BioSkryb
GENOMICS

Cells explored. Answers revealed.

RNAseq Pipeline v1.2.1 Documentation

Overview

This document describes the bioinformatics pipeline used within BaseJumper to provide various biomarker and QC output from RNA Sequencing output. The main scope of this file is to review the methods and to explain the output file formats leveraged by the pipeline. The RNAseq pipeline is a scalable, portable, and reproducible bioinformatics pipeline to process RNAseq data and assess transcript-level and gene-level quantification. The pipeline supports both single-end and paired-end data. The pipeline's summary and tool usage are explained on Figure 1:

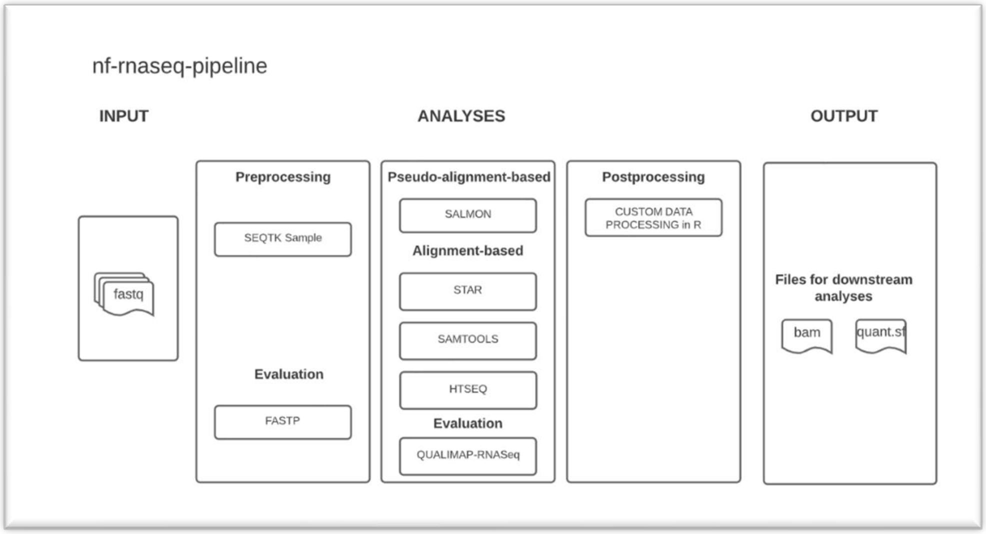

Figure 1. RNA-Seq pipeline components

The pipeline is built using Nextflow[1], a workflow tool to run tasks across multiple compute infrastructures in a very portable manner. At BioSkryb, we currently deploy these pipelines within our computing structure within AWS but it is flexible to run locally or other cloud providers. All the processes in the pipeline run inside docker containers which makes it easy to reproduce the environment and highly reproducible results. The pipeline takes raw sequencing data in form of FASTQ files and performs down-sampling (randomly selecting a fixed, smaller number from the full set of reads) and adapter trimming of FASTQ files. The pipeline then performs transcript-level quantification using the pseudo-alignment method, Salmon[2], splice-aware alignment (STAR[3]) and perform gene-level quantification (using HT-Seq[4]), and performs PCA/heatmap visualization natively.

Pipeline Workflow

1. **Preprocessing**
    - **a.** (optional) Subsample of the paired-end and single-end reads to 100,000 using SEQTK SAMPLE to compare metrics across samples. Input for this step is raw FASTQ files, and output are subsampled FASTQ files.
    - **b.** Trimming reads by adapter, or custom sequences, improves evaluation and alignment. It is performed using FASTP tool. Input for this step is raw FASTQ files, and output are trimmed FASTQ files.
    - **c.** Default adapter sequences. These currently work natively with the ResolveOME chemistry but you will need to supply your own if you are using a separate library preparation chemistry:
        - **i.** `adapter_sequence = "AAGCAGTGGTATCAACGCAGAGTACA"`
        - **ii.** `adapter_sequence_r2 = "AAGCAGTGGTATCAACGCAGAGTACAT"`
2. **Pseudo-alignment-based transcript quantification**
    - **a.** In this step transcript-level quantification is performed using the pseudo-alignment method implemented in Salmon. Input for this step are trimmed FASTQ files, referenced index file for the transcriptome and output are alignment files (BAM[5]). Currently, the reference index file corresponds to the Ensembl Human Genome GRChr38 (version 104).
3. **Alignment based gene level quantification**
    - **a.** In this step, reads are aligned to reference genome and extracting of solely primary reads (uniquely aligned) is performed, in order to achieve better analysis precision. The input are trimmed FASTQ files, referenced index genome file and the main outputs are alignment .bam file. This step consists of several substeps:
        - **i.** **STAR SPLICE-AWARE ALIGNMENT**- Spliced Transcripts Alignment to a Reference (STAR) is a fast RNA-seq read mapper, with support for splice-junction and fusion read detection. The input to the tool is a FASTQ file, its output is a BAM file.
        - **ii.** **QUALIMAP QUALITY CONTROL**- RNA-seq QC reports quality control metrics and bias estimations which are specific for whole transcriptome sequencing, including reads genomic origin, junction analysis, transcript coverage and 5'-3' bias computation. The input to the tool is an aligned BAM file, its output is a report file (which is further summarized in the final MultiQC[6] output) containing:
            1. total number of mapped reads (left/right in case of paired-end reads, secondary alignments are ignored)
            2. total number of alignments
            3. number of secondary alignments

