Weekly logs of Master thesis

Zhang Leyi[1]

[1]MSc in BiRC, Aarhus University, Aarhus, Denmark

2023-02

Contents

1 Week 1 **4**

1.1 Software Download . . . 4

1.2 QC . . . 4

1.3 Plink . . . 4

1.4 REGENIE for association study . . . 5

1.5 GEMMA . . . 6

1.6 fastGWA . . . 7

1.7 Bolt-lmm . . . 8

1.8 Question . . . 8

2 Week 2 **9**

2.1 fastGWA, 1000g . . . 9

2.2 PCA . . . 9

2.3 PCA for population prediction, K-means . . . 10

2.4 Simulate Binary Phenotype with LDAK . . . 11

2.5 Regenie for UKBB, binary . . . 11

2.6 Bolt-LMM, 1000g . . . 12

2.7 Ancestry Inference . . . 13

2.8 Week 2 Conclusion . . . 13

3 Week 3 **14**

3.1 Bolt-lmm, UKBB Binary . . . 14

3.2 On batch . . . 15

3.3 LDAK, ukbb . . . 15

3.4 Region 1 15

3.5 UKBB, Whole, Height 16

3.6 Binary 17

3.7 Type 1 error 18

3.8 Plan 20

1 Week 1

Date: 2023/2/8

1.1 Software Download

REGENIE: Stacked block ridge regression method for Mixed Linear Model.

Advantages:

1. Fast and Memory friendly.
2. Can process both quantitative and binary traits.
3. When Case-control unbalanced, h^2_{SNP} will get too high due to the too low MAC(minor allele count). REGENIE can fix this.

Plink

Bolt-LMM

GEMMA

SAIGE problems occur

1.2 QC

Filter out SNPs with genotype missingness > 10%, samples with > 10% missingness, MAF < 5%, minor allele count(MAC) < 100.

Cmd:

```
./software/plink --bfile data --geno 0.1 --mind 0.1 --maf 0.05 --mac 100 \
--make-bed --out data_qc
```

Output: data_qc

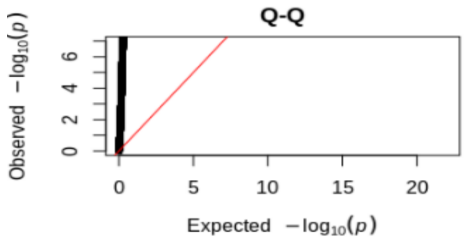
1.3 Plink

Worst result ever.

Cmd:

```
./software/plink --bfile data_qc --linear --pheno height.pheno --allow-no-sex\
--out data_plink_height
```

Q-Q plot: Figure 1



(a) Manhattan plot of REGENIE on height

Figure 1: **QQ plot of plink**

1.4 REGENIE for association study

I took height as the phenotype.

Step 1:

Input files: data_qc(3 files), covar1.covars, height1.pheno. Since REGENIE required labels in each column, I modified covar.covars and height.pheno with labels.

Cmd:

```
    regenie \
--step 1 \
--bed data_qc \
--covarFile covar1.covars \
--phenoFile height1.pheno \
--bsize 100 \
--out data_regenie_out
```

Output: A predicting matrix W for h^2_{SNP}, in file: data_regenie_out_pred.list

Step 2: Association test and LRT

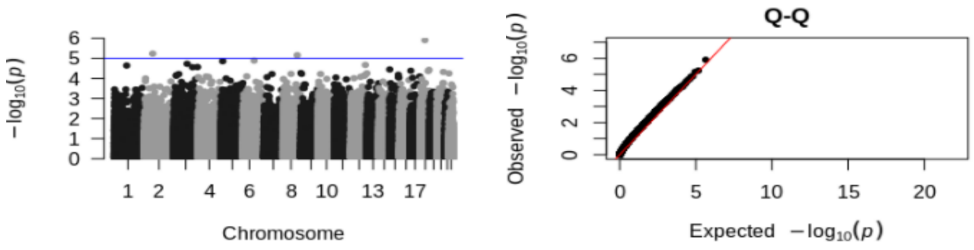
Cmd:

```
    regenie \
--step 2 \
--bgen data_qc.bgen \
```

```
--covarFile covar1.covars \
--phenoFile height1.pheno \
--bsize 200 \
--qt \
--firth --approx \
--pThresh 0.01 \
--pred data_regenie_out_pred.list \
--out data_regenie_out_firth
```

Output: data_regenie_out_firth_Phenotype.regenie

The result can be seen in Figure 2



(a) Manhattan plot of REGENIE on height (b) QQ plot of REGENIE on height

Figure 2: **GWAS results from REGENIE on height (a)**Manhattan plot. **(b)**QQ plot.

