

## 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:

SNP rs158691

Save search Advanced

Display Settings: ☒ Summary, Sorted by SNP\_ID

Results: 2

1. rs158691 [Homo sapiens]

Sequence: `tgagggtggcaattcaaaactgttgg[C/T]taggtgtataggagagtcacaat`

Chromosome: 19:23017251

Gene: LOC101929164 (GeneView)

Functional Consequence: **intron variant**

Validated: **by 1000G by 2ht 2allele by cluster by frequency by hapmap**

Global MAF: C=0.3135/1570

HGVs: NC\_000019.10:g.23017251T>C, NC\_000019.9:g.23200053T>C, NR\_110746.1:n.544+467A>G, XR\_244111.1:n.544+467A>G

2. rs60992747 has merged into rs158691 [Homo sapiens]

Sequence: `tgagggtggcaattcaaaactgttgg[C/T]taggtgtataggagagtcacaat`

Chromosome: 19:23017251

Gene: LOC101929164 (GeneView)

Functional Consequence: **intron variant**

Validated: **by 1000G by 2ht 2allele by cluster by frequency by hapmap**

Global MAF: C=0.3135/1570

HGVs: NC\_000019.10:g.23017251T>C, NC\_000019.9:g.23200053T>C, NR\_110746.1:n.544+467A>G, XR\_244111.1:n.544+467A>G

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

NCBI dbSNP Short Genetic Variations

dbVar ClinVar GGP PubMed Nucleotide Protein

Search small variations in dbSNP or large structural variations in dbVar

Search Entry: dbSNP for

Go

ANNOUNCEMENT

dbSNP dbVar

List of 53 organisms with variant annotation on their genomes available for web search and FTP download.

Search by IDs on All Assemblies

Note: **ref** and **ss** must be prefixed with "rs" or "ss", respectively (i.e. rs25, ss1).

ID: rs158691 Reference cluster ID(rs)

Search Reset

Submission Information

- By Submitter
- New Submitted Batches
- Method
- Population
- Publication

Batch

- Enter List
  - NCBI Assay ID(ss)
  - Reference SNP ID(rs)
  - Local SNP ID
- Upload List
  - NCBI Assay ID(ss)
  - Reference SNP ID(rs)
  - Local SNP ID

Batch Query Help

- 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),

Reference SNP (refSNP) Cluster Report: rs158691

RefSNP	Allele	HGVs Names
Organism: human (Homo sapiens)	Variation Class: SNV: single nucleotide variation	NC_000019.10:g.23017251T>C
Molecule Type: Genomic	RefSNP Alleles: C/T (FWD)	NC_000019.9:g.23200053T>C
Created/Updated in build: 79/149	Allele Origin:	NR_110746.1:n.544+467A>G
Map to Genome Build: 108/Weight 1	Ancestral Allele: C	XR_244111.1:n.544+467A>G
Validation Status: <b>Validated by 1000G by 2ht 2allele by cluster by frequency by hapmap</b>	Variation Viewer: <b>VarView</b>	
	Clinical Significance: NA	
	MAF/MinorAlleleCount: C=0.3135/1570 (1000 Genomes)	


SNP Details are organized in the following sections:

GeneView Map Submission FaSta Resource Diversity **Validation**

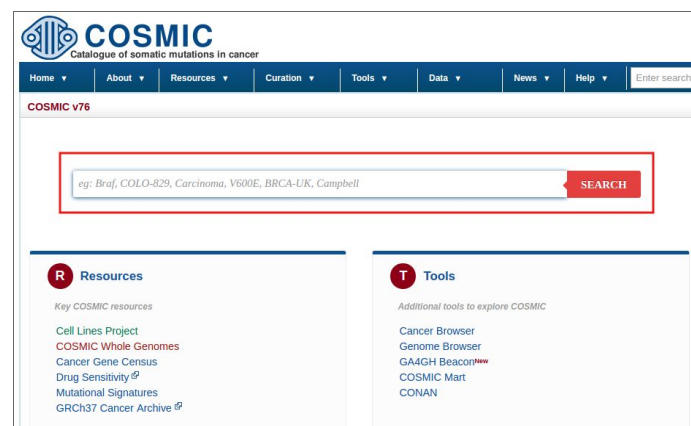
Integrated Maps (Hint: click on 'Chr Pos' to see variant in the new NCBI variation viewer)

