Power & Sample Size assessment for differential expression analysis of RNA-Seq experiments using PROPER

Jessica Randall

Briefly, PROspective Power Estimation for RNA-Seq (PROPER) simulates experimental conditions to assess adequete sample size and statistical power for differential expression (DE) analysis using two-group comparison. RNA-seq experimental designs require consideration of the desired False Discovery Rate (FDR), the choice of sequencing depth, and the procedure used to determine differential expression. PROPER accounts for all of these considerations and even provides pretty graphs to explain it.

Linked is the original 2014 paper from Wu and Wang introducing the concepts implemented in PROPER.

Definition of terms

0.0.0.1 * Unadjusted p values vs adjusted p-values/(FDR):

In DE analysis, a single p-value tells you how likely it is that a single gene is differential expressed between at least two groups (ex: a control and a treatment group) due to some actual difference between the groups as opposed to random chance. False Discovery Rate (FDR) tells you how likely it is that all genes identified as DE are false positives. A FDR of 5% means that among all genes called DE, an average of 5% of those are truly not DE. DE genes are only considered significantly so if they meet the adjusted p value, not only the unadjusted p-value. FDRs for each individual gene are called q-values or local FDRs.

Generating data

Our very first step is to load the libraries we'll need to assess the functions required for analysis and graphing. Please see Bioconductor for information about initial installation and use of Bioconductor and its packages.

Our scenario is based on the following simplified experimental conditions:

- Assuming 5% of genes are DE

The validity of this assumption depends on your specific experiment, this is just a guess for the sake of the example.

- Testing 24000 genes

This is just a nice round number similar to the numbers of genes we typically see in our RNA-Seq projects at EICC.

- Using the Cheung database

The Cheung database best simulates the inherent biological variation between unique, unrelated individuals. This will provide estimates of sample size assuming the greatest amount of biological variation. If your experiment has samples which are more similar to each other than unique, unrelated individuals, we would select a different database to better simulate the biological variation.

- 50M read sequencing depth

This is the sequencing depth we typically recommend but your mileage may vary depending on how different you expect your groups to be.

DE Gene Detection

Next, we run the simulations to detect DE genes. We chose the minimum recommended 20 simulations. This is to balance the accuracy with the time this portion of the analysis takes. We have included the time stamp to show how long even this simplified example scenario can take at the minimum number of simulations. More simulations may lead to more accurate estimates with the tradeoff being that this step takes longer. This is something we can test on a by-project basis.

We have chosen to simulate power for samples of 3,5,7,and 10 samples per treatment group (i.e. 10 cases and 10 controls) using DESeq2 since this is the tool we use most commonly for DE gene analysis. If you would like a power and sample size assessment from EICC, the choice of simulated sample sizes and the choice of DE analysis package are ones we would tailor to your specific experiment.

```
start_time <- Sys.time()

simulations <-{runSims(Nreps = c(3,5,7,10),
                       sim.opts=ourscenario,
                       DEmethod ="DESeq2",
                       nsims=20) }
```

```
## Simulation number 1
## Simulation number 2
## Simulation number 3
## Simulation number 4
## Simulation number 5
## Simulation number 6
## Simulation number 7
## Simulation number 8
## Simulation number 9
## Simulation number 10
## Simulation number 11
## Simulation number 12
## Simulation number 13
## Simulation number 14
## Simulation number 15
## Simulation number 16
## Simulation number 17
## Simulation number 18
## Simulation number 19
## Simulation number 20
```

```
end_time <- Sys.time()

end_time - start_time
```

```
## Time difference of 6.251048 mins
```

Power Assessment

Here we parameterize our power analysis. The chosen parameters will vary by experiment and often include a certain amount of filtering of lowly expressed genes since this is a common step. If you do a power assessment with EICC we will discuss the appropriate parameters for your project.

```
power <- comparePower(simulations,
                      alpha.type = "fdr",
                      alpha.nominal = 0.05,
                      strata = c(0, 10, 2^(1:7)*10, Inf),
                      filter.by = "expr",
                      strata.filtered = 1,
                      stratify.by = "expr",
                      delta = 1)

summaryPower(power)
```

```
##      SS1 SS2 Nominal FDR Actual FDR Marginal power Avg # of TD Avg # of FD  FDC
## [1,]   3   3        0.05       0.37           0.71          98          68 0.69
## [2,]   5   5        0.05       0.24           0.82         110          44 0.39
## [3,]   7   7        0.05       0.17           0.87         120          33 0.27
## [4,]  10  10        0.05       0.14           0.92         130          28 0.22
```

Given the experimental conditions of our scenario and assumptions we made about the percentae of DE genes, we would need to have 5 samples per experimental condition for a total of 10 samples in order to accurately detect 80% of DE genes with a standardized log fold change of 1.

