Genome wide association studies

Gibran Hemani

This worksheet will take you through:

1. Performing an association test for a single SNP
2. Scaling up to calculating associations for each SNP
3. Visualising the results
4. The consequences of population stratification
5. Finding out about our GWAS hits

We'll be using plink software, and also the genotype data that you cleaned in the "Genetic data" practical (`data/genetic/geno_qc`).

We'll be storing results in the `results/gwas` folder. Navigate there:

```
cd results/gwas
```

Performing an association test for a single SNP

In the phenotype file (`../../data/genetic/phen.txt`) we have phenotypic information for

- BMI (body mass index)
- DBP (diastolic blood pressure)
- SBP (systolic blood pressure)
- CRP (C reactive protein)
- HT (hypertension)

```
head ../../data/genetic/phen.txt
```

```
FID IID BMI DBP SBP CRP HT
id1 id1 31.8664742886873 87.2331807479429 152.290193244483 1.44468000003555 2
id2 id2 35.4253325607032 86.059532896277 143.082692573686 3.75456412209453 2
id3 id3 29.7780898802738 98.7465737573734 234.123838391363 0.514857650792626 2
id4 id4 34.7971502072983 102.912235367749 198.501027761789 4.93651456826931 2
id5 id5 26.9278590031599 77.8172291593599 151.038730334376 3.38852808547912 2
...
```

And we have a list of SNPs that are in our data, e.g.

```
head ../../data/genetic/geno_qc.bim
```

```
1   rs12562034  0   768448  A   G
1   rs9442372   0   1018704 A   G
1   rs3737728   0   1021415 A   G
1   rs6687776   0   1030565 T   C
1   rs9651273   0   1031540 A   G
1   rs4970405   0   1048955 G   A
```

```
1   rs12726255  0   1049950 G   A
1   rs11807848  0   1061166 C   T
1   rs9442373   0   1062638 C   A
...
```

Let's just take the first SNP - **rs12562034** - as an example. Every individual in the data has genotype data for this SNP, which might look something like this:

```
id138 1
id139 2
id140 0
id141 0
id142 0
id143 0
id144 0
id145 2
id146 1
...
```

The BMI values for these individuals look like this:

```
id138 28.0
id139 41.7
id140 35.9
id141 30.8
id142 35.7
id143 31.8
id144 34.3
id145 30.7
id146 26.5
...
```

Our question is whether there is an association between the genotype values and the trait (BMI) values for this SNP. We can calculate the association for this SNP using plink:

```
../../software/plink_mac \
--bfile ../../data/genetic/geno_qc \
--pheno ../../data/genetic/phen.txt \
--pheno-name BMI \
--assoc \
--snp rs12562034 \
--out bmi_rs12562034
```

This command has the following flags:

- `--pheno ../../data/genetic/phen.txt` is saying use the phenotype data in this file
- `--pheno-name BMI` is saying use the BMI column from the phenotype file
- `--assoc` is saying perform an association test between trait and genotypes
- `--snp rs12562034` is saying only perform the association for this SNP
- `--out bmi_rs12562034` is saying save the results with this prefix

This command produces a new file called `bmi_rs12562034.qassoc`. It looks like this:

```
CHR         1
SNP         rs12562034
BP          768448
NMISS       8237
BETA        -0.03543
SE          0.1437
R2          7.379e-06
T           -0.2465
P           0.8053
```

(note - this is transposed here for ease of reading).

The SNP is associated with BMI with a slope of `-0.03543`, but with a large standard error `0.1437`. The p-value is `0.8053` which suggests that this is not an important SNP for BMI.

Performing the genome-wide association study

We have seen what is happening in an association test between a SNP and a trait. We can just as easily perform this same operation for every SNP against the same trait. In plink it is simply done like this:

```
../../software/plink_mac \
--bfile ../../data/genetic/geno_qc \
--pheno ../../data/genetic/phen.txt \
--pheno-name BMI \
--assoc \
--out bmi_assoc
```

The difference here is we have simply removed the `--snp rs12562034` flag - it will now automatically perform the test for every SNP against the trait (e.g. it will perform 463017 tests). This produces a file called `bmi_assoc.qassoc`.

