Projeto Final

Henrique Tostes de Sousa

21/11/2020

Introdução

Projeto

As distrofias musculares são um grupo de distúrbios clínicos e geneticamente heterogêneos, caracterizados por alterações degenerativas progressivas nas fibras musculares esqueléticas. A distrofia muscular de Duchenne (DMD), a forma mais comum de distrofia muscular, é causada por mutações no gene da distrofina, muitas vezes levando à redução dos níveis transcritos de mRNA e invariavelmente à ausência da proteína. A distrofina é uma grande proteína do citoesqueleto associada a um grande complexo (conhecido como complexo de proteína associada à distrofina ou DAPC) que liga o citoesqueleto à matriz extracelular. Este complexo é interrompido em muitas formas de distrofia muscular e mutações em genes que codificam alguns componentes da DAPC além da distrofina estão associadas a outras distrofias musculares do membro do membro da cintura.

Plataforma

A plataforma utilizada foi a matriz de oligonucleotídeos, onde cada microarray tem até 500.000 células-sonda individuais, cada uma com 18 m e contendo milhões de moléculas de DNA idênticas. Nesse caso foi usado o modelo *Affymetrix Human Genome U95E Array*. O conjunto do genoma humano U95 (HG-U95), consistindo em cinco matrizes GeneChip, contém quase 63.000 conjuntos de sondas interrogando aproximadamente 54.000 clusters derivados do banco de dados UniGene.

Amostras

Dez biópsias de quadríceps de pacientes com DMD foram comparadas com 10 biópsias de quadríceps de controles não afetados. As 10 biópsias DMD eram de homens jovens (5-7 anos). As biópsias não afetadas eram principalmente de homens jovens (1–10 anos) (7 biópsias), mas incluíam 3 biópsias de homens adultos. Os controles não afetados foram feitas biópsias devido à suspeita de fraqueza muscular, mas não mostraram sinais de patologia no exame histológico e expressaram níveis normais de distrofina. Os pacientes com DMD mostraram sintomas clínicos consistentes com um diagnóstico de DMD e as biópsias mostraram ser deficientes em distrofina por imunofluorescência.

Objetivo

Identificar genes que estão significativamente associados a condição do estudo, tanto para afetado quanto para normal.

Análise

Carregando os dados

```
# Carregando os dados da plataforma GEO

gset <- getGEO("GDS612", GSEMatrix =TRUE) # Baixando o estudo
eset <- GDS2eSet(gset,do.log2=TRUE)
#eset1 <- eset[[1]] #extract the first element in the list
dat <- exprs(eset) # Extrai os dados da expressão em uma matrix
dat <- as.data.frame(dat)
dat.anno <- pData(eset) # Extrai os dados de anotação em um data frame.
```

Obersevando os dados

```
dat[1:5,1:5] # Dados da expressão
```

```
##          GSM16287 GSM16288 GSM16289  GSM16290  GSM16298
## 67022_at 6.061776 6.100137 5.722466  6.004501  6.965784
## 67024_at 8.868205 9.782343 9.471878 10.195003 10.353808
## 67026_at 7.762880 6.738768 7.454505  7.177918  7.315602
## 67028_at 8.605480 7.706669 8.342964  8.803162  8.378512
## 67030_at 9.186114 9.373953 9.868205  8.886611  8.773469
```

```
dim(dat)
```

```
## [1] 12639    21
```

```
dat.anno[10:14,] # Dados de anotação
```

