Exercises

Biological Data Analysis Using InterMine
workshop exercises with answers

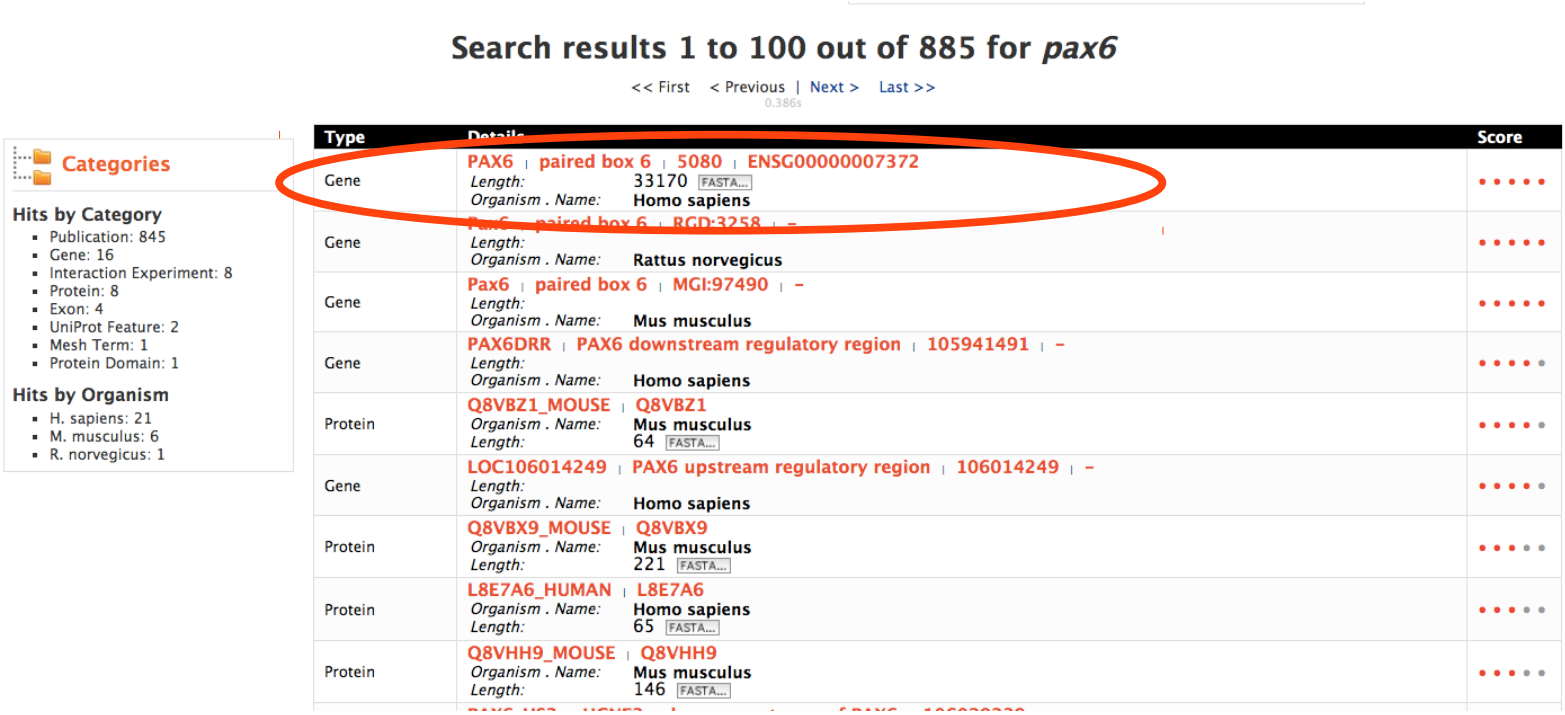
Exercise1: Faceted Search

Use HumanMine for this exercise

1. Search for one or more of the following using the keyword search (result for Pax6 only shown below):

- *Pax6*
- rs10509540
- *insulin*

Human *Pax6* is the first gene result returned when HumanMine is searched for *Pax6. Click on this gene to be taken to the report page.*

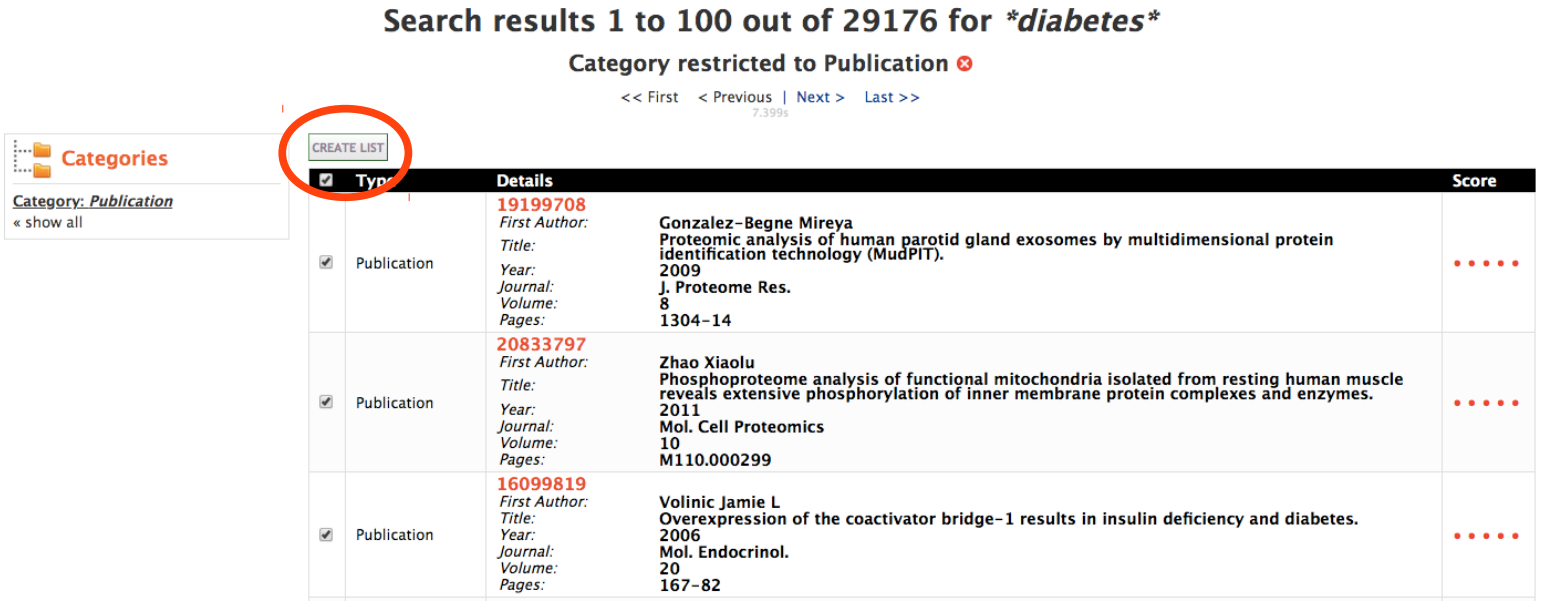


Exercise1: Faceted Search:

2. Filter and create a list:

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

Use the checkbox in the header to select all the publications and make a list:



Exercise2: Exploring a Gene:

You are interested in the Human *Pax6* gene and want to know the following things about it:

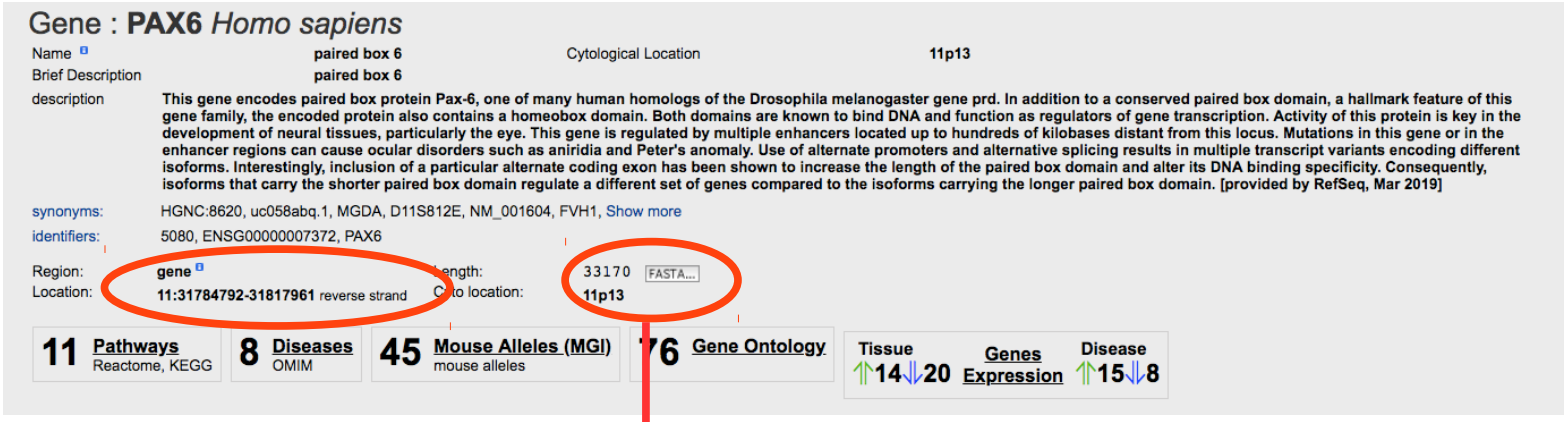
Use HumanMine for this exercise

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 know 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 with CDX2, what was the original experiment and publication that determined this interaction

Exercise2: Exploring a Gene:

1. **On which chromosome is *Pax6* located?**
2. **Can I access the sequence for the *Pax6* gene?**

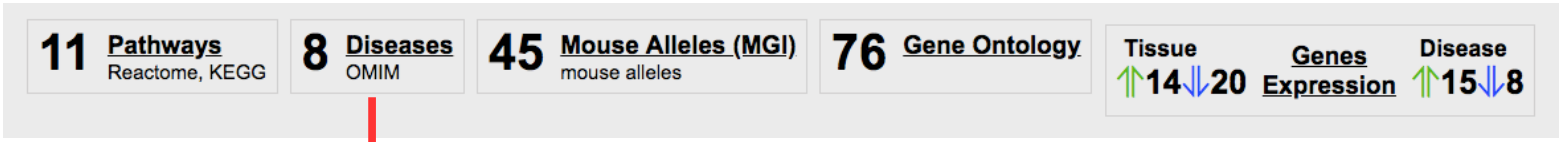
The first section on the gene report page provides information about the chromosome location of the gene along with identifiers and synonyms and a link to the FASTA DNA sequence:



Exercise2: Exploring a Gene:

3. With which diseases is *Pax6* associated?

In HumanMine a summary of data is provided at the top of the report page (note this feature is not available in all InterMine's).



