Sample Size Calculation for RNA-Seq Experimental Design – the ssizeRNA package

Ran Bi, Peng Liu

Department of Statistics, Iowa State Univeresity

February 11, 2016

Abstract

Sample size calculation is a crucial issue when designing an RNA-seq experiment. This vignette explains the use of the package ssizeRNA, which is designed to provide an estimation of sample size while controlling false discovery rate (FDR) for RNA-seq experimental design.

ssizeRNA version: 1.2.4

Contents

1 Introduction

RNA-seq technologies have been popularly applied in transcriptomic studies. In the statistical analysis of RNA-seq data, identifying differentially expressed (DE) genes across treatments or conditions is a major step or main focus. Many statistical methods have been proposed for the detection of DE genes with RNA-seq data, such as *edgeR* [1], *DESeq* [2], *DESeq2* [3] and *QuasiSeq* [4].

Due to the genetic complexity, RNA-seq experiments are rather costly. Many experiments only employ a small number of replicates, which may lead to unreliable statistical inference. Thus, one of the principal questions in designing an RNA-seq experiment is: how large of the sample size do we need?

Many of the current sample size calculation methods are simulation based, which are quite time-consuming. We propose a much less computationally intensive method and R package *ssizeRNA* for sample size calculation in designing RNA-seq experiments [5].

2 Using `ssizeRNA`

We first load the *ssizeRNA* package.

```
library(ssizeRNA)
```

To determine the sample size for an RNA-seq experiment, users need to specify the following parameters:

- G: total number of genes for testing;
- $pi0$: proportion of non-DE genes;
- fdr: FDR level to control;
- $power$: desired average power to achieve;
- mu: average read count for each gene in control group (without loss of generality, we assume that the normalization factors are equal to 1 for all samples);
- $disp$: dispersion parameter for each gene;
- $logfc$: log fold change for each gene.

We will give several examples of using *ssizeRNA* sample size estimation as follows.

2.1 Sample size calculation for a single set of parameter

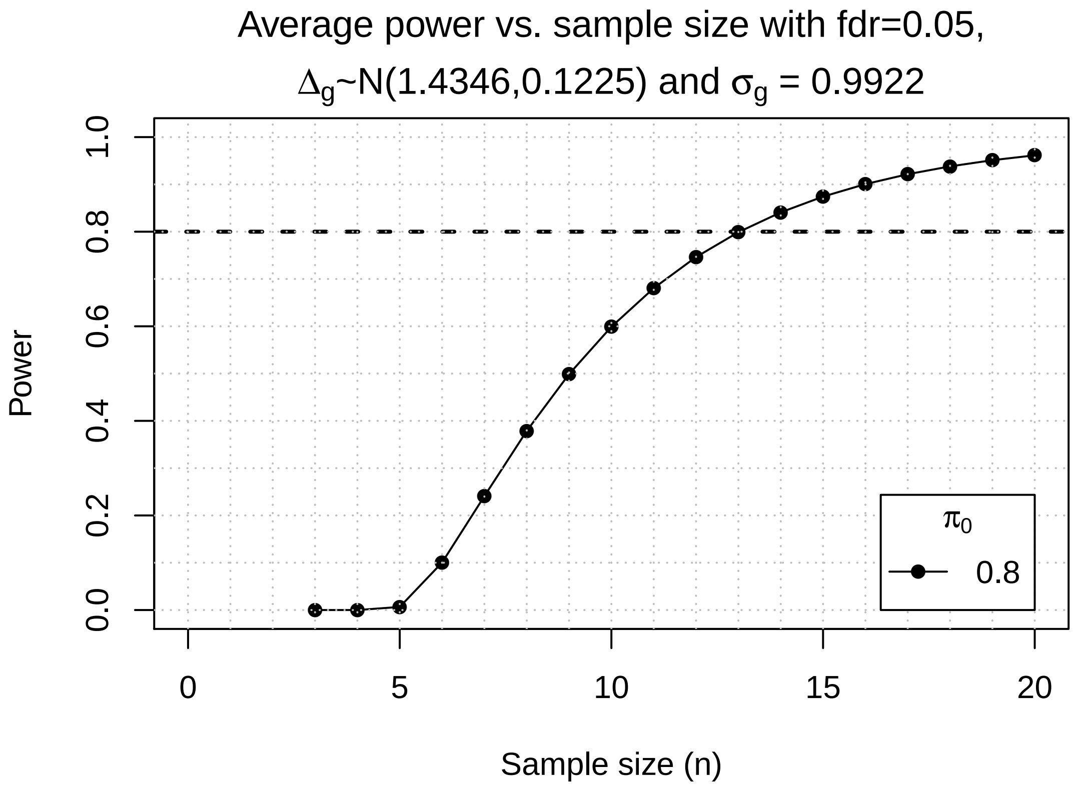
Here we consider the situation of single set of parameter, i.e. all genes share the same set of average read count in control group, dispersion parameter, and log fold change.