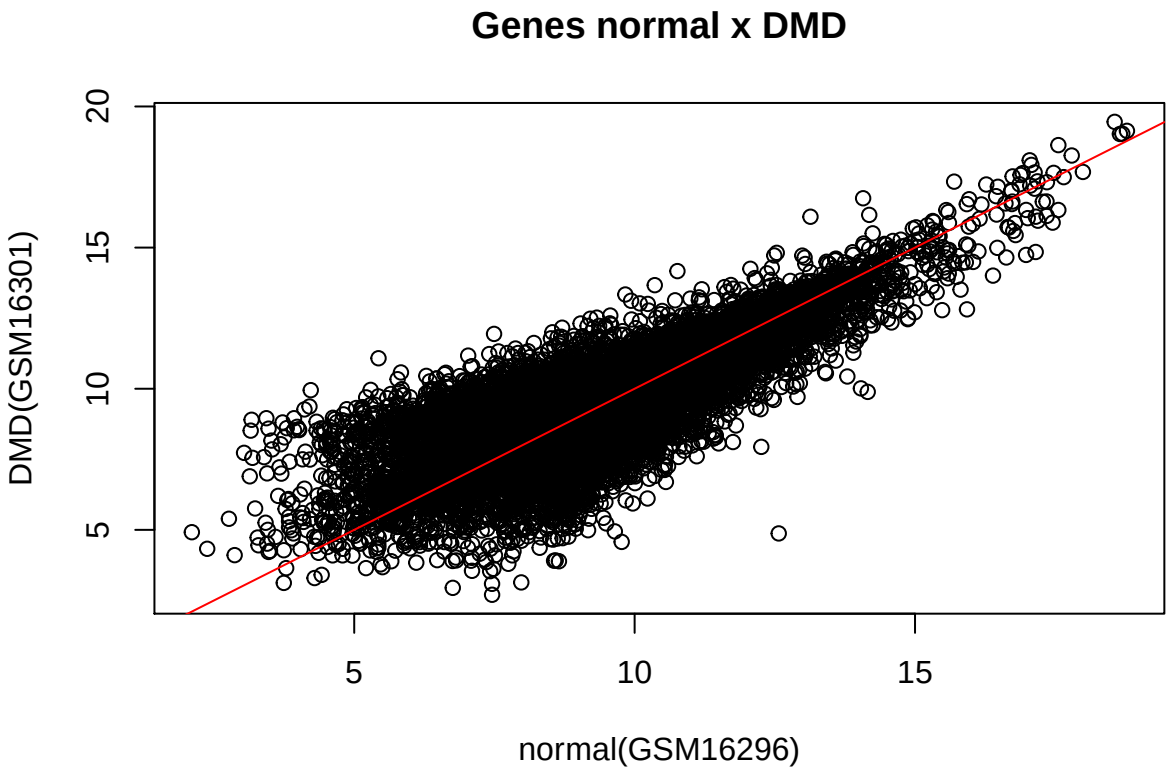
```
##            sample               disease.state
## GSM16296 GSM16296                      normal
## GSM16297 GSM16297                      normal
## GSM16299 GSM16299 Duchenne muscular dystrophy
## GSM16301 GSM16301 Duchenne muscular dystrophy
## GSM16302 GSM16302 Duchenne muscular dystrophy
##
## GSM16296                                          Value for GSM16296: normal Human skeletal muscle samp
## GSM16297                                         Value for GSM16297: normal Human skeletal muscle sampl
## GSM16299         Value for GSM16299: duchenne Muscular Dystrophy (DMD) patient's skeletal muscle samp
## GSM16301 Value for GSM16301: duchenne Muscular Dystrophy (DMD) patient's skeletal muscle sample 1(rep
## GSM16302         Value for GSM16302: duchenne Muscular Dystrophy (DMD) patient's skeletal muscle samp
```

```
dim(dat.anno)
```

```
## [1] 21  3
```

Primeira análise

No gráfico abaixo, os pontos acima da reta vermelha são genes que estão mais expressos na amostra com DMD, os pontos abaixo da reta são genes mais expressos na amostra normal.



Podemos notar que alguns se destacam por estarem bem acima ou bem abaixo da reta.

