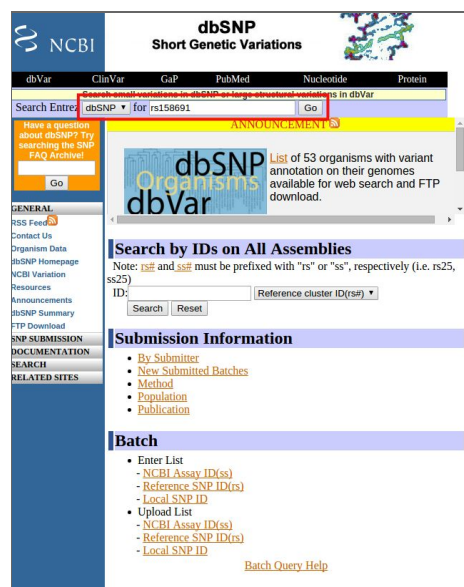


SOLVED BIOLOGICAL AND CLINICAL DATABASES EXERCISES. GDA2016

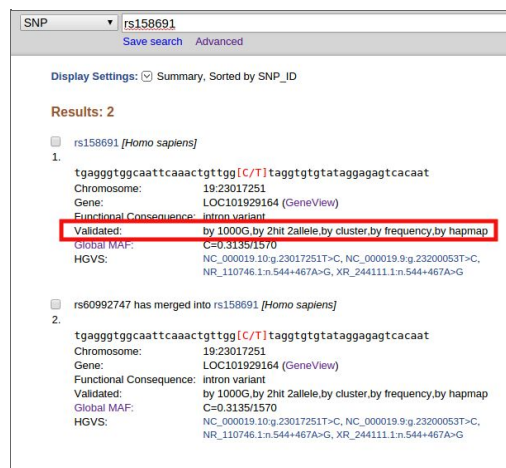
Exercise 1. Search information for specific SNVs in different databases.

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,



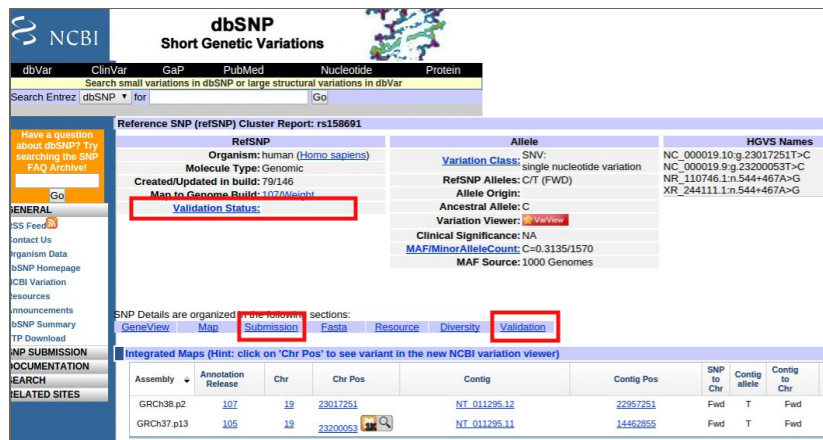
- The result of the first field search gives information about how rs158691 has been validated:



- 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 is empty and it differs from what we found when searching with the first option. Looking at other sections does not clarify the question. For example, in the **Validation section** of the report we find the following information,

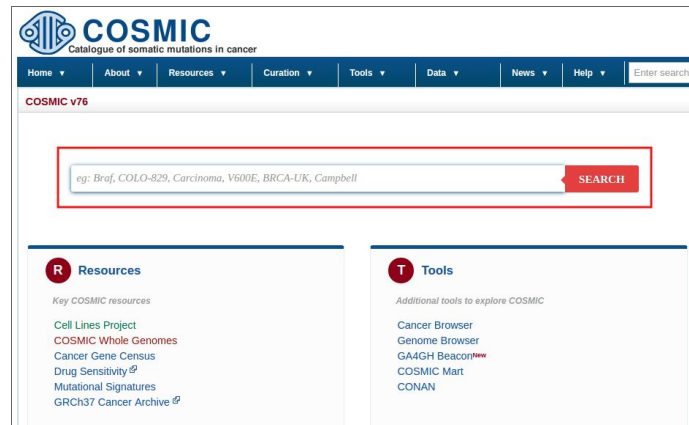
Validation Summary:			
Validation status	Marker displays Mendelian segregation	PCR results confirmed in multiple reactions	Homozygotes detected in individual genotype data
UNKNOWN	UNKNOWN	UNKNOWN	UNKNOWN

