FAANGMine Documentation

Release 1.6

Elsik Lab

Jul 24, 2023

Table of Contents

FAANGMine is a data mining resource that integrates reference genome assemblies for cattle, horse, pig, sheep, chicken, cat, dog and water buffalo with many other biological data sets. Powered by InterMine, this platform provides access to a number of datasets from a variety of sources. It also provides customized bioinformatics tools that researchers can use to create their own custom datasets. FAANGMine is part of FAANGMine.org. The FAANG (Functional Annotation of ANimal Genomes) Consortium is "a coordinated international action to accelerate genome to phenome" and aims to generate comprehensive maps of functional elements in genomes of domesticated animals. FAANGMine will integrate data generated by the FAANG Consortium for animal researchers with or without bioinformatic programming skills to use in their own research projects.

Main site: http://faangmine.org/faangmine

Link to the available datasets in FAANGMine: http://faangmine.org/faangmine/dataCategories.do

FAANGMine.org is based upon work supported by the National Science Foundation under Award Number 1759896. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. FAANGMine is developed and hosted at the University of Missouri. If you have comments or if you wish to report a problem, please contact the Database Administrator.

CHAPTER 1

Overview of FAANGMine

This section provides a brief overview of the layout for FAANGMine.

The navigation panel highlights different functionalities of FAANGMine.

Home - The home page for FAANGMine

MyMine - The MyMine serves as a portal for account management. When logged in to FAANGMine Users can access their saved templates, most recent queries and saved lists.

Templates - List of templates that users can select from based on the nature of their query. Each template is a predefined query with a simple form containing a description of what input is expected and the type of output that will be generated.

Lists - Allows users to upload lists of genes on which they can perform enrichment analyses and export the results. Users that log in to FAANGMine can save their lists for future use.

QueryBuilder - A flexible interface that allows users to create their own custom query template while browing the FAANGMine data models. Queries can be exported in a variety of formats to share with other users.

Regions - The Genomic Region Search tool where users can enter a series of genomic coordinates, specify flanking regions and fetch all features that fall within the given interval. The result can be exported or saved as a list for further analyses.

Data Sources - Provides a summary of all the data loaded into FAANGMine including their sources, associated publications and links to source sites.

Help - Links to the FAANGMine help docs and tutorials

API - Describes the InterMine API that allows users to programmatically access FAANGMine.

CHAPTER 2

Navigation and Searching in FAANGMine

There are several ways for users to query FAANGMine.

2.1 Quick Search

Quick Search allows users to search keywords from any of the FAANGMine datasets. There is a Quick Search box on the FAANGMine home page or a smaller search box in the upper right corner of all pages.

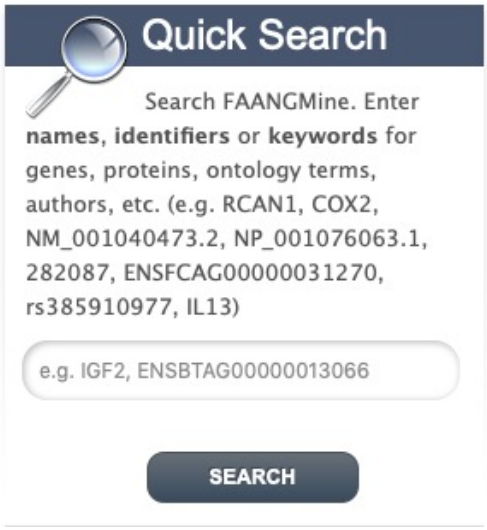


Fig. 1: Quick Search from home page

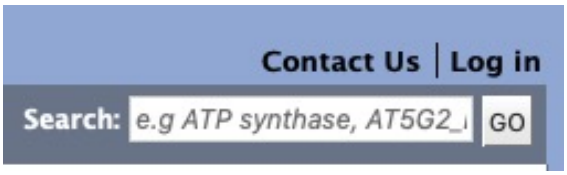


Fig. 2: Quick Search from any page

Quick Searches can be conducted with a number of identifiers including gene names, transcripts, pathways, gene identifiers or organisms. The wildcard character * can be used to retrieve all results that match a particular search query.

For an example, we will use the lysozyme gene *LYZ* as an example. Enter LYZ in the search box and click **Search**. The results page is tabulated and displays a summary about your query, as shown below.

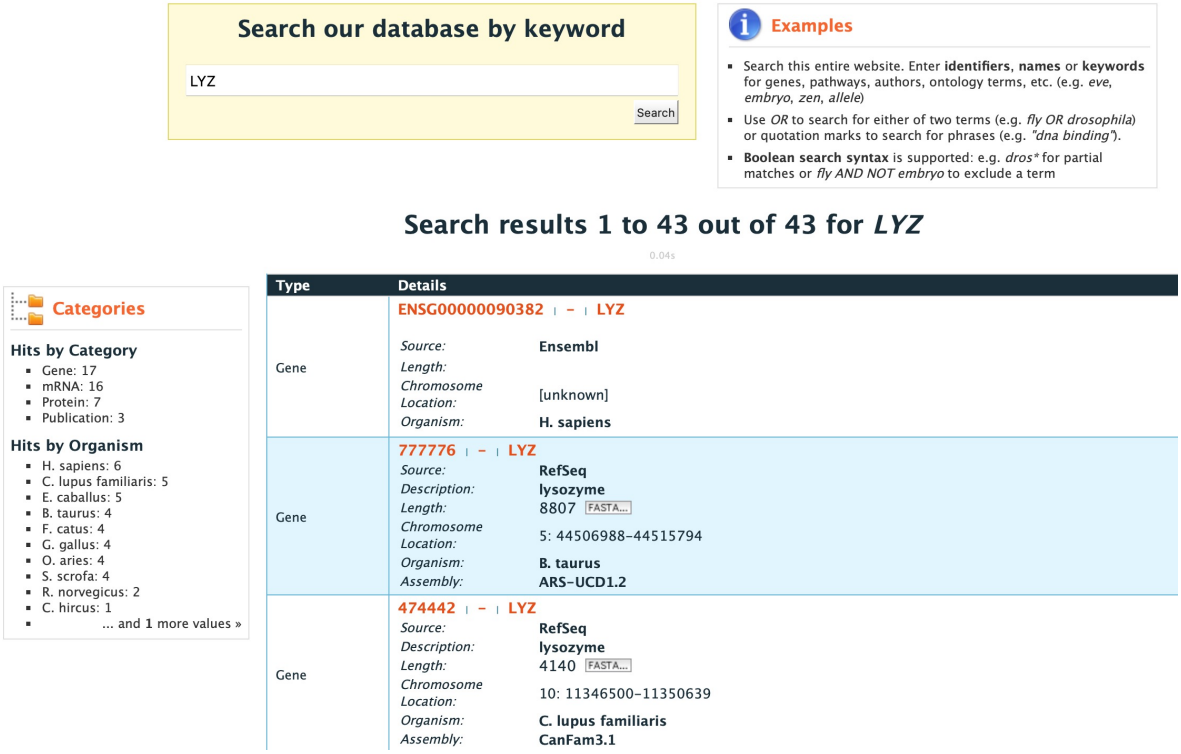


Fig. 3: Results table for Quick Search with the gene *LYZ*

Hits are summarized in the box to the left of the results tables and can be filtered based on **Category** and **Organism**. Clicking on any of them will filter by the selected category. Note that for results with sequence data available, the sequence can be downloaded in FASTA format by clicking on the FASTA box within the hit Details box. The score column in the result table indicates the similarity of your query to each of the hits. The results page can also be converted to a list (and saved if users are logged in). To enable this feature click on **Gene** in the **Hits by Category** then click on *C. lupus familiaris* in **Hits by Organism**.

After the table has been filtered for gene and organism, checkboxes will be available for users to select genes they would like to add to their list. Once the genes are selected, click on **CREATE LIST**. See the lists section for more detail on creating and saving lists.

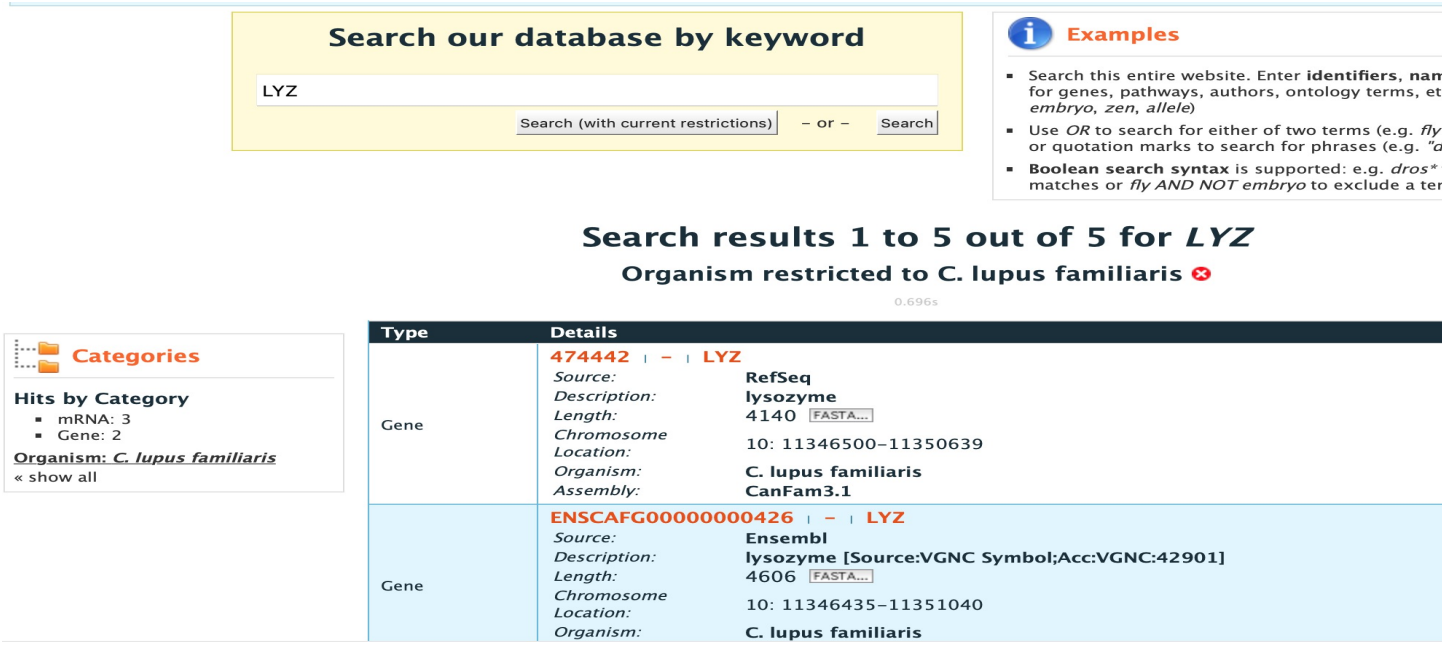


Fig. 4: Results table for Quick Search with the gene *LYZ* filtered by gene then by organism ..

2.2 Templates

Templates or predefined queries are another search method within FAANGMine. Popular templates are displayed on the home page, grouped by category (e.g., Genes, Proteins, Interactions) and the complete list can be seen by clicking the **Templates** menu tab.

As an example, the **Gene -> Homologues** template queries FAANGMine to retrieves all homologue for a given gene. Here, we will do a search for the gene *GSTM1*.

The results page displays all of the homologues for that query gene. When logged in to FAANGMine, users can save their results as a list for further analyses by clicking on the **Save as List** button above the results table then choosing columns to save in their list. See the lists section for more detail on creating and saving lists. Note the "Trail: Query" text at the upper left of the results table. Clicking on the "Query" link will bring you back to the query that generated the table to allow for edits without having to start with a new template.

Example: Gene –> Homologue template search results, identifier for Gene GSTM1

2.2.1 Generate query code

The code for each template query can be retrieved by clicking on the arrow next to **Generate Python Code** and choosing the desired language from the pull-down menu. The language options are Python, Perl, Java, Ruby, JavaScript, and XML.

