Gene set analysis of RNA-Seq data from Jaffe et al.

Nima Hejazi

2016 Nov 12 (Sat), 16:25:15

Abstract

Having detected differentially expressed genes between neural cells extracted from human adult and fetal samples, here, we report on findings from efforts to assess the degree to which these differentially expressed genes may be associated with the **H3K4me3** histone modification by examining promoter regions of a conservative set of reference genes (*RefSeq*) displaying **H3K4me3** changes across human brain and liver cell lines. Here, we make use of a permutation-based approach to examine the proportion of the differentially expressed genes that are present in H3K4me3 promoters of the reference genes across liver and brain cell lines, deriving exact p-values for the proportion of genes in each of these tissue classes.

```
library(AnnotationHub)
```

```
##
## Attaching package: 'AnnotationHub'

## The following object is masked from 'package:Biobase':
##
##     cache

## The following object is masked from 'package:dplyr':
##
##     query
```

```
library(pander)
library(ggplot2)
genes.sig <- tt_out_ranked %>%
  subset(fdrBH < 0.25)
pander(head(genes.sig))
```

geneID	lowerCI	FoldChange	upperCI	pvalue	fdrBH
NKX6-2	7.436e-11	4.675e-09	2.939e-07	1.124e-06	0.01307
TMEM235	6.422e-10	3.004e-08	1.405e-06	1.457e-06	0.01307
OPALIN	5.932e-15	4.287e-12	3.098e-09	4.55e-06	0.02721
HHATL	1.58e-10	2.782e-08	4.9e-06	1.998e-05	0.06366
HOXD1	7.971e-15	1.224e-11	1.88e-08	1.712e-05	0.06366
KLK6	3.004e-15	6.668e-12	1.48e-08	2.129e-05	0.06366

First, let us merely examine the first few genes marked as differentially expressed between human fetal and adult neurons. Note that these genes were isolated from performing alignment on RNA-Seq transcripts using the **Kallisto** pseudo-aligner and performing statistical analysis with the "voom" method of the popular `Limma` package for linear modeling in genomics. These genes were marked as being differentially expressed by using an *cutoff of the Benjamini-Hochberg procedure to control the False Discovery Rate of 25%*. The association of the full set of differentially expressed genes with H3K4me3 promoters in brain and liver tissue will be assessed below via permutation testing.

```
# set up Annotation Hub for extracting metadata
ah <- AnnotationHub()
```

```
## Warning: database may not be current
##   database: '/Users/nimahejazi//.AnnotationHub/annotationhub.sqlite3'
##   reason: Couldn't resolve host name
```

```
## snapshotDate(): 2016-11-10
```

```
ah <- subset(ah, species == "Homo sapiens")
qhs <- query(ah, "H3K4me3")
roadmap <- subset(qhs, dataprovider == "BroadInstitute")
```

Above, we merely set up an `AnnotationHub` object and extract relevant data by a set of queries to narrow our search.

```
# get genes with H3K4me3 in Brain cells
brain <- query(roadmap, "Brain")
gr.brain <- subset(brain, title == brain$title[[11]])[[1]]
```

```
## require("rtracklayer")
```

```
## loading from cache '/Users/nimahejazi//.AnnotationHub/35814'
```

```
## Warning in download.file(url, destfile, quiet = TRUE): unable to resolve
## 'hgdownload.cse.ucsc.edu'
```

```
## using guess work to populate seqinfo
```

```
# get genes with H3K4me3 in Liver cells
liver <- query(roadmap, "Liver")
gr.liver <- subset(liver, title == liver$title[[2]])[[1]]
```

```
## loading from cache '/Users/nimahejazi//.AnnotationHub/35807'
```

```
## Warning in download.file(url, destfile, quiet = TRUE): unable to resolve
## 'hgdownload.cse.ucsc.edu'
```

```
## using guess work to populate seqinfo
```