- So, 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:

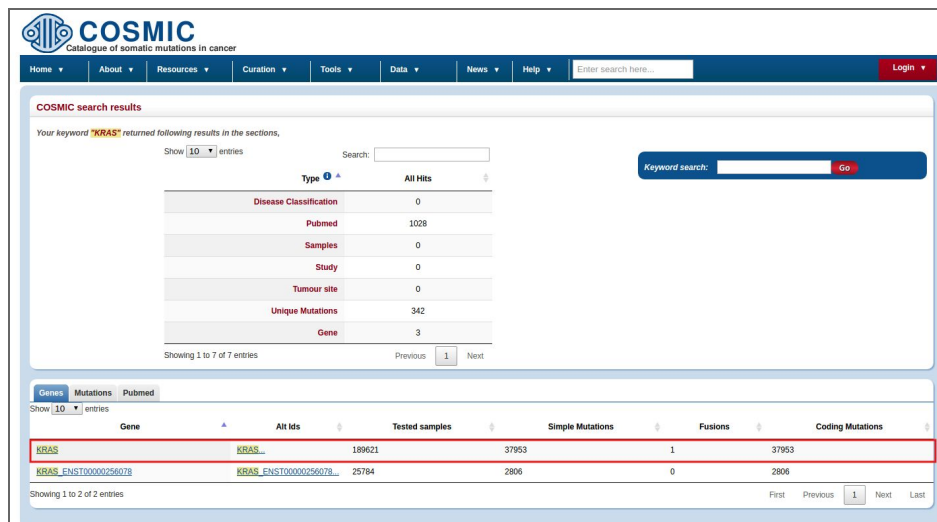
http://www.ensembl.org/info/genome/variation/data_description.html#evidence_status

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?

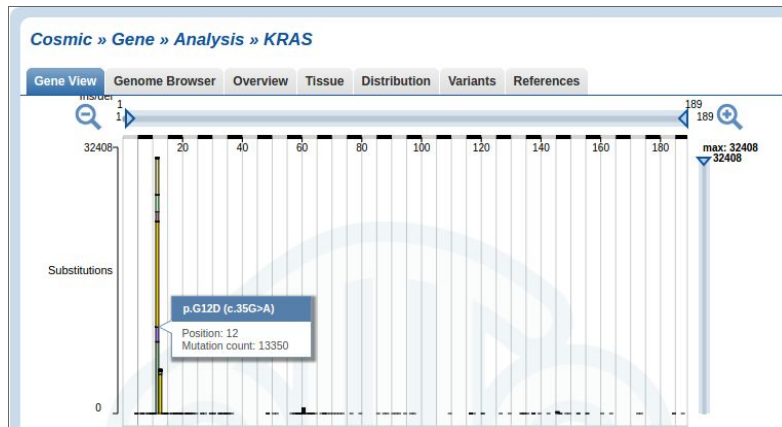
- Type the COSMIC URL on your browser (<http://cancer.sanger.ac.uk/cosmic>) and search for KRAS gene in the “Search” field.



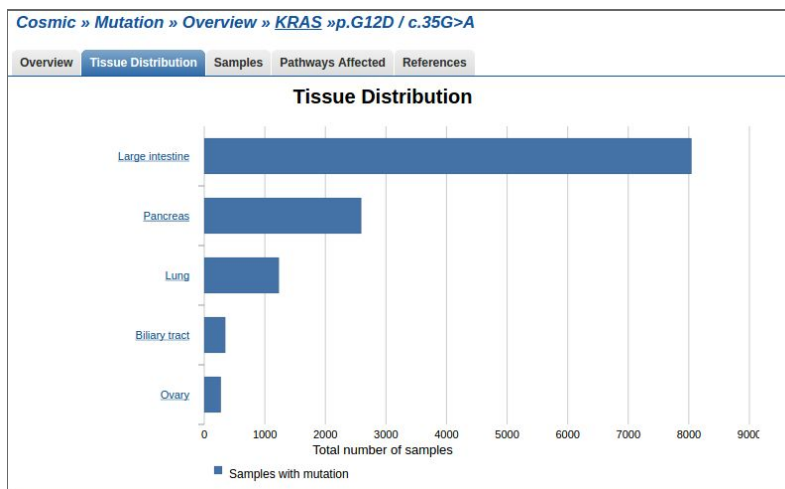
- Select the first Gene ID (“KRAS”) in the results page.



