Bioinformatics, 2016, 1–3
doi: 10.1093/bioinformatics/btw475
Advance Access Publication Date: 10 July 2016
Applications Note

OXFORD

Gene expression

samExploreR: exploring reproducibility and robustness of RNA-seq results based on SAM files

Alexey Stupnikov[1], Shailesh Tripathi[1,2], Ricardo de Matos Simoes[1], Darragh McArt[3], Manuel Salto-Tellez[3], Galina Glazko[4], Matthias Dehmer[5,6] and Frank Emmert-Streib[1,7,8,*]

[1]Computational Biology and Machine Learning Laboratory, Center for Cancer Research and Cell Biology, School of Medicine, Dentistry and Biomedical Sciences, Faculty of Medicine Health and Life Sciences, Queen's University Belfast, BT9 7AE Belfast, UK, [2]School of Mathematics and Physics, Queen's University Belfast, BT7 1NN Belfast, UK, [3]Northern Ireland Molecular Pathology Laboratory, Centre for Cancer Research and Cell Biology, Queen's University Belfast, BT9 7AE Belfast, UK, [4]Division of Biomedical Informatics, University of Arkansas for Medical Sciences, Little Rock, AR 72205, USA, [5]Department of Biomedical Computer Science and Mechatronics, UMIT, Hall in Tirol, Austria, [6]College of Computer and Control Engineering, Nankai University, Tianjin, P.R. China, [7]Predictive Medicine and Analytics Lab, Department of Signal Processing, Tampere University of Technology, 33720 Tampere, Finland and [8]Institute of Biosciences and Medical Technology, 33720 Tampere, Finland

*To whom correspondence should be addressed.

Associate Editor: Ivo Hofacker

Received on May 6, 2015; revised on June 29, 2016; accepted on June 30, 2016

Abstract

Motivation: Data from RNA-seq experiments provide us with many new possibilities to gain insights into biological and disease mechanisms of cellular functioning. However, the reproducibility and robustness of RNA-seq data analysis results is often unclear. This is in part attributed to the two counter acting goals of (i) a cost efficient and (ii) an optimal experimental design leading to a compromise, e.g. in the sequencing depth of experiments.

Results: We introduce an R package called samExploreR that allows the subsampling (m out of n bootstraping) of short-reads based on SAM files facilitating the investigation of sequencing depth related questions for the experimental design. Overall, this provides a systematic way for exploring the reproducibility and robustness of general RNA-seq studies. We exemplify the usage of samExploreR by studying the influence of the sequencing depth and the annotation on the identification of differentially expressed genes.

Availability and Implementation: samExploreR is available as an R package from Bioconductor.

Contact: v@bio-complexity.com

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

1. Introduction

RNA-seq data (Mortazavi *et al.*, 2008) generated with next-generation sequencing (NGS) platforms are offering many new and exciting opportunities from basic biology to translational clinical research. However, one important problem that have these diverse applications in common is the question regarding the reproducibility and the robustness of obtained results (Peng, 2011). Due to the novelty of RNA-seq data, there are so far only a few studies investigating either the optimal sequencing depth in general

(Sims *et al.*, 2014) or for context specific problems, e.g. for identifying differentially expressed (DE) genes (Liu *et al.*, 2014 Rapaport *et al.*, 2013).

In this paper, we introduce an R package called samExploreR that allows the convenient exploration of reproducibility and robustness of RNA-seq data analysis results. We demonstrate the utility of samExploreR by a case study for identifying DE genes.

2 Methods

The identification of DE genes from RNA-seq data requires the following 4 analysis steps. Step 1: Alignment of short-reads, e.g. with Bowtie2 (Langmead and Salzberg, 2012). Step 2: Annotation dependent matching and summarization, e.g. CuffLinks (Trapnell *et al.*, 2010), (Anders *et al.*, 2014) or featureCounts (a function available in Rsubread (Liao *et al.*, 2013)). Step 3: Normalization (Dillies *et al.*, 2013). Step 4: Statistical analysis including multiple testing correction (Soneson and Delorenzi, 2013).

In principle, the subsampling of reads can be implemented as an additional step at any position before step 3. In this paper, we introduce the R package samExploreR (which is our modification of featureCounts, that integrates the matching and summarization of reads (Steps 2) with a subsampling step. Thus, the application of samExploreR provides 'a shortcut' mapping SAM files directly to subsampled count vectors (cv) representing the number of reads assigned per gene, see Figure 2. A significant advantage of our package is its ability to perform a direct subsampling analysis for various versions of genomic annotations, as matching reads to genes loci and reads resampling, in one procedure. This is in contrast to procedures like SAMtools (Li *et al.*, 2009), Picard (http://broadinstitute.github.io/picard/) or subSeq (Robinson and Storey, 2014), see Figure 1A for a visualization.

3 Results

To demonstrate the applicability of samExploreR for studying reproducibility and robustness we use an RNA-seq dataset from *Mus musculus* (mouse) (Fiorenza *et al.*, 2015). This dataset consists of two groups with three samples in each. Aligning was performed with Bowtie2 allowing 1 mismatch. For our analysis, we consider three different annotations (Flicek *et al.*, 2013). For obtaining information about the robustness and reproducibility of this dataset w.r.t. different metrics, we draw m out of n bootstrap short reads from the original SAM files, whereas $f = m/n$ corresponds to the fraction of subsampled reads, i.e.

$$f : \text{fraction of subsampled reads} = \frac{\#\ \text{subsampled reads}}{\#\ \text{total reads}}. \quad (1)$$

The parameter f has the interpretation of a 'simulated sequencing depth' of a virtual sequencing experiment. Furthermore, for every value of f, we generate $R = 25$ replicates (see loop in Fig. 1). Each of these datasets is then analyzed with DESeq2 (Love *et al.*, 2014) to obtain a list of DE genes.

