ENSEMBL TOOLS: BIOMART (course booklet pages 50-52)

Answer 1 – Export sequences in FASTA format from mouse

On the Ensembl homepage (http://www.ensembl.org/), click on the 'BioMart' link on the toolbar.

Start with all genes in mouse by choosing the 'Ensembl Genes 83' database, then 'Mus musculus genes (GRCm38p.4)' dataset.

Now, filter for the genes on the 17 chromosome only:

Click on 'Filters' in the left panel, expand the 'REGION' section by clicking on the + box. Select 'Chromosome – 17'. Make sure to check the box in front of the filter is ticked, otherwise the filter won't work.

Now click the 'Count' button on the toolbar.

> This will give you 1718/ 47400 Genes.

Now filter further for genes that are protein-coding by expanding the 'GENE' section (simply click on the + box). Then select 'Gene type – protein_coding' and click again on 'Count'. This now gives you 1056 / 46925 Genes.

Finally, filter for genes that encode proteins that contain transmembrane domains. Expand the 'PROTEIN DOMAINS' section by clicking on the + box and 'Limit to genes … ' with 'Transmembrane domain (tmhmm).

> There are 342 genes on the murine 17 that are protein coding and contain transmembrane domains.

Now you can specify the attributes to be included in the output (note that a number of attributes will already be selected by default). Click on ‘Sequences’, then ‘Protein’. The sequence will be exported as FASTA format.

Have a look at a preview of the results (only 10 rows of the results will be shown):

Click the [Results] button on the toolbar.

If you are happy with how the results look in the preview, output all the results by selecting ‘Export all results to’, then choose the ‘Compressed web file (notify by email), click on ‘Unique results only’, enter your email address in the appropriate box and click on ‘Go’.

Remember to tick the Unique results box to export your data as a file and to view them on the internet browser.

Answer 2 – Convert IDs of human genes

Once in BioMart, click the ‘New’ button on the toolbar.

Choose the ‘Ensembl Genes 83’ database and the ‘Homo sapiens genes (GRCh38.p5)’ dataset.

Click on ‘Filters’ in the left panel, expand the ‘GENE’ section by clicking on the + box, select ‘Input external references ID list (Max 500 advised]. Enter the IDs in the text box (either comma separated or as a list).

Now, click on ‘Attributes’ in the left panel and expand the ‘GENE’ section by clicking on the + box. Deselect ‘Ensembl Transcript ID’, select ‘Description’, expand the ‘EXTERNAL’ section by clicking on the + box and select ‘HGNC symbol’ and ‘UniProt/SwissProt Accession’.

Click the [Results] button on the toolbar.
Select ‘View All rows as HTML’ or export all results to a file. Tick the box ‘Unique results only’.

Note: BioMart is 'transcript-centric', which means that it will often give a separate row of output for each transcript of a gene, even if you don't include the Ensembl Transcript ID in your output. To get rid of redundant rows, use the 'Unique results only' option.

Your results will show 19 genes.

Extra exercise 4 – Convert UniProt IDs into Ensembl IDs for Arabidopsis proteins

Go the plants.ensembl.org and click on BioMart. Click the 'New' button on the toolbar, choose the 'Plants Mart' database and 'Arabidopsis thaliana genes (TAIR10 (2010-09-TAIR10))' dataset.

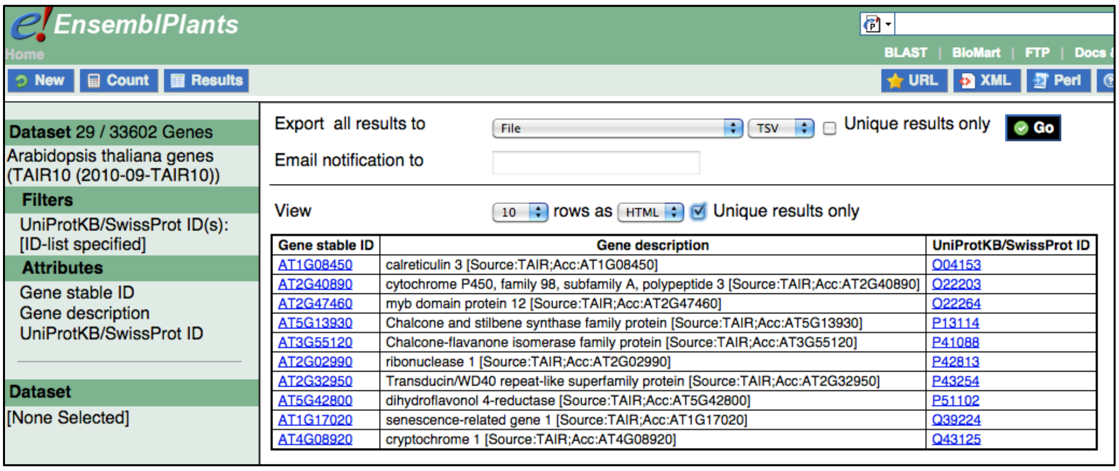
Click on Filters in the left panel and expand the GENE section.
Select ID list limit – UniProt/SwissProt ID(s).
Enter the list of IDs in the text box (either comma separated or as a list).

