Module 5 – BioMart

You will learn about

- BioMart, a joint project developed and maintained at EBI and OiCR www.biomart.org
- How to use BioMart to quickly obtain lists of gene information from Ensembl
- Specifically, how to export a table of gene information in Microsoft Excel format, HTML or text, and/or sequence in Fasta format.

BioMart can be used to directly access the data in Ensembl and export tables of gene information or sequences. Any user can obtain gene-associated data in tabular format without the need for knowing any programming. The 'query' or the initial input can be an entire set of genes for a species, or a smaller more limited set (e.g. a list of IDs or a specific region of a chromosome). Information about the gene set defined by the user can be exported as txt, html, or in Microsoft Excel format (XLS). This information can range from chromosomal position to associated IDs in other databases to a short description of the gene. Other supported export formats are Fasta and **GFF**. These are only some examples of the information that can be obtained through this fast and user-friendly interface.

The following is a 'worked example' or web-site walkthrough of BioMart. It is probably the best way to learn how to use it! Read along, or follow on the web using the archive site for version 52 so that the layout is identical and results match up.

BioMart may have been updated since the time of this worked-example, but the concepts and basic layout should be the same. You can also find BioMart at www.biomart.org (click on MartVew). Not only the Ensembl genes are accessible from BioMart; this tool can also be used to access data in **MSD** (Macromolecular Structures Database), **Wormbase**, **HapMap** and others.

BioMart Walkthrough:

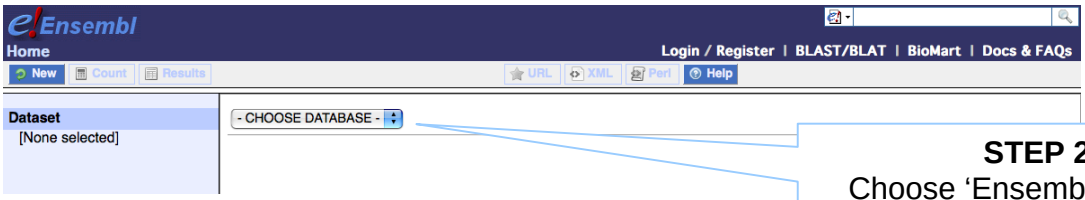
The human gene encoding Glucose-6-phosphate dehydrogenase (G6PD) is located on chromosome X in cytogenetic band q28.

Which other genes related to human diseases locate to the same band? What are their Ensembl Gene IDs and Entrez Gene IDs?

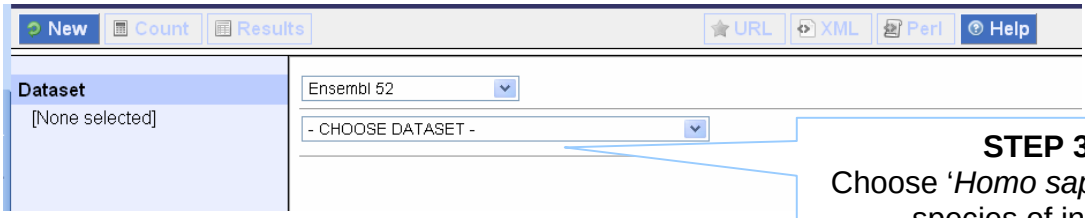
What are their cDNA sequences?

Follow the worked example below to answer these questions.

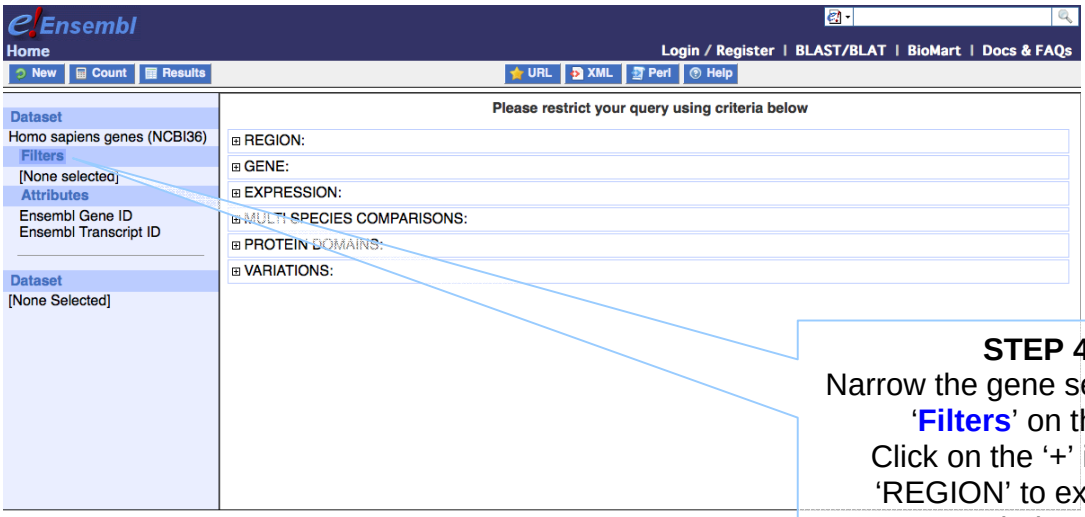
Step 1: Either click on 'BioMart' in the top right header bar of the Ensembl home page, or go to http://www.biomart.org/ and click on the 'MartView' tab.



