UTM UNIVERSITI TEKNOLOGI MALAYSIA SCHOOL OF COMPUTING Faculty of Engineering

Semester II 2021/2022

Subject : Bioinformatics I (SECB2103)
Section : 01 – Dr Haslina Hashim
Topic : Lab 02 – Access to Biological Databases

Name : Gui Yu Xuan
Matric ID : A20EC0039

1. The purpose of this problem is to introduce you to using Entrez and related NCBI resources.

a) How many human proteins are bigger than 300,000 daltons?
There are 2744 human proteins are bigger than 300,000 daltons.

b) What is the longest human protein?
Titin is the longest human protein.

There are several different ways to solve these questions.

(1) Try to first limit your search to human by using TaxBrowser. From the home page of NCBI select the alphabetical list of resources and find the Taxonomy Browser and the entry for human. Then follow the link to Entrez Protein, where all the results will be limited to human.

(2) Enter a command in the format xxxxxx:yyyyyy[molwt] to restrict the output to a certain number of Daltons; for example, 002000:010000[molwt] will select proteins of molecular weight 2,000 to 10,000.

(3) As a different approach, search 30000:50000[Sequence Length]

(4) You can read more about titin (NP_596869.4), the longest human protein, in an NCBI newsletter (WebLink 2.73 ftp://ftp.ncbi.nih.gov/genbank/gbrel.txt). While the average protein has a length of several hundred amino acids, incredibly titin is 34,423 amino acids in length.

(5) Explore additional ways to limit Entrez searches by using an NCBI Handbook chapter (http://www.ncbi.nlm.nih.gov/books/NBK44864/)

Write a report of how you would answer the question in this week's session.

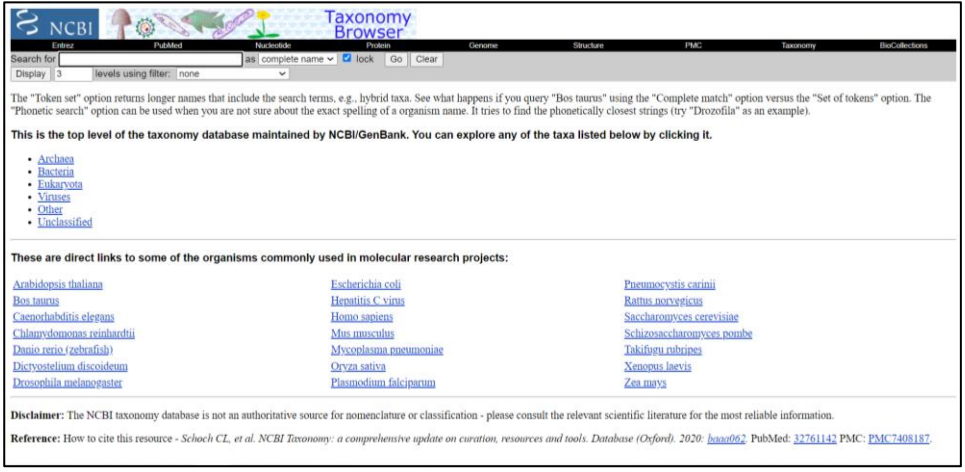
Include screen captures as well as detailed explanation and observations.

All screen captures should be labelled as figures and numbered accordingly.

All tables should be labelled as tables and numbered accordingly.

Hint: What is the Latin name for humans? Homo sapiens.

Here is the screen shot of the Taxonomy Browser.



The "Token set" option returns longer names that include the search terms, e.g., hybrid taxa. See what happens if you query "Bos taurus" using the "Complete match" option versus the "Set of tokens" option. The "Phonetic search" option can be used when you are not sure about the exact spelling of a organism name. It tries to find the phonetically closest strings (try "Drozofila" as an example).

This is the top level of the taxonomy database maintained by NCBI/GenBank. You can explore any of the taxa listed below by clicking it.

- Archaea
- Bacteria
- Eukaryota
- Viruses
- Other
- Unclassified

These are direct links to some of the organisms commonly used in molecular research projects:

Arabidopsis thaliana
Bos taurus
Caenorhabditis elegans
Chlamydomonas reinhardtii
Danio rerio (zebrafish)
Dictyostelium discoideum
Drosophila melanogaster

Escherichia coli
Hepatitis C virus
Homo sapiens
Mus musculus
Mycoplasma pneumoniae
Oryza sativa
Plasmodium falciparum

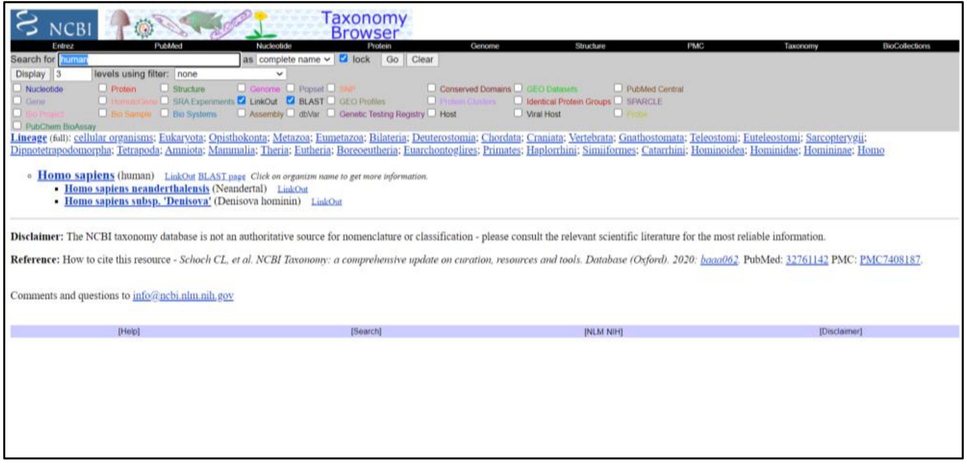
Pneumocystis carinii
Rattus norvegicus
Saccharomyces cerevisiae
Schizosaccharomyces pombe
Takifugu rubripes
Xenopus laevis
Zea mays

Disclaimer: The NCBI taxonomy database is not an authoritative source for nomenclature or classification - please consult the relevant scientific literature for the most reliable information.

Reference: How to cite this resource - *Schoch CL, et al. NCBI Taxonomy: a comprehensive update on curation, resources and tools. Database (Oxford). 2020: baaa062*. PubMed: 32761142 PMC: PMC7408187.

Figure 1

After typing "human" in the searching bar, then select the "home sapiens(human).


NCBI
Taxonomy Browser
Entrez
PubMed
Nucleotide
Protein
Genome
Structure
PMC
Taxonomy
BioCollections
Search for human
as complete name
☑ lock
Go
Clear
Display 3 levels using filter: none
□ Nucleotide □ Protein □ Structure □ Genome □ Popset □ SNP □ Conserved Domains □ GEO Datasets □ PubMed Central
□ Gene □ HomoloGene □ SRA Experiments ☑ LinkOut ☑ BLAST □ GEO Profiles □ Protein Clusters □ Identical Protein Groups □ SPARCLE
□ Bio Project □ Bio Sample □ Bio Systems □ Assembly □ dbVar □ Genetic Testing Registry □ Host □ Viral Host □ Probe
□ PubChem BioAssay
[Help]
[Search]
[NLM NIH]
[Disclaimer]

Lineage (full): cellular organisms; Eukaryota; Opisthokonta; Metazoa; Eumetazoa; Bilateria; Deuterostomia; Chordata; Craniata; Vertebrata; Gnathostomata; Teleostomi; Euteleostomi; Sarcopterygii; Dipnotetrapodomorpha; Tetrapoda; Amniota; Mammalia; Theria; Eutheria; Boreoeutheria; Euarchontoglires; Primates; Haplorrhini; Simiiformes; Catarrhini; Hominoidea; Hominidae; Homininae; Homo

- **Homo sapiens** (human) LinkOut BLAST page *Click on organizm name to get more information.*
  - **Homo sapiens neanderthalensis** (Neandertal) LinkOut
  - **Homo sapiens subsp. 'Denisova'** (Denisova hominin) LinkOut