- 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 13350 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?

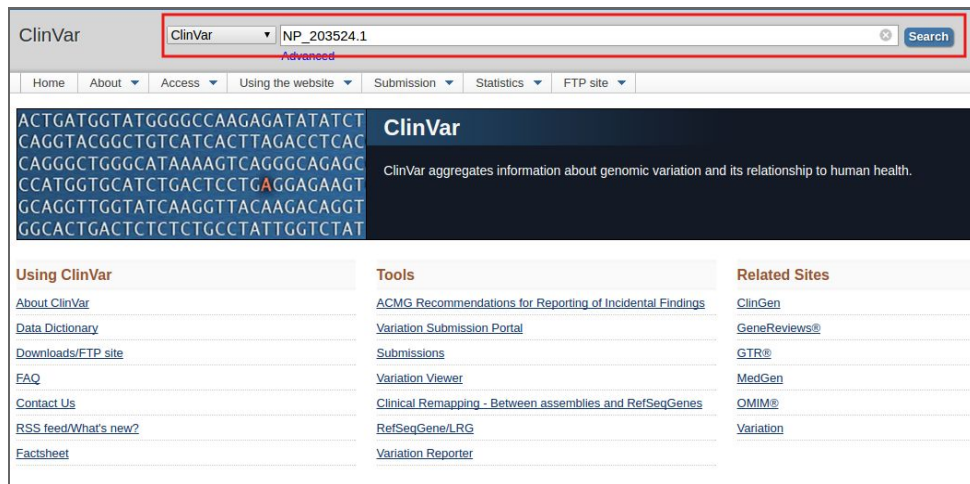
- Type the humsaVar URL on your browser (<http://www.uniprot.org/docs/humsavar>). The information of this database is contained in a text file that you can download from its web page. You can search for rs158691 either within the text file or directly in the web page using the search option of the browser.

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.

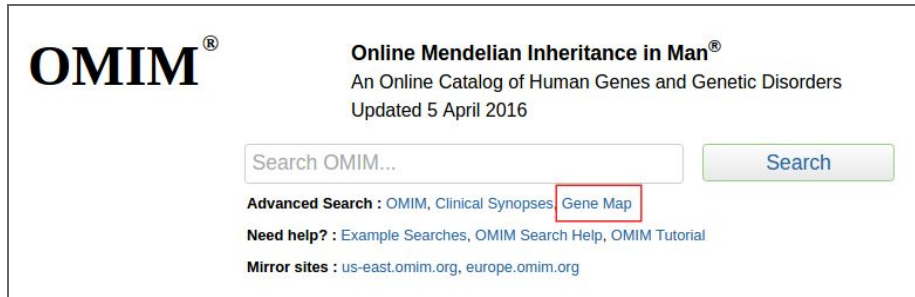


- The results page reports 62 items for NP_203524.1. In this page, you can search for c.35G>A, which is the CDS mutation ID from Exercise 1B. Then, you can find that its clinical significance states that is pathogenic and no assertion criteria is provided.

