

SCALLOP sequencing

Analysis plan

28 Sep 2020

Overview

- Meta-analysis of sequencing-based studies (both WGS and WES)
- All panels, starting with CVD panels
- Single variant-based association meta-analysis
- Gene-based rare burden testing using SMMAT (MAF<5%)
- Complemented by downstream bioinformatics
 - eQTL colocalisation, pheWAS, MR, drug target evaluation

Cohorts

- So far 6 cohorts with WGS, **N~10,733**

Cohort	Design	N	Sequencing
HELIC MANOLIS	Population-based	1457	WGS
HELIC Pomak	Population-based	1611	WGS
INTERVAL	Population-based	3724 (WGS)	WGS/WES
ORCADES	Population-based	1000 (WGS)	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	

Proteomic assays

- Include all measurements, including those below the limit of detection (LOD)
- Include all proteins regardless of proportion of <LOD values
- Olink NPX values should be rank-based inverse normal transformed (INT), with covariates regressed out and then renormalised
 - Covariates including: sex, age, age-squared, average NPX value per sample across all Olink measurements, additional study-specific covariates

Data QC

- Variant QC
 - Ideally gold standard pipeline from GATK, including Variant Quality Score Recalibration (VQSR)
 - Not filtered for minor allele count or frequency (MAC/MAF) unless low depth warrants it
- Sample QC
 - According to study-specific criteria

- 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.

Single variant-based association

- Usage of linear mixed model that accounts for population structure (e.g. GEMMA)
- Usage of genetic relatedness matrix (GRM)
 - Also required for rare variant analyses

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.

Rare variant analysis

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

Rare variant analysis – GMMAT/SMMAT

- Rationale behind choosing SMMAT:
 - Not P value-based
 - Allows for specification of variant weights
 - Accounts for relatedness (works with a GRM)
 - Implements unified optimal tests (SKAT-O)
 - Optimises rho automatically
 - Computationally efficient

- Alternatives were: RAREMETAL, MASS, MetaSKAT, rvtests, but all of them are missing one or more of the above criteria


AJHG

Volume 104, Issue 2, 7 February 2019, Pages 260-274



Article

Efficient Variant Set Mixed Model Association Tests for Continuous and Binary Traits in Large-Scale Whole-Genome Sequencing Studies

Han Chen ^{1, 2}, Jennifer E. Huffman ³, Jennifer A. Brody ⁴, Chaolong Wang ⁵, Seunggeun Lee ⁶, Zilin Li ⁷, Stephanie M. Gogarten ⁸, Tamar Sofer ^{9, 10}, Lawrence F. Bielak ¹¹, Joshua C. Bis ⁴, John Blangero ¹², Russell P. Bowler ¹³, Brian E. Cade ^{9, 10}, Michael H. Cho ^{14, 15}, Adolfo Correa ¹⁶, Joanne E. Curran ¹², Paul S. de Vries ¹, David C. Glahn ^{17, 18} ... Xihong Lin ^{7, 37} 

[Show more](#) 