Benchmark:

See in Table 1

Method	Step	CPU time	Elapsed time(s)	Memory usage(GB)
REGENIE(null)	1		1492.8	10
REGENIE-Firth	2		4992.47	0.289

Table 1: **Computational performance of REGENIE-Firth**

Refer to: Mbatchou et al. (2021)

1.5 GEMMA

1. Before using GEMMA, the 6th column of .fam should be replaced by real phenotypes.

2. Calculate Kinship Matrix

```
gemma -bfile data_qc -gk 2 -o data_gemma_height
```

While it stuck here for hours, as in Figure 3

(a) Problem with GEMMA

Figure 3: **Problem with GEMMA**

1.6 fastGWA

1. After download, see: **gcta** in ./software folder

2. GCTA-GRM: calculating the genetic relationship matrix (GRM) from all the autosomal SNPs:

```
./software/gcta --bfile data_qc --chr 1 --maf 0.01 --make-grm \
--out data_qc_chr1 --thread-num 10
./software/gcta --bfile data_qc --chr 2 --maf 0.01 --make-grm \
--out data_qc_chr2 --thread-num 10
...
./software/gcta --bfile data_qc --chr 22 --maf 0.01 --make-grm \
--out data_qc_chr22 --thread-num 10
```

Output: .grm.bin, .grm.N.bin, .grm.id

3. To generate a sparse GRM from SNP data:

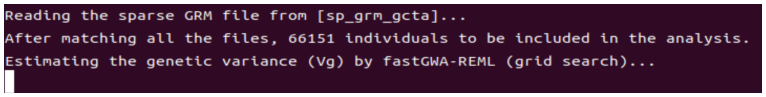
```
./software/gcta --bfile data_qc --autosome --maf 0.01 --make-grm \
--out data_qc_gcta --thread-num 10
./software/gcta --grm data_qc_gcta --make-bK-sparse 0.05 \
--out sp_grm_gcta
```

4. Association study

I didn't use PCs

```
./software/gcta --bfile data_qc --grm-sparse sp_grm_gcta \
--fastGWA-mlm --pheno height.pheno --qcovar covar.covars \
--thread-num 10 --out data_fastgwa_height
```

Problem comes from step 4, Figure 4:



(a) Problem with GEMMA

Figure 4: **Problem with fastGWA**

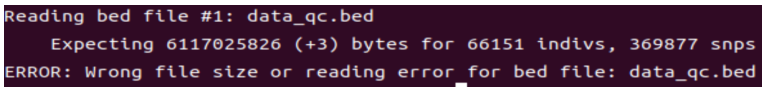
It is stuck here also for hours, is it normal? I'll open the computer for the night and see.

Still stuck there in the morning...

1.7 Bolt-lmm

```
./software/BOLT-LMM_v2.4/bolt --bfile=data_qc --phenoFile=height1.pheno --phenoCol=I
--covarFile=covar1.covars --qCovarCol --lmmForceNonInf --statsFile=data_bolt_height
```

The problem occurs at Figure 5.



(a) Problem

Figure 5: **Problem with Bolt-LMM**

1.8 Question

1. What should I do with CV and LOOCV? What for?

2. How can I detect the Type I errors? By prediction? 3. When I tried SAIGE, I found it need me to use Rscript, how to do?

```
Rscript createSparseGRM.R          \
--plinkFile=${LD pruned PLINK file} \
--nThreads=72   \
--outputPrefix=${OUTNAME}          \
--numRandomMarkerforSparseKin=5000          \
--relatednessCutoff=0.05
```

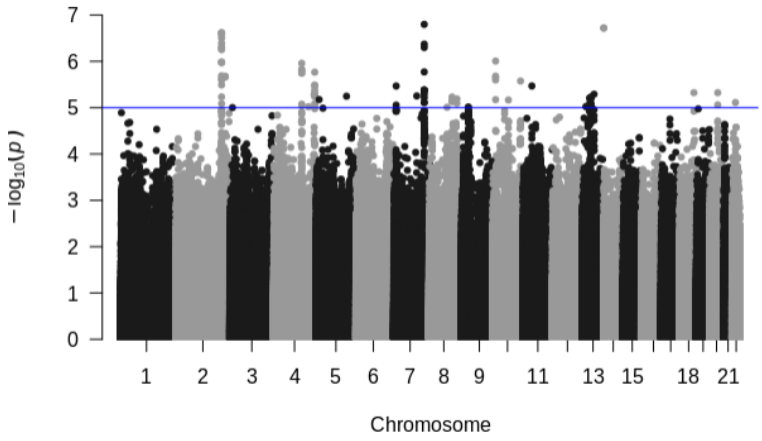
2 Week 2

2.20 - 2.23

Since there are a lot problems occuring with UKBB data, I use 1000g as a demo. If it can work, I'll run on UKBB.