Above, we set up separate `GRanges` objects holding data from queries on brain and liver tissue from the *roadmap epigenomics project*.

```
# get promoters of genes from RefSeq (conservative) annotation
qhs <- query(ah, "RefSeq")
refseq <- qhs[qhs$genome == "hg19" & qhs$title == "RefSeq Genes"]
refseq <- refseq[[1]]
```

```
## loading from cache '/Users/nimahejazi//.AnnotationHub/5040'
```

```
## Warning in download.file(url, destfile, quiet = TRUE): unable to resolve
## 'hgdownload.cse.ucsc.edu'

## using guess work to populate seqinfo
```

```
promoters <- promoters(refseq)
prom.reds <- reduce(promoters, ignore.strand = TRUE)
```

Now, we extract annotation information on promoters from the `RefSeq` collection, a conservative set of genes curated for reference.

```
# for peaks found in brain cell lines
ov.brain <- findOverlaps(promoters, gr.brain)
peaks.brain <- reduce(gr.brain)
int.brain <- intersect(prom.reds, peaks.brain)
```

For permutation testing, we will need a list of `RefSeq` promoters that are associated with H3K4me3 `narrowPeak` data from brain cells.

```
# for peaks found in liver cell lines
ov.liver <- findOverlaps(promoters, gr.liver)
peaks.liver <- reduce(gr.liver)
int.liver <- intersect(prom.reds, peaks.liver)
```

For comparison with the above, we will also need a list of `RefSeq` promoters that are associated with H3K4me3 `narrowPeak` data from liver cells.

```
# permute labels in full gene list, re-compute proportions
n_perm = 10000
props.brain <- vector("list", n_perm)
props.liver <- vector("list", n_perm)
for (i in 1:n_perm) {
  genes_fdr_perm_ind <- sample(nrow(tt_out_ranked), nrow(genes.sig))
  genes_fdr_perm <- tt_out_ranked[genes_fdr_perm_ind, ]
  prop.brain[[i]] <- sum(genes_fdr_perm$geneID %in% int.brain) / nrow(genes.sig)
  prop.liver[[i]] <- sum(genes_fdr_perm$geneID %in% int.liver) / nrow(genes.sig)
}
```

Here, in order to perform permutation testing, we randomly select sets of genes (each set being the size of the actual set of differentially expressed genes at the 25% FDR level – that is, 71 genes) that we treat as being differentially expressed. We then compute the proportion of these genes that are associated with (fall near) the set of H3K4me3 promoter regions in brain cells and in liver cells independently. We store these proportions in lists structures so that we can examine them against the proportions computed from the "true" list of differentially expressed genes.

```
# assess exact p-values of proportions
n_perm = 10000
prop.brain.sig <- no_geneSig_brain / nrow(genes.sig)
prop.liver.sig <- no_geneSig_liver / nrow(genes.sig)
pval_perm_brain <- sum(prop.brain > prop.brain.sig) / n_perm
pval_perm_liver <- sum(prop.liver > prop.liver.sig) / n_perm
print(paste("Exact p-value from permutation for H3K4me3 brain promoters:",
            pval_perm_brain))
```

```
## [1] "Exact p-value from permutation for H3K4me3 brain promoters: 0.1289"
```