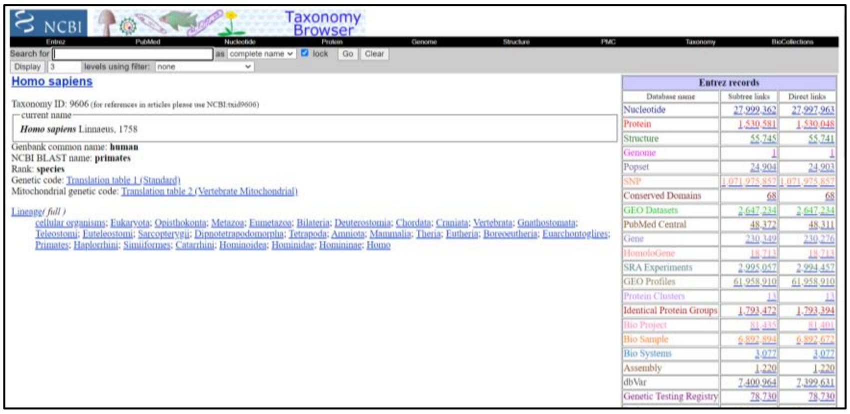
Disclaimer: The NCBI taxonomy database is not an authoritative source for nomenclature or classification - please consult the relevant scientific literature for the most reliable information.

Reference: How to cite this resource - *Schoch CL, et al. NCBI Taxonomy: a comprehensive update on curation, resources and tools. Database (Oxford). 2020: baaa062*. PubMed: 32761142 PMC: PMC7408187.

Comments and questions to info@ncbi.nlm.nih.gov

Figure 2

At the right side of the website, click the Subtree links that under the Entrez Records.



Display 3 levels using filter: none

Homo sapiens

Taxonomy ID: 9606 (for references in articles please use NCBI:txid9606)

current name

Homo sapiens Linnaeus, 1758

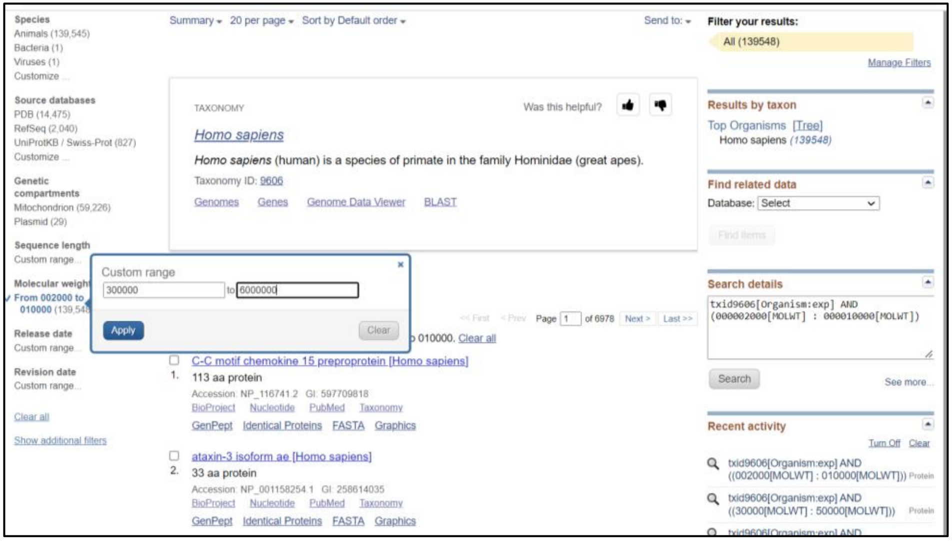
Genbank common name: **human**
NCBI BLAST name: **primates**
Rank: **species**
Genetic code: Translation table 1 (Standard)
Mitochondrial genetic code: Translation table 2 (Vertebrate Mitochondrial)

Lineage(*full*)
cellular organisms; Eukaryota; Opisthokonta; Metazoa; Eumetazoa; Bilateria; Deuterostomia; Chordata; Craniata; Vertebrata; Gnathostomata; Teleostomi; Euteleostomi; Sarcopterygii; Dipnotetrapodomorpha; Tetrapoda; Amniota; Mammalia; Theria; Eutheria; Boreoeutheria; Euarchontoglires; Primates; Haplorrhini; Simiiformes; Catarrhini; Hominoidea; Hominidae; Homininae; Homo

Entrez records		
Database name	Subtree links	Direct links
Nucleotide	27,999,362	27,997,963
Protein	1,530,581	1,530,048
Structure	55,745	55,741
Genome	1	1
Popset	24,904	24,903
SNP	1,071,975,857	1,071,975,857
Conserved Domains	68	68
GEO Datasets	2,647,234	2,647,234
PubMed Central	48,372	48,311
Gene	230,349	230,276
HomoloGene	18,713	18,713
SRA Experiments	2,995,057	2,994,457
GEO Profiles	61,958,910	61,958,910
Protein Clusters	13	13
Identical Protein Groups	1,793,472	1,793,394
Bio Project	81,435	81,401
Bio Sample	6,892,894	6,892,672
Bio Systems	3,077	3,077
Assembly	1,220	1,220
dbVar	7,400,964	7,399,631
Genetic Testing Registry	78,730	78,730

Figure 3

At the left side of the website, select the molecular weight and enter the range from 300,000 to 6,000,000.



Species
Animals (139,545)
Bacteria (1)
Viruses (1)
Customize ...

Source databases
PDB (14,475)
RefSeq (2,040)
UniProtKB / Swiss-Prot (827)
Customize ...

Genetic compartments
Mitochondrion (59,226)
Plasmid (29)

Sequence length
Custom range...

Molecular weight
✓ From 002000 to 010000 (139,548)

Release date
Custom range...

Revision date
Custom range...

Clear all

Show additional filters

Custom range
300000 to 6000000

TAXONOMY

Homo sapiens

Homo sapiens (human) is a species of primate in the family Hominidae (great apes).

Taxonomy ID: 9606

Genomes Genes Genome Data Viewer BLAST

<< First < Prev Page 1 of 6978 Next > Last >>

...o 010000. Clear all

☐ C-C motif chemokine 15 preproprotein [Homo sapiens]
1. 113 aa protein
Accession: NP_116741.2 GI: 597709818
BioProject Nucleotide PubMed Taxonomy
GenPept Identical Proteins FASTA Graphics

☐ ataxin-3 isoform ae [Homo sapiens]
2. 33 aa protein
Accession: NP_001158254.1 GI: 258614035
BioProject Nucleotide PubMed Taxonomy
GenPept Identical Proteins FASTA Graphics

Filter your results:
All (139548)

Results by taxon
Top Organisms [Tree]
Homo sapiens *(139548)*

Find related data
Database: Select

Search details

```
txid9606[Organism:exp] AND
(000002000[MOLWT] : 000010000[MOLWT])
```

Recent activity

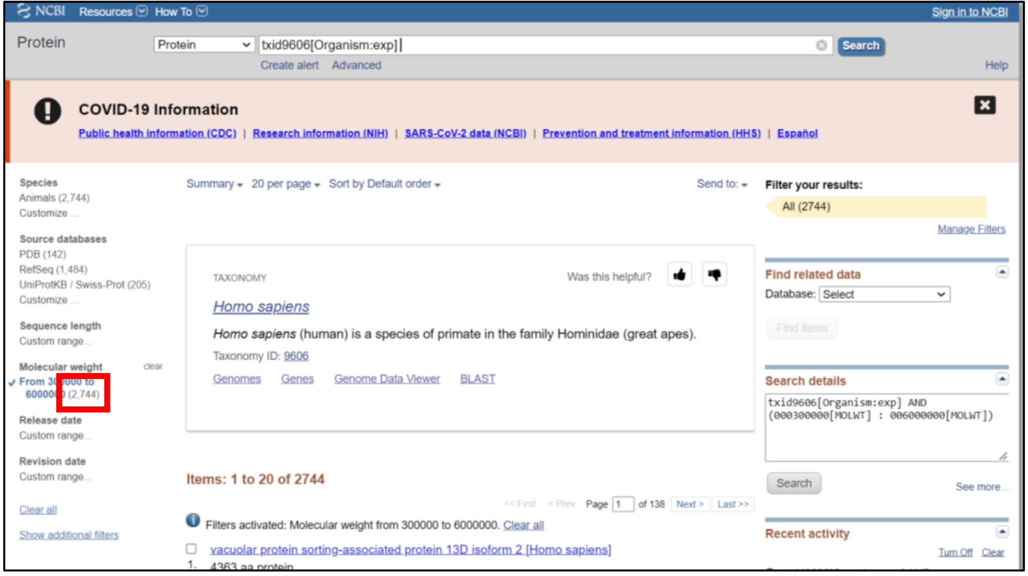
txid9606[Organism:exp] AND ((002000[MOLWT] : 010000[MOLWT])) Protein

txid9606[Organism:exp] AND ((30000[MOLWT] : 50000[MOLWT])) Protein

txid9606[Organism:exp] AND

Figure 4

The result is shown. 2744 human proteins are bigger than 300,000 daltons.