"Diseases" links to a table for OMIM data

8 Diseases

Trail: Gene

Manage Columns | Manage Filters | Manage Relationships | Save as List | Generate Python code | Export

Showing 1 to 8 of 8 rows

Diseases Primary Identifier	Diseases Name
OMIM:106210	ANIRIDIA 1
OMIM:120200	COLOBOMA, OCULAR, AUTOSOMAL DOMINANT
OMIM:120430	COLOBOMA OF OPTIC NERVE
OMIM:136520	FOVEAL HYPOPLASIA 1
OMIM:148190	KERATITIS, HEREDITARY
OMIM:165550	OPTIC NERVE HYPOPLASIA, BILATERAL
OMIM:194072	WILMS TUMOR, ANIRIDIA, GENITOURINARY ANOMALIES, AND MENTAL RETARDATION SYNDROME
OMIM:604229	ANTERIOR SEGMENT DYSGENESIS 5

Further disease information is sometimes available from the "Curated comments from Uniprot":

Curated comments from UniProt — Show proteins

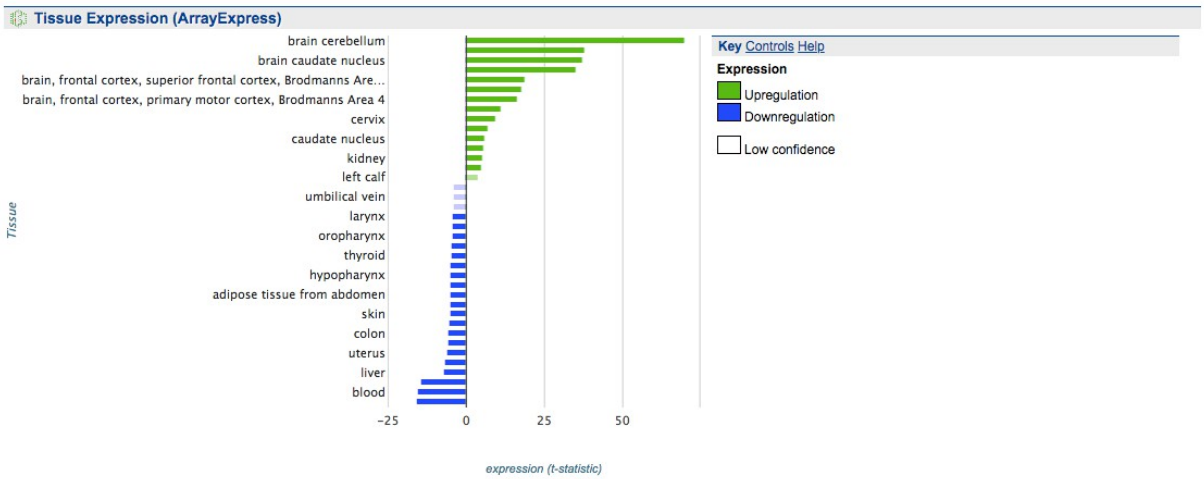
Type	Comment
developmental stage	Expressed in the developing eye and brain. Expression in the retina peaks at fetal days 51-60. At 6-week old, in the retina, is predominantly detected in the neural layer (at protein level). At 8- and 10-week old, in the retina, the expression is strongest in the inner and middle layer of the neural part (at protein level).
disease	MIM:106210; Aniridia 1; AN1; A congenital, bilateral, panocular disorder characterized by complete absence of the iris or extreme iris hypoplasia. Aniridia is not just an isolated defect in iris development but it is associated with macular and optic nerve hypoplasia, cataract, corneal changes, nystagmus. Visual acuity is generally low but is unrelated to the degree of iris hypoplasia. Glaucoma is a secondary problem causing additional visual loss over time. The disease is caused by mutations affecting the gene represented in this entry.
disease	MIM:120200; Coloboma, ocular, autosomal dominant; COAD; A set of malformations resulting from abnormal morphogenesis of the optic cup and stalk, and the fusion of the fetal fissure (optic fissure). The clinical presentation is variable. Some individuals may present with minimal defects in the anterior iris leaf without other ocular defects. More complex malformations create a combination of iris, uveoretinal and/or optic nerve defects without or with microphthalmia or even anophthalmia. The disease is caused by mutations affecting the gene represented in this entry.
disease	MIM:120430; Coloboma of optic nerve; COLON; An ocular defect that is due to malclosure of the fetal intraocular fissure affecting the optic nerve head. In some affected individuals, it appears as enlargement of the physiologic cup with severely affected eyes showing huge cavities at the site of the disk. The disease is caused by mutations affecting the gene represented in this entry.
disease	MIM:136520; Foveal hypoplasia 1; FVH1; An isolated form of foveal hypoplasia, a developmental defect of the eye defined as the lack of foveal depression with continuity of all neurosensory retinal layers in the presumed foveal area. Clinical features include absence of foveal pit on optical coherence tomography, absence of foveal hyperpigmentation, absence of foveal avascularity, absence of foveal and macular reflexes, decreased visual acuity, and nystagmus. Anterior segment anomalies and cataract are observed in some FVH1 patients. The disease is caused by mutations affecting the gene represented in this entry.

Exercise2: Exploring a Gene:

4. In which tissues is *Pax6* most highly expressed?

Data on tissue expression can be found from three sources:

A. Human gene expression atlas of 5372 samples representing 369 different cell and tissue types, disease states and cell lines: from http://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-62/



B. Human Protein Atlas: http://www.proteinatlas.org/

Protein Atlas Tissue Expression

Overall	Organ	Cell types	Antibody staining
☐	Blood and immune system (Hematopoietic)	8	
☐	Breast and female reproductive system (Female tissues)	9	
☐	Cardiovascular system (Heart and blood vessels)	1	
☐	Central nervous system (Brain)	9	
☐	Digestive tract (GI-tract)	11	
☐	Endocrine glands	3	
☐	Liver and pancreas	5	
☐	Male reproductive system (Male tissues)	5	
☐	Placenta	2	
☐	Respiratory system (Lung)	4	
☐	Skin and soft tissues	2	
☐	Urinary tract (Kidney and bladder)	3	

Key Help

Reliability: **Supported**

Level of antibody staining*

High
Medium
Low
None

C. Curated comments from Uniprot:

disease	MIM:617141; Aniridia 2; AN2; A form of aniridia, a congenital, bilateral, panocular disorder characterized by complete absence of the iris or extreme iris hypoplasia. Aniridia is not just an isolated defect in iris development but it is associated with macular and optic nerve hypoplasia, cataract, corneal changes, nystagmus. Visual acuity is generally low but is unrelated to the degree of iris hypoplasia. Glaucoma is a secondary problem causing additional visual loss over time. The gene represented in this entry is involved in disease pathogenesis. A mutation in a PAX6 long-range cis-regulatory element, known as SIMO, affects PAX6 expression in the developing eye and has pathological consequences. The mutation is located in ELP4 intron 9, 150 kb downstream of PAX6.
function	Transcription factor with important functions in the development of the eye, nose, central nervous system and pancreas. Required for the differentiation of pancreatic islet alpha cells (By similarity). Competes with PAX4 in binding to a common element in the glucagon, insulin and somatostatin promoters. Regulates specification of the ventral neuron subtypes by establishing the correct progenitor domains (By similarity). Isoform 5a appears to function as a molecular switch that specifies target genes.
similarity	Belongs to the paired homeobox family.
tissue specificity	Fetal eye, brain, spinal cord and olfactory epithelium. Isoform 5a is less abundant than the PAX6 shorter form.

Exercise2: Exploring a Gene:

5. Does the *PAX6* protein have any know isoforms?

6. Does the PAX6 protein have any known domains?

a. Navigate to the proteins table in the report page:

Proteins — Toggle Proteins from other Data Sets

Primary Identifier	Primary Accession	Organism	Is Uniprot Canonical	Is SwissProt Curate?	Length
PAX6_HUMAN	P26367	Homo sapiens	true	true	422
PAX6_HUMAN-2	P26367-2	Homo sapiens	false	true	436

Show all in a table »

b. Select the PAX6_HUMAN protein (the canonical SwissProt annotated protein) to be taken to the protein report page. Note: we can also see from this table that PAX6 has one isoform: PAX6_HUMAN-2.

c. Navigate to the Isoforms table on the protein report page. Note that this table links to a report page for each of the isoforms.

1 Isoforms — Add tags

Manage Columns | Manage Filters | Manage Relationships | Save as List | Generate Python code | Export

Showing 1 to 1 of 1 row

Isoforms Primary Identifier	Isoforms Primary Accession	Isoforms Organism . Name	Isoforms Length
PAX6_HUMAN-2	P26367-2	Homo sapiens	436

d. Navigate to the Domains table on the protein report page:

Protein-> Domains (18 rows)

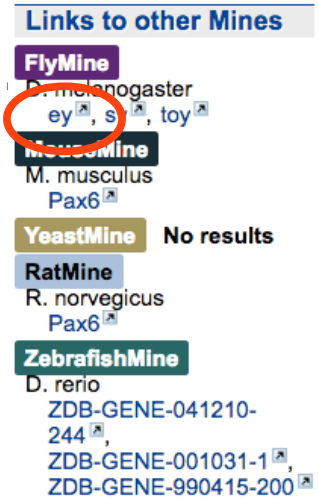
Manage Columns | Manage Filters | Manage Relationships | Save as List | Generate Python code | Export

Showing 1 to 18 of 18 rows — Rows per page: 10

Protein UniProt Accession	Protein UniProt Name	Protein Domain Name	Protein Domain Primary Identifier	Protein Domain Type	Protein Domain Regions Start	Protein Domain Regions End	Protein Domain Regions Identifier	Protein Domain Regions Database
P26367	PAX6_HUMAN	Homeobox, conserved site	IPR017970	Conserved_site	243	266	PS00027	PROSITE
P26367	PAX6_HUMAN	Homeobox domain	IPR001356	Domain	208	268	PS50071	PROSITE
P26367	PAX6_HUMAN	Homeobox domain	IPR001356	Domain	210	272	SM00389	SMART
P26367	PAX6_HUMAN	Homeobox domain	IPR001356	Domain	212	267	PF00046	Pfam
P26367	PAX6_HUMAN	Homeobox domain	IPR001356	Domain	212	269	cd00086	Conserved Domain

Exercise2: Exploring a Gene:

