

## 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 screenshot shows the NCBI dbSNP website interface. At the top, there are navigation tabs for dbVar, ClinVar, GaP, PubMed, Nucleotide, and Protein. A search bar is present with the text "Search Entire dbSNP for rs158691" and a "Go" button. Below the search bar, there is a section titled "Search by IDs on All Assemblies" with a note: "Note: **rs#** and **ss#** must be prefixed with 'rs' or 'ss', respectively (i.e. rs25, ss25)". There is an "ID:" input field and a "Reference cluster ID(rsaf)" dropdown menu. Below this, there is a "Submission Information" section with links for "By Submitter", "New Submitted Batches", "Method", "Population", and "Publication". At the bottom, there is a "Batch" section with options for "Enter List" (including NCBI Assay ID(ss), Reference SNP ID(rs), and Local SNP ID) and "Upload List" (including NCBI Assay ID(ss), Reference SNP ID(rs), and Local SNP ID). A "Batch Query Help" link is also visible.

- 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),

Assembly	Annotation Release	Chr	Chr Pos	Contig	Contig Pos	SNP to Chr	Contig allele	Contig to Chr	Neighbor SNP	Map Method
GRCh38.p7	108	19	23017251	NT_011295.12	22957251	Fwd	T	Fwd	view	mapup
GRCh37.p13	105	19	2320053	NT_011295.11	14462855	Fwd	T	Fwd	view	blast

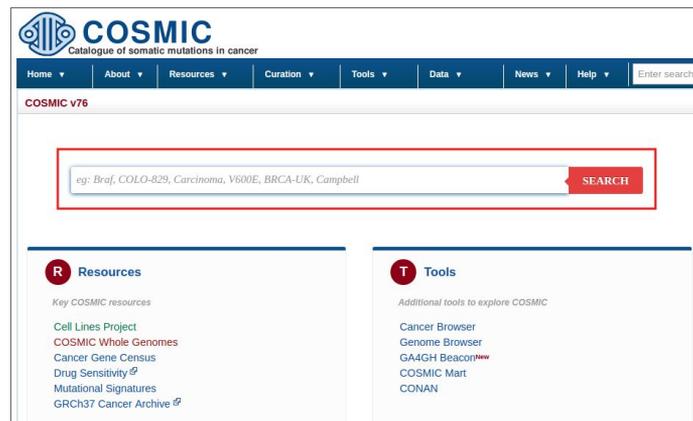
- When looking at the Validation status in the dbSNP report, the field includes some information which is slightly extended in the **Validation section** of the report,

Validation Summary:			
Validation status	Marker displays Mendelian segregation	PCR results confirmed in multiple reactions	Homozygotes detected in individual genotype data
	UNKNOWN	UNKNOWN	UNKNOWN

- 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](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](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.

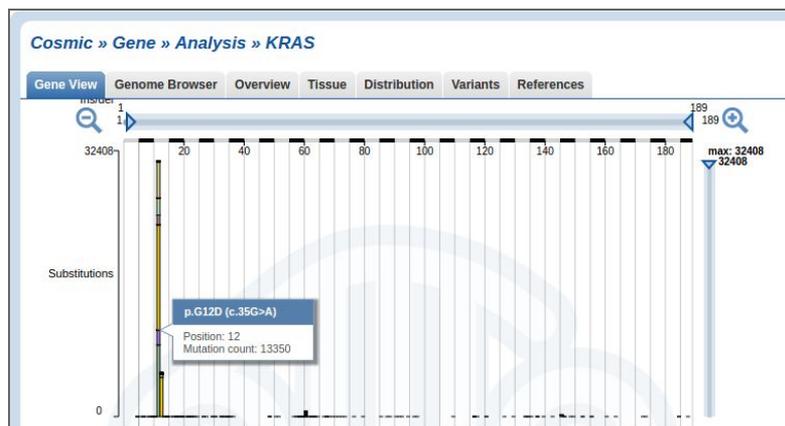
The screenshot shows the COSMIC website interface. At the top, there is a navigation bar with links for Home, About, Resources, Curation, Tools, Data, News, and Help. A search bar is present with the text "Enter search here...". Below the navigation bar, the page displays "COSMIC search results" for the keyword "KRAS". It indicates that 10 entries are shown. A summary table lists the following categories and counts:

Type	All Hits
Disease Classification	0
Pubmed	1028
Samples	0
Study	0
Tumour site	0
Unique Mutations	342
Gene	3

