Data Cleanup: Contaminant Removal

Learning objectives:

1. Identify the need for contaminant sequence removal step in raw sequence files, and what general types of contaminants need to be removed.
2. Recognize the principles underlying software performance of the contaminant removal step by performing a manual contaminant removal on a simplified .fasta data set.

Level of Difficulty: 2/5

When you receive your raw data from the sequencer, it contains a few types of contaminating sequences that should be removed before proceeding to data analysis. The preparation procedure for DNA sequencing with an Illumina sequencer involves adding adaptor nucleotide sequences to the pieces of DNA being sequenced. These will cause problems for your data analysis, so they should be removed. Additionally, a library of calibration sequences called “phiX” (aka Coliphage phi-X174) are often added to check that the sequencer was working correctly, you’ll have to take those out as well. The contaminant removal step is essentially a search and find mission for types of data that don’t belong in your data files.

To do: Locate the Illumina adaptor and Phix contaminant sequences in the below set of 16 reads and cross them out to remove them from your data set. How many reads remain after filtering? What percentage of the original read set does this amount to?

Illumina Adaptor Sequence	**Phix Sequences**
CTGTCT	CAAACATTGGGCCAAATGA
	GGGCGTTGTATGGTTGCCA
	AATACCCCCAGACGTCGGT
	TACCGAGTTTCCGATTCGC

READ ID	SEQUENCE																								
1	C	T	G	T	C	T	A	T	G	A	T	A	A	A	A	G	T	T	C	A	A	T	T	G	A
2	C	T	G	T	C	T	T	G	C	A	A	T	G	C	G	C	T	T	A	T	T	G	A	A	A
3	C	T	G	T	C	T	G	G	G	C	G	T	T	G	T	A	T	G	G	T	T	G	C	C	A
4	C	T	G	T	C	T	A	A	C	G	G	C	T	C	G	G	A	T	C	C	C	T	G	A	A
5	C	T	G	T	C	T	T	A	C	C	G	A	G	T	T	T	C	C	G	A	T	T	C	G	C
6	C	T	G	T	C	T	A	T	C	G	G	T	C	A	A	T	C	T	T	G	G	G	G	G	G
7	C	T	G	T	C	T	T	T	C	A	G	G	G	G	T	C	G	G	T	A	T	A	T	T	G
8	C	T	G	T	C	T	T	G	G	T	C	A	T	G	G	T	T	T	T	G	T	T	G	A	G
9	C	T	G	T	C	T	G	G	G	A	G	T	T	A	A	T	C	C	T	C	C	G	A	T	C
10	C	T	G	T	C	T	T	G	T	A	A	G	G	T	A	A	A	G	A	A	T	G	G	T	A
11	C	T	G	T	C	T	C	A	A	A	C	A	T	T	G	G	G	C	C	A	A	A	T	G	A
12	C	T	G	T	C	T	A	G	T	C	T	T	T	C	T	T	T	C	C	A	A	T	T	T	G
13	C	T	G	T	C	T	G	T	A	G	G	T	T	T	A	G	T	A	A	A	G	C	A	T	G
14	C	T	G	T	C	T	A	C	T	T	G	A	C	C	C	G	G	G	G	C	A	A	A	G	T
15	C	T	G	T	C	T	A	A	T	A	C	C	C	C	C	A	G	A	C	G	T	C	G	G	T
16	C	T	G	T	C	T	A	C	G	A	T	A	G	G	G	A	C	A	C	C	A	G	C	A	A

# of sequences remaining post-contaminant removal: __________

$$\frac{\text{\# of sequences remaining}}{\text{\# of original sequences}} \text{ x } 100\% = \underline{\hspace{3em}} \text{ \% original read set remaining}$$