6. Is there a *Pax6* orthologue in *D. melanogaster*?

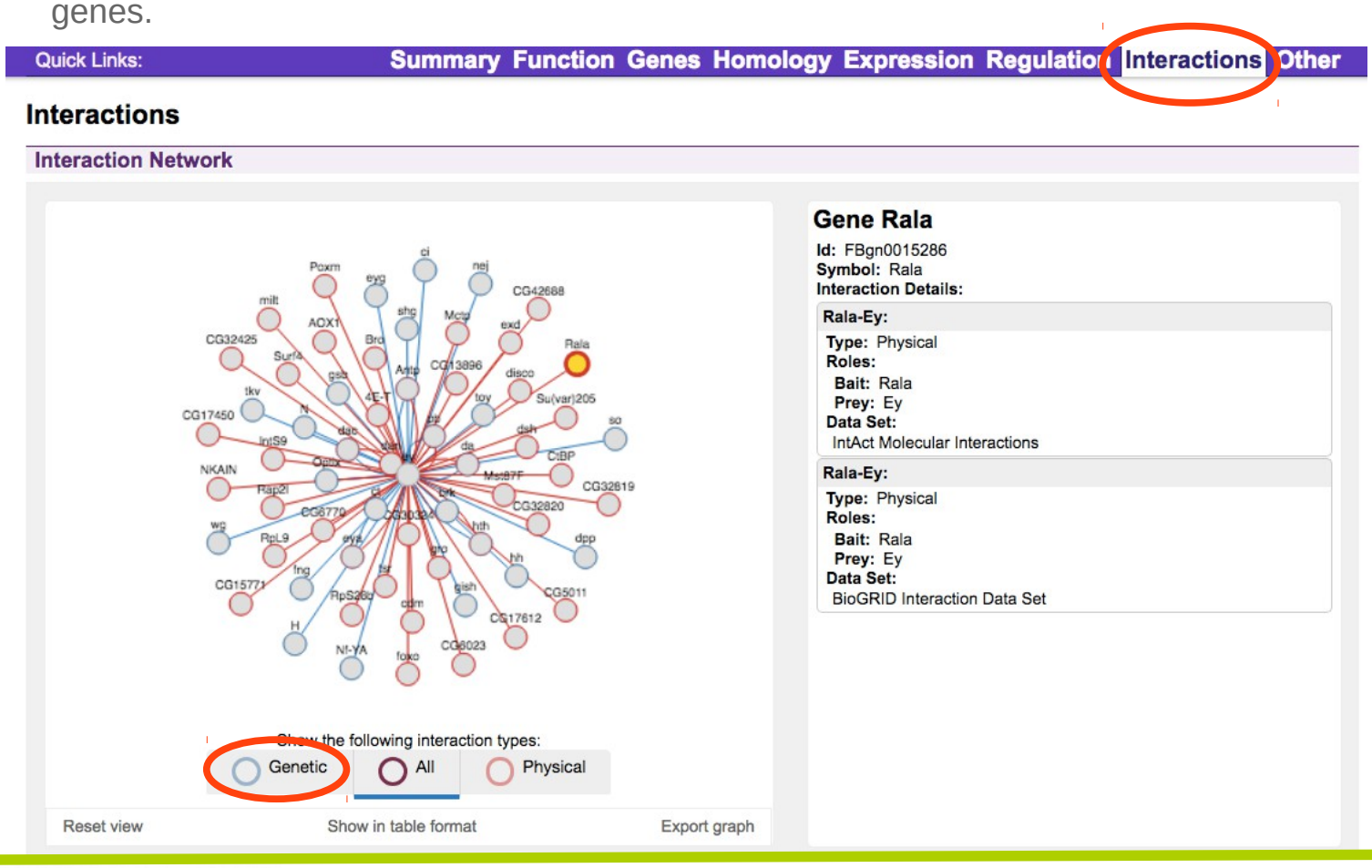


Use the “Links to other Mines” to navigate to the *D. melanogaster* orthologue in FlyMine

Note that there are three othologous fly genes. For this exercise select the first (ey).

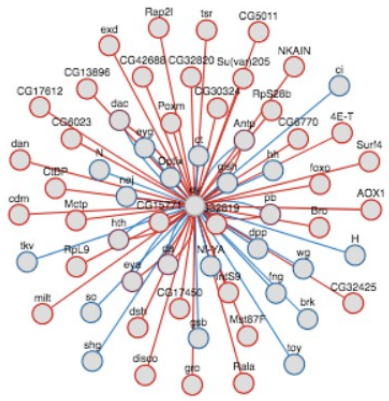
7. Does this orthologue interact with any other proteins? Identify the interaction type (genetic/physical).

Use the “Interactions” quick link to navigate to protein and genetic Interaction data. Ey has both genetic and physical interactions with a number of genes.



Exercise2: Exploring a Gene:

8. For the interaction with 4E-T, what was the original experiment and publication that determined this interaction



a. Select “Show in table format”

b. Select the interaction name for the 4E-T interaction:

Rows per page: 25

Showing 1 to 25 of 153 rows

Gene DB identifier	Gene Symbol	Participant 2 Symbol	Interactions Participant 2 . Primary Identifier	Details Name	Details Role 1	Details Role 2	Details Type	Data Sets Name
FBgn0005558	ey	4E-T	FBgn0052016	ey-4E-T	bait	prey	physical	BioGRID interaction data set
FBgn0005558	ey	AOX1	FBgn0267408	aox1-ey-1	prey	bait	physical	IntAct molecular interactions

c. Select the experiment name

1 Experiment

Manage Columns | Manage Filters | Generate Python code | Export

Manage Relationships | Save as List

Showing rows 1 to 1 of 1

Experiment Description	Experiment Host Organism	Experiment Name
A protein interaction map of Drosophila melanogaster.	NO VALUE	Giot L (2003)

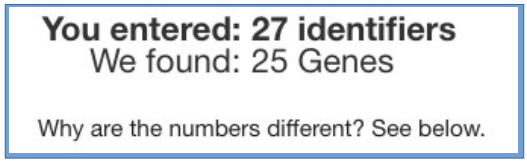
d. Information about the publication is on the experiment page

Publication First Author	Publication Title	Publication Year	Publication Journal	Publication Volume	Publication Pages	Publication PubMed ID
Giot L	A protein interaction map of Drosophila melanogaster.	2003	Science	302	1727-36	14605208

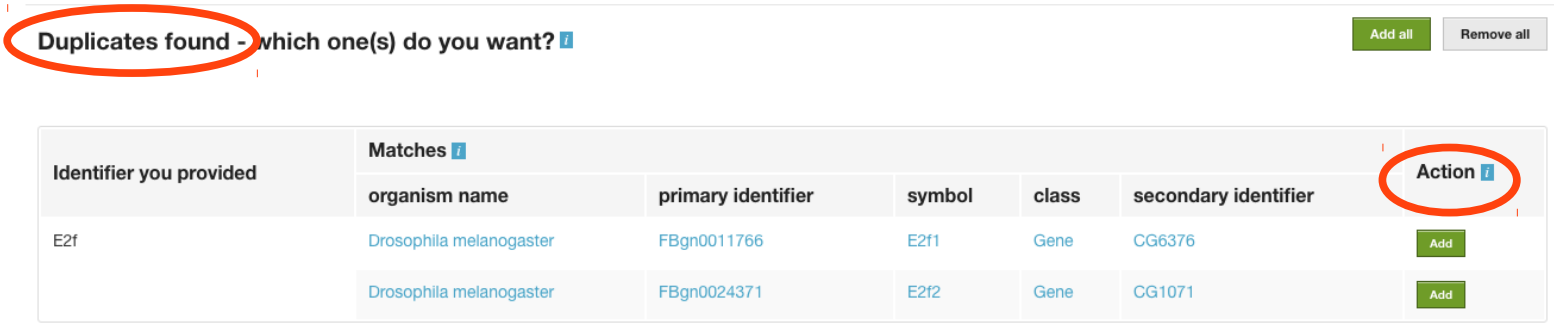
Exercise3: List Upload:

Use FlyMine for this exercise

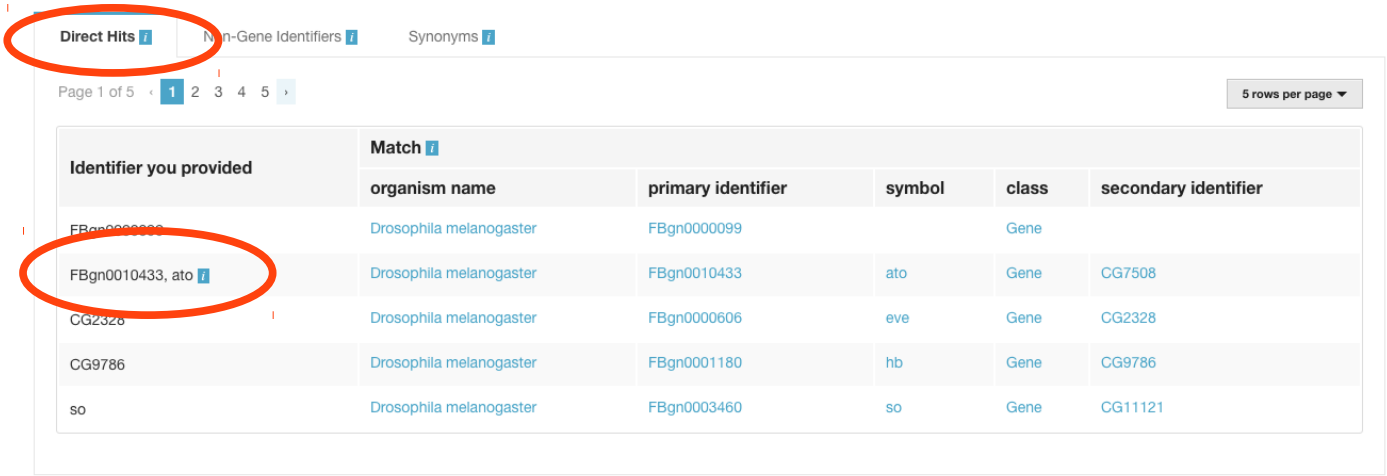
1. Navigate to the lists tab and upload sub-tab
2. Select the example list (leave type and organism as the default values).
3. Click “Create list”.
4. Examine and understand the list page, name and save your list.



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**.

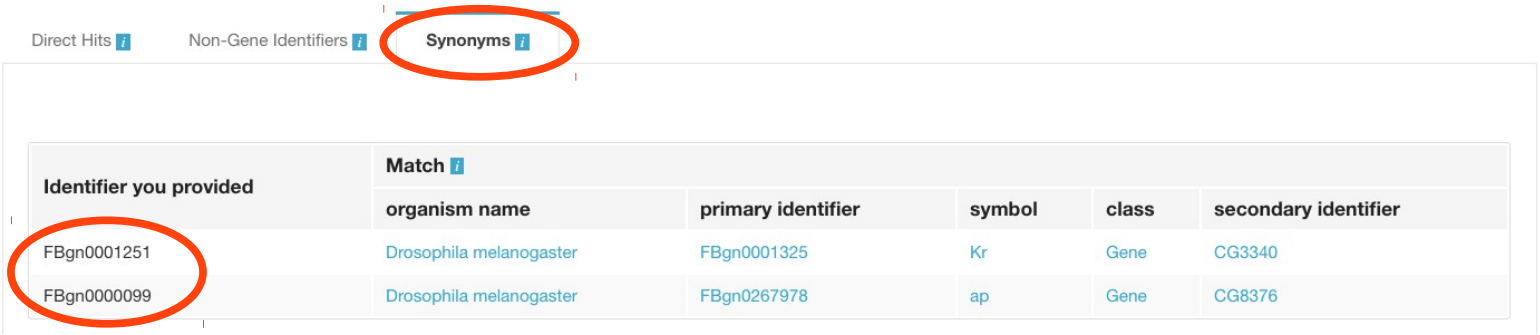


Exercise3: List Upload:

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**.

