

RNAseq pipeline example

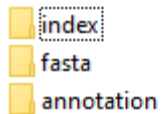
Submit shell scripts

```
>qsub runSTAR_INDEX.sh
```

1. Genome information and index

- fasta file
- index
- gtf-file

/data/projects/2020/AML_KLH/hg38/



runSTAR_INDEX.sh

```
#!/bin/sh
#$ -S /bin/sh
#$ -pe smp 16
#$ -cwd
#$ -V
#### Jobdescription at qstat
#$ -N STAR
#### Error Outputfile
#$ -e STAR.err
#$ -o STAR.log
#### Resubmit
#$ -r y
STAR --runMode genomeGenerate --runThreadN 16 --genomeDir /data/projects/2020/AML_KLH/hg38/index/
--genomeFastaFiles /data/projects/2020/AML_KLH/hg38/fasta/hg38_o.fa --sjdbGTFfile
/data/projects/2020/AML_KLH/hg38/annotation/hg38refGene_o.gtf --sjdbOverhang 100
Done
```

2. Trimming (Trimmomatic) optional

runTRIMMOMATIC.sh

```
#!/bin/sh
#$ -S /bin/sh
#$ -pe smp 10
#$ -cwd
#$ -V
#### Jobdescription at qstat
#$ -N TRIMMO
#### Error Outputfile
#$ -e TRIMMO.err
#$ -o TRIMMO.log
#### Resubmit
#$ -r y
#get files
folder_in=/data/projects/2020/AML_KLH/01_RAW/

ff=.fastq.gz
for files in /data/projects/2020/AML_KLH/01_RAW/*.fastq.gz;
do
  strstart=${#folder_in}
  strl=${#files}-${#ff}-${strstart}
  filename=${files:$strstart:$strl}
  filename+="_trimmed.fastq.gz"
  outfile=/data/projects/2020/AML_KLH/01_TRIMMO/
  outfile+=$filename
  trimmomatic SE -trimlog trimlog -threads 10 $files $outfile
  ILLUMINACLIP:TRUESEQ_adapters.fa:2:30:10 CROP:50 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15
  MINLEN:36
done
```

3. Sequencing quality (FASTQC)

/data/projects/2020/AML_KLH/01_RAW/

```
ml40354_ngs_rna_targrna_rnaaccess_121_bm_fr1_cd4+_20191112_1.clipped.fastq.gz
ml40354_ngs_rna_targrna_rnaaccess_121_bm_fr1_cd4+_20191112_2.clipped.fastq.gz
ml40354_ngs_rna_targrna_rnaaccess_121_bm_fr1_cd8+_20191112_1.clipped.fastq.gz
ml40354_ngs_rna_targrna_rnaaccess_121_bm_fr1_cd8+_20191112_2.clipped.fastq.gz
```

