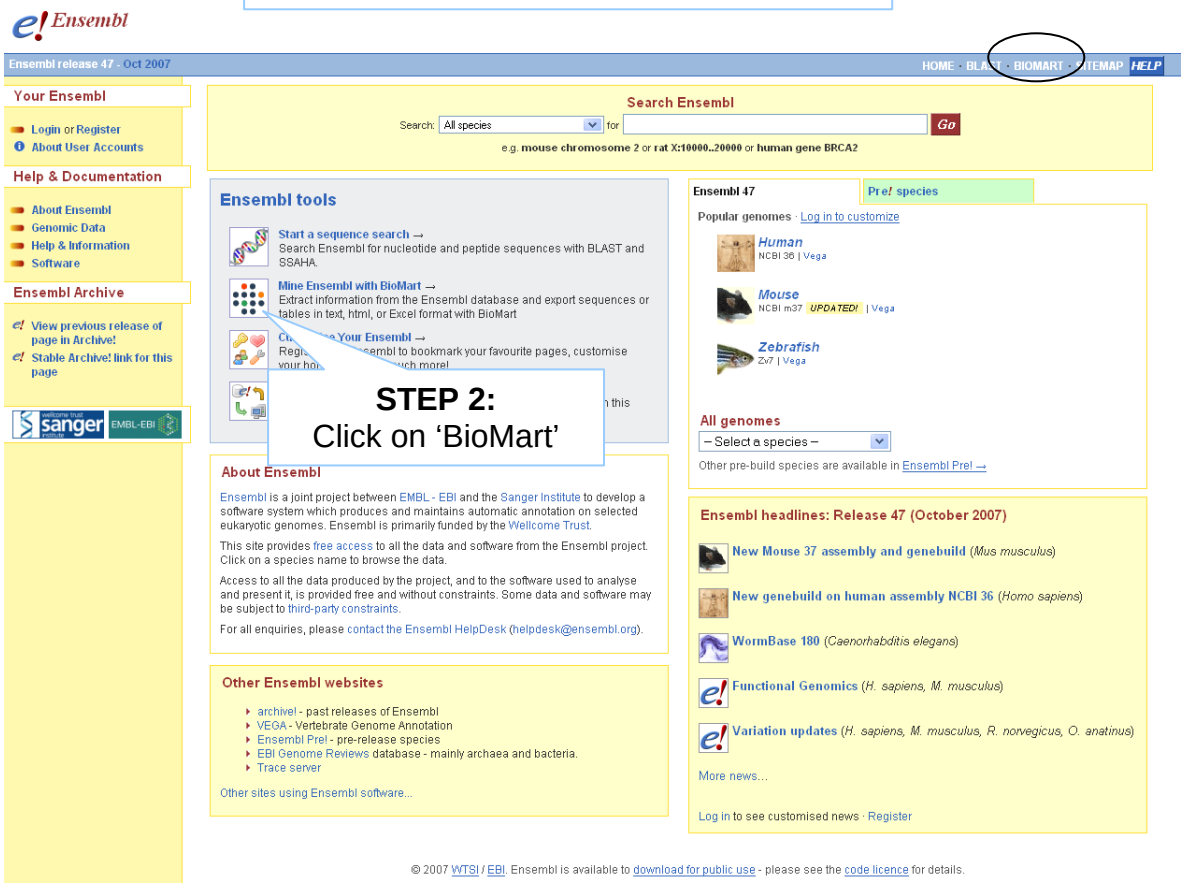
Data mining in Ensembl with BioMart
Worked Example

The human gene encoding Glucose-6-phosphate dehydrogenase (G6PD) is located on chromosome X in cytogenetic band q28.

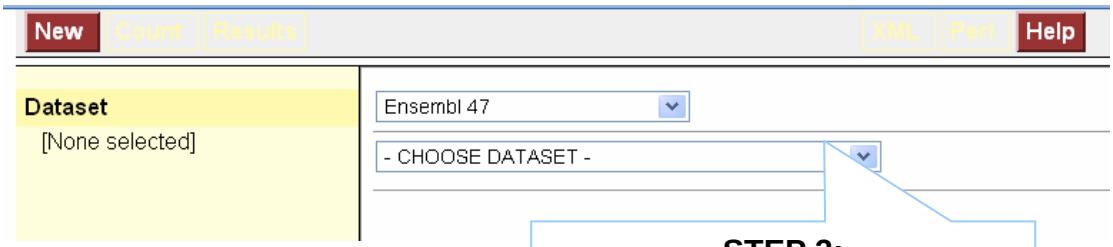
Which other genes related to human diseases locate to the same band? What are their Ensembl Gene IDs and Entrez Gene IDs? Do they have any functions predicted by Interpro?

What are their cDNA sequences?

