BIOINFORMATICS APPLICATIONS NOTE Vol. 30 no. 23 2014, pages 3424–3426
doi:10.1093/bioinformatics/btu552

Gene expression

Advance Access publication September 3, 2014

subSeq: Determining Appropriate Sequencing Depth Through Efficient Read Subsampling

David G. Robinson[1,*] and John D. Storey[1,2,*]

[1]Lewis-Sigler Institute for Integrative Genomics and [2]Department of Molecular Biology, Princeton University, Princeton, NJ 08544, USA

Associate Editor: Inanc Birol

ABSTRACT

Motivation: Next-generation sequencing experiments, such as RNA-Seq, play an increasingly important role in biological research. One complication is that the power and accuracy of such experiments depend substantially on the number of reads sequenced, so it is important and challenging to determine the optimal read depth for an experiment or to verify whether one has adequate depth in an existing experiment.

Results: By randomly sampling lower depths from a sequencing experiment and determining where the saturation of power and accuracy occurs, one can determine what the most useful depth should be for future experiments, and furthermore, confirm whether an existing experiment had sufficient depth to justify its conclusions. We introduce the subSeq R package, which uses a novel efficient approach to perform this subsampling and to calculate informative metrics at each depth.

Availability and Implementation: The subSeq R package is available at http://github.com/StoreyLab/subSeq/.

Contact: dgrtwo@princeton.edu or jstorey@princeton.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on March 26, 2014; revised on July 26, 2014; accepted on August 11, 2014

1 INTRODUCTION

Many next-generation sequencing technologies have been developed to answer important biological questions. One property these technologies have in common is that they depend on read depth or coverage: increasing the number of reads typically increases the power and accuracy. For instance, in RNA-Seq greater read depth is known to increase the power of differential expression testing and the accuracy of expression estimates (Liu *et al.*, 2013; Tarazona *et al.*, 2011). The advent of multiplexed sequencing means that researchers should consider their read depth as a trade-off against cost and replication when designing experiments (Liu *et al.*, 2014), which means knowing the relationship between read depth and power is essential to designing sequencing experiments. Similarly, many researchers need to demonstrate that they have adequate depth in an existing experiment to support their biological conclusions.

One valuable approach that multiple studies have used is to randomly subsample reads (sometimes called downsampling) and perform an identical analysis on each subsample. This is in contrast to methods that fit a parametric model to calculate power, such as Scotty (Busby *et al.*, 2013). By determining where metrics of power and accuracy 'saturate' with increasing depth, one can both determine recommendations for future experiments and demonstrate whether an existing experiment has sufficient depth. Studies have used random subsampling to propose guidelines for future experiments (Black *et al.*, 2014; Liu *et al.*, 2014), to perform a survey of different RNA-Seq analysis methods at varying read depths (Labaj *et al.*, 2011; Liu *et al.*, 2013; Rapaport *et al.*, 2013), or to demonstrate that they had achieved adequate read depth (Daines *et al.*, 2011; Toung *et al.*, 2011; Wang *et al.*, 2011). However, all took the approach of randomly subsampling from either the fastq or alignment file, and then reperforming the analysis, including the computationally intensive step of matching reads to genes, on each file. This process is slow, demanding of disk space, and requires possessing the original reads or mappings, which limits the number of subsamples that can be performed and the ease of performing this analysis on existing experiments.

We introduce the subSeq R package, which instead subsamples sequencing reads with binomial sampling *after* they have been matched to genes and assembled into a count matrix. Because the step of matching reads to genes is independent and deterministic, this approach is functionally identical to the common approach of subsampling the read alignment files, but requires only the count matrix rather than the read alignment file. It also takes negligible time and computing resources even on large datasets, as the steps downstream of the read subsampling are much faster than the upstream steps. A similar approach is used to generate saturation figures in the NOISeq package (Tarazona *et al.*, 2011), but subSeq is designed to be used with any RNA-Seq analysis method. subSeq could be performed immediately on any experiment in the ReCount resource of analysis-ready datasets (Frazee *et al.*, 2011), and on any RNA-Seq experiment that provides a matrix of read counts per gene. An early version of this software was used in Robinson *et al.* (2014), on Bar-Seq measurements of the yeast deletion set, to determine the effect of read depth on detection of differential abundance.

