dse

Okwir Julius

2/23/2022

Loading the required packages

```
library(DESeq2)
library(tximport)
library(rhdf5)
library(ggplot2)
library(apeglm)
library(org.Hs.eg.db)
```

Part 1:Hisat2

Data import

```
# import hisat2 counts data
countdata <- read.table("./hisat2_counts.txt", header = TRUE, skip = 1, row.names = 1)

# import metadata
metadata <- read.delim("practice.dataset.metadata.tsv", row.names = 1)
```

Data processing

```
# process counts data
# Remove length/char columns
countdata <- countdata[ ,c(-1:-5)]

# rename columns of countdata with sample names
colnames(countdata) <- paste0("sample", 37:42)

head(countdata)
```

```
##                 sample37 sample38 sample39 sample40 sample41 sample42
## ENSG00000223972        0        0        0        0        0        0
## ENSG00000227232       52       48      187       56       59       69
## ENSG00000278267        7       11       36        2        3        4
## ENSG00000243485        0        0        0        1        0        0
## ENSG00000284332        0        0        0        0        0        0
## ENSG00000237613        0        0        0        1        1        0
```

Deferential gene expression analysis with deseq2

```
# compare colnames of count data to rownames of metadata
# the two should be the same
colnames(countdata) == rownames(metadata)
```

```
## [1] TRUE TRUE TRUE TRUE TRUE TRUE
```

```
# create deseq2 data object
ddsMat <- DESeqDataSetFromMatrix(countData = countdata,
                                 colData = metadata,
                                 design = ~Condition)


# Find differential expressed genes
ddsMat <- DESeq(ddsMat)

# obtain the results
results <- results(ddsMat)

# head of results
head(results)
```

```
## log2 fold change (MLE): Condition normal vs disease
## Wald test p-value: Condition normal vs disease
## DataFrame with 6 rows and 6 columns
##                  baseMean log2FoldChange     lfcSE      stat    pvalue
##                 <numeric>      <numeric> <numeric> <numeric> <numeric>
## ENSG00000223972  0.000000             NA        NA        NA        NA
## ENSG00000227232 71.880462       0.617815   0.54810  1.127193 0.2596610
## ENSG00000278267  9.329872       2.540255   1.06852  2.377365 0.0174368
## ENSG00000243485  0.152147      -0.780352   4.08047 -0.191241 0.8483371
## ENSG00000284332  0.000000             NA        NA        NA        NA
## ENSG00000237613  0.333436      -1.771494   4.03022 -0.439553 0.6602608
##                      padj
##                 <numeric>
## ENSG00000223972        NA
## ENSG00000227232  0.574518
## ENSG00000278267  0.137609
## ENSG00000243485        NA
## ENSG00000284332        NA
## ENSG00000237613        NA
```

```
# Generate summary of the results.
summary(results)
```

```
##
## out of 38258 with nonzero total read count
## adjusted p-value < 0.1
## LFC > 0 (up)       : 1006, 2.6%
## LFC < 0 (down)     : 1493, 3.9%
## outliers [1]       : 293, 0.77%
## low counts [2]     : 12932, 34%
## (mean count < 3)
## [1] see 'cooksCutoff' argument of ?results
## [2] see 'independentFiltering' argument of ?results
```

store results from dse

```
# store results in text file
# store all results
write.table(x = as.data.frame(results),
            file = "results.txt",
            sep = '\t',
            quote = F,
            col.names = NA)

# store only statistically significant results(padj < 0.05)
results_sig <- subset(results, padj < 0.05)

write.table(x = as.data.frame(results_sig),
            file = "results_significant.txt",
            sep = '\t',
            quote = F,
            col.names = NA)
```