See in commands_Week2.md

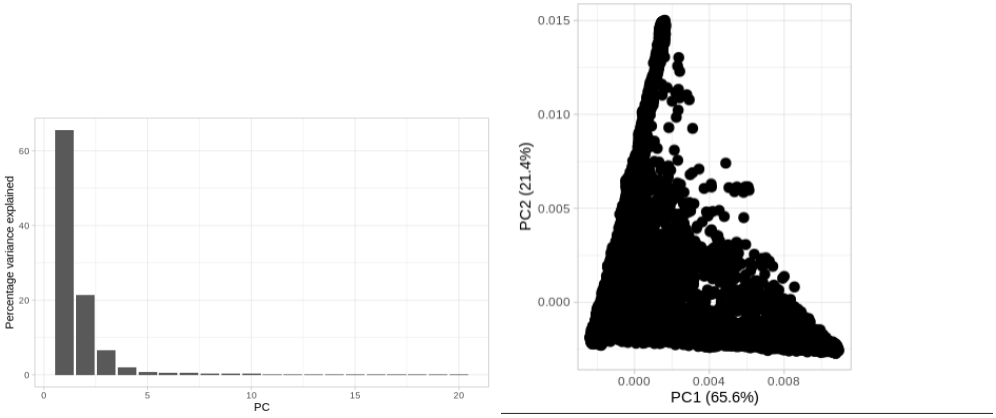
2.1 fastGWA, 1000g



(a) Manhattan plot of fastGWA on 1000g, qt

Figure 6: **fastGWA on 1000g**

2.2 PCA

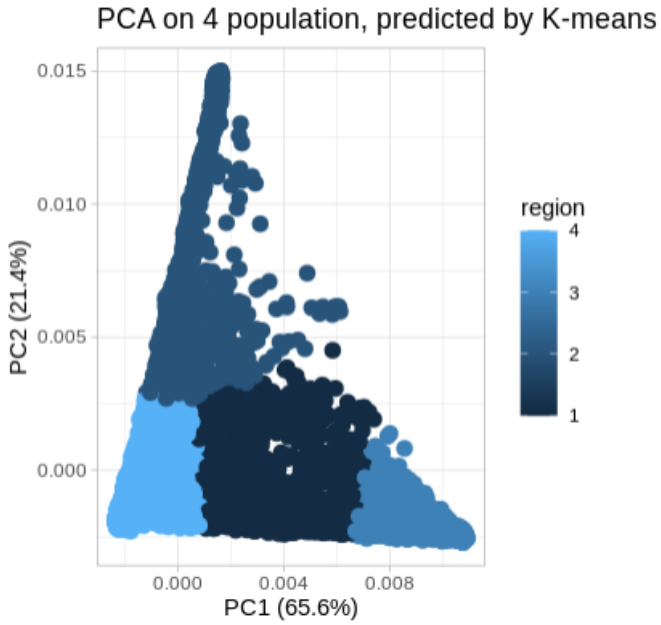


(a) PCs explained (b) PCA

Figure 7: **GWAS results from REGENIE on height**

2.3 PCA for population prediction, K-means

$y <- > PC1 + PC2$

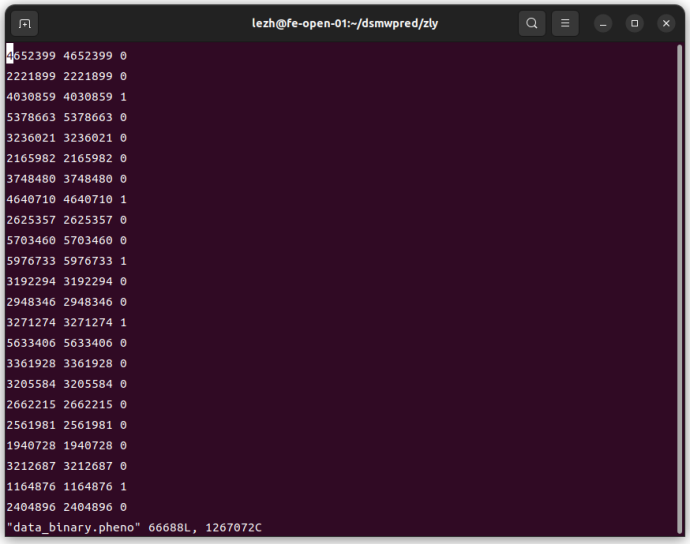


(a) 4 populations

Figure 8: **K-means clustering on PCA**

Problem: I cannot evaluate the result.

