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.

STARTED EDITING 004\_CONCAT\_COLOC\_RESULTS.R FOR EQTLGEN DATA. OK, done for coloc results.

TO DO COLOC\_SUSIE RESULTS WITH EQTLGEN DATA.

**3/11/2023**

(Morning at UHL LRI)

Found coloc\_susie with eqtlGen. Added eqtlGen colocalised genes into table var2genes\_raw.xlsx.

Reading about locus2gene pipeline of OPENTARGET: machine-learning based approached to assign a causal genes to a genomic locus based on a model trained on 445 ‘gold standard positive’ genes from several sources: ‘(i) expert domain knowledge of strong orthogonal evidence or biological plausibility; (ii) known drug target-disease pairs; (iii) experimental alteration from literature reports (e.g. nucleotide editing); (iv) observational functional data (e.g. colocalising molecular QTLs, colocalising epigenetic marks, reporter assays).’ And derived from: ChEMBL III, ChEMBL IV, Eric Fauman Twitter, ProGeM, T2D Knowledge Portal, Open Targets Curated (<https://genetics-docs.opentargets.org/our-approach/prioritising-causal-genes-at-gwas-loci-l2g>).

<https://community.opentargets.org/t/how-to-interpret-variant-to-gene-v2g-and-locus-to-gene-l2g-scores-in-open-targets-genetics/266>: ‘Interpreting the L2G score

The score is calibrated so that a gene’s score indicates the fraction of genes at or above that score threshold that would be expected to be true positives. For example, we expect that 80% of genes with a score >= 0.8 would be causal genes, assuming that the characteristics of the chosen GWAS locus are similar to those in the training dataset.

**In other words, the score can be interpreted as reflecting an FDR threshold of 1 - L2G\_score. For example, among all genes with L2G > 0.8, 20% will likely be false positives. Note that this definition means that for a gene with a score exactly 0.8, the probability that it is causal would be slightly less than 80%**, just as the last items “discovered” at an FDR threshold of 20% have a greater than 20% chance of being false positives.

If you have a disease/trait in mind, it is much better to use the L2G to make causal gene inferences, rather than the V2G score. For the most robust interpretation, when there are multiple GWAS for a given trait (or related traits), we advise to look at the L2G results for the equivalent locus in each GWAS. You may also observe cases where there are multiple independent signals at a locus, and you can evaluate the results for those distinct signals.’

I can also identify how many loci are shared in all the asthma studies presented in OPEN TARGET. 🡪 for my introduction in the thesis!! 🡪 added figures and shared loci across 10 studies published in 2018-2023 (root study Han Y 2020).

**I can try to use the L2G pipeline as such:**

Query each asthma study with fine-mapped sentinel variant for each locus (variant +/-500Kb):

So, I have 17 fine-mapped sentinel variants;

‘Asthma’ or ‘Severe asthma’ studies in OPEN TARGET, published 2018-2023, as of 03/11/2023 are:

|  |  |
| --- | --- |
| [GCST010042](https://genetics.opentargets.org/study/GCST010042) | Asthma (Han Y 2020) |
| Nat Commun |
| [GCST90038616](https://genetics.opentargets.org/study/GCST90038616) | Asthma (Donertas HM 2021) |
| Nat Aging |
| [FINNGEN\_R6\_J10\_ASTHMA](https://genetics.opentargets.org/study/FINNGEN_R6_J10_ASTHMA) | Asthma (FINNGEN\_R6 2022) |
| [GCST007798](https://genetics.opentargets.org/study/GCST007798) | Asthma (Ferreira MAR 2019) |
| Am J Hum Genet |
| [SAIGE\_495](https://genetics.opentargets.org/study/SAIGE_495) | Asthma (UKB SAIGE 2018) |
| [GCST007995](https://genetics.opentargets.org/study/GCST007995) | Asthma (childhood onset) (Pividori M 2019) |
| Lancet Respir Med |
| [GCST006911](https://genetics.opentargets.org/study/GCST006911) | Asthma (moderate or severe) (Shrine N 2018) |
| Lancet Respir Med |
| [GCST90018795](https://genetics.opentargets.org/study/GCST90018795) | Asthma (Sakaue S 2021) |
| Nat Genet |
| [GCST009798](https://genetics.opentargets.org/study/GCST009798) | Asthma (Olafsdottir TA 2020) |
| Nat Commun |
| [GCST90014325](https://genetics.opentargets.org/study/GCST90014325) | Asthma (Valette K 2021) |
| Commun Biol |

Create a folder and a script for this analysis, but left it there, because not able to run the command and tired.

**06/11/2023**

Chase AMS UKBiobank again for the bridging file.

I think I should start either the UBCLung eQTL, pQTL or the rare variant analysis.

The analysis with OpenTarget L2G data: it is not a priority task, so I will leave it there for now.

LET’S START WITH pQTL analysis: deCODE pQTL.

deCODE: variants are significant if p-value < 1.8x10-9: so, if in the credible set region there is at least one variant with pQTL significant p-value, I will include this region in the coloc/coloc.susie analyses.

Need to ask people the scripts to analyse deCODE !

**09/11/2023**

I spoke with Jing about UBC Lung eQTL data preparation: the p-value needs to be corrected for LD structure as well as for multiple testing. Genome-wide p-value cannot be used because eQTL has less power than GWAS due to the smaller sample size. Therefore, we can use other correction p-value method based on LD structure, as implemented by eigenMT; after this, an additional correction for benjamini-hockberg is needed to account for multiple testing; finally, significant associations are defined based on a FDR 0.05 threshold. Jing has done this steps (eigenMT, benjamini-hockberg- FDR threshold) for the multi-ancestry lung function paper so that I can use the significant list of variant-gene she obtained.

She forwarded me the codes to prepare the data for colocalisation for UBC Lung eQTL: 1) do a lookup of the region of interest – if no overlap, no reason to run the analysis. And then a script to create UBC Lung eQTL input for colocalisation using coloc/coloc.susie. Scripts saved in folder src/coloc\_UBClung; I rename the file to have a clearer order in my mind. NB: I deleted scripts related to eigenMT because I do not need to run it.

