BovineMine Documentation

Release 1.0

Deepak Unni, Aditi Tayal, Colin Diesh, Christine Elsik, Darren Hag

Jan 26, 2018

Contents

1 Tutorial **3**

1.1 Overview . . . 3
1.2 Navigation and Searching in BovineMine . . . 4
1.3 Report Page . . . 18
1.4 BLAST . . . 20
1.5 Genomic Region Search . . . 20
1.6 Lists . . . 22
1.7 MyMine . . . 26
1.8 API . . . 27
1.9 Data Sources . . . 28

2 Indices and tables **29**

BovineMine is a data mining warehouse for Bos taurus genome (UMD3.1.1 assembly), powered by InterMine.

Link to main site: http://bovinegenome.org/bovinemine

Link to our YouTube Channel: Bovine Genome Database on YouTube

Presentations:

- Generating Customized Integrated Functional Annotation Datasets with BovineMine - ISAG 2017

BovineMine integrates datasets from various data sources such as NCBI RefSeq, ENSEMBL, UniProt, GO annotations, EnsemblCompara, Homologene, Reactome. Additionally, BovineMine also integrates datasets from http://www.bovinegenome.org.

Link to the available datasets in BovineMine: http://bovinegenome.org/bovinemine/dataCategories.do

The Bovine Genome Database is supported by the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no. 613689, and has been supported by grants 2007-35616-17882 and 2010-65205-20407 from the USDA National Institute of Food and Agriculture. BGD is hosted at the University of Missouri. If you have comments or if you wish to report a problem, please contact the Bovine Genome Database Administrator.

Contents:

CHAPTER 1

Tutorial

BovineMine is powered by InterMine and provides a user-friendly way to access genomic, proteomic, interaction and literature data. This tutorial is aimed at giving users an introduction to different features of BovineMine.

1.1 Overview

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

The navigation panel highlights different functionalities of BovineMine based on what the user is searching for.

1. **Home** - The homepage for BovineMine
2. **Templates** - Shows the list of templates that users can use based on the nature of their query. Each template is associated with a description which describes what input is expected from the user and what will be the output.
3. **Lists** - The list page offers users a form which they can use to submit or upload a list of genes. Once the list is saved, users can perform enrichment analyses of these genes and export the results.
4. **QueryBuilder** - The QueryBuilder is an interface that provides the user to directly interact with the BovineMine data model. Some users might find the QueryBuilder difficult at first, but this tutorial is aimed at familiarizing users on the functionality of QueryBuilder, how to build a custom template and save their own templates.
5. **BLAST** - The BLAST page is where users can search their input sequence(s) against the several databases provided.
6. **Regions** - The Genomic Region Search tool provides a form where users can enter a series of genomic coordinates, specify flanking regions (if any) and fetch all features that fall within the given interval. The result can be exported, saved as a list for further analyses.

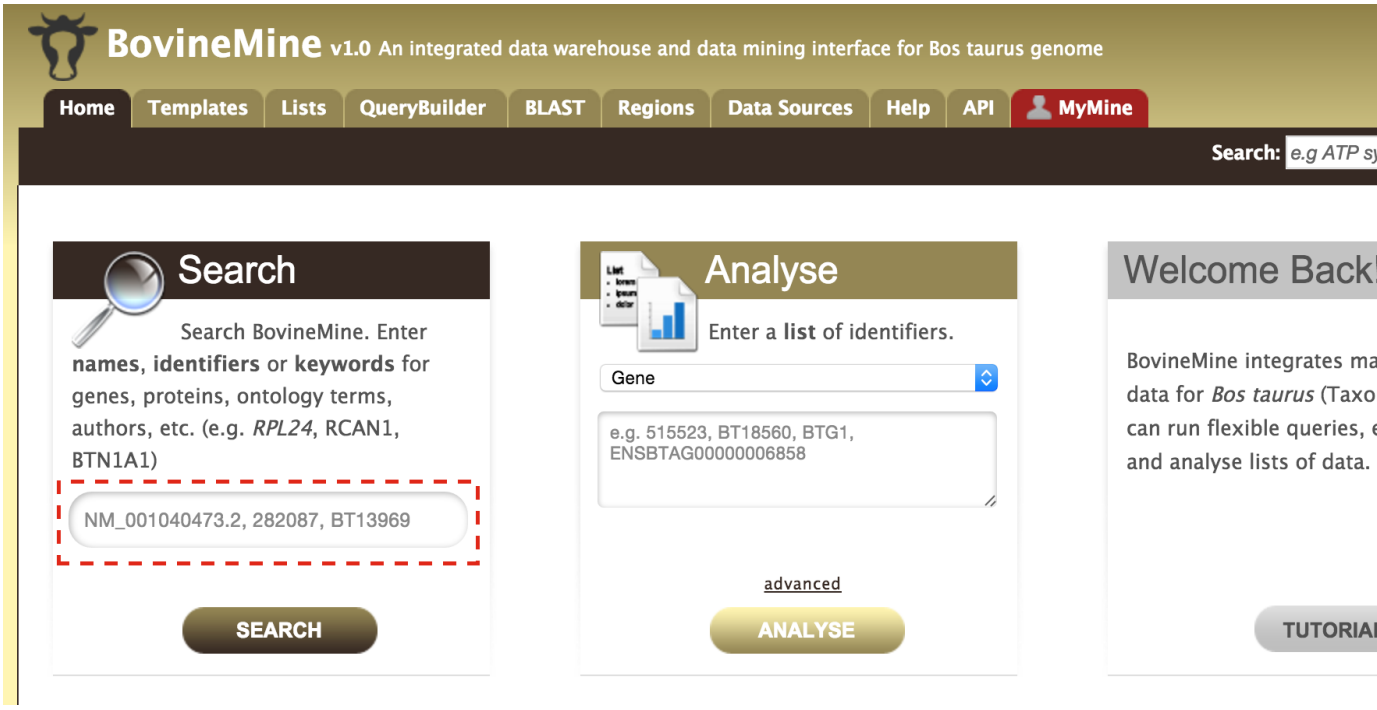
7. **Data Sources** - The Data Sources page provides a summary of all the data that is loaded into BovineMine. It lists a combination of datasets from other databases as well as in-house prepared datasets.
8. **API** - The API provides users an option for programmatically accessing BovineMine by using the InterMine API.
9. **MyMine** - The MyMine serves as a portal for users (those who are logged in) for account management. Users can access their saved templates, most recent queries and saved lists.

1.2 Navigation and Searching in BovineMine

There are three ways for users to query BovineMine.

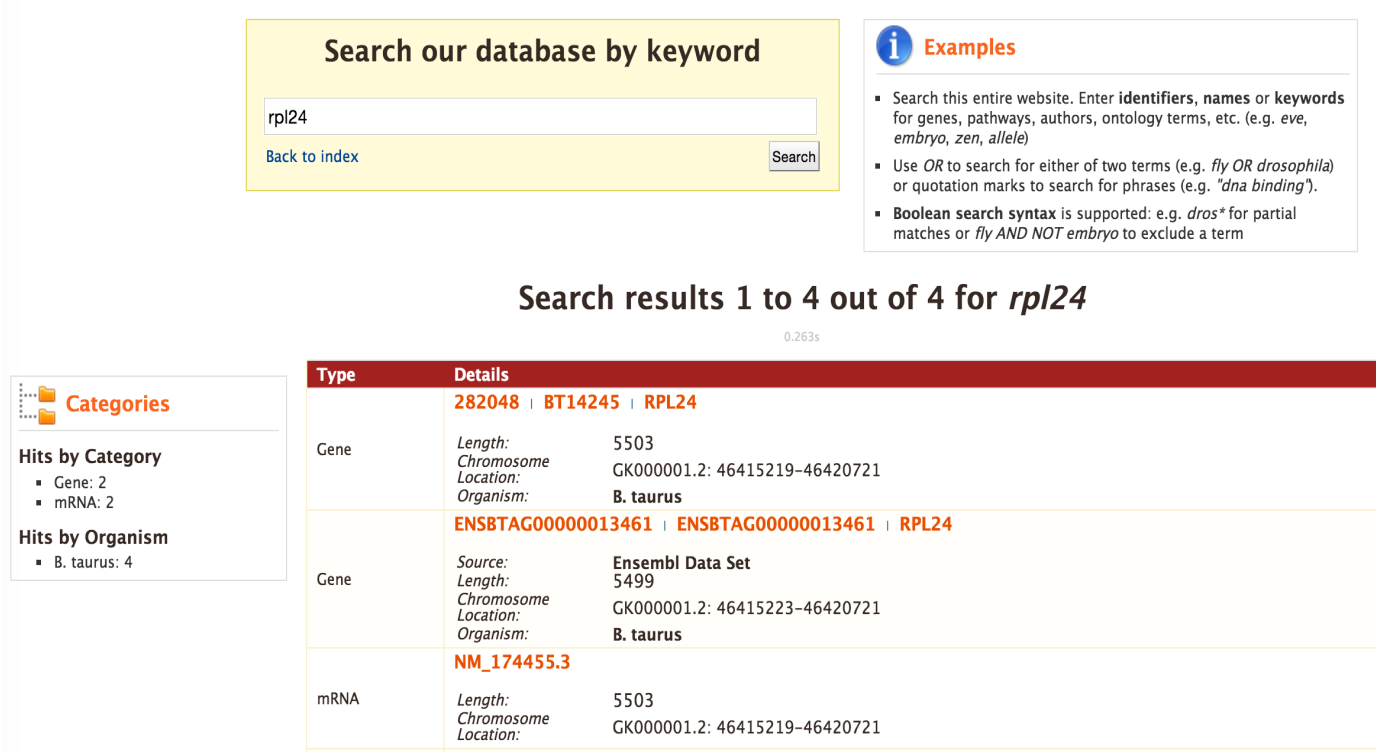
1.2.1 a. 'Search' Box

The Search box is located on the BovineMine home page.



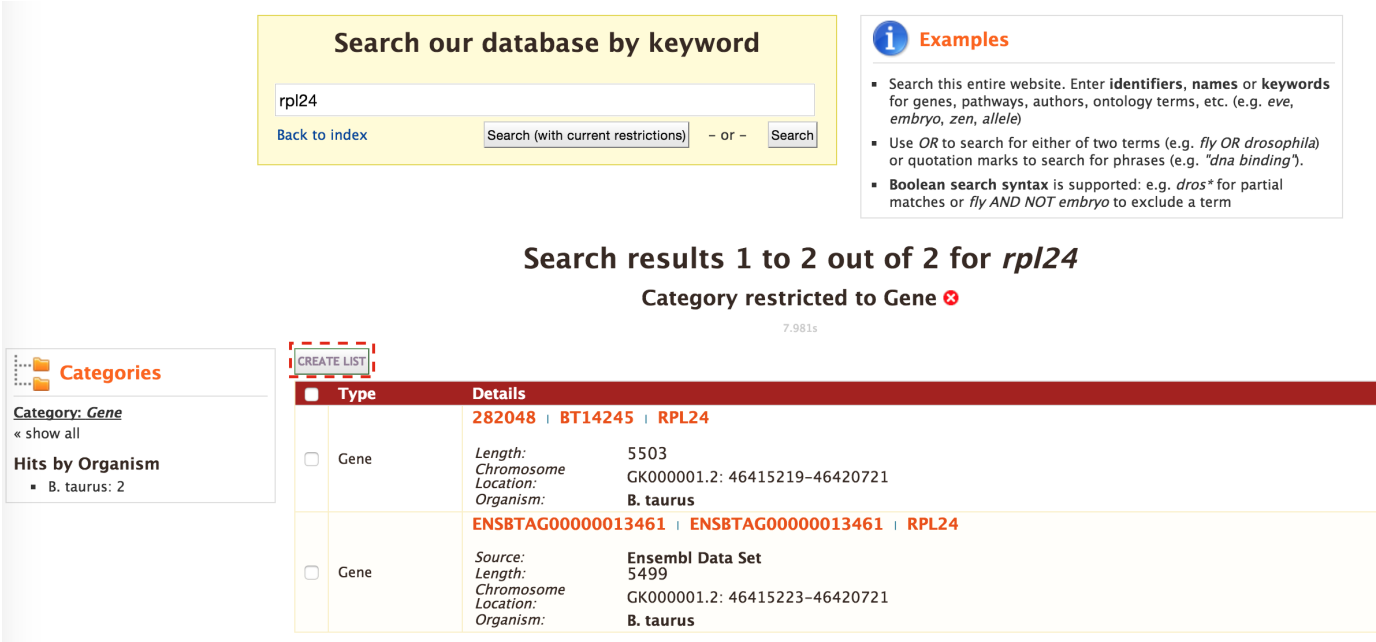
There are two locations (marked) where users can input the name, Gene ID, or keywords related to their gene of interest. The search box is not just limited to Gene but a variety of features such as Transcript, Exons, Interaction, Pathways, GO Annotations and more. Users can make use of wildcard character '*' to get all results that match a particular prefix.

Lets consider gene **rpl24** as an example. Enter rpl24 in the search box and click `Search`. The results page is tabulated and shows a summary about your query, as shown below.



