1.

Create a directory naming the genome and build identical to ELM

go into to the genome directory and create directory naming “gtf”

$cd /mnt/projects/rpd/genomes.testing/

$mkdir dm3

$cd dm3

$mkdir gtf

2.

Download genome fasta file from ensembl/ucsc/gencode

wget -q [http://hgdownload.soe.ucsc.edu/goldenPath/dm3/bigZips/chromFa.tar.gz](https://smtp.a-star.edu.sg/OWA/redir.aspx?REF=mEPPD-FLYWx-3zX9pJVFY5fQl6eJZ0dDrBMm-kf5kuqVtzBYG6nTCAFodHRwOi8vaGdkb3dubG9hZC5zb2UudWNzYy5lZHUvZ29sZGVuUGF0aC9kbTMvYmlnWmlwcy9jaHJvbUZhLnRhci5neg..).

OR

<http://www.gencodegenes.org/data.html>

OR

<http://www.ensembl.org/info/data/ftp/index.html> (choose organism and fasta/gtf links to go ftp site)

tar -xvzf dm3.tar.gz if multiple.fa files then do

cat all the chromosome \*.fa > dm3.fa

If you need to create the fasta file in lexicographical order sort and list all the .fa files and write to a file (ls \*.fa >list.fa) and arrange it in lexicographical order or sort and save. cat the file to create one .fa file.

cat list.fa | xargs cat >> MacFas5.0.fa

3.

To create bowtie2 and bwa index and fai, dict files run the below command give the prefix of the genome name

sh /mnt/AnalysisPool/libraries/genomes/bin/buildIndices4genomeDB.sh dm3

create star and rsem index

STAR --runMode genomeGenerate --runThreadN 20 --genomeDir /mnt/projects/rpd/genomes.testing/hg19/star99/ --genomeFastaFiles /mnt/projects/rpd/genomes.testing/hg19/hg19.fa --sjdbGTFfile /mnt/projects/rpd/genomes.testing/hg19/gtf/hg19\_annotation.gtf --sjdbOverhang 99

rsem-prepare-reference --gtf hg10\_annotation.gtf  hg19.fa RSEMtr\_hg19

4.

Go into gtf dir and download gtf file from the source to create the below files

hg19\_RNASeqQC\_annotation.gtf

hg19\_annotation.gtf

hg19\_annotation\_mask.gtf

hg19\_annotationdexseq.gff

eg: wget <ftp://ftp.ensembl.org/pub/release-75/gtf/drosophila_melanogaster/Drosophila_melanogaster.BDGP5.75.gtf.gz> .

Gunzip Drosophila\_melanogaster.BDGP5.75.gtf.gz

5.

Remove the lines which contain gene, transcript AND selenocysteine in the gtf files.

awk '{if ($3 != "gene" && $3 != "transcript" && $3 != "Selenocysteine" )  print $0}' gencode.v19.annotation.gtf >gencode.v19.annotation\_filtered.gtf

ln -s gencode.v19.annotation\_filtered.gtf > hg19\_RNASeqQC\_annotation.gtf

egrep -v “rRNA|tRNA”  gencode.v19.annotation\_filtered.gtf > gencode.v19.annotation\_filtered1.gtf

ln -s gencode.v19.annotation\_filtered1.gtf hg19\_annotation.gtf

6.

grep -e rRNA -e tRNA gencode.v19.annotation\_filtered.gtf > gencode.v19.annotation\_filtered\_OnlyrRNAtRNA.gtf

ln -s gencode.v19.annotation\_filtered\_OnlyrRNAtRNA.gtf hg19\_annotation\_mask.gtf

7.

To create dexseq reference

dexseq\_prepare\_annotation.py /mnt/projects/rpd/genomes.testing/hg19/gtf/hg19\_annotation.gtf hg19\_annotationdexseq.gff

Check all the chr names in gtf and genome file matches

========================================================

[http://ensemblgenomes.org/info/access/ftp](https://smtp.a-star.edu.sg/OWA/redir.aspx?REF=QbSmnGhlC7cQMULX942E07R7fAv8RjhJrQB-Sa5I0RiVtzBYG6nTCAFodHRwOi8vZW5zZW1ibGdlbm9tZXMub3JnL2luZm8vYWNjZXNzL2Z0cA..)

[http://asia.ensembl.org/info/data/ftp/rsync.html](https://smtp.a-star.edu.sg/OWA/redir.aspx?REF=vyvTxW27Nyfcy0St80-pxuVIUsUHBZPXsMcj5IzPU6KVtzBYG6nTCAFodHRwOi8vYXNpYS5lbnNlbWJsLm9yZy9pbmZvL2RhdGEvZnRwL3JzeW5jLmh0bWw.)

<ftp://ftp.ensembl.org/pub/current_embl/>

<ftp://ftp.ensembl.org/pub/release-84/fasta/homo_sapiens/dna/>

<ftp://ftp.ensembl.org/pub/release-84/gtf/homo_sapiens/>

Download genome file from ensembl “Homo\_sapiens.GRCh37.dna.primary\_assembly.fa.gz “consists chromosomes and regions not assembled into chromosomes (scaffolds).

OR

Download the individual chromosomes and combine them “Homo\_sapiens.GRCh37.dna.chromosome.\*.fa.gz”