

# Differential Gene Expression Analysis of LGRC Data

J Fah Sathirapongsasuti

October 13, 2014

## 1 Introduction

Chronic Obstructive Pulmonary Disease (COPD) is the third leading cause of death in the United States. Since the year 2000 the number of females dying from COPD has surpassed the number of males, and there is an increasing body of research suggesting females may be biologically more susceptible to COPD. The goal of the study is to explore which molecular pathways might be associated with sexual dimorphism in COPD. This vignette uses gene expression data from the Lung Genomics Research Consortium to identify 959 genes with sexually-dimorphic differential expression in the presence of COPD ("sexually dimorphic and COPD differential" or "SDCD" genes).

## 2 Preprocessing

Load the necessary packages and datasets.

```
> library(COPDSexualDimorphism)
> `+%` <- function(x,y) paste(x,y,sep="")
> p.cutoff = 0.01
> data(lgrc.expr)
> data(lgrc.expr.meta)
> data(lgrc.genes)
```

## 3 Sexually Dimorphic and COPD Differential Gene Expression Analysis

Sexually Dimorphic and COPD Differential (SDCD) analysis comprises of two stratifications: by sex and by COPD status. These stratified analyses are multivariate linear model performed by `limma`. In each of the analyses, the function `sdcd` contrasts the linear models from the two treatments and outputs a list of genes with SDCD expression. The results from the two stratification analyses are combined at the end.

### 3.1 Model 1: expression = COPD + Age + pkyrs

Stratified by sex, then compare the betas.

```
> design.mtx = cbind(ctrl=1,
+                   copd=as.integer(grepl("COPD", colnames(expr))),
+                   age=expr.meta$age,
+                   pkyr=expr.meta$pkyrs)
> good.idx = apply(design.mtx, 1, function(x){!any(is.na(x))}) & (expr.meta$gender == "1-Male")
> male.fit = lmFit(log(expr)[,good.idx], design.mtx[good.idx,])
> male.fit = eBayes(male.fit)
> good.idx = apply(design.mtx, 1, function(x){!any(is.na(x))}) & (expr.meta$gender == "2-Female")
> female.fit = lmFit(log(expr)[,good.idx], design.mtx[good.idx,])
```

```

> female.fit = eBayes(female.fit)
> male.female.copd.beta.diff.genes = sdc(male.fit, female.fit, "copd", lgrc.genes, fdr.cutoff=0.25, fil

[1] "Number of probes with sexual dimorphic differential expression: 1551"

```

### 3.2 Model 2: expression = Gender + Age + pkyrs

Male vs female analysis for COPD cases only.

```

> design.mtx = cbind(ctrl=1,
+                   gender=expr.meta$gender,
+                   age=expr.meta$age,
+                   pkyr=expr.meta$pkyrs)
> good.idx = apply(design.mtx,1,function(x){!any(is.na(x))}) & grepl("COPD",colnames(expr))
> copd.fit = lmFit(log(expr)[,good.idx], design.mtx[good.idx,])
> copd.fit = eBayes(copd.fit)
> good.idx = apply(design.mtx,1,function(x){!any(is.na(x))}) & grepl("CTRL",colnames(expr))
> ctrl.fit = lmFit(log(expr)[,good.idx], design.mtx[good.idx,])
> ctrl.fit = eBayes(ctrl.fit)
> copd.ctrl.gender.beta.diff.genes = sdc(copd.fit, ctrl.fit, "gender", lgrc.genes, fdr.cutoff=0.25, fil

[1] "Number of probes with sexual dimorphic differential expression: 1656"