4. number of non-unique alignments
5. number of reads aligned to genes
6. number of ambiguous alignments (belong to several genes, ignored during counting procedure)
7. number of alignments without any feature (intronic and intergenic)
8. number of ignored alignments when the chromosome is not found in annotation
9. number of unmapped reads

iii. **SAMTOOLS EXTRACTION OF PRIMARY ALIGNMENTS**- In this step `samtools` is used to extract only primary alignments. The input to the tool is an aligned BAM file, its output is a sorted BAM file having only primary aligned reads which will then are used for the following post-processing steps.

4. **Postprocessing**

   Post-alignment processing involves the following:

   a. **Perform gene-level quantification using the STAR-based primary alignment bam using HTseq**. The input to the tool is an aligned BAM file containing only primary aligned reads, its output are data frames containing gene-level quantification information. The quantification files are saved as tsv (Tab-separated value) file per sample and the merging per biosample is performed as well in this step.

   b. **Aggregate the Salmon pseudo-aligned quantification into transcript-level and gene-level tables:** The input to this module is the Salmon output folder containing the pseudo-alignment quantification. The R package `tximport` is used to convert Salmon raw output into data frames containing transcript and gene-level quantification information utilizing all the following scaling methods:

      i. `countsFromAbundanceNo`
      ii. `countsFromAbundancescaledTPM`
      iii. `countsFromAbundancelengthScaledTPM`
      iv. `countsFromAbundancedtuScaledTPM`
      v. For more information regarding the scaling methods visit (https://bioconductor.org/packages/devel/bioc/vignettes/tximport/inst/doc/tximport.html). The quantification files are saved as .tsv file per sample. In addition, we provide a merged table collecting all biosamples quantified tables into a single master table.

   c. **PCA.** The project-wise gene-level count matrix created using HTseq is subjected to Principal Component analysis. Briefly, the count matrix is first Log normalized and then the Top 500 most highly variable genes are extracted. Later, PCA is performed over this top 500 subset matrix. In addition, we create a heatmap using the matrix containing the top 500 most highly variable genes.

   - **d. Heatmap**. By default, both rows (genes) and samples (columns) are subjected to unsupervised hierarchical clustering. Coloring of the genes is based on z-score transformation of the quantification of each gene across samples, this allows to visualize obvious differences among samples
   - **e. Cell typing**. Inputting the HTseq based gene-level count matrix we performed unbiased cell type recognition by leveraging reference transcriptomic datasets via the R package `SingleR`[7]. We utilized the following transcriptomic reference dataset. Human primary cell atlas (HPCA), Genotype-Tissue Expression (GTEx), and (The Cancer Gene Atlas Program) TCGA datasets. In addition, we estimate the cell stage of each cell utilizing Seurat[8] Cell stage markers.

5. **Reporting**

   In this step, all metrics are collected, the custom report about which samples passed evaluation is created, and finally outputs of those substeps are forwarded to MultiQC, to create the final .html report.

   - **a. CUSTOM QC REPORT** - The tool aggregates the QC metrics across all the samples to create a summary metrics file and prepares to display as a summary table in MultiQC. The summary metrics file is containing number of reads aligned, total aligned reads, number of not aligned reads, SSP estimation, overlapping exon reads, information about reads genomic origin, transcript coverage profile information and percentage of read aligned. The input is the metrics files; its output is a summary metrics file.
   - **b. MultiQC** - The tool aggregates results of bioinformatics analyses into a single HTML report. The input is the output files from FASTP, STAR, QUALIMAP, SALMON, PCA/HEAT and CUSTOM QC REPORT; its output is HTML report which can be viewed in any browser and specific tables can be copy and pasted in your preferred spreadsheet app or as plain txt files.

Pipeline Parameters

Required Inputs

a. FASTQ pairs or csv file
b. [Pulled from BaseJumper]
   a. Organization UUID
   b. Workspace UUID
   c. Project UUID
c. STAR reference index file
d. SALMON reference index file
e. Gene transfer format (GTF) file format holding information about gene structure
f. The `tx2gene` table file holding information about the mapping of transcripts to genes.
g. celltype_ref. Custom created databases of GTEx, HPCA and TGCA ready for cell typing.

