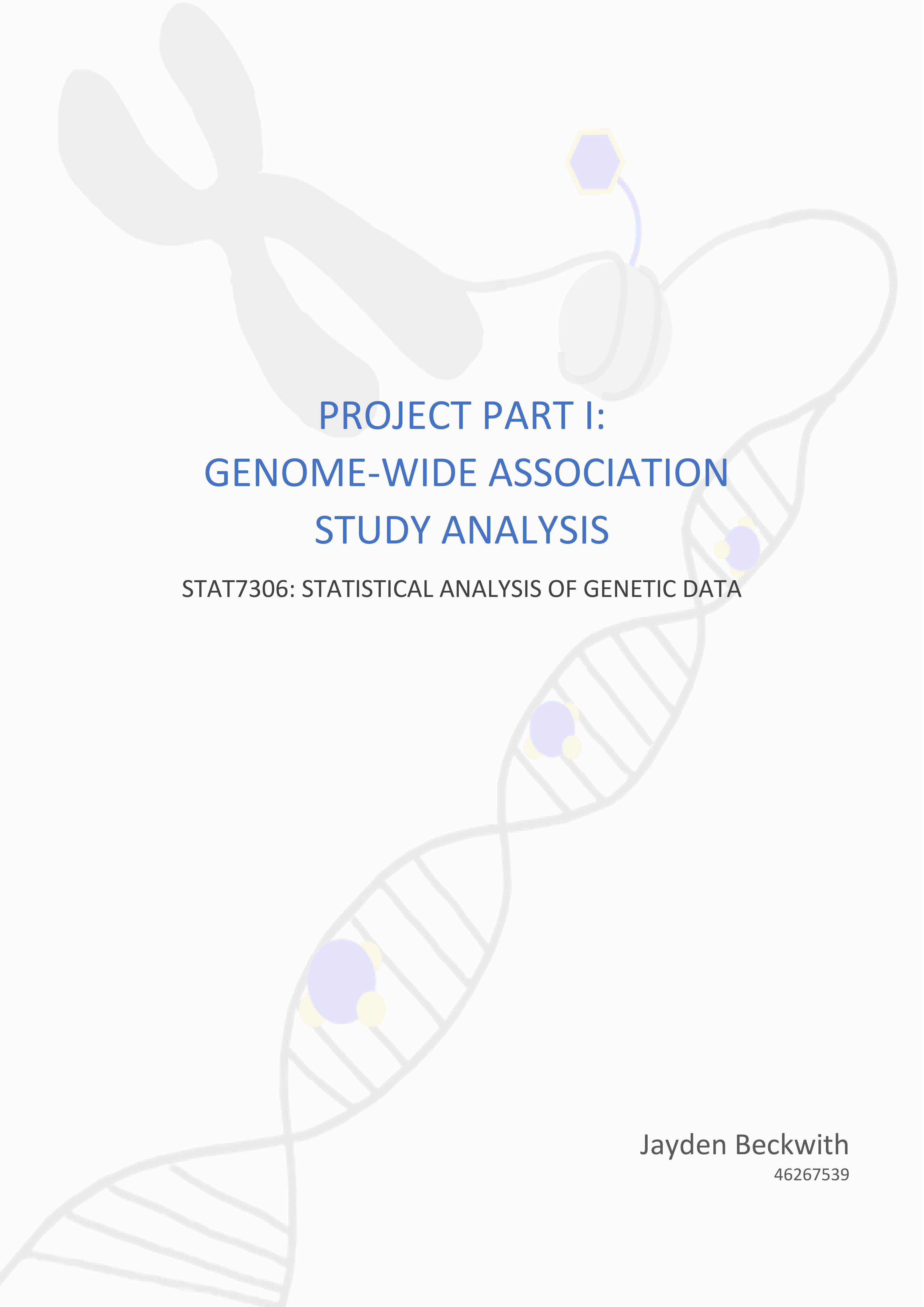


PROJECT PART I: GENOME-WIDE ASSOCIATION STUDY ANALYSIS

STAT7306: STATISTICAL ANALYSIS OF GENETIC DATA

Jayden Beckwith

46267539

Contents

Introduction

The aim of this investigation was to analyse HapMap 3 singular nucleotide polymorphisms (SNPs) for ~ 11,000 people from the UK Biobank study and identify genomic regions that are linked to the fasting glucose (FG) phenotype. The FG phenotype was of targeted interest as it may be useful in identifying underlying genomic information for metabolic conditions. Genome-wide association study (GWAS) analysis is a popular and effective technique for identifying genetic variations within population studies and was used to find genomic regions associated with the FG phenotype. The raw SNP data was represented in 3 formats consisting of test.bim, test.fam and test.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

Figure #1 represents the summary breakdown of the methods performed in the analysis.

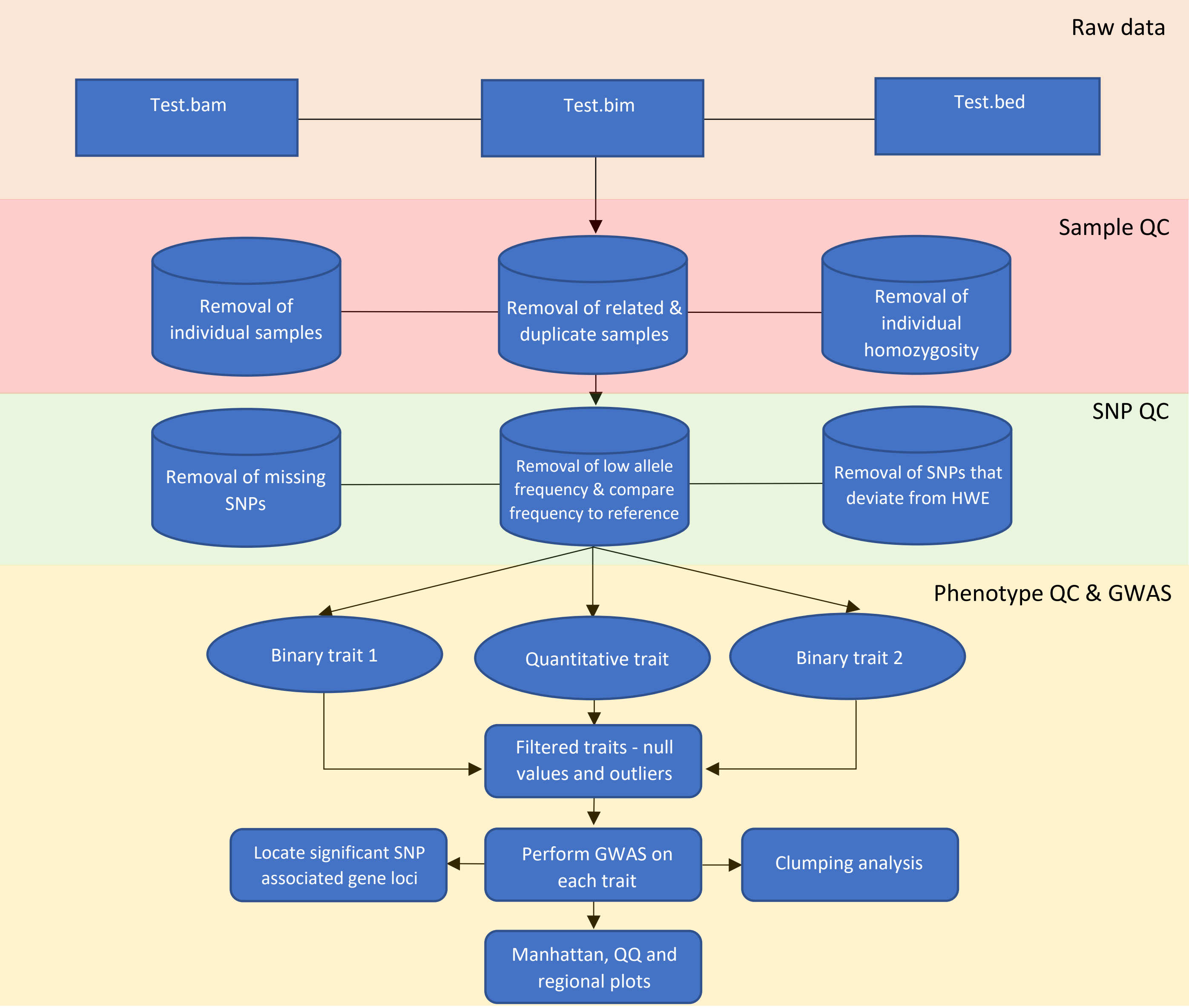

Raw data
Test.bam
Test.bim
Test.bed
Sample QC
Removal of individual samples
Removal of related & duplicate samples
Removal of individual homozygosity
SNP QC
Removal of missing SNPs
Removal of low allele frequency & compare frequency to reference
Removal of SNPs that deviate from HWE
Phenotype QC & GWAS
Binary trait 1
Quantitative trait
Binary trait 2
Filtered traits - null values and outliers
Locate significant SNP associated gene loci
Perform GWAS on each trait
Clumping analysis
Manhattan, QQ and regional plots

Figure #1 – Breakdown of the methods performed on the SNP genotype and fasting glucose phenotype data.

All QC and GWAS steps were performed using Plink and R on UQ's cluster. Sample QC was performed initially to remove any individual samples that had excess missing genotypes followed by removal of individuals that had outlying homozygosity values. The removal of outlying homozygous values is important due to possible variations across populations and genotyping platforms. Histograms and index significance plots were developed at a threshold > 0.05 for the homozygosity filter and missing genotypes (refer to supplementary materials). One major limitation observed during the sample QC phase was the inability to remove discordant sex. This was due to the HapMap genotypes not containing any polymorphic sex chromosomes (often coded as chromosome 23 in humans). Therefore, it was not possible to perform any filtering per sex genotype as the data only ranged up to chromosome 22. Removing relational samples was performed and set at a relative identity-by-descent (IBD) < 0.05, ensuring that all samples were unrelated and did not share any relatedness bias. Additionally, it was also assumed that removal of ancestral outliers was already performed prior to receiving the dataset.

After sampling QC was completed, the newly filtered subset data was used for SNP QC. SNP QC consisted of 3 filtering steps including removal of missing SNPs, low allele frequency and removal of SNPs that deviate from Hardy-Weinberg equilibrium (HWE). Similarly, removal of excessive missing SNP genotypes was performed based on an F_MISS threshold > 0.05. Furthermore, removal of SNPs that deviate from HWE was set at a threshold value P < 0.001. HWE is an important filter as departing from HWE is usually indicative of genotyping errors or the Wahlund effect. Filtering by minor allele frequency (MAF) at a threshold value of < 0.01 and was compared to the known allele reference frequency values. MAF is an important criterion for GWAS as the loci with low MAF have low heterozygosity and are less informative for the study. Therefore, this data was further filtered to remove any resulting null values within the given dataset.

Phenotype QC was performed on the 3 traits to remove any outliers within the datasets. Boxplots were developed to detect if any outliers were present in the quantitative phenotype. Subsequently, once the data was properly filtered GWAS was performed on the SNP dataset. GWAS results were illustrated using Manhattan, QQ and regional association plots for the 3 targeted traits (refer to supplementary materials for QQ plots). Lambda values were also determined for each trait to identify if principal component covariates needed to be added. However, due to the genomic inflation (lambda coefficients) being close to 1, no covariates were added to the dataset. To find the regions with the high association signals, SNPs were clumped into groups. Clumping analysis was performed using plink to provide a linkage disequilibrium-based procedure to determine how many association signals are based on the correlation between variants and p-values for each trait (clumping parameters were chosen based on a trial-and-error approach: p1 = 0.001, p2 = 0.01, R^2 = 0.50, clump distance = 250 kb). Once significant associated signals were identified, NCBI and genome UCSC browser was used to target the subsequent gene within the SNP loci.

Results

Significant signals were identified within the fasting glucose quantitative trait. Table #1 represents a breakdown of the most significant SNP signals identified from Figure 2A.

