eQTL Analysis of SDCD Genes

J. Fah Sathirapongsasuti

April 12, 2014

1 Introduction

In the vignette `lgrc_sdcd_expression` we identify 959 genes with sexually-dimorphic differential expression in the presence of COPD ("sexually dimorphic and COPD differential" or "SDCD" genes). We used genotyping information (available from dbGaP accession number: phs000624.v1.p1) to perform eQTL analysis. Here we take the eQTL results and identify eQTLs that suggest sex-specific regulation by contrasting the coefficients from the linear models.

```
> library(COPDSexualDimorphism)
> `%+%` <- function(x,y) paste(x,y,sep="")
```

2 eQTL on PLINK

There are a number of ways to do eQTL analysis. Here we chose to use PLINK, the process of which is not described here. This vignette starts at the output of PLINK commands:

```
plink -bfile lgrc_eqtl_qc -pheno lgrc_expr.txt -all-pheno -linear -covar lgrc_eQTL_covar.txt
-filter-males
```

and

```
plink -bfile lgrc_eqtl_qc -pheno lgrc_expr.txt -all-pheno -linear -covar lgrc_eQTL_covar.txt
-filter-females
```

Because of the expensive computational burden, we need to parallel process the eQTL fitting by splitting the gene expression profile into multiple files and submit jobs to an LSF cluster.

In order to do further analysis, we selected only eQTLs that associate SNPs within 100kb up and 10kb downstream of the transcription start sites of SDCD genes and gather all of the PLINK results in one file. The combined results is read into an R `data.frame` named `eqtl`, which can be loaded by `data(lgrc.eqtl)`.

3 Results

3.1 Multiple hypothesis testing adjustment

We first adjust for multiple hypothesis testing by Benjamini-Hochberg FDR.

```
> data(lgrc.eqtl)
> dim(eqtl)

[1] 29307    23

> print("There are " %+% length(unique(eqtl$SNP)) %+% " cis SNPs of SDCD genes.")

[1] "There are 28301 cis SNPs of SDCD genes."

> fdr.cutoff = 0.05
> eqtl$FDR_male = p.adjust(eqtl$P_male, "BH")
> eqtl$FDR_female = p.adjust(eqtl$P_female, "BH")
> print(sum(eqtl$FDR_male < fdr.cutoff, na.rm=T) %+% " male, " %+% sum(eqtl$FDR_female < fdr.cutoff, na.
```

```
[1] "500 male, 190 female, 127 both."
```

```
> fisher.test(eqtl$FDR_male < fdr.cutoff, eqtl$FDR_female < fdr.cutoff)

        Fisher's Exact Test for Count Data

data:  eqtl$FDR_male < fdr.cutoff and eqtl$FDR_female < fdr.cutoff
p-value < 2.2e-16
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 111.5483 215.8018
sample estimates:
odds ratio
  154.9136
```

Sometimes reference allele for male and female are different, leading to opposite signs of the regression coefficients.

```
> discord.ref.allele = which(eqtl$A1_male != eqtl$A1_female)
> eqtl$STAT_female[discord.ref.allele] = -eqtl$STAT_female[discord.ref.allele]
> eqtl$BETA_female[discord.ref.allele] = -eqtl$BETA_female[discord.ref.allele]
```

Now we are ready to identify sexually dimorphic eQTL, using the function `sdcd.core`.

```
> # package the info as limma fit object to pass to sdcd.core
> eqtl.male = list(
+   coefficients = data.frame(copd=eqtl$BETA_male),
+   stdev.unscaled = data.frame(copd=eqtl$BETA_male/eqtl$STAT_male),
+   sigma = 1,
+   df.residual = eqtl$NMISS_male - 4,
+   df.prior = eqtl$NMISS_male - 4
+   )
> eqtl.female = list(
+   coefficients = data.frame(copd=eqtl$BETA_female),
+   stdev.unscaled = data.frame(copd=eqtl$BETA_female/eqtl$STAT_female),
+   sigma = 1,
+   df.residual = eqtl$NMISS_female - 4,
+   df.prior = eqtl$NMISS_female - 4
+   )
> # The SDCD analysis
> eqtl.sdcd = sdcd.core(eqtl.male, eqtl.female, "copd")
> eqtl = cbind(eqtl, eqtl.sdcd)
> all.eqtl = eqtl
> eqtl = subset(eqtl, beta.diff.pval.adj < fdr.cutoff & !is.na(beta.diff.pval.adj))
> print("Male-female difference: " %+% nrow(eqtl) %+% " eQTL are significant at level " %+% fdr.cutoff %
```

```
[1] "Male-female difference: 860 eQTL are significant at level 0.05, covering 209 genes."
```

Note here that in the paper by Sathirapongsasuti et al., the eQTL results were further filtered for SNPs with more than five samples with homozygous recessive alleles. We cannot demonstrate that here as the SNPs data cannot be distributed through the R package.

4 Session Information

```
> sessionInfo()
```

```
R version 3.1.0 (2014-04-10)
Platform: x86_64-unknown-linux-gnu (64-bit)

locale:
 [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
 [3] LC_TIME=en_US.UTF-8        LC_COLLATE=C
 [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
 [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
 [9] LC_ADDRESS=C               LC_TELEPHONE=C
[11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C

attached base packages:
[1] parallel  stats     graphics  grDevices utils     datasets  methods
[8] base

other attached packages:
 [1] COPDSexualDimorphism_1.0.0       gtools_3.3.1
 [3] gplots_2.13.0                    GenomicRanges_1.16.0
 [5] GenomeInfoDb_1.0.0               IRanges_1.21.45
 [7] BiocGenerics_0.10.0              limma_3.20.0
 [9] beeswarm_0.1.6                   RColorBrewer_1.0-5
[11] NCBI2R_1.4.5                     COPDSexualDimorphism.data_0.99.0

loaded via a namespace (and not attached):
[1] KernSmooth_2.23-12 XVector_0.4.0      bitops_1.0-6       caTools_1.16
[5] gdata_2.13.3       stats4_3.1.0       tools_3.1.0
```