SOLVED BIOLOGICAL AND CLINICAL DATABASES EXERCISES. GDA2017

Exercise 1. Go to the NAR molecular biology database collection and find a database about germline *de novo* variants identified in the human genome.

- Type the NAR database collection URL on your browser (http://www.oxfordjournals.org/nar/database/a/)
- Select "Category list" option from the right menu and go to "Human diseases and genes" option.
- Then, select "General polymorphism databases" and find denovo-db.

Exercise 2. Go to the Gene Expression Omnibus repository browser (http://www.ncbi.nlm.nih.gov/geo/browse/) and search data for lung cancer. How many samples of lung cancer do you find?

- Type the GEO browser URL on your browser (http://www.ncbi.nlm.nih.gov/geo/browse/)
- Type "lung cancer" in the search field. Then, check the number of series and samples. Do you know the difference between them?

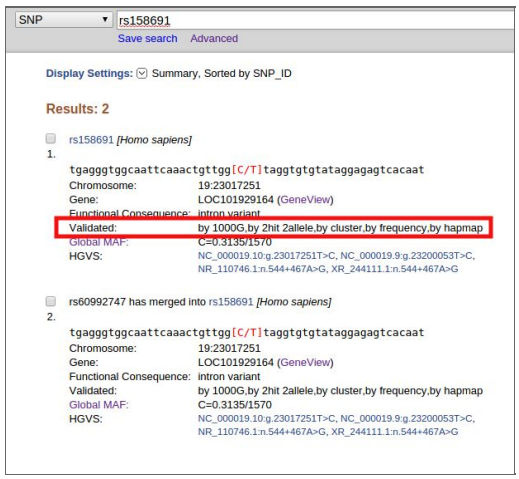
Exercise 3. Search information for specific SNVs in different databases.

Questions:

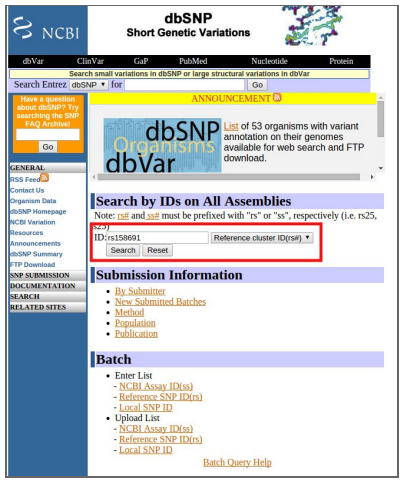
A) dbSNP database: what can you say about dbSNP id rs158691 from dbSNP database? has it been validated? how?

- Type the dbSNP URL on your browser (http://www.ncbi.nlm.nih.gov/SNP/). There are two fields for searching using dbSNP id: the first one at the upper part of the web page and the second one at the "Search by IDs on All Assemblies" section.
- Search for dbSNP id rs158691 through the first option,

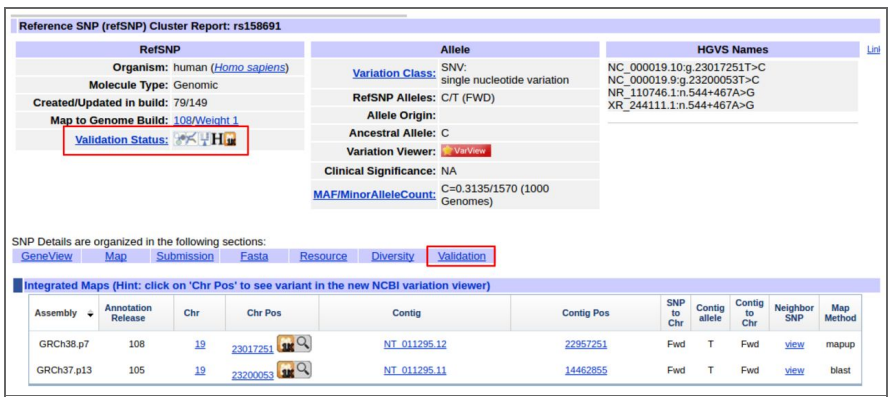
- The result of the first field search gives information about how rs158691 has been validated:



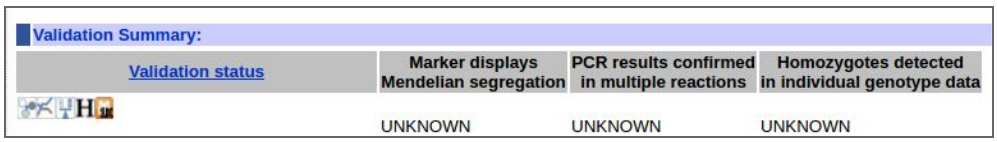
- Then, search for dbSNP id rs158691 through the second option,



- With this search field, you go straightforward to the report page where you can find more information about this SNP. Specifically the information about validation can be found in several parts of the web page (most relevant are highlighted in red),



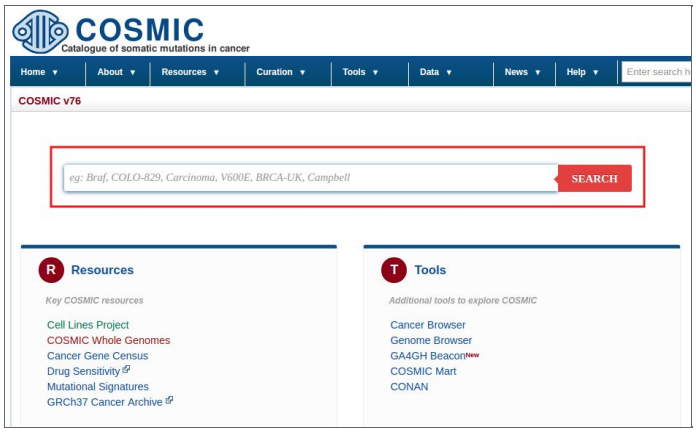
- When looking at the Validation status in the dbSNP report, the field includes some information which is slightly extended in the **Validation section** of the report,



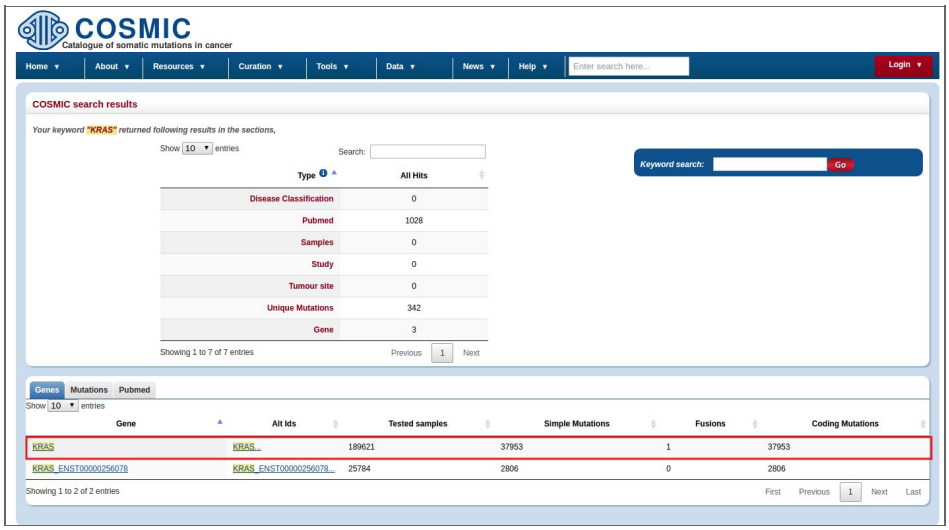
- In some cases the information in both sections could differ or be missing. Then, how can we consider this information reliable? One option is to search for the rs158691 in other databases such as Ensembl or check its population frequency in different human variation catalogs.
- More information about validation status in dbSNP: http://www.ncbi.nlm.nih.gov/books/NBK44476/#Reports.what_exactly_does_it_mean_when_a
- More information about validation status in Ensembl: http://www.ensembl.org/info/genome/variation/data_description.html#evidence_status

B) COSMIC database: which is the KRAS gene position with highest substitution rate found in cancers? which is the most common substitution in this position? Is there any specific tissue distribution for this mutation?