Filtrando Genes

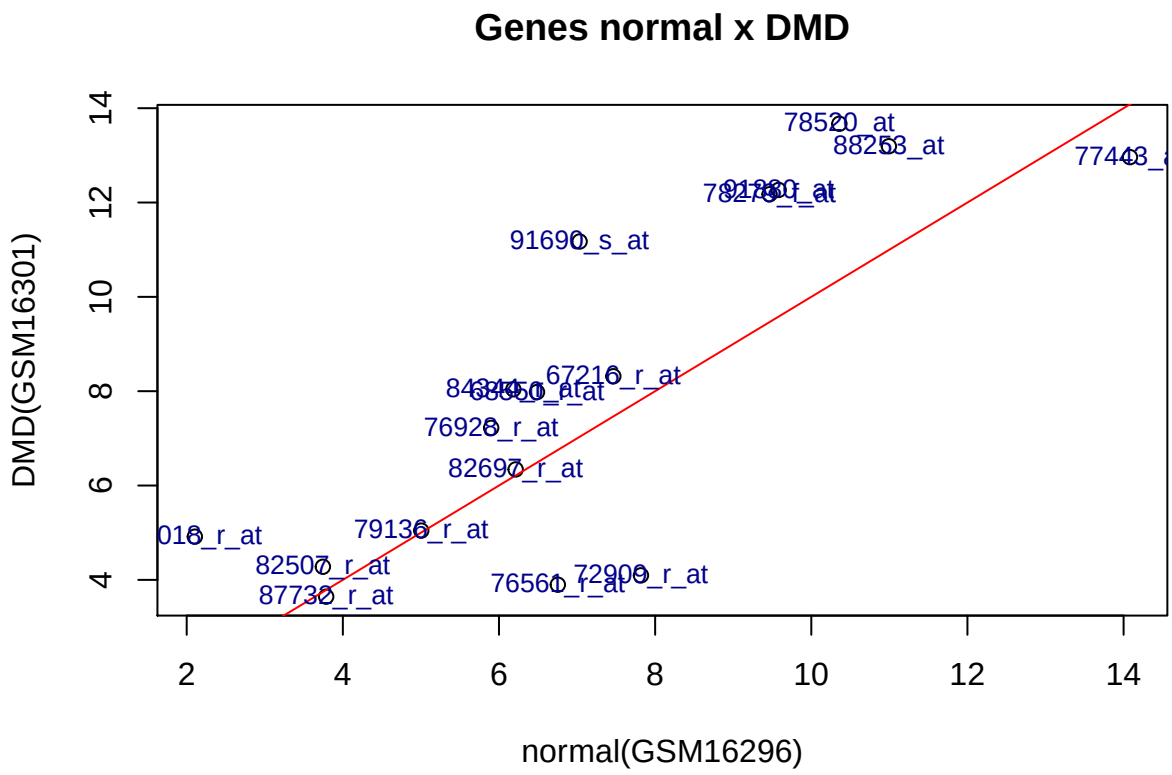
No conjunto de dados inicial temos um total de 12639 genes, qualquer análise feita em todo esse conjunto não seria esclarecedora. Por isso, vamos filtrar através do desvio padrão através de todas as amostras. Se o desvio padrão para um determinado gene através de todas as amostras for maior que 2, ele será mantido, do contrário o ignoraremos. E em seguida vamos plotar o mesmo gráfico anterior, mas dessa vez colocando o nome dos genes e ver o resultado.

```
# Filtrando os genes
x <- cbind(dat, apply(dat,1,sd))
dat.fil <- subset(x, x$`apply(dat, 1, sd)` > 2, select=c(GSM16287:GSM16309))
dim(dat.fil)
```

```
## [1] 17 21
```

```
# Construindo o grafico
plot(dat.fil$GSM16296, dat.fil$GSM16301, xlab = "normal(GSM16296)", ylab="DMD(GSM16301)", main = "Genes
abline(a = 0, b = 1, col = "red")
```

```
nomes <- row.names(dat.fil)
text(dat.fil$GSM16296, dat.fil$GSM16301,labels=nomes,cex=.8, col = "dark blue")
```

