RNA-seq Analysis Tool Standard Operating Procedure

Bioinformatics and Genomics Master's Program at University of Oregon

January 19, 2020

RNA-Seq Differential Gene Expression Analysis

NemaMetrix

1. Purpose

This procedure outlines how to operate the BGMP's RNA-seq analysis tool from the command line. The goal of the tool is to generate a report describing differential gene expression statistics. The purpose of this analysis tool is to provide NemaMetrix with the capacity to analyze raw RNA-seq reads for differential gene expression. Additionally, this analysis can be used to validate library prep and sequencing assays.

2. Scope

This tool is intended for use as a differential gene expression analysis.

3. Procedure

Installations:

See section 4 for a list of software that need to be installed prior to the execution of the analysis.

Set up:

1. Set up a Unix/Linux environment.
2. For every unique analysis, copy the files contained in "rna-seq-package" folder to a new folder labeled accordingly. Do not modify the locations of any of the files contained in the rna-seq-package folder.
3. Save the desired Ensembl reference FASTA and GTF file to be used in the analysis to a local folder.
4. Put the raw FASTQ files to be analyzed in a unique folder. This will be known as the "IN.DIR" in the input sample sheet.

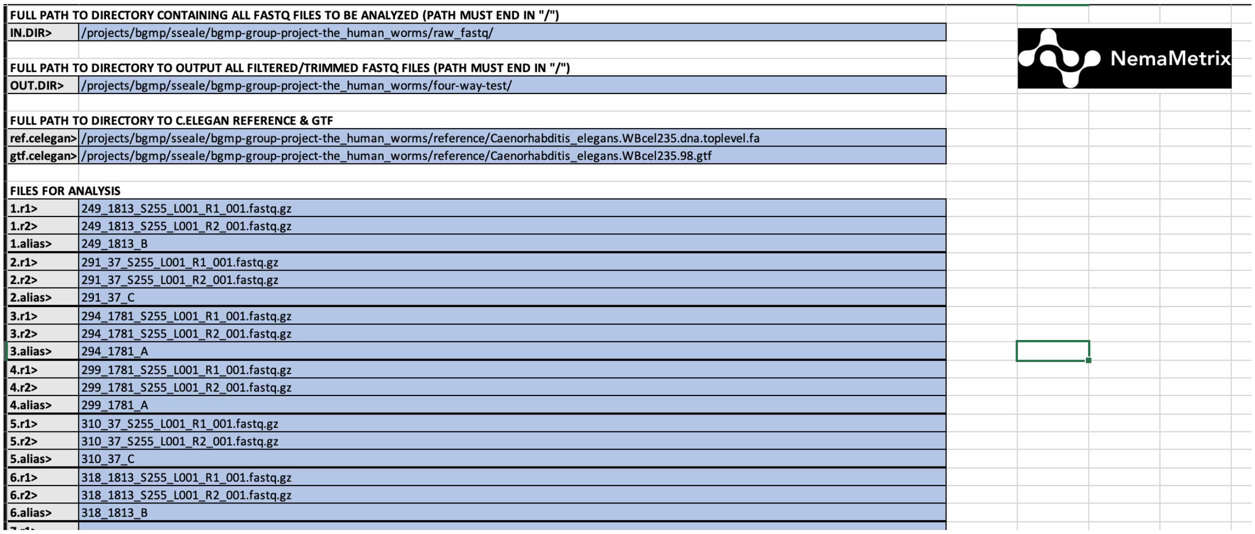
Usage:

Input sample sheet set up:

1. Open the Excel document titled “SAMPLE_SHEET.xlsx”. Do not modify this file name. The script requires this file to be identified as it has been named.
2. Fill out the Excel sheet. A sample of the completely filled out sample sheet is provided below as a guide. Be sure that the sections named “IN.DIR”, “OUT.DIR”, “ref.celegan”, and “gtf.celegan” are filled out. Note: “IN.DIR” and “OUT.DIR” file paths **must** have a trailing “/”.

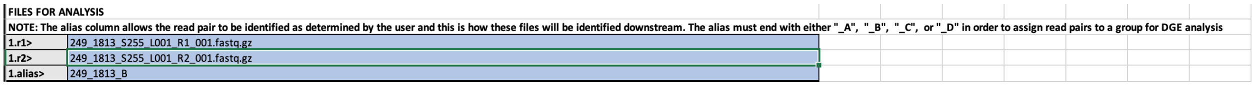
- IN.DIR: the folder containing the raw FASTQ files to be analyzed
- OUT.DIR: the folder for the filtered and trimed FASTQ files to be outputted to
- ref.celegan: the full path to the Ensembl reference FASTA file
- gtf.celegan: the full path to the Ensembl GTF annotation file

Excel input sheet sample:



