Gene identifier matching to join publicly available databases for the generation of a Mammalian Ortholog and Annotation Database with access from Galaxy-server

Jochen Bick[1], Mark Robinson[2], Susanne E. Ulbrich[1] and Stefan Bauersachs[1]

[1]**Animal Physiology, ETH Zurich;** [2]Institute of Molecular Life Sciences, University Zurich (UZH)

Introduction

So far there is a number of well-organized databases that contain useful information regarding orthologous genes, e.g., EnsemblCompara ortholog database. The main problem when using information derived from different databases is to correctly assign different gene, transcript or protein identifiers. However, because NCBI annotation is for most species the most comprehensive, we need to map information from other databases to EntrezGene IDs. This is an important issue for the generation of a Mammalian Ortholog and Annotation Database (MOA-Db) which will be partially based on information from publicly available databases, which needs to be collected, analyzed, and connected. Since each public source database uses own unique identifiers, it is necessary to assign the corresponding database-specific identifiers. Existing lists that assign corresponding genes, e.g. between Ensembl and EntrezGene are incomplete and/or contain errors. R BioConductor packages were used to find overlapping gene and exon positions which were integrated as a lookup table into the MySQL database to handle the comparison of different database sources. Finally, this database will be integrated into our local Galaxy-server to give easy access to all our research groups and provide a useful interface with various options to parse information via SQL queries. The MOA-Db provides a basis for optimal across-species comparisons of transcriptome datasets from different mammalian species accessible within a Galaxy-server.

AnnoOverlappeR

Three steps overlap algorithm:

1. Overlap on gene level

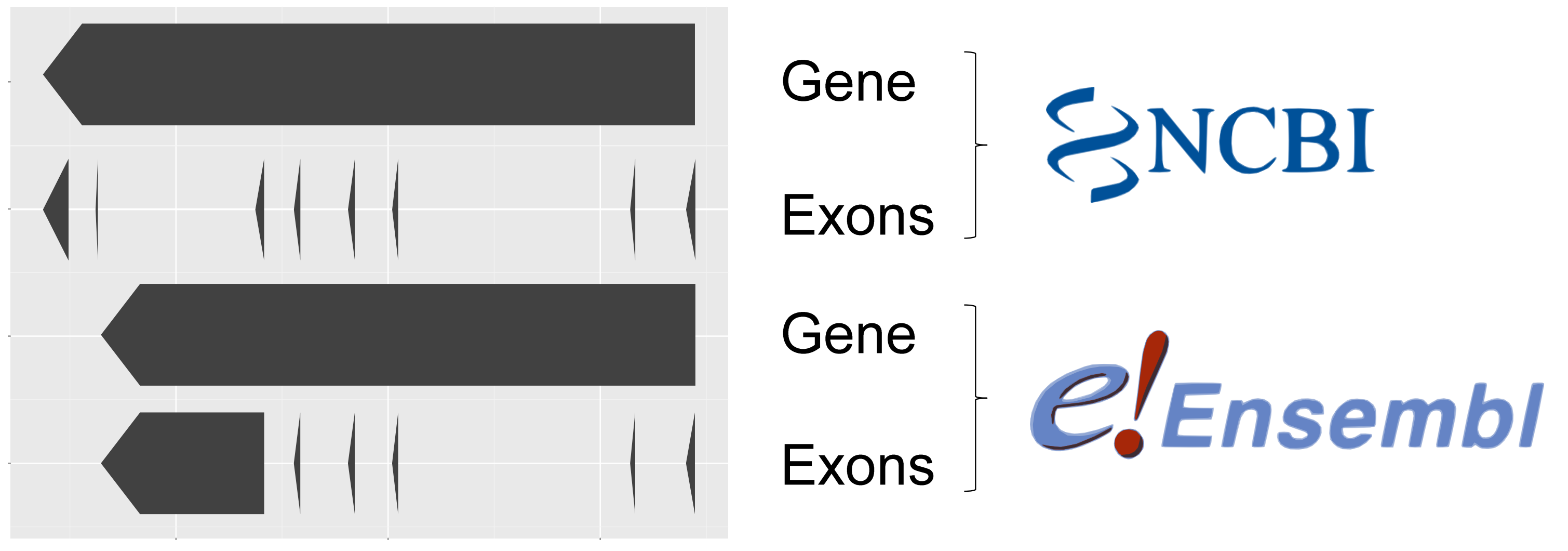
- Position-based overlap (chromosome, start, end, strand)
- Catch differently annotated genes (multiple hits)

