VERSION 4

OCT 17, 2023

OPEN ACCESS

DOI:
dx.doi.org/10.17504/protocols.io.rm7vzxx9rgx1/v4

Protocol Citation: Tyler Stahl 2023. Bulk RNASeq Delivery. **protocols.io** https://dx.doi.org/10.17504/protocols.io.rm7vzxx9rgx1/v4Version created by Tyler Stahl

License: This is an open access protocol distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited

Protocol status: Working
We use this protocol and it's working

Created: Oct 17, 2023

Bulk RNASeq Delivery V.4

Tyler Stahl[1]

[1]Genomics Research Center

Tyler Stahl

ABSTRACT

This protocol will give an overview of the grc data delivery structure and results files

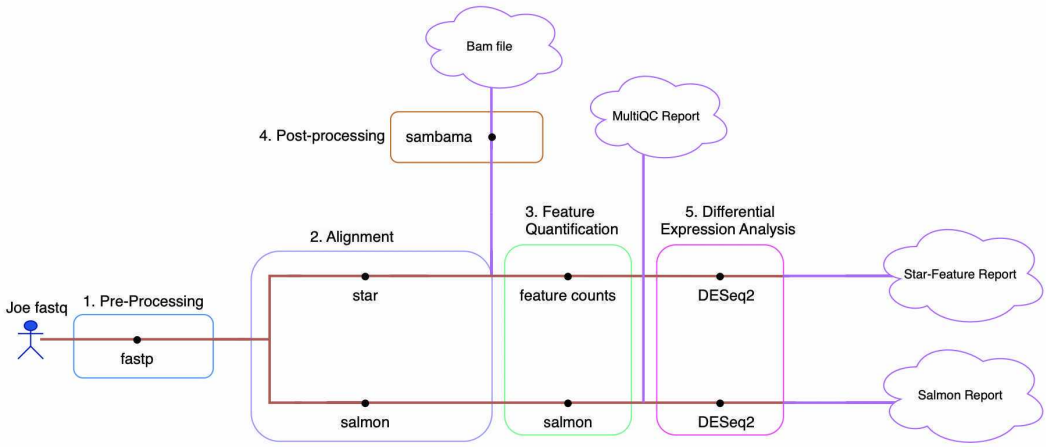
Last Modified: Oct 17, 2023

PROTOCOL integer ID:
89465

Analysis Overview

1 In brief, the RNASeq analysis follows pre-processing (quality control/filtering/trimming), alignment, post-processing, and differential expression analysis between sample groups with DESeq2. A detailed overview of the analysis can be found in the README.txt and methods.txt file. Also, please see the workflow below for those curious about data processing.

README.txt RNASeq_methods.txt



RNASeq Workflow

Delivery Structure

2 The first thing you will notice in the delivery email is three download links corresponding to 1) **Raw (fastq) files** 2) **Aligned (bam/bigWig) files** 3) **Results files**

Download your RNA-Seq Analysis Results and QC:
https://grcweb.circ.rochester.edu/pickup/230821130441-10817/deliv_NHD13_GEO_results.tar.gz
Checksum = 12ee1a41926b5d45ef24c103c7a5f48e
This URL will expire in 10 days.

RNA-Seq Analysis Aligned Data:
https://grcweb.circ.rochester.edu/pickup/230821130431-10715/deliv_NHD13_GEO_align.tar.gz
Checksum = 3c57a3de7218cae0ded7f842b9b2fdd7
This URL will expire in 10 days.

RNA-Seq Raw Data (for GEO submission):
https://grcweb.circ.rochester.edu/pickup/230821130423-10606/deliv_NHD13_GEO_raw.tar.gz
Checksum = 59b103b968ebc9174a475ef809dbc1e2
This URL will expire in 10 days.

Uncompress delivery directory with FREE compression software (http://www.7-zip.org).

If you are on a PC, you will need to have compression software downloaded such as 7zip to unzip the folders. Macs have built-in zip software.

The fastq and aligned are typically large and may take a few hours to download.

1) .fastq files contain nucleotide and quality information generated from the Illumina Sequencer. There is an unlikely reason that you would ever need to open these files directly; however, per the NIH data management policy fastq files need to be deposited online at the time of publication or end of the performance period and stored for 3 years after the grant. For more information on the NIH data management policy, please visit the following website: https://sharing.nih.gov/

2) .bam files store alignment data and mapping quality scores of reads in a binary format. BigWig files contain continuous alignment data in a binary format. BigWigs provided are normalized for sequencing depth and have a bin size of 1

3) The results folder contains quality control information and results files from DESeq2. We will go through each in more detail below.

MultiQC Report

3 MultiQC aggregates QC information from multiple different analysis outputs into a single interactive report. In the case of our RNASeq workflow**:** fastp, star, feature counts, and salmon.

A copy of the MultiQC report can be found attached to the GRC delivery email and in the results folder. Additionally, you may notice within the deliv_PROJECT_results folder; there is a multiqc folder containing individual text files for the individual analyses and some additional MultiQC

info.