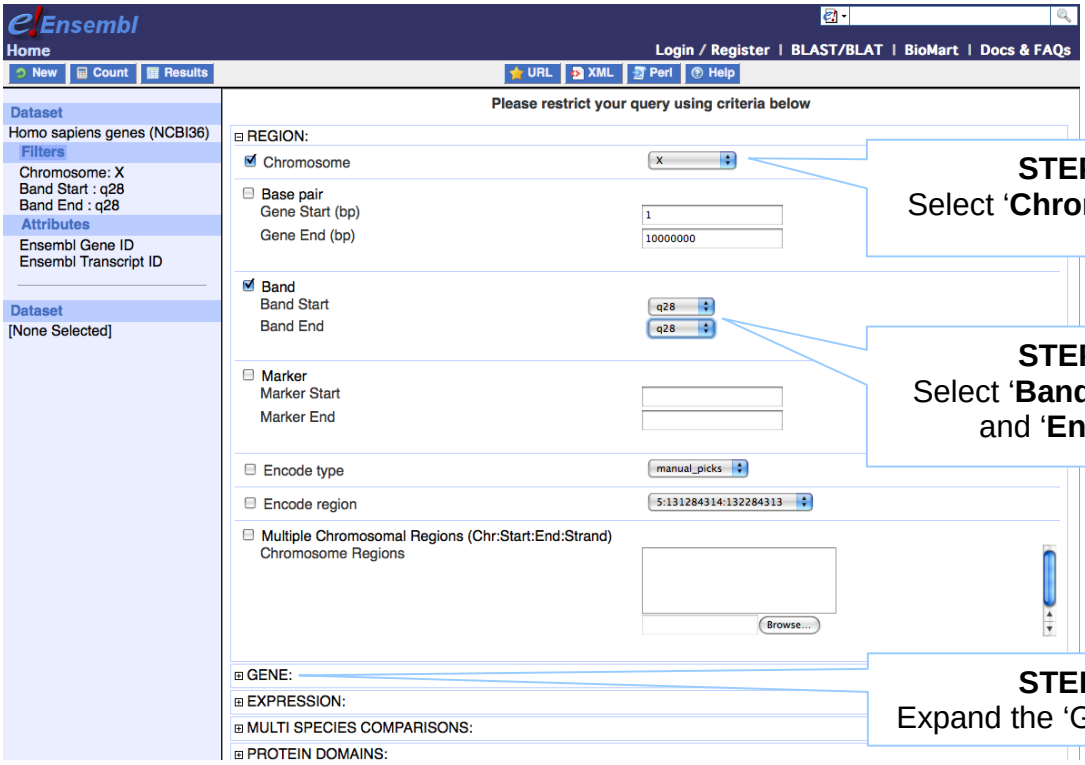
STEP 2:
Choose ‘Ensembl 52’ as the primary database.



STEP 3:
Choose ‘*Homo sapiens*’ as the species of interest.



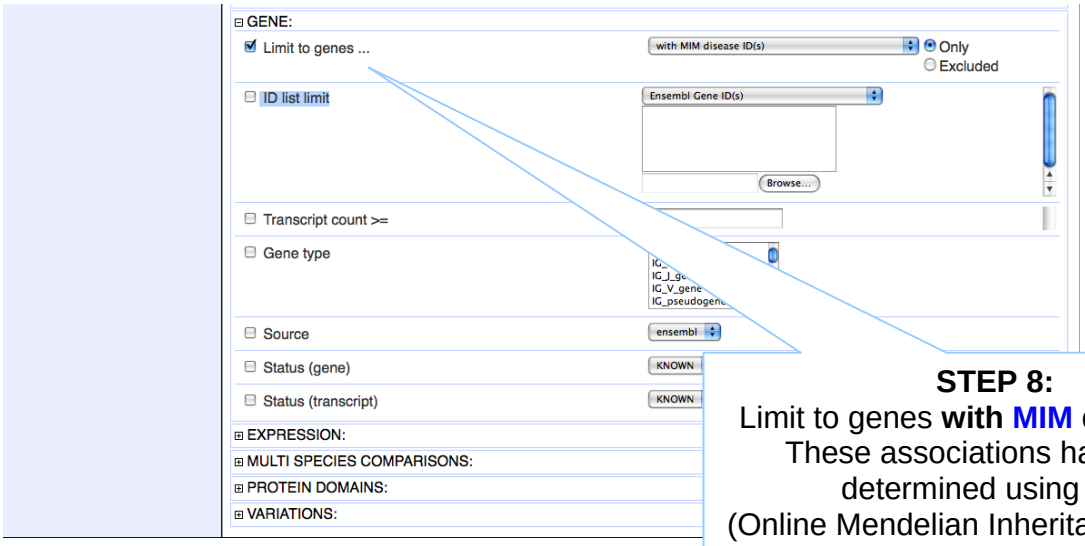
STEP 4:
Narrow the gene set by clicking ‘**Filters**’ on the left.
Click on the ‘+’ in front of ‘REGION’ to expand the choices.



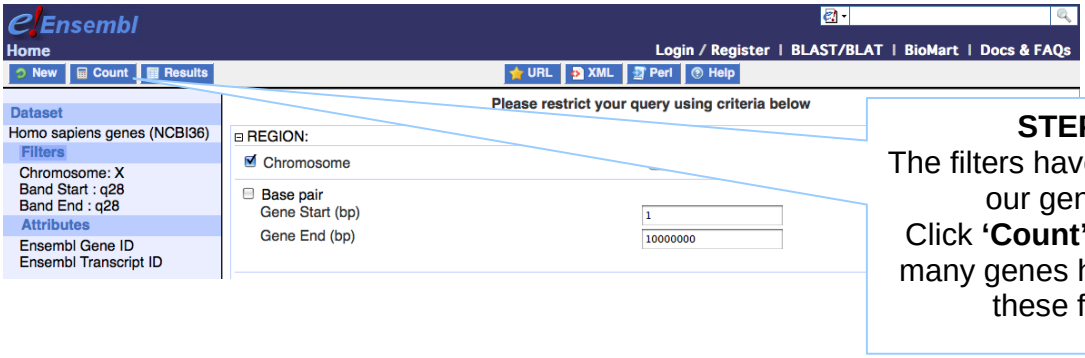
STEP 5:
Select '**Chromosome X**'

STEP 6:
Select '**Band Start q28**' and '**End q28**'