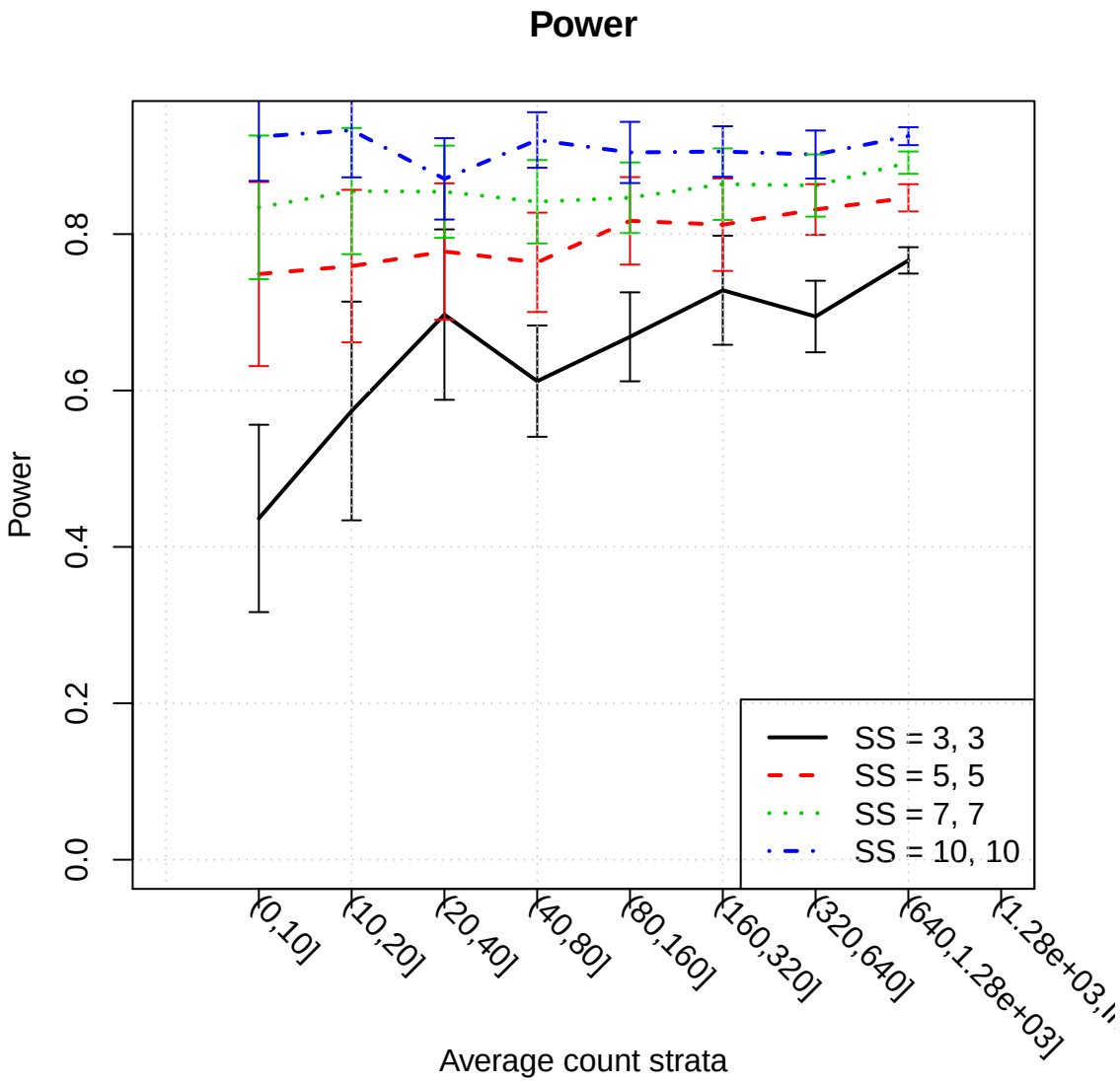
Notes on sequencing depth

Do we really need more people or is greater resolution (i.e. sequencing depth) the answer? This is something that PROPER can assess and a question EICC would be happy to discuss with you.

Visualizing

PROPER provides us with many options for visualizing the results of our analysis with publication-ready graphs and straight forward interpretations. Graphs could include the power, the number of true and false discoveries we could expect to find, estimated FDR, and false positive rate given various sample sizes along with customized graphs for your specific project.

