BioMart

Mining data- worked example

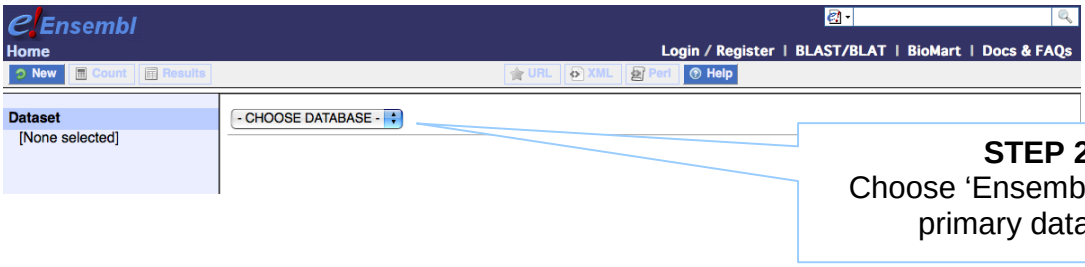
The human gene encoding Glucose-6-phosphate dehydrogenase (G6PD) is located on chromosome X in cytogenetic band q28.

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

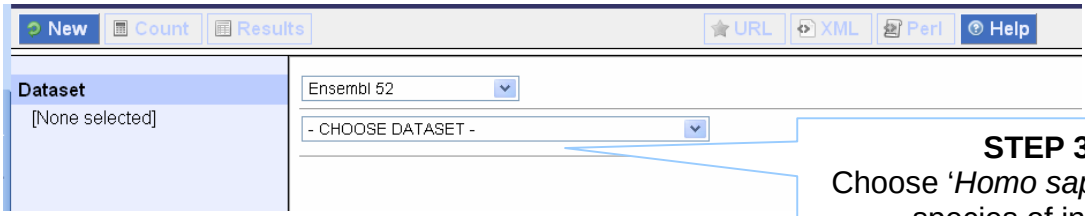
What are their cDNA sequences?

Follow the worked example below to answer these questions.

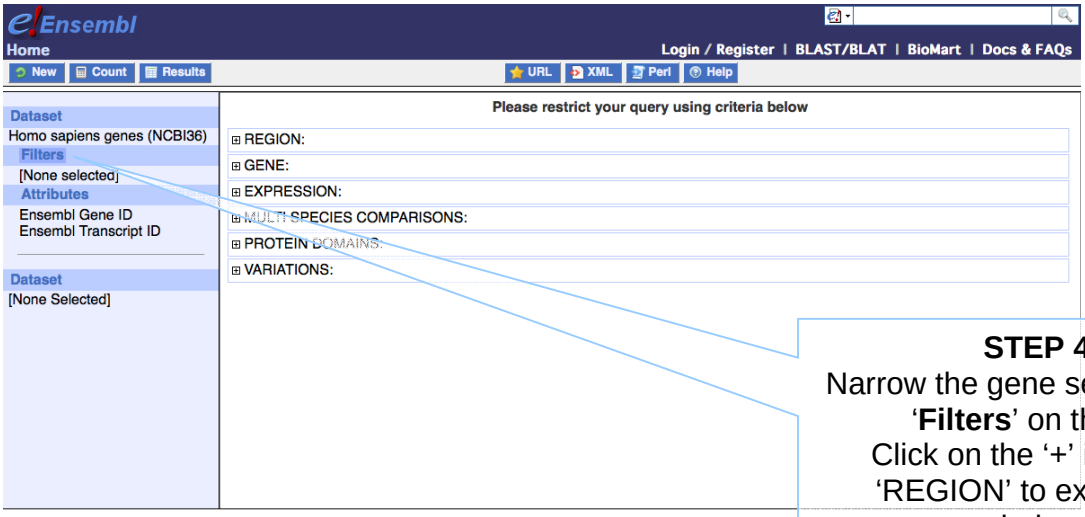
Step 1: Either click on 'BioMart' in the top right header bar of the Ensembl home page, or go to http://www.biomart.org/ and click on the 'MartView' tab.