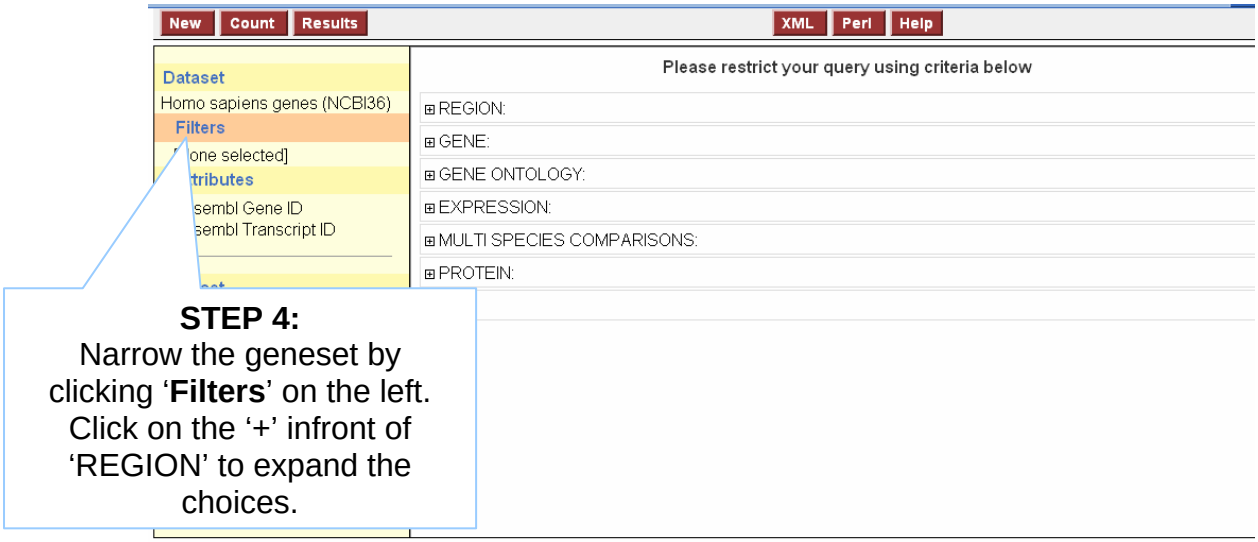
STEP 1:
Go to the Ensembl main page
www.ensembl.org



STEP 2:
Click on 'BioMart'

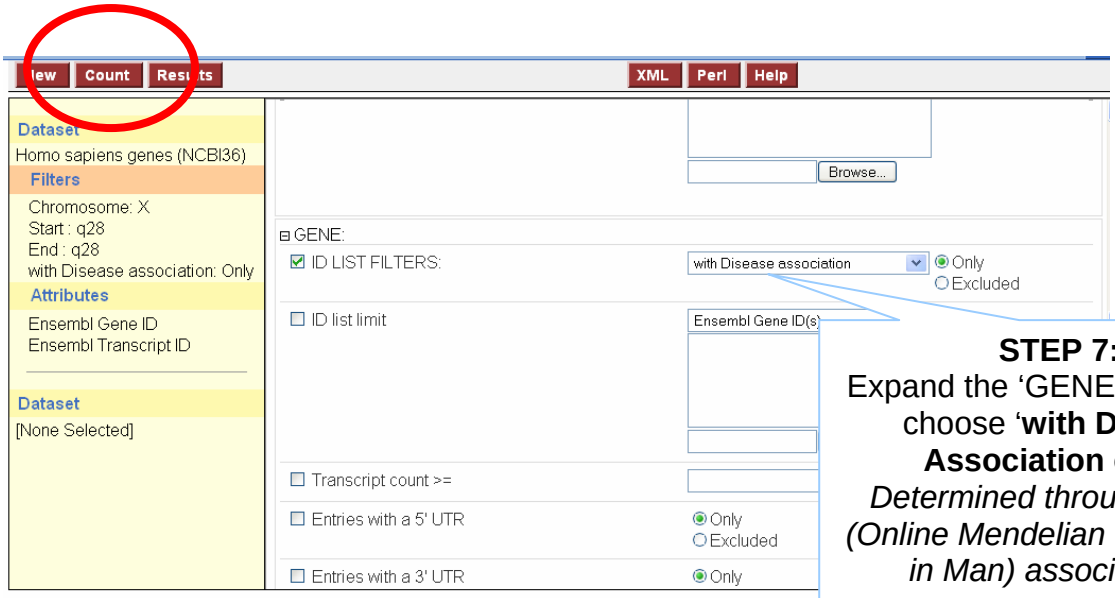


STEP 3:
Select the database: Ensembl genes (version 47) and the species of interest under ‘Choose Dataset’. (*Homo sapiens*)