2. Overlap on corresponding splice sites

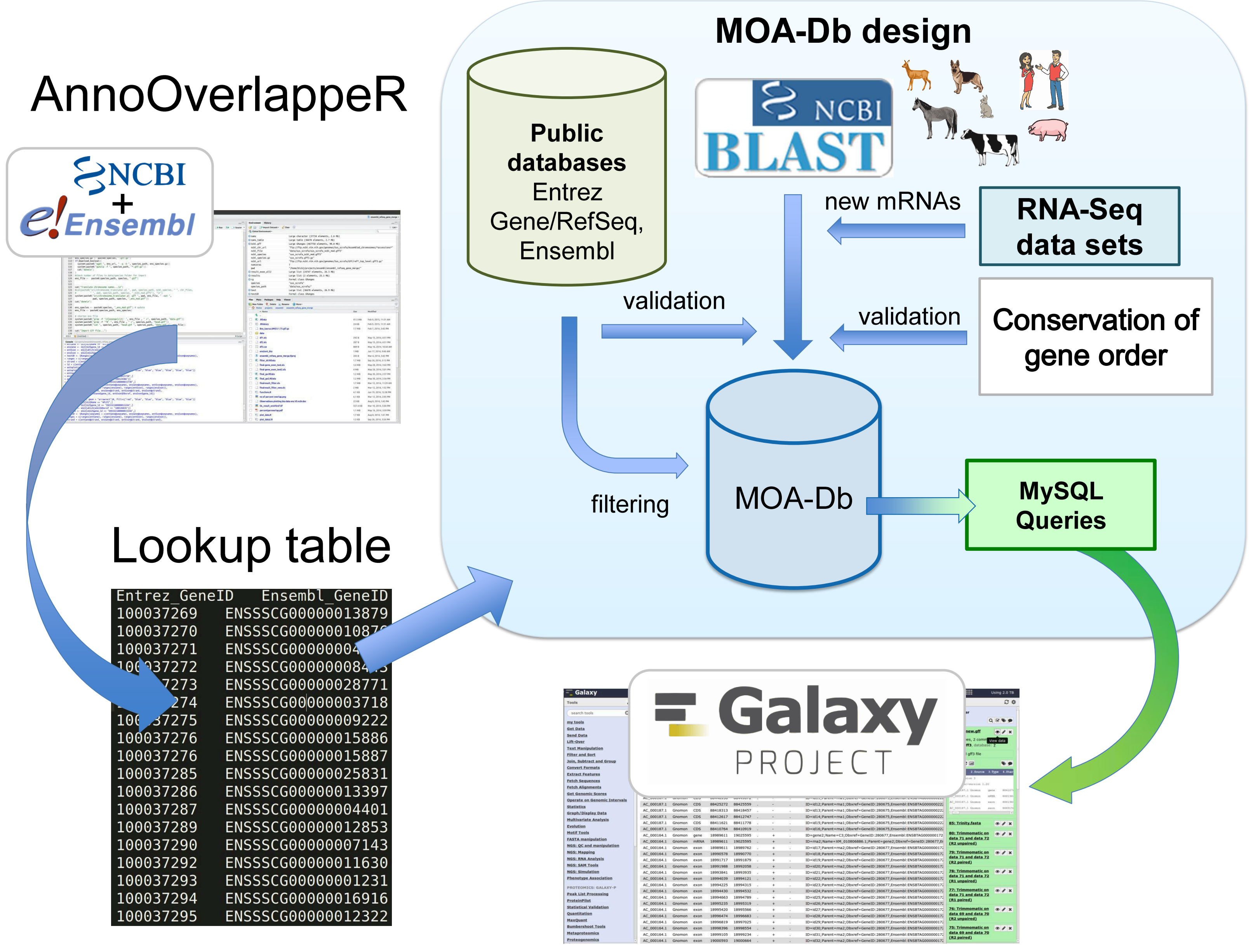
- Cutoff > 50 % overlap
- If < 50% ➡ check exon overlap, if > 50% keep it