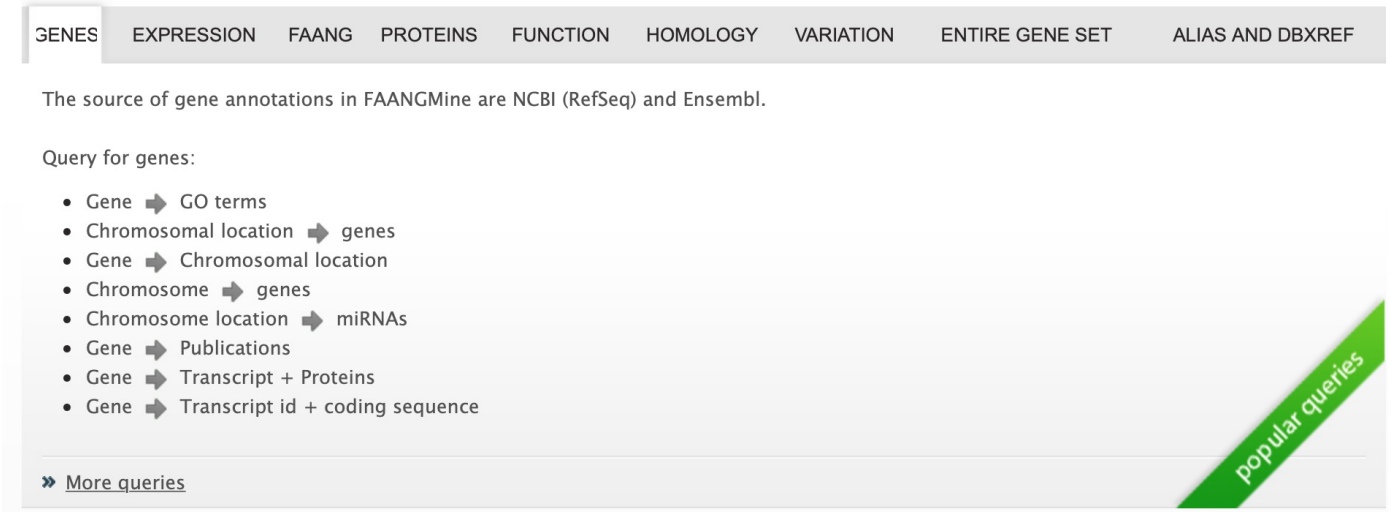


Fig. 5: Popular templates

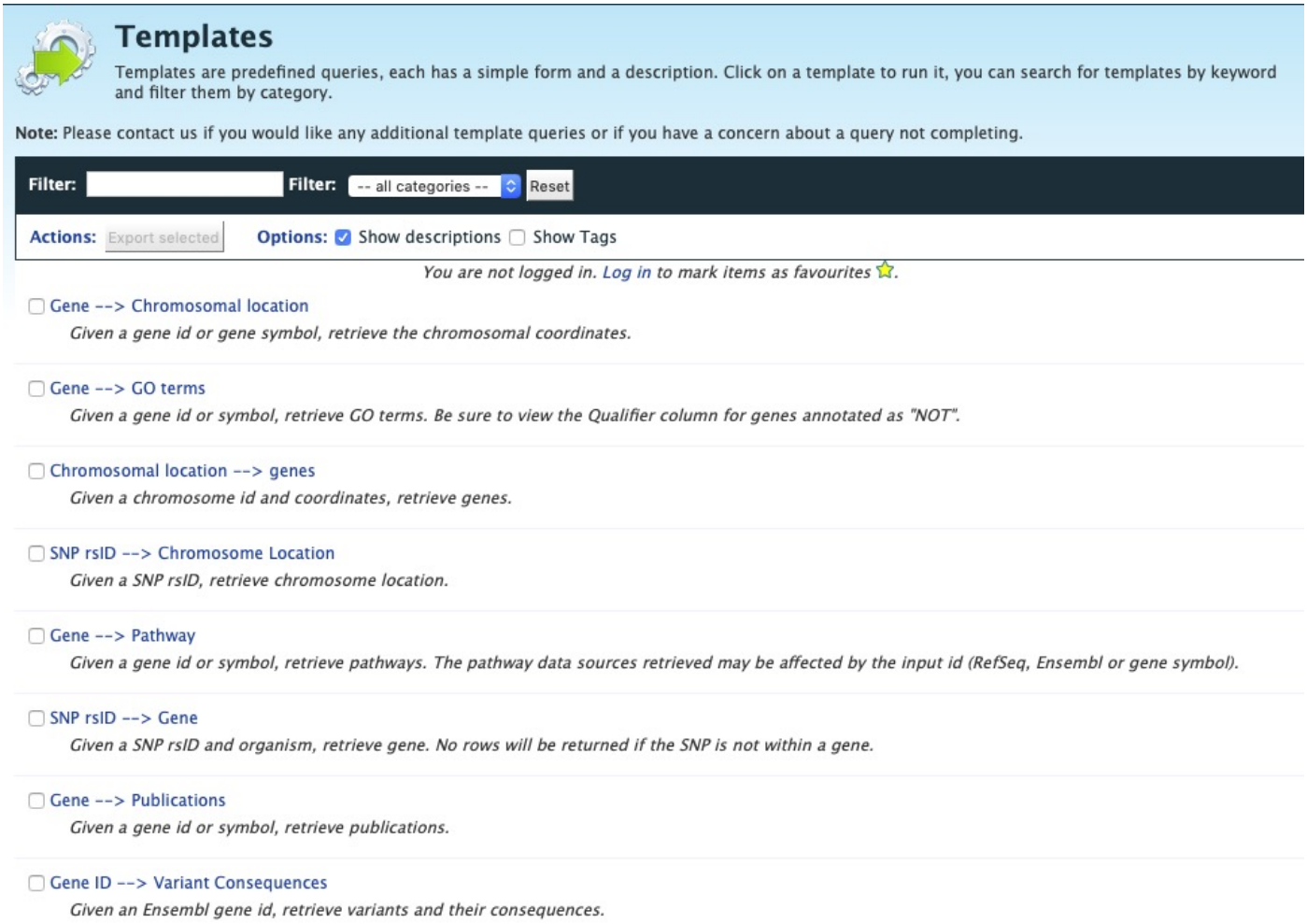


Fig. 6: Full list of templates on Templates page

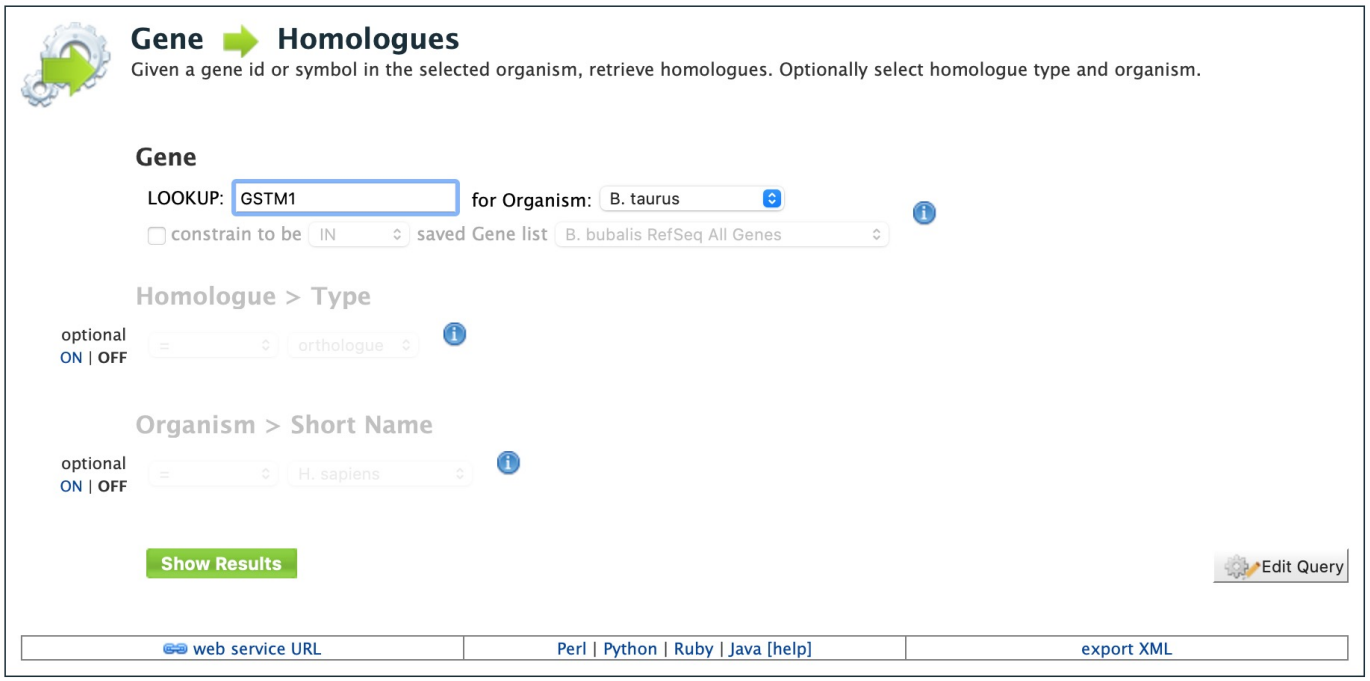


Fig. 7: Example: Gene –> Homologue

Trail: Query

Gene → Homologues

Given a gene id or symbol in the selected organism, retrieve homologues. Optionally select homologue type and organism.

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

Showing 1 to 25 of 289 rows

Rows per page: 25 | page 1

Gene Organism	Gene Gene ID	Gene Source	Homologues Homologue . Organism . Short Name	Homologues Last Common Ancestor	Homologues Type	Homologues Homologue . Primary Identifier	Homologue Source	Data Sets Name
B. taurus	327709	RefSeq	B. bubalis	Cetartiodactyla	orthologue	102396303	RefSeq	OrthoDB data set
B. taurus	327709	RefSeq	B. bubalis	Cetartiodactyla	orthologue	102397004	RefSeq	OrthoDB data set
B. taurus	327709	RefSeq	B. bubalis	Cetartiodactyla	orthologue	102398085	RefSeq	OrthoDB data set
B. taurus	327709	RefSeq	B. bubalis	Mammalia	orthologue	102395774	RefSeq	OrthoDB data set
B. taurus	327709	RefSeq	B. bubalis	Mammalia	orthologue	102396303	RefSeq	OrthoDB data set
B. taurus	327709	RefSeq	B. bubalis	Mammalia	orthologue	102397004	RefSeq	OrthoDB data set

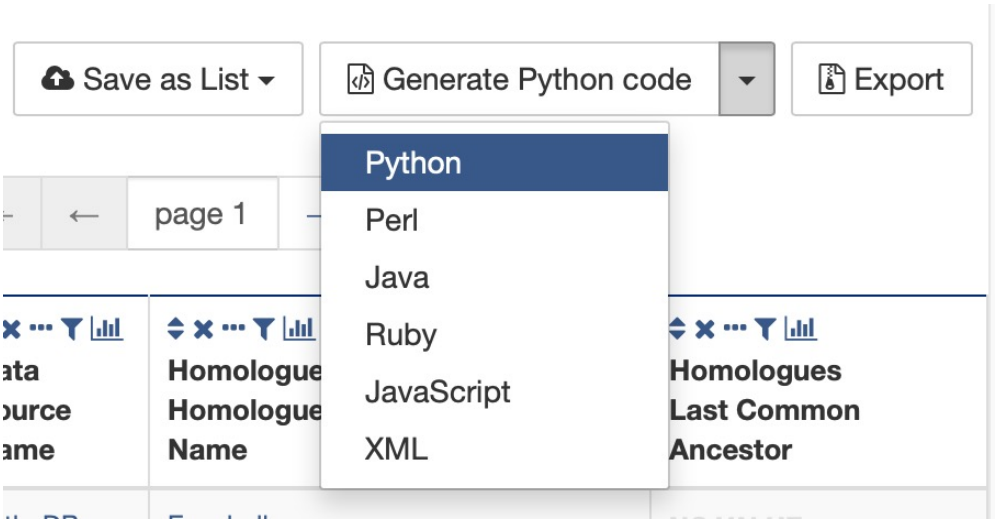


Fig. 8: Options for generating code from template query

2.2.2 Download results

The search results from a template query can be downloaded by clicking the **Export** button above the table and choosing the desired format from the pull-down menu to the right of the File name field. Available formats are tab-separated values, comma-separated values, XML, and JSON. When the results contain genomic features, they may also be downloaded in FASTA, GFF3, or BED format. Other options may be specified in the submenu to the left of the download box. By default, all rows and all columns are downloaded, but individual columns may be included or excluded by clicking on the toggles next to the column headers in the **All Columns** submenu. The number of rows and row offset are set in the **All Rows** submenu. Download the results as a compressed file by choosing GZIP or ZIP format in the **Compression** submenu (default is **No Compression**). Column headers are not added by default but may be included under the **Column Headers** submenu. Finally, the **Preview** submenu displays the first three rows of the file to be downloaded so that the desired format and options may be finalized before beginning the download. When ready, click the **Download file** button to download the results.

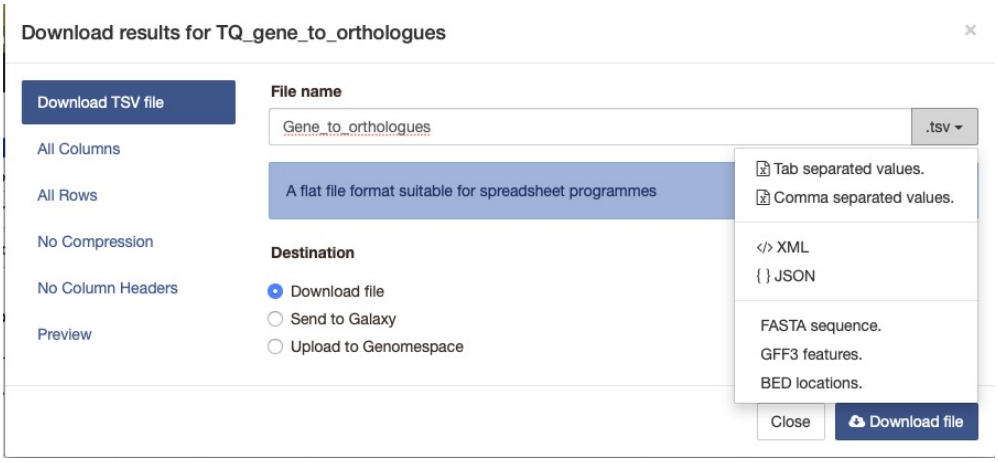


Fig. 9: Options for downloading results from template query

2.2.3 Customize output and manage columns

To customize the results table layout, click the **Manage Columns** button. This allows users to rearrange, remove or order columns. Filters can be edited by clicking the **Manage Filters** button. To specify the entity relationships within the query and change the way the results are presented in the table, click **Manage Relationships**. Clicking on the **Manage Relationships** option bring also brings up a blue information panel ("What does this do?") that provides more detailed information.

Fig. 10: Options for customizing the results of a template query.

To further manage column data, each column has a set of icons in its header. Mousing over any icon will reveal what it does. The sort icon resembles two triangles and allows you to sort a column in either direction (ascending/descending; a->z, z->a). To delete any column from your table, click on the "x" icon. Alternatively, the three dots "…" can be selected to "Toggle column visibility" and hide the column. Once hidden, the icon turns into a double arrow that can be selected to expand the column into its original form. The next icon resembling a funnel can be selected to edit or remove any currently active filters. A summary of the data within a column can be viewed by clicking on the icon that resembles a graph. The data within each summary can be selected for further filtering or downloading.

Fig. 11: Column header icons available to edit the results of a template query.

2.2.4 Optional filters

****If you are using a template with an optional filter and want to edit the template, the "Query" link at the top left of the results page is not functioning properly. Additionally, if you click on "Edit Query" in a template, the optional settings becomes non-optional and the default constraint is applied. To edit the template query, navigate back to the template by using the "Templates" tab on the home page menu bar.**

Some templates have optional filters that are disabled by default. For the Gene –> Homologue query, there is an optional filter to specify the organism used in the query. To enable the filter, click **ON** below the **Organism > Short Name** label.

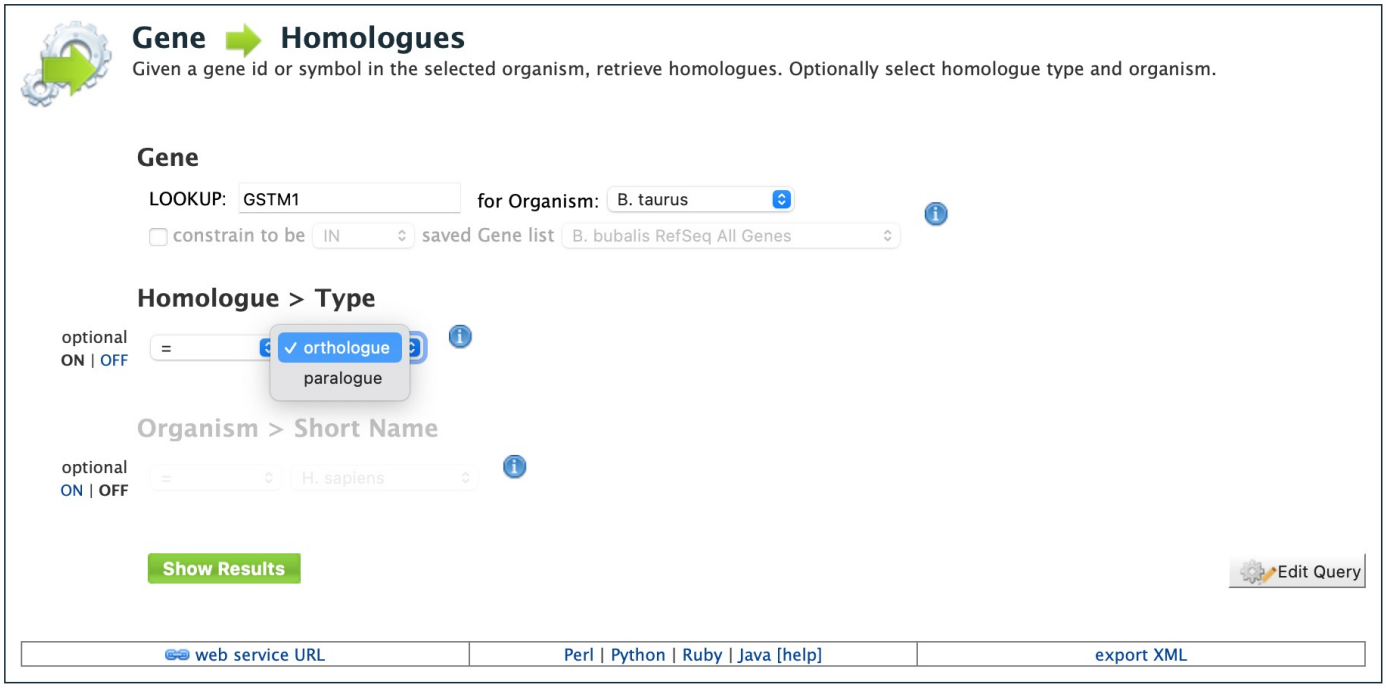
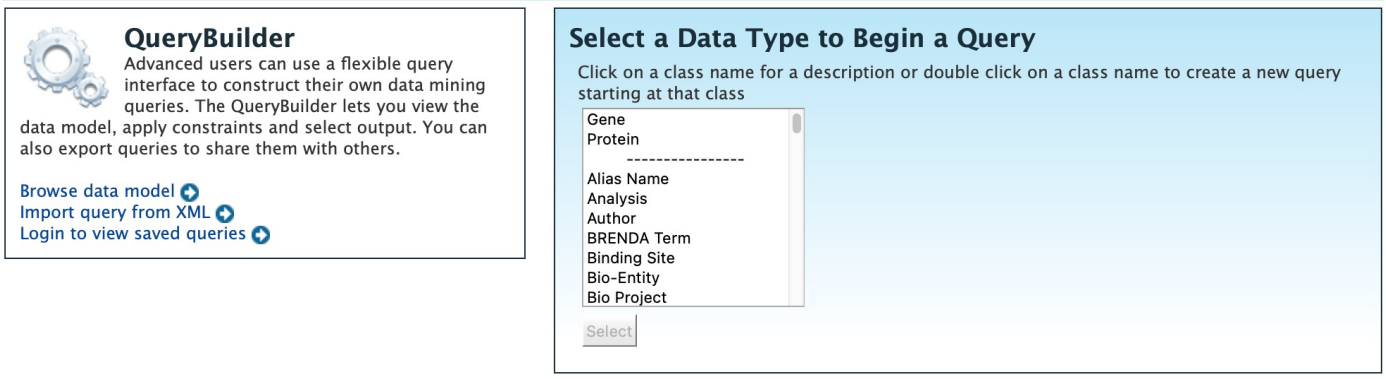


Fig. 12: Using the optional organism filter in the Gene –> Homologue query template.