Users can filter the results based on `Category` and `Organism`. The score column in the result table indicates the similarity of your query to each of the result fetched by BovineMine.

The results page can also be converted to a list. To enable this feature click on `Gene` in `Hits by Category`.

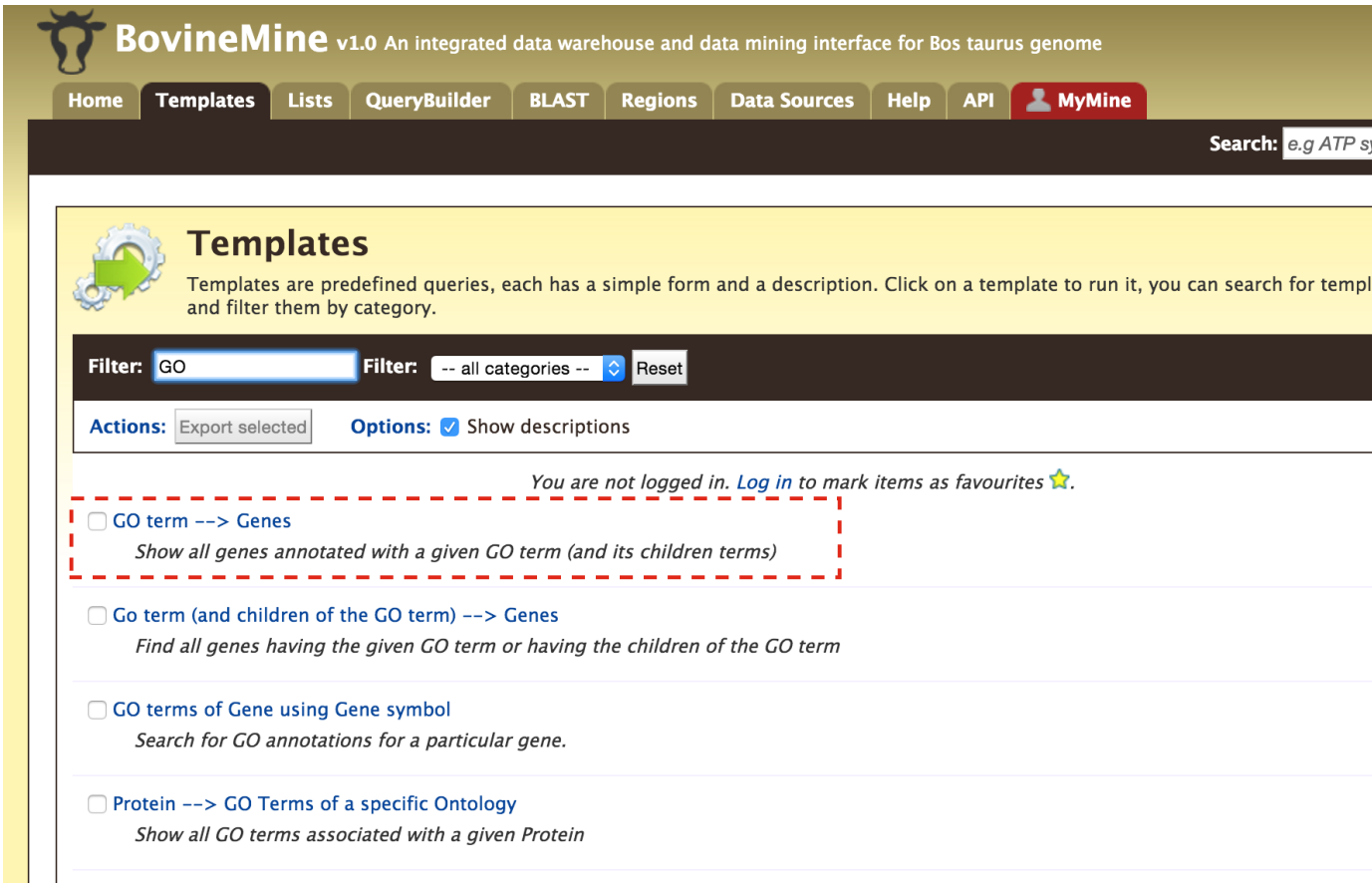


This will filter the results for the feature type `Gene` and checkboxes will be available for users to select genes they would like to add to their list. Once the genes are selected, click on `CREATE LIST`.

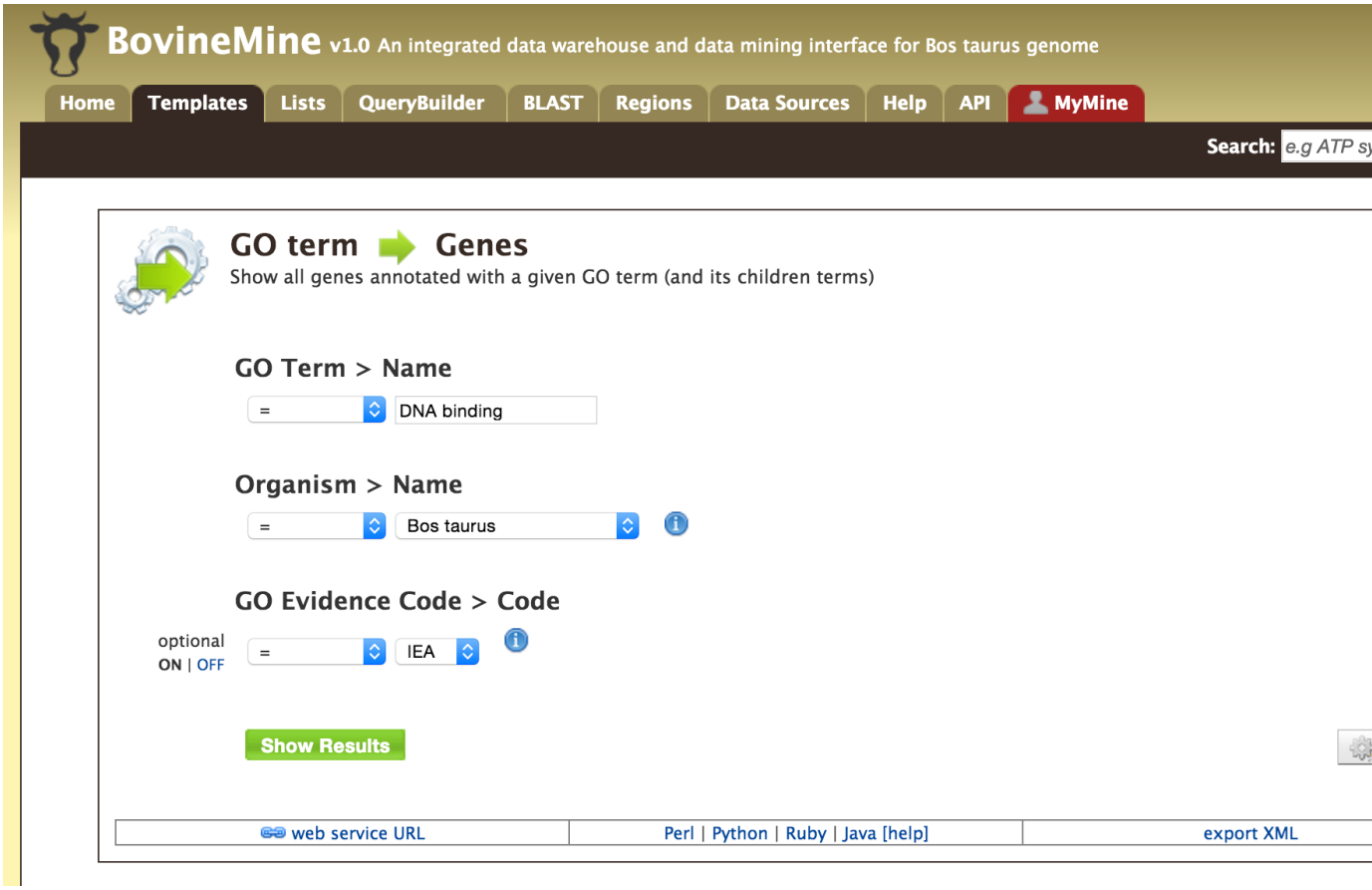
1.2.2 b. Templates

Apart from the search box, users can make use of predefined templates by clicking on the `Templates` tab.

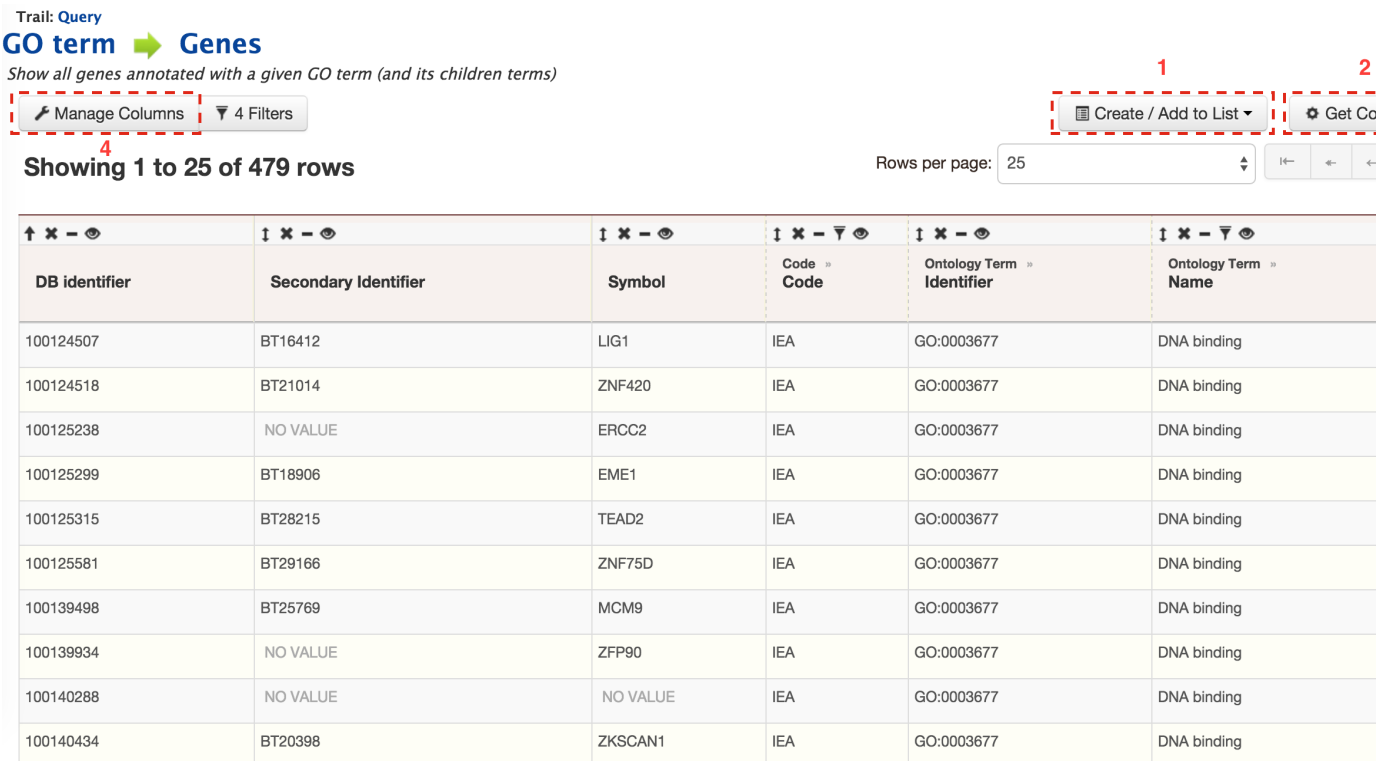
Users can choose from a list of templates based on the nature of their query.



As an example, lets click on `GO terms -> Genes` template. The goal of this template is to query BovineMine to find all genes having a given GO term.



Users can also specify a filter for GO evidence code. Lets search for the GO term `DNA binding` with GO evidence code `IEA`.