Protein

Protein: txid9606[Organism:exp]

COVID-19 Information

Public health information (CDC) | Research information (NIH) | SARS-CoV-2 data (NCBI) | Prevention and treatment information (HHS) | Español

Species
Animals (2,744)
Customize ...

Source databases
PDB (142)
RefSeq (1,484)
UniProtKB / Swiss-Prot (205)
Customize ...

Sequence length
Custom range...

Molecular weight clear
✓ From 300000 to 6000000 (2,744)

Release date
Custom range...

Revision date
Custom range...

Clear all

Show additional filters

TAXONOMY

Homo sapiens

Homo sapiens (human) is a species of primate in the family Hominidae (great apes).

Taxonomy ID: 9606

Genomes Genes Genome Data Viewer BLAST

Items: 1 to 20 of 2744

<< First < Prev Page 1 of 138 Next > Last >>

Filters activated: Molecular weight from 300000 to 6000000. Clear all

☐ vacuolar protein sorting-associated protein 13D isoform 2 [Homo sapiens]
1. 4363 aa protein

Filter your results:
All (2744)

Find related data
Database: Select

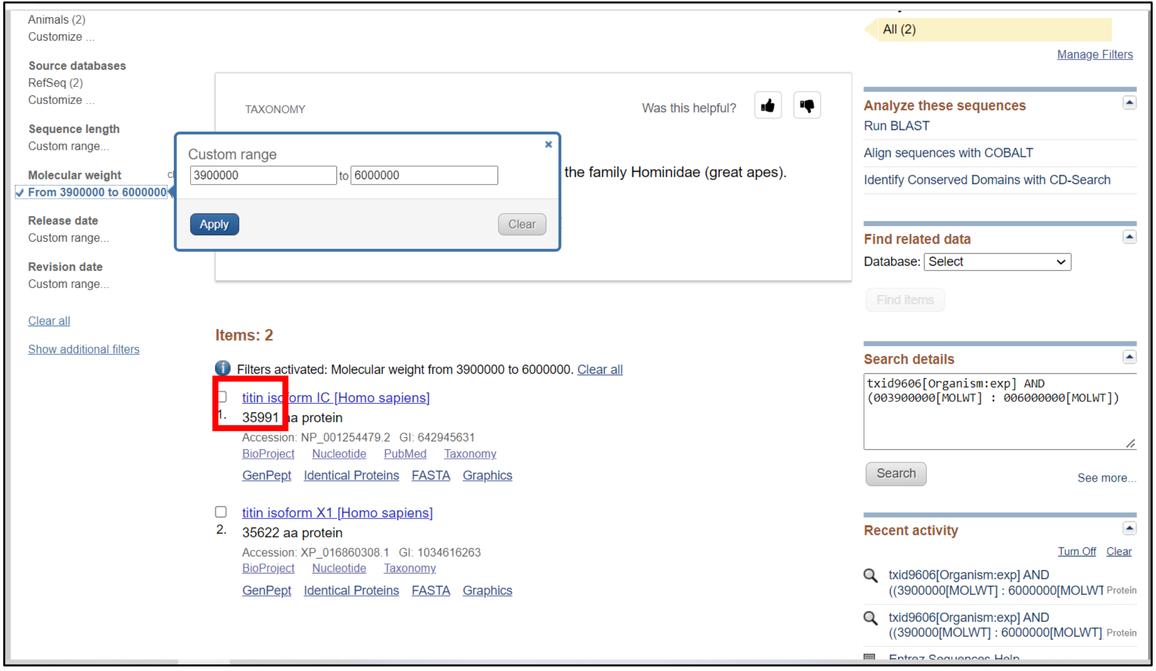
Search details

```
txid9606[Organism:exp] AND
(000300000[MOLWT] : 006000000[MOLWT])
```

Recent activity

Figure 5

To see the largest protein, the molecular weight needs to change from 3,900,000 to 6,000,000. Then, you may see the result where titin is the longest human protein.



Animals (2)
Customize ...

Source databases
RefSeq (2)
Customize ...

Sequence length
Custom range...

Molecular weight
✓ From 3900000 to 6000000

Release date
Custom range...

Revision date
Custom range...

Clear all

Show additional filters

TAXONOMY

Was this helpful?

Custom range
3900000 to 6000000
Apply Clear

the family Hominidae (great apes).

Items: 2

Filters activated: Molecular weight from 3900000 to 6000000. Clear all

□ 1. titin isoform IC [Homo sapiens]
35991 aa protein
Accession: NP_001254479.2 GI: 642945631
BioProject Nucleotide PubMed Taxonomy
GenPept Identical Proteins FASTA Graphics

□ 2. titin isoform X1 [Homo sapiens]
35622 aa protein
Accession: XP_016860308.1 GI: 1034616263
BioProject Nucleotide Taxonomy
GenPept Identical Proteins FASTA Graphics

All (2)
Manage Filters

Analyze these sequences
Run BLAST
Align sequences with COBALT
Identify Conserved Domains with CD-Search

Find related data
Database: Select
Find items

Search details

```
txid9606[Organism:exp] AND
(003900000[MOLWT] : 006000000[MOLWT])
```

Search
See more...

Recent activity
Turn Off Clear

txid9606[Organism:exp] AND ((3900000[MOLWT] : 6000000[MOLWT] Protein

txid9606[Organism:exp] AND ((390000[MOLWT] : 6000000[MOLWT] Protein

Entrez Sequences Help

Figure 6

2. The purpose of this problem is to obtain information from the NCBI website.

The RefSeq accession number of human beta globin protein is NP_000509. Go to NCBI

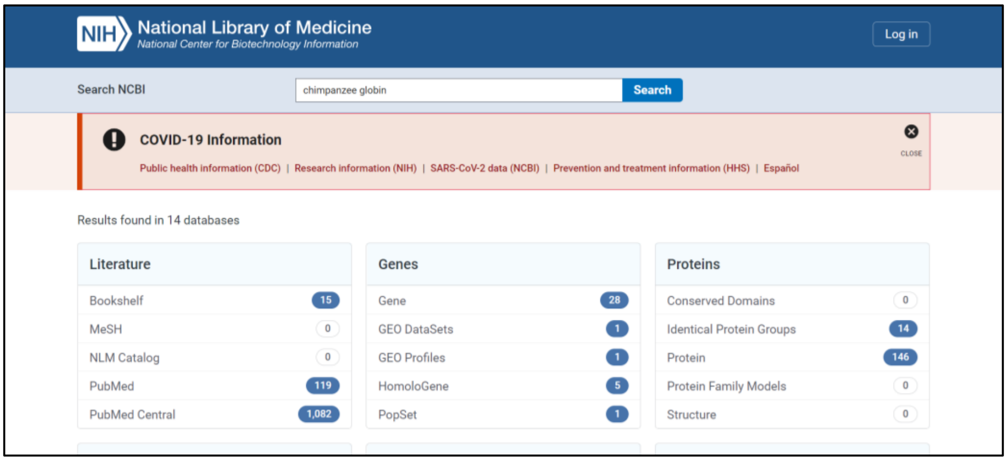
(http://www.ncbi.nlm.nih.gov/).

a) What is the RefSeq accession number of beta globin protein from the chimpanzee (Pan troglodytes)?
XP_508242.1.