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	23200053	NT_011295.11	14462855	Fwd	T	Fwd	view	blast

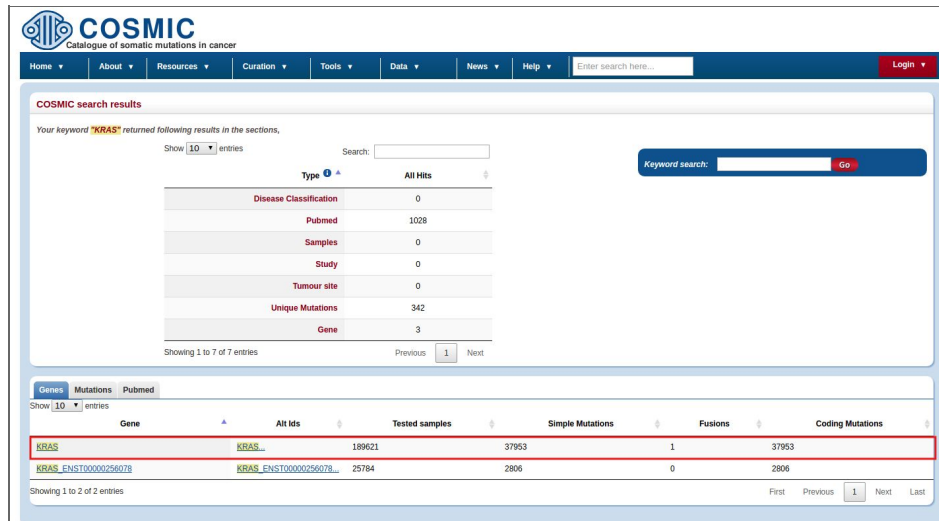
- When looking at the Validation status in the dbSNP report, the field includes some information which is slightly extended in the **Validation section** of the report,

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

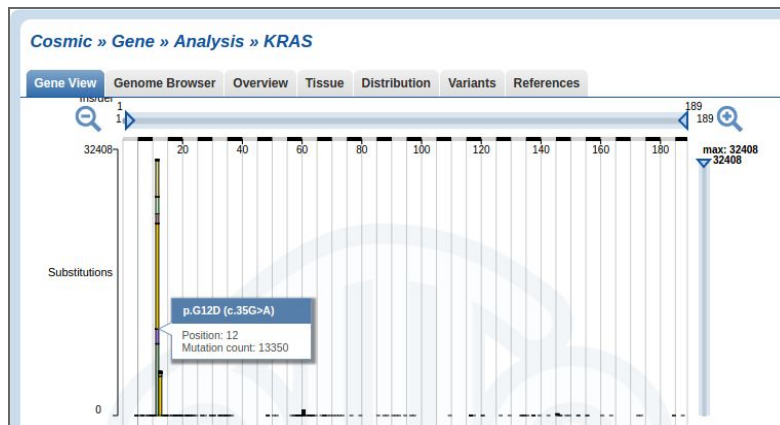
- In some cases the information in both sections could differ or be missing. Then, how can we consider this information reliable? One option is to search for the rs158691 in other databases such as Ensembl or check its population frequency in different human variation catalogs.
  - More information about 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.



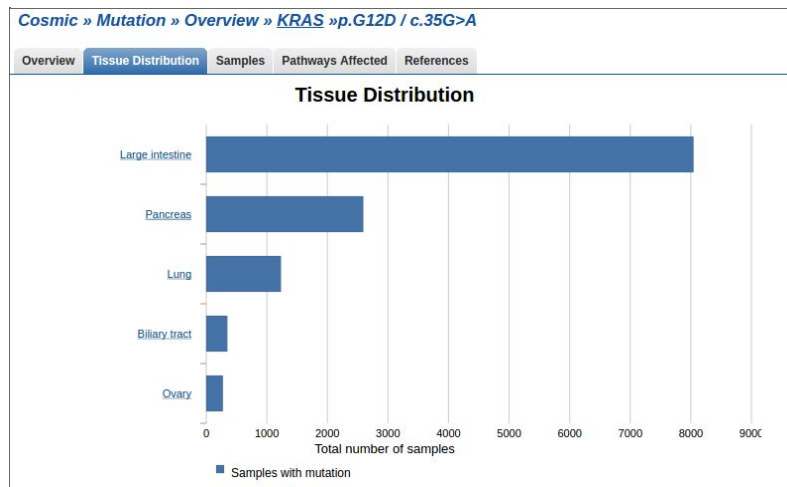
- Select the first Gene ID (“KRAS”) in the results page.