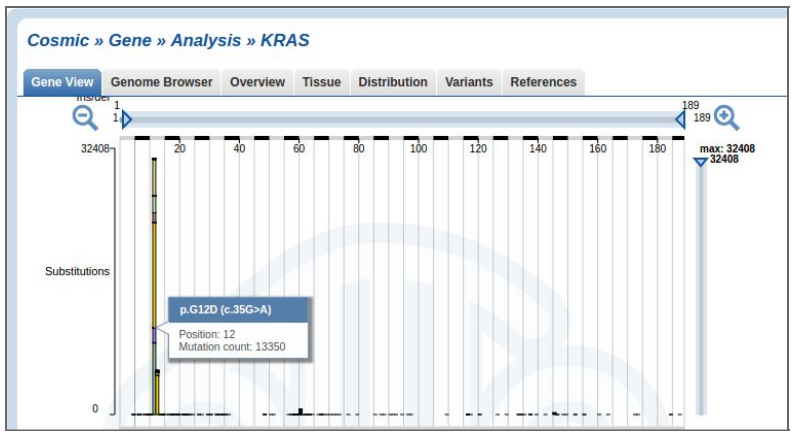
- Type the COSMIC URL on your browser (http://cancer.sanger.ac.uk/cosmic) and search for KRAS gene in the “Search” field.



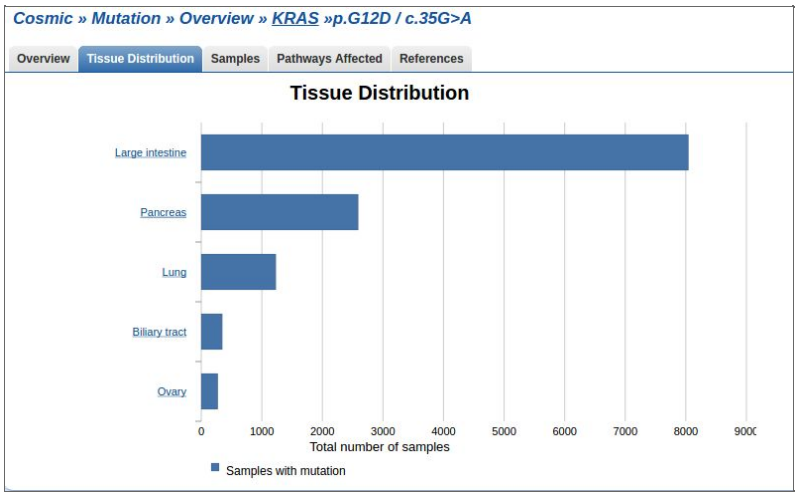
- Select the first Gene ID (“KRAS”) in the results page.



- In the next page you can find different bar plots with gene information. The first plot includes the counts of substitutions along the gene. Here, you can find that the position with the highest number of substitutions is position 12. Passing the mouse over the bars in the plot, some pop-up information appear. If you pass the mouse over the widest bar of the position 12, you can see that substitution p.G12D/c.35G>A has been observed 13879 times.



- Click on the region of the previous bar at position 12 (p.G12D/c.35G>A). There you can find information about the selected substitution. Click on the “Tissue Distribution” at the tab menu on the top to see its tissue frequency.



C) humsaVar database: could you find the previous rs158691 SNP in this file? why?

- Type the humsaVar URL on your browser (http://www.uniprot.org/docs/humsavar). The information of this database is contained in a text file that you can download from its web page. You can search for rs158691 either within the text file or directly in the web page using the search option of the browser.

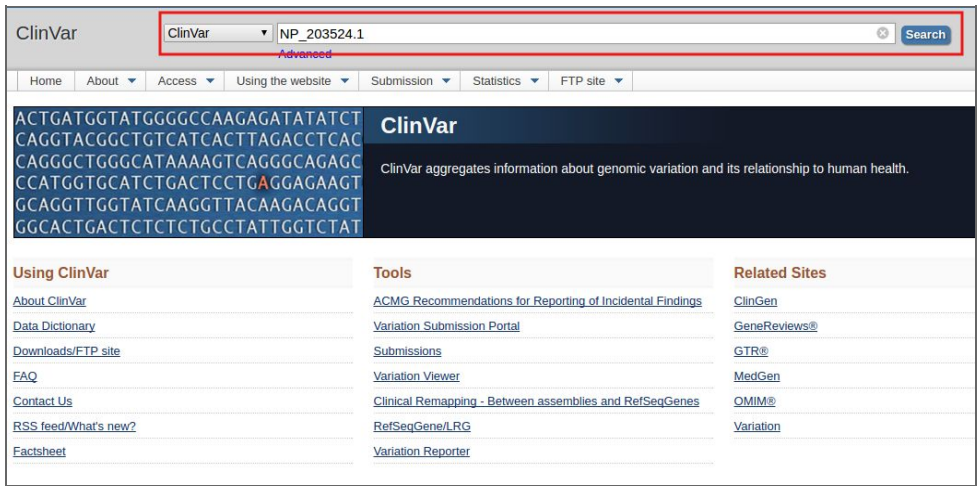
rs158691 0 of 0

Main gene name	Swiss-Prot AC	FTId	AA change	Type of variant	dbSNP	Disease name
A1BG	P04217	VAR_018369	p.His52Arg	Polymorphism	rs893184	-
A1BG	P04217	VAR_018370	p.His395Arg	Polymorphism	rs2241788	-
A1CF	Q9NQ94	VAR_052201	p.Val555Met	Polymorphism	rs9073	-
A1CF	Q9NQ94	VAR_059821	p.Ala558Ser	Polymorphism	rs11817448	-
A2ML1	A8K2U0	VAR_055463	p.Gly207Arg	Polymorphism	rs11047499	-
A2ML1	A8K2U0	VAR_055464	p.Cys970Tyr	Polymorphism	rs1558526	-
A2ML1	A8K2U0	VAR_055465	p.Thr1131Met	Polymorphism	rs7959680	-
A2ML1	A8K2U0	VAR_055466	p.Thr1412Ala	Polymorphism	rs7315591	-
A2ML1	A8K2U0	VAR_059083	p.Asp850Glu	Polymorphism	rs1860926	-
A2ML1	A8K2U0	VAR_059084	p.His1229Arg	Polymorphism	rs10219561	-
A2ML1	A8K2U0	VAR_071854	p.Arg1122Trp	Polymorphism	rs1860967	-
A2ML1	A8K2U0	VAR_071855	p.Met1257Val	Polymorphism	rs7308811	-
A2ML1	A8K2U0	VAR_071856	p.Thr1312Met	Polymorphism	rs201083574	-

