PLINK Script

Mari Johnson

24/08/2020

Preprocess data in R

Files located in X drive, folder PLINK_Mari

Code key:

- Outcome Typhoid = 2, No Typhoid = 1
- Study P1 = 1 Tyger = 2 t1 = 3, t2 =4
- Sex M = 1, F = 2
- Vaccine: Placebo = 0, M01ZH09 = 1, Ty21a = 2
- Dose: Paratyphi A 0.5-1x10^3 = 1, Paratyphi A 1-5x10^3 = 2, TN = 3
- Quailes 1-5x10^4 = 4, Quailes 1-5x10^3" = 5

Set working directory

- Set wd in windows command promt to a file containg your data and the plink files so they run from the same location

```
C:\Users\mjohnson> cd Documents\genetics\plink-1.07-dos
```

GWAS data set up

Convert .map and .ped files to .bed files for quicker processing:

```
./plink2 --file mydata --make-bed
```

- Make bed files with QC filters for FCGR region
- Select region of chromosome to scan for SNPs
- Add QC filters for maf, geno and hwe
- To do for miRNA locus put in coordinates for the miRNAregion (hg37 build)

Select SNPs FCGR region for P1Tyger cohort:

```
./plink2 --bfile P1Tyger_bed --allow-no-sex  --chr 1 --from-bp 161284166 --to-bp 161697933 --mind 0.1 --
```

Select SNPs FCGR region for T1T2 cohort:

```
./plink2 --bfile P1Tyger_bed --allow-no-sex  --chr 1 --from-bp 161284166 --to-bp 161697933 --mind 0.1 --
```

Merge files:

- merge genotype files from different studies (so includes P1, Tyger, T1, T2)

```
./plink2 --bfile T1T2_FCGRv2 --bmerge P1Tyger_FCGRv2.bed P1Tyger_FCGRv2.bim P1Tyger_FCGRv2.fam --make-b
```

Modelling SNPs with enteric fever outcome

Basic Association Testing:

- only availble in PLINK v1
- add phenotype file from the relevant pathogens (participants from each filtered by using R Script)
- out flag writes the name of the output file

Paratyphoid Assoc:

```
./plink --bfile FCGRv2merge --allow-no-sex --pheno paratyphoid_pheno.txt --assoc counts --adjust --out
```

Typhoid Assoc:

```
./plink --bfile FCGRv2merge --allow-no-sex --pheno typhoid_pheno.txt --assoc counts --adjust --out typh
```

Typhoid Assoc, no vaccine:

- Excluded those who were vaccinated in T2

```
./plink --bfile FCGRv2merge --allow-no-sex --pheno no_vaccine.txt --assoc counts --adjust --out no_vacc
```

Generalised linear mixed effect model (glm):

- Takes into account covariates when predicting modelling the effect of genotype at predicting outcome
- covariates can be supplied or use pca
- can assess output by looking at the genomic inflation factor before and after additon of covars, lower the better
- Run in Plink 2

```
./plink2 --bfile FCGRv2merge --pheno fullmerge2.txt --pheno-name Outcome --covar p1t1cov.txt --covar-na
```

Principle component analysis

Plink PCA:

- Produdces eigenvec and vals that can be explored in R

```
./plink2 --bfile FCGRv2merge --pheno fullmerge2.txt --pheno-name Outcome --covar p1t1cov.txt --covar-na
```

R set up:

```
library(tidyverse)
setwd("C:/Users/mjohnson/Documents/genetics/plink-1.07-dos")
pca <- read_table2("./Plink2.eigenvec", col_names = TRUE)
eigenval <- scan("./Plink2.eigenval")
```

Make table:

```
Sex <- rep(NA, length(pca$IID))
Sex[grep("1", pca$IID)] <- "Male"
Sex[grep("2", pca$IID)] <- "Female"
# location
 Study <- rep(NA, length(pca$IID))
Study[grep("1", pca$IID)] <- "P1"
Study[grep("2", pca$IID)] <- "Tyger"
Study[grep("3", pca$IID)] <- "T1"
Study[grep("4", pca$IID)] <- "T2"

# outcome
Outcome <- rep(NA, length(pca$IID))
Outcome[grep("1", pca$IID)] <- "No Typhoid"
Outcome[grep("2", pca$IID)] <- "Typhoid"

# combine - if you want to plot each in different colours
Sex_Study <- paste0(Sex, "_", Study)

# remake data.frame
pca <- as_tibble(data.frame(pca, Outcome, Study))

# Convert to percentage variance explained
pve <- data.frame(PC = 1:10, pve = eigenval/sum(eigenval)*100)
```

Plots

```
# make plot
a <- ggplot(pve, aes(PC, pve)) + geom_bar(stat = "identity")
a + ylab("Percentage variance explained") + theme_light()
```

```
# plot pca
b <- ggplot(pca, aes(PC1, PC2, col = Outcome, shape = Study)) + geom_point(size = 3)
b <- b + scale_colour_manual(values = c("red", "blue"))
b <- b + coord_equal() + theme_light()
b + xlab(paste0("PC1 (", signif(pve$pve[1], 3), "%)")) + ylab(paste0("PC2 (", signif(pve$pve[2], 3), "%
```

Linkage disequilibrium pruning

- Remove snps that are in LD with eachother to include in final model

```
./plink2 --bfile FCGRv2merge --indep-pairwise 50 5 0.2 --out FCGRv2_LD
```

Generalised linear mixed effect model updated

- additional flags to the script include –adjust which ranks the snps and corrects for multiple testing
- counts which shows number of individuals with snp genotype rather than frequency

```
#linkage disequilibrium glm
./plink2 --bfile FCGRv2merge --exclude  FCGRv2_LD.prune.out --pheno fullmerge2.txt --pheno-name Outcome

#Typhoid  pruned
./plink2 --bfile FCGRv2merge --exclude  FCGRv2_LD.prune.out --pheno typhoid_pheno.txt --pheno-name Outc

#Typhoid no pruning
./plink2 --bfile FCGRv2merge --pheno typhoid_pheno.txt --pheno-name Outcome --covar plink2.eigenvec --c


#Paratyphoid
./plink2 --bfile FCGRv2merge --exclude  FCGRv2_LD.prune.out--pheno paratyphoid_pheno.txt --pheno-name O

./plink2 --bfile FCGRv2merge --exclude  FCGRv2_LD.prune.out --pheno paratyphoid_pheno.txt --pheno-name
```

Output genotype information

- Significant typhoid SNPs are rs4657039, rs11590932
- Recode output to ped files which show the genotype for that SNP
- Alternatively output entire ped file and filter columns for significant snps in R
- Options to recode alleles to 12 or 01

```
./plink2 --bfile T1T2_FCGRv2  --pheno typhoid_pheno.txt --snp rs4657039  --recode --out t1t2snp_rs46
./plink2 --bfile T1T2_FCGRv2  --pheno typhoid_pheno.txt --snp rs11590932  --recode --out t1t2snp_rs11
```