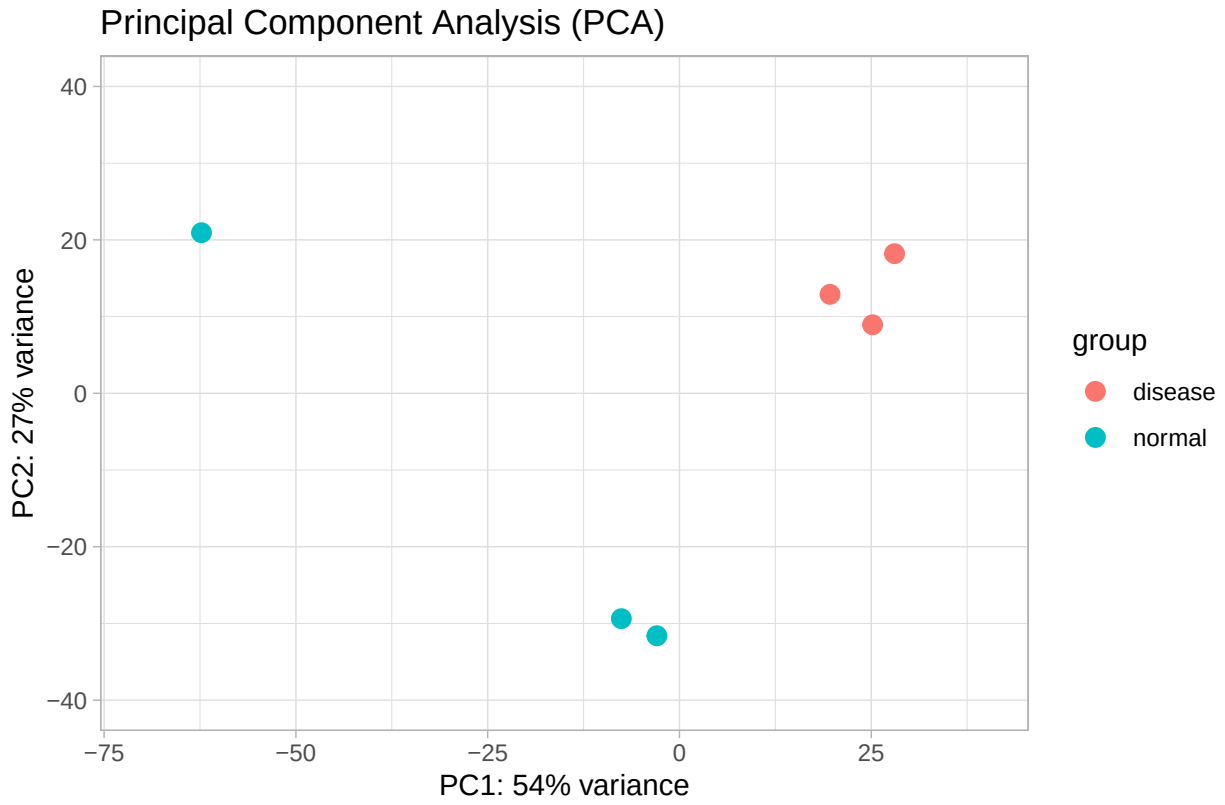
plots of gene expression data

```
# Convert all samples to rlog(regularized logarithm) for visualization
ddsMat_rlog <- rlog(ddsMat, blind = FALSE)

# head
head(assay(ddsMat_rlog))
```

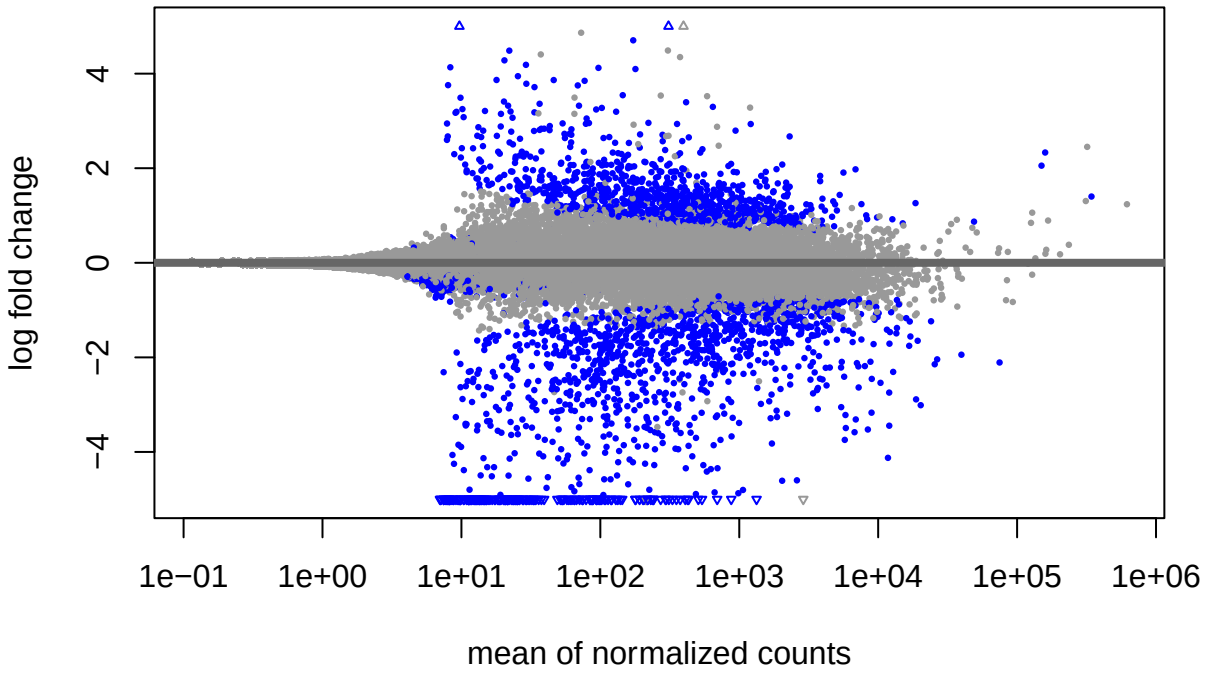
```
##                  sample37  sample38  sample39  sample40  sample41  sample42
## ENSG00000223972  0.000000  0.000000  0.000000  0.000000  0.000000  0.000000
## ENSG00000227232  6.210540  5.891545  6.701774  5.835019  6.038259  5.905677
## ENSG00000278267  3.095628  3.190564  3.618236  2.547785  2.671263  2.661103
## ENSG00000243485 -1.978997 -1.982317 -1.988425 -1.960549 -1.982833 -1.986540
## ENSG00000284332  0.000000  0.000000  0.000000  0.000000  0.000000  0.000000
## ENSG00000237613 -1.503528 -1.510520 -1.523153 -1.464670 -1.460163 -1.519416
```

```
# pca
plotPCA(ddsMat_rlog, intgroup = "Condition") +
  ggtitle(label = "Principal Component Analysis (PCA)") +
  scale_y_continuous(limits = c(-40, 40)) +
  scale_x_continuous(limits = c(-70, 40)) +
  theme_light()
```




```
# ma plot
# remove noise using apeglm
resultslfc <- lfcShrink(ddsMat, coef="Condition_normal_vs_disease", type="apeglm")

#  plot ma
plotMA(resultslfc, ylim=c(-5, 5))
```