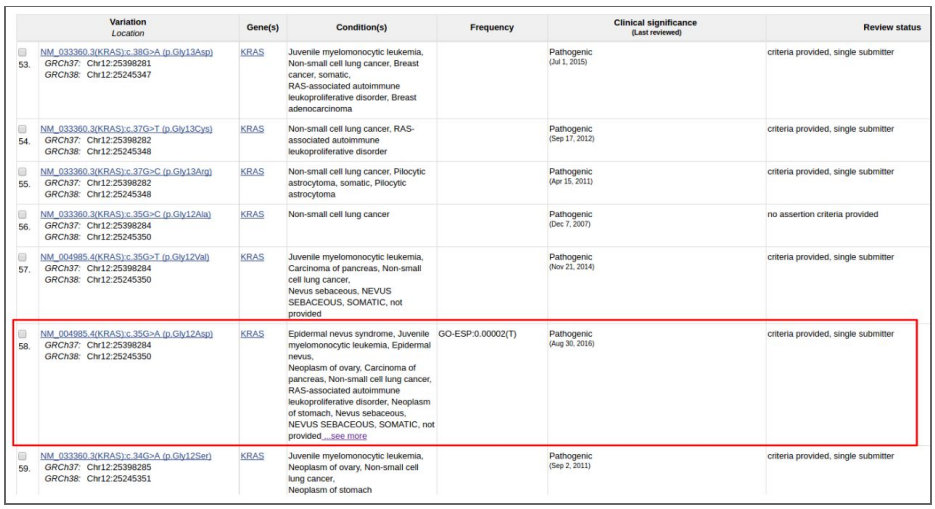
- Searching directly in the web page, you can't find any result for rs158691 because it is an intron variant. Note that humsaVar has been developed by UNIPROT, which is a well known and curated database for proteins (gene exons).

D) ClinVar database: browse the clinical information reported for the conserved domain database (CDD) id NP_203524.1. Does it include the variant detected in B? which is its clinical significance? ant its review status? Note: CDS Mutation ID c.35G>A

- Type the ClinVar URL on your browser (http://www.ncbi.nlm.nih.gov/clinvar/) and search for NP_203524.1.



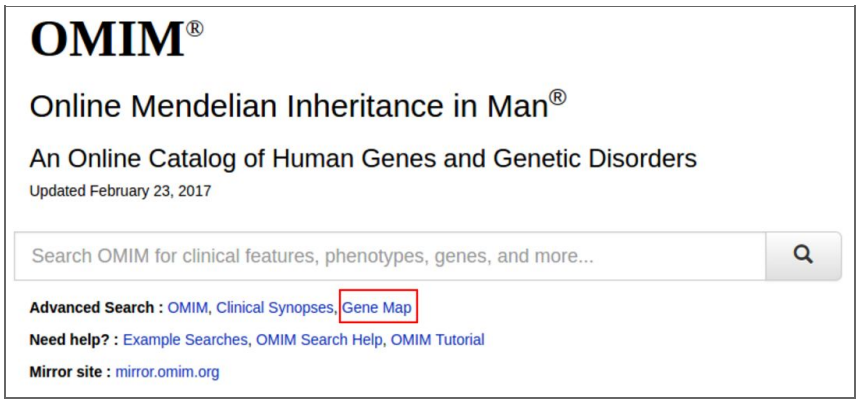
- The results page reports 67 items for NP_203524.1. In this page, you can search for c.35G>A, which is the CDS mutation ID from Exercise 1B. Then, you can find that its clinical significance states that is pathogenic and criteria is provided by single submitter.



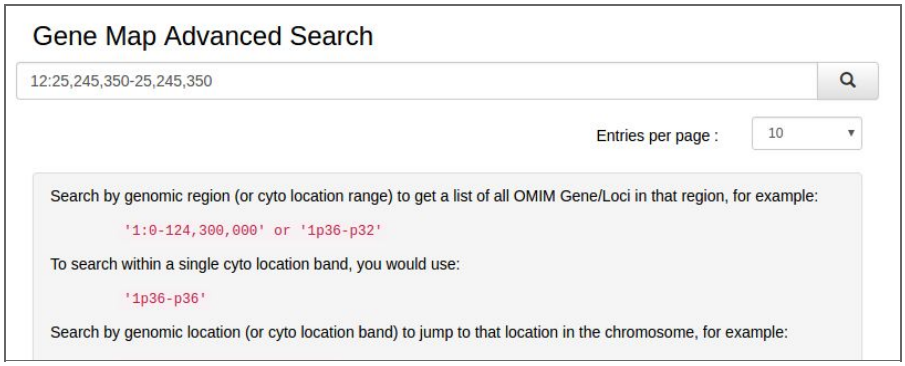
	Variation Location	Gene(s)	Condition(s)	Frequency	Clinical significance (Last reviewed)	Review status
53.	NM_033360.3(KRAS):c.38G>A (p.Gly13Asp) GRCh37: Chr12:25398281 GRCh38: Chr12:25245347	KRAS	Juvenile myelomonocytic leukemia, Non-small cell lung cancer, Breast cancer, somatic, RAS-associated autoimmune leukoproliferative disorder, Breast adenocarcinoma		Pathogenic (Jul 1, 2015)	criteria provided, single submitter
54.	NM_033360.3(KRAS):c.37G>T (p.Gly13Cys) GRCh37: Chr12:25398282 GRCh38: Chr12:25245348	KRAS	Non-small cell lung cancer, RAS-associated autoimmune leukoproliferative disorder		Pathogenic (Sep 17, 2012)	criteria provided, single submitter
55.	NM_033360.3(KRAS):c.37G>C (p.Gly13Arg) GRCh37: Chr12:25398282 GRCh38: Chr12:25245348	KRAS	Non-small cell lung cancer, Pilocytic astrocytoma, somatic, Pilocytic astrocytoma		Pathogenic (Apr 15, 2011)	criteria provided, single submitter
56.	NM_033360.3(KRAS):c.35G>C (p.Gly12Ala) GRCh37: Chr12:25398284 GRCh38: Chr12:25245350	KRAS	Non-small cell lung cancer		Pathogenic (Dec 7, 2007)	no assertion criteria provided
57.	NM_004985.4(KRAS):c.35G>T (p.Gly12Val) GRCh37: Chr12:25398284 GRCh38: Chr12:25245350	KRAS	Juvenile myelomonocytic leukemia, Carcinoma of pancreas, Non-small cell lung cancer, Nevus sebaceous, NEVUS SEBACEOUS, SOMATIC, not provided		Pathogenic (Nov 21, 2014)	criteria provided, single submitter
58.	NM_004985.4(KRAS):c.35G>A (p.Gly12Asp) GRCh37: Chr12:25398284 GRCh38: Chr12:25245350	KRAS	Epidermal nevus syndrome, Juvenile myelomonocytic leukemia, Epidermal nevus, Neoplasm of ovary, Carcinoma of pancreas, Non-small cell lung cancer, RAS-associated autoimmune leukoproliferative disorder, Neoplasm of stomach, Nevus sebaceous, NEVUS SEBACEOUS, SOMATIC, not provided ...see more	GO-ESP:0.00002(T)	Pathogenic (Aug 30, 2016)	criteria provided, single submitter
59.	NM_033360.3(KRAS):c.34G>A (p.Gly12Ser) GRCh37: Chr12:25398285 GRCh38: Chr12:25245351	KRAS	Juvenile myelomonocytic leukemia, Neoplasm of ovary, Non-small cell lung cancer, Neoplasm of stomach		Pathogenic (Sep 2, 2011)	criteria provided, single submitter