DB identifier	Secondary Identifier	Symbol	Code » Code	Ontology Term » Identifier	Ontology Term » Name
100124507	BT16412	LIG1	IEA	GO:0003677	DNA binding
100124518	BT21014	ZNF420	IEA	GO:0003677	DNA binding
100125238	NO VALUE	ERCC2	IEA	GO:0003677	DNA binding
100125299	BT18906	EME1	IEA	GO:0003677	DNA binding
100125315	BT28215	TEAD2	IEA	GO:0003677	DNA binding
100125581	BT29166	ZNF75D	IEA	GO:0003677	DNA binding
100139498	BT25769	MCM9	IEA	GO:0003677	DNA binding
100139934	NO VALUE	ZFP90	IEA	GO:0003677	DNA binding
100140288	NO VALUE	NO VALUE	IEA	GO:0003677	DNA binding
100140434	BT20398	ZKSCAN1	IEA	GO:0003677	DNA binding

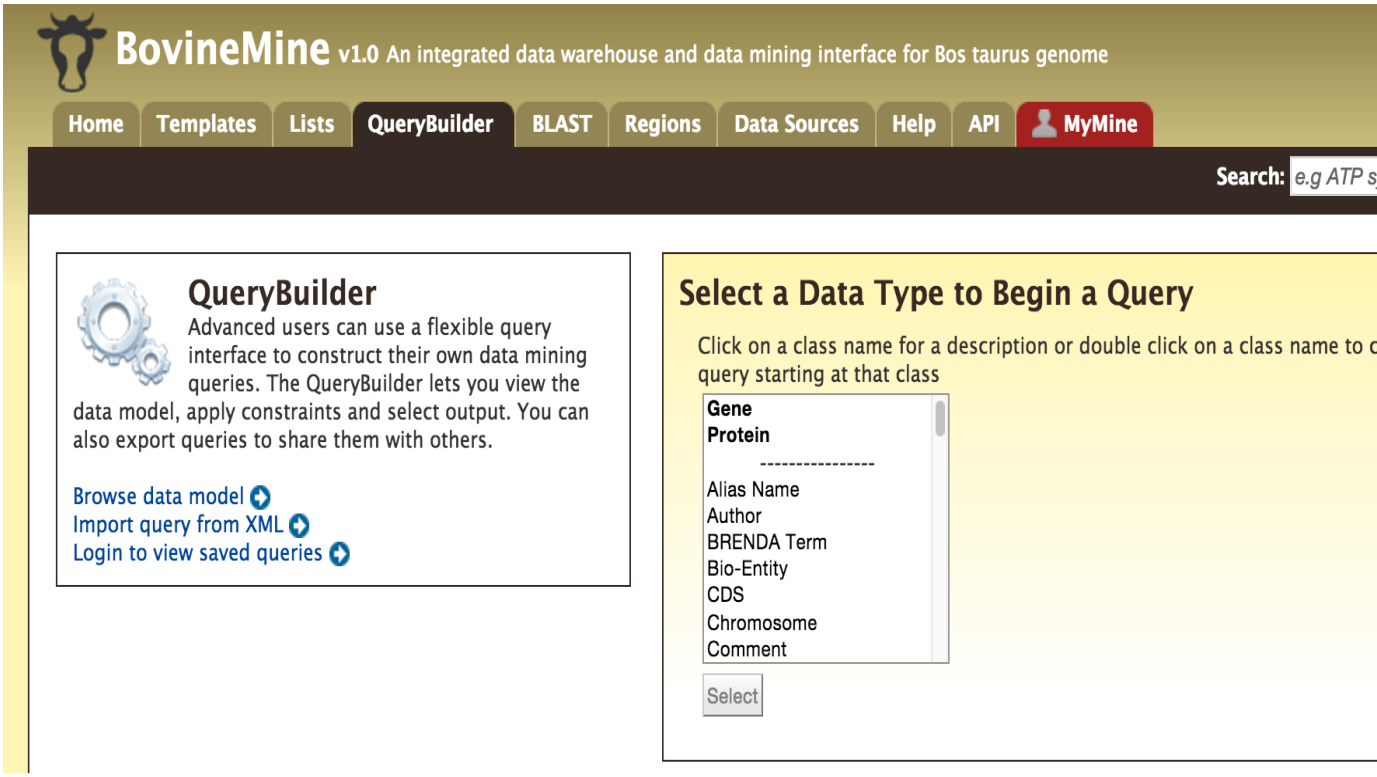
The result page shows all the genes that have been annotated with the Gene Ontology term `DNA binding`.

Users can,

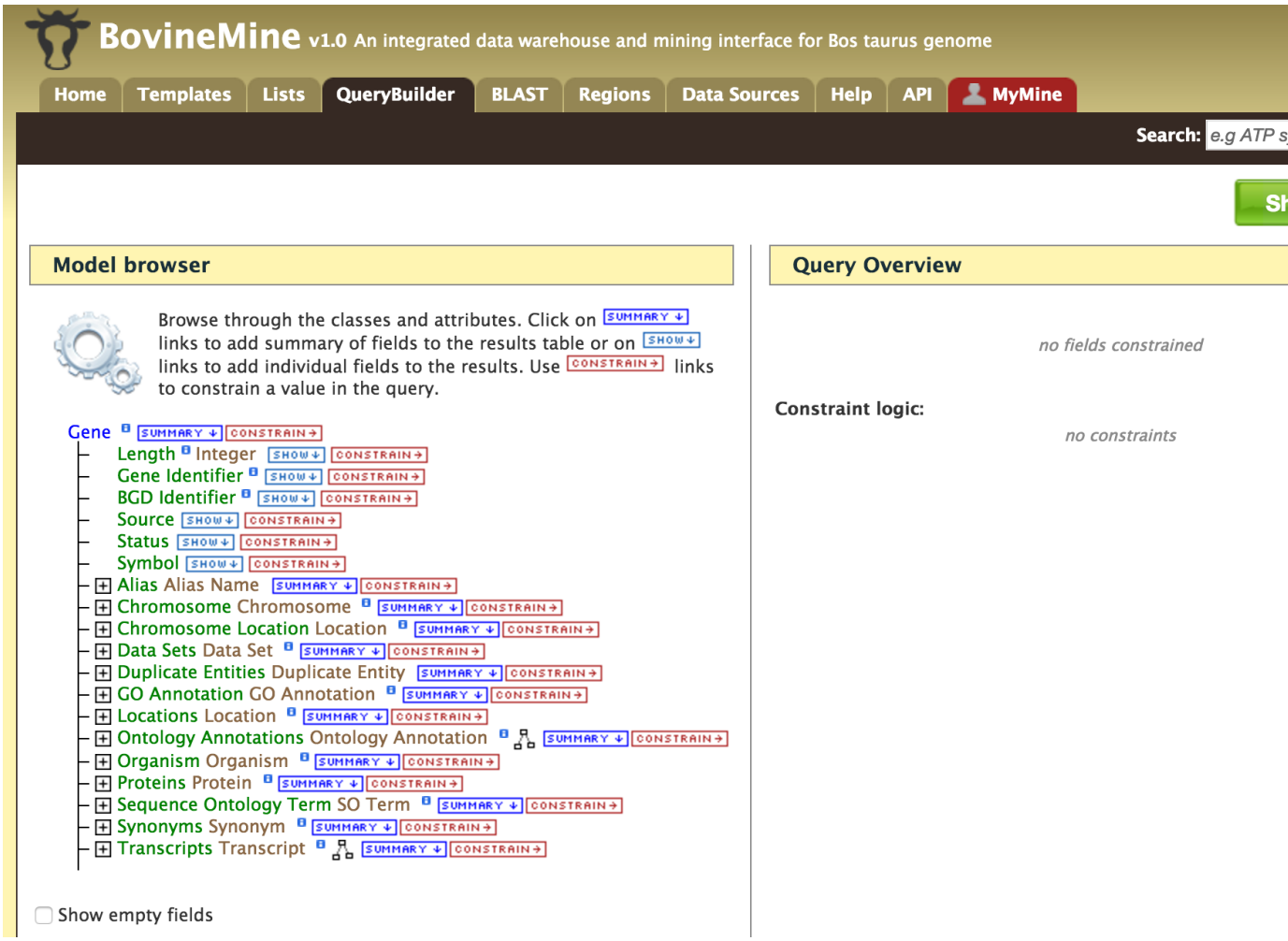
1. create or add these genes to a list, by clicking on `Create/Add to List` (Box 1) to perform further analyses.
2. get the code for the query in Perl, Python, Java, Ruby, JavaScript or XML by clicking on `Get Code` (Box 2).
3. download the search results, by clicking on `Download` (Box 3), as tab-delimited, comma-separated values, XML or JSON. If the results are genomic features, which is true for the current example, then users can download the results in GFF3 and BED format.
4. customize the layout of the result page by clicking on `Manage Columns` (Box 4).

1.2.3 c. QueryBuilder

We have provided templates suitable for several common types of queries but if users need more fidelity in their search they can make use of the QueryBuilder. The possibilities of queries using the QueryBuilder are endless. You can format the output the way you want and constrain your queries to perform complex search operations.



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.

