

Biological Data Analysis using InterMine

Rachel Lyne
Julie Sullivan
Yo Yehudi
Kevin Reierskog
June 19th 2019



Workshop Schedule

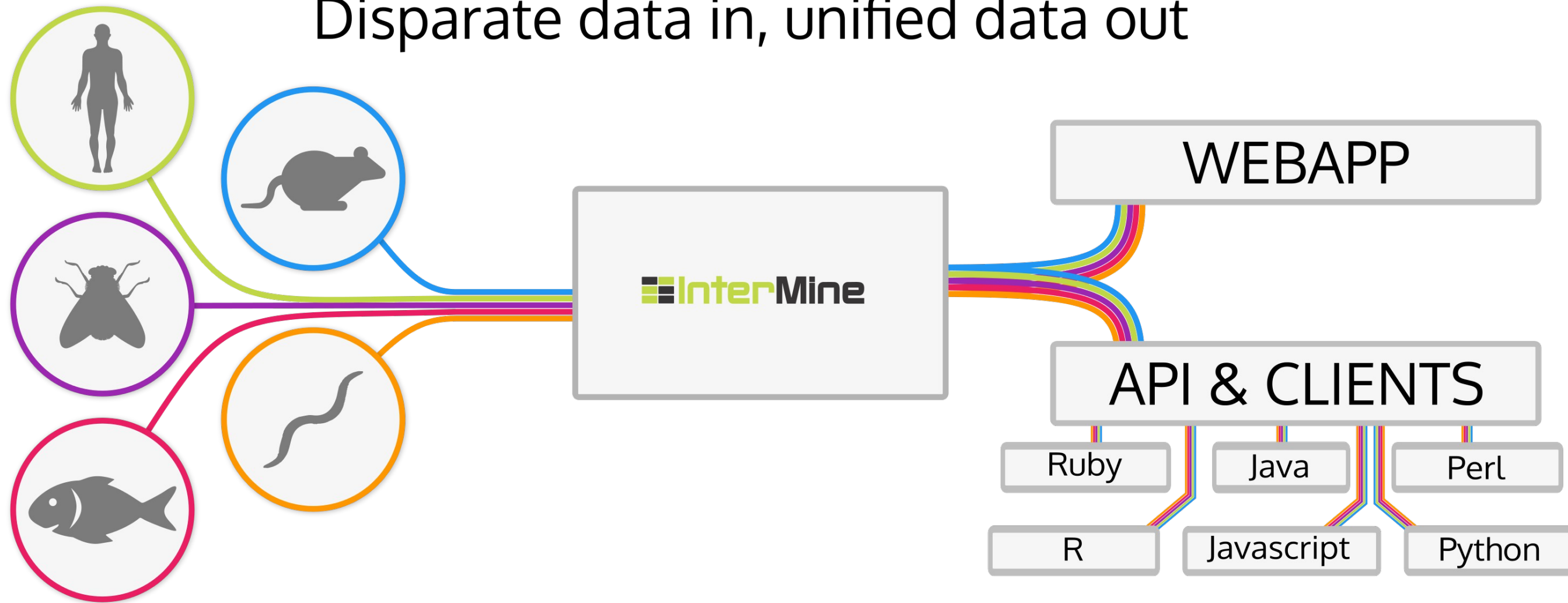
- Introduction
- Walk through interface with hands-on exercises
- Analysis workflows
- Questions and Advanced exercises

www.flymine.org

www.humanmine.org

What is InterMine

Disparate data in, unified data out




















Model organism images Designed by Freepik and distributed by Flaticon

Who Uses InterMine?



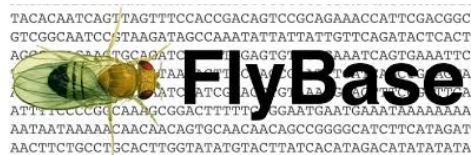
<http://registry.intermine.org/>

InterMine Registry All InterMine instances up-to-date information in one place.

	Name	Description	Organisms
	BMAP	Brassicales Map Alignment Project	
	BeanMine	A mine with common bean data from the Legume Info tripal.chado database	A. ipaensis, A. duranensis, A. thalia...
	BovineMine	An integrated data warehouse for the Bovine Genome Database	B. taurus, C. hircus, O. aries
	CHOmine	An integrated database for Cricetulus griseus and CHO cells	C. griseus
	ChickpeaMine	A mine with chickpea data (both desi and kabuli varieties) from the Legume...	A. ipaensis, A. duranensis, A. thalia...
	CowpeaMine	A mine containing both cowpea genetic and genomic data, courtesy UC-Riv...	A. duranensis, A. ipaensis, C. arietin...
	FlyMine	An integrated database for Drosophila genomics	D. melanogaster
	GrapeMine	An integrated database for grapevine data	
	HumanMine	HumanMine integrates many types of data for Homo sapiens and Mus mus...	H. sapiens
	HymenopteraMine	An integrated data warehouse for the Hymenoptera Genome Database	A. dorsata, A. echinatio, A. florea, A...
	IndigoMine	INDIGO enables the integration of annotations for the exploration and analy...	Archae
	LegumeMine	Multi-organism mine integrates data from legume species: string bean, soy...	A. duranensis, A. ipaensis, C arietin...
	MaizeMine	An integrated data warehouse for MaizeGDB	
	MedicMine	MedicMine integrates many types of data for Medicago truncatula. You can...	A. thaliana, M. truncatula, M. trunca...
	MitoMiner	MitoMiner is an integrated web resource of mitochondrial localisation evide...	D. rerio, H. sapiens, M. musculus, R...
	ModMine	A data warehouse for the modENCODE project	
	MouseMine	MouseMine is a powerful new system for online access to mouse data fro...	M. musculus



Data



TRANSCRIPTS

ORTHOLOGUES

GENES UTRs

GENE ONTOLOGY

INTERACTIONS

EXONS

PATHWAYS

PROTEINS

Protein Domains

GWAS

REGULATORY

VARIANTS

SNPs MICROARRAY

RNA-seq

ALLELES PHENOTYPES

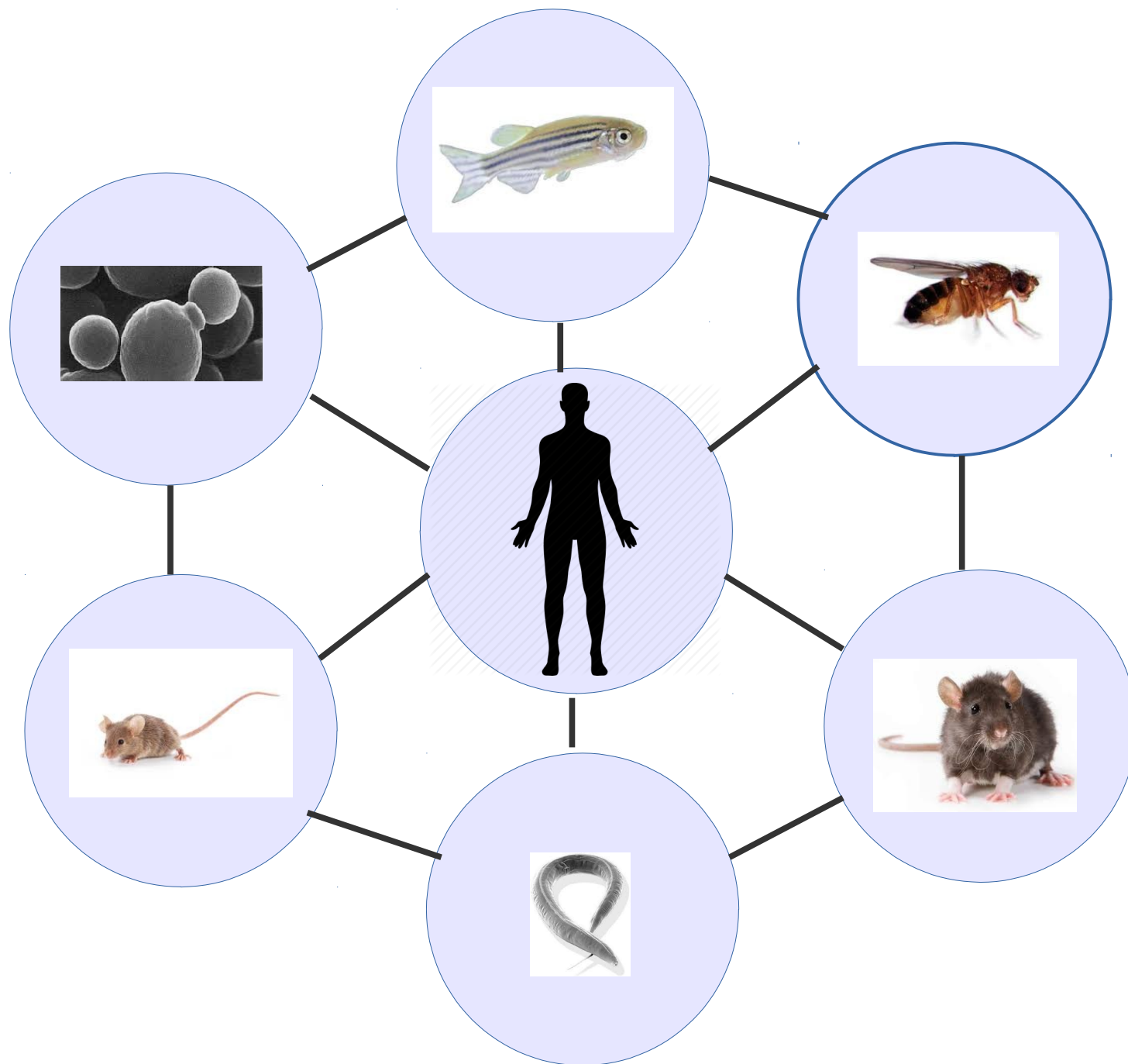
DISEASE

EXPRESSION

Why use InterMine?

