SCAP-T RNA Sequencing Analysis Protocol
Nov 2014

This document provides the details on of how the processing of the next generation sequencing data from the RNA-sequencing of the single cell samples analysis was carried out at University of Pennsylvania using the *PennSCAP-T Pipeline*.

To get the more information and to download the *PennSCAP-T pipeline* please visit: https://github.com/safisher/ngs/wiki

Software versions used:

PennSCAP-T pipeline: 2.0.1-alpha
STAR: 2.4.0d
HTSeq: 0.6.1
ncbi-blast: 2.2.29+
FastQC: 0.11.2
Samtools: 0.1.19

Human Genome Assembly:

hg38 (soft mask) from UCSC Genome Browser
Date downloaded: September 2014

Gene models used:

Gencode GTF Release 21 (reference chromosomes only) was downloaded and only transcripts assigned confidence level 1 (transcripts with experimentally validated splice junctions or pseudogenes predicted by three different groups within the Gencode consortium) or level 2 (manual annotation from Havana). The other transcripts with level 3 (automatic annotation from Ensembl) were filtered out. Doing so also resulted in the removal of all mitochondrial genes since gencode used Ensembl annotation for these genes and hence their pipeline assigns level 3 to these genes. Since mitochondrial genes are very well known and annotated we added them back to our GTF file
Date downloaded: September 2014

Exonic Quantifications:

HTSeq was run to quantify the exons as annotated in the above GTF file in an un-stranded manner.

Intronic Quantifications:

Using the exon annotations in the above GTF file, we generated a GTF for introns and used that as input to HTSeq for intron quantifications.

STAR Genome

To carry out the alignments we first generated the STAR genome using the hg38 genome sequences as well as the ERCC sequences. The ERCC GTF was appended to original Gencode GTF Release 21 (reference chromosomes only) and this was used to build the splice junction database.

Commands:

Listed below are the commands that are executed for each sample by the *PennSCAP-T pipeline*. The samples from all three centers (PENN, USC and UCSD) are processed in the exact same way except for differences in the contaminants that need to be trimmed or filtered as a result of the differences in the experimental protocols.