For example, if we are estimating the sample size for an RNA-seq experiment with

- Total number of genes: $G = 10000$;
- Proportion of non-DE genes: $pi0 = 0.8$;
- FDR level to control: $fdr = 0.05$;
- Desired average power to achieve: $power = 0.8$;
- Average read count for each gene in control group: $mu = 10$;
- Dispersion parameter for each gene: $disp = 0.1$;
- Log fold change for each gene: $logfc = log(2)$.

The estimated sample size is 14 with anticipated power 0.84 by *ssizeRNA_single* function. The function also gives the power vs. sample size curve estimated by our method.

```
set.seed(2016)
size1 <- ssizeRNA_single(nGenes = 10000, pi0 = 0.8, m = 200, mu = 10,
                         disp = 0.1, logfc = log(2), fdr = 0.05,
                         power = 0.8, maxN = 20)
```



```
size1$ssize

##      pi0 ssize    power
## [1,] 0.8    14 0.840891
```

To check whether desired power would be achieved at the calculated sample size 14 for voom and limma pipeline [6, 7], we could use the *check.power* function, which gives the observed power and true FDR by Benjamini and Hochberg's method [8] and Storey's q-value procedure [9] respectively. The results below are based on 10 simulations, indicating that desired power is achieved and FDR is controlled successfully.

```
check.power(m = 14, mu = 10, disp = 0.1, logfc = log(2), sims = 10)

## $pow_bh_ave
## [1] 0.86235
##
## $fdr_bh_ave
## [1] 0.03496079
##
## $pow_qvalue_ave
## [1] 0.88335
##
## $fdr_qvalue_ave
## [1] 0.04628789
```

2.2 Sample size calculation for gene-specific mean and dispersion with fixed log fold change

Now we will give an example of sample size calculation for gene-specific mean and dispersion. Here we use the real RNA-seq dataset from Hammer, P. et al., 2010 [10] to generate gene-specific mean and dispersion parameters.

```
data(hammer.eset)
counts <- exprs(hammer.eset)[, phenoData(hammer.eset)$Time == "2 weeks"]
counts <- counts[rowSums(counts) > 0,]   ## filter zero count genes
trt <- hammer.eset$protocol[which(hammer.eset$Time == "2 weeks")]

## average read count in control group for each gene
mu <- apply(counts[, trt == "control"], 1, mean)

## dispersion for each gene
d <- DGEList(counts)
d <- calcNormFactors(d)
```

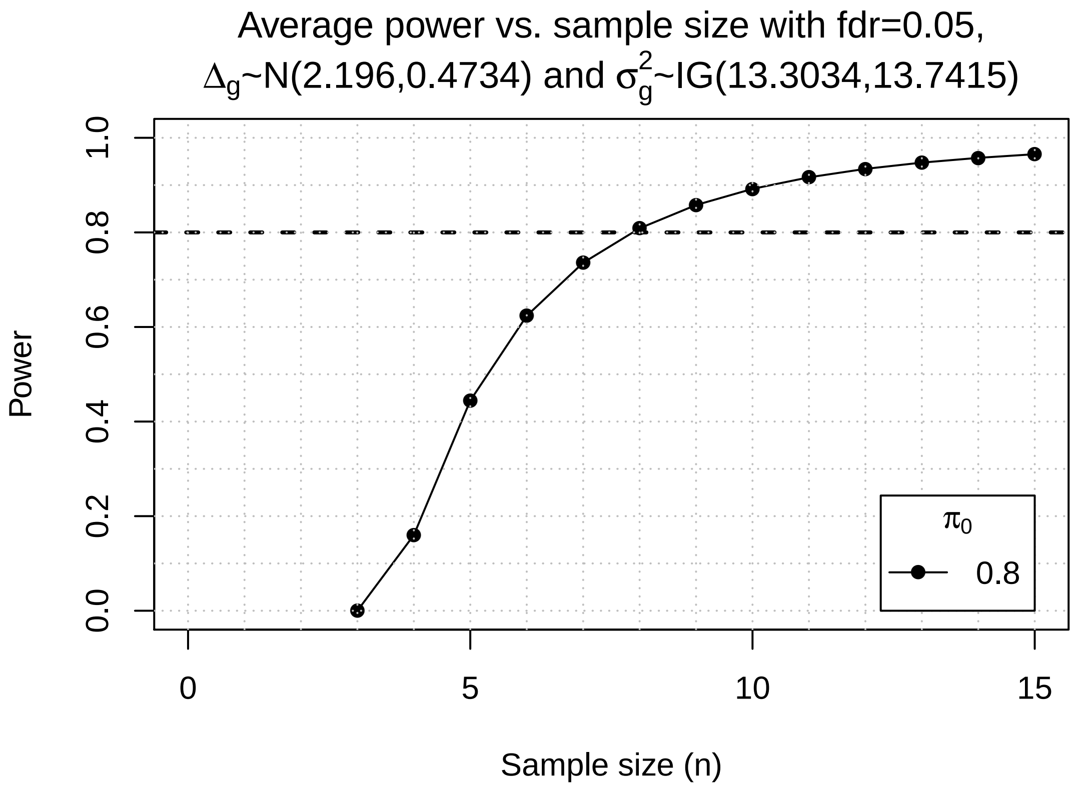
```
d <- estimateCommonDisp(d)
d <- estimateTagwiseDisp(d)
disp <- d$tagwise.dispersion
```