- Query across several data sources at once
- Data formatting issues resolved
- Identifier resolution system
- Collate information about items and sets
- Common platform to many organisms and type of data
- Information without visiting several sites

Cross-organism analysis



InterMine Data Integration

Your own InterMine:

- Sophisticated build system
- Set integration keys
- Core library of data loaders
- Custom data loaders - own data

Help docs, mailing lists, etc: <http://www.intermine.org/>

Twitter: @intermineorg

Github: <http://www.github.com/intermine>

InterMine Accounts

InterMine is free to use without creating an account.

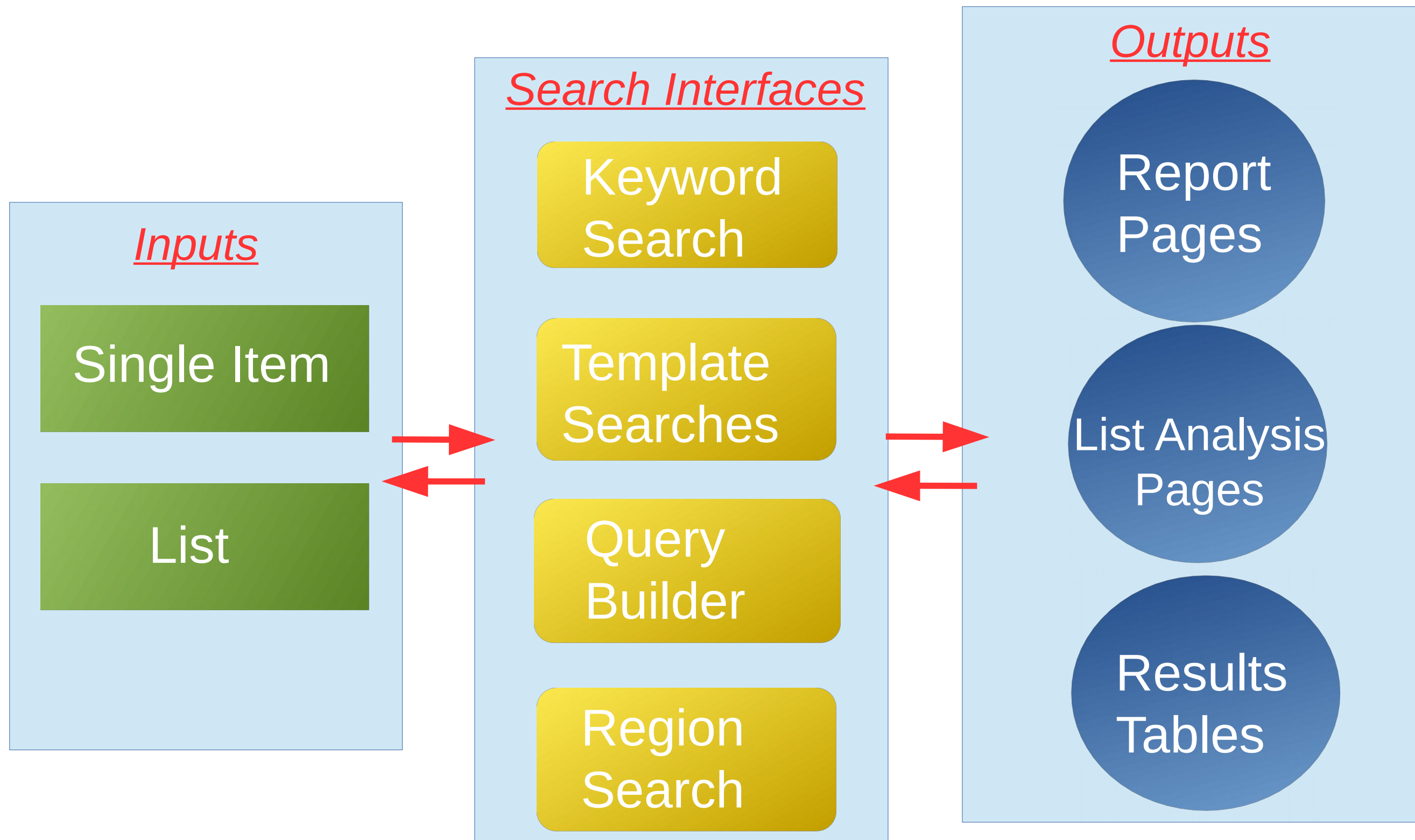
Creating an account allows you to save lists and searches permanently and share lists with your colleagues.

At the moment you have to make a separate account for each InterMine database

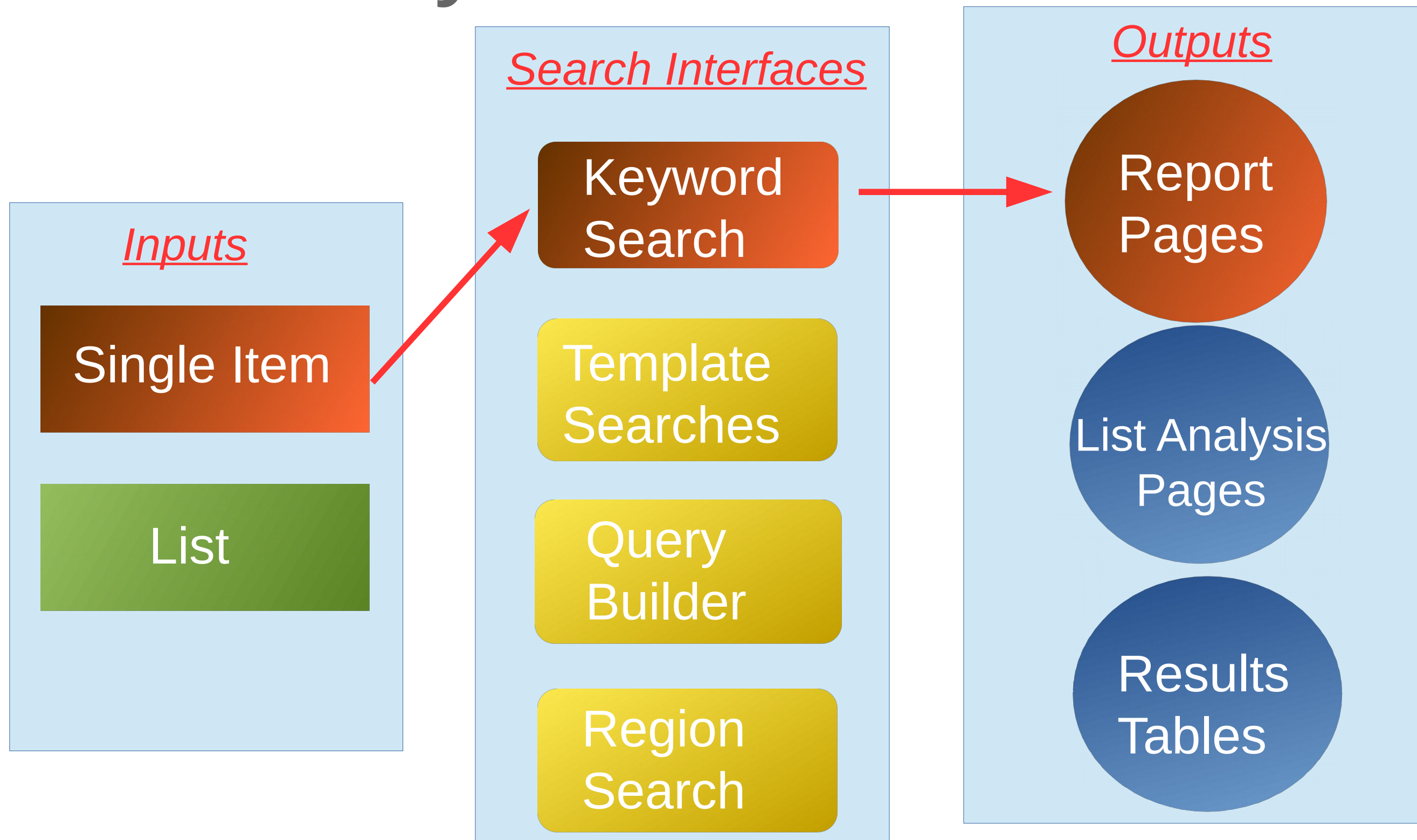
Using InterMine

Search
Explore
Analyse

The Web Interface



Keyword Search



Search and Explore

Faceted Keyword Search:

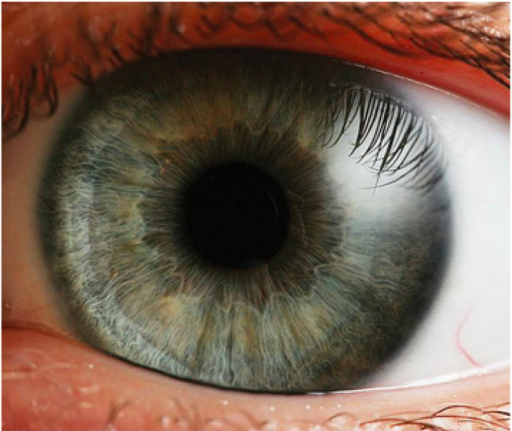
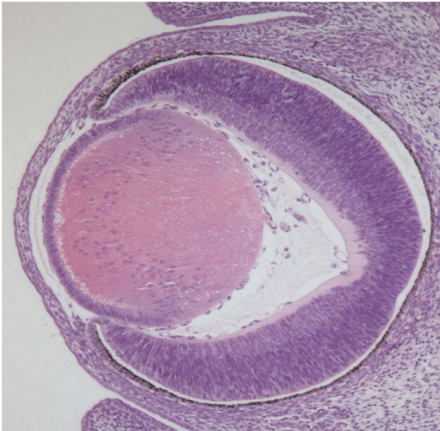