3. Handling of multiple hits

- One gene maps to multiple genes



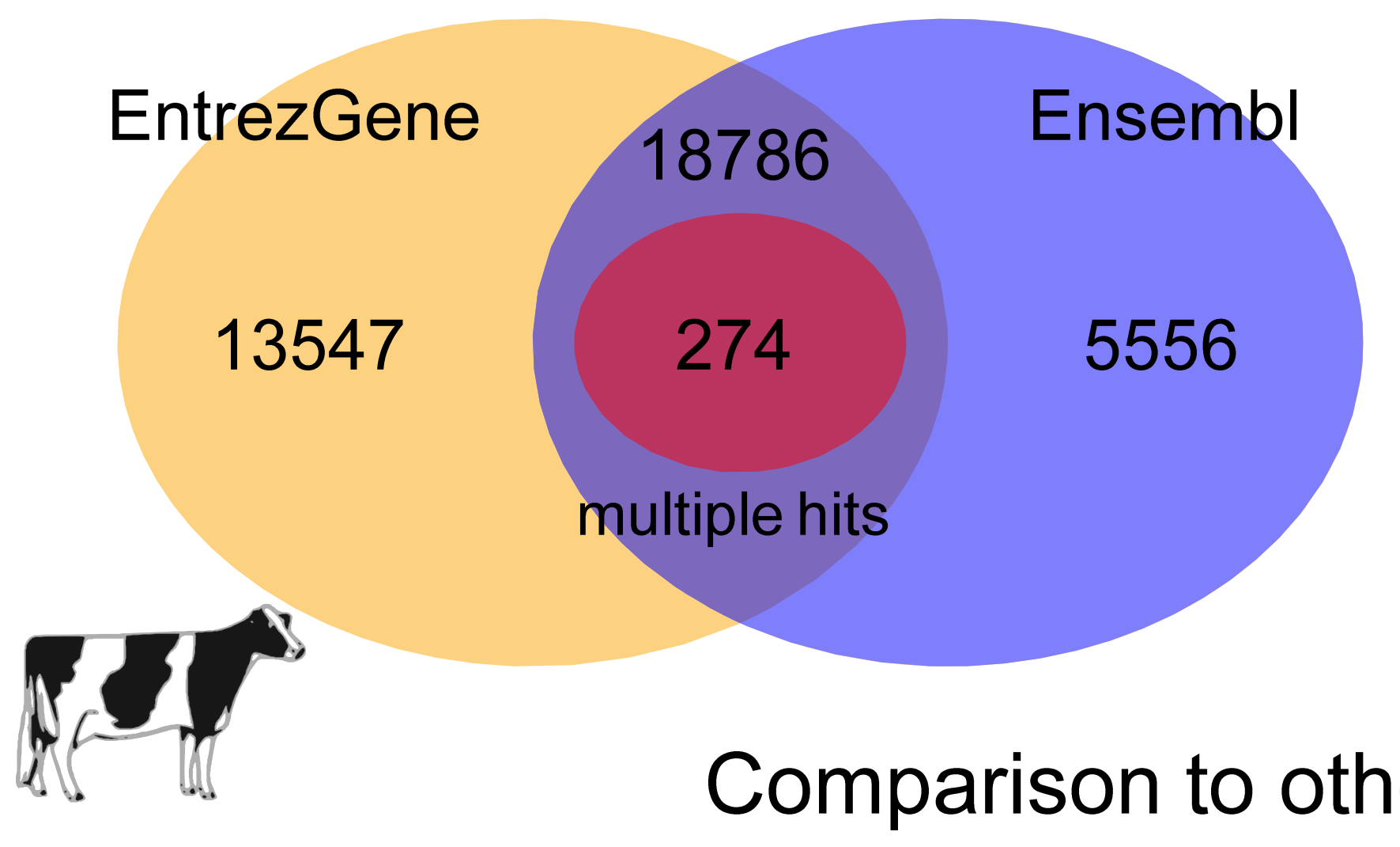
Workflow



Entrez_GeneID	Ensembl_GeneID
100037269	ENSSSCG00000013879
100037270	ENSSSCG0000001087
100037271	ENSSSCG00000004
10 037272	ENSSSCG0000000084 3
7273	ENSSSCG00000028771
274	ENSSSCG00000003718
10 37275	ENSSSCG00000009222
100037276	ENSSSCG00000015886
100037276	ENSSSCG00000015887
100037285	ENSSSCG00000025831
100037286	ENSSSCG00000013397
100037287	ENSSSCG00000004401
100037289	ENSSSCG00000012853
100037290	ENSSSCG00000007143
100037292	ENSSSCG00000011630
100037293	ENSSSCG00000001231
100037294	ENSSSCG00000016916
100037295	ENSSSCG00000012322