Instance of running the PennSCAP-T pipeline on U. Penn sample SC_PENN_1:

```
# COMMAND: ngs.sh PIPELINE -t RNASeq -p 6 -s hg38.gencode21 -c PENN
SC_PENN_1
#################################################################
# BEGIN: PAIRED-END, RNASeq PIPELINE
#################################################################
# BEGIN: INIT PAIRED-END
mkdir SC_PENN_1/init
zcat raw/SC_PENN_1/CM27_ATCACG_L001_R1_001.fastq.gz
raw/SC_PENN_1/CM27_ATCACG_L001_R1_002.fastq.gz
raw/SC_PENN_1/CM27_ATCACG_L001_R1_003.fastq.gz raw/SC_PE
NN_1/CM27_ATCACG_L001_R1_004.fastq.gz
raw/SC_PENN_1/CM27_ATCACG_L001_R1_005.fastq.gz
raw/SC_PENN_1/CM27_ATCACG_L001_R1_006.fastq.gz > SC_PENN_1/init/unalig
ned_1.fq
zcat raw/SC_PENN_1/CM27_ATCACG_L001_R2_001.fastq.gz
raw/SC_PENN_1/CM27_ATCACG_L001_R2_002.fastq.gz
raw/SC_PENN_1/CM27_ATCACG_L001_R2_003.fastq.gz raw/SC_PE
NN_1/CM27_ATCACG_L001_R2_004.fastq.gz
raw/SC_PENN_1/CM27_ATCACG_L001_R2_005.fastq.gz
raw/SC_PENN_1/CM27_ATCACG_L001_R2_006.fastq.gz > SC_PENN_1/init/unalig
ned_2.fq;
# INIT ERROR CHECKING: RUNNING
# INIT ERROR CHECKING: DONE
# FINISHED: INIT PAIRED-END
#################################################################

#################################################################
# BEGIN: FASTQC
mkdir SC_PENN_1/fastqc
# fastqc -v | awk '{print $2}' | sed s/v//
fastqc --OUTDIR=SC_PENN_1/fastqc SC_PENN_1/init/unaligned_1.fq
mv SC_PENN_1/fastqc/unaligned_1.fq_fastqc.html
SC_PENN_1/fastqc/SC_PENN_1.fastqc.html
# FINISHED: FASTQC
#################################################################

#################################################################
```

```
# BEGIN: BLAST
mkdir SC_PENN_1/blast
randomSample.py 5000 4 SC_PENN_1/init/unaligned_1.fq
SC_PENN_1/blast/raw.fq > SC_PENN_1/blast/sampling.out.txt
awk 'BEGIN{P=1}{if(P==1||P==2){gsub(/^[@]/,">");print}; if(P==4)P=0;
P++}' SC_PENN_1/blast/raw.fq > SC_PENN_1/blast/raw.fa
blastn -query SC_PENN_1/blast/raw.fa -db nt -num_descriptions 10 -
num_alignments 10 -word_size 15 -gapopen 3 -gapextend 1 -evalue 1e-15 -
num_threads 6 -out
 SC_PENN_1/blast/blast.txt
parseBlast.py hg38.gencode21 SC_PENN_1/blast/raw.fa
SC_PENN_1/blast/blast.txt
mv SC_PENN_1/blast/speciesCounts.txt
SC_PENN_1/blast/SC_PENN_1.blast.stats.txt
# BLAST ERROR CHECKING: RUNNING
# BLAST ERROR CHECKING: DONE
# blastn version: blastn -version | tail -1 | awk '{print $3}' | sed
s/,//
# parseBlast.py version: grep parseBlast
SC_PENN_1/blast/SC_PENN_1.blast.stats.txt | awk -F: '{print $2}'
# FINISHED: BLAST
#####################################################################
ngsArgs_TRIM -m 20 -q 53 -rAT 26 -rN -c
/home/mugdhak/repo/resources/trim/contaminants_PENN.fa SC_PENN_1
#####################################################################
# BEGIN: TRIMMING PAIRED-END
mkdir SC_PENN_1/trim
# trimReads.py version: trimReads.py -v 2>&1
trimReads.py -m 20 -q 53 -rN -rAT 26 -c
/home/mugdhak/repo/resources/trim/contaminants_PENN.fa -f
SC_PENN_1/init/unaligned_1.fq -r SC_PENN_1/init/unaligned
_2.fq -o SC_PENN_1/trim/unaligned >
SC_PENN_1/trim/SC_PENN_1.trim.stats.txt
cp /home/mugdhak/repo/resources/trim/contaminants_PENN.fa
SC_PENN_1/trim/.
# TRIM ERROR CHECKING: RUNNING
# TRIM ERROR CHECKING: DONE
# FINISHED: TRIMMING PAIRED-END
#####################################################################

#####################################################################
# BEGIN: STAR PAIRED-END ALIGNMENT
mkdir SC_PENN_1/star
STAR --genomeDir
/home/mugdhak/repo/resources/star_hg38softmask_ERCC/hg38.gencode21 --
outFilterScoreMin 0 --outFilterScoreMinOverLread 0 --outFilterMatchNm
