Diary – Variant to Gene Mapping analysis (since 20/10/2023)

**20/10/2023**

Updates scripts for GTEx colocalisation. Worked on the 004\_concat\_results.R

Created a .xlsx file in the src/report/Variant\_to\_Gene\_Tables.xlsx; for tables on V2G analysis. Put a table with colocalisation data: locus, tissue, N\_gene\_sign/N\_gene\_tot

I still have to modify files for eQTLGen colocalisation; still to obtain GTExV8 files for Colon\_Transverse and Colon\_Sigmoid

Did colocalisation for ‘Stomach’ and ‘Small\_Intestine\_Terminal\_Ileum’ **\***

**23/10/2023**

We have OK for using U-BIOPRED eQTL data with genotyped data.

We have OK for using UBC Lung eQTL data.

There was an error in the script to run colocalisation with GTExV8. So I had to run colocalisation again for ‘Stomach’ and ‘Small\_Intestine\_Terminal\_Ileum’ **\***

After discussion with team, I do colocalisation only if the eQTL data for the tissue-gene-credset region contains significant association, aka pvalue <= 5x10-6. 🡪 Updated 003\_run\_coloc\_susie\_GTEx.R to integrate this step.

Update scripts with some quality checks as well on the number of genes analysed, analysed by colo, analysed by coloc.susie

**24/10/2023**

Add additional checks in the colocalisation with GTExV8.

Removed the additional checks line form the Var\_to\_Gene\_pipeline.sh and some from 003\_run\_coloc\_susie\_GTExV8.R

Submit coloc for GTExV8 ‘Lung’, ‘Small Intestine Terminal Ileum’, ‘Stomach’, ‘Esophagus Muscularis’.

**30/10/2023**

Re-read the report as it is up today, and updated it a little.

Put a new check to find if all the genes for each tissue have been analysed.

Run colocalisation for 'Esophagus\_Gastroesophageal\_Junction', ‘Artery\_Tibial’, ‘Artery\_Coronary’

STILL NEED TO RUN COLOCALISATION FOR ‘ARTERY\_AORTA’ -DONE 31/10/2023

STILL NEED TO CREATE eQTL FILES FOR COLON\_TRANSVERSE AND COLON\_SIGMOID. -DONE 31/10/2023

NEED TO CODE 004\_concat\_coloc\_results.R FOR COLOC.SUSIE RESULTS

**31/10/23**

Run colocalisation for 'Artery\_Aorta’

Started working on the liftOver of eQTL data, for Colon Transverse and Colon Sigmoid, with script 000\_liftover\_b38\_to\_b37\_GTExV8.sh, based on Chiara’s script. I wanted to use liftOverPlink, but I can’t with bed file only. Excursus: needed to modify the exe file of liftOverPlink for python3 – print command wants parenthesis for the argument. I needed to download liftOver, apparently not installed in ALICE3.

So, Chiara explained me that the script 000\_liftover\_b38\_to\_b37\_GTExV8.sh does the liftOver on all individuals of GTExV8, meanwhile I am interested in European ancestry individuals. So, I have to start from a different set of GTExV8 .parquet data, as downloaded from the website (<https://www.gtexportal.org/home/downloads/adult-gtex#qtl>) and present in ALICE folder: /data/gen1/ACEI/colocalisation\_datasets/eQTL/GTeX

I have to 1)convert hg38 .parquet file into hg38 .gz file; 2)liftOver hg38 .gz file into hg19 .gz file. In this way, I will obtain the same data Kayesha did for the other tissues (/data/gen1/ACEI/colocalisation\_datasets/eQTL/GTeX/${tissue}.v8.EUR.allpairs.chr${chr}.hg19.txt.gz).

So, I am now looking at Kayesha’s scripts.

Created 000A\_submit\_eqtl\_gtex\_extraction.sh: ok, problem with chromosome X (segmentation issue ?, need to understand) \*\*The problem is with chromosome 23 or X. Anyway, since I did not analyse the sex chromosome for my GWAS, I did not include it in this analysis as well. So, ok, I did not resolve the issue, but I did not have necessity to do it, so it is fine like this.

Other scripts are: 000A\_eqtl\_gtex\_extraction.R; 000B\_eqtl\_gtex\_liftover.sh; 000C\_eqtl\_gtex\_conversion.R

STILL NEED TO ADD THE SCRIPTS ON THE REPORT AND TO RUN THEM. Added them !