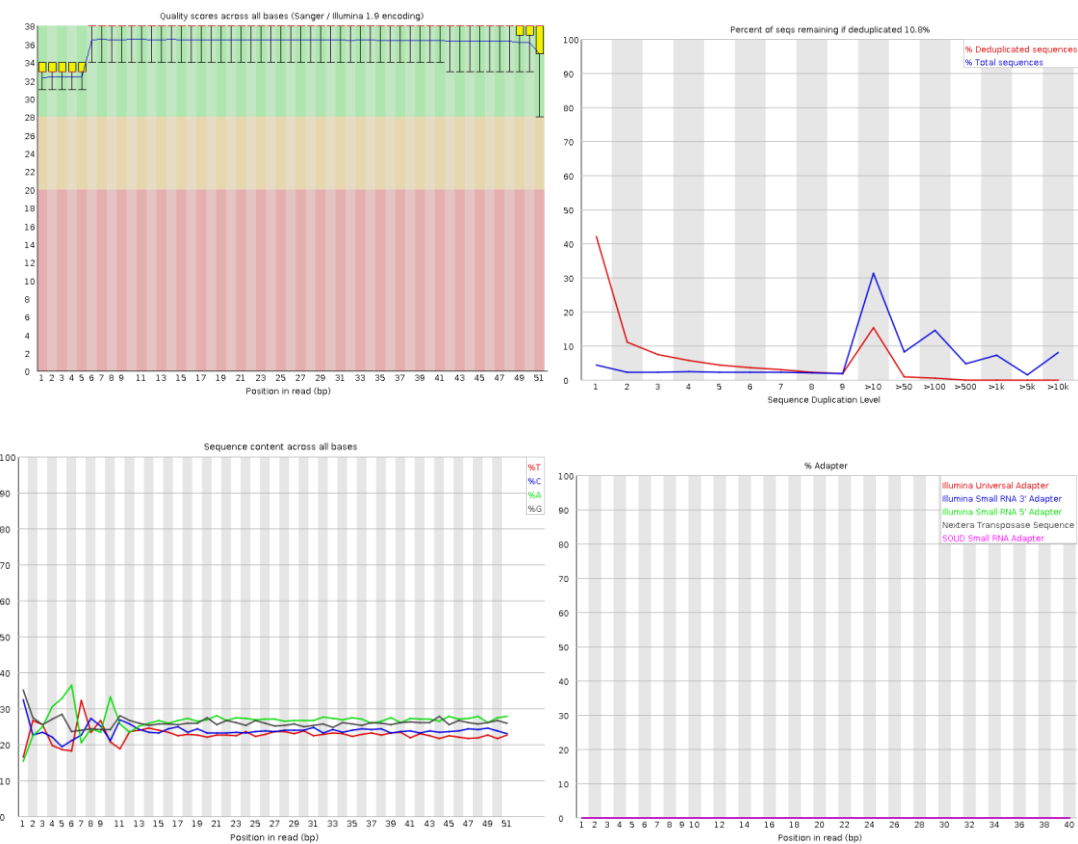
runFASTQC.sh

```
#!/bin/sh
#$ -S /bin/sh
#$ -pe smp 40
#$ -cwd
#$ -V
#### Jobname
#$ -N FASTQC
#### Error Outputfile
#$ -e FASTQC.err
#$ -o FASTQC.log
#### Resubmit
#$ -r y
#working directory
folder="/data/projects/2020/AML_KLH/"
#input files
allfiles=$folder
allfiles+="01_RAW/*.clipped.fastq.gz"
#output folder
outfolder=$folder
outfolder+="02_FASTQC/"
for files in $allfiles;
do
fastqc -t 40 -o $outfolder $files
done
```

/data/projects/2020/AML_KLH/02_FASTQC/

```
ml40354_ngs_rna_targrna_rnaaccess_121_bm_fr1_cd4+_20191112_1.clipped_fastqc.html
ml40354_ngs_rna_targrna_rnaaccess_121_bm_fr1_cd4+_20191112_1.clipped_fastqc.zip
ml40354_ngs_rna_targrna_rnaaccess_121_bm_fr1_cd4+_20191112_2.clipped_fastqc.html
ml40354_ngs_rna_targrna_rnaaccess_121_bm_fr1_cd4+_20191112_2.clipped_fastqc.zip
```

