PROJECT PART II: GENETIC ARCHITECTURE ANALYSES

STAT7306: STATISTICAL ANALYSIS OF GENETIC DATA

Jayden Beckwith
46267539

Contents

Introduction

Following "Part I: Genome-wide association study analyses", study data consisting of HapMap 3 genome-wide SNPs for approximately 11,000 people from the UK Biobank was used for genetic architecture analyses. The aim of this investigation was to determine the genetic heritability of the fasting glucose (FG) phenotype across non-related study participants. Following QC procedure and GWAS analyses, the SNP data was represented in 3 formats consisting of testFiltered.bim, testFiltered.fam and testFiltered.bed. Additionally, the FG data was represented in the form of three traits in .phen text format. This included a quantitative trait, "binary 1" trait (scoring in the top 20% of the phenotype are scored 1 = case and the remainder 0 =control) and the "binary 2" trait (scoring in the top 20% of the phenotype are scored 1 = case and those scoring in the bottom 30% of the phenotype are scored 0 = control).

Methods

Despite genome-wide association studies (GWAS) being highly effective at detecting genetic variation in human traits, the genetic architecture of human complex traits remains largely unexplained. Complex traits are determined by numerous genetic and environment factors as well as their interactions (Yang et al., 2013). Several statistical methods were used in this investigation to explain the genetic variance between genome-wide SNPs amongst the targeted complex traits. Genome wide complex trait analysis (GCTA) was performed and is based on the variance explained by all SNPs of a phenotypical trait rather than testing associates of each individual SNP like in GWAS. GCTA comprises of two main steps: estimation via the Genetic Relationship Matrix (GRM), which is calculated to measure the genetic correlation between study participants and SNP data and the use of genome-based restricted maximum likelihood (GREML), to maximise the likelihood of the target traits given by the GRM. This allows to effectively compare the phenotypic similarity to genetic similarity in individuals. Once the GRM is developed, GREML involves the fitting of a mixed linear model (MLM) to measure the effects of SNPs (Yang et al., 2013). This is represented below by the following equation:

$$y = X\beta + Z\mu + \epsilon$$

Where, fixed effects are denoted by β (Special case: X = 1 and $\beta = \mu$), Z represents the standardised SNP genotype matrix, μ is a vector of the random effects with approximately normal distribution $\mu \sim N(0, I\sigma^2{}_u)$ which can be described as the phenotype effect on the SNP data and ϵ is the vector of error or "noise" (Yang, 2022). Furthermore, an equivalent model for when $g = Z\mu$, the total genetic effect for the nth individuals in the analysis and the GRM representation, $G = \frac{ZZ^T}{N}$ (where *N* is the number of SNPs) and *I* is the identity matrix, which gives the following equation:

$$y = X\beta + g + \epsilon \ \ with\ var(y)\ or\ V = G\sigma^2{}_g + I\sigma^2{}_e$$

Furthermore, the main purpose of GCTA was to identify $\sigma^2{}_e$ and $\sigma^2{}_g$ variance components through the fit of MLM (Yang et al., 2013). Prior to fitting, the GRM (G) was calculated for each trait and drives the functionality of GCTA for variance determination through GREML

analysis. Where parameter estimation for variance components in GREML analysis can be denoted by the log likelihood (Visscher, 2022; Yang, 2022):

$$logL = -\frac{1}{2}(log|V| + log|\mathrm{X}^{\mathrm{T}}\mathrm{V}^{-1}X| + \mathrm{y}^{\mathrm{T}}Py)$$

$$P = \mathrm{V}^{-1} - \mathrm{V}^{-1}X(\mathrm{X}^{\mathrm{T}}\mathrm{V}^{-1}X)^{-1}\mathrm{X}^{\mathrm{T}}\mathrm{V}^{-1}$$

GREML analysis consisted of singular SNP heritability estimates for the three traits and partitioning SNP heritability estimation based on the provided genomic annotations of "0 – non-target SNP" and "1 – targeted SNP" requiring a multi-GRM approach. It was hypothesised that the target SNPs via genome annotation explain a greater amount of variance than the non-targeted annotated SNPs. To test this hypothesis, genomic annotation labels were separated in R and collated into a single text file to be utilised for the multi-GRM.