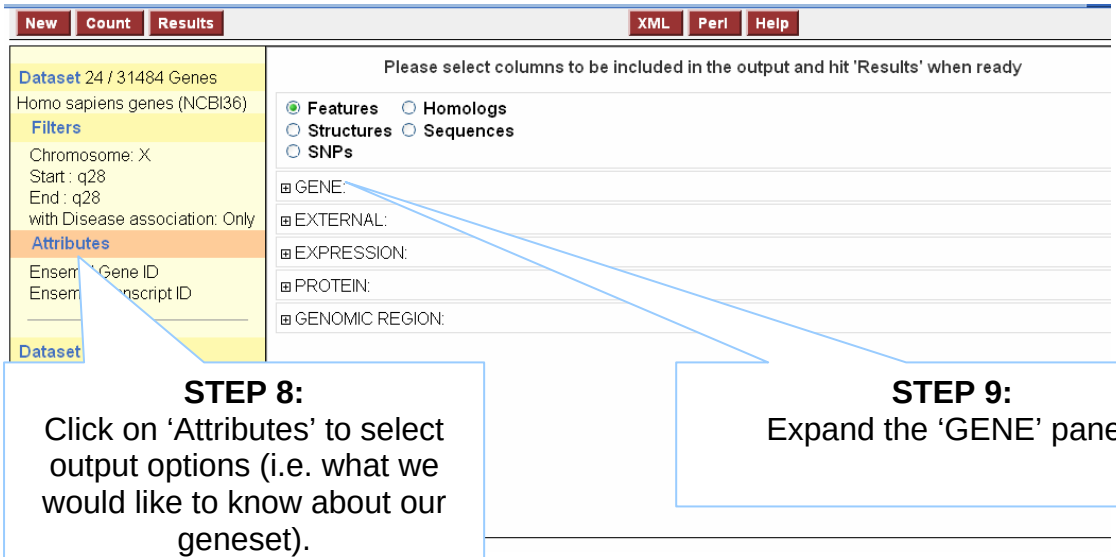
STEP 4:
Narrow the geneset by clicking ‘**Filters**’ on the left. Click on the ‘+’ infront of ‘REGION’ to expand the choices.


New
Count
Results
XML
Perl
Help
Dataset
Homo sapiens genes (NCBI36)
Filters
Chromosome: X
Start : q28
End : q28
Attributes
Ensembl Gene ID
Ensembl Transcript ID
Dataset
[None Selected]
Please restrict your query using criteria below
REGION:
Chromosome
X
Base pair
Gene Start (bp)
1
Gene End (bp)
10000000
Band
Start
q28
End
q28
Marker
Start
End
Encode type
manual_picks
STEP 5:
Select 'Chromosome X'
STEP 6:
Select 'Band Start q28' and 'End q28'



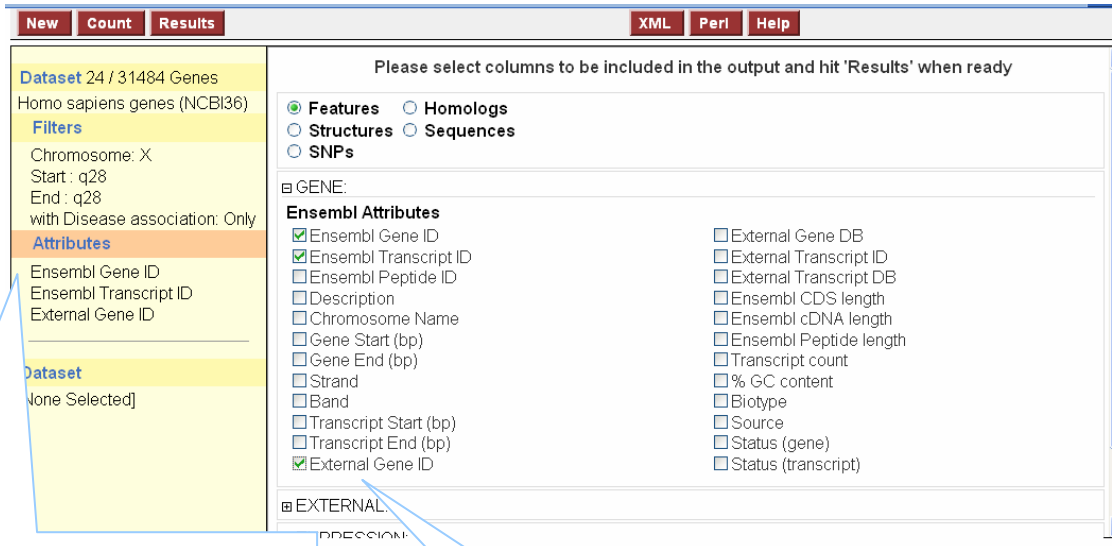
STEP 7:
Expand the 'GENE' panel and choose '**with Disease Association only**'.
Determined through OMIM (Online Mendelian Inheritance in Man) associations.

The filters have determined our gene set. Click **'Count'** (at the top) to see how many genes have passed these filters.



STEP 8:
Click on 'Attributes' to select output options (i.e. what we would like to know about our geneset).

STEP 9:
Expand the 'GENE' panel.