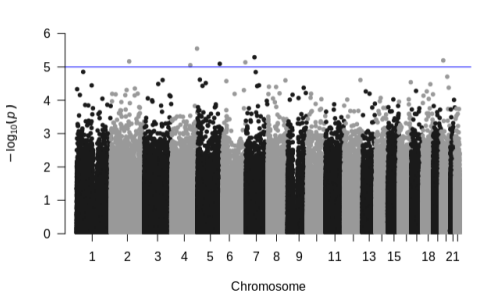
2.4 Simulate Binary Phenotype with LDAK



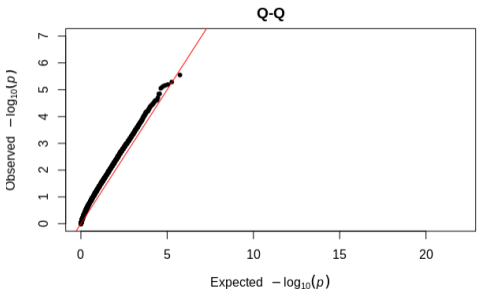
(a) ukbb, simulation

Figure 9: **Binary Phenotypes UKBB simulation**

2.5 Regenie for UKBB, binary



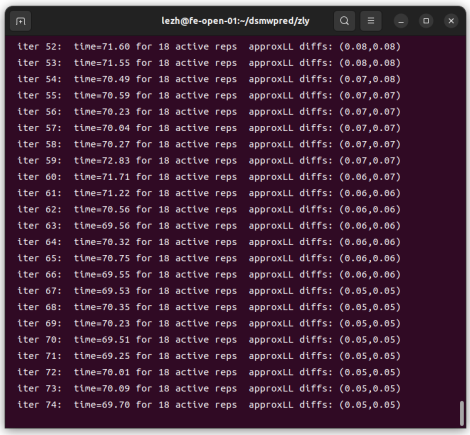
(a) Manhattan plot, UKBB, Binary



(b) QQ plot, UKBB, Binary

Figure 10: **Regenie results of UKBB on binary traits**

2.6 Bolt-LMM, 1000g



(a) CV processing

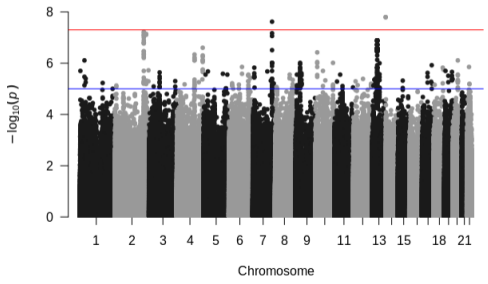
Figure 11: **Bolt result awaiting**

The code has run for 24 hours, not consistent with the brilliant result from the paper, Loh et al. (2015).

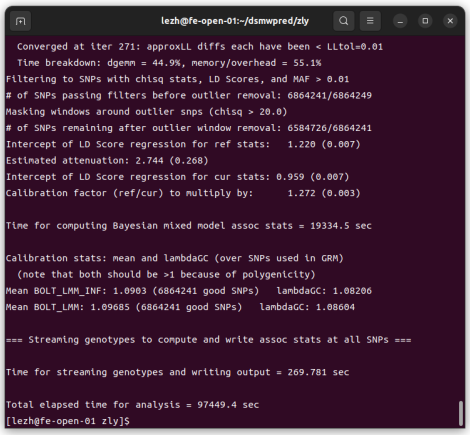
Update: 2023/2/23

SNP Number: 6M

Time elapsed: 97449.4 sec



(a) Bolt result manhattan



(b) Bolt time elapsed

Figure 12: **Bolt result and performance**

2.7 Ancestry Inference

1. as shown in the section 2.3 [PCA for population prediction, K-means].

2. Using software: ADMIXTURE and fastSTRUCTURE for clustering.

3. **Finished**: extract **Region 1** out from .fam., as file: data_region1

4. **processing**: Do Regenie based on ukbb data: data_region1 and **binary traits**.

2.8 Week 2 Conclusion

1. Manhattan from ukbb are not very ideal, as the P value thresholds are almost close to 10^{-5} as shown in the blue lines. While the result from Bolt-lmm + 1000g is good, the threshold is 10^{-7}.

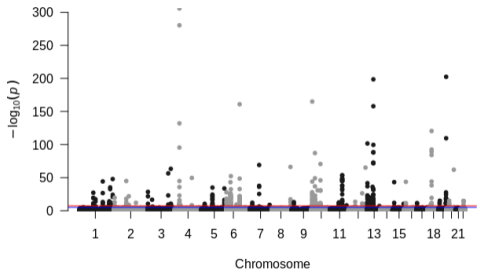
2. I'll take a look on the data after ancestry inference, to see if it's due to cryptic relatedness. But I am also questioning this inference because the 1000g data also has population structure confounding.