- 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? 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

NP\_203524.1

Advanced

Home About Access Using the website Submission Statistics FTP site

ACTGATGGTATGGGGCCAAGAGATATATCT  
CAGGTACGGCTGTCATCACTTAGACCTCAC  
CAGGGCTGGGCATAAAAGTCAGGGCAGAGC  
CCATGGTGCATCTGACTCCTGAGGAGAAGT  
GCAGGTTGGTATCAAGGTTACAAGACAGGT  
GGCACTGACTCTCTGCTATTGGTCTAT

**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.	NM_033360.3(KRAS):c.35G>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, <a href="#">see note</a>	GO:ESP:0.00002(1)	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 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...

Q

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

Q

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.

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 (gdb)
1:	12p	KAR	Aromatic alpha keto acid reductase	107520					Thane in MCH1	
2:	12p	PKS	Pallister-Killian syndrome	601803	Pallister-Killian syndrome	601803	SMo	4		
3:	12p11.000.000	DNBS62	Duchenne, autosomal recessive 62	610163	Duchenne, autosomal recessive 62	610163	AD	2	between D12S158 and D12S164	
4:	12p11.000.000	IBD2	Inflammatory bowel disease 2	601458	Inflammatory bowel disease 2	601458		2	mainly ulcerative colitis	
5:	12p11.000.000	HYT4	Hypertension, essential, susceptibility to, 4	608742	Hypertension, essential, susceptibility to, 4	145000	Mo	2		
6:	12p11.000.000	KRAS, KRAS2, RAS2, RAS, CFC2, RALD	Neurotrophic factor 2 (v-Ki-2) oncogene homolog	100070	Bladder cancer, somatic Breast cancer, somatic Cardiovascular disease 2 Gastric cancer, somatic Leukemia, acute myeloid Lung cancer, somatic Neurotrophic factor 2 Pancreatic cancer, somatic RAS-associated autoimmune leukoproliferative disorder Schwannoma-Paraneoplastic syndrome, somatic, mosaic	100070 114460 611270 137215 605620 211980 609942 200910 614470 145200	AD AD AD AD AD AD AD AD AD AD	prolonged KRAS/RAF1 at 12p11	Kras	
7:	12p	TRIM41	Trisomy, alpha 41	601130						

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.

7

**The Human Gene Mutation Database**  
at the Institute of Medical Genetics in Cardiff

Home Search Help Statistics Download What's New Background Publications Contact Us Login Logout Other Links

Gene symbol:  Go!

Symbol:  Missense/non-sense  Go!

The Human Gene Mutation Database (HGMD®) represents an attempt to collate known (published) gene lesions responsible for human inherited disease, and is maintained in Cardiff by D.N. Cooper, K.V. Ball, P.D. Stenson, A.D. Phillips, K. Howells, S. Heywood, M.J. Hayden, M.E. Mort and M.P. Hearn.

**Get HGMD Professional** Please note that this less up-to-date public version of our database is freely available only to [academic/institutional/non-profit](#) organisations. All commercial users are required to purchase a license from [QIAGEN®](#), our commercial partner. A license to [HGMD Professional](#) is available to both commercial and academic/non-profit users wishing to access the most up-to-date version of the database. HGMD mutation data are made available via the public site **3 years** after initial inclusion. Please read the [DISCLAIMER!](#)

For legal reasons, only users who give an email address that can be **CLEARLY** assigned to an academic or non-profit organization will be allowed to register successfully (ie - a university or hospital email address etc). Please **DO NOT** use your personal email address from hotmail, yahoo, gmail, MSN or any other commercial web-based system. You **WILL NOT** be allowed to access HGMD. Once registered, users will receive a password via the academic/non-profit email address supplied when registering. To use the HGMD login, users must have session cookies enabled, and will require their email address, country and password to log in.

