Quick User Guide
RNA-Seq Ver2

In a nutshell

- *This pipeline, has been designed to work for Paired-End RNA-Seq data generated and tested on an Illumina platform, with the primary goal of analyzing the data for differentially expressed genes (DEG) between any two of the specified sample groups. The pipeline is implemented using snakemake and could be run in parallel on cluster of nodes using a batch system.*
- *There are only two inputs by the user: (a) Compressed raw fastq (fastq.gz) files for each of the concatenated R1 and R2 reads per sample, and (b) a complete config file in JSON format (details provided further below)*
- *Based on publications and internal discussions, STAR is the aligner (custom option settings) implemented in this pipeline.*
- *The pipeline allows for the choice of doing read trimming (Trimmomatic)*
- *If trimming is performed, FastQC is run on the post-trimmed data followed by summarization of the FastQC results*
- *Another choice the user has is whether to continue with DEG workflow or pause after post-alignment QC (RNASeQC package) and gene counting*
- *The workflow implements three tools – DESeq2, EdgeR, Limma Voom – to perform analysis for DEGs.*
- *Frequency distribution plots are generated before and after normalization, which is performed as a default setting for each DEG tool.*
- *PCA and MA plots are generated for each of the specified contrasts*

Files under the CCBR/RNASeq Pipeline Repository:

1. **README.md:** general instructions to run the pipeline
2. **dryrun.sh:** fake run to confirm if settings are as intended and functional
3. **submit.sh:** main script to submit to the batch system (qsub)
4. **jobscript.sh:** script called in the main script
5. **rnaseq_wf_v2.py:** the complete code for the RNA-Seq workflow
6. **summarizeFastQCver2:** add-on perl script to summarize FastQC results
7. **TruSeq_and_nextera_adapters_new.fa:** file with the newest set of adapters for Illumina's TruSeq v4 and Nextera protocols
8. **config_example_v2.json:** custom user file that contains directory information, sample metadata, contrasts, tool settings, decision points, etc. (see below)

Configuration file (config.json)

This file should be named config.json. It includes:

- ✓ Samples and contrasts information:
  - ✓ MYFOLDER: path to the working directory
  - ✓ INPUTDIR: path to the location of input files fastq gz compressed files. The name of the files should have the following format: sample_name_R1_all.fastq.gz or sample_name_R2_all.fatsq.gz
  - ✓ MYSAMPLES: list of sample names
  - ✓ GROUPS: list of groups or conditions for the samples (main factors)
  - ✓ CONTRASTS: This a list of paired groups where each pair represent a contrast
- ✓ Genome/GTF files locations
  - ✓ GENOMEFILE: path to the genome reference file
  - ✓ GTFFILE: path to the GTFfile
- ✓ Branching (optional tasks):
  - ✓ TRIM: "yes" or "no" to call TRIMMOMATIC
  - ✓ RESUME: "yes" or "no". When trimming is requested, it is followed by a QC call by running FASTQC and assessing the results using a pre-defined rule. If the QC fails, the pipeline will stop. The user could ignore the current status and resume the pipeline to the next step by calling the pipeline again with option RESUME set to yes
  - ✓ DEG: "yes" or "no". The user has the option to limit the pipeline to Alignment and gene counting by setting this option to "no" otherwise differentially expressed genes will be computed using 3 methods: Deseq2, EdgeR and Limma Voom.
- ✓ Tools versions and settings: the pipeline include the current tools or packages: Trimmomatic, Fastqc, Star (2 passes), Subread, Picard, Rnaseqc, DEseq2, EdgeR and Limma Voom. The user needs to specify the tool version as well as the main settings (please see below for current parameters)

