Lindgren Group: Bioinformatician Work Test

Hazel.A.Fernando

Last compiled on 03 September, 2024

Methodology

Quality Control

Used PLINK to perform QC.

The genotype and phenotype data were converted into into PLINK compatible file formats.

Note, for the phenotypic data to be compatible for PLINK, the sample column was duplicated and given the necessary column names, IID and FID.

```
vcf_file <- "StatGen_work_test.vcf.gz"
vcf_data <- read.vcfR(vcf_file)

data <- "data"

system(paste("plink --vcf", vcf_file, "--make-bed --out", data))
```

```
phenotype_file <- "StatGen_work_test_phenotype.txt"
phenotype_data <- fread(phenotype_file)

write.table(phenotype_data, file = "phenotype_data.plink.txt", quote = FALSE,
            row.names = FALSE, col.names = FALSE, sep = "\t")
```

Combined the genotypic and phenotypic data together.

```
system(paste("plink --bfile", data,
             "--pheno phenotype_data.plink.txt --make-bed --out combined_data"))
```

QC steps

The standard QC steps were taken to filter out low quality variants and individuals.

```
combined_data <- "combined_data"
```

1. Removed SNPs missingness > 5%.

```
system(paste("plink --bfile", combined_data,
             "--geno 0.03 --make-bed --out QC1"))
```

2. Removed individuals/samples with missingness $> 3\%$.

```
system(paste("plink --bfile QC1 --mind 0.03 --make-bed --out QC2"))
```

3. Removed SNPs with a MAF $< 1\%$.

```
system(paste("plink --bfile QC2 --maf 0.01 --make-bed --out QC3"))
```

4. Removed SNPs with a p-value for HWE $< 1 \times 10^{-6}$

```
system(paste("plink --bfile QC3 --hwe 1e-6 --make-bed --out QC4"))
```

Association Analysis

Used PLINK to perform GWAS analysis.

```
filtered_data <- "QC4"

gwas_results <- "gwas_results"
```

```
system(paste("plink --bfile", filtered_data, "--allow-no-sex", "--assoc --out",
             gwas_results))
```

```
gwas_results <- fread("gwas_results.qassoc")
```

Filtered the GWAS results to only include GWS SNPs with p-value $< (5 \times 10^{-8})$.

```
sig_snps <- gwas_results[gwas_results$P < 5e-8,]

number_sig_signals <- nrow(sig_snps)
```

```
## [1] "Number of significant independent signals: 9"
```

```
sig_snps
```

