Spring April, 2016

Making Sense of Genomic Variation: Part 1 SNP Annotation

Rolando Garcia-Milian

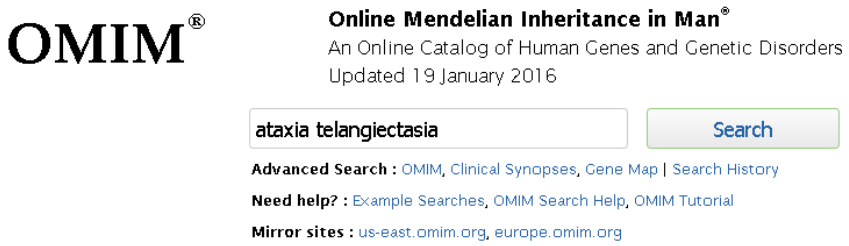
Making Sense of Genomic Variation: SNP Annotation -Part 1

Rolando Garcia-Milian
rolando.milian@yale.eu
Biomedical Sciences Research Support
Yale Cushing/Whitney Medical Library
Yale University.

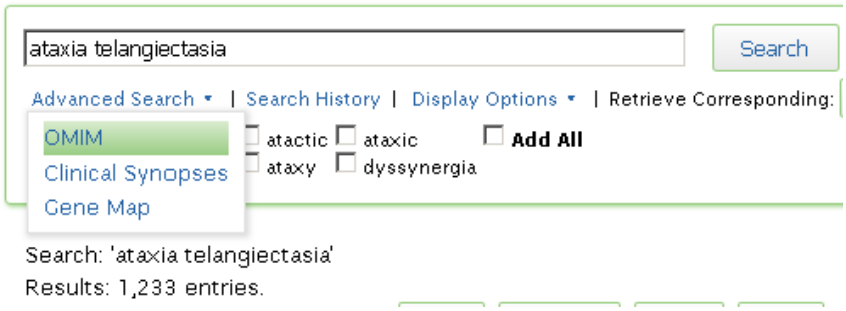
Contents

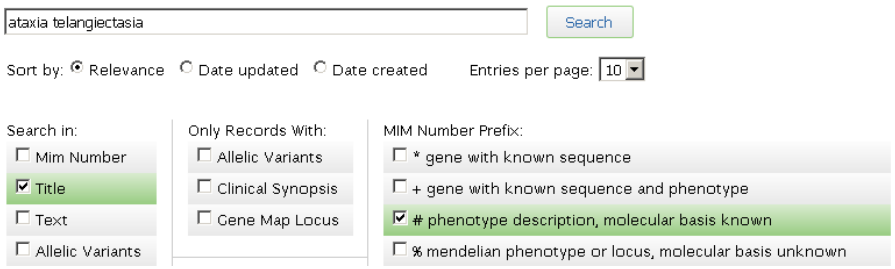
Online Mendelian Inheritance in Man

For this example, we will found out the mode of inheritance of the genetic disorder ataxia telangiectasia and the gene associated with this disorder, genetic tests, and labs in US offering these tests. Go to the OMIM website http://www.omim.org/ Enter "ataxia telangiectasia" in the search box and click on the "Search" button.



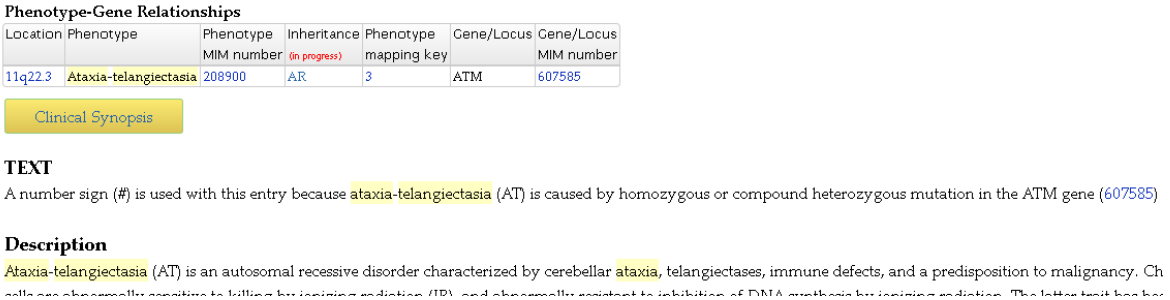
Look at the number of the results (1,233 entries). Use the Advance search to limit the number and specificity of your results. Click on the "Advance search" link located under the search box and click on the "OMIM" button. A new page will open showing all the limits available. Check the "Title" and the "#phenotype description, molecular basis known" limits and click on the "Search" button.



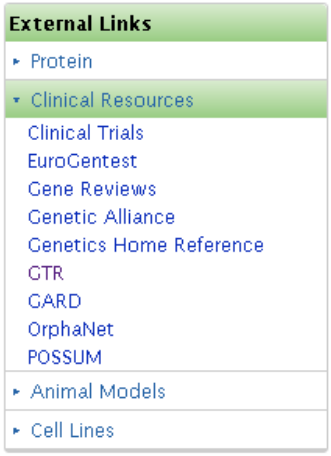


The number of results will go down to 100. Click on the first record. " # 208900. ATAXIA-TELANGIECTASIA; AT"

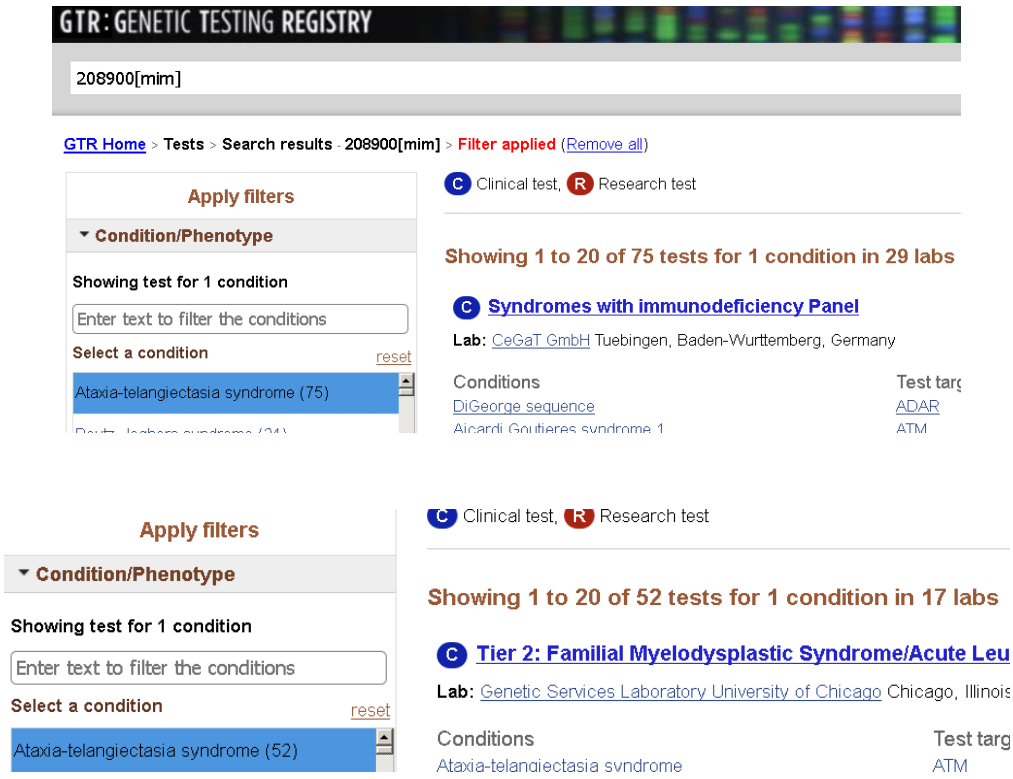
Note the table "Phenotype-Gene Relationship" showing the "Gene/Locus" associated with this disorder (ATM gene).
Under "TEXT", it also says that "ataxia-telangiectasia (AT) is caused by homozygous or compound heterozygous mutation in the ATM gene (607585) on chromosome 11q22"
Under "Description": is an autosomal recessive disorder