SNP heritability was represented as variance explained where, $\mathrm{h}^2_{\mathrm{SNP}} = \sigma^2_{\mathrm{g}} / (\sigma^2_{\mathrm{g}} + \sigma^2_{\mathrm{e}})$ and is the ratio of phenotypic variance on the basis of SNP effects (on an observed scale between 0-1). Significance was determined by likelihood ratio testing to determine the goodness-of-fit between the two models via $LRT = 2[L(H1) - L(H0)]\ at\ df = 1$. Degrees of freedom (df) was 1 in LRT because of the distribution between half the probability of 0 and a half the probability of chi-squared (Yang et al., 2013). Significance was denoted by the null and alternative hypotheses:

$$\mathrm{H}_0\text{: } \mathrm{h}^2_{\mathrm{SNP}} = 0$$

$$\mathrm{H}_1\text{: } \mathrm{h}^2_{\mathrm{SNP}} \neq 0$$

A test statistic for SNP heritability was assumed at $p < 0.05$ and the higher the result of the log likelihood ratio, the better the model fit for the SNP dataset. Since GREML analysis works off the basis of the assumptions of linear regression, GREML assumptions follow accordingly. For example, it can be assumed that there exists independence between random effects when estimating variance components. Furthermore, GREML was also limited by the assumptions of population stratification and cryptic relatedness. Therefore, the initial GRM generated was set to 0.05 to minimise related individuals used in the analysis that could create potential genetic bias. By not using relatives who may share a similar environment, this provides results that are not reliant on family study assumptions and therefore, give a more accurate representation of SNP heritability. Furthermore, quantitative covariates from the QC dataset were also investigated for population stratification. However, it was observed to create collinearity problems in the GRM matrix between the genetic and environmental variables. It was decided to not include the covariates in the analysis to ensure p-values were captured for the total genetic effect. Lastly, it was also assumed that there was no underlying batch, plate or other technical artefact that would affect the quality of the data used for analysis.

Results

SNP Heritability Estimation

Table #1 represents a summary of the variance calculated for genetic SNP heritability of the FG concentration phenotype traits (to 2 decimal places), where QFG is quantitative fasting glucose trait, B1FG is binary 1 fasting glucose trait and BF2G is Binary 2 Fasting glucose trait.

Table #1 – SNP heritability and standard error for each FG trait

Trait	Genetic Heritability (V(G)/Vp)	Standard Error
QFG	0.18	0.03
B1FG	0.11	0.03
BF2G	0.33	0.06

Table #2 represents a summary of the results calculated from LRT and the significance of the heritability captured by each trait based on their sample size.

Table #2 – Results of Log Ratio Testing and statistical significance

Source	QFG	B1FG	B2FG
LRT	31.54	12.6	28.15
p-value	9.77×10^{-9}	1.93×10^{-4}	5.62×10^{-8}
n	11721	11731	5853

Partitioning SNP Heritability Estimation

Partitioning SNP heritability of the annotated SNPs via using a multi-GRM approach were displayed in table #3.

Table #3 – Partitioning heritability of targeted vs non-targeted SNPs

Trait	Heritability of non-target (V(G1)/Vp)	Heritability of target (V(G2)/Vp)	Standard Error non-target	Standard Error target
QFG	0.02	0.15	0.03	0.03
B1FG	0.007	0.10	0.026	0.025
B2FG	0.03	0.28	0.05	0.05

Discussion

The results from GCTA indicated that the FG traits were largely influenced by environmental variance. However, the partitioning SNP heritability showed that the targeted SNPs displayed a greater amount of variance in comparison to the non-target SNPs across the three traits. Subsequently, the results were suggestive that these targeted SNPs accounted for approximately 90% of the total SNP variance explained as genetic heritability:

$$Target\ SNP\ h^2 = \frac{\Sigma\ V(\frac{G_2}{Vp})}{\Sigma\ V\left(\frac{G_2}{Vp}\right) + V\left(\frac{G_1}{Vp}\right)} = \sim 0.9 = explains\ 90\%\ of\ genetic\ heritability$$

Comparing the SNP heritability of the three traits, it was evident that B2FG trait had the highest degree of heritability. This was supported by an LRT result of 28.15 for singular SNP heritability ($P = 5.62 \times 10^{-8}$) indicative that the model had a good fit and was statistically significant to the SNP data. This was also captured in partitioning heritability between the three traits, with the targeted annotated SNPs in the B2FG trait resulting in the highest genetic heritability ratio. B1FG was observed to display the lowest captured degree of heritability, but still displayed statistically significant variance ($P = 1.93 \times 10^{-4}$) . This is likely due to the nature of the data since the cases only represent 20% of the data in the trait, substantially skewing the results due to the loss of most of the original data. Conversely, the highest LRT result was 31.54 ($P = 9.77 \times 10^{-9}$) and was captured in the QFG trait and is likely a better representation of the genetic variance in the of the FG traits due to the more significant test statistic and larger sample size.