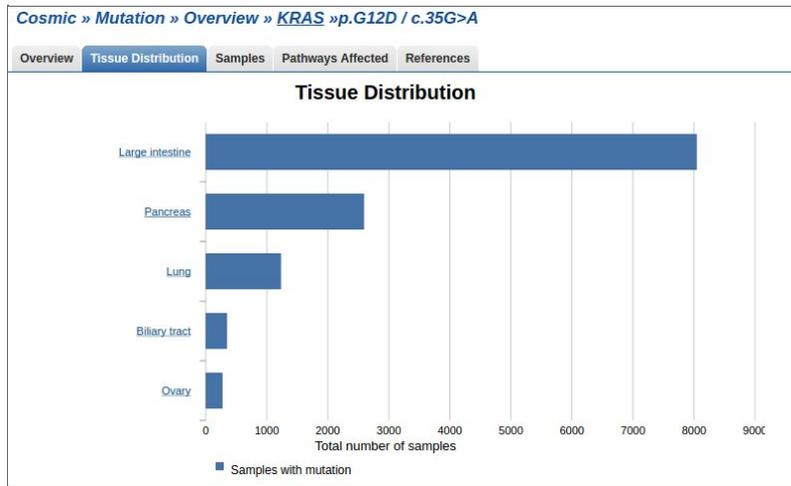
Below the summary table, there is a table with columns: Gene, Alt ids, Tested samples, Simple Mutations, Fusions, and Coding Mutations. The first two rows are highlighted in red:

Gene	Alt ids	Tested samples	Simple Mutations	Fusions	Coding Mutations
KRAS	KRAS_189621	37953	1	37953	
KRAS_ENST0000026078	KRAS_ENST0000026078	25784	2806	0	2806

- 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	Q9N094	VAR_052201	p.Val555Met	Polymorphism	rs9073	-
A1CF	Q9N094	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? and 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.

ClinVar

Home About Access Using the website Submission Statistics FTP site

**ClinVar**  
ClinVar aggregates information about genomic variation and its relationship to human health.

**Using ClinVar**  
[About ClinVar](#)  
[Data Dictionary](#)  
[Downloads/FTP site](#)  
[FAQ](#)  
[Contact Us](#)  
[RSS feed/What's new?](#)  
[Factsheet](#)

**Tools**  
[ACMG Recommendations for Reporting of Incidental Findings](#)  
[Variation Submission Portal](#)  
[Submissions](#)  
[Variation Viewer](#)  
[Clinical Remapping - Between assemblies and RefSeqGenes](#)  
[RefSeqGene/LRG](#)  
[Variation Reporter](#)

**Related Sites**  
[ClinGen](#)  
[GeneReviews®](#)  
[GTR®](#)  
[MedGen](#)  
[OMIM®](#)  
[Variation](#)

- 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. <a href="#">NM_033360.3(KRAS):c.38G&gt;A (p.Gly134Ser)</a> 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, 2019)	criteria provided, single submitter
54. <a href="#">NM_033360.3(KRAS):c.37G&gt;T (p.Gly133Cys)</a> 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. <a href="#">NM_033360.3(KRAS):c.37G&gt;C (p.Gly134Arg)</a> 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. <a href="#">NM_033360.3(KRAS):c.35G&gt;C (p.Gly128Ala)</a> GRCh37: Chr12:25398284 GRCh38: Chr12:25245350	KRAS	Non-small cell lung cancer		Pathogenic (Dec 7, 2007)	no assertion criteria provided
57. <a href="#">NM_004985.4(KRAS):c.35G&gt;T (p.Gly129Val)</a> 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. <a href="#">NM_004985.4(KRAS):c.35G&gt;A (p.Gly128Ser)</a> 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	GO:ESP:0.00002(1)	Pathogenic (Aug 30, 2016)	criteria provided, single submitter
59. <a href="#">NM_033360.3(KRAS):c.34G&gt;A (p.Gly129Ser)</a> GRCh37: Chr12:25398285 GRCh38: Chr12:25245351	KRAS	Juvenile myelomonocytic leukemia, Neoplasm of ovary, Non-small cell lung cancer, Neoplasm of stomach		Pathogenic (Sep 4, 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.

**OMIM®**  
**Online Mendelian Inheritance in Man®**  
 An Online Catalog of Human Genes and Genetic Disorders  
 Updated February 23, 2017

Search OMIM for clinical features, phenotypes, genes, and more...