Expand "Clinical Resources" tab- under the "External Links table located to the right of the page and click on the GTR (Genetic Testing Registry) link.

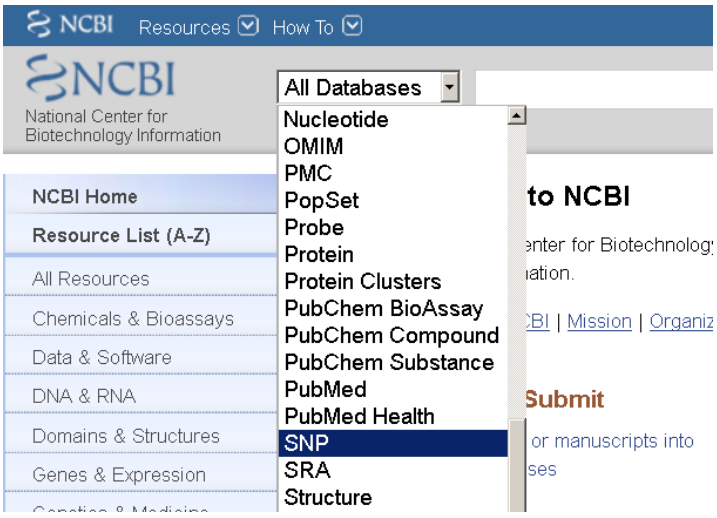


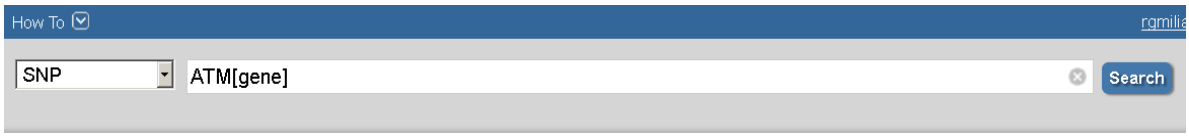
A new page will open showing the Genetic Testing Registry results for Ataxia-telangiectasia syndrome- 75 tests available in 29 labs. Scroll down the page to see the filtering options. Under "Lab location", check United States: your results will narrow down to "52 tests for 1 condition in 17 labs".



dbSNP

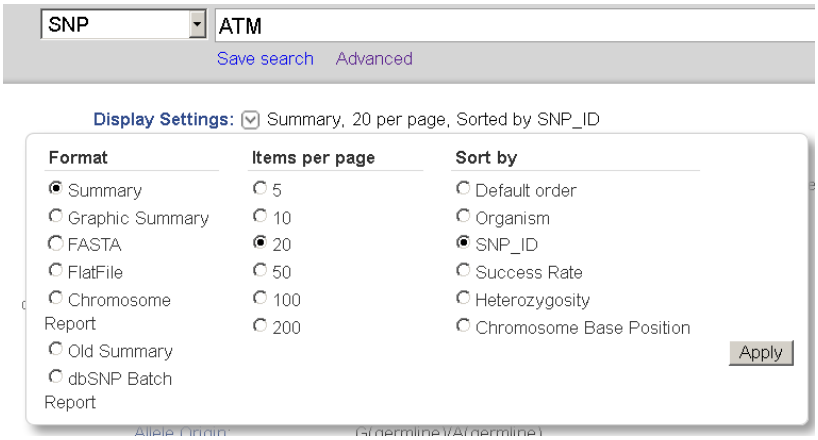
The ATM gene has been associated with Ataxia-telangiectasia, B-cell non-Hodgkin lymphoma, T-cell prolymphocytic leukemia, and susceptibility to Breast cancer. Find all missense variants reported for this gene. How many of theses are pathogenic?
Go to the NCBI database main page: http://www.ncbi.nlm.nih.gov/ From the drop-down menu select the SNP database. On the search box type ATM[gene], and click on the search button.





Click on “Human” located on the right side of the page under the filter “Organism”. Click on “Pathogenic” under the Clinical significance” filter, and click on “missense” under the “Function class” filter.

The format, number of results per page, and organization of results can be changed by clicking on “Display setting” menu. Results can be downloaded

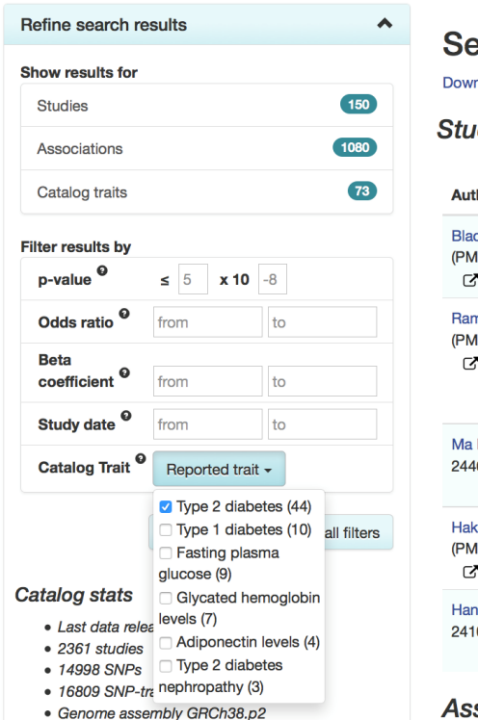


The same results can be obtained by searching the NCBI Gene database: http://www.ncbi.nlm.nih.gov/gene or from the Variation Viewer tool http://www.ncbi.nlm.nih.gov/variation/view/