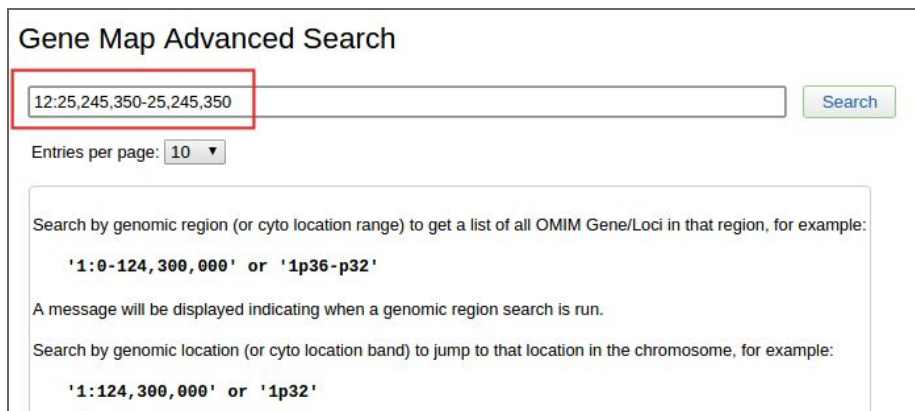
Variation Location	Gene(s)	Condition(s)	Frequency	Clinical significance (Last reviewed)	c.35G>A
49. NM_033360.3(KRAS):c.38G>A (p.Gly135Asp) GRCh37: Chr12:25398281 GRCh38: Chr12:25245347	KRAS	Juvenile myelomonocytic leukemia, Non-small cell lung cancer, Breast cancer, somatic, RAS-associated autoimmune leukoproliferative disorder, Breast adenocarcinoma	Pathogenic (Jul 1, 2015)	criteria provided, single submitter	
50. NM_033360.3(KRAS):c.37G>T (p.Gly133Cys) 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	
51. NM_033360.3(KRAS):c.37G>C (p.Gly133Arg) 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	
52. NM_033360.3(KRAS):c.35G>C (p.Gly126Ile) GRCh37: Chr12:25398284 GRCh38: Chr12:25245350	KRAS	Non-small cell lung cancer	Pathogenic (Dec 7, 2007)	no assertion criteria provided	
53. NM_004985.4(KRAS):c.35G>T (p.Gly126Val) GRCh37: Chr12:25398284 GRCh38: Chr12:25245350	KRAS	Juvenile myelomonocytic leukemia, Carcinoma of pancreas, Non-small cell lung cancer, Nevus sebaceous, NEVUS SEBACEOUS, SOMATIC, Rasopathy	Pathogenic (Mar 25, 2013)	criteria provided, single submitter	
54. NM_033360.3(KRAS):c.35G>A (p.Gly126Asp) 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	Pathogenic (Jun 10, 2012)	no assertion criteria provided	
55. NM_033360.3(KRAS):c.34G>A (p.Gly125Ser) GRCh37: Chr12:25398284 GRCh38: Chr12:25245350	KRAS	Juvenile myelomonocytic leukemia,	Pathogenic	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.

Genomic content table	Location (genomic start, cyt location)	Gene/Locus	Gene/Locus name	Gene/Locus MIM number	Phenotype	Phenotype MIM number	Inheritance	Pheno map key	Comments	Miscellaneous
1	12p12	KAR	Aromatic amino acid oxidase	107520					Name in MIM#1	
2	12p12	PKS	Pallister-Killian syndrome	601803	Pallister-Killian syndrome	601803	SN	4		
3	12:10,000,000-12p12-p11.23	DFNB2	Deafness, autosomal recessive 62	610143	Deafness, autosomal recessive 62	610143	AR	2	between D12S338 and D12S1042	
4	12:10,000,000-12p12-p11.23	IBD2	Inflammatory bowel disease 2	601456	Inflammatory bowel disease 2	601456		2	mainly ulcerative colitis	
5	12:10,000,000-12p12-p11.23	HYT4	Hypertension, essential, susceptibility to, 4	606742	Hypertension, essential, susceptibility to, 4	145500	M	2		
6	12:25,204,788-12p12.1	KRAS, KRAS2, KRAS2, NS, CFC1, KALD	Kirsten rat sarcoma-2 viral (v-Ki-6) oncogene homolog	108070	Bladder cancer, somatic Breast cancer, somatic Carcinoma, endometrial Cervical intraepithelial neoplasia 2 Genetic cancer, somatic Leukemia, acute myeloid Lung cancer, somatic Nasopharyngeal carcinoma Pancreatic carcinoma, somatic Rectal adenocarcinoma Schwannoma Sclerosing papillitis Tuberculosis, somatic Uterine leiomyosarcoma	108070 114480 81279 107215 601826 211000 605642 292250 614470 103200	3 3 3 3 AD 3 3 3 3 3	prostaglandin synthase 2 on 12p11	None	
7	12p12 Chr.12	TERTAC1	Tubulin, alpha-1A-1	291120						

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:
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.	12768	17925
cdna sequence	cdna reference sequences are provided, numbered by codes.	4605	7199
Genomic coordinates	Genomic (chromosomal) coordinates have been calculated for missense/nonsense, splicing, regulatory, small deletions, small insertions and small indels.	0	12768

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

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"/>
Address 1*:	<input type="text"/>
Address 2:	<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