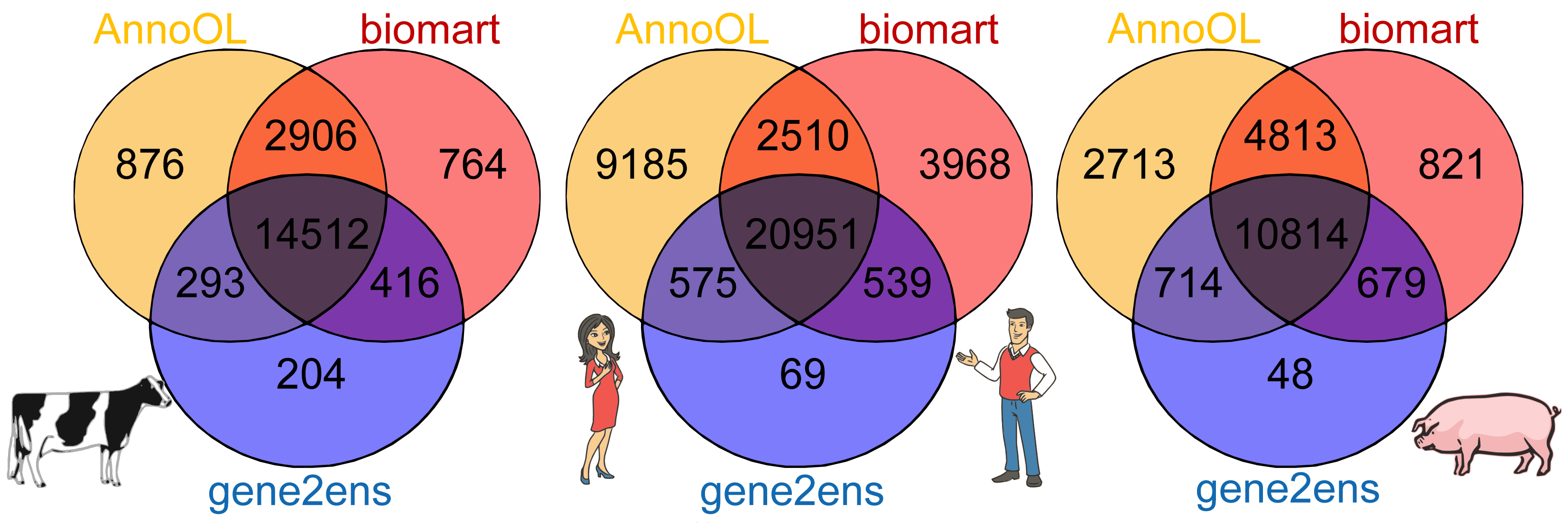
Results

Number of mapped identifiers



The Venn diagram (left) shows the mapped gene identifiers from Bos taurus (cattle). Around 70% of the EntrezGene and ~80% of Ensembl gene identifiers are assigned to each other. The remaining unmapped genes are due to the different annotation pipelines and gene definitions.

