Day 3 AM: Practical 1

Adjusting for population structure using multidimensional scaling in PLINK

In this practical, we will analyse a simulated GWAS of 500 cases and 500 controls from a population with Northern European ancestry. The samples and SNPs have been previously cleaned using the QC methods discussed earlier in this course. In reality, we would analyse genotype data from all autosomes. However, because of the time constraints of the practical session, we will demonstrate the analysis protocol using genotype data from chromosome 1 only.

In this session, we will go through the following steps:

1. Perform a test of association at each SNP using PLINK, and evaluate the genomic control inflation factor to assess evidence for the presence of population structure using R.
1. Perform MDS of the genotype data from the case-control study combined with that from 180 HapMap samples using PLINK, and identify population outliers by plotting the first two of the axes of genetic variation.
2. Perform a test of association at each SNP using PLINK, excluding the population outliers, and re-evaluate the genomic control inflation factor to re-assess evidence for the presence of population structure using R.
3. Perform MDS of the genotype data from the case-control study (excluding population outliers) using PLINK (this time without HapMap samples).
4. Perform a test of association at each SNP using PLINK, excluding the population outliers, adjusting for the "important" axes of genetic variation from the MDS, and re-evaluate the genomic control inflation factor to assess evidence for fine-scale population structure that we have not yet accounted for.

Testing for association without accounting for population structure

Begin by going to the directory for the DAY3-AM-1 practical. The files `psdata.ped` and `psdata.map` contain the sample and genotype information for more than 20,000 SNPs on chromosome 1. We can perform a test of association at each SNP, in a logistic regression model, assuming additive genetic effects using PLINK with the command:

```
plink --file psdata --logistic --out stage1
```

The output of the analysis is reported in the file `stage1.assoc.logistic`. Look at the first few lines of the file, and check that you understand the output.

We can use R to evaluate the genomic control inflation factor from the PLINK output. In another window, go to the DAY3-AM-1 practical directory and start R. Read in the PLINK output using the command:

```
snpdat <- read.table("stage1.assoc.logistic",header=T)
```

The *p*-values from the trend test are stored in column 9, so we can calculate the genomic control inflation factor using the command:

```
median(qchisq(snpdat[,9],df=1,lower.tail=F),na.rm=T)/0.456
```

What do you conclude about population structure? You can produce a QQ plot for the unadjusted analysis using the commands:

```
index <- -log10(seq(1,nrow(snpdat))/nrow(snpdat))
logp <- -log10(snpdat[,9])
qqplot(index,logp,xlab="Expected -log10 p",ylab="Observed -log10 p")
lines(index,index)
```

Is there any evidence of association of SNPs with disease, over and above that which we would expect by chance? Which SNPs have the strongest evidence of association with disease? Hint: if you want to see the kth smallest *p*-value (stored in column 9), and the corresponding SNP ID (stored in column 2), you can type the commands:

```
snpdat[order(logp,decreasing=TRUE)[k],9]
snpdat[order(logp,decreasing=TRUE)[k],2]
```

where you replace `k` by the appropriate numeric value.

Identification of population outliers

We will identify population outliers by applying MDS to the study samples, combined with 180 individuals from the international HapMap project. We have already generated the HapMap files for the same set of chromosome 1 SNPs in the files `hapmap.ped` and `psdata.map`. To merge the genotype data of the study samples and HapMap individuals across the SNPs, we can use the following command in PLINK:

```
plink --file psdata --merge hapmap.ped hapmap.map --recode --out merge
```

Before performing MDS, it is important to prune SNPs for linkage disequilibrium. This can be done with the following command in PLINK:

```
plink --file merge --indep-pairwise 50 5 0.2 --out merge
```

This command uses a sliding window of 50 SNPs, moving across 5 SNPs to the next window, and retains only those SNPs with $r^2<0.2$. The list of independent SNPs is stored in the file `merge.prune.in`.

We next calculate the "genome" file, which contains the pairwise IBS between samples, and measures the extent of their genetic similarity across the set of independent SNPs. This can be achieved with the following command in PLINK:

```
plink --file merge --extract merge.prune.in --genome --out merge
```

We then perform MDS of the pairwise IBS using the following command in PLINK:

```
plink --file merge --read-genome merge.genome --cluster --mds-plot 2 --out
merge
```

The `--mds-plot 2` part of the command tells PLINK to generate two axes of genetic variation, which are reported in the file `merge.mds`. Look at the first few lines of the file, and check that you understand the output.