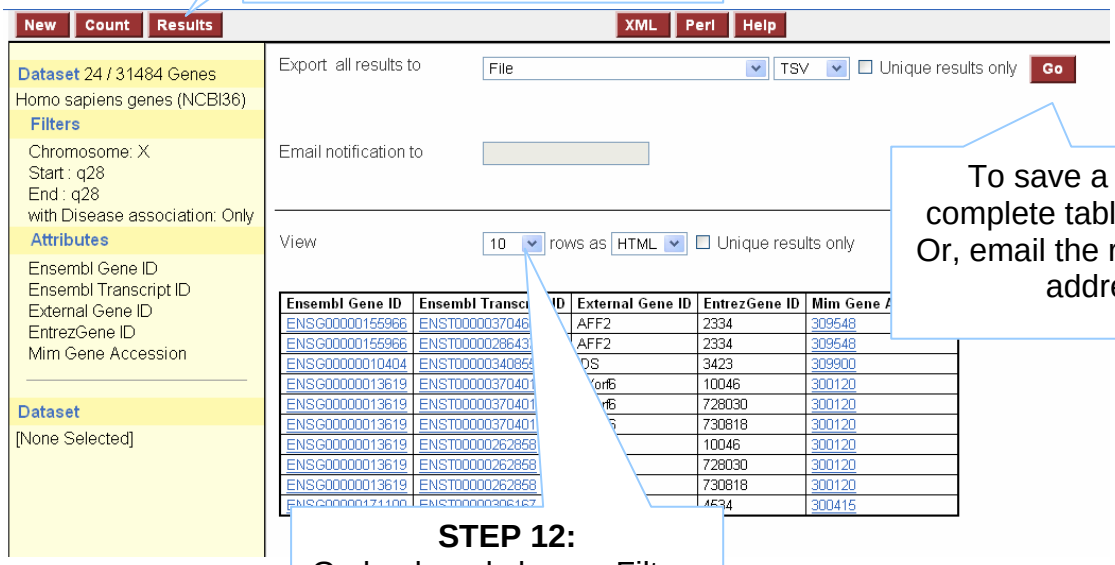


Note the summary of selected options.

The order of attributes determines the order of columns in the result table.

STEP 10:
Select, along with the default options, '**External Gene ID**' (this shows the gene symbol from HGNC).
Expand the 'EXTERNAL' panel to select '**EntrezGene ID**' and '**Mim Gene Accession**' (this is the ID from OMIM)
www.ncbi.nlm.nih.gov/**omim**/

STEP 11:
Click 'RESULTS' at the top to preview the output.



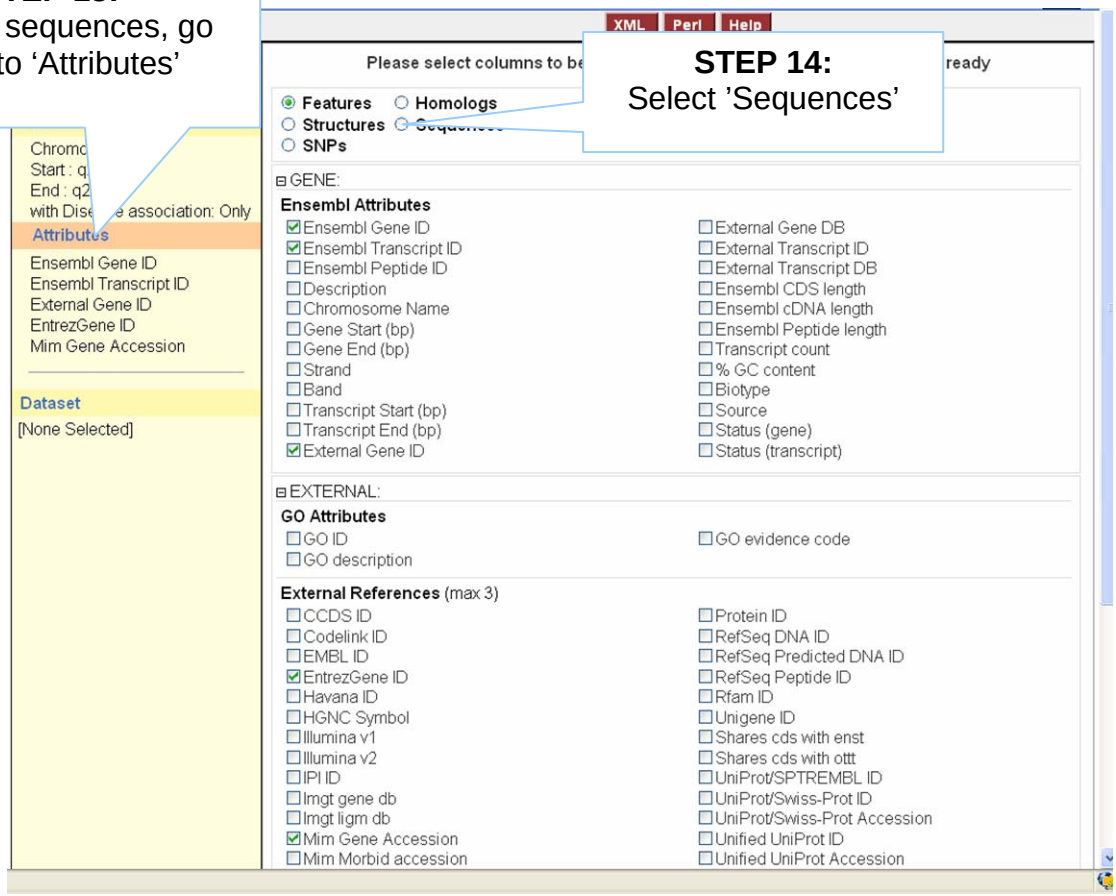
To save a file of the complete table, click 'Go'. Or, email the results to any address.