Comparison to other approaches

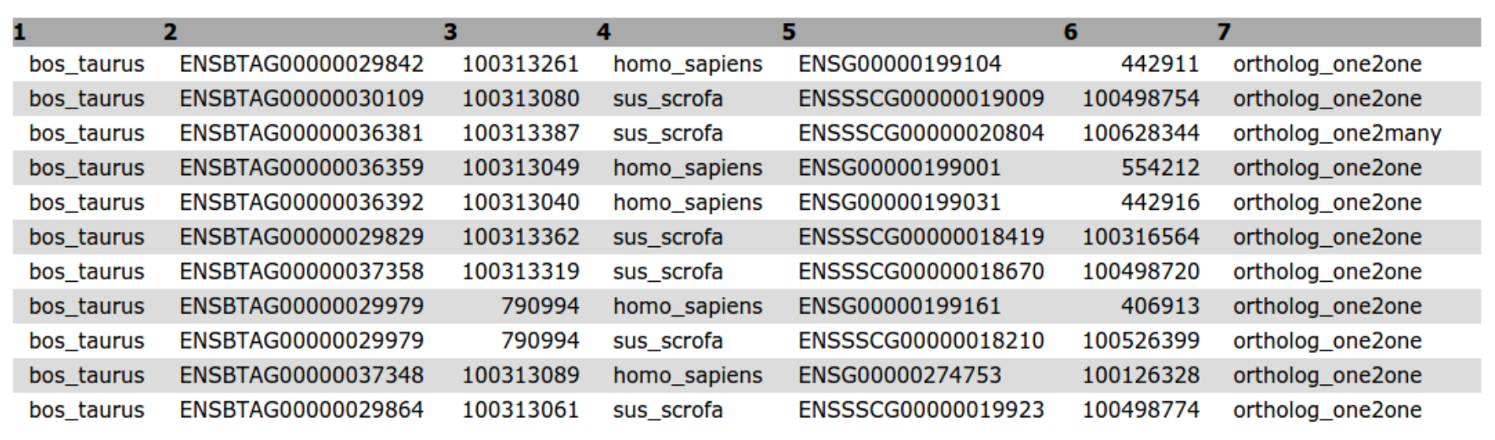


The Venn diagrams show three different lookup tables which are available at Ensembl BioMart (biomart), NCBI gene2ensembl (gene2ens), and the AnnoOverlappeR (AnnoOL). Ensembl BioMart uses a sequence-based comparison while NCBI gene2ensembl uses a position-based algorithm like AnnoOverlappeR but more stringent and without multiple hits handling.

Visualisation in Galaxy

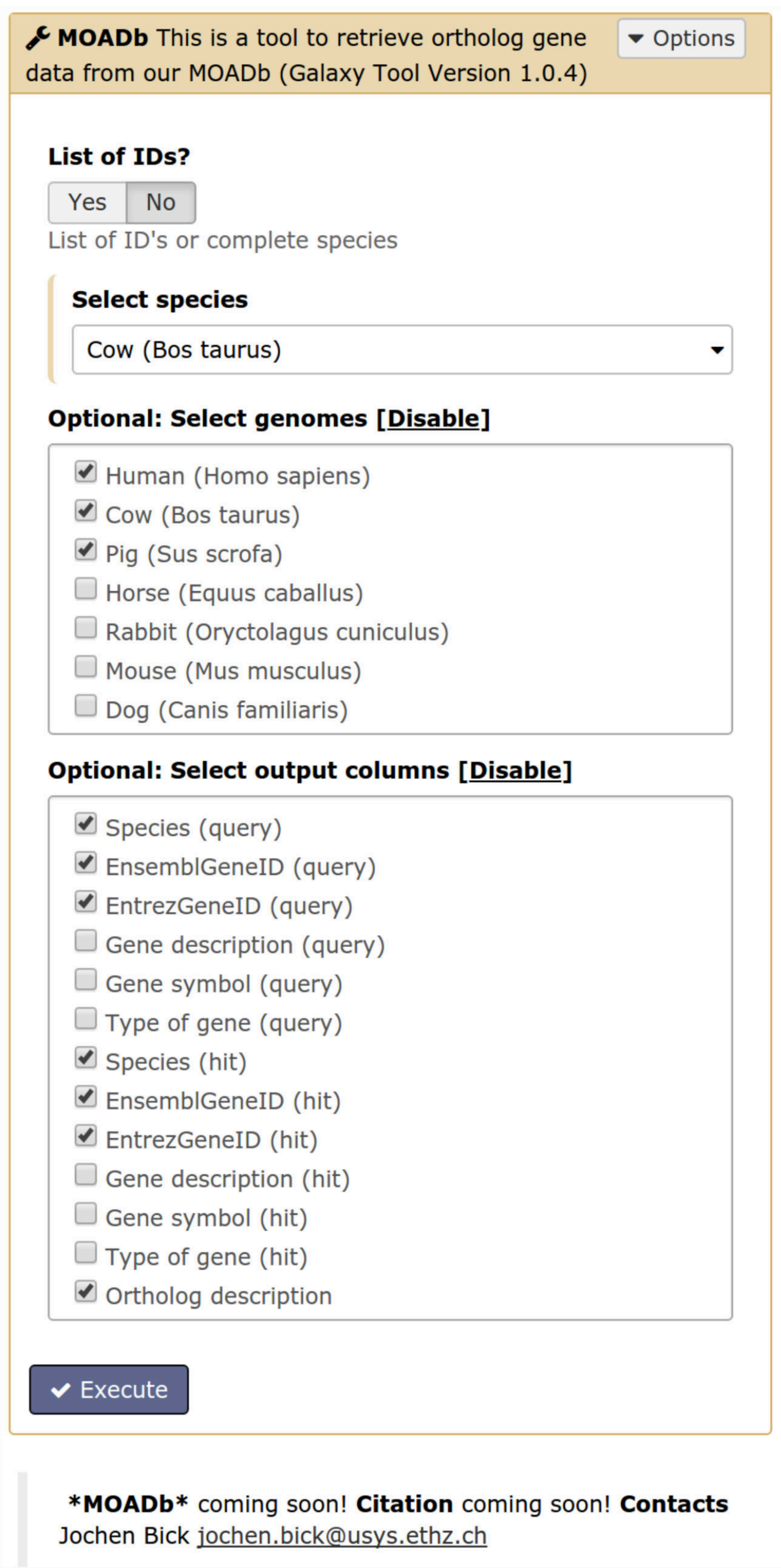
The figure on the right shows a screenshot of the MOA-Db interface on our Galaxy server. For now you can chose between a list of identifiers (IDs) provided by your history or a complete species like selected in the example. 'Select genomes' will specify different mammals which will be matched inside the database. 'Select output columns' decides the columns to be shown in the output file.