All loci except SA\_10\_9064716\_C\_T have overlaps with UBCLung eQTL data.

Updated github with UBCLung scripts.

TO DO: WORK ON 001\_run\_coloc\_input\_lung\_eQTL.r (done on the 16/11/2023)

Not understanding how I am supposed to do prepare the data for UBC Lung, frustrated – left it there- next time will be better.

**16/11/2023**

Ok, so, I asked Jing, there is another script that I needed for this analysis !

For UBCLung, do the analysis by probeset: when significant results, map the probeset to gene with the file /data/gen1/reference/lung\_eQTL/tabMerged\_anno.txt.

Ok, so, I was able to run coloc for UBCLung eQTL: 003\_submit\_coloc\_susie\_kung\_eQTL.sh and 003\_run\_coloc\_susie\_lung\_eQTL.r

TO DO: UPDATE REPORT WITH SCRIPTS AND ANALYSIS FOR UBCLUNG!

**17/11/2023**

Wrote scripts for colocalisation UBCLung eQTL ! 003\_submit\_coloc\_susie\_lung\_eQTL.sh and 003\_run\_ coloc\_susie\_lung\_eQTL.r, and 004\_concat\_coloc\_susie\_ubclung.R

TO DO: ONCE JOBS FINISHED, RUN 004\_concat\_coloc\_susie\_ubclung.R

Updated github with new scritps and updated the Report.

Analyses left to do:

pQTL colocalisation UKBiobank

pQTL colocalisation SCALLOP

pQTL colocalisation deCODE

ExWAS – collapsing

ExWAS – individual variants

Opent Target (variant to gene ? locus to genes ?)

PoPS

Nearby Mouse knockout orthologs genes

NB! This study has such nice tables of functional analysis from asthma in UK Biobank that I could use ! <https://pubmed.ncbi.nlm.nih.gov/34103634/> They did TWAS, DEPICT, PoPS, Mendelian Randomisation for asthma They also used GARFIELD: supplementary table SD10 has results for GARFIELD, I can compare with mine. They did a PheWAS for genes then

**18/11/2023**

Library(xlsx) does nto work any more, problem with rJava. So, I saved the genes form UBCLung colocalisation as a tsv file and then added a table in /Var\_to\_Gene/input/var2genes\_raw.xlsx.

UBCLung done!

Re run UBCLung eQTL colocalisation with window +/- 1MB (I did it for 500,000bases before). Genes are the same.

Started UKB-pQTL: look-up of significant proteins. Check that the p-value threshold I am using for significance is the right one (will look at people answer on Monday) 🡪 Nick confirmed the threshold is good.

**20/11/2023**

Kayesha-Alex: they did a gene-based rare variants analysis for ExWAS in UK Biobank. They made some assumptions – so good to have a talk with them about their strategy.

Be in contact with people who have done:

pQTL colocalisation with SCALLOP, with UKBiobank, with deCODE: ok, so, look-up from the original files and then use eQTL colocalisation scripts to run pQTL colocalisation.

**Abril and Jing:**

-Nearby Mendelian rare disease-genes

-Nearby Mouse knockout orthologs genes

-Polygenic Priority Score (PoPS)

**For pQTL:** as per the TSH paper (https://www.nature.com/articles/s41467-023-42284-5#Sec10): ‘We performed approximate colocalisation between our GWAS and pQTL sentinel variants by identifying whether a sentinel pQTL SNP was in one of our 95% credible sets.’ 🡪 they performed a look-up for pQTL signals basically. I could follow the same approach as it will be less time consuming! LOOK-UP IN pQTL done also for credible set in MHC region!.

* So, let’s look at all the variants in the credible sets: if they have pQTL significant association in pQTL UKBiobank. I was limiting this analysis to the sentinel variants with highest PIP, but actually for a comprehensive picture, let’s do a look-up for all the variants in credible sets:

So, I modified the src/pQTL\_coloc/000\_submit\_lookup\_ukbpqtl.sh in order to do the look up for each variant; IF NO RESULTS FROM THE LOOK-UP, THEN THERE IS NO REASON TO DO A FORMAL COLOCALISATION WITH COLOC/COLOC.SUSIE.

IF RESULTS, NEED TO DO: IDENTIFY AS CIS- OR TRANS- THE PROTEINS THAT SHOW INTERACTION.

To do so, I am following the scripts for orion pain: /data/gen1/UKBiobank/olink/pQTL/orion\_pain:

Looked at

To do the look-up: run\_lookup.sh

To combine files with a lookup: combine\_pqtl.awk

To identify cis and trans pQTL (as per +/-1MB from variant): cistrans.R

Cistrans.R: NO RUN AT THE MOMENT, THERE IS A PROBLEM WITH ucsc = browserSession("UCSC") and library(rtracklayer); so for now, I leave it.

TO DO: WAIT RESULTS FOR CHROMOSOME 6 MHC CREDSET LOOK-UP AND RUN combine\_pqtl.awk FOR pQTL UKB

TO DO: A LOOK-UP ANALYSIS FOR SCALLOP AND deCODE. 🡪 useful scripts in github repository with Chiara’a and Kayesha’s codes for look-up in SCALLOP and deCODE: <https://github.com/legenepi/tsh_project/tree/main/pQTL>

20-24 WEEK:

pQTL approximate colocalisation (look-up) analysis in 3 datasets – do it ! – NEED TO WAIT FOR NICK TO BE BACK – HE KNOWS HOW TO DO IT.

**21/11/2023**

The github repository of Chiara is out to date and she addressed me to Kayesha and Alex for the final analyses about look-up in deCODE and SCALLOP pQTL. I emailed them asking to have access to these analyses – hoping that they will send me scripts for these.

In the meantime, I received the UK Biobank bridging file for “application (88144) to those of application 56607, as requested. The password to decrypt the file is the 64-character encryption key associated with application 56607's last approved basket (released on 10.12.2022).” Finally! After more than year asking for it !.