Table #1 – Top 3 Highest SNP Association Signals for Fasting Glucose Quantitative Trait

SNP Id	Chromosomal loci	P-value	Associated gene
1. rs560887	Chr 2: q24.3	7.19×10^{-44}	G6PC2
2. rs569805	Chr 2: q24.3	3.72×10^{-39}	ABCB11
3. rs12719860	Chr 5: q35.3	2.29×10^{-36}	ZNF454

Clumping analysis revealed key genomic markers with significant association across chromosome 2, 6, 8 and 19. Chromosome 8 had the highest clumped region with 10 signals found at the loci of 145 Mb (refer to Figure 2B and Table #2). The top SNP identified in that region was rs11777402 which was associated within the PLEC gene. However, the highest statistically significant SNP associated across all chromosomes was observed to be rs560887 and was determined to be in the region of G6PC2 gene loci.

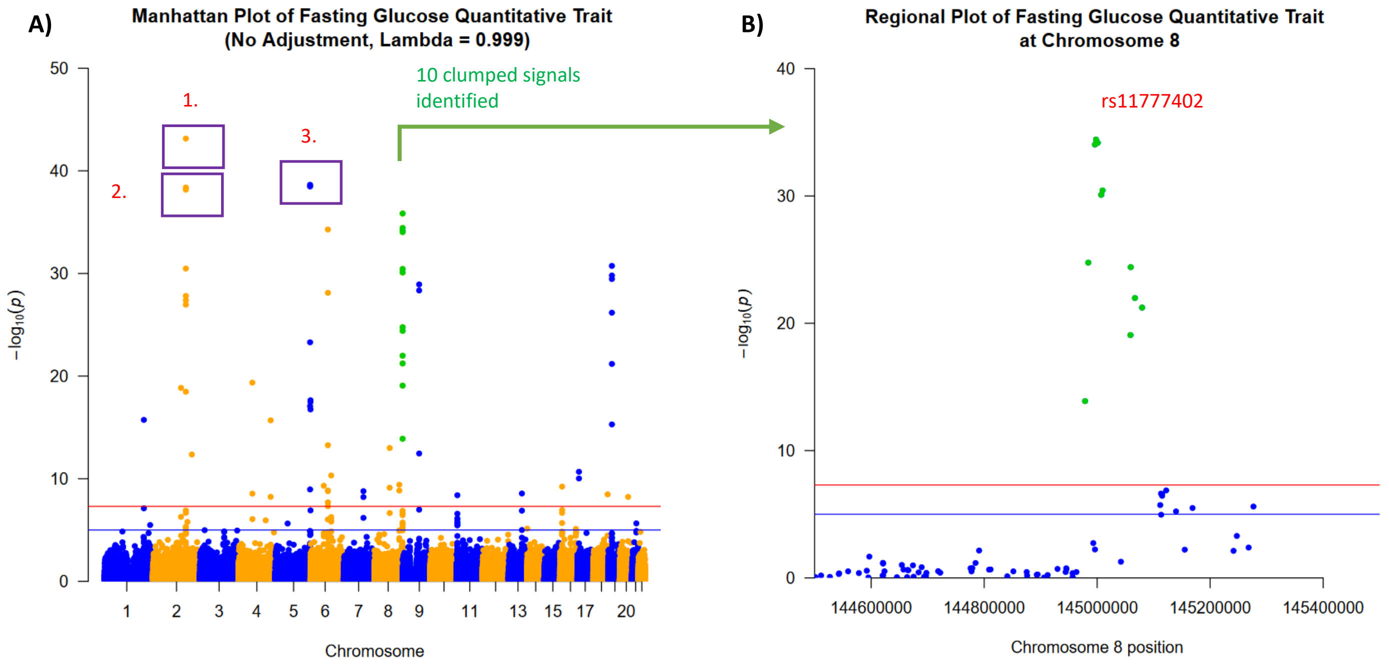


Figure #2 – A) Manhattan plot of fasting glucose quantitative trait with highlighted SNPs of interest. B) Regional plot of fasting glucose quantitative trait at loci chromosome 8 q24.3, 145 Mb.

Table #2 represents a breakdown of the SNPs identified to have high association from clumping analysis, with the corresponding genes and loci of the quantitative trait.

Table #2 – Top 3 significant associated SNP signals of the quantitative traits from clumping analysis

Top SNP Id in associated region	Chromosomal loci	P-value	Associated gene/s	Total significant signals
rs11777402	Chr 8: q24.3	1.45×10^{-36}	PLEC	10
rs73379171	Chr 8: q24.3	1.34×10^{-7}	EXOSC4	7
rs7648	Chr 19: p13.11	1.84×10^{-31}	KXD1, UBA52	5

Similarly, Manhattan plots were developed for both binary traits with corresponding lambda values as depicted in Figure 3. Subsequently, clumping analysis was performed on both binary traits and compared against the quantitative trait. It was found that the most statistically significant SNPs followed closely to the quantitative trait with only minor differences in log scores.

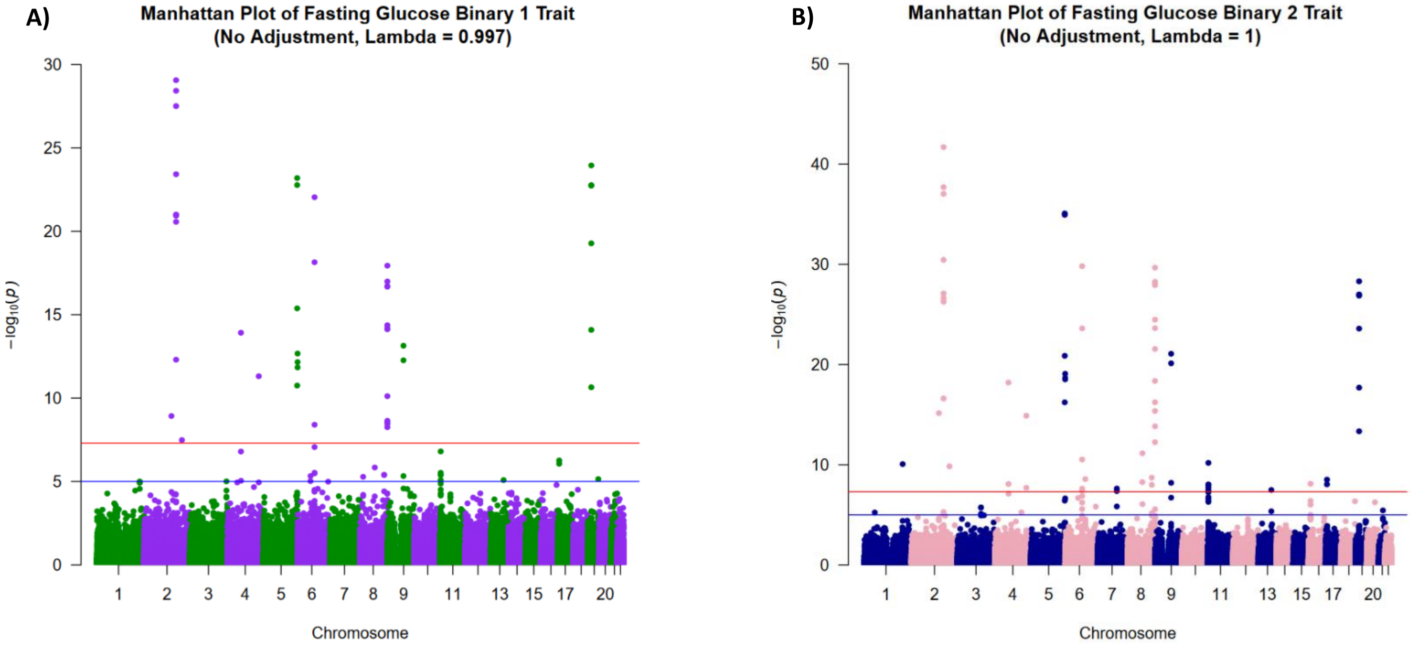


Figure #3 – A) Manhattan plot of fasting glucose binary trait 1. B) Manhattan plot of fasting glucose binary trait 2.

Clumping analysis was also performed on both binary traits and it was found that the top 3 significant signals was the same as the quantitative trait, this included rs11777402, rs7648, rs17324609 SNPs.

Discussion

GWAS analysed 11,793 participants from the UK biostudy sample (female = 6442, male = 5351) and found significant associations for the FG phenotype. It was evident that the binary traits followed closely to the quantitative trait indicating that there was little difference between them. It was observed that binary trait 1 had weaker SNP signals in comparison with binary trait 2 and the quantitative trait. Since binary trait 2 represents the top 20% of the phenotype and bottom 30%, it covers a wider distribution of the dataset than binary trait 1. Therefore, based on this and the developed Manhattan plots, it was suggestive that binary trait 2 was a better binary representation of the FG trait as it followed more closely to the quantitative counterpart.

It was found that the top statistically significant SNPs were predominately within the loci of G6PC2, ABCB11, ZNF454 and PLEC genes across the quantitative and binary traits. The rs569805 ($P = 7.19 \times 10^{-44}$) and rs560887 ($P = 3.72 \times 10^{-39}$) SNPs had the most significant p-values and were located on G6PC2 and ABCB11 chromosome 2, q24.3. In combination with data from literature, it is suggestive that G6PC2 is exclusively expressed in the pancreatic isle cells and ABCB11 is expressed in the liver (Chen et al., 2008). Subsequently, there is supporting evidence that these variants may have underlying influence in FG concentration. It was identified that G6PC2 encodes the islet-specific glucose-6-phosphatase (G6P) related peptide (IGRP). From literature, it has been found that IGRP activity in beta cells dephosphorylates G6P to form glucose, opposing the action of glucokinase (Rose et al., 2009). A GWAS conducted by Rose et al. (2009) suggested that the SNP variants found in GC6PC2 and ABCB11 increase the expression of IGRP, decreasing the glucose into the glycolytic pathway and thus upregulating the impact of the concentration of plasma glucose on insulin secretion in the fasting state. Furthermore, variations in ABCB11 have been identified to play a role in the interruption between bile acid re-absorption and decrease plasma glucose to improve insulin sensitivity (Takeuchi et al., 2010). Additionally, a study published by Zhu et al. (2020) identified that ZNF454 is a protein-coding gene that is involved in various functional pathways. However, other zinc finger protein coding genes have been identified to regulate insulin stimulation and promote glucose uptake, such as ZFP407 (Buchner et al., 2015). Therefore, ZNF454 may have a similar association and play a role in the FG phenotype. Clumping analysis revealed that there exist significant signals between SNPs associated within the PLEC gene loci on chromosome 8, which may play a role in FG concentration. PLEC codes for the plectin, a protein which maintains tissue integrity and has direct sarcolemma interactions that effect glucose reuptake (Raith et al., 2013). Therefore, this may have an association with FG concentration within the sample population. Ultimately, additional replication studies would need to be conducted with a larger population size to validate the identified genetic variants for FG concentration.

Conclusion

The aim of this investigation was to analyse HapMap 3 SNPs from the UK Biobank study and identify genomic regions that are linked to the FG phenotype. Several QC steps were used to perform GWAS on the targeted 3 FG traits. Ultimately, the most statistically significant SNPs identified were related to the G6PC2 and ABCB11 genes which have been referenced throughout literature to be highly associated with the FG phenotype. Several other genes were also identified, including ZNF454 and PLEC which may have questionable relevance towards FG concentration. Therefore, additional replication studies are necessary to rule out any false positive signals and validate the results obtained from this investigation.

References

Buchner, D. A., Charrier, A., Srinivasan, E., Wang, L., Paulsen, M. T., Ljungman, M., Bridges, D., & Saltiel, A. R. (2015). Zinc finger protein 407 (ZFP407) regulates insulin-stimulated glucose uptake and glucose transporter 4 (Glut4) mRNA. *J Biol Chem*, *290*(10), 6376-6386. DOI: 10.1074/jbc.M114.623736.