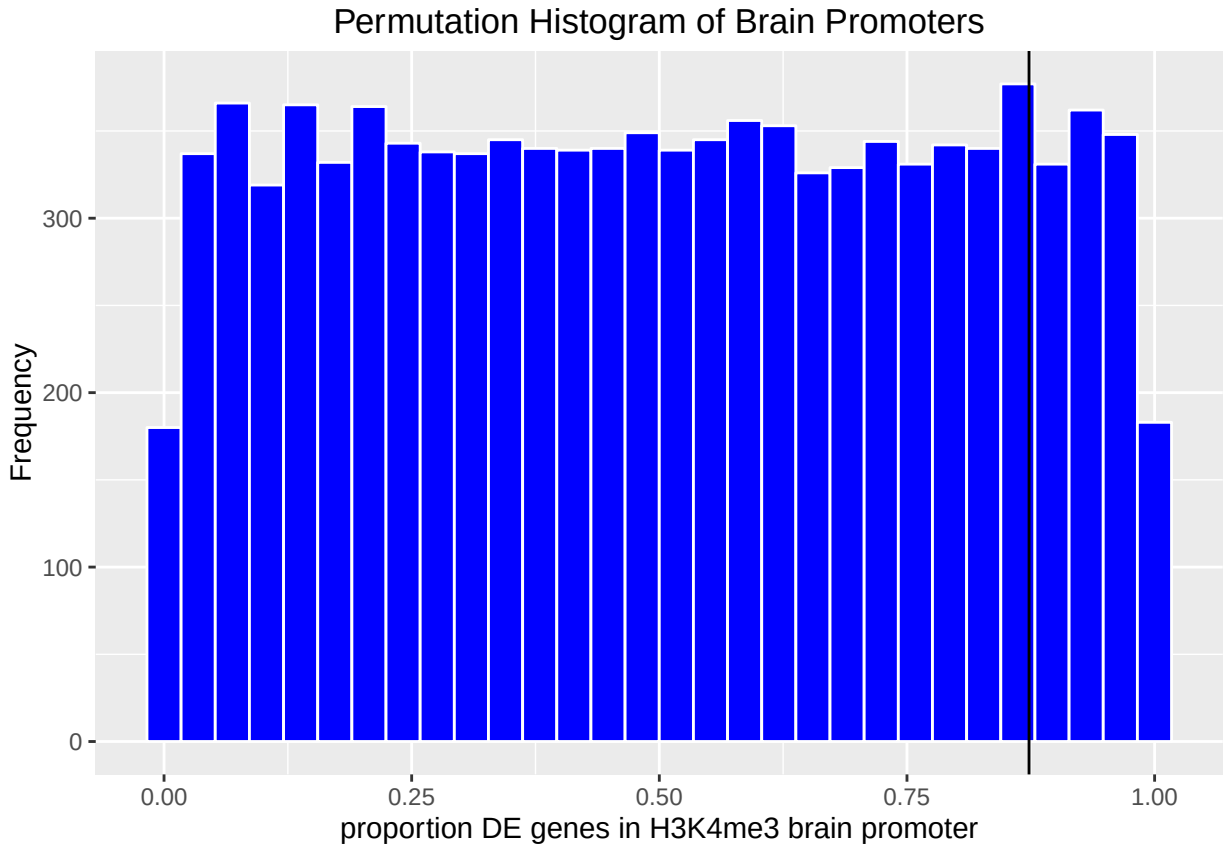
```
print(paste("Exact p-value from permutation for H3K4me3 liver promoters:",
            pval_perm_liver))
```

```
## [1] "Exact p-value from permutation for H3K4me3 liver promoters: 0.4756"
```