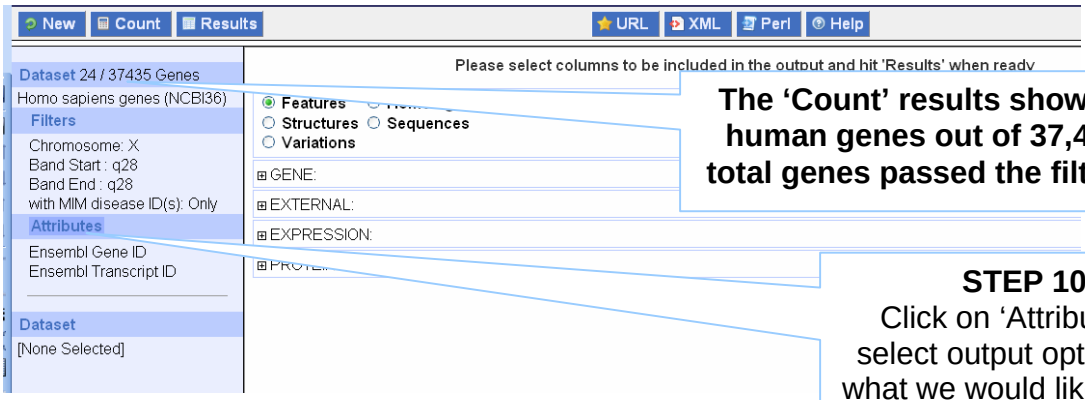
STEP 7:
Expand the 'GENE' panel.



STEP 8:
Limit to genes **with MIM disease ID**'. These associations have been determined using MIM (Online Mendelian Inheritance in Man).
http://www.ncbi.nlm.nih.gov/omim/

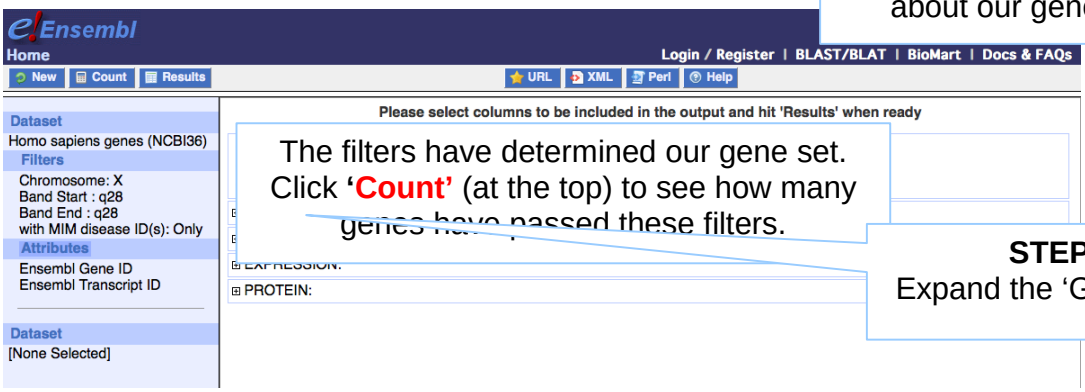


STEP 9:
The filters have determined our gene set.
Click **'Count'** to see how many genes have passed these filters.



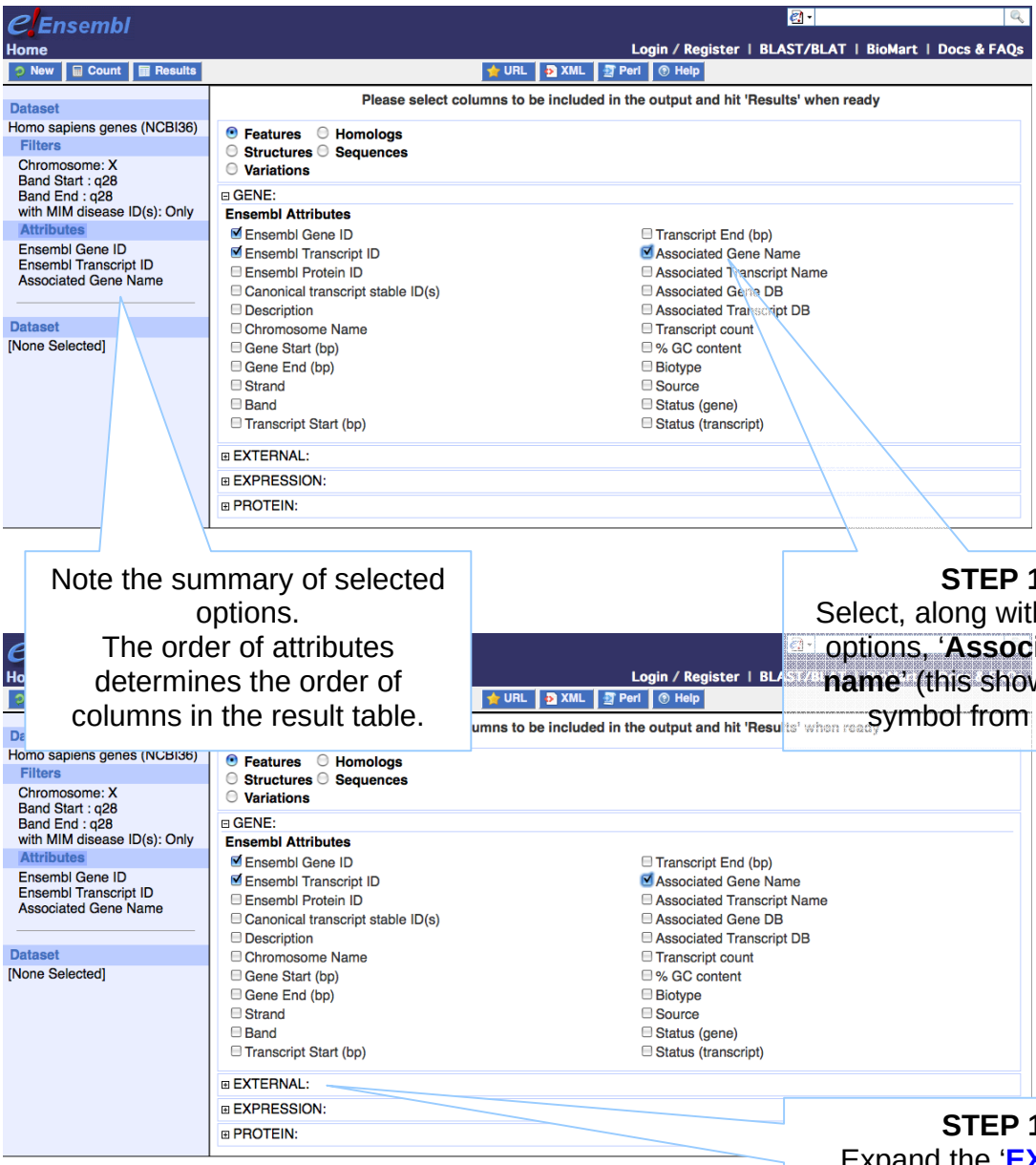
The 'Count' results show 24 human genes out of 37,435 total genes passed the filters.

