Advanced Gene Mapping Course: 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 show evidence of biological and/or mediated pleiotropy.

A population-based dataset with 3000 subjects and two quantitative traits (Trait 1 and Trait 2) along with 2000 SNPs on one chromosome were simulated. Let's assume that Trait 1 was measured 20 years prior to Trait 2 (i.e. Trait 1 will act as the mediator in our mediation analysis). The two quantitative traits are correlated and there are markers associated with one or both phenotypes as well as unassociated.

The dataset has been QC'd. The files for the initial analyses are:

pleiotropy_exercise.bed, .bim, .fam and pleiotropy_exercise_phenotypes.txt

I have included a summary table that you will want to fill out as you are working through this exercise. This will help keep track of the SNPs you select for the mediation analysis as well as the interpretation of the results at the end of the exercise

Univariate analyses

a. Conduct a univariate analysis (using --linear) in PLINK for both datasets and both traits
   Note: You will need to use the --pheno/--pheno-name commands to specify the phenotype file and phenotype name.

   ```
   plink\
    --bfile pleiotropy_exercise\
    --pheno pleiotropy_exercise_phenotypes.txt\
    --pheno-name Trait1\
    --sex\
    --linear\
    --out Trait1
   ```

   For use in several downstream steps, let's create files with only the header and SNP results for each of the univariate analyses:

   ```
   grep 'TEST' Trait1.assoc.linear > Trait1_snp.assoc.linear
   grep 'ADD' Trait1.assoc.linear >> Trait1_snp.assoc.linear
   grep 'TEST' Trait2.assoc.linear > Trait2_snp.assoc.linear
   grep 'ADD' Trait2.assoc.linear >> Trait2_snp.assoc.linear
   ```

b. Try visualizing the data by creating a Hudson plot in R. This will give you some sense of the overlapping signals between the two association analyses.

```
library(hudson)
dat1<-read.table("Trait1_snp.assoc.linear",header=T)
dat2<-read.table("Trait2_snp.assoc.linear",header=T)
names(dat1_snps)<-c("CHR", "SNP", "POS", "A1", "TEST", "NMISS", "BETA",
+"STAT", "pvalue")
names(dat2_snps)<-(names(dat1_snps)
gmirror(top=dat1_snps, bottom=dat2_snps, tline=5e-08, bline=5e-08,

+ toptitle="Trait11", bottomtitle = "Trait2",
+ highlight_p = c(0.00000005,0.00000005), highlighter="green",
+ file = 'pleiotropy_hudson', res = 300, type = 'pdf')
```

c. Now Identify genome-wide significant SNPs ($p<5x10^{-8}$) that overlap for both traits. This can be done using some simple R code:

```
Trait1 <- read.table("Trait1_snp.assoc.linear", header = T)
Trait2 <- read.table("Trait2_snp.assoc.linear", header = T)
SigTrait1 <- subset(Trait1, P<0.00000005)
SigTrait2 <- subset(Trait2, P<0.00000005)
intersect(SigTrait1$SNP, SigTrait2$SNP)
```

d. As you can see, there are some genome-wide significant SNPs that are adjacent or close to each other. To explore whether or not these are independent associations, let's perform some simple LD clumping. You will want to carry through the index SNP identified for each clumped region. You will also want to carry through any SNPs from 1c above that were not part of a clumped region.

```
plink\
--bfile pleiotropy_exercise\
--clump Trait1_snp.assoc.linear,Trait2_snp.assoc.linear\
--clump-kb 250\
--clump-p1 5e-8\
--clump-p2 5e-8\
--clump-r2 0.2\
--clump-replicate\
--clump-verbose\
--out Trait1_Trait2_clump
```

Multivariate analysis

a. Before moving on to dissecting the cross-phenotype associations, let's see if we can include a few additional SNPs/regions to explore by using multivariate analysis. But let's only consider additional regions that are genome-wide suggestive for both phenotypes.

   First run a multivariate analysis on Traits 1 and 2.

```
plink.multivariate\
 --noweb\
 --bfile pleiotropy_exercise\
 --mult-pheno pleiotropy_exercise_phenotypes.txt\
 --sex\
 --mqfam\
 --out Trait1_Trait2
```

   Please note: You should use the --noweb flag due to this program being built on an old version of PLINK.

b. Now let's identify the intersection of SNPs that are genome-wide significant in the multivariate analysis and at least suggestive for each trait in the univariate analysis, i.e. we want to make sure that both traits are contributing to the multivariate signal.

```
Trait1<-read.table("Trait1_snp.assoc.linear", header=T)
Trait2<-read.table("Trait2_snp.assoc.linear", header=T)
multi<-read.table("Trait1_Trait2.mqfam.total", header=T)
sigMulti<-subset(multi, P<0.00000005)
suggTrait1<-subset(Trait1, P<0.000005)
suggTrait2<-subset(Trait2, P<0.000005)
Reduce(intersect, list(suggTrait1$SNP, suggTrait2$SNP, sigMulti$SNP))
```

   Select the additional SNPs that are identified from the intersection of the multivariate analysis and genome-wide suggestive lists for both traits that were not in your original list.

c. You may want to re-run the LD clumping with a suggestive threshold to see if these additional SNPs clump with your existing clumps or are new potential regions to explore.

```
plink\
 --bfile pleiotropy_exercise\
 --clump Trait1_snp.assoc.linear,Trait2_snp.assoc.linear\
```

```
--clump-p1 0.000005\
--clump-p2 0.000005\
--clump-r2 0.2\
--clump-replicate\
--clump-verbose\
--out Trait1_Trait2_clump_suggestive
```

Mediation analyses

a. For each SNPs that you have identified as a cross phenotype association (evidence of overlapping association signals as well as incorporating results from LD clumping and multivariate association) you will need to extract this data from the original plink files and create a genotype file that is coded as 0|1|2 for the genotypes. This can be done in PLINK using the --recodeA command and the --extract command by providing a file with the list of snps. This will give you a .raw genotype file with only the snps that you will 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 that you have selected):

```
library(mediation)
genotypes <- read.table("snps_for_mediation.raw", header=T)
phenotypes<-read.table("pleiotropy_exercise_phenotypes.txt", header=T)
combined<-merge(genotypes,phenotypes)
med.fit<-lm(Trait1~rs125_0, data=combined)
out.fit<-lm(Trait2~Trait1+rs125_0, data=combined)
med.out<-mediate(med.fit,out.fit,treat="rs125_0", mediator="Trait1", boot=TRUE,
+boot.ci.type="bca", sims=1000)
summary(med.out)
```

This will print out a summary of the mediation analysis.

Please note: The more simulations (sims) you specific in the med.out step the more 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 traits?

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 a significant direct effect which is indicative of some level of biological pleiotropy. Do any of the SNPs exhibit complete mediation?

4) Why do some of the SNPs have negative values for the proportion mediated?

Summary table of pleiotropy results

SNP	Beta (Trait 1)	P (Trait 1)	Beta (Trait 2)	P (Trait 2)	MV (P)	MV Loading (Trait 1)	MV Loading (Trait 2)	ADE	ADE (P)	ACME	ACME (P)	Total Effect	Total Effect (P)	Prop Mediated	Prop Mediated (P)