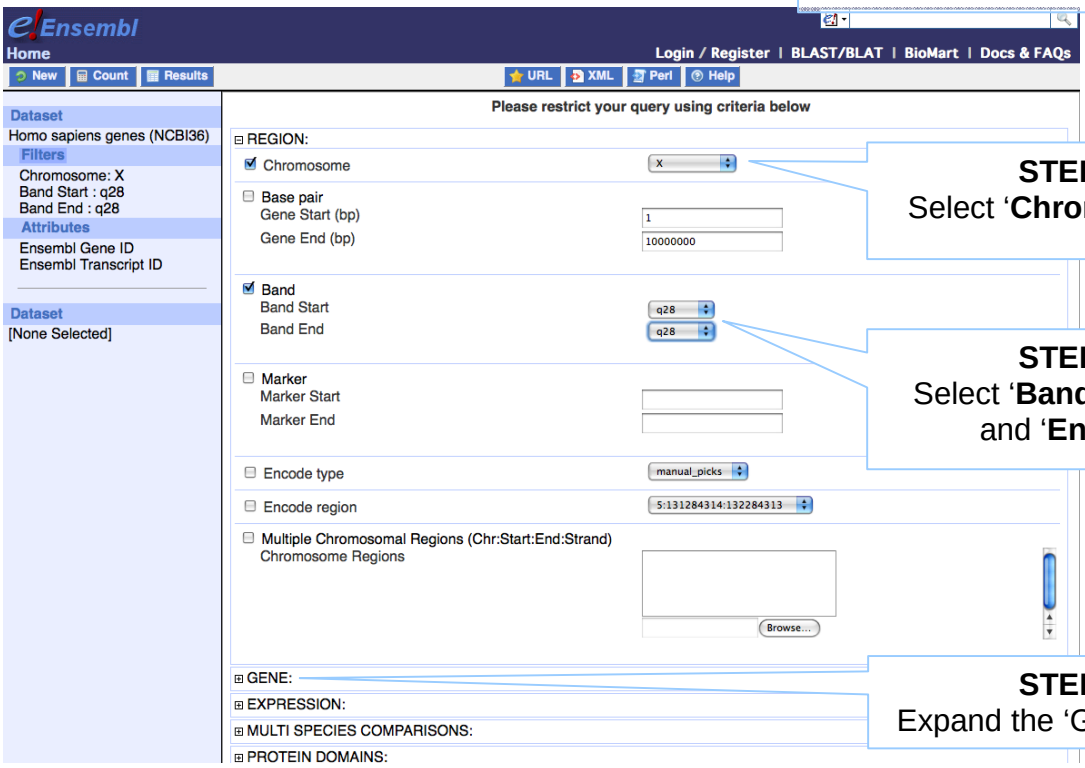
STEP 2:
Choose 'Ensembl 52' as the primary database.