Click on Attributes in the left panel and expand the GENE section.
Deselect Transcript Stable ID, select Gene name and Gene description and expand the EXTERNAL section.

Now select UniProtKB/SwissProt ID(s).

Click the Results button on the toolbar.
Select View All rows as HTML or export all results to a file. Tick the box Unique results only.

Your results should show 28 / 33602 genes. This is a preview of them:



Dataset 29 / 33602 Genes
Arabidopsis thaliana genes (TAIR10 (2010-09-TAIR10))

Filters
UniProtKB/SwissProt ID(s): [ID-list specified]

Attributes
Gene stable ID
Gene description
UniProtKB/SwissProt ID

Dataset
[None Selected]

Export all results to File TSV □ Unique results only Go

Email notification to

View 10 rows as HTML ☑ Unique results only

Gene stable ID	Gene description	UniProtKB/SwissProt ID
AT1G08450	calreticulin 3 [Source:TAIR;Acc:AT1G08450]	O04153
AT2G40890	cytochrome P450, family 98, subfamily A, polypeptide 3 [Source:TAIR;Acc:AT2G40890]	O22203
AT2G47460	myb domain protein 12 [Source:TAIR;Acc:AT2G47460]	O22264
AT5G13930	Chalcone and stilbene synthase family protein [Source:TAIR;Acc:AT5G13930]	P13114
AT3G55120	Chalcone-flavanone isomerase family protein [Source:TAIR;Acc:AT3G55120]	P41088
AT2G02990	ribonuclease 1 [Source:TAIR;Acc:AT2G02990]	P42813
AT2G32950	Transducin/WD40 repeat-like superfamily protein [Source:TAIR;Acc:AT2G32950]	P43254
AT5G42800	dihydroflavonol 4-reductase [Source:TAIR;Acc:AT5G42800]	P51102
AT1G17020	senescence-related gene 1 [Source:TAIR;Acc:AT1G17020]	Q39224
AT4G08920	cryptochrome 1 [Source:TAIR;Acc:AT4G08920]	Q43125

Note: Remember to tick the 'Unique Results only' box!

Answers extra exercise 5 - Retrieve a list of SNPs from the tomato genome (Solanum lycopersicum)

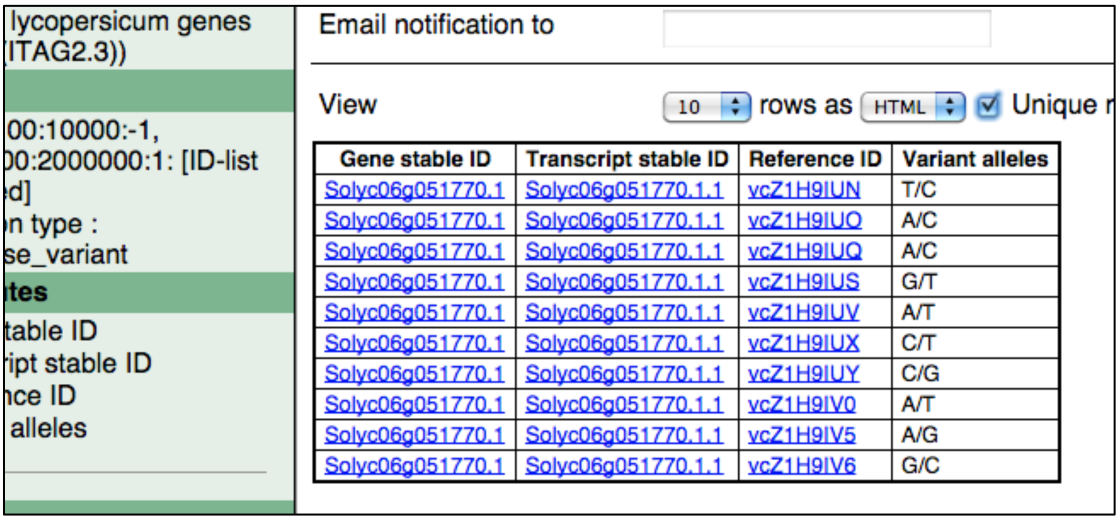
Click the New button on the toolbar, choose the Plants Mart database and Solanum lycopersicum genes (SL2.40 (ITAG2.3)) dataset.

Click on Filters in the left panel and expand the REGION section. Under 'Multiple regions' enter 6:21394819-21397868.

Stay in the same page (Filters) and scroll down to see the field 'VARIATION'. Select 'missense_variant' under 'Variation type'.

Click on Attributes in the left panel, select the Variation attribute and under 'GERMLINE VARIATION INFORMATION', tick 'Reference ID' and 'variant alleles'.

