



Gene expression

RNA-SeQC 2: efficient RNA-seq quality control and quantification for large cohorts

Aaron Graubert^{1,†}, François Aguet ^{1,†}, Arvind Ravi¹, Kristin G. Ardlie¹ and Gad Getz ^{1,2,3,*}

¹Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA, ²Cancer Center and Department of Pathology, Massachusetts General Hospital, Boston, MA 02114, USA and ³Department of Pathology, Harvard Medical School, Boston, MA 02115, USA

*To whom correspondence should be addressed.

[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Associate Editor: Anthony Mathelier

Received on August 9, 2020; revised on January 20, 2021; editorial decision on February 9, 2021; accepted on February 26, 2021

Abstract

Summary: Post-sequencing quality control is a crucial component of RNA sequencing (RNA-seq) data generation and analysis, as sample quality can be affected by sample storage, extraction and sequencing protocols. RNA-seq is increasingly applied to cohorts ranging from hundreds to tens of thousands of samples in size, but existing tools do not readily scale to these sizes, and were not designed for a wide range of sample types and qualities. Here, we describe RNA-SeQC 2, an efficient reimplement of RNA-SeQC (DeLuca *et al.*, 2012) that adds multiple metrics designed to characterize sample quality across a wide range of RNA-seq protocols.

Availability and implementation: The command-line tool, documentation and C++ source code are available at the GitHub repository <https://github.com/getzlab/rnaseqc>. Code and data for reproducing the figures in this paper are available at <https://github.com/getzlab/rnaseqc2-paper>.

Contact: gadgetz@broadinstitute.org

Supplementary information: [Supplementary data](#) are available at *Bioinformatics* online.

1 Introduction

Identification of high-quality samples is a crucial step in generating and analyzing RNA sequencing (RNA-seq) data and in optimizing RNA-seq protocols. To gain statistical power to detect genetic and environmental effects on the transcriptome, RNA-seq data is being generated for increasingly large cohorts of samples. For example, the Genotype-Tissue Expression (GTEx) project and the Trans-Omics for Precision Medicine (TOPMed) program have generated tens of thousands of RNA-seq measurements across diverse tissue and cell types (GTEx Consortium, 2020; Taliun *et al.*, 2021). Such large-scale studies frequently contain samples of variable RNA quality, including lower-quality archival samples from biobanks. Likewise, research and clinical sequencing efforts in cancer often generate RNA-seq data from formalin-fixed and paraffin-embedded (FFPE) tumor samples, typically using capture-based protocols (e.g. Van Allen *et al.*, 2015), with a wide range of RNA degradation and data quality. Discarding and resequencing samples on the basis of quality filters may not be feasible, and a diverse set of quality metrics is needed to guide the interpretation of the data and analysis results.

Here, we present RNA-SeQC 2, an efficient new version of RNA-SeQC (DeLuca *et al.*, 2012) that computes a comprehensive set of metrics for characterizing samples processed by a wide

range of protocols. It also quantifies gene- and exon-level expression, enabling effective quality control of large-scale RNA-seq datasets.

2 Quality control metrics

RNA-SeQC 2 generates over 70 metrics that characterize the quality of the RNA, sequencing data, alignments and expression profile of the sample. The output metrics are described in detail in [Supplementary Tables S1–S3](#). RNA-SeQC 2 calculates metrics at the gene level and does not take into account transcript isoforms ([Supplementary Methods](#)). Cohort-level analyses are supported by aggregating individual sample outputs into metrics and gene expression tables, and generating a graphical report that displays the distribution of key metrics across samples. The metrics are compatible with MultiQC (Ewels *et al.*, 2016), which can aggregate results from multiple quality control tools. Details of the read filtering steps, insert size distribution estimation, and depth and coverage bias calculations are provided in [Supplementary Methods](#).