The larger proportion of SNP heritability of FG in the B2FG trait was due to the high variance distribution of the dataset, since the trait comprised of the top 20% and bottom 30% of all SNPs. Therefore, accounting for the largest spread in comparison to the other two traits. Consequently, this created a greater predisposition to have a larger genetic heritability ratio than the other two traits. Drawing back to the phenotypic liability threshold models of complex traits, the reported estimates of heritability are the heritability of liability. If B2FG were to be attempted to be modelled on a liability scale, this would create a general bias in the trait when modelling for the population. Additionally, it would violate the assumption of normality for liability modelling, due to only taking into consideration the top 20% and bottom 30% of SNP data. The observed scale traits are binary, and the quantitative traits can only be modelled on a liability scale. Ultimately, given the higher statistical significance and the ability to be modelled on a liability scale, the QFG trait would be the preferable trait to be modelled for the population.

Conclusion

Ultimately, this investigation aimed to determine the genetic heritability for the FG phenotype in study participants from the UK Biobank study. It was observed from partitioning heritability estimation that the targeted SNPs explained approximately 90% of the total genetic variance and B2FG had the highest genetic heritability across the three observed traits. It was questionable whether B2FG is an accurate model representation give the high predisposition in variance present in the trait that may drive the increased SNP heritability ratio. Although FG concentration remains largely an environmental effected phenotype, this investigation showed that a significant proportion of SNPs do play a role in FG concentration.

References

Visscher, P. M. (2022). *Variance component estimation*. The University of Queensland. https://learn.uq.edu.au/webapps/blackboard/execute/content/file?cmd=view&content_id=_8259128_1&course_id=_161736_1

Yang, J. (2022). *GREML: estimation of genetic variance in unrelated individuals*. The University ofQueensland.https://learn.uq.edu.au/webapps/blackboard/execute/content/file?cmd=view&content_id=_8292637_1&course_id=_161736_1&framesetWrapped=true

Yang, J., Lee, S. H., Goddard, M. E., & Visscher, P. M. (2013). Genome-wide complex trait analysis (GCTA): methods, data analyses, and interpretations. *Methods Mol Biol*, *1019*, 215-236. DOI: 10.1007/978-1-62703-447-0_9.

Supplementary Materials

Make GRM

```
/data/STAT3306/gcta --bfile Data_QC/testFiltered --make-grm --autosome --out
results/grmdata --thread-num 8

/data/STAT3306/gcta --grm results/grmdata --grm-cutoff 0.05 --make-grm --out
results/data_nr
```

Non-relational

GREML Quantitative trait

```
/data/STAT3306/gcta --grm results/data_nr --pheno
Phenotypes_QC/Fasting_Glucose_QC.phen --mpheno 1 --reml --out results/FG_quant_nr -
-thread-num 4
```

```
[s4626753@fac-login-1 results]$ cat FG_quant_nr.hsq
Source  Variance        SE
V(G)    0.809326        0.147465
V(e)    3.751549        0.150240
Vp      4.560876        0.059795
V(G)/Vp 0.177450        0.032052
logL    -14738.013
logL0   -14753.783
LRT     31.540
df      1
Pval    9.7679e-09
n       11721
```

GREML binary 1 trait

```
/data/STAT3306/gcta --grm results/data_nr --pheno
Phenotypes_QC/Fasting_Glucose_binary1.phen --mpheno 1 --reml --out
results/FG_binary1_nr --thread-num 4
```

```
[s4626753@fac-login-1 results]$ cat FG_binary1_nr.hsq
Source  Variance        SE
V(G)    0.018035        0.005069
V(e)    0.141666        0.005281
Vp      0.159701        0.002089
V(G)/Vp 0.112930        0.031621
logL    4897.733
logL0   4891.434
LRT     12.598
df      1
Pval    1.9308e-04
n       11731
```