FASTQC Results



Overrepresented sequences

Sequence	Count	Percentage	Possible Source
TTTCTATAGTGTGTACTAGAGAGTTTCTCTGAAGTGTAGAGCACCAGAA	261041	0.5066831787089221	No Hit
CATTTCCTATAGTGTGTACTAGAGAGTTTCTCTGAAGTGTAGAGCACC	247820	0.4810210561421963	No Hit
GTTCCTCTGAAGTGTGTACTAGAGAGTTTCTCTGAAGTGTAGAGCACC	151280	0.29365539553891085	No Hit
CCGAAACACGAGAGAGAGTGTGTCTCTGAAGTGTAGAGCACCAG	149472	0.2905148414376354	No Hit
TTCTATAGTGTGTACTAGAGAGTTTCTCTGAAGTGTAGAGCACCAGAA	134783	0.26163455437158045	No Hit
TTTCAGGAGTATTTCCTATAGTGTGTACTAGAGAGTTTCTCTGAAGTGT	119280	0.2315234669637308	No Hit
TTGAGGATCTTTCCTATAGTGTGTACTAGAGAGTTTCTCTGAAGTGT	117432	0.2283248825965573	No Hit
ATTTCCTATAGTGTGTACTAGAGAGTTTCTCTGAAGTGTAGAGCACCAG	109944	0.2134612182864763	No Hit
GGATCTTTCCTATAGTGTGTACTAGAGAGTTTCTCTGAAGTGTAGAG	103983	0.20183203777065614	No Hit
GATCTTTCCTATAGTGTGTACTAGAGAGTTTCTCTGAAGTGTAGAGC	101142	0.1963564486979324	No Hit
CAGGATCTTTCCTATAGTGTGTACTAGAGAGTTTCTCTGAAGTGTAG	93795	0.18205702838635826	No Hit
GGATCTTTCCTATAGTGTGTACTAGAGAGTTTCTCTGAAGTGTAGAG	84946	0.1648910313269107	No Hit
TCTATAGTGTGTACTAGAGAGTTTCTCTGAAGTGTAGAGCACCAGAA	83515	0.1621034460865367	No Hit
CTGAGAGTCTCTCTCCAGAGTCTCTCTGAAGTGTAGAGCACCAGAA	81935	0.15903665036341239	No Hit
CTATAGTGTGTACTAGAGAGTTTCTCTGAAGTGTAGAGCACCAGAA	79740	0.1547761335202112	No Hit
GAGTTCCTCTGAAGTGTGTAGAGCACCAGAAACACGAGAGAGAGTAG	77181	0.14980908905472057	No Hit
AGCAGGAGAGAGAGAGAGAGAGTGTGTCTCTGAAGTGTAGAGC	67434	0.13089006505896567	No Hit
CACAGAGAGAGAGAGAGTGTGTCTCTGAAGTGTAGAGCACCAGAA	58349	0.11329286854007433	No Hit
GCACGAGAGAGAGAGAGAGTGTGTCTCTGAAGTGTAGAGC	57493	0.11159445513244988	No Hit
CTCTGAAGTGTGTAGAGCACCAGAAACACGAGAGAGAGTGTGTCT	57078	0.11137126098624526	No Hit
GAGAGTTCCTCTGAAGTGTGTAGAGCACCAGAAACACGAGAGAGAGT	55053	0.10685842085136929	No Hit
AGAGCAGCAGAAACACGAGAGAGAGTGTGTCTCTGAAGTGTAGAG	52634	0.10216311777906692	No Hit

4. Read Mapping (STAR Aligner)

/data/projects/2020/AML_KLH/01_RAW/

```
ml40354_ngs_rna_targrna_rnaaccess_121_bm_fr1_cd4+_20191112_1.clipped.fastq.gz
ml40354_ngs_rna_targrna_rnaaccess_121_bm_fr1_cd4+_20191112_2.clipped.fastq.gz
ml40354_ngs_rna_targrna_rnaaccess_121_bm_fr1_cd8+_20191112_1.clipped.fastq.gz
ml40354_ngs_rna_targrna_rnaaccess_121_bm_fr1_cd8+_20191112_2.clipped.fastq.gz
```

runSTAR.sh

```
#!/bin/sh
#$ -S /bin/sh
#$ -pe smp 16
#$ -cwd
#$ -V
#### Jobdescription at qstat
#$ -N STAR
#### Error Outputfile
#$ -e STAR.err
#$ -o STAR.log
#### Resubmit
#$ -r y
#input folder
folder=/data/projects/2020/AML_KLH/01_RAW/
ff= 1.clipped.fastq.gz
fullnames=$folder
fullnames+="*"
fullnames+=$ff
for files in $fullnames;
do
  strstart=${#folder}
  strl=${#files}-${#ff}-strstart
  filename=${files:$strstart:$strl}
  filename+="_"

  STAR --runMode alignReads --runThreadN 16 --genomeDir /data/projects/2020/AML_KLH/hg38/index --
  readFilesIn $files ${files%_1.clipped.fastq.gz}_2.clipped.fastq.gz --readFilesCommand zcat --
  outFilePrefix $filename --outSAMtype BAM SortedByCoordinate --outReadsUnmapped Fastx

Done
```