Advanced Search : [OMIM](#), [Clinical Synopses](#), [Gene Map](#)

Need help? : [Example Searches](#), [OMIM Search Help](#), [OMIM Tutorial](#)

Mirror site : [mirror.omim.org](#)

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

**Gene Map Advanced Search**

12:25,245,350-25,245,350

Entries per page : 10

Search by genomic region (or cyto location range) to get a list of all OMIM Gene/Loci in that region, for example:  
 '1:0-124,300,000' or '1p36-p32'

To search within a single cyto location band, you would use:  
 '1p36-p36'

Search by genomic location (or cyto location band) to jump to that location in the chromosome, for example:

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

12:25,245,350-25,245,350

Search: OMIM | Advanced Search | Display Options | Phenotype Only Entries | All Entries

Search: 12:25,245,350-25,245,350  
 Results: 7 entries

Genomic context table	Location (genomic start cyto location)	Gene/Locus	Gene/Locus name	Gene/Locus MIM number	Phenotype	Phenotype MIM number	Inheritance (in progress)	Pheno map key	Comments	Mouse ortholog
1:	12p12p	KAR	Aromatic alpha-keto acid reductase	107500					None in MIM#1	
2:	12p12p	PKS	Pallister-Killian syndrome	601803	Pallister-Killian syndrome	601803	SMo	4		
3:	12:25,000,000-12p13.3p11.23	DPNBG2	Deafness, autosomal recessive 62	610163	Deafness, autosomal recessive 62	610163	AR	2	between D12S158 and D12S1042	
4:	12:19,000,000-12p13.3q24.1	IBD2	Inflammatory bowel disease 2	601458	Inflammatory bowel disease 2	601458		2	mainly ulcerative colitis	
5:	12:19,000,000-12p13.3q24.1	HY14	Hypertension, essential, susceptibility to, 4	608742	Hypertension, essential, susceptibility to, 4	145000	Ms	2		
6:	12:25,204,788-12p12.1	KRAS, KRAS2, RAS2, NS, CFC2, RALD	Kinase of sarcoma 2 viral (v-Ki-ras2) oncogene homolog	190070	Bladder cancer, somatic Breast cancer, somatic Cardiofacioscapular syndrome 2 Cervical cancer, somatic Leukemia, acute myeloid Lung cancer, somatic Nosean syndrome 3 Pancreatic carcinoma, somatic RAS-associated autoimmune leukoproliferative disorder Schwannoma/Paraganglioma-Mina syndrome, somatic, mosaic	109000 114400 611270 137215 606420 211900 609942 260900 614470 102200	S S S S AD S S S S	prostaglandin synthase 2 (gp12p11)	Kras	
7:	12p12.1 Chr12	HEXA1	Hexokinase 1	101120						

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.

Table	Description	Public entries	Total entries
Mutation data (as of 2015-04-05)		12768	17925
Gene symbol	The gene description, gene symbol (as recommended by the HUGO Nomenclature Committee) and chromosomal location is recorded for each gene. In cases where a gene symbol has not yet been made official, a provisional symbol has been adopted which is denoted by lower-case letters.	4605	7199
cDNA sequence	cDNA reference sequences are provided, numbered by codes.	4708	7425
Genomic coordinates	Genomic (chromosomal) coordinates have been calculated for missense/nonsense, splicing, regulatory, small deletions, small insertions and small indels.	0	12768
HGVS			

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

**Registration data (\*required)**

First name\*:

Last name\*:

Background\*:

Role/title\*:

Company/Organisation\*:

Department:

Address 1\*:

Address 2:

City\*:

Post/Zip code\*:

Country\*:

Telephone\*:

Fax:

Email\*:

[Privacy policy & disclaimer](#) Accept and

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

Gene symbol: KRAS

Description: The gene description, gene symbol (as recommended by the HU which is denoted by lower-case letters

- Select KRAS in the following table.

Gene symbol	
<a href="#">KRAS</a>	V-ki-ras2 kirsten rat sarcoma viral oncogene homologue