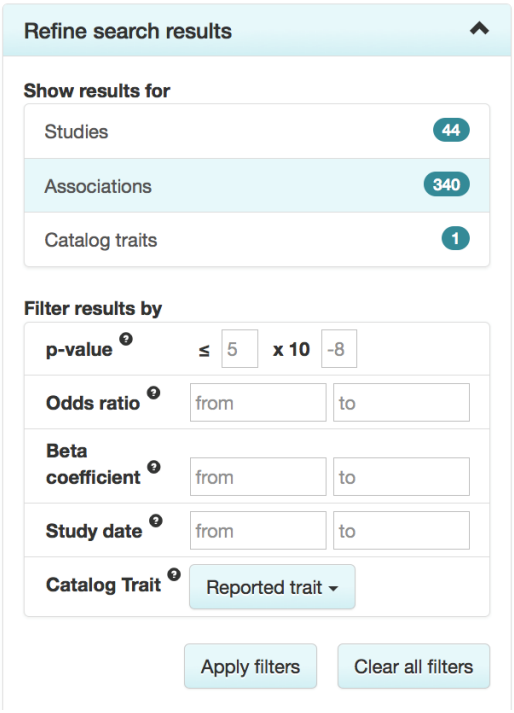
SNPs – Trait Relationship: GWAS Catalog

In this exercise we will find SNPs associated with "type 2 diabetes" by using the National Human Genome Research Institute (NHGRI) Catalog of Published Genome-Wide Association Studies (GWAS) (Welter et al., 2014)
Go to the main page of the GWAS Catalog: https://www.ebi.ac.uk/gwas/home
Type "diabetes" in the search box and click on the search icon.

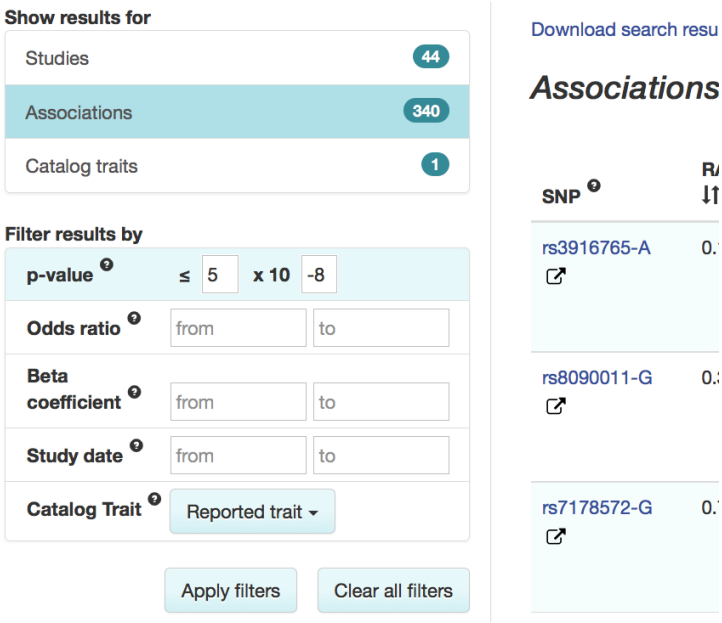
Click on the "Reported trait" drop-down menu button next to Catalog Trait located in the "Refine search results" –left of the page. Select "Type 2 diabetes"[1]. Note the number of studies in the catalog (44). Click on the "Apply filter" button located right below the "Reported trait" button.
The page will update to show the Studies listed in the Catalog. From the left bar, click on "Associations" (340) located under the "Show results for" section.



[1] The GWAS Catalog use Experimental Factor Ontology to map phenotypes/reported traits. In this case, "type 2 diabetes" is mapped to the term "type II diabetes mellitus". This page http://www.ebi.ac.uk/efo/EFO_0001360 provides more information on this term as well as synonyms, parental term, etc.



The page will update to show the 340 SNP associated with "type 2 diabetes". Please notice that the resulting table of associations can be downloaded from the "Download search results" link at the top of the table. However for this exercise, we will filter and download only those associations with a p value < 5 x 10e-8 (default). Click on the "Apply filters" button again.

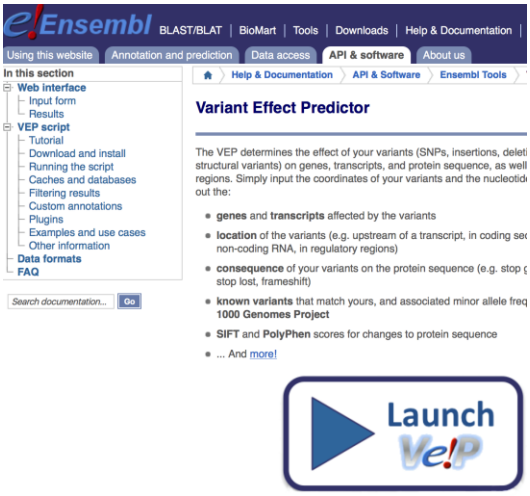


The table will update to show only those SNPs with association value p< 5 x 10e-8. The total number of associations will go down from 340 to 134. There are other relevant filters that you would like to apply depending on your specific case (e.g. Odds ratio, Study date, and Beta coefficient)

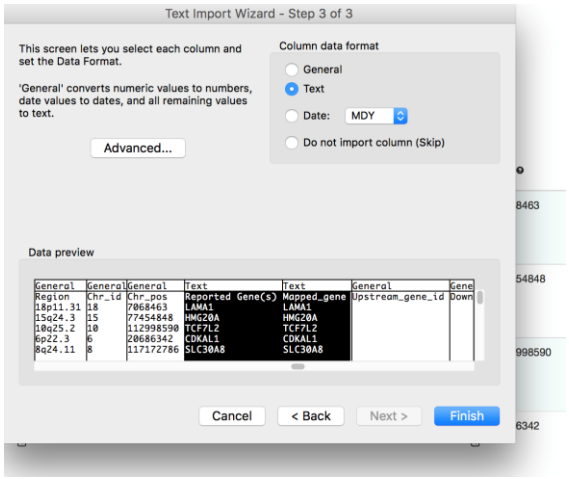
Click on the “Download search result” to save the result. We will use these in the next section to annotate a list of SNPs.

