New Data Mining Interfaces at the Bovine Genome Database

Colin Diesh[1], Aditi Tayal[1], Deepak Unni[1], Darren Hagen[1], Christine G. Elsik[1]

[1] *Division of Animal Sciences, University of Missouri, Columbia, MO 65211, USA*

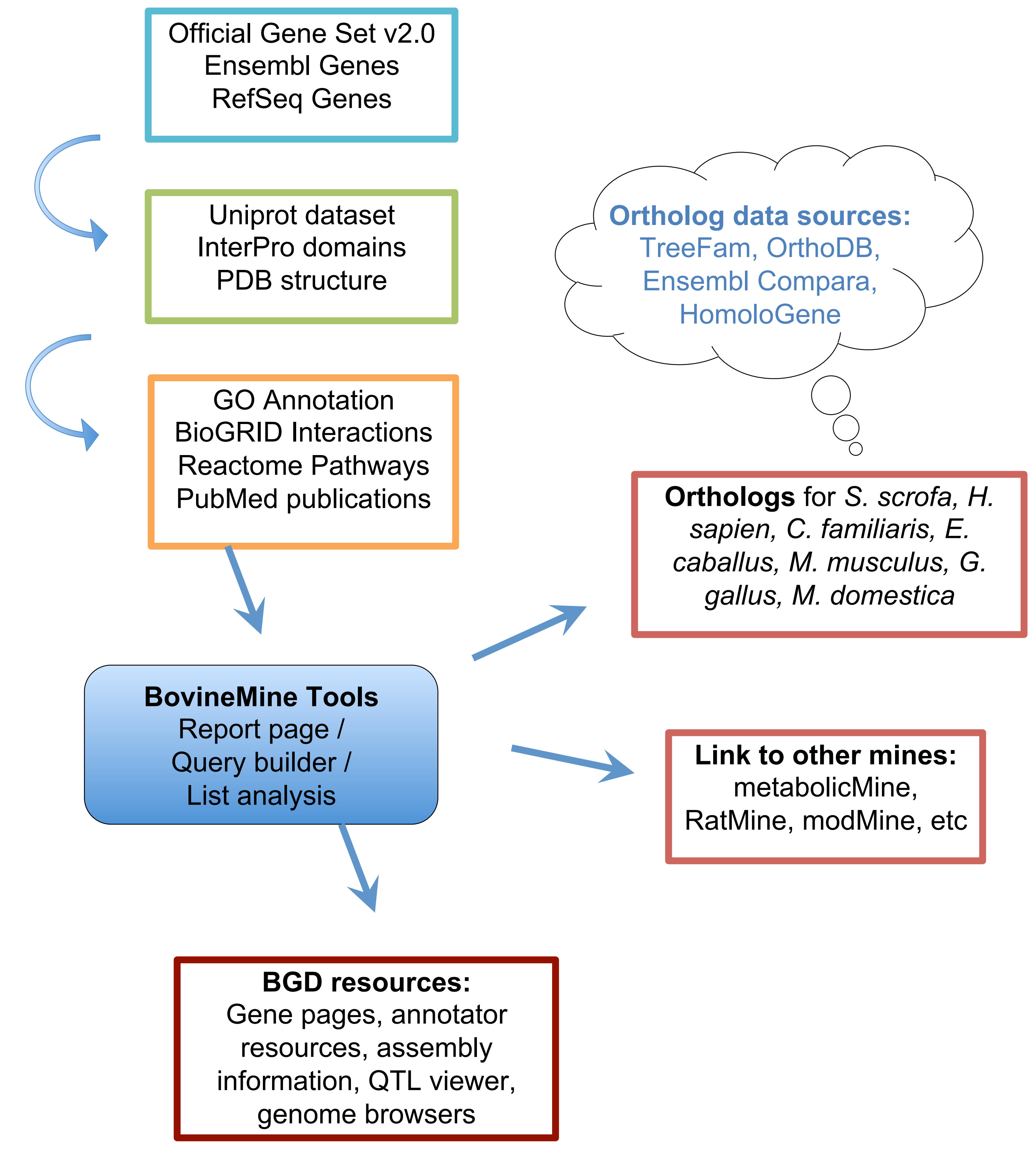
United States Department of Agriculture
National Institute of Food and Agriculture