If we would like to estimate the sample size for the above RNA-seq experiment with

- Total number of genes: $G = 10000$;
- Proportion of non-DE genes: $pi0 = 0.8$;
- FDR level to control: $fdr = 0.05$;
- Desired average power to achieve: $power = 0.8$;
- Log fold change for each gene: $logfc = log(2)$.

The estimated sample size is 8 with anticipated power 0.81 by *ssizeRNA_vary* function. The function also gives the power vs. sample size curve estimated by our method.

```
set.seed(2016)
size2 <- ssizeRNA_vary(nGenes = 10000, pi0 = 0.8, m = 200, mu = mu,
                       disp = disp, logfc = log(2), fdr = 0.05,
                       power = 0.8, maxN = 15, replace = FALSE)
```



```
size2$ssize

##      pi0 ssize     power
## [1,] 0.8     8 0.8086813
```

The observed power and true FDR by Benjamini and Hochberg's method and Storey's q-value procedure could also be checked by the *check.power* function.

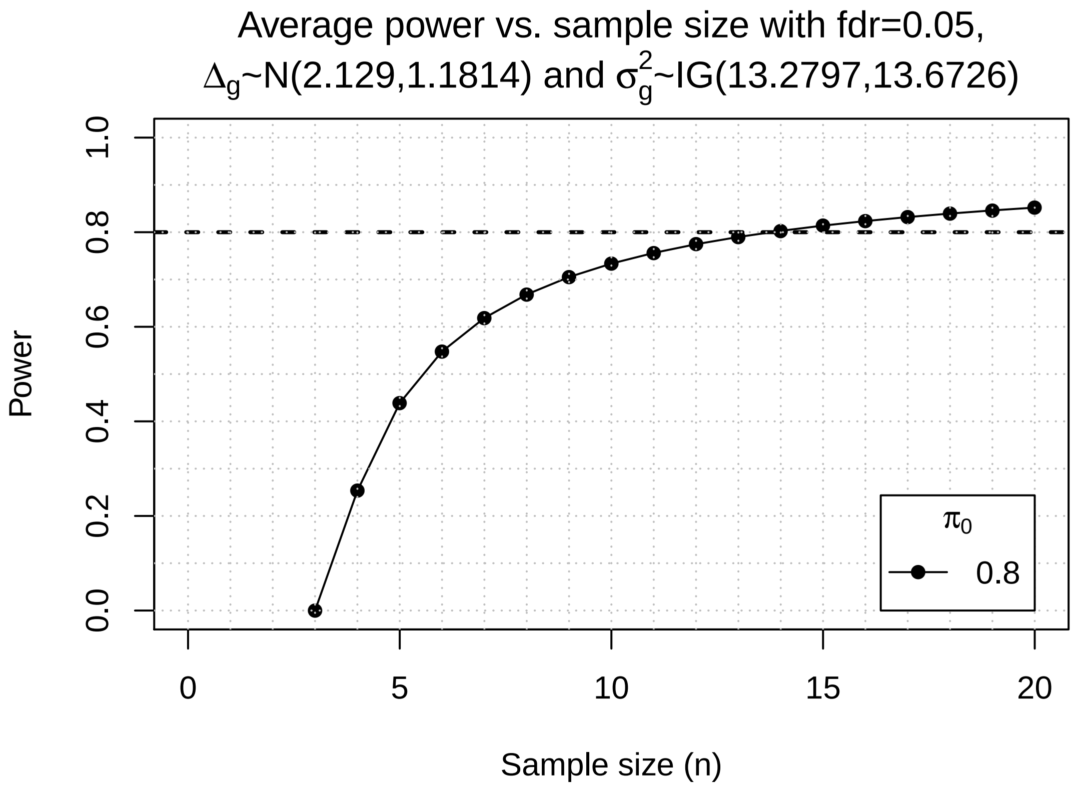
2.3 Sample size calculation for gene-specific mean and dispersion with different log fold change