2.3 QueryBuilder

The provided templates are suitable for many different types of searches, new queries may be built from scratch using the **QueryBuilder**. The possibilities of queries using the QueryBuilder are endless. The output may be formatted exactly as desired, and the query constraints may be chosen to perform complex search operations.



To begin, select a **Data Type**. For example, select **Gene** as a Data Type then click the **Select** button to be taken to the Model browser.

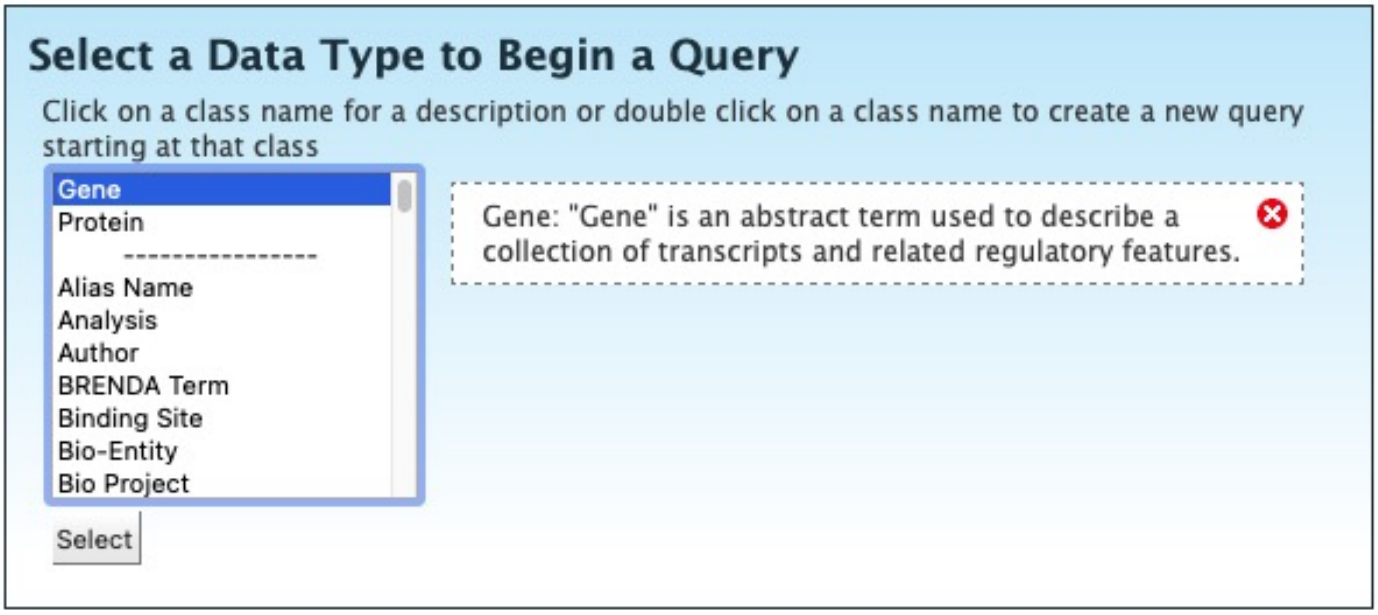


Fig. 13: Selecting data type as **Gene** in QueryBuilder.

2.3.1 Model browser

After selecting a data type, the **Model builder** appears displaying the attributes for the chosen feature class **Gene**.

First lets select `Gene` as a Data Type in the QueryBuilder. Then click on `Select`. This will take you to a Model browser where you can select the attributes for the feature class 'Gene', which you would want to be shown in your results.

2.3.2 QueryBuilder Examples

The following three examples provide details as to how to use the QueryBuilder using "Gene" as the selected data type.

2.3.2.1 Example 1: Querying for protein coding genes

In the Model browser, click **Show** next to **Biotype**, **Gene ID** and **Symbol**, which will add these fields to the query. Notice that these two fields appear below the data type **Gene** in the Query Overview section.

Then click **Constrain** next to **Biotype**. The first drop-down menu defaults to = (equals sign). In the second drop-down menu, select **Protein Coding**, then click the **Add to query** button. This adds a constraint to the query to search only for protein coding genes. Notice that the Query Overview section now shows "Biotype = Protein Coding". Also, two types of icons appear next to the attributes. Clicking on the red "X" icon next to an attribute will remove that field or constraint from the query. Clicking on the blue pencil icon next to a constraint brings up the constraint editing window where changes may be made to the query filters.

Lastly, click on **Show Results** above the Model Browser. The resulting table contains all protein coding genes in the database, with Gene ID, Gene Symbol and Gene Biotype as the table columns. Because the Biotypes should all be the same (protein coding), that column can be deleted by clicking the "x" above it. Alternatively, the three dots "…" can

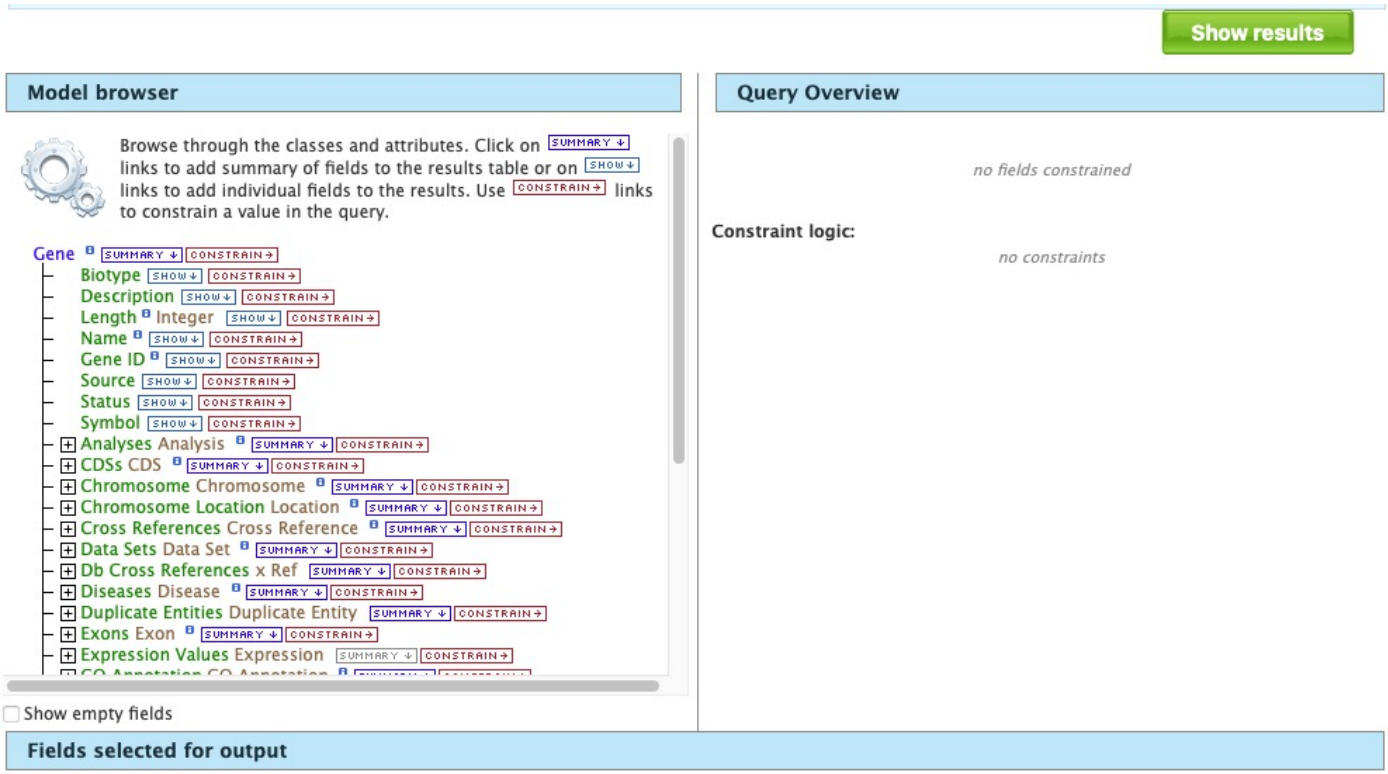


Fig. 14: Model browser with “Gene” selected as the data type.

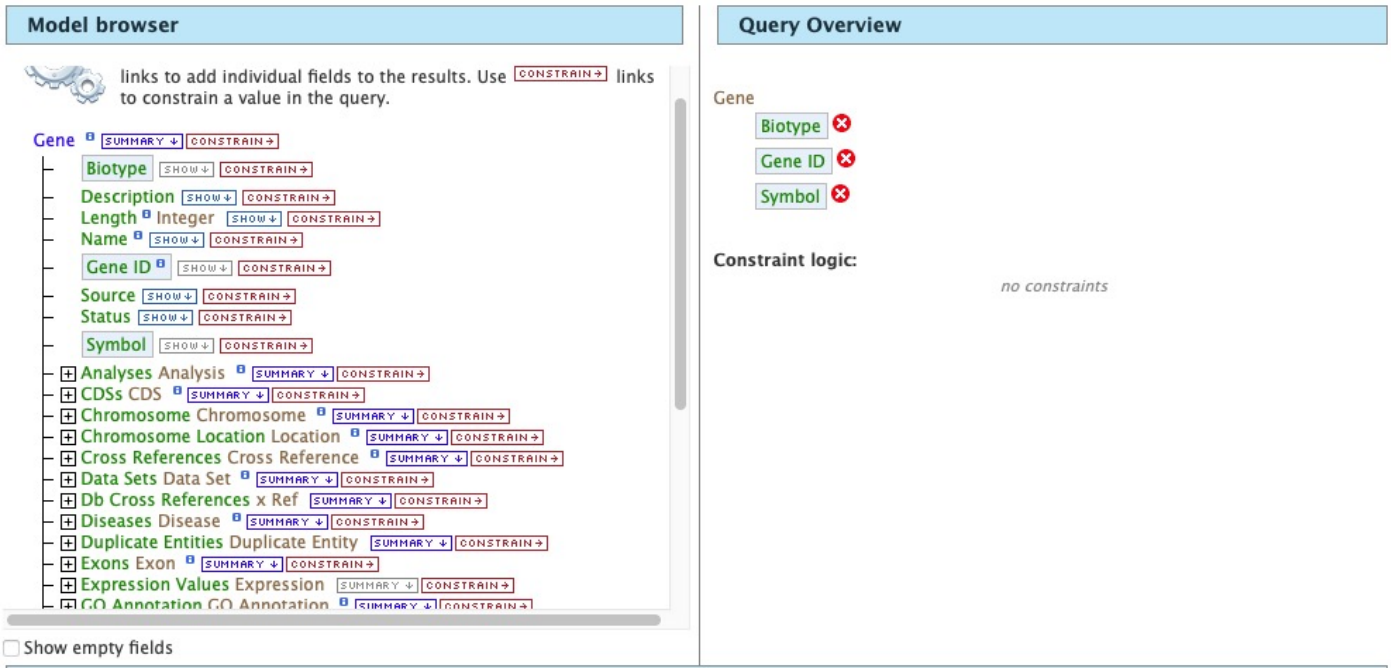


Fig. 15: Model browser with “Gene” selected as the data type.
Example 1, Step 1: Select fileds to be added to the query

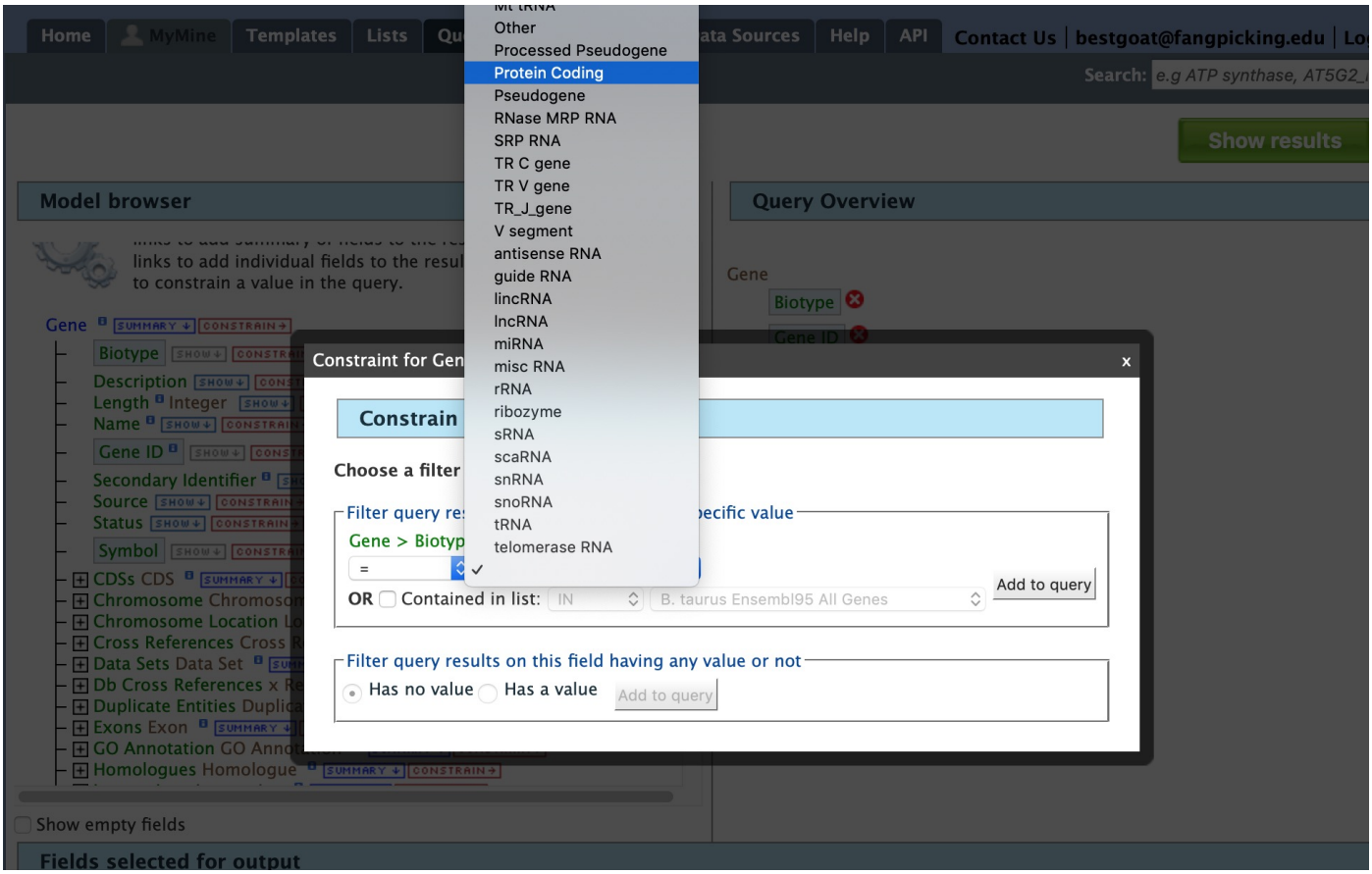
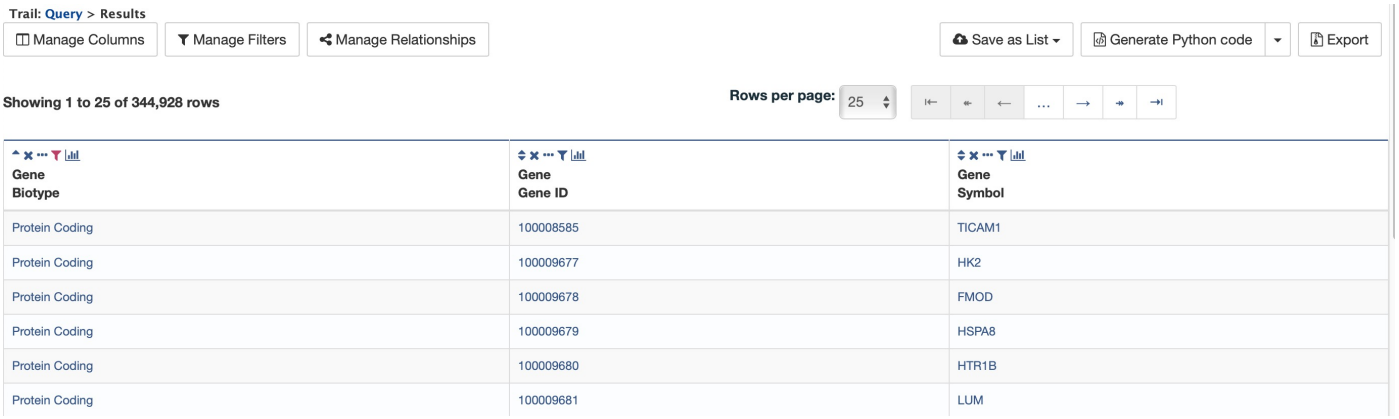


Fig. 16: Example 1, Step 2 Adding a constraint to the query on Biotype.

be selected to "Toggle column visibility" and hide the column. See the *Customize output and manage columns* section to review ways to manage column data.