Abstract

The Bovine Genome Database (BGD, http://BovineGenome.org) is an informatics resource for Bos taurus. We have used InterMine to deploy a new data warehouse called BovineMine to provide a fast and flexible query interface. BovineMine integrates information from many data sources including RefSeq, Ensembl, UniProt, InterPro, OrthoDB, Homologene, Pubmed, Gene Ontology and BioGRID. BovineMine also includes data that we provide on our genome browsers, such as the Bovine Official Gene Set, RNAseq data, SNP and QTL. Users may perform a “Quick Search” or use the “Query Builder” for specialized searches. The “Genome Region Search” and “Overlapping Feature Search” allow users to download annotations with a specified genomic context. Users may download query results in various formats, such as tab-delimited files, GFF, Fasta, BED, JSON and XML. In addition to BovineMine, users may access data through genome browsers (GBrowse and JBrowse) and .

We have also created specialized search interfaces that focus on differences in annotations and assembly versions, and provide easy navigation to novel genes on genome browsers. The Annotation Assembly Comparison Tool allows users to lookup locations of genes on two bovine assemblies (UMD3.1 and Btau_4.6.1). The Ensembl-NCBI Comparison Tool allows users to investigate disagreements in gene models across gene sets. The Predicted Transcript RNAseq Read Count Tool provides counts of spliced RNAseq read alignments to predicted transcripts to aid selection of RNAseq tracks in JBrowse. The Candidate Novel Protein Coding Gene Search Tool allows users to easily navigate to locations candidate novel genes in JBrowse.

BovineMine Datasets


Official Gene Set v2.0
Ensembl Genes
RefSeq Genes

Uniprot dataset
InterPro domains
PDB structure

GO Annotation
BioGRID Interactions
Reactome Pathways
PubMed publications

Ortholog data sources: TreeFam, OrthoDB, Ensembl Compara, HomoloGene

Orthologs for *S. scrofa, H. sapien, C. familiaris, E. caballus, M. musculus, G. gallus, M. domestica*

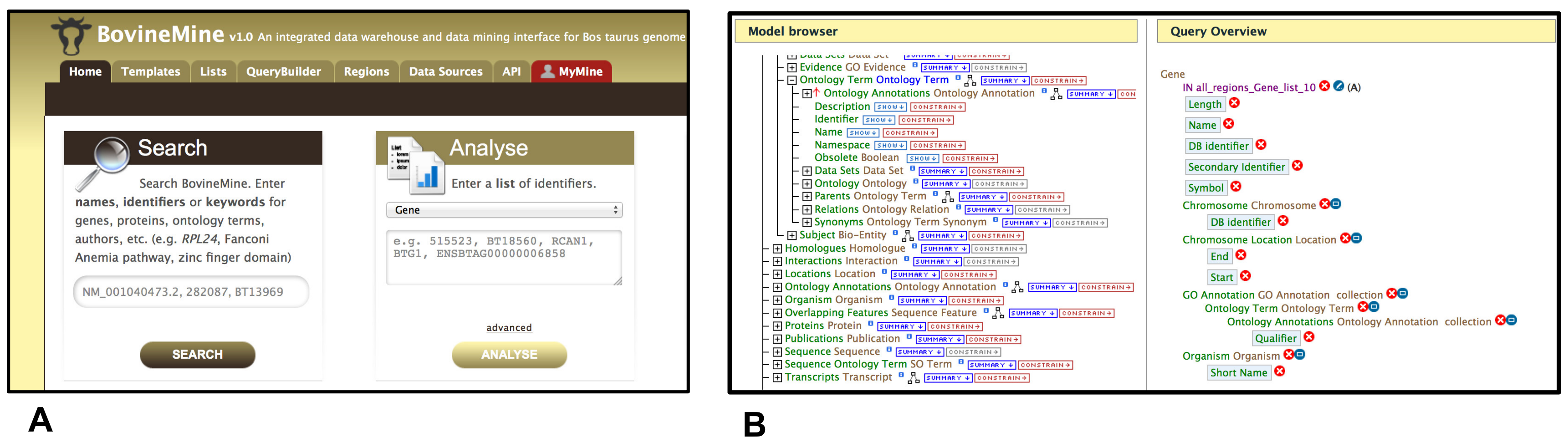
BovineMine Tools
Report page /
Query builder /
List analysis