STEP 12:
Go back and change Filters or Attributes if desired.
Or, View 'ALL' as HTML…

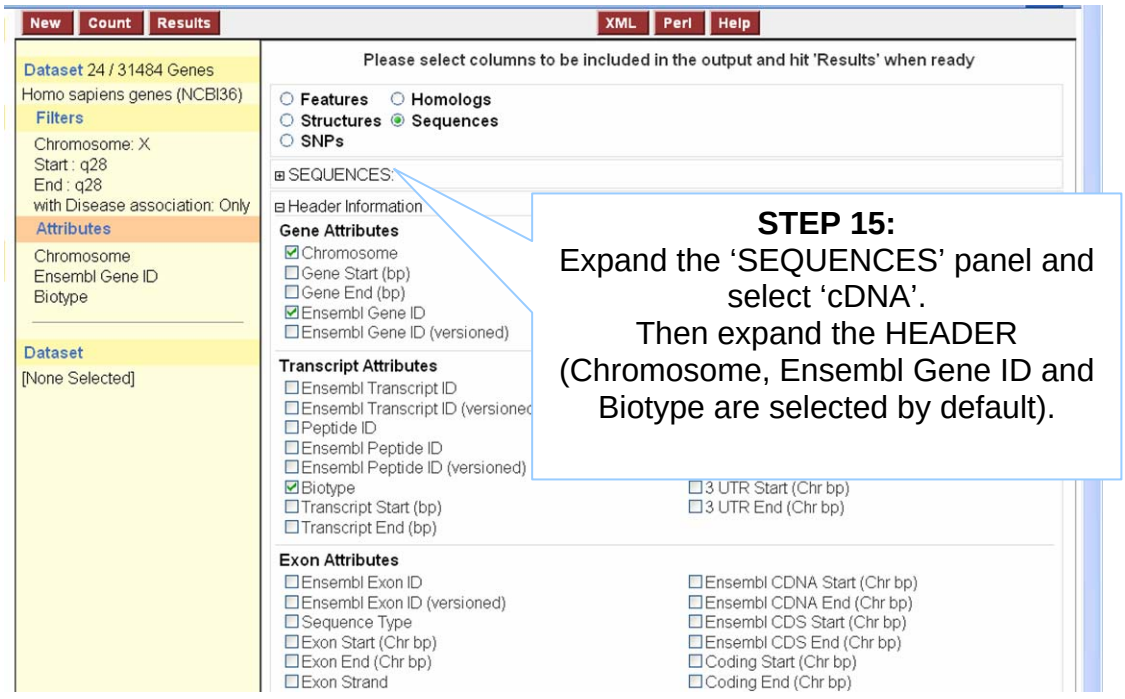
Result Table 1

Ensembl Gene ID	Ensembl Transcript ID	External Gene ID	EntrezGene ID	Mim Gene Accession
ENSG00000155966	ENST00000370460	AFF2	2334	309548
ENSG00000155966	ENST00000286437	AFF2	2334	309548
ENSG00000010404	ENST00000340855	IDS	3423	309900
ENSG00000013619	ENST00000370401	CXorf6	10046	300120
ENSG00000013619	ENST00000370401	CXorf6	728030	300120
ENSG00000013619	ENST00000370401	CXorf6	730818	300120
ENSG00000013619	ENST00000262858	CXorf6	10046	300120
ENSG00000013619	ENST00000262858	CXorf6	728030	300120
ENSG00000013619	ENST00000262858	CXorf6	730818	300120
ENSG00000171100	ENST00000306167	MTM1	4534	300415
ENSG00000147383	ENST00000370274	NSDHL	50814	300275
ENSG00000130821	ENST00000330048	SLC6A8	6535	300036
ENSG00000130821	ENST00000253122	SLC6A8	6535	300036
ENSG00000185825	ENST00000345046	BCAP31	10134	300398
ENSG00000185825	ENST00000370133	BCAP31	10134	300398
ENSG00000101986	ENST00000218104	ABCD1	215	300371
ENSG00000101986	ENST00000218104	ABCD1	642762	300371
ENSG00000198910	ENST00000370060	L1CAM	3897	308840
ENSG00000198910	ENST00000361699	L1CAM	3897	308840
ENSG00000126895	ENST00000358927	AVPR2	554	300538
ENSG00000126895	ENST00000337474	AVPR2	554	300538
ENSG00000169057	ENST00000369964	MECP2	4204	300005
ENSG00000169057	ENST00000303391	MECP2	4204	300005
ENSG00000102076	ENST00000369951	OPN1LW	5956	303900
ENSG00000147380	ENST00000369935	OPN1MW	2652	303800
ENSG00000147380	ENST00000369935	OPN1MW	728458	303800
ENSG00000166160	ENST00000369929	OPN1MW2	2652	303800
ENSG00000166160	ENST00000369929	OPN1MW2	728458	303800
ENSG00000007350	ENST00000369915	TKTL1	8277	300044
ENSG00000196924	ENST00000369850	FLNA		300017
ENSG00000102119	ENST00000369842	EMD	2010	300384
ENSG00000147403	ENST00000369817	RPL10	6134	312173
ENSG00000147403	ENST00000369817	RPL10	647074	312173
ENSG00000147403	ENST00000344746	RPL10	6134	312173
ENSG00000147403	ENST00000344746	RPL10	647074	312173

STEP 13:
To view sequences, go back to ‘Attributes’



STEP 14:
Select ’Sequences’



STEP 15:
Expand the ‘SEQUENCES’ panel and select ‘cDNA’.
Then expand the HEADER (Chromosome, Ensembl Gene ID and Biotype are selected by default).

