K-mer Analysis and Contamination Exercises.

Starting Note: Please do not copy and paste the commands. Characters in this document may not be copied correctly. Please type the commands and use **<tab> complete** for commands, directories and long names.

Loading Modules:

First do **`module load bioinfo-tools`** and then

KAT: **`module load KAT/2.1.1`**

Kraken: **`module load Kraken/0.10.5-beta`**

Krona: **`module load Krona/2.7`**

1. What is a k-mer?

2. How many 4-mers are in the following sequence?
   `ACGTTTATCCTATACGGTAATAC`

3. What are the frequencies of the 4-mers in the sequence above.

ACGT	CGTT	GTTT	TTTA	TTAT
TATC	ATCC	TCCT	CCTA	CTAT
TATA	ATAC	TACG	ACGG	CGGT
GGTA	GTAA	TAAT	AATA	

4. Use the following commands to get a list of all the k-mers in **`Bacteria/bacteria_R{1,2}.fastq.gz`**.

```
gunzip Bacteria/bacteria_R1.fastq.gz;
gunzip Bacteria/bacteria_R2.fastq.gz;
kat hist -t 8 -d -o bac.hist
```

```
Bacteria/bacteria_R{1,2}.fastq;
jellyfish dump bac.hist-hash.jf27 > kmer.lst
```

The **kmer.lst** file has the following format.

```
>frequency
kmer_sequence
```

How many distinct k-mers were found?

5. How many k-mers have a frequency of 1? Use the following command to find out.

```
paste - - < kmer.lst | cut -c2- | awk '$1 == 1 {
sum++ } END { print sum+0 }'
```

6. How many k-mers have a frequency greater than 5?

7. **kat hist** plotted a histogram in **bac.hist.png**. Open this using **eog**. What is the estimated mean k-mer frequency (k-mer coverage)?

8. The following command prints the frequency of each k-mer frequency between 5 and 45. What is the mean k-mer frequency?

```
paste - - < kmer.lst | cut -c2- | awk '$1 > 5 && $1
< 45 {sum[$1]++ } END { for (freq in sum) {print
freq" "sum[freq]} }' | sort -k1,1n
```

9. Use the following command to plot the gc content vs k-mer frequency.

```
kat gcp -t 8 -o bac.gcp
Bacteria/bacteria_R{1,2}.fastq
```

Open the plot of GC vs coverage using **eog**. On what scale is the GC content measured?

10. Use **kat comp** to compare **Bacteria/bacteria_R{1,2}.fastq**.

```
kat comp -t 8 -o bac_r1vr2 --density_plot
Bacteria/bacteria_R{1,2}.fastq;
kat plot spectra-mx -x 50 -y500000 -n -o
bac_r1vr2-main.mx.spectra_mx.png bac_r1vr2-main.mx
```

Why is there a difference in the distribution means between the two datasets?

11. Run Kraken on **Bacteria/bacteria_R{1,2}.fastq**. What is identified here and why?

```
MINIKRAKEN_DB=/proj/g2016024/nobackup/minikraken_201
41208;
kraken --threads 8 --db $MINIKRAKEN_DB --fastq-input
--paired Bacteria/bacteria_R{1,2}.fastq >
bacteria.kraken.out;
kraken-report --db $MINIKRAKEN_DB
bacteria.kraken.out > bacteria.kraken.rpt;
cut -f2,3 bacteria.kraken.out > bacteria.krona.in;
ktImportTaxonomy bacteria.krona.in -o
bacteria.krona.html
```

12. Run Kraken on **Ecoli/E01_1_135x.fastq.gz**. What is identified here and how does the Pacific Biosciences sequence error rate effect classification?