Introduction

The series of practicals this morning will introduce you to analyzing Genome Wide Association Study (GWAS) datasets using a program called PLINK, which is a freely available GWAS analysis toolkit. PLINK has many functions including those related to data organization, formatting, quality control, association testing, population stratification and much more. Details about PLINK and its documentation are available at the reference section at the end of these practicals.

We will use the example files in the */faculty/barrett/2011/gwas-practical* directory. These files have are based on HapMap data with simulated disease effects. They serve to illustrate important points in working with GWAS data. We will use `plink` commands in the terminal to perform analyses. We can use the text editor, UNIX commands, and R (as introduced earlier this week) to help interpret results. Checkpoint questions in the practicals are **shown in bold**.

Start by making a working folder and copying the practical files to it. In the terminal:

```
mkdir gwas-practical
cp /faculty/barrett/2011/gwas-practical/* gwas-practical/
cd gwas-practical
```

PLINK is run from the terminal command line by typing, “plink”. In these practicals we will show commands to be typed in the terminal like this:

```
plink --file small-example
```

In this case we’ve asked PLINK to simply load some input data and give us a basic summary (since no additional options were specified). The data come in two files: a *.map* file which contains information about the SNPs included (in this case, just one) and a *.ped* file which contains information about the individuals and their genotypes, with one line per individual. You can look at these files in your text editor. **What is the name of the SNP in this dataset? How many cases and controls are included?**

We can do something more interesting with these files by running:

```
plink --file small-example --assoc
```

Now PLINK has generated two files in the directory we are working in: *plink.log* and *plink.assoc*. The log file simply captures the status information that PLINK reports with each run. If we look in the assoc file we will see that this single SNP is not associated with our phenotype.

The final step in this introduction is to learn to rename PLINK’s output files, since we’ll be generating lots of them in the practicals.

```
plink --file small-example --assoc --out getting-started
```

Now the output files will be named *getting-started.log* and *getting-started.assoc*. This is the basic pattern to working with PLINK: specifying input files and analyses, along with an output name to save results.

Quality Control Practical

We will now work with a set of data files containing many SNPs from chromosome 20 genotyped on a small number of cases and controls. Data from a GWAS would contain genotypes for thousands of individuals at SNPs across the entire genome, but we will focus on just one chromosome for fewer than 200 samples to make the exercises more tractable.

The key files are:

gwas-example.bed
gwas-example.bim
gwas-example.fam

Like the *.ped* and *.map* above, these files contain information about the samples and SNPs, as well as the genotypes for each of the samples at each of the SNPs. Unlike the human-readable text *.ped* file we used before, these data are in “binary ped” format (*.bed*). This format is a compressed format, which saves space and speeds up analysis. Information on samples can be found in the *.fam* file and information on SNPs in the *.bim* file. We can load data in these formats using the `--bfile` option.

```
plink --bfile gwas-example
```

How many SNPs and samples are in this dataset?

Reformatting between .bed/.bim/.fam and .ped/.map is easy, using the `--recode` option in PLINK. Similarly, converting the other way (i.e. from .ped to .bed):

```
plink --file small-example --make-bed --out small-example
```

Now we can look at some quality control statistics about our dataset:

```
plink --bfile gwas-example --missing --out miss-info
```

This will produce a *.imiss* file with information about individuals and *.lmiss* with information about loci (SNPs). You can load this output into theHaploview program to look at it in more detail.

```
java -jar /faculty/barrett/2011/gwas-practical/Haploview.jar
```

Use the 'PLINK format' option to load your `.lmiss` file (tick the 'Integrated Map Info' option, as well). **What SNP has the highest missing rate?**

Similarly, we can examine the frequencies of the SNPs in our data:

```
plink --bfile gwas-example --freq --out freq-info
```

We can use R to visualize these results. Launch R as in previous practicals. Commands to be typed in R (rather than in the terminal) will be shaded like this:

```
IN R WINDOW:

data<-read.table("freq-info.frq",h=T)

hist(data$MAF,breaks=20)
```

We can see that in our sample the SNPs are relatively evenly distributed across allele frequencies.