Visualising the results

There are two important aspects to visualising the results.

1. Identifying regions of the genome that are associated with the trait
2. Checking that there is no evidence for technical artifacts driving our results

Identifying significant regions

The standard method for doing this is to make a Manhattan plot. Let's do this in R.

The following commands are to be run in R Studio

First, open up R Studio and set the working directory to where our results are (`results/gwas`). Go to `Session -> Set working directory -> Choose directory...`.

Let's read in the results:

```
bmi <- read.table("bmi_assoc.qassoc", header=TRUE)
```

How many rows are there?

```
nrow(bmi)
```

What does the data look like?

```
head(bmi)
```

This is a function for creating a Manhattan plot:

```
manhattan_plot <- function(p, chr, pos, filename=NULL, width=15, height=7, threshold=-log10(0.05/1000000
{
    require(ggplot2)
    dat <- data.frame(chrom=as.numeric(chr), bp=pos, pval=-log10(p))
    dat <- dat[order(dat$chrom, dat$bp), ]
    dat$col <- dat$chr %% 2 + 1
    dat <- subset(dat, !is.na(pval))

    pl <- ggplot(dat, aes(x=bp, y=pval)) +
    geom_point(aes(colour=factor(col))) +
    facet_grid(. ~ chrom, scale="free_x", space="free_x") +
    theme(legend.position="none") +
    scale_colour_manual(values=c("#404040", "#ca0020")) +
    theme(axis.text.x=element_blank(), axis.ticks.x=element_blank()) +
    ylim(0, max(c(threshold, dat$pval, na.rm=TRUE))) +
    labs(y=expression(-log[10]*p), x="Position") +
    geom_hline(yintercept=threshold)

    if(!is.null(filename))
    {
        ggsave(filename, pl, width=width, height=height)
    } else {
        print(pl)
    }
}
```

```
manhattan_plot(bmi$P, bmi$CHR, bmi$BP, filename="bmi_assoc_manhattan.png")
```

This produces a file called `bmi_assoc_manhattan.png`. Please open this file. It looks like there are some pretty big signals for certain regions of the genome!

Checking the validity of the GWAS

Under the null hypothesis, that no SNPs are associated with the trait, we expect a set of p-values that **are uniformly distributed**. Because we are performing 500000 tests, we might expect that the majority of the p-values follow this uniform distribution, with only a few of the SNPs departing from this distribution to have more extreme values. This can be visualised using a Q-Q plot. The R package **GenABEL** has a convenient function for doing this.

```
library(GenABEL)
```

Now make the Q-Q plot

```
png("bmi_assoc_qq.png")
estlambda(bmi$P, plot=TRUE, method="median")
dev.off()
```

This produces a file called `bmi_assoc_qq.png` and the following output in R:

```
$estimate
[1] 1.094061
```

Let's first look at the plot. The black line represents the line of expectation under the null - a slope of 1. The red line shows the actual relationship between expected p-values and observed p-values - here the slope looks much higher than 1. Each point represents a SNP's p-value (ordered).

The `estimate` value printed above represents a value known as **lambda**. This is calculated as follows:

```
median observed test statistic / median expected test statistic
```

A value of 1 indicates no artificial inflation. If the median value of our p-values is more extreme than what we expect by chance it is a strong indication that all of our test statistics are being systematically inflated.

The value of 1.09 is a cause for concern here - what could be causing this?

