Analysis plan for sequencing-based meta-analysis of Olink panels

1. Background

We will carry out a meta-analysis of sequencing-based studies (WGS and WES) across large cross-sectional cohorts with Olink proteomics panels available.

We will perform single-point meta-analysis for each panel. Then, we will perform gene-based tests using summary-statistics-based rare-variant burden testing (SMMAT) on low-frequency and rare variants (MAF<0.05), using an established protocol which defines four variant selection and weighting schemes [Gilly et al., 2018; PMID: 30405126]. In particular, for cohorts where WGS is available, we will include non-coding variant annotation in order to account for regulatory effects in rare variant burdens.

The analysis will be complemented by extensive downstream bioinformatics analyses. This includes, but is not limited to, eQTL overlap, pathway analysis, pheWAS, two-sample Mendelian randomisation (MR), and drug target overlap.

2. Motivation and aims

Sequencing data provides us with access to understudied low-frequency and rare variants that genotype array-based studies do not have. This applies in both coding regions (WGS and WES) as well as intergenic and regulatory regions (WGS). By carrying out a pQTL meta-analysis and downstream bioinformatics analyses in sequenced cohorts, we aim to:

a. Provide a more complete picture of the genetic architecture of the human plasma proteome by characterising *cis*- and *trans*-acting pQTLs across the allele frequency spectrum;
b. Identify novel disease biomarkers for risk prediction, diagnosis, or as targets for drug development or repurposing;
c. Deliver novel insights into disease pathways by establishing links between and within the human genome, proteome, and phenome.

3. SCALLOP cohorts with WGS/WES and Olink data

Cohort	**Design**	**N**	**Sequencing**
HELIC MANOLIS	Population isolate	1457	WGS
HELIC Pomak	Population isolate	1611	WGS
INTERVAL	Population-based	3724	WGS/WES
ORCADES	Population-based	1000	WGS/WES
Estonian Biobank	Population-based	500	WGS
NSPHS	Population-based	1041	WGS
WHI	Postmenopausal women (sub-sample of two case groups for stroke and venous thromboembolism + controls)	1400	WGS
Total		**10,733**	

4. Proteomic assays

- All protein data should be included in the analysis, including values below the lower limit of detection (LOD). Note that these should not be missing or imputed, but measured values as provided by Olink;
- All proteins should be analysed, regardless of proportion of below-LOD values;
- Rank-based inverse normal transformation (INT) should be performed on NPX values
  e.g. in R:
  invnormal <- function(x) qnorm((rank(x,na.last="keep")-0.5)/sum(!is.na(x)))
- Covariates should be regressed out of the INT transformed NPX values and the residuals renormalised. Covariates should include sex, age, age-squared and average value per sample across all Olink measurements. Additional covariates, such as season, (x,y) coordinate on the plate or time of storage should be adjusted for in a study-specific manner if warranted by residual inflation.

5. Data Quality Control (QC)

Genomic positions for all study data should be aligned to the GRCh38/hg38 build of the human genome.

5.1 Variant QC

Appropriate variant QC will be performed by the individual study groups. Ideally, the gold standard pipeline from the Genome Analysis Toolkit (GATK), including Variant Quality Score Recalibration tool (VQSR), should be applied on the data.

Variants should not be filtered for minor allele count (MAC) or frequency (MAF) unless a low depth warrants it. Singletons are considered reliable for depth equal to or above 15x.

5.2 Sample QC

Appropriate sample QC will be performed by individual study groups according to study-specific criteria. Indicators to consider for sample exclusion include:

- Duplication, relatedness and sample swap metrics;
- Sex mismatches between genetic and phenotypic data;
- Average sequencing depth;
- Heterozygosity rate;
- Population outliers via dimensionality reduction techniques;
- Per-sample missingness rate;
- Sample contamination;
- Where data from other genotyping methods are available, genotype concordance.

6. Single variant-based association

- We suggest the usage of a linear mixed model that accounts for population structure;
- The GRM should be tailored in a study-specific way so as to reduce genomic inflation;
- No genomic control (lambda) correction should be applied to the summary statistics.

Note: The rare variant analyses require the input of a genetic relatedness matrix (GRM). We therefore advise analysts to compute a GRM using a reduced set of independent variants at this stage. Example variant sets are: LD-pruned, Hardy-Weinberg filtered variants above 1% MAF, or pre-determined sets such as those provided by AKT (https://github.com/Illumina/akt). Exact criteria for variant selection will vary depending on the level of relatedness present in the cohorts. We suggest benchmarking the GRM calculation against a quantitative trait single-point analysis, and checking for genomic inflation.