In Figure 1B.A, we show the number of DE genes (y-axis) as a function of the simulated sequencing depth, f, (x-axis) for three annotations. The three included P-values correspond to results from a Friedman test for dependent groups, comparing the results from $r = 3$ user-specified values of f (horizontal slices; see arrows) for a fixed annotation, testing the null hypothesis of equal means. Ideally, a nonsignificant Friedman test corresponds to statistically robust results because a further increase of the (real) sequencing depth is unlikely to change the results. Practically, the Friedman test may give significant results indicating an imperfection in the robustness of the results w.r.t. variations of the parameter f, as is the case for the examples in Figure 1B.A. Next, we study the reproducibility of different annotations for fixed f values (vertical slices; see arrows). As one can see from Figure 1B.A, the results for different annotations are diverging for increasing values of f making the results less reproducible w.r.t. different genome annotations. The choice of f values is explored in the Suppl.file

In order to emphasize that 'robustness' as well as 'reproducibility' are always defined w.r.t. to a specific metric, we repeat the above analysis for the 'common number of DE genes' between two annotations; results shown in Figure 1B.B. Further questions that could be studied with samExploreR in a similar way would be the reproducibility of results of RNA-seq data from different wet labs, sequencing platforms, gene summarizations (including introns), tissue preparations (FFPE versus FF samples) or statistical analysis methods. Also the robustness of the summarization parameter for multiple gene hits or the number of mismatches for the alignments could be explored.

4 Conclusion

A cornerstone of any scientific study is the question regarding the reproducibility and robustness of obtained results. Unfortunately, for high-throughput data from RNA-seq experiments such questions

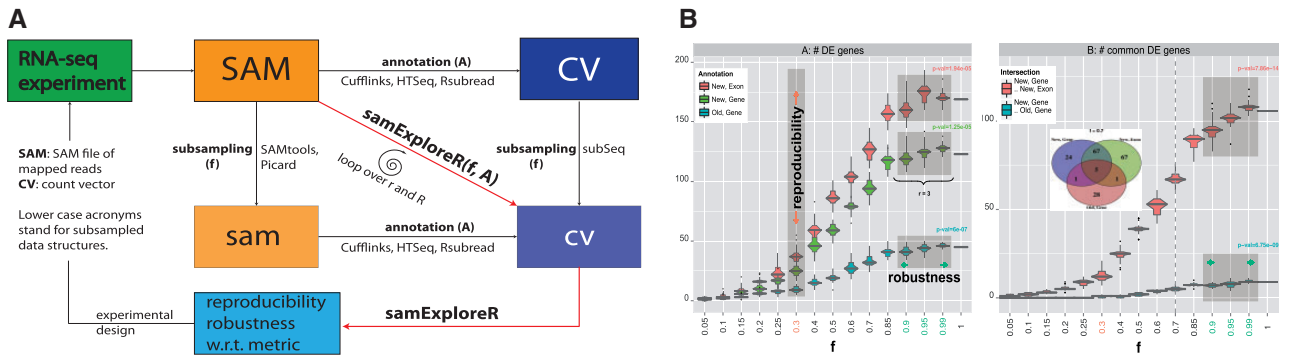

Fig. 1. Figure 1A: Schematic working mechanism of samExploreR with *f*: fraction of reads; *R*: number of resamplings. Figure 1B: Results from samExploreR. (A) Number of DE genes for three annotations and (B) their pairwise intersections of common DE genes. The Venn diagram is for intersections at $f = 0.7$ (vertical dashed line) whereas r is the number of used f values

are highly non-trivial to answer, which may be an explanation for the severe underrepresentation of this topic in the literature. samExploreR provides a flexible exploratory tool for investigating general RNA-seq datasets w.r.t. user-defined metrics.

Funding

A.S. is supported by an international studentship by the CCRCB (Belfast, UK). M.D. thanks the Austrian Science Funds for supporting this work (project P26142). F.E.-S. is supported by the Tampere University of Technology (Finland).

Conflict of Interest: none declared.

References

Anders,S. *et al.* (2014) HTSeq–A Python framework to work with high-throughput sequencing data. *Bioinformatics*, btu638.

Dillies,M.A. *et al.* (2013) A comprehensive evaluation of normalization methods for illumina high-throughput RNA sequencing data analysis. *Brief. Bioinf.*, **14**, 671–683.

Fiorenza,A. *et al.* (2015) Blocking miRNA biogenesis in adult forebrain neurons enhances seizure susceptibility, fear memory, and food intake by increasing neuronal responsiveness. *Cerebral Cortex*, bhu332.

Flicek,P. *et al.* (2013) Ensembl 2014. *Nucleic Acids Res.*, gkt1196.

Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–359.

Li,H. *et al.* (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Liao,Y. *et al.* (2013) featurecounts: an efficient general-purpose read summarization program. *arXiv*, **1305**, 16.

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. *Genome Biol.*, **15**, 550.

Mortazavi,A. *et al.* (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods*, **5**, 621–628.

Peng,R.D. (2011) Reproducible research in computational science. *Science (New York, NY)*, **334**, 1226.

Rapaport,F. *et al.* (2013) Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biol.*, **14**, R95.

Robinson,D.G. and Storey,J.D. (2014) subseq: determining appropriate sequencing depth through efficient read subsampling. *Bioinformatics*, **30**, 3424–3426.

Sims,D. *et al.* (2014) Sequencing depth and coverage: key considerations in genomic analyses. *Nat. Rev. Genet.*, **15**, 121–132.

Soneson,C. and Delorenzi,M. (2013) A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinf.*, **14**, 91.

Trapnell,C. *et al.* (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.*, **28**, 511–515.