Identifier you provided	Match				
	organism name	primary identifier	symbol	class	secondary identifier
TWIST_DROME	Drosophila melanogaster	FBgn0003900	twi	Gene	CG2956

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.



Identifier you provided	Match				
	organism name	primary identifier	symbol	class	secondary identifier
FBgn0001251	Drosophila melanogaster	FBgn0001325	Kr	Gene	CG3340
FBgn0000099	Drosophila melanogaster	FBgn0267978	ap	Gene	CG8376

Exercise4: List Analysis Pages:

Use HumanMine for this exercise

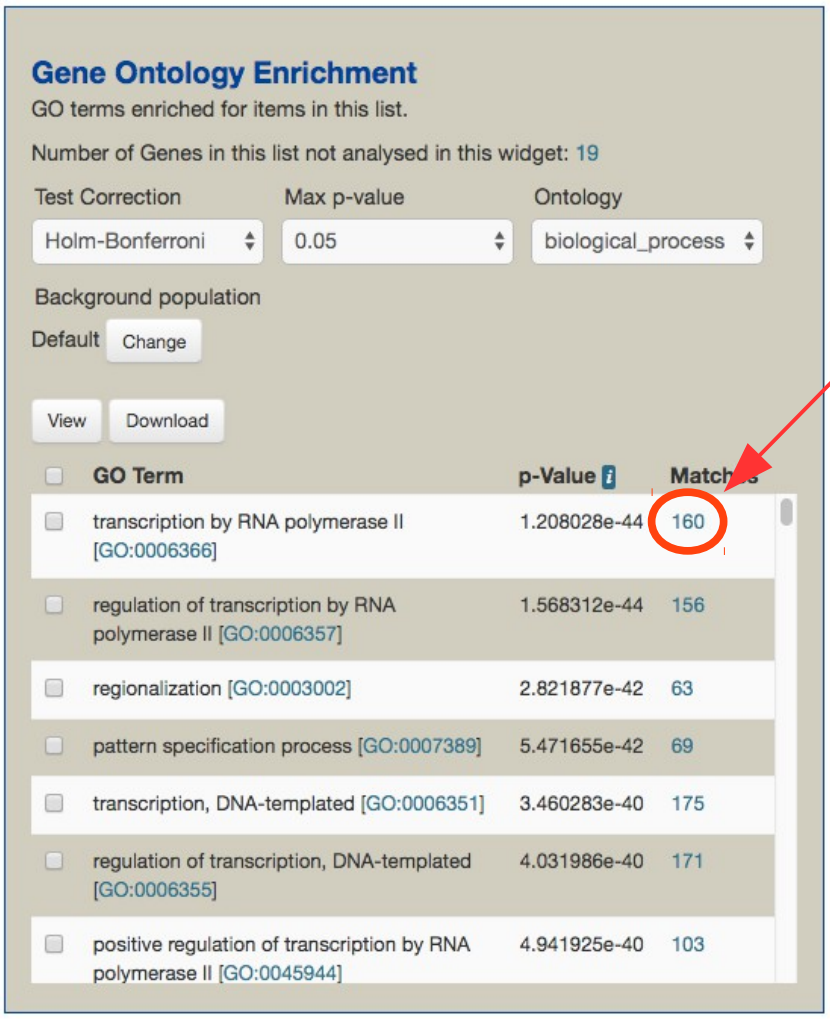
Examine the HumanMine public list: PL_Pax6_Targets (319 genes)

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 (Mammalian Phenotype Ontology)?

Exercise4: List Analysis Pages:

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



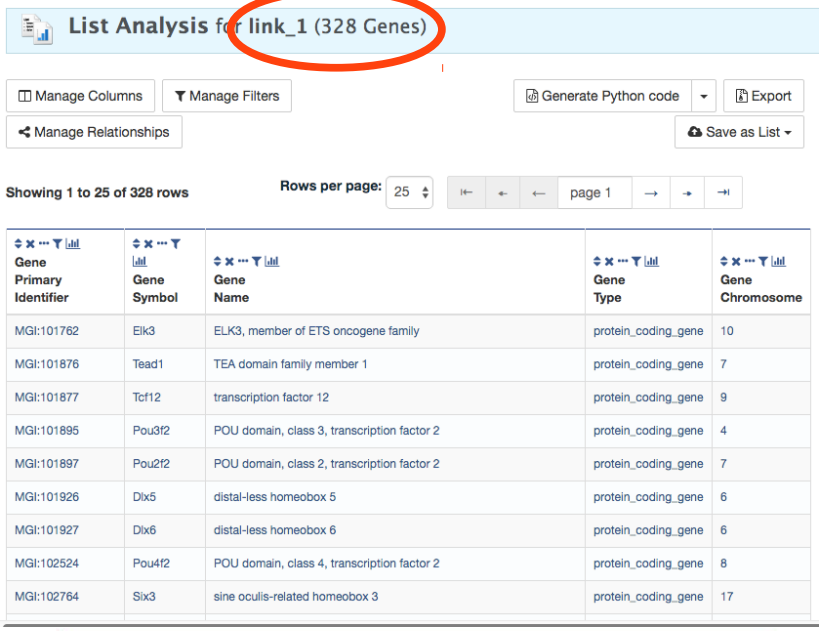
160 genes in the list are annotated with the GO term “transcription by RNA polymerase II”.
This is the most enriched GO term

You can click on this number to create A sub-list of just these 160 genes

Exercise4: List Analysis Pages:

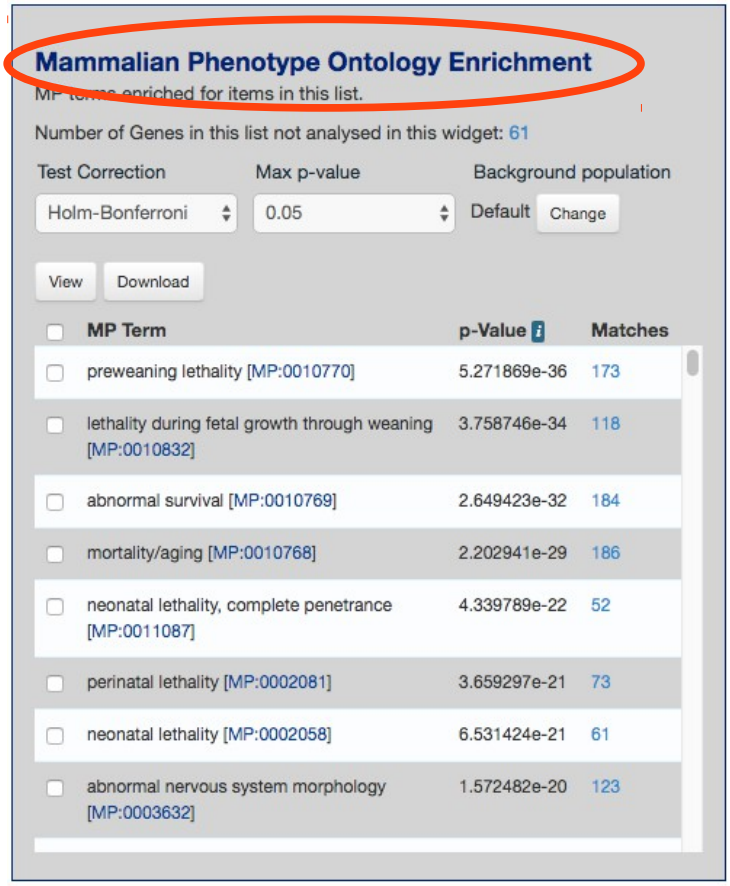
3. Navigate to the MouseMine database to examine the mouse orthologues for this list.

4. How many mouse orthologues are there for this list?



Gene Primary Identifier	Gene Symbol	Gene Name	Gene Type	Gene Chromosome
MGI:101762	Elk3	ELK3, member of ETS oncogene family	protein_coding_gene	10
MGI:101876	Tead1	TEA domain family member 1	protein_coding_gene	7
MGI:101877	Tcf12	transcription factor 12	protein_coding_gene	9
MGI:101895	Pou3f2	POU domain, class 3, transcription factor 2	protein_coding_gene	4
MGI:101897	Pou2f2	POU domain, class 2, transcription factor 2	protein_coding_gene	7
MGI:101926	Dlx5	distal-less homeobox 5	protein_coding_gene	6
MGI:101927	Dlx6	distal-less homeobox 6	protein_coding_gene	6
MGI:102524	Pou4f2	POU domain, class 4, transcription factor 2	protein_coding_gene	8
MGI:102764	Six3	sine oculis-related homeobox 3	protein_coding_gene	17

