Practical 3 - Genome wide association studies (GWAS) for quantitative traits

In this practical we will perform a genome wide association study (GWAS) using the PLINK software. This will involve a power calculation to investigate the probability that we reject the null given the alternative is true. We will then outline key components of the PLINK software and perform an association analysis for height and serum transferrin level. We will investigate the output of the GWAS through visualisation.

Power calculation for GWAS

Power calculations are often a prospective component of doing GWAS. We will again use R to do our calculation but there are many avenues to estimate power for different modelling scenarios, for example http://pngu.mgh.harvard.edu/~purcell/gpc/ is one key resource. The power calculation done in listing 1 is based on (Shi et al., 2009).

```
> # Define the genome wide significance level
> alpha     <- 5e-8
> # Calculate the threshold given alpha
> threshold <- qchisq(1 - alpha, 1)
> threshold
  29.71679
> qsqr      <- 0.005
> n         <- 5000
> # Calculate the non-centrality parameter
> ncp          <- n * qsqr / (1 - qsqr)
> ncp
  25.12563
> # Estimate power
> power     <- 1 - pchisq(threshold, 1, ncp)
> power
  0.3304165
> # Keep the same parameters but increase the sample size
> alpha     <- 5e-8
> threshold <- qchisq(1 - alpha, 1)
> qsqr      <- 0.005
> n         <- 10000
> ncp       <- n * qsqr / (1 - qsqr)
> ncp
  50.25126
> power     <- 1 - pchisq(threshold, 1, ncp)
> power
  0.9492371
```

Listing 1 GWAS power calculation

We can see that we need quite a few individuals to detect a locus that explains 0.5 % of the phenotypic variance.

Exercise 1

- How many individuals would you need to detect a locus that explains 1% with a power of 80%?

The **PLINK** software

To perform the GWAS, we will use the excellent and widely used PLINK command line program (http://pngu.mgh.harvard.edu/~purcell/plink/). If you haven't used PLINK before is may be worth your time exploring the website a little. The PLINK software has recently been updated to include incredible speed ups in PLINK 1.9 (https://www.cog-genomics.org/plink2). Thanks to a heavy use of bitwise operators, sequential memory access patterns, multithreading, and higher-level algorithmic improvements, PLINK 1.9 is much, much faster than PLINK 1.07. We will use the PLINK software (feel free to use PLINK 1.9 implemented as plink2) to perform data management and an association analysis. This is only the tip of the iceberg for what PLINK can do.

The PLINK program has a standard data format that allows for fast reading and writing of genotype/phenotype data. PLINK supports two file formats with the first consisting of two files the .ped and .map. When using this data type the associated PLINK command requires the use of the --file flag, for example, if we were to run PLINK with this data format we would type plink --file /path/to/genotype/file at the command line. To save space and time, PLINK allows for a binary .ped file to be made. This file format includes three files the .bed, .bim, and .fam files. We will take a quick look at these files to understand a little more about how PLINK makes use of them.

The .ped file usually contains six columns with the following content respectively, family ID, individual ID, paternal ID, maternal ID, sex (1 = male, 2 = female, other = NA) and the phenotype. The .map file has four columns with column 1 indicating the chromosome (1-22, X, Y, or 0), column 2 the rs ID or SNP identifier, column 3 the genetic distance in morgans, and column 4 the base-pair (BP) position (note BP position restarts for each chromosome).