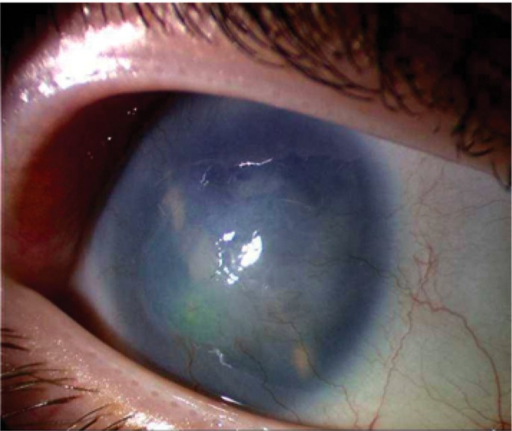
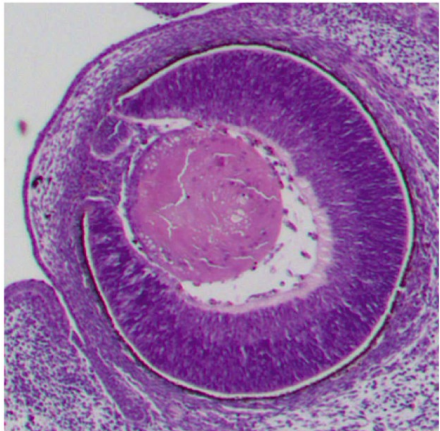
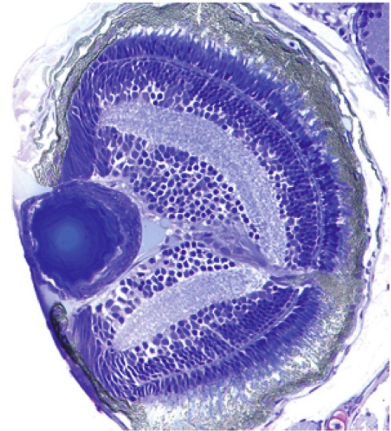
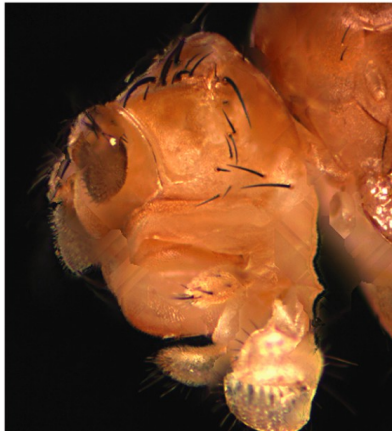
- Specific e.g. pax6, pparg
 - Report pages
- Exploratory e.g. Insulin, Diabetes
 - Report pages
 - Lists
- Filter results by data type

Data Exploration: report pages

Every object (item) in InterMine has a report page

- Collate all the data available for that object
- Contain a mixture of interactive tables, search results and graphical displays.
- Links to related data

Explore: Pax6

	Human	Mouse	Zebrafish	Drosophila
WT				
mut				
	<i>PAX6^{+/-}</i>	<i>Pax6^{-/-}</i>	<i>pax6b^{-/-}</i>	<i>ey^{-/-}</i>
EQs	cornea opaque iris absent retina degenerate lens opaque aqueous humor of eyeball increased pressure	eye decreased size lens fused_to cornea iris morphology anterior chamber absent	eye decreased size lens decreased size retina malformed	eye absent

Washington NL, Haendel MA, Mungall CJ, Ashburner M, Westerfield M, Lewis SE. - Figure 1 of Washington et al.:
 "Linking Human Diseases to Animal Models Using Ontology-Based Phenotype Annotation." PLoS Biol 7(11): e1000247. doi:10.1371/journal.pbio.1000247

Exercise 1: Faceted Search

1. Search for one or more of the following in HumanMine:

- *Pax6*
- rs10509540
- *diabetes*

2. Filter and create a list:

- Search for *diabetes*
- Filter for publications
- Make a list of these publications

Note: If you filter for genes note that this is not a comprehensive search for diabetes genes as it does not check functional annotations.

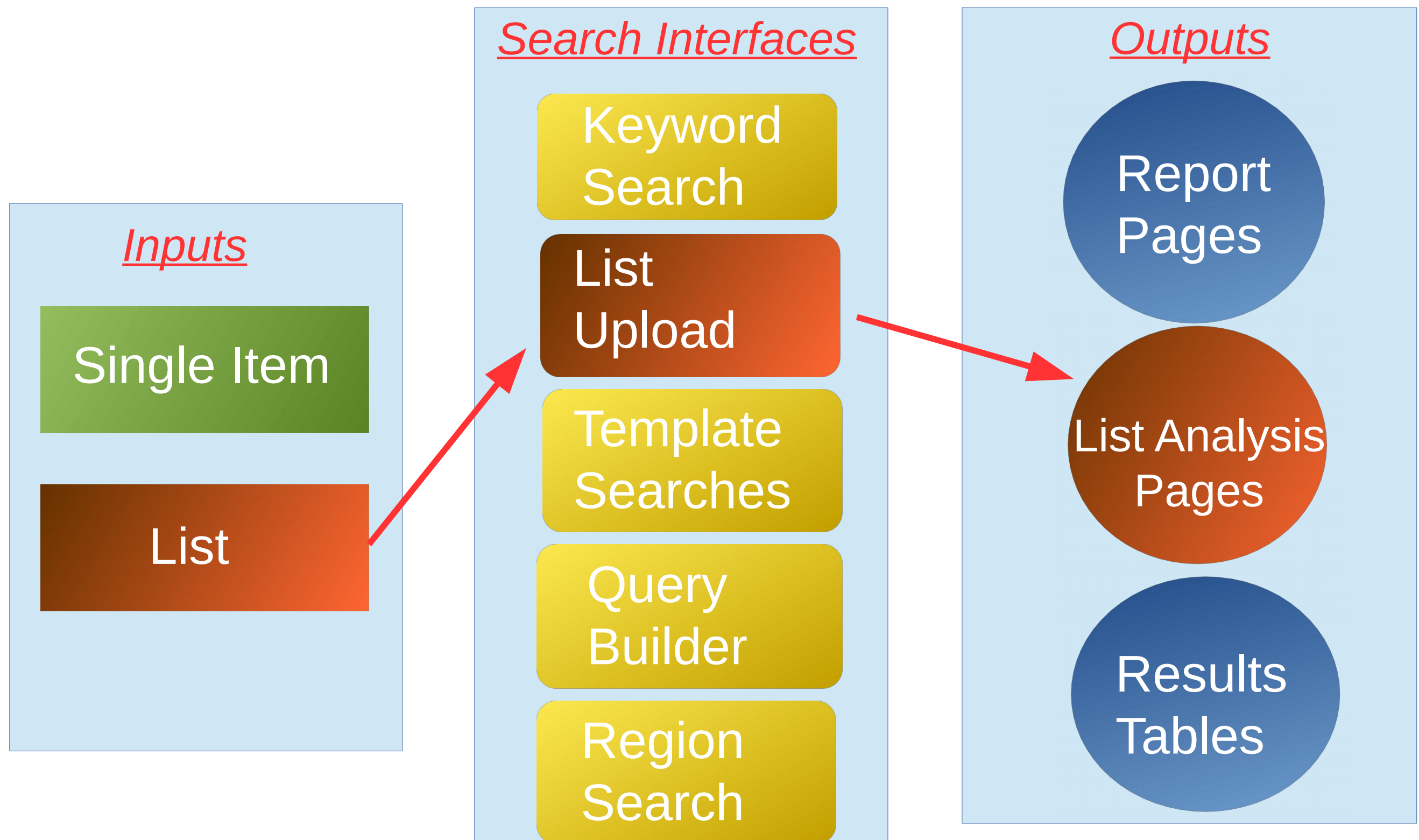
Exercise2: Exploring a Gene

You are interested in the Human PAX6 gene and want to know the following things about it:

1. On which chromosome is *PAX6* located?
2. Can I access the sequence for the *PAX6* gene?
3. With which diseases is *PAX6* associated?
4. In which tissues is *PAX6* most highly expressed?
5. Does the *PAX6* protein have any known isoforms?
6. Does the Pax6 protein have known domains?
7. Is there a *PAX6* orthologue in *D. melanogaster*?
8. Does this orthologue interact with any other Genes/proteins? Identify the interaction type (genetic/physical)
9. For the interaction of Ey with 4E-T, what was the original experiment and publication that determined this interaction



Lists



List Analysis: Uploading a list

- Upload your own lists to InterMine
- Powerful identifier resolution system
- Convert old identifiers into an up-to-date set

Exercise 3: Uploading a list:

1. Use FlyMine: navigate to the lists tab and list upload sub-tab
2. Select the example list (leave type and organism as the default values).
3. Click “Create list”.
4. Examine and understand the list page, name and save your list.

Exercise 3: Uploading a list:

- E2f has matched two genes (**duplicates**) - in this case you need to decide which of the two genes you want in your list (or both). The action column allows you to do this.
- Two of the identifiers in the list matched the same gene: FBgn0010433 and ato. This is indicated in the **direct hits**.
- One of the identifiers is a protein identifier (TWIST_DROME). As the associated gene could be identified, this has been added to the list. This is shown under **non-gene identifiers**.
- Two of the identifiers matched a **synonym** (rather than a current identifier). As the synonyms matched only one gene, these are automatically added to the list.