in 30 --outFilterMismatchNmax 100 --outFilterMismatchNoverLmax 0.3 --
outSAMunmapped Within --genomeLoad LoadAndRemove --
outFilterMatchNminOverLread 0.4 --r
eadFilesIn SC_PENN_1/trim/unaligned_1.fq SC_PENN_1/trim/unaligned_2.fq
--runThreadN 6 --outFileNamePrefix SC_PENN_1/star/
# STAR POST PROCESSING
CUR_DIR=/home/mugdhak/repo/u01/PENN/Penn_dbGaP_Ver1_Submission
JOURNAL_SAV=SC_PENN_1/log/2014-10-18_16-57.PIPELINE.log
cd SC_PENN_1/star
JOURNAL=../../../../SC_PENN_1/log/2014-10-18_16-57.PIPELINE.log
# converting SAM output to sorted BAM file
samtools view -h -b -S -o STAR.bam Aligned.out.sam
```

```
samtools sort STAR.bam SC_PENN_1.star.sorted
samtools index SC_PENN_1.star.sorted.bam
# generating STAR_Unique.bam file
samtools view -H -S Aligned.out.sam > header.sam
samtools view -S -F 0x4 Aligned.out.sam | grep -P 'NH:i:1
rm header.sam
rm STAR.bam Aligned.out.sam
mv Log.final.out SC_PENN_1.star.stats.txt
cd /home/mugdhak/repo/u01/PENN/Penn_dbGaP_Ver1_Submission
JOURNAL=SC_PENN_1/log/2014-10-18_16-57.PIPELINE.log
# STAR ERROR CHECKING: RUNNING
# STAR ERROR CHECKING: DONE
# STAR version: head -1 SC_PENN_1/star/Log.out | awk -F= '{print $2}'
# samtools version: samtools 2>&1 | grep 'Version:' | awk '{print $2}'
# FINISHED: STAR PAIRED-END ALIGNMENT
#####################################################################
#####################################################################
# BEGIN: HTSEQ
mkdir SC_PENN_1/htseq
# HTSeq version: python -c 'import HTSeq, pkg_resources; print
pkg_resources.get_distribution("HTSeq").version'
python -m HTSeq.scripts.count --format=bam --order=pos --
mode=intersection-nonempty --stranded=no --type=exon --idattr=gene_id
SC_PENN_1/star/SC_PENN_1.sta
r.unique.bam /home/mugdhak/repo/resources/htseq/hg38.gencode21.gz >
SC_PENN_1/htseq/SC_PENN_1.htseq.out 2>&1
(my_python)[mugdhak@node061 u01]$ awk -F\\t '{print $2}'
PENN/Penn_dbGaP_Ver1_Submission/SC_PENN_1/log/2014-10-18_23-
59.PIPELINE.log |more
# COMMAND: ngs.sh PIPELINE -t RNASeq -p 6 -s hg38.gencode21 -c PENN
SC_PENN_1
#####################################################################
# BEGIN: PAIRED-END, RNASeq PIPELINE
#####################################################################
# BEGIN: HTSEQ
# HTSeq version: python -c 'import HTSeq, pkg_resources; print
pkg_resources.get_distribution("HTSeq").version'
python -m HTSeq.scripts.count --format=bam --order=pos --
mode=intersection-nonempty --stranded=no --type=exon --idattr=gene_id
SC_PENN_1/star/SC_PENN_1.sta
r.unique.bam /home/mugdhak/repo/resources/htseq/hg38.gencode21.gz >
SC_PENN_1/htseq/SC_PENN_1.htseq.out 2>&1
# splitting output file into counts, log, and error files
grep 'Warning' SC_PENN_1/htseq/SC_PENN_1.htseq.out >
SC_PENN_1/htseq/SC_PENN_1.htseq.err.txt
grep -v 'Warning' SC_PENN_1/htseq/SC_PENN_1.htseq.out >
SC_PENN_1/htseq/tmp.txt
echo -e 'gene
grep -P '
grep -P -v '
grep -P
'no_feature|ambiguous|too_low_aQual|not_aligned|alignment_not_unique'
SC_PENN_1/htseq/tmp.txt >> SC_PENN_1/htseq/SC_PENN_1.htseq.log.txt
rm SC_PENN_1/htseq/SC_PENN_1.htseq.out SC_PENN_1/htseq/tmp.txt
# HTSEQ ERROR CHECKING: RUNNING
# HTSEQ ERROR CHECKING: DONE
# FINISHED: HTSEQ
#####################################################################
```

```
#####################################################################
# BEGIN: POST PROCESSING
rm SC_PENN_1/trim/unaligned_1.fq SC_PENN_1/trim/unaligned_2.fq
# FINISHED: POST PROCESSING
#####################################################################

# FINISHED: PAIRED-END, RNASeq PIPELINE
#####################################################################
```