- 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 <small>(More available in public)</small>	12p12.1	V-Ki-ras2 Kirsten rat sarcoma viral oncogene homolog <small>(More available in public)</small>	NM_004855.4	Not available	BIOMASE Feature available in subscribers
Mutation type					
		Number of mutations	Mutation data by type (collapse 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			BIOMASE Feature available in subscribers		
Public total (HGMD Professional 2015.4 total)		27 (33)			
Disease/phenotype					
		Number of mutations	Mutation data by disease/phenotype		
Notoan syndrome		14	BIOMASE Feature available in subscribers		
Carlo-facio-cutaneous syndrome		6	BIOMASE Feature available in subscribers		
Costello syndrome		2	BIOMASE Feature available in subscribers		
Carlo-facio-cutaneous syndrome ?		1	BIOMASE Feature available in subscribers		
Gallbladder carcinoma, increased risk, assoc with		1	BIOMASE Feature available in subscribers		
Lung cancer, risk, association with		1	BIOMASE Feature available in subscribers		
Multiple mole melanoma syndrome		1	BIOMASE Feature available in subscribers		
Myelodysplastic/myeloproliferative disease ?		1	BIOMASE Feature available in 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
28 mutations in HGMD professional 2015.4	No mutations	2 mutations in HGMD professional 2015.4	3 mutations in HGMD professional 2015.4	No mutations	No mutations	3 mutations in HGMD professional 2015.4	No mutations	No mutations	No mutations
Further options available in BIOMASE professional 2015.4									
Accession Number	Codon change	Amino acid change	Codon number	Genes (condition & HGVS nomenclature)	Phenotype	Reference	Comments		
CM5070963	RRH>RRT	Lys>Phe	5	BIOMASE Feature available in subscribers	Costello syndrome	Zentgraf (2007) J Med Genet 44, 131 Facioscapular humeral dysplasia report available in subscribers Additional report available in subscribers			
CM5073168	RRH>GRH	Lys>Glu	5	BIOMASE Feature available in subscribers	Costello syndrome	Beretta (2007) J Hum Genet 52, 324 Additional phenotype report available in subscribers Additional report available in subscribers Additional phenotype report available in subscribers			
CM5076251	GCT>RGT	Gly>Ser	12	BIOMASE Feature available in subscribers	Carlo-facio-cutaneous syndrome	Noya (2007) J Med Genet 44, 283 Additional report available in subscribers			
CM507372	GCT>GRT	Gly>Arg	12	BIOMASE Feature available in subscribers	Multiple mole melanoma syndrome	Koppelman (2008) Am J Surg Pathol 32, 1905 Additional report available in subscribers			
CM125166	GCC>GRC	Gly>Arg	13	BIOMASE Feature available in subscribers	Myelodysplastic/myeloproliferative disease ?	Bonaldi (2012) Br J Haematol 138, 320 Additional report available in subscribers Additional phenotype report available in subscribers		Neutrophilic leuko...	
CM501082	GTR>RTR	Val>Leu	14	BIOMASE Feature available in subscribers	Notoan syndrome	Schubert (2006) Nat Genet 38, 331 Facioscapular humeral dysplasia report available in subscribers Additional report available in subscribers			
CM5070966	GRR>GGG	Gly>Arg	22	BIOMASE Feature available in subscribers	Notoan syndrome	Zentgraf (2007) J Med Genet 44, 131 Facioscapular humeral dysplasia report available in subscribers Additional report available in subscribers			
CM5070964	GRR>GRG	Gly>Glu	22	BIOMASE Feature available in subscribers	Carlo-facio-cutaneous syndrome	Zentgraf (2007) J Med Genet 44, 131 Facioscapular humeral dysplasia report available in subscribers Additional report available in 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:

- 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](http://ws.bioinfo.cipf.es/cellbase/rest/latest/hsa/feature/snp/rs158691/population_frequency)

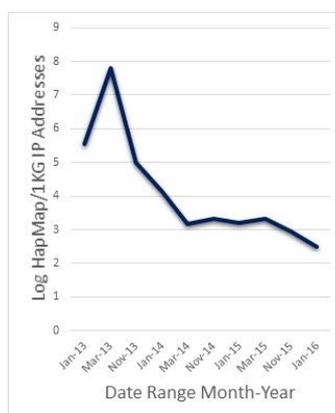
C) We have obtained a SNP of interest (rs28937313, location GRCh37 9:107584801) in our analysis and we want to know if it has been related with any disease.

- 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).

