superSeq: Assessing the limits of sequencing depth through read subsampling

Andrew J. Bass, David G. Robinson, and John D. Storey
Princeton University

Abstract

RNA-Seq is a standard gene expression profiling technology for differential expression analysis. In RNA-Seq studies, the read depth strongly affects the power of the test statistics, such that larger read depths induce higher statistical power. After a certain read depth, the power of the test statistics begins to asymptote, at which point there are only marginal improvements in power. Although existing methods, such as subSeq, can help determine if the read depth of an experiment is saturated, they are limited in that they do not provide a way of estimating the appropriate read depth for under-saturated experiments. We provide a new method called *superSeq* that models and estimates the increase in statistical power that would result in increasing the read depth for a given experiment. We then apply the superSeq framework to 38 RNA-Seq experiments in the Expression Atlas. In the majority of the studies, the method accurately predicts the relationship between the power of the test statistics and the read depth. Researchers can thus use this method, implemented in the forthcoming R package `superSeq`, to determine the appropriate read depth for a completed experiment in order to maximize the statistical power.

Background

Motivation

It has been previously shown that increasing the read depth in RNA-Seq studies increases both the accuracy and the power of the study. Understanding the role of appropriate read depth is important when designing studies, especially when considering the tradeoff between available resources and experimental design. Previous methods approach this problem by randomly subsampling the reads from either the fastq or alignment file, and then perform the analysis on each subsample, but this is a computationally slow approach. Recent work by Robinson and Storey (2014) suggests a computationally efficient approach to subsampling by dealing directly with the count matrix.

subSeq Approach

subSeq allows users to perform subsampling on an unnormalized *I x N* matrix ***X*** of read counts:

- The user inputs a vector of subsampling proportions, p.
- For each k subsample proportions,

$$X^{k}_{(i,n)} \sim Binom(X_{(i,n)}, p_k)$$

- The analysis is run on each subsample using user chosen available methods for RNA-Seq differential expression.

Example

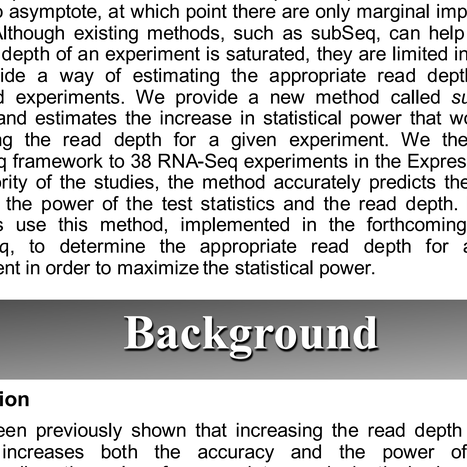

Figure 1: The default plot generated by subSeq on subsamples from the Hammer et al. (2010) data. This shows the number of significant genes at each depth (top left), the estimated FDP (top right), the Spearman correlation (bottom left), and mean-squared error (bottom right) comparing the estimates at each depth with the full experiment.

Model

$$log(D_i + b) \sim N(\mu, \sigma)$$

$$G(p) = \Phi\left(\frac{log(p+b) - \mu}{\sigma}\right)$$

- D_i : Per-gene read depth for gene i
- p : subsambling proportion
- b : offset term to make per-gene read depths approximately normal
- μ : mean of the per-gene read depth distribution
- σ: standard deviation of the per-gene read depth distribution
- $G(p)$: Total significant genes at p

Results

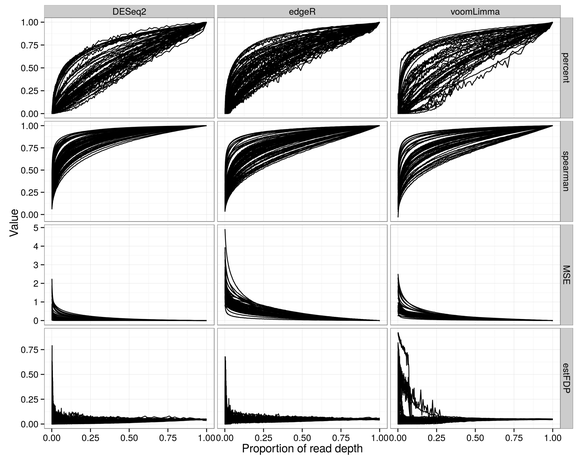

Figure 2: Metrics showing power and accuracy across varying read depths for 38 differential expression experiments (all those with more than 100 significant genes at their full depth), using DESeq2, edgeR and voom to detect significance. Each value was averaged across seven replicates.