E) OMIM database: search for the chromosome location of the B result. Is there any nearby clinical annotation that makes sense with the KRAS gene? (Note that OMIM mapping uses build GRCh38)

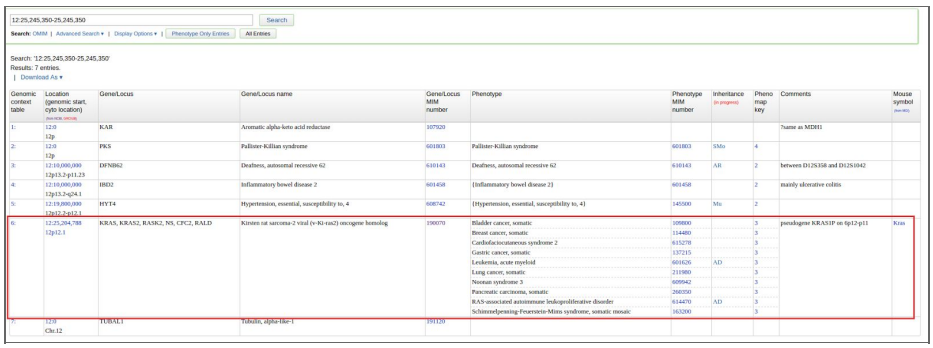
- Type the OMIM URL on your browser (http://www.omim.org/) and click on “Gene Map” at “Advanced Search” section.



- Then, search for the location 12:25,245,350-25,245,350. Note the OMIM special format with commas.

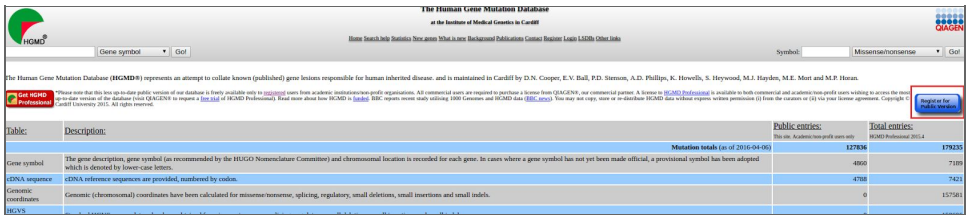


- In the next results page, you can find 12:25,204,788 as the nearest KRAS position to the selected substitution in Exercise 1B.

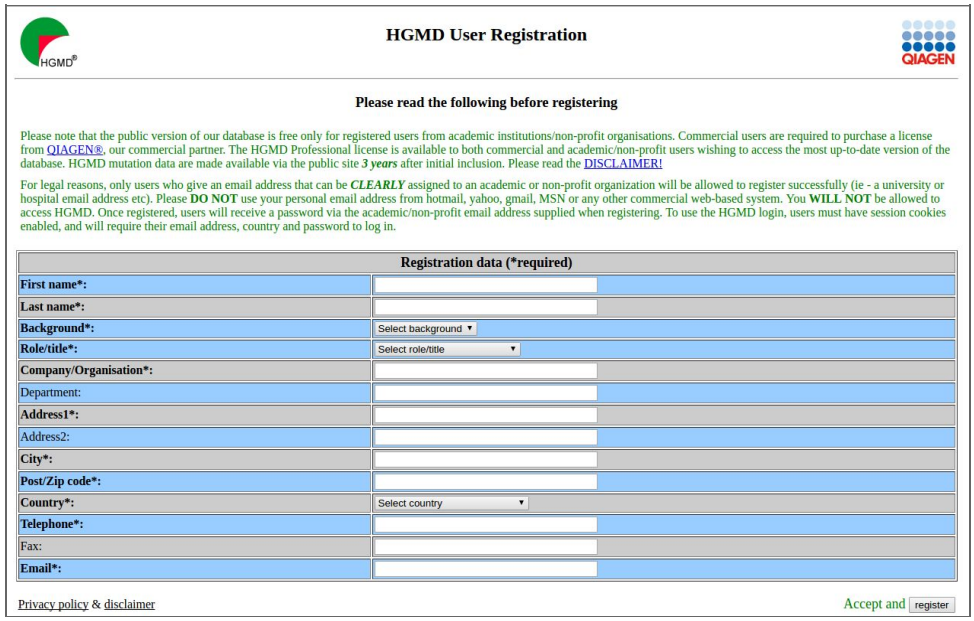


F) HGMD database: register for the public version and try it at home.

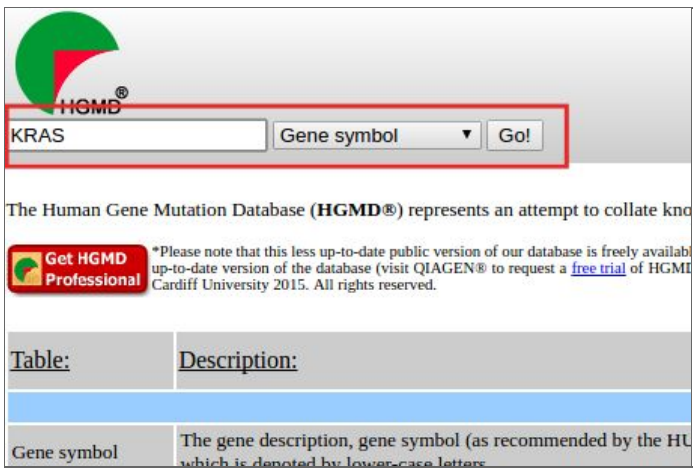
- Type the HGMD URL on your web browser (http://www.hgmd.cf.ac.uk/ac/index.php). Click on “Register for public version” button.



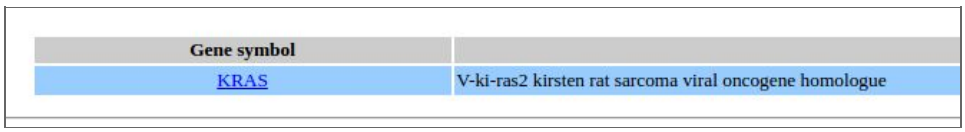
- Then, fill the form to get access to the public version of HGMD.



