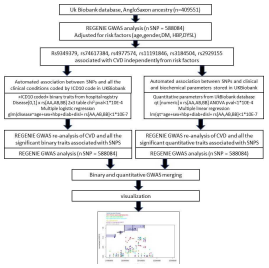
Nov 05, 2024

A new method to refine GWAS results based on the UKBiobank phenotype database



DOI
dx.doi.org/10.17504/protocols.io.ewov1o5nklr2/v1

Davide Noto[1]

[1]University of Palermo

Davide Noto
University of Palermo

OPEN ACCESS

DOI: dx.doi.org/10.17504/protocols.io.ewov1o5nklr2/v1

Protocol Citation: Davide Noto 2024. A new method to refine GWAS results based on the UKBiobank phenotype database. **protocols.io https://dx.doi.org/10.17504/protocols.io.ewov1o5nklr2/v1**

Protocol status: Working
We use this protocol and it's working

Created: January 31, 2023

Last Modified: November 05, 2024

Protocol Integer ID: 76145

Keywords: UKBiobank, GWAS, REGENIE, R

Disclaimer

All the necessary files are contained in the atteched ZIP file. For all the necessary files, routines, related to UK Biobank and RAP contact the UKBiobank organization for details.

Abstract

Genome wide association studies (GWAS) is an untargeted methodology able to identify novel gene variants associated with diseases. Sometimes the gene variants identified by GWAS are located within genes whose function is unknown or not related to the investigated trait. This paper describes a novel methodology based on GWAS filtering, aimed to find novel phenotypes associated to genetic loci that may determine cardiovascular disease (CVD), not conventionally listed among CVD canonical risk factors. UKBiobank, the largest clinical, laboratory, instrumental and genetic phenotypes database, was interrogated by an automated routine. Six gene variants associated with CVD, independently of canonical risk factors, were identified using a variants database of more than 400k genotyped subjects, and many novel clinical and biochemical phenotypes have been associated to the variants. The phenotypical characterisation of the loci resolved some ambiguities regarding gene loci lacking a clear CVD-associated identification.

Attachments

ZIP

UK_Biobank_phenotypi..

15.3MB

Safety warnings

There is no plan to upgrade the R routines in the future. Please check DNAnexus documentation for database creation and manipulation, REGENIE instruction for the RAP platform, and all other workflows within RAP.