We can use these quality control metrics (and others described in the lectures) to create a clean dataset (note that although we need to break long commands in this document over two lines, all options should be typed on one continuous line in the terminal without breaks):

```
plink --bfile gwas-example --geno 0.05 --mind 0.05
      --hwe 1e-6 --maf 0.01 --make-bed --out gwas-clean
```

What filters have we applied here? (the PLINK website can help understand some of these options)

Association Practical

Now we will test for association between the SNPs in our dataset and disease.

Basic association tests can be done as follows:

```
plink --bfile gwas-example --assoc --out basic-test
```

Load these results in Haploview to investigate them further. **How many SNPs are associated? Is this what you expect?**

We can again use R to visualize these data:

```
IN R WINDOW:
data<-read.table("basic-test.assoc",h=T)
plot(data$BP,-log(data$P)/log(10),ylim=c(0,15))
```

What does this plot of association p-values across our data tell you?

You can save a picture of this analysis for later reference:

```
IN R WINDOW:
png("basic-association.png")
plot(data$BP,-log(data$P)/log(10),ylim=c(0,15))
dev.off()
```

We can now test for association within the cleaned dataset we created earlier:

```
plink --bfile gwas-clean --assoc --out clean-test
```

Read this new dataset into R (as above) and look at the plot of association p-values. **How has cleaning the data affected our signals of association? What does that imply about the associations seen in the previous analysis?**

In addition to using data cleaning to remove strong false positive associations, we are also interested in the overall distribution of test statistics. One way to look at this is to compare the observed data to our expectation under the null (remember to type each command as one long entry):

```
IN R WINDOW:
median(data$CHISQ)


expected<-qchisq(seq(0,1-1/length(data$CHISQ),
          by=1/length(data$CHISQ)),1)


plot(expected,sort(data$CHISQ))
abline(0,1)
```

This figure is called a Q-Q plot and can be very useful in evaluating GWAS data for systematic bias.

The expected median of the chi-square distribution with one degree of freedom is 0.455. The diagonal line shows where the points should fall under the null. **Given this information, what can you infer about our current association analysis?**

Using the PLINK website, can you identify how to run more general tests of association (beyond the basic test we've done already)?

Population Structure Practical

We have seen how applying appropriate quality control filters to our data eliminated many false positives, but a systematic inflation remained. We can use PLINK's multidimensional scaling procedure to extract information about ancestry which might correct for the inflation.

```
plink --bfile gwas-clean --cluster --mds-plot 2 --out gwas-mds
```

We can visualize this analysis in R. At first the clustering doesn't seem informative, but then try coloring the individual samples by affection status:

```
IN R WINDOW:
mds<-read.table(“gwas-mds.mds”,h=T)

plot(mds$C1,mds$C2)

samples<-read.table(“gwas-example.fam”)

case_status<-samples$V6

plot(mds$C1,mds$C2,col=case_status)
```

This analysis tells us that the MDS components do seem to be correlated with disease status and may be relevant to the inflation we observe. Run a logistic regression corrected for these two dimensions:

```
plink --bfile gwas-clean --logistic --covar gwas-mds.mds
      --covar-number 2,3 --out mds-corrected --hide-covar
```

Note that this analysis produces a logistic regression, rather than chi-square tests as above. Hint: squaring the STAT column should produce a chi-square distributed test statistic.

Has stratifying by disease group corrected our inflation?

Are there any disease associations in this dataset?

Of course, if we have information about the ancestral groups our samples are descended from (e.g. in the file *gwas-example.pop*) we can correct more directly:

```
plink --bfile gwas-clean --mh
      --within gwas-example.pop --out stratified-test
```

We have now produced an analysis stratified by group membership.

Which method of correcting more completely removes the inflation? Why?

How are the QC statistics for any associated SNPs? Do you believe these associations?

What could explain differences in association p-values in the different analyses we've done?

Produce Q-Q plots, genome-wide p-value plots and a summary of your results.

Reference Information

You can download PLINK, and find much more information at the website:

http://pngu.mgh.harvard.edu/~purcell/plink/

There is detailed documentation about all the options available, file formats and examples of commands. A detailed tutorial (similar to work we have done here) is available at:

http://pngu.mgh.harvard.edu/~purcell/plink/tutorial.shtml