Samtools and FeatureCounts

BMS Bootcamp 2019

Schedule

- Recap
- Samtools
- FeatureCounts
- Python
- TPM/RPKM

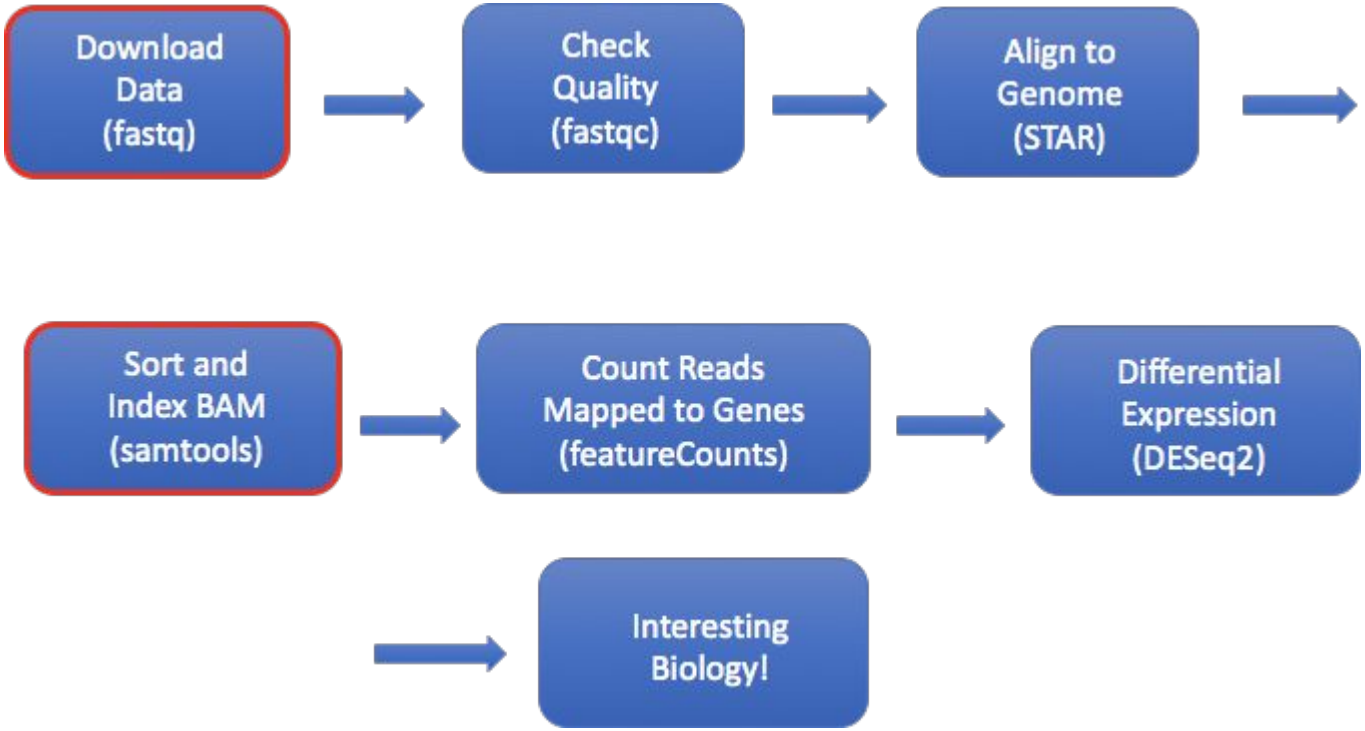
Learning Objectives

Reading documentation

Writing and submitting scripts

Use Python to calculate RPKM

RNA-Seq Pipeline



Samtools

Documentation: http://www.htslib.org/doc/samtools.html

We use this package to sort the DNA sequence alignments we obtained from doing STAR alignment

Can use and convert between SAM (Sequence Alignment/Map) and BAM (Binary Alignment/Map) formats

BAM are compressed and smaller to work with

FeatureCounts

We use FeatureCounts to assign mapped reads to genes and count how many we see in each of the conditions

Documentation:
http://gensoft.pasteur.fr/docs/subread/1.4.6-p3/SubreadUsersGuide.pdf#targetText=The%20featureCounts%20program%20is%20designed,for%20genomic%20features%5B7%5D.

Setting up Jupyter Notebooks

- “The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text. Uses include: data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more.”
  https://jupyter.org/

TPM/RPKM

From RNA-seq data we can normalize our data by sequencing depth and gene length

- **RPKM (Reads Per Kilobase Million)**
- **TPM (Transcripts Per Kilobase Million)**

To analyze the expression changes across samples we use different calculations to normalize the data. We will use out outputs and Python to calculate these different methods.

The sum of TPM in each sample is the same, easier to compare.

RPKM

Steps:

1. Count up total reads per sample and divide by 1,000,000
2. Divide read counts by this scaling factor
3. Divide those values by length of gene

TPM

Steps:

1. Divide the counts by gene length
2. Count up all of the reads and divide by 1,000,000
3. Divide values in 1 by scaling factor in 2

Gene name	Rep 1 read counts	Rep 2	Rep 3
A (2kb)	10	12	30
B (4kb)	20	25	60
C (1kb)	5	8	15
D (10kb)	0	0	1

RPKM

1. Count up total reads per sample and divide by 1,000,000
2. Divide read counts by this scaling factor in Step 1
3. Divide those values by length of gene

TPM

1. Divide the counts by gene length
2. Count up all of the reads and divide by 1,000,000
3. Divide values in 1 by scaling factor in Step 2