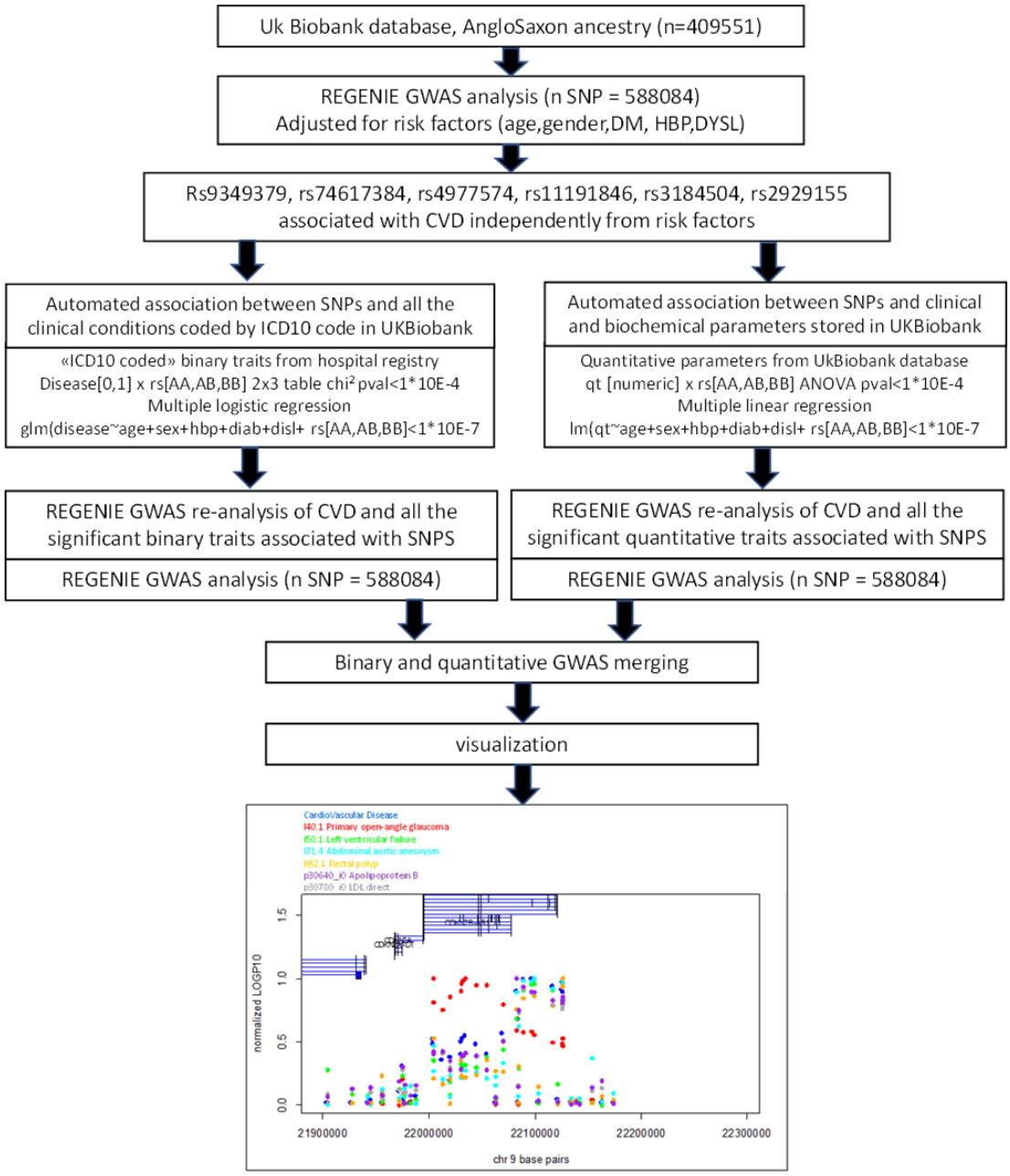
1 **Description of the Protocol**

This protocol describes an automated workflow able to enrich the results of a Genome Wide Association Study (GWAS) obtained from the UK Biobank data. In particular the workflow uses the Single Nucleotide Polymorphisms (SNP) that resulted associated with the investigated trait (Cardiovascular disease in the current example) and search for further associations with the hundreds of phenotypic variables stored in theUK Biobank Database.Phenotypic association are evaluated by an automated set of multiple regression analyses (for numerical variables) and multiple logistic analyses (for binary variables).
All the procedures, from GWAS to multivariate analyses, are adjusted for canonical risk factors (of CVD in our example) in order to let emerge novel loci associated to the trait (CVD) independent from canonical risk factor. The association of other phenotypes to the SNP has some advantages:

1.1 i)If the SNP is located in proximity of a gene with unknown function, or with functions not directly related to the investigated trait, the enrichment of the phenotype could supply clues of how the gene is linked to the trait.

1.2 ii)If the locus contains different genes with different functions, the colocalization of the investigated trait with other phenotypes could help to discriminate which gene could be responsible for the association with the trait, if some associated phenotype points to a gene with known function.

The Figure shows the proposed protocol in graphical format.



