# RNA-Seq Analysis Pipeline

This repository provides instructions for processing RNA-Seq data from *Drosophila sechellia*. The workflow includes data acquisition, genome indexing, mapping to the reference genome, and calculating gene expression levels using FPKM.

This is an example on how to get the FPKM Tracking files from the raw data (example: GSM1650069: Dsec male TW rep2). Similar for all other files.

## Tools Required

Ensure the following tools are installed before proceeding:

- **STAR**: For genome indexing and read alignment. [sudo apt install rna-star]
- **Cufflinks**: For calculating gene expression levels. [sudo apt install cufflinks]

Optional:

- **SRA Toolkit**: For downloading and converting SRA data.
- **tmux**: For running long processes in a terminal session.
- **wget**: For downloading files from FTP servers.

## Data Acquisition

### RNA-Seq Data

Retrieve RNA-Seq FASTQ files using either `wget` or SRA tools:

#### Option 1: Using `wget`

```
wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR195/003/SRR1952773/SRR1952773_1.fastq.gz
wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR195/003/SRR1952773/SRR1952773_2.fastq.gz
```

#### Option 2: Using SRA Tools

```
prefetch SRR1952773
fastq-dump --split-files --gzip SRR1952773
```

### Reference Genome and Annotations

Retrieve the reference genome and annotation files:

```
wget ftp://ftp.flybase.net/releases/FB2012_06/dsec_r1.3/fasta/dsec-all-chromosome-r1.3.fasta.gz
wget ftp://ftp.flybase.net/releases/FB2012_06/dsec_r1.3/gff/dsec-all-r1.3.gff.gz
```

## Steps

### 1. Create Genome Indexes

Use `STAR` to generate genome indexes:

```
STAR --runThreadN 4 \
     --runMode genomeGenerate \
     --genomeDir star_genome_index \
     --genomeFastaFiles dsec-all-chromosome-r1.3.fasta \
     --sjdbGTFfile dsec-all-r1.3.gff \
     --sjdbOverhang 100 \
     --genomeSAindexNbases 12 \
     --limitGenomeGenerateRAM 16000000000
```

### 2. Map Reads to Reference Genome

Run `STAR` for read alignment. It is recommended to run in a `tmux` session for long processes:

```
STAR --runThreadN 8 \
     --genomeDir star_genome_index \
     --readFilesIn SRR1952773_1.fastq.gz SRR1952773_2.fastq.gz \
     --readFilesCommand zcat \
     --outFileNamePrefix SRR1952773_ \
     --outSAMtype BAM SortedByCoordinate \
     --outSAMattributes NH HI AS nM MD XS
```

#### Useful Commands for `tmux`

- Detach: `Ctrl+b` then `d`
- View resource usage: `top`
- Reattach: `tmux attach`

### 3. Calculate Gene Expression (FPKM)

Use `cufflinks` to compute expression levels:

```
cufflinks -p 8 \
          -G dsec-all-r1.3.gff \
          -o cufflinks_out \
          SRR1952773_Aligned.sortedByCoord.out.bam
```

## Notes

- Ensure sufficient computational resources are available, especially for genome indexing and mapping.
- Adjust parameters like thread counts ( `--runThreadN` ) and memory limits ( `--limitGenomeGenerateRAM` ) based on your system's specifications.

## References

- STAR Manual
- Cufflinks Documentation