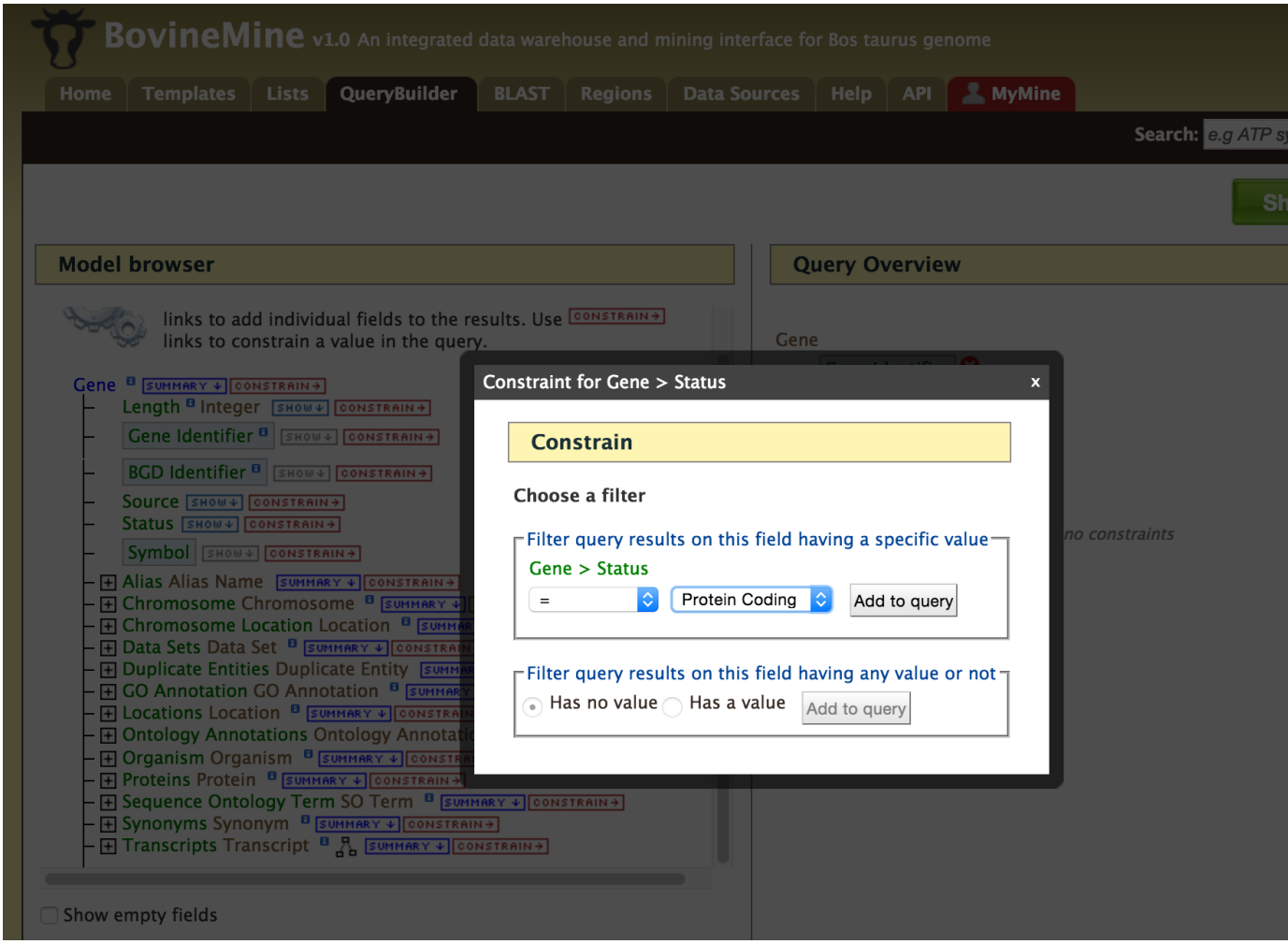


Lets consider three scenarios for using the QueryBuilder,

i. Querying for Protein Coding genes

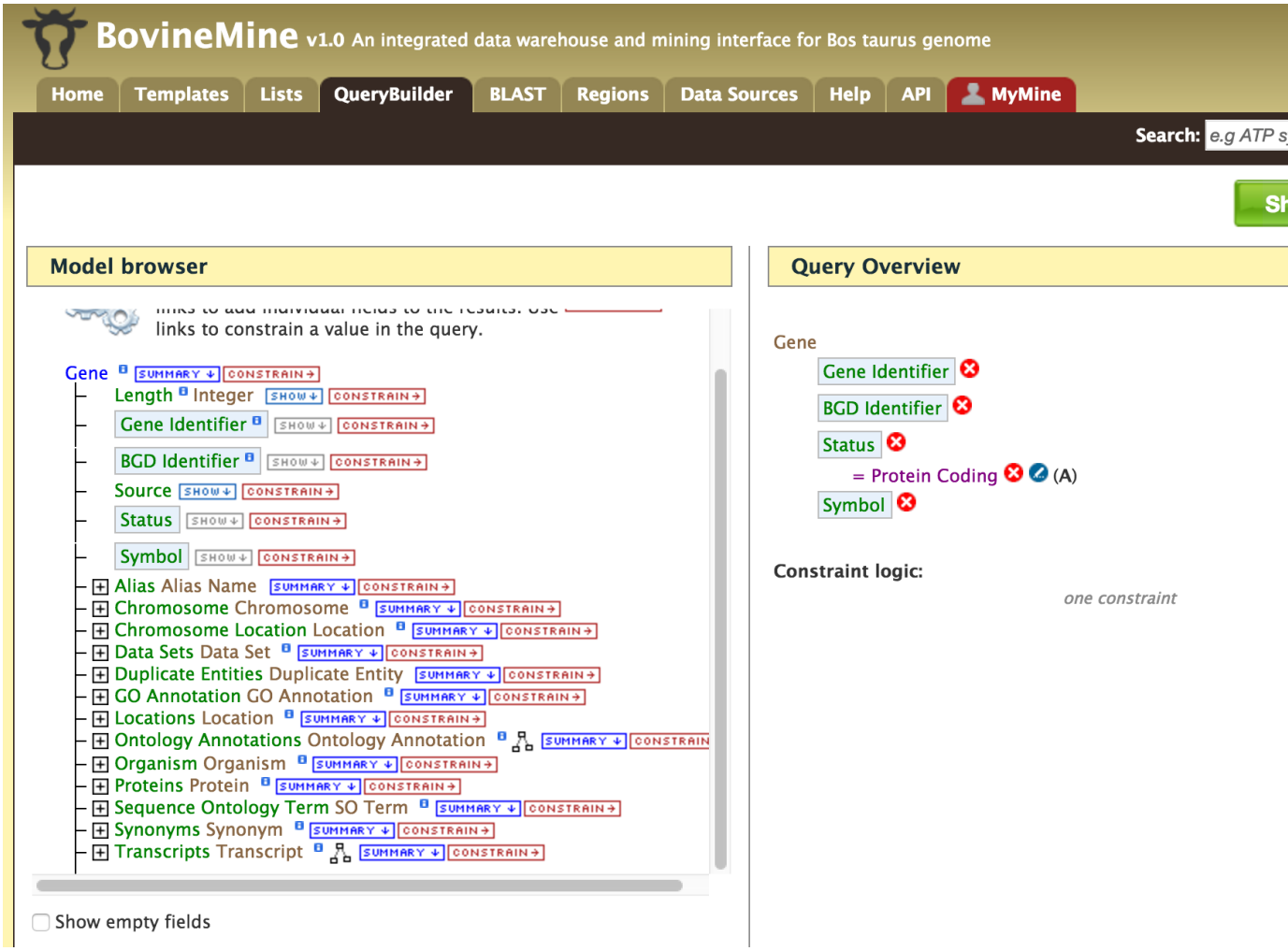
Click on `Show` tab for the attributes `Gene Identifier`, `BGD Identifier`, `Symbol` and `Status`. This tells the QueryBuilder to show the Gene ID, Bovine Genome Database (BGD) ID and the Symbol for `Gene`.

Then click on `Constrain` tab for the attribute `Status`.

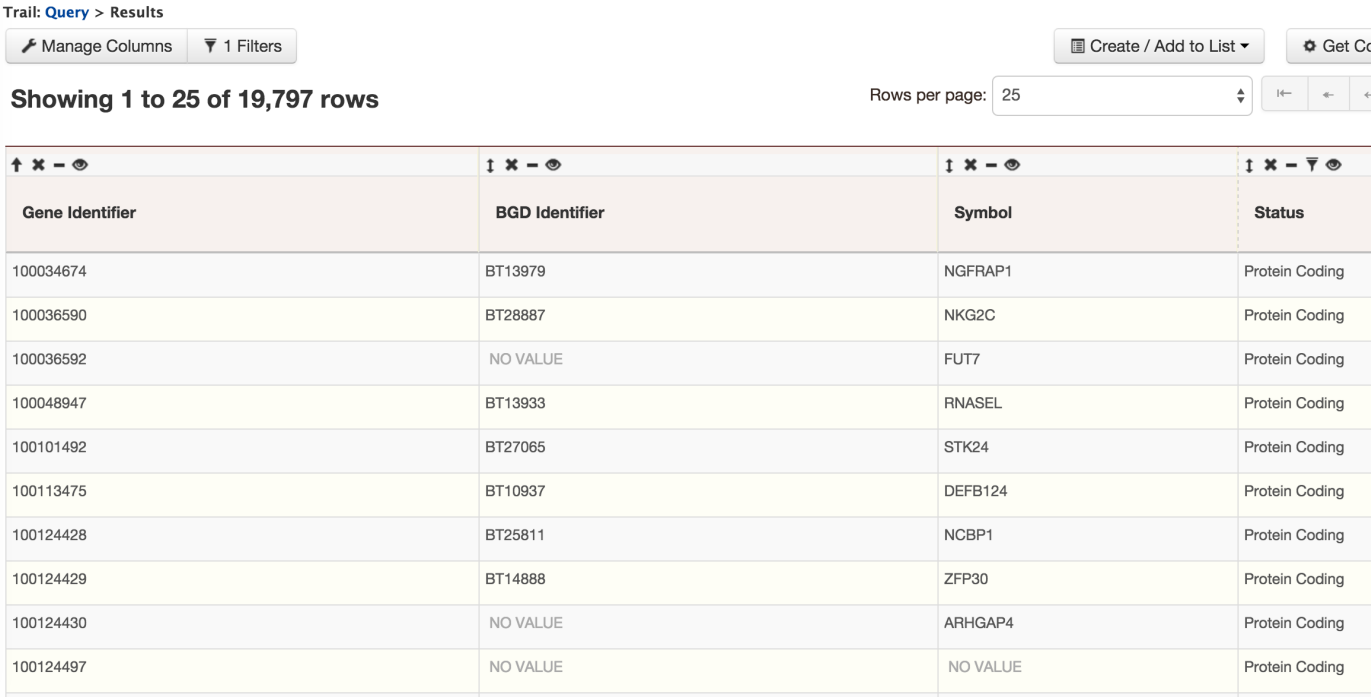


Click on the drop down list for `Status` and select `Protein Coding`. Then click on `Add to Query`.

The Query Overview should resemble the image below,



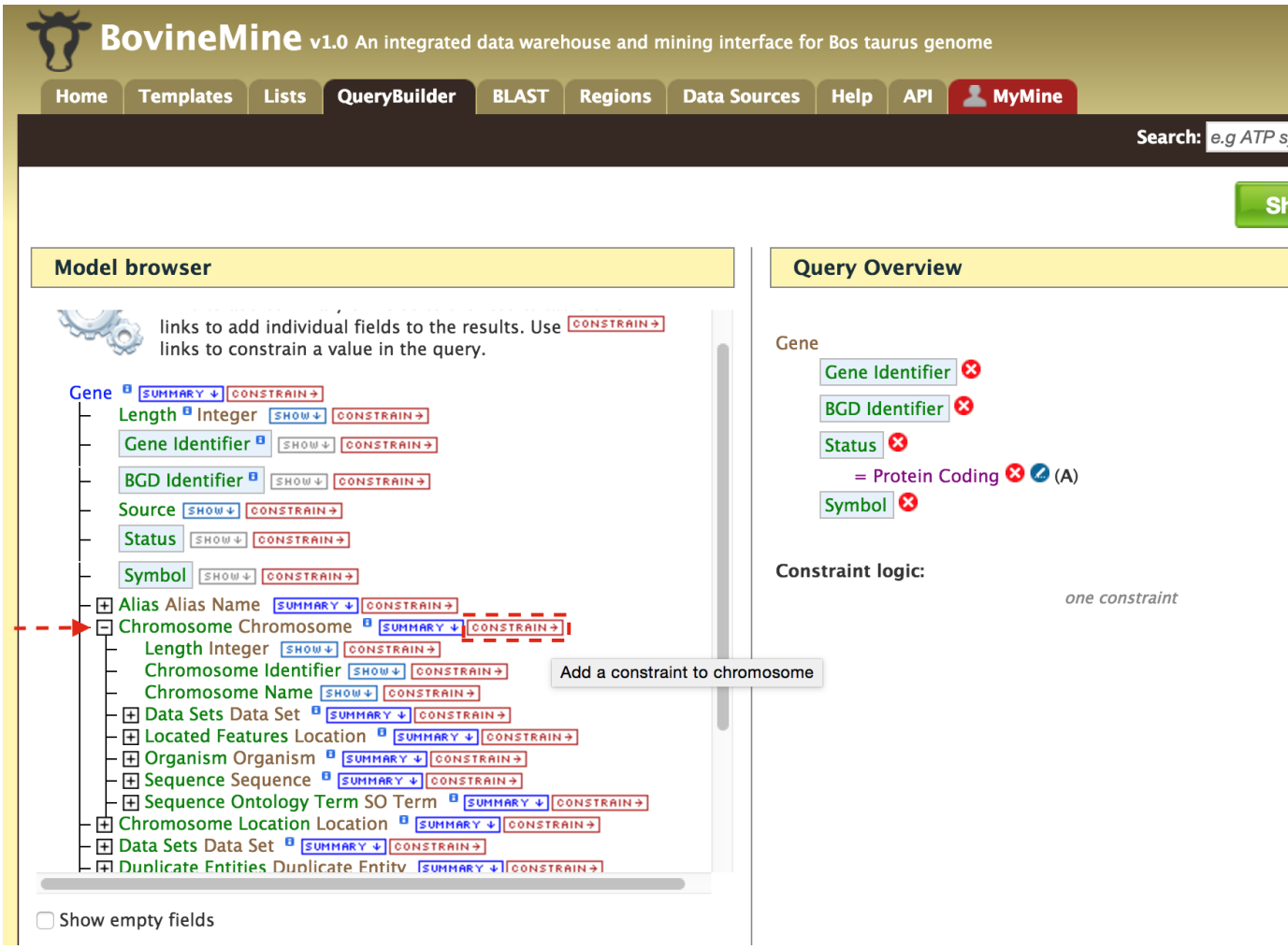
Now, click on `Show results` to see all the genes that have the status `Protein Coding`.



Gene Identifier	BGD Identifier	Symbol	Status
100034674	BT13979	NGFRAP1	Protein Coding
100036590	BT28887	NKG2C	Protein Coding
100036592	NO VALUE	FUT7	Protein Coding
100048947	BT13933	RNASEL	Protein Coding
100101492	BT27065	STK24	Protein Coding
100113475	BT10937	DEFB124	Protein Coding
100124428	BT25811	NCBP1	Protein Coding
100124429	BT14888	ZFP30	Protein Coding
100124430	NO VALUE	ARHGAP4	Protein Coding
100124497	NO VALUE	NO VALUE	Protein Coding

ii. Querying for Protein Coding genes on a particular chromosome

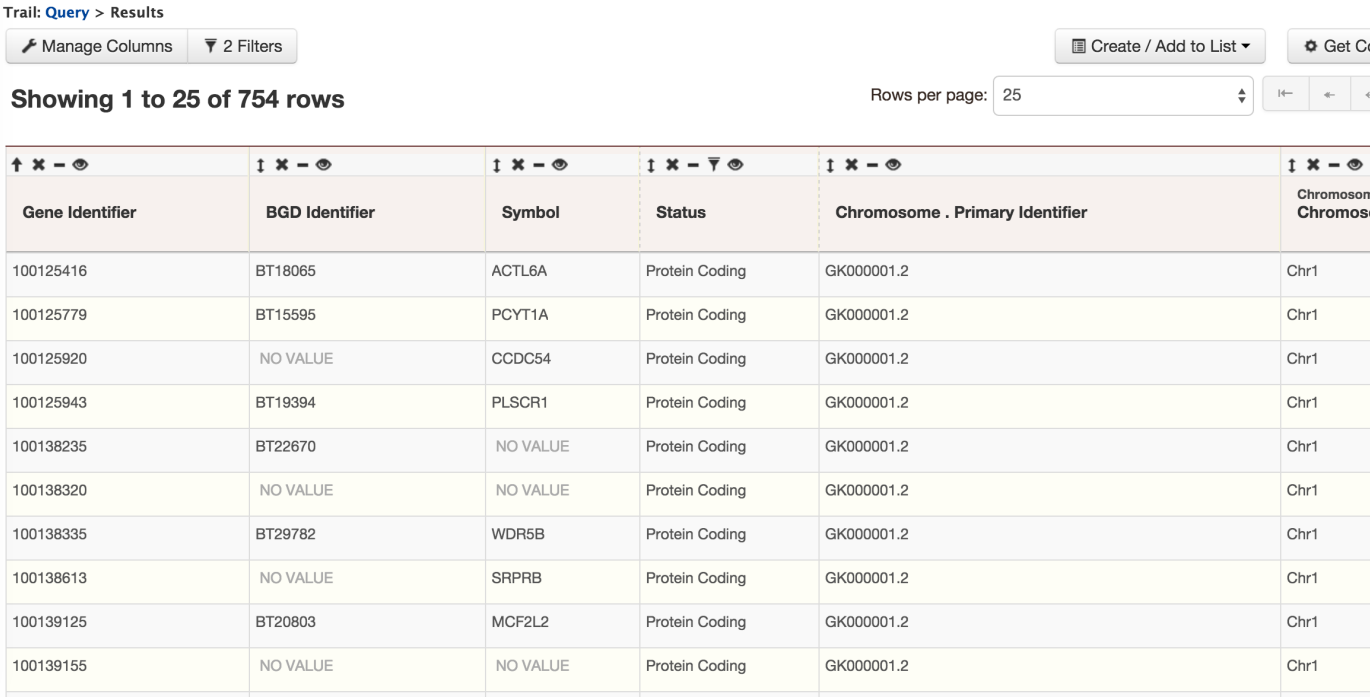
Users can customize the previously run query by adding another constraint for `Chromosome`.