The figure shows the diagram flow of the proposed study workflow. The first GWAS identified six loci associated with CVD independently from canonical risk factors. For every locus, the representative SNP were selected considering the SNP with highest LOD score after pruning of the variables in Linkage Disequilibrium. The six SNP were then used as predictors in a series of multiple logistic regression analyses using the clinical conditions stored as ICD10 codes in the UK Biobank database as dependent variables, while a set of multiple regression analyses were used to deal with numerical (clinical and Biochemical) parameters stored in the database. The clinical conditions and the numerical parameters associated with the SNPs with p-values < 1E07 were then subjected to a new round of GWAS. The results were then merged and the colocalizations of the CVD GWAS with the novel binary/quantitative trait loci were investigated by a merged plot of the relative LOD scores, normalized between 0 and 1 for comparison purposes.

2 STEP2: GWAS analysis

2.1 The first step is the preparation of the database for the GWAS analysis of UKBiobank. The following procedures describe the workflow in the Research Analysis Platform (RAP) online platform from DNAnexus. (https://ukbiobank.dnanexus.com/landing).

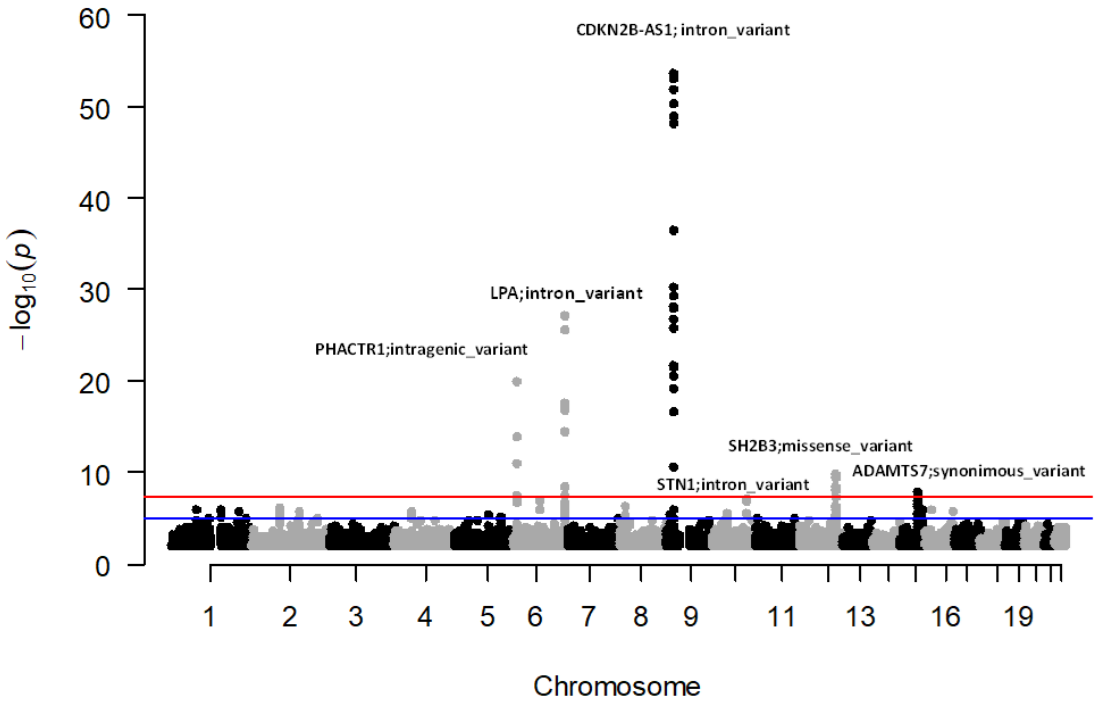
2.2 The second step is the creation of the cohorts of cases and controls for the analysis. This procedure is realized with the Cohort Browser application of RAP. Check the DNAnexus documentation for details (https://documentation.dnanexus.com/user/cohort-browser).

2.3 **The third step is the preparation of the phenotype file. Again check the documentation. (https://dnanexus.gitbook.io/uk-biobank-rap/working-on-the-research-analysis-platform/ukb-rap)**
In particular, extraction of the traits and other database manipulation can be executed using the JupiterNotebook pipeline within RAP. Explanation can be found in youtube DNAnexus channels (https://www.youtube.com/watch?v=762PVlyZJ-U&ab_channel=DNAnexus).