Saved the bridging file in: /rfs/TobinGroup/data/UKBiobank/application\_88144/Bridge\_eids\_88144\_56607.csv

Organised meeting with Jing on Thursday to talk about PoPS - Nearby Mendelian rare disease-genes, Nearby Mouse knockout orthologs genes. She ran these analyses for the TSH paper based on Nick’s scripts.

I wanted to modify the command to save gene list from Var\_to\_Gene/src/Variant\_annotation\_FAVOR.R: since xlsx does not work, I save them as txt file and then add them into the Var\_to\_Gene/input/var2genes\_raw.xlsx file. BUT NOT POSSIBLE: I AM USING ANOTHER LIBRARY – QDAP THAT NEEDS RJAVA SO BETTER IF I FIX THE ISSUE.

Decided not to do sQTL colocalisation: time-wise decision.

I found a look-up script for deCODE data : /data/gen1/TSH/coloc\_susie/lookup\_decode.awk

I am trying to run it, but not results… Maybe I am running in the wrong way: wait for Nick to be back and ask a meeting with him to help me with these analyses.

Ok, so now I can have a look into PoPS, nearby mendelian rare disease-gene (ORPHANET) and Nearby Mouse knockout orthologs genes (International Mouse Phenotyping consortium).

PoPS 🡪 using v0.1 which is the one that other lab members have used. (<https://github.com/FinucaneLab/pops/tree/add-license-1>) and it can be found here:

/data/gen1/LF\_HRC\_transethnic/PoPS/

Ok, let’s have a look at PoPS: 🡪 write basic pipeline of PoPS in the .Rmd report.

Ok, following some log files and using my stalking skills, I was able to understand how to run MAGMA and obtained .genes.raw and .genes.out. Starting writing src/PoPS/PoPS.sh.

I coded step2 as a job file, it is running – let’s see tomorrow !

For the interpretation of final POP score: ‘We currently suggest taking the highest scoring gene in each GWAS locus. You could further filter this set to only include genes in top 10% of PoP scores across all genes. Negative scores generally mean low evidence -- this is the predicted MAGMA z-score!!’ (<https://github.com/FinucaneLab/pops/issues/4>) From Kayesha’s work and multi-ancestry paper (think so): ‘gene with highest polygenic priority score within ±250 kb of a the top PIP SNP for credset’.

I downloaded TableS21 of Zhou et al. 2022 in which they released the prioritised genes from 5 different gene prioritisation analysis: they showed that there is little agreement between the different methods. Anyway, I would use this table to verify the genes I found by my var-to-gene strategy: ‘have been already prioritised for asthma ?’ 🡪 look-up with this table

**22/11/2023**

I was a little angry yesterday and I did not notice that Jing shared her codes for PoPS - Nearby Mendelian rare disease-genes, Nearby Mouse knockout orthologs genes! Although, I learnt about PoPS and what I did was correct, so it’s a win win at the end !

Submit PoPS Job with feature selection and score calculation; job ID: 306557; logfile: SA\_POPS-306557.out

For PoPS results – created table Var\_to\_Gene/output/PoPS\_results.xlsx as they did for the TSH paper upplementary table S19:

|  |  |  |  |
| --- | --- | --- | --- |
| **Gene** | **Sentinel** | **PoPS\_score** | **Window** |
|  |  |  |  |

I like the idea of it.

In the TSH paper – they looked at a 500kb total window (+/-250Kb); if no genes, they opened up to 1Mb total window (+/-500Kb). Started looking into src/PoPS/PoPS\_summary.R – but I need the results from submit\_pops.sh to finale the script.

I learnt how to remove a file form github if I do not want it to be seen in the github repo:

1. add the file in the .gitignore file

2. git rm --cached <filename>

3. git commit -m "<Message>"

4. git push

Et voila !

Started with ‘Nearby mouse knockout orthologs with asthma related phenotype’ (title taken from the TSH paper ‘Nearby mouse knockout orthologs with thyroid related phenotype’). for the multi-ancestry study lung function paper, they used "respiratory system phenotype". I think this term is too broad for my analysis. 🡪

The issue is that there is no term for ‘asthma’ of substring of ‘asthma’. So I thought to use some key terms to pull out asthma-related MP terms: ‘lung’/’airway’/’muscle’/’imm’/’epith’/’ bronchoconstri’/’pulm’. From this search, I obtained a total of 35 MP terms, here attached with the respective key term. ‘Do you think this is a reasonable approach? Or should I decide not to run this analysis based on the fact that there is no ‘asthma’ MP term?’ Asked this to Kath and Ian! Waiting for their answer.

* New terms can used – although Kath pointed out to a review in which they highlighted the differences btw mice and human asthma:

|  |
| --- |
| Distribution of lung inflammation different from human asthma |
| Lack of chronicity of the response to allergen |
| Tolerance after repeated allergen exposure |

In the meantime, PoPS Job with feature selection and score calculation - job ID: 30655 completed !

And I run PoPS\_summary.R and obtained the list of prioritised genes from PoPS !!!! Yuppyyy !!!! 😊

Question to Jing:

Why hypergeometric test in mouse\_ko analysis?

Analysis left to do:

1.mouse knock out –

2.nearby mendelian rare disease genes –

3.pQTL lookup in SCALLOP –

4.pQTL lookup in deCODE –

5.rare variant analysis look-up in AZ PheWASPortal and other database ? –

6.gene-based collapsing rare variant analysis in the RAP.

DAI !!!! Positive !!!

**23/11/2023**

Spoke with Jing ! So, for the TSH paper for rare mendelian and mouse knock out analyses, she used the scripts Nick used for the multi-ancestry paper. So, for more in depth questions on the code ask Nick directly.

During next team meeting ask: look-up in SCALLOP and deCODE for TSH paper, who did this analysis ?