Chen, W. M., Erdos, M. R., Jackson, A. U., Saxena, R., Sanna, S., Silver, K. D., Timpson, N. J., Hansen, T., Orru, M., Grazia Piras, M., Bonnycastle, L. L., Willer, C. J., Lyssenko, V., Shen, H., Kuusisto, J., Ebrahim, S., Sestu, N., Duren, W. L., Spada, M. C., Watanabe, R. M. (2008). Variations in the G6PC2/ABCB11 genomic region are associated with fasting glucose levels. *J Clin Invest*, *118*(7), 2620-2628. DOI: 10.1172/JCI34566.

Raith, M., Valencia, R. G., Fischer, I., Orthofer, M., Penninger, J. M., Spuler, S., Rezniczek, G. A., & Wiche, G. (2013). Linking cytoarchitecture to metabolism: sarcolemma-associated plectin affects glucose uptake by destabilizing microtubule networks in mdx myofibers. *Skelet Muscle*, *3*(1), 14. DOI: 10.1186/2044-5040-3-14.

Rose, C. S., Grarup, N., Krarup, N. T., Poulsen, P., Wegner, L., Nielsen, T., Banasik, K., Faerch, K., Andersen, G., Albrechtsen, A., Borch-Johnsen, K., Clausen, J. O., Jorgensen, T., Vaag, A., Pedersen, O., & Hansen, T. (2009). A variant in the G6PC2/ABCB11 locus is associated with increased fasting plasma glucose, increased basal hepatic glucose production and increased insulin release after oral and intravenous glucose loads. *Diabetologia*, *52*(10), 2122-2129. DOI: 10.1007/s00125-009-1463-z.

Takeuchi, F., Katsuya, T., Chakrewarthy, S., Yamamoto, K., Fujioka, A., Serizawa, M., Fujisawa, T., Nakashima, E., Ohnaka, K., Ikegami, H., Sugiyama, T., Nabika, T., Kasturiratne, A., Yamaguchi, S., Kono, S., Takayanagi, R., Yamori, Y., Kobayashi, S., Ogihara, T., Kato, N. (2010). Common variants at the GCK, GCKR, G6PC2-ABCB11 and MTNR1B loci are associated with fasting glucose in two Asian populations. *Diabetologia*, *53*(2), 299-308. DOI: 10.1007/s00125-009-1595-1.

Zhu, Q., Wang, J., Zhang, Q., Wang, F., Fang, L., Song, B., Xie, C., & Liu, J. (2020). Methylationdriven genes PMPCAP1, SOWAHC and ZNF454 as potential prognostic biomarkers in lung squamous cell carcinoma. *Mol Med Rep*, *21*(3), 1285-1295. DOI: 10.3892/mmr.2020.10933.

Supplementary Materials

Sample Quality Control

Plink and Unix commands

Filter missing samples – 18 samples

```
/data/STAT3306/plink --bfile raw_data/test --missing --out plink_out/plink
```

Filter homozygosity – 13 samples

```
/data/STAT3306/plink --bfile raw_data/test --het --out plink_out/plink
```

Filter duplicates – 245 samples

```
cut -f 4 raw_data/test.bim | sort | uniq -d > dup.snps
wc -l dup.snps
/data/STAT3306/plink --bfile raw_data/test --exclude dup.snps --make-bed --out filt_dup
```

Filter related SNPs with relative cut off < 0.05

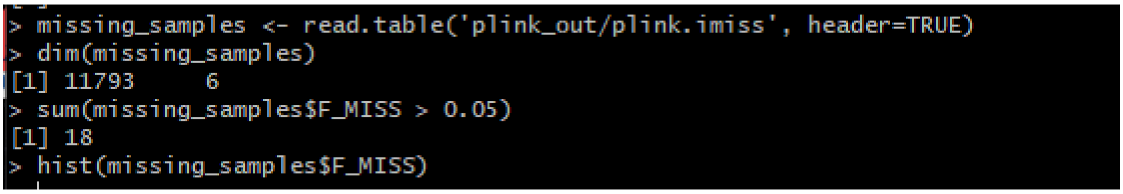
```
/data/STAT3306/plink --bfile filt_dup/filt_dup --genome --grm-cutoff 0.05 --make-bed --out
filt_dup/filt_dup
```

Apply sample QC filter

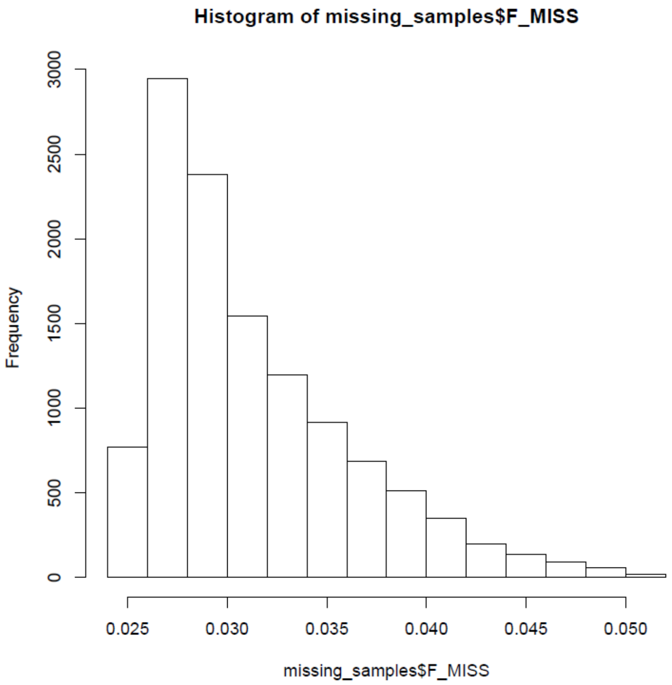
```
/data/STAT3306/plink --bfile filt_dup/filt_dup --exclude plink_out/remove.samples.for.qc.txt --
make-bed --out post_sample_qc/samp_filt
```

R commands and plots for sample QC

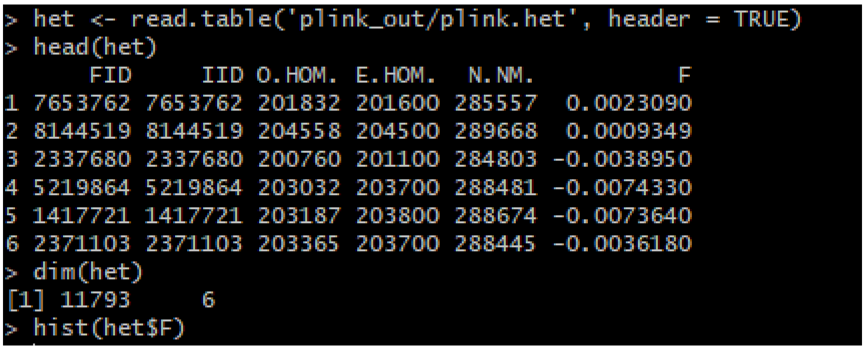
18 individual missing samples and plot histogram.



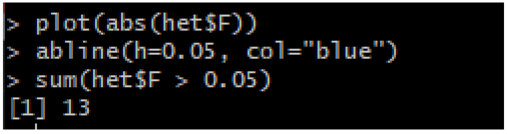
```
> missing_samples <- read.table('plink_out/plink.imiss', header=TRUE)
> dim(missing_samples)
[1] 11793     6
> sum(missing_samples$F_MISS > 0.05)
[1] 18
> hist(missing_samples$F_MISS)
```



Removal of individuals with outlying homozygosity values.

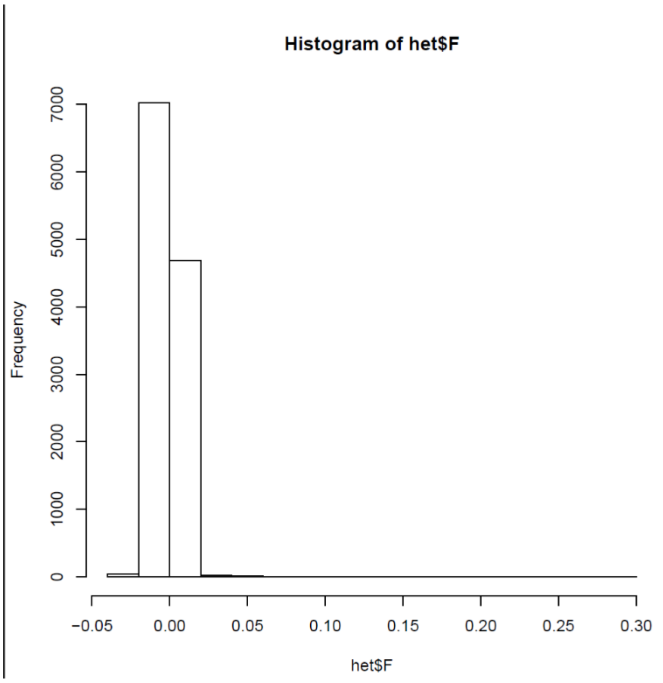


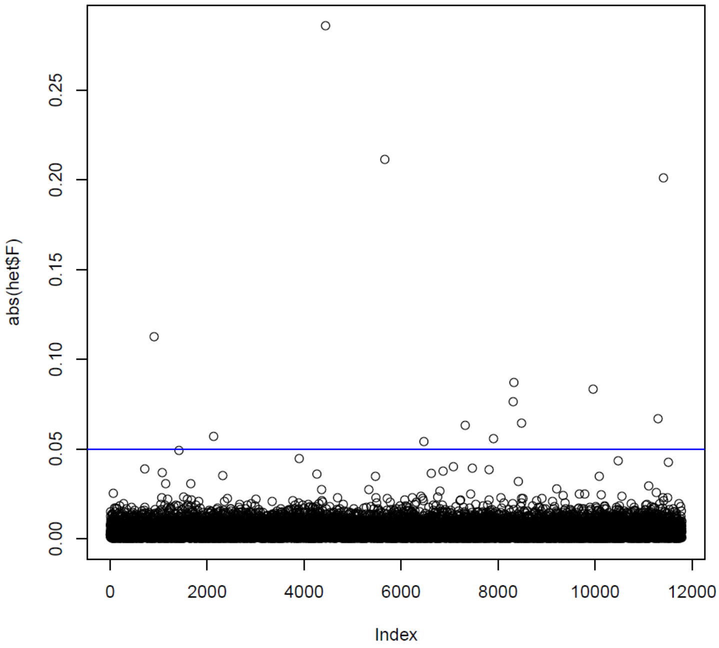
```
> het <- read.table('plink_out/plink.het', header = TRUE)
> head(het)
      FID     IID O.HOM. E.HOM.   N.NM.          F
1 7653762 7653762 201832 201600 285557  0.0023090
2 8144519 8144519 204558 204500 289668  0.0009349
3 2337680 2337680 200760 201100 284803 -0.0038950
4 5219864 5219864 203032 203700 288481 -0.0074330
5 1417721 1417721 203187 203800 288674 -0.0073640
6 2371103 2371103 203365 203700 288445 -0.0036180
> dim(het)
[1] 11793     6
> hist(het$F)
```



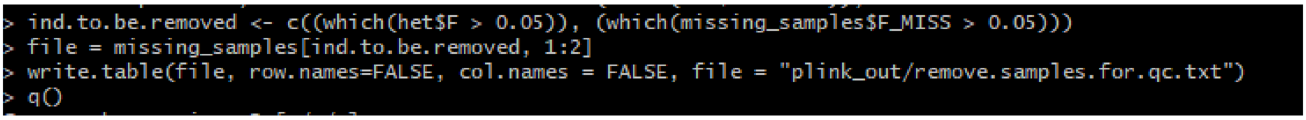
```
> plot(abs(het$F))
> abline(h=0.05, col="blue")
> sum(het$F > 0.05)
[1] 13
```

Plot of significant homozygous values to be removed = 13 in total.