Link to other mines: metabolicMine, RatMine, modMine, etc

BGD resources: Gene pages, annotator resources, assembly information, QTL viewer, genome browsers

Figure 1. Integration of Data and Tools in BovineMine

BovineMine Search Tools


BovineMine v1.0 An integrated data warehouse and data mining interface for Bos taurus genome

Search

Search BovineMine. Enter **names, identifiers** or **keywords** for genes, proteins, ontology terms, authors, etc. (e.g. *RPL24*, Fanconi Anemia pathway, zinc finger domain)

NM_001040473.2, 282087, BT13969

SEARCH

Analyse

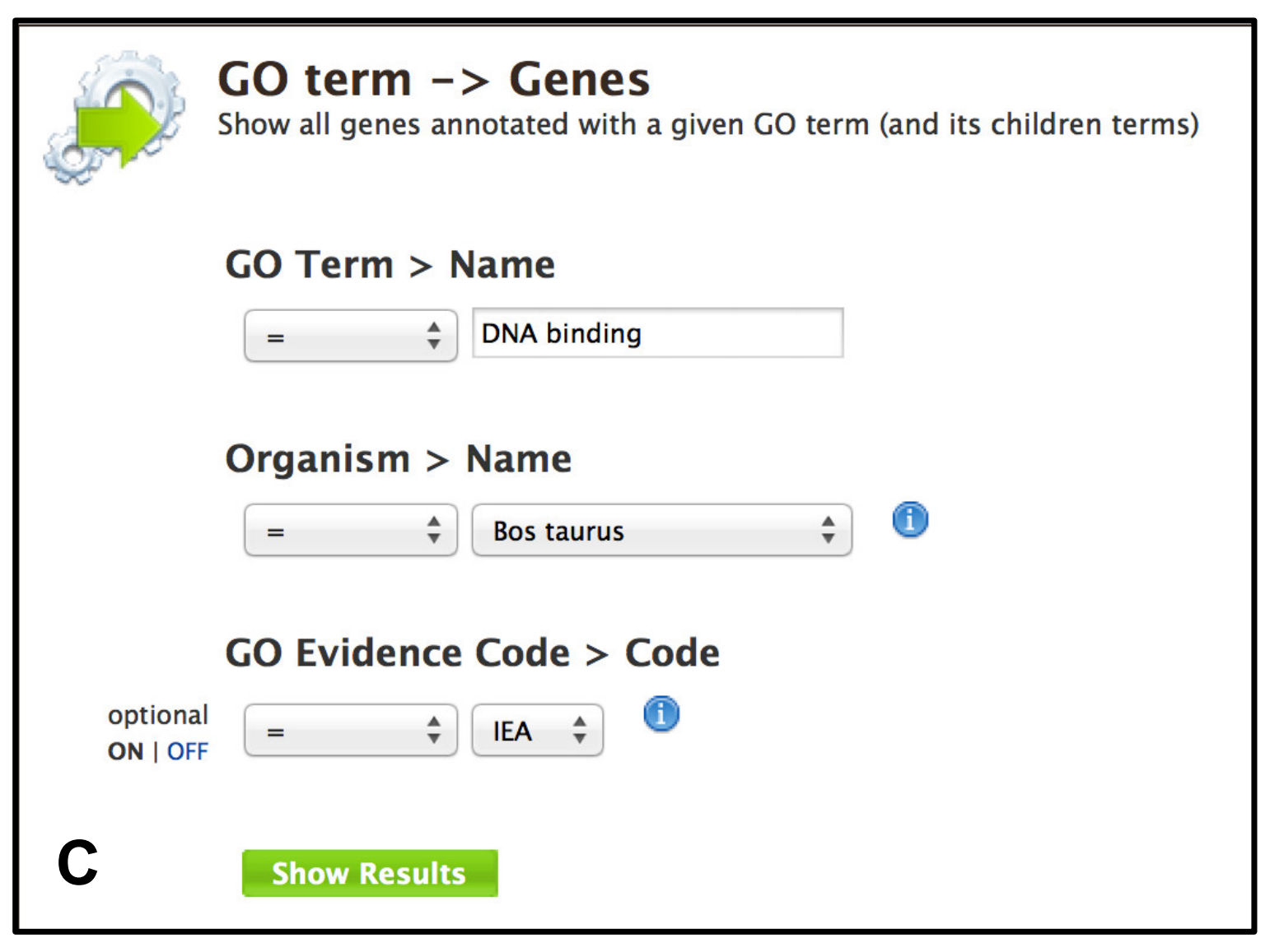
Enter a **list** of identifiers.