2.4 **The final product of the REGENIE pipeline is the result file with extension ".regenie", (CVD_out_firth_pheno_CVD.regenie in our example). The structure is:**

CHROM	GENPOS	ID	ALLELE0	ALLELE1	A1FREQ	N	TEST	BETA	SE	CHISQ	LOG10P	EXTRA
1	756604	rs3131962	G	A	0.129851	274753	ADD	0.002367	0.013297	0.031682	0.066145	NA
1	768448	rs1256203	G	A	0.10497	274779	ADD	-0.00533	0.01454	0.134281	0.146281	NA
1	779322	rs4040617	A	G	0.127709	274538	ADD	-0.00059	0.013393	0.001953	0.015583	NA
1	801536	rs7937392	T	G	0.014963	275110	ADD	-0.03947	0.036597	1.16313	0.551577	NA
1	808631	rs1124077	A	G	0.225196	273069	ADD	-0.00094	0.010754	0.007607	0.031286	NA
1	809876	rs5718170	A	G	0.100134	275032	ADD	0.001533	0.014815	0.010713	0.037365	NA
1	835499	rs4422948	A	G	0.240068	269946	ADD	0.004136	0.010547	0.153771	0.158042	NA
1	838555	rs4970383	C	A	0.245642	274652	ADD	0.008042	0.010359	0.602654	0.358955	NA
1	840753	rs4970382	T	C	0.399942	274687	ADD	0.003936	0.009111	0.186622	0.176693	NA
1	846864	rs950122	G	C	0.198748	273811	ADD	0.009248	0.011217	0.679741	0.38756	NA
1	849998	rs1330322	G	A	0.178502	274798	ADD	-0.00047	0.011652	0.001632	0.014225	NA
1	850780	rs6657440	T	C	0.396477	274837	ADD	0.006942	0.00913	0.578046	0.349615	NA
1	851390	rs7263188	G	T	0.041607	274739	ADD	-0.00181	0.022282	0.006587	0.029044	NA
1	858051	rs4970459	C	T	0.220076	274775	ADD	0.01021	0.010782	0.89683	0.463905	NA

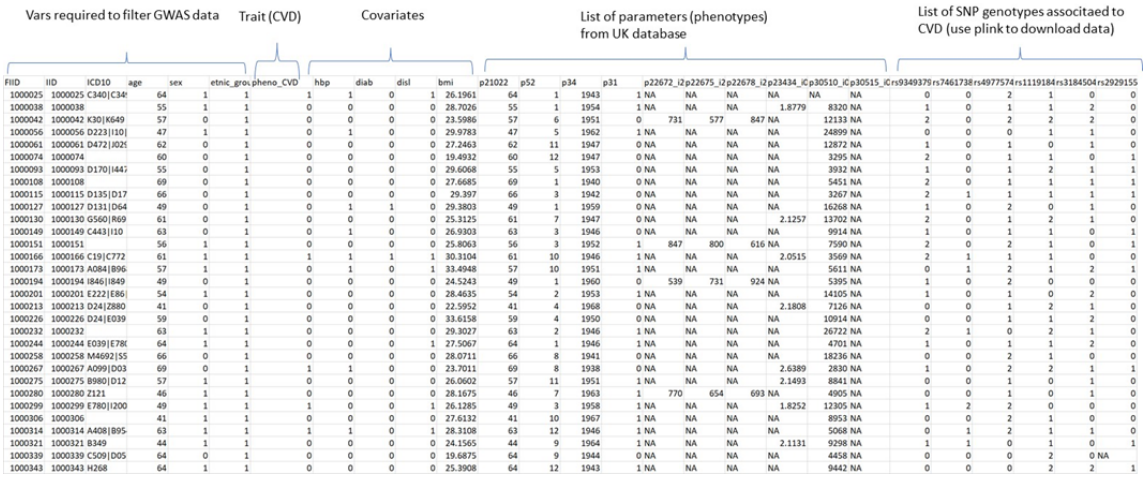
The result con be plotted with the R “qqman”package. The code to plot our results is in the “plot GWAS_base.R” script.