FULL PATH TO DIRECTORY CONTAINING ALL FASTQ FILES TO BE ANALYZED (PATH MUST END IN "/")	
IN.DIR>	/projects/bgmp/sseale/bgmp-group-project-the_human_worms/raw_fastq/
FULL PATH TO DIRECTORY TO OUTPUT ALL FILTERED/TRIMMED FASTQ FILES (PATH MUST END IN "/")	
OUT.DIR>	/projects/bgmp/sseale/bgmp-group-project-the_human_worms/four-way-test/
FULL PATH TO DIRECTORY TO C.ELEGAN REFERENCE & GTF	
ref.celegan>	/projects/bgmp/sseale/bgmp-group-project-the_human_worms/reference/Caenorhabditis_elegans.WBcel235.dna.toplevel.fa
gtf.celegan>	/projects/bgmp/sseale/bgmp-group-project-the_human_worms/reference/Caenorhabditis_elegans.WBcel235.98.gtf
FILES FOR ANALYSIS	
1.r1>	249_1813_S255_L001_R1_001.fastq.gz
1.r2>	249_1813_S255_L001_R2_001.fastq.gz
1.alias>	249_1813_B
2.r1>	291_37_S255_L001_R1_001.fastq.gz
2.r2>	291_37_S255_L001_R2_001.fastq.gz
2.alias>	291_37_C
3.r1>	294_1781_S255_L001_R1_001.fastq.gz
3.r2>	294_1781_S255_L001_R2_001.fastq.gz
3.alias>	294_1781_A
4.r1>	299_1781_S255_L001_R1_001.fastq.gz
4.r2>	299_1781_S255_L001_R2_001.fastq.gz
4.alias>	299_1781_A
5.r1>	310_37_S255_L001_R1_001.fastq.gz
5.r2>	310_37_S255_L001_R2_001.fastq.gz
5.alias>	310_37_C
6.r1>	318_1813_S255_L001_R1_001.fastq.gz
6.r2>	318_1813_S255_L001_R2_001.fastq.gz
6.alias>	318_1813_B

3. For each read pair being analyzed, add its file name to the sample list. IF more than 16 samples are being analyzed, add additional rows below following the designated row naming convention.



FILES FOR ANALYSIS	
NOTE: The alias column allows the read pair to be identified as determined by the user and this is how these files will be identified downstream. The alias must end with either "_A", "_B", "_C", or "_D" in order to assign read pairs to a group for DGE analysis	
1.r1>	249_1813_S255_L001_R1_001.fastq.gz
1.r2>	249_1813_S255_L001_R2_001.fastq.gz
1.alias>	249_1813_B

4. For each read pair add an alias that will be used to identify the sample in the outputted report. **Important**: The aliases should follow the naming convention shown in the table below based on the number of comparisons you would like the tool to execute. Make sure there are no spaces in the alias name.

- Example: For an experiment containing three treatment groups with six samples to be compared against one another the naming convention should follow “Sample1_A”, “Sample2_A”, “Sample3_B”, “Sample4_B”, “Sample5_C”, “Sample6_C”

Number of Comparisons	Group IDs appended to alias
2 treatments	"_A","_B"
3 treatments	"_A","_B","_C"
4 treatments	"_A","_B","_C","_D"

5. Export the Excel into a .csv file named "SAMPLE_SHEET.csv". Save the file to the same folder where the analyses are going to be performed from.

Running the analysis:

1. Open a command line prompt and navigate to the folder containing the RNA-seq analysis package.
2. To run the analysis, ensure script is executable and enter this command into the terminal:

```
./rna-analysis-master.sh
```

- the script will now continuously run through quality control, alignment, feature counting, and differential expression analysis.

Once the analysis is complete:

- the quality control reports will be in a folder titled "QC_REPORTS"
- the generated report will outputted as an HTML file labeled "dge.html"
- text files containing differential gene expression tables will be outputted to the current folder as "…_l2fc_values.txt"
- a file containing the list of FPKM normalized counts will be outputted as "FPKM_gene_data.tsv"

Once the pipeline has run to completion, enter this command into the terminal and an interactive application will run and open to allow for greater manipulation of the outputted data statistics. The application will take approximately 10 minutes to open.

```
R -e "shiny::runApp('shiny', launch.browser=TRUE)"
```

4. Systems and Documentation

- fastp version 0.20.0 (https://github.com/OpenGene/fastp)
- STAR version 2.5.3a (https://github.com/alexdobin/STAR)
- HTSeq version 0.9.1-Python-3.6.1 (https://htseq.readthedocs.io/en/release_0.11.1/install.html)
- Samtools version1.5 (http://www.sthda.com/english/wiki/install-samtools-on-unix-system)
- R/ R Studio version 3.6.1 (https://www.r-project.org/)

R System Libraries to install when installing R (https://cran.r-project.org/doc/manuals/r-release/R-admin.html)

- Base
- datasets
- graphics
- grDevices
- grid
- grid Extra
- methods

- parallel
- stats
- stats4
- utils

R Packages (Where a package location is not given, the package and documentation are found through Bioconductor and have documentation there. https://www.bioconductor.org/ The others are mostly available through Cran. If you have a license for R, you can use those. One is from github, but is also open source)

- Basic Bioconductor packages to install, when installing Bioconductor. (https://www.bioconductor.org/)
- BiocParallel
- BiocGenerics
- Biobase
- devtools (https://cran.r-project.org/web/packages/devtools/index.html)
- BiocManager (https://cran.r-project.org/web/packages/BiocManager/vignettes/BiocManager.html)
- edgeR
- EnhancedVolcano (https://github.com/kevinblighe/EnhancedVolcano)
- DESeq2
- GO.db
- org.Ce.eg.db
- VennDiagram (https://cran.r-project.org/web/packages/VennDiagram/VennDiagram.pdf)
- RcolorBrewer (https://cran.r-project.org/web/packages/RColorBrewer/index.html)
- ensemble (useMart) (https://www.rdocumentation.org/packages/biomaRt/versions/2.28.0/topics/useMart)
- futile.logger (https://cran.r-project.org/web/packages/futile.logger/futile.logger.pdf)
- ggrepel (https://cran.r-project.org/web/packages/ggrepel/index.html)
- GenomeInfoDb
- GenomicRanges
- ggplot2 (https://cran.r-project.org/web/packages/ggplot2/index.html)
- IRanges
- KnitR (https://cran.r-project.org/web/packages/knitr/index.html)
- limma
- matrixStats (https://cran.rstudio.com/web/packages/matrixStats/index.html)
- S4Vectors
- SummarixedExperiment
- AnnotationDbi