**Pipeline overview**

The TaRGET II RNA-seq pipeline is designed for standardized data processing of mouse RNA-seq samples. It incorporates automatic quality control, generates user friendly files for computational analysis and outputs genome browser tracks for data visualization. To ensure consistent and reproducible data processing, the entire workflow, associated software and libraries are built into a singularity image, which can be run on computational clusters with job submission as well as on stand-alone machines. Pipeline installation requires minimal user input. All the software and genome references used for TaRGET II RNA-seq data processing are included in the pipeline image. The pipeline supports both single- and paired-end data, it accepts FASTQ files, performs alignments, gene features summary and data visualization. The major outputs produced by the pipeline are:

1. JSON file with quality control measurement of user supplied dataset.
2. Log file with processing status of each step.
3. Processed files after alignment (.bam) and summary files of gene features.
4. Bigwig file for visualization purpose.

**Usage of the pipeline**

An easy 2-step guideline for TaRGET II RNA-seq pipeline usage:

1. Download the singularity image of the pipeline.
2. Run the code below from the same directory as your FASTQ file:

singularity run -B ./:/process <path-to-image> -r <SE/PE> -g <mm10 > -o <read\_file1> -p <read\_file2>

**Software Used in the Pipeline**

cutadapt (v1.16) was used to find and remove adapter sequence, and other types of unwanted sequence from high-throughput sequencing reads.

fastqc (v0.11.7) was used to provide quality control check s on raw sequence data.

star (v2.5.4b) was used to map sequence reads against the mouse reference genome.

RSeQC was used to comprehensively evaluate high throughput sequence data.

featureCounts (v1.5.1) was used to count reads to genomic features such as genes.

RSEM (v1.3.0) was used to estimate gene and isoform expression levels from RNA-Seq data.

samtools (v1.9) and bedtools (v2.29) were used to manipulate the sequence alignments and genome features.

**Genomic References Used in the Pipeline**

The following mouse genome and gene references were built into the singularity image and used for TaRGET RNA-seq data processing:

1. STAR indexes of mouse genome (mm10).
2. Gtf file of vM20 gene annotation from GENCODE.
3. RSEM references using GENCODE vM20 annotations.
4. Chromosome size file of mm10.