Data Exploration: Lists

InterMine allows you to explore data for a whole list of objects

- Uploaded or created from searches
- Identifier resolution
- List analysis pages
- Summary tables and graphs, enrichment statistics and search results
- List set operations - union, intersect, subtract
- Workflows through iterative querying and set operations
- Public lists also available

List Analysis

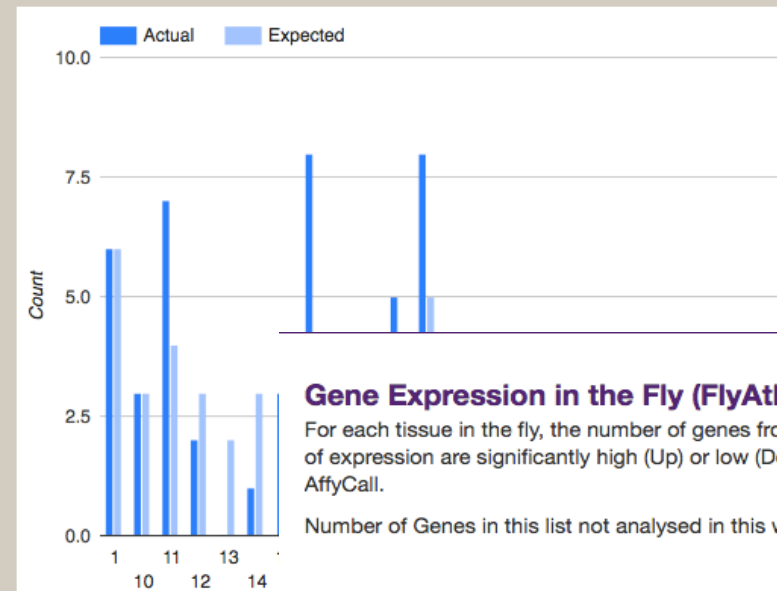
Chromosome Distribution

Actual: number of items in this list found on each chromosome. Expected: given the total number of items on the chromosome and the number of items in this list, the number of items expected to be found on each chromosome.

All items in your list have been analysed.

Organism

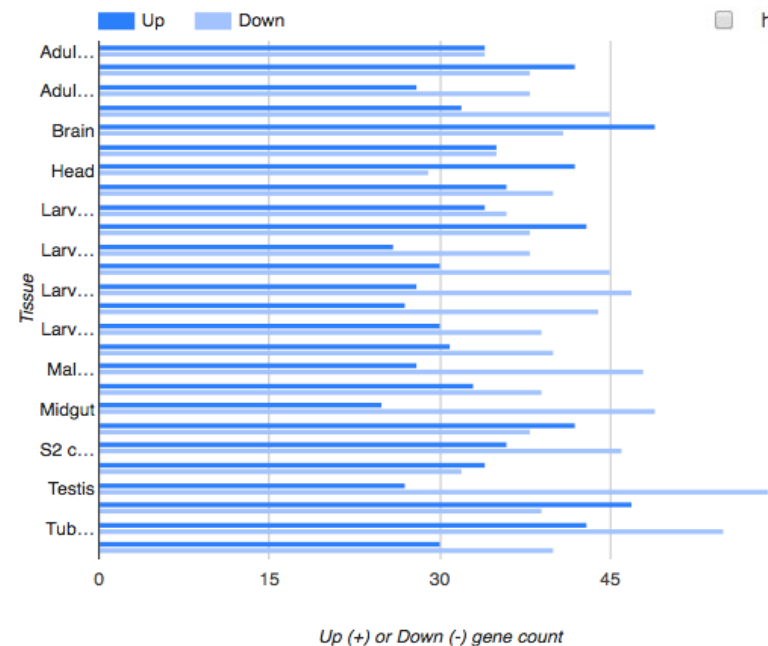
Homo sapiens



Gene Expression in the Fly (FlyAtlas)

For each tissue in the fly, the number of genes from this list for which of expression are significantly high (Up) or low (Down) according to FlyAtlas.

Number of Genes in this list not analysed in this widget: 3



Gene Ontology Enrichment

GO terms enriched for items in this list.

Number of Genes in this list not analysed in this widget: 17

Test Correction: Holm-Bonferroni, Max p-value: 0.05, Ontology: biological_process

Background population

Default Change

View Download

GO Term	p-Value	Matches
cell adhesion [GO:0007155]	8.077144e-5	13
biological adhesion [GO:0022610]		
Wnt signaling pathway [GO:0016055]		
cell-cell adhesion [GO:0098609]		
wing disc development [GO:0035220]		
cell-matrix adhesion [GO:0007160]		
tissue development [GO:0009888]		
heart development [GO:0007507]		

View homologues in other Mines:

RatMine

R. norvegicus

YeastMine

S. cerevisiae

MouseMine

M. musculus

HumanMine

H. sapiens

ZebrafishMine

D. rerio

Mammalian Phenotype Ontology Enrichment

MP terms enriched for items in this list.

Number of Genes in this list not analysed in this widget: 72

Test Correction: Holm-Bonferroni, Max p-value: 0.05, Background population: Default Change

View Download

MP Term	p-Value	Matches
abnormal DNA repair [MP:0008058]	1.831867e-66	52
increased sensitivity to induced cell death [MP:0008943]	1.383230e-62	59
abnormal induced cell death [MP:0008942]	1.175928e-60	65
abnormal chromosome stability [MP:0010094]	4.405108e-58	48
chromosomal instability [MP:0008866]	1.743731e-56	47
abnormal cell physiology [MP:0005621]	2.197659e-53	163
cellular phenotype [MP:0005384]	1.305734e-51	168
chromosome breakage [MP:0004028]	4.007892e-50	37
abnormal cell death [MP:0000313]	1.184952e-44	106

Exercise 4: List Analysis:

Examine the HumanMine public list: **PL_Pax6_Targets (319 genes)**

1. What is the most enriched GO term for this list?
2. How many genes in the list are annotated with this GO term?

Note: you could make a sub-list containing only genes from this list annotated with this term by clicking on the matches number

3. Navigate to the MouseMine database to examine the mouse orthologues for this list.
4. How many mouse orthologues are there for this list?
5. Are these mouse genes enriched for any phenotypes?

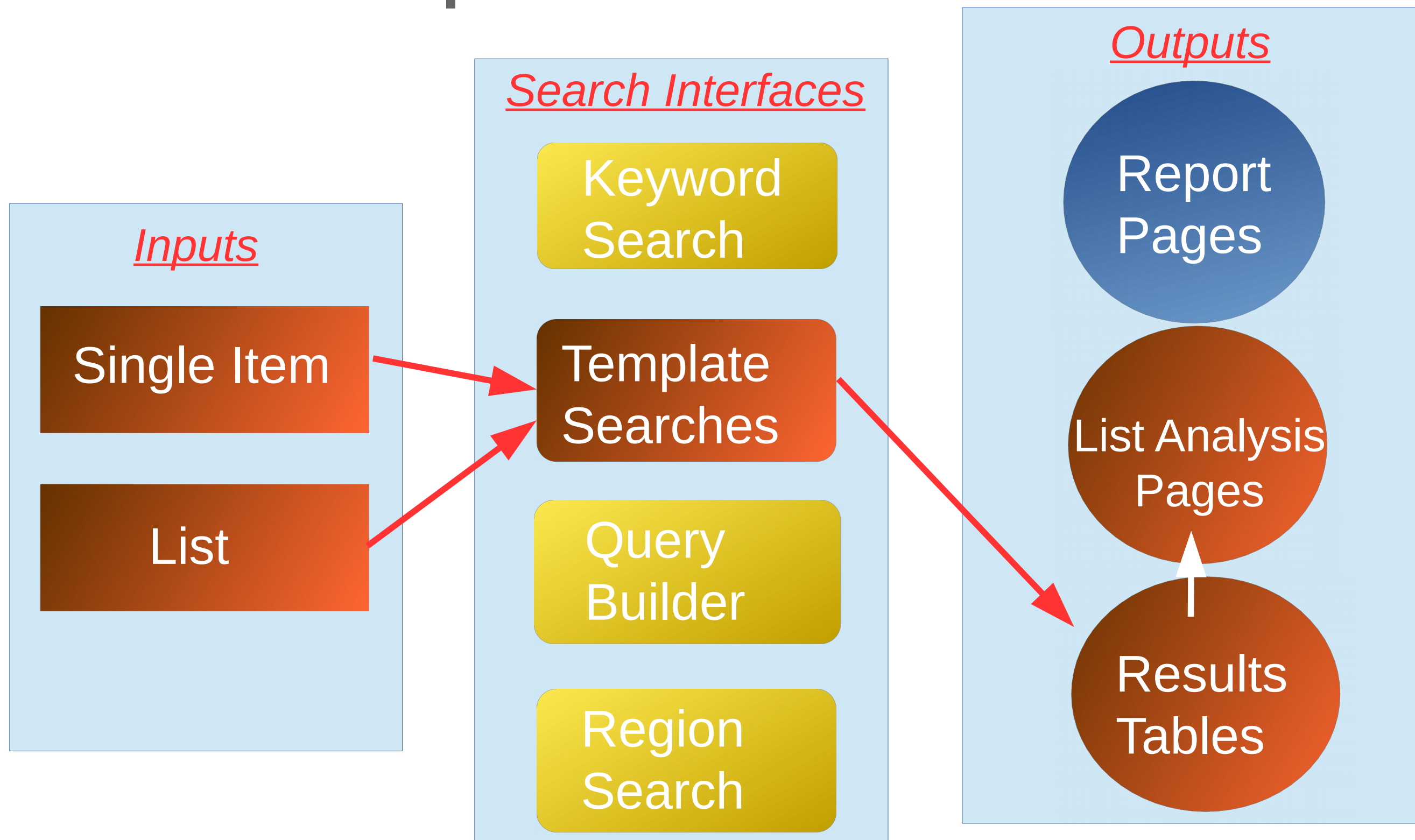
Using InterMine

Search

Explore

Analyse

Template Searches



Data Analysis: Template Searches

The template searches allow a more refined search than the keyword search and report pages but are still quick and easy to access.