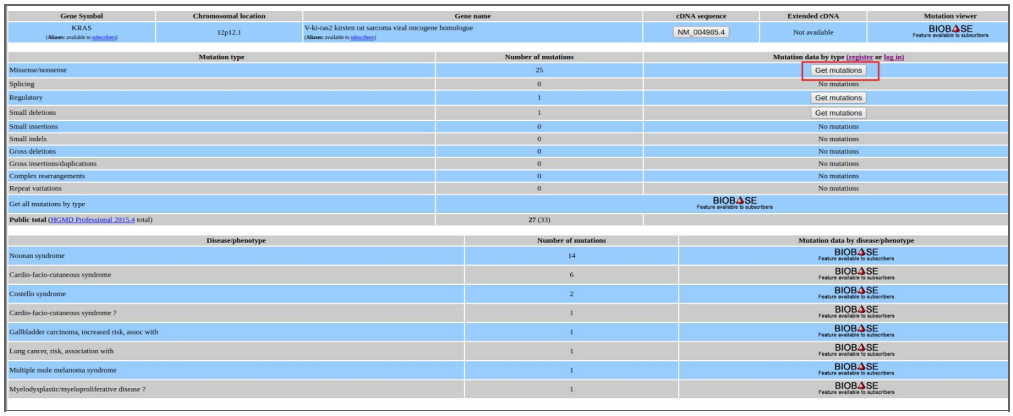
- Once you have logged in, search for KRAS gene on the upper left.



- Select KRAS in the following table.



- The next results page includes several information about KRAS, but some of it is only accessible from the professional version. Click on “Get mutations” of missense/nonsense type.

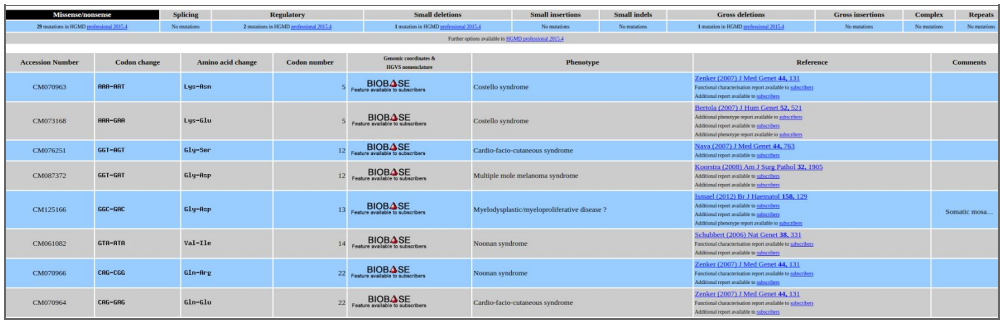


Gene Symbol	Chromosomal location	Gene name	cDNA sequence	Extended cDNA	Mutation viewer
KRAS (Aliases: available to subscribers)	12p12.1	V-ki-ras2 kirsten rat sarcoma viral oncogene homologue (Aliases: available to subscribers)	NM_004985.4	Not available	BIOBASE Feature available to subscribers

Mutation type	Number of mutations	Mutation data by type (register or log in)
Missense/nonsense	25	Get mutations
Splicing	0	No mutations
Regulatory	1	Get mutations
Small deletions	1	Get mutations
Small insertions	0	No mutations
Small indels	0	No mutations
Gross deletions	0	No mutations
Gross insertions/duplications	0	No mutations
Complex rearrangements	0	No mutations
Repeat variations	0	No mutations
Get all mutations by type	BIOBASE Feature available to subscribers	
Public total (HGMD Professional 2015.4 total)	27 (33)	

Disease/phenotype	Number of mutations	Mutation data by disease/phenotype
Noonan syndrome	14	BIOBASE Feature available to subscribers
Cardio-facio-cutaneous syndrome	6	BIOBASE Feature available to subscribers
Costello syndrome	2	BIOBASE Feature available to subscribers
Cardio-facio-cutaneous syndrome ?	1	BIOBASE Feature available to subscribers
Gallbladder carcinoma, increased risk, assoc with	1	BIOBASE Feature available to subscribers
Lung cancer, risk, association with	1	BIOBASE Feature available to subscribers
Multiple mole melanoma syndrome	1	BIOBASE Feature available to subscribers
Myelodysplastic/myeloproliferative disease ?	1	BIOBASE Feature available to subscribers

- Here, you can find the codon and amino acid changes, as well as the phenotype it has been associated with.



Missense/nonsense	Splicing	Regulatory	Small deletions	Small insertions	Small indels	Gross deletions	Gross insertions	Complex	Repeats
29 mutations in HGMD professional 2015.4	No mutations	2 mutations in HGMD professional 2015.4	1 mutation in HGMD professional 2015.4	No mutations	No mutations	1 mutation in HGMD professional 2015.4	No mutations	No mutations	No mutations

Further options available to HGMD professional 2015.4

Accession Number	Codon change	Amino acid change	Codon number	Genomic coordinates & HGVS nomenclature	Phenotype	Reference	Comments
CM070963	AAA→AAT	Lys→Asn	5	BIOBASE Feature available to subscribers	Costello syndrome	Zenker (2007) J Med Genet 44, 131 Functional characterisation report available to subscribers Additional report available to subscribers	
CM073168	AAA→GAA	Lys→Glu	5	BIOBASE Feature available to subscribers	Costello syndrome	Bertola (2007) J Hum Genet 52, 521 Additional phenotype report available to subscribers Additional report available to subscribers Additional phenotype report available to subscribers	
CM076251	GGT→AGT	Gly→Ser	12	BIOBASE Feature available to subscribers	Cardio-facio-cutaneous syndrome	Nava (2007) J Med Genet 44, 763 Additional report available to subscribers	
CM087372	GGT→GAT	Gly→Asp	12	BIOBASE Feature available to subscribers	Multiple mole melanoma syndrome	Kourstra (2008) Am J Surg Pathol 32, 1905 Additional report available to subscribers Additional report available to subscribers	
CM125166	GGC→GAC	Gly→Asp	13	BIOBASE Feature available to subscribers	Myelodysplastic/myeloproliferative disease ?	Ismpel (2012) Br J Haematol 158, 129 Additional report available to subscribers Additional report available to subscribers Additional phenotype report available to subscribers	Somatic mosa...
CM061082	GTA→ATA	Val→Ile	14	BIOBASE Feature available to subscribers	Noonan syndrome	Schubbert (2006) Nat Genet 38, 331 Functional characterisation report available to subscribers Additional report available to subscribers	
CM070966	CAG→CGG	Gln→Arg	22	BIOBASE Feature available to subscribers	Noonan syndrome	Zenker (2007) J Med Genet 44, 131 Functional characterisation report available to subscribers Additional report available to subscribers	
CM070964	CAG→GAG	Gln→Glu	22	BIOBASE Feature available to subscribers	Cardio-facio-cutaneous syndrome	Zenker (2007) J Med Genet 44, 131 Functional characterisation report available to subscribers Additional report available to subscribers	

Exercise 4. Retrieve genomic variation data from CellBase using its web services API. Note that the main host is http://ws.bioinfo.cipf.es/ (GRCh37) but there is another mirror in http://bioinfo.hpc.cam.ac.uk/cellbase/webservices/rest (GRCh38)

Some examples:
Get species included in CellBase:
http://ws.bioinfo.cipf.es/cellbase/rest/latest
Get all the mutations from BRCA2 gene:
http://ws.bioinfo.cipf.es/cellbase/rest/latest/hsa/feature/gene/BRCA2/mutation
Get all the genes within a specific genomic region:
http://ws.bioinfo.cipf.es/cellbase/rest/latest/hsa/genomic/region/1:3972105-12973105/gene
Get the phenotype from rs3934834 SNP:
http://ws.bioinfo.cipf.es/cellbase/rest/latest/hsa/feature/snp/rs3934834/phenotype

Questions:

A) We are interested in a particular region of the human genome chr12:25,350,000-25,245,000 (GRCh37), and we want to know if this region contains

mutations already catalogued. Help: latest (version), hsa (species), genomic (category), region (subcategory), 12:25350000-25450000 (id), mutation (resource).

