**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.

**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?

**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?

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

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

Questions:

A) We are interested in a particular region of the human genome (chromosome:start-end) 12:25,350,000-25,245,000 (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 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?

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

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?

**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).

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

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

## References

- 1000 Genomes Project Consortium (2015) A global reference for human genetic variation. Nature, 526(7571):68-74. doi: 10.1038/nature15393
- Bush, WS and Moore, JH (2012) Chapter 11: Genome-wide association studies. PLOS Computational Biology, 8(12):e1002822. doi: 10.1371/journal.pcbi.1002822
- Cooper, GM and Shendure, J (2011) Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. Nature Review Genetics, 12(9):628-40. doi: 10.1038/nrg3046
- Dopazo, J *et al.* (2016) 267 Spanish Exomes Reveal Population-Specific Differences in Disease-Related Genetic Variation. Molecular Biology and Evolution, 33(5):1205-18. doi: 10.1093/molbev/msw005

- Koonin, EV (2012) Does the central dogma still stand? Biology Direct, 7:27. doi: 10.1186/1745-6150-7-27
- Lek, M *et al.* (2016) Analysis of protein-coding genetic variation in 60,706 humans. Nature, 536(7616):285-91. doi: 10.1038/nature19057
- MacArthur, DG *et al.* (2014) Guidelines for investigating causality of sequence variants in human disease. Nature, 508(7497):469-76. doi: 10.1038/nature13127