- Pre-defined searches with simple filters
- Range from simple searches to more complex searches spanning several data types
- Run with single item or list
- Results are returned in sophisticated results tables
- Easy to add - just ask

Many Many Searches.....

- Which other genes have this GO annotation?
- Are there mutant phenotypes for this gene?
- Where is this gene expressed?
- What does this gene interact with?
- Do any of the interacting genes share the mutant phenotypes?
- Does this gene have a human orthologue with a disease association?
- Have any variants been associated with this gene/disease?
- Which organisms have models for this disease/gene?


Data Analysis: Template Searches

- What other genes are involved in pancreatic function?
- Are there potential targets of pax6 in pancreatic tissue?
- Have these genes been implicated in pancreatic disease?
- What published data is there about these genes?

Data Analysis: Template Searches

Are there potential targets of pax6 in pancreatic tissue?

Gene + Tissue Expression → Interactors that are expressed in that tissue

**Gene + Tissue Expression** → **Interactors that are expressed in that tissue**
For a specified gene or genes, show genes that interact with this gene that are also expressed in the specified tissue. (Interactors could be physical or genetic; Expression is based on tissue expression from the Protein Atlas Project and should be further interrogated for reliability level).

Gene – For the gene:
LOOKUP: for Organism:
☐ constrain to be saved Gene list

Tissue > Name – Show interactors that are expressed in:

[web service URL](#) [Perl](#) | [Python](#) | [Ruby](#) | [Java \[help\]](#) [export XML](#)

Data Analysis: Results Tables

Results tables allow further interactive analysis of the data through:

- Column summaries
- Column sorting
- Adding additional columns of data
- Filtering
- List creation
- Export

The power of column summaries

- Find the number of unique items in a column
- Find the number of items with a particular property
- Filter the table for one or more specific properties

Exercise 5: Template searches:

1. Browse the template searches in FlyMine and HumanMine - try running a few or changing the filters.
2. Use the search box to find template searches for interactions
3. Filter the FlyMine template searches to show only “expression” templates.

Exercise 6: Using template searches:

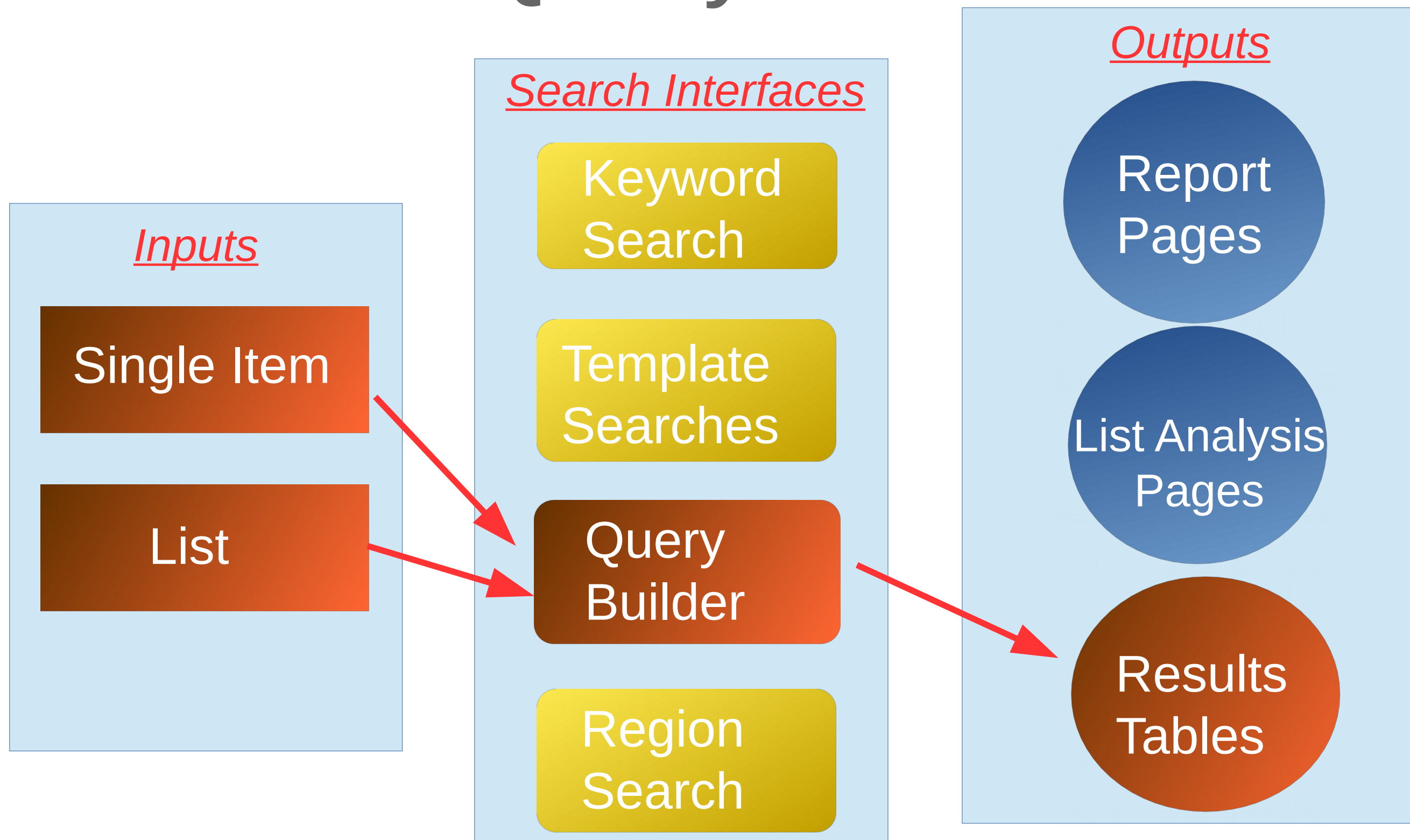
We will continue our exploration of the Pax6 gene in Pancreatic tissue. Use the template searches in HumanMine to answer the following question:

Are any of the known targets of Pax6 expressed in the pancreas. You will find a public list of known Pax6 target genes in HumanMine (PL_Pax6_Targets).

Use the Protein Atlas dataset for the expression measurement.

Save the list of target genes expressed in the pancreas as a list.

The Query Builder



Advanced Search: Query Builder

The Query Builder is InterMine's custom query builder, allowing you to create and save your own searches.

- Build your own Searches
- Modify template searches
- Combine any data:
 - And, Or,
 - Intersect; Union

Data Analysis: Query Builder

Three steps to construct a query:

1. Navigate the data model to find the class or attribute you need
2. Add the appropriate constraint (filter) to the class/attribute
3. Decide on the columns you want to view in your results

Exercise 7: Query Builder:

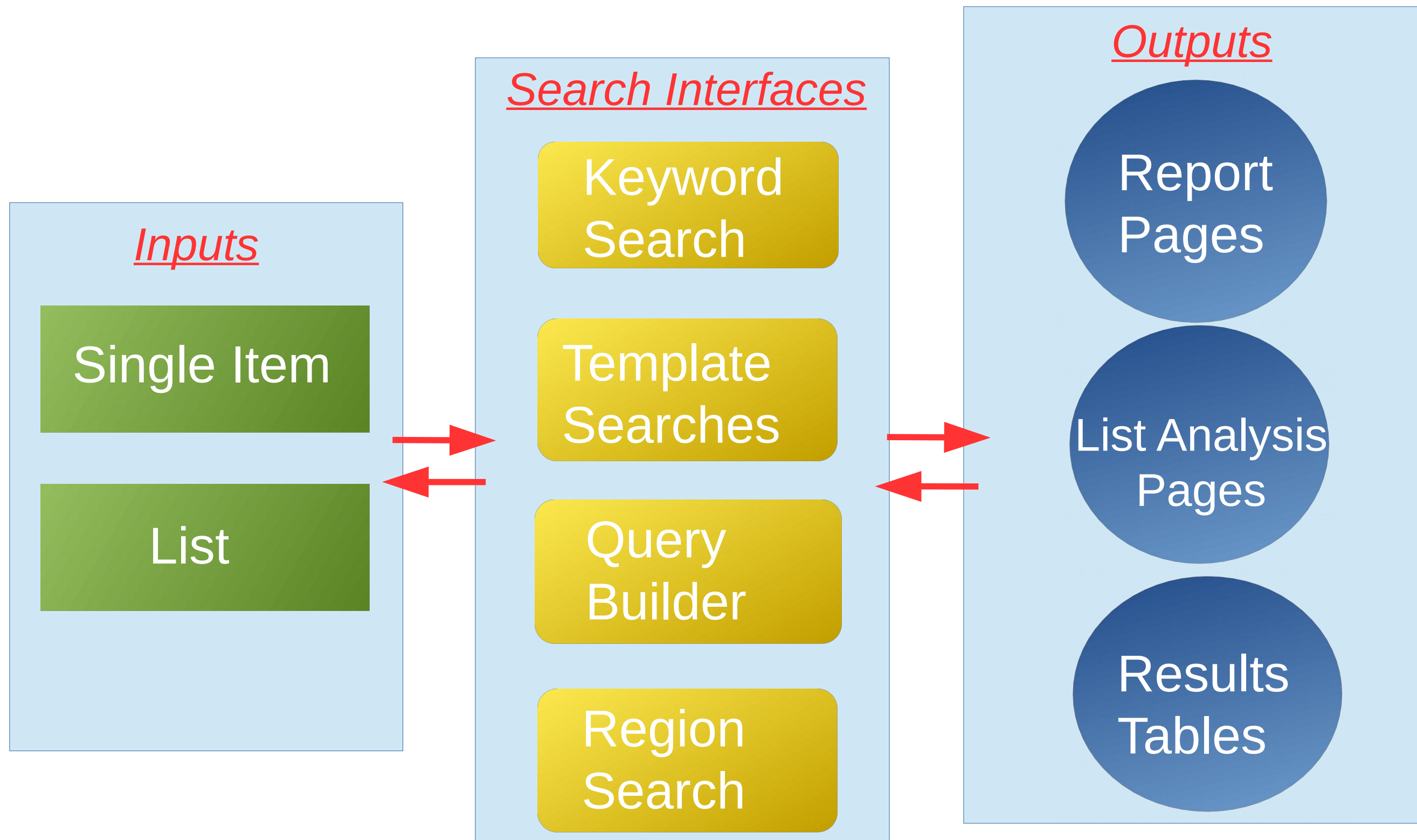
Using HumanMine: we will build a query to show Human genes and OMIM diseases, and then add a further constraint to show genes associated with all types of Diabetes.

