7659 HW5

Guannan Shen

October 22, 2018

Contents

```r
## set up workspace
library(knitr)
library(tidyverse)
library(RNASeqPower)
library(edgeR)
library(cqn)
library(EDASeq)
library(yeastRNASeq)
library(kableExtra)
options(stringsAsFactors = F)
options(dplyr.width = Inf)
getwd()
```

```
## [1] "/home/guanshim/Documents/Stats/CIDA_OMICs/7659Stats_Genetics/HW5"
```

```r
## not in function
"%nin%" <- Negate("%in%")
```

1 HW5

1.1 1. Next Generation Sequencing: Sample Size Estimates

1.1.1 (a) Using rnapower(), recreate Figure 3 from the journal club paper, Hart

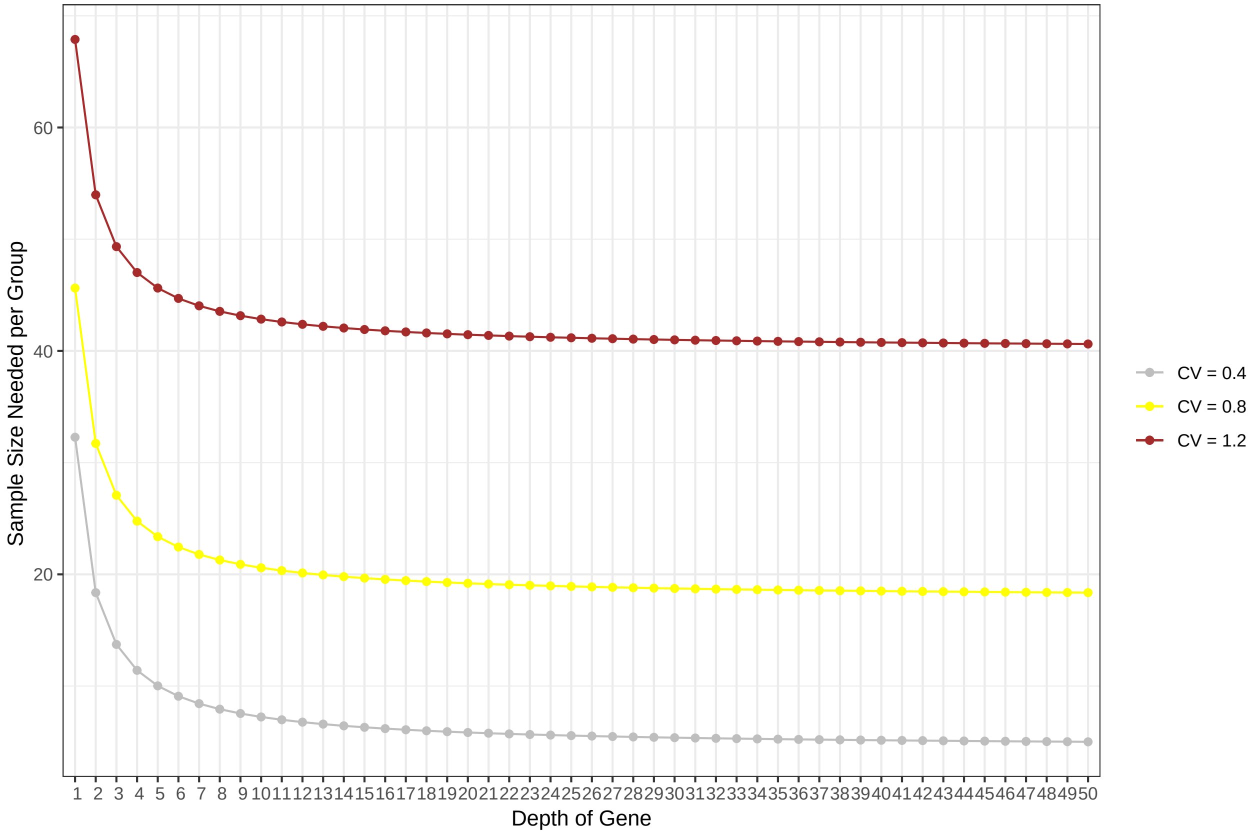
```r
## montgomery data from cqn
data(montgomery.subset)
## GC and gene length of montgomery
data(uCovar)
## vector of length 10 containing the number of mapped reads
## for each sample
data(sizeFactors.subset)

########## Understand the dataset ######## help(montgomery) number of
########## genes genes that have zero counts in all 10 samples were
########## already excluded
ng_mont <- nrow(montgomery.subset)

############## Question 1 figure 3 ############### sample size (ss) vs
############## depth sample size per group
```

```
ssize_depth <- sapply(c(0.4, 0.8, 1.2), function(y) {
    sapply(1:50, function(x) {
        rnapower(depth = x, cv = y, effect = 2.5, alpha = 0.01,
            power = 0.8)
    })
})
ssize_depth <- data.frame(ssize_depth)
colnames(ssize_depth) <- c("V1", "V2", "V3")

### Plot
ggplot(ssize_depth, aes(x = 1:50)) + geom_line(aes(y = V1, color = "CV = 0.4")) +
    geom_point(aes(y = V1, color = "CV = 0.4")) + geom_line(aes(y = V2,
    color = "CV = 0.8")) + geom_point(aes(y = V2, color = "CV = 0.8")) +
    geom_line(aes(y = V3, color = "CV = 1.2")) + geom_point(aes(y = V3,
    color = "CV = 1.2")) +
scale_x_discrete(name = "Depth of Gene", limits = c(1:50)) +
    scale_y_continuous(name = "Sample Size Needed per Group ") +
    theme_bw() + scale_colour_manual("", breaks = c("CV = 0.4",
    "CV = 0.8", "CV = 1.2"), values = c(`CV = 0.4` = "grey",
    `CV = 0.8` = "yellow", `CV = 1.2` = "brown"))
```



```
########## average of sequence reads aligning to the gene/ depth
########## ############ how many reads are assigned to a particular
########## gene / depth ## is a data frame with 23552 observations on
########## 10 different samples ##
N_total <- sum(sizeFactors.subset)
## number of genes genes that have zero counts in all 10
```

```r
## samples were already excluded
ng_mont
```

```
## [1] 23552
```

```r
counts_gene_million <- rowSums(montgomery.subset)/N_total * 1e+06

mont_counts <- data.frame(Sample = "Montgomery", n = 10, Reads = round(N_total/(ng_mont *
    10), 2), mapped = "100%", a = round(sum(counts_gene_million <
    0.01)/ng_mont, 2), b = round(sum(0.01 <= counts_gene_million &
    counts_gene_million < 0.1)/ng_mont, 2), c = round(sum(0.1 <=
    counts_gene_million & counts_gene_million < 1)/ng_mont, 2),
    d = round(sum(1 <= counts_gene_million & counts_gene_million <
        10)/ng_mont, 2), e = round(sum(10 <= counts_gene_million &
        counts_gene_million < 100)/ng_mont, 2), f = round(sum(100 <=
        counts_gene_million & counts_gene_million < 1000)/ng_mont,
        2), g = round(sum(1000 <= counts_gene_million)/ng_mont,
        2))
colnames(mont_counts) <- c("Sample", "n", "Avg Reads", "% mapped",
    "<0.01", "0.01-0.1", "0.1-1", "1-10", "10-100", "100-1000",
    ">1000")
kable(mont_counts)
```

Sample	n	Avg Reads	% mapped	<0.01	0.01–0.1	0.1–1	1–10	10-100	100-1000	>1000
Montgomery	10	164.66	100%	0	0.23	0.22	0.18	0.27	0.09	0