Gene Biotype	Gene Gene ID	Gene Symbol
Protein Coding	100008585	TICAM1
Protein Coding	100009677	HK2
Protein Coding	100009678	FMOD
Protein Coding	100009679	HSPA8
Protein Coding	100009680	HTR1B
Protein Coding	100009681	LUM

Fig. 17: Example 1, Step 3 Display query results.

2.3.2.2 Example 2: Querying for protein coding genes on a particuler chromosome

Users can customize the previously run query by adding another constraint for **Chromosome**. Note the "Trail: Query" text at the upper left of the results table. Clicking on the "Query" link will bring you back to the query that generated the table to allow for edits without having to build a new query. In the Model browser, click on the "+" (plus sign) next to the Chromosome feature class to display its attributes.

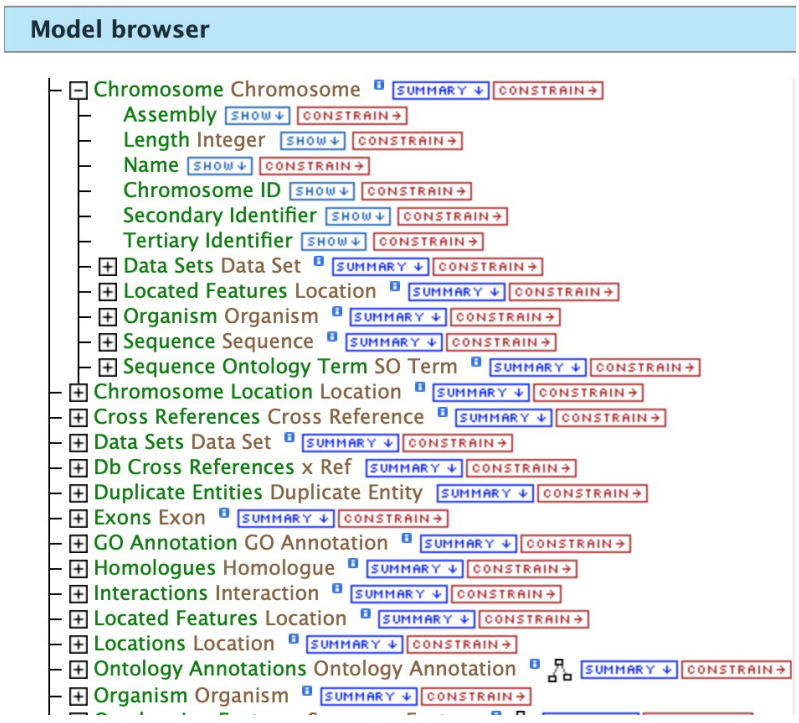


Fig. 18: Example 2, Step 1 View attributes of Chromosome feature class

Next click on **Constrain** next to the **Chromosome ID** attribute and in the text box of the pop-up window select "=" (equals). Enter **18** for chromosome number 18then click on **Add to Query**, which adds the additional constraint to the query. Just clicking chromosome 18 will bring up that chromosome in multiple organisms so we will add another organism constraint. Click on the **Organism** to reveal its attributes then click on **constrain** next to Genus. In the resulting pop-up window, slect "=" (equals) then "Equus" for the chromosome 18 of horse.

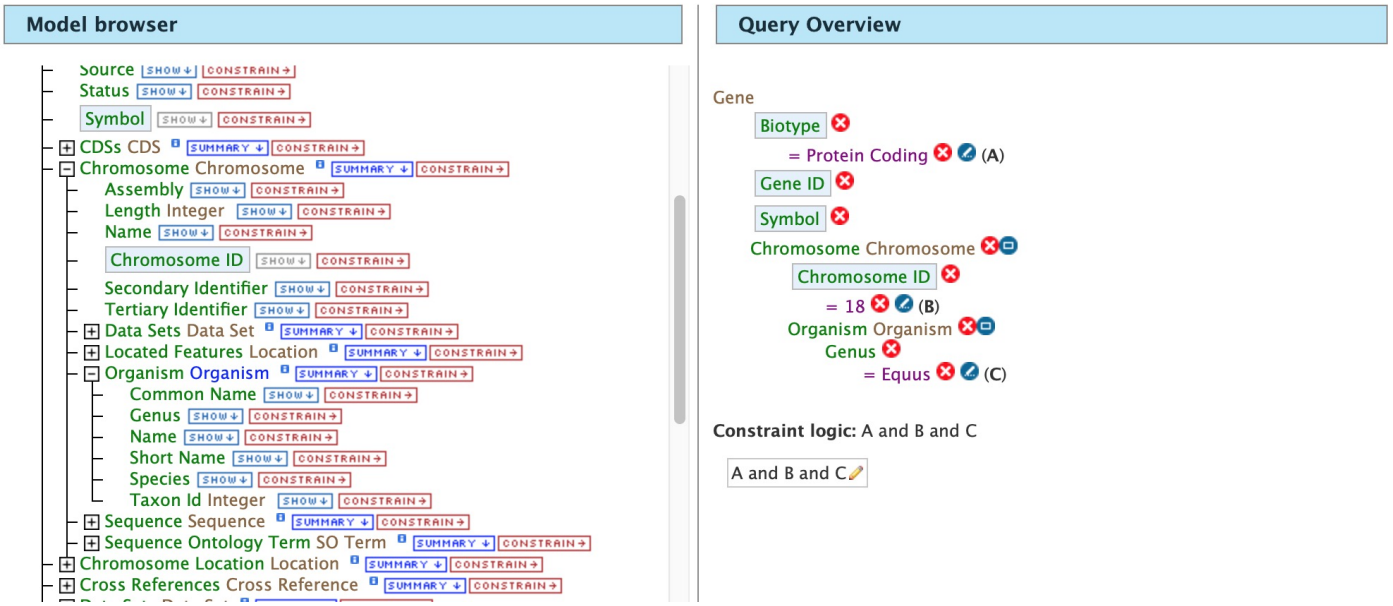
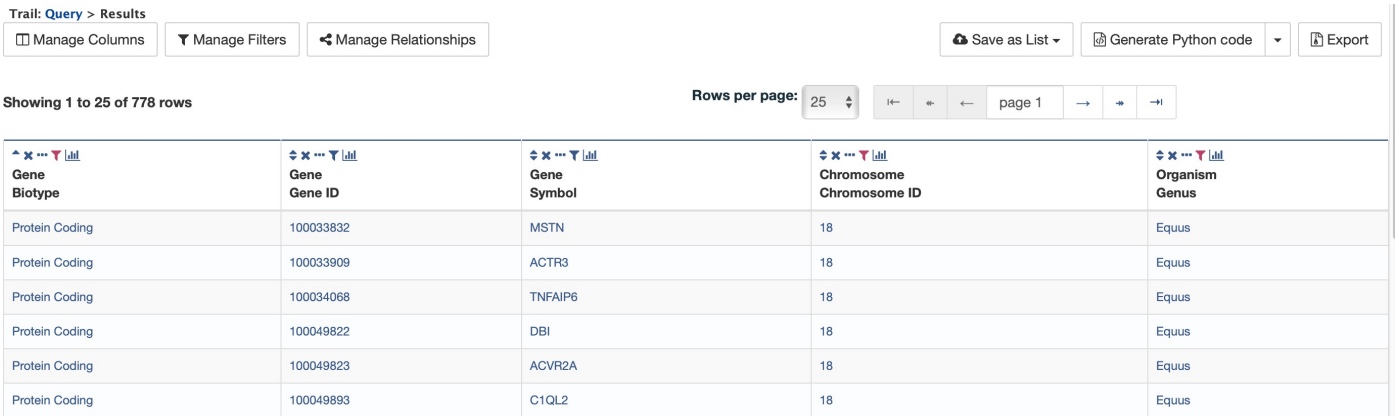


Fig. 19: Example 2, Step 2 Adding constraints to the Chromosome ID and Organism

Click on **Show results** and the query will result in all protein-coding genes on the Chromosome with constrained ID and organism. Note that the number of results has been reduced with the addition of the constraint.



Gene Biotype	Gene Gene ID	Gene Symbol	Chromosome Chromosome ID	Organism Genus
Protein Coding	100033832	MSTN	18	Equus
Protein Coding	100033909	ACTR3	18	Equus
Protein Coding	100034068	TNFAIP6	18	Equus
Protein Coding	100049822	DBI	18	Equus
Protein Coding	100049823	ACVR2A	18	Equus
Protein Coding	100049893	C1QL2	18	Equus

Fig. 20: Example 2, Step 3 Result table after constraining by Chromosome ID and Organism

2.3.2.3 Example 3: Querying for Protein Coding genes on a particular chromosome and their exons

This last example builds upon the previous queries to display all exons for each gene on a particular chromosome. Again, note the "Trail: Query" text at the upper left of the results table. Clicking on the "Query" link will bring you back to the query that generated the table to allow for edits without having to build a new query. Scroll down to **Exon** feature class and select the "+" (plus sign) next to Exons to display its attributes. Click **Show** next to **Length** and **Exon Identifier**.

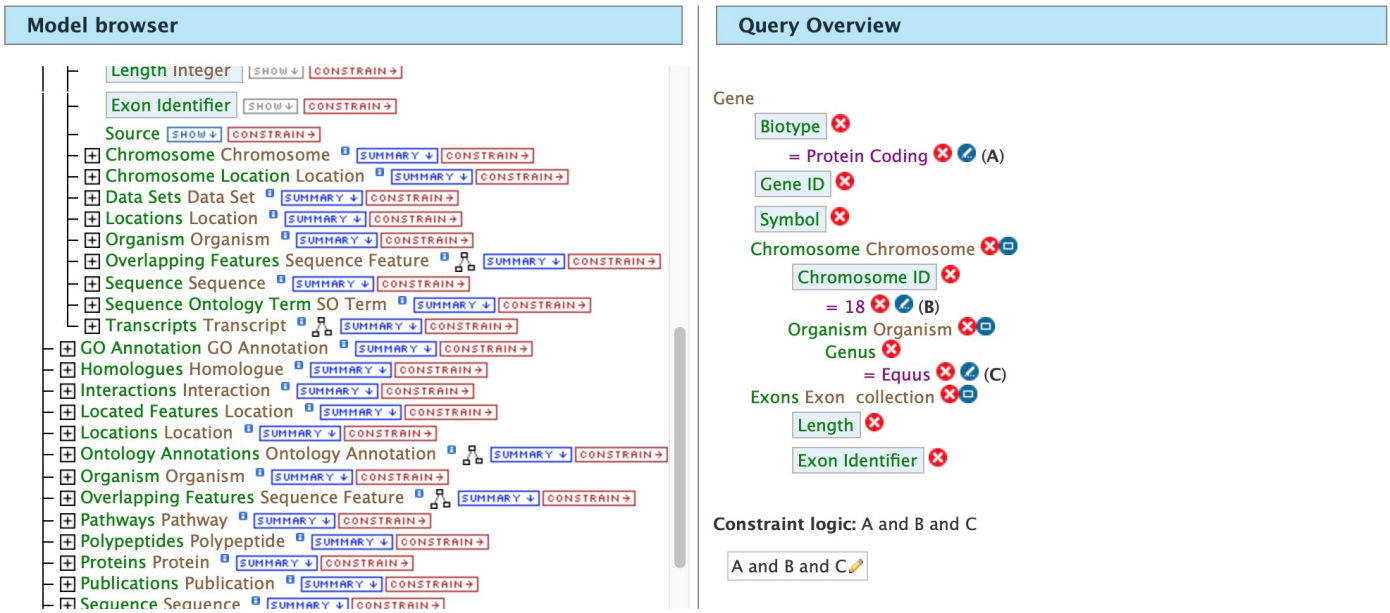


Fig. 21: Example 3, Step 1 Expand exon attributes and add fields to the query

The Query Overview shows the query in progress with the selected fields. Also notice that a third type of icon, a blue square, appears next to some attributes. Clicking on a blue square icon brings up a window where the query Join Style may be modified. When adding a constraint, you can decide whether you want to show only those results with the information (genes with exons) or all results and the constrained feature if it exists (e.g., all genes and indicate exons if they exist). Click on the blue square icon next to **Exon collection** to bring up the Switch Join Style window. The default option is to show only Genes if they have a exon (inner join). Change this to **Show all Genes and show Exons if they are present** (outer join) then click **Add to query**.

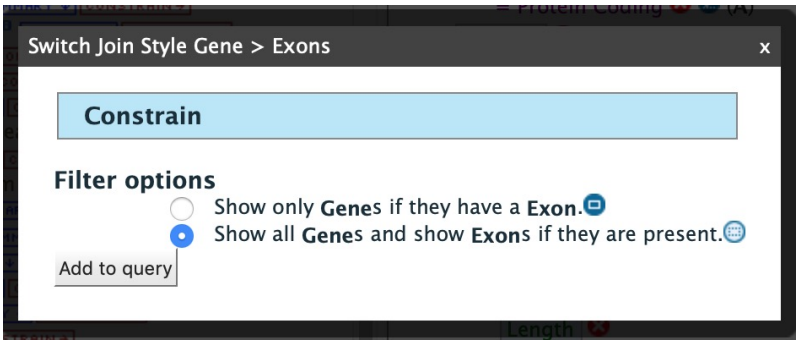
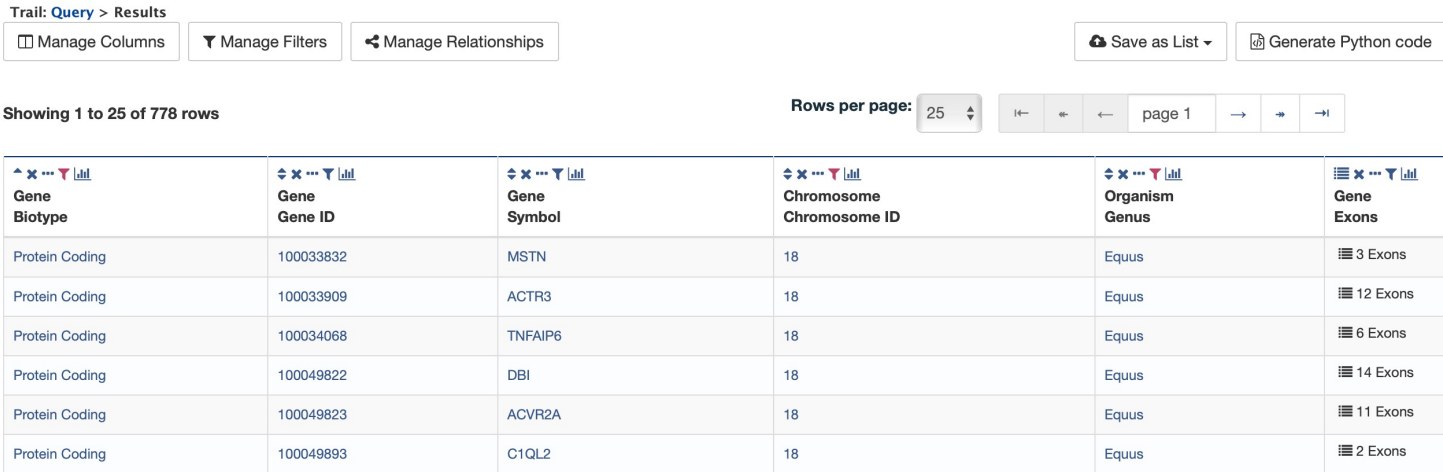