Let's attempt to use PLINK to convert the .ped and .map files to binary format. Navigate to the terminal or command prompt and execute the following command. IMPORTANT - remember to export your path again to the binaries if you are on Mac or Linux

```
$ # Export path
$ export PATH=$PATH:~/Desktop/SISG_AQG_2015/bin
$ # Convert to .ped and .map PLINK format from binary
$ plink2 --bfile practical_3/data/QIMRX --recode --out QIMRX
```

We can use the basic plink2 --bfile in combination with many many other options to perform a variety of key quantitative genetics tasks. PLINK has a vast array of data management tools with the following being used for subsetting data. Please take some time to read through the options as we will use these later.

```
--keep    # Retains a set of individuals
--remove  # Removes a set of individuals
--extract # Retains a set of SNPs
--exclude # Removes a set of SNPs
--chr     # Retains the chromosome given after the flag
--from SNP1 --to SNP2 # Retains the SNPs between these two SNPs. Uses rs ID
--out     # Specifies the file name to write out
```

Listing 2 Data subsetting examples using PLINK

Some other key data management flags are

```
--make-bed  # Makes binary files. Need to include with subsetting flags if want have binary files as output
--recode    # Recodes the allele labels as they appear in the original. Makes .ped and .map files from binary
            # files
--bmerge    # Merge two PLINK files in binary format
--pheno     # Followed by a file name that specifies alternate phenotype
--all-pheno # Performs association analysis for all phenotypes in file
--mpheno    # Specify which column in phenotype file (if >1)
```

Some examples using these flags together. Note that these are only dummy commands.

```
plink --bfile test  --remove individual_subset.txt --chr 7 --make-bed --out test_subset
plink --file mydata --pheno  pheno2.txt --pheno-name bmi --assoc
plink --file mydata --pheno  pheno2.txt --mpheno 4
```

Quality control filters

```
--maf  # Filter on minor allele frequency
--geno # Filter on SNP missing rate
--mind # Filter on individual missing rate
--hwe  # Filter on Hardy-Weinberg equilibrium
```

Listing 3 PLINK quality control filters

Summary statistics

```
--freq    # Calculates and reports the MAF for each SNP in .frq file
--missing # Reports SNP missing rate and individual missing rate in .lmiss and .imiss files
--hwe     # Calculates and reports the Hardy-Weinberg equilibrium test statistics for each SNP
```

Association analyses

```
--assoc  # Performs a basic association analysis
--linear # Performs association analyses but with extra functionality
--within # Performs a stratified analysis with a separate my cluster.dat file
--covar  # Includes covariates in the model using a mycov.txt file
--gxe    # Includes a GxE interaction term in the analysis
```

Running a GWAS

Given this short overview of PLINK we will now attempt to perform a GWAS on data from human height and serum transferrin levels. This will be done in three steps:

- File inspection
- QC
- GWAS for human height and serum transferrin level

File inspection

Firstly, we will use R to inspect the .fam, .bim, and HT_T_X.pheno files. Let's have a quick look around these files and establish some of the properties of the data using R. As always, let's firstly read in the data.

```
> # Make sure you have set the working directory to the ~/Desktop/SISG_AQG_2015 folder
> setwd("")
> fam <- read.table("practical_3/data/QIMRX.fam")
> bim <- read.table("practical_3/data/QIMRX.bim")
> pheno <- read.table("practical_3/data/HT_T_X.pheno")
```

```
> # Count the number of genotyped individuals. Number of rows
> dim(fam)
  4861    6
> # Take a look at the .fam file
> head(fam)
    V1  V2  V3    V4 V5 V6
   355 883 681 10680  2 -9
   355 884 681 10680  1 -9
  3004 885 682 10681  2 -9
  3155 886 683 10682  1 -9
  1629 887 684 10683  2 -9
  1747 888 685 10684  2 -9
```

Listing 4 The .fam file

```
> # Count the number of genotyped SNPs
> dim(bim)
  281313      6
> head(bim)
  V1         V2 V3      V4 V5 V6
   1  rs3934834  0  995669  T  C
   1  rs3737728  0 1011278  A  G
   1  rs6687776  0 1020428  T  C
   1  rs9651273  0 1021403  A  G
   1  rs4970405  0 1038818  G  A
   1 rs12726255  0 1039813  G  A
```

