Hands on identifier mapping

# Webtools

ID mapping webtools are easy to use without any programming experience.

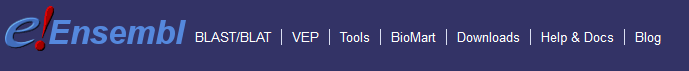
## BioMart

The most commonly used ID mapping tool is BioMart provided by Ensembl database. BioMart is mainly designed as data mining or data extraction tool but provides also functionalities that allows to use it as identifier mapping tool.

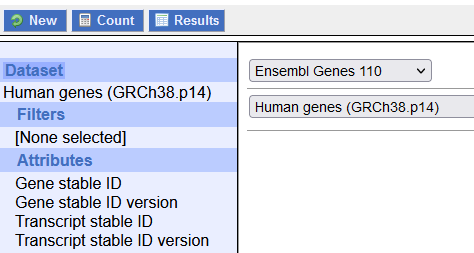
1. **Task: Add a column with matching HGNC symbols to the SigUP\_F.csv file**

**Extract a mapping file from BioMart:**

* Go to <https://www.ensembl.org> and select BioMart



* Choose Database “Ensembl Genes 110”
* Choose Dataset “Human Genes” (current version is GRCh38.p14)

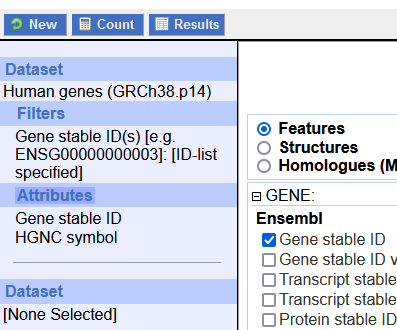


“Filter” is where you define your input or your query and “Attribute” is where you define what you want as output. These are the standard settings.

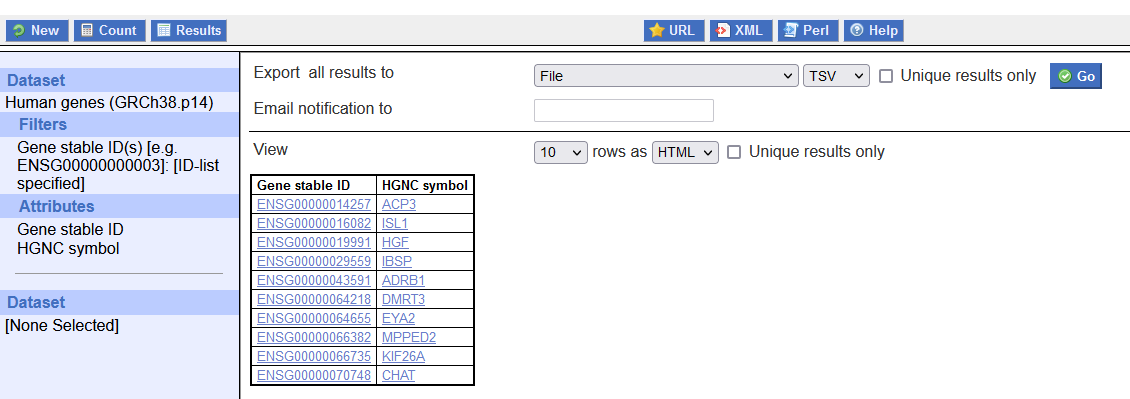
* Open “SigUP\_F.csv” in Excel
* Copy the list of Ensembl identifiers (without header)
* In Biomart, click on “Filter” – “Gene” – “Input external references ID list”
* Paste the list of Ensembl IDs in the window
* Make sure “Gene stable ID(s)” is selected
* Define your output by clicking on “Attributes” – “Gene” and check only “Gene stable ID” (version and transcript IDs are automatically checked but for this task we don’t need them).

BioMart will deliver the results alphabetically sorted, which may be a different order than the input. Therefore, it’s important to have the results with both, input and output identifier system.

* Click on “External” and check “HGNC Symbol”. The list on the left side should look like this:



* Click on “Results” and download your full list of results by clicking on “Go”

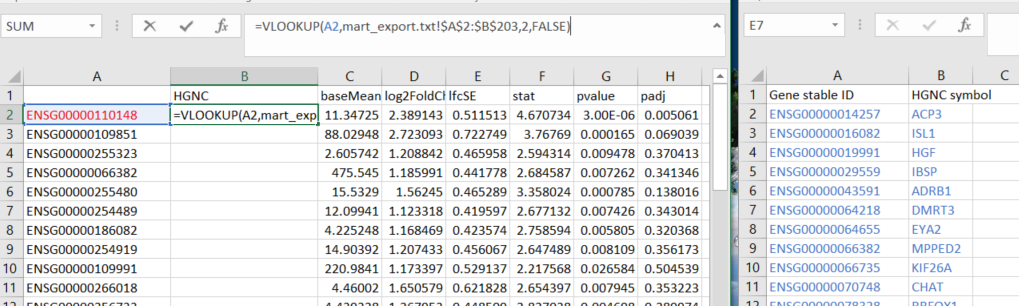


**Use the mapping file to add a matching column with HGNC identifiers in your dataset**

* Open the file “mart\_export” in Excel. For some identifiers, you will see that for some IDs there is HGNC symbol available. This is partly because they are new, or not longer used by Ensembl - but still found in data analysis, because they used older versions when the data was created.