Fig. 22: Example 3, Step 2 Select join style for exons

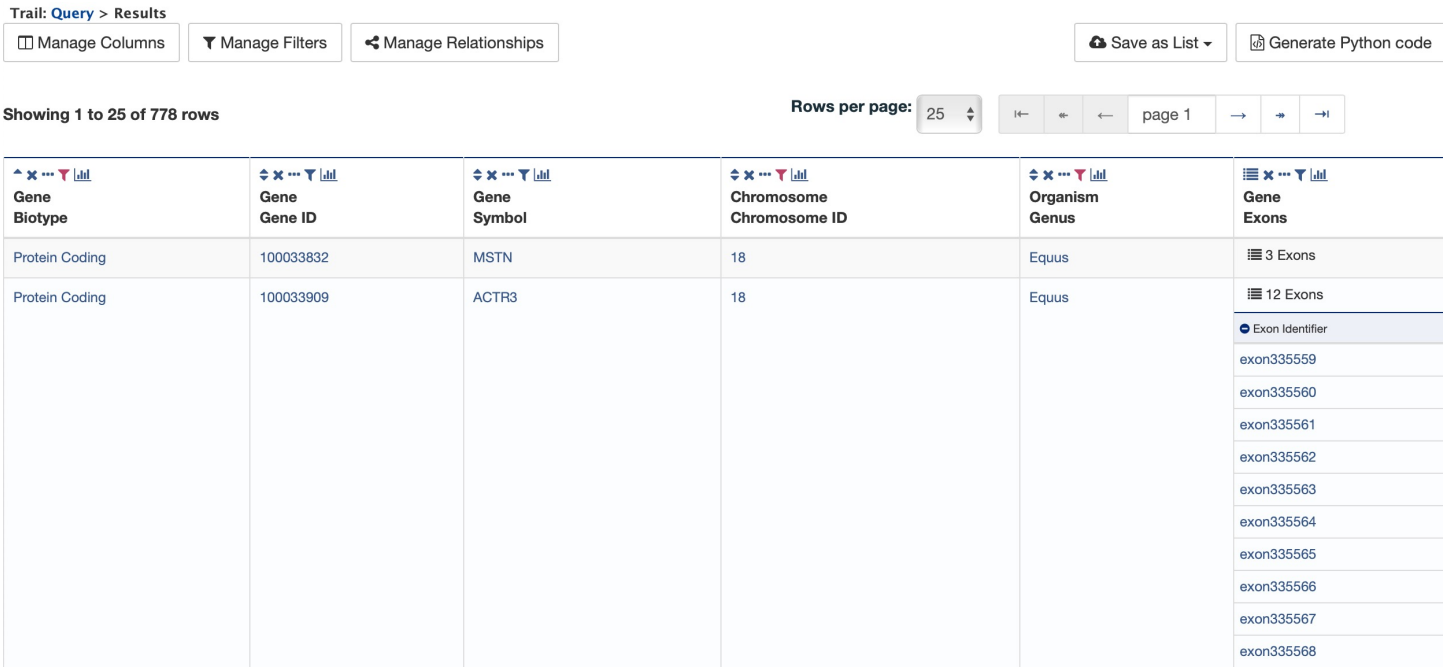
Then click "Show results" to run the new query.



Gene Biotype	Gene Gene ID	Gene Symbol	Chromosome Chromosome ID	Organism Genus	Gene Exons
Protein Coding	100033832	MSTN	18	Equus	3 Exons
Protein Coding	100033909	ACTR3	18	Equus	12 Exons
Protein Coding	100034068	TNFAIP6	18	Equus	6 Exons
Protein Coding	100049822	DBI	18	Equus	14 Exons
Protein Coding	100049823	ACVR2A	18	Equus	11 Exons
Protein Coding	100049893	C1QL2	18	Equus	2 Exons

Fig. 23: Example 3, Step 3 Query results with exon constraint

The results table now lists a new column **Gene Exons**, which we added to the query. If we look at the second row in the table, it lists 12 exons. Click on the **12 exons** link to expand that entry. That column now has additional rows containing the **Exon identifier** and **Length** for each of the 12 exons.



Gene Biotype	Gene Gene ID	Gene Symbol	Chromosome Chromosome ID	Organism Genus	Gene Exons
Protein Coding	100033832	MSTN	18	Equus	3 Exons
Protein Coding	100033909	ACTR3	18	Equus	12 Exons
					Exon Identifier
					exon335559
					exon335560
					exon335561
					exon335562
					exon335563
					exon335564
					exon335565
					exon335566
					exon335567
					exon335568

Fig. 24: Example 3, Step 4 Query results with exon column expanded

In changing the join style to an outer join, the exons have been grouped together by gene making it easier to determine how many exons are contained in each gene. If the same query is run with the default join (outer join) of **Show only Genes if they have an Exon**, the results table adds a new row for each new exon rather than grouping exons by gene.

Trail: Query > Results

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

Showing rows 1 to 25 of 51,208 — Rows per page: 25

Gene Biotype	Gene Gene ID	Gene Symbol	Chromosome Chromosome ID	Exons Length	Exons Exon Identifier
Protein Coding	100033832	MSTN	18	373	exon370698
Protein Coding	100033832	MSTN	18	374	exon370699
Protein Coding	100033832	MSTN	18	381	exon370700
Protein Coding	100033909	ACTR3	18	56	exon335560
Protein Coding	100033909	ACTR3	18	84	exon335569
Protein Coding	100033909	ACTR3	18	93	exon335567
Protein Coding	100033909	ACTR3	18	96	exon335563
Protein Coding	100033909	ACTR3	18	108	exon335564
Protein Coding	100033909	ACTR3	18	111	exon335562
Protein Coding	100033909	ACTR3	18	125	exon335561

Fig. 25: Example 3, Step 5 Query results with default join style run for genes containing exons

CHAPTER 3

Report Pages

All objects in FAANGMine (e.g., gene, protein, transcript, publication) have report pages that can viewed after running a query. It allows users to view all available information for that object while providing links to related objects. As an example, we can revisit the templates example. In the list of templates under the **Templates** tab on the FAANGMine home page, select Gene -> Homologues to query FAANGMine to retrieve all homologues for a given gene. Enter "GSTM1" into the the LOOKUP search box then click **Show Results**. In the results table, note that every entry is contains a link. You can mouse over any link to bring up a summary of that object. If we hover over the first Gene ID, we can see a summary box that includes

Clicking on that same item will bring up its report page that includes a comprehensive for gene GSTM1. The report page header shows the Gene ID and its Biotype, for this example, protein coding. The tabs at the top of the page in the Quick Links menu bar quickly bring you to the data listed. The column on the right side of the report page displays external links to other Mines and databases.

The content of the report page is divided into categories based on the type of information provided for that particular object. Clicking on links within each category bring up more details about the objects of interest.

3.1 Summary

The **Summary** section near the top of the report provides information on the gene such as its length, chromosome location, and strand information. Users can also get the complete FASTA sequence of the gene by clicking on the FASTA tab.

3.2 Transcripts

The **Transcripts** section contains information about the gene model, such as transcripts and exons. Links to FASTA files are included where applicable.

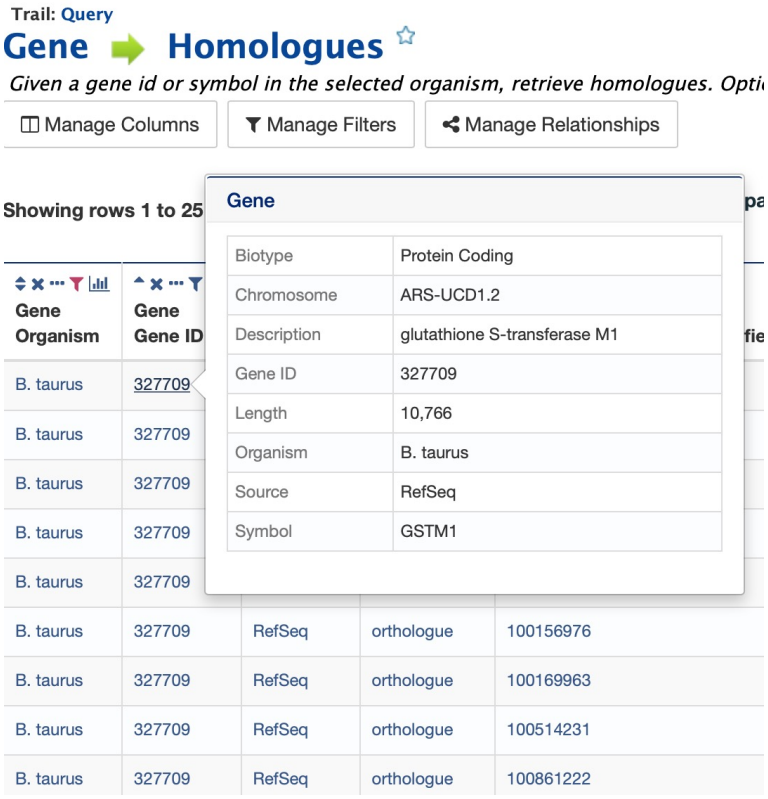


Fig. 1: Summary for gene entry in query results table

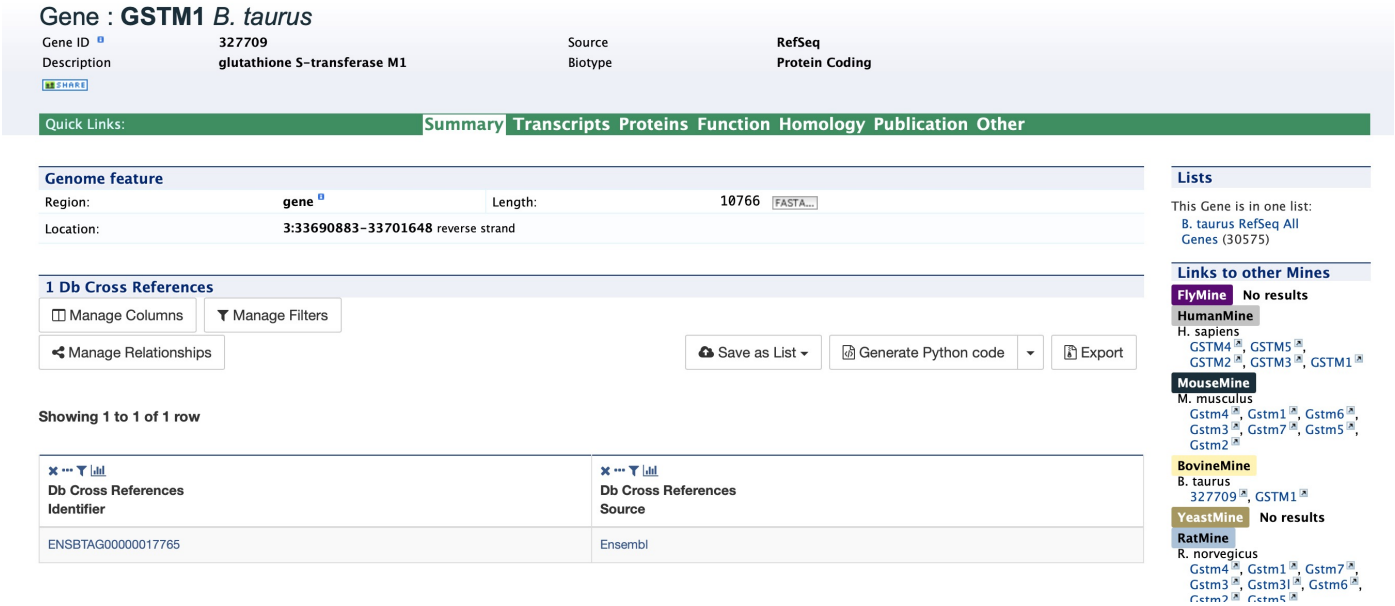
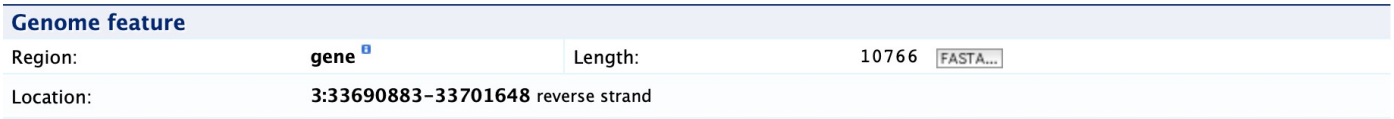


Fig. 2: Report page for protein-coding gene



Transcripts

All Transcripts for Gene – GSTM1 327709
Transcripts: 1 Exons: 8 Coding Sequence: 1

Transcript	Exons		Coding Sequence
GSTM1 NM_175825.3 1141 FASTA...	exon632091	100 FASTA...	NM_175825.3-CDS 657 FASTA...
	exon632092	76 FASTA...	
	exon632093	65 FASTA...	
	exon632094	82 FASTA...	
	exon632095	101 FASTA...	
	exon632096	96 FASTA...	
	exon632097	111 FASTA...	
	exon632098	510 FASTA...	

3.3 Proteins

The **Proteins** section provides information about the protein product of the gene. The comments section gives a brief description about the protein along with the UniProt accession and links to any outside data sets.

Proteins

Curated comments from UniProt Show proteins

Type	Comment
function	Conjugation of reduced glutathione to a wide number of exogenous and endogenous hydrophobic electrophiles. Protects against the thiol-mediated metal-catalysed oxidative inactivation of enzymes.
similarity	Belongs to the GST superfamily. Mu family.

1 Proteins

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

Showing rows 1 to 1 of 1

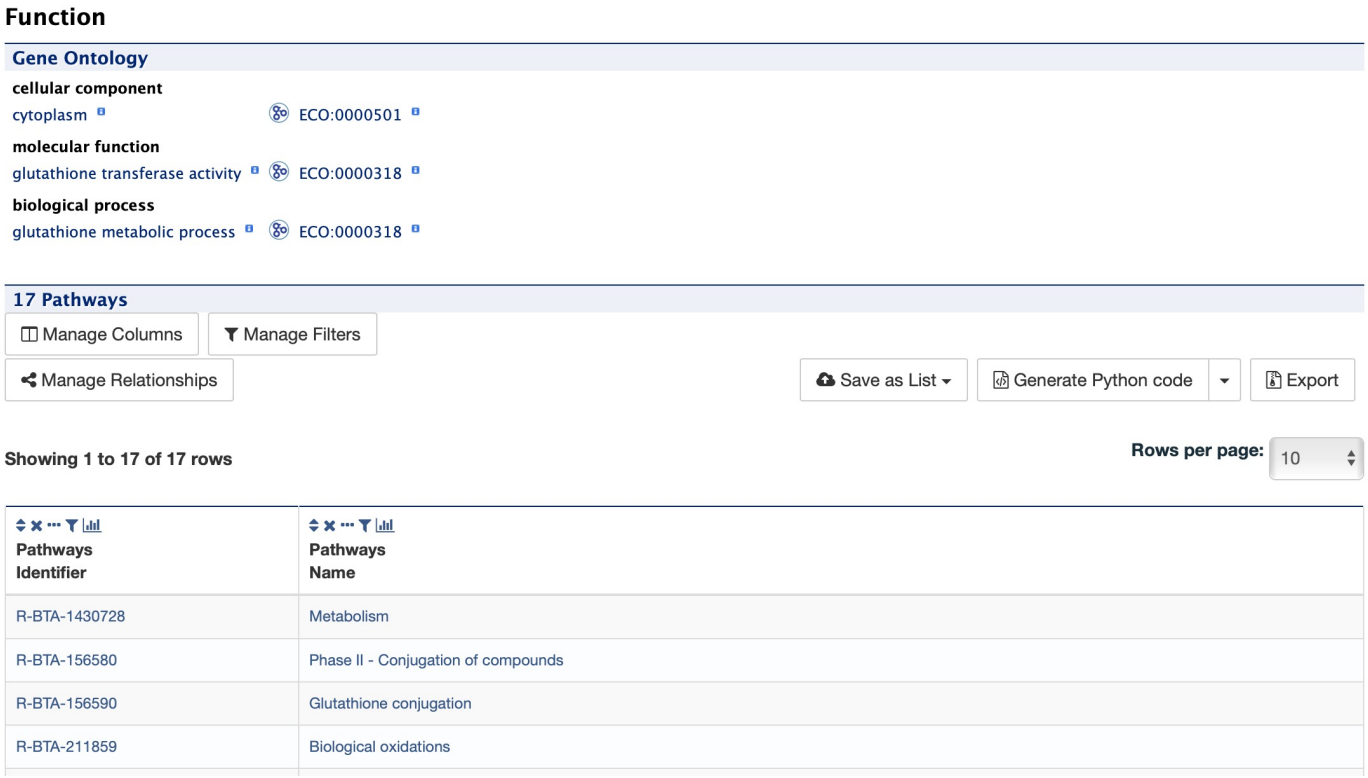
Proteins DB identifier	Proteins Primary Accession	Proteins Organism . Name	Proteins Length	Proteins datasets
GSTM1_BOVIN	Q9N0V4	Bos taurus	218	3 Data Set datasets

3.4 Function

The **Function** section displays Gene Ontology annotations for a gene. Annotations are divided into three categories:

- Cellular Component
- Molecular Function
- Biological Process

The GO terms are displayed along with the evidence code indicating how the annotations were derived. A results with Pathway information is also displayed if applicable.