Mouse knock out gene: from Ian: ‘Mouse models of allergic airway disease are not great for translation to humans and are really just an inflammatory model. Regarding search terms, I would keep this relatively broad with any airway related trait and/or inflammation. There is also value in look up search terms related to cells types that may be relevant as we have done for the lung function papers e.g. smooth muscle, inflammatory cell numbers.’ 🡪 so I proposed to use ‘respiratory’ as in the lung function paper and do a manual filtering for phenotype relevant to asthma at the end.

Rare-variant ExWAS look-up in AZ-PhewWAS Portal: first thing to do is to define the phenotype to look at –

Do the look-up in other summary statistics for rare-variant ExWAS for severe asthma ?

Work on mouse knock out analysis: I broadened up the search for top level MP term: "respiratory system phenotype", "immune system phenotype", "muscle phenotype". Obtained: 180 subtypes MP terms. Start with this and then at the end do manual filter of the subtypes MP term relevant for asthma.

* From Kath: “The absence of smooth muscle seems odd to me when I can find so many smooth muscle phenotypes in the MGI database: <https://www.informatics.jax.org/>. I wonder whether a better strategy might be to use their human-mouse disease connection tool: <https://www.informatics.jax.org/humanDisease.shtml> Perhaps we can ask in team meeting if there was a reason this wasn’t used for the lung function and thyroid papers…”

Need to troubleshoot error: ucsc <- browserSession("UCSC")

Error in .local(.Object, ...) : Failed to obtain 'hguid' cookie

Resolved the issue! Running:

#if (!require("BiocManager", quietly = TRUE))

# install.packages("BiocManager")

#BiocManager::install(version = "3.18")

And installing new version of BiocManager – it took ~2hrs, but then it all worked ! Nice.

Mouse\_ko.r works. I obtained a list of genes – but I wanted to double run it again to see if it works fine from terminal

**27/11/2023**

Ask in team meeting who did:

In the TSH paper:

1-2 Mouse knock out and nearby mendelian genes: ask nick to explain the rationale of the analysis – is 500Kb reasonable or should I reduce it to 250Kb ?

3.pQTL lookup in SCALLOP –

4.pQTL lookup in deCODE –

5.rare variant analysis look-up in AZ PheWASPortal ?

Ask Nick to help us with the:

6.gene-based collapsing rare variant analysis in the RAP.

TO DO: Do the phenotype-covariate file to use severe asthma application – convert the IDs for the new application

* iPathwayGuide: for gene pathway analysis – available at UoL?

Preparing for Nick meeting:

**Mouse knock out and rare disease** 🡪 Mouse\_ko.r worked on terminal. Another analysis done BUT need to speak with Nick about which data are relevant !

Started working on rare\_disease.r ! Ok, it works, BUT need to speak with Nick about this analysis: for example, I removed the disease column because there is no match for ‘asthma’ – is that ok ?

* **mouse knock out and rare disease – results\_by\_gene give the genes – hypergeometric test maybe not needed for my purposes** –look at the term for rare disease: same as used in the lung function paper ?

“respir, lung, pulm, asthma, COPD, pneum, eosin, immunodef, cili, autoimm, leukopenia, neutropenia, Alagille syndrome”

* **after meeting with Nick, I decided to use the key term: "asthma","eosin","immunodef","cili","autoimm","leukopenia","neutropenia" to filter in HPOTerm and Disease columns. Run analysis again, completed. Wrote genes in the xlxs table.**

**Look-up in pQTL deCODE and SCALLOP:**

I found a look-up script for deCODE data : /data/gen1/TSH/coloc\_susie/lookup\_decode.awk

I am trying to run it, but not results… Maybe I am running in the wrong way: wait for Nick to be back and ask a meeting with him to help me with these analyses.

After meeting with Nick:

deCODE – modify sentinel – allele in alphabetical order

SCALLOP: we were not able to find a script to do the look-up in SCALLOP – anyway, I did this for the smoking paper – try to look at that script ?? Found some scripts from Chiara !

**Analyses left to do:**

**~~3.pQTL lookup in SCALLOP –~~**

**4.pQTL lookup in deCODE –**

**5.rare variant ExWAS in the RAP.**

**6.gene-based collapsing rare variant ExWAS in the RAP.**

**(7.rare variant analysis look-up in AZ PheWASPortal)**

After speaking with Chiara, she pointed out to a script for lookup in SCALLOP on the backbone of a Nick’s script. 🡪 I was able to run them and now I am waiting for the results and to filter out genes with significant pQTL in the region of my fine-mapping credible set SNPs !. 🡪 obtained genes for SCALLOP !.

NB:They filtered for credset variants with PIP > 0.1; I did not do this.

TO DO FOR deCODE: create snp\_list with allele in alphabetical order – create .sh job to run over all files

So, today I finalised three analyses – mouse knock out and nearby rare disease mendelian gene and SCALLOP !.

Tomorrow -Tuesday- to do: deCODE lookup and setting up for the RAP

**28/11/2023**

Doing the decode pQTL look-up adapting the script from scallop lookup! The other one of Nick was not working properly, so just needed to adjust the scallop one. **Decode pQTL look-up is on b38, in fact it needs dependency of the script src/pQTL\_coloc/000\_preprocess\_cs\_b38.R TO CREATE B38 FILES**.

With this script, I need chr and pos, no need alleles.

The job is running – waiting for final results

Ok, so now I can move on the rare variant analysis in ExWAS data in UKBiobank!

I need to create the pheno-covariate file with application 88144 – created folder in /data/gen1 for application 88144:

*mkdir /data/gen1/UKBiobank/application\_88144*

*cd /data/gen1/UKBiobank/application\_88144*

*cp /rfs/TobinGroup/data/UKBiobank/application\_88144/Bridge\_eids\_88144\_56607.csv ./*

I created the pheno-cov file and uploaded in */rfs/TobinGroup/data/UKBiobank/application\_88144/* so that Kath can access it as well.

Now I have to read the script of Nick for gene-based collapsing rare variant analysis and set up my RAP !