```

## 4 Combine the Results

We use set intersection to combine the results from the two stratification analyses.

```

> male.female.copd.beta.diff.genes.all = sdc(male.fit, female.fit, "copd", lgrc.genes, fdr.cutoff=10, fil

[1] "Number of probes with sexual dimorphic differential expression: 13870"
> copd.ctrl.gender.beta.diff.genes.all = sdc(copd.fit, ctrl.fit, "gender", lgrc.genes, fdr.cutoff=10, fil

[1] "Number of probes with sexual dimorphic differential expression: 13870"
> all.beta.diff.genes = cbind(copd.ctrl.gender.beta.diff.genes.all, male.female.copd.beta.diff.genes.all)
> rename.col = grep("beta.diff", names(all.beta.diff.genes))
> names(all.beta.diff.genes)[rename.col[1:2]] = names(all.beta.diff.genes)[rename.col[1:2]] %+"copd.c"
> names(all.beta.diff.genes)[rename.col[3:4]] = names(all.beta.diff.genes)[rename.col[3:4]] %+"male.f"

> sdc.genes = merge(copd.ctrl.gender.beta.diff.genes, male.female.copd.beta.diff.genes, by=setdiff(inte
> sdc.genes = unique(sdc.genes)

> data(lgrc.sdc.genes)
> print("There are " %+% nrow(sdc.genes) %+% " SDCD genes")

[1] "There are 959 SDCD genes"

```

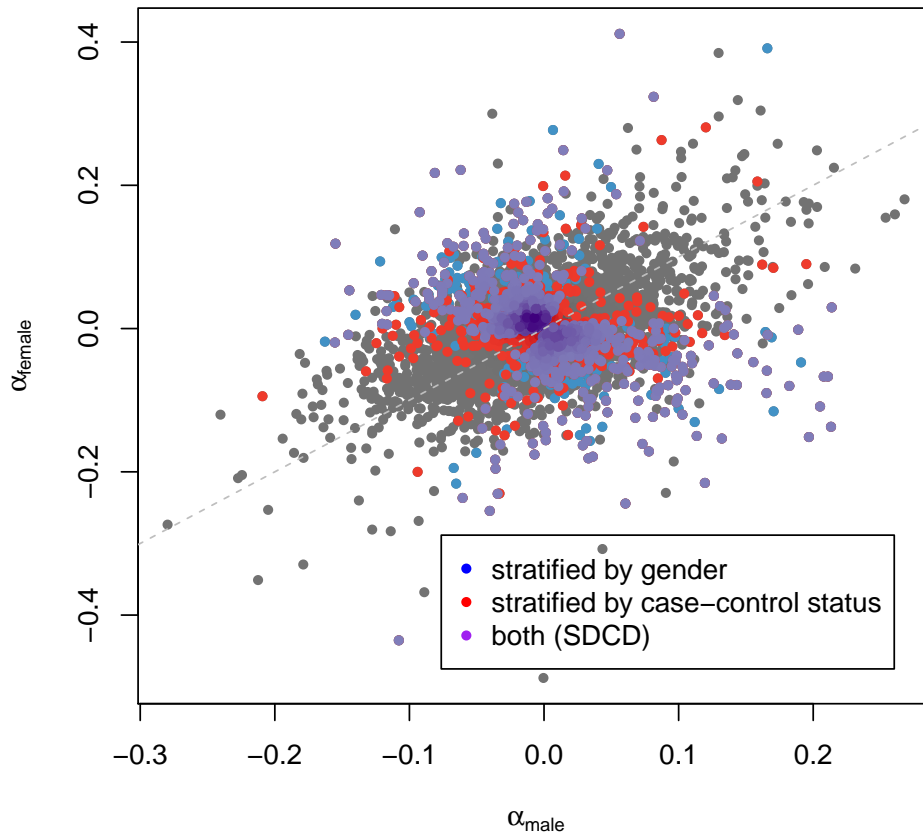
Then we can plot the results:

```

> # FIGURE 1B
> my.smart.plot(male.fit$coefficients[, "copd"], female.fit$coefficients[, "copd"], main="Coefficients of
> my.smart.plot(male.fit$coefficients[male.female.copd.beta.diff.genes$ensembl_gene_id, "copd"], female.f
> my.smart.plot(male.fit$coefficients[copd.ctrl.gender.beta.diff.genes$ensembl_gene_id, "copd"], female.f
> my.smart.plot(male.fit$coefficients[sdc.genes$ensembl_gene_id, "copd"], female.fit$coefficients[sdc.g
> abline(0,1,lty=2,col="gray")
> abline(h=c(qnorm(0.025),qnorm(0.975)),v=c(qnorm(0.025),qnorm(0.975)),lty=3,col="gray")
> smartlegend("right","bottom",c("stratified by gender","stratified by case-control status","both (SDCD)