3.5 Homology

The **Homology** section provides information for all homologues. The first portion displays a summarized view of the homologues reported in different organisms. The next portion provides more detailed information about the homologue, the type of homologue and from which dataset the information was obtained all displayed in a results table.

3.6 Interactions

The **Interactions** section provides interaction information. For GSTM1 there are no interaction information available but for genes that do have interaction information, a network is displayed showing all interactors for the current gene.

3.7 Publications

The **Publications** section displays a table of publications related to the gene with links to full citations.

3.8 Other

This last section provides miscellaneous information that do not fit into any of the above categories. This example lists protein coding annotations and their sources.

Homology

Homologues

B. bubalis	C. hircus	C. lupus familiaris	E. caballus	F. catus	G. gallus	H. sapiens	M. musculus	O. aries	R. norvegicus	S. scrofa
LOC102397004 LOC102398085 GSTM3 LOC102396303	LOC102185621 LOC106503993 LOC102189813 LOC100861222 LOC108633298 LOC102190481 GSTM3	LOC479912 LOC479911	LOC100058329 LOC100061761 GSTM3 LOC100058290	LOC101100824 GSTM3	GSTM2 GST2L	GSTM2 GSTM3 GSTM4 GSTM1 GSTM5	Gstm1 Gstm5 Gstm4 Gstm3 Gstm7 Gstm2 Gstm6	LOC101107401 LOC101108705 LOC101108092 LOC101107831 GSTM3	Gstm4 Gstm6 Gstm2 Gstm3l Gstm7 Gstm3 Gstm1 Gstm5	GSTM3 LOC780435 LOC106510200 LOC110260348 LOC110260350 LOC110260351 LOC100156976

Gene --> Homologues for report page (135 rows)

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

Showing 1 to 25 of 135 rows

Rows per page: 25 | page 1

Data Sets Name	Homologues Type	Homologues Last Common Ancestor	Homologues Homologue . Primary Identifier	Homologue Symbol	Homologues Homologue . Organism . Short Name
OrthoDB data set	orthologue	Cetartiodactyla	100156976	LOC100156976	S. scrofa
OrthoDB data set	orthologue	Cetartiodactyla	100861222	LOC100861222	C. hircus
OrthoDB data set	orthologue	Cetartiodactyla	101107831	LOC101107831	O. aries

Interactions

Gene --> Interacting genes (0 rows)

Publication

Publications (3 rows)

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

Showing rows 1 to 3 of 3

Publications Year	Publications First Author	Publications Title	Publications Journal	Publications Volume	Publications Pages	Publications PubMed ID
2015	Hering D M	Missense mutation in glutathione-S-transferase M1 gene is associated with sperm motility and ATP content in frozen-thawed semen of Holstein-Friesian bulls.	Anim. Reprod. Sci.	159	94-7	26091956

Other

6 Data Sets

Swiss-Prot data set, OrthoDB data set, PubMed to gene mapping, Gene RNASeq Expression data, KEGG pathways data set, Bos taurus RefSeq Protein Coding Genes

CHAPTER 4

Genomic Regions Search

The **Genomic Regions Search** is a tool to fetch features that are within a given set of genomic coordinates or are within a given number of bases flanking the coordinates.

To begin this type of search, click the **Regions** tab on the menu bar. A form will appear asking for the search parameters (organism, feature types, genomic coordinates, etc.)

The coordinates must have one of three formats:

1. chromosome_number:start..end
2. chromosome_number:start-end
3. chromosome_number start end (tab delimited)

Click on the input examples above the text input box (number 4) to view a representative set of coordinates in each format. Click the **Genome coordinates help** link near the top of the form for more detailed information on the input format requirements.

During a search, regions may be extended on either side of the genomic coordinates using the slider or by entering text in the field to the left of the slide bar. There is also the option to perform a strand-specific region search using the checkbox at the bottom of the form (number 6)

As an example, select *B. taurus* from the Select Organism drop-down, and ARS-UCD1.2 as the Assembly. Slick the box next to Select Feature Types to uncheck all of the boxes, then check the box next to Gene, and enter the following coordinates into the genomic regions search text field:

```
14:2000000..2800000
```

Click the search box to conduct the genomic regions search. If there are no overlaps within your search coordinates, the search can be done again with the search region extended using the slide bar or entering text into the search box (e.g., 10k).

The search results page presents a list of features present within the genomic interval that was searched. In this case, the feature type was limited to Gene. The results may be exported as tab-separated or comma-separated values. If they

Fig. 1: Genomic Regions search form

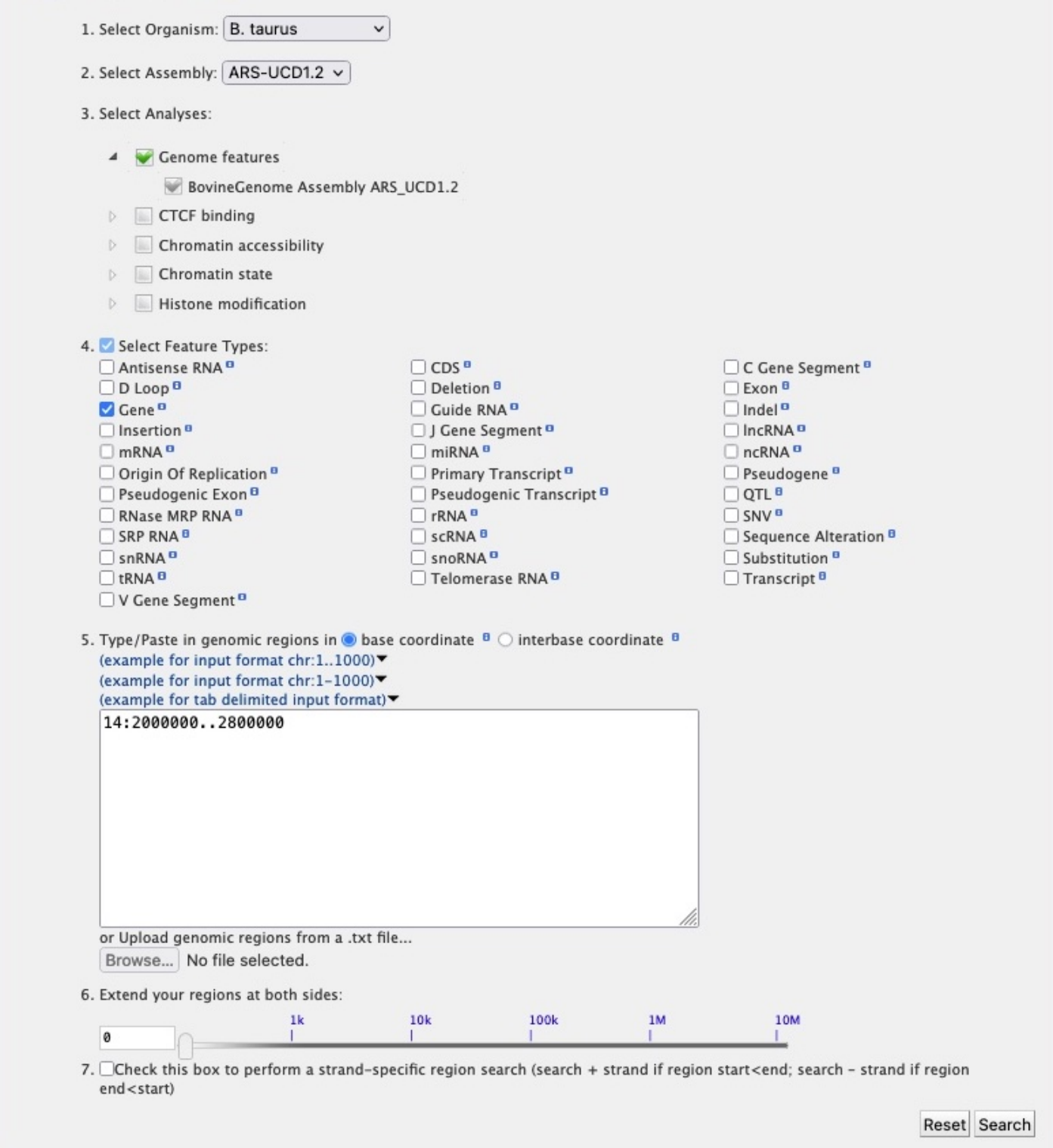
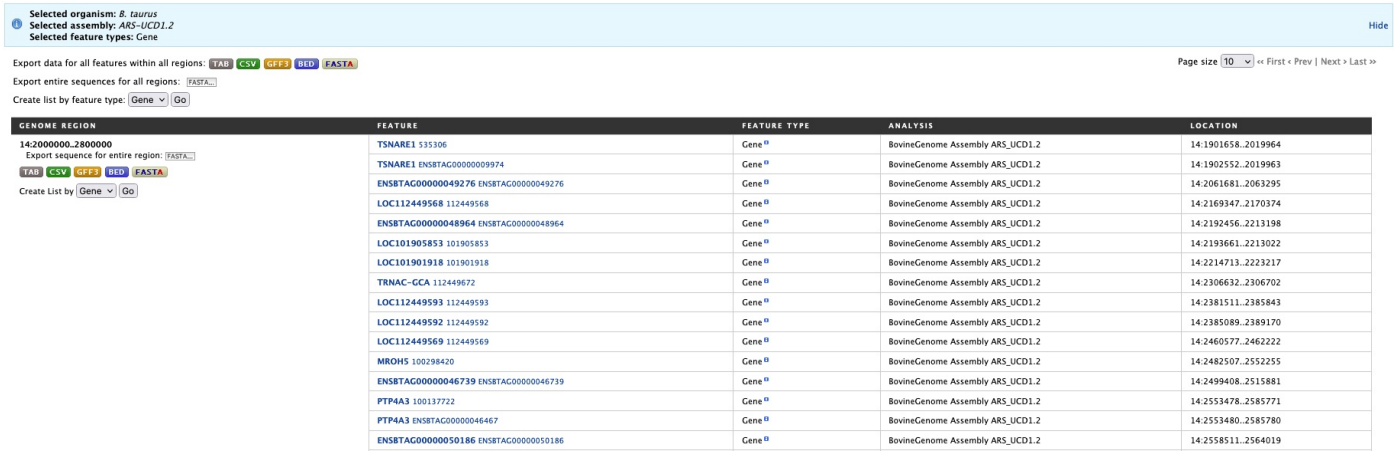


Fig. 2: Genomic Regions search example with *Bos taurus*

contain genomic features, there is also the option to saved the results in GFF3 or BED format. The FASTA sequences of the features may also be downloaded. Links within the features provide detailed reports. If users are interested in creating a list of particular features from the result page then they can filter based on feature type (if applicable), shown in red box, and click on **Go**.



Selected organism: *B. taurus*
Selected assembly: *ARS-UCD1.2*
Selected feature types: Gene

Hide

Export data for all features within all regions: TAB CSV GFF3 BED FASTA

Export entire sequences for all regions: FASTA...

Create list by feature type: Gene Go

Page size 10 « First ‹ Prev | Next › Last »

GENOME REGION	FEATURE	FEATURE TYPE	ANALYSIS	LOCATION
14:2000000..2800000 Export sequence for entire region: FASTA... TAB CSV GFF3 BED FASTA Create List by Gene Go	TSNARE1 535306	Gene	BovineGenome Assembly ARS_UCD1.2	14:1901658..2019964
	TSNARE1 ENSBTAG00000009974	Gene	BovineGenome Assembly ARS_UCD1.2	14:1902552..2019963
	ENSBTAG00000049276 ENSBTAG00000049276	Gene	BovineGenome Assembly ARS_UCD1.2	14:2061681..2063295
	LOC112449568 112449568	Gene	BovineGenome Assembly ARS_UCD1.2	14:2169347..2170374
	ENSBTAG00000048964 ENSBTAG00000048964	Gene	BovineGenome Assembly ARS_UCD1.2	14:2192456..2213198
	LOC101905853 101905853	Gene	BovineGenome Assembly ARS_UCD1.2	14:2193661..2213022
	LOC101901918 101901918	Gene	BovineGenome Assembly ARS_UCD1.2	14:2214713..2223217
	TRNAC-GCA 112449672	Gene	BovineGenome Assembly ARS_UCD1.2	14:2306632..2306702
	LOC112449593 112449593	Gene	BovineGenome Assembly ARS_UCD1.2	14:2381511..2385843
	LOC112449592 112449592	Gene	BovineGenome Assembly ARS_UCD1.2	14:2385089..2389170
	LOC112449569 112449569	Gene	BovineGenome Assembly ARS_UCD1.2	14:2460577..2462222
	MROH5 100298420	Gene	BovineGenome Assembly ARS_UCD1.2	14:2482507..2552255
	ENSBTAG00000046739 ENSBTAG00000046739	Gene	BovineGenome Assembly ARS_UCD1.2	14:2499408..2515881
	PTP4A3 100137722	Gene	BovineGenome Assembly ARS_UCD1.2	14:2553478..2585771
	PTP4A3 ENSBTAG00000046467	Gene	BovineGenome Assembly ARS_UCD1.2	14:2553480..2585780
	ENSBTAG00000050186 ENSBTAG00000050186	Gene	BovineGenome Assembly ARS_UCD1.2	14:2558511..2564019

Fig. 3: Genomic Regions search results

CHAPTER 5

Lists

5.1 Creating Lists

Users may create and save lists of features, such as gene IDs, transcript IDs, gene symbols, etc. The list tool searches the database for the list items and attempts to convert each identifier to the selected type. Click on the Lists tab from the menu to access the full list upload form. A short version of the form is also in the Quick List box on the home page.

As an example, enter the following comma-separated identifiers into the Lists upload form under the **Lists** tab. Notice that they do not have to be in the same format. A Summary table is displayed with the results of searching for each of the five identifiers in the list.

```
CAPN2, ENSCHIG00000014802, BTG1, XDH, 101107826
```

Leave the **Select Type** drop-down menu to **Gene** and the **Organism** drop-down to **Any**. Click on **Create List**. Note that you can also upload a list from a .txt file.

The summary table provides information regarding those identifiers that had a direct hit without any duplicates. If there are any duplicates, users can decide to add the relevant entries individually by clicking on the **Add** button under the **Action** column or choosing the **Add all** tab. Here we will click **Add all**. Once the selections have been added, the list can be saved by clicking the **Save a list of 66 Genes** button on the top of the summary table. Name the list by entering text into the **Choose a name for the list** box at the top of the results page.

After the list is saved, users are presented with a **List Analysis** page. This page provides users with widgets to perform analyses on gene lists that they have created.