Although this permutation approach to computing exact p-values for the association of the differentially expressed genes with H3K4me3 promoters in human liver and brain cells yields different p-values as shown above, we are unable to conclude that there is a statistically significant difference. That is, although there are distributional differences between the proportion of differentially expressed genes associated with the promoter modification of interest, the use of the exact p-value does not allow for an inferential conclusion to be drawn as to the magnitude of this difference.

```
suppressMessages(qplot(prop.brain, geom = "histogram",
                       xlab = "proportion DE genes in H3K4me3 brain promoter",
                       ylab = "Frequency",
                       main = "Permutation Histogram of Brain Promoters",
                       fill = I("blue"),
                       col = I("white")
                       ) + geom_vline(xintercept = prop.brain.sig))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

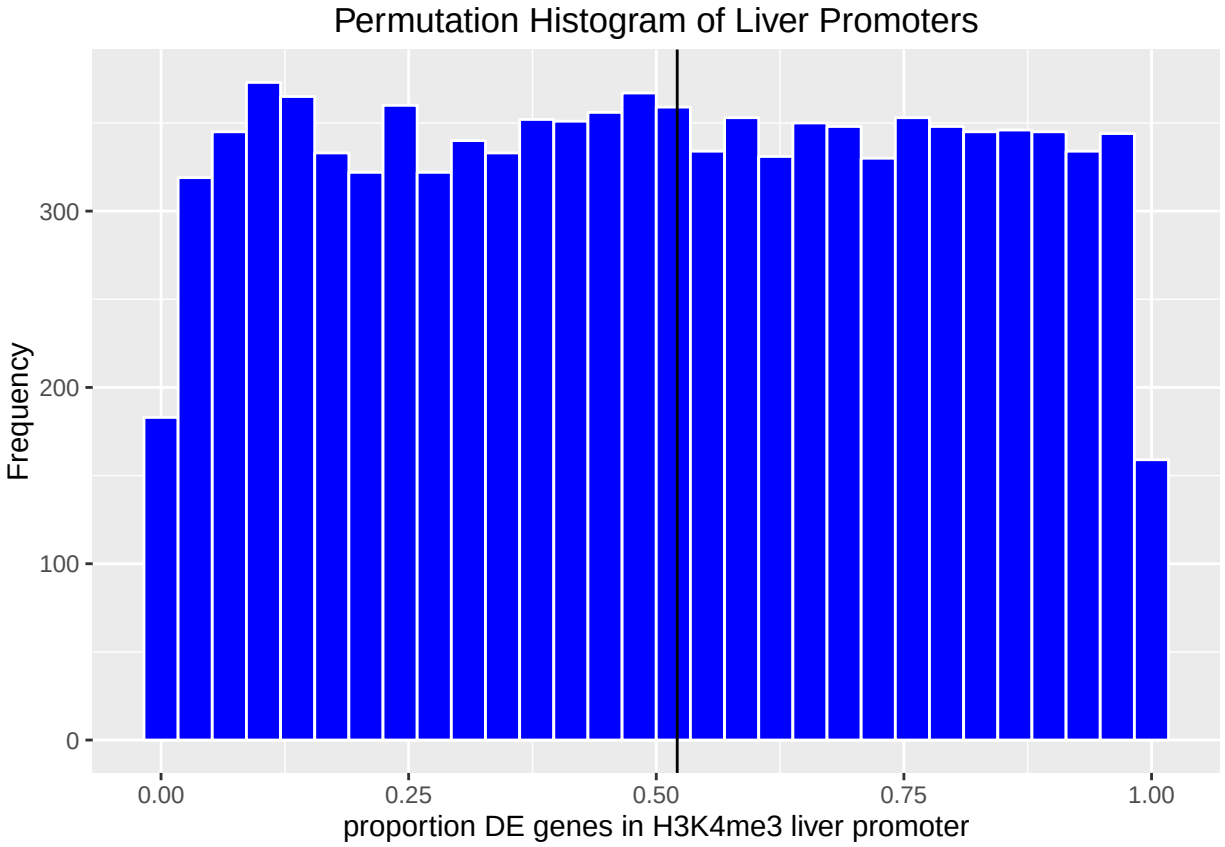


The histogram above provides the permutation distribution of the proportions of differentially expressed genes falling near/in H3K4me3-associated promoter regions in brain cells from the roadmap epigenomics

project. As expected of a permutation distribution, it is roughly uniform, though we note that the value of the proportion of differentially expressed genes from the "true" list (as analyzed by the methods described earlier in this report) fall fairly high in this distribution (marked by the black line) – that is, a relatively few number of proportions from the permuted gene list display more extreme values. This leads to the conclusion that there may be an associated between the differentially expressed genes and H3K4me3 promoter sequences in brain cells, though the exact permutation test may not have enough power to adequately capture such a nuanced association.

```
suppressMessages(qplot(prop.liver, geom = "histogram",
                       xlab = "proportion DE genes in H3K4me3 liver promoter",
                       ylab = "Frequency",
                       main = "Permutation Histogram of Liver Promoters",
                       fill = I("blue"),
                       col = I("white")
                       ) + geom_vline(xintercept = prop.liver.sig))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



The histogram above provides the permutation distribution of the proportions of differentially expressed genes falling near/in H3K4me3-associated promoter regions in liver cells from the roadmap epigenomics project. As expected of a permutation distribution, it is roughly uniform, though we note that the value of the proportion of differentially expressed genes from the "true" list (as analyzed by the methods described earlier in this report) fall towards the middle of this distribution (marked by the black line) – that is, many of the proportions from the permuted gene lists display more extreme values. This leads to the (fairly conservative) conclusion that there is likely not an association between differentially expressed genes and H3K4me3 promoter sequences in liver cells.