(1) There are several different ways to solve this. Try typing chimpanzee globin into the home page of NCBI; or use the Taxonomy Browser to find chimpanzee Entrez Gene entries.

(2) HomoloGene (http://www.ncbi.nlm.nih.gov/homologene)(WebLink 2.38) is a great resource to learn about sets of related eukaryotic proteins. Use HomoloGene to find a set of beta globins including chimpanzee.

Entering chimpanzee globin into the search bar of the NCBI home page. Then the results in both Gene and HomoloGene will be seen.


NIH National Library of Medicine
National Center for Biotechnology Information
Log in
Search NCBI
Search
CLOSE

Search NCBI: chimpanzee globin

COVID-19 Information

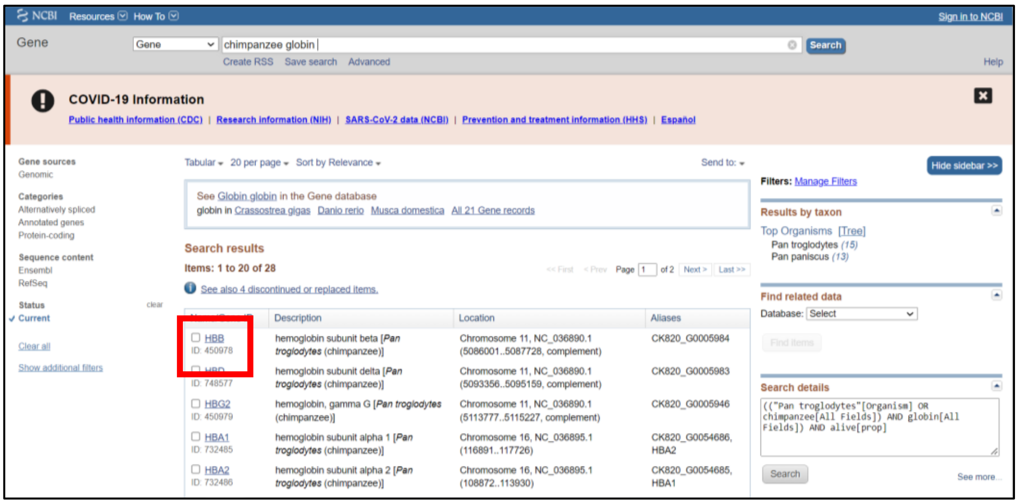
Public health information (CDC) | Research information (NIH) | SARS-CoV-2 data (NCBI) | Prevention and treatment information (HHS) | Español

Results found in 14 databases

Literature		Genes		Proteins	
Bookshelf	15	Gene	28	Conserved Domains	0
MeSH	0	GEO DataSets	1	Identical Protein Groups	14
NLM Catalog	0	GEO Profiles	1	Protein	146
PubMed	119	HomoloGene	5	Protein Family Models	0
PubMed Central	1,082	PopSet	1	Structure	0

Figure 7

Then, by clicking the Gene, the gene results sort the HBB to the top. You may see the RefSeq accession NC_036890., this corresponds to the entire chromosome 11 of the chimp.



Gene: chimpanzee globin

COVID-19 Information

Public health information (CDC) | Research information (NIH) | SARS-CoV-2 data (NCBI) | Prevention and treatment information (HHS) | Español

See Globin globin in the Gene database
globin in Crassostrea gigas Danio rerio Musca domestica All 21 Gene records

Search results

Items: 1 to 20 of 28

Page 1 of 2

See also 4 discontinued or replaced items.

	Description	Location	Aliases
□ HBB ID: 450978	hemoglobin subunit beta [*Pan troglodytes* (chimpanzee)]	Chromosome 11, NC_036890.1 (5086001..5087728, complement)	CK820_G0005984
□ HBD ID: 748577	hemoglobin subunit delta [*Pan troglodytes* (chimpanzee)]	Chromosome 11, NC_036890.1 (5093356..5095159, complement)	CK820_G0005983
□ HBG2 ID: 450979	hemoglobin, gamma G [*Pan troglodytes* (chimpanzee)]	Chromosome 11, NC_036890.1 (5113777..5115227, complement)	CK820_G0005946
□ HBA1 ID: 732485	hemoglobin subunit alpha 1 [*Pan troglodytes* (chimpanzee)]	Chromosome 16, NC_036895.1 (116891..117726)	CK820_G0054686, HBA2
□ HBA2 ID: 732486	hemoglobin subunit alpha 2 [*Pan troglodytes* (chimpanzee)]	Chromosome 16, NC_036895.1 (108872..113930)	CK820_G0054685, HBA1

Filters: Manage Filters

Results by taxon

Top Organisms [Tree]
Pan troglodytes (15)
Pan paniscus (13)

Find related data

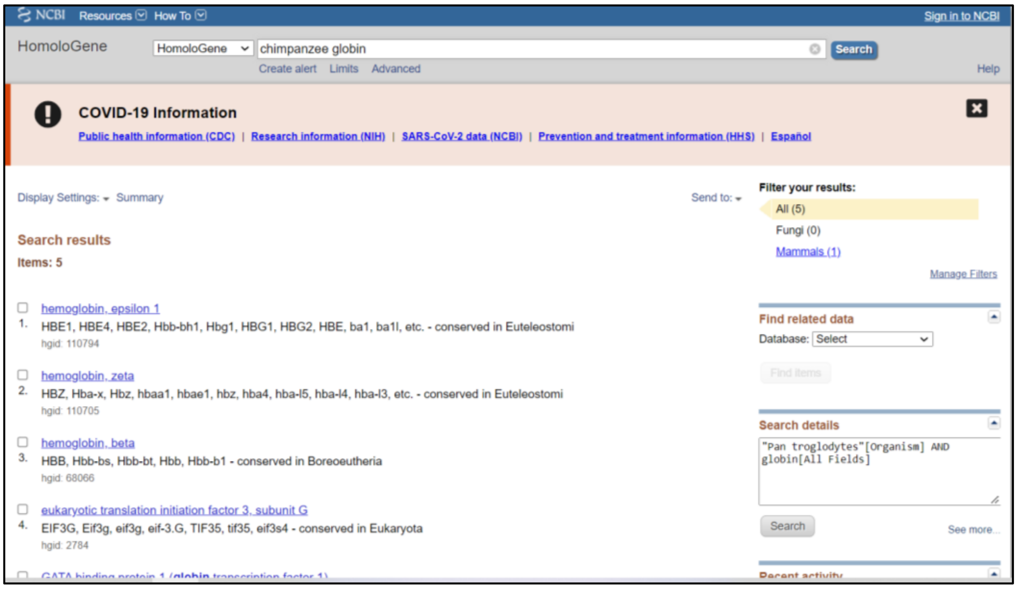
Database: Select

Search details

```
(("Pan troglodytes"[Organism] OR
chimpanzee[All Fields]) AND globin[All
Fields]) AND alive[prop]
```

Figure 8

By following the HomoloGene, you may select the hemoglobin beta, and it will lead you to the accession for the chimp HBB protein.



