# SCALLOP consortium analysis plan for INF panel proteins

***Adapted from SCALLOP/CVD1 analysis plan, last updated 6/12/2018***

Timeline for completing cohort-specific analyses and uploading the results for this project:



## 1. Overview

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 Olink CVD1 panel data from participating cohorts; consequent requests were sent and contributions made on the Olink INF panel. This document follows closely the SCALLOP/CVD1 analysis plan for the analysis, and in particular highlights relevant information required to facilitate the meta-analysis.

As with the CVD1 meta-analysis, the tasks will involve

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

## 2. Data and analysis

### 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, build 37 (hg19) positions.
* SNPs filtering on imputation quality at time of meta-analysis.
* Quality control on aspects such as SNP/sample call rates, gender mismatch, abnormal inbreeding coefficient, failed cryptic relatedness test, ancestry outlier, heterozygosity and Hardy-Weinberg equilibrium test.

### Association analysis

* Rank-based inverse normal transformation on the raw measurement of proteins including those below lower limit of detection, e.g., via invnormal function,

invnormal <- function(x)  
 qnorm((rank(x,na.last="keep")-0.5)/sum(!is.na(x)))

* Multiple linear regression for all samples including sex, age, principal components and other cohort specific covariates.
* Additive genetic model
* For case-control data, cases and controls are analysed separately – results will be merged at meta-analysis stage

### Software

It is preferable to use software which account for genotype uncertainty, such as SNPTEST, QUICKTEST, and BOLT-LMM.

## 3. Descriptive statistics

Please fill out the spreadsheet as with SCALLOP/CVD1 with 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 “DDMMYYYY”.

## 4. File formats for GWAS results

### SNP table for GWAS results

Please include the following columns. Missing values are coded as “NA”.

|  |  |  |
| --- | --- | --- |
| No | Variable name | Description of variable |
| 1 | SNPID | CHR:POS\_A1\_A2 (such that A1<A2) or rsid |
| 2 | CHR | Chromosome number |
| 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\_analyst\_inf1\_protein\_UniProtID\_date.gz, see https://www.uniprot.org/ for additional information on UniProt IDs.

### Notes on PLINK

Due possibly to the large number of proteins for GWAS, some cohorts employed PLINK to expedite analysis with which one may see the following information:

|  |  |  |  |
| --- | --- | --- | --- |
| No | Name | Description | Additional comment |
| 1 | BP | Position in base pairs |
| 2 | CHR | Chromosome |
| 3 | SNP | CHR:POS\_A1\_A2 or rsid |
| 4\* | HWE | Hardy-Weinberg equilibrium P-value |
| 5\* | MAF | Minor allele frequency | Please indicate if this is the effect allele frequency |
| 6 | A1 | Allele 1 | Please indicate if this is the effect/reference allele |
| 7\* | A2 | Allele 2 | Please indicate if this is the effect/reference allele |
| 8 | NMISS | Sample size |
| 9 | BETA | Regression coefficient |
| 10 | STAT | Regression test statistic |
| 11 | P | P value |

\* may be taken from the PLINK –hardy option and .bim file.

In this case, please provide for each SNP information on strand, effect allele, effect allele frequency, and the information measures for imputation – the information measure can be on the genotype level obtained once for a cohort rather than from phenotype-genotype regression through software such as SNPTEST. SNP and sample based statistics can be greatly facilitated with software qctool, http://www.well.ox.ac.uk/~gav/qctool\_v2/. As is the case with INTERVAL.bgen and INTERVAL.sample, one can obtain the SNP-based statistics as follows,

qctool -g INTERVAL.bgen -s INTERVAL.sample -snp-stats -osnp INTERVAL.snp-stats -sample-stats -osample INTERVAL.sample-stats -os INTERVAL.os

See also the full SLURM sbatch script in the appendix.

## 5. Meta-analysis

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

Genomic control and appropriate marker filters will be applied at this stage.

* **Marker exclusion filters**: we will apply imputation quality filters at the meta-analysis stage, so provide unfiltered results.
* **Genomic control (GC)**: genomic control will be applied to each study at the meta-analysis stage (single GC), so GC-correction is needed for each cohort.
* **Significance**: the Bonferroni threshold for the genome-wide analyses will be set at 5 x 10-10. The results will be replicated in independent cohorts.

## 6. Uploading of results

See the CVD1 analysis plan.

## 7. Contact information

For general questions about SCALLOP, please contact Anders Malarstig (anders.malarstig@ki.se). For technical issues about TRYGGVE, please contact Lasse Folkersen (lasfol@cbs.dtu.dk).

For questions regarding SCALLOP/INF, please contact Jing Hua Zhao (jhz22@medschl.cam.ac.uk) and James Peters (jp549@medschl.cam.ac.uk).

## Appendix

**SLURM script for qctool 2.0.1**

This is called with sbatch qctool.sb.

#!/bin/bash --login  
# 6-12-2018 JHZ  
  
#SBATCH -J qctool  
#SBATCH -o qctool.log  
#SBATCH -p long  
#SBATCH -t 4-0:0  
#SBATCH --export ALL  
#SBATCH --nodes=1  
#SBATCH --ntasks-per-node=1  
#SBATCH --cpus-per-task=8  
  
export DIR=/scratch/bp406/data\_sets/interval\_subset\_olink/genotype\_files/unrelated\_4994\_pihat\_0.1875\_autosomal\_typed\_only  
export INTERVAL=$DIR/interval\_olink\_subset\_unrelated\_4994\_pihat\_0.1875\_autosomal\_typed\_only  
ln -sf $INTERVAL.bgen INTERVAL.bgen  
ln -sf $INTERVAL.sample INTERVAL.sample  
  
# to obtain SNP-specific statistics as in .bgen and .sample format with qctool, tested with qctool 2.0.1  
  
qctool -g INTERVAL.bgen -s INTERVAL.sample -snp-stats -osnp INTERVAL.snp-stats -sample-stats -osample INTERVAL.sample-stats  
  
# Note in particular: the # option allows for chromosome-specific analysis; the -strand option will enable results in positive strand.