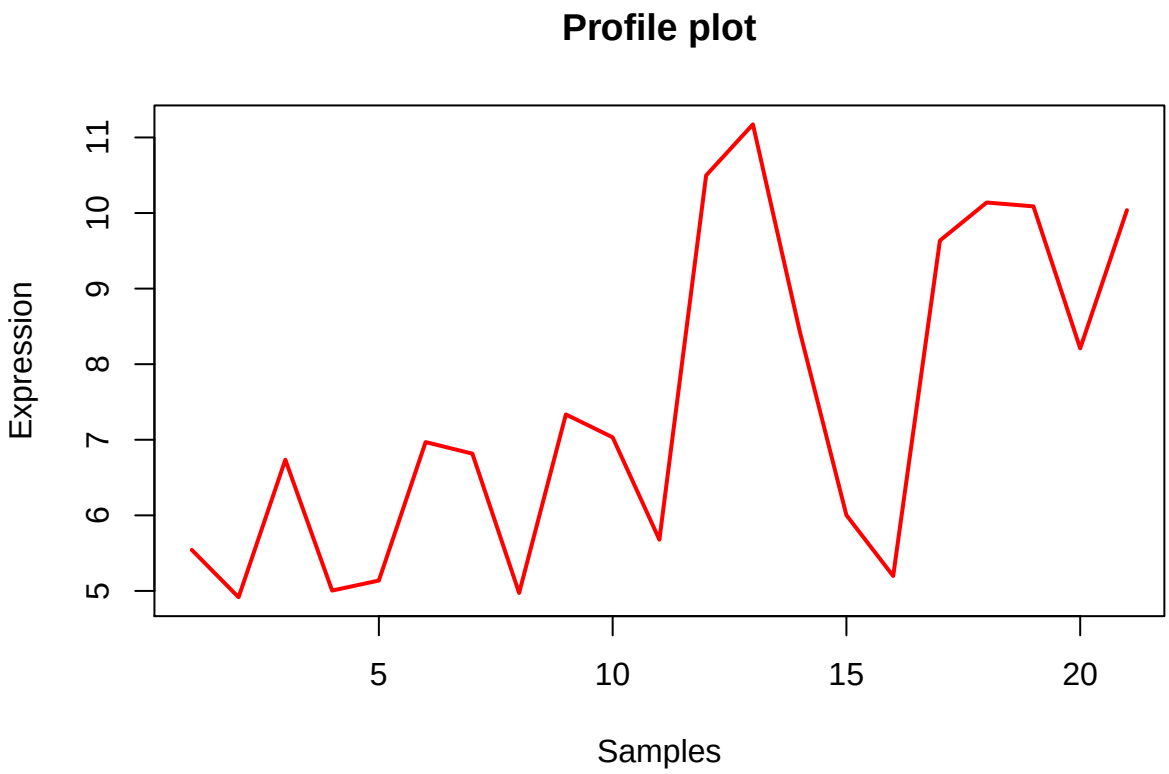


Ficamos com 17 genes para estudar após o filtro. Estamos observando a amostra normal(GSM16296) e a amostra afetada(GSM16301), por isso há 3 genes em que a expressão é a mesma em ambas as condições, mas todos os outros 14 genes estão com expressões diferentes, ou maiores ou menores.

Profile Plot

Agora vamos pegar o gene 91690_s_at que está bem mais expresso na amostra afetada do que na normal e vamos construir um *Profile Plot* para ver como está sua expressão através de todas as amostras.