You can now return back to the Terminal

Including covariates in the GWAS

So far our GWAS did not include covariates. We have a covariates file, located here (run this in the Terminal, not in R):

```
less ../../data/genetic/covs.txt
```

It contains principal components 1-10, sex, age and smoking status. To perform linear regression of SNP against BMI, adjusting for these covariates, we can use the following plink command:

```
../../software/plink_mac \
--bfile ../../data/genetic/geno_qc \
--pheno ../../data/genetic/phen.txt \
--pheno-name BMI \
--covar ../../data/genetic/covs.txt \
--covar-name PC1-PC10, age, sex \
--linear \
--snp rs12562034 \
--out bmi_adjusted_rs12562034
```

The flags here are:

- `--covar ../../data/genetic/covs.txt` specifies the new covariate file to use
- `--covar-name PC1-PC10, age, sex` specifies which covariates to use
- `--linear` specifies that a full linear regression should be performed. Previously we were using `--assoc` which is a fast approximation, and which doesn't handle covariates.

The command creates a new file called `bmi_adjusted_rs12562034.assoc.linear`. It contains the association for the SNP and the trait, but it also contains the association of the covariates for the trait when they are fitted together in the following model:

```
bmi ~ rs12562034 + pc1 + pc2 + ... + pc10 + age + sex
```

It is clear from these results that some of the principal components are strongly associated with the trait

TEST	NMISS	BETA	STAT	P
ADD	8237	-0.005021	-0.04026	0.9679
PC1	8237	123	25.12	3.352e-134
PC2	8237	130.9	26.73	5.195e-151
PC3	8237	-2.022	-0.4131	0.6796
PC4	8237	118.3	24.16	1.141e-124
PC5	8237	0.3503	0.07155	0.943
PC6	8237	123.3	25.19	6.504e-135
PC7	8237	4.308	0.8799	0.3789
PC8	8237	-1.695	-0.3462	0.7292
PC9	8237	6.9	1.409	0.1589
PC10	8237	2.282	0.4662	0.6411
age	8237	-0.05786	-1.074	0.2831
sex	8237	1.333	12.35	9.853e-35

(only the relevant columns shown here). We should perform the GWAS again, making sure to adjust for PCs and covariates. The code to do this is similar to above but without the `--snp` flag,

```
../../software/plink_mac \
--bfile ../../data/genetic/geno_qc \
--pheno ../../data/genetic/phen.txt \
--pheno-name BMI \
--covar ../../data/genetic/covs.txt \
--covar-name PC1-PC10, age, sex \
--linear \
--out bmi_adjusted
```

NOTE THAT THIS MAY TAKE A LONG TIME TO RUN To cancel it press `ctrl+c`.

In the interests of time, the results from this command have been precomputed. Let's look at how the Manhattan and Q-Q plots look now.

Perform the following commands in R

Read in the **adjusted** BMI GWAS results

```
bmi_adj <- read.table("precomputed/bmi_adjusted.assoc.linear.add.gz", header=TRUE)
```

Create the Manhatton plot:

```
manhattan_plot(bmi_adj$P, bmi_adj$CHR, bmi_adj$BP, filename="bmi_adjusted_manhattan.png")
```

Notice that the p-values are much less extreme now, and there are fewer peaks. Estimate lambda, and create the Q-Q plot:

```
png("bmi_adjusted_qq.png")
estlambda(bmi_adj$P, plot=TRUE, method="median")
dev.off()
```

The lambda estimate is now much lower - 1.03. Still not perfect but this is much more acceptable.

Finding out about our GWAS hits

How many SNPs are below the threshold of `5e-8`?

```
table(bmi_adj$P < 5e-8)
```

Which is our top SNP in this GWAS? In R:

```
bmi_adj[which.min(bmi_adj$P),]
```

We can use online tools to find out about any particular SNP. There are many to choose from, which we will cover later on, but for now let's just use the UCSC Genome Browser

http://genome.ucsc.edu/cgi-bin/hgGateway

Put an rs ID into the search. When the result comes up, try zooming out a few times. Which is the closest gene? Click on the gene and find out more information about it.

Questions

1. Load the CRP GWAS results into R.
    - How many SNPs are "significant" for the CRP GWAS?
    - Which SNP has the smallest p-value?
    - Which gene is closest to this SNP?
2. Repeat question 1, but for the hypertension GWAS.