5. Are these mouse genes enriched for any phenotypes?

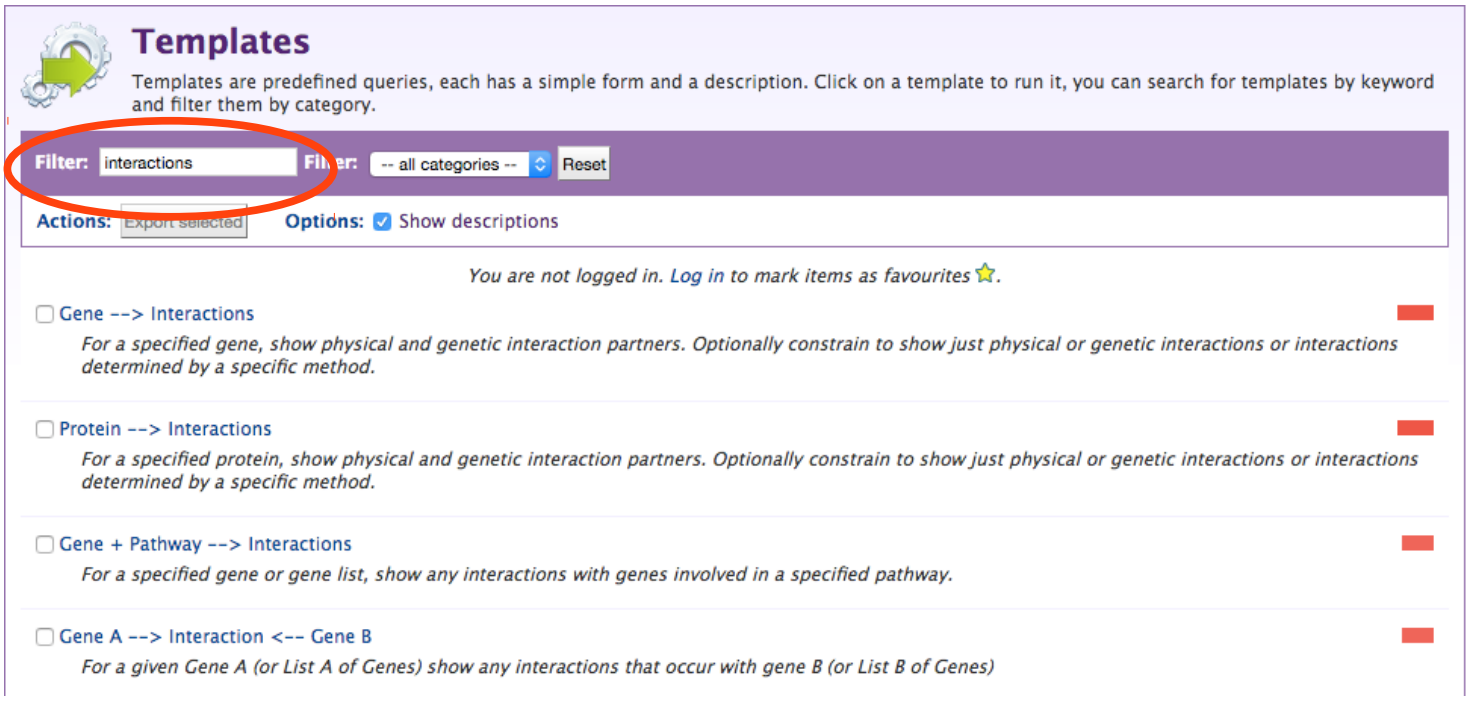


MP Term	p-Value	Matches
preweaning lethality [MP:0010770]	5.271869e-36	173
lethality during fetal growth through weaning [MP:0010832]	3.758746e-34	118
abnormal survival [MP:0010769]	2.649423e-32	184
mortality/aging [MP:0010768]	2.202941e-29	186
neonatal lethality, complete penetrance [MP:0011087]	4.339789e-22	52
perinatal lethality [MP:0002081]	3.659297e-21	73
neonatal lethality [MP:0002058]	6.531424e-21	61
abnormal nervous system morphology [MP:0003632]	1.572482e-20	123

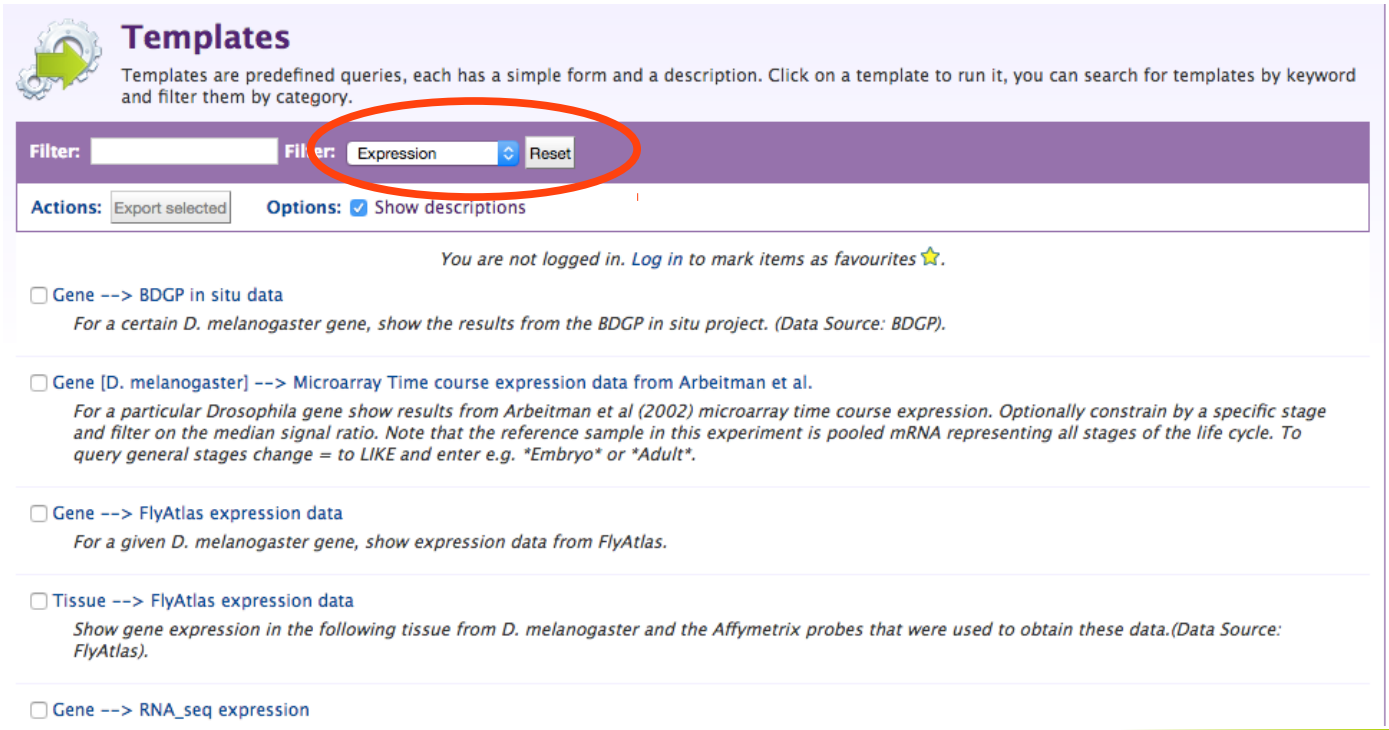
Exercise 5: Template searches:

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 a template search in HumanMine to answer the following question:

Are any of the known targets of Pax6 expressed in the pancreas. Use the public list of known Pax6 target genes in HumanMine (PL_Pax6_Targets).

Use the Protein Atlas dataset for the expression measurement

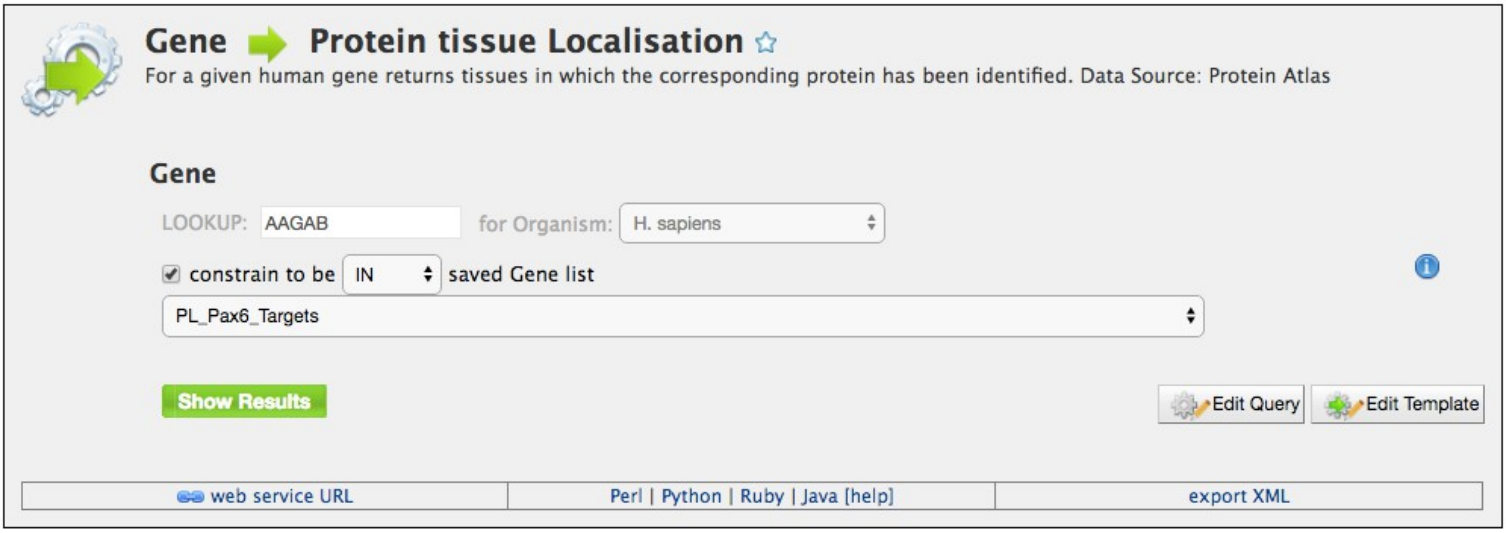
Save the list of target genes expressed in the pancreas as a list.

Exercise 6: Using template searches:

Use HumanMine for this exercise

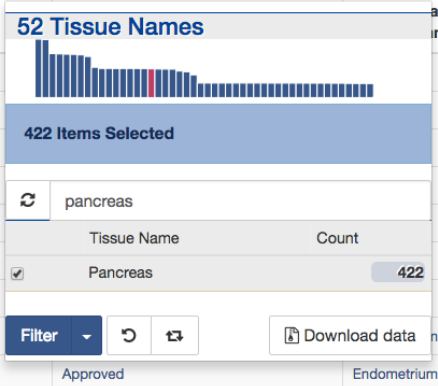
Are any of the known targets of Pax6 expressed in the pancreas.

1. Find the following template and for Gene, select the PL_Pax6_Targets list



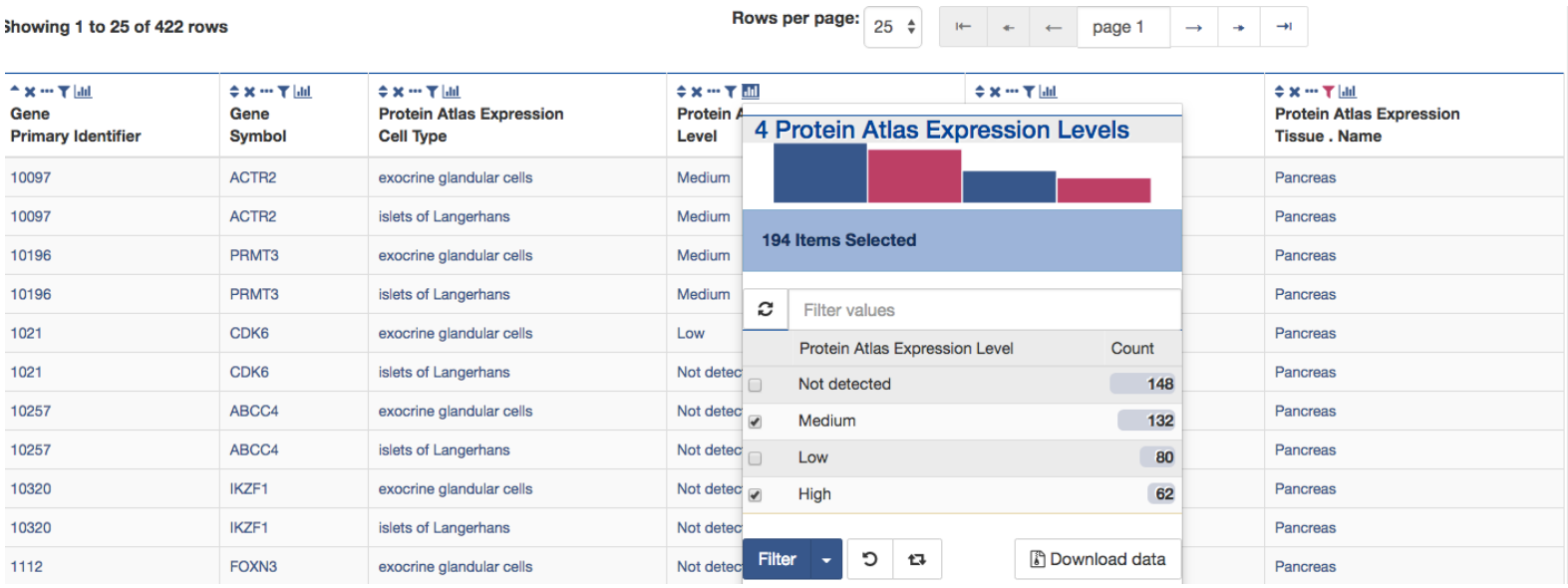
2. Filter the column “Tissue.name” for “Pancreas”

