K-mer Analysis and Contamination

Exercises and Solutions.

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?

   A sequence of nucleotides of length k.

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

   `20 (L-k+1)=(23-4+1)`

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

ACGT 1	CGTT 1	GTTT 1	TTTA 1	TTAT 1
TATC 1	ATCC 1	TCCT 1	CCTA 1	CTAT 1
TATA 1	ATAC 2	TACG 1	ACGG 1	CGGT 1
GGTA 1	GTAA 1	TAAT 1	AATA 1	

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?

```
$ paste - - < kmer.lst | wc -l
```

41531545 (distinct k-mers)

Other methods to calculate this number are viable too e.g.

```
grep -c "^>" kmer.lst
```

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 }'
```

35701246 (distinct k-mers)

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

```
$ paste - - < kmer.lst | cut -c2- | awk '$1 > 5 {
sum++ } END { print sum+0 }'
```

4969071 (distinct k-mers)

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)?

25, since this is approximately where the peak of the distribution is.

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
```

26 has the highest frequency

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?

The GC content scale is the absolute GC count per k-mer.

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?

R2 has lower quality reads meaning more errors or ambiguous bases, and therefore lower k-mer frequency counts shifting the mean k-mer

frequency down.

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
```

No species is identified because the database is not comprehensive enough to classify the bacteria in the sample.

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?

```
$ kraken --threads 4 --db
/opt/byod-data/minikraken_20141208 --fastq-input
--gzip-compressed E01/E01_1_135x.fastq.gz >
E01.kraken.out
$ kraken-report --db
/opt/byod-data/minikraken_20141208 E01.kraken.out >
E01.kraken.rpt
$ cut -f2,3 E01.kraken.out > E01.krona.in
$ ktImportTaxonomy E01.krona.in -o E01.krona.html
```

E. coli is identified as the primary organism. The higher error rate of the

platform reduces classification specificity and increases mis-classification by a small amount.