STEP 10:
Click on 'Attributes' to select output options (i.e. what we would like to know about our gene set).



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

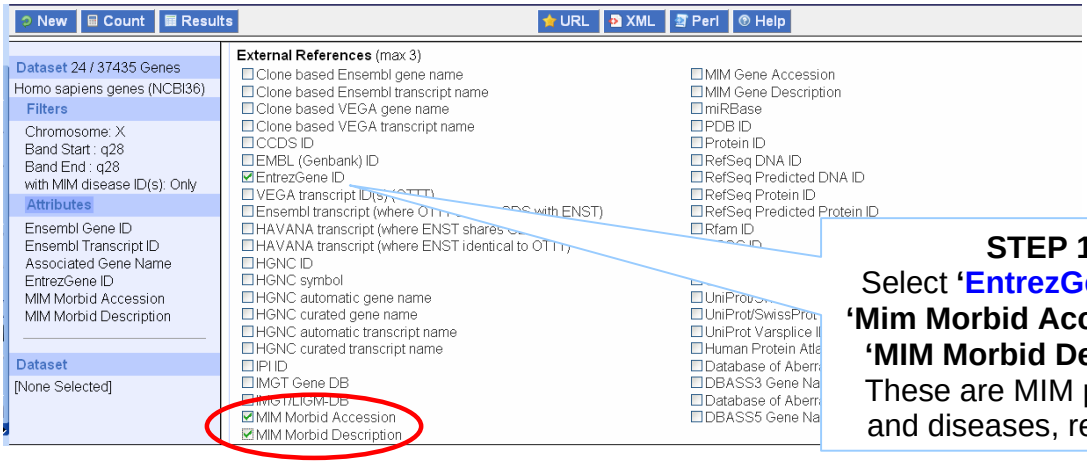
STEP 11:
Expand the 'GENE' panel.



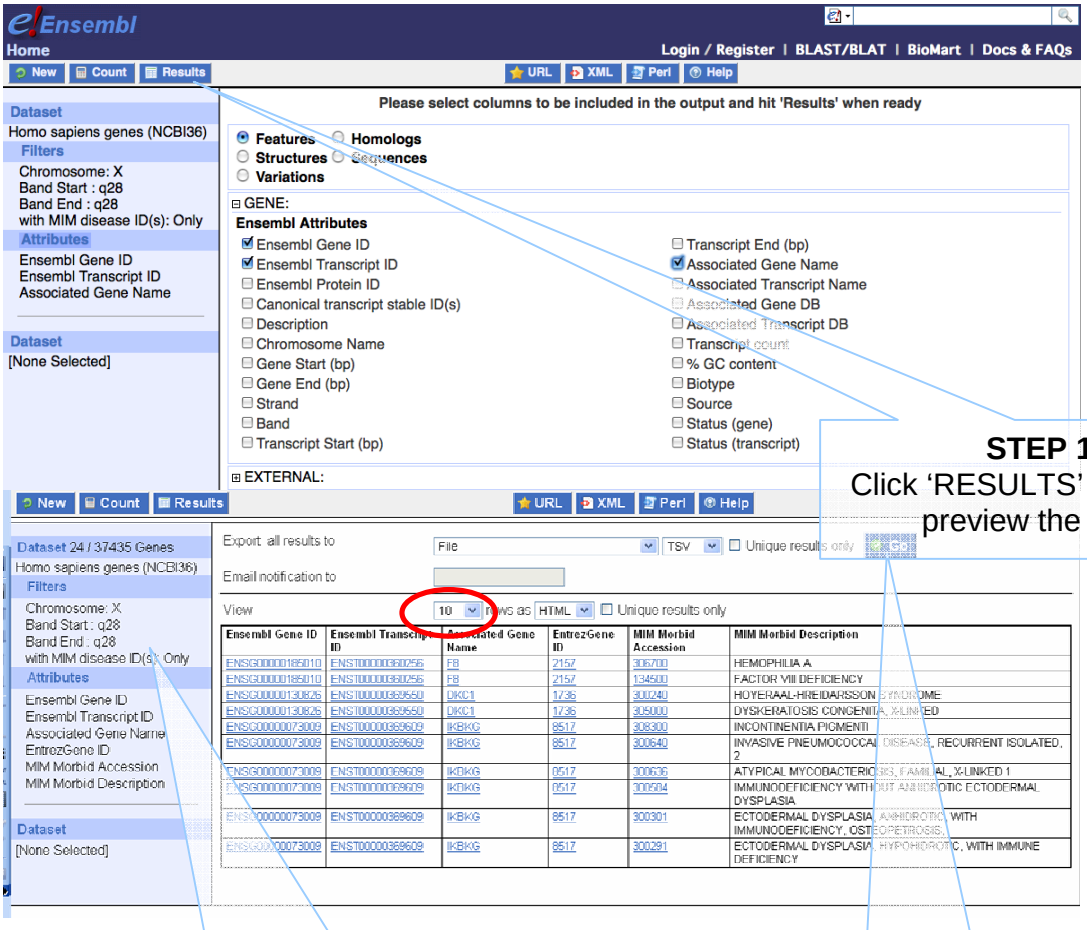
Note the summary of selected options.
The order of attributes determines the order of columns in the result table.