Gene Primary Identifier	Gene Symbol	Protein Atlas Expression Cell Type	Protein Atlas Expression Level		
10097	ACTR2	Langerhans	Not detected		
10097	ACTR2	Leydig cells	Medium		
10097	ACTR2	Purkinje cells	Medium		
10097	ACTR2	adipocytes	Low		
10097	ACTR2	adipocytes	Low		
10097	ACTR2	adipocytes	Not detected		
10097	ACTR2	bile duct cells	Medium		
10097	ACTR2	cells in endometrial stroma	Not detected		
10097	ACTR2	cells in endometrial stroma	Not detected	Approved	Endometrium 2

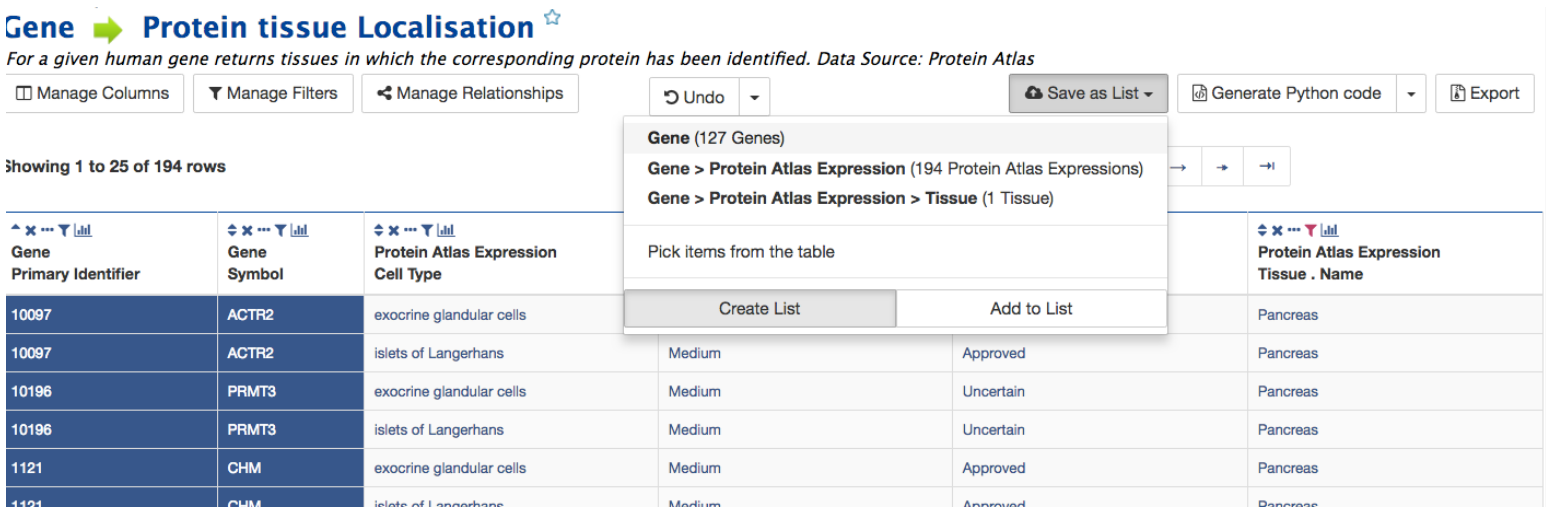


Exercise 6: Using template searches:

3. Filter the Expression.level column for “Medium” and “High”.



4. Save the genes from the resulting table as a list (127 genes)



Exercise 7: Query Builder:

Using HumanMine: we will build a query to show Human genes and associated 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:

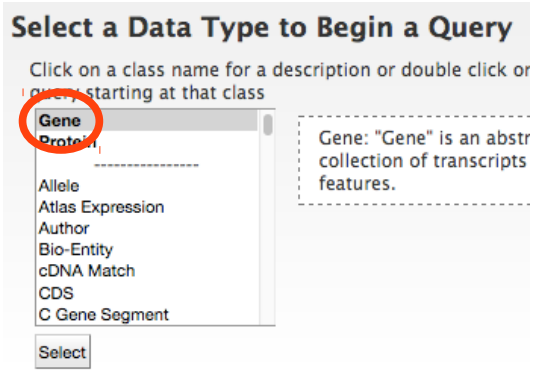
Gene: Primary identifier and Symbol
Disease: name

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

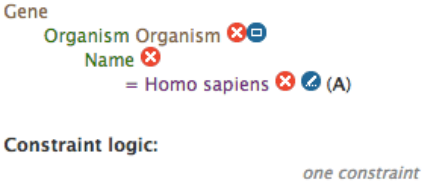
Exercise 7: Query Builder:

Use HumanMine for this exercise

1. Start your query from Gene



2. Constrain “Organism.name” to Homo sapiens



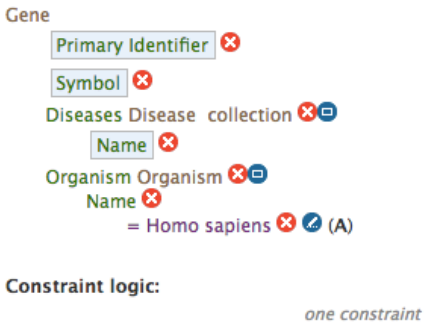
UNIVERSITY OF CAMBRIDGE

InterMine

Exercise 7: Query Builder:

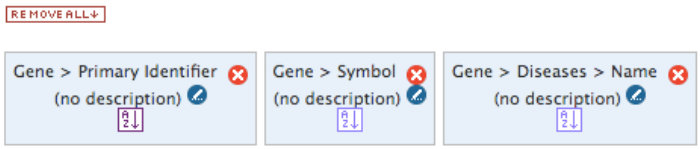
3. Add the columns of data we want in our results:

Gene: Primary identifier and Symbol
Disease: name

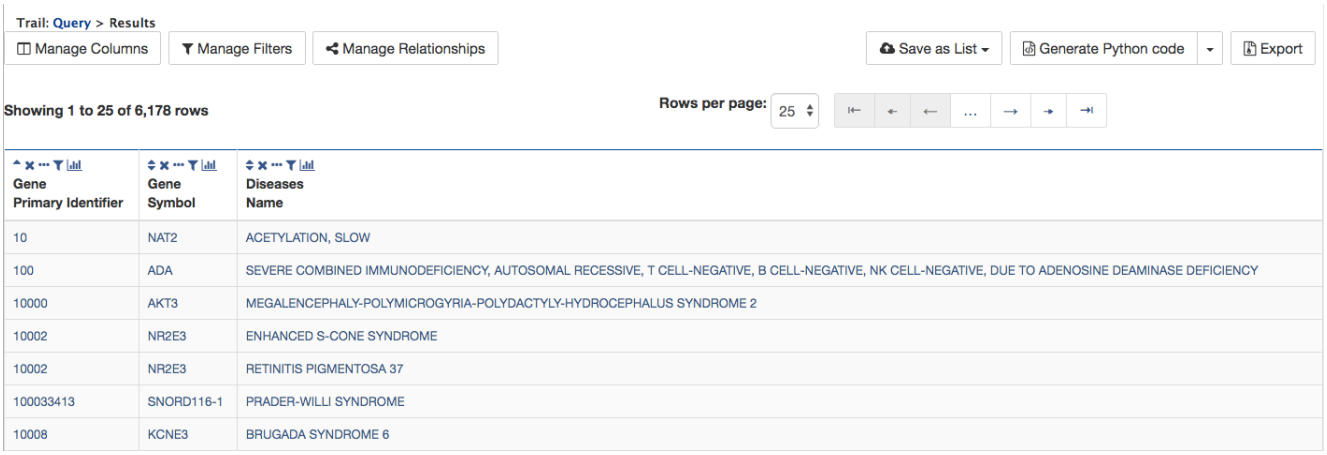


Columns to Display

Use the SHOW or SUMMARY links to add fields to the results table. Click and drag the blue output boxes to choose the output column order.Click to choose a column to sort results by, click again to select ascending or descending .Use the REMOVEALL link to remove all fields from the results table.