3 STEP3: Phenotypes association to SNPs (representative of GWAS loci)

All the data can be downloaded by the “Table exporter” app from RAP. SNP genotypes in (0,1,2 minor allele format can be downloaded by plink. Data can be merged within R with the “merge” function using the “IID” column as index. Data are reduced in width. The list of the phenotypes is usually from 100 to 1000 column long

3.1 The main database needs to be arranged to be analyzed by the R script. The final structure should be:



FID	IID	ICD10	age	sex	etnic_gro	pheno_CVD	hbp	diab	disl	bmi	p21022	p52	p34	p31	p22672_i2	p22675_i2	p22678_i2	p23434_iC	p30510_iC	p30515_iC	rs9349379	rs7461738	rs4977574	rs1119184	rs3184504	rs2929155
1000025	1000025	C340|C34	64	1	1	1	1	0	1	26.1961	64	1	1943	1	NA	NA	NA	NA	NA	NA	0	0	2	1	0	0
1000038	1000038		55	1	1	0	0	0	0	28.7026	55	1	1954	1	NA	NA	NA	1.8779	8320	NA	1	0	1	1	2	0
1000042	1000042	K30|K649	57	0	1	0	0	0	0	23.5986	57	6	1951	0	731	577	847	NA	12133	NA	2	0	2	2	2	0
1000056	1000056	D223|I10|	47	1	1	0	1	0	0	29.9783	47	5	1962	1	NA	NA	NA	NA	24899	NA	0	0	0	1	1	0
1000061	1000061	D472|J02|	62	0	1	0	0	0	0	27.2463	62	11	1947	0	NA	NA	NA	NA	12872	NA	1	0	1	0	1	0
1000074	1000074		60	0	1	0	0	0	0	19.4932	60	12	1947	0	NA	NA	NA	NA	3295	NA	2	0	1	1	0	1
1000093	1000093	D170|I447	55	0	1	0	0	0	0	29.6068	55	5	1953	0	NA	NA	NA	NA	3932	NA	1	0	1	2	1	1
1000108	1000108		69	0	1	0	0	0	0	27.6685	69	1	1940	0	NA	NA	NA	NA	5451	NA	2	0	1	1	1	1
1000115	1000115	D135|D17	66	0	1	0	0	0	0	29.397	66	3	1942	0	NA	NA	NA	NA	3267	NA	2	1	1	1	1	1
1000127	1000127	D131|D64	49	0	1	0	1	1	0	29.3803	49	1	1959	0	NA	NA	NA	NA	16268	NA	1	0	2	0	1	0
1000130	1000130	G560|R69	61	0	1	0	0	0	0	25.3125	61	7	1947	0	NA	NA	NA	2.1257	13702	NA	2	0	1	2	1	0
1000149	1000149	C443|I10	63	0	1	0	1	0	0	26.9303	63	3	1946	0	NA	NA	NA	NA	9914	NA	1	0	1	1	0	1
1000151	1000151		56	1	1	0	0	0	0	25.8063	56	3	1952	1	847	800	616	NA	7590	NA	2	0	2	1	0	1
1000166	1000166	C19|C772	61	1	1	1	1	1	1	30.3104	61	10	1946	1	NA	NA	NA	2.0515	3569	NA	2	1	1	1	0	0
1000173	1000173	A084|B96	57	1	1	0	1	0	1	33.4948	57	10	1951	1	NA	NA	NA	NA	5611	NA	0	1	2	1	2	1
1000194	1000194	I846|I849	49	0	1	0	0	0	0	24.5243	49	1	1960	0	539	731	924	NA	5395	NA	1	0	2	0	0	0
1000201	1000201	E222|E86	54	1	1	0	0	0	0	28.4635	54	2	1953	1	NA	NA	NA	NA	14105	NA	1	0	1	0	2	0
1000213	1000213	D24|Z880	41	0	1	0	0	0	0	22.5952	41	4	1968	0	NA	NA	NA	2.1808	7126	NA	0	0	1	2	1	0
1000226	1000226	D24|E039	59	0	1	0	0	0	0	33.6158	59	4	1950	0	NA	NA	NA	NA	10914	NA	0	0	1	1	2	0
1000232	1000232		63	1	1	0	0	0	0	29.3027	63	2	1946	1	NA	NA	NA	NA	26722	NA	2	1	0	2	1	0
1000244	1000244	E039|E78(	64	1	1	0	0	0	1	27.5067	64	1	1946	1	NA	NA	NA	NA	4701	NA	1	0	1	1	2	0
1000258	1000258	M4692|S5	66	0	1	0	0	0	0	28.0711	66	8	1941	0	NA	NA	NA	NA	18236	NA	0	0	2	1	0	0
1000267	1000267	A099|D03	69	0	1	1	1	0	0	23.7011	69	8	1938	0	NA	NA	NA	2.6389	2830	NA	1	0	2	2	1	1
1000275	1000275	B980|D12	57	1	1	0	0	0	0	26.0602	57	11	1951	1	NA	NA	NA	2.1493	8841	NA	0	0	1	0	1	0
1000280	1000280	Z121	46	1	1	0	0	0	0	28.1675	46	7	1963	1	770	654	693	NA	4905	NA	0	0	1	0	1	0
1000299	1000299	E780|I200	49	1	1	1	0	0	1	26.1285	49	3	1958	1	NA	NA	NA	1.8252	12305	NA	1	2	2	0	0	0
1000306	1000306		41	1	1	0	0	0	0	27.6132	41	10	1967	1	NA	NA	NA	NA	8953	NA	0	0	2	1	0	0
1000314	1000314	A408|B95	63	1	1	1	1	0	1	28.3108	63	12	1946	1	NA	NA	NA	NA	5068	NA	0	1	2	1	1	0
1000321	1000321	B349	44	1	1	0	0	0	0	24.1565	44	9	1964	1	NA	NA	NA	2.1131	9298	NA	1	1	0	1	0	1
1000339	1000339	C509|D05	64	0	1	0	0	0	0	19.6875	64	9	1944	0	NA	NA	NA	NA	4458	NA	0	0	0	2	0	NA
1000343	1000343	H268	64	1	1	0	0	0	0	25.3908	64	12	1943	1	NA	NA	NA	NA	9442	NA	0	0	0	2	2	1