subSeq also streamlines the process of performing a differential expression analysis on each subsample, and of calculating relevant biological metrics for each to determine how they vary depending on read depth. In particular, subSeq reports metrics representing (i) the power to detect differential expression or abundance, (ii) the accuracy of effect size estimation and (iii)

*To whom correspondence should be addressed.

the estimated rate of false discoveries relative to the full experiment.

2 METHODS

The user provides an unnormalized $M \times N$ matrix X of read counts, where each row represents one of M genes, each column represents one of N samples and each value denotes the number of reads aligned to each gene within each sample. The user also specifies a vector of K subsampling proportions $\boldsymbol{p}$, each in the interval (0, 1], and the number of replications to perform at each proportion. For each p_k, a subsampled matrix $Y^{(k)}$ is generated such that $Y^{(k)}_{m,n} \sim \text{Binom}(X_{m,n}, p_k)$ for $m = 1, \ldots, M$ and $n = 1, \ldots, N$. This is equivalent to allowing each original mapped read to have probability p_k of being included in the new counts, as done, for example, by the Picard DownsampleSam function.

For each subsample, we perform the same analysis that is performed on the full set of reads. Multiple approaches for the determination of RNA-Seq differential expression from a matrix of counts, including edgeR (Robinson *et al.*, 2010) and DESeq2 (Love *et al.*, 2014), are built into subSeq, as is DEXSeq for differential exon usage detection (Anders *et al.*, 2012). The user can also provide a custom method to be applied to each subsample.

Here we use subSeq to examine the effect of depth on the RNA-Seq dataset from Hammer *et al.* (2010), testing for differential expression between rats with induced chronic neuropathic pain and a control group. The mapped read counts were downloaded from ReCount, only samples from the 2-month time point were used, and genes with fewer than five mapped reads were filtered out. We subsampled 11 proportions on a logarithmic scale from 0.01 to 1, performing five replications at each proportion.

3 RESULTS

As an illustrative example, we show the results of subsampling of an RNA-Seq dataset from Hammer *et al.* (2010), using edgeR or DESeq2 to normalize and test each subsample for differential expression. To perform these subsamples manually, it would have required downloading 11.4 Gb of reads, mapping them to the mouse genome, downsampling to produce an additional 95Gb of alignments, matching each read to the gene annotations and only then performing the differential expression analysis. Using subSeq, the subsampling requires only the 4.9 Mb matrix from the ReCount database, can be performed entirely in memory in R and takes a negligible amount of time (<1 s to perform the 55 subsamplings, ~2–8 minutes to perform the analysis at each step, depending on the method chosen).

After constructing subsamples and performing an analysis on each, subSeq calculates and visualizes summary metrics about each sequencing depth (Fig. 1); these plots aid in determining saturation of depth (Supplementary Fig. S1). As the plots show how read depth changes the conclusions of the analysis, the 'oracle' is defined as the *P*-values and estimates at the full depth. To estimate the power, subSeq determines the number of genes found significant at a given false discovery rate. To determine whether the decrease in read depth affects specificity, we also estimate the false discovery proportion (FDP) at each depth. subSeq does this by using the qvalue package to estimate the local false discovery rate for each gene in the oracle, then calculating the average of the oracle local FDR values among the genes found significant at each depth. To determine how depth

