Methodology and protocol design for identifying candidate drugs for Asthma

# 1. Primary Data Input

The primary data source will be GWAS summary statistics data which can be sourced from the GWAS Catalog (<https://www.ebi.ac.uk/gwas/>). This will be standardised to the GRCh37 build for compatibility.

# 2. Publicly Accessible Dataset Dependencies

## Linkage Disequilibrium (LD) Matrices:

This will be obtained from LDlink (<https://ldlink.nci.nih.gov/>) via the ***LDlinkR*** R package. The LD matrix must be representable of the GWAS summary statistic demographic data.

## eQTL Data:

To investigate how genetic variants affect gene expression across different tissues. Representable Single-cell sequencing data will be downloaded from the eQTL Catalog (<https://www.ebi.ac.uk/eqtl/>). This will be achieved by developing an API in R using the ***seqminer*** package. Here it will be possible to specify gene expression profile, tissue, genomic region and source eg study. If neccessairy this will be converted to GRCh37 build for compatibility.

# 3. Analysis Protocol

## Standardise GWAS summary statistics file:

Using R, the downloaded GWAS summary statistics file (refer to: <https://www.ebi.ac.uk/gwas/studies/GCST006862>) will be imported and standardized into a GenomicRange format. If required, the MungeStats package can be employed to convert this data into a different genomic build.

**R script:** 1.GWAStoGrange.R

**Input**: GWAS file

**Output**: Genomic Range File

## Define locus of interest

In R, visualise summary statistic data using a Manhattan plot which can be achieved using the **karyoploteR** package. From this plot, identify locus of interest that shows significant association with Asthma. Zoom into this region to clearly define the locus for subsequent fine-mapping analysis. Include gene track data using **TxDb.Hsapiens.UCSC.hg19.knownGene** R package to provide reference to the locus.

In example code a major locus on chromosome 17 was identified from summary statistics data. Locus was defined between 17:39500000-40250000 in GRCh37 build.

**Rscript:** 2.DefineLocus

**Input:** Genomic Range File

**Output: i)** Manhatten plot of summary statistic file and ii) Manhatten plot of Locus, iii) GenomicRange file of Locus

## Statistical Fine Mapping:

In R, identify putative causal variants to locus of interest. This will be achieved with the SuSiE fine-mapping tool via the **SuSiE** R package. The susie\_rss() function will be implemented which allows the identification of multiple causal variants per locus. This requires a representable Linkage disequilibrium panel which is obtained from LDlink using the **LDlinkR** package. To visualize the degree of linkage disequilibrium between genetic variants within the defined locus an LD heat map can be plotted using the **gaston** package. This provides a more comprehensive understanding of the genetic architecture of the locus, aiding in the interpretation of the fine mapping results.

In the given example statistical fine-mapping was performed on a locus on Chrm17. Although in practice you would probably do this on numerous loci

**Rscript:** 3.statisticalFineMap

**Input:** GenomicRange file of Locus, ii) Representable LD reference panel (obtained in script)

**Output:** i)Plot highlighting putative causal variants, ii) plot showing credible sets, iii) table of putative causal variants, iv) LD heat map

## eQTL Analysis:

In R, eQTLs were identified from the putative causal variants previously identified from fine-mapping analysis. Their impact on gene expression was examined across different tissue types. The eQTLs are downloaded in the script from the eQTL catalog online database via an API constructed using the **seqminer** package. The user can i) specify relevant gene expression profile to examine ii) which tissue eQTL data is from, iii) the origin study where the data was from and iv) specific genomic region. If necessary the imported eQTL profile should be converted to the sample genomic build as summary statistics data using the **Mungestats** package

The tissue of interest is dependent on the disease, whereas the genes to examine can be informed by searching eQTLs using the putative causal variants into vanno portal online database (<http://www.mulinlab.org/vportal/>).

With the putative causal variants identified from the fine-mapping analysis, their influence on gene expression across various tissues will be investigated using R. eQTL data will be downloaded directly from the eQTL Catalog online database through an API constructed using the **seqminer** package within the R script. This approach enables users to specify the gene expression profile to examine, select the tissue origin for the eQTL data, cite the original study the data was derived from, and designate a specific genomic region. If needed, the imported eQTL profile should be converted to the same genomic build as the summary statistics data using the **MungeStats** package.