Gene

e.g. 515523, BT18560, RCAN1, BTG1, ENSBTAG00000006858

advanced

ANALYSE


GO term -> Genes
Show all genes annotated with a given GO term (and its children terms)

GO Term > Name
= DNA binding

Organism > Name
= Bos taurus

GO Evidence Code > Code
optional ON | OFF
= IEA

Show Results

Figure 2. Search Tools

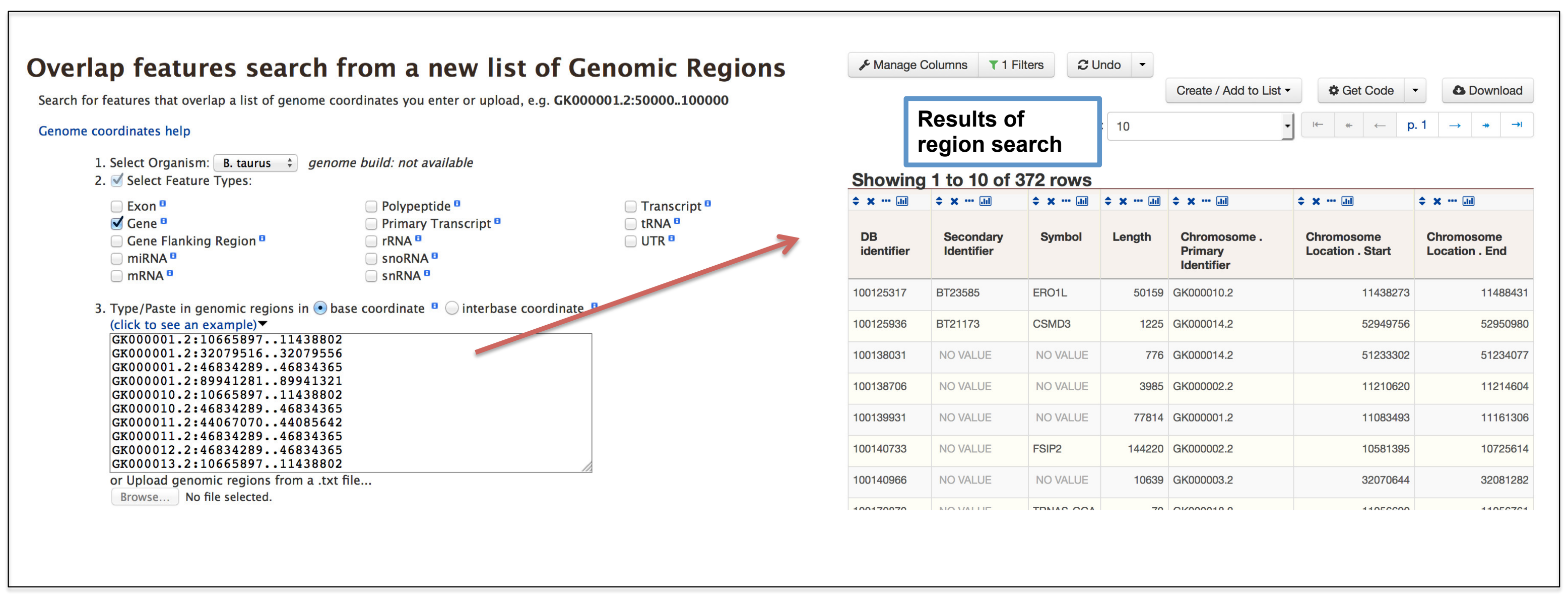
A. The **Quick Search** tool can be used to find keywords from multiple datasets at once. With the **List Analyse** tool, lists of IDs can be saved and analyzed using multiple queries.