Part 2: kallisto

import kallisto data

```
# create path to the abundance files
# sample names
samples <- paste0("sample", 37:42)

# file path
files <- file.path(".", samples, "abundance.h5")

# file names
names(files) <- paste0("sample", 37:42)

# import abundance data
txi.kallisto <- tximport(files, type = "kallisto", txOut = TRUE, ignoreAfterBar = TRUE)

# view first few lines of count data
head(txi.kallisto$counts)
```

```
##                   sample37 sample38 sample39  sample40  sample41 sample42
## ENST00000456328.2 10.24791     0.00   0.0000  3.598943  1.498626  0.00000
## ENST00000450305.2  0.00000     0.00   0.0000  0.000000  0.000000  0.00000
## ENST00000488147.1 55.65363    50.81 161.9119 45.050681 62.597347 64.58395
```

```
## ENST00000619216.1  0.00000     0.00   0.0000  0.000000  0.000000  0.00000
## ENST00000473358.1  0.00000     0.00   0.0000  0.000000  0.000000  0.00000
## ENST00000469289.1  0.00000     0.00   0.0000  0.000000  0.000000  0.00000
```

differential gene expression analysis

```
# create deseq2 data object
kallisto_dds <- DESeqDataSetFromTximport(txi.kallisto,
                                         colData = metadata,
                                         design = ~ Condition)

# Find differential expressed genes
kallisto_diff <- DESeq(kallisto_dds)

# obtain results
kallisto_results <- results(kallisto_diff)

# results
head(kallisto_results)
```

```
## log2 fold change (MLE): Condition normal vs disease
## Wald test p-value: Condition normal vs disease
## DataFrame with 6 rows and 6 columns
##                     baseMean log2FoldChange     lfcSE      stat    pvalue
##                    <numeric>      <numeric> <numeric> <numeric> <numeric>
## ENST00000456328.2   3.15905       1.465475  2.864086  0.511673  0.608880
## ENST00000450305.2   0.00000             NA        NA        NA        NA
## ENST00000488147.1  69.28047       0.448005  0.678288  0.660493  0.508937
## ENST00000619216.1   0.00000             NA        NA        NA        NA
## ENST00000473358.1   0.00000             NA        NA        NA        NA
## ENST00000469289.1   0.00000             NA        NA        NA        NA
##                        padj
##                   <numeric>
## ENST00000456328.2  0.850728
## ENST00000450305.2        NA
## ENST00000488147.1  0.795915
## ENST00000619216.1        NA
## ENST00000473358.1        NA
## ENST00000469289.1        NA
```

```
# summary
summary(kallisto_results)
```

```
##
## out of 156982 with nonzero total read count
## adjusted p-value < 0.1
## LFC > 0 (up)       : 3391, 2.2%
## LFC < 0 (down)     : 8425, 5.4%
## outliers [1]       : 1992, 1.3%
## low counts [2]     : 59420, 38%
## (mean count < 3)
## [1] see 'cooksCutoff' argument of ?results
## [2] see 'independentFiltering' argument of ?results
```

proceed as with hisat2