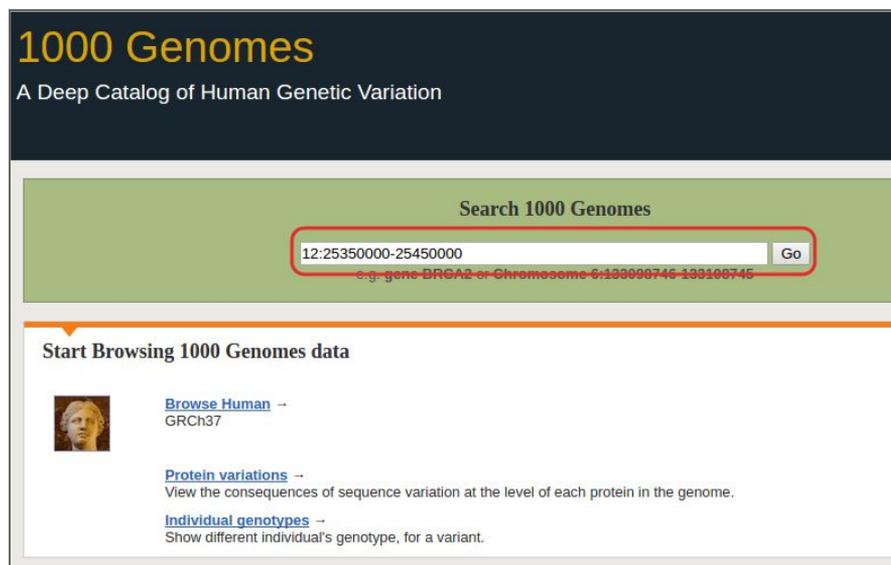


- B) Now, go to the 1,000 Genomes browser and search for the KRAS genomic region (example: 12:25350000-25450000). Can you find the global MAFs of the SNPs in this region from the 1,000 Genome populations?

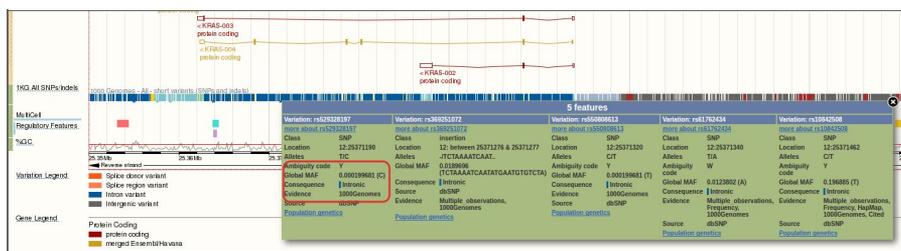
Note: this project is already finished but there are several available browsers.

More information: <http://www.internationalgenome.org/1000-genomes-browsers>.

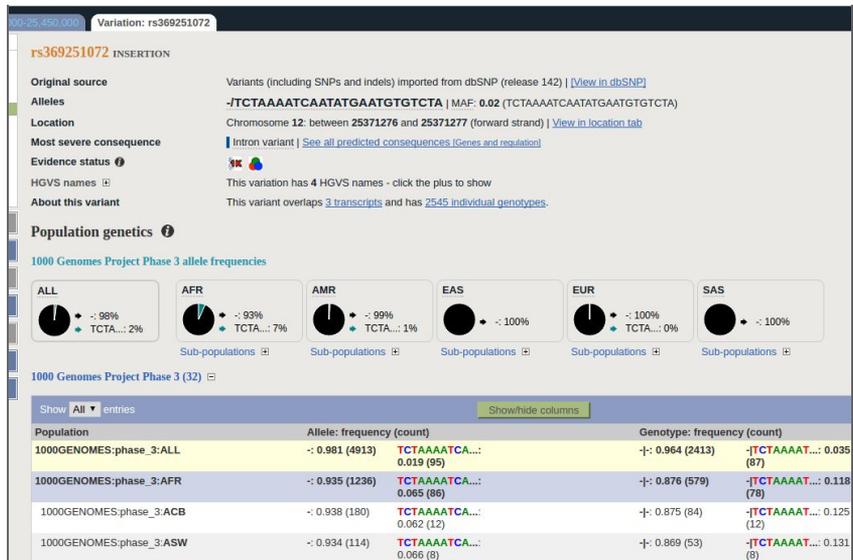
- Type the 1,000 Genomes URL in your browser (<http://phase3browser.1000genomes.org/index.html>) and search for KRAS region.



- 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.

Home **Data Browser** Data Usage and Release How to Use What's New Contact and FAQ Downloads

Target:   →

*examples of valid input for targets (one target per query):*

Gene HUGO: ACTB

Gene ID: 60

Chr. Region: 1:1000000-1100000

Single Chr. Location: 7:5567417

rsID: rs71531321

- 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.

