"subfraction analysis : prepare data for sharing"

Heejung Shim

August 23, 2015

Contents

1 Plan

- Effect size estimates and standard error for every pairs of 'one' SNP and 'one' quantile transformed phenotype.
  - Use software GEMMA with -lm optiion.
- (identified) genotypes and corresponding phenotypes for just the handful of SNPs we highlight in the manuscript.
  - get normalized phenotypes
  - extract particular genotypes

2 Prepare directory to save data

I create directory to save data:

```
cd /mnt/lustre/home/shim/subfraction/
mkdir data_shared
cd data_shared
mkdir normalized_phenotypes
mkdir genotypes
mkdir summary_statistics
```

3 Get normalized phenotypes

I modified a script here:

```
cd /mnt/lustre/home/shim/subfraction/revision/code/
vi revise_correlation_final_figure.R
```

Normalized averages of pre-treatment measurements and post-treatment measurements are saved in:

```
vi /mnt/lustre/home/shim/subfraction/data_shared/normalized_phenotypes/average.
   txt
```

and normalized differences between post-treatment and pre-treatment measures are saved in:

```
vi /mnt/lustre/home/shim/subfraction/data_shared/normalized_phenotypes/
   difference.txt
```

3.1 Script

```
path.out = "/mnt/lustre/home/shim/subfraction/data_shared/normalized_phenotypes
   /"

PhenoNames_o = c("A", "Ld", "i1", "i2", "i3", "l1", "l2a", "l2b", "l3a", "l3b",
     "l4a", "l4b", "H", "L", "Tg", "T")
new_IX = c(14,2,3:12, 15, 13, 16, 1)
PhenoNames = PhenoNames_o[new_IX]
```

```
PhenoNames
# [1] "L" "Ld" "i1" "i2" "i3" "l1" "l2a" "l2b" "l3a" "l3b" "l4a" "l4b"
#[13] "Tg" "H" "T" "A"

## Average
dat15 = read.table("/mnt/lustre/home/shim/subfraction/data/final_pheno_cor_v2/
    phenotype_ready/Sum_II_1_a", as.is = TRUE)
dat1 = read.table("/mnt/lustre/home/shim/subfraction/data/final_pheno_cor_v2/
    phenotype_ready/TC_sum", as.is = TRUE)
numIND = dim(dat15)[1]
data = matrix(data=NA, nc=16, nr = numIND)
data[,1:15] = as.matrix(dat15)
data[,16] = as.matrix(dat1)
new_data = data[,new_IX]

write.table(new_data, file = paste0(path.out, "average.txt"), quote=FALSE, col.
    names = PhenoNames, row.names=FALSE)

## Difference
dat15 = read.table("/mnt/lustre/home/shim/subfraction/data/final_pheno_cor_v2/
    phenotype_ready/Diff_II_1_a", as.is = TRUE)
dat1 = read.table("/mnt/lustre/home/shim/subfraction/data/final_pheno_cor_v2/
    phenotype_ready/TC_diff", as.is = TRUE)
numIND = dim(dat15)[1]
data = matrix(data=NA, nc=16, nr = numIND)
data[,1:15] = as.matrix(dat15)
data[,16] = as.matrix(dat1)
new_data = data[,new_IX]

write.table(new_data, file = paste0(path.out, "difference.txt"), quote=FALSE,
    col.names = PhenoNames, row.names=FALSE)
```

4 Get effect size estimates and standard error

SNPs included in our analysis is only goodSNPs (7,836,525). Genotype files are in:

```
cd /mnt/lustre/home/shim/subfraction/data/allSNPs/goodSNPs/
```

Create directory to run scripts:

```
cd /mnt/lustre/home/shim/subfraction/data_shared/
mkdir runGEMMA
mkdir runGEMMA/pheno.dat
mkdir runGEMMA/com
mkdir runGEMMA/err
```

Prepare phenotypes:

```
## average
path.input = "/mnt/lustre/home/shim/subfraction/data_shared/normalized_
    phenotypes/average.txt"
pheno.dat = read.table(path.input, header = TRUE)
pheno.name = names(pheno.dat)

path.output = "/mnt/lustre/home/shim/subfraction/data_shared/runGEMMA/pheno.dat
    /average."
for(i in 1:length(pheno.name)){
  write.table(pheno.dat[,i], file = paste0(path.output, pheno.name[i]), quote=
      FALSE, row.names = FALSE, col.names = FALSE)
}

## difference
path.input = "/mnt/lustre/home/shim/subfraction/data_shared/normalized_
    phenotypes/difference.txt"
pheno.dat = read.table(path.input, header = TRUE)
pheno.name = names(pheno.dat)

path.output = "/mnt/lustre/home/shim/subfraction/data_shared/runGEMMA/pheno.dat
    /difference."
for(i in 1:length(pheno.name)){
  write.table(pheno.dat[,i], file = paste0(path.output, pheno.name[i]), quote=
      FALSE, row.names = FALSE, col.names = FALSE)
}
```