STEP 12:
Select, along with the default options, '**Associated Gene name**' (this shows the gene symbol from HGNC).

STEP 13:
Expand the '**EXTERNAL**' panel to select External References.



STEP 14:
Select '**EntrezGene ID**' and '**Mim Morbid Accession**' and '**MIM Morbid Description**'. These are MIM phenotypes and diseases, respectively.



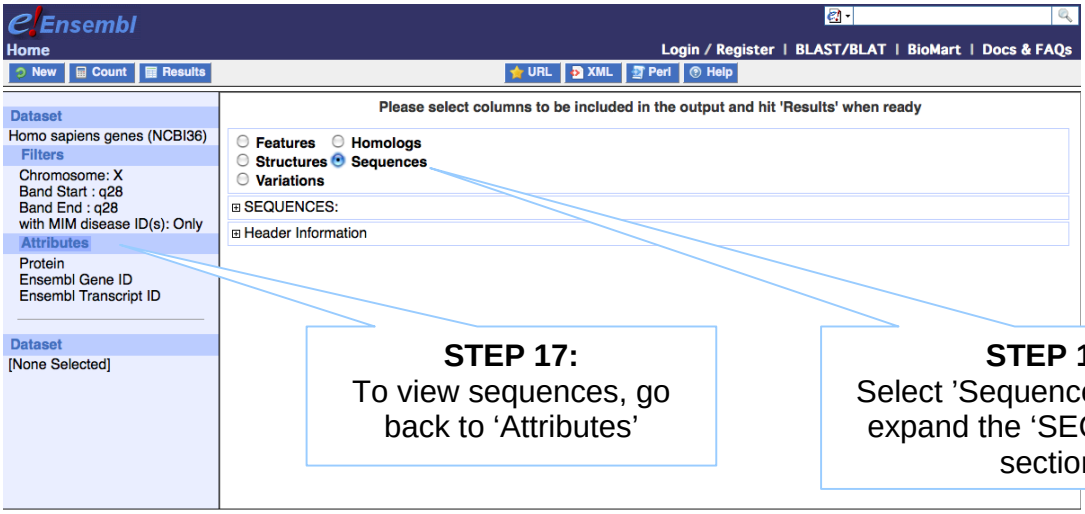
STEP 15:
Click 'RESULTS' at the top to preview the output.

STEP 16:
Go back and change Filters or Attributes if desired.
Or, View ALL rows as HTML…