**Variant Results** Coverage Results

Chromosome 12: 25350000 - 25450000

Genes in this region: KRAS(-) LYRM5(+)

**Select Data Set(s)**  
Check at least one data set below.

Select	Number Variations	Population
<input checked="" type="checkbox"/>	33	EuropeanAmerican
<input checked="" type="checkbox"/>	28	AfricanAmerican

**Display Results**

→

Chromosome 12:25350000-25450000

Genes in this region: KRAS(5) LYRMS(1)

Population: AfricanAmerican

GIWAS Catalog: KRAS LYRMS

KEGG Pathway: KRAS

Sanger COSMIC: KRAS LYRMS

PPV STRING 9.0: KRAS LYRMS

OMIM: KRAS LYRMS

Variation Color Code:

- splice or nonsense or frameshift
- missense
- coding-synonymous
- coding
- UTR
- codingComplex

Download Option:

File Format: Text

Zip Format: gzip

Download

Add or Remove Columns (Description of Columns)

rsID  Alleles  EA Allele Count  AA Allele Count  AA Allele Count  Allele Count  EA Genotype Count  AA Genotype Count

Genotype Count  MAF (%)  Sample Read Depth  Genes  Gene Accession #  GVS Function  cDNA Change

cDNA Size  Protein Change  Conservation (GERP)  Conservation (phastCons)  Grantham Score  PolyPhen Prediction  Clinical Link

NCBI 37 Allele  Chmp Allele  Illumina HumanExome Chip  GWAS Hits  EA Est. Age (yrs)  AA Est. Age (yrs)  GRCh38 Position

Sort Variants by: Allele Count

Select Population: AfricanAmerican

Select Transcript: Union of Transcripts

If "Select Transcript" above is set to "Union of Transcripts", and if multiple transcripts of a gene are involved in a variant and the function annotations for the variant are the same, only one representative transcript listed in the downloaded file if one chooses to download the data.

Show 10 entries

Variant GRCh37 Pos	rs ID	Alleles	EA Allele #	AA Allele #	All Allele #	EA Genotype #	AA Genotype #	All Genotype #	Avg. Sample Read Depth	Genes	mRNA Accession #
12:25362805	rs372792780	C>T	T=0C=8578	T=1C=4401	T=1C=12979	TT=0TC=0CC=4289	TT=0TC=1CC=2200	TT=0TC=1CC=6489	62	KRAS	NM_004865.3
12:25368413	rs174961132	G>A	A=0G=8596	A=1G=4403	A=1G=12999	AA=0AG=0GG=4298	AA=0AG=1GG=2201	AA=0AG=1GG=6499	137	KRAS	NM_033360.2
12:25379575	rs189989124	A>T	T=0A=8600	T=1A=4403	T=1A=13003	TT=0TA=0AA=4300	TT=0TA=1AA=2201	TT=0TA=1AA=6501	196	KRAS	NM_033360.2
12:25389226	rs377364475	T>C	C=0T=8600	C=1T=4405	C=1T=13005	CC=0CT=0TT=4300	CC=0CT=1TT=2202	CC=0CT=1TT=6502	66	KRAS	NM_033360.2
12:25392805	rs372792780	C>T	T=0C=8578	T=1C=4401	T=1C=12979	TT=0TC=0CC=4289	TT=0TC=1CC=2200	TT=0TC=1CC=6489	62	KRAS	NM_033360.2
12:25398227	rs170230528	G>A	A=0G=8600	A=1G=4403	A=1G=13003	AA=0AG=0GG=4300	AA=0AG=1GG=2201	AA=0AG=1GG=6501	57	KRAS	NM_033360.2
12:25392801	rs373149272	T>C	C=0T=8560	C=1T=4391	C=1T=12951	CC=0CT=0TT=4280	CC=0CT=1TT=2195	CC=0CT=1TT=6475	35	KRAS	NM_033360.2
12:25362805	rs136348382	C>T	T=0C=8558	T=1C=4389	T=1C=12947	TT=0TC=0CC=4279	TT=0TC=1CC=2194	TT=0TC=1CC=6473	33	KRAS	NM_033360.2
12:25388413	rs174961132	G>A	A=0G=8596	A=1G=4403	A=1G=12999	AA=0AG=0GG=4298	AA=0AG=1GG=2201	AA=0AG=1GG=6499	137	KRAS	NM_004865.3
12:25379510	rs174961132	G>T	T=0G=8594	T=1G=4403	T=1G=12997	TT=0TG=0GG=4297	TT=0TG=1GG=2201	TT=0TG=1GG=6496	120	KRAS	NM_033360.2

Showing 1 to 10 of 32 entries

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.

ExAC Browser (Beta) | Exome Aggregation Consortium

12:25350000-25450000

Examples: Gene: PCSK9, Transcript: ENST00000407236, Variant: 22-46615880-T-C, Multi-allelic variant: rs1800234, Region: 22:46615715-46615880

About ExAC

The Exome Aggregation Consortium (ExAC) is a coalition of investigators seeking to aggregate and harmonize exome sequencing data from a wide variety of large-scale sequencing projects, and to make summary data available for the wider scientific community.

