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
```

Pleiotropy Exercise - Answers

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.

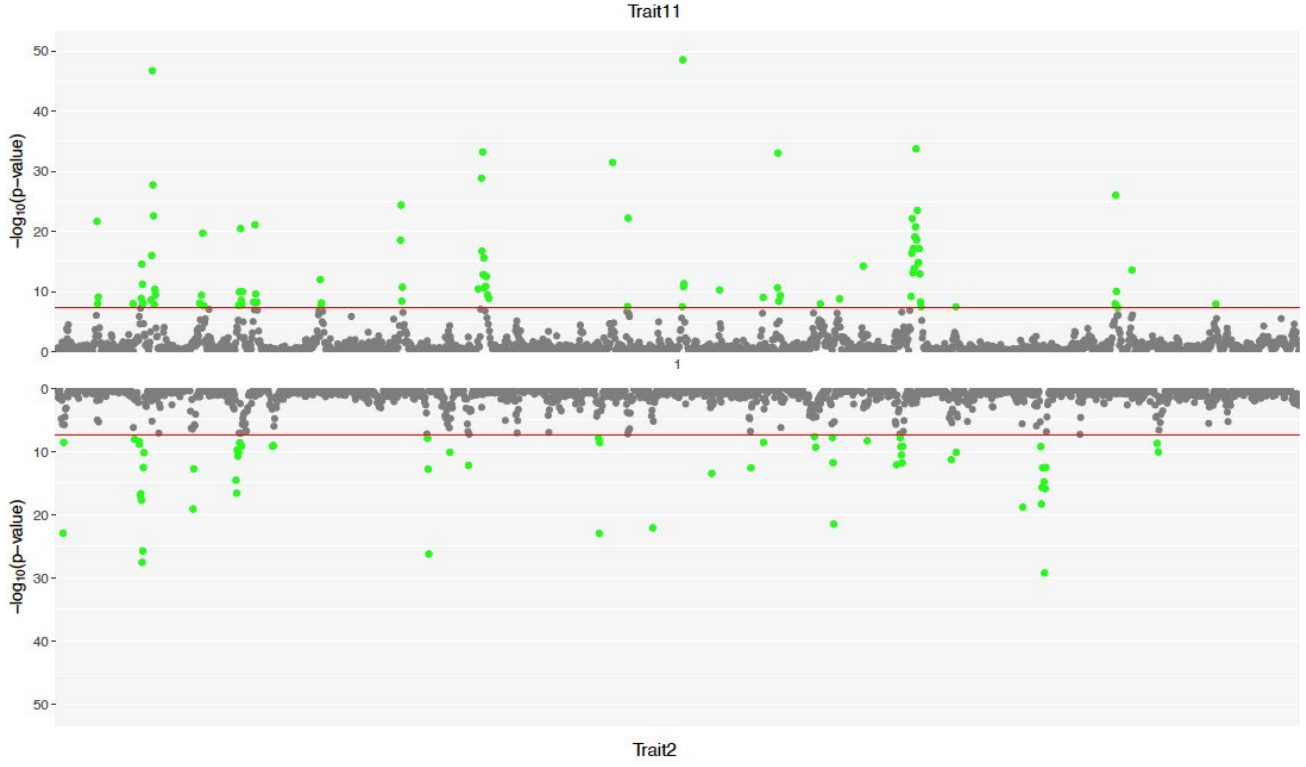
```
plink\
 --bfile pleiotropy_exercise\
 --pheno pleiotropy_exercise_phenotypes.txt\
 --pheno-name Trait1\
 --sex\
 --linear\
 --out Trait1
```

   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.

```
devtools::install_github('anastasia-lucas/hudson')
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)
```

The overlapping genome-wide significant variants are: rs138, rs139, rs140, rs141, rs296, rs299, rs1138, rs1448

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
```

OUTPUT:

```
CHR    F     SNP          BP           P     TOTAL   NSIG    S05    S01
S001  S0001
   1    2   rs139      139000   2.86e-28         9      0      0      0
0      9

                           KB       RSQ  ALLELES    F            P
  (INDEX)   rs139           0     1.000         0    2     2.86e-28

            rs137          -2     0.247     00/00    2     2.17e-17
            rs138          -1     0.399     00/00    1     1.34e-09
            rs138          -1     0.399     00/00    2     1.91e-18
            rs139           0         1     00/00    1     2.77e-15
            rs140           1     0.229     00/00    1     6.05e-12
            rs140           1     0.229     00/00    2     1.85e-26
            rs141           2     0.235     00/00    1      9.9e-09
            rs141           2     0.235     00/00    2     2.98e-13

          RANGE: chr1:137000..141000
           SPAN: 4kb

-------------------------------------------------------------------------


 CHR    F     SNP          BP           P     TOTAL   NSIG    S05    S01
S001  S0001
   1    1   rs296      296000   1.15e-10         5      0      0      0
0      5

                           KB       RSQ  ALLELES    F            P
  (INDEX)   rs296           0     1.000         0    1     1.15e-10

            rs295          -1     0.429     00/00    1     2.01e-08
            rs296           0         1     00/00    2      2.6e-09
            rs299           3     0.267     00/00    1     2.77e-09
            rs299           3     0.267     00/00    2     7.29e-10

          RANGE: chr1:295000..299000
           SPAN: 4kb
```

From clump 1, let's choose rs139 (and not rs138, rs140 and rs141) and from clump 2 let's choose rs296 (and not rs295 and rs299) to carry forward to our mediation analysis. We will also carry forward rs1138 and rs1448 since these two SNPs are not part of any other clumps but are genome-wide significant for both traits.