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



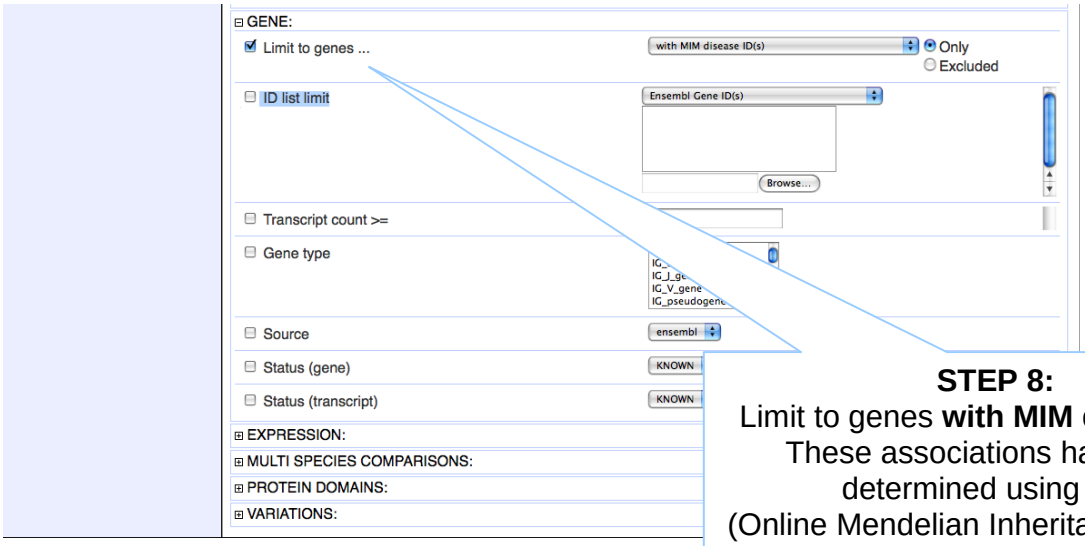
STEP 4:
Narrow the gene set by clicking '**Filters**' on the left.
Click on the '+' in front of 'REGION' to expand the choices.



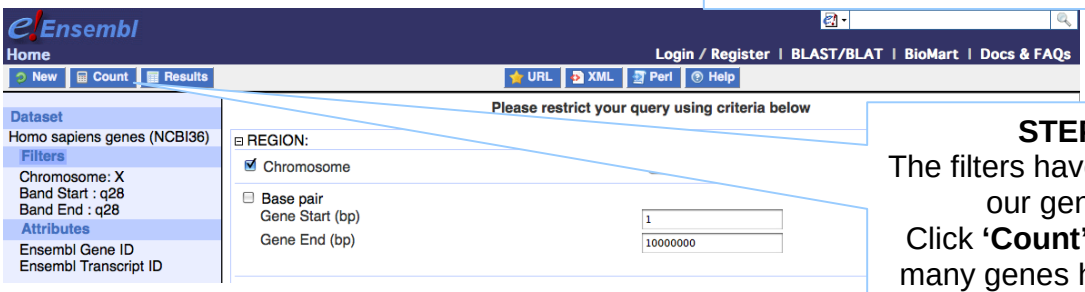
STEP 5:
Select '**Chromosome X**'

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

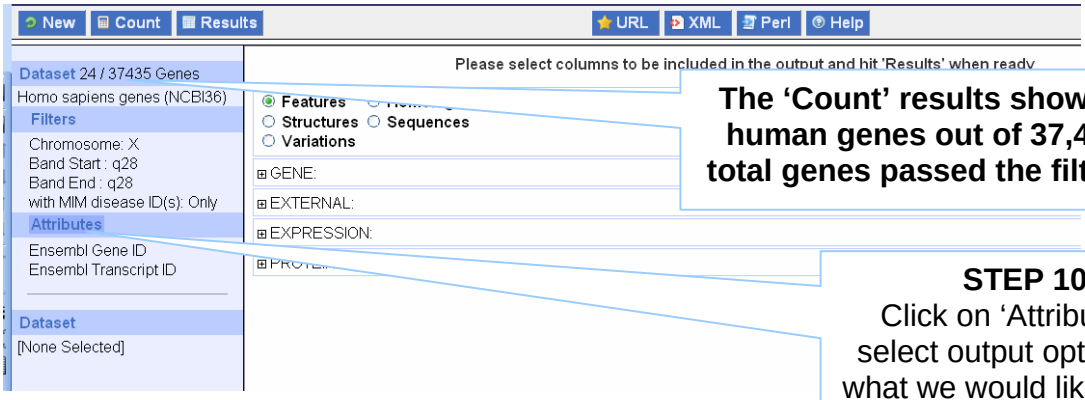
STEP 7:
Expand the 'GENE' panel.