NCBI
Resources
How To
Sign in to NCBI
HomoloGene
Search
Create alert
Limits
Advanced
Help
COVID-19 Information
Public health information (CDC) | Research information (NIH) | SARS-CoV-2 data (NCBI) | Prevention and treatment information (HHS) | Español
Display Settings: Summary
Send to:
Manage Filters
Find items
See more...
Recent activity

HomoloGene: chimpanzee globin

Search results

Items: 5

□ hemoglobin, epsilon 1
1. HBE1, HBE4, HBE2, Hbb-bh1, Hbg1, HBG1, HBG2, HBE, ba1, ba1l, etc. - conserved in Euteleostomi
hgid: 110794

□ hemoglobin, zeta
2. HBZ, Hba-x, Hbz, hbaa1, hbae1, hbz, hba4, hba-l5, hba-l4, hba-l3, etc. - conserved in Euteleostomi
hgid: 110705

□ hemoglobin, beta
3. HBB, Hbb-bs, Hbb-bt, Hbb, Hbb-b1 - conserved in Boreoeutheria
hgid: 68066

□ eukaryotic translation initiation factor 3, subunit G
4. EIF3G, Eif3g, eif3g, eif-3.G, TIF35, tif35, eif3s4 - conserved in Eukaryota
hgid: 2784

□ GATA binding protein 1 (globin transcription factor 1)

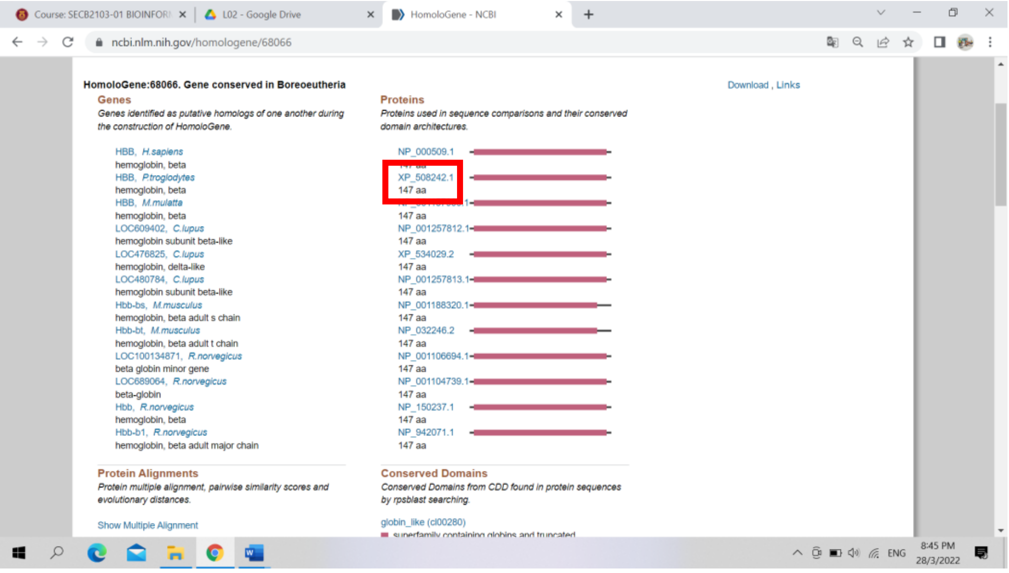
Filter your results:
All (5)
Fungi (0)
Mammals (1)

Find related data
Database: Select

Search details
```
"Pan troglodytes"[Organism] AND
globin[All Fields]
```

Figure 9

Hence, you may see the result of the RefSeq accession number of beta globin protein from the chimpanzee (Pan troglodytes).



HomoloGene:68066. Gene conserved in Boreoeutheria

Genes
Genes identified as putative homologs of one another during the construction of HomoloGene.

Proteins
Proteins used in sequence comparisons and their conserved domain architectures.

Genes	Proteins
HBB, *H.sapiens* hemoglobin, beta	NP_000509.1 147 aa
HBB, *P.troglodytes* hemoglobin, beta	XP_508242.1 147 aa
HBB, *M.mulatta* hemoglobin, beta	[illegible] 147 aa
LOC609402, *C.lupus* hemoglobin subunit beta-like	NP_001257812.1 147 aa
LOC476825, *C.lupus* hemoglobin, delta-like	XP_534029.2 147 aa
LOC480784, *C.lupus* hemoglobin subunit beta-like	NP_001257813.1 147 aa
Hbb-bs, *M.musculus* hemoglobin, beta adult s chain	NP_001188320.1 147 aa
Hbb-bt, *M.musculus* hemoglobin, beta adult t chain	NP_032246.2 147 aa
LOC100134871, *R.norvegicus* beta globin minor gene	NP_001106694.1 147 aa
LOC689064, *R.norvegicus* beta-globin	NP_001104739.1 147 aa
Hbb, *R.norvegicus* hemoglobin, beta	NP_150237.1 147 aa
Hbb-b1, *R.norvegicus* hemoglobin, beta adult major chain	NP_942071.1 147 aa

Protein Alignments
Protein multiple alignment, pairwise similarity scores and evolutionary distances.

Conserved Domains
Conserved Domains from CDD found in protein sequences by rpsblast searching.

globin_like (cl00280)
superfamily containing globins and truncated

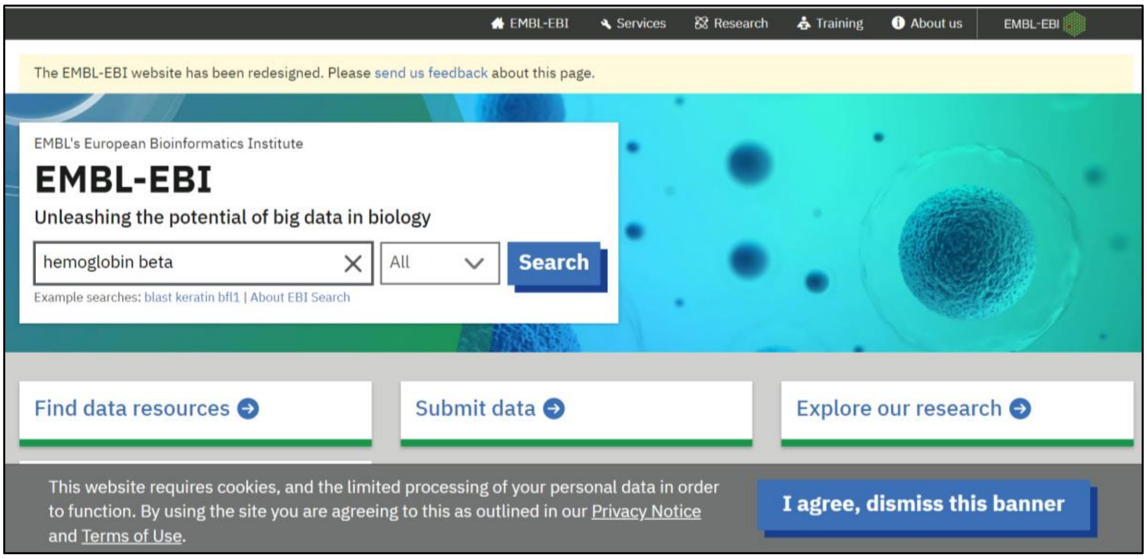
Figure 10

3. The purpose of this exercise is to become familiar with the EBI website and how to use it to access information.

(1) Visit the site (http://www.ebi.ac.uk/)(WebLink 2.5). Enter hemoglobin beta in the main query box (alternatively use the query human hemoglobin beta).

(2) Inspect the results. Explore the various links to information about pathways, genomes, nucleotide and protein sequences, structures, protein families, and more.

Here is the EBI home page. Typing the hemoglobin beta at the search bar.



The EMBL-EBI website has been redesigned. Please send us feedback about this page.

EMBL's European Bioinformatics Institute

EMBL-EBI

Unleashing the potential of big data in biology

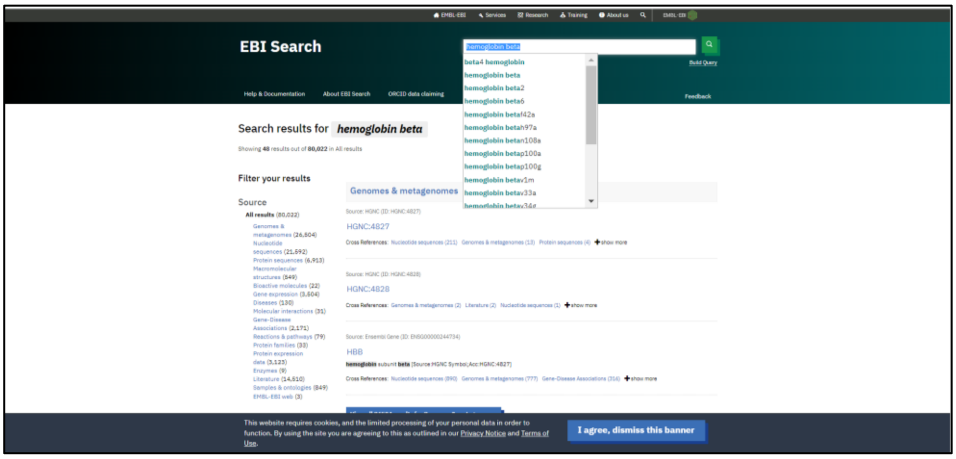
hemoglobin beta

Example searches: blast keratin bfl1 | About EBI Search

This website requires cookies, and the limited processing of your personal data in order to function. By using the site you are agreeing to this as outlined in our Privacy Notice and Terms of Use.