Prepare submission files:

```
get.com.run.GEMMA <- function(phenoT, phenoN){
output.dir.name = paste0(phenoT, ".", phenoN)
com.path = "/mnt/lustre/home/shim/subfraction/data_shared/runGEMMA/"
```

```
file.name = paste0(com.path, "com/", output.dir.name, ".sh")

com = "#!/bin/bash"
cat(com, file = file.name)
cat("\n", file = file.name, append = TRUE)

com = "#$␣-t␣1-22"
cat(com, file = file.name, append = TRUE)
cat("\n", file = file.name, append = TRUE)

com = paste("#$␣-o␣", com.path, "err/out.$JOB_ID", sep="")
cat(com, file = file.name, append = TRUE)
cat("\n", file = file.name, append = TRUE)

com = paste("#$␣-e␣", com.path, "err/err.$JOB_ID", sep="")
cat(com, file = file.name, append = TRUE)
cat("\n", file = file.name, append = TRUE)

com = paste0("cd␣", com.path)
cat(com, file = file.name, append = TRUE)
cat("\n", file = file.name, append = TRUE)

geno.path = paste0("/mnt/lustre/home/shim/subfraction/data/allSNPs/goodSNPs/sub
    _chr$SGE_TASK_ID.geno")
pheno.path = paste0(com.path, "pheno.dat/", phenoT, ".", phenoN)
com = paste0("~shim/bin/gemma␣-g␣", geno.path, "␣-p␣", pheno.path, "␣-maf␣0␣-r2
    ␣1␣-miss␣1␣-lm␣-o␣", output.dir.name, ".$SGE_TASK_ID")
cat(com, file = file.name, append = TRUE)
cat("\n", file = file.name, append = TRUE)

}

path.input = "/mnt/lustre/home/shim/subfraction/data_shared/normalized_
    phenotypes/average.txt"
pheno.name = names(read.table(path.input, header = TRUE))

phenoT = "average"
for(pp in 1:length(pheno.name)){
  phenoN = pheno.name[pp]
  get.com.run.GEMMA(phenoT, phenoN)
```

```
}
phenoT = "difference"
for(pp in 1:length(pheno.name)){
  phenoN = pheno.name[pp]
  get.com.run.GEMMA(phenoT, phenoN)
}
```

I submitted jobs here:

```
cd /mnt/lustre/home/shim/subfraction/data_shared/runGEMMA/com/
for file in *.sh ; do echo $file; done
for file in *.sh ; do qsub -l h_vmem=5g -V $file; done
```

Collect outputs and put them into one file:

```
path.output = "/mnt/lustre/home/shim/subfraction/data_shared/summary_statistics
    /"
path.input = "/mnt/lustre/home/shim/subfraction/data_shared/runGEMMA/output/"

path.temp = "/mnt/lustre/home/shim/subfraction/data_shared/normalized_
    phenotypes/average.txt"
pheno.name = names(read.table(path.temp, header = TRUE))

## Make names for output
name.list = rep(NA, length(pheno.name)*2)
ix = 1
for(pp in 1:length(pheno.name)){
  name.list[ix:(ix+1)] = paste0(pheno.name[pp], c(".beta", ".se"))
  ix = ix + 2
}

phenoT = "average"
for(chr in 1:22){
  pp = 1
  phenoN = pheno.name[pp]
  path.file = paste0(path.input, phenoT, ".", phenoN, ".", chr, ".assoc.txt")
  dat = read.table(path.file, header = T, as.is = TRUE)
  dat.info = data.frame(rs = dat$rs, allele1 = dat$allele1, allele0 = dat$
      allele0, af = dat$af)

  path.each.output = paste0(path.output, phenoT, ".", "chr", chr, ".summary.txt
      ")

  res.dat = matrix(data=NA, nr = dim(dat)[1], nc = length(pheno.name)*2)
  for(pp in 1:length(pheno.name)){
    phenoN = pheno.name[pp]
    path.file = paste0(path.input, phenoT, ".", phenoN, ".", chr, ".assoc.txt")
    dat = read.table(path.file, header = T, as.is = TRUE)
    res.dat[,((2*pp -1):(2*pp))] = as.matrix(dat[,9:10])
  }

  colnames(res.dat) = name.list
  final.out = cbind(dat.info, res.dat)

  write.table(final.out, file = path.each.output, quote=FALSE, row.names =
      FALSE, col.names = TRUE)
}


phenoT = "difference"
for(chr in 20:22){
  pp = 1
  phenoN = pheno.name[pp]
  path.file = paste0(path.input, phenoT, ".", phenoN, ".", chr, ".assoc.txt")
  dat = read.table(path.file, header = T, as.is = TRUE)
  dat.info = data.frame(rs = dat$rs, allele1 = dat$allele1, allele0 = dat$
      allele0, af = dat$af)

  path.each.output = paste0(path.output, phenoT, ".", "chr", chr, ".summary.txt
      ")

  res.dat = matrix(data=NA, nr = dim(dat)[1], nc = length(pheno.name)*2)
  for(pp in 1:length(pheno.name)){
    phenoN = pheno.name[pp]
    path.file = paste0(path.input, phenoT, ".", phenoN, ".", chr, ".assoc.txt")
    dat = read.table(path.file, header = T, as.is = TRUE)
    res.dat[,((2*pp -1):(2*pp))] = as.matrix(dat[,9:10])
  }

  colnames(res.dat) = name.list
  final.out = cbind(dat.info, res.dat)

  write.table(final.out, file = path.each.output, quote=FALSE, row.names =
      FALSE, col.names = TRUE)
}
```