Using Variant Effect Predictor (Ve!P) to Annotate a List of SNPs

In this exercise we will annotate a list of SNPs associated with type 2 diabetes- obtained from the previous section by using the Variant Effect Predictor (McLaren et al., 2010)
Open the Variant Effect Predictor (part of the Ensembl genome browser tools) main page http://www.ensembl.org/info/docs/tools/vep/index.html and click on the “Launch Ve!P” button. A new page will open

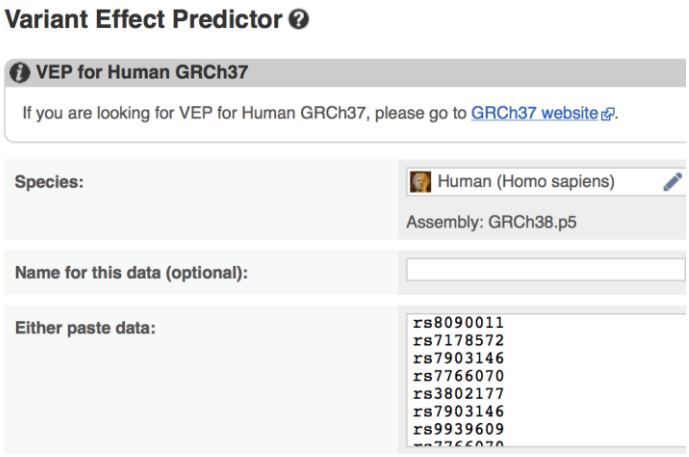


Use Microsoft Excel to open the table downloaded in the previous section containing the SNPs associated with type 2 diabetes. Define the column data format as “Text” for those columns containing the gene symbols. Excel automatically converts some gene symbols (MARCH3, SEPT1, DEC1) into dates (Zeeberg et al., 2004)
From the Excel sheet- SNPs column-, copy the 184 SNPs “rs” identifiers, as shown in the figure below.

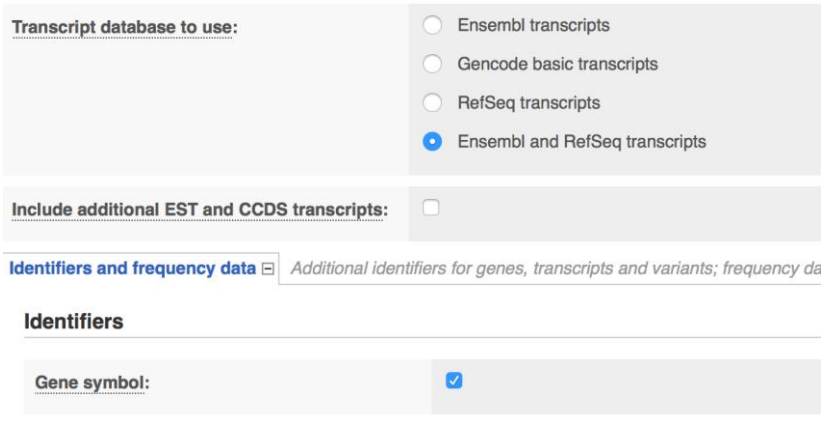


	U	V	W	X	Y
eam	Strongest SNP-Risk Allel	SNPs	Merged	Snp_id_curre	Context
	rs8090011-G	rs8090011		8090011	intron
	rs7178572-G	rs7178572		7178572	intron
	rs7903146-T	rs7903146		7903146	intron
	rs7766070-A	rs7766070		7766070	intron
	rs3802177-G	rs3802177		3802177	UTR-3
	rs7903146-T	rs7903146		7903146	intron
	rs9939609-A	rs9939609		9939609	intron
	rs7766070-A	rs7766070		7766070	intron
029	rs5015480-C	rs5015480		5015480	
	rs4402960-T	rs4402960		4402960	intron
	rs864745-T	rs864745		864745	intron
502	rs12779790-G	rs12779790		12779790	

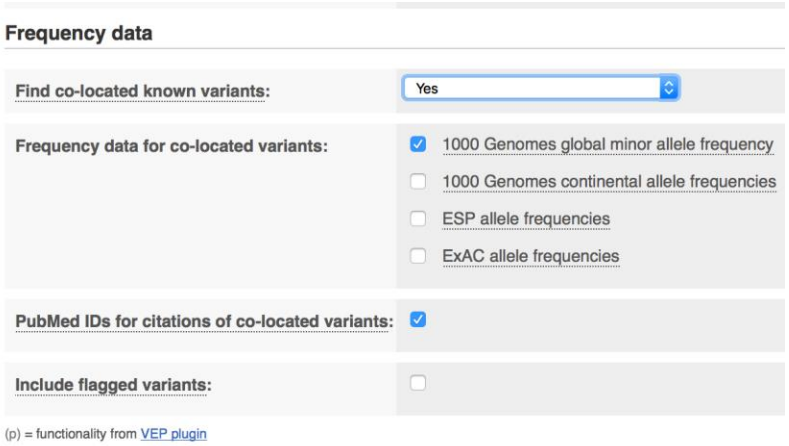
Pates these identifiers into the Ve!P box next to “Either paste data”. Please note that one can change the current GRCh38.p5 to the previous assembly GRCh37 or the species. For this example, we will use the GRCh38 assembly and Human (Homo sapiens) as species.



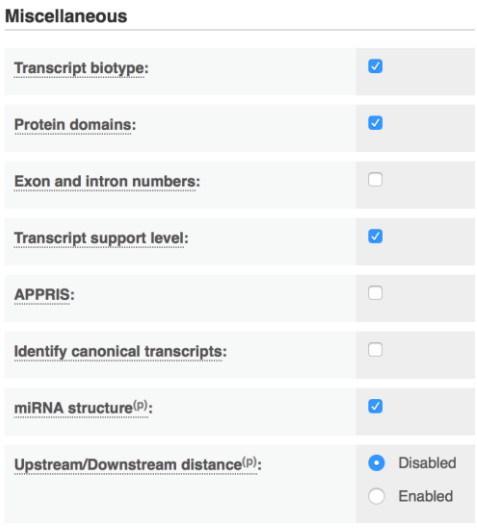
Provide a name for your work in the box next to “Name for this data (optional). Select “the Transcript database to use” as “Ensembl and RefSeq transcripts”. Under the “Identifiers” tab, select the “Gene symbol” (default). Please note that there are many other identifiers available (e.g. UniProt, etc.).



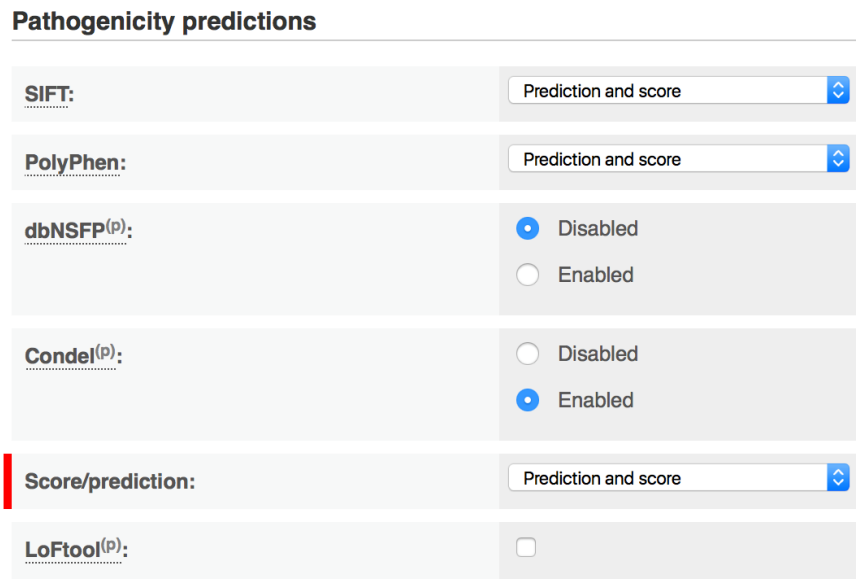
Under "Frequency data" section. Select "Yes" from the drop-down menu next to "Find co-located known variants" and check the "1000 Genomes global minor allele frequency" (default).