Table:	Description:	Public entries:	Total entries:
Mutation totals (as of 2015-04-05)		12764	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.	4805	7399
cDNA sequence	cDNA reference sequences are provided, numbered by order.	4708	7425
Genomic coordinates	Genomic (chromosomal) coordinates have been calculated for missense, insertion, splicing, regulatory, small deletions, small insertions and small indels.	0	15750

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

**HGMD User Registration**

Please read the following before registering

Please note that the public version of our database is free only for registered users from academic institutions/non-profit organisations. Commercial users are required to purchase a license from [QIAGEN®](#), our commercial partner. The HGMD Professional license is available to both commercial and academic/non-profit users wishing to access the most up-to-date version of the database. HGMD mutation data are made available via the public site **3 years** after initial inclusion. Please read the [DISCLAIMER!](#)

For legal reasons, only users who give an email address that can be **CLEARLY** assigned to an academic or non-profit organization will be allowed to register successfully (ie - a university or hospital email address etc). Please **DO NOT** use your personal email address from hotmail, yahoo, gmail, MSN or any other commercial web-based system. You **WILL NOT** be allowed to access HGMD. Once registered, users will receive a password via the academic/non-profit email address supplied when registering. To use the HGMD login, users must have session cookies enabled, and will require their email address, country and password to log in.

**Registration data (\*required)**

First name*:	<input type="text"/>
Last name*:	<input type="text"/>
Background*:	Select background ▼
Role/title*:	Select role/title ▼
Company/Organisation*:	<input type="text"/>
Department:	<input type="text"/>
Address1*:	<input type="text"/>
Address2:	<input type="text"/>
City*:	<input type="text"/>
Post/Zip code*:	<input type="text"/>
Country*:	Select country ▼
Telephone*:	<input type="text"/>
Fax:	<input type="text"/>
Email*:	<input type="text"/>

[Privacy policy & disclaimer](#) [Accept and register](#)

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

**HGMD®**

KRAS  Gene symbol ▼ Go!

The Human Gene Mutation Database (HGMD®) represents an attempt to collate known (published) gene lesions responsible for human inherited disease, and is maintained in Cardiff by D.N. Cooper, K.V. Ball, P.D. Stenson, A.D. Phillips, K. Howells, S. Heywood, M.J. Hayden, M.E. Mort and M.P. Hearn.

**Get HGMD Professional** \*Please note that this less up-to-date public version of our database is freely available only to [academic/institutional/non-profit](#) organisations. All commercial users are required to purchase a license from [QIAGEN®](#), our commercial partner. A license to [HGMD Professional](#) is available to both commercial and academic/non-profit users wishing to access the most up-to-date version of the database. HGMD mutation data are made available via the public site **3 years** after initial inclusion. Please read the [DISCLAIMER!](#)

For legal reasons, only users who give an email address that can be **CLEARLY** assigned to an academic or non-profit organization will be allowed to register successfully (ie - a university or hospital email address etc). Please **DO NOT** use your personal email address from hotmail, yahoo, gmail, MSN or any other commercial web-based system. You **WILL NOT** be allowed to access HGMD. Once registered, users will receive a password via the academic/non-profit email address supplied when registering. To use the HGMD login, users must have session cookies enabled, and will require their email address, country and password to log in.

Table:	Description:
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.