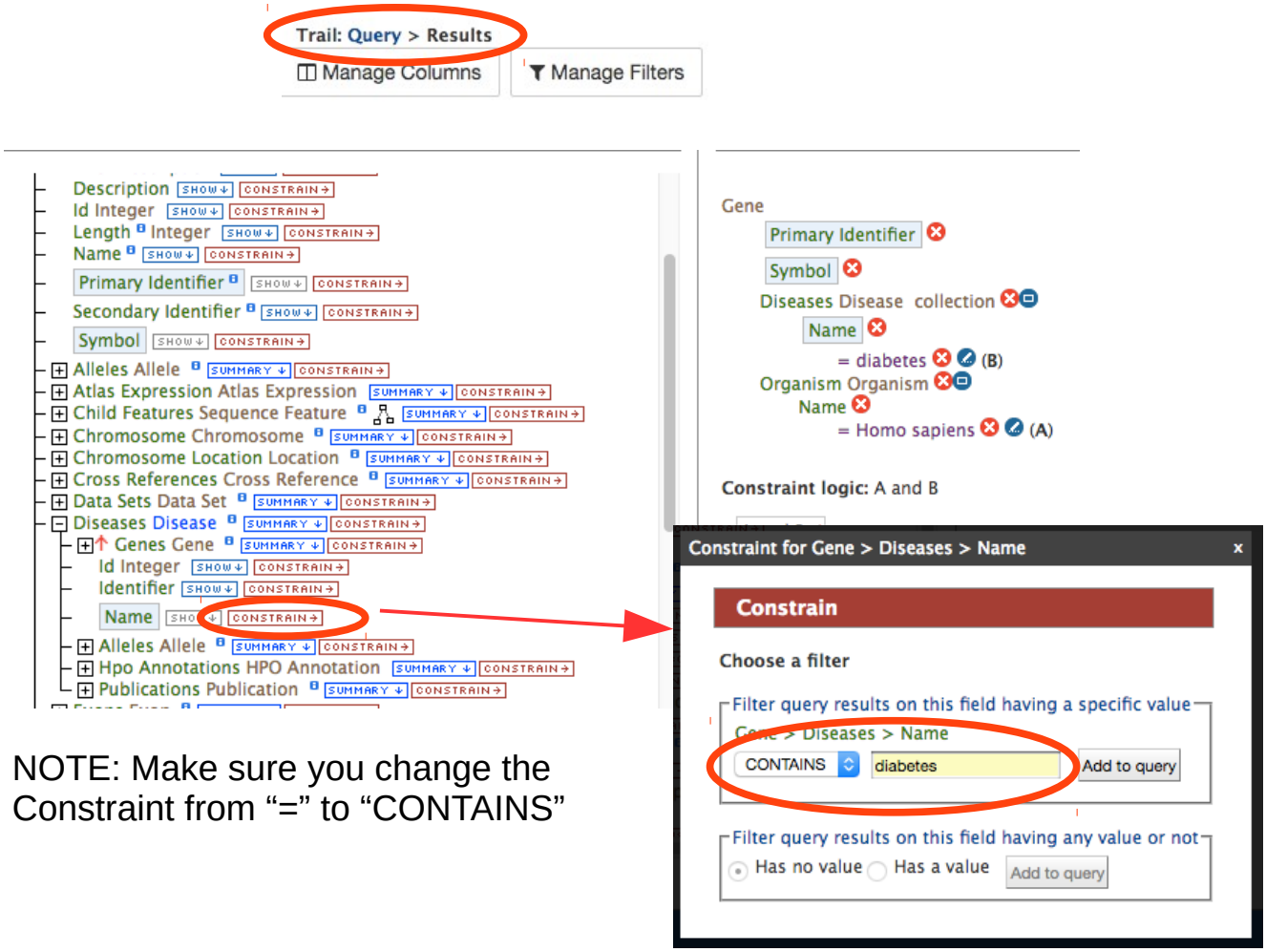
4. Run this search - 'Show results'.



Gene Primary Identifier	Gene Symbol	Diseases Name
10	NAT2	ACETYLATION, SLOW
100	ADA	SEVERE COMBINED IMMUNODEFICIENCY, AUTOSOMAL RECESSIVE, T CELL-NEGATIVE, B CELL-NEGATIVE, NK CELL-NEGATIVE, DUE TO ADENOSINE DEAMINASE DEFICIENCY
10000	AKT3	MEGALENCEPHALY-POLYMICROGYRIA-POLYDACTYLY-HYDROCEPHALUS SYNDROME 2
10002	NR2E3	ENHANCED S-CONE SYNDROME
10002	NR2E3	RETINITIS PIGMENTOSA 37
100033413	SNORD116-1	PRADER-WILLI SYNDROME
10008	KCNE3	BRUGADA SYNDROME 6

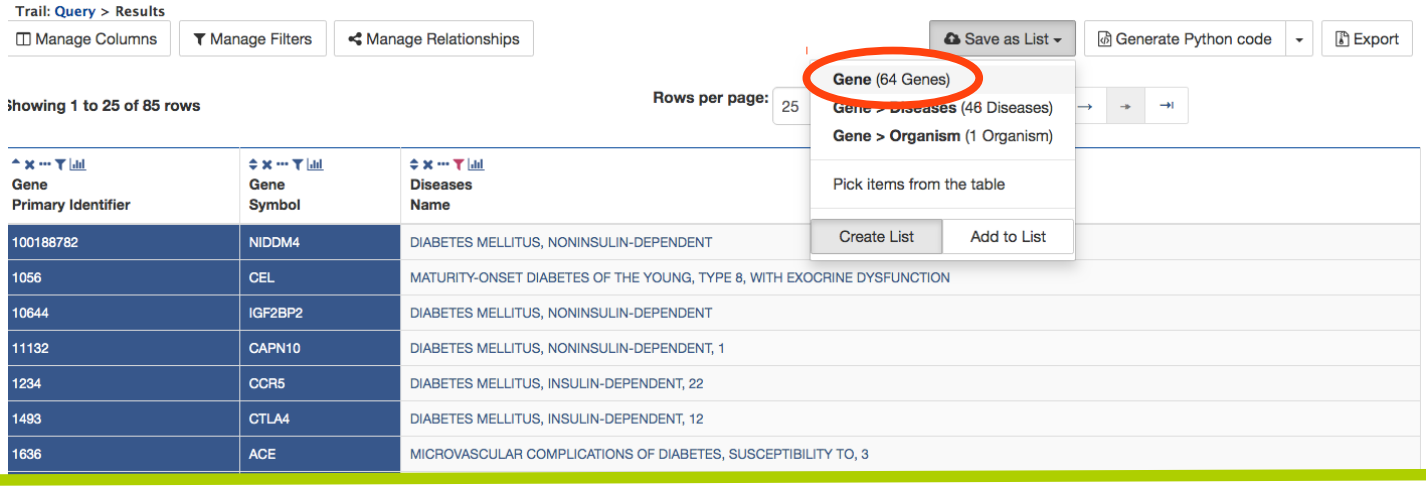
Exercise 7: Query Builder:

5. Return to the query (Use the "Trail" in the top left) and add a constraint to Disease name for "CONTAINS *Diabetes*"



NOTE: Make sure you change the Constraint from "=" to "CONTAINS"

6. Run the search and save the set of genes (64 genes)



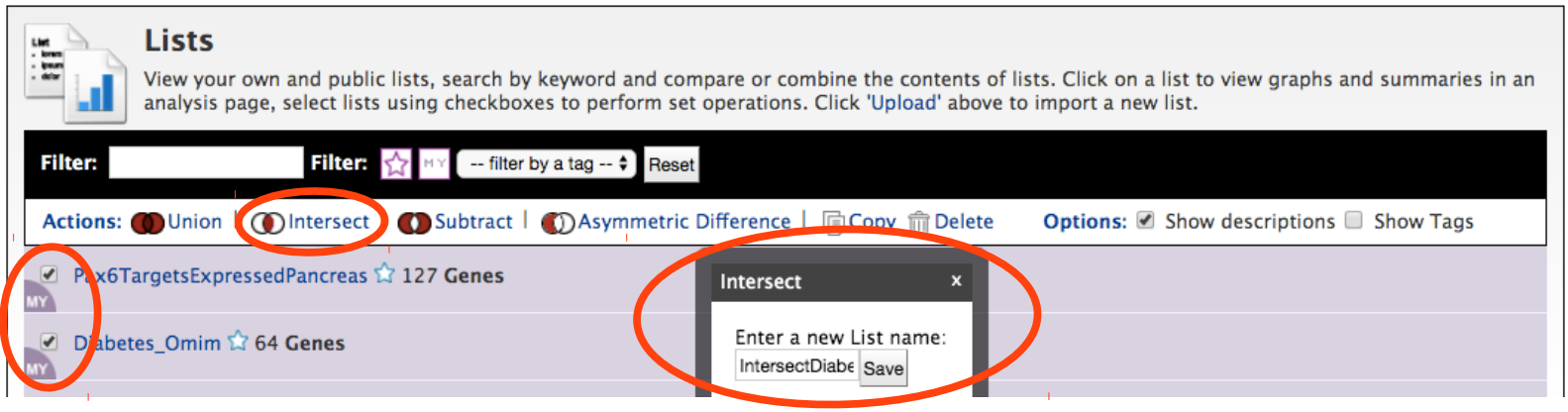
Gene Primary Identifier	Gene Symbol	Diseases Name
100188782	NIDDM4	DIABETES MELLITUS, NONINSULIN-DEPENDENT
1056	CEL	MATURITY-ONSET DIABETES OF THE YOUNG, TYPE 8, WITH EXOCRINE DYSFUNCTION
10644	IGF2BP2	DIABETES MELLITUS, NONINSULIN-DEPENDENT
11132	CAPN10	DIABETES MELLITUS, NONINSULIN-DEPENDENT, 1
1234	CCR5	DIABETES MELLITUS, INSULIN-DEPENDENT, 22
1493	CTLA4	DIABETES MELLITUS, INSULIN-DEPENDENT, 12
1636	ACE	MICROVASCULAR COMPLICATIONS OF DIABETES, SUSCEPTIBILITY TO, 3

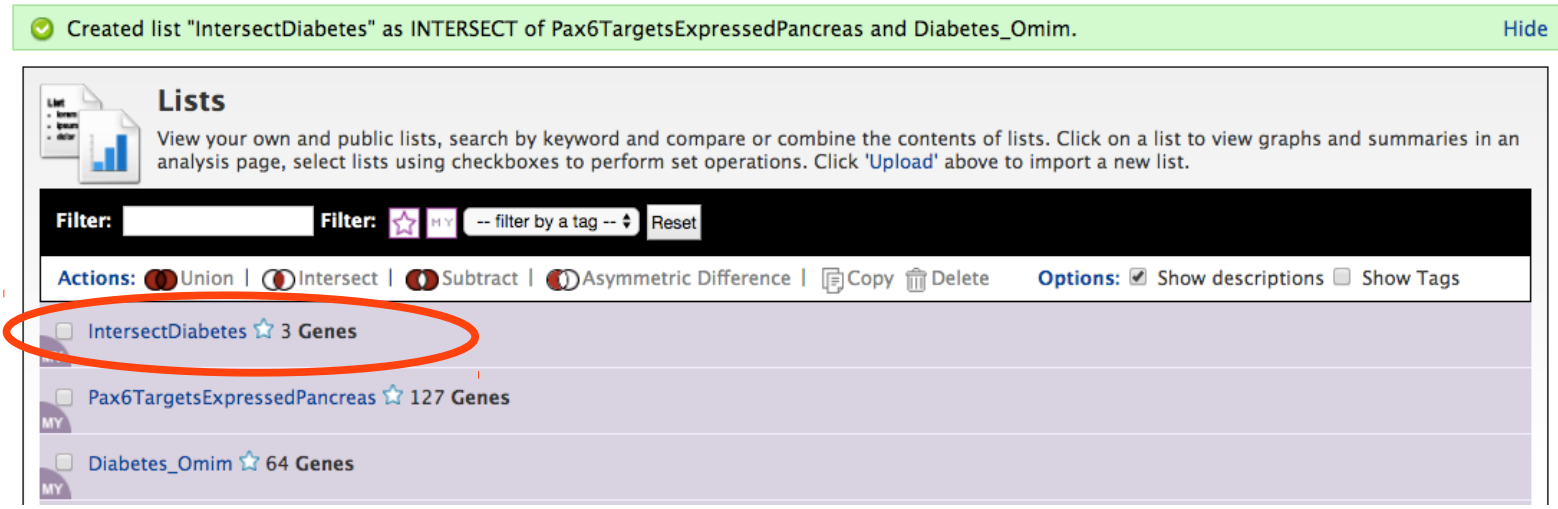
Exercise 8: Analysis Workflows:

Use HumanMine for this exercise

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.