Listing 5 The .bim file

```
> # Count the number of individuals with height and transferrin
> # measurements
> dim(pheno)
  4861    4
> dim(pheno[complete.cases(pheno), ])
  337   4]
> head(pheno)
   355 883          NA -0.815
   355 884 -1.01219122     NA
  3004 885 -1.11122366     NA
  3155 886          NA -0.299
  1629 887 -0.04663134     NA
  1747 888  1.59969343  1.182
```

Listing 6 The `.pheno` file

Quality Control

The quality control steps are critical to any analysis and often take more time then the analyses themselves. We will now work with `PLINK 1.9` to do some QC on the binary files for the QIMR data on human height and transferrin levels. These QC step will include

- Estimate the allele frequencies for all SNPs
- Calculate SNP and individual missingness
- Calculate p-values for Hardy-Weinberg (HW)

Each of the below commands should produce a file with an extension that is indicative of the process used. Unlike previous programs `PLINK` allows for an `out` directory to be specified with `--out`. We will put all our results in the `practical_3/results` folder. Execute the following command line arguments to produce three files for allele frequency, missingness, and HW.

```
$ plink2 --bfile practical_3/data/QIMRX --freq --out practical_3/results/QIMRX
$ plink2 --bfile practical_3/data/QIMRX --missing --out practical_3/results/QIMRX
$ plink2 --bfile practical_3/data/QIMRX --hardy --out practical_3/results/QIMRX
```

Listing 7 `PLINK` combination commands

Mini Exercise - Run these same commands but with `PLINK 1` to recognise the incredible differences in speed that `PLINK 2` offers. For large analyses these speed ups make all the difference between being able to complete analyses or not – additionally, RAM is often the limiting factor rather than CPU time and `PLINK v1.9` has a much lower memory profile when compared to the first version.

Read the resultant files into R and attempt to answer the following questions

- How many SNPs have MAF > 0.05?
- How many individuals have missingness > 10%?
- How many SNPs have missingness of > 1%?
- How many SNPs have a HWE p-value < 0.001?

Below are some hints on how to do this

```
> # Read in the .frq file
> frq <- read.table("practical_3/results/QIMRX.frq", header = T, na = "NA")
> head(frq)
  CHR        SNP A1 A2    MAF NCHROBS
    1  rs3934834  T  C 0.1776     366
    1  rs3737728  A  G 0.2838     370
    1  rs6687776  T  C 0.1811     370
    1  rs9651273  A  G 0.2703     370
    1  rs4970405  G  A 0.1081     370
    1 rs12726255  G  A 0.1346     364
> # Find the proportion of those with missingness less than 0.05
> prop.lw.maf <- sum((frq$MAF < 0.05)) / length(frq$MAF)
> # The sum component above is summing up the true values
> # Look at what (frq$MAF < 0.05) does to get a better feel for
> # this solution. Can be done in many other ways
> prop.lw.maf
  0.02557649
> # Do something similar for the rest of the questions
```

Listing 8 PLINK .frq file

Running a GWAS

The work above was designed to investigate what PLINK can do on-the-fly with the filtering commands in listing 7. We can do all of these steps and the GWAS in one command with PLINK. In the terminal execute a similar command as in listing 7 using the quality control flags from listing 3 along with the --assoc flag to run a GWAS with MAF filter 0.05, individual missing rate 0.1, SNP missing rate 0.01, and Hardy-Weinberg of 0.001. Note that this will be quite a long terminal command. Remember to give the out path to the practical_3/results folder.