Figure 11

Here is the search results of the hemoglobin beta.



hemoglobin beta

beta4 hemoglobin
hemoglobin beta
hemoglobin beta2
hemoglobin beta6
hemoglobin betaf42a
hemoglobin betah97a
hemoglobin betan108a
hemoglobin betap100a
hemoglobin betap100g
hemoglobin betav1m
hemoglobin betav33a
hemoglobin betav34g

Search results for *hemoglobin beta*

Showing **48** results out of **80,022** in All results

Filter your results

Source

All results (80,022)

Genomes & metagenomes (26,804)
Nucleotide sequences (21,692)
Protein sequences (6,913)
Macromolecular structures (849)
Bioactive molecules (22)
Gene expression (3,504)
Diseases (130)
Molecular interactions (31)
Gene-Disease Associations (2,171)
Reactions & pathways (79)
Protein families (33)
Protein expression data (3,123)
Enzymes (9)
Literature (14,510)
Samples & ontologies (849)
EMBL-EBI web (3)

Genomes & metagenomes

Source: HGNC (ID: HGNC:4827)

HGNC:4827

Cross References: Nucleotide sequences (211) Genomes & metagenomes (13) Protein sequences (4) ✚show more

Source: HGNC (ID: HGNC:4828)

HGNC:4828

Cross References: Genomes & metagenomes (2) Literature (2) Nucleotide sequences (1) ✚show more

Source: Ensembl Gene (ID: ENSG00000244734)

HBB

hemoglobin subunit **beta** [Source:HGNC Symbol;Acc:HGNC:4827]

Cross References: Nucleotide sequences (890) Genomes & metagenomes (777) Gene-Disease Associations (316) ✚show more

This website requires cookies, and the limited processing of your personal data in order to function. By using the site you are agreeing to this as outlined in our Privacy Notice and Terms of Use.

Figure 12

4. Accessing information from BioMart: the beta globin locus

(1) Go to www.ensembl.org and follow the link to BioMart.

(2) First choose a database; we will select Ensembl Genes 71.

(3) Choose a dataset: Homo sapiens genes (GRCh37.p10). Note the other available datasets.

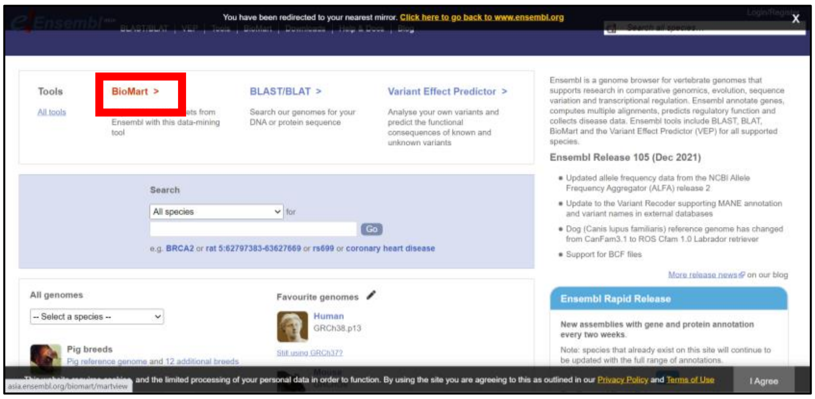
(4) Choose a filter. Here the options include region, gene, transcript event, expression, multispecies comparisons, protein domains, and variation. Select "region", chromosome 11, and enter 5240000 for the Gene Start (bp) and 5300000 (bp) for the Gene End. (Note that this region spans 60 kilobases and corresponds to chr11:5,240,001-5,300,000.)

(5) Choose attributes. Select the following features. Under "Gene" select Ensembl Gene ID and %GC content; under "External" select the external references CCDS ID, HGNC symbol (this is the official gene symbol) and HGNC ID(s).

(6) At the top left select "Count." Currently there are 8 genes matching these criteria.

(7) To view these results press "Results. "Note that you can export your results in several formats (including a comma separated values or CSV file) that can be further manipulated (e.g., converted to a BED file).

Here is the Ensembl home page. At the home page, select the BioMart.



You have been redirected to your nearest mirror. Click here to go back to www.ensembl.org

Ensembl with this data-mining tool

Search our genomes for your DNA or protein sequence

Analyse your own variants and predict the functional consequences of known and unknown variants

Ensembl is a genome browser for vertebrate genomes that supports research in comparative genomics, evolution, sequence variation and transcriptional regulation. Ensembl annotate genes, computes multiple alignments, predicts regulatory function and collects disease data. Ensembl tools include BLAST, BLAT, BioMart and the Variant Effect Predictor (VEP) for all supported species.

Ensembl Release 105 (Dec 2021)

- Updated allele frequency data from the NCBI Allele Frequency Aggregator (ALFA) release 2
- Update to the Variant Recoder supporting MANE annotation and variant names in external databases
- Dog (Canis lupus familiaris) reference genome has changed from CanFam3.1 to ROS Cfam 1.0 Labrador retriever
- Support for BCF files

More release news on our blog

Ensembl Rapid Release

New assemblies with gene and protein annotation every two weeks.

Note: species that already exist on this site will continue to be updated with the full range of annotations.

and the limited processing of your personal data in order to function. By using the site you are agreeing to this as outlined in our Privacy Policy and Terms of Use

Tools
All tools
BioMart >
BLAST/BLAT >
Variant Effect Predictor >
Search
All species
for
Go
e.g. BRCA2 or rat 5:62797383-63627669 or rs699 or coronary heart disease
All genomes
-- Select a species --
Pig breeds
Pig reference genome and 12 additional breeds
Favourite genomes
Human
GRCh38.p13
Still using GRCh37?
I Agree
asia.ensembl.org/biomart/martview

Figure 13

Then, choose a Ensembl Genes 105 database.

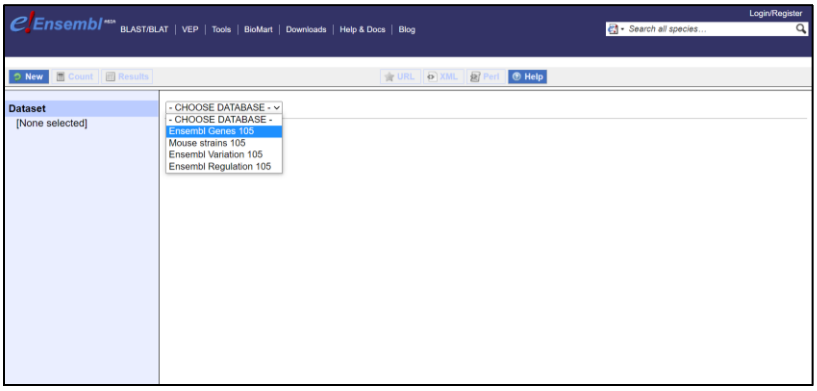
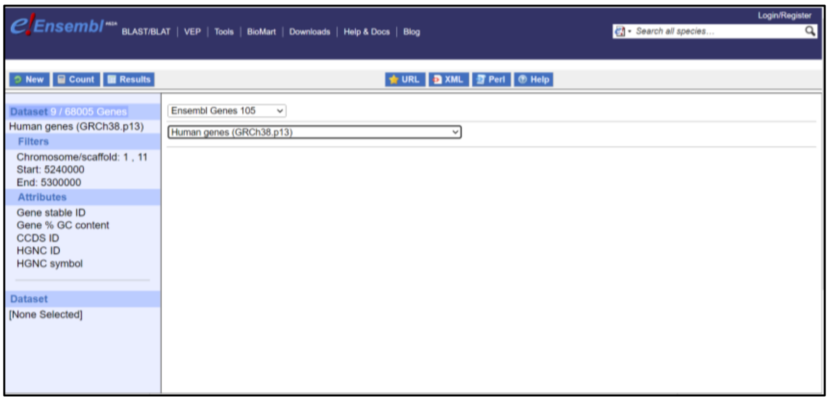


Figure 14

Then, select a Human genes (GRCh38.p13) dataset.