3 Week 3

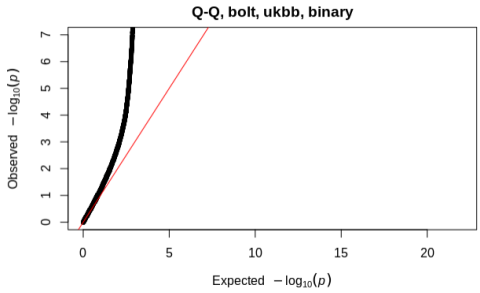
2/24 - 3/01

3.1 Bolt-lmm, UKBB Binary

Time elapsed: 76193.3 sec



(a) Bolt result manhattan



(b) Bolt result QQ-plot

```
        Increasing --maxIters may improve phenotype model and statistical power
  Time breakdown: dgemm = 74.6%, memory/overhead = 25.4%
Filtering to SNPs with chisq stats, LD Scores, and MAF > 0.01
# of SNPs passing filters before outlier removal: 369877/369877
Masking windows around outlier snps (chisq > 66.7)
# of SNPs remaining after outlier window removal: 354065/369877
Intercept of LD Score regression for ref stats:   1.232 (0.016)
Estimated attenuation: 1.516 (0.265)
Intercept of LD Score regression for cur stats: 0.978 (0.044)
Calibration factor (ref/cur) to multiply by:      1.260 (0.044)

Time for computing Bayesian mixed model assoc stats = 43535.6 sec

Calibration stats: mean and lambdaGC (over SNPs used in GRM)
  (note that both should be >1 because of polygenicity)
Mean BOLT_LMM_INF: 1.29132 (369877 good SNPs)   lambdaGC: 1.13876
Mean BOLT_LMM: 1.41723 (369877 good SNPs)   lambdaGC: 1.23414

=== Streaming genotypes to compute and write assoc stats at all SNPs ===

Time for streaming genotypes and writing output = 272.409 sec

Total elapsed time for analysis = 76193.3 sec
```

(c) Bolt result time elapsed

Figure 13: **UKBB, binary, Bolt result and performance**

Comment: Bad result, in qq-plot the black points are far from the red line. And in Manhattan plot, there are also a lot significant associated SNPs.

3.2 On batch

See leyi_week3.txt ukbb_whole_height: fastGWA, regenie, bolt

Results:

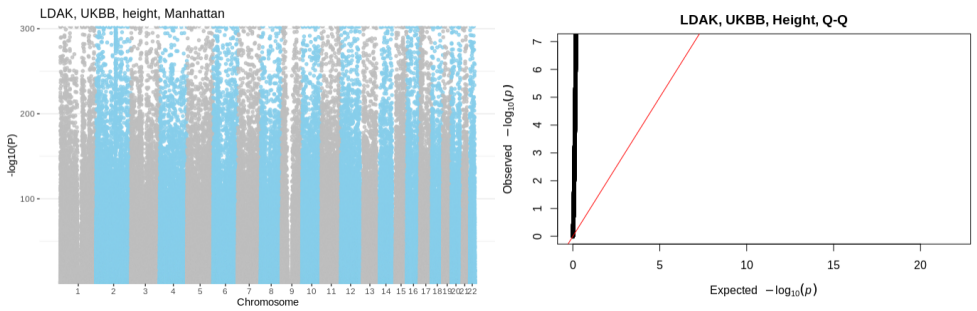
ADMIXTURE: Ran for 3 days, and stopped by system...

3.3 LDAK, ukbb

60 K inds in whole ukbb.

Binary

Height: Figure 14



(a) LDAK, Height, Manhattan (b) LDAK, Height, QQ

Figure 14: **LDAK performance on UKBB, Height.**

Comment: I guess I used a wrong method in LDAK? It's more like the basic Plink?

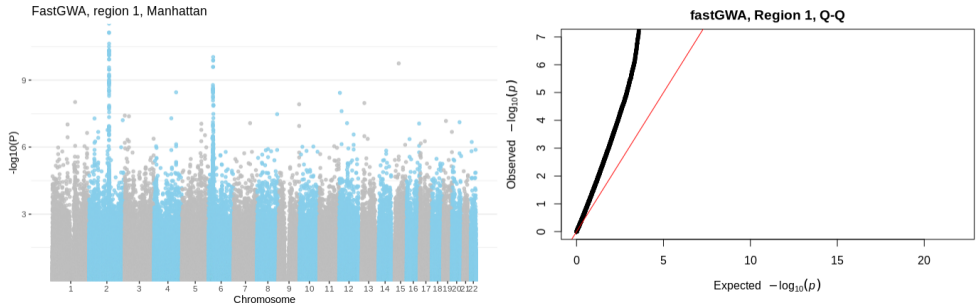
3.4 Region 1

10 K inds in region 1.

Binary

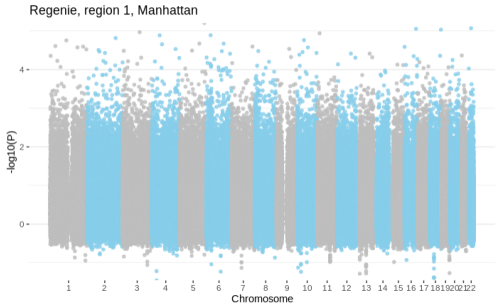
Height: bolt, fastGWA, regenie

Results are shown below: Figure 15:

