# SCALLOP consortium - Analysis plan for INF panel proteins

***Adapted from CVD analysis plan***

## 1. Background

The SCAndinavian coLLaboration for Olink plasma Protein genetics (SCALLOP) consortium, https://www.olink.com/scallop/, is a collaborative framework for discovery and follow-up of genetic associations with proteins on the Olink Proteomics platform. A meta-analysis has been conducted on data Olink CVD panela from participating cohorts and consequent contributions have been made on Olink INF panel. This document therefore follows closely the analysis plan for the analysis, with highlight of relevant information to facilitate the meta-analysis.

## 2. Aims

As with the CVD meta-analysis, the tasks will involve

* Identification of pQTLs in SCALLOP discovery cohorts
* Replication of pQTLs in SCALLOP replication cohorts
* Investigation of the mechanistic basis of identified cis- and trans-pQTL by functional annotation
* Investigation of pleiotropic effects of the pQTLs
* Evaluation over the causal role of INF proteins disease outcomes such as CHD and stroke
* Other downstream analysis

among others.

## 3. Data analysis

* Use standard linear regression for assays with 80% of samples above the lower detection limit.
* Dichotomize proteins with more than 20% of samples below the lower detection limit and code values below the detection limit as 0 and those above as 1.
* Rank-based inverse normal transformation, e.g., invnormal function from https://github.com/jinghuazhao/R/tree/master/gap

### Proteins

The Olink INFlammation panel of 92 proteins, e.g, https://github.com/jinghuazhao/INF/blob/master/doc/olink.inf.panel.annot.tsv.

### SNPs

* 1000 genomes imputation
* SNPs will be filtered for imputation quality at time of meta-analysis, but please filter out SNPs with IMPUTE INFO quality less than 0.2
* Standard QC, including call rate < 95% or failed Illumina genotype calling, gender mismatch, abnormal inbreeding coefficient, failed cryptic relatedness test, ancestry outlier, sample call rate < 95%, Bonferroni corrected Hardy-Weinberg Equilibrium test.

### Association analysis

* Linear regression with adjustment for study-specific covariates. These should always include age at time of sample collection, gender and adjustment for population structure / geography if applicable (e.g across countries). Sample storage time and season of collection if applicable.
* Use imputation-dosages
* Additive genetic model
* Separate the analyses for men and women for X chromosome SNPs (exception for cohorts that have already performed analyses)

### Stratification

* Analyse patients and controls separately – results will be merged at meta-analysis stage

## 4. Descriptive statistics

Please fill out the attached descriptive statistics spreadsheet and use the naming convention:

* STUDY.descriptives.DATE.xls
* Where, STUDY is a short (14 characters or less) identifier for the population studied, which is the same for all files provided by your study.
* DATE is the date on which the file was prepared, in the format “YYYYMMDD”.

## 5. GWAS results submission and file formats

SNP table for association results. Please include the following columns. Missing values are coded as “NA”.

|  |  |  |
| --- | --- | --- |
| No | Variable name | Description |
| 1 | SNPID | SNP ID as rs number |
| 2 | CHR | Chromosome number (1-22) |
| 3 | POS | Physical position for the reference sequence (please indicate NCBI build in descriptive file) |
| 4 | STRAND | Indicator of strand direction. Please specify “+” if positive or forward strand and “-” if negative or reverse strand. |
| 5 | N | Number of non-missing observations |
| 6 | EFFECT\_ALLELE | Allele for which the effect (beta coefficient) is reported. For example, in an A/G SNP in which AA = 0, AG=1, and GG=2, the coded allele is G. |
| 7 | REFERENCE\_ALLELE | Second allele at the SNP (the other allele). In the example above, the non-coded allele is A. |
| 8 | CODE\_ALL\_FQ | Allele frequency for the coded allele – “NA” if not available |
| 9 | BETA | Effect size for the coded allele, beta estimate from the genotype-phenotype association, with at least 5 decimal places. Note: if not available, please report “NA” for this variable. |
| 10 | SE | Standard error of the beta estimate, to at least 5 decimal places - “NA” if not available. |
| 11 | PVAL | p-value of Wald test statistic – “NA” if not available |
| 12 | RSQ | Residual phenotypic variance explained by SNP. “NA” if not available |
| 13 | RSQ\_IMP | Observed divided by expected variance for imputed allele dosage. |
| 14 | IMP | Please specify whether the SNP was imputed or genotyped: 1: imputed SNP, 0: directly genotyped SNP |

### File-naming convention

It is recommended to use format STUDY\_inf1\_protein\_UnitProtID\_date.gz.

Note that due possibly to the large number of proteins for GWAS, some cohorts employed PLINK to expedite analysis in which case one sees the following information:

|  |  |  |  |
| --- | --- | --- | --- |
| No | Name | Description | Comment |
| 1 | BP | Position in basepairs |
| 2 | CHR | Chromosome |
| 3 | SNP | SNP name/chr:pos\_a1\_a2 |
| 4 | HWE | Hardy-Weinberg equilibrium P |
| 5 | MAF | Minor allele frequency |
| 6 | A1 | Allele 1 |
| 7 | A2 | Allele 2 |
| 8 | N | Sample size |
| 9 | BETA | Regression coefficient |
| 10 | CHI2 | Regression statistics |
| 11 | P | P value |

In this case, if is preferable to provide strand, effect allele, effect allele frequency, and the information measures.

## 6. Meta-analysis

Meta-analysis will be performed using the inverse-N weighted analysis of regression betas and standard errors, as implemented in the software METAL (https://github.com/statgen/METAL).

We will apply genomic control and the appropriate marker filters at this stage (i.e. please provide unfiltered results).

*. Marker exclusion filters: we will apply imputation quality filters at the meta-analysis stage. Please do not apply these filters yourself and provide unfiltered results.* . Genomic control (GC): genomic control will be applied to each study at the meta-analysis stage (single GC). Please do not apply GC to GWAS results and provide uncorrected standard errors, as (double) GC will be applied at the meta-analysis stage. \*. Significance: the threshold for the genome-wide analyses will be set at 5 x 10-8. The results will be replicated in independent cohorts so no need for additional correction.

If you have any questions, please contact Jing Hua Zhao via jhz22@medschl.cam.ac.uk or James Peters at jp549@medschl.cam.ac.uk.

## 7. Uploading of results data to TRYGGVE server

See CVD analysis plan.

Cambridge, 8/11/2018