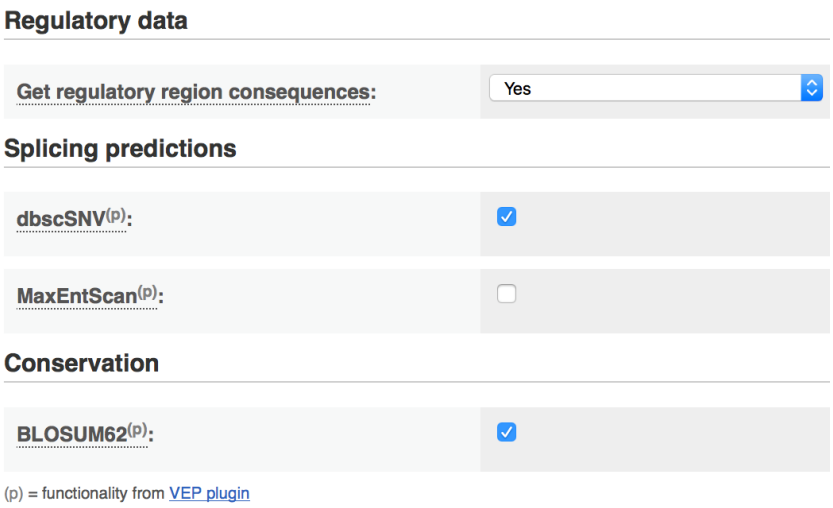
Under the "Extra options" tab select the annotations of your interest. For this example: "Transcript biotype", "Protein domain", "Transcript support level", and "miRNA structure".



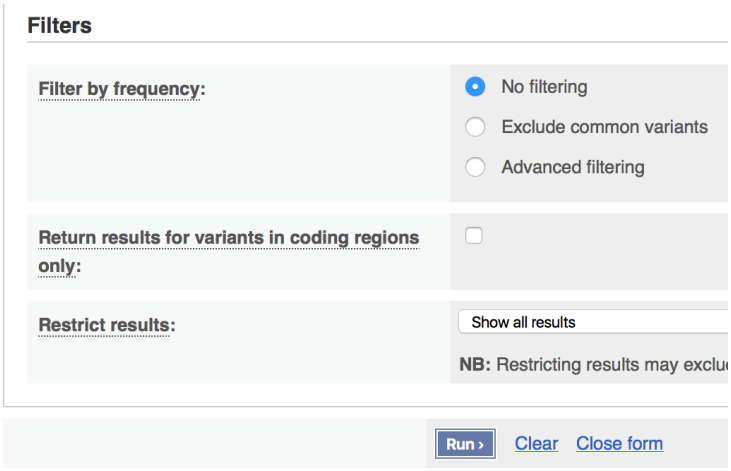
Under the “Pathogenicity predictions” tab, select “SIFT: Prediction and score”, “PolyPhen: Prediction and score”, “Condel: Enabled – Prediction and score”.



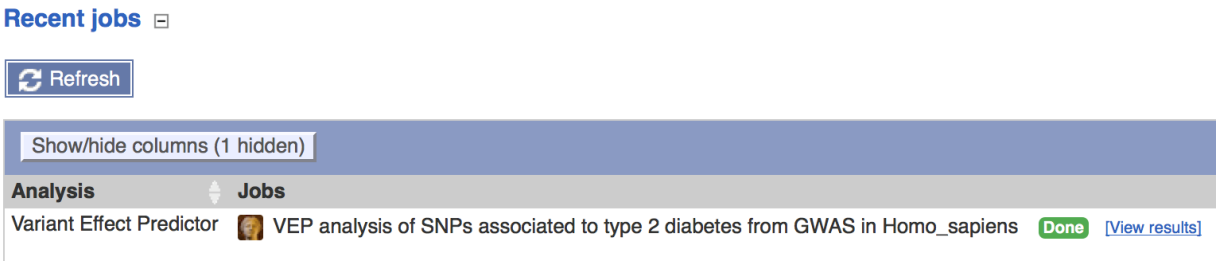
Under “Regulatory data”, select “dbscSNV”. Under “Conservation”, select BLOSUM62 for this example.



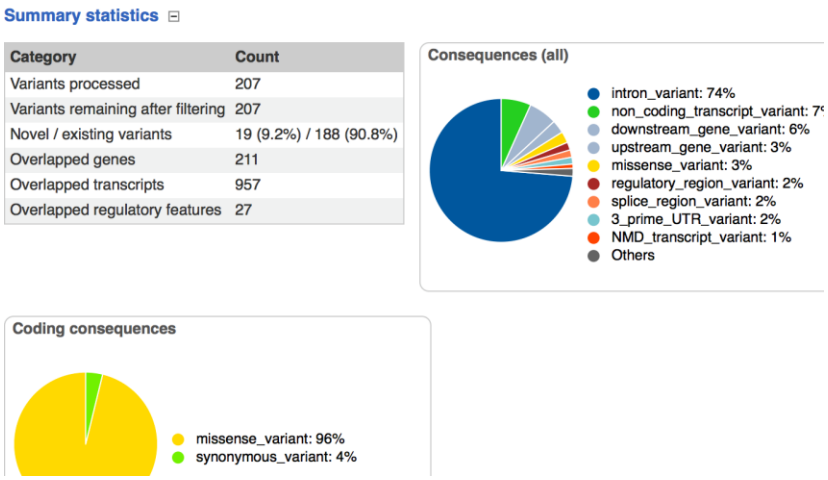
We will not filter the SNP list for any of the options available under “Filters”. Select “No filtering” (default) and click on the “Run” button.