B. The **Query Builder** is a powerful GUI for making complex structured queries.

C. Template queries are prepared queries to accomplish common tasks (genes matching GO term, homologues of given gene, etc).

Data Mining Example

The following is an example using QTL regions that are associated to milk fat yield. To begin, we use the “Region search” and paste regions corresponding to QTL locations on base pair coordinates in the format “chr:start..end”:


Overlap features search from a new list of Genomic Regions

Search for features that overlap a list of genome coordinates you enter or upload, e.g. **GK000001.2:50000..100000**

Genome coordinates help

1. Select Organism: B. taurus *genome build: not available*
2. ☑ Select Feature Types:

☐ Exon
☑ Gene
☐ Gene Flanking Region
☐ miRNA
☐ mRNA
☐ Polypeptide
☐ Primary Transcript
☐ rRNA
☐ snoRNA
☐ snRNA
☐ Transcript
☐ tRNA
☐ UTR

3. Type/Paste in genomic regions in ● base coordinate ○ interbase coordinate
(click to see an example)

```
GK000001.2:10665897..11438802
GK000001.2:32079516..32079556
GK000001.2:46834289..46834365
GK000001.2:89941281..89941321
GK000010.2:10665897..11438802
GK000010.2:46834289..46834365
GK000011.2:44067070..44085642
GK000011.2:46834289..46834365
GK000012.2:46834289..46834365
GK000013.2:10665897..11438802
```

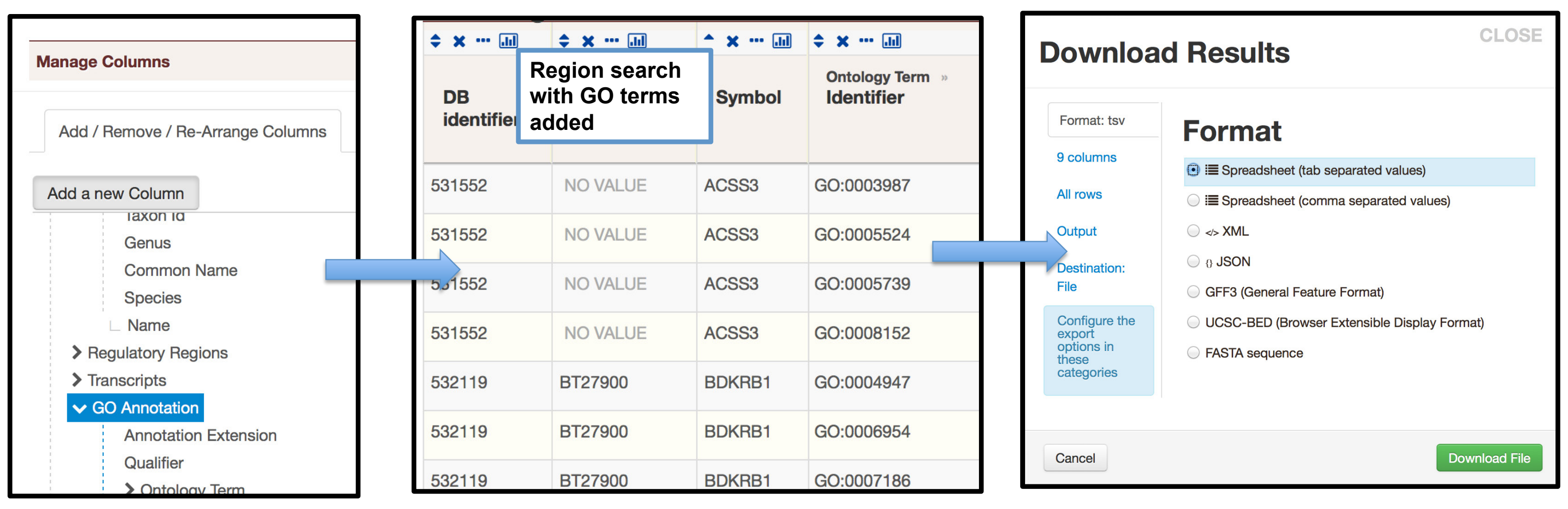
or Upload genomic regions from a .txt file...
Browse... No file selected.