Writing new table to remove outlying values



```
> ind.to.be.removed <- c((which(het$F > 0.05)), (which(missing_samples$F_MISS > 0.05)))
> file = missing_samples[ind.to.be.removed, 1:2]
> write.table(file, row.names=FALSE, col.names = FALSE, file = "plink_out/remove.samples.for.qc.txt")
> q()
```

SNP Quality Control

Plink and Unix commands

Removal of SNPs with excess genotype missing – 18.2k SNPs

```
/data/STAT3306/plink --bfile post_sample_qc/samp_filt --missing --out plink_out/plink
```

Removal of SNPs that deviate from HWE – 3486 SNPs

```
/data/STAT3306/plink --bfile post_sample_qc/samp_filt --hardy --out plink_out/plink
```

Removal of SNPs with low minor Allele frequency – 5.2k SNPs

```
/data/STAT3306/plink --bfile post_sample_qc/samp_filt --freq --out plink_out/plink
```

Applying SNP QC

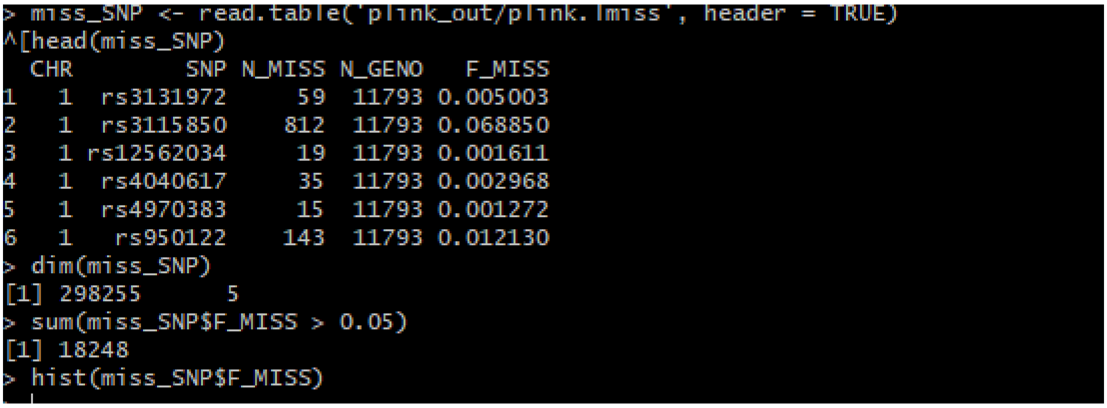
```
/data/STAT3306/plink --bfile post_sample_qc/samp_filt --make-bed --exclude
plink_out/remove.SNPs.txt --out post_SNP_qc/final_QC
```

Filter low allele frequency

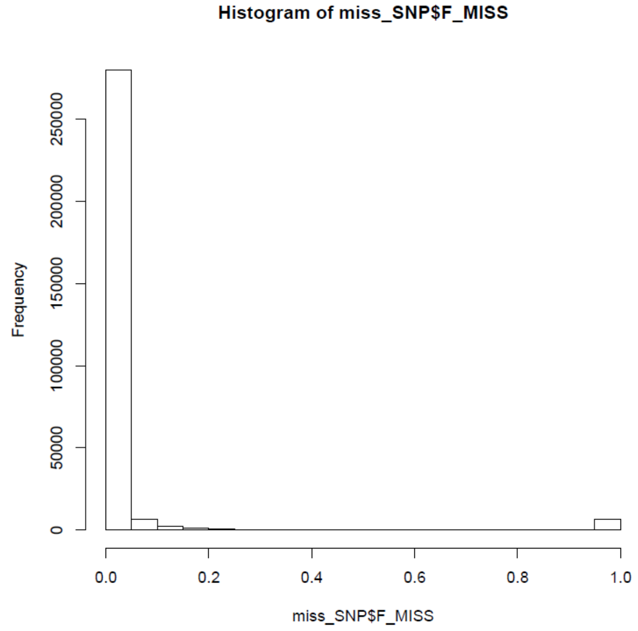
```
/data/STAT3306/plink --bfile post_SNP_qc/final_QC --freq --out plink_out/f_plink
```

R commands and plots

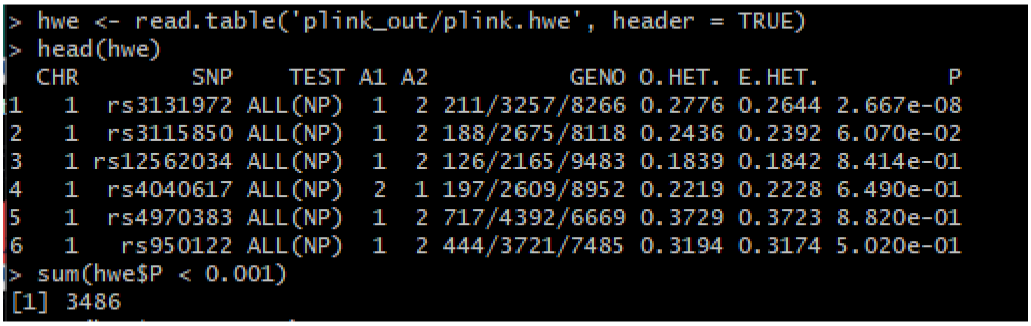
Reading in plink output for excessive genotypes missing and plotting the values.



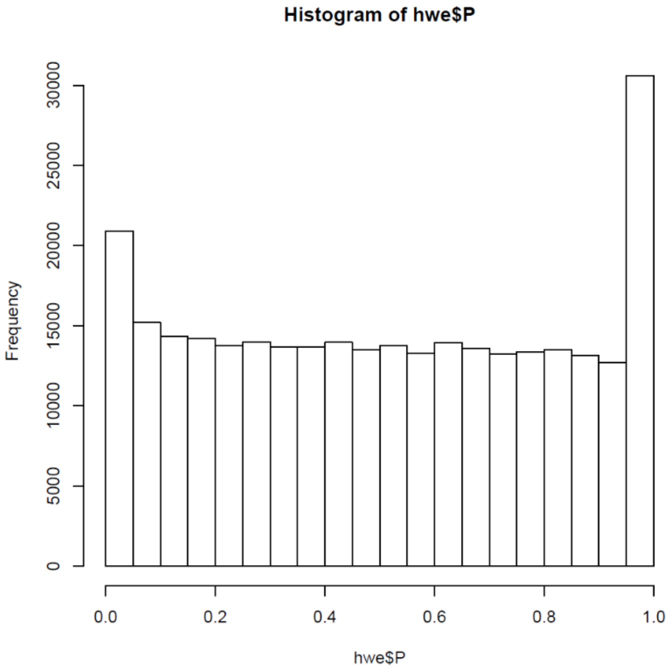
```
> miss_SNP <- read.table('plink_out/plink.lmiss', header = TRUE)
^[head(miss_SNP)
  CHR        SNP N_MISS N_GENO   F_MISS
1   1  rs3131972     59  11793 0.005003
2   1  rs3115850    812  11793 0.068850
3   1 rs12562034     19  11793 0.001611
4   1  rs4040617     35  11793 0.002968
5   1  rs4970383     15  11793 0.001272
6   1   rs950122    143  11793 0.012130
> dim(miss_SNP)
[1] 298255      5
> sum(miss_SNP$F_MISS > 0.05)
[1] 18248
> hist(miss_SNP$F_MISS)
```



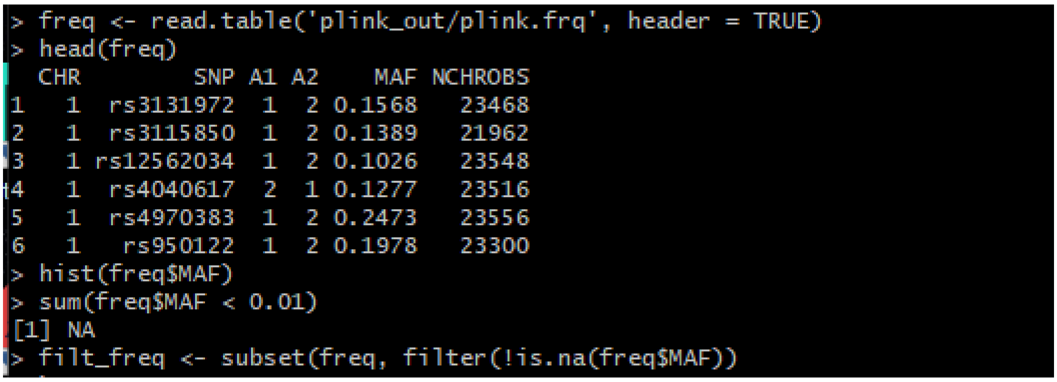
Reading in plink output for HWE and plotting the output



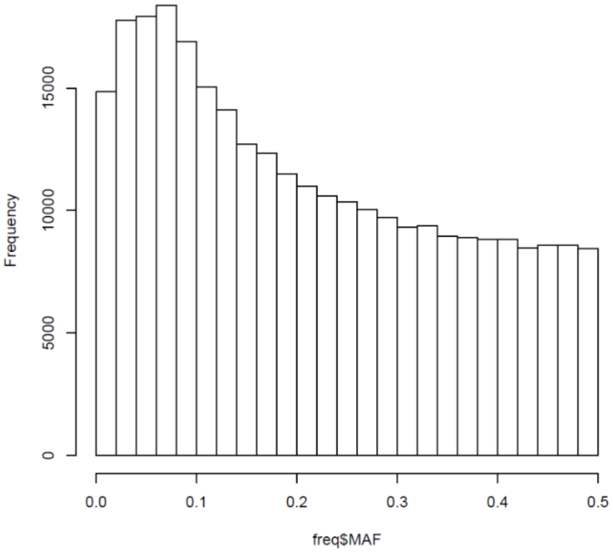
```
> hwe <- read.table('plink_out/plink.hwe', header = TRUE)
> head(hwe)
  CHR        SNP    TEST A1 A2           GENO O.HET. E.HET.         P
1   1  rs3131972 ALL(NP)  1  2 211/3257/8266 0.2776 0.2644 2.667e-08
2   1  rs3115850 ALL(NP)  1  2 188/2675/8118 0.2436 0.2392 6.070e-02
3   1 rs12562034 ALL(NP)  1  2 126/2165/9483 0.1839 0.1842 8.414e-01
4   1  rs4040617 ALL(NP)  2  1 197/2609/8952 0.2219 0.2228 6.490e-01
5   1  rs4970383 ALL(NP)  1  2 717/4392/6669 0.3729 0.3723 8.820e-01
6   1   rs950122 ALL(NP)  1  2 444/3721/7485 0.3194 0.3174 5.020e-01
> sum(hwe$P < 0.001)
[1] 3486
```



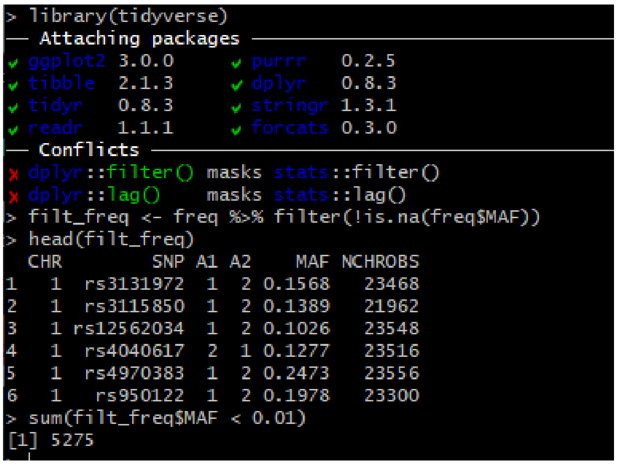
Reading in allele frequency output from plink and plotting the output



```
> freq <- read.table('plink_out/plink.frq', header = TRUE)
> head(freq)
  CHR        SNP A1 A2    MAF NCHROBS
1   1  rs3131972  1  2 0.1568   23468
2   1  rs3115850  1  2 0.1389   21962
3   1 rs12562034  1  2 0.1026   23548
4   1  rs4040617  2  1 0.1277   23516
5   1  rs4970383  1  2 0.2473   23556
6   1   rs950122  1  2 0.1978   23300
> hist(freq$MAF)
> sum(freq$MAF < 0.01)
[1] NA
> filt_freq <- subset(freq, filter(!is.na(freq$MAF))
```

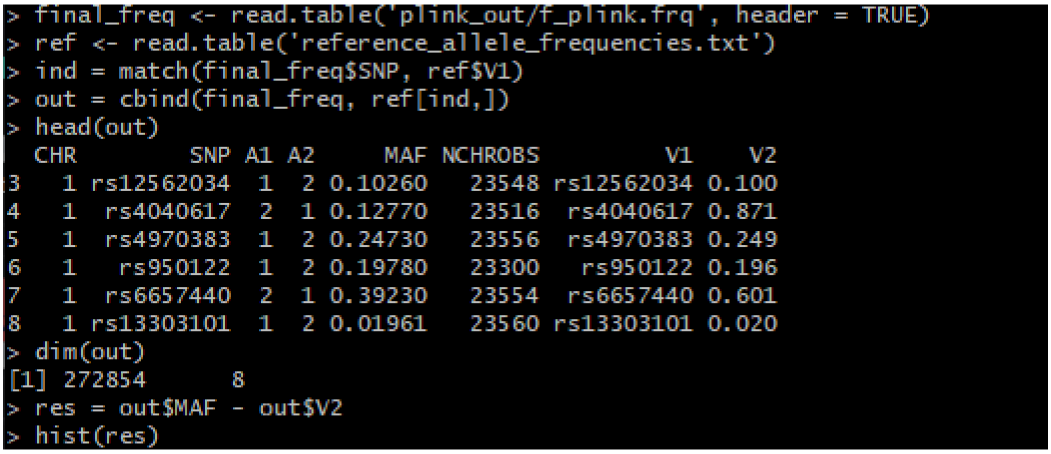


Null values present, filter with tidyverse and get count for MAFs < 0.01.