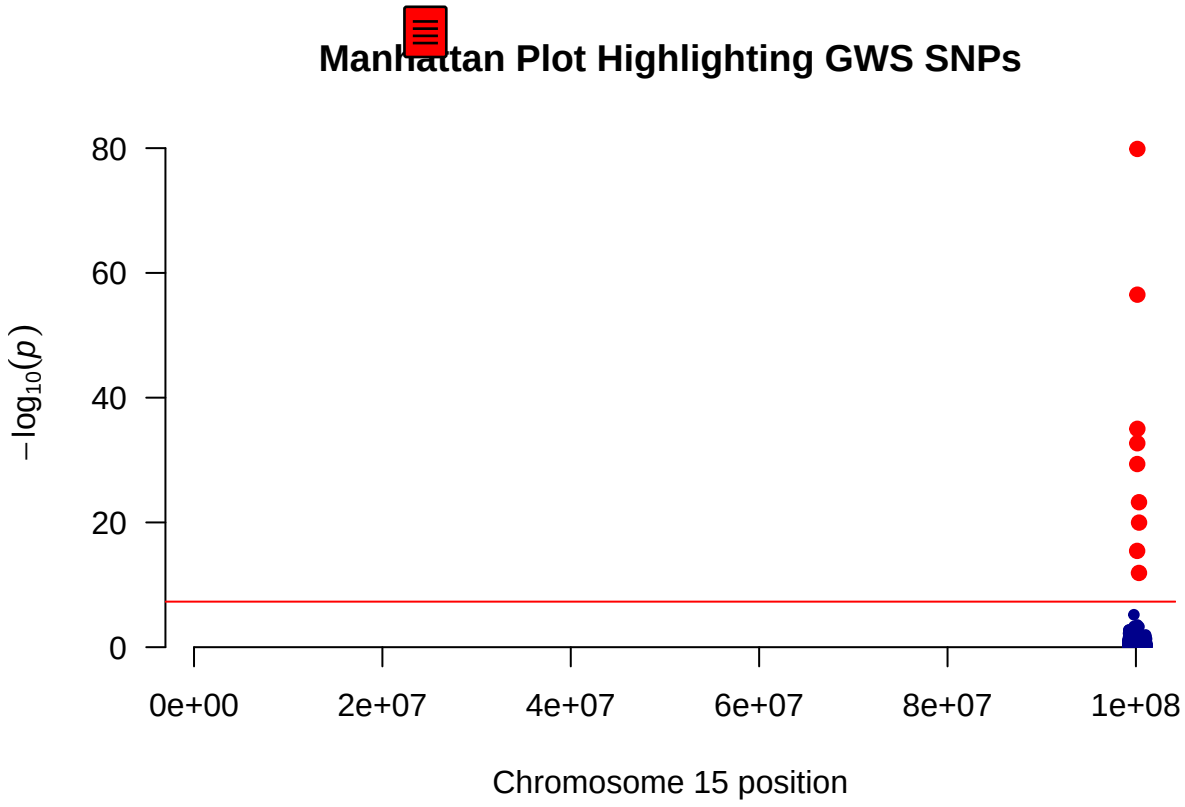
```
##      CHR        SNP        BP NMISS    BETA      SE       R2      T         P
##    <int>     <char>     <int> <int>   <num>   <num>    <num>  <num>     <num>
## 1:    15 rs45960959 100130220  9907 0.1292 0.01583 0.006683  8.163 3.656e-16
## 2:    15 rs31187023 100140220  9907 0.1953 0.01708 0.013030 11.430 4.351e-30
## 3:    15 rs48338810 100144139  9907 0.1980 0.01638 0.014540 12.090 2.036e-33
## 4:    15 rs72755233 100152748  9907 0.3920 0.02043 0.035840 19.190 1.341e-80
## 5:    15 rs38965480 100153138  9907 0.2221 0.01772 0.015600 12.530 9.820e-36
## 6:    15 rs10611351 100154139  9907 0.2986 0.01861 0.025340 16.050 3.090e-57
## 7:    15 rs49128648 100325138  9907 0.1168 0.01643 0.005076  7.109 1.254e-12
## 8:    15 rs27298763 100327445  9907 0.1822 0.01800 0.010240 10.120 5.781e-24
## 9:    15 rs24254005 100334139  9907 0.1666 0.01782 0.008749  9.350 1.065e-20
```

```r
gws_snps <- gwas_results$SNP[gwas_results$P < 5e-8]

gwas_results$highlight <- ifelse(gwas_results$P < 5e-8, "red", "black")

manhattan(gwas_results,
          chr = "CHR",
          bp = "BP",
          snp = "SNP",
          p = "P",
          genomewideline = -log10(5e-8),
          suggestiveline = FALSE,
          col = "blue4",
          main = "Manhattan Plot Highlighting GWS SNPs",
          ylim = c(0, max(-log10(gwas_results$P), na.rm = TRUE) + 2))

with(gwas_results[gwas_results$P < 5e-8,], {
  points(BP, -log10(P), col = "red", pch = 19)
})
```



Manhattan plot of the GWAS results of the index variants along chromosome 15. GWS < 5e-08 is marked by the red line. GWS independent loci are highlight as blue. -log10(p) values are plotted against the chromosomal positions.

Aggregated the significant signals together to identify independent signals to eliminate related signals that possibly arise from LD.