Recent News

March 14, 2016

- Version 0.3.1 ExAC data and browser (beta) is released! ([Release notes](#))

- 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".

Region: 12 / 25350000 / 25450000

Coverage summary

Coverage metric: Average Individuals over X

Metric: mean

Genes

- ENSG00000205707
- ENSG00000298076
- ENSG00000133703

Variants

All Missense + LoF LoF Include filtered (non-PASS) variants Invert (highlight rare variants)

Export table to CSV

Variant	Chrom	Position	Consequence	Filter	Annotation	Flags	Allele Count	Allele Number	Number of Homozygotes	Allele Frequency
12:25356849 T/C	12	25356849	c-37-5T-C	PASS	splice region		1	12162	0	0.0000622
12:25356873 G/A (rs147480774)	12	25356873	PASS	5' UTR			3	12164	0	0.0002407
12:25356905 T/C	12	25356905	p.Asn54n	PASS	synonymous		1	14512	0	0.0000891

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

12:25398284 A / G	12	25398284	p.Leu19Leu	PASS	synonymous	1	102040	0	0.000009800
12:25398288 A / G	12	25398288	p.Ser175Ser	PASS	synonymous	1	102060	0	0.000009798
12:25398279 C / T (rs104994365)	12	25398279	p.Val124Ile	PASS	missense	1	101898	0	0.000009814
12:25398284 C / T (rs121913529)	12	25398284	p.Gly124Asp	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.Val9Val	PASS	synonymous	36	98618	0	0.00038650
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).

### Variant: 12:25398284 C / T

**Filter Status** PASS

**dbSNP** rs121913529

**Allele Frequency** 1.976e-05

**Allele Count** 2 / 101204

**UCSC** [12-25398284-C-T](#)

**ClinVar** [Click to search for variant in ClinVar](#)

Genotype Quality Metrics

Site Quality Metrics

---

#### Annotations

This variant falls on 4 transcripts in 1 genes:

missense

• KRAS Transcripts

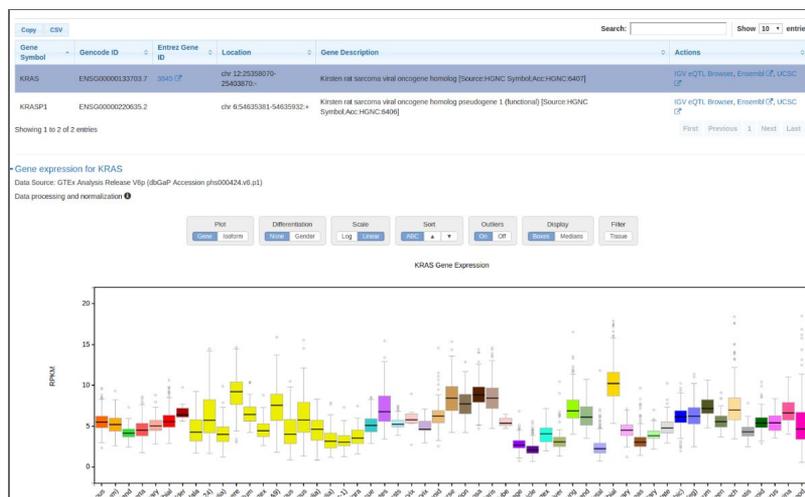
Note: This list may not include additional transcripts in the same gene that the variant does not overlap.

#### Population Frequencies

Population	Allele Count	Allele Number	Number of Homozygotes	Allele Frequency
African	1	8994	0	0.0001112
European (Non-Finnish)	1	53382	0	1.873e-05
East Asian	0	7960	0	0
European (Finnish)	0	5844	0	0
Latino	0	10162	0	0
Other	0	772	0	0
South Asian	0	14090	0	0
<b>Total</b>	<b>2</b>	<b>101204</b>	<b>0</b>	<b>1.976e-05</b>

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?

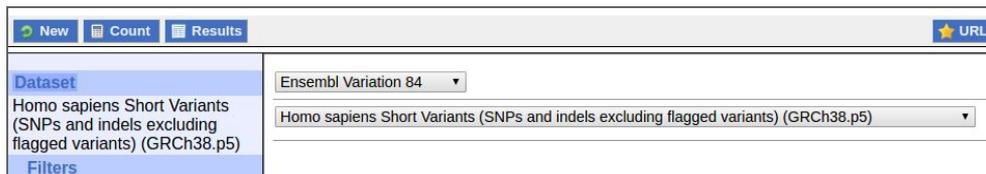
- Type the GTEx portal URL on your browser (<http://www.gtexportal.org>) and search the KRAS gene. In the graph with the boxplots you can find that “Nerve-tibial” is the one with the greatest expression and “Heart-left ventricle” together with “Muscle-skeletal”.