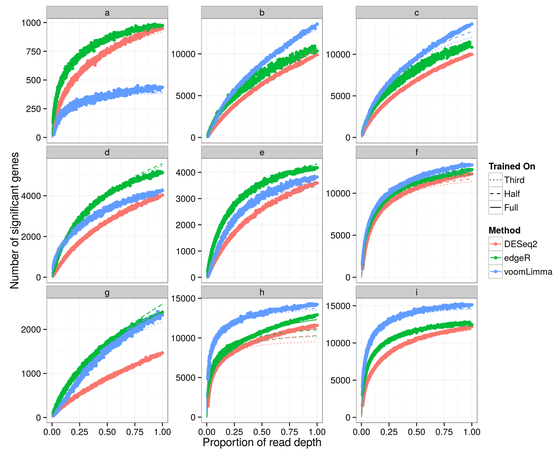

Figure 3: The number of significant genes at each depth, along with predicted parametric curves for each experiment/contrast individually. Each fit was performed three times: on the full experiment and on data that had previously been sampled to one half or one third of the depth, to determine how well the fit could predict the saturation curve at greater depths.

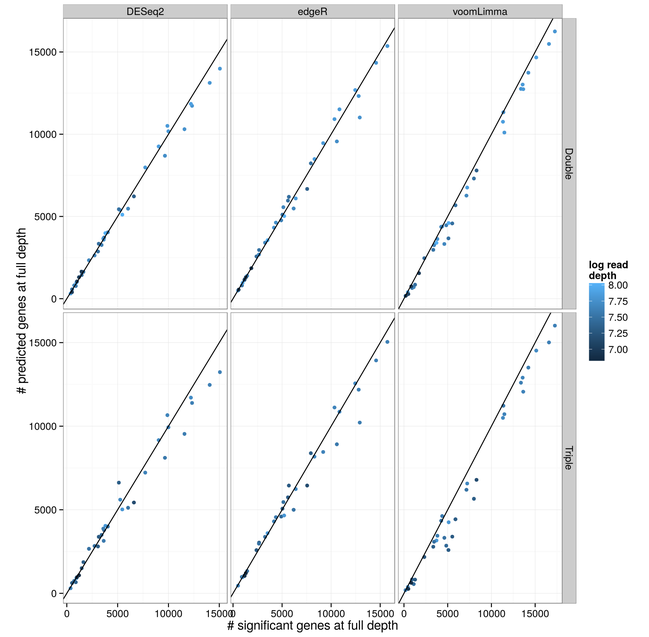

Figure 4: Accuracy of the predicted number of significant genes from double or tripling read depth. Our model is capturing the power trend when the read depth is increasing. In particular, edgeR and DESeq2 predictions are accurate while voom is generally conservative.

Conclusion

Using RNA-Seq studies from the Expression Atlas, we find that our method accurately predicts the effect of double or tripling read depth when using DESeq2, edgeR, and to a lesser extent, voom. We recommend using all three methods when determining an appropriate read depth. The superSeq method is implemented in a forthcoming R package called `superSeq`.

Acknowledgements

This work was supported in part by NIH grant R01 HG002913.

References

- Hammer, P. *et al.* (2010) mRNA-seq with agnostic splice site discovery for nervous system transcriptomics tested in chronic pain. *Genome Research*, 20, 847–860.
- Petryszak, R. *et al.* (2013) Expression Atlas update - a database of gene and transcript expression from microarray and sequencing-based functional genomics experiments. *Nucleic Acids Research*, 42, 10.1093/nar/gkt1270.
- Robinson, D.G. and Storey, J.D. (2014) subSeq: Determining appropriate sequencing depth through efficient read subsampling. *Bioinformatics*, 30, 3424–3426.

https://github.com/StoreyLab/

CENTER FOR STATISTICS AND MACHINE LEARNING