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

Table:	Description:
Gene symbol	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	
KRAS	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 pubmed)</small>	12p12.1	V-Ki-ras2 Kirsten rat sarcoma viral oncogene homolog <small>(More available in pubmed)</small>	NM_004853.4	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					
Public total (HGMD Professional 2015.4 total)					
		27 (33)	BIOMASE Feature available in subscribers		
Disease/phenotype					
		Number of mutations	Mutation data by disease/phenotype		
Noonan syndrome		14	BIOMASE Feature available in subscribers		
Cardio-facio-cutaneous syndrome		6	BIOMASE Feature available in subscribers		
Costello syndrome		2	BIOMASE Feature available in subscribers		
Cardio-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
Feature system available in HGMD professional 2015.4									
Accession Number	Codon change	Amino acid change	Codon number	Genes (condition & HGVS nomenclature)	Phenotype	Reference	Comments		
CM070963	RRH>RRT	Lys>Phe	5	BIOMASE Feature available in subscribers	Costello syndrome	Zentgraf (2007) J Med Genet 44, 131 Facioscapulohumeral upper limb anomalies Additional report available in pubmed			
CM073168	RRH>GRH	Lys>Glu	5	BIOMASE Feature available in subscribers	Costello syndrome	Beretta (2007) J Hum Genet 52, 324 Additional phenotype report available in pubmed Additional report available in pubmed			
CM076251	GCT>RGT	Gly>Ser	12	BIOMASE Feature available in subscribers	Cardio-facio-cutaneous syndrome	Noya (2007) J Med Genet 44, 283 Additional report available in pubmed			
CM007372	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 pubmed			
CM125166	GCC>GRC	Gly>Arg	13	BIOMASE Feature available in subscribers	Myelodysplastic/myeloproliferative disease ?	Shaw (2012) Br J Haematol 138, 320 Additional report available in pubmed Additional phenotype report available in pubmed	Neutrophilic meta...		
CM061082	GTR>RTR	Val>Leu	14	BIOMASE Feature available in subscribers	Noonan syndrome	Schubert (2006) Nat Genet 38, 331 Facioscapulohumeral upper limb anomalies Additional report available in pubmed			
CM070966	GRR>GRR	Gly>Arg	22	BIOMASE Feature available in subscribers	Noonan syndrome	Zentgraf (2007) J Med Genet 44, 131 Facioscapulohumeral upper limb anomalies Additional report available in pubmed			
CM070964	GRR>GRR	Gly>Glu	22	BIOMASE Feature available in subscribers	Cardio-facio-cutaneous syndrome	Zentgraf (2007) J Med Genet 44, 131 Facioscapulohumeral upper limb anomalies Additional report available in pubmed			

Exercise 2. 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

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>

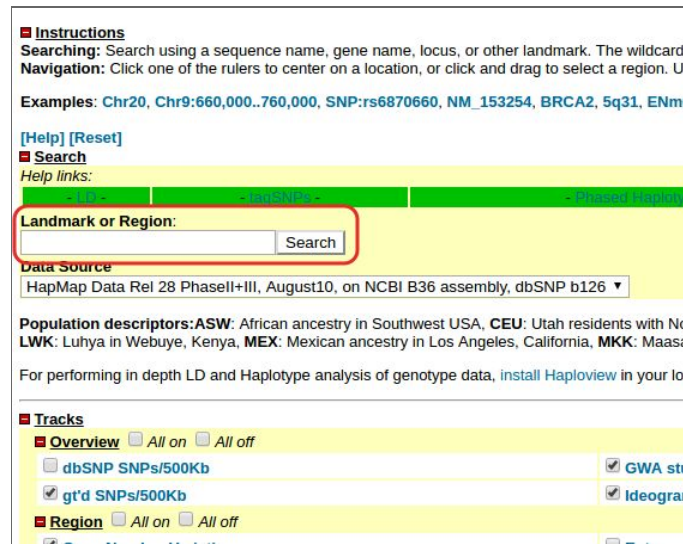
Exercise 3. Browse different catalogs of human genetic variation.