STEP 16:
Click on ‘Results’.

View all rows as FASTA…

RESULTS

Header: chromosome, Ensembl Gene ID, Biotype

>X|ENSG00000130821|protein_coding

cDNA 1

GCCTCCGCGGGCCCCGGCCGGGGCGGGGGGCGCGGGCCACAGGCCCCTGCTCCGGCCGCC
GCTTGCAGACCGCGGGCGCCGATGTCGCCCGCGCCCCGCTAGGCTGAGCCTCGGGTCGGG
CGAGGAGCCGCCGCAGCCGCCGCCGCCCGAGCCGCGGGCAGGAGCCTCGGGAGCCGCCGC
CGCCGCCGCCGCCGCCCGGCCGGGCCCCGCCGCCGCCCGCGCGCCCCCGGGCCCCCGACA
CACATGAGATTCTTCAGGCTCACTTTCAAGTGCTTCGTGGACTGCTTCTGACTGCGCCGC
CCGCGCCCCGCACCCCGCCGCCCGCCCGCCGCCCCGTCCCCCGGCCCGGCCGCCCCCCGG
CCCCCGGCCGGCCCGCGCCCTCGGGGCCCTCCCCGGTGCCGCCGGTGCCCCCCGCCTGAC
CGCCGCCCCCCGTGAGGCGCCGCGACCCCGGCCCGGCCGTGCGGCCCGCCGAGGCCATGG
CGAAGAAGAGCGCCGAGAACGGCATCTATAGCGTGTCCGGCGACGAGAAGAAGGGCCCCC
TCATCGCGCCCGGGCCCGACGGGGCCCCGGCCAAGGGCGACGGCCCCGTGGGCCTGGGGA
CACCCGGCGGCCGCCTGGCCGTGCCGCCGCGCGAGACCTGGACGCGCCAGATGGACTTCA
TCATGTCGTGCGTGGGCTTCGCCGTGGGCTTGGGCAACGTGTGGCGCTTCCCCTACCTGT
GCTACAAGAACGGCGGAGGTGTGTTCCTTATTCCCTACGTCCTGATCGCCCTGGTTGGAG
GAATCCCCATTTTCTTCTTAGAGATCTCGCTGGGCCAGTTCATGAAGGCCGGCAGCATCA
ATGTCTGGAACATCTGTCCCCTGTTCAAAGGCCTGGGCTACGCCTCCATGGTGATCGTCT
TCTACTGCAACACCTACTACATCATGGTGCTGGCCTGGGGCTTCTATTACCTGGTCAAGT
CCTTTACCACCACGCTGCCCTGGGCCACATGTGGCCACACCTGGAACACTCCCGACTGCG
TGGAGATCTTCCGCCATGAAGACTGTGCCAATGCCAGCCTGGCCAACCTCACCTGTGACC
AGCTTGCTGACCGCCGGTCCCCTGTCATCGAGTTCTGGGAGAACAAAGTCTTGAGGCTGT
CTGGGGGACTGGAGGTGCCAGGGGCCCTCAACTGGGAGGTGACCCTTTGTCTGCTGGCCT
GCTGGGTGCTGGTCTACTTCTGTGTCTGGAAGGGGGTCAAATCCACGGGAAAGATCGTGT
ACTTCACTGCTACATTCCCCTACGTGGTCCTGGTCGTGCTGCTGGTGCGTGGAGTGCTGC
TGCCTGGCGCCCTGGATGGCATCATTTACTATCTCAAGCCTGACTGGTCAAAGCTGGGGT
CCCCTCAGGTGTGGATAGATGCGGGGACCCAGATTTTCTTTTCTTACGCCATTGGCCTGG
GGGCCCTCACAGCCCTGGGCAGCTACAACCGCTTCAACAACAACTGCTACAAGGACGCCA
TCATCCTGGCTCTCATCAACAGTGGGACCAGCTTCTTTGCTGGCTTCGTGGTCTTCTCCA
TCCTGGGCTTCATGGCTGCAGAGCAGGGCGTGCACATCTCCAAGGTGGCAGAGTCAGGGC
CGGGCCTGGCCTTCATCGCCTACCCGCGGGCTGTCACGCTGATGCCAGTGGCCCCACTCT
GGGCTGCCCTGTTCTTCTTCATGCTGTTGCTGCTTGGTCTCGACAGCCAGTTTGTAGGTG
TGGAGGGCTTCATCACCGGCCTCCTCGACCTCCTCCCGGCCTCCTACTACTTCCGTTTCC
AAAGGGAGATCTCTGTGGCCCTCTGTTGTGCCCTCTGCTTTGTCATCGATCTCTCCATGG

