Differential Methylation Analysis of LGRC Data

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). Here we focus on methylated regions in the promoter regions of SDCD genes. Using methylation profile also from Lung Genomic Research Consortium (LGRC), we identify regions with sexually dimorphic differential methylation.

2 Preprocessing

We load in all the necessary packages and format the data.

```
> library(COPDSexualDimorphism)
> `%+%` <- function(x,y) paste(x,y,sep="")
> p.cutoff = 0.01
> data(lgrc.methp)
> data(lgrc.meta)
> data(lgrc.sdcd.genes)
> sampleID = names(methp)[grepl("^LT",names(methp),perl=TRUE) & (names(methp) %in% meta$tissueid)]
> only.methp = as.matrix(methp[,sampleID])
> row.names(only.methp) = methp$name
> colnames(only.methp) = sampleID
```

3 VMR for SDCD genes

We annotate the VMRs by linking them to genes within 10kb, using functions in the package GenomicRanges.

```
> sdcd.genes = subset(sdcd.genes, chromosome_name != "HSCHR6_MHC_QBL")
> sdcd.genes.bed = GRanges(seqnames=Rle("chr" %+% sdcd.genes$chromosome_name),
+                          ranges=IRanges(start=sdcd.genes$start_position,
+                                         end=sdcd.genes$end_position,
+                                         names=sdcd.genes$ensembl_gene_id),
+                          strand=Rle(strand(sdcd.genes$strand)))
> methp = subset(methp, !(chr %in% c("chrX", "chrY")))
> methp.bed = GRanges(seqnames=Rle(methp$chr),
+                     ranges=IRanges(start=methp$start,
+                                    end=methp$end,
+                                    names=methp$name))
> window = 2e4
> sum(countOverlaps(sdcd.genes.bed, resize(methp.bed, window, fix="center")) != 0) # 397 SDCD genes have

[1] 395

> sum(countOverlaps(sdcd.genes.bed, resize(methp.bed, 2e6)) != 0)
```

```
[1] 931

> sum(countOverlaps(resize(methp.bed, window, fix="center"), sdcd.genes.bed) != 0) # 892 VMR have SDCD g

[1] 888

> sdcd.genes.vmr.bed = subsetByOverlaps(resize(methp.bed, window, fix="center"), sdcd.genes.bed)
> # 892 VMRs have SDCD genes
>
> sdcd.genes.vmr = names(sdcd.genes.vmr.bed)
```

4 Sexually Dimorphic Differential Methylation Analysis

We first stratify the data by COPD status, fit linear models, and contrast the coefficients.

```
> design = cbind(ctrl=1,
+                   gender=as.integer(meta[sampleID,"GENDER"] == "1-Male"),
+                 age=meta[sampleID,"age"],
+                 pkyr=meta[sampleID,"pkyrs"])
> good.idx = apply(design,1,function(x){!any(is.na(x))}) & meta[sampleID,"diagmaj"] == "2-COPD/Emphysema
> copd.fit = lmFit(logit(only.methp)[sdcd.genes.vmr,good.idx], design[good.idx,])
> copd.fit = eBayes(copd.fit)
> good.idx = apply(design,1,function(x){!any(is.na(x))}) & meta[sampleID,"diagmaj"] == "3-Control"
> ctrl.fit = lmFit(logit(only.methp)[sdcd.genes.vmr,good.idx], design[good.idx,])
> ctrl.fit = eBayes(ctrl.fit)
```

And here is the SDCD analysis on the methylation data. We have a specialize function `sdcd.vmr` to help annotate the results with SDCD genes.

```
> copd.ctrl.gender.beta.diff.genes = sdcd.vmr(copd.fit, ctrl.fit, "gender", sdcd.genes, annotate=TRUE, a

[1] "Number of probes with sexual dimorphic VMR: 387"

> copd.ctrl.gender.beta.diff.vmr = copd.ctrl.gender.beta.diff.genes$vmr
```

5 Boxplots for the VMRs

We now plot the percent methylation for each of the sexually dimorphic VMRs.

```
> vmr.sdcd.gene = sapply(as.character(copd.ctrl.gender.beta.diff.genes$genesymbol), function(g) {
+                 this.vmr.genes = unlist(strsplit(g,","))
+                 this.vmr.sdcd = this.vmr.genes[which(this.vmr.genes %in% sdcd.genes$hgnc_symbol)]
+                 if (length(this.vmr.sdcd) > 1) {
+                         print(g %+% " has more than one SDCD")
+                         better.sdcd = as.character(sdcd.genes[sdcd.genes$hgnc_symbol %in% this.vmr.sdc
+                         print("Keeping " %+% better.sdcd %+% " because of copd.ctrl.p.adj")
+                         this.vmr.sdcd = better.sdcd
+                 }
+                 if (length(this.vmr.sdcd) == 0) this.vmr.sdcd = NA
+                 return(this.vmr.sdcd)
+         } )
> interesting.vmrs = copd.ctrl.gender.beta.diff.genes$vmr[vmr.sdcd.gene %in% sdcd.genes$hgnc_symbol]
> interesting.vmrs.genes = vmr.sdcd.gene[vmr.sdcd.gene %in% sdcd.genes$hgnc_symbol]
> names(interesting.vmrs.genes) = interesting.vmrs
> copd.bool = meta[sampleID,"diagmaj"] == "2-COPD/Emphysema"
```

```
> male.bool = meta[sampleID,"GENDER"] == "1-Male"
> for (ivmr in interesting.vmrs) {
+         this.gene = interesting.vmrs.genes[ivmr]
+         do.sdcd.boxplot(ivmr, only.methp, copd.bool, male.bool, symbol=this.gene, filename=this.gene %
+ }
```

6 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
```