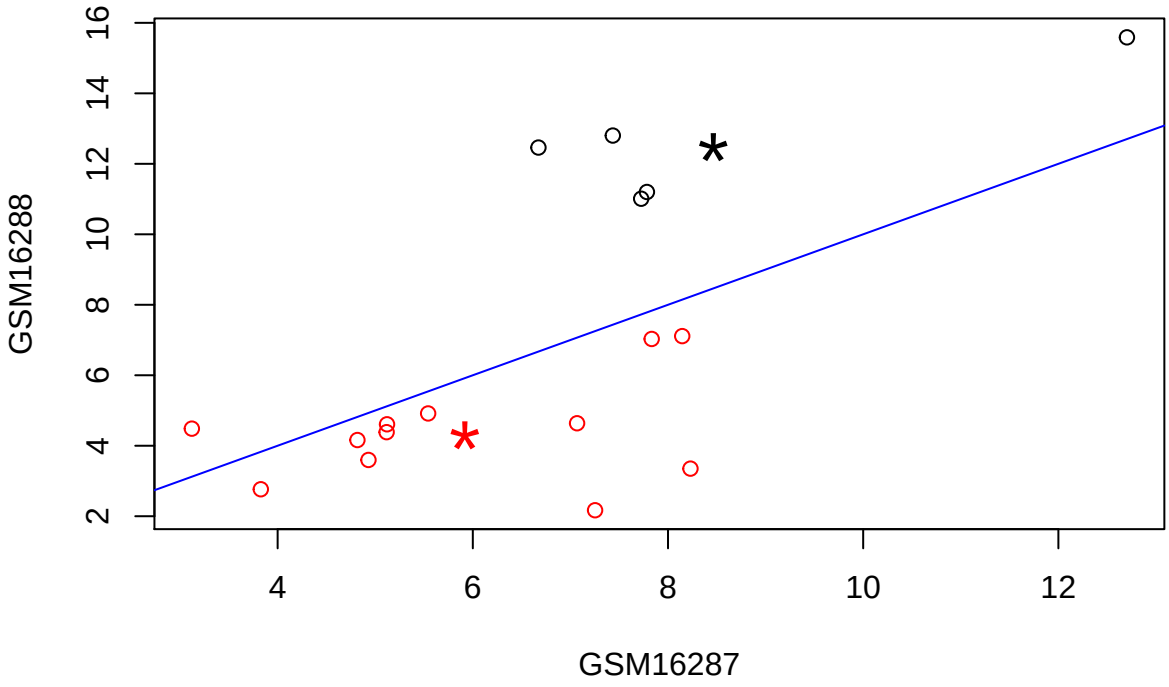
```
plot(x = 1:ncol(dat.fil),
     y = as.numeric(dat.fil["91690_s_at",]),
     xlab = "Samples", ylab = "Expression",
     type = "l", col = "red", main= "Profile plot", lwd = 2)
```



Até a amostra 10 são todas normais e a partir da amostra 11 são todas afetadas. Sendo assim, podemos ver que em média o gene 91690_s_at está mais expresso nas amostras afetadas.

Métodos não-supervisionados

```
# Cluster
matriz <- as.matrix(dat.fil)
clus <- kmeans(matriz, centers=2, iter.max=20)
plot(matriz, col = clus$cluster, cex = 1)
points(clus$centers, col = 1:2, pch = "*",cex=2.5)
abline(0,1, col = "blue")
```