The selection of widgets provided on the List Analysis page depend on the contents of the list. The available widgets for this list example include:

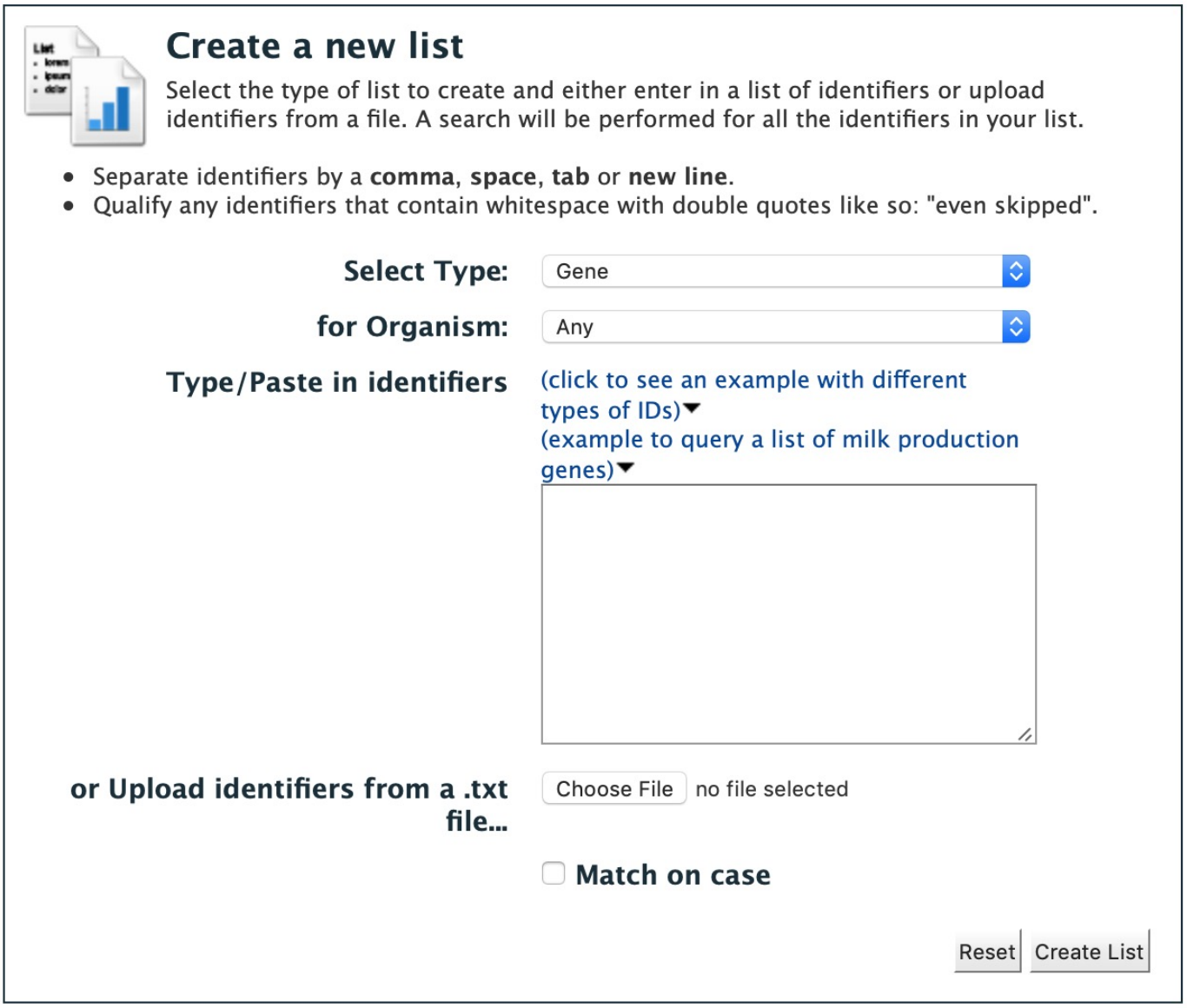


Fig. 1: List upload form

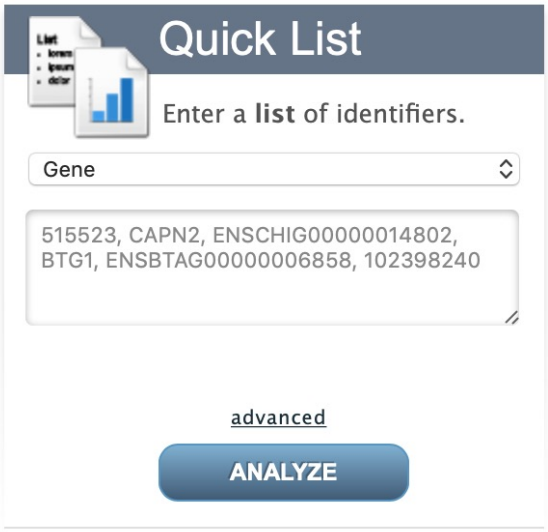


Fig. 2: Quick list from FAANGMine home page

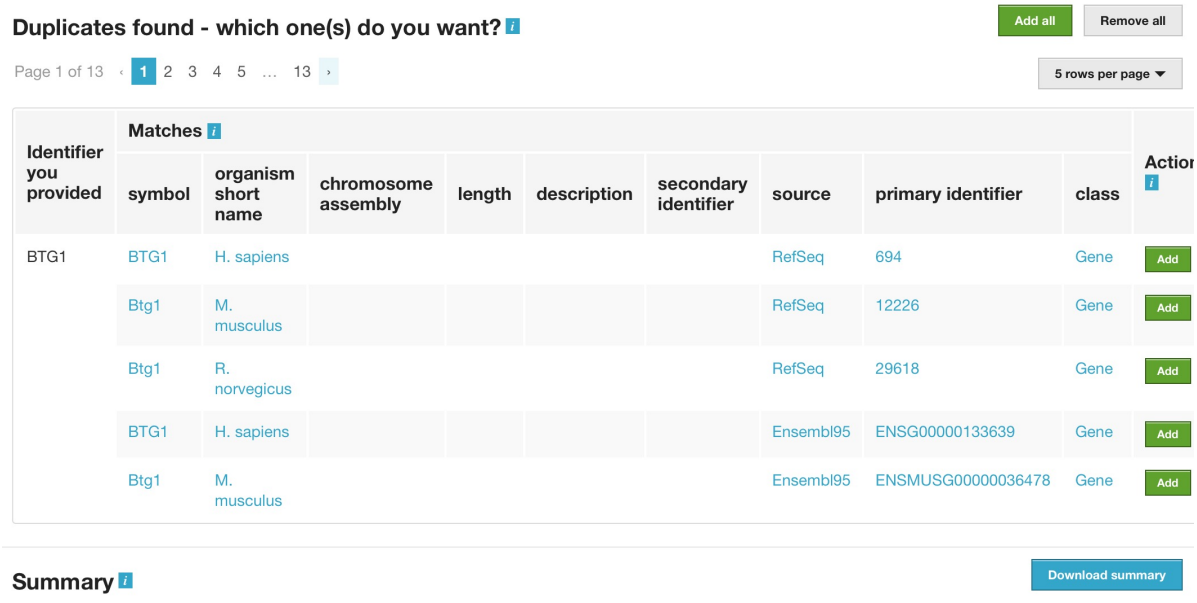


Fig. 3: List Example: Search results for list of identifiers

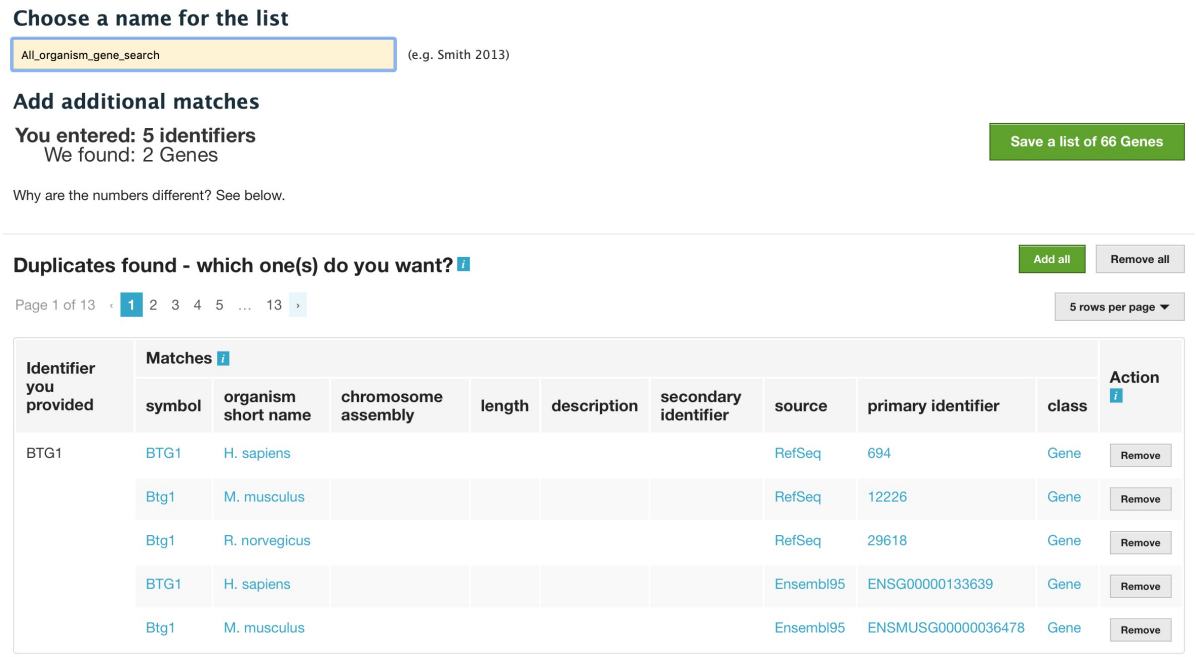
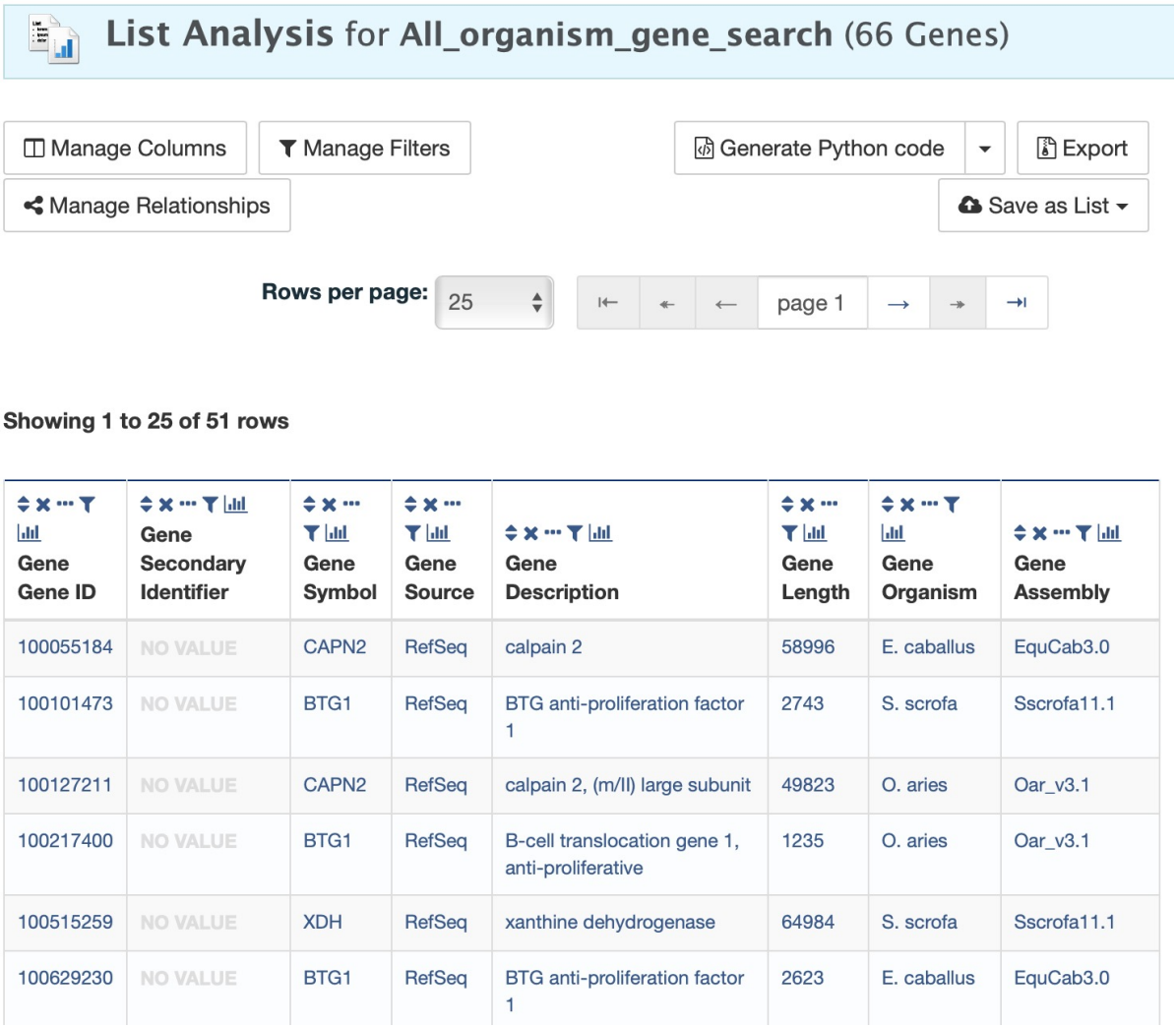


Fig. 4: List Example: Saving list of identifiers

1. Gene Ontology Enrichment
2. Publication Enrichment
3. Pathway Enrichment
4. Orthologues

5.2 Saving Lists

To see your saved lists, click the **View** tab on the **Lists** page. If not logged in, lists will be saved temporarily during your current session. However, you must be logged in to save your lists permanently. Further analyses of lists can be done with the **Actions** links at the top of the list. The links become active once lists are selected for analyses. Saved lists may also be accessed from the **MyMine** menu tab.



Gene Gene ID	Gene Secondary Identifier	Gene Symbol	Gene Source	Gene Description	Gene Length	Gene Organism	Gene Assembly
100055184	NO VALUE	CAPN2	RefSeq	calpain 2	58996	E. caballus	EquCab3.0
100101473	NO VALUE	BTG1	RefSeq	BTG anti-proliferation factor 1	2743	S. scrofa	Sscrofa11.1
100127211	NO VALUE	CAPN2	RefSeq	calpain 2, (m/II) large subunit	49823	O. aries	Oar_v3.1
100217400	NO VALUE	BTG1	RefSeq	B-cell translocation gene 1, anti-proliferative	1235	O. aries	Oar_v3.1
100515259	NO VALUE	XDH	RefSeq	xanthine dehydrogenase	64984	S. scrofa	Sscrofa11.1
100629230	NO VALUE	BTG1	RefSeq	BTG anti-proliferation factor 1	2623	E. caballus	EquCab3.0

Fig. 5: List Example: Analysis for gene list

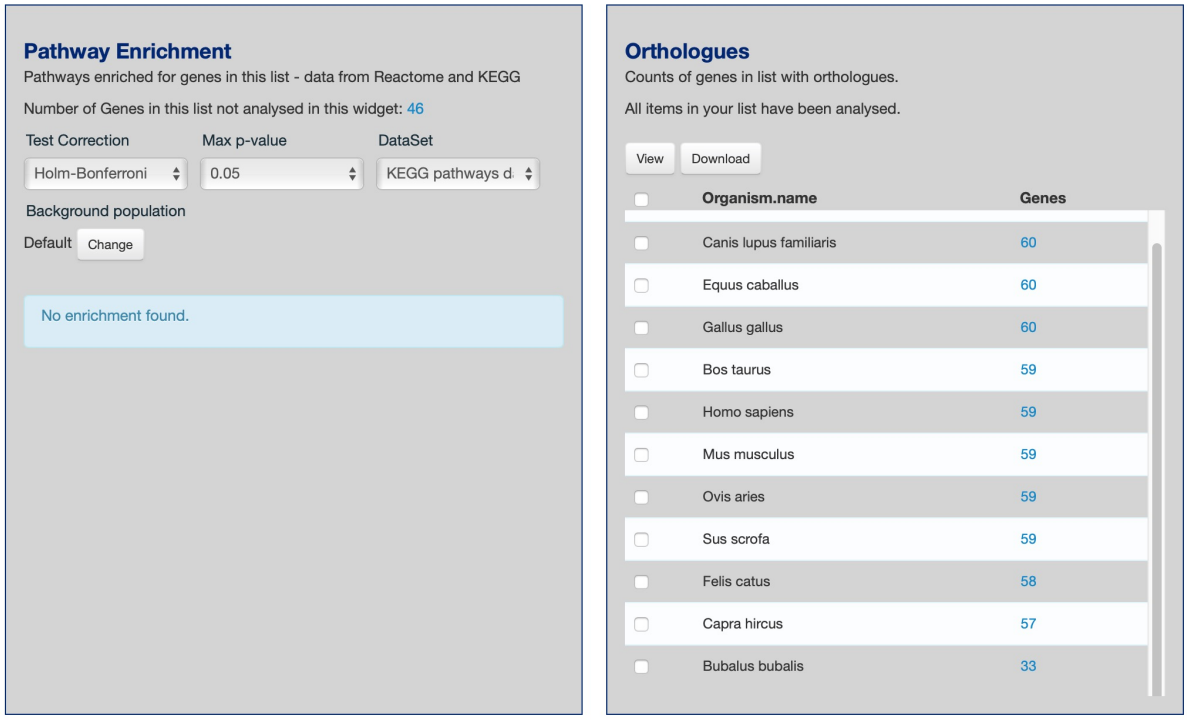


Fig. 6: List Example: Displayed widgets for list analysis

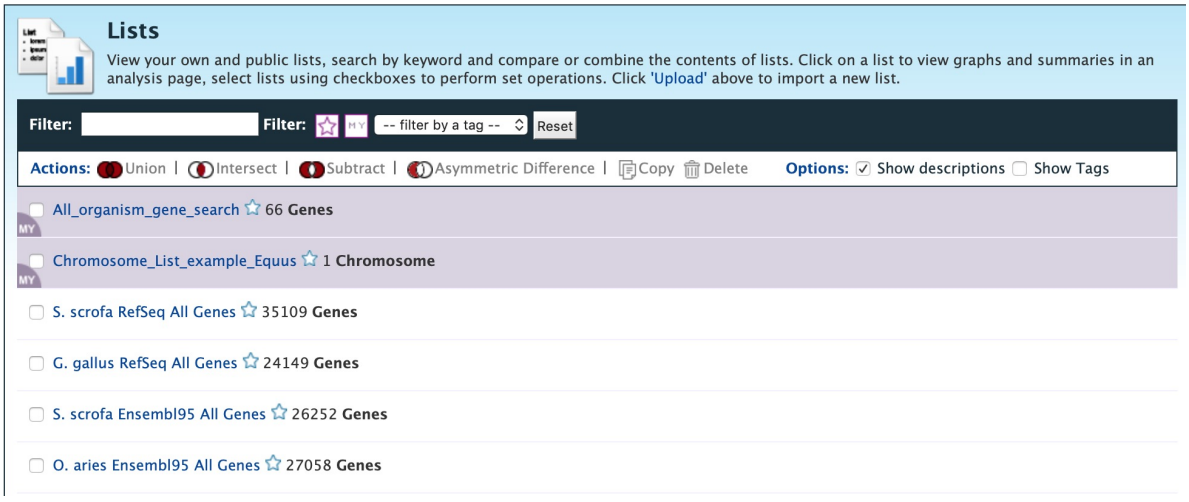


Fig. 7: List Example: Saved user lists

CHAPTER 6

MyMine

MyMine serves as a portal where logged-in users may manage their lists, queries, templates, and account details.

