Leiden University Medical Center, Frontiers of Science, Translational Genomics

Your team has received a set of 7 single nucleotide polymorphisms (SNPs). Your task is to investigate these SNPs using various online resources. Here is a list of the resources:

- GWAS Catalog (https://www.ebi.ac.uk/gwas/)
- ENSEMBL (https://www.ensembl.org/Homo_sapiens/Tools/VR?db=core and https://grch37.ensembl.org/Homo_sapiens/Tools/VR?db=core)
- GTEx Portal (https://www.gtexportal.org/home/)
- LD Link (https://ldlink.nih.gov/?tab=ldproxy)
- CADD (https://cadd.gs.washington.edu/score)
- RegulomeDB (https://regulomedb.org/regulome-search/)
- Enrichr (https://maayanlab.cloud/Enrichr/)
- String-db (https://string-db.org/)
- Sei (https://hb.flatironinstitute.org/sei/)
- ExPecto (https://hb.flatironinstitute.org/expecto?tabId=3)

Your objective is to combine information about these SNPs and explore their potential implications in the context of the respective diseases. To assist you in getting started, we have provided some guiding questions that can be addressed using these online resources.

GWAS Catalog:

- What disease is associated with the provided SNPs?
- Does the GWAS Catalog provide information on genes mapped to these SNPs?

ENSEMBL

- Some tools need a VCF file as input instead of the SNPs. Using ENSEMBL you can convert the list of SNPs to a VCF file.
- Pay attention: some tools use GRCh37 instead of GRCh38. Double check whether you use the correct version for each tool.

GTEx Portal:

- Are the SNPs eQTLs (expression quantitative trait loci)? If so, for which genes?

LD Link:

- Are there other SNPs in high linkage disequilibrium (LD) that are functionally relevant?

CADD/RegulomeDB:

- What are the functional consequences of these SNPs?
- Do these functional consequences relate to specific genes?

Enrichr/String-db:

- Are the genes identified in the previous steps involved in a specific biological process?

Sei:

- Which chromatin profile or which cell type or tissue is most likely affected by a variant?
- Sei is a tool that uses the DNA sequence to predict 21,907 chromatin profiles (i.e. chromatin accessibility, transcription factor binding sites, etc) that were measured in a diverse set of cell types and tissues.
- To understand the output better, you can use their tutorial, tool and documentation (https://hb.flatironinstitute.org/sei/) to answer the following questions:
    1. Run the tool using 'Show VCF Demo'
    2. Use the scatter plot shown on the top. To answer the questions, you can change the view.
        - What sequence class label do the four variants have?
        - Is the predicted effect positive or negative for the four variants? What does that mean?
    3. Scroll down to the heatmap. You can switch between the 'Sequence class scores' and 'Probability diffs' if necessary.
        - What do these sequence classes mean?
        - For the variants with the same sequence class label, do they belong to the same sequence class?
        - Which tissue(s) do the variants most likely affect?

ExPecto:

- Is the expression of a gene affected by a variant? If so, in which tissue or cell type?
- ExPecto is a tool to predict tissue-specific gene expression levels. First, it uses DeepSEA (the precursor of Sei) to predict chromatin profiles. The chromatin profiles are input to a simpler model to predict gene expression.
- To understand the output better, you can use their tutorial, tool and documentation (https://humanbase.readthedocs.io/en/latest/expecto.html) to answer the following questions:
    1. Run the tool using 'Show VCF Demo' to make predictions for the same variants as before.
        - Why does ExPecto only predict the effect of 3 variants?
        - Do the tissue(s) you found using Sei correspond to the tissues you found using ExPecto.