```

## Coefficients of differential gene expression in males and females



```

> # FIGURE 1C
> all.beta.diff.genes$copd.ctrl.beta.diff = all.beta.diff.genes$copd.beta - all.beta.diff.genes$ctrl.beta
> this.pch = 20
> my.smart.plot(all.beta.diff.genes$copd.ctrl.beta.diff, -log10(all.beta.diff.genes$copd.ctrl.p), main="")
> my.smart.plot(all.beta.diff.genes[sdcd.genes$ensembl_gene_id,"copd.ctrl.beta.diff"], -log10(all.beta.c
> smartlegend("right","top",c("SDCD Genes"),pch=this.pch,col=c("purple"))
> CIpercent = 0.9
> abline(v=quantile(all.beta.diff.genes$beta.diff.copd.ctrl, c((1-CIpercent)/2, (1+CIpercent)/2)), col="")
> extreme.betas.idx = abs(sdcd.genes$beta.diff.x) > 0.25 | (abs(sdcd.genes$beta.diff.x) > 0.2 & sdcd.gen
> extreme.betas = cbind(sdcd.genes[extreme.betas.idx, c("hgnc_symbol","beta.diff.x","copd.ctrl.p","male
+ n.log.p=-log10(sdcd.genes[extreme.betas.idx, c("copd.ctrl.p")]))
> print("Extreme beta_diff points are: ")

[1] "Extreme beta_diff points are: "

> print(extreme.betas)

  hgnc_symbol beta.diff.x  copd.ctrl.p male.female.p.adj copd.ctrl.p.adj
244      TBX18  0.2625562 3.576277e-04      0.028464545  0.0280109784
260      CD207 -0.2537996 4.160540e-05      0.125745744  0.0098877620
298     ACVR1C -0.2435254 5.584837e-06      0.010899375  0.0028915492
304       EDN3 -0.3337424 3.564346e-07      0.006448865  0.0004697484
377     HMGCS2 -0.2716475 3.437951e-03      0.102258238  0.0839056712

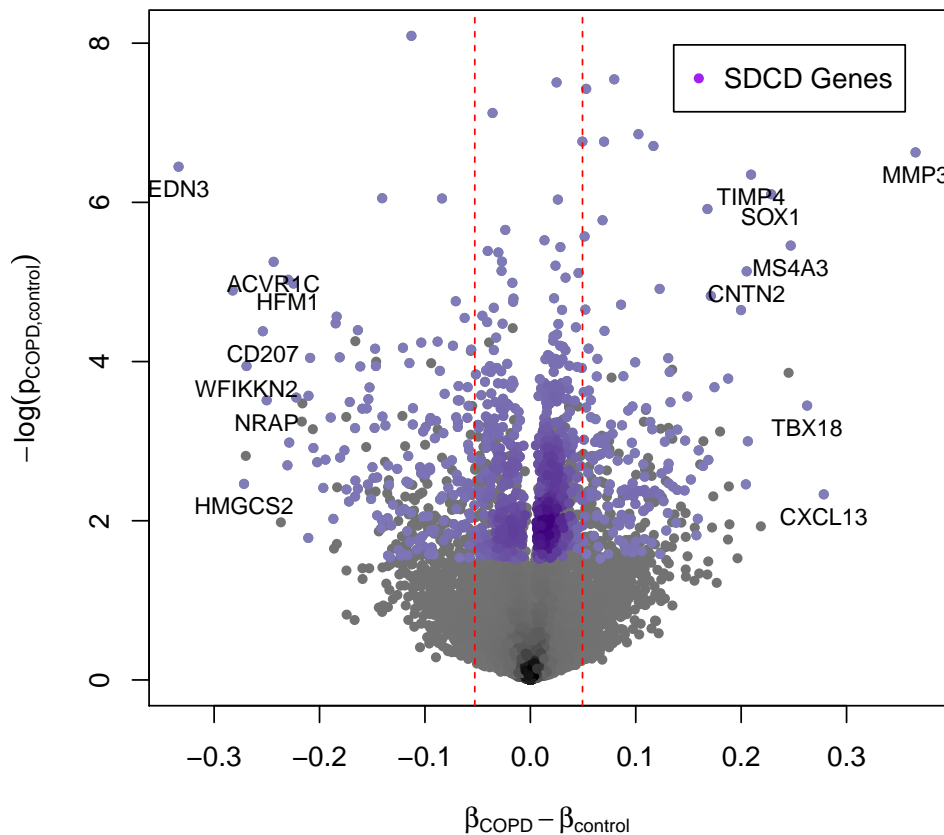
```