To access MyMine, click on the MyMine menu tab. A submenu appears with six options:

Lists - Lists saved by the user when logged in.

History - List of most recently run queries.

Queries - List of saved queries.

Templates - Templates created or marked as “favorite” by the user.

Password - Password reset form.

Account Details - User preferences form.

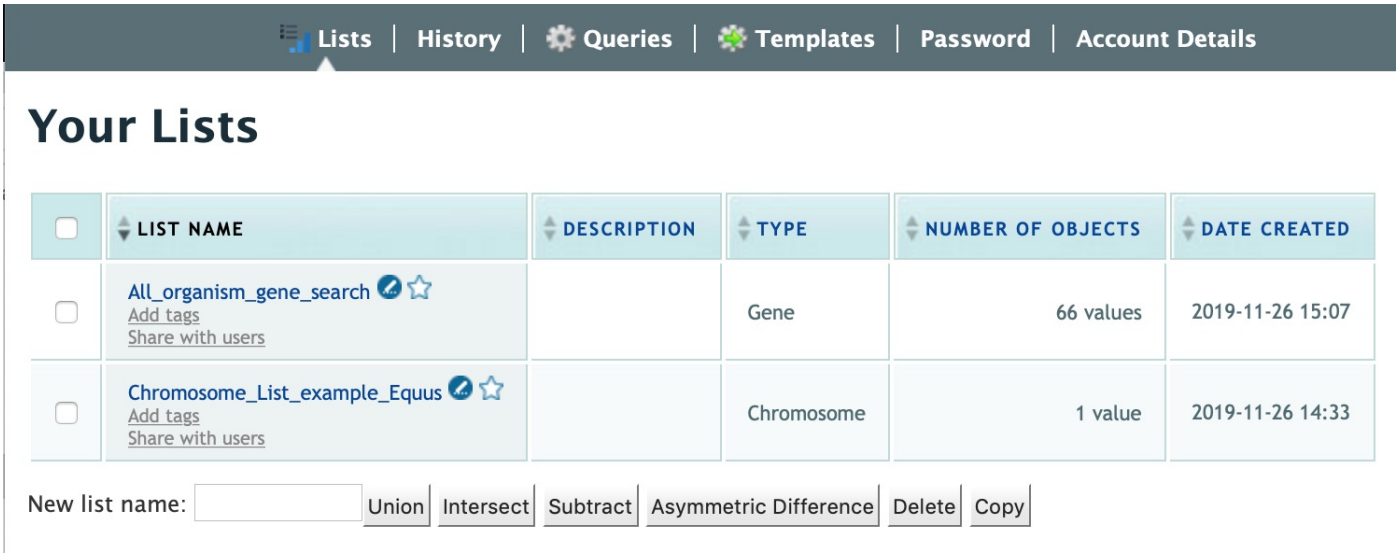
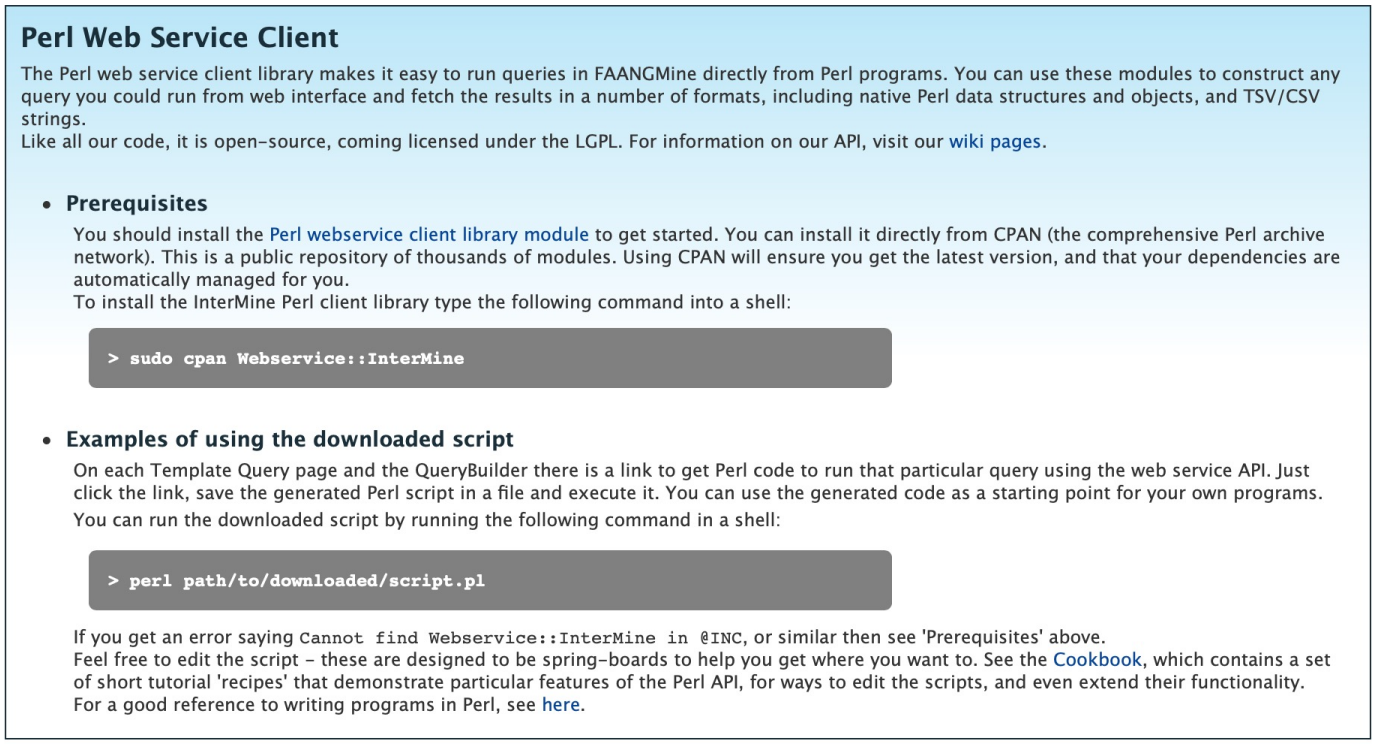


Fig. 1: Saved lists found under MyMine. Note that currently saved lists can be selected for analyses to contribute to new lists.

CHAPTER 7

API

An API is available for users who would like to programmatically access FAANGMine.



Perl Web Service Client

The Perl web service client library makes it easy to run queries in FAANGMine directly from Perl programs. You can use these modules to construct any query you could run from web interface and fetch the results in a number of formats, including native Perl data structures and objects, and TSV/CSV strings.
Like all our code, it is open-source, coming licensed under the LGPL. For information on our API, visit our wiki pages.

- **Prerequisites**

  You should install the Perl webservice client library module to get started. You can install it directly from CPAN (the comprehensive Perl archive network). This is a public repository of thousands of modules. Using CPAN will ensure you get the latest version, and that your dependencies are automatically managed for you.
  To install the InterMine Perl client library type the following command into a shell:

  ```
  > sudo cpan Webservice::InterMine
  ```

- **Examples of using the downloaded script**

  On each Template Query page and the QueryBuilder there is a link to get Perl code to run that particular query using the web service API. Just click the link, save the generated Perl script in a file and execute it. You can use the generated code as a starting point for your own programs.
  You can run the downloaded script by running the following command in a shell:

  ```
  > perl path/to/downloaded/script.pl
  ```

  If you get an error saying `Cannot find Webservice::InterMine in @INC`, or similar then see 'Prerequisites' above.
  Feel free to edit the script - these are designed to be spring-boards to help you get where you want to. See the Cookbook, which contains a set of short tutorial 'recipes' that demonstrate particular features of the Perl API, for ways to edit the scripts, and even extend their functionality.
  For a good reference to writing programs in Perl, see here.

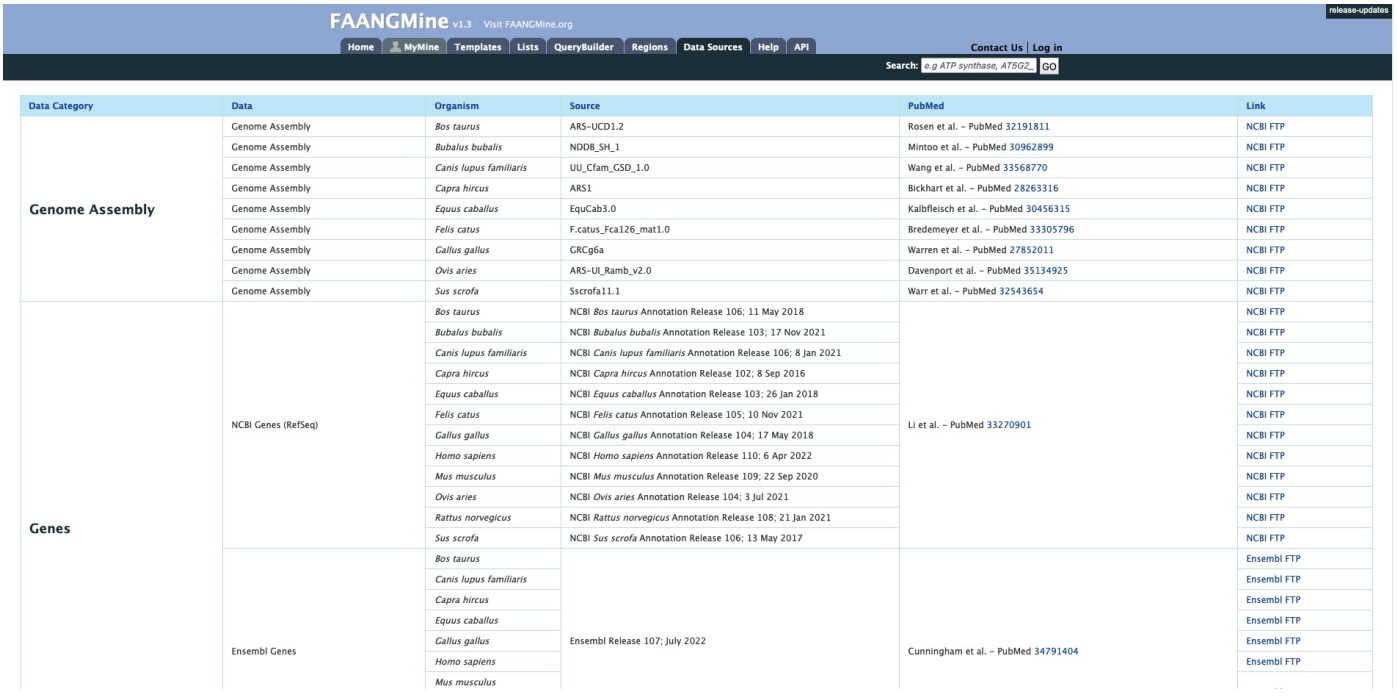
Perl, Python, Ruby, and Java are the languages supported by the InterMine API.

For more detailed information, view the InterMine documentation.

CHAPTER 8

Data Sources

The Data Sources table provides a description of the datasets that are integrated into FAANGMine, along with their download location, version or release, citations wherever applicable, and any additional comments.



Data Category	Data	Organism	Source	PubMed	Link
Genome Assembly	Genome Assembly	*Bos taurus*	ARS-UCD1.2	Rosen et al. - PubMed 32191811	NCBI FTP
	Genome Assembly	*Bubalus bubalis*	NDDB_SH_1	Mintoo et al. - PubMed 30962899	NCBI FTP
	Genome Assembly	*Canis lupus familiaris*	UU_Cfam_GSD_1.0	Wang et al. - PubMed 33568770	NCBI FTP
	Genome Assembly	*Capra hircus*	ARS1	Bickhart et al. - PubMed 28263316	NCBI FTP
	Genome Assembly	*Equus caballus*	EquCab3.0	Kalbfleisch et al. - PubMed 30456315	NCBI FTP
	Genome Assembly	*Felis catus*	F.catus_Fca126_mat1.0	Bredemeyer et al. - PubMed 33305796	NCBI FTP
	Genome Assembly	*Gallus gallus*	GRCg6a	Warren et al. - PubMed 27852011	NCBI FTP
	Genome Assembly	*Ovis aries*	ARS-UI_Ramb_v2.0	Davenport et al. - PubMed 35134925	NCBI FTP
	Genome Assembly	*Sus scrofa*	Sscrofa11.1	Warr et al. - PubMed 32543654	NCBI FTP
Genes	NCBI Genes (RefSeq)	*Bos taurus*	NCBI *Bos taurus* Annotation Release 106; 11 May 2018	Li et al. - PubMed 33270901	NCBI FTP
		Bubalus bubalis	NCBI *Bubalus bubalis* Annotation Release 103; 17 Nov 2021		NCBI FTP
		Canis lupus familiaris	NCBI *Canis lupus familiaris* Annotation Release 106; 8 Jan 2021		NCBI FTP
		Capra hircus	NCBI *Capra hircus* Annotation Release 102; 8 Sep 2016		NCBI FTP
		Equus caballus	NCBI *Equus caballus* Annotation Release 103; 26 Jan 2018		NCBI FTP
		Felis catus	NCBI *Felis catus* Annotation Release 105; 10 Nov 2021		NCBI FTP
		Gallus gallus	NCBI *Gallus gallus* Annotation Release 104; 17 May 2018		NCBI FTP
		Homo sapiens	NCBI *Homo sapiens* Annotation Release 110; 6 Apr 2022		NCBI FTP
		Mus musculus	NCBI *Mus musculus* Annotation Release 109; 22 Sep 2020		NCBI FTP
		Ovis aries	NCBI *Ovis aries* Annotation Release 104; 3 Jul 2021		NCBI FTP
		Rattus norvegicus	NCBI *Rattus norvegicus* Annotation Release 108; 21 Jan 2021		NCBI FTP
		Sus scrofa	NCBI *Sus scrofa* Annotation Release 106; 13 May 2017		NCBI FTP
	Ensembl Genes	*Bos taurus*	Ensembl Release 107; July 2022	Cunningham et al. - PubMed 34791404	Ensembl FTP
		Canis lupus familiaris			Ensembl FTP
		Capra hircus			Ensembl FTP
		Equus caballus			Ensembl FTP
		Gallus gallus			Ensembl FTP
		Homo sapiens			Ensembl FTP
		Mus musculus			

Fig. 1: FAANGMine Data Sources table

CHAPTER 9

How to cite

FAANGMine is a project supported by the National Science Foundation to address the need for a high performance data mining resource that enables fine-grained querying and integrating the heterogeneous FAANG data with existing information, such as functions of known genes and research datasets.

For more generic examples on how to use InterMine, click here. These are tutorials created by FlyMine that showcase the different features of InterMine.