The signals with p-value $< 5 \times 10^{-8}$ are clumped together based on R^2 threshold of above 0.1 (moderate

LD) and a standard distance of 250kb.

```
system(paste("plink --bfile", filtered_data,"--allow-no-sex",
             "--clump gwas_results.qassoc --clump-p1 5e-8 --clump-r2 0.1
             --clump-kb 250 --out clumped_results"))

clumped_results <- fread("clumped_results.clumped")

n_independent_signals <- nrow(clumped_results)
```

```
## [1] "Number of independent signals: 2"
```

```
clumped_results
```

```
##      CHR     F        SNP        BP        P TOTAL  NSIG   S05   S01  S001
##    <int> <int>     <char>     <int>    <num> <int> <int> <int> <int> <int>
## 1:    15     1 rs72755233 100152748 1.34e-80     5     0     0     0     0
## 2:    15     1 rs27298763 100327445 5.78e-24     2     0     0     0     0
##    S0001                                                                SP2
##    <int>                                                              <char>
## 1:     5 rs45960959(1),rs31187023(1),rs48338810(1),rs38965480(1),rs10611351(1)
## 2:     2                                          rs49128648(1),rs24254005(1)
```

Short Summary Report

The raw genotypic data was first converted from VCF format to PLINK compatible formats. Concurrently, the phenotypic data was manually reformatted to ensure compatibility with PLINK, which involved adding Family ID (FID) and Individual ID (IID) columns and contains the same sample identifier information. The genotypic and phenotypic datasets were then combined and prepared for quality control (QC).

An initial review of the datasets revealed the following: there were 1,100 single nucleotide polymorphisms (SNPs) and 10,000 individual samples, all variants were from chromosome 15, and the sex information for both variants and individuals was unspecified. The phenotypic data was quantitative.

Standard QC filters were applied using PLINK[1] to exclude low-quality or poorly genotyped variants and individuals. Specifically, SNPs and individuals with missingness greater than 0.03, minor allele frequency (MAF) less than 0.01, and Hardy-Weinberg Equilibrium (HWE) p-values less than 1×10^{-6} were removed.

Sex discrepancy checks were not possible due to the lack of sex information. Relatedness and population stratification analyses were also not performed since the simulated data assumed these factors wouldn't impact the association analysis. In real-world datasets, however, controlling for these is crucial. Normally, a principal component analysis (PCA) would be used to control for population stratification, and relatedness would be addressed to avoid confounding.

Following QC, association analysis was conducted with PLINK[1]. Significant SNPs were identified using a genome-wide significance threshold of 5×10^{-8}. A total of 9 SNPs met this threshold and were considered genome-wide significant (GWS) (Table 1).

Table 1: Table of the genome-wide significant independent signals. Abbrev: CHR-chromosome, BP-basepair, NMISS-number of missing values, SE-standard error, R2-R squared, T-T statistic, P-p value.

CHR	SNP	BP	NMISS	BETA	SE	R2	T	P
15	rs45960959	100130220	9907	0.1292	0.01583	0.006683	8.163	3.656e-16
15	rs31187023	100140220	9907	0.1953	0.01708	0.013030	11.430	4.351e-30
15	rs48338810	100144139	9907	0.1980	0.01638	0.014540	12.090	2.036e-33
15	rs72755233	100152748	9907	0.3920	0.02043	0.035840	19.190	1.341e-80
15	rs38965480	100153138	9907	0.2221	0.01772	0.015600	12.530	9.820e-36
15	rs10611351	100154139	9907	0.2986	0.01861	0.025340	16.050	3.090e-57
15	rs49128648	100325138	9907	0.1168	0.01643	0.005076	7.109	1.254e-12
15	rs27298763	100327445	9907	0.1822	0.01800	0.010240	10.120	5.781e-24
15	rs24254005	100334139	9907	0.1666	0.01782	0.008749	9.350	1.065e-20

The analysis uses a GWS threshold of 5×10^{-8}, which is standard in genome-wide analysis studies (GWAS). However, to more accurately account for multiple testing and the risk of false positives, alternative methods such as a Bonferroni correction or controlling the false discovery rate (FDR) could be considered.

The Manhattan plot (Figure 1) demonstrates that the dataset is highly focused, with all variants within a specific region on chromosome 15. Notably, all variants are located between 9.90×10^7bp and 1.015×10^8bp. The plot highlights 9 significant SNPs, which are marked in orange, with p-values less than 5×10^{-8}.

The QQ plot (figure 2) also suggests that the GWAS analysis has identified several SNPs significantly associated with the phenotype of interest, as indicated by the upward deviation from the expected distribution.