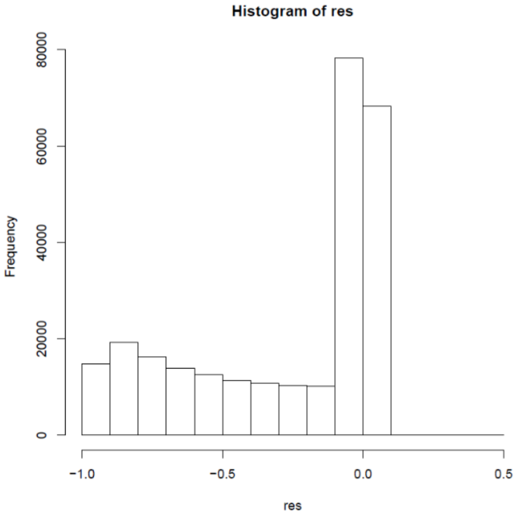


```
> library(tidyverse)
── Attaching packages ──────────────────────────
✔ ggplot2 3.0.0     ✔ purrr   0.2.5
✔ tibble  2.1.3     ✔ dplyr   0.8.3
✔ tidyr   0.8.3     ✔ stringr 1.3.1
✔ readr   1.1.1     ✔ forcats 0.3.0
── Conflicts ───────────────────────────────────
✖ dplyr::filter() masks stats::filter()
✖ dplyr::lag()    masks stats::lag()
> filt_freq <- freq %>% filter(!is.na(freq$MAF))
> head(filt_freq)
  CHR        SNP A1 A2    MAF NCHROBS
1   1  rs3131972  1  2 0.1568   23468
2   1  rs3115850  1  2 0.1389   21962
3   1 rs12562034  1  2 0.1026   23548
4   1  rs4040617  2  1 0.1277   23516
5   1  rs4970383  1  2 0.2473   23556
6   1   rs950122  1  2 0.1978   23300
> sum(filt_freq$MAF < 0.01)
[1] 5275
```

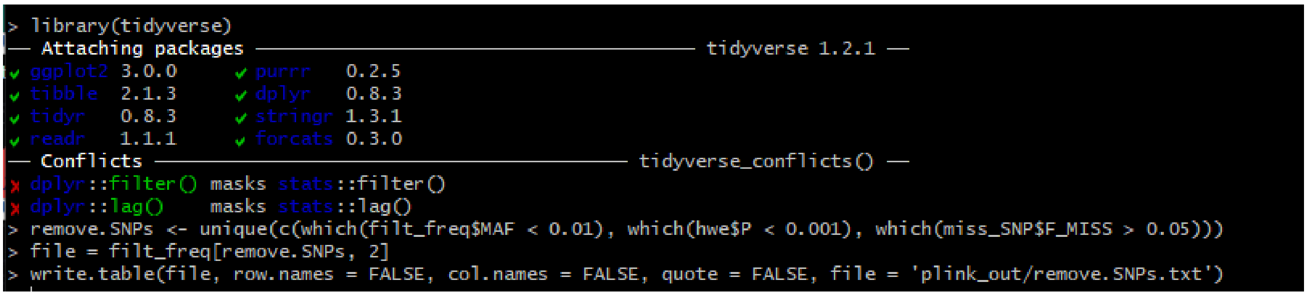
Combine allele frequency with reference allele frequency text file and plot output difference.



```
> final_freq <- read.table('plink_out/f_plink.frq', header = TRUE)
> ref <- read.table('reference_allele_frequencies.txt')
> ind = match(final_freq$SNP, ref$V1)
> out = cbind(final_freq, ref[ind,])
> head(out)
  CHR        SNP A1 A2     MAF NCHROBS         V1    V2
3   1 rs12562034  1  2 0.10260   23548 rs12562034 0.100
4   1  rs4040617  2  1 0.12770   23516  rs4040617 0.871
5   1  rs4970383  1  2 0.24730   23556  rs4970383 0.249
6   1   rs950122  1  2 0.19780   23300   rs950122 0.196
7   1  rs6657440  2  1 0.39230   23554  rs6657440 0.601
8   1 rs13303101  1  2 0.01961   23560 rs13303101 0.020
> dim(out)
[1] 272854      8
> res = out$MAF - out$V2
> hist(res)
```



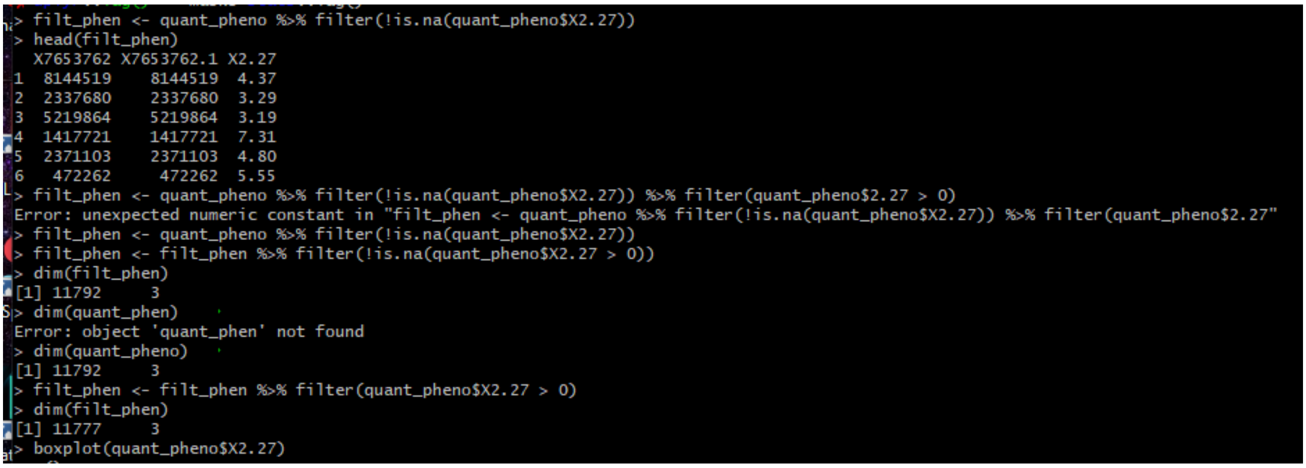
Apply SNP filter to dataset



```
> library(tidyverse)
── Attaching packages ─────────────────────────────────────── tidyverse 1.2.1 ──
✔ ggplot2 3.0.0     ✔ purrr   0.2.5
✔ tibble  2.1.3     ✔ dplyr   0.8.3
✔ tidyr   0.8.3     ✔ stringr 1.3.1
✔ readr   1.1.1     ✔ forcats 0.3.0
── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
✖ dplyr::filter() masks stats::filter()
✖ dplyr::lag()    masks stats::lag()
> remove.SNPs <- unique(c(which(filt_freq$MAF < 0.01), which(hwe$P < 0.001), which(miss_SNP$F_MISS > 0.05)))
> file = filt_freq[remove.SNPs, 2]
> write.table(file, row.names = FALSE, col.names = FALSE, quote = FALSE, file = 'plink_out/remove.SNPs.txt')
```

Phenotype Quality Control

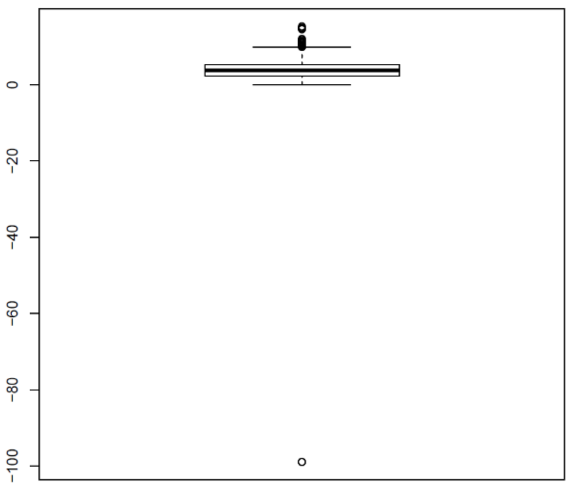
Filtering the quantitative phenotype file to remove outliers and null values using a boxplot for visualisation

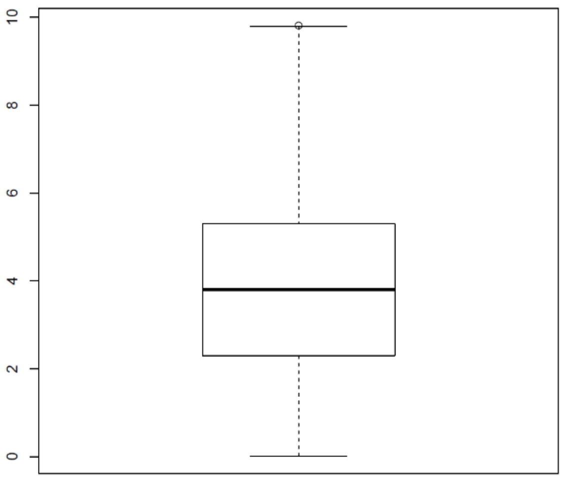


```
> filt_phen <- quant_pheno %>% filter(!is.na(quant_pheno$X2.27))
> head(filt_phen)
  X7653762 X7653762.1 X2.27
1  8144519    8144519  4.37
2  2337680    2337680  3.29
3  5219864    5219864  3.19
4  1417721    1417721  7.31
5  2371103    2371103  4.80
6   472262     472262  5.55
> filt_phen <- quant_pheno %>% filter(!is.na(quant_pheno$X2.27)) %>% filter(quant_pheno$2.27 > 0)
Error: unexpected numeric constant in "filt_phen <- quant_pheno %>% filter(!is.na(quant_pheno$X2.27)) %>% filter(quant_pheno$2.27"
> filt_phen <- quant_pheno %>% filter(!is.na(quant_pheno$X2.27))
> filt_phen <- filt_phen %>% filter(!is.na(quant_pheno$X2.27 > 0))
> dim(filt_phen)
[1] 11792     3
> dim(quant_phen)
Error: object 'quant_phen' not found
> dim(quant_pheno)
[1] 11792     3
> filt_phen <- filt_phen %>% filter(quant_pheno$X2.27 > 0)
> dim(filt_phen)
[1] 11777     3
> boxplot(quant_pheno$X2.27)
```