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.

   We identify the following overlapping SNPs: rs125, rs135, rs137, rs138, rs139, rs140, rs141, rs295, rs296, rs298, rs299, rs300, rs920, rs921, rs923, rs1138, rs1166, rs1361, rs1448. Of course, this list includes the original set of 8 variants that were genome-wide significant for both Traits 1 and 2.

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
```

OUTPUT:

```
 CHR    F     SNP         BP           P    TOTAL   NSIG    S05    S01
S001  S0001
   1    2   rs139     139000   2.86e-28        9      0      0      0
0      9

                           KB      RSQ  ALLELES    F           P
  (INDEX)   rs139           0    1.000        0    2     2.86e-28

            rs137          -2    0.247    00/00    1     6.05e-08
            rs137          -2    0.247    00/00    2     2.17e-17
            rs138          -1    0.399    00/00    1     1.34e-09
            rs138          -1    0.399    00/00    2     1.91e-18
            rs139           0        1    00/00    1     2.77e-15
            rs140           1    0.229    00/00    1     6.05e-12
            rs140           1    0.229    00/00    2     1.85e-26
            rs141           2    0.235    00/00    1      9.9e-09
            rs141           2    0.235    00/00    2     2.98e-13

          RANGE: chr1:137000..141000
           SPAN: 4kb

-------------------------------------------------------------------


 CHR    F     SNP         BP           P    TOTAL   NSIG    S05    S01
S001  S0001
   1    1   rs921     921000   6.29e-23        5      0      0      0
1      4

                           KB      RSQ  ALLELES    F           P
  (INDEX)   rs921           0    1.000        0    1     6.29e-23

            rs920          -1    0.224    00/00    1     3.11e-08
            rs920          -1    0.224    00/00    2     6.25e-08
            rs921           0        1    00/00    2     1.94e-07
            rs922           1    0.202    00/00    1     4.52e-07

          RANGE: chr1:920000..922000
           SPAN: 2kb

-------------------------------------------------------------------


 CHR    F     SNP         BP           P    TOTAL   NSIG    S05    S01
S001  S0001
   1    2   rs136     136000    1.3e-17        5      0      1      0
0      4

                           KB      RSQ  ALLELES    F           P
  (INDEX)   rs136           0    1.000        0    2      1.3e-17

            rs134          -2    0.229    00/00    2     3.97e-09
            rs135          -1    0.379    00/00    1     1.41e-06
```

```
            rs135         -1    0.379    00/00    2     1.47e-09

          RANGE: chr1:134000..136000
           SPAN: 2kb

-----------------------------------------------------------------


 CHR    F     SNP         BP          P    TOTAL   NSIG    S05    S01
S001  S0001
   1    2  rs1361    1361000   1.68e-12        9      0      1      2
1      5

                           KB      RSQ  ALLELES    F           P
  (INDEX)  rs1361           0    1.000        0    2     1.68e-12

           rs1359         -2    0.238    00/00    2     5.98e-10
           rs1360         -1    0.281    00/00    2      2.8e-11
           rs1361          0        1    00/00    1     2.65e-07
           rs1362          1    0.271    00/00    2     6.54e-10
           rs1363          2    0.204    00/00    2     1.64e-07

          RANGE: chr1:1359000..1363000
           SPAN: 4kb

-----------------------------------------------------------------


 CHR    F     SNP         BP          P    TOTAL   NSIG    S05    S01
S001  S0001
   1    1   rs296     296000   1.15e-10        5      0      0      0
0      5

                           KB      RSQ  ALLELES    F           P
  (INDEX)   rs296           0    1.000        0    1     1.15e-10

           rs295         -1    0.429    00/00    1     2.01e-08
           rs295         -1    0.429    00/00    2     8.62e-08
           rs296          0        1    00/00    2      2.6e-09
           rs299          3    0.267    00/00    1     2.77e-09
           rs299          3    0.267    00/00    2     7.29e-10

          RANGE: chr1:295000..299000
           SPAN: 4kb

-----------------------------------------------------------------


 CHR    F     SNP         BP          P    TOTAL   NSIG    S05    S01
S001  S0001
   1    1  rs1138    1138000   9.58e-10        3      0      1      0
0      2

                           KB      RSQ  ALLELES    F           P
  (INDEX)  rs1138           0    1.000        0    1     9.58e-10
```

```
rs1137          -1    0.315    00/00    1      4.09e-07
rs1138           0        1    00/00    2       2.9e-09