Nick’s script for rare variant gene collapsing analysis:

<https://github.com/legenepi/rare_collapsing>

**29/11/2023**

Morning with Kath – set up rare variant and gene-based collapsing analysis ! It was useful and fun !

NB: ‘sandpit’ to call folders for testing purposes.

To understand masks for gene-based collapsing analysis, this paper: <https://www.nature.com/articles/s41586-021-04103-z> *‘For each gene, we considered two categories of masks: a strict burden of rare pLOFs (M1) and a more permissive burden of rare pLOFs and likely deleterious missense variants (M3). For each of these groups, we considered five separate burden masks per gene, based on the frequency of the alternative allele of the variants that were screened in that group: MAF ≤ 1%, MAF ≤ 0.1%, MAF ≤ 0.01%, MAF ≤ 0.001%, and singletons only. Thus, overall, up to 10 burden tests were performed for each gene’*

SBAT paper for gene-based collapsing analysis:

[https://www.biorxiv.org/content/10.1101/2023.02.22.529560v2](https://eur03.safelinks.protection.outlook.com/?url=https%3A%2F%2Fwww.biorxiv.org%2Fcontent%2F10.1101%2F2023.02.22.529560v2&data=05%7C01%7Cnnp5%40leicester.ac.uk%7Cf9b0e842e7f04874225b08dbf0d8f77d%7Caebecd6a31d44b0195ce8274afe853d9%7C0%7C0%7C638368586871230323%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=6YkWGzTD4zmOzPl0vedu7j9nuXUNosodbw1f5hgXKb4%3D&reserved=0)

Tutorial ‘running regenie’ here: <https://dnanexus.gitbook.io/uk-biobank-rap/getting-started/research-analysis-platform-training-webinars>

No need to have same genome build for step1 and step2 of REGENIE as explained here:

[https://github.com/rgcgithub/regenie/issues/82](https://eur03.safelinks.protection.outlook.com/?url=https%3A%2F%2Fgithub.com%2Frgcgithub%2Fregenie%2Fissues%2F82&data=05%7C01%7Cnnp5%40leicester.ac.uk%7C5e050a0a9aa140b42f0d08dbf0d9d36f%7Caebecd6a31d44b0195ce8274afe853d9%7C0%7C0%7C638368590560385153%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=HuCY4N7QMtHHU7LtoN62Nf8e8zp1Twsj%2BrVMoXm69wE%3D&reserved=0)

So, for the single-variant rare-variant ExWAS, in the tutorial of DNANexus, they suggest to use good quality exome data for REGENIE step2. Since I want rare variant, I will filter for the ukb23158\_500k\_OQFE.90pct10dp\_qc\_variants.txt in regenie step2. For the plink qc step ahead of step2, I will use filtering parameters:

--autosome \  
--mac 3 \  
--geno 0.1 \  
--hwe 1e-15 \  
--mind 0.1 \

The reason for the ukb23158\_500k\_OQFE.90pct10dp\_qc\_variants.txt file:

*“As noted in the UKB WES 200k FAQ (here, section 23.d), we suggest the inclusion of a batch covariate in association tests on these data, to account for differences in oligo lots between Phase 1 and Phase 2. These coverage heterogeneities can also be mitigated by a single variant-level filter requiring that at least 90% of all genotypes for a given variant - independent of variant allele zygosity - have a read depth of at least 10 (i.e. DP>=10). When this filter is applied to the UKB WES 200k data prior to association analysis, the results are largely devoid of the spurious hits (Fig. 1B). Application of this depth filter (“90pct10dp”) is consistent across the UKB 200k and UKB 300k WES sets with respect to numbers of variants removed (Table 1).”* [*https://dnanexus.gitbook.io/uk-biobank-rap/science-corner/whole-exome-sequencing-oqfe-protocol/generation-and-utilization-of-quality-control-set-90pct10dp-on-oqfe-data/details-on-processing-the-300k-exome-data-to-generate-the-quality-control-set*](https://dnanexus.gitbook.io/uk-biobank-rap/science-corner/whole-exome-sequencing-oqfe-protocol/generation-and-utilization-of-quality-control-set-90pct10dp-on-oqfe-data/details-on-processing-the-300k-exome-data-to-generate-the-quality-control-set)

Run qc for exome data in plink.

Waiting for regenie step1 to finish– once finished:

1 run regenie step2 for single-variant rare-variant ExWAS analysis.

2 run regenie step2 for gene-based collapsing analysis rare-variant ExWAS analysis.

Copied /data/gen1/LF\_HRC\_transethnic/V2G/new\_heatmap.R script so that I have the script for the table and visualisation of the variant-to-gene mapping and prioritisation analysis.

cp new\_heatmap.R /home/n/nnp5/PhD/PhD\_project/Var\_to\_Gene/src/

The input file for this command looks like: 🡪 Need to create a file like: (To put in Supplementary as full list of mapped genes, then in results the list of prioritised genes with the evidence, locus and variant(s) associated)

head v2g\_full.txt

**evidence signal gene locus start end chr pos trait effect other eaf Z P Novel width**

eQTL 3\_49869158\_A\_C RBM6 FEV1\_locus60 48869158 50869158 3 49869158 FEV1 A C 0.553 7.534 4.91e-14 novel 2

eQTL 17\_62010270\_A\_G CD79B FEV1\_locus237 61010270 63010270 17 62010270 FEV1 A G 0.399 -6.733 1.66e-11 novel 2

eQTL 1\_113054659\_C\_T ST7L FVC\_locus12 112054659 114054659 1 113054659 FVC T C 0.616 -6.628 3.41e-11 novel 2

eQTL 1\_1822209\_C\_T CDK11A FVC\_locus1 1160390 3160390 1 1822209 FVC T C 0.25 6.47 9.78e-11 novel 2

eQTL 1\_1822209\_C\_T NADK FVC\_locus1 1160390 3160390 1 1822209 FVC T C 0.25 6.47 9.78e-11 novel 2

eQTL 17\_38352743\_A\_ATT MED24 FF\_locus232 35915540 39265196 17 38352743 FF A ATT 0.844 7.447 9.53e-14 tier3 3.349656

eQTL 1\_41544279\_G\_T SCMH1 FVC\_locus7 40544279 42544279 1 41544279 FVC T G 0.676 -6.715 1.89e-11 tier3 2

eQTL 12\_3390197\_C\_T TSPAN9 FVC\_locus174 329170 5384844 12 3390197 FVC T C 0.0614 6.167 6.96e-10 novel 5.055674

eQTL 12\_2908330\_A\_C TSPAN9 FVC\_locus174 329170 5384844 12 2908330 FVC A C 0.0295 -7.477 7.62e-14 previous 5.055674