While building the query, click on the `Constrain` tab for the `Chromosome` feature class, enter `Chr1` in the text box of the pop-up window and click on `Add to Query`. Then click on `Show` for `Chromosome Identifier` and `Chromosome Name`.

Now, click on `Show results` and the query will result in all the genes that are of type `Protein Coding` on Chromosome with ID `Chr1`.

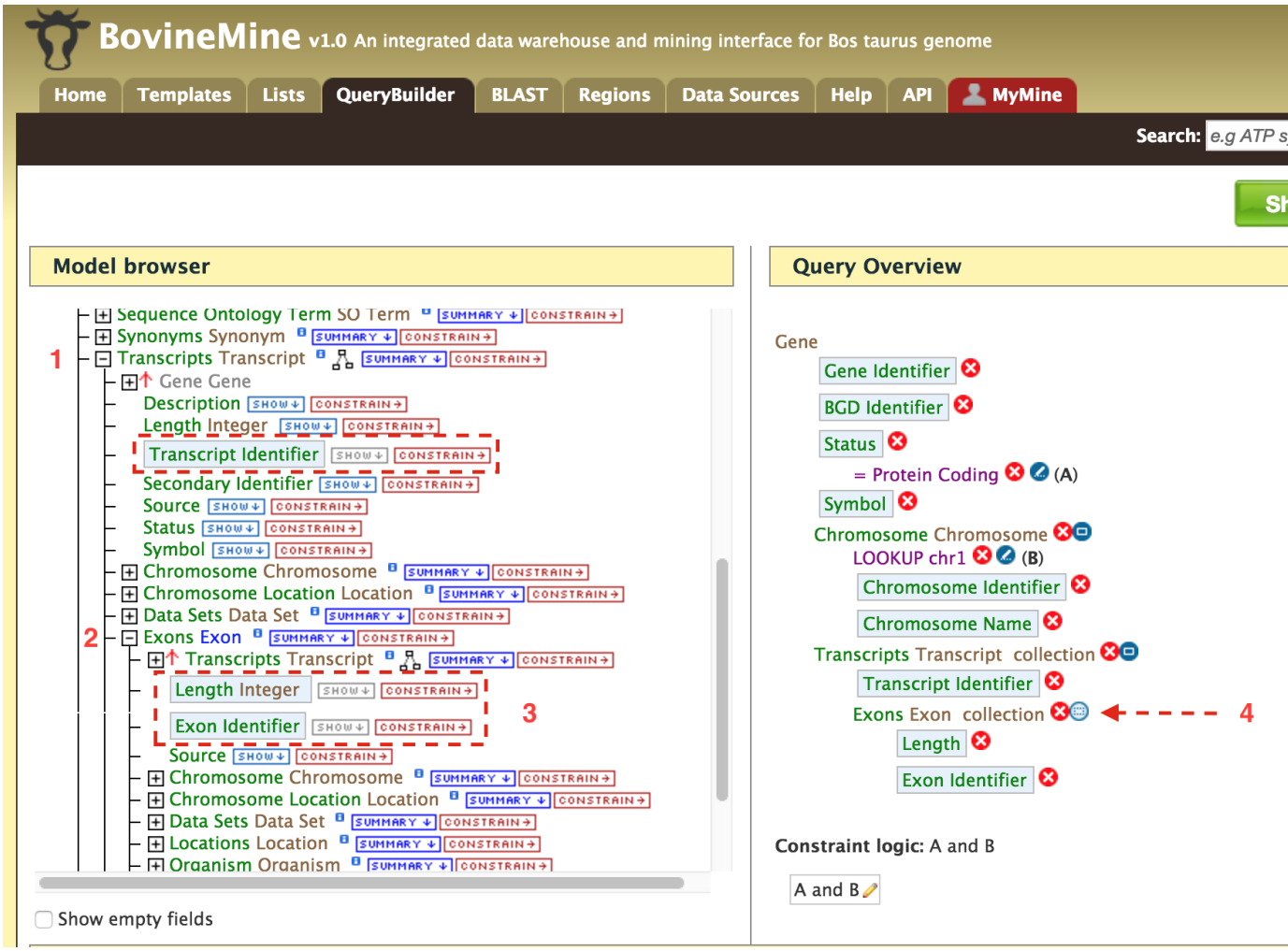
Note: Since here we are specifying a constrain at a class level, instead of at an individual attribute level, the QueryBuilder will look up the input ID `chr1` in all attributes.



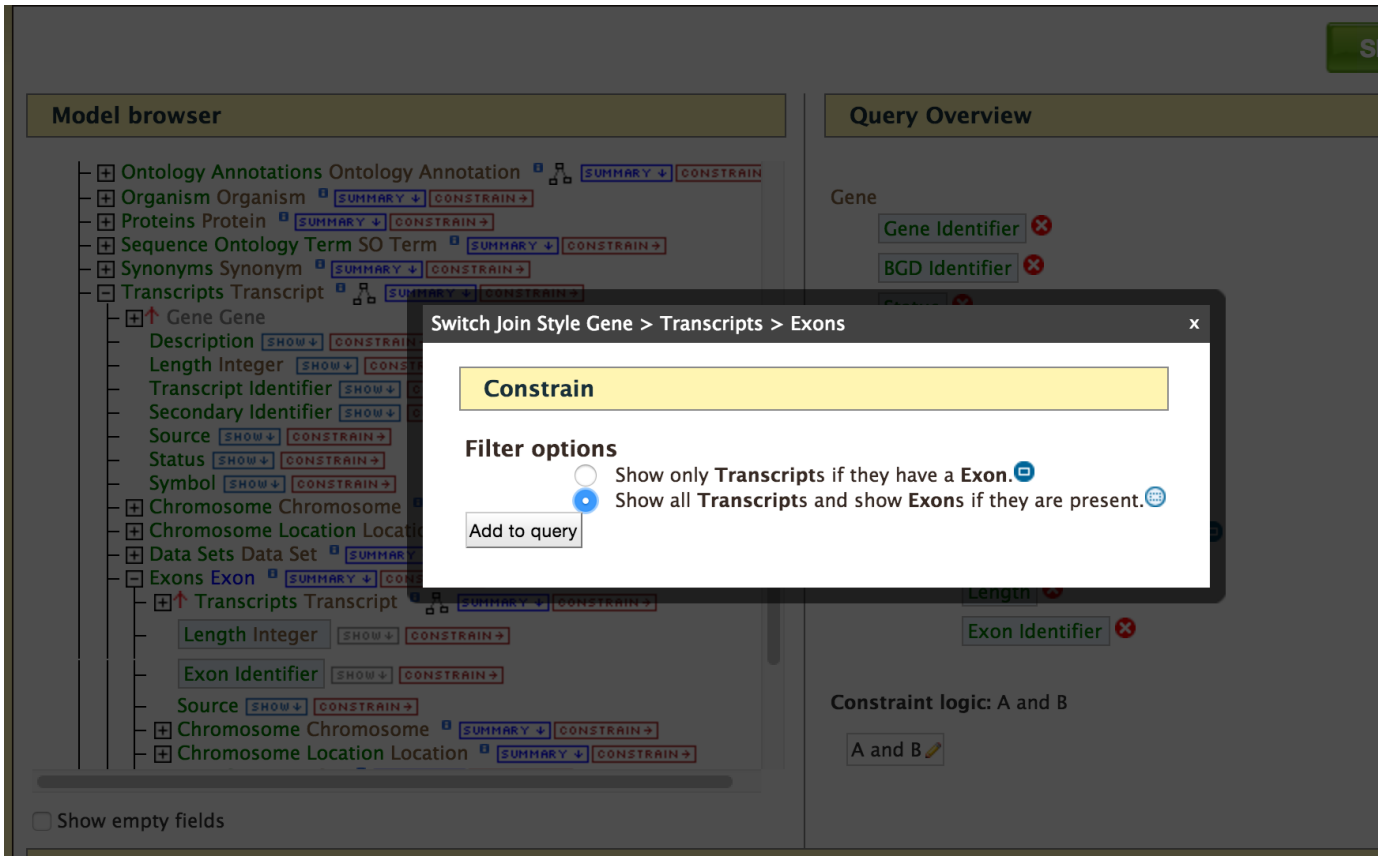
Gene Identifier	BGD Identifier	Symbol	Status	Chromosome . Primary Identifier	Chromoson Chromos
100125416	BT18065	ACTL6A	Protein Coding	GK000001.2	Chr1
100125779	BT15595	PCYT1A	Protein Coding	GK000001.2	Chr1
100125920	NO VALUE	CCDC54	Protein Coding	GK000001.2	Chr1
100125943	BT19394	PLSCR1	Protein Coding	GK000001.2	Chr1
100138235	BT22670	NO VALUE	Protein Coding	GK000001.2	Chr1
100138320	NO VALUE	NO VALUE	Protein Coding	GK000001.2	Chr1
100138335	BT29782	WDR5B	Protein Coding	GK000001.2	Chr1
100138613	NO VALUE	SRPRB	Protein Coding	GK000001.2	Chr1
100139125	BT20803	MCF2L2	Protein Coding	GK000001.2	Chr1
100139155	NO VALUE	NO VALUE	Protein Coding	GK000001.2	Chr1

iii. Querying for Protein Coding genes on a particular chromosome and theirs exons

Users can customize this query further by configuring the query to show all the exons for each Gene.

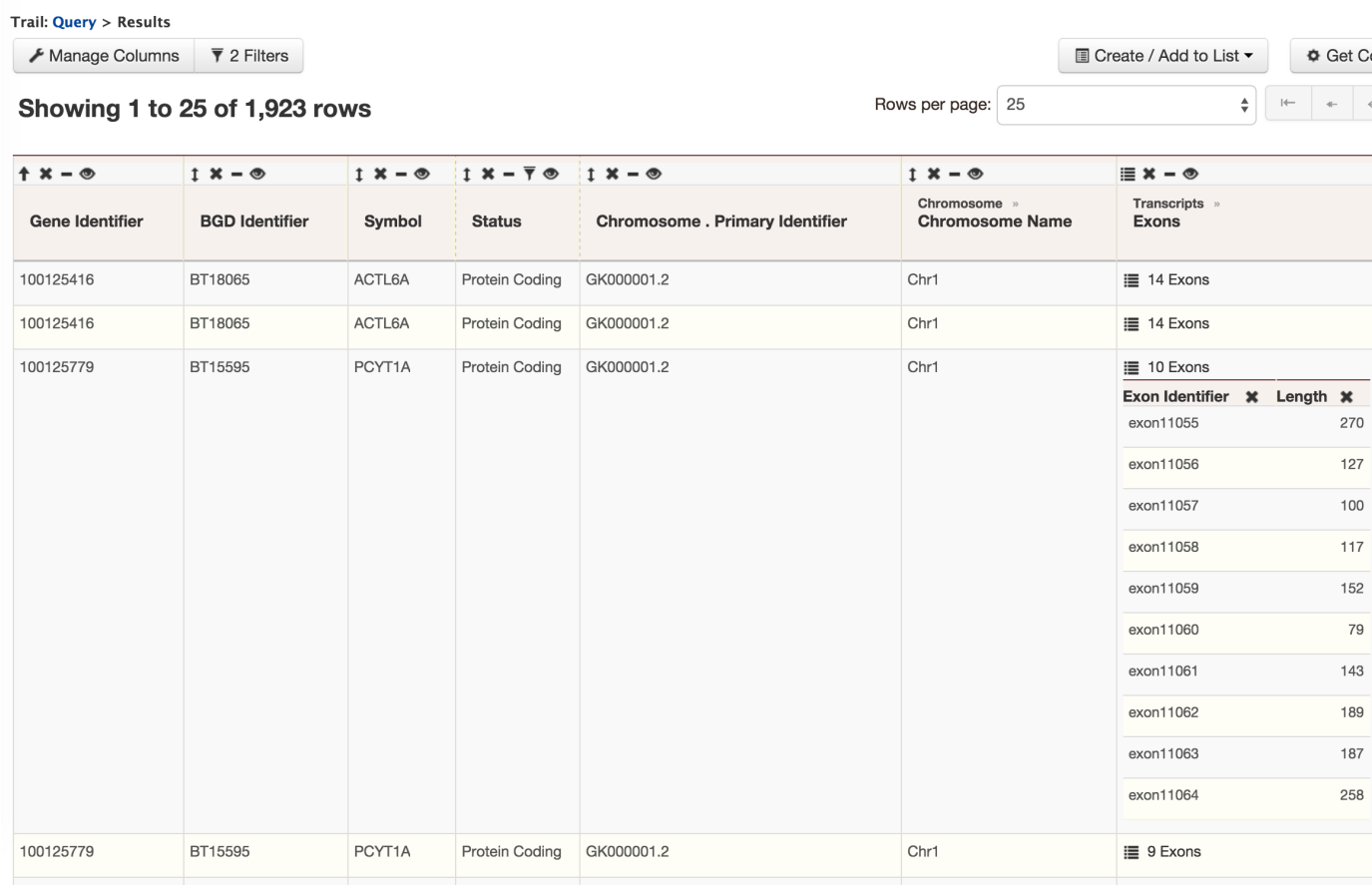


1. Expand the Transcript subclass for Gene
2. Expand the Exon subclass for Transcript
3. Select `Exon Identifier` and `Length`.
4. To make the query more relaxed, click on the blue square icon (labelled with No. 4) near the Exon collection in the Query Overview. You will see a pop as shown below,



Select `Show all Genes and show Exons if they are present` and click on `Add to query`.