/data/projects/2020/AML_KLH/03_STAR_UCSC/

```
ml40354_ngs_rna_targrna_rnaaccess_v07_pre_nr_cd8+_20200408_Aligned.sortedByCoord.out.bam
```

5. Read Counting (featureCounts)

/data/projects/2020/AML_KLH/03_STAR_UCSC/

 ml40354_ngs_rna_targrna_rnaaccess_v07_pre_nr_cd8+_20200408_Aligned.sortedByCoord.out.bam

runFEATURE_COUNTS.sh

```
#!/bin/sh
#$ -S /bin/sh
#### -pe smp 10
#$ -cwd
#$ -V
#### Jobdescription at qstat
#$ -N FC
#### Error Outputfile
#$ -e FC.err
#$ -o FC.log
#### Resubmit
#$ -r y
#working directory
folder="/data/projects/2020/AML_KLH/03_STAR_UCSC/"
#input files
allfiles=$folder
ff=".bam"
allfiles+="*"
allfiles+=$ff
gff_file="/data/genomes/hg38/annotation/ucsc/hg38refGene.gtf"
for files in $allfiles;
do
  strstart=${#folder}
  strl=${#files}-${#ff}-${strstart}
  filename=${files:$strstart:$strl}
  filename+="_count"
  featureCounts -T 10 -t exon -s 2 -p -g gene_id -O -a $gff_file -o $filename $files
done
```

/data/projects/2020/AML_KLH/04_FEATURE_COUNTS/

 ml40354_ngs_rna_targrna_rnaaccess_v07_pre_nr_cd8+_20200408_Aligned.sortedByCoord.out_count.summary
 ml40354_ngs_rna_targrna_rnaaccess_v07_pre_nr_cd8+_20200408_Aligned.sortedByCoord.out_count
 ml40354_ngs_rna_targrna_rnaaccess_v07_pre_nr_cd4+_20200408_Aligned.sortedByCoord.out_count.summary
 ml40354_ngs_rna_targrna_rnaaccess_v07_pre_nr_cd4+_20200408_Aligned.sortedByCoord.out_count

6. Generate Count Matrix

targets1.txt

FILENAME	SAMPLE	NAMES	DATE	PAT	SITE	TIME	CELL	RESPONSE	PAIRS					
m140354_1.out_count	176_bm_pre_cd4+	176_BM_PRE_CD4_NR					20200408	176	BM	PRE	CD4	NR	1	
m140354_2.out_count	176_bm_frl_cd4+	176_BM_FRL_CD4_NR					20200408	176	BM	FRL	CD4	NR	1	
m140354_3.out_count	176_pb_pre_cd4+	176_PB_PRE_CD4_NR					20200408	176	PB	PRE	CD4	NR	1	

m140354_1.out_count

```
# Program:featureCounts v2.0.0; Command:"featureCounts" "-T
Geneid Chr      Start  End      Strand Length /data/proje
SLC35A3 chr1;chr1;chr1;chr1;chr1;chr1;chr1;chr1;chr1;c
MFSD14A chr1;chr1;chr1;chr1;chr1;chr1;chr1;chr1;chr1;c
RBP7     chr1;chr1;chr1;chr1;chr1;chr1;chr1;chr1;chr1 999
SASS6    chr1;chr1;chr1;chr1;chr1;chr1;chr1;chr1;chr1;c
RPM113   chr1;chr1;chr1;chr1;chr1;chr1;chr1;chr1;chr1;c
```