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/



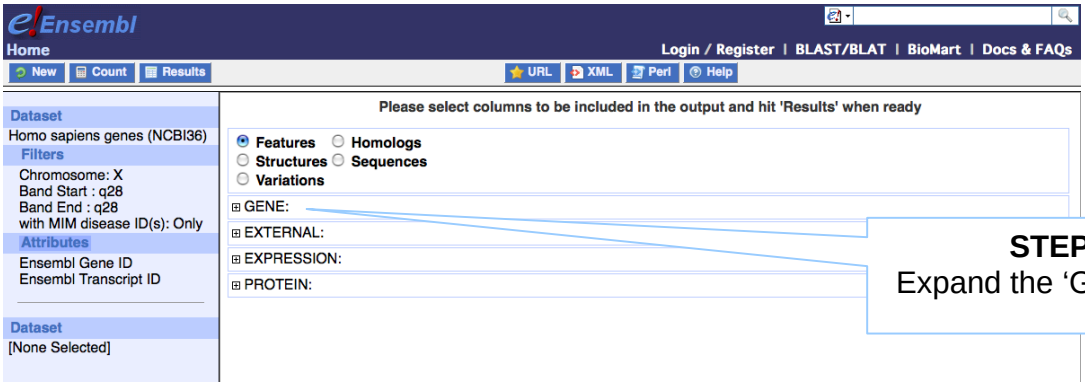
STEP 9:
The filters have determined our gene set.
Click **'Count'** to see how many genes have passed these filters.



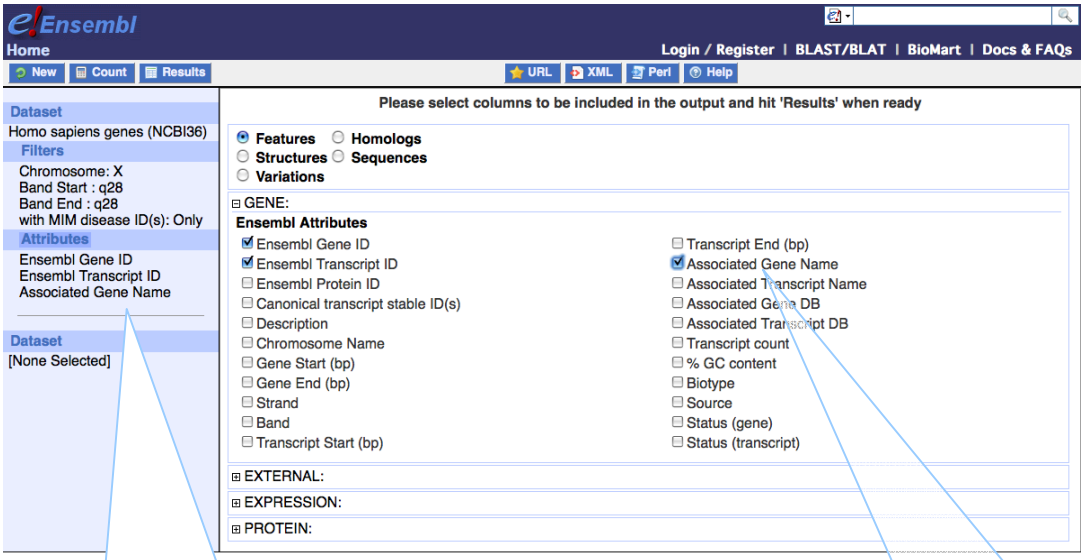
The 'Count' results show 24 human genes out of 37,435 total genes passed the filters.

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.