The choice of tissue of interest will be determined by the disease under study. Meanwhile, the selection of genes for investigation can be guided by the vanno portal online database (<http://www.mulinlab.org/vportal/>) by inputting the putative causal variants to search for associated eQTLs.

In the example code the following eQTL profiles were examined: IKZF3, GSDMB, ORLD3, GSDMA, PGAP3, MSL1, PNMT. These were explored in different single-cell tissues including: CD4\_T-cell, Macrophage, lung, esophagus\_mucosa, esophagus\_muscularis. By examining the effects of these putative causal variants on gene expression in various tissues, we can gain further insights into the genetic underpinnings of Asthma.

**Rscript:** 4.eQTL.R

**Input:** i) eQTL data (obtained in script)

**Output:** i)eQTL profiles across different tissues

## Colocalization Analysis:

A colocalization analysis will be performed to validate if the eQTLs and GWAS signals (putative causal variants) share the same causal variant. This was performed using the **COLOC** R package which assesses the likelihood of independent or shared causative variants between two traits. Expression profiles were obtained from ***eQTL catalogue*** obtained In previous section the *runsusie()* and *susie.coloc()* functions were used to relax the single causal variant per locus assumption. This requires a representable LD matrix which as previously described in obtained from LDlink using the **LDlinkR** package. However the eQTL does not come with RSID information required by LDlink to get LD matrix, this will need to be obtained via the **BioMart** package.

To validate the commonality of causal variants between the eQTLs and GWAS signals, a colocalization analysis will be conducted. This analysis uses the **COLOC** R package, which is designed to estimate the probability of independent or shared causal variants between two distinct traits.

The expression profiles used in this analysis will be obtained from the eQTL catalogue which was downloaded in the previous section. The runsusie() and susie.coloc() functions which use functions from the **SuSiE** R package will be used in this step, as these functions allow for the assumption of multiple causal variants per locus, rather than restricting it to just one.

As with the earlier steps, this analysis requires a representative LD matrix, which is obtained from LDlink via the **LDlinkR** package. The results of this colocalization analysis will enhance our understanding of shared genetic underpinnings between gene expression and Asthma, and provide crucial insights for further drug target identification and validation.

In the provided example code, colocalization was examined between the previously identified locus on chromosome 17 and eQTL data for GSDMB in lung tissue. While the script is incomplete due to the inability to procure an LD matrix for eQTL data, the code structure remains insightful for understanding the overall process. This colocalization analysis should be repeated for each eQTL profile to validate shared causal variant between putative causal asthma variant and gene expression.

**Rscript:** 5.Coloc.R

**Input:** i) eQTL data (obtained in script), ii) GWAS Locus, iii) LD matrix for GWAS data (obtained in script), iv) LD matrix for eQTL data (obtained in script)

**Output:** i)data table detailing colocalization variants.

## Gene Set Enrichment Analysis (GSEA):

Upon validating the causal genes, a Gene Set Enrichment Analysis (GSEA) can be used to identify the potentially causal pathways or biological processes. This step enables a deeper understanding of the biological context in which these genes operate, providing valuable insights into the underlying mechanisms of Asthma which can provide further information to identify drug targets.

GSEA analysis will be performed in R using the **clusterProfiler** package using the predefined gene curated by Gene Ontology (GO) and and Kyoto Encyclopedia of Genes and Genomes (KEGG) databases. The insights derived from this analysis could enhance the selection of potential drug targets, for Asthma.

In the sample code provided, GSEA analysis was carried out on a list of genes recognized as co-expressed with ORMDL3, data for which was retrieved from the CoxpresDB database (<https://coxpresdb.jp/gene_coexpression/>). This step presumes that ORMDL3 was previously validated as a causal gene during the colocalization analysis stage.

**Rscript:** 6.GSEA\_analysis.R

**Input:** i) gene list

**Output:** i)data table of colocalization variants.

# 4. Identification of Candidate Drugs

The final step of this protocol is to identify potential druggable targets from the validated causal variants and biological pathways previously identified in the colocalization and GSEA analysis steps. This can be achieved using the DrugBank data base (<https://go.drugbank.com/>), where a list of drugs known to interact with them is returned.

Outputs from this analysis will inform potential drugs for further in vitro and in vivo testing, while clinical trials will be required to confirm these drugs' safety and efficacy.