Showing 1 to 10 of 372 rows

DB identifier	Secondary Identifier	Symbol	Length	Chromosome . Primary Identifier	Chromosome Location . Start	Chromosome Location . End
100125317	BT23585	ERO1L	50159	GK000010.2	11438273	11488431
100125936	BT21173	CSMD3	1225	GK000014.2	52949756	52950980
100138031	NO VALUE	NO VALUE	776	GK000014.2	51233302	51234077
100138706	NO VALUE	NO VALUE	3985	GK000002.2	11210620	11214604
100139931	NO VALUE	NO VALUE	77814	GK000001.2	11083493	11161306
100140733	NO VALUE	FSIP2	144220	GK000002.2	10581395	10725614
100140966	NO VALUE	NO VALUE	10639	GK000003.2	32070644	32081282

Figure 3. (left) BovineMine’s Region search page and (right) results page containing genes that overlap the regions of interest.

The results from the region search can then be augmented with GO Annotations.


Manage Columns

Add / Remove / Re-Arrange Columns

Region search with GO terms added

DB identifier		Symbol	Ontology Term Identifier
531552	NO VALUE	ACSS3	GO:0003987
531552	NO VALUE	ACSS3	GO:0005524
531552	NO VALUE	ACSS3	GO:0005739
531552	NO VALUE	ACSS3	GO:0008152
532119	BT27900	BDKRB1	GO:0004947
532119	BT27900	BDKRB1	GO:0006954
532119	BT27900	BDKRB1	GO:0007186