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

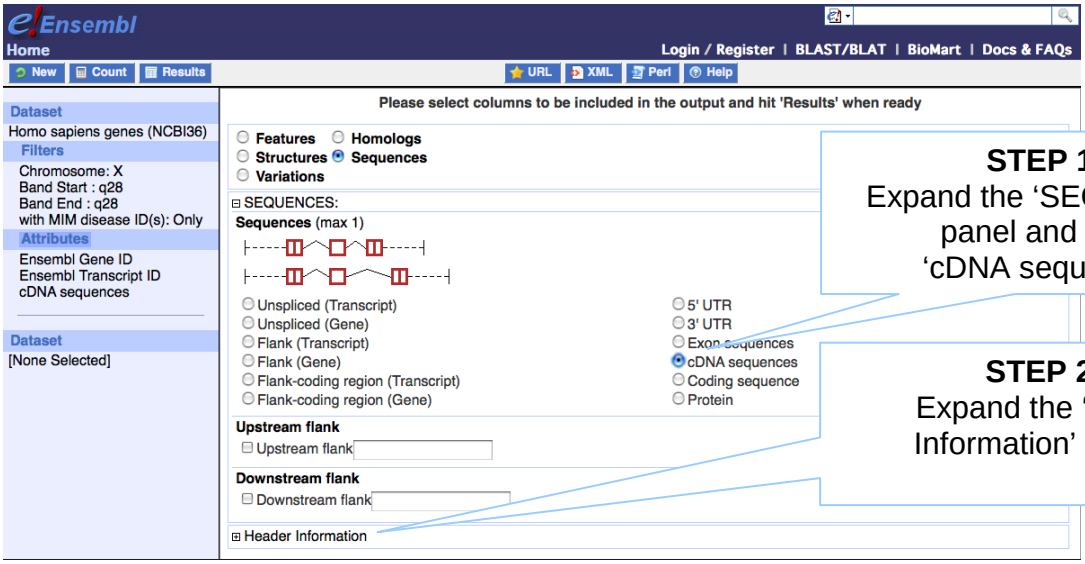
Ensembl Gene ID	Ensembl Transcript ID	Associated Gene Name	EntrezGene ID	MIM Morbid Accession	MIM Morbid Description
ENSG00000185010	ENST00000360256	F8	2157	306700	HEMOPHILIA A
ENSG00000185010	ENST00000360256	F8	2157	134500	FACTOR VIII DEFICIENCY
ENSG00000130826	ENST00000369550	DKC1	1736	300240	HOYERAAL-HREIDARSSON SYNDROME
ENSG00000130826	ENST00000369550	DKC1	1736	305000	DYSKERATOSIS CONGENITA, X-LINKED
ENSG00000073009	ENST00000369609	IKBKG	8517	308300	INCONTINENTIA PIGMENTI
ENSG00000073009	ENST00000369609	IKBKG	8517	300640	INVASIVE PNEUMOCOCCAL DISEASE, RECURRENT ISOLATED, 2
ENSG00000073009	ENST00000369609	IKBKG	8517	300636	ATYPICAL MYCOBACTERIOSIS, FAMILIAL, X-LINKED 1
ENSG00000073009	ENST00000369609	IKBKG	8517	300584	IMMUNODEFICIENCY WITHOUT ANHIDROTIC ECTODERMAL DYSPLASIA
ENSG00000073009	ENST00000369609	IKBKG	8517	300301	ECTODERMAL DYSPLASIA, ANHIDROTIC, WITH IMMUNODEFICIENCY, OSTEOPETROSIS,
ENSG00000073009	ENST00000369609	IKBKG	8517	300291	ECTODERMAL DYSPLASIA, HYPOHIDROTIC, WITH IMMUNE DEFICIENCY
ENSG00000073009	ENST00000369601	IKBKG	8517	308300	INCONTINENTIA PIGMENTI
ENSG00000073009	ENST00000369601	IKBKG	8517	300640	INVASIVE PNEUMOCOCCAL DISEASE, RECURRENT ISOLATED, 2
ENSG00000073009	ENST00000369601	IKBKG	8517	300636	ATYPICAL MYCOBACTERIOSIS, FAMILIAL, X-LINKED 1
ENSG00000073009	ENST00000369601	IKBKG	8517	300584	IMMUNODEFICIENCY WITHOUT ANHIDROTIC ECTODERMAL DYSPLASIA
ENSG00000073009	ENST00000369601	IKBKG	8517	300301	ECTODERMAL DYSPLASIA, ANHIDROTIC, WITH IMMUNODEFICIENCY, OSTEOPETROSIS,
ENSG00000073009	ENST00000369601	IKBKG	8517	300291	ECTODERMAL DYSPLASIA, HYPOHIDROTIC, WITH IMMUNE DEFICIENCY
ENSG00000073009	ENST00000369606	IKBKG	8517	308300	INCONTINENTIA PIGMENTI
ENSG00000073009	ENST00000369606	IKBKG	8517	300640	INVASIVE PNEUMOCOCCAL DISEASE, RECURRENT ISOLATED, 2
ENSG00000073009	ENST00000369606	IKBKG	8517	300636	ATYPICAL MYCOBACTERIOSIS, FAMILIAL, X-LINKED 1
ENSG00000073009	ENST00000369606	IKBKG	8517	300584	IMMUNODEFICIENCY WITHOUT ANHIDROTIC ECTODERMAL DYSPLASIA
ENSG00000073009	ENST00000369606	IKBKG	8517	300301	ECTODERMAL DYSPLASIA, ANHIDROTIC, WITH IMMUNODEFICIENCY, OSTEOPETROSIS,
ENSG00000073009	ENST00000369606	IKBKG	8517	300291	ECTODERMAL DYSPLASIA, HYPOHIDROTIC, WITH IMMUNE DEFICIENCY
ENSG00000073009	ENST00000369607	IKBKG	8517	308300	INCONTINENTIA PIGMENTI
ENSG00000073009	ENST00000369607	IKBKG	8517	300640	INVASIVE PNEUMOCOCCAL DISEASE, RECURRENT ISOLATED, 2
ENSG00000073009	ENST00000369607	IKBKG	8517	300636	ATYPICAL MYCOBACTERIOSIS, FAMILIAL, X-LINKED 1
ENSG00000073009	ENST00000369607	IKBKG	8517	300584	IMMUNODEFICIENCY WITHOUT ANHIDROTIC ECTODERMAL DYSPLASIA
ENSG00000073009	ENST00000369607	IKBKG	8517	300301	ECTODERMAL DYSPLASIA, ANHIDROTIC, WITH IMMUNODEFICIENCY, OSTEOPETROSIS,

Result Table 1



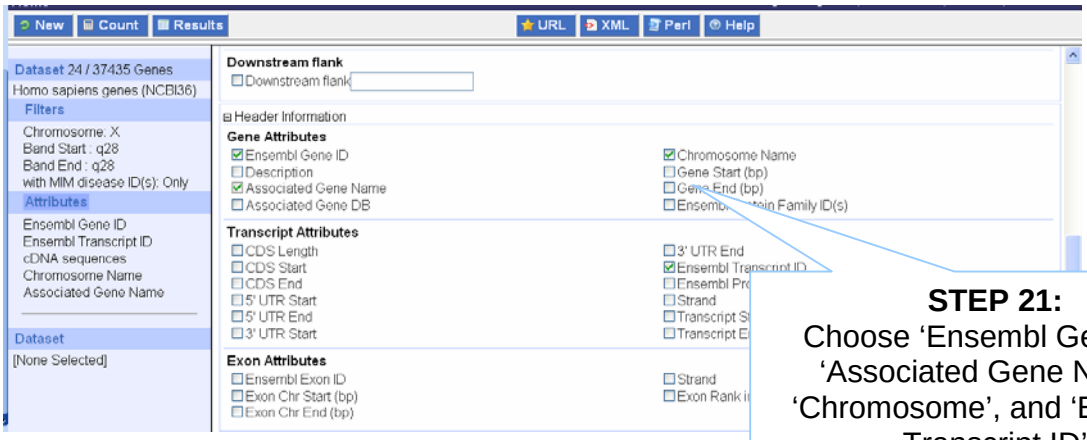
STEP 17:
To view sequences, go back to ‘Attributes’

STEP 18:
Select ’Sequences’ and then expand the ‘SEQUENCES’ section.

