
**Computational & Comparative Genomics
Cold Spring Harbor, New York, United States
9-15 November 2011**

**BROWSING GENES AND GENOMES
WITH ENSEMBL**



EXERCISES AND ANSWERS

Note: These exercises are based on Ensembl version 64 (September 2011). After in future a new version has gone live, version 64 will still be available at <http://e64.ensembl.org>. If your answer doesn't correspond with the given answer, please consult the instructor.

BIOMART

Exercise 1

Generate a list of all human genes that are located on the Y chromosome, that are protein-coding and that encode for a protein containing one or more transmembrane domains. Do a count after selection of each filter to check the number of genes remaining in your dataset.

Answer

- 🔗 Go to the Ensembl homepage (<http://www.ensembl.org>).
- 🔗 Click on the 'BioMart' link on the toolbar.

Start with all human Ensembl genes:

- 🔗 Choose the 'Ensembl Genes 64' database.
- 🔗 Choose the 'Homo sapiens genes (GRCh37.p5)' dataset.

Now filter for the genes on the Y chromosome:

- 🔗 Click on 'Filters' in the left panel.
- 🔗 Expand the 'REGION' section by clicking on the + box.
- 🔗 Select 'Chromosome - Y'. Make sure the check box in front of the filter is ticked, otherwise the filter won't work.
- 🔗 Click the [Count] button on the toolbar.

This should give you 554 / 54345 Genes.

Now filter further for genes that are protein-coding:

- 🔗 Expand the 'GENE' section by clicking on the + box.
- 🔗 Select 'Gene type - protein_coding'.
- 🔗 Click the [Count] button on the toolbar.

This should give you 72 / 54345 Genes.

Finally filter for genes that encode proteins containing one or more transmembrane domains:

- ☞ Expand the 'PROTEIN DOMAINS' section by clicking on the + box.
- ☞ Select 'Transmembrane domains - Only'.
- ☞ Click the [Count] button on the toolbar.

This should give you 12 / 54345 Genes.

Specify the attributes to be included in the output (note that a number of attributes will already be default selected):

- ☞ Click on 'Attributes' in the left panel.
- ☞ Expand the 'GENE' section by clicking on the + box.
- ☞ Select, in addition to the attributes 'Ensembl Gene ID' and 'Ensembl Transcript ID' that are already default selected, for instance 'Associated Gene Name' and 'Description'.

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:

- ☞ Select 'View All rows as HTML' or export all results to a file.

Note: When you select 'View All rows as HTML', your results will be shown under a new tab or in a new window in your internet browser.

Although you have filtered for only 12 genes, your results will contain more than 12 rows. This is because several of the genes have more than one transcript and consequently the results contain a separate row for each of these transcripts.

Exercise 2

BioMart is a very handy tool when you want to map between different databases. The following is a list of accession numbers from the UniProtKB/Swiss-Prot database (<http://www.uniprot.org/>) of mouse proteins that are supposedly involved in the sensory perception of pain (http://amigo.geneontology.org/cgi-bin/amigo/term_details?term=GO:0019233):

Q9ESQ8, Q02013, Q04573, P30548, P22387, P35363, P09470, Q61614, P09632, P12968, P47746, P51683, P34971, P47936, Q61125, Q9QXK8, Q9QYS2, P41539, Q63844, P22005, Q8VE91, Q01727, P49615, P35438, P35436, Q8BHK2, P63085, P35383, Q9Z1M0, Q9Z2D6, P25233, P01139, O70174, P49650, Q704Y3, P48999, Q9QXT8, P97772, Q99JA0, P70160, P43136, Q6X1Y6, P70202, Q60612, P39688, P70380, Q8BLA8

Generate a list that shows to which Ensembl Gene IDs and to which MGI symbols these UniProtKB/Swiss-Prot accession numbers map. Also include the gene description.

Hint: Instead of doing a lot of typing (which is not only a waste of time but also introduces the risk of making typos), copy-paste the above list from the digital version of this document, which you can find on the course website.

Answer

- 🔗 Go to the Ensembl homepage (<http://www.ensembl.org>).
- 🔗 Click on the 'BioMart' link on the toolbar.

or if you are already in BioMart:

- 🔗 Click the [New] button on the toolbar.
- 🔗 Choose the 'Ensembl Genes 64' database.
- 🔗 Choose the 'Mus musculus genes (NCBIM37)' dataset.
- 🔗 Click on 'Filters' in the left panel.
- 🔗 Expand the 'GENE' section by clicking on the + box.
- 🔗 Select 'ID list limit - UniProt/Swissprot Accession(s)'.
- 🔗 Enter the list of 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 'Description'.
- 🔗 Expand the 'EXTERNAL' section by clicking on the + box.
- 🔗 Select 'MGI 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 should show 46 genes.

Exercise 3

The paper 'Fine mapping of the usher syndrome type IC to chromosome 11p14 and identification of flanking markers by haplotype analysis' (Ayyagari *et al.* Mol Vis. 1995 Oct 25;1:2) describes the mapping of the human Usher Syndrome type I C to the genomic region between the markers D11S1397 and D11S1310.

Confirm this finding by generating a list of the genes located in the region between D11S1397 and D11S1310. Include the Ensembl Gene ID, name and description.

Answer

☞ Go to the Ensembl homepage (<http://www.ensembl.org>).

☞ Click on the 'BioMart' link on the toolbar.

or if you are already in BioMart:

☞ Click the [New] button on the toolbar.