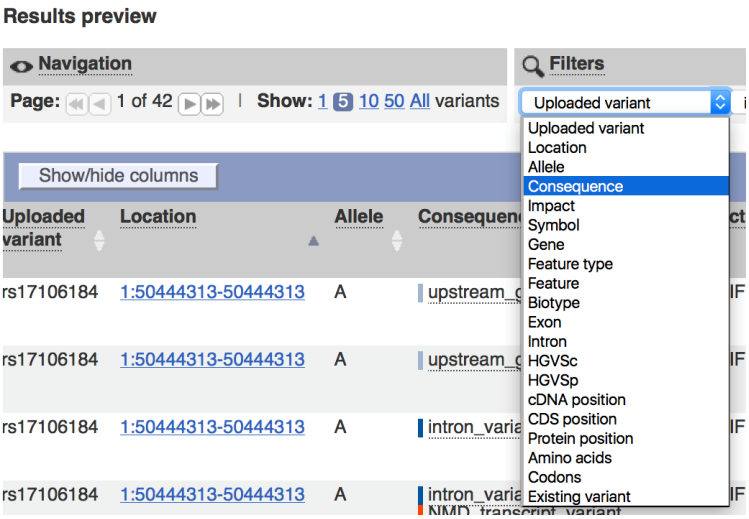
Your job will be queued until done. Once it is done, click on the "View results" link.



A new page will open showing the annotation results of the SNP list consisting of a summary statistics, variant functional and coding consequences followed by a table.

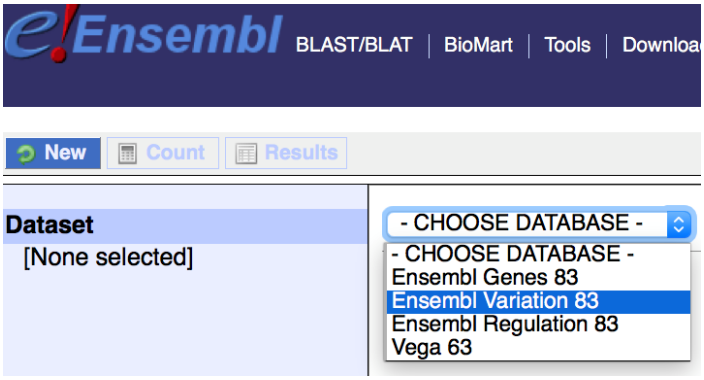


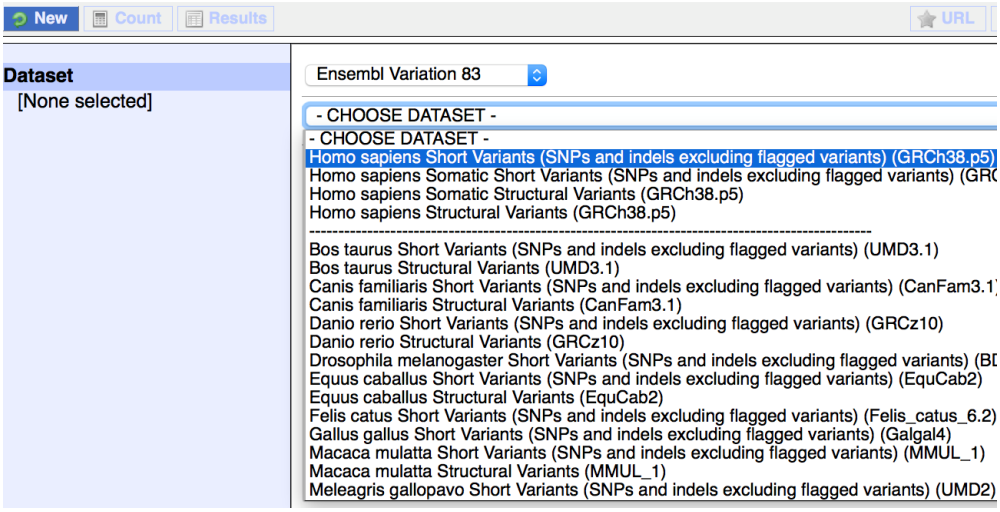
The results in the table can be further filtered by location, allele, consequence, allele frequency, clinical significance, etc. This table can be downloaded as VCF, VEP format, or TEX (best for Excel) for further analysis. Help on Ensembl prediction and variant classification can be found here:
http://www.ensembl.org/info/genome/variation/predicted_data.html



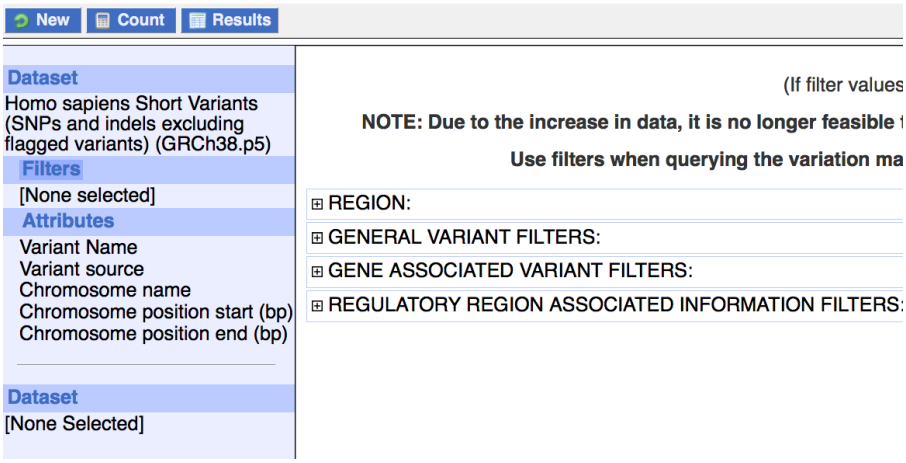
Additional Annotation of Variants: Ensembl BioMart

In this example, we will use BioMart (Kinsella et al., 2011) to find whether these variants have a other phenotype associated and a Mendelian Inheritance in Man entry and description. In addition, BioMart can be use to add additional annotations related not only to the variant but also to associated genes and regulatory regions.
Open Ensembl BioMart http://www.ensembl.org/biomart
Select "Ensembl Variation 83" from the CHOOSE DATABASE drop-down menu. Select "Homo sapiens Short Variants (SNPs and indels excluding flagged variants) GRCh38.p5" from the CHOOSE DATASET drop-down menu.