Each of the 9 GWS independent signals was cross-referenced with existing databases and literature to assess whether these signals had been previously reported. The SNP identifiers (rs number) for each signal were checked across NCBI databases, including PubMed, ClinVar, and dbSNP. Additionally, the GWAS Catalog was queried to determine if these SNPs had been identified in previous GWAS studies.

Of the 9 GWS signals, 8 appear to be novel. The SNP rs72755233, located at chromosome 15:100152748, was an exception. According to the GWAS Catalog, this SNP has been reported in 39 previous GWAS studies[2], and it has also appeared in 12 PubMed published studies. These studies link rs72755233 to various traits, including the risk of pancreatitis[3], height and weight[4], carpal tunnel syndrome[5], and ocular conditions[6][7]. Despite these associations, the variant is primarily classified as benign.

Notably, all the GWS signals were linked to the *ADAMTS17* gene, likely due to their close proximity to this gene.

A potential next step would be to perform functional annotation of the significant SNPs to investigate their potential regulatory or functional roles.

To ensure the independence of GWS signals and to account for potential linkage disequilibrium (LD), signals located in close proximity were carefully looked at. GWS signals that clustered together were identified using an R^2 threshold of 0.1 (indicating moderate LD) and a standard distance of 250 kb. After applying these criteria, only 2 independent GWS signals remained (Table 2).

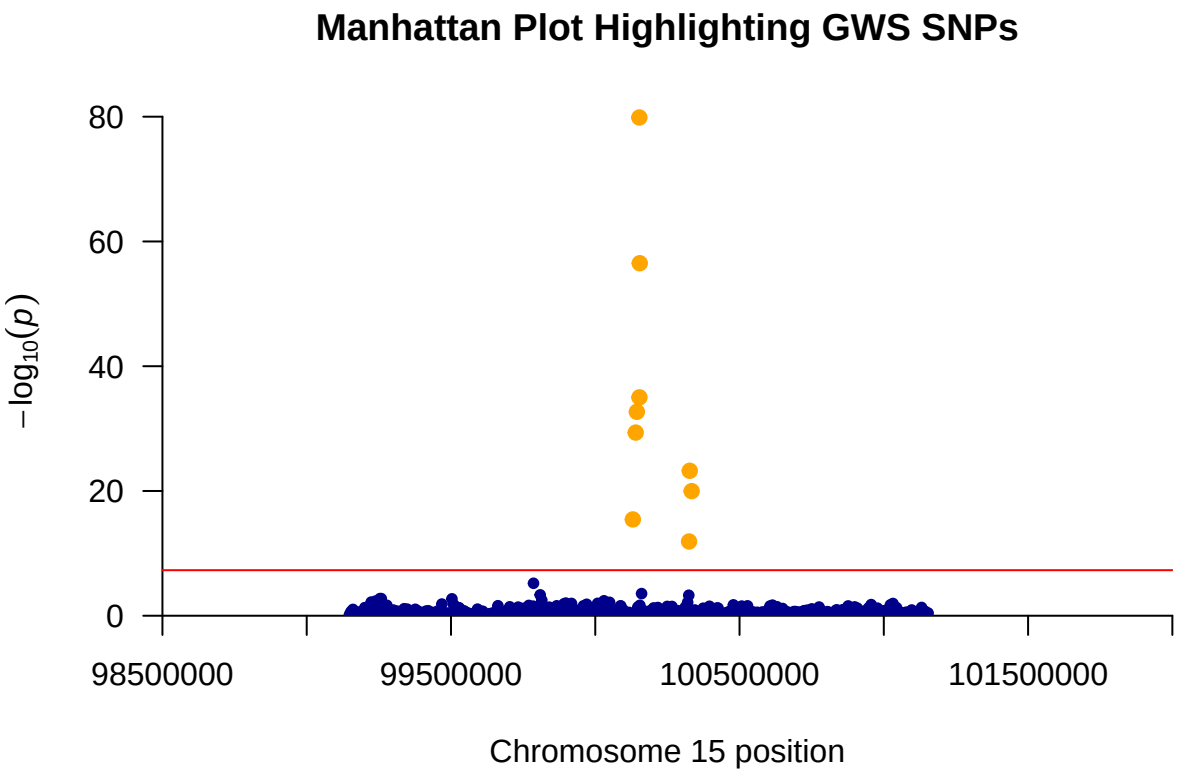


Figure 1: Manhattan plot of the GWAS results of the index variants along chromosome 15. GWS < 5e-08 is marked by the red line. GWS independent loci are highlight as orange. -log10(p) values are plotted against an amplified view of the chromosomal positions (chromosome 15 between 9.85e7bp - 1.02e8bp).

