Introduction

The series of practicals this afternoon and tomorrow morning will introduce you to analyzing Genome Wide Association Study (GWAS) datasets using a program called PLINK, a freely available GWAS analysis toolkit. PLINK has many functions for: 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 this document.

We will use the example files in the */faculty/barrett/2017/gwas-practical* directory. These files are based on HapMap data with simulated disease effects. We will use `plink` commands in the terminal to perform analyses. The examples illustrate some important points in working with GWAS data, and the practical will hopefully get you comfortable working with genetic data in UNIX.

We will use the text editor, UNIX commands, RStudio and a couple of other programs to help interpret results. You can copy and paste specific commands from this document to your UNIX terminal, or into the R window.

```
UNIX terminal commands are shown in a monospaced font.
```

```
Commands for R are in monospace on a grey background.
```

Checkpoint questions are shown in bold.

1. The basics

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

```
mkdir gwas-practical
cp –r /faculty/barrett/2017/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 --assoc
```

This pattern is typical of PLINK commands will have multiple options (words beginning with “--”), some of which take arguments, such as a filename, and others which tell PLINK what kind of analysis we want to do. We’ll see lots of examples of this in action, below.

The data for this small example 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, or using the `less`

command in the terminal. **What is the name of the SNP in this dataset? How many cases and controls are included?**

Let's look at that command we introduced earlier:

```
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. The assoc file has information about a basic case-control association test: if we look at this 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.

2. Power and sample size

This power calculator from the University of Michigan is a helpful tool for understanding GWAS power:

http://csg.sph.umich.edu/abecasis/CaTS/gas_power_calculator/index.html

You can play around with the values in the calculator to match what you might expect for a variant associated with a complex disease, or to match some data you might have yourself.

A 2007 GWAS of 988 Crohn's disease cases (prevalence about 1 in 400 people) and 1007 controls by Rioux *et al.* reported associations near the genes *ATG16L1* and *PHOX2B*, with relevant parameters shown below.

SNP	**Disease allele frequency**	**Genotype relative risk**
rs2241880	0.547	1.47
rs16853571	0.923	1.45

Calculate their power to detect these associations at genome-wide significance ($5x10^{-8}$).

Does either surprise you?

Which might you expect has been consistently replicated in subsequent studies?

Change the main plot to show power vs. controls. How might projects like the UK Biobank (with 500,000 GWAS'd individuals) be useful?

3. 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 it possible to run the exercises on a laptop.

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 --write-snplist
```

This command somewhat boringly just writes a list of the SNPs in our file. **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:

```
plink --bfile gwas-example --recode --out gwas-example
```

Use `ls -l` on the terminal to look at the new files. **How different are the file sizes for the .ped and .bed versions?**

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 the *Haploview* program (in the Workshop dropdown menu at the top of the screen) to look at it in more detail. Choose the 'PLINK format' option and select your `.lmiss` file as the "Results File" and select the `.map` file for the "Map File". **What SNP has the**

highest missing rate? You could also browse results in this way in Excel or a similar program, or directly from the command line. **(Bonus: can you identify the SNP with the highest missing rate with a one line UNIX command?)**

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

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

We can use R to visualize these results. Launch RStudio from the Workshop menu, and then from the *Session* menu choose *Set Working Directory*, and choose the folder you're working in. 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. **Is this surprising? Why or why not?**

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)

While we aim to use QC thresholds like those applied above to filter out badly genotyped SNPs, these filters aren't always perfect. It's a good idea to always visualize the raw intensity data, or "cluster plots" for any SNP that you're particularly interested in (e.g. those that show strong association).

Launch the *Evoker* from the Workshop drop-down menu. From the *File* menu choose, *Open directory*, select the *rawdata* folder from the folder you're working on, and click *Open*. You can click *Random* to see a randomly chosen SNP (note these data aren't the same as what we've worked with so far, but have been chosen to be illustrative). Try mousing over data points, or clicking and dragging to zoom.

Load the list of SNPs, *qc-snps.txt*, to check by choosing from the *File* menu, *Load marker list*. You can "vote" on each SNP using the buttons at the bottom, which will automatically show the next one. **Which SNPs have problems? Which would be most worrying if you had observed association?**

4. 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 (the “Plot” button might be helpful). **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$BP,-log10(data$P),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$BP,-log10(data$P),ylim=c(0,15))
dev.off()
```

We can now test for association with 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 of these three commands 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)?

PLINK can be used to analyze quantitative traits as well as case-control data. **Try using the `--linear` option to analyze the *gwas-example.qt* file which describes a trait measured in our same set of individuals. What do you find?**

5. 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 (analogous to principal components analysis) 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
      --ci 0.95
```

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 ancestry 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 find detailed documentation about PLINK, including information about all the options available, file formats, and examples of commands at the website:

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

You should use version 1.9 of PLINK, which has been rewritten to be much faster. It's linked from the page above, or available at:

https://www.cog-genomics.org/plink2/

A detailed tutorial (similar to work we have done here) is available at:

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