1. Start your query from Gene
2. Constrain “Organism” to Homo Sapiens
3. Add the columns of data we want in our results:

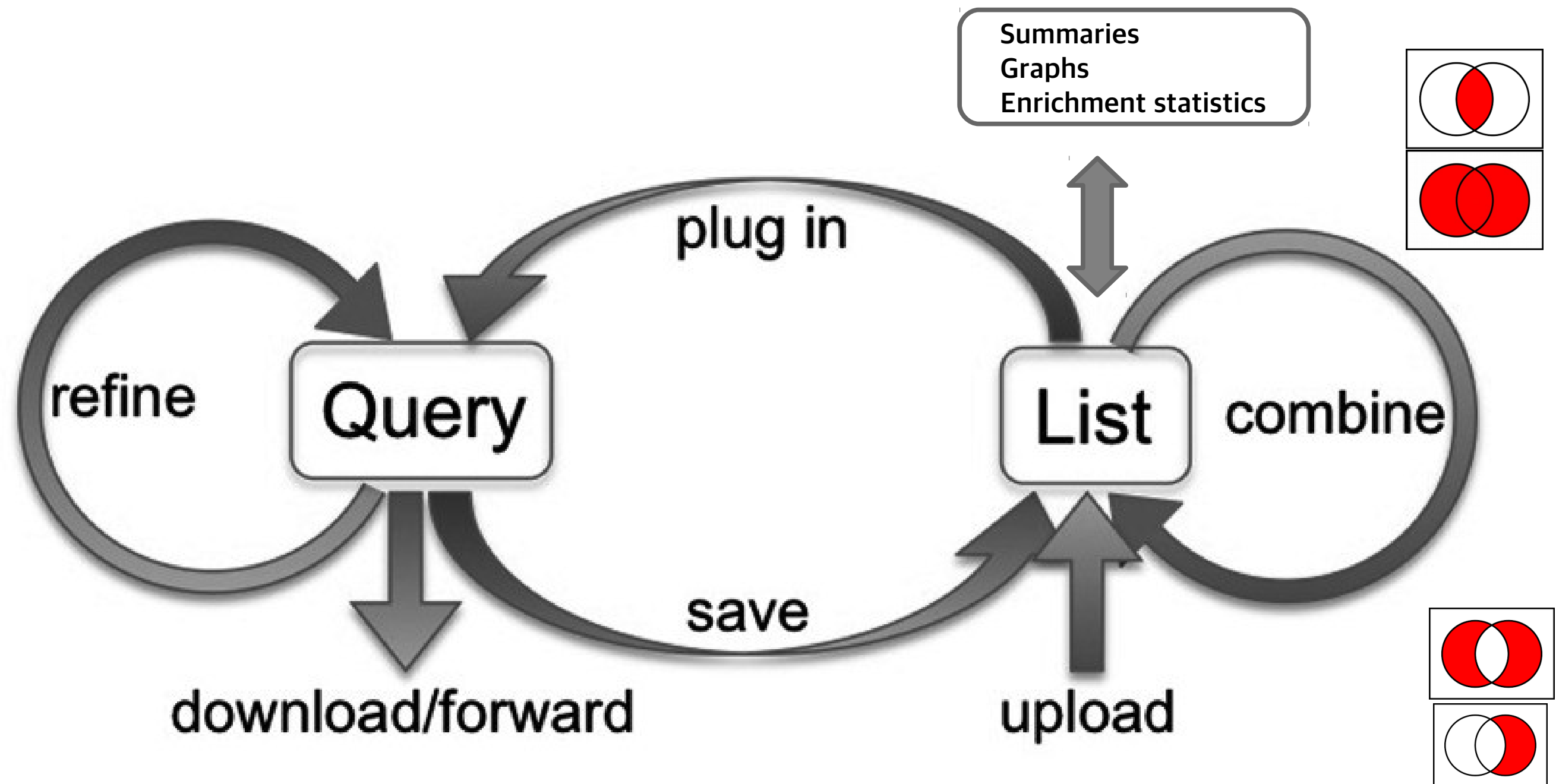
Gene: Primary identifier and Symbol
Disease: name

4. Run this search - ‘Show results’.
5. Return to the query (Use the “Trail” in the top left) and add a constraint to Disease name for “CONTAINS *Diabetes*”
6. Run the search and save the set of genes

The Web Interface

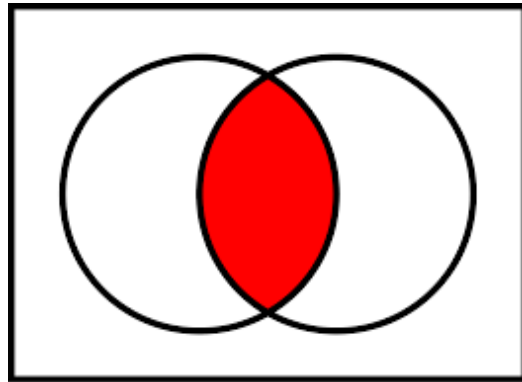


Analysis Workflows

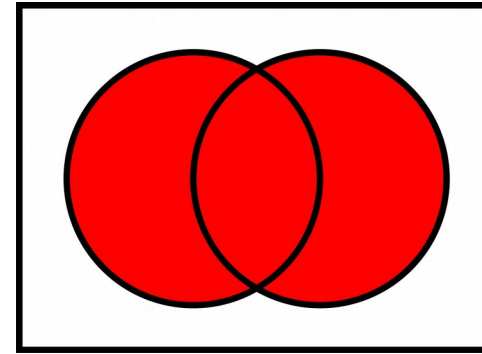


Motenko H, Neuhauser SB, O'Keefe M, Richardson JE. MouseMine: a new data warehouse for MGI. Mamm Genome. 2015 Aug;26(7-8):325-30. doi: 10.1007/s00335-015-9573-z. PubMed PMID: 26092688; PubMed Central PMCID: PMC4534495

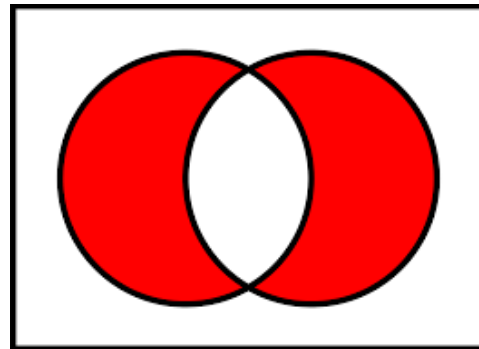
Lists: Set Analysis



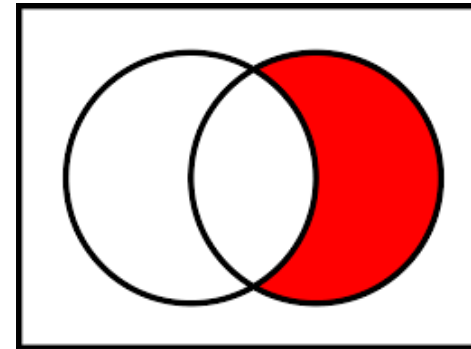
Intersect



Union



Subtraction



Asymmetric difference

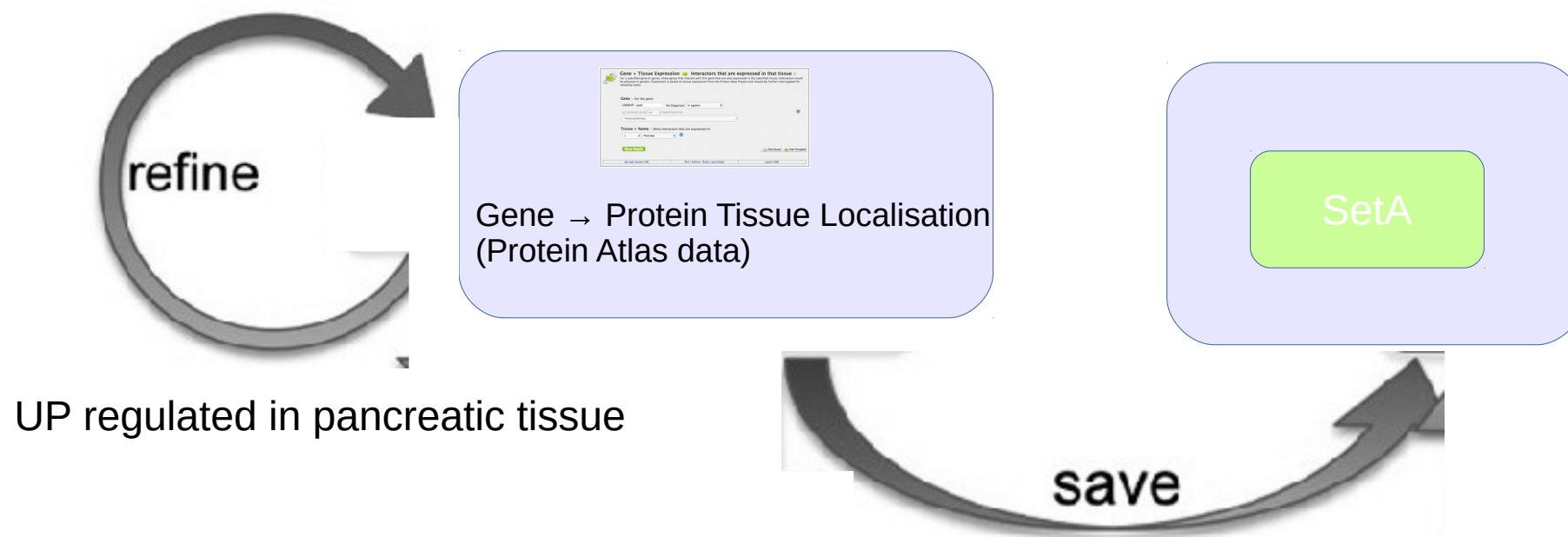
Exercise 8: Analysis Workflows

In exercise 6 we ran a template search and saved a set of genes. In this exercise we will combine this set of genes from exercise 6 with our set of diabetes genes from exercise 7. We will then feed the resulting set into a further query.