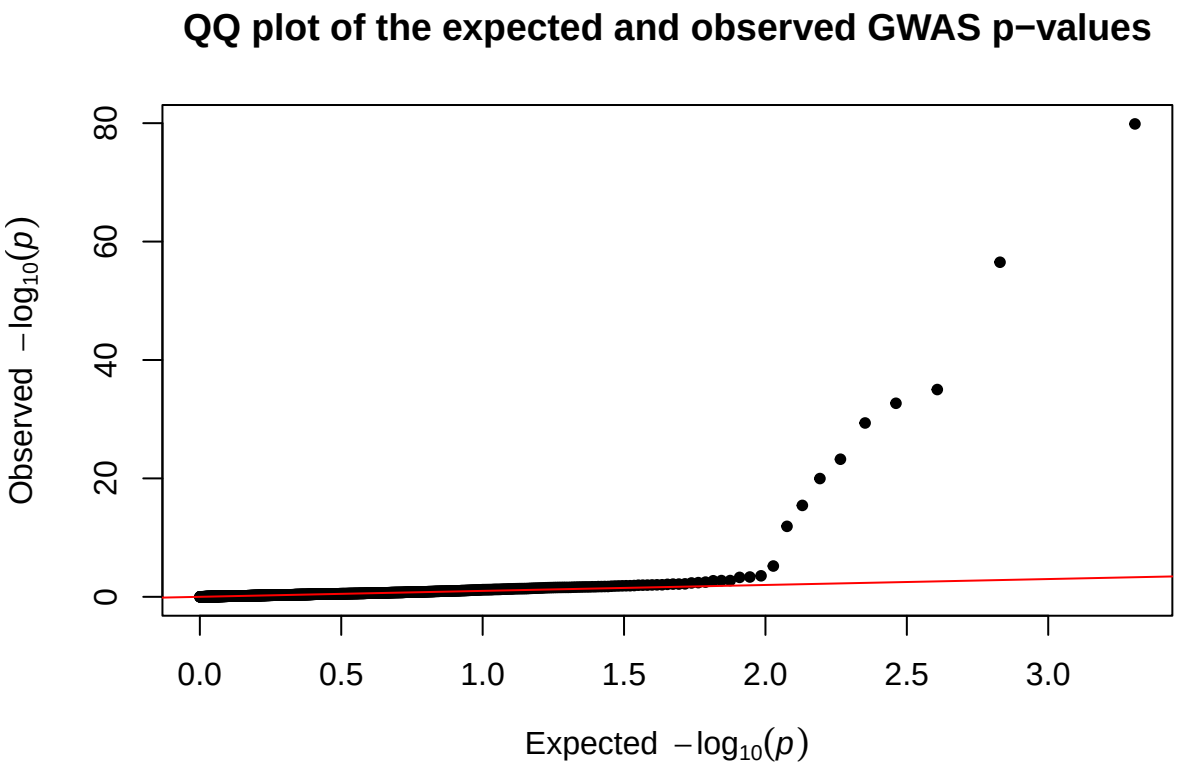


Figure 2: QQ plot of the observed -log10(p) generated by the GWAS analysis against the expected values. The red line shows the expected distribution of SNP p-values under the null hypothesis (disbribution of the expected vs the observed -log10(p) are the same).

Table 2: Table of the grouped genome-wide significant independent signals and associated p-values. Abbrev: CHR-chromosome, BP-basepair, S001-number of SNPs with p-values < 0.001, SP2-list of SNPs clustered into the locus.