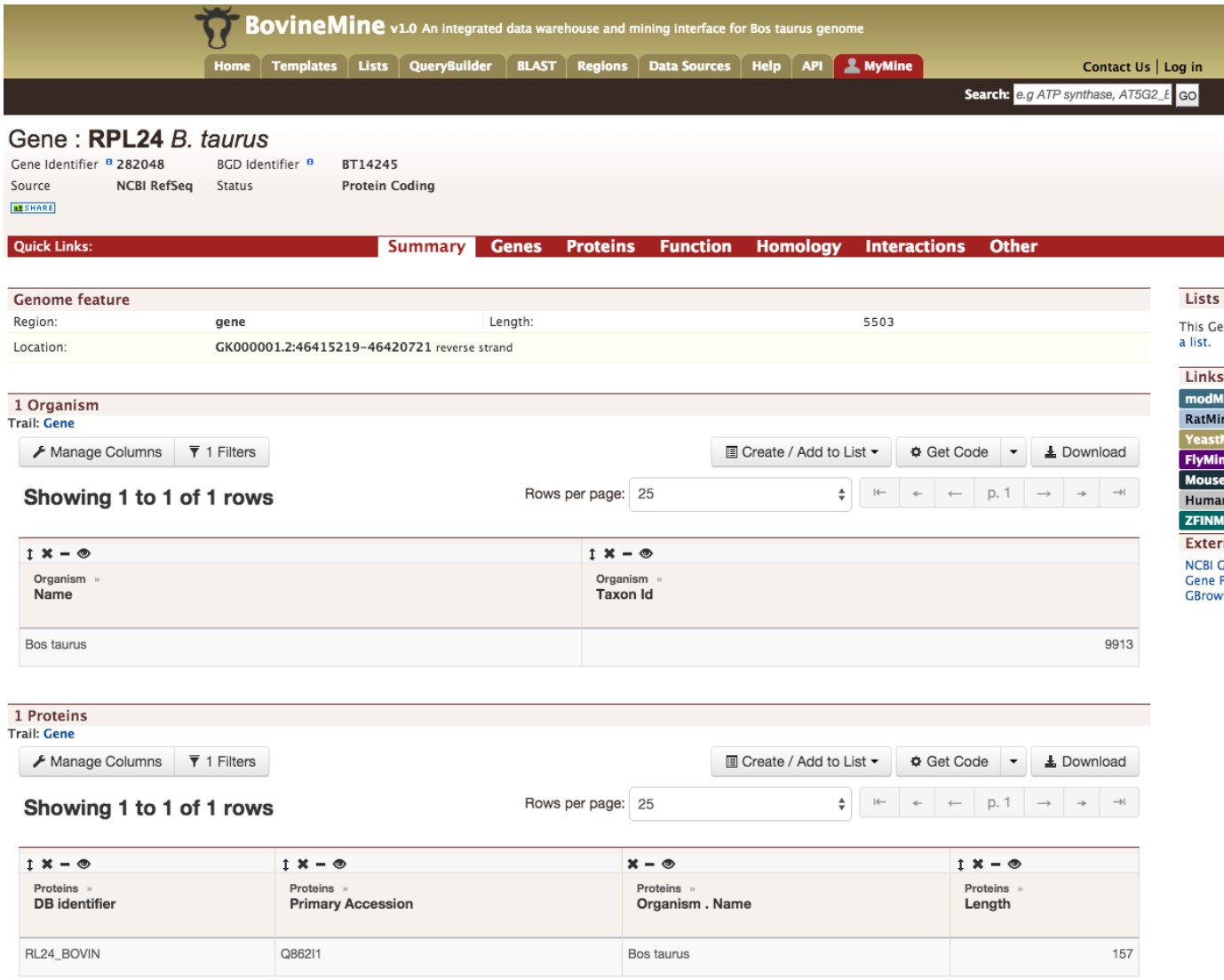
Now run the query and you should see the following results page,



There are 1923 genes on Chromosome 1 and the gene with NCBI Gene ID 100125416 has two transcripts, each with 14 exons. Users can click on the `14 exons` to expand the table with additional rows describing the Exon ID and Length for each exon.

1.3 Report Page

Every query result has a report page and the layout of the report page depends on the data available for a given query. Continuing with the example of **rpl24**, the report page for this gene is shown below.



The report page provides a complete description for gene rpl24. The header of the report page shows the `Gene Identifier` (NCBI Gene ID) and `status` indicating the type of gene, in this case a protein coding gene.

On the right hand side of the report page there are external links that links out to other Mines and databases.

The contents of the report page is divided into categories based on the type of information provided,

Summary

Provides a summary about a gene such as length, chromosome location and strand information. Users can also get the complete FASTA sequence of the gene by clicking on the FASTA tab.

Genes

This section provides information about the gene model. It displays all the transcripts and exons for a gene. The FASTA sequence of each transcript or exon can be downloaded by clicking on the FASTA tab. JBrowse is used to visualize the gene model.

Proteins

This section provides information about the protein product of gene rpl24. The comments section gives a brief description about the protein along with the UniProt accession.

Function

This section provides Gene Ontology annotations for a gene and annotations are divided into three categories,

- Biological process
- Molecular function
- Cellular Component

The GO terms are displayed along with the evidence code indicating how the annotations were derived.

Homology

This section provides information on all the homologues for rpl24. The first part shows a summarized view of the homologues in different organisms. The table below gives a detailed information about the homologue, the type of homologue and from which dataset the information was obtained.

Interaction

This section provides information about interactions. For rpl24 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. The network displayer is a Cytoscape plugin.

Others

This section provides additional information such as,

- Child features – lists all the features that are sub features of the current gene
- Flanking regions – lists all the features flanking the current gene
- Overlapping features – lists all the features that overlap with current gene
- Pathways – lists all the pathways in which the current gene is a part of
- Publications – Publications related to the current gene

1.4 BLAST

Users can perform BLAST against the Bos taurus genomic, CDS or protein sequences using the BLAST page.

1.5 Genomic Region Search

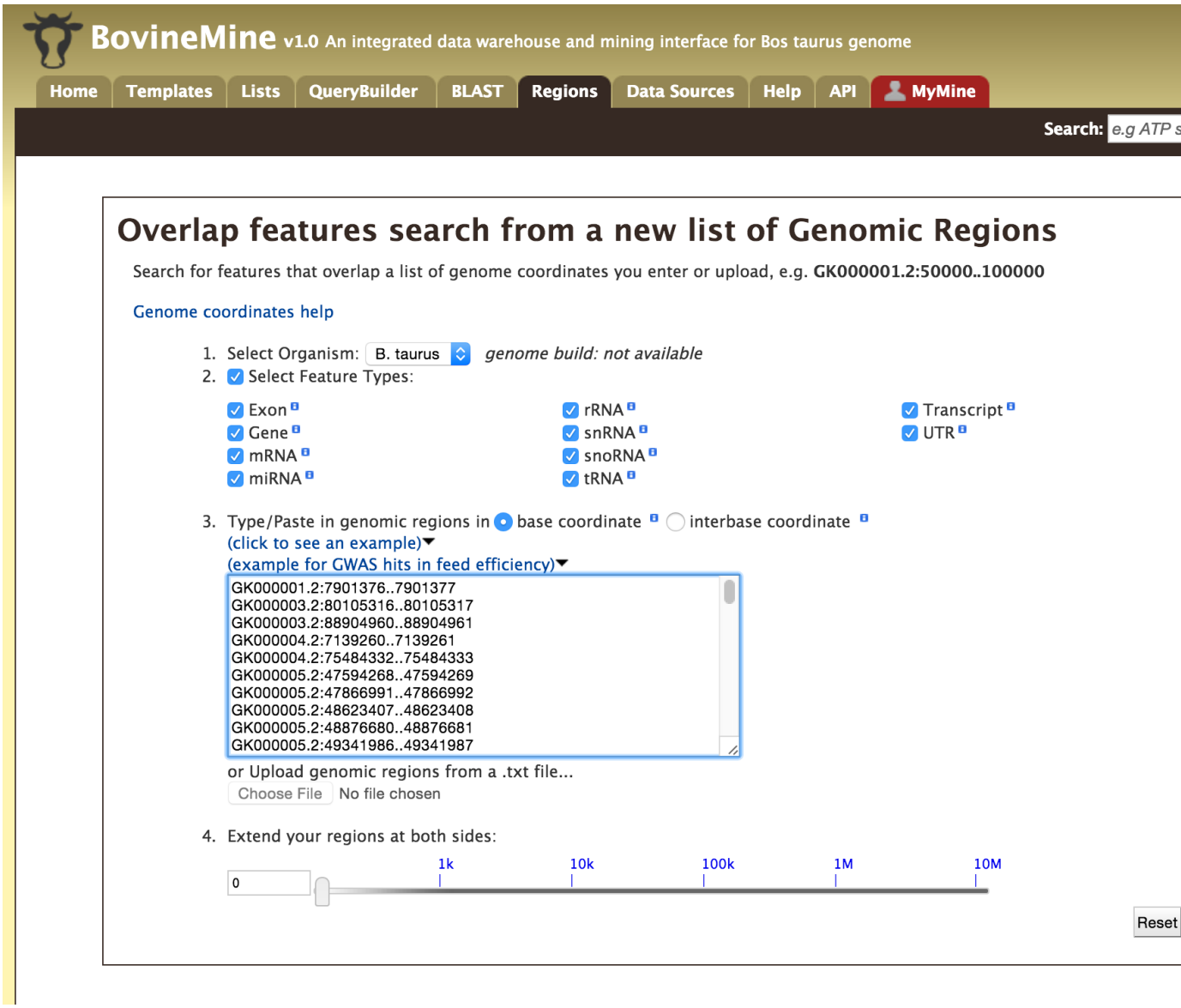
The Genomic Region Search is a tool to fetch features that are within a given set of genomic coordinates or to fetch features that are within a given number of bases flanking a given set of genomic coordinates.

The coordinates have to be of the format,

```
chromosome_name:start..end
```

```
chromosome_name:start-end
```

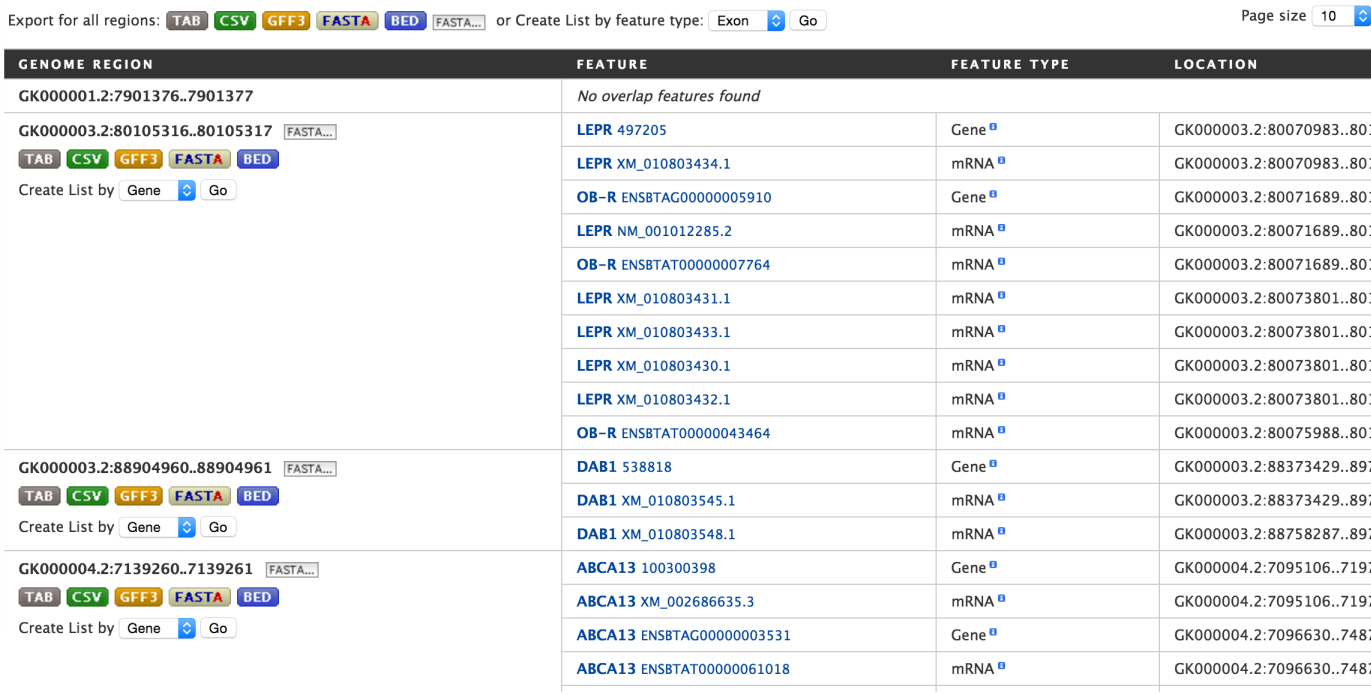
```
chromosome_name start end
```



Click on `click to see an example` for a representative set of genomic coordinates. Users can extend the regions on either side of a given coordinate using the slider or using the textbox.