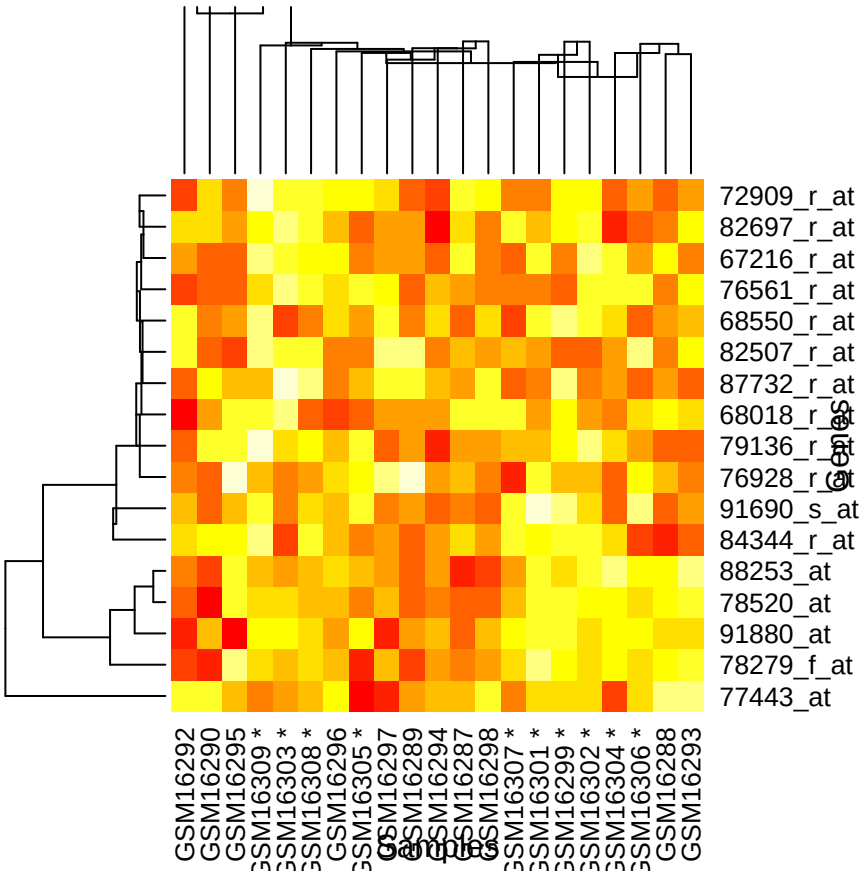


Consiguimos ver pelo método não-supervisionado de cluster que ele também encontrou dois grupos. Utilizando a mesma reta do primeiro gráfico vemos que o grupo vermelho são os mais expressos em condições normais e o grupo preto mais expresso em condições afetadas.

Vamos agora construir um *Heatmap* e ver se descobrimos padrões ocultos.

```
# Marcando os afetados com *
for(x in colnames(matriz)){
  if(dat.anno[x,"disease.state"]!="normal"){
    novo_nome <- paste(x,"*")
    colnames(matriz)[colnames(matriz) == x] <- novo_nome
  }
}

# Heatmap
heatmap(matriz, distfun = function(matriz) dist(matriz,method='manhattan'), hclustfun =
          function(matriz) hclust(dist(matriz,method='manhattan'), method= 'median'), xlab="Samples", y
```



Quanto a separação das amostras, o método não conseguiu distinguir e agrupar os afetados dos normais, sendo assim eles ficaram misturados. Por outro lado, o método encontrou uma certa relação entre alguns genes. Nas linhas inferiores, do gene 88253_at até o gene 78279_f_at parecem ter alguma relação, visto a árvore construída à esquerda.

Conclusão

Conseguimos inicialmente filtrar mais de 12 mil genes em 17 genes baseado na discrepância encontrada(alto desvio padrão) entre as expressões através das amostras normais e afetadas. Em seguida confirmamos com um *Profile Plot* que o gene 91690_s_at aparece mais expresso em média nas amostras afetadas, sendo assim um possível gene marcador.

Em seguida, os métodos não-supervisionados nos mostraram alguns padrões não óbvios. O método *Cluster* conseguiu dividir os genes em dois grupos, que são os mais expressos em afetados e os mais expressos em amostras normais. Poderiamos ainda analisar cada um desse genes construindo um *Profile Plot* para cada um deles e ver como suas expressões se dão ao longo de todas as amostras, como fizemos com o gene 91690_s_at. Na sequência, o *Heatmap* não revelou tanto quanto o esperado, mas ainda sim nos mostrou que existe alguma relação entre os genes citados na seção anterior.