Optional

a. Number of reads
b. Read length
c. Adapter sequence
d. Processes to skip

Main Pipeline Deliverables

Analyses Step	Folder	Raw Output Name	Data Type	Description
Alignment	**RNASEQ_WF_STARALIGN**	<BIOSAMPLE_NAME>_Aligned.sortedByCoord.out.bam	BAM	Aligned bam for each biosamples
Quantification gene level	**RNASEQ_WF_MERGE_HTseq_SUMMARY/**	merged_df_num_detected_gene_starhtseq.tsv	TSV	TSV file data frame file having information about gene biotype, and counts for each project
Quantification transcript level	**RNASEQ_WF_MERGE_TXIMPORT_SALMON_TX_GENE/**	merged_df_num_detected_tx_salmon.tsv	TSV	Transcript level TSV file data frame file having information about gene biotype, and counts for each project
Quantification transcript level	**RNASEQ_WF_MERGE_TXIMPORT_SALMON_TX_GENE/**	merged_df_num_detected_gene_salmon.tsv	TSV	Gene level TSV file data frame file having information about gene biotype, and counts for each project
Cell typing	**RNASEQ_WF_CELL_TYPING/**	df_summary_celltype_singler_hpca_gtex_tcga.tsv	TSV	
Reporting	**RNASEQ_WF_MULTIQC_WF_MULTIQC/**	MultiQC_report.html	csv	MultiQC report providing summary of the key metrics in the analyses
PCA/Heatmap	**RNASEQ_WF_PLOTTER_PCAHEATMAP_HTseq_SUMMARY**	heatmap_mqc.png pca_mqc.png mt_mqc.png	PNG	
Quantification transcript level	**RNASEQ_WF_MERGE_TXIMPORT_SALMON_TX_GENE/**	merged_df_mtcounts_salmon.tsv	TSV	Transcript level TSV file data frame file having information about mitochondrial gene biotype, and counts for each project

Other Pipeline Outputs

Analyses Step	Folder	Raw Output Name	Data Type	Description
Preprocessing	**RNASEQ_WF_FastpFull_WF_FASTP**	<BIOSAMPLE_NAME>_{1,2}_trim FASTQ.gz - FASTQ files after trimming of adapter sequences provide by client	FASTQ	Trimmed FASTQ files for each biosample
Preprocessing	**RNASEQ_WF_FastpFull_WF_FASTP**	<BIOSAMPLE_NAME>.fastp.json	json	metric json file containing summary of statistics data before and after filtering, filtering results containing number of reads, insert size.
Alignment	**RNASEQ_WF_STARALIGN**	star_outdir_<BIOSAMPLE_NAME>	log	folder containing star alignment metrics file <sample_id>_Aligned.sortedByCoord.out.bam
Pseudo alignment	**RNASEQ_WF_SALMON QUANT/**	salmon_outdir_<BIOSAMPLE_NAME>	log	folder containing metrics data from pseudo alignment process
Pseudo alignment	**RNASEQ_WF_SALMON QUANT/**	star_outdir_<BIOSAMPLE_NAME>	SAM	SAM alignment file for each biosample
Quantification gene level	**RNASEQ_WF_MERGE_HTseq_SUMMARY/**	merged_df_mtcounts_starhtseq.tsv	TSV	TSV file data frame file having information about mitochondrial gene biotype, and counts for each project
Quantification transcript level	**RNASEQ_WF_MERGE_TXIMPORT_SALMON_TX_GENE/**	merged_df_mtcounts_starhtseq.tsv	TSV	Transcript level TSV file data frame file having information about mitochondrial gene biotype, and counts for each project
Reporting	**RNASEQ_WF_CREATE_QC_REPORT/**	qualimap.stats.csv qualimap_percents_stats.csv	csv	The summary metrics files containing number of reads aligned, total aligned reads, number of not aligned reads, SSP estimation, overlapping exon reads, information about reads genomic origin, transcript coverage profile information and percentage of read aligned.

References

1. Di Tommaso, P. *et al.* Nextflow enables reproducible computational workflows. *Nat. Biotechnol.* **35**, 316–319 (2017).
2. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **14**, 417–419 (2017).
3. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinforma. Oxf. Engl.* **29**, 15–21 (2013).
4. Anders, S., Pyl, P. T. & Huber, W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).
5. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *GigaScience* **10**, giab008 (2021).
6. Ewels, P., Magnusson, M., Lundin, S. & Käller, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**, 3047–3048 (2016).
7. Aran, D. *et al.* Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nat. Immunol.* **20**, 163–172 (2019).
8. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411–420 (2018).