Instance of running the PennSCAP-T pipeline on USC sample SC_USC_1:

```
# COMMAND: ngs.sh PIPELINE -t RNASeq -p 6 -s hg38.gencode21 -se -c USC
SC_USC_1
#####################################################################
# BEGIN: SINGLE-END, RNASeq PIPELINE
#####################################################################
# BEGIN: INIT SINGLE-END
mkdir SC_USC_1/init
zcat raw/SC_USC_1/C4DRMACXX-RD-NRW27T10-EB19_GTGGCC_L002_R1_.fastq.gz >
SC_USC_1/init/unaligned_1.fq
# INIT ERROR CHECKING: RUNNING
# INIT ERROR CHECKING: DONE
# FINISHED: INIT SINGLE-END
#####################################################################

#####################################################################
# BEGIN: FASTQC
mkdir SC_USC_1/fastqc
# fastqc -v | awk '{print $2}' | sed s/v//
fastqc --OUTDIR=SC_USC_1/fastqc SC_USC_1/init/unaligned_1.fq
mv SC_USC_1/fastqc/unaligned_1.fq_fastqc.html
SC_USC_1/fastqc/SC_USC_1.fastqc.html
# FINISHED: FASTQC
#####################################################################

#####################################################################
# BEGIN: BLAST
mkdir SC_USC_1/blast
randomSample.py 5000 4 SC_USC_1/init/unaligned_1.fq
SC_USC_1/blast/raw.fq > SC_USC_1/blast/sampling.out.txt
awk 'BEGIN{P=1}{if(P==1||P==2){gsub(/^[@]/,">");print}; if(P==4)P=0;
P++}' SC_USC_1/blast/raw.fq > SC_USC_1/blast/raw.fa
blastn -query SC_USC_1/blast/raw.fa -db nt -num_descriptions 10 -
num_alignments 10 -word_size 15 -gapopen 3 -gapextend 1 -evalue 1e-15 -
num_threads 6 -out
SC_USC_1/blast/blast.txt
parseBlast.py hg38.gencode21 SC_USC_1/blast/raw.fa
SC_USC_1/blast/blast.txt
mv SC_USC_1/blast/speciesCounts.txt
SC_USC_1/blast/SC_USC_1.blast.stats.txt
# BLAST ERROR CHECKING: RUNNING
# BLAST ERROR CHECKING: DONE
# blastn version: blastn -version | tail -1 | awk '{print $3}' | sed
s/,//
```

```
# parseBlast.py version: grep parseBlast
SC_USC_1/blast/SC_USC_1.blast.stats.txt | awk -F: '{print $2}'
# FINISHED: BLAST
#####################################################################

ngsArgs_TRIM -m 20 -q 53 -rAT 26 -rN -c
/home/mugdhak/repo/resources/trim/contaminants_USC.fa SC_USC_1
#####################################################################
# BEGIN: TRIMMING SINGLE-END
mkdir SC_USC_1/trim
# trimReads.py version: trimReads.py -v 2>&1
mkdir SC_USC_1/fastqc
# fastqc -v | awk '{print $2}' | sed s/v//
fastqc --OUTDIR=SC_USC_1/fastqc SC_USC_1/init/unaligned_1.fq
mv SC_USC_1/fastqc/unaligned_1.fq_fastqc.html
SC_USC_1/fastqc/SC_USC_1.fastqc.html
# FINISHED: FASTQC
#####################################################################

#####################################################################
# BEGIN: BLAST
mkdir SC_USC_1/blast
randomSample.py 5000 4 SC_USC_1/init/unaligned_1.fq
SC_USC_1/blast/raw.fq > SC_USC_1/blast/sampling.out.txt
awk 'BEGIN{P=1}{if(P==1||P==2){gsub(/^[@]/,">");print}; if(P==4)P=0;
P++}' SC_USC_1/blast/raw.fq > SC_USC_1/blast/raw.fa
blastn -query SC_USC_1/blast/raw.fa -db nt -num_descriptions 10 -
num_alignments 10 -word_size 15 -gapopen 3 -gapextend 1 -evalue 1e-15 -
num_threads 6 -out
SC_USC_1/blast/blast.txt
parseBlast.py hg38.gencode21 SC_USC_1/blast/raw.fa
SC_USC_1/blast/blast.txt
mv SC_USC_1/blast/speciesCounts.txt
SC_USC_1/blast/SC_USC_1.blast.stats.txt
# BLAST ERROR CHECKING: RUNNING
# BLAST ERROR CHECKING: DONE
# blastn version: blastn -version | tail -1 | awk '{print $3}' | sed
s/,//
# parseBlast.py version: grep parseBlast
SC_USC_1/blast/SC_USC_1.blast.stats.txt | awk -F: '{print $2}'
# FINISHED: BLAST
#####################################################################