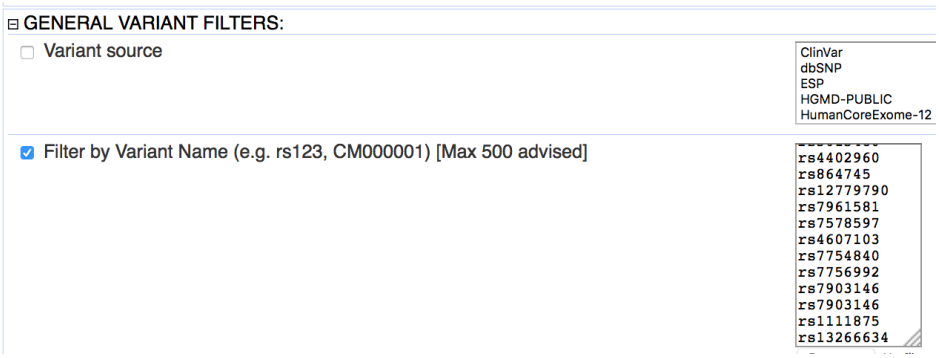


Click on the "Filter" link located on the left side of the page. A new menu will appear in the right pane. Expand the "General Variant Filters" heading by clicking on the plus sign next to it. Check the "Filter by Variant Name" option and paste your 124 variants associated with type 2 diabetes found in the GWAS Catalog - listed below:

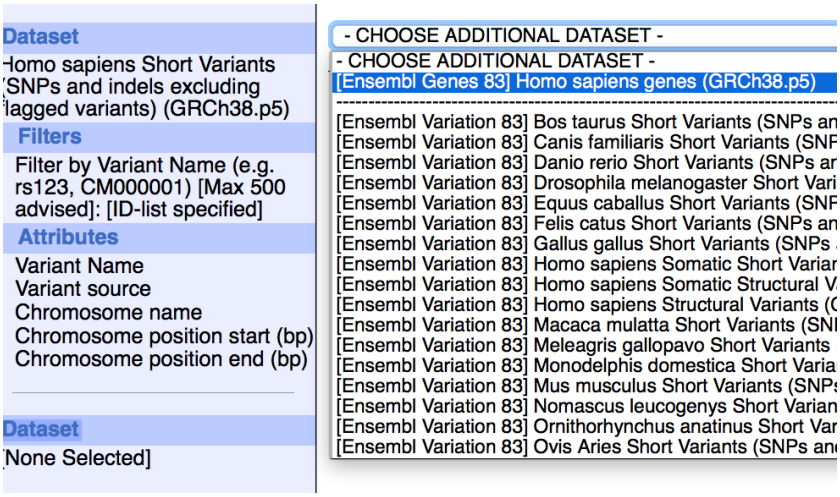


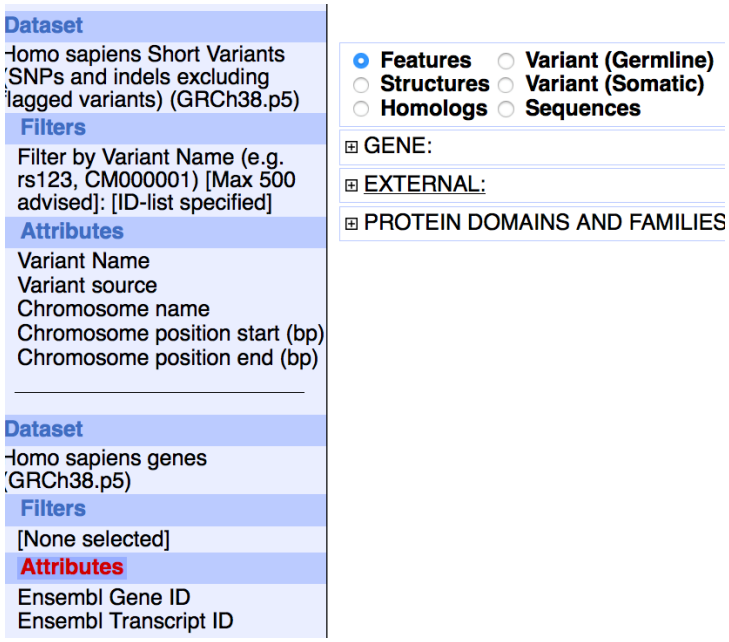
rs8090011	rs12779790	rs10923931	rs7178572
rs7178572	rs7961581	rs6931514	rs2028299
rs7903146	rs7578597	rs7903146	rs4812829
rs7766070	rs4607103	rs5215	rs7593730
rs3802177	rs7754840	rs1048886	rs243021
rs7903146	rs7756992	rs1333051	rs4457053
rs9939609	rs7903146	rs7305618	rs972283
rs7766070	rs7903146	rs163182	rs896854
rs5015480	rs1111875	rs3923113	rs13292136
rs4402960	rs13266634	rs16861329	rs231362
rs864745	rs7903146	rs1802295	rs1552224

rs1531343	rs10946398	rs10811661	rs7903146
rs7957197	rs10811661	rs7903146	rs9552911
rs11634397	rs4402960	rs1111875	rs7903146
rs8042680	rs7903146	rs2237892	rs7903146
rs5945326	rs4402960	rs5945326	rs1470579
rs7578326	rs7754840	rs7754840	rs7903146
rs1387153	rs10811661	rs7903146	rs1470579
rs1470579	rs8050136	rs7903146	rs7903146
rs1801214	rs13266634	rs3842770	rs1470579
rs10440833	rs1111875	rs2283228	rs7903146
rs849134	rs5219	rs343092	rs7756992
rs3802177	rs4402960	rs2244020	rs17791513
rs10965250	rs13266634	rs7903146	rs1111875
rs5015480	rs7901695	rs2283228	rs11257655
rs7903146	rs391300	rs343092	rs163184
rs11642841	rs17584499	rs231356	rs849135
rs7172432	rs2237895	rs6815464	rs5215
rs10906115	rs515071	rs7041847	rs4812829
rs1359790	rs7656416	rs6017317	rs2261181
rs8050136	rs7903146	rs6467136	rs9936385
rs5219	rs7560163	rs831571	rs4402960
rs7903146	rs10886471	rs9470794	rs1801282
rs7903146	rs7403531	rs3786897	rs4430796
rs2237892	rs10814916	rs1535500	rs10811661
rs4712524	rs2383208	rs10229583	rs3802177
rs6769511	rs11257655	rs75493593	rs4458523
rs2237897	rs4430796	rs6813195	rs2943640
rs4712523	rs12010175	rs9502570	rs7612463
rs2383208	rs5945326	rs17106184	rs7178572
rs2237892	rs791595	rs3132524	rs12571751
rs13266634	rs11787792	rs6808574	rs11717195
rs1111875	rs312457	rs702634	rs12970134
rs7903146	rs8181588	rs1727313	rs4506565
rs7903146	rs1470579	rs7903146	rs7018475
rs8050136	rs7754840	rs2237897	rs7766070

