**FUMA Practical Trondheim 2023: PART1**

**Part 1: Functional annotation without FUMA**

In the past week you carried out your first GWAS and have run several of the necessary quality control steps. Today we will focus on interpreting results from GWAS studies.

So far, GWAS have produced large amounts of disease-associated genetic variants. However, these genetic variants typically explain only a relatively small part of the individual differences in the trait, and for most traits many (100s or 1000s) associated variants have been found. Crudely stated, the result of a GWAS is a list of *P*-values for all SNPs included in the analysis, with lower P-values indicating stronger evidence that that SNP is associated with your trait. However, not only truly causal SNPs will show a statistically significant association, but also all SNPs that are correlated with that causal SNP due to linkage disequilibrium. It is thus very difficult, based on P-values alone, to pinpoint the actual causal SNP.

One way to overcome this is to annotate all associated SNPs, and prioritize those SNPs that have a known functional impact. For example, a SNP in a gene-desert with no known regulatory functions is less likely to be the actual causal SNP, than a SNP that changes the protein structure or the level of translated protein.

In order to learn more about which variants are likely to have functional impact on a trait, we will have to do post-GWAS functional association of all associated SNPs. We need to know where our significant SNPs are located, what kind of variants they are (intergenic, intronic, exonic), if exonic; what kind (missense, nonsense etc.), what SNPs they are in LD with, if they are known to influence expression of a gene, etc. etc. This information is available from many different databases. Today we will work with some of these databases.

Let’s assume you have run a GWAS on 36,989 schizophrenia cases and 113,075 controls, all European. You identified 128 independent loci.

We will start with one significant SNP from this analysis:

**Location: 12:2344960**, intergenic

**Effect allele: G**

**Reference allele: A**

**rs-ID: rs2007044**

**p-value: 2.625e-17**

**gene: CACNA1C**

First, find out something about the location of this SNP

* Go to Ensembl: <http://grch37.ensembl.org/index.html>

The GRCh37 build matches the build in which the GWAS was done, always make sure this is the case! [*a ‘build’ is a map of the genome describing the locations of all known variants. With new technology, these maps are constantly updated and variants may change exact position across different builds*]

* Select ‘human’, and input the rs-ID
* Go to ‘Linkage disequilibrium’ in the left panel under Variant displays, go to European - 1000GENOMES:phase\_3:**CEU** (Utah residents with Northern and Western European ancestry) - select ‘Variants in high LD’

***Q1: according to this source, how many variants could be the variant that is actually associated schizophrenia (using a threshold of r2 > 0.8)? [note that r2 is a measure of LD: r2 of 1 is perfect LD r2 of 0 indicates 2 independent SNPs. Usually r2> 0.2 is considered not independent]***

***ANSWER=***

* Any of these SNPs could be the causal SNP, but we’ll stay with the genotyped SNP for now.
* As your variant is located in the *CACNA1C* gene, it could be that it influences the gene function or expression. We can use GTEx, which has data on human expression per tissue to find out where this gene is expressed and see whether this SNP is known to influence expression of CACNA1C.
* Go to <https://www.gtexportal.org/home/>
* Enter the gene name in ‘Browse by gene ID’. Scroll down to the figure.

***Q2: In which five tissues is the gene most highly expressed? Does that make this gene a likely candidate for schizophrenia? What would you have expected?***

* However, the SNP might also influence expression of a gene that is further away, as an eQTL. Go back to <https://www.gtexportal.org/home/> and input the rsID at ‘Browse by variant or rs ID’.

***Q3: why is the location of your variant different than in the output of your GWAS results?***

* The variant is an eQTL for the same gene, but apparently influences gene expression in the brain (click multi-tissue eQTL plot). Interesting!
* Let’s try to find out more about this gene. Useful sites are:
* dbSNP: under ‘Clinical Significance’ you can see whether this variant is in Clin Var, which would mean that this variant is a disease causing candidate. Furthermore, under ‘Publications’, you can see publications that mentioned this SNP specifically.
* Genetics Home Reference: <https://ghr.nlm.nih.gov/gene/CACNA1C>.

***Q4: under ‘health conditions related to genetic changes’ you can see that variants in this gene can cause Timothy syndrome. Why is it unlikely that the SNP we found would cause Timothy syndrome?***

* PubMed (<https://pubmed.ncbi.nlm.nih.gov> ; have other groups identified an association between the gene and schizophrenia?)
* OMIM (<https://www.ncbi.nlm.nih.gov/omim> ; mainly used for monogenic diseases, but does list knock-out mouse models and information about other gene function studies)
* ***Q5: list information you’ve found on these sites to make the association between this gene and schizophrenia more or less unlikely.***

Of course there are many more:

* Sources to annotate your SNPs, and map them to genes
* Significant SNPs in your GWAS and
* SNPs in LD with significant SNPs

To fully interpret our results we have to go through all of the above steps and more for all of these resources and all of the SNPs... You can imagine how laborious, error-prone and subjective the process of trying to extract something functionally relevant from your GWAS results can be. Luckily, there is FUMA, which does all of this for you in 30 minutes and produces results that can be interactively interpreted.