ngsArgs_TRIM -m 20 -q 53 -rAT 26 -rN -c
/home/mugdhak/repo/resources/trim/contaminants_USC.fa SC_USC_1
#####################################################################
# BEGIN: TRIMMING SINGLE-END
mkdir SC_USC_1/trim
# trimReads.py version: trimReads.py -v 2>&1
trimReads.py -m 20 -q 53 -rN -rAT 26 -c
/home/mugdhak/repo/resources/trim/contaminants_USC.fa -f
SC_USC_1/init/unaligned_1.fq -o SC_USC_1/trim/unaligned >
SC_USC_1/trim/SC_USC_1.trim.stats.txt
cp /home/mugdhak/repo/resources/trim/contaminants_USC.fa
SC_USC_1/trim/.
# TRIM ERROR CHECKING: RUNNING
# TRIM ERROR CHECKING: DONE
# FINISHED: TRIMMING SINGLE-END
```

```
#####################################################################
#####################################################################
# BEGIN: STAR SINGLE-END ALIGNMENT
STAR --genomeDir
/home/mugdhak/repo/resources/star_hg38softmask_ERCC/hg38.gencode21 --
outFilterScoreMin 0 --outFilterScoreMinOverLread 0 --outFilterMatchNm
in 30 --outFilterMismatchNmax 100 --outFilterMismatchNoverLmax 0.3 --
outSAMunmapped Within --genomeLoad LoadAndRemove --
outFilterMatchNminOverLread 0.6 --r
eadFilesIn SC_USC_1/trim/unaligned_1.fq --runThreadN 6 --
outFileNamePrefix SC_USC_1/star/
# STAR POST PROCESSING
CUR_DIR=/home/mugdhak/repo/u01/USC/USC-Version1-Submission-2014-10-01
JOURNAL_SAV=SC_USC_1/log/2014-10-16_09-24.PIPELINE.log
cd SC_USC_1/star
JOURNAL=../../../../SC_USC_1/log/2014-10-16_09-24.PIPELINE.log
# converting SAM output to sorted BAM file
samtools view -h -b -S -o STAR.bam Aligned.out.sam
samtools sort STAR.bam SC_USC_1.star.sorted
samtools index SC_USC_1.star.sorted.bam
# generating STAR_Unique.bam file
samtools view -H -S Aligned.out.sam > header.sam
samtools view -S -F 0x4 Aligned.out.sam | grep -P 'NH:i:1
rm header.sam
rm STAR.bam Aligned.out.sam
mv Log.final.out SC_USC_1.star.stats.txt
cd /home/mugdhak/repo/u01/USC/USC-Version1-Submission-2014-10-01
JOURNAL=SC_USC_1/log/2014-10-16_09-24.PIPELINE.log
# STAR ERROR CHECKING: RUNNING
# STAR ERROR CHECKING: DONE
# STAR version: head -1 SC_USC_1/star/Log.out | awk -F= '{print $2}'
# samtools version: samtools 2>&1 | grep 'Version:' | awk '{print $2}'
# FINISHED: STAR SINGLE-END ALIGNMENT
#####################################################################
# BEGIN: HTSEQ
# HTSeq version: python -c 'import HTSeq, pkg_resources; print
pkg_resources.get_distribution("HTSeq").version'
python -m HTSeq.scripts.count --format=bam --order=pos --
mode=intersection-nonempty --stranded=no --type=exon --idattr=gene_id
SC_USC_1/star/SC_USC_1.star.
unique.bam /home/mugdhak/repo/resources/htseq/hg38.gencode21.gz >
SC_USC_1/htseq/SC_USC_1.htseq.out 2>&1
# splitting output file into counts, log, and error files
mv SC_USC_1/htseq/SC_USC_1.htseq.out SC_USC_1/htseq/tmp.txt
echo -e 'gene
grep -P '
grep -P -v '
grep -P
'no_feature|ambiguous|too_low_aQual|not_aligned|alignment_not_unique'
SC_USC_1/htseq/tmp.txt >> SC_USC_1/htseq/SC_USC_1.htseq.log.txt
rm SC_USC_1/htseq/SC_USC_1.htseq.out SC_USC_1/htseq/tmp.txt
# HTSEQ ERROR CHECKING: RUNNING
# HTSEQ ERROR CHECKING: DONE
# FINISHED: HTSEQ
#####################################################################

#####################################################################
# BEGIN: POST PROCESSING
```

```
rm SC_USC_1/trim/unaligned_1.fq
# FINISHED: POST PROCESSING
#####################################################################

# FINISHED: SINGLE-END, RNASeq PIPELINE
#####################################################################
```