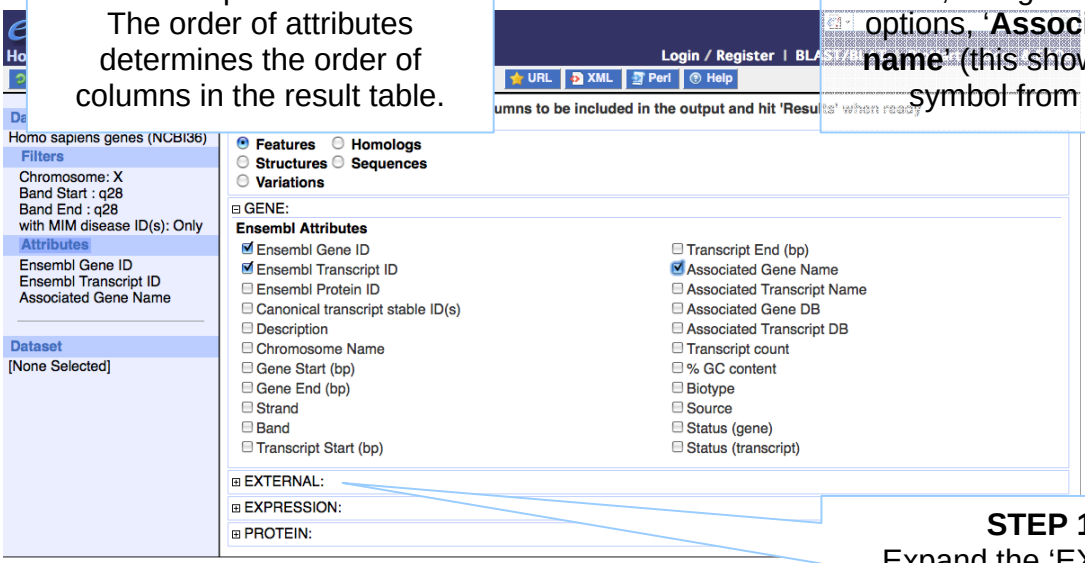


STEP 11:
Expand the 'GENE' panel.

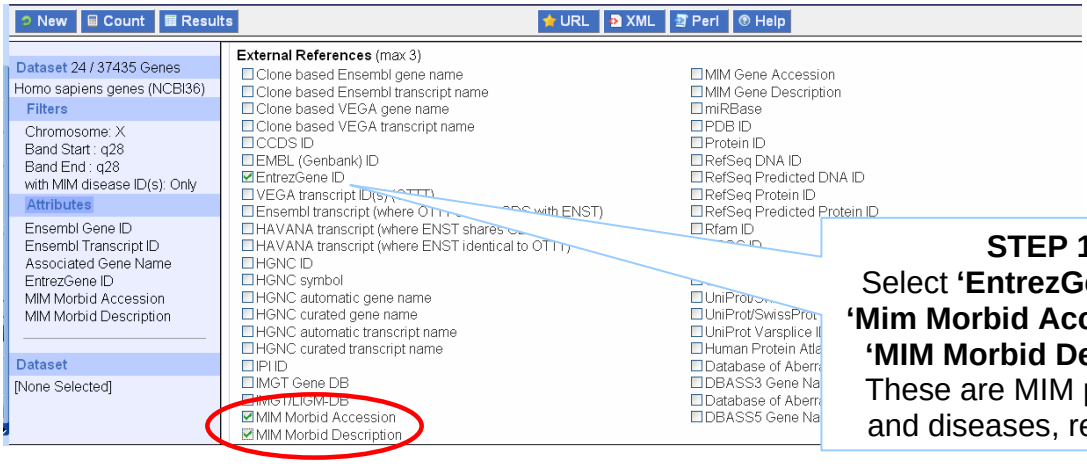


Note the summary of selected options.
The order of attributes determines the order of columns in the result table.

