Notes on Running STAR on Galaxy

Since I do not have access to the Harvard cluster to run STAR, I am planning to use galaxy.

- Reference for this page is: https://training.galaxyproject.org/training-material/topics/transcriptomics/tutorials/rna-seq-bash-star-align/tutorial.html

Organism

- Drosophila melanogaster

Data

- Original source: GSE18508, converted by Galaxy folks to .fastqsanger files
- Zenodo - .fastqsanger files have been provided
- They are using 2 samples, one untreated and one sample where PS gene has been knocked out via RNA interference.
- Both samples are paired end.
- Also using the gene annotation file .gtf and reference genome file .fa for drosophila from ncbi genbank https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000001215.4/
    - Rename the .fna file to .fa and use the .fa and .gtf file in galaxy

Notes:

- Initially using usegalaxy.org website
- Using the subsampled smaller files.
- The files are coming with extension .fastqsanger. The convert to datatype is not showing the option to convert them to .fastq. So I have simply changed the

assigned datatype to fastq and changed the name to reflect the .fastq extension

- The files are running fine with fastqc and cutadapt.
- However, I am unable to run the genome index builder with STAR. It's not running correctly and exits without generating all the files. Unable to figure out what the issue is. I am trying with new reference genome and annotation files from NCBI.
- It was clear from looking at several STAR github issues that this was a memory error. The memory available in the usegalaxy.org instance was showing 21GB and number of processors was 8. I tried all the parameters mentioned in the solutions to make STAR genomeGenerate work with smaller memory, like - - limitGenomeGenerateRAM 20000000000, - - genomeSAsparseD 3, but nothing solved the issue.
- Switched to usegalaxy.eu, where number of processors `nproc --all` = 36 and available memory `free -h` is >40GB and the genome generation is working.

Steps

1. Upload the paired end RNA-seq data as .fastq files on the galaxy server
2. Upload the reference genome as .fa file and gene annotations as .gtf file in the galaxy server
3. Launch Rstudio from galaxy tools
4. Import data from galaxy history into Rstudio environment. The following steps are executed in the R console of Rstudio in usegalaxy.eu. The following command downloads the files in the galaxy history and makes it accessible to Rstudio as /import/2, /import/3, /import/4, /import/5, /import/7 and /import/8

```
gx_get(c(2,3,4,5,7,8)
```

The following steps are executed in the terminal of Rstudio

5. Create a conda environment and install the required packages

```
conda create -n name_of_your_env fastqc cutadapt star csamtoo
conda activate name_of_your_env
```

6. Genome_generate_command:

```
mkdir index
STAR --runThreadN 16 --runMode genomeGenerate --genomeDir ~/i
```

7. FastQC and multiqc on raw reads:

```
mkdir qualityRaw
fastqc /import/2 -o qualityRaw
fastqc /import/3 -o qualityRaw
fastqc /import/4 -o qualityRaw
fastqc /import/5 -o qualityRaw
multiqc qualityRaw/ --outdir qualityRaw/ --filename multiqc_r
```

8. Trimming the data using cutadapt

```
cutadapt /import/2 /import/3 -o trimmedData/GSM461177_R1 -p t
cutadapt /import/4 /import/5 -o trimmedData/GSM461180_R1 -p t
```

9. FastQC and multiQC on trimmed Data

```
mkdir qualityTrimmed
fastqc trimmedData/GSM461177_R1 -o qualityTrimmed/
fastqc trimmedData/GSM461177_R2 -o qualityTrimmed/
fastqc trimmedData/GSM461180_R1 -o qualityTrimmed/
fastqc trimmedData/GSM461180_R1 -o qualityTrimmed/
multiqc qualityTrimmed/ --outdir qualityTrimmed/ --filename r
```

10. Running alignment on trimmedData with STAR

```
STAR --genomeDir ~/index --runThreadN 16 -- readFilesIn trimm
STAR --genomeDir ~/index --runThreadN 16 -- readFilesIn trimm
```

11. Convert SAM files to BAM files

```
samtools view -S -b GSM461177Aligned.out.sam > GSM461177Alig
samtools view -S -b GSM461180Aligned.out.sam > GSM461180Alig
```

12. Sort BAM files by co-ordinates

```
samtools sort -o GSM461177Aligned.out.sorted.bam GSM461177Al
samtools sort -o GSM461180Aligned.out.sorted.bam GSM461180Al
```

13. Use featureCounts to count the number of reads/fragments mapped to each annotated gene

```
featureCounts -a /import/8 -T 8 -o featurecounts.txt -p GSM4
# - a is the annotation file
# -T is the number of CPU threads to use
# -o is the name of the output file
# -p specifies that the files contain paired end reads
```

14. Save any required files back to Galaxy history using gx_put() function in R console. make sure the files get saved in your history.
15. Close Rstudio.