|  |  |
| --- | --- |
|  |  |
|  |  |

* Note that the order of the identifiers is different in both files, so you cannot just copy them! Use the VLOOKUP function in Excel to add the appropriate HGNC symbol to the Ensembl identifier. If you use Excel in another language than English, please look up the name of the function!



2

1

The formula is: =VLOOKUP(A2,mart\_export.txt!$A$2:$B$203,2,FALSE)

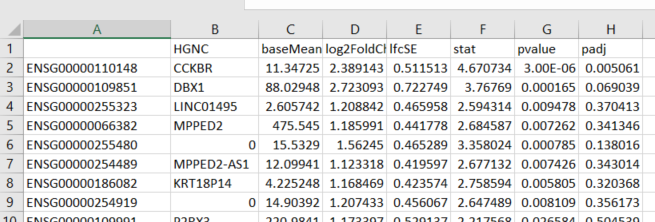
A2 = the identifier you want to map

mart\_export.txt!$A$2:$B$203 = the field in the file mart\_export.txt, from A2 to B203

2 = column 2 if the one with the matching identifier can be found in column 1

FALSE = only map when there is an exact match

* If you double-click on the green dot (indicated by orange arrow) the formula will be applied to the whole row.

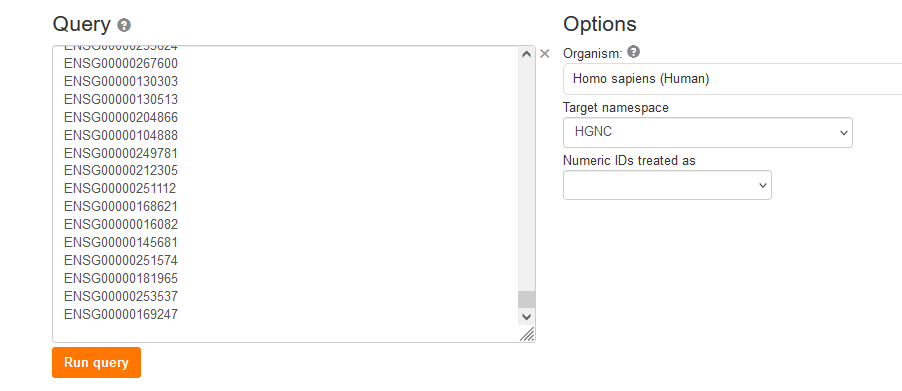


* Save as e.g. SigUP\_F\_HGNC.csv – use a different name in order not to overwrite the previous file.

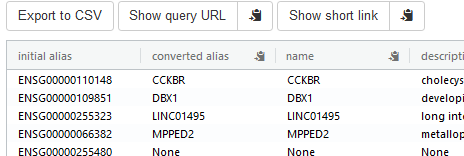
## g:Convert

g:Convert is the ID conversion tool provided by g:Profiler. We have used this tool before to get HGNC symbols for EnrichR, so this is just a quick repetition of the steps:

* 1. Open g:Convert website
  2. Paste the Ensembl IDs in the query field
  3. Select HGNC in the Target namespace



* 1. Hit “Query”
  2. You can now “Export to CSV” and proceed with adding them using Excel and VLOOKUP function like described above for BioMart.



## CTD

BioMart and g:Convert provide only data extraction and ID conversion services for genes, gene products and their associated information. CTD, the “Chemical Translation Service” is a webtool for chemical compounds and therefore suitable for any kind of metabolomics data.

* Go to the website <https://cts.fiehnlab.ucdavis.edu/> “Simple Conversion” converts one ID/name per query, “Batch Conversion” allows to either copy a list of IDs or to upload a file with IDs for batch conversion.
* Use the “Simple Conversion” to complete this table:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ID | Database | ID | Database | Chemical name |
| CHEMBL14249 | ChEMBL |  | Human Metabolome Database (HMDB) |  |
| LMST01010001 | LipidMaps |  | ChEMBL |  |
| Ethanol | Chemical Name |  | ChEBI | Ethanol |

# R packages

## BiomaRt

* See biomaRt.R script

## BridgeDb

* See bridgeDb R script

# Further reading:

- RDMkit <https://rdmkit.elixir-europe.org/identifiers>

- The FAIR cookbook <https://faircookbook.elixir-europe.org/content/recipes/interoperability/identifier-mapping.html>

- The bridgeDb paper by van Iersel et al. 2010 <https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-11-5>