Users can also select the type of coordinate system they would like to use: **base coordinate** or **interbase coordinate**.

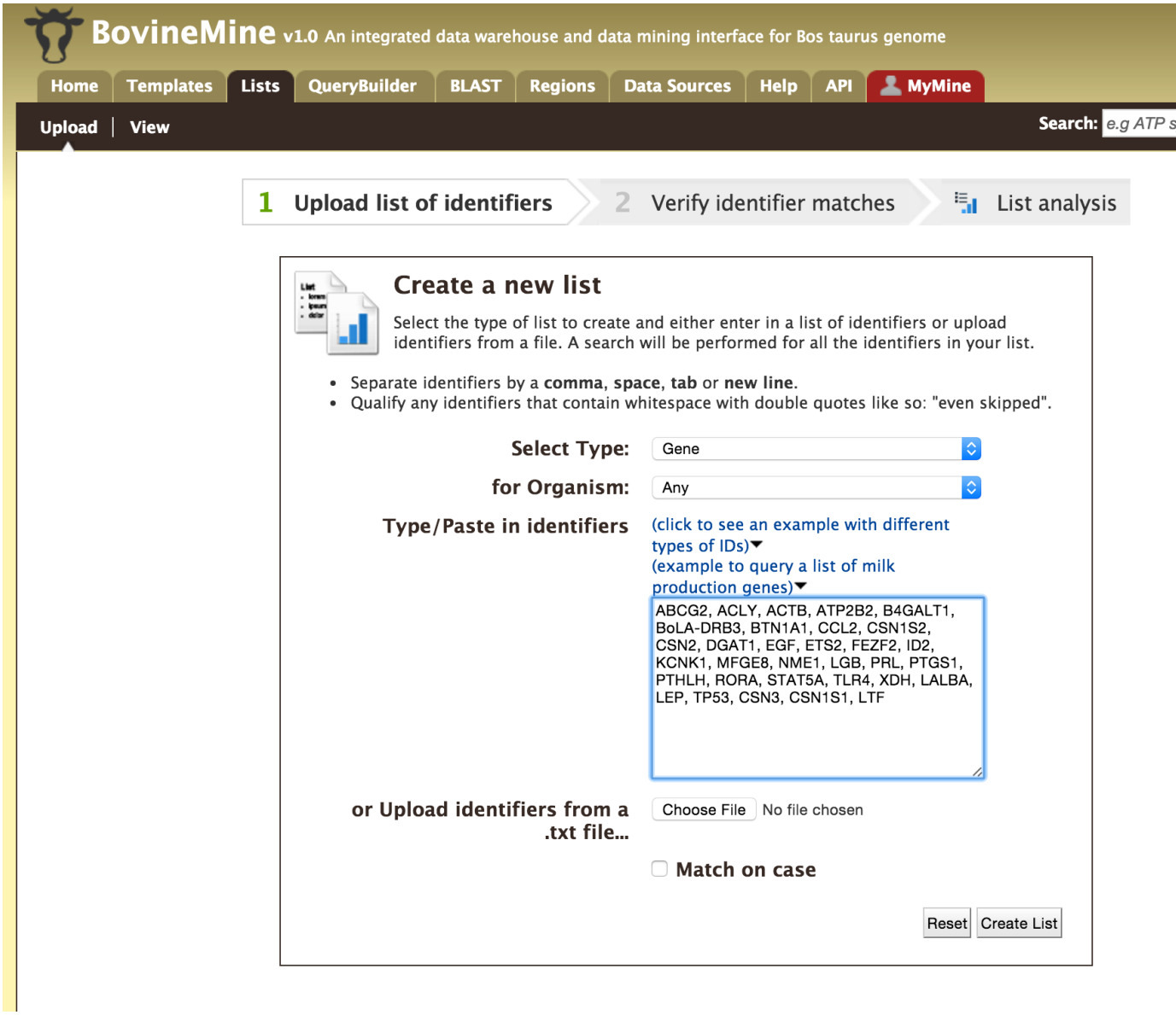
Lets try and example. Click on the `click to see an example` and extend the region search by 500 bp and click on `Search`. The result page will give a list of features that are present in each of the genomic interval provided in the input.



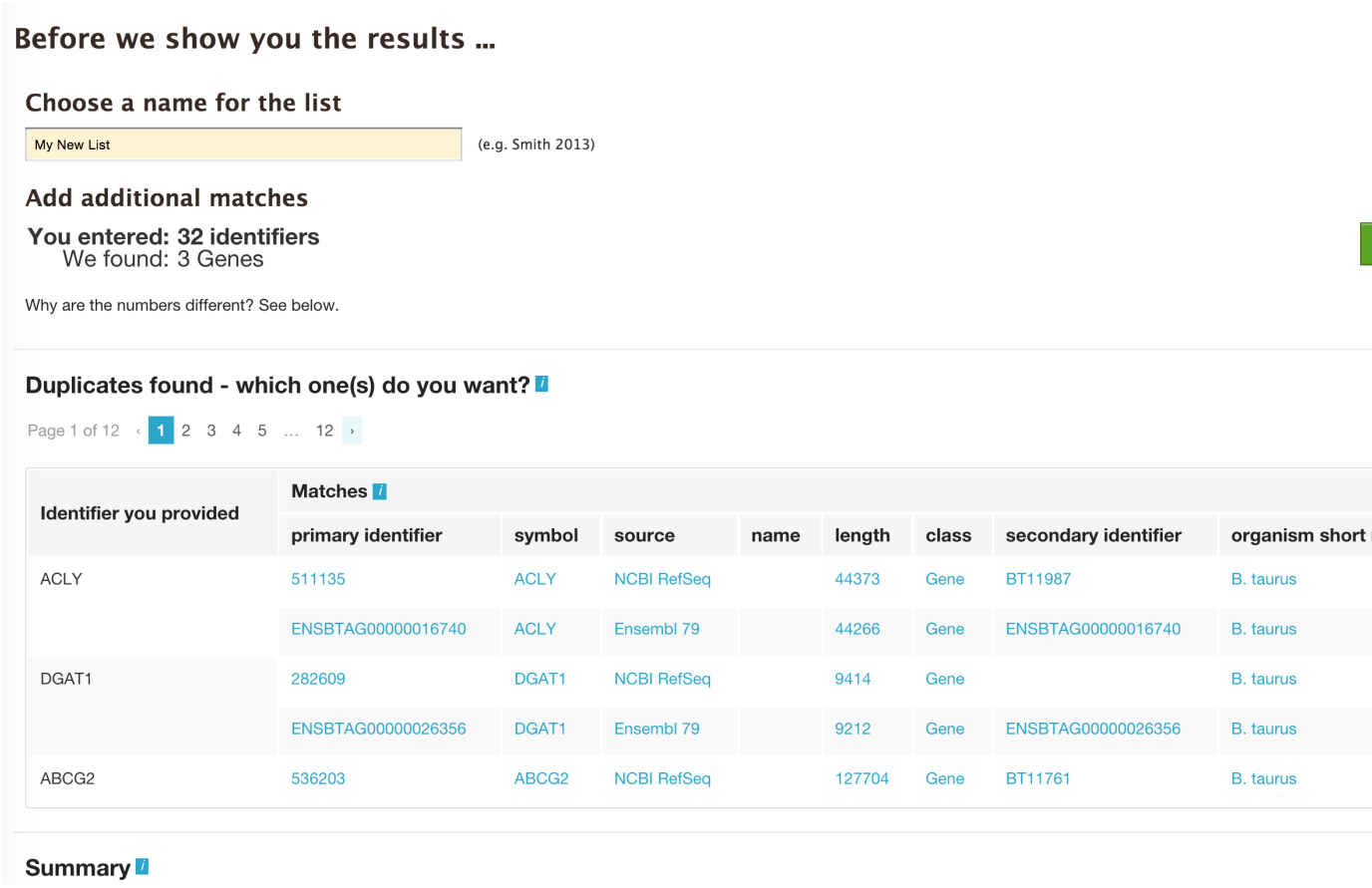
The results can be exported as tab-delimited and comma-separated values. If the results have genomic features then they can be exported in GFF3 or BED format. Users can also export FASTA sequences of the features. If users are interested in creating a list of particular features from the result page then they can filter based on feature type, shown in red box, and click on 'Go'.

1.6 Lists

Users can create a list of features. The input can either be gene IDs, transcript IDs, gene symbols, etc.



The list tool tries to lookup the query throughout the database and will attempt to convert the identifiers to the type selected in the list 'Select Type' option. Lets try the examples provided. Click on 'Click to see example' link (highlighted in red box) and click on 'Create List'.



Before we show you the results ...

Choose a name for the list

My New List (e.g. Smith 2013)

Add additional matches

You entered: 32 identifiers
We found: 3 Genes

Why are the numbers different? See below.

Duplicates found - which one(s) do you want?

Page 1 of 12 ‹ 1 2 3 4 5 … 12 ›

Identifier you provided	Matches							
	primary identifier	symbol	source	name	length	class	secondary identifier	organism short
ACLY	511135	ACLY	NCBI RefSeq		44373	Gene	BT11987	B. taurus
	ENSBTAG00000016740	ACLY	Ensembl 79		44266	Gene	ENSBTAG00000016740	B. taurus
DGAT1	282609	DGAT1	NCBI RefSeq		9414	Gene		B. taurus
	ENSBTAG00000026356	DGAT1	Ensembl 79		9212	Gene	ENSBTAG00000026356	B. taurus
ABCG2	536203	ABCG2	NCBI RefSeq		127704	Gene	BT11761	B. taurus

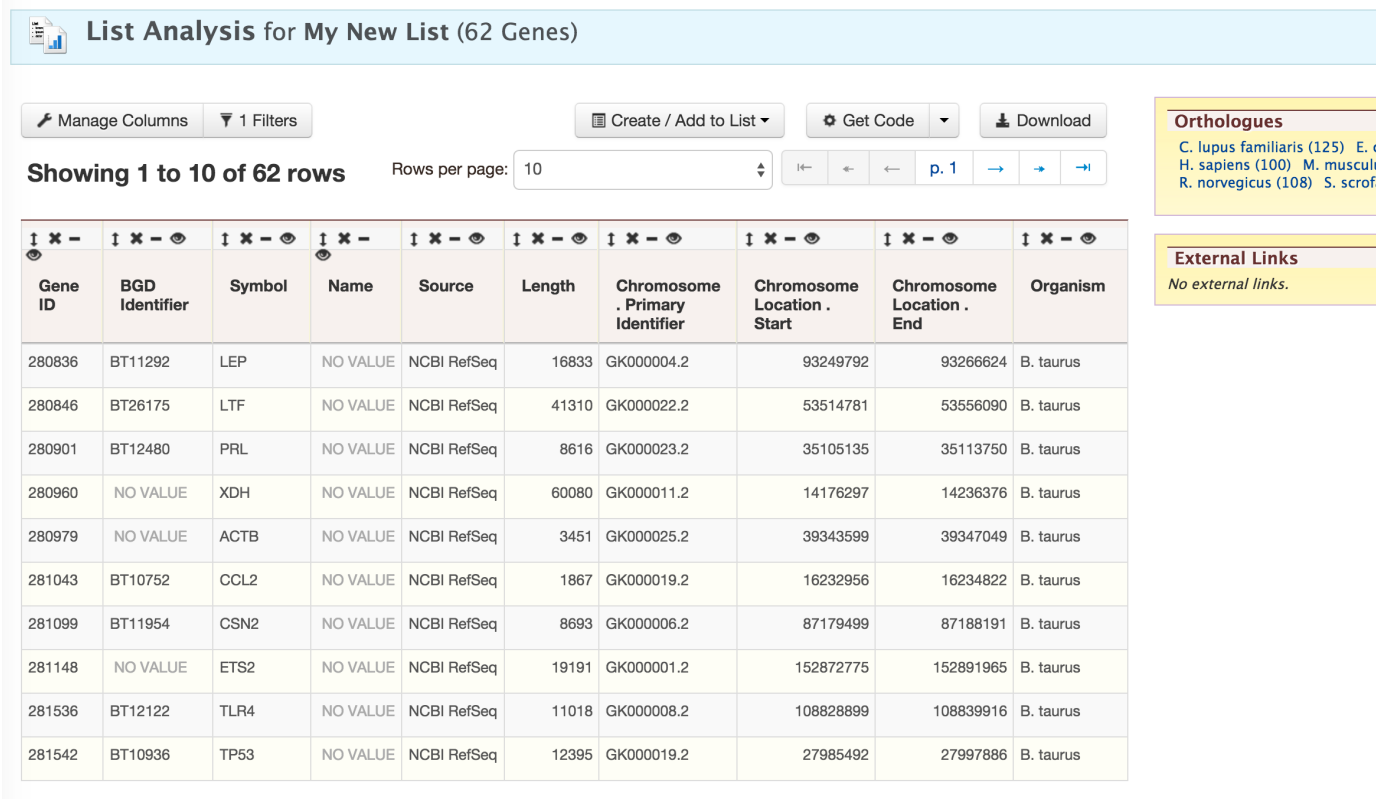
Summary

The list tool does a lookup of the identifiers and shows you the results. If there are any duplicates, users can decide to add the relevant entries individually.