CHR	SNP	BP	S001	SP2
15	rs72755233	100152748	5	rs45960959(1),rs31187023(1),rs48338810(1),rs38965480(1),rs10611351(1)
15	rs27298763	100327445	2	rs49128648(1),rs24254005(1)

One of these signals was the SNP rs72755233, which stood out due to its exceptionally low p-value (1.341×10^{-80}). This SNP was retained in the final analysis, not only because of its significant association but also due to the overwhelming statistical evidence, suggesting that it represents a genuine signal rather than one arising from LD with nearby variants (or from the random generation of data). This strong signal is likely the reason why multiple studies have noted and found associations with rs72755233, linking it to various traits and conditions.

However, considering the variants are focused around such a specific region of the chromosome, there's a risk of covering only a limited genomic context. This concentration might lead to an overemphasis on local linkage disequilibrium patterns. Therefore, it may be more appropriate to consider the individual GWS signals as independent signals rather than as part of a clustered locus. By analysing each signal separately, it can better assess their distinct contributions to the phenotype, reducing the risk of conflating nearby variants that may have independent effects.

A limitation of this analysis is the small subset of genome-wide data, which may result in reduced statistical power and increase the risk of false negatives, potentially missing true associations. Expanding the dataset would likely yield reliable findings.

While the simulated data has helped certain aspects of the analysis by providing a controlled environment, it also limits the applicability of the findings to real-world scenarios. In contrast, with real-world data, findings would be validated through the use of additional independent datasets and replication studies.

To summarise, QC methods were applied to filter the raw genomic and phenotypic data. Subsequently, an association analysis was performed to identify GWS independent signals. A total of 9 significant signals were detected and then compared with the current literature to determine if they had been previously reported in other studies. Out of these, only one signal (rs72755233) was found to have been associated with traits in prior studies.

Bibliography

1. Purcell, S. M. 2009. PMID: 19571811.
2. Sollis, E. 2022. PMID: 36350656.
3. Abaji, R. 2017. PMID: 28574850.
4. Tachmazidou, I. 2017. PMID: 28552196.
5. Wiberg, A. 2019. PMID: 30833571.
6. Simcoe, M. J. 2020. PMID: 32716492.
7. Choquet. H. 2020. PMID: 32528159.

```
gc()
```

```
##            used  (Mb) gc trigger  (Mb) max used  (Mb)
## Ncells  2023764 108.1    3807540 203.4  2935626 156.8
## Vcells 14631838 111.7   32617805 248.9 27419759 209.2
```

```
length(getLoadedDLLs())
```

```
## [1] 36
```

```
sessionInfo()
```

```
## R version 4.3.1 (2023-06-16 ucrt)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 11 x64 (build 22631)
##
## Matrix products: default
##
##
## locale:
## [1] LC_COLLATE=English_United Kingdom.utf8
## [2] LC_CTYPE=English_United Kingdom.utf8
## [3] LC_MONETARY=English_United Kingdom.utf8
## [4] LC_NUMERIC=C
## [5] LC_TIME=English_United Kingdom.utf8
##
## time zone: Europe/London
## tzcode source: internal
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] kableExtra_1.4.0  qqman_0.1.9       data.table_1.16.0 vcfR_1.15.0
##
## loaded via a namespace (and not attached):
##  [1] Matrix_1.6-5      highr_0.10        vegan_2.6-4       compiler_4.3.1
##  [5] Rcpp_1.0.11       xml2_1.3.6        stringr_1.5.1     parallel_4.3.1
##  [9] pinfsc50_1.3.0    cluster_2.1.4     splines_4.3.1     systemfonts_1.0.5
## [13] scales_1.3.0      yaml_2.3.7        fastmap_1.1.1     lattice_0.21-8
## [17] R6_2.5.1          knitr_1.45        MASS_7.3-60       munsell_0.5.0
## [21] svglite_2.1.3     rlang_1.1.2       calibrate_1.7.7   stringi_1.8.1
## [25] xfun_0.41         viridisLite_0.4.2 cli_3.6.1         magrittr_2.0.3
## [29] mgcv_1.9-1        digest_0.6.33     grid_4.3.1        rstudioapi_0.15.0
## [33] permute_0.9-7     lifecycle_1.0.4   nlme_3.1-162      evaluate_0.23
## [37] glue_1.6.2        memuse_4.2-3      ape_5.7-1         colorspace_2.1-0
## [41] rmarkdown_2.25    tools_4.3.1       htmltools_0.5.7
```