

# A simple tutorial using R/Bioconductor to analyze RNA sequencing data

*Chiachun Chiu*

*Oct 26, 2015*

## Download reference genome

Using mouse genome as an example

- [NCBI](#)
- [Ensembl](#)

## Align reads to reference genome (using STAR as aligner)

Downloading and installation of STAR

- Download STAR from [github](#)
- Requirement of STAR:
  - x86-64 compatible processors
  - 64 bit Linux/Mac OS X
  - More than 30 GB of RAM

```
git clone https://github.com/alexdobin/STAR.git
cd STAR
make STAR
# To include STAR-fusion
git submodule update --init --recursive
```

Generating genome indices

- The option `--runMode` directs **STAR** to run genome indices generation job.
- The option `--genomeDir` specifies the path to the directory where the genome indices are stored. This directory has to be created before **STAR** run.
- The option `--sjdbOverhang` specifies the length of the genomic sequence around the annotated junction to be used in constructing the splice junction database, where `ReadLength` is the length of the reads.
- Including chromosome/scaffolds/patches is strongly recommended.
- Using annotations:
  - **GTF**
  - **GFF**
  - **user-defined format**
- For very small genome, the parameter `--genomeSAindexNbases` could be set according to the result:  $\min(14, \log_2(\text{GenomeLength})/2) - 1$
- For the genome with a large number of references ( $>5,000$ ), the parameter `--genomeChrBinNbits` should be reduced (set to  $\min(18, \log_2(\text{GenomeLength}/\text{NumberOfReferences}))$ ) to reduce RAM consumption.

```

refGenomeInfoDir="RefGenome/Ensembl"
genomeVersion="Mus_musculus.GRCm38.82"
annotFile=${genomeVersion}.gtf
refgenomeFile=${genomeVersion}.fa
numThreads=12
ReadLength=75
ReadLength_1=`expr $ReadLength - 1`
runMode="genomeGenerate"
refgenomeParameterDir="ENSEMBL.Mus_musculus.release-82"
readDirs="Whole_Transcriptome"

if [[ ! -d "$refgenomeParameterDir" ]]; then
    mkdir $refgenomeParameterDir
fi

STAR --runThreadN $numThreads \
    --runMode $runMode \
    --genomeDir $refgenomeParameterDir \
    --genomeFastaFiles $refGenomeInfoDir/$refgenomeFile \
    --sjdbGTFfile $refGenomeInfoDir/$annotFile \
    --sjdbOverhang $ReadLength_1

```

## Running mapping jobs

- The option `--readFilesIn` specifies the names of the files containing the sequences to be mapped. **STAR** supports both **FASTA** and **FASTQ** files. If the read files are compressed, use the option `--readFilesCommand UncompressionCommand`.

```

for f in `cat $filelist`;
do
    echo "Now, STAR is processing ${f}.fastq....."
    if [[ $f =~ 'LPS' ]]; then
        exp=`echo $f | awk 'BEGIN {FS="_"} {print $1"_"$2}'`
    else
        exp=`echo $f | awk 'BEGIN {FS="_"} {print $1}'`
    fi

    STAR --genomeDir $refgenomeParameterDir \
        --readFilesIn $readDirs/$exp/${f}.fastq \
        --runThreadN 12 \
        --outFileNamePrefix Aligned/$f
done

```

## Converting SAM files into BAM files using samtools

```

for f in `cat $filelist`;
do
    samtools view -bS Aligned/${f}Aligned.out.sam -o Aligned/${f}.bam
done

```