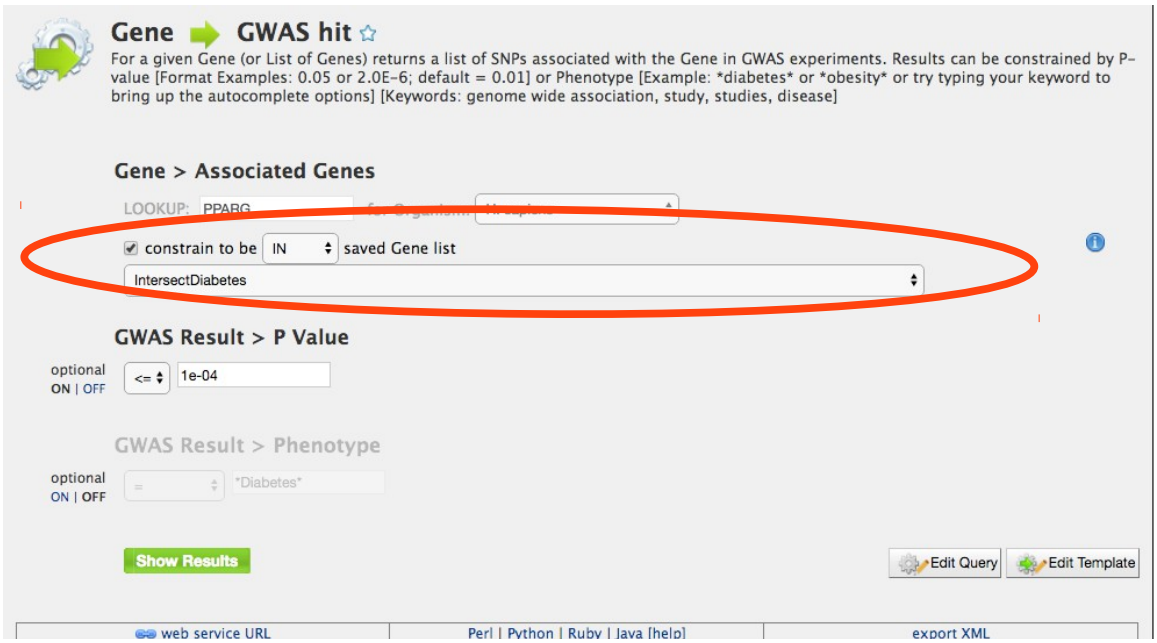


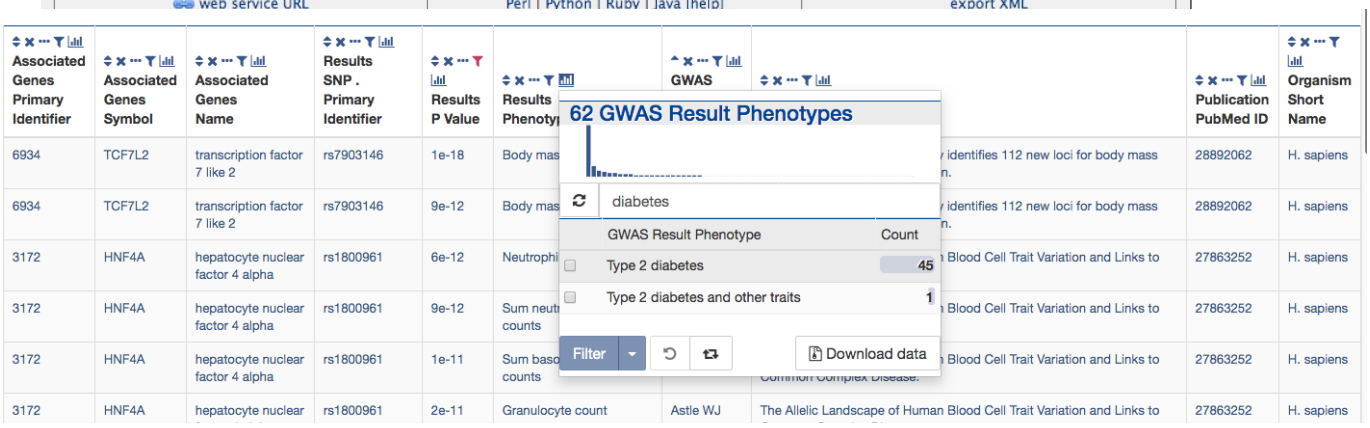
Three genes are found that are Pax6 targets expressed in the pancreas, and that are associated with diabetes. You can find out more about these genes by examining the list analysis page and their report pages.

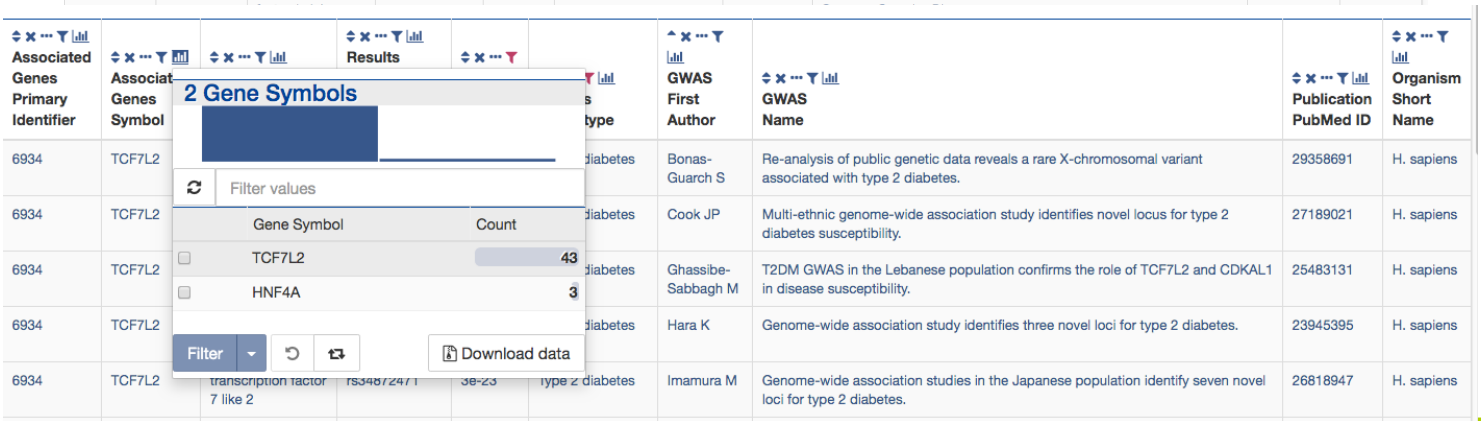
Exercise 8: Analysis Workflows:

3. We now want to know if this gene has been identified in GWAS studies. Run a template on your intersected list.

4. Use the column summary to find if any of the GWAS phenotypes are related to diabetes.







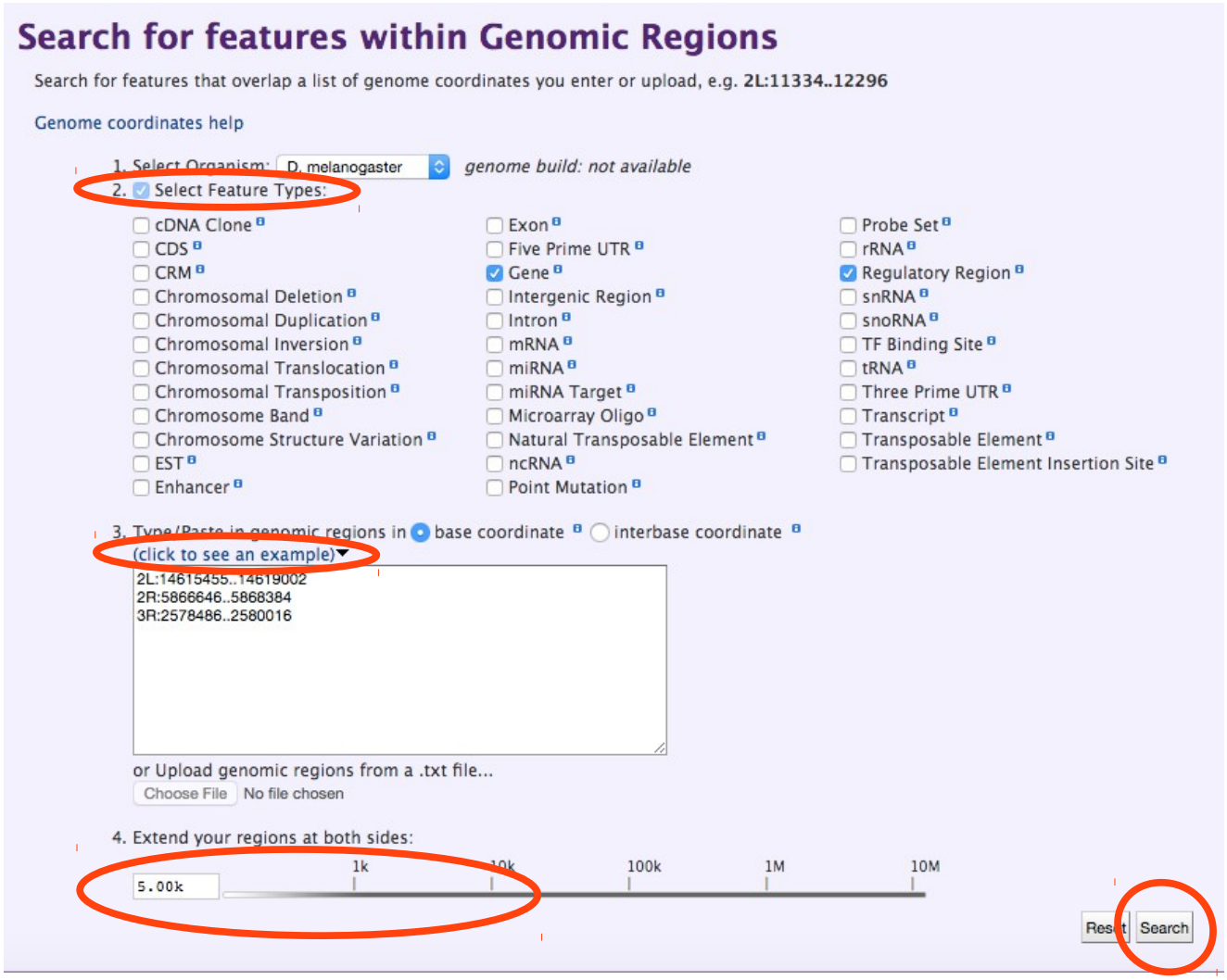
UNIVERSITY OF CAMBRIDGE

InterMine

Exercise 9: Region Search:

Use FlyMine for this exercise

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



Exercise 9: Region 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.
7. Notice that you can also export the sequences for your regions in fasta format.