- 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 (Show available to subscribers)	12p12.1	V-ki-ras2 Kirsten rat sarcoma viral oncogene homolog (Show available to subscribers)	NM_004895.4	Not available	BIOMASE Feature available to subscribers
Mutation type		Number of mutations	Mutation data by type (single 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 to subscribers		
Public total (ref:391 Professional (2015-4 total))		27 (33)			
Disease/phenotype		Number of mutations	Mutation data by disease/phenotype		
Neonatal syndrome		14	BIOMASE Feature available to subscribers		
Cardio-facio-cutaneous syndrome		6	BIOMASE Feature available to subscribers		
Crohn's syndrome		2	BIOMASE Feature available to subscribers		
Cardio-facio-cutaneous syndrome ?		1	BIOMASE Feature available to subscribers		
Gallbladder carcinoma, increased risk, assoc with		1	BIOMASE Feature available to subscribers		
Lung cancer, risk, association with		1	BIOMASE Feature available to subscribers		
Multiple mole melanoma syndrome		1	BIOMASE Feature available to subscribers		
Myelodysplastic/myeloproliferative disease ?		1	BIOMASE 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 <small>28 mutations in NCBI (ref:391) (2015-4)</small>	Splicing <small>No mutations</small>	Regulatory <small>2 mutations in NCBI (ref:391) (2015-4)</small>	Small deletions <small>1 mutation in NCBI (ref:391) (2015-4)</small>	Small insertions <small>No mutations</small>	Small indels <small>No mutations</small>	Gross deletions <small>3 mutations in NCBI (ref:391) (2015-4)</small>	Gross insertions <small>No mutations</small>	Complex <small>No mutations</small>	Repeats <small>No mutations</small>
Further options available in BIOMASE Professional (2015-4)									
Accession Number	Codon change	Amino acid change	Codon number	Genes (mutations in 391) (2015-4)	Phenotype	Reference	Comments		
CM070963	RRH-HIT	Lys-His	5	BIOMASE Feature available to subscribers	Crohn's syndrome	Zentgraf (2007) J Med Genet 44, 131 Facioscapulohumeral report available in pubmed Additional report available in pubmed			
CM073168	RRH-GRR	Lys-Glu	5	BIOMASE Feature available to subscribers	Crohn's syndrome	Berthiaume (2007) J Hum Genet 50, 324 Additional phenotype report available in pubmed Additional phenotype report available in pubmed			
CM076251	GCT-HCT	Gly-Ser	12	BIOMASE Feature available to subscribers	Cardio-facio-cutaneous syndrome	Nova (2007) J Med Genet 44, 283 Additional report available in pubmed			
CM007372	GCT-HCT	Gly-Ser	12	BIOMASE Feature available to subscribers	Multiple mole melanoma syndrome	Komatsu (2008) Ann J Surg Pathol 35, 1905 Additional report available in pubmed			
CM125166	GCC-GCC	Gly-Gly	13	BIOMASE Feature available to subscribers	Myelodysplastic/myeloproliferative disease ?	Brown (2012) Br J Haematol 146, 520 Additional report available in pubmed Additional phenotype report available in pubmed			
CM061082	GTH-HTR	Val-His	14	BIOMASE Feature available to subscribers	Neonatal syndrome	Schubert (2006) Nat Genet 38, 331 Facioscapulohumeral report available in pubmed			
CM070966	GCG-GGG	Gly-Gly	22	BIOMASE Feature available to subscribers	Neonatal syndrome	Zentgraf (2007) J Med Genet 44, 131 Facioscapulohumeral report available in pubmed Additional report available in pubmed			
CM070964	GCG-GGG	Gly-Gly	22	BIOMASE Feature available to subscribers	Cardio-facio-cutaneous syndrome	Zentgraf (2007) J Med Genet 44, 131 Facioscapulohumeral report available in pubmed Additional report available in pubmed			

**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)

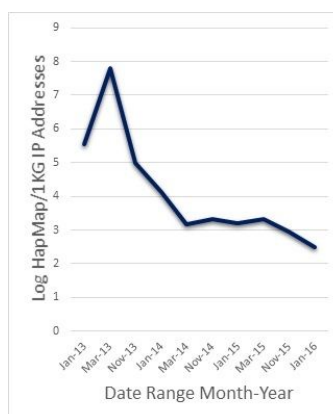
C) We have obtained a SNP of interest (rs28937313, location GRCh37 9:107584801) in our analysis and we want to know if it has been related with any disease.

- 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).

