OUTLINE

This analysis pipeline takes **de-multiplexed**, **single- or paired-end**, **directional** **RNA-seq** reads, tagged with the **PCR duplicates’ marker**, and produces a map of where these reads originated from in the genome.

DIRECTIONS

1) The read files must be de-multiplexed (DGseq demultiplexing) and have the .fastq suffix. Everything that precedes the fastq suffix is the name that will identify all the output files produced.

2) Log in your NYU HPC account and create a folder containing all the demultiplexed read files you want to analyze. This folder will be the /source/folder.

3) (Optional) Make a second folder that will become the /destination/folder. Alternatively, you can use the same directory every time you are asked for either the /source/folder or /destination/folder.

4) Go to the folder /data/cgsb/gresham/LABSHARE/SHARED/CODE/. You need to be listed as a member of the Gresham lab, registered with the NYU HPC team managing /data/cgsb/, to be able to read, write or execute the programs in this folder.

5) To submit the job and start the analysis of your results type:

a. Single-end reads

qsub –M {your NetID}@nyu.edu RNAseq\_1end.pbs –v source={source/folder},destination={ destination/folder}, ref={reference genome}, name={name}

|  |  |
| --- | --- |
| -M | [Optional] Enables automatic emails, upon initiation, termination or if the run is aborted (an exit status=0 indicates the job was terminated without problems). |
| -v | *There should not be any spaces between n the commas* |
| {source/folder} | This is the folder you created in step 2). |
| {destination/folder} | [Optional] This is the folder you created in step 3). |
| {reference genome} | Currently only yeast is supported. The possible values are **Scer3** (w/o spike-ins), **ScERCC** (w/ ERCC spike-ins). |
| {name} | This is *everything* that precedes the .fastq suffix. This name will follow and identify all the results concerning this sample. |

b. Paired-end reads

qsub –M {your NetID}@nyu.edu RNAseq\_2end.pbs –v source={source/folder},destination={ destination/folder}, ref={reference genome}, name={name}

|  |  |
| --- | --- |
| -M | [Optional] Enables automatic emails, upon initiation, termination or if the run is aborted (an exit status=0 indicates the job was terminated without problems). |
| -v | *There should not be any spaces between the commas* |
| {source/folder} | This is the folder you created in step 2). |
| {destination/folder} | [Optional] This is the folder you created in step 3). |
| {reference genome} | Currently only yeast is supported. The possible values are **Scer3** (w/o spike-ins), **ScERCC** (w/ ERCC spike-ins). |
| {name} | This is *everything* that precedes the \_R1.fastq or \_R2.fastq suffix. This name will follow and identify all the results concerning this sample. |

OUTPUT

1) Two folders within the **/scratch/NetID** of the logged-in user, in the following format:NNNNNNN.dRNAseq\_1end.jobout and NNNNNNN.dRNAseq\_1end.joberr for single-end reads or NNNNNNN.dRNAseq\_2end.jobout and NNNNNNN.dRNAseq\_2end.joberr for paired-end reads. In the previous example, NNNNNNN is the unique job ID allocated automatically by the scheduler. These files contain some useful records of the analysis process and possible alerts.

2) At the assigned **/destination/folder/:**

1. The reads that failed to align in {name}\_unaligned\_1.fastq and {name}\_unaligned\_2.fastq for each read direction.
2. The alignment files containing any PCR duplicate cases that may have occurred in the process. Files {name}\_MAPQ15.bam and {name}\_MAPQ15.sam
3. The alignment files after removing PCR duplicates in the “coordinate method” alone {name}\_MAPQ15.noDUP.bam and {name}\_MAPQ15.noDUP.sam
4. [Recommended for further use] The alignment files after removing PCR duplicates in the “indel method” {name}.rmdup.bam and {name}.rmdup.sam
5. The **subfolder**  metrics/ that contains useful information about the various stages of the process you should expect tho see the folders {name}.rmdp\_stat, {name}.ALN\_Metrics and {name}.RNA\_metrics

DETAILS (pseudo-pseudo-code)

**STEP1**

Fastq files, produced from an Illumina run with the our custom-made adaptors, demoultiplexed are read in, and any residual rRNA contained in this digital sample is removed. The report of this alignment can be found at /scratch/NetID/NNNNNNN.dRNAseq1end.joberr (or /scratch/NetID/NNNNNNN.dRNAseq2end.joberr for paired-end)

**STEP2**

The cleaned-up reads are now aligned with optimized parameters for yeast RNA-seq experiments. Spliced reads are not taken into account in this step. The report of this alignment can be found at /scratch/NetID/NNNNNNN.dRNAseq1end.joberr (or /scratch/NetID/NNNNNNN.dRNAseq2end.joberr for paired-end)

**STEP3**

The few (comparatively) reads that failed to align in the previous step is now tested for splice intron-exon junctions and aligned taking splicing into account. Whatever read still fails to align after this step, is reported in {name}\_unaligned\_1.fastqand {name}\_unaligned\_2.fastq.

**STEP4**

The aligned reads from the two processes are concatenated into a single alignment file wich is further trimmed by removing reads with poor mapping quality (MAPQ<15, can not be changed by the user at the moment. If you have concerns you could slightly modify the relevant samtools script within the RNAseq\_1end\_looseRibo.pbs or RNAseq\_2end\_looseRibo.pbs and save it in a private folder.

**STEP5**

The last step is the removal of possible PCR duplicates from the report and has three outputs:

1. Possible PCR duplicates are not removed. These alignment report are named {name}\_MAPQ15.sam (the .bam version is also available) and can imported into cufflinks and/or Rnubread/EdgeR for further analysis.
2. Possible PCR duplicates removed with the long-published “coordinates” method. These alignment files are {name}\_MAPQ15.noDUP.sam (the .bam version is also available)
3. [Recommended dataset] Possible PCR duplicates removed with our lab’s new method with custom-made adaptors incorporating PCR-counters. These alignment files are {name}.rmdup.sam (the .bam version is also available) .If the “sorted” version is required for certain downstream analysis software use the {name}.rmdup.sort sam (the .bam version is also available) file.

.