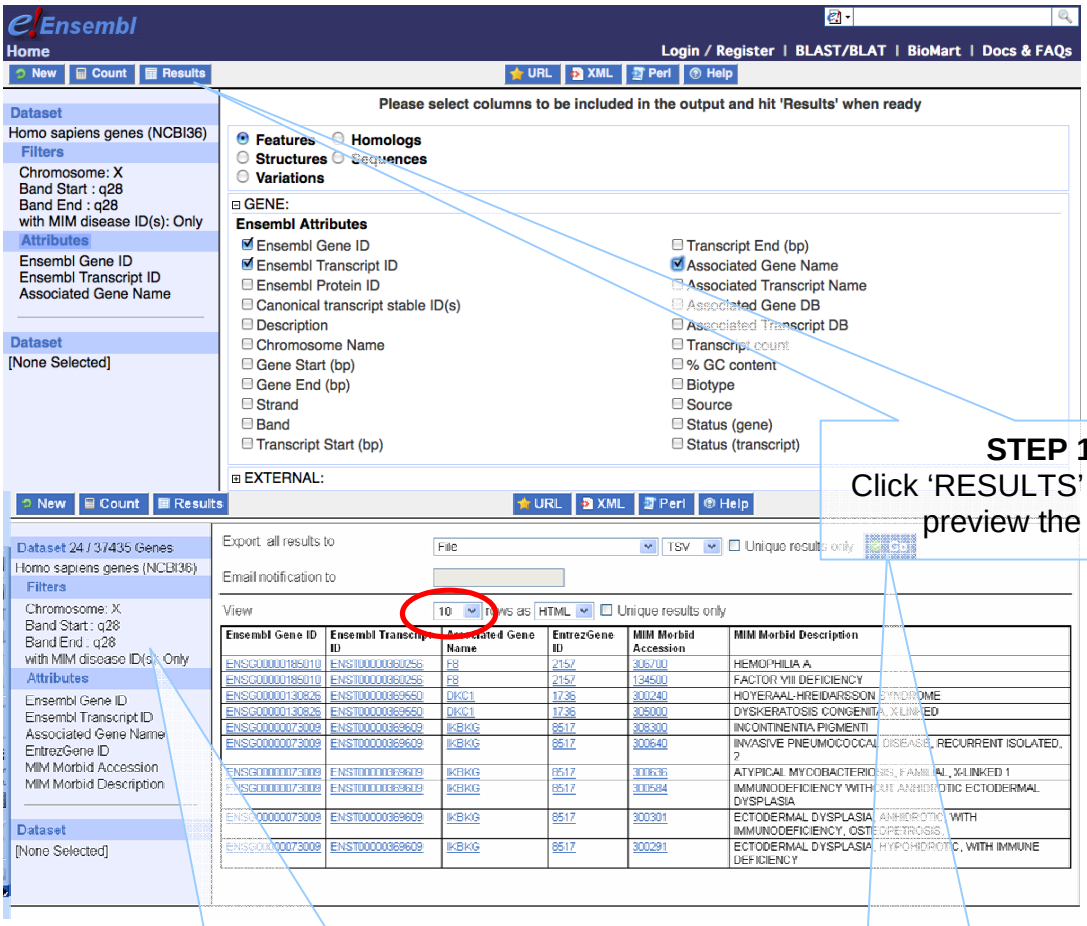
STEP 12:
Select, along with the default options, '**Associated Gene name**' (this shows the gene symbol from HGNC).



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.