3.2 Once ready the file is analyzed by the main R script **“UK_Biobank_phenotypizer_ ver2.0.R”**. The script checks if the all the supporting files are present in the working directory. Then, the first analysis is executed, and the association of diseases stored in ICD10 column with the SNP is evaluated for every ICD10 term. This ICD10 column corresponds to the p41202 field of the UK Biobank database, and it stores all information in a single string format, whose elements are separated by a “|” sign. Example:

IID	FID.x	ICD10	age	sex	genetic_se
1000017	1000017	E780|I10|I200|I251|K210|K409|M720|T810|Z82	54	1	1
1000025	1000025	C340|C349|E780|E835|E876|F171|I10|I251|I252	64	1	1
1000038	1000038		55	1	1
1000042	1000042	K30|K649	57	0	0
1000056	1000056	D223|I10|Z121	47	1	1
1000061	1000061	D472|J029|K30|M1300|N840|N950|S430|Z530	62	0	0
1000074	1000074		60	0	0
1000089	1000089	K768|L299|R932	46	0	0
1000093	1000093	D170|I447|J330|J339|J459|K029|R073|T818|Z53	55	0	0

The function **"ICDxSNP"** ask for some information about the column position of the trait, covariate, phenotypes and SNP and start to calculate the association of phenotypes to SNP adjusting for covariates. The routine extracts every ICD term detected in more than a x number of subjects, where x is a user defined threshold, and . Then the routine performs a series of chi^2 test on a 6 cells table (ICD10 yes/no) x (SNP 0,1,2) . If the chi^2 testp-value is < 10E-04 it proceeds to a logistic regression with format:
The it perform a series of logistic regression with format:

glm[ICD10 (0,1)~ covariates + SNP rs(0,1,2)].