- Query result: http://ws.bioinfo.cipf.es/cellbase/rest/latest/hsa/genomic/region/12:25350000-25450000/mutation

B) We want to know the allelic and genotypic frequencies for a SNP, rs158691, across populations. Help: latest (version), hsa (species), feature (category), snp (subcategory), rs158691 (id), population_frequency (resource).

- Query result: http://ws.bioinfo.cipf.es/cellbase/rest/latest/hsa/feature/snp/rs158691/population_frequency

C) We have obtained a SNP of interest (rs28937313, location GRCh37 9:107584801) in our analysis and we want to know if it has been related with any disease.

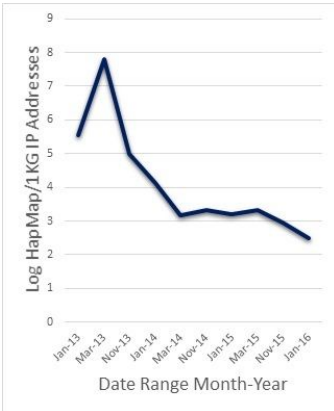
- Query result: http://ws.bioinfo.cipf.es/cellbase/rest/latest/hsa/feature/snp/rs28937313/phenotype

Exercise 5. Browse different catalogs of human genetic variation.

Questions:

A) The HapMap project (http://hapmap.ncbi.nlm.nih.gov) was a multi-country effort to identify and catalog genetic similarities and differences in human beings. The NCBI decided to retire this resource last year due to the observed decline of usage. Nevertheless, the HapMap data sets are still available via FTP. Which project has been established as the current standard for population genetics and genomics?

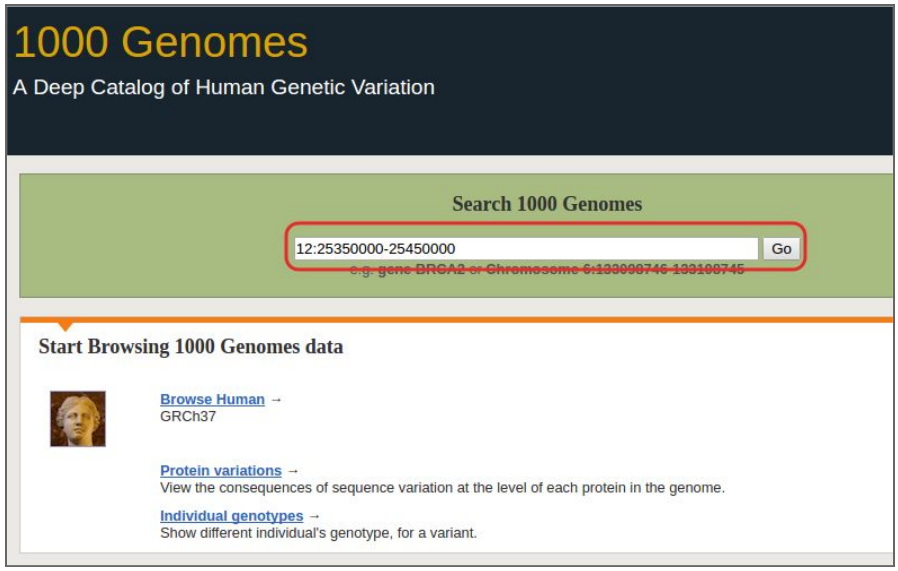
- Type the HapMap URL on your web browser (https://hapmap.ncbi.nlm.nih.gov/) and find out the standard project (1000 genomes project).



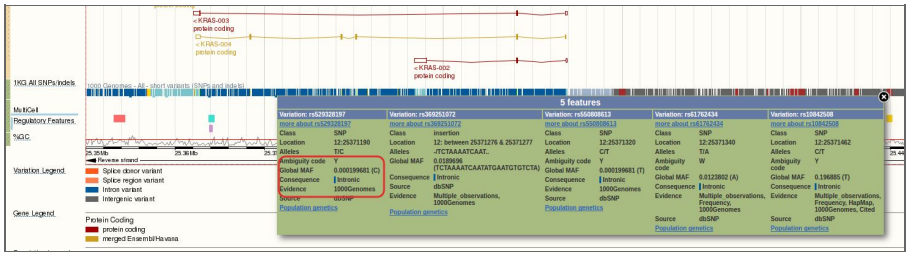
B) Now, go to the 1,000 Genomes browser and search for the KRAS genomic region (example: 12:25350000-25450000). Can you find the global MAFs of the SNPS in this region from the 1,000 Genome populations?

Note: this project is already finished but there are several available browsers. More information: http://www.internationalgenome.org/1000-genomes-browsers.

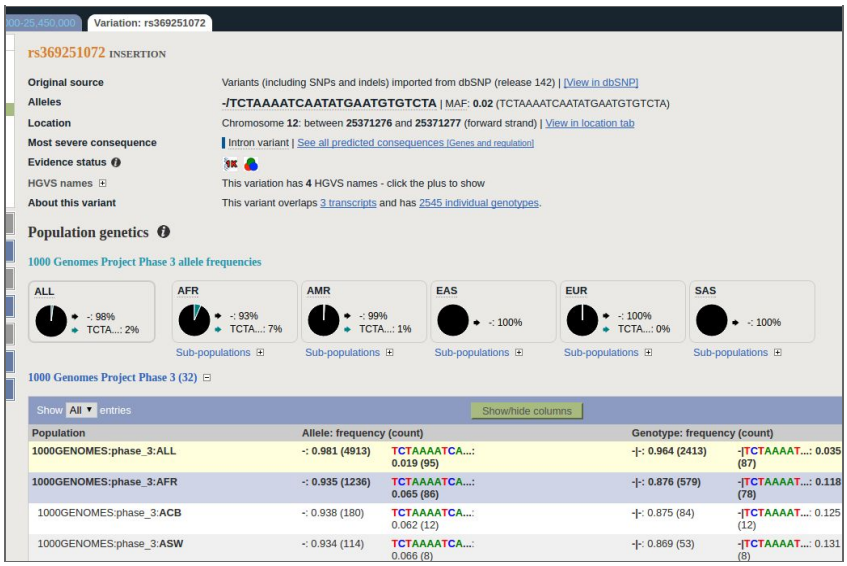
- Type the 1,000 Genomes URL in your browser (http://phase3browser.1000genomes.org/index.html) and search for KRAS region.