>X|ENSG00000155966|protein_coding

cDNA 2

CGCCGCCTGTGCAGCCGCTGCCGCCGCCGCCGCCGCCGCCGCCGCCGCCGCCGCCGCCGC
CGCCGCTGCCGCCCCGGCTGCCGCGCCGCGCCGCTGCCTCTGCCCCGGCCGCCCCCGCCG
CCGCTGCCGCCGCCGGCCCGCAGCCAGCCAGGCGGGCGGCCCAGCCCGCCTGAGCCCGCA
GCGGCTGCCGCCGCAGCGTCGGGTCGCTGGGTGCGCGGGCTACCGCGGACCGAGCGGACC
CGAGTGGGCGACCAGGCGCTTGCCCGCCCAGTGCCACTGCCGCCGCTTCCTCGCCGGAGC
ACAGGACCAGACACCTCCAGCGCCCGCTGCTGCTGCCGATGCGGCCCGGACACTTTTAGC
TGGGCGGGAGGGCTGGAGAGCCGGGGGCCGCCGAGAACCGCCAGCGAGCTGTGCCGAGAG
CCGCGCCGACCCGCTGCGATCAGGGACAGGCGCCCGCCCGCCGCCGCCGCCTGGCCGCTA
TGGATCTATTCGACTTTTTCAGAGACTGGGACTTGGAGCAGCAGTGTCACTATGAACAAG
ACCGTAGTGCACTTAAAAAAAGGGAATGGGAGCGGAGGAATCAAGAAGTCCAGCAAGAAG
ACGATCTCTTTTCTTCAGGCTTTGATCTTTTTGGGGAGCCATACAAGGTAGCTGAATATA
CAAACAAAGGTGATGCACTTGCCAACCGAGTCCAGAACACGCTTGGAAACTATGATGAAA
TGAAGAATTTGCTAACTAACCATTCTAATCAGAATCACCTAGTGGGAATTCCAAAGAATT
CTGTGCCCCAGAATCCCAACAACAAAAATGAACCAAGCTTTTTTCCAGAACAAAAGAACA
GAATAATTCCACCTCACCAGGATAATACCCATCCTTCAGCACCAATGCCTCCACCTTCTG
TTGTGATACTGAATTCAACTCTAATACACAGCAACAGAAAATCAAAACCTGAGTGGTCAC
GTGATAGTCATAACCCTAGCACTGTACTGGCAAGCCAGGCCAGTGGTCAGCCAAACAAGA
TGCAGACTTTGACACAGGACCAGTCTCAAGCCAAACTGGAAGACTTCTTTGTCTACCCAG
CTGAACAGCCCCAGATTGGAGAAGTTGAAGAGTCAAACCCATCTGCAAAGGAAGACAGTA

V) BIOMART - Exercises

These exercises have been designed to familiarise you with different questions you can answer with this tool, and the types of data you can retrieve with BioMart.

1. Retrieve all SNPs for 'novel' human G-protein coupled receptor genes (GPCRs – Use the InterPro domain ID: IPR000276) on chromosome 2.

Note: As this is the first exercise we walk you this time through BioMart step-by-step (but of course you can also try to do this exercise without our help!)

Start a new BioMart session by clicking 'New', or go back to the Ensembl homepage and click on 'Mine Ensembl with Biomart' under 'Ensembl tools'.

Choose the **database** and the **dataset** for your query as follows:
- Select 'Ensembl 47'
- Select 'Homo sapiens genes (NCBI36)'.

Click on '**Filters**' at the left. Filter this dataset to select your genes of interest as follows:
- Expand the 'REGION' section at the right by clicking on the '+'. Select 'Chromosome 2'. Click [count] at the top of the panel and note the number of Ensembl genes on *Homo sapiens* chromosome 2.
- In the 'GENE' section, select 'Status (gene)' 'NOVEL'.
- In the 'PROTEIN' section, select the second 'Limit to genes with these family or domain IDs' option. Select 'Interpro ID(s)' and enter 'IPR000276' in the box. Click [count] again and note that the number of genes is updated.

Click on '**Attributes**' (at the left). Select the output for your gene list as follows:
- Select the 'SNPs' Attribute Page.
- In the 'GENE' section 'Ensembl Gene ID' and 'Ensembl Transcript ID' are selected by default – also select 'Ensembl Peptide ID and 'Ensembl Peptide length'.
- In the 'GENE ASSOCIATED SNPs' section select 'Reference ID', 'Allele', 'Peptide location (aa)', 'Location in Gene (coding etc)', 'Synonymous Status' and 'Peptide Shift'.

Click on '**Results**' (at the top) to obtain the first 10 rows of your table. To obtain the entire table select 'View all rows as HTML' or export a file by clicking 'Go'.

Note that the output for this query gives you one row for each SNP, and if there are alternative transcripts then SNP data is given for each. This means that a particular SNP may appear more than once.

Find the coding SNPs, and note that you have information about the effect of the SNP, and its location within the protein. Synonymous status is 'yes' for

silent mutations. Two amino acids will be shown in the 'Peptide Shift' column if there are two alleles on the protein level. The Peptide location (aa), Synonymous Status and Peptide Shift will all be blank if the SNP is not in a coding region.