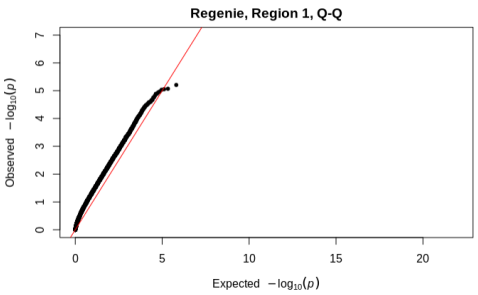


(a) fastGWA, Manhattan

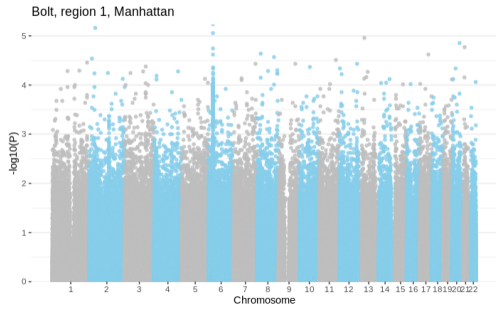
(b) fastGWA, QQ



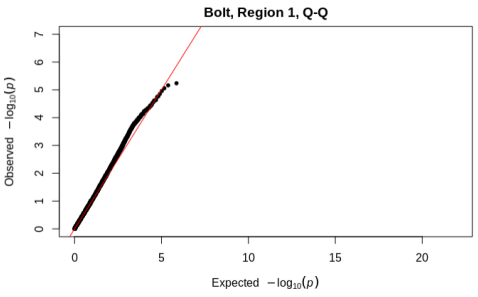
(c) Regenie, Manhattan



(d) Regenie, QQ



(e) Bolt, Manhattan



(f) Bolt, QQ

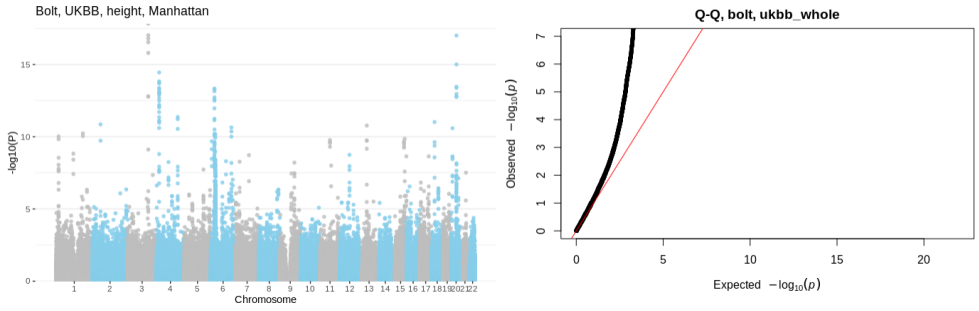
Figure 15: **Different software performance on UKBB region 1.**

In the next subsection, I'll compare results from region 1 with the whole UKBB data.

3.5 UKBB, Whole, Height

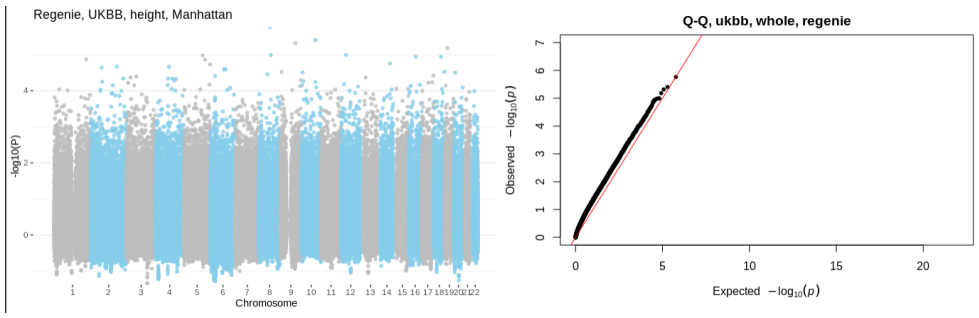
fastGWA is still awaiting. Only results from Bolt and Regenie now.