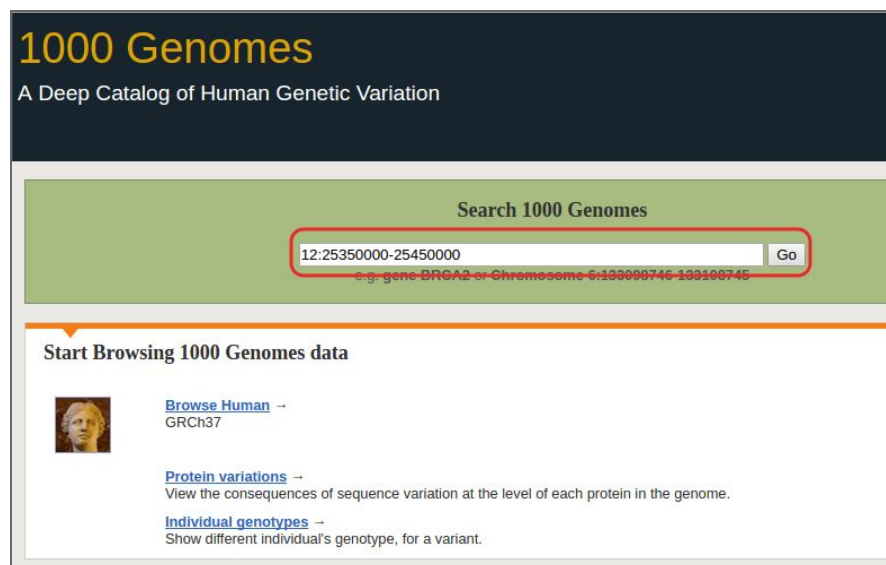


- B) Now, go to the 1,000 Genomes browser and search for the KRAS genomic region (example: 12:25350000-25450000). Can you find the global MAFs of the SNPs in this region from the 1,000 Genome populations?

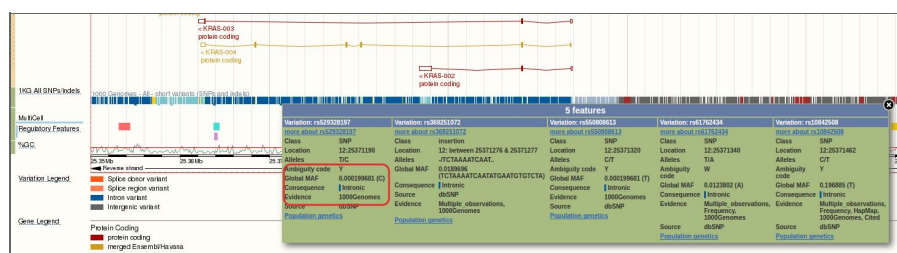
Note: this project is already finished but there are several available browsers.

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

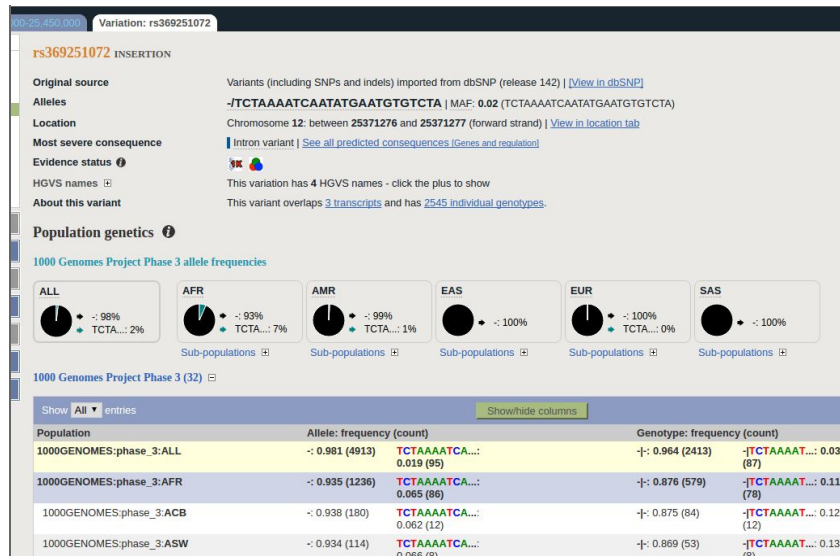
- Type the 1,000 Genomes URL in your browser (<http://phase3browser.1000genomes.org/index.html>) and search for KRAS region.