511	MS4A3	0.2470051	3.487462e-06	0.014957343	0.0022980789
515	MMP3	0.3654424	2.353935e-07	0.221377877	0.0003412500
546	CXCL13	0.2783933	4.655509e-03	0.063294179	0.0984226964
551	TIMP4	0.2093865	4.481220e-07	0.001322135	0.0005413688
590	HFM1	-0.2297657	9.393241e-06	0.128370483	0.0040051121
744	WFIKK2	-0.2692408	1.139020e-04	0.065013255	0.0159592624
820	SOX1	0.2284689	7.902981e-07	0.068103876	0.0008357564
830	CNTN2	0.2052041	7.347437e-06	0.001349746	0.0034359932
861		-0.2824065	1.279387e-05	0.063908211	0.0048808618
892	NRAP	-0.2501514	3.063447e-04	0.014957343	0.0262491592

	chromosome_name	n	log.p
244	6	3	3.446569
260	2	4	3.380850
298	2	5	2.52990
304	20	6	4.48020
377	1	2	2.463700
511	11	5	4.57491
515	11	6	6.28206
546	4	2	3.32033
551	3	6	3.48604
590	1	5	5.027185
744	17	3	3.943468
820	13	6	6.102209
830	1	5	5.133864
861	2	4	4.892998
892	10	3	5.13790

```
> text(extreme.betas$beta.diff.x, extreme.betas$n.log.p, extreme.betas$hgnc_symbol, pos=1, cex=0.8)
```

## Volcano plot for COPO-control differential expression



```
> # Figure S2
> all.beta.diff.genes$male.female.beta.diff = all.beta.diff.genes$male.beta - all.beta.diff.genes$female.beta
> this.pch = 20
> my.smart.plot(all.beta.diff.genes$male.female.beta.diff, -log10(all.beta.diff.genes$male.female.p), main="Volcano plot for COPO-control differential expression",
> my.smart.plot(all.beta.diff.genes[sdcd.genes$ensembl_gene_id,"male.female.beta.diff"], -log10(all.beta.diff.genes[sdcd.genes$ensembl_gene_id,"male.female.p"]),
> smartlegend("right", "top", c("SDCD Genes"), pch=this.pch, col=c("purple"))
> CIpercent = 0.9
> abline(v=quantile(all.beta.diff.genes$beta.diff.male.female, c((1-CIpercent)/2, (1+CIpercent)/2)), col="red", lty=2)
```

## 5 Session Information

```
> sessionInfo()
```

R version 3.1.1 Patched (2014-09-25 r66681)

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] stats4    parallel  stats      graphics  grDevices  utils      datasets  
[8] methods   base
```

```
other attached packages:
```

```
[1] COPDSexualDimorphism_1.2.0    gtools_3.4.1  
[3] gplots_2.14.2                 GenomicRanges_1.18.0  
[5] GenomeInfoDb_1.2.0           IRanges_2.0.0  
[7] S4Vectors_0.4.0              BiocGenerics_0.12.0  
[9] limma_3.22.0                  beeswarm_0.1.6  
[11] RColorBrewer_1.0-5           NCBI2R_1.4.6  
[13] COPDSexualDimorphism.data_1.1.0
```

```
loaded via a namespace (and not attached):
```

```
[1] KernSmooth_2.23-13 XVector_0.6.0    bitops_1.0-6    caTools_1.17.1  
[5] gdata_2.13.3      tools_3.1.1
```