- In the results page, there is a section named “1KG All SNPs/indels” where you can find the Global MAFs of variations by clicking on each position.



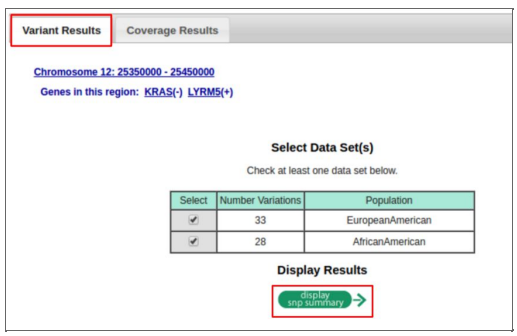
- Then, from the pop-up box, you can click on “Population genetics” and get more information about allele and genotype frequencies of each variant and population.

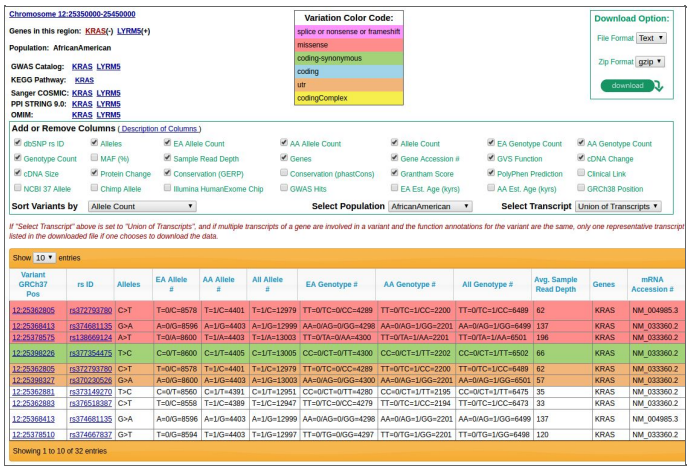


C) Check the allele frequencies of same genomic region in the ESP 6,500 samples.

- Type the ESP URL on your browser (http://evs.gs.washington.edu/EVS/) and search for 12:25350000-25450000 region.

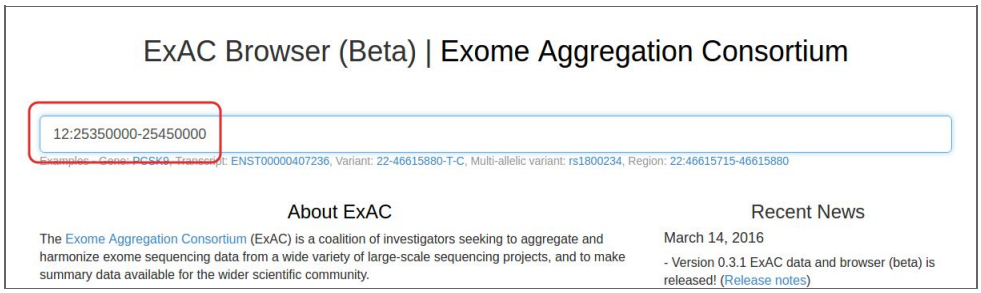
- Then, you will find that there are two genes in this region and two populations with different number of variants. Select “display snp summary” to check the variant information.



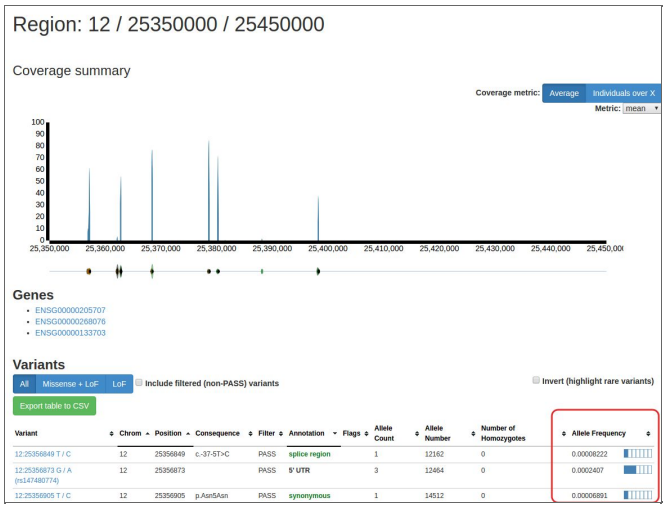


D) Check the genetic variation of KRAS in ExAC browser. Which is the allele frequency of rs121913529 in the European (Non-Finnish) population?

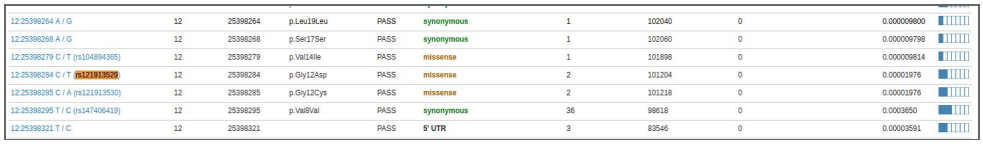
- Type the ExAC URL on your browser (http://exac.broadinstitute.org/) and search for 12:25350000-25450000 region.



- In the results page, you can find information about the variants located in the selected region. One of the columns of the table shown is called “Allele frequency”.

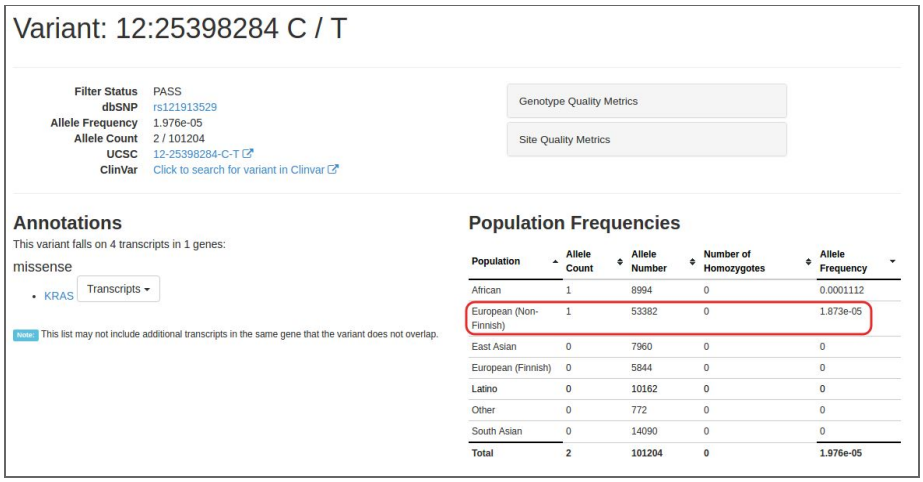


- Select the second gene Ensembl ID (ENSG00000133703, KRAS gene) from the previous web page and search for “rs121913529” in the results web page.