If not all genes share the same log fold change, for example, if log fold change comes from a normal distribution,

$$logfc \sim Normal(log(2), 0.5 * log(2))$$

other parameters remain the same as in subsection 2.2, then the estimated sample size is 14 with anticipated power 0.80 by *ssizeRNA_vary* function.

```
set.seed(2016)
logfc <- function(x){rnorm(x, log(2), 0.5*log(2))}
size3 <- ssizeRNA_vary(nGenes = 10000, pi0 = 0.8, m = 200, mu = mu,
                       disp = disp, logfc = logfc, fdr = 0.05,
                       power = 0.8, maxN = 20, replace = FALSE)
```



```
size3$ssize

##      pi0 ssize     power
## [1,] 0.8    14 0.8028241
```

By the following command, we verified that the desired power 0.8 is achieved at the calculated sample size 14 for voom and limma pipeline.

```
check.power(m = 14, mu = mu, disp = disp, logfc = logfc, sims = 10,
            replace = FALSE)

## $pow_bh_ave
## [1] 0.82995
##
## $fdr_bh_ave
## [1] 0.03722581
##
## $pow_qvalue_ave
## [1] 0.8375
##
```

```
## $fdr_qvalue_ave
## [1] 0.04590851
```

3 Conclusion

ssizeRNA provides a quick calculation for sample size, and an accurate estimate of power. Examples in section 2 demonstrate that our proposed method offers a reliable approach for sample size calculation for RNA-seq experiments.

4 Session Info

```
sessionInfo()
```

```
## R version 3.2.0 (2015-04-16)
## Platform: x86_64-apple-darwin14.3.0 (64-bit)
## Running under: OS X 10.10.5 (Yosemite)
##
## locale:
## [1] C/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
## [1] parallel  stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] ssizeRNA_1.2.4      edgeR_3.12.0        limma_3.26.7        Biobase_2.30.0
## [5] BiocGenerics_0.16.1 knitr_1.12.3
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_0.12.3      MASS_7.3-45      plyr_1.8.3       grid_3.2.0       gtable_0.1.2
##  [6] formatR_1.2.1    magrittr_1.5     scales_0.3.0     evaluate_0.8     highr_0.5.1
## [11] ggplot2_2.0.0    stringi_1.0-1    reshape2_1.4.1   qvalue_2.2.2     splines_3.2.0
## [16] BiocStyle_1.8.0  tools_3.2.0      stringr_1.0.0    munsell_0.4.2    colorspace_1.2-
## [21] ssize.fdr_1.2
```

References

[1] Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2010.

[2] Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Genome biol*, 11(10):R106, 2010.

[3] Michael Love, Simon Anders, and Wolfgang Huber. Differential analysis of count data–the deseq2 package. *Genome Biology*, 15:550, 2014.

[4] S Lund, Dan Nettleton, D McCarthy, G Smyth, et al. Detecting differential expression in rna-sequence data using quasi-likelihood with shrunken dispersion estimates. *Statistical applications in genetics and molecular biology*, 11(5):8, 2012.

[5] Ran Bi and Peng Liu. Sample size calculation while controlling false discovery rate for differential expression analysis with rna-sequencing experiments. *In Revision*.

[6] Charity W Law, Yunshun Chen, Wei Shi, and Gordon K Smyth. Voom: precision weights unlock linear model analysis tools for rna-seq read counts. *Genome Biol*, 15(2):R29, 2014.

[7] Gordon K Smyth. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical applications in genetics and molecular biology*, 3(1), 2004.

[8] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300, 1995.

[9] John D Storey, Jonathan E Taylor, and David Siegmund. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(1):187–205, 2004.

[10] Paul Hammer, Michaela S Banck, Ronny Amberg, Cheng Wang, Gabriele Petznick, Shujun Luo, Irina Khrebtukova, Gary P Schroth, Peter Beyerlein, and Andreas S Beutler. mrna-seq with agnostic splice site discovery for nervous system transcriptomics tested in chronic pain. *Genome research*, 20(6):847–860, 2010.