Example MultiQC

NHD13_GEO_multiqc_report.html

3.1 **General Statistics:**

fastp: % Duplication, GC content, % PF, % Adapter
Star: % aligned, M aligned
Feature counts: % assigned, M assigned
Salmon: % aligned, M aligned

We will go through each section in a bit more detail below:

Fastp:

Filtering statistics: In addition to duplication removal and adapter trimming, fastp has a number of default filtering metrics, including read quality, read length, N-Content,

Sequencing Quality: Here, you will see the Phred quality scores assigned to each base. A higher Phred score means a higher confidence and lower error rate. A phred score of 30 means an error rate of 1:1000.

N Content: "N content" refers to the proportion or percentage of ambiguous or unknown bases present in the DNA or RNA sequences. These 'N' bases indicate that the actual nucleotide at that position is uncertain or could not be determined during sequencing. The presence of 'N' bases can be due to various factors, including sequencing errors, poor quality reads, and regions of the genome that are difficult to sequence accurately.

GC content: "GC Content" is the proportion or percentage of nucleotide bases in a DNA or RNA sequence that are either guanine (G) or cytosine (C).

STAR:

Alignment:

Uniquely mapped: reads aligned to a single loci
Mapped to multiple loci: read has been aligned to multiple locations (loci) in the reference genome.
Mapped to too many loci: indicates that a read has been aligned to an excessive number of different locations in the reference genome
Mapped too short: refers to reads that have been aligned to the genome but fall short of basic filtering metrics for alignment
Unmapped: other: This category includes reads that were not able to be aligned to the reference genome or alignable transcripts

Feature Counts:

Assignments

Assigned: Number of reads assigned to a genomic feature (i.e gene)
Unassigned: Multi Mapping: This category includes reads that align to multiple genomic features.
Unassigned: No Features This category includes reads that could not be aligned to any of the defined genomic features.
Unassigned: Ambiguity: Reads that align to multiple features but none of the features clearly dominate in terms of read count are categorized as "Ambiguity."

Salmon:

Fragment length distribution: "Fragment length distribution" refers to the distribution of fragment lengths generated in the sample. Salmon leverages this information to accurately quantify transcript abundances by accounting for the effects of fragment lengths and other factors.

DESeq2 Results

4 When you open the deSeq2 folder, you will notice two reports: star-feature (gene-level) and salmon (transcript level).

While most investigators find gene counts to be sufficient for their experiment, there may be specific cases where salmon may be valuable. In theory, transcript-level quantifications can accurately represent expression and biological changes between conditions. For example, while gene level expression may increase or decrease, these changes are driven by a non-functional transcript.

NHD13_GEO_starFeat_SequencingReport.html
NHD13_GEO_salmon_SequencingReport.html

4.1 In the respective star and salmon folders, there will be a few different files:
*Note all text files can be opened in Excel

DESeq Counts:
deSeq2_counts.txt- Raw count values
deSeq2_NormCounts.txt- Count values normalized with DESeq2's median of ratio
deSeq2_rlog_NormCounts.txt- log of the normalized counts

Comparisons files:
You will find a text file containing a full gene list for each comparison (e.g deSeq2_NHD13_vs_WT.txt)

	A	B	C	D	E	F
		BaseMean	log2FoldChange	stat	pvalue	padj
	Hoxa9	636.168	2.557	21.331	5.92E-101	8.68E-97
	Pbx3	456.879	3.091	20.932	2.76E-97	2.03E-93
	Pbx1	401.557	-3.27	-16.477	5.38E-61	2.63E-57

Example DeSeq2 comparison file

BaseMean: The base mean is an estimate of the average expression level for a gene across all samples in a dataset.
log2FoldChange: The log2 fold change represents the gene's expression level change between two conditions or groups. It's calculated by taking the logarithm base 2 of the normalized expression levels (counts) ratio between the conditions or groups. For the comparison EXP vs WT, a positive log2FoldChange
stat: "stat" refers to the test statistic used to assess the significance of differential expression. This statistic is calculated based on the differences in expression levels between conditions and is used to quantify the strength of evidence for differential expression.
p-value: DESeq2 uses a negative binomial distribution to calculate p-values.
p-adj: *The a*djusted p-value, often referred to as the false discovery rate (FDR), corrects for multiple hypothesis testing using the Benjamini and Hochberg method. When conducting thousands of statistical tests (as in gene expression analysis), the likelihood of obtaining false positives (false discoveries) increases. The adjusted p-value provides a more conservative measure of statistical significance.

EnrichR Files:

EnichR is a package to find, based on a list of genes, the enrichment of certain biological pathways and functions. While there are a number of libraries on the EnrichR website, we query 4 common libraries for gene set enrichment: KEGG, GO, Wiki Pathways, and ChEA

Since EnrichR only takes a list of gene symbols, it does not account for directionality. To gage up or downregulated pathways, we separate significant (p-adj <.05) upregulated and downregulated genes before running the analysis.

