

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?
- 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?
- C) humsaVar database: could you find the previous rs158691 SNP in this file? why?
- 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
- 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)
- F) HGMD database: register for the public version and try it at home.

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:

- A) We are interested in a particular region of the human genome chr12:25,350,000-25,245,000 (GRCh37), and we want to know if this region contains mutations already catalogued. Help: latest (version), hsa (species), genomic (category), region (subcategory), 12:25350000-25450000 (id), mutation (resource).
- 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).
- 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.

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?
- 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?
- C) Check the allele frequencies of same genomic region in the ESP 6,500 samples.
- D) Finally, check the genetic variation of KRAS in ExAC browser. Which is the allele frequency of rs121913529 in the European (Non-Finnish) population?

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?
- B) Which information can we obtain searching the 1:24536 position? (Effect, phenotype, etc.)
- C) Now, search for BRCA2 gene only in the MGP population. Is there any variant that could be characteristic of the Spanish population?