☞ Choose the 'Ensembl Genes 64' database.

☞ Choose the 'Homo sapiens genes (GRCh37.p5)' dataset.

☞ Click on 'Filters' in the left panel.

☞ Expand the 'REGION' section by clicking on the + box.

☞ Enter 'Marker Start: d11s1397' and 'Marker End: d11s1310'.

☞ 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'.

☞ Click the [Results] button on the toolbar.

☞ Select 'View All rows as HTML' or export all results to a file.

Your results should show 32 genes. Among these there should be one gene (ENSG00000006611) named '*USH1C*' with the description 'Usher syndrome 1C (autosomal recessive, severe) [Source:HGNC Symbol;Acc:12597]'. This confirms that Ayyagari *et al.* correctly mapped Usher Syndrome type I C to this genomic region.

Exercise 4

In the paper 'Discovery of novel biomarkers by microarray analysis of peripheral blood mononuclear cell gene expression in benzene-exposed workers' (Forrest *et al.* Environ Health Perspect. 2005 June; 113(6): 801–807) the effect of benzene exposure on peripheral blood mononuclear cell gene expression in a population of shoe factory workers with well-characterized occupational exposures was examined using microarrays. The microarray used was the Affymetrix U133A/B GeneChip (also called 'U133 plus 2'). The top 25 probe sets up-regulated by benzene exposure were:

207630_s_at, 221840_at, 219228_at, 204924_at 227613_at, 223454_at, 228962_at, 214696_at, 210732_s_at, 212371_at, 225390_s_at, 227645_at, 226652_at, 221641_s_at, 202055_at, 226743_at, 228393_s_at, 225120_at, 218515_at, 202224_at, 200614_at, 212014_x_at, 223461_at, 209835_x_at, 213315_x_at

- (a) Generate a list of the genes to which these probesets map. Include the Ensembl Gene ID, name and description as well as the probeset name.
- (b) As a first step towards analysing them for possible regulatory features they have in common, retrieve the 250 bp upstream of the transcripts of these genes. Include the Ensembl Gene and Transcript ID, name and description in the sequence header.
- (c) In order to be able to study these human genes in mouse, generate a list of the human genes and their mouse orthologs. Include the Ensembl Gene ID for both the human and mouse genes and the homology type in your list.
- (d) Generate the same list as in (c), but now also include the name and description of both the human and mouse genes. As the name and description of the mouse genes are not available as attributes in the Ensembl human genes dataset, you have to add the Ensembl mouse genes as a second dataset.

Answer

(a)

- ☞ Go to the Ensembl homepage (<http://www.ensembl.org>).
- ☞ Click on the 'BioMart' link on the toolbar.

or if you are already in BioMart:

- ☞ Click the [New] button on the toolbar.
- ☞ Choose the 'Ensembl Genes 64' database.
- ☞ Choose the 'Homo sapiens genes (GRCh37.p5)' 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 ID(s)'.
- ☞ Enter the list of 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.
- ☞ Select 'Affy HG U133-PLUS-2'.

- ☞ 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 23 genes. In most cases one probeset maps to one gene. Exceptions are 209835_x_at and 212014_x_at, that both map to ENSG0000026508 (CD44), 219228_at and 227613_at, that both map to ENSG00000130844 (ZNF331), and 213315_x_at, that maps to both ENSG00000197620 (CXorf40A) and ENSG00000197021 (CXorf40B). One probeset (226652_at) doesn't map to an Ensembl gene at all and is therefore not shown in the list of results.

(b)

You can leave the dataset and filters the same, so you can directly specify the attributes:

- ☞ Click on 'Attributes' in the left panel.
- ☞ Select the 'Sequences' attributes page.
- ☞ Expand the 'SEQUENCES' section by clicking on the + box.
- ☞ Select 'Flank (Transcript)'.
- ☞ Enter '250' in the 'Upstream flank' text box.
- ☞ Expand the 'Header Information' section by clicking on the + box.
- ☞ 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 the [Results] button on the toolbar.
- ☞ Select 'View All rows as FASTA' or export all results to a file.

(c)

You can leave the dataset and filters the same, so you can directly specify the attributes:

- ☞ Click on 'Attributes' in the left panel.
- ☞ Select the 'Homologs' attributes page.
- ☞ Expand the 'GENE' section by clicking on the + box.
- ☞ Deselect 'Ensembl Transcript ID'.
- ☞ Expand the 'ORTHOLOGS' section by clicking on the + box.
- ☞ Select 'Mouse Ensembl Gene ID' and 'Homology Type'.

- ☞ 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 that for most of the 23 human genes a one-to-one ortholog in mouse has been identified, while ENSG00000123130 has two mouse orthologs and ENSG00000172716 has three mouse orthologs. ENSG00000197620 and ENSG00000197021 map to the same mouse gene. For two human genes (ENSG00000130844 and ENSG00000186594) no mouse ortholog has been identified.

(d)

You can leave the Ensembl human genes dataset and filters the same, so you can directly specify the attributes:

- ☞ Click on 'Attributes' in the left panel.
- ☞ 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:

- ☞ Click on 'Dataset' at the bottom of the left panel.
- ☞ Choose the '[Ensembl Genes 64] Mus musculus genes (NCBIM37)' dataset.

Specify the same attributes for mouse as for human:

- ☞ 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'.

☞ 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 the same list as in (c), but now with the name and description added for both the human and mouse genes.

Exercise 5

Design your own query!