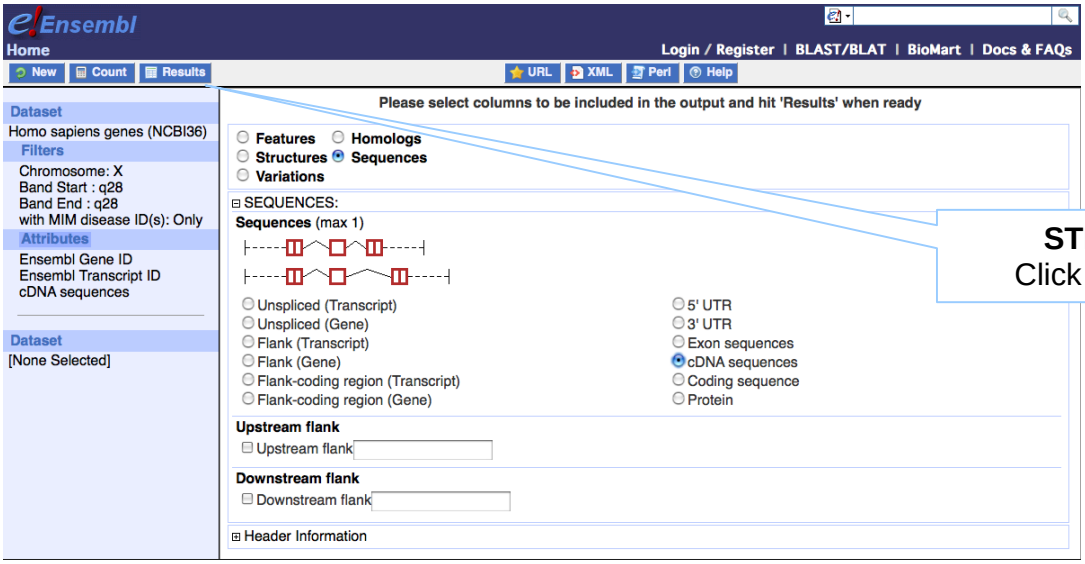


STEP 19:
Expand the 'SEQUENCES' panel and select 'cDNA sequences'.

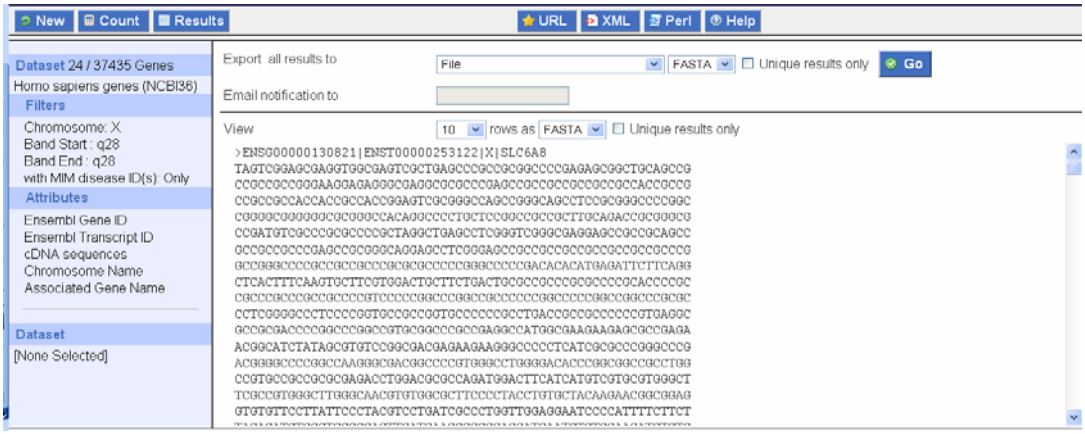
STEP 20:
Expand the 'Header Information' section.



STEP 21:
Choose 'Ensembl Gene ID', 'Associated Gene Name', 'Chromosome', and 'Ensembl Transcript ID'



STEP 22:
Click ‘Results’



Again, View ALL rows as FASTA for the full list… (make sure pop-up moduleer is off).