If the glm p-value is < 10E-07, it stores the prevalence of the ICD10 condition according to the SNP (0,1,2) , the odds ratios and confidence intervals.

The function **"NUMPARxSNP"** ask for some information about the column position of the trait, covariate, phenotypes and SNP and start to calculate the association of numeric phenotypes to SNP adjusting for covariates. The routine performs a series of ANOVAtest on a numerical variables x (SNP 0,1,2) . If the ANOVAtestp-value is < 10E-04 it proceeds to a series of multiple regression with format:

lm[numeric_parameters~ covariates + SNP rs(0,1,2)].

If the glm p-value is < 10E-07, it stores mean ± SD according to the SNP, multiple regression coefficient beta coefficient and multiple regression p-value.

4 STEP4 : REGENIE analysis of ICD10 conditions and numerical parameters obtained in STEP3.

A new set of REGENIE analyses is performed on the variables resulted in STEP3.
The aim of this step is to refine and enrich the loci found in STEP1 by adding new phenotypes. Then REGENIE is set to work with single chromosomes, those containing the loci associated with the main trait (CVD in out example).

4.1 For binary phenotypes (as ICD10 condition) REGENIE option is set to “bt”:
Example for chrom 12.

```
regenie \
--step 1 \
--bed ukb22418_c12_b0_v2 \                #chrom 12
--extract snps_qc_pass.snplist \
--covarFile phenofile_to_refine_eg1.txt \
--phenoFile phenofile_to_refine_eg1.txt \
--phenoCol pheno_CVD \                    # main trait
--phenoCol E039 \                         # ICD10 condition
--phenoCol K900 \                         # ICD10 condition
--covarCol age --covarCol sex --covarCol diab --covarCol hbp --covarCol disl \
--bsize 750 \
--bt --lowmem --loocv \                   # bt for binary trait
--lowmem-prefix tmp_rg \
--out CVD_fit_bin_out_refine_c12          #result file

regenie \
--step 2 \
--bed ukb22418_c12_b0_v2 \
--extract snps_qc_pass.snplist \
--covarFile phenofile_to_refine_eg1.txt \
--phenoFile phenofile_to_refine_eg1.txt \
--phenoCol pheno_CVD \
--phenoCol E039 \
--phenoCol K900 \
--covarCol age --covarCol sex --covarCol diab --covarCol hbp --covarCol disl \
--bsize 750 \
--bt \
--strict \
--firth --approx \
--pThresh 0.01 \
--pred CVD_fit_bin_out_refine_c12_pred.list \
--out test_bin_out_firth_c12_refine
```

4.2 For numeric phenotypes (as p30010_i0, plasma Albumin concentration) REGENIE option is set to “qt”:
Example for chrom 10.

```
regenie \
--step 1 \
```

