Pleiotropy Exercise

Andrew DeWan, PhD, MPH

This exercise was designed to give you practical experience identifying cross phenotype associations using both univariate and multivariate methods and then dissecting these cross phenotype associations to determine if they are examples of biological or mediated pleiotropy. Two population-based datasets have been simulated (dataset1 and dataset2) each with 100,000 subjects. Each dataset contains two correlated phenotypes; there are markers associated with one or both phenotypes as well as unassociated. For practical reasons (file size and minimizing run times), only markers on a small number of chromosomes are provided in each dataset for this exercise.

Dataset1 contains two dichotomous phenotypes, W1 and W2, each with a population prevalence of 0.2. They have a correlation in the population of ~0.2. When the study was conducted, information about W1 was ascertained by asking about a doctor's diagnosis of W1 at least 20 years prior to enrollment in the study. Information about W2 was ascertained at enrollment (i.e. W1 may potentially be the mediator between a genetic variant and W2).

Similarly, Dataset2 contains two dichotomous phenotypes, X1 and X2, each with a population prevalence of 0.2. They have a correlation in the population of ~0.35. When the study was conducted, information about X1 was ascertained by asking about a doctor's diagnosis of X1 at least 10 years prior to enrollment in the study. Information about X2 was ascertained at enrollment (i.e. X1 may potentially be the mediator between a genetic variant and X2).

Both datasets have been QC'd and for the initial analyses no covariates are needed. The files for the initial analyses are:

Dataset1: dataset1.bed, dataset1.bim, dataset1.fam, dataset1_phenotypes.txt
Dataset2: dataset2.bed, dataset2.bim, dataset2.fam, dataset2_phenotypes.txt

1.) Univariate analyses

   a. Conduct a univariate analysis (using --assoc) in PLINK for both datasets and both phenotypes
      Note: You will need to use the --pheno/--pheno-name commands to specify the phenotype file and phenotype name. The phenotypes are coded 0 (controls) and 1 (cases), so you will also need to use the --1 flag.

   b. Within each dataset, pull out SNPs that have $p<1x10^{-5}$ for both phenotypes. This can be done using some simple R code (will need to edit for each dataset and depending on your output file names):

```
>phenW1 <- read.table("<W1 output>", header = T)
>phenW2 <- read.table("<W2 output>", header = T)
>SuggphenW1 <- subset(phenW1, P<0.00001)
>SuggphenW2 <- subset(phenW2, P<0.00001)
>intersect(SuggphenW1$SNP, SuggphenW2$SNP)
```

c. This code will print a list of the SNPs to explore in the multivariate analyses. In each dataset, create a subset of these genome-wide suggestive cross phenotype SNPs. This can be done in PLINK using the --extract command.

2. Multivariate analyses

a. Conduct a multivariate analysis using a PLINK extension program called MV-PLINK on the subset of SNPs from each dataset. Below is an example of the command (see all the multivariate plink manual):

```
>plink.multivariate --noweb --bfile dataset1_subset --mult-pheno
dataset1_phenotypes.txt --1 --mqfam --out dataset1_subset
```

Please note: You should use the --noweb flag due to this program being built on an old version of PLINK. The --1 flag indicates cases are coded as 1 and controls are 0, instead of the default coding (cases = 2, controls = 1).

3. Mediation analyses

a. In each dataset (Dataset1 or Dataset2): for the SNPs that are genome-wide significant cross phenotype associations in each dataset you will need to create a genotype file that is coded as 0|1|2 for the genotypes. This can be done in PLINK using the --recodeA command. This will give you a .raw genotype file that can be using in the mediation analysis.

b. Conduct a mediation analysis in R using the *mediation* R library. Sample code for this is below (Note: replace <SNP> with the variable name for the SNP you are investigating. You will need to repeat this for each SNP in both datasets):

```
>library(mediation)
>genotypes <- read.table("dataset1_subset.raw", header=T)
>phenotypes <- read.table("dataset1_phenotype.txt", header=T)
>combined <- merge(genotypes, phenotypes)
>head(combined)                    #to see variable names in combined dataset
>med.fit<-glm(W1~<SNP>, data=combined, family=binomial("logit"))
>out.fit<-glm(W2~W1+<SNP>, data=combined, family=binomial("logit"))
```

>med.out<-mediate(med.fit, out.fit, treat="<SNP>", mediator = "W1", boot = TRUE, boot.ci.type = "bca", sims = 1000)
>summary(med.out)

This will print out a summary of the mediation analysis. As noted during the lecture, you want to focus on the following four rows of the summary output: Total Effect, ACME (average), ADE (average), Prop. Mediated (average).

Please note: The more simulations (sims) you specific in the med.out step the more accurate the CI and p-value estimates will be, however, this can also be time-consuming. If this step is taking a substantial amount of time (>20 minutes) you may want to reduce the number of simulations for the purposes of completing the exercise.

Questions:

1) Which of the SNPs have genome-wide significant ($p<5x10^{-8}$) associations for both phenotypes within a dataset?
2) Did the multivariate analyses result in additional SNPs that had genome-wide significant cross phenotype associations? Which SNP(s)?
3) For each SNP analyzed in the mediation analysis, determine if there is evidence of biological or mediated pleiotropy. If mediated, is the mediation complete or incomplete?