

Methods

[Jump to bottom](#)

John Vivian edited this page on Mar 13, 2019 · 6 revisions

Tools

Tool	Version	Description
FastQC	0.11.5	Obtains quality metrics on each FASTQ input file.
CutAdapt	1.9	Adapter trimming and quality checking by enforcing fastq samples are properly paired.
STAR	2.4.2a	Aligns fastq samples to the genome. Produces transcriptome bam for RSEM, and can optionally generate a genome-aligned bam and BigWig files.
RSEM	1.2.25	Performs quantification of RNA-seq data to produces count values for genes and isoforms.
Kallisto	0.43.1	Performs quantification of RNA-seq data to produces counts for isoforms directly from fastq data.
Hera	1.1	Performs quantification of RNA-seq data to produces counts for isoforms directly from fastq data.

All tool containers can be found on our [quay.io account](#).

Reference Data

HG38 (no alternative sequences) was downloaded from [NCBI](#). The PAR locus on the Y chromosome, which has duplicate sequences relative to the X chromosome, were removed. chrY:10,000-2,781,479 chrY:56,887,902-57,217,415. This was a requirement in order to run Kallisto. This locus is not removed by the pipeline, and was manually removed. To get this manually modified reference genome, use the `s3cmd` tool with the `requester-pays` option and download: `s3://cgl-pipeline-inputs/rnaseq_cgl/hg38_no_alt.fa`.

Gencode v23 annotations were downloaded from [Gencode](#). Comprehensive gene annotation (Regions=CHR) GTF was used to generate reference input data.

STAR index was created using the reference genome and annotation file with the following Docker command: `sudo docker run -v $(pwd):/data quay.io/ucsc_cgl/star --runThreadN 32 --runMode genomeGenerate --genomeDir /data/genomeDir --genomeFastaFiles hg38.fa --sjdbGTFfile gencode.v23.annotation.gtf`

RSEM reference was created using the reference genome and annotation file with the following Docker command: `sudo docker run -v $(pwd):/data --entrypoint=rsem-prepare-reference quay.io/ucsc_cgl/rsem -p 4 --gtf gencode.v23.annotation.gtf hg38.fa hg38`

Kallisto index was created using the transcriptome and annotation file with the following Docker command: `sudo docker run -v $(pwd):/data quay.io/ucsc_cgl/kallisto index -i hg38.gencodeV23.transcripts.idx transcriptome_hg38_gencodev23.fasta`

Tool Options

- FastQC is run with default options
- CutAdapt is run with default options
- Kallisto is run with `bootstraps` set to 100 and with the `--fusion` flag
- STAR parameters came from ENCODE's [DCC pipeline](#)
- Hera is run with `bootstraps` set to 100 and and bam output suppressed (`-w 1`)

STAR

```
'--outFileNamePrefix', 'rna',
'--outSAMtype', 'BAM', 'SortedByCoordinate',
'--outSAMunmapped', 'Within',
'--quantMode', 'TranscriptomeSAM',
'--outSAMattributes', 'NH', 'HI', 'AS', 'NM', 'MD',
'--outFilterType', 'BySJout',
'--outFilterMultimapNmax', '20',
'--outFilterMismatchNmax', '999',
'--outFilterMismatchNoverReadLmax', '0.04',
'--alignIntronMin', '20',
'--alignIntronMax', '1000000',
'--alignMatesGapMax', '1000000',
'--alignSJoverhangMin', '8',
'--alignSJDBoverhangMin', '1',
'--sjdbScore', '1'
```

RSEM

```
'--quiet',  
'--no-qualities',  
'-p', str(cores),  
'--forward-prob', '0.5',  
'--seed-length', '25',  
'--fragment-length-mean', '-1.0',  
'--bam', '/data/transcriptome.bam',
```

▼ Pages 7[Home](#)[Autoscaling Example Using AWS](#)[Examples](#)[Methods](#)[Troubleshooting](#)[Wiki](#)[Workflow Inputs](#)**Clone this wiki locally**