Results are shown below: Figure 16:



(a) Bolt, height, Manhattan

(b) Bolt, height, QQ



(c) Regenie, height, Manhattan

(d) Regenie, height, QQ

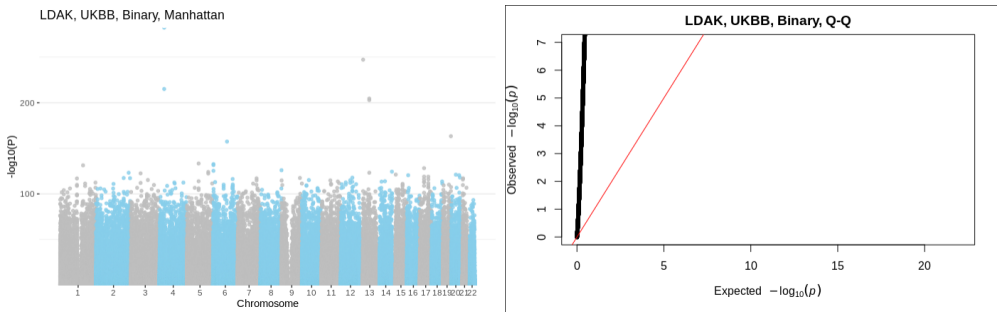
Figure 16: **Results from UKBB, height: Bolt and Regenie.**

3.6 Binary

60 K inds.

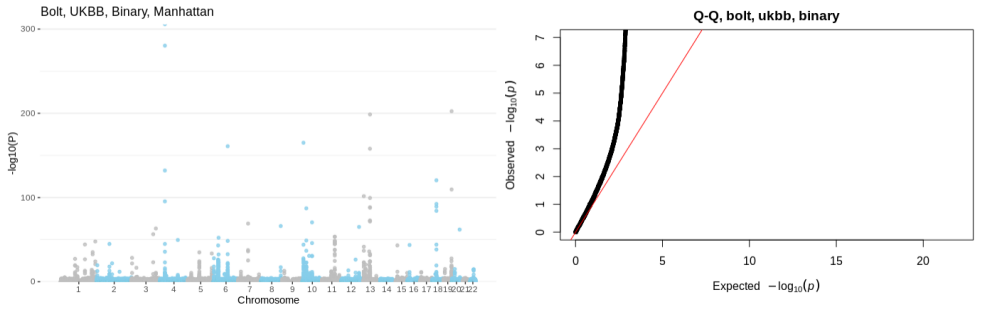
Binary

Results are shown below: Figure 17:



(a) LDAK, Manhattan

(b) LDAK, QQ



(c) Bolt, Manhattan

(d) Bolt, QQ

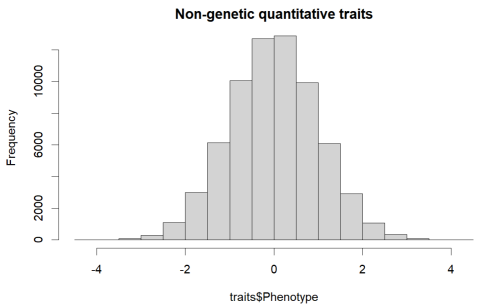
Figure 17: **LDAK and Bolt performance on UKBB, Binary.**

3.7 Type 1 error

1. To generate **complete non-genetic traits**, Sabourin et al. (2019).

2. Calculate the proportion of P-value below a threshold, Berrandou, Balding, and Speed (2023).

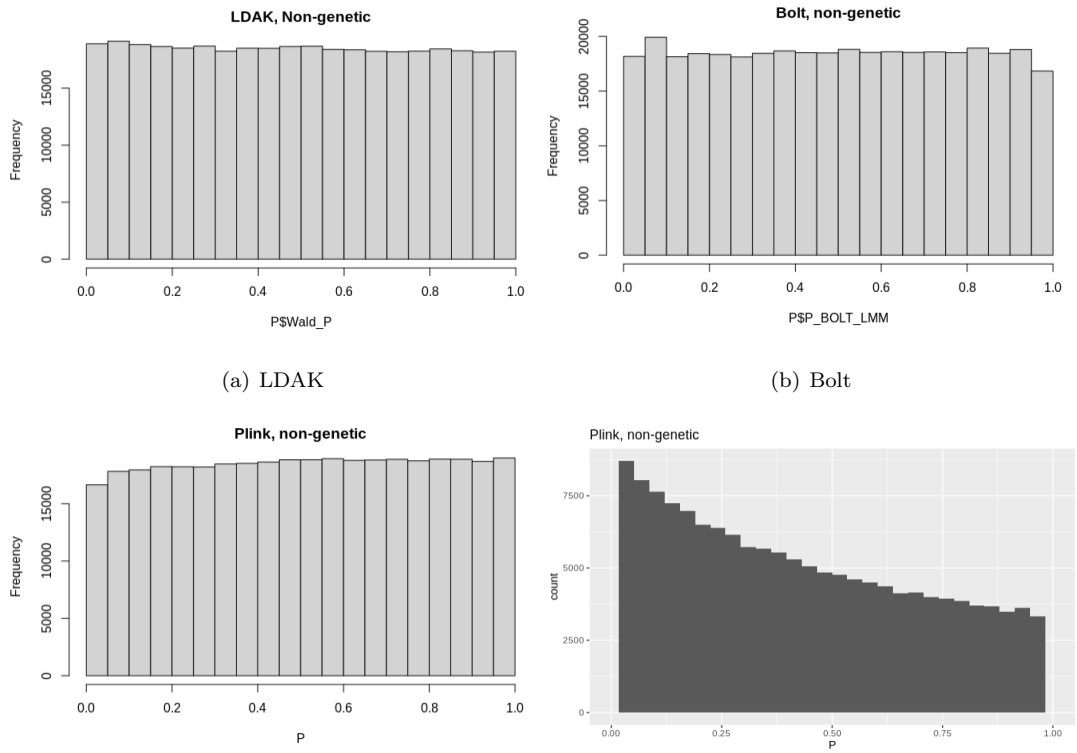
The phenotypes simulated look as shown in Figure 18.