Questions:

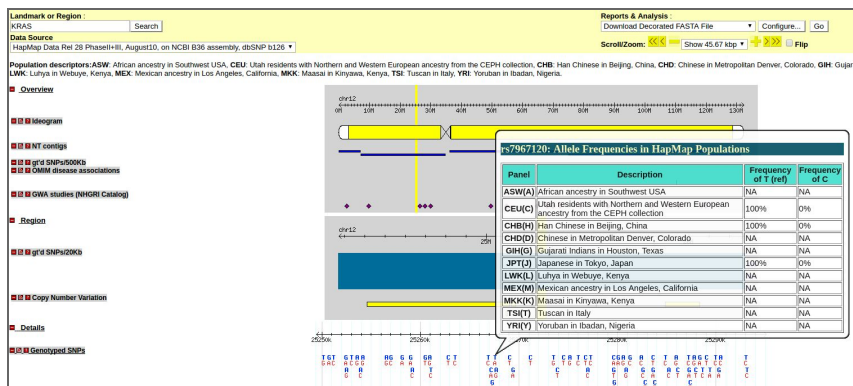
- A) Go to the latest release of the HapMap project and check the KRAS gene region (Note that HapMap uses NCBI build 36). Can you find the allele frequencies of genotyped SNPs in the HapMap populations?
- Type the HapMap URL on your web browser (<https://hapmap.ncbi.nlm.nih.gov/>). Select HapMap release #28 on the left menu (Project Data).

The screenshot shows the International HapMap Project website. The header includes the project name and navigation links. The left sidebar contains a 'Project Data' menu with a red box highlighting the link for 'HapMap Genome Browser release #28 (Phases 1, 2 & 3 - merged genotypes & frequencies)'. The main content area displays a list of news items, including 'HapMap data conversion tool', 'Downtime for hardware maintenance', and 'HapMap help desk announcement'.

- Then, search for KRAS gene on the landmark or region search field.



- In the results page, there is a section named “Genotyped SNPs” where you can pass the mouse over the letters and check the frequency of the alleles. If you click directly on the letter, you can see a frequency report with population genotype and allele frequencies.



- 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?

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

1000 Genomes

A Deep Catalog of Human Genetic Variation

Search 1000 Genomes

12:25350000-25450000 Go

e.g. gene BRCA2 of Chromosome 6:1320008746-132108745

Start Browsing 1000 Genomes data

[Browse Human](#) --
GRCh37

[Protein variations](#) --
View the consequences of sequence variation at the level of each protein in the genome.

[Individual genotypes](#) --
Show different individual's genotype, for a variant.

- 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.

5 features

Version: rs369251072	Version: rs369251072	Version: rs369251072	Version: rs369251072	Version: rs369251072
Class	Class	Class	Class	Class
Location	Location	Location	Location	Location
Alleles	Alleles	Alleles	Alleles	Alleles
Antipathy code	Antipathy code	Antipathy code	Antipathy code	Antipathy code
Global MAF	Global MAF	Global MAF	Global MAF	Global MAF
Consequence	Consequence	Consequence	Consequence	Consequence
Evidence	Evidence	Evidence	Evidence	Evidence
Source	Source	Source	Source	Source
Population genetics	Population genetics	Population genetics	Population genetics	Population genetics

- 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.

Variation: rs369251072

rs369251072 INSERTION

Original source Variants (including SNPs and indels) imported from dbSNP (release 142) | [View in dbSNP](#)

Alleles **-TCTAAAATCAATGAATGTGCTA** | MAF: 0.02 (TCTAAAATCAATGAATGTGCTA)

Location Chromosome 12: between 25371276 and 25371277 (forward strand) | [View in location tab](#)

Most severe consequence Intron variant | [See all predicted consequences \(Genes and regulation\)](#)

Evidence status

HGVS names This variation has 4 HGVS names - click the plus to show

About this variant This variant overlaps 3 transcripts and has 2545 individual genotypes.