RESULTS

>Header: Gene ID, Transcript ID, Chromosome and Gene Name

```
>ENSG00000073009|ENST00000369601|X|IKBKG
AGCCCGTTCCTGCTCCGCGCTTCTGGAGCACTGGCCAAGGCGGGCCGATTCAGGACCCAG
GTTACTTGGGCGGCGAGCTGGACTGTTTCTACTCCTCCCTCCTCCTCCACTGCGGGGTCT
GACCCTACTCCTTGTGTGAGGACTCCTCTAGTTCAGAGACATATTCTGTTCACCAAACTT
GACTGCGCTCTATCGAGGTCGTTAAATTCTTCGGAAATGCCTCACATATAGTTTGGCAGC
TAGCCCTTGCCCTGTTGGATGAATAGGCACCTCTGGAAGAGCCAACTGTGTGAGATGGTG
CAGCCCAGTGGTGGCCCGGCAGCAGATCAGGACGTACTGGGCGAAGAGTCTCCTCTGGGG
AAGCCAGCCATGCTGCACCTGCCTTCAGAACAGGGCGCTCCTGAGACCCTCCAGCGCTGC
CTGGAGGAGAATCAAGAGCTCCGAGATGCCATCCGGCAGAGCAACCAGATTCTGCGGGAG
CGCTGCGAGGAGCTTCTGCATTTCCAAGCCAGCCAGAGGGAGGAGAAGGAGTTCCTCATG
TGCAAGTTCCAGGAGGCCAGGAAACTGGTGGAGAGACTCGGCCTGGAGAAGCTCGATCTG
AAGAGGCAGAAGGAGCAGGCTCTGCGGGAGGTGGAGCACCTGAAGAGATGCCAGCAGCAG
ATGGCTGAGGACAAGGCCTCTGTGAAAGCCCAGGTGACGTCCTTGCTCGGGGAGCTGCAG
GAGAGCCAGAGTCGCTTGGAGGCTGCCACTAAGGAATGCCAGGCTCTGGAGGGTCGGGCC
CGGGCGGCCAGCGAGCAGGCGCGGCAGCTGGAGAGTGAGCGCGAGGCGCTGCAGCAGCAG
CACAGCGTGCAGGTGGACCAGCTGCGCATGCAGGGCCAGAGCGTGGAGGCCGCGCTCCGC
ATGGAGCGCCAGGCCGCCTCGGAGGAGAAGAGGAAGCTGGCCCAGTTGCAGGTGGCCTAT
CACCAGCTCTTCCAAGAATACGACAACCACATCAAGAGCAGCGTGGTGGGCAGTGAGCGG
AAGCGAGGAATGCAGCTGGAAGATCTCAAACAGCAGCTCCAGCAGGCCGAGGAGGCCCTG
GTGGCCAAACAGGAGGTGATCGATAAGCTGAAGGAGGAGGCCGAGCAGCACAAGATTGTG
ATGGAGACCGTTCCGGTGCTGAAGGCCCAGGCGGATATCTACAAGGCGGACTTCCAGGCT
GAGAGGCAGGCCCGGGAGAAGCTGGCCGAGAAGAAGGAGCTCCTGCAGGAGCAGCTGGAG
CAGCTGCAGAGGGAGTACAGCAAACTGAAGGCCAGCTGTCAGGAGTCGGCCAGGATCGAG
GACATGAGGAAGCGGCATGTCGAGGTCTCCCAGGCCCCCTTGCCCCCCGCCCCTGCCTAC
CTCTCCTCTCCCCTGGCCCTGCCCAGCCAGAGGAGGAGCCCCCCCGAGGAGCCACCTGAC
TTCTGCTGTCCCAAGTGCCAGTATCAGGCCCCTGATATGGACACCCTGCAGATACATGTC
ATGGAGTGCATTGAGTAGGGCCGGCCAGTGCAAGGCCACTGCCTGCCGAGGACGTGCCCG
GGACCGTGCAGTCTGCGCTTTCCTCTCCCGCCTGCCTAGCCCAGGATGAAGGGCTGGGTG
GCCACAACTGGGATGCCACCTGGAGCCCCACCCAGGAGCTGGCCGCGGCACCTTACGCTT
CAGCTGTTGATCCGCTGGTCCCCTCTTTTGGGGTAGATGCGGCCCCGATCAGGCCTGACT
CGCTGCTCTTTTTGTTCCCTTCTGTCTGCTCGAACCACTTGCCTCGGGCTAATCCCTCCC
TCTTCCTCCACCCGGCACTGGGGAAGTCAAGAATGGGGCCTGGGGCTCTCAGGGAGAACT
GCTTCCCCTGGCAGAGCTGGGTGGCAGCTCTTCCTCCCACCGGACACCGACCCGCCCGCT
GCTGTGCCCTGGGAGTGCTGCCCTCTTACCATGCACACGGGTGCTCTCCTTTTGGGCTGC
ATGCTATTCCATTTTGCAGCCAGACCGATGTGTATTTAACCAGTCACTATTGATGGACAT
TTGGGTTGTTTCCCATCTTTTTGTTACCATAAATAATGGCATAGTAAAAATCCTTGTGCA
TT
>ENSG00000126895|ENST00000358927|X|AVPR2
TTCACGCCACCGCCCAGCTGCCCAGGAGCCCAGCCAGGACTGGCCATACTGCCACCGACA
CGTGCACACACGCCAACAGGCATCTGCCATGCTGGCATCTCTATAAGGGCTCCAGTCCAG
AGACCCTGGGCCATTGAACTTGCTCCTCAGGCAGAGGCTGAGTCCGCACATCACCTCCAG
GCCCTCAGAACACCTGCCCCAGCCCCACCATGCTCATGGCGTCCACCACTTCCGCTGTGC
CTGGGCATCCCTCTCTGCCCAGCCTGCCCAGCAACAGCAGCCAGGAGAGGCCACTGGACA
CCCGGGACCCGCTGCTAGCCCGGGCGGAGCTGGCGCTGCTCTCCATAGTCTTTGTGGCTG
TGGCCCTGAGCAATGGCCTGGTGCTGGCGGCCCTAGCTCGGCGGGGCCGGCGGGGCCACT
GGGCACCCATACACGTCTTCATTGGCCACTTGTGCCTGGCCGACCTGGCCGTGGCTCTGT
TCCAAGTGCTGCCCCAGCTGGCCTGGAAGGCCACCGACCGCTTCCGTGGGCCAGATGCCC
TGTGTCGGGCCGTGAAGTATCTGCAGATGGTGGGCATGTATGCCTCCTCCTACATGATCC
```

cDNA 1

cDNA 2

Glossary

Attributes *(In BioMart)* Information to attach to the geneset selected, either sequence or column headers

Biotype *(In BioMart)* Gene type (i.e. coding, non-coding)

Entrez Gene NCBI searchable database of gene sequences. Accession numbers in GenBank agree with DDBJ and EMBL.

External Gene ID An ID or accession number in a database apart from Ensembl

Filters *(In BioMart)* Information applied to narrow the selection, such as filtering the entire geneset for a species down to only genes on a specific chromosomal region with a GO term or Interpro domain. IDs can be used as 'filters' to select a gene set by an ID list.

GFF A file format often used in genetics, applicable across programs and databases

HapMap An international partnership committed to the development of a haplotype map describing common patterns of the human genome. http://www.hapmap.org/

HGNC – The HUGO Gene Nomenclature Committee A committee focused on the determination of one unique symbol for every human gene. www.genenames.org/

InterPro A database of common protein motifs and domains, accesses information across a large number of protein databases. http://www.ebi.ac.uk/interpro/

OMIM 'Online Mendelian Inheritance in Man'. A database of phenotypic information for human. http://www.ncbi.nlm.nih.gov/omim/

MSD 'Macromolecular Structure Database' A collection of protein and other macromolecular structures, in part from the PDB (Protein Data Bank). http://www.ebi.ac.uk/msd/

WormBase An in-depth look at the *C. elegans* genome and other worm genomes. A manually curated gene set, genome browser and WormMart are available. http://www.wormbase.org/

What to do next

Watch the related video!

www.ensembl.org/info/website/tutorials/index.html

'Introduction to BioMart'