Assignment-1

1. In your DSG docker folder, make a soft link for fastq file of RNA-seq profile

Using commans ln –s /storage/vibhor/RNA-seq/SRR16292072.fastq.gz

a. Align it to hg19 genome using Tophat and STAR aligner tools

The indexes for Tophat are in
/storage/vibhor/genomes/hg19/hg19/

The indexes for STAR aligner are at
/storage/vibhor/genomes/hg19/starIndex

the gtf file for genes is here
/storage/vibhor/genomes/hg19/hg19-Gencode.gtf

b. After making bam files with aligner tools run cufflink to get the FPKM based expression of genes (see http://cole-trapnell-lab.github.io/cufflinks/cufflinks/)
c. Run htseq-count to get the read count.

d. Plot the distribution of FPKM of genes in one sample.
Tell what kind of distribution it fits best. Give argument/justification for your answer
e. run cufflink to discover novel transcripts and promoters.

Write all the commands in the answer sheet and write a report on the stats like
Total number of reads , number of reads with alignment.

Compare STAR aligner and Tophat in terms of speed and alignment percentage.

Note : first use command on the docker for the course, first type $bash then type
$ source /storage/vibhor/addpath.sh
to get access to different pre-installed tools, use tophat2

2. a)Make derivations and show relationship between binomial distribution and Poisson distribution
b) Make derivation and show relationship between negative binomial distribution and Poisson distribution
c) Make derivation for variance of poisson distribution
d) Make derivation for variance of negative binomial distribution

3. You are given a set of numbers, assumed to be from a population of values following Poisson distribution

2, 3, 4, 4, 2, 1, 1, 2, 3, 3, 4 , 5, 6, 7, 2, 2, 1, 1, 2

Now find the likelihood and p-value for occurrence of a number 6 , given Possion distribution corresponding the above numbers.

Similarly do it for numbers , 7 and 8.

4. What is central limit theorem, give 2 examples where it helps in analysis in genomic data-science.

5. What is the 3-prime bias in RNA-seq data? How it could be created ?

6. What could be the limitations of bridge PCR methods that are not present in emulsion PCR?

7. What is the difference between MAPQ and CIGAR fields in sam/bam file?