Power analysis and sample size estimation for RNA-Seq differential expression

TRAVERS CHING,[1,2] SIJIA HUANG,[1,2] and LANA X. GARMIRE[1,2]

[1]University of Hawaii Cancer Center, Honolulu, Hawaii 96813, USA
[2]Graduate Program of Molecular Biosciences and Bioengineering, University of Hawaii–Manoa, Honolulu, Hawaii 96822, USA

ABSTRACT

It is crucial for researchers to optimize RNA-seq experimental designs for differential expression detection. Currently, the field lacks general methods to estimate power and sample size for RNA-Seq in complex experimental designs, under the assumption of the negative binomial distribution. We simulate RNA-Seq count data based on parameters estimated from six widely different public data sets (including cell line comparison, tissue comparison, and cancer data sets) and calculate the statistical power in paired and unpaired sample experiments. We comprehensively compare five differential expression analysis packages (DESeq, edgeR, DESeq2, sSeq, and EBSeq) and evaluate their performance by power, receiver operator characteristic (ROC) curves, and other metrics including areas under the curve (AUC), Matthews correlation coefficient (MCC), and *F*-measures. DESeq2 and edgeR tend to give the best performance in general. Increasing sample size or sequencing depth increases power; however, increasing sample size is more potent than sequencing depth to increase power, especially when the sequencing depth reaches 20 million reads. Long intergenic noncoding RNAs (lincRNA) yields lower power relative to the protein coding mRNAs, given their lower expression level in the same RNA-Seq experiment. On the other hand, paired-sample RNA-Seq significantly enhances the statistical power, confirming the importance of considering the multifactor experimental design. Finally, a local optimal power is achievable for a given budget constraint, and the dominant contributing factor is sample size rather than the sequencing depth. In conclusion, we provide a power analysis tool (http://www2.hawaii.edu/~lgarmire/RNASeqPowerCalculator.htm) that captures the dispersion in the data and can serve as a practical reference under the budget constraint of RNA-Seq experiments.

Keywords: RNA-Seq; sample size; power analysis; simulation; bioinformatics