Population genetics

1000 Genomes Project Phase 3 allele frequencies

ALL

● - : 98%
● TCTA... : 2%

AFR

● - : 93%
● TCTA... : 7%

AMR

● - : 99%
● TCTA... : 1%

EAS

● - : 100%

EUR

● - : 100%
● TCTA... : 0%

SAS

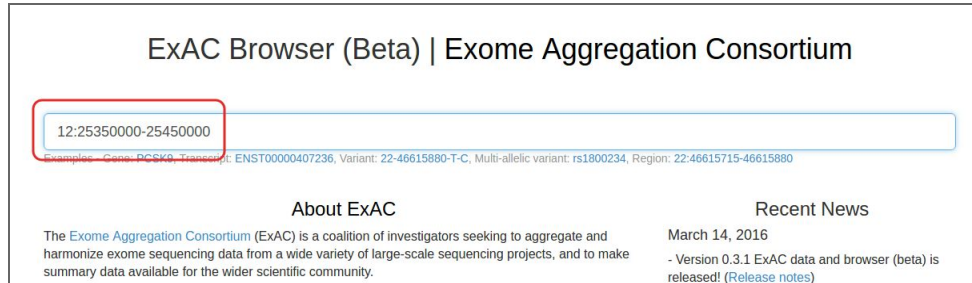
● - : 100%

1000 Genomes Project Phase 3 (32)

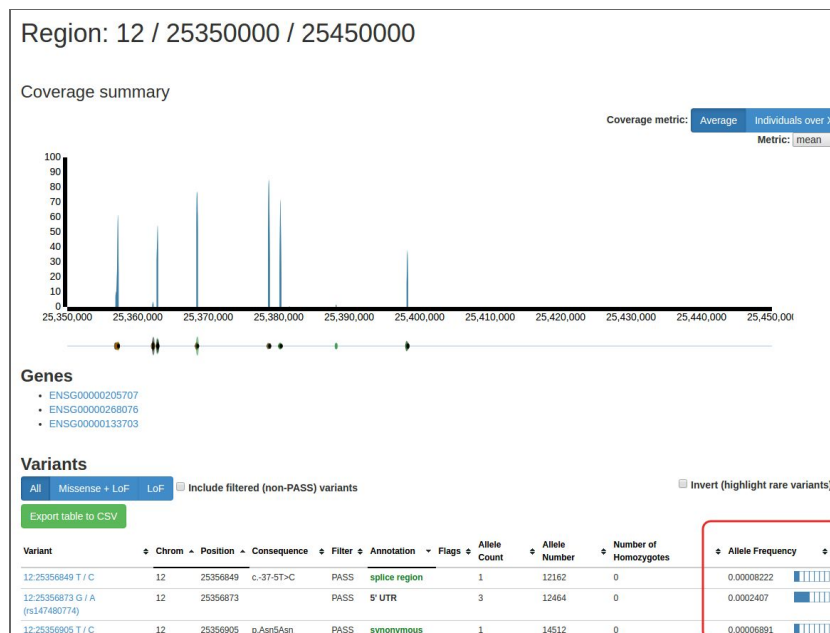
Population	Allele: frequency (count)	Genotype: frequency (count)
1000GENOMES:phase_3-ALL	- : 0.981 (4913) TCTAAAATCA... : 0.019 (95)	- : 0.964 (2413) -TCTAAAAT... : 0.035 (87)
1000GENOMES:phase_3-AFR	- : 0.935 (1236) TCTAAAATCA... : 0.065 (86)	- : 0.876 (579) -TCTAAAAT... : 0.118 (78)
1000GENOMES:phase_3-ACB	- : 0.938 (180) TCTAAAATCA... : 0.062 (12)	- : 0.875 (84) -TCTAAAAT... : 0.125 (12)
1000GENOMES:phase_3-ASW	- : 0.934 (114) TCTAAAATCA... : 0.065 (8)	- : 0.869 (53) -TCTAAAAT... : 0.131 (8)

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

- 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”.



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

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

12:25396264 A / G	12	25396264	p.Leu19Leu	PASS	synonymous		1	102040	0	0.00000800
12:25396268 A / G	12	25396268	p.Ser17Ser	PASS	synonymous		1	102050	0	0.00000978
12:25396279 C / T (rs104894385)	12	25396279	p.Val14Ile	PASS	missense		1	101898	0	0.000009814
12:25396284 C / T (rs121913529)	12	25396284	p.Gly12Asp	PASS	missense		2	101204	0	0.00001976
12:25396285 C / A (rs121913530)	12	25396285	p.Gly12Cys	PASS	missense		2	101218	0	0.00001976
12:25396295 T / C (rs147406419)	12	25396295	p.Val8Val	PASS	synonymous		36	98618	0	0.0003650
12:25396321 T / C	12	25396321		PASS	5' UTR		3	83546	0	0.00003991

- 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

Exercise 4. Browse genomic variation using the CIBERER Spanish Variant Server.