I am obtaining GTExV8 data for Colon and Skin (4 tissues in tot). Colon because multi-ancestry paper used it, Skin because other asthma studies have used it.

Having a look to Variant\_annotation\_FAVOR.R: added lines to save corrplots for integrative scores, and polished gene name – and obtained final list of 40 genes found by functional annotation.

Updated the .Rmd with the corrplot and venn diagram from functional annotation.

Obtained .hg19.gz GTExV8 eQTL files for Colon\_Sigmoid, Colon\_Transverse, Skin\_Sun\_Exposed\_Lower\_leg, and Skin\_Not\_Sun\_Exposed\_Suprapubic.

Run eQTL colocalisation for Colon\_Sigmoid, Colon\_Transverse, Skin\_Sun\_Exposed\_Lower\_leg, and Skin\_Not\_Sun\_Exposed\_Suprapubic.

Obtained list of genes found by coloc with GTExV8 eQTL.

NEED TO CODE 004\_concat\_coloc\_results.R FOR COLOC.SUSIE RESULTS

To find gene symbol for genes found by eQTL coloc, uses the webtool <https://www.biotools.fr/human/ensembl_symbol_converter>

I tried with a package on R (gprofiler2) but it does not have info about genes names such as ‘AC’ something, e.g. AC004466.3

**1/11/2023**

Coded 004\_concat\_coloc\_results.R to extract values for coloc.susie with PP.H4.abf > 0.8. It was difficult to figure it out how to do it ! But I managed to do it 😊 Also, added lines to save genes for GTExV8 eQTL colocalisation in the var2gene.xlsx file.

I wrote to Ian, Kath and Mike, because I could use the results form Portelli et al.2021 (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8974692/>) in a systematic way in my V2G analysis. Let’s see what they say.

Updated on git the developments of today. It’s good, so: I have finished the analysis with GTExV8 eQTL colocalisation – STILL NEED TO LOOK INTO WHICH TISSUE AND VARIANTS WERE INVOLVED 🡪 DO IT IN THE REPORT (RESULTS SECTION FOR THIS ANALYSIS).

I think I can move into eqtlGen then ! I still have to look into sQTL from GTExV8… let’s see… I also have to draw a line on the amount of analyses I will do!

[went to the event on Indian healthy cusine]

Tidying up Vars\_to\_gene\_analysis\_tools\_data\_after\_10232023.xlsx file. Updated with the analysis done so far.

Started modifying 000\_run\_edit\_eQTLGen.R and 000\_submit\_edit\_eQTLGen.sh. ADD these scripts into Report.Rmd !

**02/11/2023**

Ok, 000\_run\_edit\_eQTLGen.R and 000\_submit\_edit\_eQTLGen.sh run successfully. Edited and run 001\_submit\_eqtl\_lookup\_eQTLGen.sh 001\_run\_eqtl\_lookup\_eQTLGen.R.

For U-BIOPRED and UBC Lung eQTL, I need to do more work. U-BIOPRED, I do not know where the data are. Which data to use for U-BIOPRED? eQTL with genotyped and RNA-seq data or eQTL with WGS and RNA-seq data ??

I think for now, do the colocalisation for GTExV8 cis-eQTL, eqtlGen cis-eQTL.

Wait for Jing’s answer on UBC Lung eQTL analysis pipeline. 🡪 got scripts for eigenMT, a preprocess step to adjust p-value taking into account of LD between variants. Scripts are in src/coloc\_UBClung/.

Wait for Kath on how to use data for U-BIOPRED for colocalisation analysis

Code analysis for other variant-to-gene mapping analysis: PoPS, rare variant analysis, pQTL, Nearby Mendelian rare disease-genes, Nearby Mouse knockout orthologs genes.

Create src/coloc/002\_prepare\_LDinput\_eqtlgen.R because I needed the pairs file for eqtlGen. And run it.

Added parameters lines in the get\_LD.sh for eqtlGen. Run get\_LD.sh for eqtlGen.

Created src/coloc/003\_run\_coloc\_susie\_eQTLGen.R and src/coloc/003\_submit\_coloc\_susie\_eQTLGen.sh.and run them. Added to Report.Rmd.

Update tables with tissue pair and number of genes for eqtlGen.