12:25398264 A / G	12	25398264	p.Leu19Leu	PASS	synonymous	1	102040	0	0.000009800
12:25398268 A / G	12	25398268	p.Ser17Ser	PASS	synonymous	1	102060	0	0.000009798
12:25398279 C / T (rs104894365)	12	25398279	p.Val14Ile	PASS	missense	1	101898	0	0.000009814
12:25398284 C / T rs121913529	12	25398284	p.Gly12Asp	PASS	missense	2	101204	0	0.00001976
12:25398285 C / A (rs121913530)	12	25398285	p.Gly12Cys	PASS	missense	2	101218	0	0.00001976
12:25398295 T / C (rs147406419)	12	25398295	p.Val8Val	PASS	synonymous	36	98618	0	0.0003650
12:25398321 T / C	12	25398321		PASS	5' UTR	3	83546	0	0.00003591

- Click on the link and check the European (Non-Finnish) population frequency (1.873e-05).



E) Finally, check the gene expression of KRAS in different tissues using the GTEx portal. Which is the tissue with the greatest expression? and the lowest?

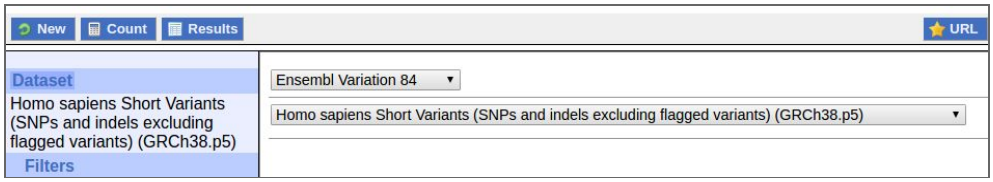
- Type the GTEx portal URL on your browser (http://www.gtexportal.org) and search the KRAS gene. In the graph with the boxplots you can find that “Nerve-tibial” is the one with the greatest expression and “Heart-left ventricle” together with “Muscle-skeletal”.

Exercise 6. Retrieve genomic variation data using Ensembl Biomart (Ensembl Variation database, http://www.ensembl.org/biomart).

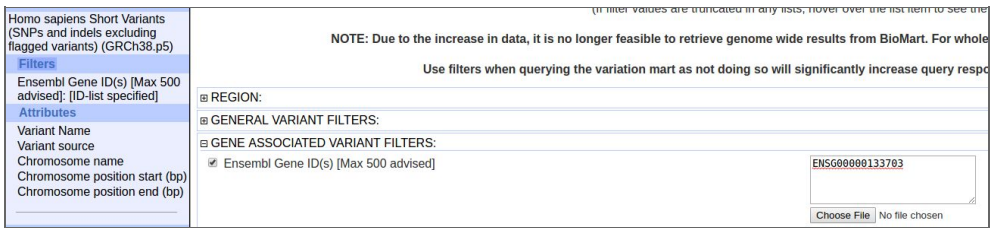
Questions:

A) Retrieve the variant alleles, the ancestral allele, the clinical significance, the SIFT and PolyPhen information about all the variants of the KRAS gene (ENSG00000133703).

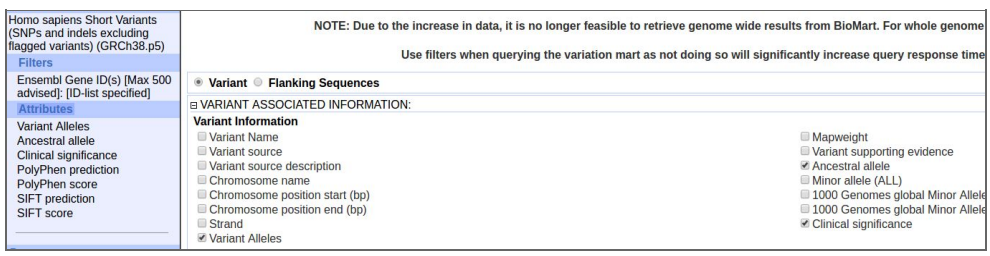
- Type the BioMart URL on your browser (http://www.ensembl.org/biomart) and choose “Ensembl Variation Database” and “Homo sapiens Short Variants (SNPs and indels excluding flagged variants)”.



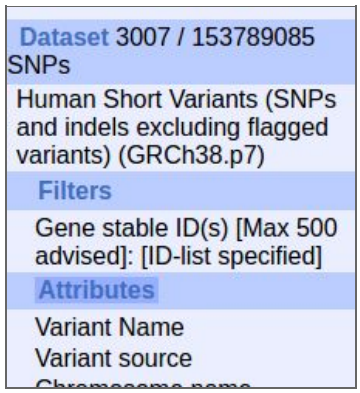
- Filter by gene ID at “Gene Associated Variant Filters” setting the Ensembl Gene ID to ENSG00000133703.



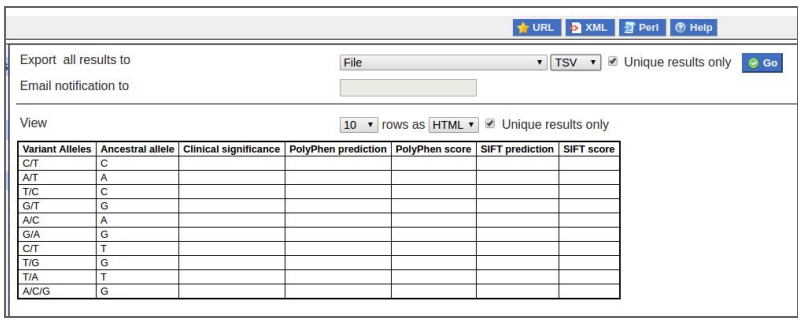
- Select the following attributes: “Variant alleles”, “Ancestral allele”, “Clinical significance” from “Variant associated information” section and “PolyPhen prediction”, “PolyPhen score”, “SIFT prediction”, “SIFT score” from “Gene Associated Information” section.



- Press “Count” button and you obtain 3007 SNPs.



- Click on “Results” button to retrieve a file with unique results.

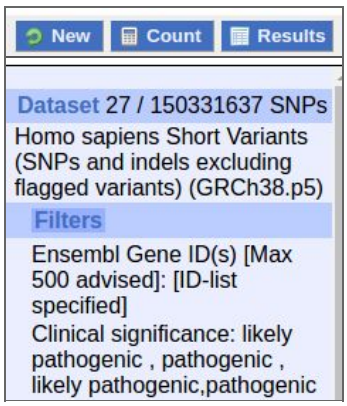


Variant Alleles	Ancestral allele	Clinical significance	PolyPhen prediction	PolyPhen score	SIFT prediction	SIFT score
C/T	C					
A/T	A					
T/C	C					
G/T	G					
A/C	A					
G/A	G					
C/T	T					
T/G	G					
T/A	T					
A/C/G	G					

B) Now, filter only the pathogenic ones using Biomart filters.

- Add a filter to the previous search. Select “Clinical Significance” from “General Variant Filters” and choose the following options: “likely pathogenic”, “pathogenic” and “likely pathogenic, pathogenic”.

- Press “Count” and now you obtain 27 SNPs.



C) Retrieve all the variants of the ABCA1 gene (ENSG00000165029) that are included in HGMD-Public database.

- Adjust the parameters for a new search filtering by Gene ID (ENSG00000165029) and HGMD-Public database as the variant source.