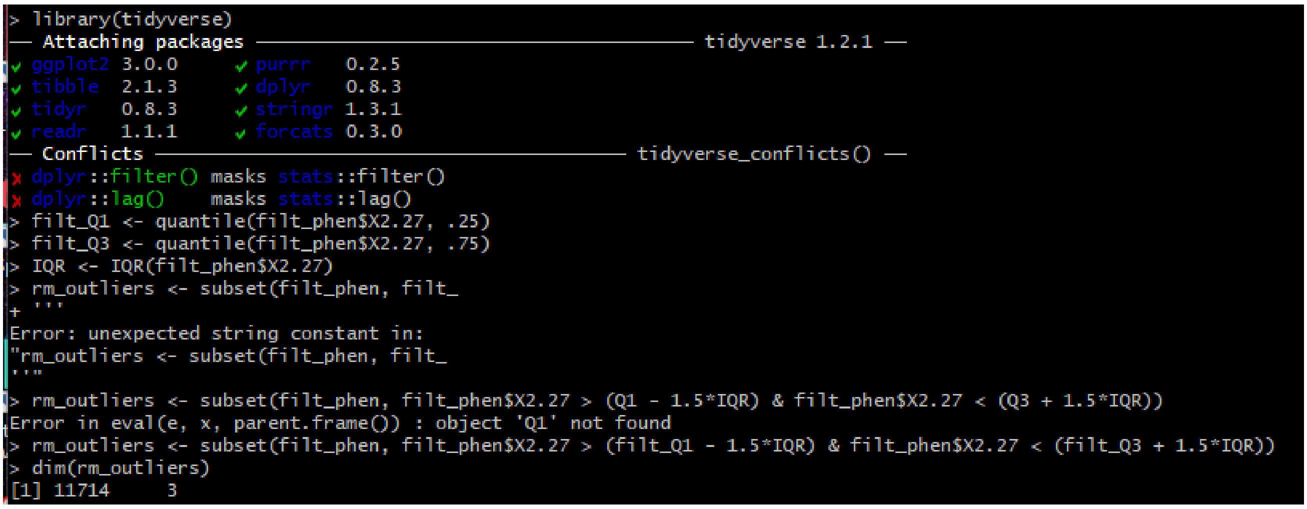
Before outlier filter

After outlier filter

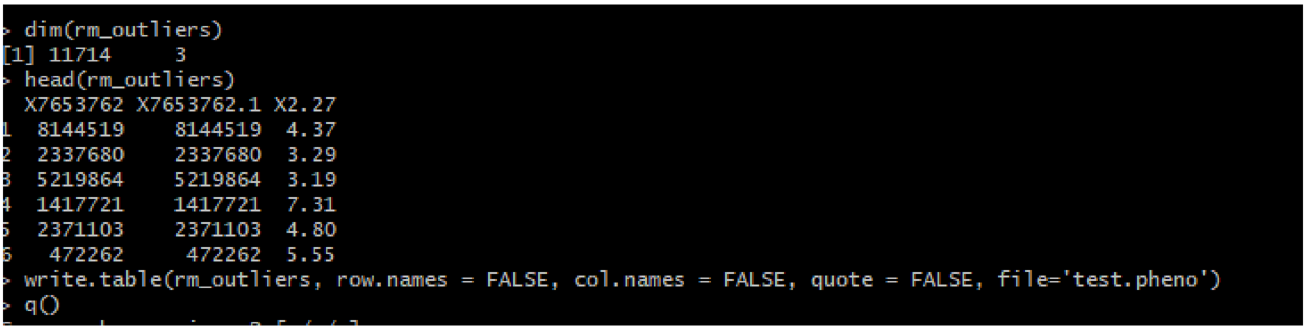




Removing outliers by using IQR, Q1 and Q3 for filtering



```
> library(tidyverse)
── Attaching packages ─────────────────────────────────────── tidyverse 1.2.1 ──
✔ ggplot2 3.0.0     ✔ purrr   0.2.5
✔ tibble  2.1.3     ✔ dplyr   0.8.3
✔ tidyr   0.8.3     ✔ stringr 1.3.1
✔ readr   1.1.1     ✔ forcats 0.3.0
── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
✖ dplyr::filter() masks stats::filter()
✖ dplyr::lag()    masks stats::lag()
> filt_Q1 <- quantile(filt_phen$X2.27, .25)
> filt_Q3 <- quantile(filt_phen$X2.27, .75)
> IQR <- IQR(filt_phen$X2.27)
> rm_outliers <- subset(filt_phen, filt_
+ '''
Error: unexpected string constant in:
"rm_outliers <- subset(filt_phen, filt_
''"
> rm_outliers <- subset(filt_phen, filt_phen$X2.27 > (Q1 - 1.5*IQR) & filt_phen$X2.27 < (Q3 + 1.5*IQR))
Error in eval(e, x, parent.frame()) : object 'Q1' not found
> rm_outliers <- subset(filt_phen, filt_phen$X2.27 > (filt_Q1 - 1.5*IQR) & filt_phen$X2.27 < (filt_Q3 + 1.5*IQR))
> dim(rm_outliers)
[1] 11714     3
```



```
> dim(rm_outliers)
[1] 11714     3
> head(rm_outliers)
  X7653762 X7653762.1 X2.27
1  8144519    8144519  4.37
2  2337680    2337680  3.29
3  5219864    5219864  3.19
4  1417721    1417721  7.31
5  2371103    2371103  4.80
6   472262     472262  5.55
> write.table(rm_outliers, row.names = FALSE, col.names = FALSE, quote = FALSE, file='test.pheno')
> q()
```

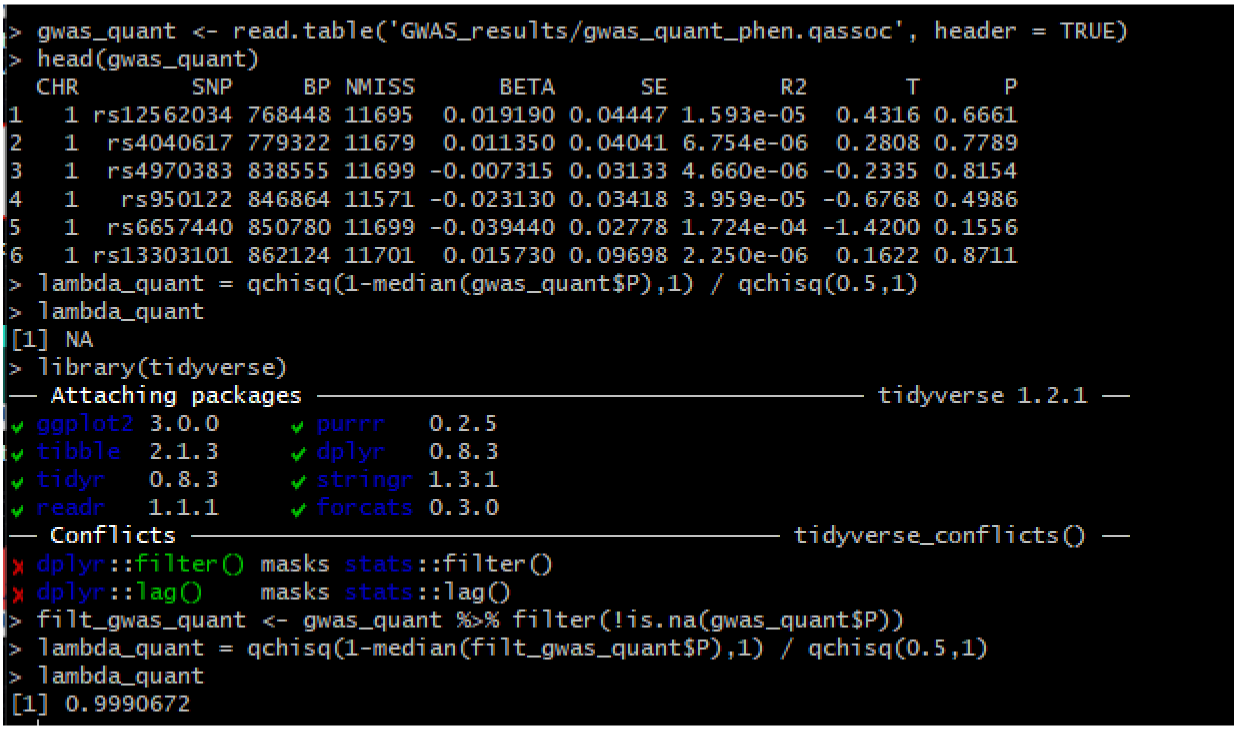
Quantitative Trait GWAS Results

Plink and Unix commands

```
/data/STAT3306/plink --bfile post_SNP_qc/final_QC --assoc --pheno filt_pheno/test.pheno --mpheno
1 --out GWAS_results/gwas_quant_phen
```

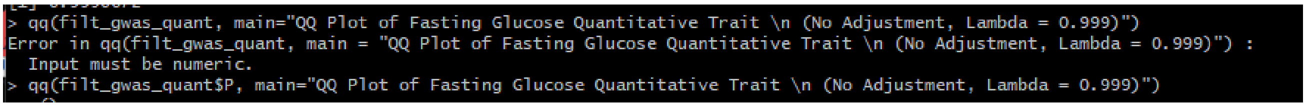
R commands and plots

Applying gwas results from plink to get corresponding lambda values



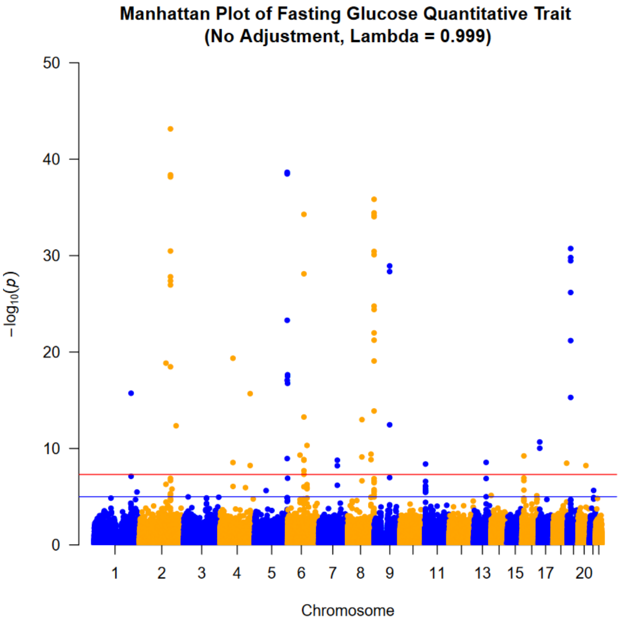
```
> gwas_quant <- read.table('GWAS_results/gwas_quant_phen.qassoc', header = TRUE)
> head(gwas_quant)
  CHR        SNP     BP NMISS      BETA      SE        R2       T      P
1   1 rs12562034 768448 11695  0.019190 0.04447 1.593e-05  0.4316 0.6661
2   1  rs4040617 779322 11679  0.011350 0.04041 6.754e-06  0.2808 0.7789
3   1  rs4970383 838555 11699 -0.007315 0.03133 4.660e-06 -0.2335 0.8154
4   1   rs950122 846864 11571 -0.023130 0.03418 3.959e-05 -0.6768 0.4986
5   1  rs6657440 850780 11699 -0.039440 0.02778 1.724e-04 -1.4200 0.1556
6   1 rs13303101 862124 11701  0.015730 0.09698 2.250e-06  0.1622 0.8711
> lambda_quant = qchisq(1-median(gwas_quant$P),1) / qchisq(0.5,1)
> lambda_quant
[1] NA
> library(tidyverse)
── Attaching packages ─────────────────────────────────────── tidyverse 1.2.1 ──
✔ ggplot2 3.0.0     ✔ purrr   0.2.5
✔ tibble  2.1.3     ✔ dplyr   0.8.3
✔ tidyr   0.8.3     ✔ stringr 1.3.1
✔ readr   1.1.1     ✔ forcats 0.3.0
── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
✖ dplyr::filter() masks stats::filter()
✖ dplyr::lag()    masks stats::lag()
> filt_gwas_quant <- gwas_quant %>% filter(!is.na(gwas_quant$P))
> lambda_quant = qchisq(1-median(filt_gwas_quant$P),1) / qchisq(0.5,1)
> lambda_quant
[1] 0.9990672
```