1. Identify the sets of genes you have created under the lists “view” tab.
2. Use the list set operations available on this page to intersect the list of diabetes genes you created with the query builder with your previous set of genes (Pax6 target genes expressed in the Pancreas) created in exercise 6.
3. We now want to know if any of these genes have been identified in GWAS studies. Run a template on your intersected list to find this out.
4. Use the column summary to find if any of the GWAS phenotypes are related to diabetes.

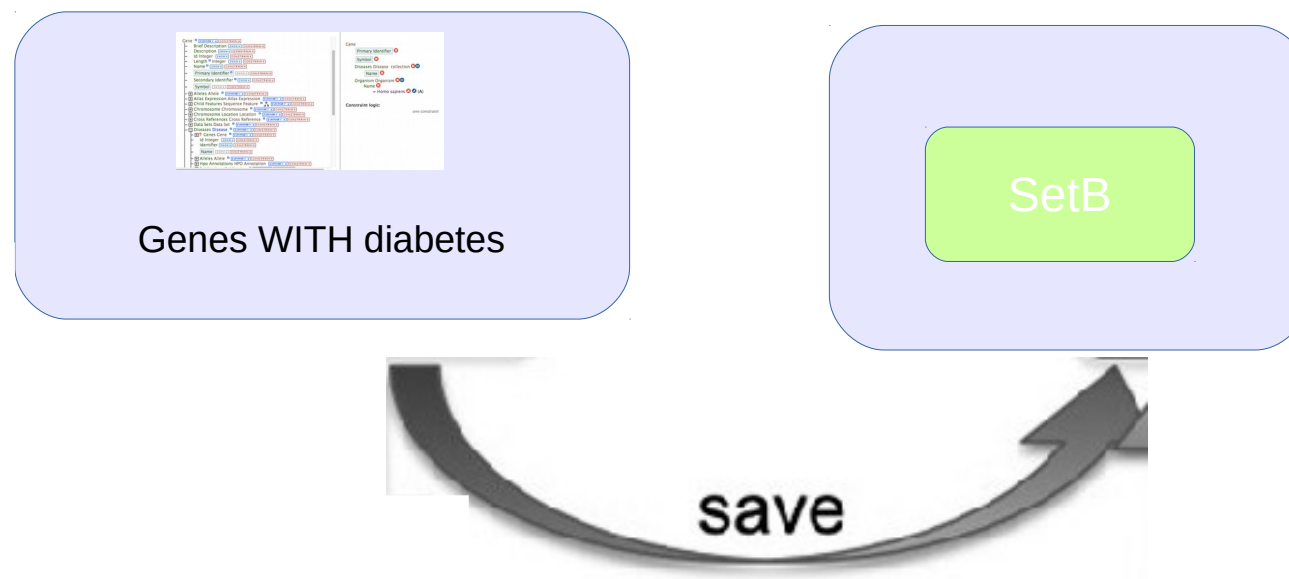
Analysis Workflow

1. We looked at gene expression of Pax6 targets in the Pancreas and refined this to genes with a high or medium level expression. We saved these as a list – Set A.



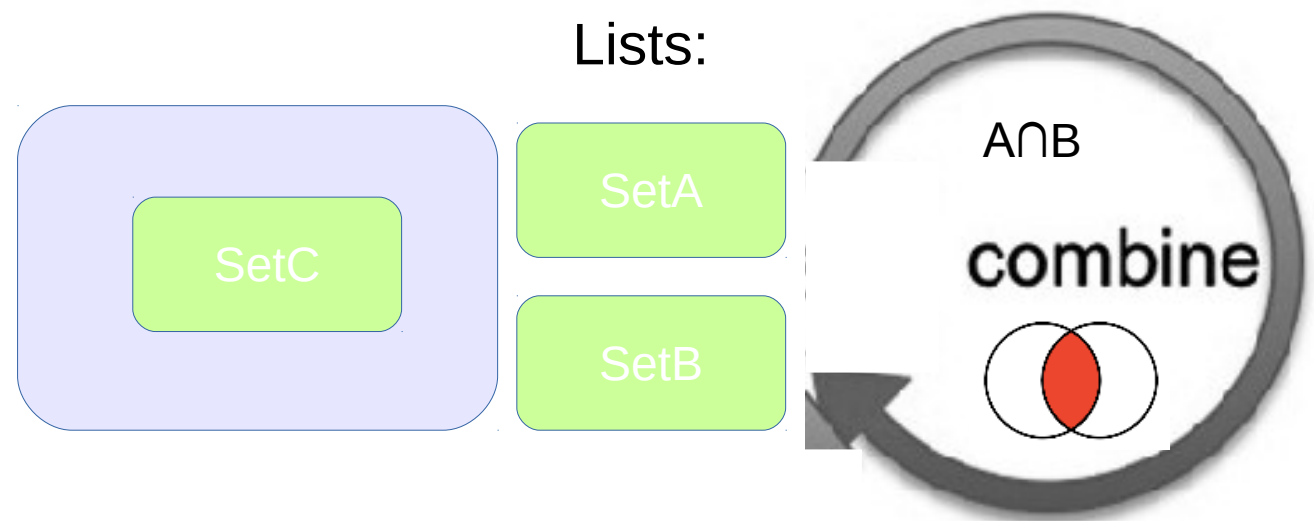
Analysis Workflow

2. A query built through the query builder identified a set of genes involved in Diabetes -> setB



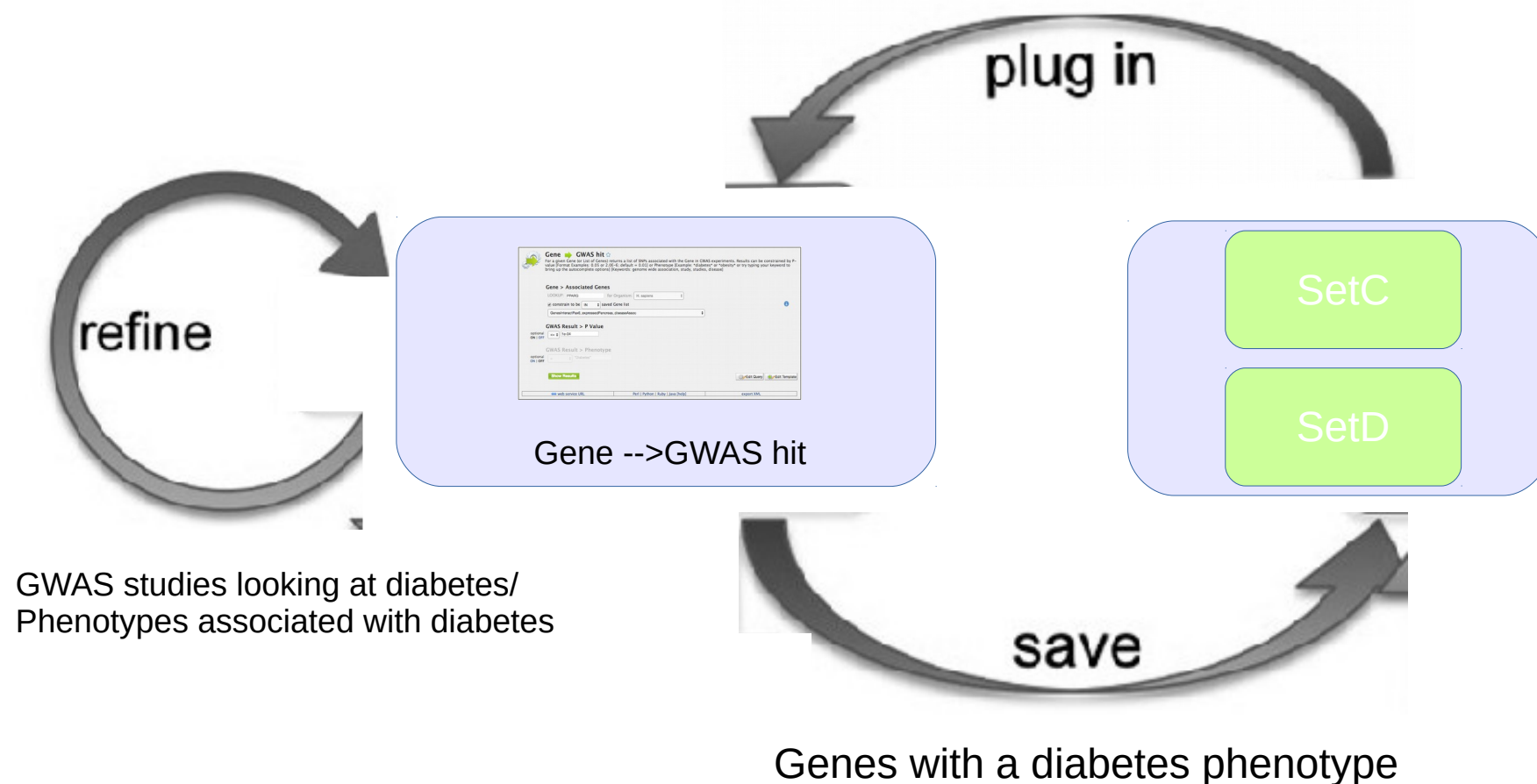
Analysis Workflow

3. A list intersection between setA and setB identified Pax6 target genes that are expressed in the pancreas that also have some association with diabetes → setC



