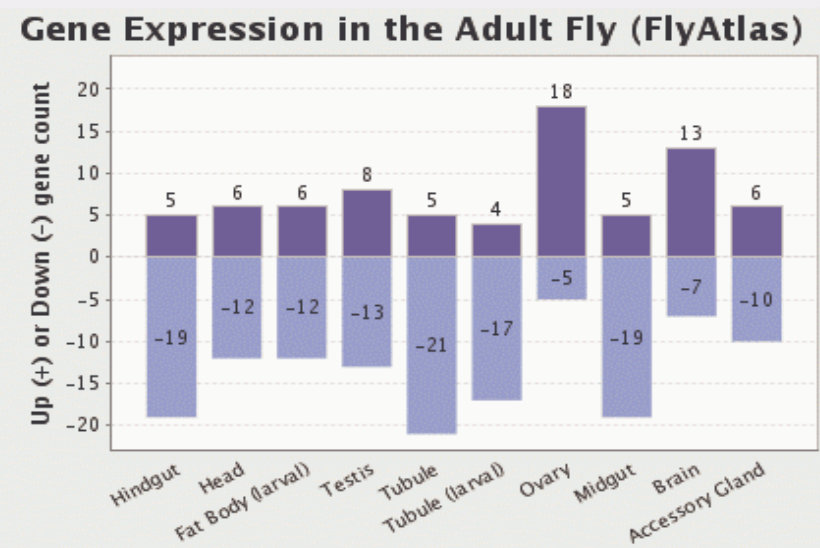
Bag Details Page



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:

GO Term	p-value	Genes
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]

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

bag for which the according to FlyAtlas

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.



www.flymine.org



UNIVERSITY OF
CAMBRIDGE