get_matrix.R

```
C<-read.table("targets1.txt", header=TRUE, sep="\t", check.names="FALSE")
F<-C
indpb<-which(F$SITE=="PB")
indbm<-which(F$SITE=="BM")
FILES<-as.vector(F$FILENAME)
NAMES<-as.vector(F$NAME)
D<-read.table(FILES[1],header=TRUE,sep="\t", check.names="FALSE", row.names=1)
DF<-D[ order(row.names(D)), ]
CH1<-gsub(".*","",as.vector(DF$Chr))
STR1<-gsub(".*","",as.vector(DF$Strand))
STA1<-gsub(".*","",as.vector(DF$Start))
END1<-gsub(".*","",as.vector(DF$End))
L1<-as.vector(DF$Length)
M<-data.frame(GENE=row.names(DF)[1:28047],CHR=CH1[1:28047],
STRAND=STR1[1:28047],START=STA1[1:28047],END=END1[1:28047],LENGTH=L1[1:28047],COUNTS=DF[1:28047,6
])
for (i in 2:length(FILES)) {
  E<-read.table(FILES[i], header=TRUE, sep="\t", check.names="FALSE", row.names=1)
  EF<-E[order(row.names(E)), ]
  M<-data.frame(M,COUNTS=EF[1:28047,6])
}
names(M)<-c("GENE","CHR","STRAND","START","END","LENGTH", NAMES)
write.table(M,file="COUNT_MATRIX.txt",sep="\t",quote=FALSE, row.names=FALSE)
N<-as.matrix(M[,7:ncol(M)])
K<-N/M$LENGTH
V1<-apply(K,2,sum)/1000000
K1<-t(t(K)/V1)
V2<-apply(K1,2,sum)

TPM<-data.frame(M[,1:6],K1)
names(TPM)<-names(M)
write.table(TPM,file="TPM_MATRIX.txt",sep="\t",quote=FALSE, row.names=FALSE)

LOGTPM1<-data.frame(M[,1:6],log2(K1+1))
names(LOGTPM1)<-names(M)
write.table(LOGTPM1,file="LOG2TPM1_MATRIX.txt",sep="\t",quote=FALSE, row.names=FALSE)
```

COUNT_MATRIX.txt

GENE	CHR	STRAND	START	END	LENGTH	176_BM_PRE_CD4_NR			176_BM_FR1_CD4_NR		
AlBG	chr19	-	58346806		58353499	1766	99		121	108	89
AlBG-AS1		chr19	+	58351970		58355183	2130		33	29	22
AlCF	chr10	-	50799409		50885675	9529	0		0	0	0
A2M	chr12	-	9067708	9116229	5073	24	10	2	0	3	17
A2M-AS1	chr12	+	9065177	9068055	2300	14	34	33	62	78	58

TPM_MATRIX.txt

GENE	CHR	STRAND	START	END	LENGTH	176_BM_PRE_CD4_NR			176_BM_FR1_CD4_NR		
AlBG	chr19	-	58346806		58353499	1766	2.2719935	2783967	2.2719935	2783967	2.2719935
AlBG-AS1		chr19	+	58351970		58355183	2130		0.6279093	22404516	
AlCF	chr10	-	50799409		50885675	9529	0		0	0	0
A2M	chr12	-	9067708	9116229	5073	0.1917383	44780262		0.0740899	577545573	

LOG2TPM1_MATRIX.txt

GENE	CHR	STRAND	START	END	LENGTH	176_BM_PRE_CD4_NR			176_BM_FR1_CD4_NR		
AlBG	chr19	-	58346806		58353499	1766	1.7101698	9456241	1.7101698	9456241	1.7101698
AlBG-AS1		chr19	+	58351970		58355183	2130		0.7030203	41028779	
AlCF	chr10	-	50799409		50885675	9529	0		0	0	0
A2M	chr12	-	9067708	9116229	5073	0.2530675	15896414		0.1031148	27747412	
A2M-AS1	chr12	+	9065177	9068055	2300	0.2101101	65666600		0.6279093	22404516	