RANGE: chr1:1137000..1138000
 SPAN: 1kb
```

Based on the multivariate analysis and additional clumping, you should add the following SNPs to your list of SNPs for mediation: rs125, rs135, rs300, rs921, rs923, rs1166, rs1361.

The final list of SNPs that were selected to carry through to the mediation analysis are:

rs125, rs135, rs139, rs296, rs300, rs921, rs923, rs1138, rs1166, rs1361, rs1448

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?

   rs138, rs139, rs140, rs141, rs296, rs299, rs1138, rs1448

2) Did the multivariate analyses result in additional SNPs that had genome-wide significant cross phenotype associations but that also had genome-wide suggestive (p<5x10-6) univariate association for each trait? Which SNP(s)?

   rs125, rs135, rs137, rs295, rs298, rs300, rs920, rs921, rs923, rs1166, rs1361

   Instead of running mediation analysis on all 19 SNPs, I suggested that you perform LD clumping to reduce this number of SNPs and only focus on the index SNP for each clump (or if the index SNP was not associated with both traits to choose another SNP from among the clumped SNPs). This reduced the set of SNPs to 11.

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?

   All SNPs show a significant direct effect on Trait 2 indicating some level of biological pleiotropy. rs923 has an ADE p-value of 0.002 but this is still less than the Bonferroni corrected p-value of 0.0045, adjusting for the 11 SNPs. No SNP shows an association with Trait 2 that is completely mediated through its association with Trait 1, i.e. an ACME estimate that is equal to (or close to) the total effect. The strongest mediated effect is for rs921 in which the mediated effect accounts for ~40% of the total effect of the SNP on Trait 2.

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

   The estimate of the proportion mediated is not the best way to interpret the mediation results, despite its seemingly obvious interpretability. In reality this proportion does not range from 0-1 but can rather be less than 0 and greater than 1. The negative proportion mediated values that we see for many of the SNPs we have analyzed is due to the fact that these SNPs have an effect estimate for the total effect and mediated effect that are in opposite directions, i.e. the effect of the SNPs on Trait 1 and Trait 2 is in opposite directions. Depending on your study question, you may want to limit your selection of SNPs to only those with effects on the two traits in the same direction. We did not see this among our SNPs, but a proportion mediated > 1 can happen when the strength of the association with the mediator (Trait 1) is much higher than the strength of the association with the outcome (Trait 2). This is often why it is recommended that the direct, indirect and total effects be used in the interpretation rather than the proportion mediated.

Summary table of pleiotropy results

SNP	Beta (Trait 1)	P (Trait 1)	Beta (Trait 2)	P (Trait 2)	MV[1] P	MV[1] Loading (Trait 1)	MV[1] Loading (Trait 2)	ADE	ADE (P)	ACME	ACME (P)	Total Effect	Total Effect (P)	Prop Mediated	Prop Mediated (P)
rs125	0.072	1.08E-08	0.062	6.45E-07	1.80E-10	0.8397	0.7096	0.045	<2e-16	0.015	<2e-16	0.0596	<2e-16	0.2516	<2e-16
rs135	-0.040	1.41E-06	0.050	1.47E-09	2.82E-17	-0.5457	0.7024	0.059	<2e-16	-0.008	<2e-16	0.0509	<2e-16	-0.3317	<2e-16
rs139	-0.065	2.77E-15	0.090	2.86E-28	2.26E-50	-0.5249	0.7196	0.103	<2e-16	-0.014	<2e-16	0.0891	<2e-16	-0.1580	<2e-16
rs296	-0.056	1.15E-10	-0.051	2.60E-09	1.78E-14	0.8100	0.7456	-0.039	<2e-16	-0.012	<2e-16	-0.0512	<2e-16	0.2306	<2e-16
rs300	-0.046	1.85E-08	-0.039	2.09E-06	1.34E-10	-0.8386	-0.7110	-0.029	<2e-16	-0.010	<2e-16	-0.0392	<2e-16	0.2507	<2e-16
rs921	0.109	6.29E-23	0.057	1.95E-07	1.16E-24	-0.9475	-0.5144	0.036	<2e-16	0.023	<2e-16	0.0595	<2e-16	0.3908	<2e-16
rs923	0.041	1.48E-06	0.042	4.04E-07	2.35E-09	0.7615	0.7957	0.034	0.002	0.009	<2e-16	0.0421	<2e-16	0.2035	<2e-16
rs1138	-0.050	9.58E-10	0.048	2.90E-09	2.44E-20	0.6511	-0.6027	0.058	<2e-16	-0.011	<2e-16	0.0468	<2e-16	-0.2319	<2e-16
rs1166	-0.051	4.77E-10	0.041	6.30E-07	1.02E-17	-0.7175	0.5275	0.049	<2e-16	-0.011	<2e-16	0.0382	<2e-16	-0.2918	<2e-16
rs1361	-0.056	2.65E-07	0.076	1.68E-12	3.52E-21	-0.5349	0.7114	0.087	<2e-16	-0.012	<2e-16	0.0751	<2e-16	-0.1614	<2e-16
rs1448	-0.079	3.46E-08	0.092	1.51E-11	2.21E-20	-0.5847	0.6679	0.108	<2e-16	-0.017	<2e-16	0.0908	<2e-16	-0.1879	<2e-16

[1]Multivariate