**Exercise**

1. **Hadoop Kmer Counting**

The exercise will be to implement a Kmer counter using Hadoop. Conceptually this is very similar to the wordcount program, but since there are no spaces in the human genome, we will count overlapping kmers instead of discrete words.

The idea is if the genome is:

>chr1

ACACACAGT

And we are counting 3-mers, your map function will output

ACA 1

CAC 1

ACA 1

CAC 1

ACA 1

CAG 1

AGT 1

The shuffle function will sort them so the same key comes right after each other

ACA 1

ACA 1

ACA 1

CAC 1

CAC 1

CAG 1Yu

AGT 1

And your reducer will output:

ACA 3

CAC 2

CAG 1

AGT 1

You can implement this in Java, using the WordCount program as an example, or you can use Hadoop Streaming to implement it in any language you would like.

The Hadoop Streaming documentation describes how to use it:

<https://hadoop.apache.org/docs/r1.2.1/streaming.html>

And here is a nice tutorial using Python:

<http://www.michael-noll.com/tutorials/writing-an-hadoop-mapreduce-program-in-python/>

The genome file is available here: [ecoli.fa.gz](http://schatz-lab.org/teaching/exercises/hadoop/data/ecoli.fa.gz)

**Question: What are the top 10 most frequently occurring 9-mers in E coli?**