You should see a similar table of Preview of the results as the one below:



lycopersicum genes (ITAG2.3))

00:10000:-1,
00:2000000:1: [ID-list
d]
n type :
se_variant

tes

table ID
ipt stable ID
nce ID
alleles

Email notification to

View 10 rows as HTML ☑ Unique r

Gene stable ID	Transcript stable ID	Reference ID	Variant alleles
Solyc06g051770.1	Solyc06g051770.1.1	vcZ1H9IUN	T/C
Solyc06g051770.1	Solyc06g051770.1.1	vcZ1H9IUO	A/C
Solyc06g051770.1	Solyc06g051770.1.1	vcZ1H9IUQ	A/C
Solyc06g051770.1	Solyc06g051770.1.1	vcZ1H9IUS	G/T
Solyc06g051770.1	Solyc06g051770.1.1	vcZ1H9IUV	A/T
Solyc06g051770.1	Solyc06g051770.1.1	vcZ1H9IUX	C/T
Solyc06g051770.1	Solyc06g051770.1.1	vcZ1H9IUY	C/G
Solyc06g051770.1	Solyc06g051770.1.1	vcZ1H9IV0	A/T
Solyc06g051770.1	Solyc06g051770.1.1	vcZ1H9IV5	A/G
Solyc06g051770.1	Solyc06g051770.1.1	vcZ1H9IV6	G/C

Answers extra exercise 6 - Find genes associated with array probes the human genome (using the Ensembl release 78)

Go to the previous version of Ensembl, release 78

e78.ensembl.org

Click on BioMart at the top of the main homepage.

a) In the BioMart page, click on the 'New' button on the toolbar. Choose the 'Ensembl Genes 78' database and then 'Homo sapiens genes (GRCh38)' dataset.

Click on 'Filters' in the left panel, expand the 'GENE' section by clicking on the + box, select 'ID list limit - Affy hg u133 plus 2 probeset ID(s)' and enter the probeset IDs in the text box (either comma separated or as a list).

Click on 'Attributes' in the left panel, expand the 'GENE' section by clicking on the + box, deselect 'Ensembl Transcript ID', select 'Associated Gene Name' and 'Description', expand the 'EXTERNAL' section by clicking on the + box and select 'Affy HG U133-PLUS-2 probeset'.

Now click on 'Results'. Select 'View All rows as HTML' or export all results to a file. Tick the box 'Unique results only'.

Your results will show 25 genes. In most cases, one probeset maps to one gene. Exceptions are 212014_x_at and 209835_x_at, that both map to ENSG00000026508 (CD44), 227613_at and 219228_at, that both map to ENSG00000130844 (ZNF331) and 213315_x_at, that maps to both ENSG00000197620 and ENSG00000197021.

b) Continue with the same dataset and filters, you can now choose different attributes:

Click on 'Attributes' in the left panel, select the 'Sequences' attributes page, expand the 'SEQUENCES' section and select 'Flank (Transcript)'. Now type '250' in the 'Upstream flank' text box and expand the 'Header Information' section. Select 'Associated Gene Name' and 'Description'.

Note: 'Flank (Transcript)' will give the flanks for all the transcripts of a gene with multiple transcripts. 'Flank (Gene)' will only give the flank for the transcript with the outermost 5' (or 3') end.

Click on 'Results' and select 'View All rows as FASTA' or export all results to a file.

c) Once again you will continue with the same dataset and filters, and choose different attributes:

Click on 'Attributes' in the left panel, select the 'Homologs' attributes page, expand the 'GENE' section, deselect 'Ensembl Transcript ID', expand the 'ORTHOLOGS' section and select 'Mouse Ensembl Gene ID' and 'Homology Type'.

Click on 'Results' and select 'View All rows as HTML' or export all results to a file. Tick the box 'Unique results only'.

Your results will show that for most of the 25 human genes, a one-to-one orthologue in mouse has been identified. However ENSG00000123130 and ENSG00000172716 have two mouse orthologues and ENSG00000197620 and ENSG00000197021 map to the same mouse gene. For four human genes (ENSG00000186594, ENSG00000269952, ENSG00000089335 and ENSG00000130844), no mouse orthologue was identified.

d) Click on ‘Attributes’ in the left panel and select the ‘Features’ attributes page.

Make sure that in the ‘GENE’ section the attributes ‘Ensembl Gene ID’, ‘Associated Gene Name’ and ‘Description’ are selected.

Deselect ‘Affy HG U133-PLUS-2’ in the ‘EXTERNAL’ section.

Add the Ensembl mouse genes as a second dataset by clicking on ‘Dataset’ at the bottom of the left panel. Choose the ‘[Ensembl Genes 78] Mus musculus genes (GRCm38.p3)’ dataset.

Specify the same attributes for mouse as for human: deselect ‘Ensembl Transcript ID’ and select ‘Associated Gene Name’ and ‘Description’. Your results will show the same list as in (c), but now with the name and description added for both human and mouse genes.