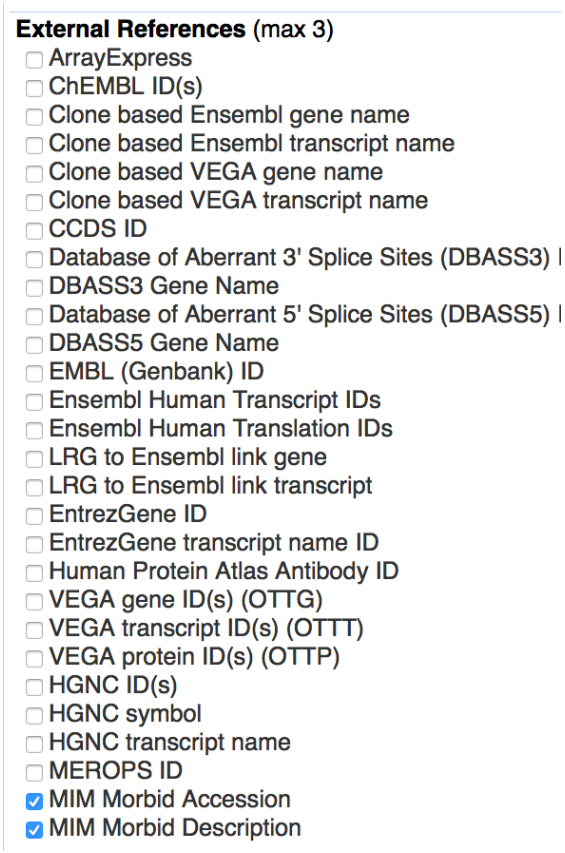


In order to find the MIM annotations for these variants we need to cross the Ensembl Variation 83 with the Ensembl Genes 83 database (containing the MIM annotations). For this, click on the "Dataset" link located at the bottom of the left pane under "Attributes". Select "(Ensembl Genes 83) Homo sapiens genes (GRCh38.p5)"

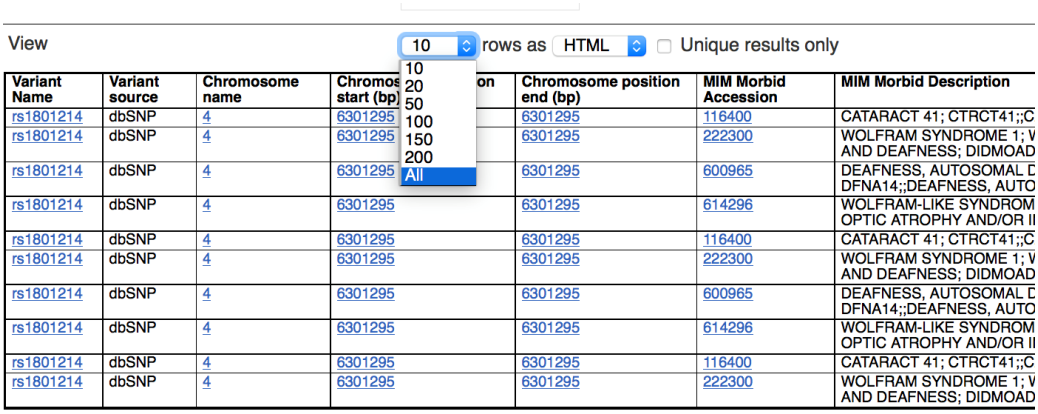




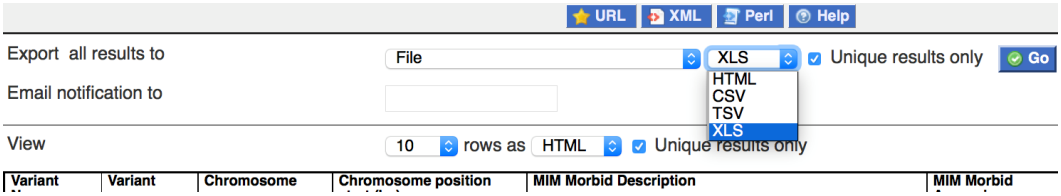
A new menu will appear on the right pane. Expand the "EXTERNAL" heading by clicking on the plus sign next to it. You might want to uncheck the "Ensembl Gene ID" and "Ensembl Transcript ID" located under the "GENE" heading if not needed since these will add additional columns to the table.



Click on the "Results" button located on top of the left pane. The first 10 rows of results will appear on the right pane. You can select "All" from the drop-down menu to see all results (will open on a new tab). Save the the web page as text.
Results can also be saved by using BioMart export function located on top of the right pane: "Export all results to: File". There is an option to export as XLS file format. Click on the "Go" button.



Variant Name	Variant source	Chromosome name	Chromos start (bp)	on	Chromosome position end (bp)	MIM Morbid Accession	MIM Morbid Description
rs1801214	dbSNP	4	6301295		6301295	116400	CATARACT 41; CTRCT41;;C
rs1801214	dbSNP	4	6301295		6301295	222300	WOLFRAM SYNDROME 1; V AND DEAFNESS; DIDMOAD
rs1801214	dbSNP	4	6301295		6301295	600965	DEAFNESS, AUTOSOMAL D DFNA14;;DEAFNESS, AUTO
rs1801214	dbSNP	4	6301295		6301295	614296	WOLFRAM-LIKE SYNDROM OPTIC ATROPHY AND/OR II
rs1801214	dbSNP	4	6301295		6301295	116400	CATARACT 41; CTRCT41;;C
rs1801214	dbSNP	4	6301295		6301295	222300	WOLFRAM SYNDROME 1; V AND DEAFNESS; DIDMOAD
rs1801214	dbSNP	4	6301295		6301295	600965	DEAFNESS, AUTOSOMAL D DFNA14;;DEAFNESS, AUTO
rs1801214	dbSNP	4	6301295		6301295	614296	WOLFRAM-LIKE SYNDROM OPTIC ATROPHY AND/OR II
rs1801214	dbSNP	4	6301295		6301295	116400	CATARACT 41; CTRCT41;;C
rs1801214	dbSNP	4	6301295		6301295	222300	WOLFRAM SYNDROME 1; V AND DEAFNESS; DIDMOAD


URL
XML
Perl
Help
Export all results to
File
XLS
HTML
CSV
TSV
XLS
Unique results only
Go
Email notification to
View
10
rows as
HTML
Unique results only
Variant
Variant
Chromosome
Chromosome position
MIM Morbid Description
MIM Morbid

References

Kinsella, R. J., Kahari, A., Haider, S., Zamora, J., Proctor, G., Spudich, G., . . . Flicek, P. (2011). Ensembl BioMarts: a hub for data retrieval across taxonomic space. *Database (Oxford), 2011*, bar030. doi:10.1093/database/bar030

McLaren, W., Pritchard, B., Rios, D., Chen, Y., Flicek, P., & Cunningham, F. (2010). Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics, 26*(16), 2069-2070. doi:10.1093/bioinformatics/btq330

Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., . . . Parkinson, H. (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res, 42*(Database issue), D1001-1006. doi:10.1093/nar/gkt1229

Zeeberg, B. R., Riss, J., Kane, D. W., Bussey, K. J., Uchio, E., Linehan, W. M., . . . Weinstein, J. N. (2004). Mistaken identifiers: gene name errors can be introduced inadvertently when using Excel in bioinformatics. *BMC Bioinformatics, 5*, 80. doi:10.1186/1471-2105-5-80