2. Retrieve the gene structure (i.e. start and end coordinates of exons) of the mouse gene ENSMUSG00000042351.

3. Retrieve all human disease genes located between p11.2 and q22 (these are bands on chromosome 1).

4. The file http://www.ebi.ac.uk/~xose/Affy_exercise.txt contains a list of probeset IDs from a microarray experiment using the Affymetrix array HG-U133 Plus 2.0 (human). Retrieve the 500 bp upstream of the transcripts matching these probeset IDs.

5. Retrieve the sequences 5kb upstream of all human 'known' genes between D1S2806 and D1S464.

6. Retrieve sequence (including reference ID in the header) of all human SNPs that have an ID from The SNP Consortium (TSC), from chromosome 6 between 15 Mb and 15.2 Mb, with 200 bases flanking sequence.

7. Retrieve the mouse homologues of *Homo sapiens* genes CASP1, CASP2, CASP3, and CASP4. (These are HGNC symbols for the genes).

8. Design your own query!

Answers (BioMart)

1. You should find ***one*** novel gene on chromosome 2 with this InterPro domain. (*Note: there can be more than one gene with one InterPro domain*). The result set has one transcript and a total of 261 rows of output (to see this, change the option from TSV to XLS under 'Export all results' and click 'Go', then open in Excel so you don't have to count the rows manually). The transcript has 8 coding SNPs ('Location in Gene' is 'coding'), most of which are non-synonymous ('Synonymous status' is 'no') and thus affect the amino acid sequence of the encoded peptide. One allele is a stop codon- can you find it?

2. Database and dataset: 'Ensembl 47' and 'Mus musculus genes (NCBIM36)'.

Filters: GENE 'ID list limit Ensembl Gene ID(s)': enter the mouse gene ID.

Attributes 'Structures': select in the **EXON** panel: 'Ensembl Exon ID', 'Exon Start' and 'Exon End'.

Click '**Results**'.

You should find ***7 exons***. Take the link from the Ensembl Gene ID in your output back to the **GeneView** page to confirm the BioMart data with the gene structure displayed on this page.

3. Database and dataset: 'Ensembl 47' and 'Homo sapiens genes (NCBI36)'.

Filters: REGION 'Chromosome 1', 'Band Start p11.2, 'Band End q22' ,
GENE: 'with Disease Association Only' (look under 'ID LIST FILTERS')

Attributes: Features: select 'GO ID' and 'GO description' along with the default options ('Ensembl Gene ID' and 'Transcript ID').

Results should show ***17 Ensembl genes*** (multiple transcripts and GO terms).

4. Database and dataset: 'Ensembl 47' and 'Homo sapiens genes (NCBI36)'.

Filters: GENE: 'ID list limit': Affy hg u133 plus 2 ID(s) and enter the list of probeset IDs.

Attributes: 'Sequences' select 'Flank (Transcript)', 'Upstream flank 500'. In the header, apart from the already default selected options, select 'Ensembl Transcript ID'.

You should find upstream sequences for the transcripts of ***31 genes*** (Hint: click 'count' to see the number of genes!)

5. Database and dataset: 'Ensembl 47' and 'Homo sapiens genes (NCBI36)'.

Filters: REGION 'Marker' : Start D1S2806' End D1S464'
GENE: 'Status: KNOWN'.

Attributes 'Sequences' and select, apart from the already default selected options, 'Flank (Gene)' and 'Upstream flank 5000'.

You should find sequences for ***26 genes***.

When you choose the option 'Flank (Gene)' you will see only one upstream sequence per gene in the output. In the case where a gene has multiple transcripts, the upstream sequence of the transcript that extends the furthest at the 5' end is shown. If you want to export the upstream sequences for each transcript you should choose the option 'Flank (Transcript)'.

'Known' genes are Ensembl gene predictions that could be matched to same-species external database entries (e.g. UniProt/SwissProt) with a high similarity score (i.e. with BLAST or a similar sequence identity-matching program)

6. Database: ‘SNP’ and **dataset:** ‘Homo sapiens SNPs (dbSNP127;HGVbase 15; TSC 1; affy GeneChip Mapping Array)’.

Filters: REGION: ‘Chromosome 6’, ‘Base pair Start 15000000’, ‘Base pair End 15200000’
GENERAL SNP FILTERS: SNP source: ‘SNPs with TSC ID(s) Only’.

Attributes ‘Sequences’: SEQUENCES : ‘SNP sequences’, ‘Upstream flank 200’, ‘Downstream flank 200’.
SNP: SNP attributes, select ‘Reference ID’.

You should find ***69 SNPs***.

7. Database: ‘Ensembl 47’ **Dataset:** Homo sapiens genes (NCBI36)

Filters: **GENE:** ‘ID list limit HGNC Symbol(s)’. Enter the human HGNC (HUGO) symbols in the box: CASP1, CASP2, CASP3, and CASP4.

Attributes: Under ‘**Homologs**’, select in the ‘**MOUSE ORTHOLOGS**’ panel ‘Mouse Ensembl Gene ID’ and ‘Mouse External ID’. Also select ‘Ensembl gene ID’ and Transcript ID (default options) and ‘Description’ in the ‘**GENE**’ panel (these will be for the starting dataset… i.e. Human.)

Results displays the mouse orthologues of the human CASP genes.