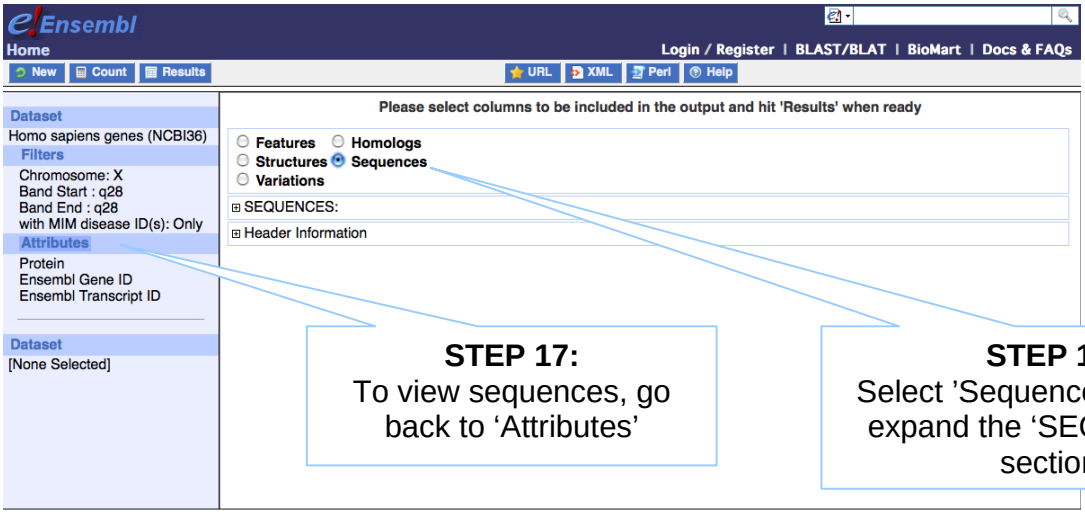
STEP 15:
Click 'RESULTS' at the top to preview the output.

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.

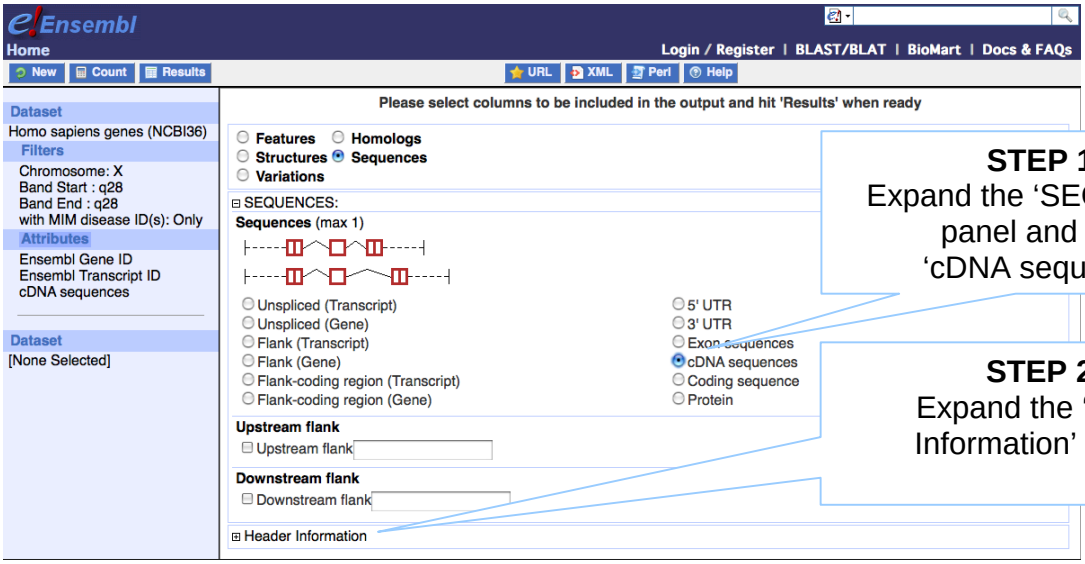
Ensembl Gene ID	Ensembl Transcript ID	Associated Gene Name	EntrezGene ID	MIM Morbid Accession	MIM Morbid Description
ENSG00000185010	ENST00000360256	F8	2157	306700	HEMOPHILIA A
ENSG00000185010	ENST00000360256	F8	2157	134500	FACTOR VIII DEFICIENCY
ENSG00000130826	ENST00000369550	DKC1	1736	300240	HOYERAAL-HREIDARSSON 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