Here is the power graph from our simplified example scenario. As the green line indicates, sample sizes of 5 samples (red line) reach over 80% power, the minimum standard for statistical practice, at higher mean count strata but sample sizes of 7 or 10 per group would likely be better to ensoure power was sufficiently high across all count strata.



We look forward to working with you to assess the necessary sample size for your project. For a more specific application and interpretation of these or additional tools and visualizations for your own data, please find our contact information on our website and examples of some previous graphs here

Session information and References

```
## [1] "Thu Mar 26 07:18:53 2020"
```

```
## R version 3.6.3 (2020-02-29)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Debian GNU/Linux 9 (stretch)
##
## Matrix products: default
## BLAS:   /usr/lib/openblas-base/libblas.so.3
## LAPACK: /usr/lib/libopenblasp-r0.2.19.so
##
```

```
## locale:
##  [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
##  [3] LC_TIME=en_US.UTF-8        LC_COLLATE=en_US.UTF-8
##  [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
##  [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
##  [9] LC_ADDRESS=C               LC_TELEPHONE=C
## [11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C
##
## attached base packages:
## [1] parallel  stats4    stats     graphics  grDevices utils     datasets
## [8] methods   base
##
## other attached packages:
##  [1] DESeq2_1.26.0               SummarizedExperiment_1.16.1
##  [3] DelayedArray_0.12.2         BiocParallel_1.20.1
##  [5] matrixStats_0.56.0          Biobase_2.46.0
##  [7] GenomicRanges_1.38.0        GenomeInfoDb_1.22.0
##  [9] IRanges_2.20.2              S4Vectors_0.24.3
## [11] BiocGenerics_0.32.0         PROPER_1.18.0
##
## loaded via a namespace (and not attached):
##  [1] bit64_0.9-7            splines_3.6.3          Formula_1.2-3
##  [4] assertthat_0.2.1       latticeExtra_0.6-29    blob_1.2.1
##  [7] GenomeInfoDbData_1.2.2 yaml_2.2.1             RSQLite_2.2.0
## [10] pillar_1.4.3           backports_1.1.5        lattice_0.20-40
## [13] glue_1.3.2             digest_0.6.25          RColorBrewer_1.1-2
## [16] XVector_0.26.0         checkmate_2.0.0        colorspace_1.4-1
## [19] htmltools_0.4.0        Matrix_1.2-18          XML_3.99-0.3
## [22] pkgconfig_2.0.3        genefilter_1.68.0      bookdown_0.18
## [25] zlibbioc_1.32.0        purrr_0.3.3            xtable_1.8-4
## [28] scales_1.1.0           jpeg_0.1-8.1           tibble_2.1.3
## [31] htmlTable_1.13.3       annotate_1.64.0        ggplot2_3.3.0
## [34] pacman_0.5.1           nnet_7.3-13            survival_3.1-11
## [37] magrittr_1.5           crayon_1.3.4           memoise_1.1.0
## [40] evaluate_0.14          foreign_0.8-76         tools_3.6.3
## [43] data.table_1.12.8      lifecycle_0.2.0        stringr_1.4.0
## [46] locfit_1.5-9.1         munsell_0.5.0          cluster_2.1.0
## [49] AnnotationDbi_1.48.0   compiler_3.6.3         rlang_0.4.5
## [52] grid_3.6.3             RCurl_1.98-1.1         rstudioapi_0.11
## [55] htmlwidgets_1.5.1      bitops_1.0-6           base64enc_0.1-3
## [58] rmarkdown_2.1          gtable_0.3.0           DBI_1.1.0
## [61] R6_2.4.1               gridExtra_2.3          knitr_1.28
## [64] dplyr_0.8.5            bit_1.1-15.2           Hmisc_4.3-1
## [67] stringi_1.4.6          Rcpp_1.0.3             geneplotter_1.64.0
## [70] vctrs_0.2.4            rpart_4.1-15           acepack_1.4.1
## [73] png_0.1-7              tidyselect_1.0.0       xfun_0.12


##
##   Hao Wu, Chi Wang, Zhijin Wu (2014): PROPER: Comprehensive Power
##   Evaluation for Differential Expression using RNA-seq. Bioinformatics.
##   doi:10.1093/bioinformatics/btu640
##
## A BibTeX entry for LaTeX users is
##
```

```
##   @Article{,
##     title = {PROPER: Comprehensive Power Evaluation for Differential Expression using RNA-seq},
##     author = {Hao Wu and Chi Wang and Zhijin Wu},
##     year = {2014},
##     journal = {Bioinformatics},
##     doi = {10.1093/bioinformatics/btu640},
##   }
```