In term of the script though, I only need 2 columns: ‘evidence’ and ‘gene’:

Need to combine together genes from eQTL and pQTL respectively. No significant for decode. 🡪 done it for pQTL

Need to merge genes for eQTL colocalisation

For GTExV8 genes, in ensemble, I used an online converter to have the gene symbol GTExV8\_eQTL\_genes\_symbol table in the var2gene.xlsx file.

**30/11/2023**

REGENIE step2 for ExWAS worked successfully on the RAP, yeee !! Need to understand if I can download the all chromosome summary stats – I hope I can!

Almost finishing with the ExWAS, in terms of codes for generating the summary stats – need to interpret the results and decide which variants and p-value threshold: MAF < 1% and P < 5 × 10−6 ?

Run the LocusZoom command to have Manhattan Plot on the RAP (cannot download data but I can do screenshot of it). Waiting the results !

TO DO rare variant gene-based collapsing analysis on the RAP

ASK NICK AGAIN ABOUT HIS CODE, NOT REALLY CLEAR

ASK KATH IF SHE OBTAINED ONE SINGLE FILE FOR REGENIE STEP1 – yes she had one.

**01/12/2023**

Gene-based collapsing analysis for chromosome 22 run successfully. I had a doubt on a warning on the mask: \* set file : [ukb23158\_500k\_OQFE.sets.txt.gz] n\_sets = 418

WARNING: Detected 418 sets with variants not in genetic data or annotation files.

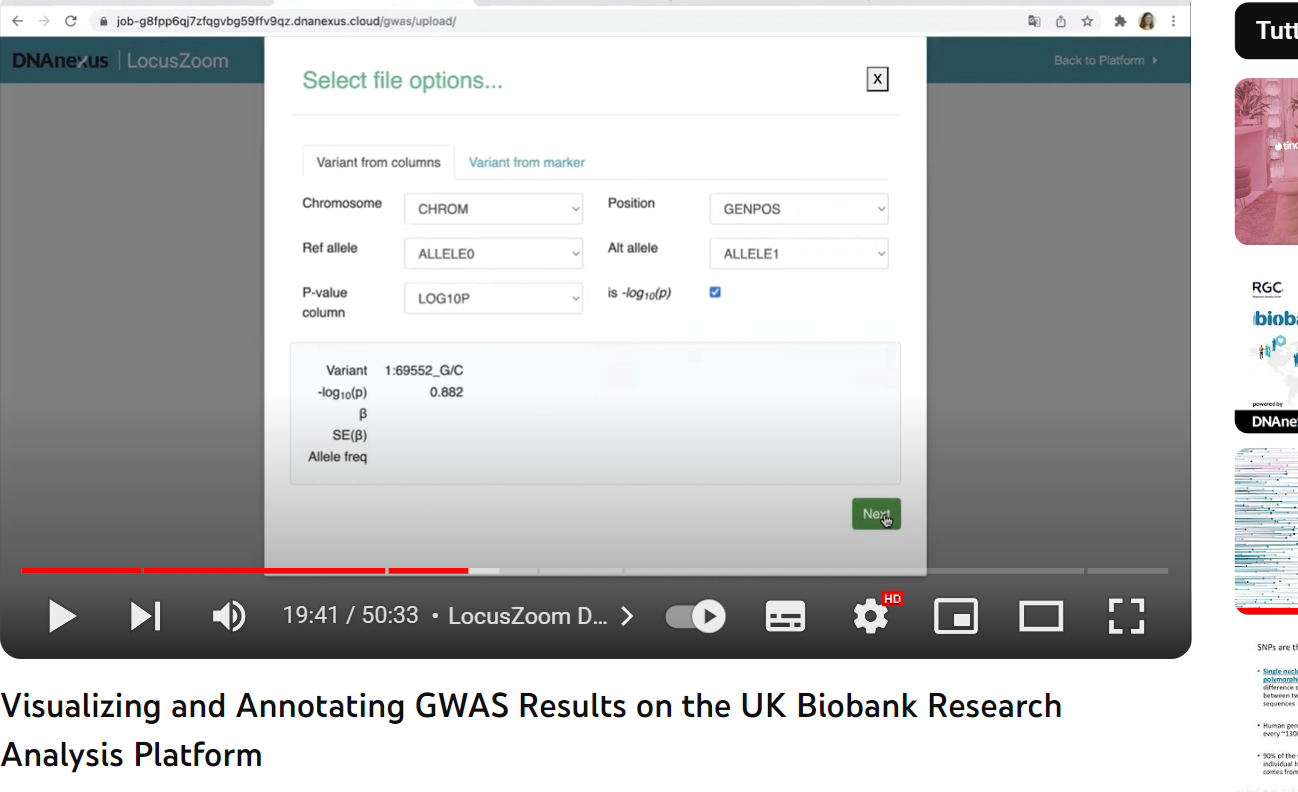
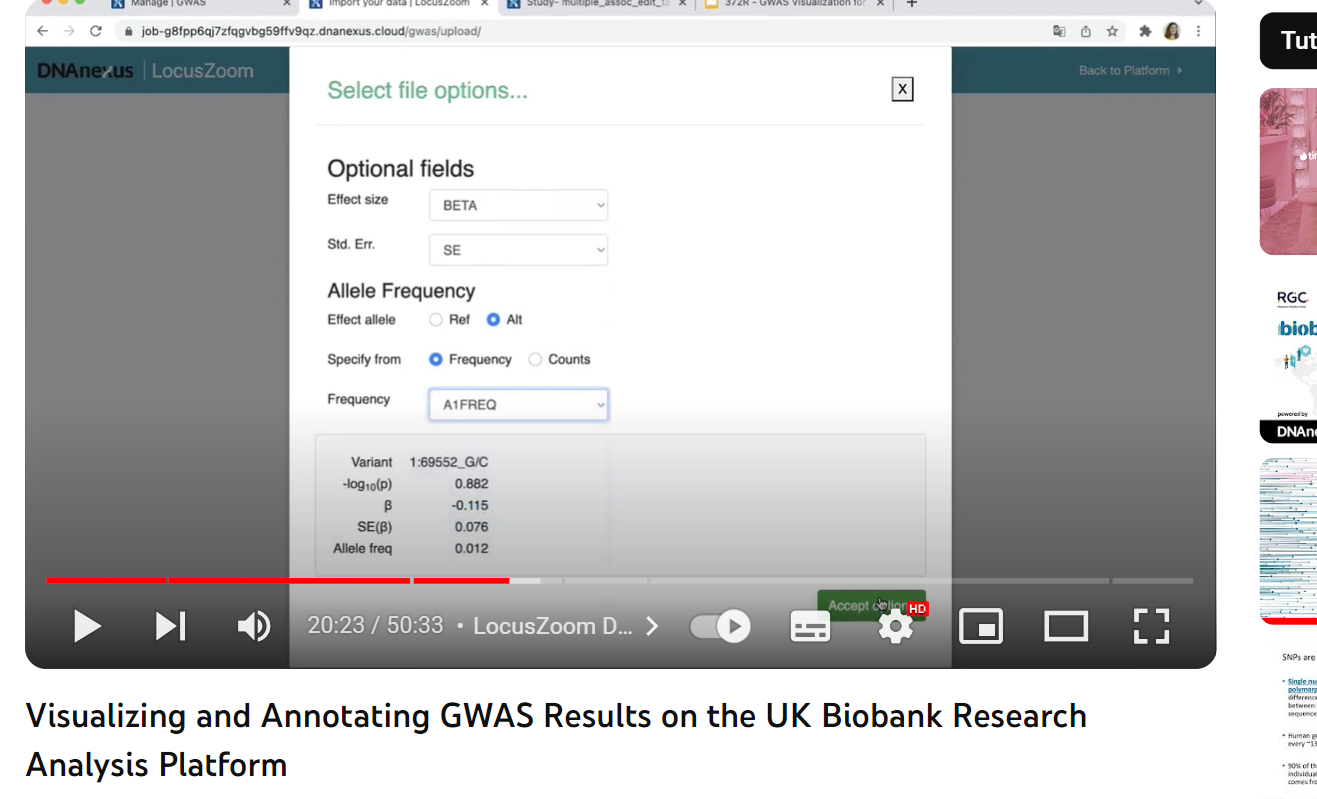
But it got clarified by Nick: ‘missing variants in all sets, not that all sets are unrecognized’