Once the files have been moved, read the association results into R. We will first draw a manhattan plot

```
> # Read in the association results
> gwas.res <- read.table("practical_3/results/gwas_pheno_1.qassoc",
                          header = T)
> dim(gwas.res)
> 273201      9
> head(gwas.res)
  CHR        SNP      BP NMISS      BETA      SE        R2         T       P
    1  rs3934834  995669  2812  0.014000 0.03857 4.690e-05  0.363100 0.71660
    1  rs3737728 1011278  2833  0.000238 0.03023 2.190e-08  0.007873 0.99370
    1  rs6687776 1020428  2834  0.086810 0.03742 1.897e-03  2.320000 0.02043
    1  rs9651273 1021403  2836 -0.024630 0.03090 2.242e-04 -0.797100 0.42550
    1  rs4970405 1038818  2832  0.083190 0.04372 1.278e-03  1.903000 0.05714
    1 rs12726255 1039813  2829  0.056490 0.03973 7.145e-04  1.422000 0.15520
> # Build the data frame that the manhattan plot function requires
> man.df <- data.frame(gwas.res$BP, gwas.res$CHR, gwas.res$P,
                        gwas.res$SNP)
> # Load the manhattan plot library. If not installed use install.packages("qqman")
> library(qqman)
> # Rename the columns for the manhattan function
>  colnames(man.df) <- c("BP", "CHR", "P", "SNP")
> # Produce the manhattan plot. NOTE THAT THIS MAY TAKE SOME TIME AND MAY CRASH YOUR COMPUTER
> # IF IT HAS POOR RESOURCES
> manhattan(man.df)
> # Drawing a qq plot
> # -------------------
> obs.p <- gwas.res$P
> # Order these
> obs.p.srt <- sort(obs.p)
> m.log10.obs.p <- -(log10(obs.p.srt))
> max.p   <- max(m.log10.obs.p)
> exp.val <- seq(1, length(obs.p))
> m.log10.exp.val <- -log10((exp.val - 0.5) / length(exp.val))
> plot(c(0, max.p), c(0, max.p), col = "red", lwd = 2, type = "l",
       xlab = "Expected -log10(p)", ylab = "Observed -log10(p)",
       xlim = c(0, max.p), ylim = c(0, max.p), las = 1, xaxs = "i",
       yaxs = "i", bty = "l", main = "Trait 1")
> points(m.log10.exp.val, m.log10.obs.p)
> # Alternatively you can use a package. This has the added component of
> # having confidence interval bounds
install.packages("Haplin")
library(Haplin)
x <- pQQ(obs.p.srt, nlabs = 6, conf = 0.95)
```

Listing 9 Drawing Manhattan and QQ plot

Hopefully you obtain plots similar to those below.

The final QQ plot implies potential systematic inflation of the p-values. This may be attributed to population structure, which could be excluded by including the first few principal components as covariates in the association analyses.

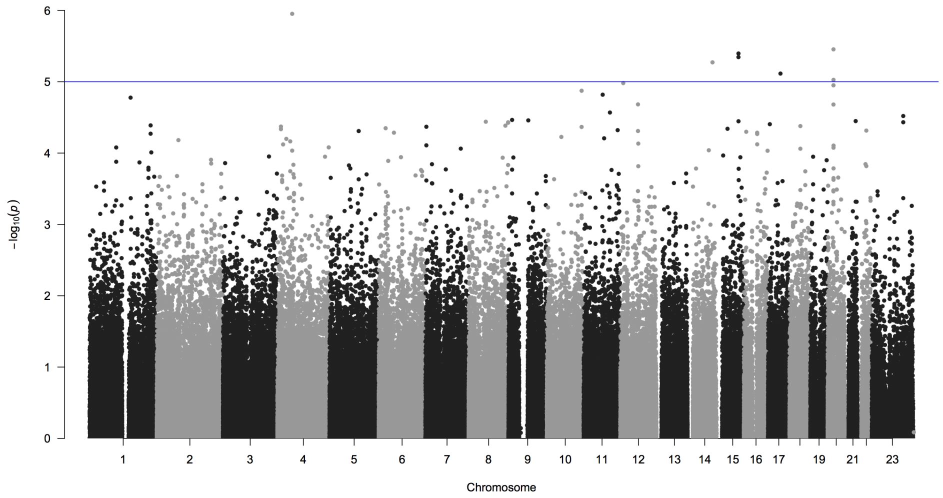

Figure 1 Manhattan plot of trait 1 from QIMR data set.