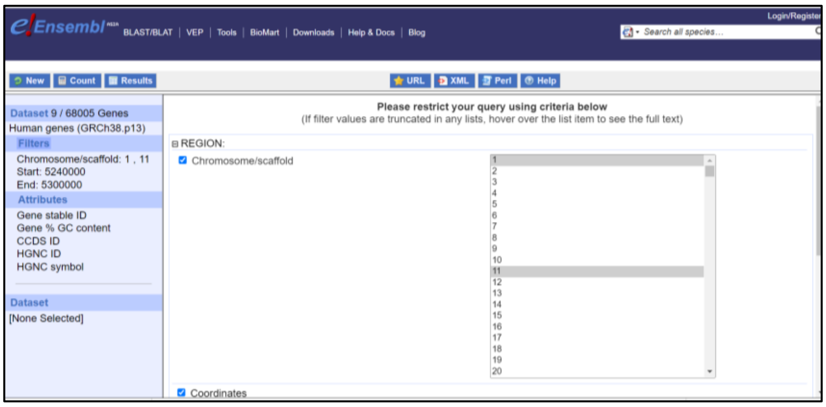
Dataset 9 / 68005 Genes
Human genes (GRCh38.p13)
Filters
Chromosome/scaffold: 1 , 11
Start: 5240000
End: 5300000
Attributes
Gene stable ID
Gene % GC content
CCDS ID
HGNC ID
HGNC symbol

Dataset
[None Selected]

Ensembl Genes 105
Human genes (GRCh38.p13)

Figure 15

After that, select the filters. Choose a region and fill the gene start and gene end.



Dataset 9 / 68005 Genes
Human genes (GRCh38.p13)
Filters
Chromosome/scaffold: 1 , 11
Start: 5240000
End: 5300000
Attributes
Gene stable ID
Gene % GC content
CCDS ID
HGNC ID
HGNC symbol

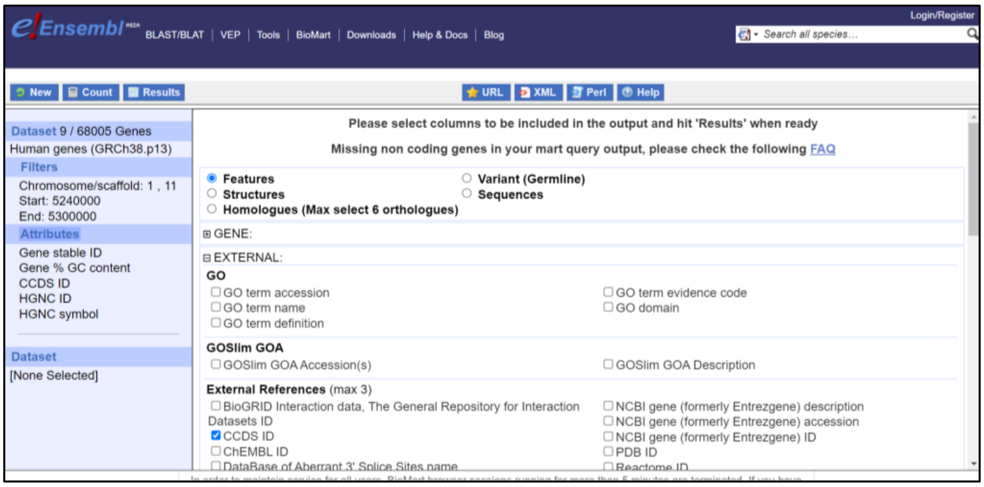
Dataset
[None Selected]

Please restrict your query using criteria below
(If filter values are truncated in any lists, hover over the list item to see the full text)

⊟ REGION:
☑ Chromosome/scaffold
☑ Coordinates

Figure 16

Then, select the attributes. Choose gene and external references.



Dataset 9 / 68005 Genes
Human genes (GRCh38.p13)
Filters
Chromosome/scaffold: 1 , 11
Start: 5240000
End: 5300000
Attributes
Gene stable ID
Gene % GC content
CCDS ID
HGNC ID
HGNC symbol

Dataset
[None Selected]

Please select columns to be included in the output and hit 'Results' when ready

Missing non coding genes in your mart query output, please check the following FAQ

● Features ○ Variant (Germline)
○ Structures ○ Sequences
○ Homologues (Max select 6 orthologues)

⊞ GENE:

⊟ EXTERNAL:

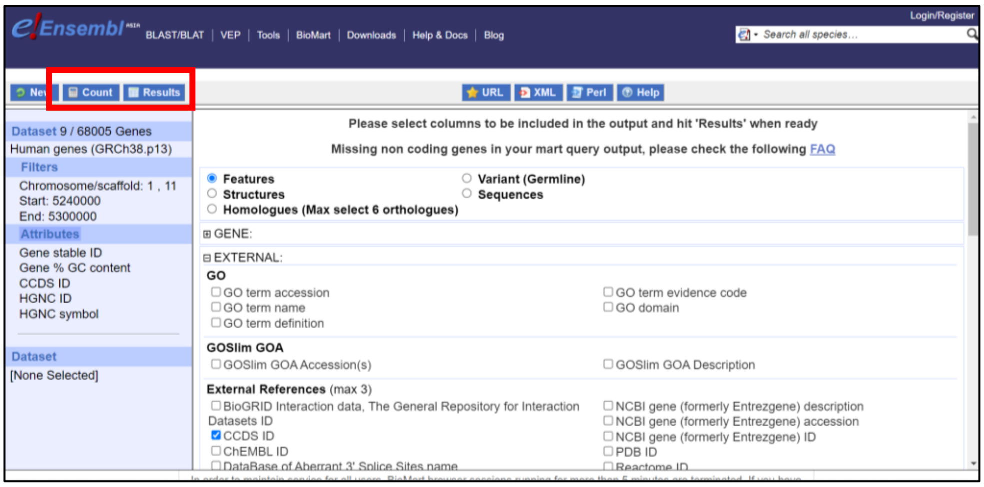
GO
□ GO term accession □ GO term evidence code
□ GO term name □ GO domain
□ GO term definition

GOSlim GOA
□ GOSlim GOA Accession(s) □ GOSlim GOA Description

External References (max 3)
□ BioGRID Interaction data, The General Repository for Interaction Datasets ID
☑ CCDS ID
□ ChEMBL ID
□ DataBase of Aberrant 3' Splice Sites name
□ NCBI gene (formerly Entrezgene) description
□ NCBI gene (formerly Entrezgene) accession
□ NCBI gene (formerly Entrezgene) ID
□ PDB ID
□ Reactome ID

Figure 17

Click “count” and “results”.