7. Rare variant analysis (RVA)

Gene-based burden testing will be carried out using the GMMAT/SMMAT R package (vignette; Chen et al., 2019), which provides an interface to unified optimal tests (SKAT-O) and allows to specify variant weights that reflect predicted consequence on gene function. The RVA pipeline will be split into three different phases, as detailed in subsections 7.2-7.4 below:

1. Generation of SNP info files ("group file" in SMMAT);
2. Study-level analysis to generate score statistics and between-variant covariance matrices;
3. Meta-analysis using above generated files;

Input will be required from participating study groups for phases 1 and 2. To maximise detection power, exonic data will be extracted from both WES and WGS cohorts (should any WES-only cohorts participate), whereas WGS will provide additional information for the Exonic & Regulatory and Regulatory Only analyses. The table below describes the different combinations of regions, variant filters, and weighting schemes that will be used.

Burden analysis condition	**Regions**	**Variant filtering criteria**	**Weighting scheme**
Exon severe	Exons only	Ensembl most severe consequence more severe than 'missense'	none
Exon CADD xtend	Exons extended by 50bp	none	CADD
Exon+Regulatory EigenPhred	Exons extended by 50bp & Regulatory	none	Phred-transformed Eigen
Regulatory only EigenPhred	Regulatory only	none	Phred-transformed Eigen

7.1 Statistical power

Statistical power is an important study design consideration in association studies. While closed formulae exist for computing power in single-variant association tests, calculating the power of rare variant tests is more challenging, especially for optimal or hybrid tests such as the one implemented in SMMAT. Nonetheless, the statistical power of our approach has been empirically demonstrated in a study with similar sample size, both for biochemical [Gilly et al., 2018; PMID: 30405126] and proteomic [Gilly et al., 2019; bioRXiv] traits. We expect that the study of a large, diverse cohort in a meta-analysis setting will further improve discovery power through the inclusion of population-specific rare haplotypes.

7.2 Phase 1: Generation of SMMAT group file

The SMMAT group file is a tab-delimited file containing six columns (gene name, chromosome, position, reference allele, alternate allele, weight). As SMMAT requires the same group file to be used across all studies, each cohort will be required to provide a list of all non-monomorphic variants in order to generate common group files. In this phase, one group file per filtering/weighting scheme containing all relevant SNPs from each participating study will be generated for use in Phase 2.

The list of variants from each study should contain the following columns, which can be easily extracted from a filtered VCF file:

Column no.	**Column name**	**Description**
1	CHR	Chromosome number (1-22)
2	POS	Physical base-pair position on the chromosome (in b38 coordinates)
3	REF	Reference allele
4	ALT	Alternate allele
5	AC	Allele count in genotypes for each ALT allele
6	AN	Total number of alleles in called genotypes

These files will then be annotated by the analysis team with the following information:

- variant consequence using Ensembl VEP
- CADD scores
- Eigen scores

7.3 Phase 2: Study-level analysis to generate score statistics and covariance matrices

The objective of this phase is to generate input files for the meta-analysis. To facilitate the rare variant burden testing, we plan to package Phase 2 into a Singularity container that should be deployable on any server without any software installation besides Singularity itself. In this phase, two types of files will be generated:

- A space-delimited file containing single variant scores
- A binary file containing between-variant covariance matrices

The following input files will be required:

- Filtered VCF file (see above section on Variant QC) OR GDS files if they already exist
- Phenotype files containing two columns: **sample ID** and **INT-transformed, covariate-adjusted and renormalised residuals** without header
- Genetic relatedness matrix (GRM) with sample IDs as row and column names

7.4 Phase 3: Meta-analysis

Individual cohorts will upload the score statistics and covariance matrices without any further QC for meta-analysis using SMMATmeta on our servers.

8. Results file formats & naming

8.1 Single point association

File naming

- Each file should contain the summary statistics of one protein per cohort
- Files should be bgzipped, tabixed and named accordingly:
  <olink_protein>_<cohort>_<date_of_analysis>_<analyst_initials>.txt.bgz
  e.g. ACE2_pooled_MANOLIS_28102019_GP.txt.bgz

File format

Column no.	Column name	Description
1	SNP	rsID, or NA if unavailable
2	CHR	Chromosome number
3	POS	Physical base pair position on the chromosome (in b38 coordinates)
4	N	Number of non-missing observations
5	EFF_ALLELE	Allele whose effect is reported (beta estimates)
6	OTHER_ALLELE	The other allele at the SNP
7	EFF_ALLELE_FREQ	Allele frequency of EFF_ALLELE
8	BETA	Effect size estimate, to at least 5 d.p.
9	SE	Standard error of the beta estimates, to at least 5 d.p.
10	P_LRT	P-value LRT
11	P_SCORE	P-value score

8.2 Rare variant analysis

As mentioned in 7.4, score statistics and covariance matrix files can be uploaded without further QC for meta-analysis using SMMATmeta.

Upload of summary statistics and contact information

Grace Png: grace.png@helmholtz-muenchen.de
Arthur Gilly: arthur.gilly@helmholtz-muenchen.de