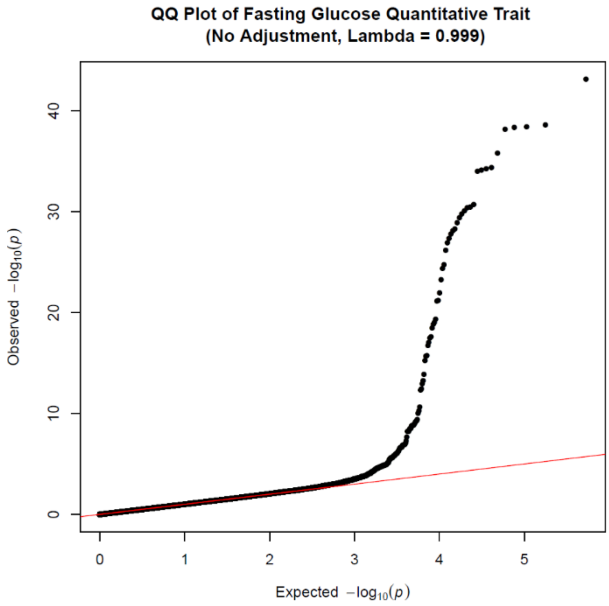
Code for QQ plot



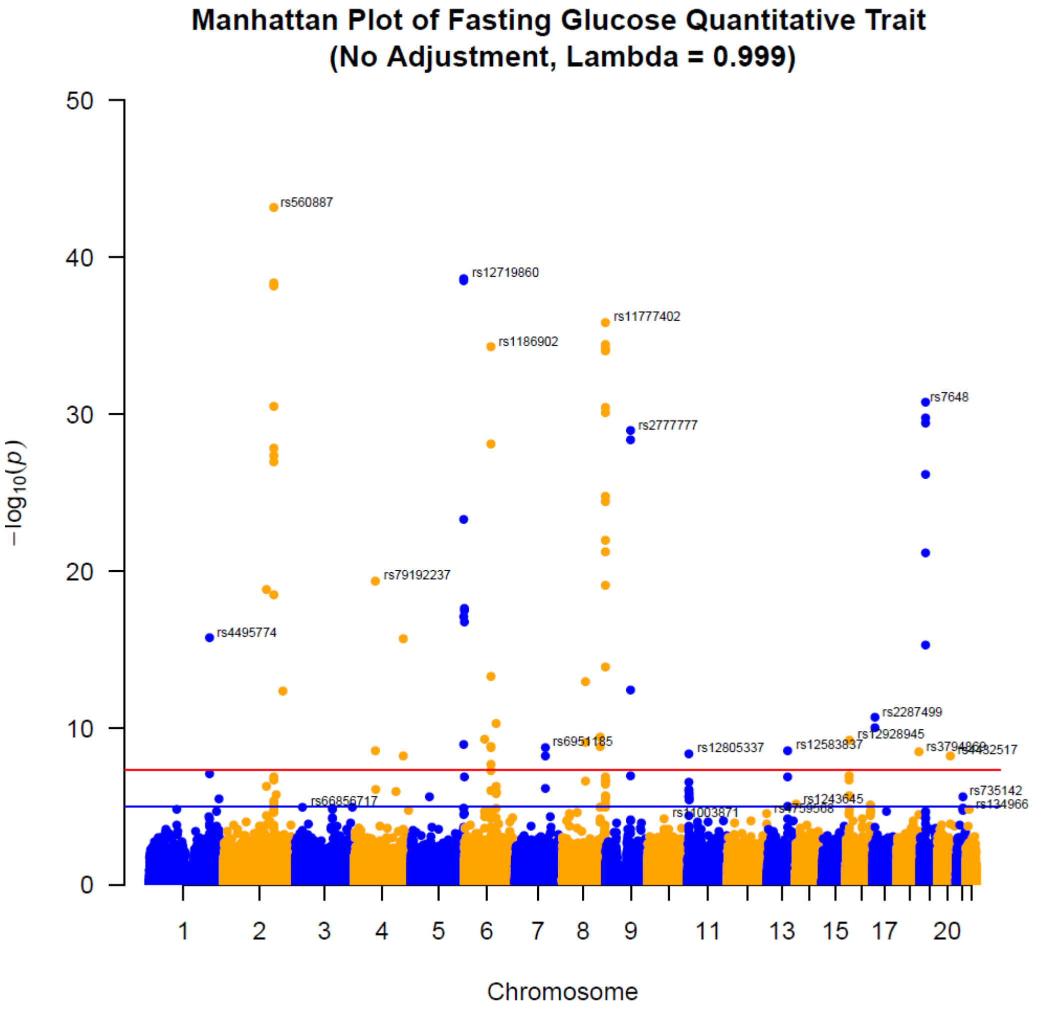
```
> qq(filt_gwas_quant, main="QQ Plot of Fasting Glucose Quantitative Trait \n (No Adjustment, Lambda = 0.999)")
Error in qq(filt_gwas_quant, main = "QQ Plot of Fasting Glucose Quantitative Trait \n (No Adjustment, Lambda = 0.999)") :
  Input must be numeric.
> qq(filt_gwas_quant$P, main="QQ Plot of Fasting Glucose Quantitative Trait \n (No Adjustment, Lambda = 0.999)")
```

QQ and Manhattan Plots





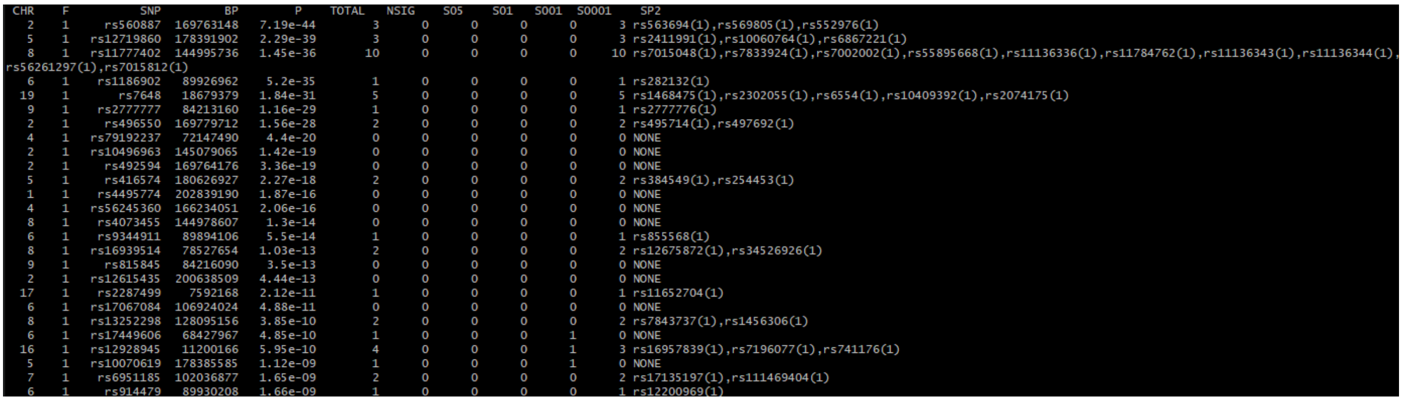
Manhattan with significant p-values = 0.0001, with top SNPs annotated



Clumping analysis

```
/data/STAT3306/plink --bfile post_SNP_qc/final_QC --clump GWAS_results/gwas_quant_phen.qassoc --clump-p1 0.001 --clump-p2 0.01 --clump-r2 0.50 --clump-kb 250 --out GWAS_results/gwas_quant_clump
```

Screenshot of clumping analysis results

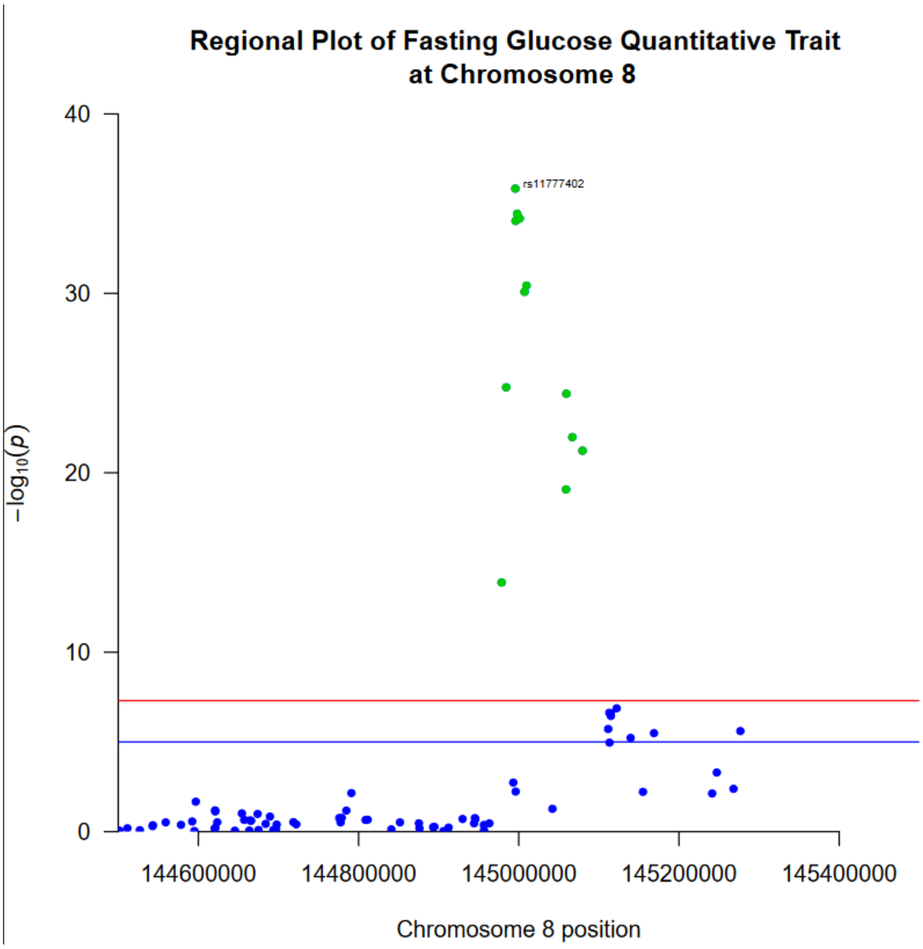


```
 CHR    F          SNP         BP          P    TOTAL   NSIG    S05    S01   S001  S0001    SP2
   2    1     rs560887  169763148   7.19e-44        3      0      0      0      0      3 rs563694(1),rs569805(1),rs552976(1)
   5    1   rs12719860  178391902   2.29e-39        3      0      0      0      0      3 rs2411991(1),rs10060764(1),rs6867221(1)
   8    1   rs11777402  144995736   1.45e-36       10      0      0      0      0     10 rs7015048(1),rs7833924(1),rs7002002(1),rs55895668(1),rs11136336(1),rs11784762(1),rs11136343(1),rs11136344(1),
rs56261297(1),rs7015812(1)
   6    1    rs1186902   89926962    5.2e-35        1      0      0      0      0      1 rs282132(1)
  19    1       rs7648   18679379   1.84e-31        5      0      0      0      0      5 rs1468475(1),rs2302055(1),rs6554(1),rs10409392(1),rs2074175(1)
   9    1    rs2777777   84213160   1.16e-29        1      0      0      0      0      1 rs2777776(1)
   2    1     rs496550  169779712   1.56e-28        2      0      0      0      0      2 rs495714(1),rs497692(1)
   4    1   rs79192237   72147490    4.4e-20        0      0      0      0      0      0 NONE
   2    1   rs10496963  145079065   1.42e-19        0      0      0      0      0      0 NONE
   2    1     rs492594  169764176   3.36e-19        0      0      0      0      0      0 NONE
   5    1     rs416574  180626927   2.27e-18        2      0      0      0      0      2 rs384549(1),rs254453(1)
   1    1    rs4495774  202839190   1.87e-16        0      0      0      0      0      0 NONE
   4    1   rs56245360  166234051   2.06e-16        0      0      0      0      0      0 NONE
   8    1    rs4073455  144978607    1.3e-14        0      0      0      0      0      0 NONE
   6    1    rs9344911   89894106    5.5e-14        1      0      0      0      0      1 rs855568(1)
   8    1   rs16939514   78527654   1.03e-13        2      0      0      0      0      2 rs12675872(1),rs34526926(1)
   9    1     rs815845   84216090    3.5e-13        0      0      0      0      0      0 NONE
   2    1   rs12615435  200638509   4.44e-13        0      0      0      0      0      0 NONE
  17    1    rs2287499    7592168   2.12e-11        1      0      0      0      0      1 rs11652704(1)
   6    1   rs17067084  106924024   4.88e-11        0      0      0      0      0      0 NONE
   8    1   rs13252298  128095156   3.85e-10        2      0      0      0      0      2 rs7843737(1),rs1456306(1)
   6    1   rs17449606   68427967   4.85e-10        1      0      0      0      1      0 NONE
  16    1   rs12928945   11200166   5.95e-10        4      0      0      0      1      3 rs16957839(1),rs7196077(1),rs741176(1)
   5    1   rs10070619  178385585   1.12e-09        1      0      0      0      1      0 NONE
   7    1    rs6951185  102036877   1.65e-09        2      0      0      0      0      2 rs17135197(1),rs111469404(1)
   6    1     rs914479   89930208   1.66e-09        1      0      0      0      0      1 rs12200969(1)
```