Dataset 9 / 68005 Genes
Human genes (GRCh38.p13)
Filters
Chromosome/scaffold: 1 , 11
Start: 5240000
End: 5300000
Attributes
Gene stable ID
Gene % GC content
CCDS ID
HGNC ID
HGNC symbol
Dataset
[None Selected]

Please select columns to be included in the output and hit 'Results' when ready

Missing non coding genes in your mart query output, please check the following FAQ

● Features ○ Variant (Germline)
○ Structures ○ Sequences
○ Homologues (Max select 6 orthologues)

⊞ GENE:

⊟ EXTERNAL:

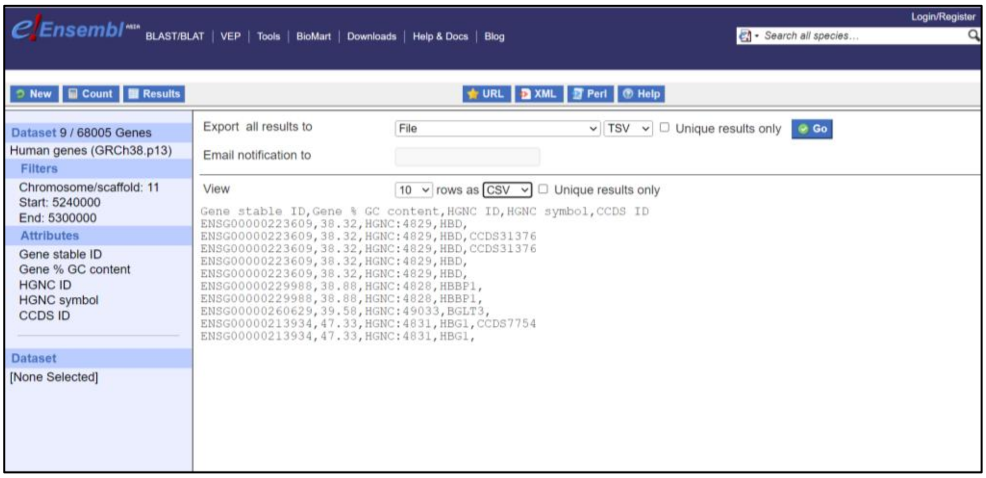
GO
□ GO term accession □ GO term evidence code
□ GO term name □ GO domain
□ GO term definition

GOSlim GOA
□ GOSlim GOA Accession(s) □ GOSlim GOA Description

External References (max 3)
□ BioGRID Interaction data, The General Repository for Interaction Datasets ID □ NCBI gene (formerly Entrezgene) description
☑ CCDS ID □ NCBI gene (formerly Entrezgene) accession
□ ChEMBL ID □ NCBI gene (formerly Entrezgene) ID
□ DataBase of Aberrant 3' Splice Sites name □ PDB ID
□ Reactome ID

Figure 18

The result will be shown.



Dataset 9 / 68005 Genes
Human genes (GRCh38.p13)
Filters
Chromosome/scaffold: 11
Start: 5240000
End: 5300000
Attributes
Gene stable ID
Gene % GC content
HGNC ID
HGNC symbol
CCDS ID
Dataset
[None Selected]

Export all results to File TSV □ Unique results only

Email notification to

View 10 rows as CSV □ Unique results only

```
Gene stable ID,Gene % GC content,HGNC ID,HGNC symbol,CCDS ID
ENSG00000223609,38.32,HGNC:4829,HBD,
ENSG00000223609,38.32,HGNC:4829,HBD,CCDS31376
ENSG00000223609,38.32,HGNC:4829,HBD,CCDS31376
ENSG00000223609,38.32,HGNC:4829,HBD,
ENSG00000223609,38.32,HGNC:4829,HBD,
ENSG00000229988,38.88,HGNC:4828,HBBP1,
ENSG00000229988,38.88,HGNC:4828,HBBP1,
ENSG00000260629,39.58,HGNC:49033,BGLT3,
ENSG00000213934,47.33,HGNC:4831,HBG1,CCDS7754
ENSG00000213934,47.33,HGNC:4831,HBG1,
```

Figure 19

5. BioMart: working with lists. The goal of this exercise is to access information in BioMart by uploading a text file listing gene identifier of interest.

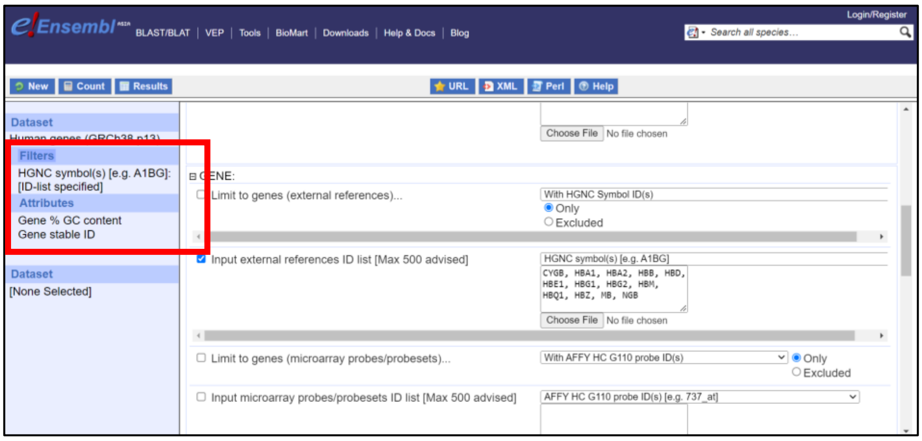
Follow the steps from problem 2-4, but for the filter set choose Gene (instead of Region), select ID list limit and adjust the pulldown menu to HGNC symbol, then browse for a text file having a list of gene symbols.

See Web Document 2.5 (WebDocument_2-5_13_humanGlobins_HGNCsymbols.txt) for a text file listing official HGNC symbols for 13 human globin genes (*CYGB*, *HBA1*, *HBA2*, *HBB*, *HBD*, *HBE1*, *HBG1*, *HBG2*, *HBM*, *HBQ1*, *HBZ*, *MB*, *NGB*).

You could also enter these gene symbols manually.

For attributes choose any set of features that is different than in problem 2-4, so that you can further explore BioMart resources.

Change the filters and attributes.



Dataset
Human genes (GRCh38.p13)
Filters
HGNC symbol(s) [e.g. A1BG]: [ID-list specified]
Attributes
Gene % GC content
Gene stable ID
Dataset
[None Selected]

Choose File No file chosen

⊟ GENE:

□ Limit to genes (external references)... With HGNC Symbol ID(s) ● Only ○ Excluded

☑ Input external references ID list [Max 500 advised] HGNC symbol(s) [e.g. A1BG]

```
CYGB, HBA1, HBA2, HBB, HBD,
HBE1, HBG1, HBG2, HBM,
HBQ1, HBZ, MB, NGB
```

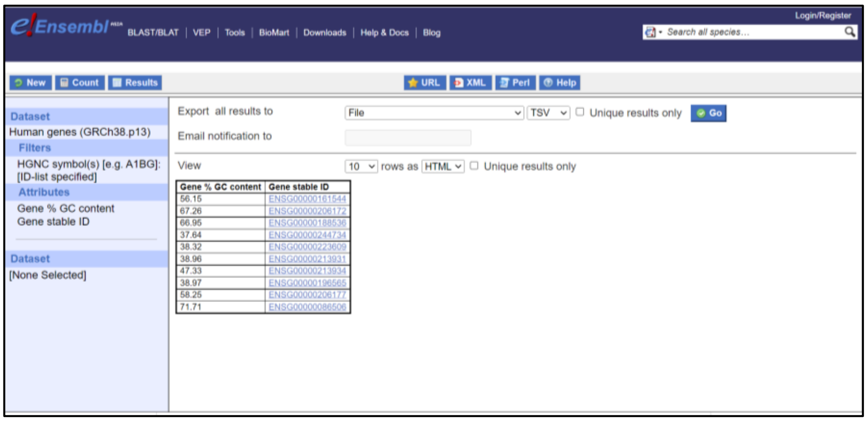
Choose File No file chosen

□ Limit to genes (microarray probes/probesets)... With AFFY HC G110 probe ID(s) ● Only ○ Excluded

□ Input microarray probes/probesets ID list [Max 500 advised] AFFY HC G110 probe ID(s) [e.g. 737_at]

Figure 20

The results will be shown after click on the results.



Dataset
Human genes (GRCh38.p13)
Filters
HGNC symbol(s) [e.g. A1BG]: [ID-list specified]
Attributes
Gene % GC content
Gene stable ID
Dataset
[None Selected]

Export all results to File TSV □ Unique results only

Email notification to

View 10 rows as HTML □ Unique results only

Gene % GC content	Gene stable ID
56.15	ENSG00000161544
67.26	ENSG00000206172
66.95	ENSG00000188536
37.64	ENSG00000244734
38.32	ENSG00000223609
38.96	ENSG00000213931
47.33	ENSG00000213934
38.97	ENSG00000196565
58.25	ENSG00000206177
71.71	ENSG00000086506

Figure 21