Example of Parameters and Options in a Config File (JSON Format):

```
{
"IOoptions": "----------Output and input locations ----------",
"MYFOLDER": "full path to directory folder for result output",
"INPUTDIR": " full path to directory folder containing input files ",
"MYSAMPLES": ["Ctrl1", "Ctrl2", "Expt1", "Expt2"],
"GROUPS": ["Control", "Control", "Experiment","Experiment"], # maintain order here between
'mysamples' and 'groups'
"CONTRASTS": ["Experiment"," Control"], # this will do Experiment vs Control comparison;
## If for whatever reason, one wanted Control vs Experiment – specify ["Control","Experiment"]

"Options": "---------- Branching options ----------",
"TRIM": "no",
"DEG": "yes",
"RESUME": "no",
```

```
"Annotations": "---------- GENOME and Annotations files----------",
"GENOMEFILE": "…/GRCm38.p3.genome.fa", # Mouse mm10 here
"GTFFILE": "./gencode.vM4.annotation.gtf",
"Trimsettings": "---------- TRIMMOMATIC version and parameters ----------",
"TRIMMOMATICVER": "trimmomatic/0.32", # here we are using the module version to load
"FASTAWITHADAPTERSETC": "pipeline/TruSeq_and_nextera_adapters_new.fa", # in the repository
"SEEDMISMATCHES": 3,
"PALINDROMECLIPTHRESHOLD": 30,
"SIMPLECLIPTHRESHOLD": 10,
"WINDOWSIZE": 4,
"WINDOWQUALITY": 27,
"LEADINGQUALITY": 10,
"TRAILINGQUALITY": 10,
"CROPLENGTH": 0,
"HEADCROPLENGTH": 0,
"MINLEN": 25,
"TARGETLENGTH": 50,
"STRICTNESS": 0.8,
"Pre-alignment QC Settings": "------FASTQC version and parameters and QC check program version-----",
"FASTQCVER": "fastqc/0.10.1",
"CHKQCVER": " pipeline/summarizeFastQCver2", # in the repository
"STARsettings": "---------- STAR version and parameters ----------",
"STARDIR": "/fdb/STAR/GRCm38.annotation.100", # path the genome directory for the STAR first pass
"STARVER": "STAR/2.4.0d",
"SJDBOVERHANG": 100,
"OUTSAMUNMAPPED":  "Within",
"ADAPTER1": "CTGTCTCTTATACACATCTCCGAGCCCACGAGAC",
"ADAPTER2": "CTGTCTCTTATACACATCTGACGCTGCCGACGA",
"FILTERINTRONMOTIFS": "RemoveNoncanonicalUnannotated",
"SAMSTRANDFIELD": "None",
"ENCODE "------- options (for now we have defaults except first) ----------",
"FILTERTYPE": "BySJout",
"FILTERMULTIMAPNMAX": 10,
"ALIGNSJOVERHANGMIN": 5,
"ALIGNSJDBOVERHANGMIN": 3,
"FILTERMISMATCHNMAX": 10,
"FILTERMISMATCHNOVERLMAX": 0.3,
"ALIGNINTRONMIN": 21,
"ALIGNINTRONMAX": 0,
"ALIGNMATESGAPMAX": 0,
"wig": "--------------- options (for now default) ----------",
"WIGTYPE": "None",
"WIGSTRAND": "Stranded",
"Picardsettings": "---------- PICARD version ----------",
"PICARDVER": "picard/1.119",
"subreadsettings": "---------- SUBREAD version and parameters ----------",
"SUBREADVER": "subread/1.4.6",
"strandinfo": "------- 0=unstranded--1=stranded--2=reverse---------- ",
"STRANDED": 0,
"PostAlignQCsettings": "---------- RNASeQC version and parameters ----------",
"BWAVER": "bwa/0.7.10",
```

```
"RNASEQCVER": "/usr/local/apps/rnaseqc/current/RNA-SeQC_v1.1.8.jar",
"RRNALIST": "-", #rRNA interval file specification
“Filtering”: “----------filtering low counts----------“
"MINCOUNT": 5, # only those genes with more than 5 reads taken into analysis
"MINSAMPLES": 1 # shared by at least one sample
}
```

Main pipeline output

- **Trimming results:** folder trim
- **Post-trimming QC:** folder postTrimQC and file FastqcSummary.xlsx
- **Star and Picard results:** folder STARINDEX, *.p2.Aligned.out.sam, and *.star_*rg*_added.sorted.dmark.bam
- **Post-Alignment QC:** folder STAR_QC
- **Subread results:** *.star.count.txt, RawCountFile.txt, distribution plot before normalization
- **DESeq2, EdgeR & Limma results:** normalized count files, list of differentially expressed genes (DEG*.txt), distribution plot after normalization, PCA and MA plots, heatmaps (*.png)