Right now we support two different database IDs EntrezGene IDs and Ensembl gene IDs. Our database also contains data from seven different species human, cow, pig, horse, rabbit, mouse and dog.



1	2	3	4	5	6	7
bos_taurus	ENSBTAG00000029842	100313261	homo_sapiens	ENSG00000199104	442911	ortholog_one2one
bos_taurus	ENSBTAG00000030109	100313080	sus_scrofa	ENSSSCG00000019009	100498754	ortholog_one2one
bos_taurus	ENSBTAG00000036381	100313387	sus_scrofa	ENSSSCG00000020804	100628344	ortholog_one2many
bos_taurus	ENSBTAG00000036359	100313049	homo_sapiens	ENSG00000199001	554212	ortholog_one2one
bos_taurus	ENSBTAG00000036392	100313040	homo_sapiens	ENSG00000199031	442916	ortholog_one2one
bos_taurus	ENSBTAG00000029829	100313362	sus_scrofa	ENSSSCG00000018419	100316564	ortholog_one2one
bos_taurus	ENSBTAG00000037358	100313319	sus_scrofa	ENSSSCG00000018670	100498720	ortholog_one2one
bos_taurus	ENSBTAG00000029979	790994	homo_sapiens	ENSG00000199161	406913	ortholog_one2one
bos_taurus	ENSBTAG00000029979	790994	sus_scrofa	ENSSSCG00000018210	100526399	ortholog_one2one
bos_taurus	ENSBTAG00000037348	100313089	homo_sapiens	ENSG00000274753	100126328	ortholog_one2one
bos_taurus	ENSBTAG00000029864	100313061	sus_scrofa	ENSSSCG00000019923	100498774	ortholog_one2one

The figure shows the output in Galaxy which is represented in “tabular” format. You can see all selected features in corresponding order as shown in the right figure.



MOADb This is a tool to retrieve ortholog gene data from our MOADb (Galaxy Tool Version 1.0.4)

List of IDs?

Yes [✓ No]

List of ID's or complete species

Select species

Cow (Bos taurus)

Optional: Select genomes [Disable]

- ☑ Human (Homo sapiens)
- ☑ Cow (Bos taurus)
- ☑ Pig (Sus scrofa)
- ☐ Horse (Equus caballus)
- ☐ Rabbit (Oryctolagus cuniculus)
- ☐ Mouse (Mus musculus)
- ☐ Dog (Canis familiaris)

Optional: Select output columns [Disable]

- ☑ Species (query)
- ☑ EnsemblGeneID (query)
- ☑ EntrezGeneID (query)
- ☐ Gene description (query)
- ☐ Gene symbol (query)
- ☐ Type of gene (query)
- ☑ Species (hit)
- ☑ EnsemblGeneID (hit)
- ☑ EntrezGeneID (hit)
- ☐ Gene description (hit)
- ☐ Gene symbol (hit)
- ☐ Type of gene (hit)
- ☑ Orthology description

MOADb coming soon! **Citation** coming soon! **Contacts** Jochen Bick jochen.bick@usys.ethz.ch