We can plot the study samples and HapMap individuals on the first two axes of genetic variation in R. Go back to the window in which you already have R opened, and read in the data from the MDS output file:

```
pca = read.table("merge.mds",header=T)
```

To distinguish the study samples from HapMap individuals from different ethnic groups, we have prepared a simple text file, `merge.txt`, that contains the "status" of each individual: Case, Control, CEU, YRI or CAJ. To combine this information with the axes of genetic variation for each individual, we can merge by individual ID (IID) in R, using the following commands:

```
status = read.table("merge.txt",header=T)
pcastat = merge(pca, status, by.x="IID", all.x=T)
```

We can then plot the first two axes of genetic variation in order to identify potential outliers by typing the following commands:

```
plot(pcastat[,4],pcastat[,5],xlab="PC 1",ylab="PC 2",type="n")
for(i in 1:nrow(pcastat)){
if(pcastat[i,6]=='CEU') points(pcastat[i,4],pcastat[i,5],col=2)
if(pcastat[i,6]=='YRI') points(pcastat[i,4],pcastat[i,5],col=3)
if(pcastat[i,6]=='CAJ') points(pcastat[i,4],pcastat[i,5],col=4)
if(pcastat[i,6]=='Case') points(pcastat[i,4],pcastat[i,5],col=1,pch=19)
if(pcastat[i,6]=='Control') points(pcastat[i,4],pcastat[i,5],col=1,pch=19)
}
```

We would expect samples of European ancestry to cluster closely with CEU HapMap samples (highlighted in red open circles). Is there evidence of individuals (solid black dots) with outlying population ancestry? If so, what are the IDs of these individuals? You can do this using the following command in R:

```
identify(pcastat[,4],pcastat[,5],labels=pcastat[,2])
```

You can then put the mouse pointer over any point in the graph and click the left mouse button to produce the corresponding sample ID. When you have finished identifying all individuals, press the right mouse button over the R graphics window.

Note that all outliers have PC 1 greater than 0.01. You can visualise the threshold on the plot with the following command in R:

```
abline(v=0.01)
```

We can then list the sample IDs with the following command in R:

```
pcastat[(pcastat$C1>0.01 & pcastat$Status %in% c('Case','Control')),]$IID
```

Testing for association excluding population outliers

Our next step is to repeat the association analysis, excluding the population outliers identified from the PC analysis with HapMap samples. To do this, we need to generate a new file called `sample.exclusions`, and put the family ID (FID) and individual ID (IID) of each excluded individual on a separate line. Note that in this example, FID and IID are the same, so to remove the sample with IID ST2000, you would add a line:

```
ST2000 ST2000
```

We can do this using the following command in R:

```
ids = pcastat[(pcastat$C1>0.01 & pcastat$Status %in%
c('Case','Control')),1:2]
write.table(ids, file="sample.exclusions", row.names=F, quote=F,
col.names=F)
```

We can perform a test of association at each SNP, in a logistic regression model, assuming additive genetic effects, and excluding the population outliers, using PLINK with the command:

```
plink --file psdata --remove sample.exclusions --logistic --out stage2
```

As before, we can use R to evaluate the genomic control inflation factor from the PLINK output. Go to the window in which you have R open, and use the same commands as before:

```
snpdat <- read.table("stage2.assoc.logistic",header=T)
median(qchisq(snpdat[,9],df=1,lower.tail=F),na.rm=T)/0.456
```

What do you now conclude about population structure? As before, we can produce a QQ plot for the unadjusted analysis using the commands:

```
index <- -log10(seq(1,nrow(snpdat))/nrow(snpdat))
logp <- -log10(snpdat[,9])
qqplot(index,logp,xlab="Expected -log10 p",ylab="Observed -log10 p")
lines(index,index)
```

Is there any evidence of association of SNPs with disease, over and above that which we would expect by chance? Which SNPs have the strongest evidence of association with disease?