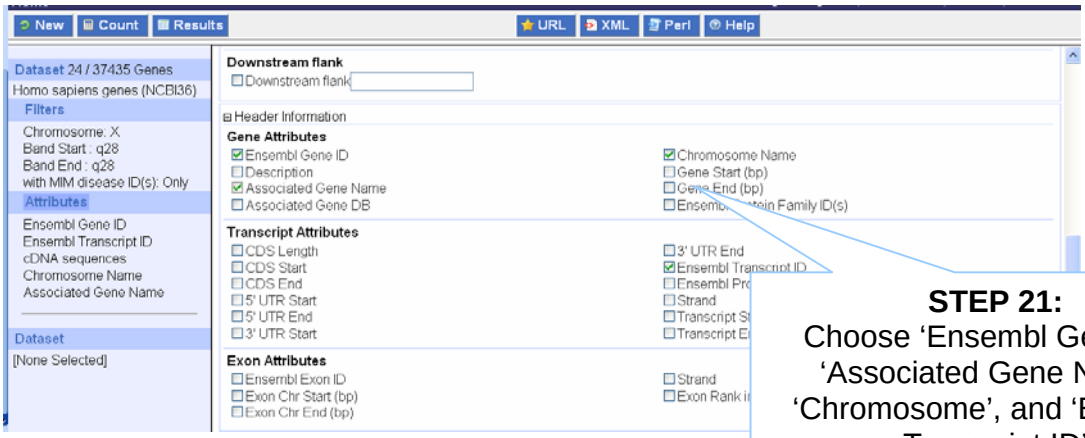
STEP 17:
To view sequences, go back to ‘Attributes’

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

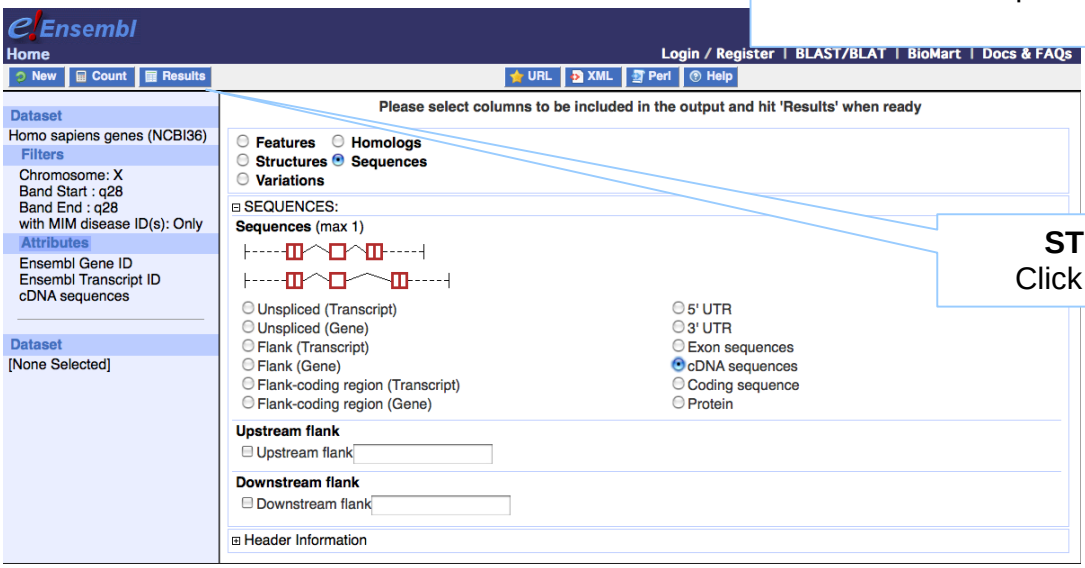


STEP 19:
Expand the 'SEQUENCES' panel and select 'cDNA sequences'.

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 blocker 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
TGGCCATGACGCTGGACCGCCACCGTGCCATCTGCCGTCCCATGCTGGCGTACCGCCATG
```

cDNA 1

cDNA 2