Now summary statistics (effect size and standard error for each SNP and phenotype pair) are saved in

```
cd /mnt/lustre/home/shim/subfraction/summary_statistics/
```

Each file contains SNPname, two alleles, MAF, effect size and standard error for each phenotype.

5 Get genotypes for SNPs reported in our paper

These are SNPs reported in GWAS either irrespective of statin exposure or of statin response.

- chr1: rs7528419
- chr19: rs7412, rs157581
- chr6: rs55730499, 6-161069320, rs10455872
- chr16: rs247616, rs11076175

First, let's extract genotype information using these commands:

```
cd ~/subfraction/data/allSNPs/goodSNPs/
grep "rs7528419" sub_chr1.geno > /mnt/lustre/home/shim/subfraction/data_shared/
    genotypes/rs7528419.geno
grep "rs7412" sub_chr19.geno > /mnt/lustre/home/shim/subfraction/data_shared/
    genotypes/rs7412.geno
grep "rs157581" sub_chr19.geno > /mnt/lustre/home/shim/subfraction/data_shared/
    genotypes/rs157581.geno
grep "rs55730499" sub_chr6.geno > /mnt/lustre/home/shim/subfraction/data_shared
    /genotypes/rs55730499.geno
grep "6-161069320" sub_chr6.geno > /mnt/lustre/home/shim/subfraction/data_
    shared/genotypes/6-161069320.geno
grep "rs10455872" sub_chr6.geno > /mnt/lustre/home/shim/subfraction/data_shared
    /genotypes/rs10455872.geno
grep "rs247616" sub_chr16.geno > /mnt/lustre/home/shim/subfraction/data_shared/
    genotypes/rs247616.geno
```

```
grep "rs11076175" sub_chr16.geno > /mnt/lustre/home/shim/subfraction/data_
    shared/genotypes/rs11076175.geno
```

Make mean genotype files:

```
genotype_name = c("rs7528419", "rs7412", "rs157581", "rs55730499", "6-161069320
    ", "rs10455872", "rs247616", "rs11076175")
path = "/mnt/lustre/home/shim/subfraction/data_shared/genotypes/"

num = length(genotype_name)

for(i in 1:num){

  geno_input = paste(genotype_name[i], ".geno", sep="")
  path_each = paste(path, geno_input, sep="")

  genoD1 = read.table(path_each, as.is = TRUE)
  numIND = 1868

  genoM = matrix(data=NA, nc = (numIND + 3), nr = 1)
  genoM[1,1:3] = as.matrix(genoD1[1,1:3])
  IX2 = (1:numIND)*2 +3
  IX1 = (1:numIND)*2 +3 - 1

  genoM[1,4:(numIND+3)] = as.matrix(2*genoD1[1,IX1] + genoD1[1,IX2])

  geno_out = paste(genotype_name[i], ".meangeno",sep="")
  path_each = paste(path, geno_out, sep="")

  write.table(genoM, path_each, quote= FALSE, row.names = FALSE, col.names =
     FALSE)

}
```