The summary section provides information regarding those identifiers that had a direct hit without any duplicates.

Click on ‘Add all’ and then click on ‘Save a list of 62 genes’.

This will take users to a List Analysis page,



List Analysis for My New List (62 Genes)

Manage Columns | 1 Filters | Create / Add to List | Get Code | Download

Showing 1 to 10 of 62 rows Rows per page: 10 p. 1

Gene ID	BGD Identifier	Symbol	Name	Source	Length	Chromosome . Primary Identifier	Chromosome Location . Start	Chromosome Location . End	Organism
280836	BT11292	LEP	NO VALUE	NCBI RefSeq	16833	GK000004.2	93249792	93266624	B. taurus
280846	BT26175	LTF	NO VALUE	NCBI RefSeq	41310	GK000022.2	53514781	53556090	B. taurus
280901	BT12480	PRL	NO VALUE	NCBI RefSeq	8616	GK000023.2	35105135	35113750	B. taurus
280960	NO VALUE	XDH	NO VALUE	NCBI RefSeq	60080	GK000011.2	14176297	14236376	B. taurus
280979	NO VALUE	ACTB	NO VALUE	NCBI RefSeq	3451	GK000025.2	39343599	39347049	B. taurus
281043	BT10752	CCL2	NO VALUE	NCBI RefSeq	1867	GK000019.2	16232956	16234822	B. taurus
281099	BT11954	CSN2	NO VALUE	NCBI RefSeq	8693	GK000006.2	87179499	87188191	B. taurus
281148	NO VALUE	ETS2	NO VALUE	NCBI RefSeq	19191	GK000001.2	152872775	152891965	B. taurus
281536	BT12122	TLR4	NO VALUE	NCBI RefSeq	11018	GK000008.2	108828899	108839916	B. taurus
281542	BT10936	TP53	NO VALUE	NCBI RefSeq	12395	GK000019.2	27985492	27997886	B. taurus

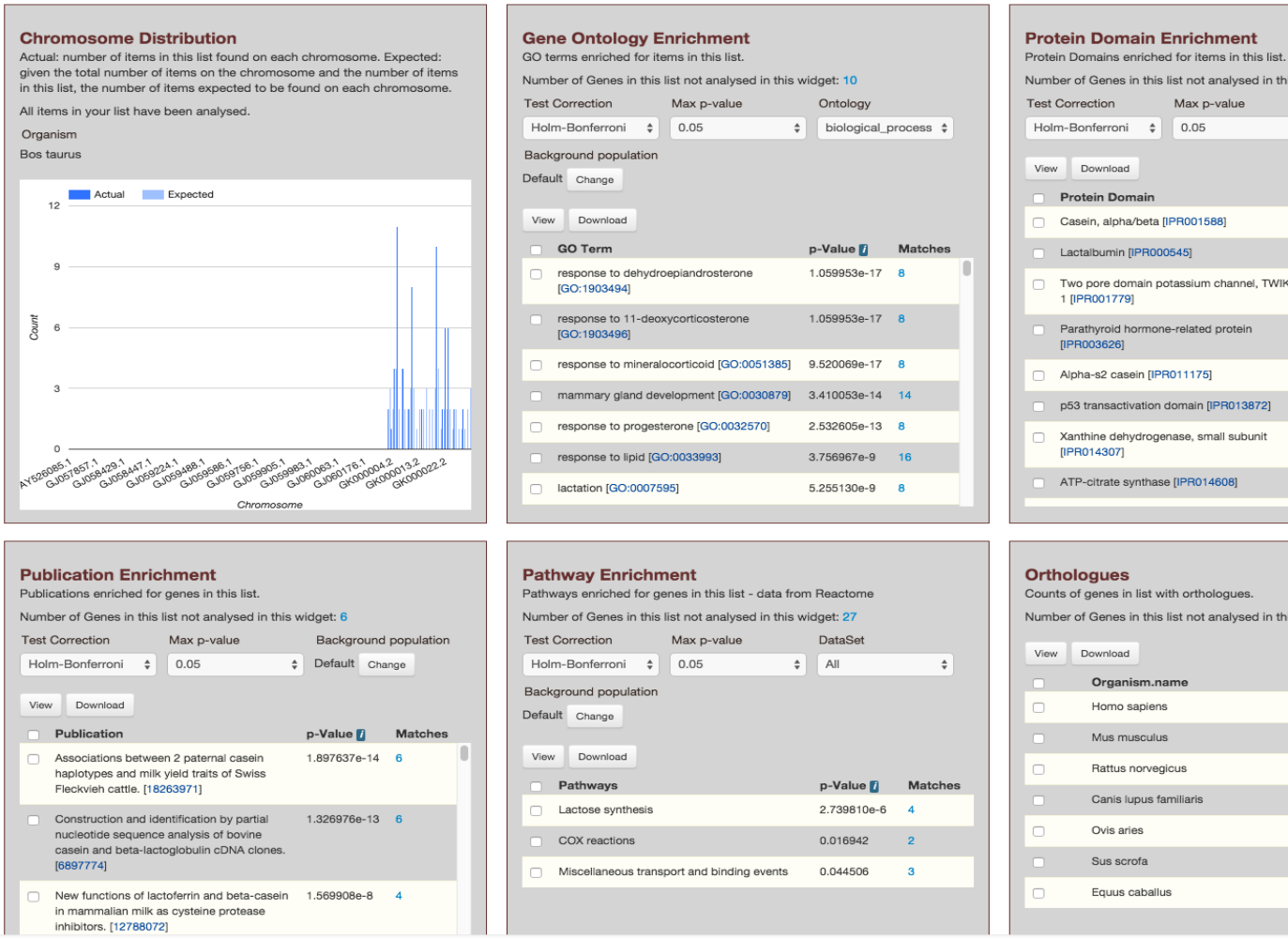
Orthologues

C. lupus familiaris (125) E.
H. sapiens (100) M. musculi
R. norvegicus (108) S. scrof

External Links

No external links.

This page provides users with widgets to perform analyses on gene lists that they have created.



Currently available widgets are,

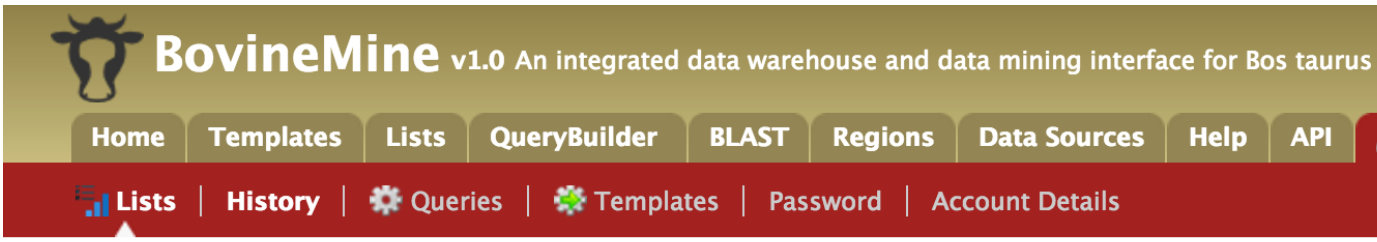
1. Chromosome Distribution
2. Gene Ontology Enrichment
3. Protein Domain Enrichment
4. Publication Enrichment
5. Pathway Enrichment
6. Orthologues

1.7 MyMine

Users can create an account on BovineMine which enables them to store their lists, custom templates and keep track of their sessions. To create a login, click on the 'Log in' on the top-right corner of any page on BovineMine.

New users can create a new account with the username being their Email ID.

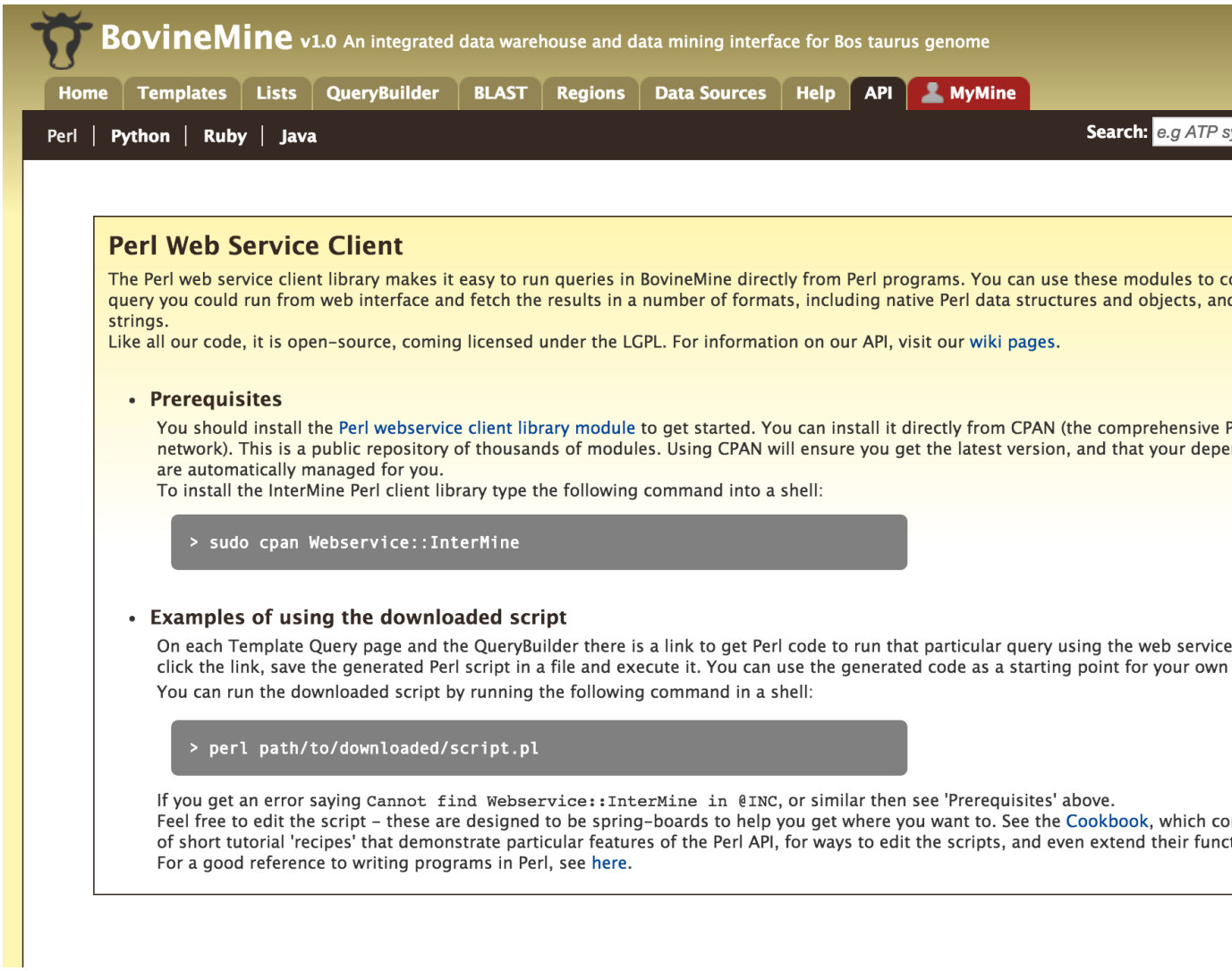
BovineMine provides MyMine for users to manage their lists, queries, templates, and account details.



1. Lists – lists saved by the user.
2. History of queries by user – shows a list of most recent queries performed by the user.
3. Templates – Templates created by the user or existing templates that are marked as 'favorite' by the user.
4. Password change – change the password for the user's account.
5. Account details – for updating user preferences.

1.8 API

For users who would like to programmatically access BovineMine they can make use of the InterMine API.



Perl, Python, Ruby and Java are the 4 languages in which the InterMine API is available. For more information on the details on the API visit the InterMine Wiki.

1.9 Data Sources

Provides a description of the datasets that are integrated into BovineMine along with the location from which these datasets were downloaded, their version or release, citations wherever applicable and any additional comments.

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

CHAPTER 2

Indices and tables

- genindex
- modindex
- search