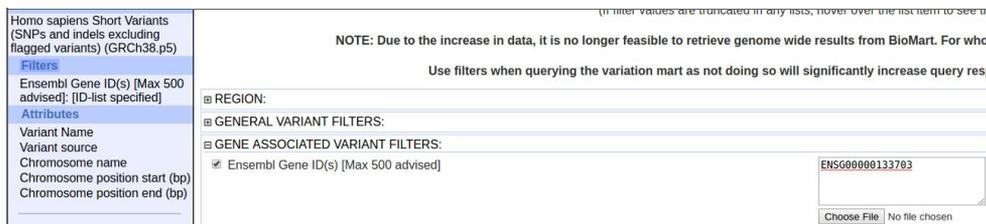
**Exercise 6.** Retrieve genomic variation data using Ensembl Biomart (Ensembl Variation database, <http://www.ensembl.org/biomart>).

Questions:

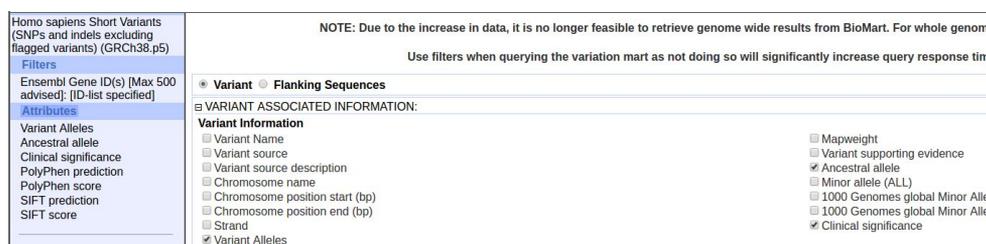
- A) Retrieve the variant alleles, the ancestral allele, the clinical significance, the SIFT and PolyPhen information about all the variants of the KRAS gene (ENSG00000133703).
- Type the BioMart URL on your browser (<http://www.ensembl.org/biomart>) and choose “Ensembl Variation Database” and “Homo 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.

**Dataset 3007 / 153789085 SNPs**  
 Human Short Variants (SNPs and indels excluding flagged variants) (GRCh38.p7)

**Filters**  
 Gene stable ID(s) [Max 500 advised]: [ID-list specified]

**Attributes**  
 Variant Name  
 Variant source  
 Chromosome name

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

Export all results to  TSV  Unique results only

Email notification to

View  rows as   Unique results only

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”.

Clinical significance

benign,likely benign  
 uncertain significance,benign,likely benign  
 not provided,benign,likely benign  
 uncertain significance,not provided,benign,likely benign  
 likely pathogenic  
 uncertain significance,likely pathogenic  
 not provided,likely pathogenic  
 uncertain significance,not provided,likely pathogenic  
 benign,likely pathogenic  
 likely benign,likely pathogenic  
 uncertain significance,likely benign,likely pathogenic  
 not provided,likely benign,likely pathogenic  
 benign,likely benign,likely pathogenic  
 pathogenic  
 uncertain significance,pathogenic  
 not provided,pathogenic

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

**Dataset 27 / 150331637 SNPs**  
 Homo sapiens Short Variants (SNPs and indels excluding flagged variants) (GRCh38.p5)

**Filters**  
 Ensembl Gene ID(s) [Max 500 advised]: [ID-list specified]  
 Clinical significance: likely pathogenic , pathogenic , likely pathogenic,pathogenic

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.

<b>Dataset</b> Homo sapiens Short Variants (SNPs and indels excluding flagged variants) (GRCh38.p5)	<input type="button" value="Choose File"/> No file chosen
<b>Filters</b> Ensembl Gene ID(s) [Max 500 advised]; [ID-list specified] Variant source: HGMD-PUBLIC	<b>GENERAL VARIANT FILTERS:</b>
<b>Attributes</b> Variant Name Variant source Chromosome name Chromosome position start (bp) Chromosome position end (bp)	<input checked="" type="checkbox"/> Variant source <input type="checkbox"/> Filter by Variant Name (e.g. rs123, CM000001) [Max 500 advised]
	<input type="button" value="Choose File"/> No file chosen