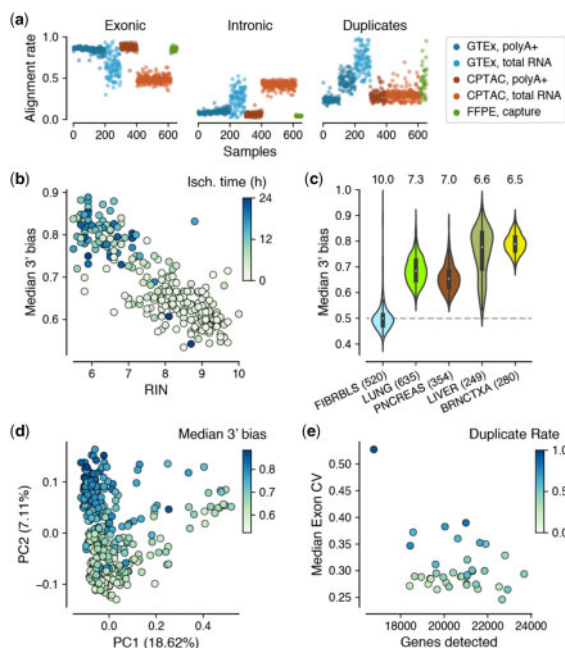


Fig. 1. RNA-SeQC 2 metrics capture a diverse range of sample qualities. (a) Proportion of reads aligning to exons and introns, and of duplicate reads, across five cohorts sequenced using polyA+, total RNA and capture-based protocols, illustrating how metrics and outliers vary across cohorts (see also [Supplementary Fig. S1](#)). (b,c) RNA quality metrics capture sources of RNA degradation. (b) 3' coverage bias resulting from polyA+ capture correlates with ischemic time and RNA fragmentation (measured by RNA integrity number, RIN), shown for 272 GTEx adrenal gland samples. (c) Tissue-specific quality differences, with minimal 3' bias in samples from cell lines (cultured fibroblasts, FIBRBLs) where average RNA quality is high (RIN at top). 3' bias values are normalized to [0,1], with 0.5 corresponding to balanced coverage. Tissue abbreviations and colors from ([GTEx Consortium, 2020](#)), with sample numbers in parentheses. (d) Top two expression principal components (PCs) for 389 GTEx sigmoid colon samples, illustrating that metrics provide insights into technical sources of expression variation ($R^2 = 0.64$ for PC2 and 3' bias). (e) FFPE samples from [Van Allen et al. \(2015\)](#), where the coefficient of variation (CV) of exon coverage facilitates identification of lower-quality samples with higher variability in coverage, higher duplication rates and fewer genes detected

3 Implementation and performance

RNA-SeQC 2 is implemented in C++ and requires a coordinate-sorted BAM or CRAM file ([Li et al., 2009](#)); it leverages the sort order to minimize memory overhead, achieving approximately constant memory usage over time and across samples (~1 GiB, varying with read density). RNA-SeQC 2 uses the SeqLib library ([Wala and Beroukheim, 2017](#)) to process over 150 000 reads/second, approximately twice the rate of version 1.1.9.

4 Results

To demonstrate the utility of the new coverage metrics, we ran RNA-SeQC 2 on ~18,500 RNA-seq samples from GTEx, which were sequenced using an unstranded polyA+ selection protocol ([GTEx Consortium, 2020](#)) and represent a diverse range of sample qualities, on colon cancer ([Vasaikar et al., 2019](#)) and lung adenocarcinoma ([Gillette et al., 2020](#)) samples from CPTAC, sequenced using stranded polyA+ and total RNA protocols, respectively, and on FFPE samples sequenced with a capture-based protocol from [Van Allen et al. \(2015\)](#). Comparisons of metrics across samples reveal significant variability across cohorts and sequencing protocols and identify low-quality samples as outliers ([Fig. 1a](#), [Supplementary Figs S1 and S2](#)). While some metrics are correlated by design (e.g. exonic and intronic alignment rates), the full set captures diverse sample

characteristics, ranging from quality of RNA and sequencing data to expression profiling ([Supplementary Figs S3–S5](#)). The new median 3' bias metric captures RNA degradation for a wide range of sample qualities ([Fig. 1b, c](#)) and is balanced (i.e. close to 0.5) for samples with high RNA quality (e.g. cell lines; [Fig. 1c](#)).