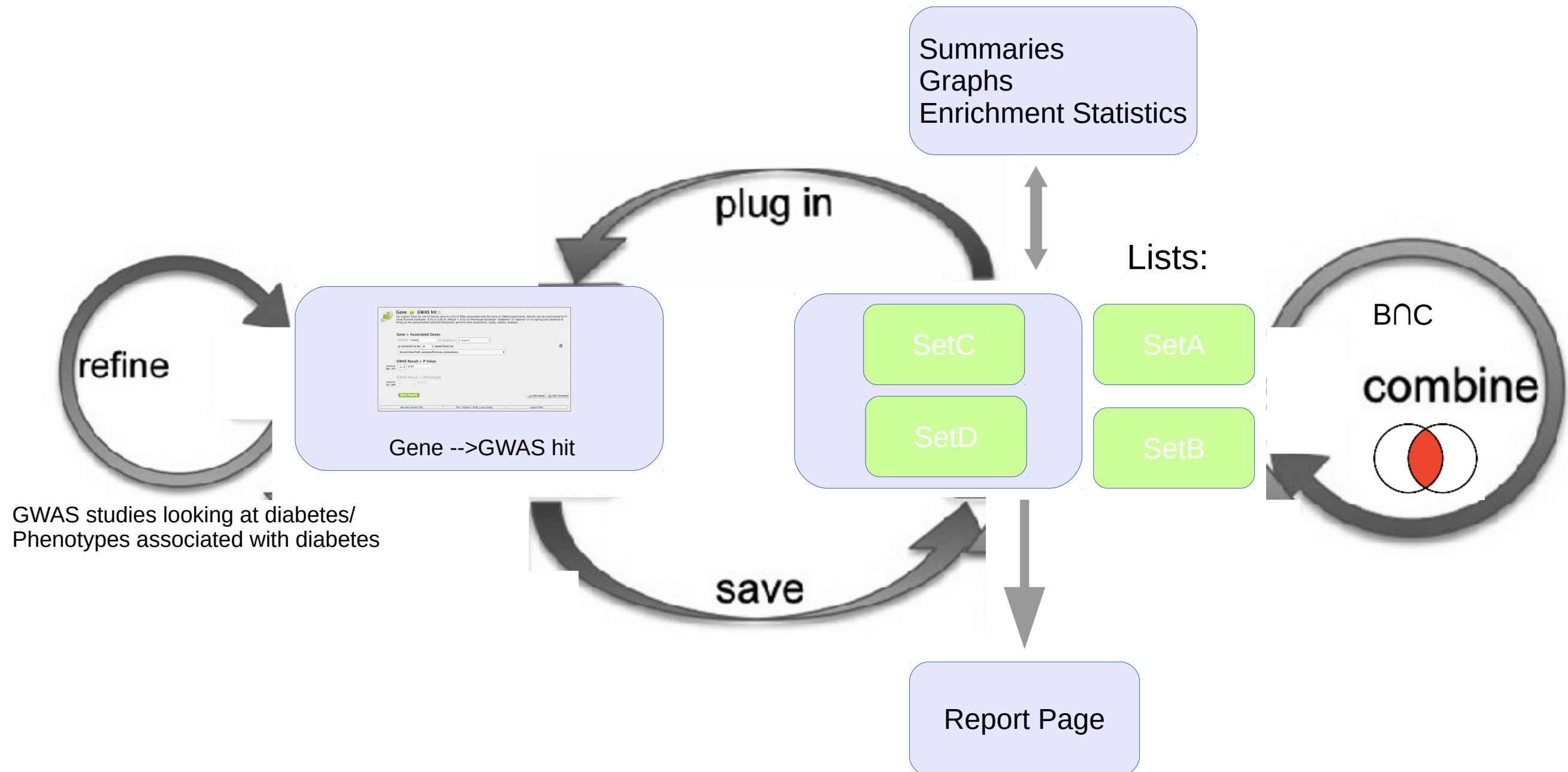
Analysis Workflow

4. Genes saved as SetC were “plugged in” to a second template to identify if there was any association of these genes with diabetes phenotypes according to GWAS studies. These genes were saved - SetD

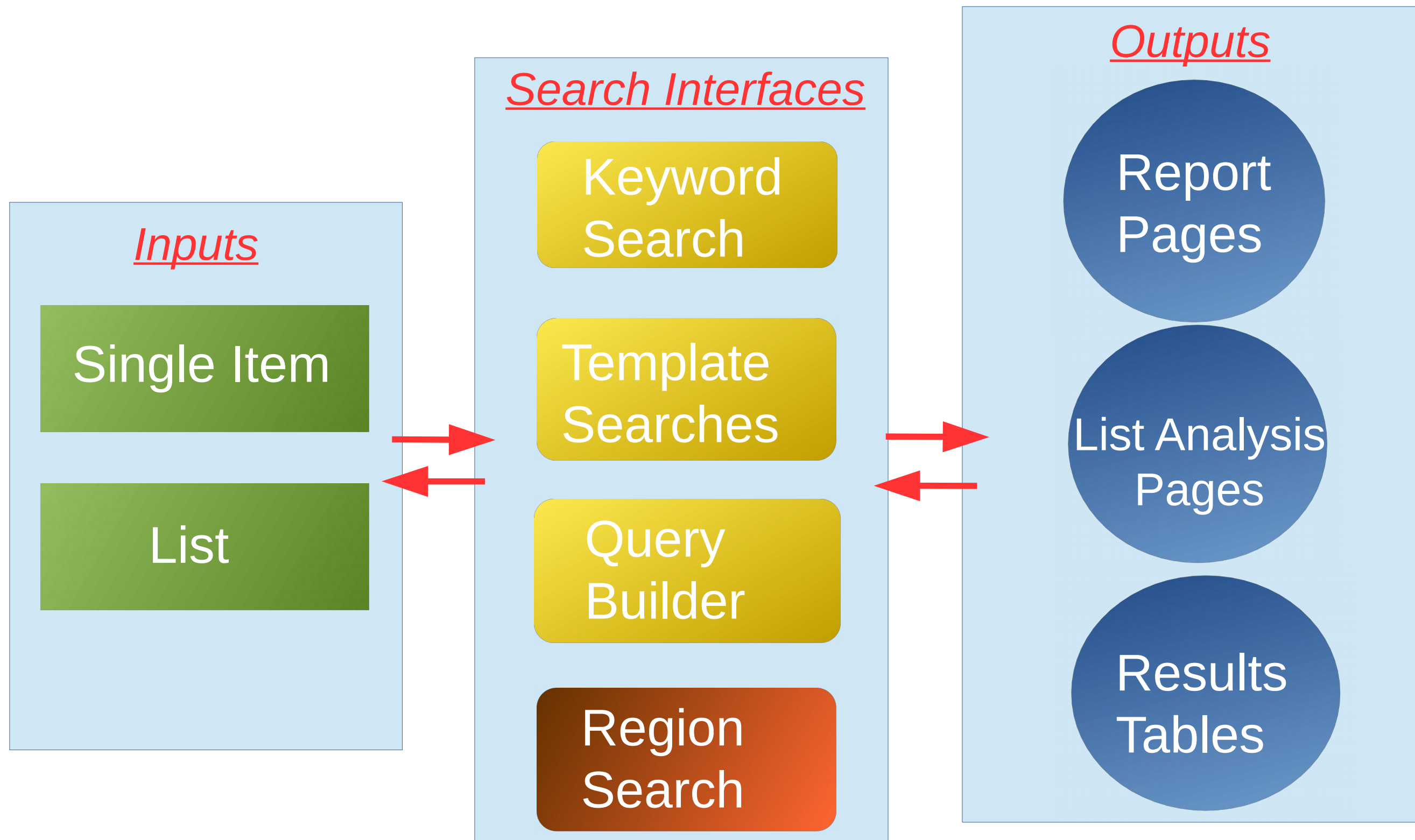


Analysis Workflow

5. ListD contained two genes that we could explore further through their report pages, the list analysis page and through further template searches



The Web Interface



Data Analysis: RegionSearch

The Region Search allows you to search for features that overlap a list of genome coordinates.

- Any or selected genome features can be searched.
- Accepts base or interbase coordinates
- Region to be searched can be extended upstream and downstream

Exercise 9: Region Search:

Using **FlyMine**:

1. Select the example set of regions
2. De-select the features and re-select Genes and Regulatory regions
3. Extend the search by 5kb
4. Run the search

Examine the results and:

5. Create a list of all genes found.
6. Create a list of the regulatory regions found in the first genomic span.



HELP?

- FlyMine - extensive manual and videos under the 'help' link. These apply to all InterMine databases
- Each InterMine has it's own help pages, videos and tutorials

Email Us: Every InterMine page has a 'questions/comments' link.

Support chat: <http://chat.intermine.org>

Support email: info@intermine.org

Twitter: @intermineorg