**I cannot download data from the RAP, not even summary stats. So I have to do the analysis on the RAP and then understand how to take the results!.**

Regarding Locus Zoom: ‘I think it can only accept files up to 1GB in size.’ From Callum.

From LOG10P to P: P=(10^(-LOG10P))

In the screenshot, the column for LocusZoom with REGENIE output:

**Apparently, from LocusZoom, I can download the summarystats !! And I can download the LocusZoom as PNG, as well as regional plots.**

Obtained sentinel rare variants for 17 different loci for single variant ExWAS – NEED TO CHECK IF INDEPENDENT AND NEED TO DO QQ PLOT, MANHATTAN PLOT, AND CALCULATE LDSC!

How to calculate LDSC for ExWAS ?

ASK NICK HOW TO DEAL WITH THE RESULTS 🡪 LOOK ONLINE AS WELL?

**Week 4-8/12/2023**

Finalise single and gene-based rare variant exWAS – find mapped genes.

Plot to visualise variant-to-gene mapping results. 🡪 After two hours to understand how it has to be done, able to have the plot – NEED TO TIDY IT FOR COLNAMES, MAYBE PUT A GRID TO SEPARATE VISUALLY AS WELL.

Single rare-variant ExWAS: significant associated variants in LD with common highest PIP and in fine-mapping loci? If yes, which is the nearest gene to these rare variants and functional annotation of these variants ?

Gene-based collapsing ExWAS: significant associated genes in the MAF bin 0.01 and below. Which variants drive the association? 🡪 do a leave-one out association to understand which variants drive the association.

**4/12/2023**

**Analyses left to do:**

**~~3.pQTL lookup in SCALLOP –~~**

**~~4.pQTL lookup in deCODE~~ –**

**5.rare variant ExWAS in the RAP. (1/2)**

**6.gene-based collapsing rare variant ExWAS in the RAP. (1/2)**

For the table visualising genes and analysis, still to finish the script

Solved it, with Nick’s help – following his new\_heatmap.R script, and making it easier.

Not able to put them in order of how many evidence for gene mapping they have, but, it is fine like this for now !

93 genes in total, 17 genes with 2+ evidence: 4 with 3 evidence, 13 with 2 evidence

TOOK THIS STUDY BACK TO DO SOME COMPARISON:

NB! This study has such nice tables of functional analysis from asthma in UK Biobank that I could use ! <https://pubmed.ncbi.nlm.nih.gov/34103634/> They did TWAS, DEPICT, PoPS, Mendelian Randomisation for asthma They also used GARFIELD: supplementary table SD10 has results for GARFIELD, I can compare with mine. They did a PheWAS for genes then

TableS6 has results of Exonic variants associated with asthma (sorted by CADD scores); if of interest !