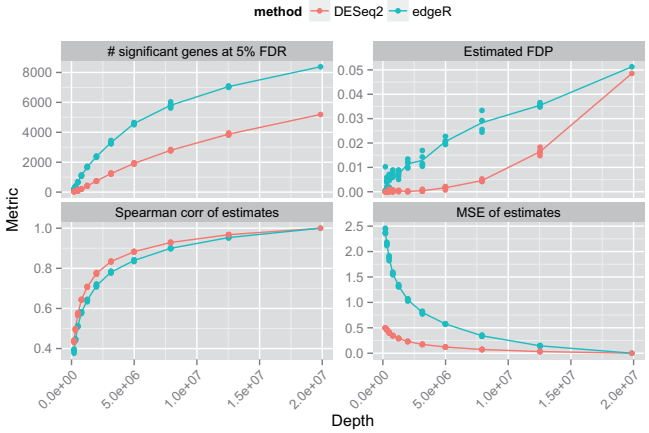

Fig. 1. The default plot generated by subSeq on subsamples of Hammer *et al.* (2010). This shows the number of significant genes at each depth (top left), the estimated FDP (top right) and the Spearman correlation (bottom left) and mean-squared error (bottom right) comparing the estimates at each depth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

affects the accuracy of effect size estimation, subSeq compares the log fold-changes estimated at each depth with the oracle estimates, reporting the mean-squared error and the Pearson and Spearman correlations.

subSeq is designed to allow any analysis to be performed on each subsample. While the example demonstrated here used RNA-Seq data, subSeq works equally well on other genomic approaches such as Bar-Seq or Tn-Seq, as demonstrated in Robinson *et al.* (2014).

ACKNOWLEDGEMENT

The authors thank A.J. Bass for helpful comments on the software and manuscript.

Funding: This work was supported in part by NIH (R01 HG002913).

Conflict of interest: none declared.

REFERENCES

Anders,S. *et al.* (2012) Detecting differential usage of exons from RNA-seq data. *Genome Res.*, **22**, 2008–2017.

Black,M.B. *et al.* (2014) Comparison of microarrays and RNA-seq for gene expression analyses of dose-response experiments. *Toxicol. Sci.*, **137**, 385–403.

Busby,M.A. *et al.* (2013) Scotty: a web tool for designing RNA-Seq experiments to measure differential gene expression. *Bioinformatics*, **29**, 656–657.

Daines,B. *et al.* (2011) The *Drosophila melanogaster* transcriptome by paired-end RNA sequencing. *Genome Res.*, **21**, 315–324.

Frazee,A.C. *et al.* (2011) ReCount: a multi-experiment resource of analysis-ready RNA-seq gene count datasets. *BMC Bioinformatics*, **12**, 449.

Hammer,P. *et al.* (2010) mRNA-seq with agnostic splice site discovery for nervous system transcriptomics tested in chronic pain. *Genome Res.*, **20**, 847–860.

Labaj,P.P. *et al.* (2011) Characterization and improvement of RNA-Seq precision in quantitative transcript expression profiling. *Bioinformatics*, **27**, i383–i391.

Liu,Y. *et al.* (2013) Evaluating the impact of sequencing depth on transcriptome profiling in human adipose. *PLoS One*, **8**, e66883.
Liu,Y. *et al.* (2014) RNA-seq differential expression studies: more sequence or more replication? *Bioinformatics*, **30**, 301–304.
Love,M.I. *et al.* (2014) Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. biorXiv preprint. doi:10.1101/002832.
Rapaport,F. *et al.* (2013) Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biol.*, **14**, R95.
Robinson,D.G. *et al.* (2014) Design and analysis of Bar-seq experiments. *G3 (Bethesda)*, **4**, 11–18.
Robinson,M.D. *et al.* (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.
Tarazona,S. *et al.* (2011) Differential expression in RNA-seq: a matter of depth. *Genome Res.*, **21**, 2213–2223.
Toung,J.M. *et al.* (2011) RNA-sequence analysis of human B-cells. *Genome Res.*, **21**, 991–998.
Wang,Y. *et al.* (2011) Evaluation of the coverage and depth of transcriptome by RNA-Seq in chickens. *BMC Bioinformatics*, **12** (**Suppl. 10**), S5.