Identification of axes of genetic variation describing population structure

In order to identify structure within our population, we can perform MDS of study samples, after exclusion of the ethnic outliers, using the following commands in PLINK:

```
plink --file psdata --remove sample.exclusions --indep-pairwise 50 5 0.2 --
out psdata
plink --file psdata --remove sample.exclusions --extract psdata.prune.in --
genome --out psdata
plink --file psdata --remove sample.exclusions --read-genome psdata.genome
--cluster --mds-plot 10 --out psdata
```

As before, we first prune SNPs for linkage disequilibrium, then calculate the genome file, and then perform MDS. This time, we have generated 10 axes of genetic variation. We can plot the study samples on the first two axes of genetic variation in R. Go back to the window in which you already have R opened, and read in the data from the MDS output file:

```
pca = read.table("psdata.mds",header=T)
```

To distinguish the cases and controls, we can use file `merge.txt`, as before, that contains the "status" of each individual. To combine this information with the axes of genetic variation for each individual, we can merge by individual ID (IID) in R, using the following commands:

```
status = read.table("merge.txt",header=T)
pcastat = merge(pca, status, by.x="IID", all.x=T)
```

We can then plot the first two axes of genetic variation in order to identify potential outliers by typing the following commands:

```
plot(pcastat[,4],pcastat[,5],xlab="PC 1",ylab="PC 2",type="n")
for(i in 1:nrow(pcastat)){
if(pcastat[i,14]=='Case') points(pcastat[i,4],pcastat[i,5],col=2,pch=19)
if(pcastat[i,14]=='Control') points(pcastat[i,4],pcastat[i,5],col=3,pch=19)
}
```

Is there any evidence of correlation between disease status and these two axes of genetic variation? You can use the same commands to plot any other pairs of axes of genetic variation, stored in columns 4-13 of the `pcastat` matrix.

We can test for association between disease status and the first ten axes of genetic variation using the command:

```
summary(glm(as.factor(pcastat[,14])~pcastat[,4]+pcastat[,5]+pcastat[,6]+
pcastat[,7]+pcastat[,8]+pcastat[,9]+pcastat[,10]+pcastat[,11]+pcastat[,12]+
pcastat[,13],family="binomial"))
```

This command uses logistic regression to model the relationship between disease status (stored in column 14 of the `pcastat` matrix) and the first ten axes of genetic variation, stored in columns 4-13 of the `pcastat` matrix. The command will provide parameter estimates, standard errors and *p*-values for each axis of genetic variation. Is there any evidence that the axes of genetic variation are associated with disease status?

Testing for association with adjustment for population structure

We can take account of population structure by including the "significant" axes of genetic variation as covariates in our trend test of association. This can be performed in PLINK using the following command, which will adjust for the first two axes of genetic variation:

```
plink --file psdata --remove sample.exclusions --logistic --covar
psdata.mds --covar-name C1 C2 --hide-covar --out stage3
```

The `--hide-covar` command instructs PLINK not to report the regression coefficients for the covariates in the logistic regression output. You can adjust for other covariates by listing them after the `--covar-name` command.

As before, we can use R to evaluate the genomic control inflation factor from the PLINK output. Go to the window in which you have R open, and use the same commands as before:

```
snpdat <- read.table("stage3.assoc.logistic",header=T)
median(qchisq(snpdat[,9],df=1,lower.tail=F),na.rm=T)/0.456
```

What do you now conclude about population structure? As before, we can produce a QQ plot for the unadjusted analysis using the commands:

```
index <- -log10(seq(1,nrow(snpdat))/nrow(snpdat))
logp <- -log10(snpdat[,9])
qqplot(index,logp,xlab="Expected -log10 p",ylab="Observed -log10 p")
lines(index,index)
```

Is there any evidence of association of SNPs with disease, over and above that which we would expect by chance? Which SNPs have the strongest evidence of association with disease?

Brief solutions

The initial analysis yields a genomic control inflation factor of 1.157, indicating evidence of population structure that has not been taken account of in the analysis. There is a single SNP with evidence of association over and above that which would be expected by chance: rs4908527 (p=8.9x10^{-8}).

MDS of the study samples together with HapMap data demonstrates that the first two axes of genetic variation separate all populations from each other. Plotting the first two components against each other highlights ten outliers: ST1991-ST2000. When you remove these ten samples from the association analysis, the genomic control inflation factor reduces to 1.082. The same SNP, rs4908527, still has strongest evidence of association (p=4.3x10^{-8}). The evidence of association is stronger than in the analysis of all samples.

MDS of the study samples identifies two axes of genetic variation that are strongly associated with disease: C1 and C2. Association analysis of the study samples, adjusting for these two axes of genetic variation reduces the genomic control inflation factor to 1.042. The same SNP, rs4908527, still has strongest evidence of association (p=1.1x10^{-7}), suggesting that this is not a false positive resulting from population structure.