Regional Manhattan plot code

```
chr8_snpsofinterest <- c(
'rs7015048','rs7833924','rs7002002','rs55895668','rs11136336','rs11784762','rs11136343','rs11136344','rs56261297','rs7015812', 'rs4073455', 'rs11777402')

manhattan(subset(filt_gwas_quant, CHR == 8), main="Regional Plot of Fasting Glucose Quantitative Trait \n at Chromosome 8", highlight = chr8_snpsofinterest, annotatePval = 0.001, annotateTop = TRUE, ylim = c(0, 40), xlim=c(144500000, 145500000), col = c('blue1', 'orange1')
```



Binary Traits GWAS Results

Plink and Unix commands

```
/data/STAT3306/plink --bfile post_SNP_qc/final_QC --assoc --pheno
raw_pheno/Fasting_Glucose_binary1.phen --mpheno 1 --1 --out GWAS_results/gwas_b1_allele
```

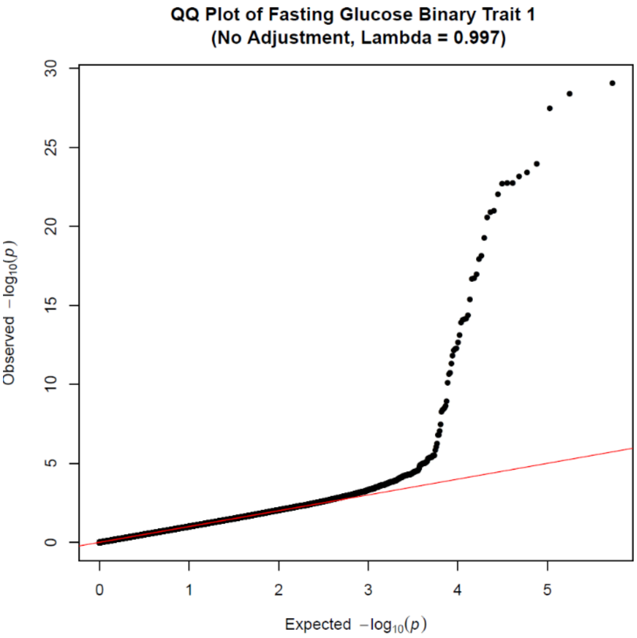
```
/data/STAT3306/plink --bfile post_SNP_qc/final_QC --assoc --pheno
raw_pheno/Fasting_Glucose_binary2.phen --mpheno 1 --1 --out GWAS_results/gwas_b2_allele
```

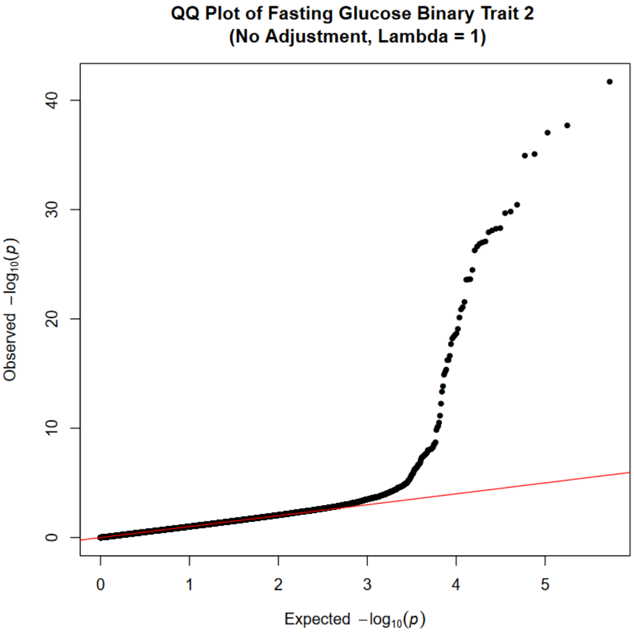
Clumping analysis

```
/data/STAT3306/plink --bfile post_SNP_qc/final_QC --clump GWAS_results/ filt_2.assoc --clump-p1
0.001 --clump-p2 0.01 --clump-r2 0.50 --clump-kb 250 --out GWAS_results/gwas_b2
```

```
/data/STAT3306/plink --bfile post_SNP_qc/final_QC --clump GWAS_results/filt_1.assoc --clump-p1
0.001 --clump-p2 0.01 --clump-r2 0.50 --clump-kb 250 --out GWAS_results/gwas_b1
```

QQ plots of binary traits





R commands for plots and GWAS filtering

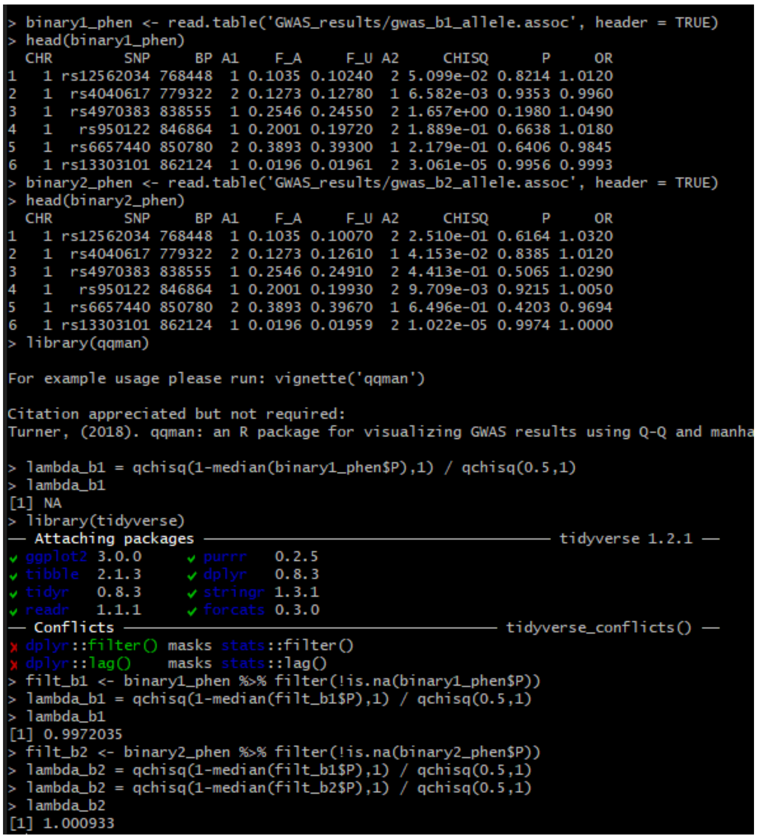
```
write.table(filt_b1, row.names = FALSE, col.names = TRUE, quote = FALSE, file =
"GWAS_results/filt_1.assoc")
```

```
write.table(filt_b2, row.names = FALSE, col.names = TRUE, quote = FALSE, file =
"GWAS_results/filt_2.assoc")
```

```
manhattan(filt_b1, main="Manhattan Plot of Fasting Glucose Binary 1 Trait \n (No Adjustment,
Lambda = 0.997)", ylim = c(0,30), col = c('green4', 'purple2'))
```

```
manhattan(filt_b2, main="Manhattan Plot of Fasting Glucose Binary 2 Trait \n (No Adjustment,
Lambda = 1)", ylim = c(0,50), col = c('blue4', 'pink2'))
```

Lambda values from GWAS filtering



```
> binary1_phen <- read.table('GWAS_results/gwas_b1_allele.assoc', header = TRUE)
> head(binary1_phen)
  CHR        SNP     BP A1    F_A     F_U A2     CHISQ      P     OR
1   1 rs12562034 768448  1 0.1035 0.10240  2 5.099e-02 0.8214 1.0120
2   1  rs4040617 779322  2 0.1273 0.12780  1 6.582e-03 0.9353 0.9960
3   1  rs4970383 838555  1 0.2546 0.24550  2 1.657e+00 0.1980 1.0490
4   1   rs950122 846864  1 0.2001 0.19720  2 1.889e-01 0.6638 1.0180
5   1  rs6657440 850780  2 0.3893 0.39300  1 2.179e-01 0.6406 0.9845
6   1 rs13303101 862124  1 0.0196 0.01961  2 3.061e-05 0.9956 0.9993
> binary2_phen <- read.table('GWAS_results/gwas_b2_allele.assoc', header = TRUE)
> head(binary2_phen)
  CHR        SNP     BP A1    F_A     F_U A2     CHISQ      P     OR
1   1 rs12562034 768448  1 0.1035 0.10070  2 2.510e-01 0.6164 1.0320
2   1  rs4040617 779322  2 0.1273 0.12610  1 4.153e-02 0.8385 1.0120
3   1  rs4970383 838555  1 0.2546 0.24910  2 4.413e-01 0.5065 1.0290
4   1   rs950122 846864  1 0.2001 0.19930  2 9.709e-03 0.9215 1.0050
5   1  rs6657440 850780  2 0.3893 0.39670  1 6.496e-01 0.4203 0.9694
6   1 rs13303101 862124  1 0.0196 0.01959  2 1.022e-05 0.9974 1.0000
> library(qqman)

For example usage please run: vignette('qqman')

Citation appreciated but not required:
Turner, (2018). qqman: an R package for visualizing GWAS results using Q-Q and manha

> lambda_b1 = qchisq(1-median(binary1_phen$P),1) / qchisq(0.5,1)
> lambda_b1
[1] NA
> library(tidyverse)
── Attaching packages ─────────────────────────────────────── tidyverse 1.2.1 ──
✔ ggplot2 3.0.0     ✔ purrr   0.2.5
✔ tibble  2.1.3     ✔ dplyr   0.8.3
✔ tidyr   0.8.3     ✔ stringr 1.3.1
✔ readr   1.1.1     ✔ forcats 0.3.0
── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
✖ dplyr::filter() masks stats::filter()
✖ dplyr::lag()    masks stats::lag()
> filt_b1 <- binary1_phen %>% filter(!is.na(binary1_phen$P))
> lambda_b1 = qchisq(1-median(filt_b1$P),1) / qchisq(0.5,1)
> lambda_b1
[1] 0.9972035
> filt_b2 <- binary2_phen %>% filter(!is.na(binary2_phen$P))
> lambda_b2 = qchisq(1-median(filt_b1$P),1) / qchisq(0.5,1)
> lambda_b2 = qchisq(1-median(filt_b2$P),1) / qchisq(0.5,1)
> lambda_b2
[1] 1.000933
```