<https://doi.org/10.1016/j.ajhg.2018.12.012>

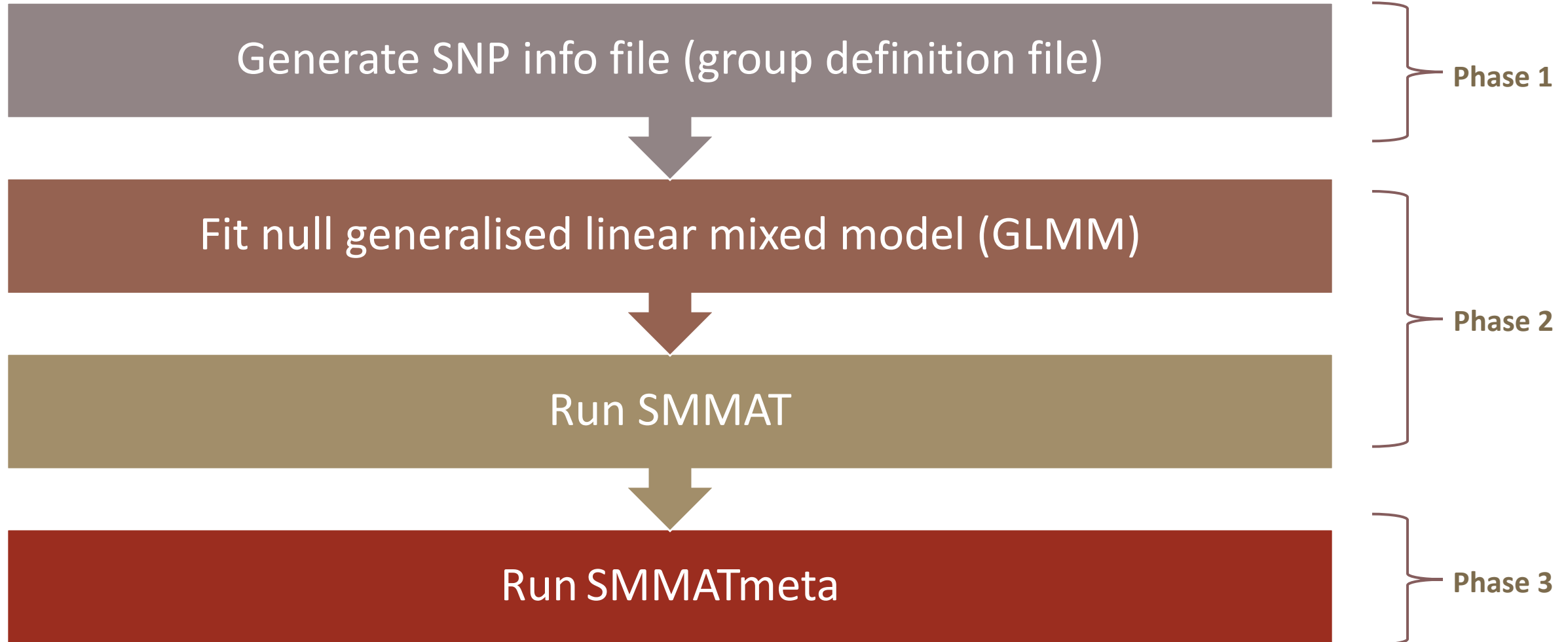
Under an Elsevier user license

[Get rights and content](#)

[open archive](#)

Rare variant analysis – SMMAT

Phase 1 and 2 to be implemented by individual cohorts, Phase 3 to be carried out centrally



RVA – Phase 1: Generation of SMMAT group file

- Tab-delimited file containing SNP info:
 - Gene name
 - Chr
 - Position
 - Ref allele
 - Alt allele
 - Weight (CADD/EigenPhred score)
- SMMAT requires that the same group files be used across all studies, meaning that the group files will be generated centrally
- Each cohort will have to provide a VCF file, which will be annotated by the analysis team with the variant consequences and CADD/EigenPhred scores
- Total of **four** group files to be generated, one for each condition, and redistributed to participating teams for Phase 2

Phase 2: Study-level analysis

- Objective: generation of single variant score statistics and covariance matrices required for SMMATmeta (Phase 3)
- The plan is to package Phase 2 into a Singularity container that should be deployable on any server
- Each cohort will have to prepare the following files:
 - Filtered VCF or GDS files
 - Phenotype files (INT-transformed, covariate adjusted, renormalised)
 - GRM

Phase 3: Meta-analysis

- Output files from Phase 2 can be uploaded without further QC for meta-analysis using SMMAT.meta, which will be carried out on our servers

```
> SMMAT.meta(meta.files.prefix = "SMMAT.meta", n.files = 1,  
+           group.file = group.file, MAF.range = c(1e-7, 0.5),  
+           miss.cutoff = 1, method = "davies", tests = "S")
```