GREML binary 2 trait

```
/data/STAT3306/gcta --grm results/data_nr --pheno
Phenotypes_QC/Fasting_Glucose_binary2.phen --mpheno 1 --reml --out
results/FG_binary2_nr --thread-num 4
```

```
[s4626753@fac-login-1 results]$ cat FG_binary2_nr.hsq
Source  Variance        SE
V(G)    0.078980        0.015293
V(e)    0.161211        0.015029
Vp      0.240190        0.004468
V(G)/Vp 0.328822        0.062704
logL    1259.898
logL0   1245.824
LRT     28.149
df      1
Pval    5.6170e-08
n       5853
```

Multi-GRM SNP heritability

Pre-processing annotation file

```
> annotation_tab <- read.table('Data_QC/annotation.txt', header =TRUE)
> head(annotation_tab)
  chr         ID Annotated
1   1  rs3131972         0
2   1  rs3115850         0
3   1 rs12562034         1
4   1  rs4040617         1
5   1  rs4970383         0
6   1   rs950122         0
```

```
> target <- which(annotation_tab$Annotated == 1)
> target_file = annotation_tab[target, 2]
> write.table(target_file, row.names = FALSE, col.names = FALSE, quote = FALSE, file = 'results/target_SNPs.txt')
```

```
> control <- which(annotation_tab$Annotated == 0)
> control_file = annotation_tab[control, 2]
> write.table(control_file, row.names = FALSE, col.names = FALSE, quote = FALSE, file = 'results/control_SNPs.txt')
>
```

Executing GRM for targeted and non-targeted SNPs

```
/data/STAT3306/gcta --bfile Data_QC/testFiltered --extract results/control_SNPs.txt
--autosome --make-grm --keep results/data_nr.grm.id --thread-num 8 --out
results/data_nr_control_snps
```

```
/data/STAT3306/gcta --bfile Data_QC/testFiltered --extract results/target_SNPs.txt
--autosome --make-grm --keep results/data_nr.grm.id --thread-num 6 --out
results/data_nr_target_snps
```

Compile paths into a singular text file for multi GRM via vim editing

```
[s4626753@smp-0-34 finalProject]$ cat multi_SNP.txt
results/data_nr_control_snps
results/data_nr_target_snps
[s4626753@smp-0-34 finalProject]$
```

Runing multi GRM using non-relational data against traits

```
/data/STAT3306/gcta --mgrm multi_SNP.txt --pheno
Phenotypes_QC/Fasting_Glucose_QC.phen --mpheno 1 --reml --thread-num 8 --out
results/data_FG_mlt_nr
```

```
[s4626753@smp-0-34 results]$ cat data_FG_mlt_nr.hsq
Source  Variance        SE
V(G1)   0.106887        0.120895
V(G2)   0.667923        0.117402
V(e)    3.786123        0.150449
Vp      4.560933        0.059847
V(G1)/Vp        0.023435        0.026500
V(G2)/Vp        0.146444        0.025505

Sum of V(G)/Vp  0.169880        0.032047
logL    -14733.486
n       11721
```

```
/data/STAT3306/gcta --mgrm multi_SNP.txt --pheno
Phenotypes_QC/Fasting_Glucose_binary1.phen --mpheno 1 --reml --thread-num 8 --out
results/data_FG_mlt_binary1_nr
```

```
[s4626753@smp-0-34 results]$ cat data_FG_mlt_binary1_nr.hsq
Source  Variance        SE
V(G1)   0.001087        0.004196
V(G2)   0.016113        0.004042
V(e)    0.142505        0.005287
Vp      0.159705        0.002090
V(G1)/Vp        0.006804        0.026275
V(G2)/Vp        0.100894        0.025195

Sum of V(G)/Vp  0.107698        0.031625
logL    4900.405
n       11731
```

/data/STAT3306/gcta --mgrm multi_SNP.txt --pheno Phenotypes_QC/Fasting_Glucose_binary2.phen --mpheno 1 --reml --thread-num 8 --out results/data_FG_mlt_binary2_nr

```
[s4626753@smp-0-34 results]$ cat data_FG_mlt_binary2_nr.hsq
Source  Variance        SE
V(G1)   0.006810        0.012699
V(G2)   0.066618        0.012098
V(e)    0.166710        0.015056
Vp      0.240138        0.004473
V(G1)/Vp        0.028357        0.052870
V(G2)/Vp        0.277417        0.049539

Sum of V(G)/Vp  0.305774        0.062699
logL    1264.579
n       5853
```