Note- we do not run enrichR for salmon outputs.

For each comparison, you will notice two enrichR files: **enrichr_*_up.txt** & **enrichr_*_down.txt**

	database	Term	Overlap	P.value	Adjusted.P.value	Combined.Score	Genes
1	GO_Biological_Process_2021	ribosome biogenesis (GO:0042254)	72/192	5.41E-29	2.21E-25	438.4374903	POP5;RPL3;DDX47;RRP1;MRPL36
2	GO_Biological_Process_2021	ncRNA processing (GO:0034470)	72/201	1.53E-27	3.12E-24	386.7264654	POP5;PUS10;RPL3;POP1;DDX47;I
3	GO_Biological_Process_2021	rRNA processing (GO:0006364)	63/173	1.04E-24	1.41E-21	353.3326573	POP5;RPL3;DDX47;RRP1;NAT10;I
4	GO_Biological_Process_2021	rRNA metabolic process (GO:0016072)	58/162	2.01E-22	2.05E-19	310.4137631	POP5;RPL3;DDX47;RRP1;PWP2;R
5	GO_Biological_Process_2021	mRNA processing (GO:0006397)	70/300	2.98E-15	2.16E-12	113.4528304	ISY1;DDX46;CELF1;YBX1;PRPF19;
6	GO_Biological_Process_2021	gene expression (GO:0010467)	78/356	3.17E-15	2.16E-12	104.6377015	HBS1L;NUP107;RPL3;NUP188;M
7	GO_Biological_Process_2021	DNA-dependent DNA replication (GO:0006261)	41/129	3.27E-14	1.91E-11	159.6510599	PRIM2;BLM;PSMD12;PSMD11;DI

Example enrichR file

Database- Database/library from which the pathway is queried from
Term- Name of biology pathway or ontology
Overlap- The number of significant genes that overlap with the ontology or pathway
P.value- The p-value is computed using Fisher's exact test
Adjusted.P.Value- The adjusted p-value is the p-value adjusted for multiple tests using the Benjamini and Hochberg method
Combined Score- The combined score is a combination of the p-values and odds ratio
Genes- Here, you can find significantly differently expressed genes that overlap with the pathway or ontology

More info on enrichR can be found here.

***Excel defaults to reading cells as dates; you may notice some overlap as dates. Please follow this guide to read the data properly in Excel.

Excel_steps.docx

Image Files:

All images contained in the reports can also be found as SVG images.

FAQ

5 **What are the salmon results?**

For salmon, a different alignment algorithm is used, allowing us to better use reads that we would otherwise discard due to multimapping within star-featurecounts. For the salmon report, it maps those reads at a transcript level rather then the whole gene, so you will see inflation of total features. We also report what gene these transcripts belong to, so that the reports are comparable. You may see differential expression within the same gene, but that is tied to what transcripts are up or downregulated. We also provide a summary of the type of transcript (protein-coding, lncRNA, etc).

Why do I not see enrichR results?

If you don't see enrichR results in your StarFeature counts report, this is because there are not enough differentially expressed upregulated or downregulated genes for a specific comparison. In order to reduce false positives and thus calculate significantly enriched pathways, enrichR needs at least 50 genes to be differentially expressed.

Why is my RNASeq data showing a weak knockdown of my gene of interest despite being validated with qRT-PCR?

One such explanation for discrepancies in knockdown expression between qRT-PCR and RNA-Seq data is the expression and alignment of a non-functional transcript. We recommend reading the aligned files into a Genome Browser to look at how the reads align to the gene.

Can I remove a sample from the analysis?

Although certain samples may seem as outliers, this may be due to biological variation. Thus, we don't recommend removing samples from an analysis based on clustering alone. If there is experimental reasoning, such as poor sample quality, that may cause technical variation, you may want to consider removing samples from the analysis.

Can the GRC re-analyze my RNA-Seq experiment?

To rerun the analysis, we recomend using the BulkDeSeq application in HyperGen©. HyperGen© is a custom no-code genomics analytics software platform developed by the Wilmot Cancer Institute's Informatics and Technology Development Group (WCII) incorporating bioinformatics tools that are either open source or customized by the University of Rochester Genomics Research Center (GRC). HyperGen© can be accessed using standard computer login credentials at http://hypergen.hyperionstack.com/ .

What counts files do I use and where?

Please use the raw counts (deSeq2_counts.txt) files within the HyperGen© BulkDeqSeq application. The application replicates our workflow and the statistical modeling is based on the assumption of unnormalized counts.

If you would like to run Gene Set Enrichment Analysis (GSEA), please use the normalized counts (deSeq2_NormCounts.txt). We have written guide to help users get started here https://www.protocols.io/view/gene-set-enrichment-analysis-kqdg3x67qg25/v1

Further Educational Resources

6 https://genomebiology.biomedcentral.com/articles/10.1186/s13059-014-0550-8
https://hbctraining.github.io/DGE_workshop/lessons/04_DGE_DESeq2_analysis.html