https://github.com/hmgu-itg/burden_testing

grace.png@helmholtz-muenchen.de

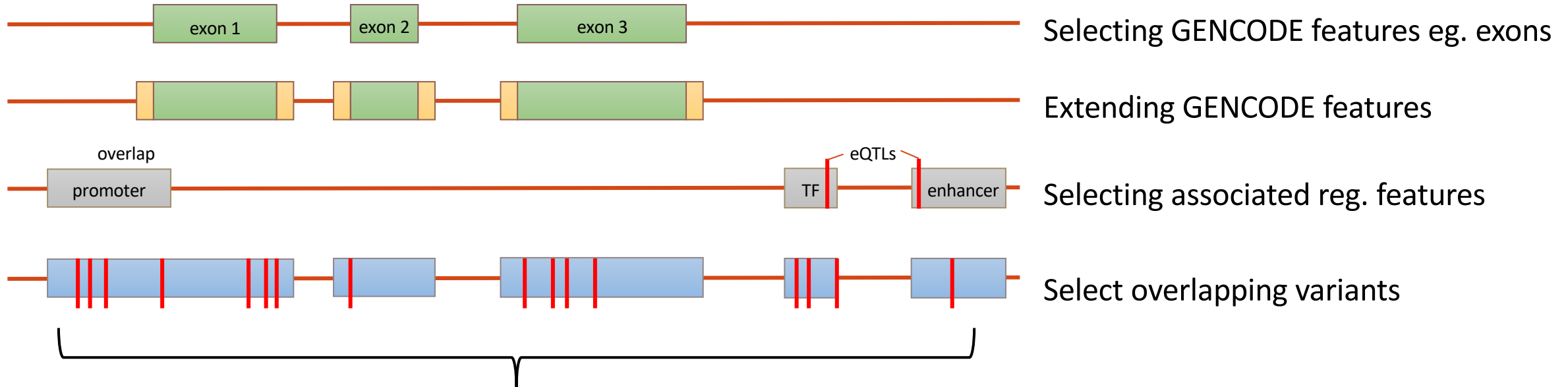
arthur.gilly@helmholtz-muenchen.de

Additional

SMMATmeta output example for 1 condition

group	n.variants	B.score	B.var	B.pval	S.pval	O.pval	O.minp	O.minp.rho	E.pval
SAMD11	77	115.3952	81497114	0.9898013	0.3939582	0.5979449	0.3939582	0.00	0.6603812
NOC2L	131	7860.6246	35597523	0.1876746	0.1059472	0.1894530	0.1043120	0.04	0.1767559
KLHL17	73	-6423.9864	58699428	0.4017666	0.3498692	0.5425452	0.3498692	0.00	0.4653773
PLEKHN1	74	-8669.3533	40691411	0.1741309	0.9244274	0.3008549	0.1741309	1.00	0.4779731
PERM1	56	-5063.7327	24131583	0.3026308	0.6295145	0.4748405	0.3026308	1.00	0.6248257
HES4	14	3847.0927	18920952	0.3764667	0.5083520	0.5620503	0.3764667	1.00	0.3736780

Selecting variants for burden testing:



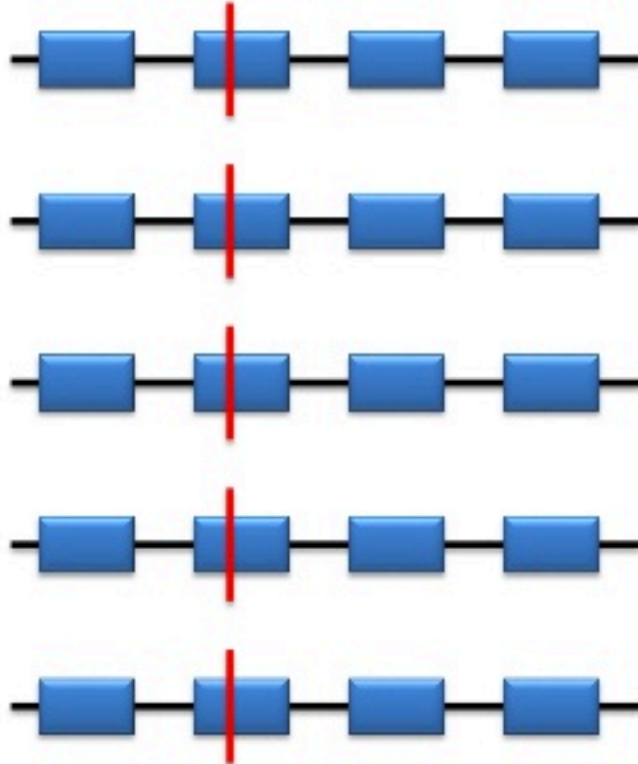
Pooling variants and apply further filters:

- Upper MAF threshold: 5%
- Upper missingness threshold: 1% (remaining missingness is imputed)
- Optional selection of variants based on predicted consequence

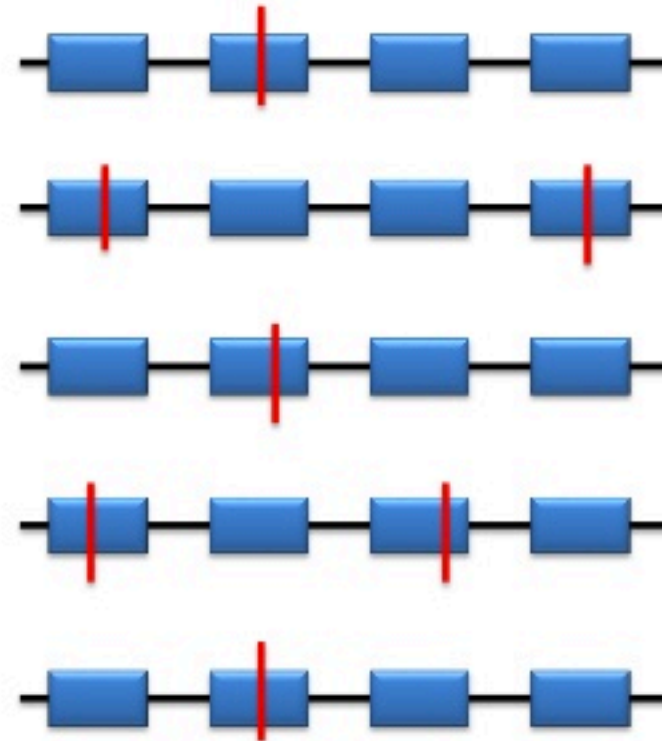
Linear Regression vs Burden Analysis

Typical Gene — exon — exon — exon — exon —

Common Single Variant:Phenotype



Rare Variant Gene:Phenotype



Singe-variant scores (S)

- G^T = n x q genotype matrix of the variant set
- y = n x 1 vector of phenotype values
- $\hat{\mu}_0$ = vector of fitted mean values under the GLMM
- $\hat{\phi}$ = dispersion parameter (residual variance)

$$\mathbf{S} = \mathbf{G}^T (\mathbf{y} - \hat{\mu}_0) / \hat{\phi}$$