CHECK 3 levels of evidence – BUT first need to check the phenotype in mouse\_ko analysis **RPS26**

**NSMCE1 IL1RL1 BACH2**

**05/12/2023**

**Today try to do:**

**~~Send agenda to Ian and Glenda~~**

**~~Read about ExWAS~~**

**~~Do one post on the PgDip (the ChatGPT one)~~**

**~~Send points of discussion to Chiara + admin for the day (snacks, coffee ?)~~**

**~~Rename output files for gene-based collapsing analysis in the RAP~~**

**~~Attend Session 2: FREE – VIRTUAL LUNCHTIME SESSION STP GENOMIC COUNSELLING is almost here St George Hospital~~**

Messed up with output names for gene-based rare-variant analysis ! Try to run it again for chromosome 1 to see if it takes same money or less. If less do it again, otherwise changed output files name with the right chromosome name ! I changed the name, re-checked I named it correctly as well and uploaded on the RAP 😊

**06/12/2023**

I think this is why Nick had all the chromosome .loco file in the script: "Of the 454,787 individuals with exome sequencing data, 413 did not have array data after quality control, and so these individuals were excluded from association analyses. For each trait, the leave-one-chromosome-out predictors obtained with step 1 were then included as covariates in step 2 for both the exome sequencing and TOPMed imputed data."

<https://www.nature.com/articles/s41586-021-04103-z#Sec1>

**Meeting with Kath:**

ExWAS – how many independent tests?

Significance threshold: 0.05/N-markers 🡪 all variants (including common vars): 1905952, or only rare variants : 1772263

How to present the results? Using common variants as well ?

Power calculation – to do

a)Look-up in the rare-variant single vars and gene-based collapsing analysis.

b)Reporting the whole results on ExWAS analyses ?

dbSNP to find the gene and the consequence

AZPheWAS Portal – based on the 200,000 UKBB so it may be that they are not included

Lots of neurological genes, theory on neuro function for contraction of smooth muscle cells.

**Genes found – do a look-up in the genes I found from the another analysis.**

Gene-based collapsing analysis:

Maybe underpower?

0.05/N.genes 🡪

Whole Genome Sequencing:

UKBB released now, do a rare-variants WGS 🡪 set up a new project in the RAP and ask for data to be dispend in this project

**12/12/2023**

It took me some days to come back and work on these analyses… Got busy with meeting for the creative project, diploma, and sorting out for the house. But for a couple of hours, let’s focus on the report for the VAR2GENE mapping analyses.

From the meeting with supervisors on the 7/12/2023, they advised to report results for each locus and do the report.

So I will do this now.

Kath wanted the nearest gene as well from the annotation analysis. 🡪 did nearest gene analysis for sentinel-highest PIP variant in each locus. Got final number of 98 genes (rare variants analysis not included yet) and re-did the genes heatmap.

I am creating the table with all the results from V2G analysis – with the script Locus\_to\_gene\_table.R

Still to do for mouse\_ko, rare\_disease, ExWAS rare variant single var and gene-collapsing analysis

**14/12/2023**

Let’s come back to the table with V2G results

TO DO AS WELL: WRITE REPORTS AND METHODS OF MANUSCRIPT !

Do a table for each variant to gene results !

‘xlsx’ keeps give problem with rJava – put there is a libray(“writexl”) so let’s try this. YES, it work!

So, created src/report/var2gene\_full.xlsx to include each analysis results with regards the locus and the sentinel.

Did the different table yeee ! DO IT: RUN IT AGAIN TO SEE IF IT GIVES SAME RESULTS ! Ok, same results.

Sent the tables and updated figures of genes prioritisation to Kath. 🡪 let’s see what she will say

Still to retrieve results form rare variants analyses !!

**15/12/2023**

Ok of tables from Kath – DO REGION PLOTS WITH V2G IDENTIFIED.

Saved in the file the sentinel with highest PIP from the rare\_disease.R script:

fwrite(sentinels, "input/highest\_PIP\_sentinels",row.names=F,quote=F)

So, also did v2g\_minimal: (locus,gene) and b2g\_full: (chr,posb37,locus,gene,evidence)

So that I can use this to fin the gene for each locus for the regionplot

For region plot, I can use the ones that I did for fine-mapping: highlight the gene found in the V2G strategy.

NEED TO DO: I downloaded the regionplto from the finemapping report into the folder for V2G. I will then look at which genes are present in the V2G analysis. 🡪 highlight genes in V2G in abode

For the region plots I did for variant-to-gene finamapping, I used maximum probability only from FINEMAP ! So, I need to run it again to have the sentinel with highest PIP and R2 accordingly to these variants ! – there are three or so locus in which the highest FINEMAP-PIP is NOT the highest AVERAGE-PIP !!.

Let’s take the codes again and re-do the region plots.

PAIN IN THE ASS – I AM TIRED, FRUSTRATED BY CAMBRIDGE HOME SEARCH- FRIDAY AFTERNOON, 5.05PM – WANNO GO HOME

**18/12/2023**

Updated the Report with the description of the analysis – NEED TO PUT THIS INTO THE METHODS OF THE MANUSCRIPT REPORT TO BE FINISHED

**19/12/2023**

Continue to write the Report for Var2Gene analysis.

How to integrate results from rare-variant analysis ?

Following Williams et al. paper (TSH paper), they did ‘(vii) identification of a rare variant (±500 kb of a TSH sentinel variant)’ 🡪 I would use the same approach for a look-up analysis in my single variant and gene-based collapsing analyses.

Look-up of exonic rare variants (MAF < 0.01) within +/- 500Kb credible set variants for each locus using single variant and gene-based results. Genes of exonic rare variants in the regions with suggestive p-value <= 5E-6 were identified.

* No suggestive significant results for single rare variants.
* No suggestive significant results for gene-based collapsing rare variants.

The script src/Locus\_to\_genes\_table.R, I need to tidy it for reported genes with special formatting and have the final number of locus-gene pair for the minimal locus-gene table.

Are they the same gene?

AC034102.1

AC034102.4

AC116366.3

AC116366.1