Instance of running the PennSCAP-T pipeline on UCSD sample SC_UCSD_1:

```
# COMMAND: ngs.sh PIPELINE -t RNASeq -p 6 -s hg38.gencode21 -se -c UCSD
SC_UCSD_1
#####################################################################
# BEGIN: SINGLE-END, RNASeq PIPELINE
#####################################################################
# BEGIN: INIT SINGLE-END
mkdir SC_UCSD_1/init
zcat raw/SC_UCSD_1/20140212-1C94_S96_L001_R1_001.fastq.gz
raw/SC_UCSD_1/20140212-1C94_S96_L002_R1_001.fastq.gz
raw/SC_UCSD_1/20140212-1C94_S96_L003_R1_001.
fastq.gz raw/SC_UCSD_1/20140212-1C94_S96_L004_R1_001.fastq.gz >
SC_UCSD_1/init/unaligned_1.fq
# INIT ERROR CHECKING: RUNNING
# INIT ERROR CHECKING: DONE
# FINISHED: INIT SINGLE-END
#####################################################################

#####################################################################
# BEGIN: FASTQC
mkdir SC_UCSD_1/fastqc
# fastqc -v | awk '{print $2}' | sed s/v//
fastqc --OUTDIR=SC_UCSD_1/fastqc SC_UCSD_1/init/unaligned_1.fq
mv SC_UCSD_1/fastqc/unaligned_1.fq_fastqc.html
SC_UCSD_1/fastqc/SC_UCSD_1.fastqc.html
# FINISHED: FASTQC
#####################################################################

#####################################################################
# BEGIN: BLAST
mkdir SC_UCSD_1/blast
randomSample.py 5000 4 SC_UCSD_1/init/unaligned_1.fq
SC_UCSD_1/blast/raw.fq > SC_UCSD_1/blast/sampling.out.txt
awk 'BEGIN{P=1}{if(P==1||P==2){gsub(/^[@]/,">");print}; if(P==4)P=0;
P++}' SC_UCSD_1/blast/raw.fq > SC_UCSD_1/blast/raw.fa
blastn -query SC_UCSD_1/blast/raw.fa -db nt -num_descriptions 10 -
num_alignments 10 -word_size 15 -gapopen 3 -gapextend 1 -evalue 1e-15 -
num_threads 6 -out
 SC_UCSD_1/blast/blast.txt
parseBlast.py hg38.gencode21 SC_UCSD_1/blast/raw.fa
SC_UCSD_1/blast/blast.txt
mv SC_UCSD_1/blast/speciesCounts.txt
SC_UCSD_1/blast/SC_UCSD_1.blast.stats.txt
# BLAST ERROR CHECKING: RUNNING
# BLAST ERROR CHECKING: DONE
# blastn version: blastn -version | tail -1 | awk '{print $3}' | sed
s/,//
# parseBlast.py version: grep parseBlast
```

```
SC_UCSD_1/blast/SC_UCSD_1.blast.stats.txt | awk -F: '{print $2}'
# FINISHED: BLAST
#####################################################################

ngsArgs_TRIM -m 20 -q 53 -rAT 26 -rN -c
/home/mugdhak/repo/resources/trim/contaminants_UCSD.fa -filter
/home/mugdhak/repo/resources/trim/contaminants_filte
r_UCSD.txt SC_UCSD_1
#####################################################################
# BEGIN: TRIMMING SINGLE-END
mkdir SC_UCSD_1/trim
# trimReads.py version: trimReads.py -v 2>&1
trimReads.py -m 20 -q 53 -rN -rAT 26 -c
/home/mugdhak/repo/resources/trim/contaminants_UCSD.fa -f
SC_UCSD_1/init/unaligned_1.fq -o SC_UCSD_1/trim/unaligned
 > SC_UCSD_1/trim/SC_UCSD_1.trim.stats.txt
rename SC_UCSD_1/trim/unaligned_1.fq SC_UCSD_1/trim/unaligned_1_bF.fq
SC_UCSD_1/trim/unaligned_1.fq
filterReads.pl
/home/mugdhak/repo/resources/trim/contaminants_filter_UCSD.txt
SC_UCSD_1/trim/unaligned_1_bF.fq SC_UCSD_1/trim/unaligned_1.fq
SC_UCSD_1/trim
/SC_UCSD_1.trim.stats.txt > SC_UCSD_1/trim/SC_UCSD_1.filter.stats.txt
rm SC_UCSD_1/trim/unaligned_1_bF.fq