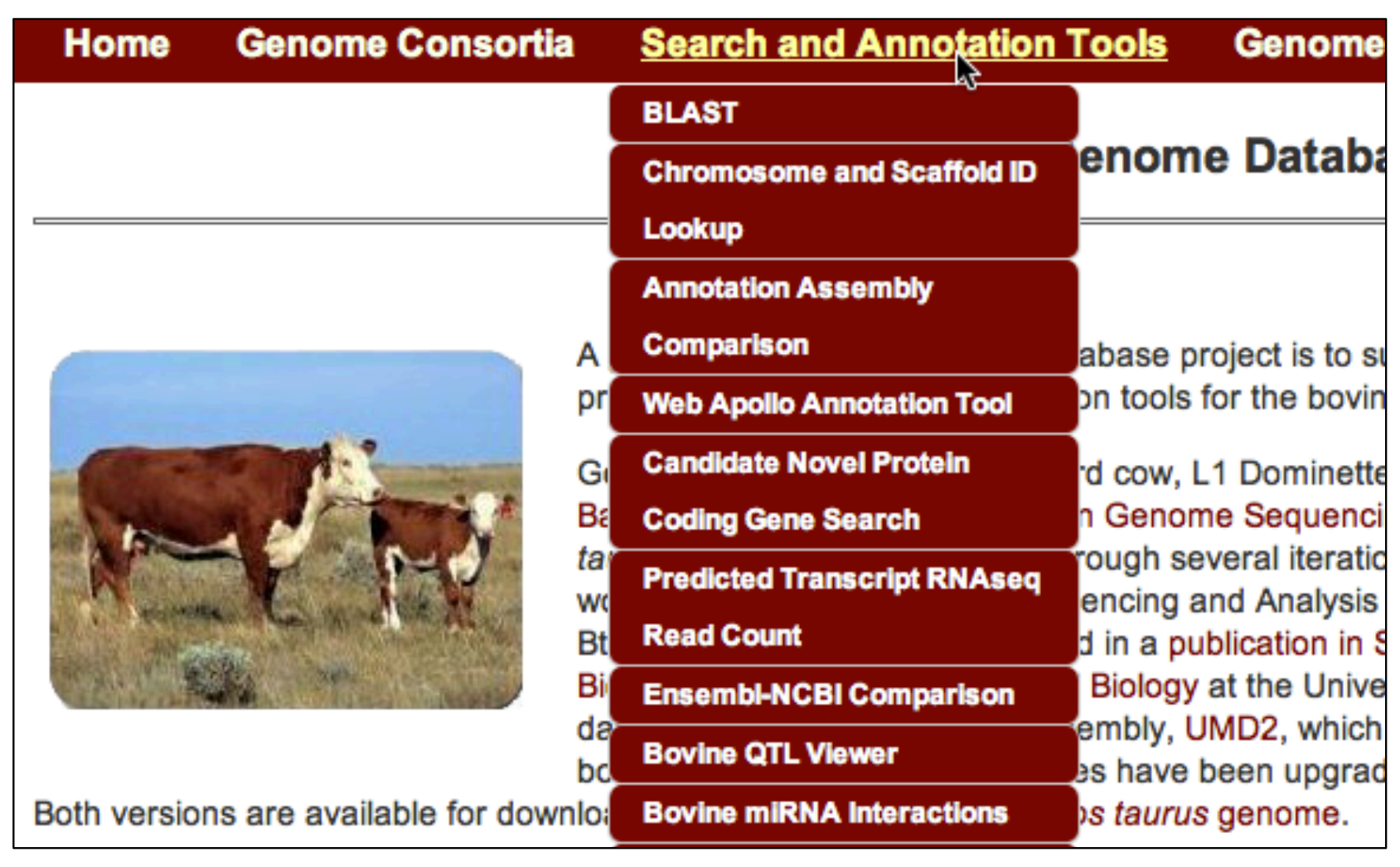
Download Results

Format

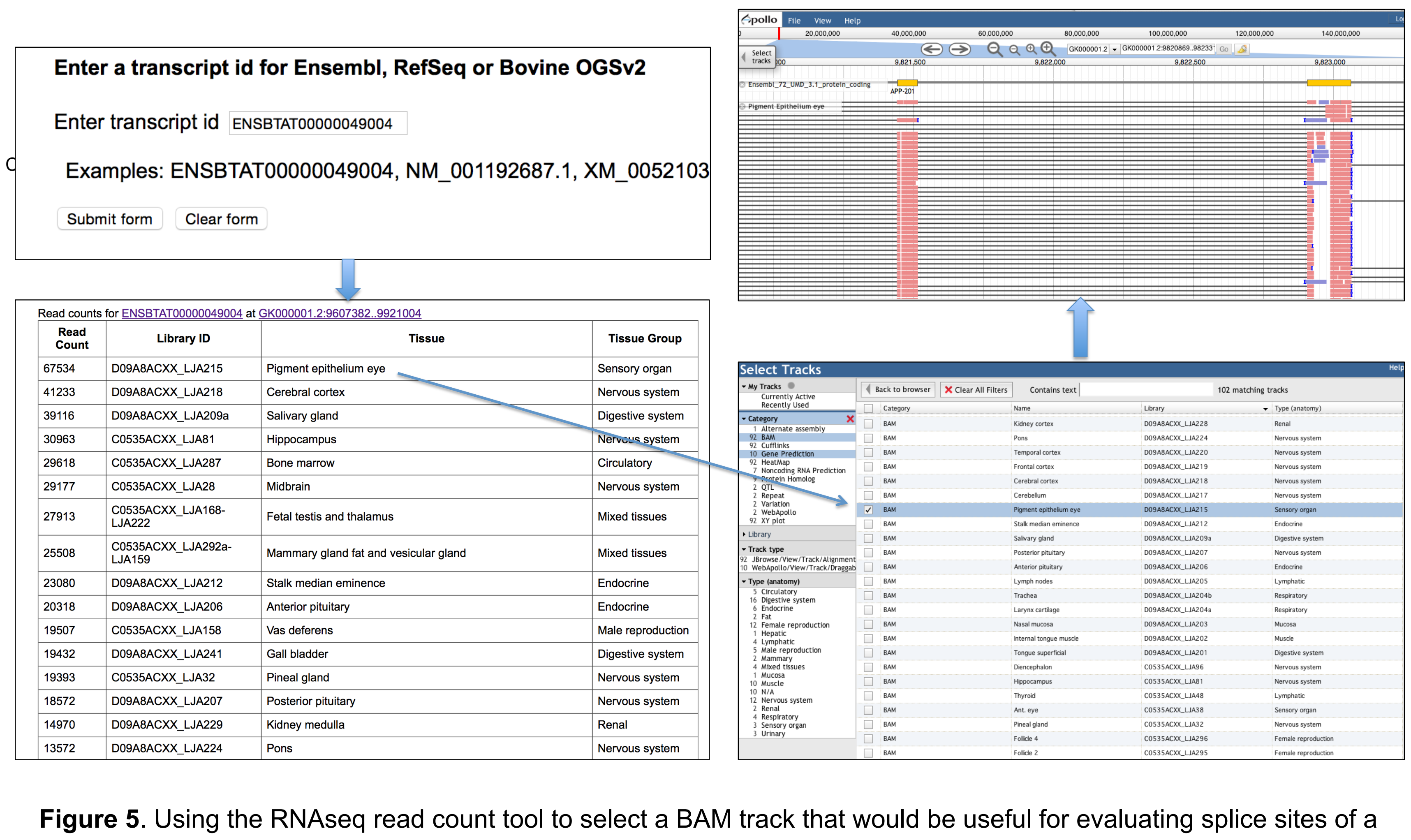
● Spreadsheet (tab separated values)
○ Spreadsheet (comma separated values)
○ XML
○ JSON
○ GFF3 (General Feature Format)
○ UCSC-BED (Browser Extensible Display Format)
○ FASTA sequence