Questions:

- A) Search all the genomic variants of KRAS gene in the Spanish population. How many variants do you find? Now, try again but selecting only the IBS population from the 1,000 Genomes project. How many variants do you find?
- Type the CSVS URL (<http://csvs.babelomics.org/>) and search for KRAS gene. 51 variants are obtained.

- Select only the IBS subpopulation and click on “Search” button. 14 variants are obtained.

B) Which information can we obtain searching the 1:24536 position? (Effect, phenotype, etc.)

- There is no information retrieved by searching for the position 1:24536. Then, we try other genomic region and search for variants included in 12:25368400-25368500.

Chr	Position	Alleles	Gene	Id	Genotypes			Freq.			1000G MAF (phase 1)				1000G MAF (phase 3)									
					D/D	D/1	1/1	0 Freq	1 Freq	MAF	ALL	AME	ASI	AFR	EUR	ALL	AME	South ASI	East ASI	AFR	EUR			
12	25368410	C>T	KRAS	rs20970347	576	2	0	0	0.998	0.002	0.002	0.001	0.000	0.000	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000
12	25368462	C>T	KRAS	rs4362222	110	1	467	0	0.191	0.809	0.191	0.000	0.000	0.000	0.000	0.000	0.000	0.002	0.003	0.000	0.000	0.000	0.008	0.000

- Select the first one (12:25368410) and explore the phenotype and effect information.

Gene Name	Ensembl Gene Id	Ensembl Transc.	Conseq. type	Relative Position	Codon	Strand	Biotype	cDna Position	cds Position	AA Position	AA Change	Sift	PolyPhen
KRAS	ENSG00000133	ENST000000311	intron_variant			-	protein_coding						
KRAS	ENSG00000133	ENST000000557	intron_variant			-	protein_coding						
KRAS	ENSG00000133	ENST000000256	missense_variant		Ggp/Agc	-	protein_coding	599	535	179	GLYSER	0.14	0.002

C) Now, search for BRCA2 gene only in the MGP population. Is there any variant that could be characteristic of the Spanish population?

- Good candidates to be characteristic of a population are those variants that can be found in that population and not in others.

Chr	Position	Alleles	Gene	Id	Genotypes			Freq.			1000G MAF (phase 1)				1000G MAF (phase 3)				ESP-ESV	SIFT	POLYPHEN	PhastCons	phyP									
					D/D	D/1	1/1	0 Freq	1 Freq	MAF	ALL	AME	AS	AFR	EUR	ALL	AME	South ASI						East ASI	AFR	EUR						
13	32953580	C>T	BRCA2	rs2995390	381	39	3	22	0.859	0.141	0.141													0.279	0.491	0.998	0.000					
13	32953590	T>	BRCA2	rs2995391	384	23	0	0	0.957	0.043	0.043																0.007	1.332				
13	32956480	A>C	BRCA2	rs764373	244	23	0	0	0.917	0.083	0.083	0.060	0.090	0.100	0.050	0.040	0.274	0.092	0.133	0.094	0.033	0.035	0.037	0.020	0.121	0.022	0.980	0.341	0.533			
13	32956571	A>C	BRCA2	rs3939572	265	2	0	0	0.996	0.004	0.004	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000			
13	32956679	A>C	BRCA2	rs148848	144	96	27	0	0.719	0.281	0.281	0.240	0.330	0.240	0.190	0.290	0.249	0.300	0.354	0.283	0.081	0.295	0.284	0.129	0.141	0.022	0.980	0.297	0.333			
13	32956680	A>G	BRCA2	rs3825439	244	23	0	0	0.917	0.083	0.083	0.060	0.090	0.100	0.050	0.040	0.274	0.092	0.133	0.094	0.033	0.035	0.037	0.020	0.121	0.022	0.980	0.231	0.435			
13	32957129	T>C	BRCA2	rs2897708	266	3	0	0	0.998	0.002	0.002	0.001	0.000	0.000	0.000	0.003	0.000	0.000	0.000	0.000	0.000	0.000	0.002	0.001	0.000	0.153	0.027	0.980	0.419	0.538		
13	32957401	G>C	BRCA2	rs3428793	258	1	0	8	0.998	0.002	0.002																	0.000	0.000	0.946	0.930	0.538
13	32957536	T>	BRCA2	rs2995392	230	35	2	0	0.889	0.111	0.111																		0.002	0.378		
13	32957536	T>	BRCA2	rs2995393	230	35	2	0	0.889	0.111	0.111																			0.002	0.378	