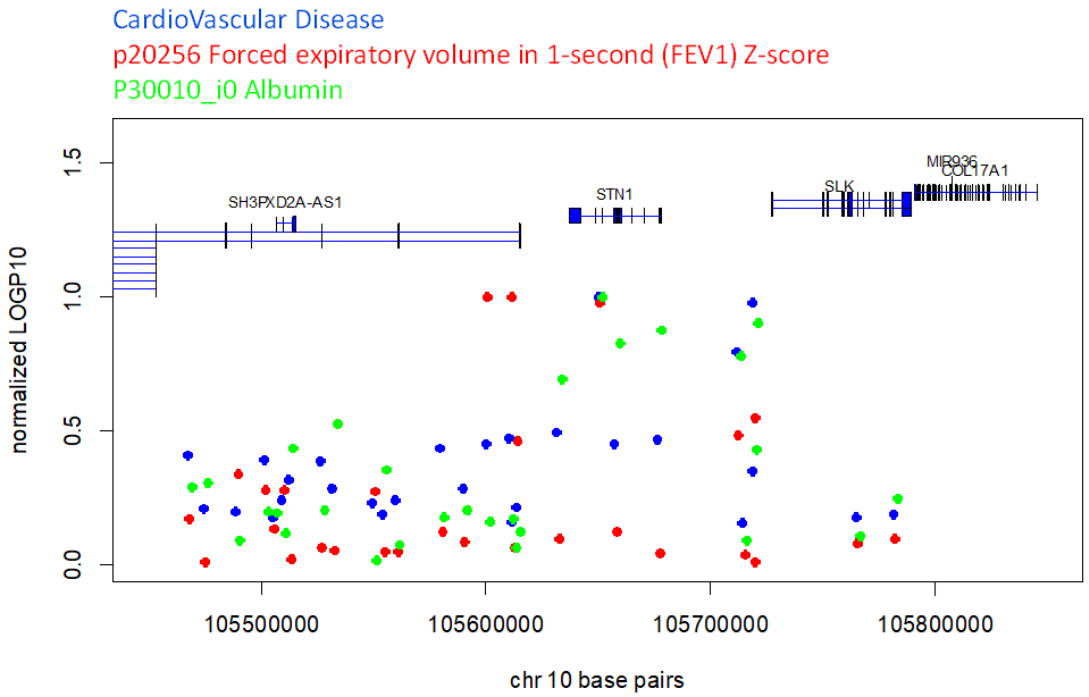
```
--bed ukb22418_c10_b0_v2 \                    #chrom 10
--extract snps_qc_pass.snplist \
--covarFile phenofile_to_refine_eg1_unix.txt \
--phenoFile phenofile_to_refine_eg1_unix.txt \
--phenoCol p30010_i0 \                        # quantitative trait
--covarCol age --covarCol sex --covarCol diab --covarCol hbp --covarCol disl \
--bsize 750 \
--qt --lowmem --loocv \                       # qt for quantitative traits
--lowmem-prefix tmp_rg \
--out CVD2_fit_bin_out_refine_c10             #result files

regenie \
--step 2 \
--bed ukb22418_c10_b0_v2 \
--extract snps_qc_pass.snplist \
--covarFile phenofile_to_refine_eg1_unix.txt \
--phenoFile phenofile_to_refine_eg1_unix.txt \
--phenoCol p30010_i0 \
--covarCol age --covarCol sex --covarCol diab --covarCol hbp --covarCol disl \
--bsize 750 \
--qt \
--strict \
--firth --approx \
--pThresh 0.01 \
--pred CVD2_fit_bin_out_refine_c10_pred.list \
--out test2_bin_out_firth_c10_refine
```

Check the REGENIE pipeline in DNAnexus documentation for details.

5 STEP5: plotting the results of REGENIE analyses

The final step is to merge the LOD score from the main REGENIE file (for CVD in out example) with the secondary REGENIE analyses on selected traits (from STEP3) together with the chromosomal positions of the gene contained in the loci.

The "**plot_multi_LOD**" functions ask for position of regenie file to merge within the result directory, extract and merge LOD scores, extract gene positions from the *"Megabed_hg19.bed"* file and arrange all the information in a single plot. This the results for chr10.



The Figure shows the chromosomal loci in chromosome 10 associated with CVD identified by the present study. LOD scores (y-axis) are plotted according to the chromosomal position (x-axis) of the SNPs. The phenotypes associated to the chromosomal loci are also superimposed to the CVD LOD scoresThe colours of the dots correspond to the label colours.

6