Figure 4. (left) The “GO Annotation” attribute is added to the results using “Manage Columns” tool. (middle) The results of region search with GO terms is shown. (right) The table download page for the region search is used to download results to plain text.

New Tools for Annotators at BGD


Home
Genome Consortia
Search and Annotation Tools
Genome
BLAST
Chromosome and Scaffold ID Lookup
Annotation Assembly Comparison
Web Apollo Annotation Tool
Candidate Novel Protein Coding Gene Search
Predicted Transcript RNAseq Read Count
Ensembl-NCBI Comparison
Bovine QTL Viewer
Bovine miRNA Interactions

New tools help annotators navigate across alternate bovine assemblies, convert identifiers, compare gene sets and select tracks to view in the browser. For example, the RNAseq Read Count tool provides raw read counts from spliced RNAseq alignments to locations of predicted transcripts (Ensembl, RefSeq and Bovine OGSv2). The read counts help users determine which of 91 RNAseq data sets to view in Jbrowse/Web Apollo. The spliced read alignments (BAM tracks) are particularly useful in determining whether splice junctions in predicted genes are correct.


Enter a transcript id for Ensembl, RefSeq or Bovine OGSv2

Enter transcript id ENSBTAT00000049004

Examples: ENSBTAT00000049004, NM_001192687.1, XM_0052103

Submit form Clear form

Read counts for ENSBTAT00000049004 at GK000001.2:9607382..9921004

Read Count	Library ID	Tissue	Tissue Group
67534	D09A8ACXX_LJA215	Pigment epithelium eye	Sensory organ
41233	D09A8ACXX_LJA218	Cerebral cortex	Nervous system
39116	D09A8ACXX_LJA209a	Salivary gland	Digestive system
30963	C0535ACXX_LJA81	Hippocampus	Nervous system
29618	C0535ACXX_LJA287	Bone marrow	Circulatory
29177	C0535ACXX_LJA28	Midbrain	Nervous system
27913	C0535ACXX_LJA168-LJA222	Fetal testis and thalamus	Mixed tissues
25508	C0535ACXX_LJA292a-LJA159	Mammary gland fat and vesicular gland	Mixed tissues
23080	D09A8ACXX_LJA212	Stalk median eminence	Endocrine
20318	D09A8ACXX_LJA206	Anterior pituitary	Endocrine
19507	C0535ACXX_LJA158	Vas deferens	Male reproduction
19432	D09A8ACXX_LJA241	Gall bladder	Digestive system
19393	C0535ACXX_LJA32	Pineal gland	Nervous system
18572	D09A8ACXX_LJA207	Posterior pituitary	Nervous system
14970	D09A8ACXX_LJA229	Kidney medulla	Renal
13572	D09A8ACXX_LJA224	Pons	Nervous system

Figure 5. Using the RNAseq read count tool to select a BAM track that would be useful for evaluating splice sites of a particular gene model.

Acknowledgement

This work was supported by Agriculture and Food Research Initiative Competitive grant no. 2010-65205-20647 from the USDA National Institute of Food and Agriculture.