cp /home/mugdhak/repo/resources/trim/contaminants_UCSD.fa
SC_UCSD_1/trim/.
# TRIM ERROR CHECKING: RUNNING
# TRIM ERROR CHECKING: DONE
# FINISHED: TRIMMING SINGLE-END
#####################################################################

#####################################################################
# BEGIN: STAR SINGLE-END ALIGNMENT
mkdir SC_UCSD_1/star
STAR --genomeDir
/home/mugdhak/repo/resources/star_hg38softmask_ERCC/hg38.gencode21 --
outFilterScoreMin 0 --outFilterScoreMinOverLread 0 --outFilterMatchNm
in 30 --outFilterMismatchNmax 100 --outFilterMismatchNoverLmax 0.3 --
outSAMunmapped Within --genomeLoad LoadAndRemove --
outFilterMatchNminOverLread 0.6 --r
eadFilesIn SC_UCSD_1/trim/unaligned_1.fq --runThreadN 6 --
outFileNamePrefix SC_UCSD_1/star/
# STAR POST PROCESSING
CUR_DIR=/home/mugdhak/repo/u01/UCSD/UCSD-Version1-Submission-2014-10-01
JOURNAL_SAV=SC_UCSD_1/log/2014-10-17_13-22.PIPELINE.log
cd SC_UCSD_1/star
JOURNAL=../../../../SC_UCSD_1/log/2014-10-17_13-22.PIPELINE.log
# converting SAM output to sorted BAM file
samtools view -h -b -S -o STAR.bam Aligned.out.sam
samtools sort STAR.bam SC_UCSD_1.star.sorted
samtools index SC_UCSD_1.star.sorted.bam
# generating STAR_Unique.bam file
samtools view -H -S Aligned.out.sam > header.sam
samtools view -S -F 0x4 Aligned.out.sam | grep -P 'NH:i:1
rm header.sam
rm STAR.bam Aligned.out.sam
mv Log.final.out SC_UCSD_1.star.stats.txt
cd /home/mugdhak/repo/u01/UCSD/UCSD-Version1-Submission-2014-10-01
```

```
JOURNAL=SC_UCSD_1/log/2014-10-17_13-22.PIPELINE.log
# STAR ERROR CHECKING: RUNNING
# STAR ERROR CHECKING: DONE
# STAR version: head -1 SC_UCSD_1/star/Log.out | awk -F= '{print $2}'
# samtools version: samtools 2>&1 | grep 'Version:' | awk '{print $2}'
# FINISHED: STAR SINGLE-END ALIGNMENT

#########################################################################
# BEGIN: HTSEQ
mkdir SC_UCSD_1/htseq
# HTSeq version: python -c 'import HTSeq, pkg_resources; print
pkg_resources.get_distribution("HTSeq").version'
python -m HTSeq.scripts.count --format=bam --order=pos --
mode=intersection-nonempty --stranded=no --type=exon --idattr=gene_id
SC_UCSD_1/star/SC_UCSD_1.sta
r.unique.bam /home/mugdhak/repo/resources/htseq/hg38.gencode21.gz >
SC_UCSD_1/htseq/SC_UCSD_1.htseq.out 2>&1
# splitting output file into counts, log, and error files
mv SC_UCSD_1/htseq/SC_UCSD_1.htseq.out SC_UCSD_1/htseq/tmp.txt
echo -e 'gene
grep -P '
grep -P -v '
grep -P
'no_feature|ambiguous|too_low_aQual|not_aligned|alignment_not_unique'
SC_UCSD_1/htseq/tmp.txt >> SC_UCSD_1/htseq/SC_UCSD_1.htseq.log.txt
rm SC_UCSD_1/htseq/SC_UCSD_1.htseq.out SC_UCSD_1/htseq/tmp.txt
# HTSEQ ERROR CHECKING: RUNNING
# HTSEQ ERROR CHECKING: DONE
# FINISHED: HTSEQ
#########################################################################

#########################################################################
# BEGIN: POST PROCESSING
rm SC_UCSD_1/trim/unaligned_1.fq
# FINISHED: POST PROCESSING
#########################################################################

# FINISHED: SINGLE-END, RNASeq PIPELINE
#########################################################################
```