- 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** LYRMS(+)

Population: AfricanAmerican

GWAS Catalog: **KRAS** LYRMS

KEGG Pathway: **KRAS**

Sanger COSMIC: **KRAS** LYRMS

PPV STRING 9.0: **KRAS** LYRMS

Gene: **KRAS** LYRMS

**Variation Color Code:**

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

**Download Option:**

File Format: **Text**

Zip Format: **gzip**

**Download**

**Add or Remove Columns (Description of Columns)**

☒ dbSNP rs ID ☒ Alleles ☒ EA 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 ☒ Chimp Allele ☒ Illumina HumanExome Chip ☒ GWAS Hits ☒ EA Est. Age (yrs) ☒ EA 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 is shown 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	rs37220780	C>T	T>G C=6576	T=1JC=4401	T=1JC=12979	TT=0TC=0CC=4289	TT=0TC=1CC=2200	TT=0TC=1CC=6489	62	62	KRAS	NM_00485.3
12:25364113	rs37681135	G>A	A=0G=8596	A=1JC=4403	A=1JC=12999	AA=0AG=0GG=4298	AA=0AG=1GG=2201	AA=0AG=1GG=6499	137	137	KRAS	NM_033360.2
12:25378575	rs38968124	A>T	T=0A=9600	T=1JC=4403	T=1JC=13003	TT=0TA=0AA=4300	TT=0TA=1AA=2201	TT=0TA=1AA=6501	196	196	KRAS	NM_033360.2
12:25389526	rs377354475	T>C	C=0T=9600	C=1JT=4405	C=1JT=13005	CC=0CT=0TT=4300	CC=0CT=1TT=2202	CC=0CT=1TT=6502	66	66	KRAS	NM_033360.2
12:25389526	rs377354475	C>T	T=0C=6576	T=1JC=4401	T=1JC=12979	TT=0TC=0CC=4289	TT=0TC=1CC=2200	TT=0TC=1CC=6489	62	62	KRAS	NM_033360.2
12:25389527	rs37022626	G>A	A=0G=8800	A=1JC=4403	A=1JC=13003	AA=0AG=0GG=4300	AA=0AG=1GG=2201	AA=0AG=1GG=6501	57	57	KRAS	NM_033360.2
12:25389581	rs373149272	T>C	C=0T=6560	C=1JT=4391	C=1JT=12951	CC=0CT=0TT=4289	CC=0CT=1TT=2195	CC=0CT=1TT=6475	35	35	KRAS	NM_033360.2
12:25389581	rs373149272	C>T	T=0C=6558	T=1JC=4389	T=1JC=12947	TT=0TC=0CC=4279	TT=0TC=1CC=2194	TT=0TC=1CC=6473	33	33	KRAS	NM_033360.2
12:25389613	rs37681135	G>A	A=0G=8596	A=1JC=4403	A=1JC=12999	AA=0AG=0GG=4298	AA=0AG=1GG=2201	AA=0AG=1GG=6499	137	137	KRAS	NM_00485.3
12:25378510	rs37681135	G>T	T=0G=8594	T=1JC=4403	T=1JC=12997	TT=0TG=0GG=4297	TT=0TG=1GG=2201	TT=0TG=1GG=6496	120	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: **T2-46615880-T-C**, Multi-allelic variant: **rs1800234**, Region: **22:46615715-46615880**

**About ExAC**

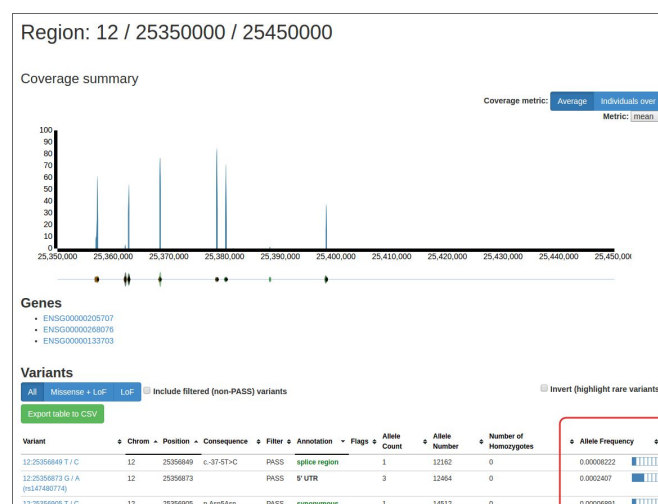
The **Exome Aggregation Consortium** (ExAC) is a coalition of investigators seeking to aggregate and harmonize exome sequencing data from a wide variety of large-scale sequencing projects, and to make summary data available for the wider scientific community.