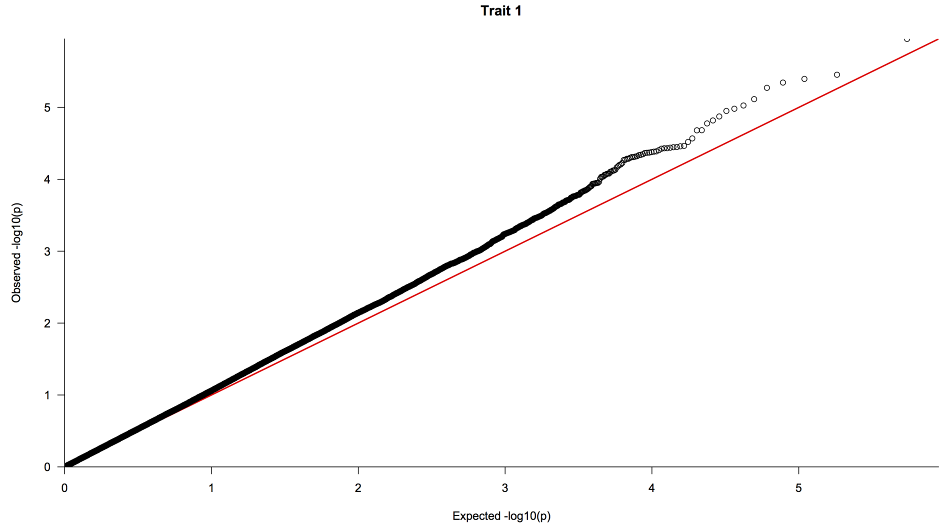

Figure 2 QQ plot of trait 1 from QIMR data set.

Exercise 2

- Repeat the GWAS for the second phenotype.
- Produce the same plots as for height
- Calculate λ_{GC} for both phenotypes.

Exercise 3

- Calculate the first 10 principal components of the genotype matrix
- `plink2 --bfile practical_3/data/QIMRX --pca 10 --out practical_3/results/QIMRX`
- Rerun the association analysis with these 10 PCs

-
  ```
  plink2 --bfile practical_3/data/QIMRX --maf 0.05 --geno 0.1 --mind 0.01
  --hwe 0.001 --linear --covar practical_3/results/QIMRX.eigenvec
  --pheno practical_3/data/HT_T_X.pheno --mpheno 2 --out practical_3/results/QIMRX_ST2
  ```
- Read the results back into R
- Subset the data to only leave the estimates for the SNPs
- `man.df <- subset(gwas.res, TEST=='ADD',c('BP','CHR','P','SNP'))`
- Alternatively you can use `grep "ADD"` over the output file.
- Re-draw the manhattan plots
- Calculate λ_{GC} for both phenotypes.
- You may find the following command useful to clump your results into ‘roughly’ independent regions
  ```
  plink --bfile practical_3/data/QIMRX
  --clump QIMRX_ST2.assoc.linear
  --clump-p1 0.5 --clump-p2 0.5
  --clump-r2 0.20 --clump-kb 500
  --out hgt_gwas_clump
  ```

Additional exercise

Use `PLINK` to do a test for dominance for the top 5 SNPs for height. Is there any evidence for dominance?

```
$ plink2 --bfile practical_3/data/QIMRX --extract practical_3/data/top_snps.txt
$        --pheno practical_3/data/HT_T_X.pheno --mpheno 1 --linear --genotypic
$        --out practical_3/results/dom_test.txt
```

Listing 10 `PLINK` dominance test

References

Jianxin Shi, Douglas F Levinson, Jubao Duan, Alan R Sanders, Yonglan Zheng, Itsik PeEr, Frank Dudbridge, Peter A Holmans, Alice S Whittemore, Bryan J Mowry, et al. Common variants on chromosome 6p22. 1 are associated with schizophrenia. *Nature*, 460(7256):753–757, 2009.