While cell type composition is a major component of expression variability in bulk tissue samples ([Kim-Hellmuth et al., 2020](#)), the diverse set of metrics from RNA-SeQC 2 enables identification of technical sources of variation ([Fig. 1d](#)) and can be used in downstream analyses as explicit covariates or to inform the selection of latent variables that capture overall sources of variation [e.g. PEER factors ([Stegle et al., 2010](#); [GTEx Consortium, 2020](#))]. The new coverage metrics were designed to represent the potentially wide variation present in the transcriptome-capture protocols typically used to sequence FFPE samples. Indeed, RNA degradation during sample preparation, followed by the selective sampling inherent to target capture, may result in increased levels of uneven amplification and decreased library complexity. These simultaneous effects are well captured by the new 'Median Exon CV' metric, which measures evenness of coverage across exonic regions ([Fig. 1e](#) and [Supplementary Fig. S6](#)).

In summary, RNA-SeQC 2 expands the scope of version 1 to a wide range of sample types and qualities, and the improved performance enables rapid and cost-effective quality control of cohorts of thousands of samples.

Acknowledgments

A.G. implemented the software. F.A. and A.G. outlined and planned development. A.R. contributed to metrics development. F.A., A.G. and G.G. prepared and reviewed the manuscript and figures. F.A., K.G.A. and G.G. conceived and supervised the project. All authors reviewed and edited the manuscript. We thank Daniel McGoldrick for help with testing RNA-SeQC 2, Yo Akiyama for processing the CPTAC data and Mendy Miller for editorial suggestions.

Funding

This work was funded by the National Institutes of Health contract HHSN268201000029C to The Broad Institute, Inc.

Conflicts of Interest: F.A. is an inventor on a patent application related to TensorQTL; F.A. and K.G.A. receive research funds from Calico Life Sciences, LLC; G.G. receives research funds from IBM and Pharmacyclics, and is an inventor on patent applications related to MuTect, ABSOLUTE, MutSig, MSMuTect, MSMuSig, POLYSOLVER and TensorQTL. G.G. is a founder, consultant and holds privately held equity in Scorpion Therapeutics.

References

- DeLuca, D.S. et al. (2012) RNA-SeQC: RNA-seq metrics for quality control and process optimization. *Bioinformatics*, **28**, 1530–1532.
- Ewels, P. et al. (2016) MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, **32**, 3047–3048.
- Gillette, M.A. et al. (2020) Clinical Proteomic Tumor Analysis Consortium. (2020) Proteogenomic characterization reveals therapeutic vulnerabilities in lung adenocarcinoma. *Cell*, **182**, 200–225.e35.
- GTEx Consortium. (2020) The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science*, **369**, 1318–1330.
- Kim-Hellmuth, S. et al.; GTEx Consortium. (2020) Cell type-specific genetic regulation of gene expression across human tissues. *Science*, **369**, eaaz8528.
- Li, H., 1000 Genome Project Data Processing Subgroup. et al. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
- Stegle, O. et al. (2010) A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. *PLoS Comput. Biol.*, **6**, e1000770.

- Taliun,D. *et al.*; NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium. (2021) Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature*, **590**, 290–299.
- Van Allen,E.M. *et al.* (2015) Genomic correlates of response to CTLA-4 blockade in metastatic melanoma. *Science*, **350**, 207–211.
- Vasaikar,S. *et al.*; Clinical Proteomic Tumor Analysis Consortium. (2019) Proteogenomic analysis of human colon cancer reveals new therapeutic opportunities. *Cell*, **177**, 1035–1049.e19.
- Wala,J. and Beroukhi,R. (2017) SeqLib: a C++ API for rapid BAM manipulation, sequence alignment and sequence assembly. *Bioinformatics*, **33**, 751–753.