(a) Distribution of the traits

Figure 18: **UKBB, non-genetic traits, quantitative, by LDAK**

I chose the softwares above: Plink, fastGWA, LDAK, Bolt, Regenie, to test the performance. The distribution of P value in each software shows in Figure 19.



(a) LDAK

(b) Bolt

(c) Plink

(d) Regenie

Figure 19: **Distribution of P values of each software.**

After getting the results, I did the bonferroni correction for p-value. Or by using X2?

Should I test them with permutation?

I tested the performance and proportion of P value < 0.01 and P value $< 5 \times 10^{-5}$ from each software, Table 2. **UNFINISHED**

Software	Time	Memory(G)	P<0.01	P<5e-5
fastGWA_step1	9:08:01	41.78	NA	NA
fastGWA_step2	0:02:50	1.1	NA	NA
fastGWA_step3				
Regenie_Step1	7:00:06	35.67	NA	NA
Regenie_Step2	2:18:15	0.513	0.007294	3.51E=05
Bolt	2:15:37	6.21	0.009798	4.05E-05
LDAK	5:50:00	2.57	0.010101	4.86E-05
Plink	0:09:17	0.0413	0.008476	4.05E-05

Table 2: **Performance to Type I error.**

3.8 Plan

In the contract: This project will compare strategies for analyzing multi-ancestry datasets. In particular, it will quantify the power increase (or decrease) of MMAA (analyzing all individuals together) with per-ancestry analyses (e.g., analyze European, African and Asian separately, then meta-analyse the three sets of results). **In particular, it will investigate whether the optimum strategy depends on consistency of genetic architecture across ancestries.**

1. Today and tomorrow: meta-analysis

2. METAL software for meta-analysis. Khunsriraksakul et al. (2023), Willer, Li, and Abecasis (2010)

3. Question: can I use some software I have used to do a meta-analysis, otherwise should I also test for the performance of METAL?

References

Berrandou, T.-E., Balding, D., & Speed, D. (2023, 1). Ldak-gbat: Fast and powerful gene-based association testing using summary statistics. *The American Journal of Human Genetics*, *110*, 23-29. DOI: 10.1016/j.ajhg.2022.11.010

Khunsriraksakul, C., Li, Q., Markus, H., Patrick, M. T., Sauteraud, R., McGuire, D., ... Liu, D. J. (2023, 2). Multi-ancestry and multi-trait genome-wide association meta-analyses inform clinical risk prediction for systemic lupus erythematosus. *Nature Communications*, *14*, 668. DOI: 10.1038/s41467-023-36306-5

Loh, P.-R., Tucker, G., Bulik-Sullivan, B. K., Vilhjálmsson, B. J., Finucane, H. K., Salem, R. M., ... Price, A. L. (2015, 3). Efficient bayesian mixed-model analysis increases association power in large cohorts. *Nature Genetics*, *47*, 284-290. DOI: 10.1038/ng.3190

Mbatchou, J., Barnard, L., Backman, J., Marcketta, A., Kosmicki, J. A., Ziyatdinov, A., ... Marchini, J. (2021, 7). Computationally efficient whole-genome regression for quantitative and binary traits. *Nature Genetics*, *53*, 1097-1103. DOI: 10.1038/s41588-021-00870-7

Sabourin, J. A., Cropp, C. D., Sung, H., Brody, L. C., Bailey-Wilson, J. E., & Wilson, A. F. (2019, 2). Compass-gwas: A method to reduce type i error in genome-wide association studies when replication data are not available. *Genetic Epidemiology*, *43*, 102-111. DOI: 10.1002/gepi.22168

Willer, C. J., Li, Y., & Abecasis, G. R. (2010, 9). Metal: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics*, *26*, 2190-2191. DOI: 10.1093/bioinformatics/btq340