**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".



- 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	0.00000800	
12-25398288 A / G	12	25398288	p.Ser175Ser	PASS	synonymous	1	102060	0	0.00000798	
12-25398279 C / T (rs104894365)	12	25398279	p.Val14Ile	PASS	missense	1	101898	0	0.00000814	
12-25398284 C / T (rs121913529)	12	25398284	p.Gly124Asp	PASS	missense	2	101204	0	0.00001976	
12-25398285 C / A (rs121013530)	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	0.00038650	
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

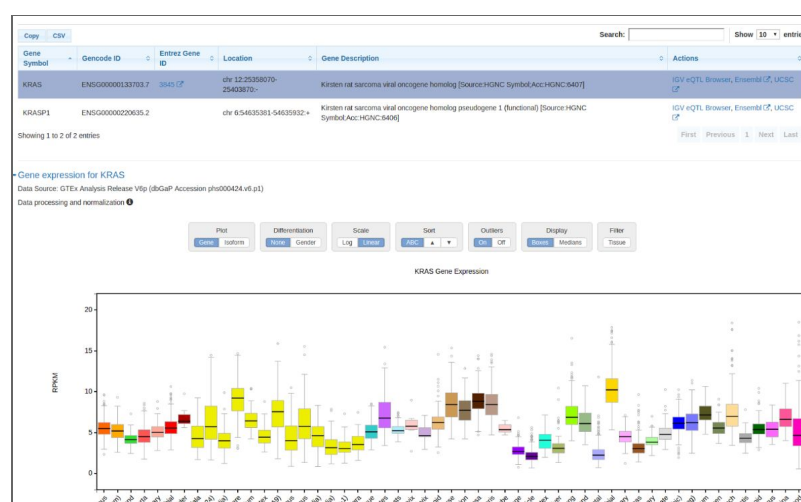
**Annotations**  
This variant falls on 4 transcripts in 1 genes:  
missense  
KRAS Transcripts

**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
Total	2	101204	0	1.976e-05

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

- 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)”.

The screenshot shows the Ensembl Biomart interface. The 'Dataset' dropdown is set to 'Ensembl Variation 84'. Below it, the 'Homo sapiens Short Variants (SNPs and indels excluding flagged variants) (GRCh38.p5)' is selected. The 'Filters' tab is active on the left.

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

The screenshot shows the 'Gene Associated Variant Filters' section. The 'Ensembl Gene ID(s) [Max 500 advised]' field is set to 'ENSG00000133703'. The 'Choose File' button is visible below the field.

- 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.

The screenshot shows the 'Variant associated information' and 'Gene Associated Information' sections. The 'Variant associated information' section is expanded, showing 'Variant Alleles', 'Ancestral allele', 'Clinical significance', 'PolyPhen prediction', 'PolyPhen score', 'SIFT prediction', and 'SIFT score'. The 'Gene Associated Information' section is also expanded, showing 'Mapweight', 'Variant supporting evidence', 'Ancestral allele', 'Minor allele (ALL)', '1000 Genomes global Minor Allele', '1000 Genomes global Minor Allele', and 'Clinical significance'.

- 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  File  ☒ 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”.

**Dataset**  
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

**Attributes**  
Variant Alleles  
Ancestral allele

☒ 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  
likely 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)	<div>Choose File No file chosen</div>	
<b>Filters</b> Ensembl Gene ID(s) [Max 500 advised]; [ID-list specified] Variant source: HGMD-PUBLIC	<div>GENERAL VARIANT FILTERS:</div> <div> <input checked="" type="checkbox"/> Variant source           <div>             ClinVar              dbSNP              ESP  <b>HGMD-PUBLIC</b>              HumanCoreExome-12           </div> </div> <div> <input type="checkbox"/> Filter by Variant Name (e.g. rs123, CM000001) [Max 500 advised]           <div> <div>Choose File No file chosen</div> </div> </div> <div> <input type="checkbox"/> Variant Source source           <div> <div>Archive dbSNP</div> </div> </div>	
<b>Attributes</b> Variant Name Variant source Chromosome name Chromosome position start (bp) Chromosome position end (bp)		