MCCRD-SOP0012: RNASeq Transciptome Data Analysis Pipeline and Specifications

Effective Date: 5/10/2022

Please check for revision status of the SOP at

https://pdmr.cancer.gov/sops/

PDMR NCI Patient-Derived Models Repository
An NCI Precision Oncology Initiative℠ Resource

TABLE OF CONTENTS

VERSION INFORMATION

1. Change History

Revision	Description
	Internal SOP used by MOCHA Laboratory
11/1/2017	Standardize SOP for posting to PDMR internal site for use by designated NCI intramural laboratories
5/20/2022	Gene Expression analysis recommendations section added

1.0 PURPOSE/SCOPE

This Standing Operating Procedure (SOP) describes the pipeline and data analysis specifications for HiSeq RNASeq Pipeline for Patient-Derived Models used/performed by the Molecular Characterization and Clinical Assay Development Laboratory (MoCha) at the Frederick National Laboratory for Cancer Research. **This SOP is for research purposes only and no clinical samples will be processed using this SOP. Any deviation from this SOP will be noted but will not be formally documented**.

2.0 REFERENCES

2.1 Langmead B, Salzberg S. Fast gapped-read alignment with Bowtie 2. Nature Methods. 2012, 9:357-359.

2.2 Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R. and 1000 Genome Project Data Processing Subgroup (2009) The Sequence alignment/map (SAM) format and SAMtools. Bioinformatics, 25, 2078-9.

2.3 Li, B. and Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics. 12:323.

2.4 Link to BBMap short read aligner on GitHub: https://github.com/BioInfoTools/BBMap

3.0 PROCEDURE

3.1 FASTQ files are generated with bcl2fastq (version: 2.17.1.14, Illumina). Adaptors are trimmed within this process using the default cutoff of the adapter-stringency option.

3.2 PDX Mouse reads are removed from the raw FASTQ files using bbsplit (bbtools v37.36).

3.2.1 bbtools bbsplit build=1 -Xmx10g path=<indexPath> in1=Sample.R1.fastq.gz in2=Sample.R2.fastq.gz basename=Sample_%_#.fastq.gz refstats=Sample_stat.txt

3.3 The fastq files are mapped to human transcriptome based on exon models from hg19 using Bowtie2 (version 2.2.6) [1]. The resulting SAM files are converted to BAM forma using samtools [2] and the coordinations in BAM are converted to the genomic (hg19) coordinations using RSEM (version 1.2.31). Gene and transcript quantifications are also done using RSEM.

3.3.1 rsem-calculate-expression --output-genome-bam --bowtie2 --forward-prob 0 --paired-end Sample_hg19_1.fastq.gz Sample_hg19_2.fastq.gz hg19_UCSC_RefGene_rsem

3.4 Removal of Small nucleolar RNAs (snoRNAs)

3.4.1 custom script, available upon request

GENE EXPRESSION ANALYSIS RECOMMENDATION

Added 5/10/2022

- Recommendation for differential gene expression analysis based on RSEM output

  TPM data may not be suitable for differential expression (DE) analysis when comparing across samples in some cases, especially when ribosomal and mitochondria RNAs constitute a very large part in sequence reads [1]. One recommended approach for DE analysis is using R package "tximport" [2] to convert sample based gene level or isoform level RSEM data to estimated count data and then use DESeq2 to perform normalization and DE analysis [3].

[1] Zhao S, Ye Z, Stanton R. "Misuse of RPKM or TPM normalization when comparing across samples and sequencing protocols." *RNA*. 2020; 26(8): 903-909. doi:10.1261/rna.074922.120
[2] Soneson C, Love MI, Robinson MD. "Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences." *F1000Research*, 2015; 4. doi: 10.12688/ f1000 research.7563.1
[3] Love MI, Huber W, Anders S . "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2." *Genome Biology*, 2014; 15, 550. doi: 10.1186/s13059-014-0550-8.