Characterizing Metagenomes Using K-mer Abundance

Beverly A. Hom, Robert A. Edwards

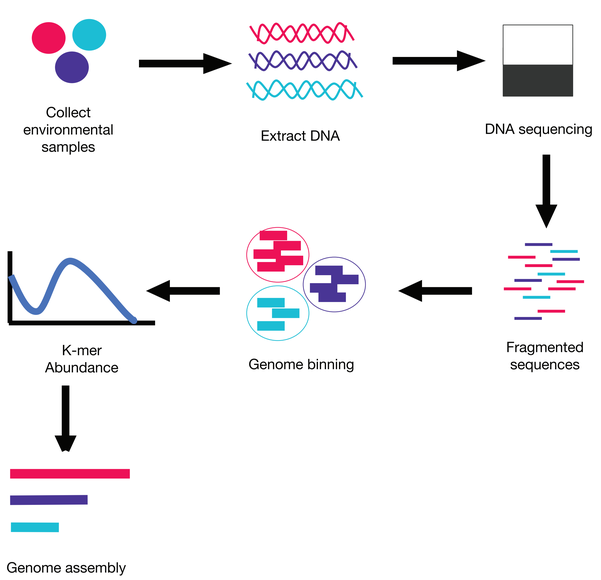
Biomedical Informatics Research Center, San Diego State University

Introduction

Metagenomic methods allow us to understand microbial communities found in environmental samples.

Methods use short sequences of length k for efficient counting. Where k-mers are all substrings of length k in a sequence. Counting k-mers is a preliminary step in bioinformatics applications to assess the data before performing assembly.

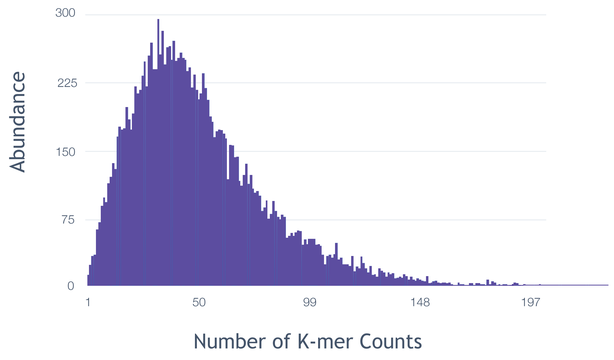
Abundance histograms can be generated from counting all k-mers in a sequence. Generating histograms can be used to estimate genome size, coverage, and estimate an important parameter for assembly.



Future Direction

- Estimate the best k value parameter for each genome present in metagenomic data
- Test the k values by perfoming assembly.

Generating Abundance Histograms



7-mers	K-mer Count
AGATTAG	1
GGAATAC	1
TAAGATC	12
AGTACGT	6
GATGACA	2

K-mer Count	Abundance
1	12
2	23
3	33
4	34
5	62

Figure 1. Abundance of an Algae Genome

DNA sequences are fragmented into k-mers and stored with a value of counts. Abundance refers to the number of k-mers that have the same number of counts in a sequence.

Effect of K-mer Size

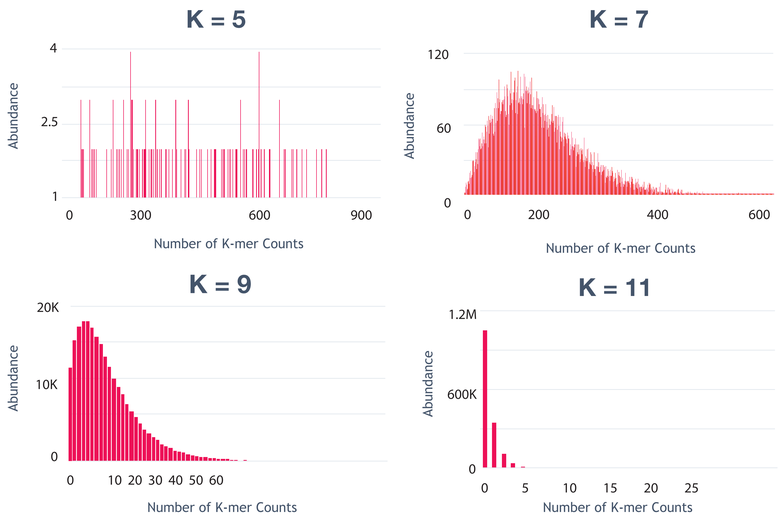


Figure 3. Effect of K = 5, 7, 9, 11

Choosing a k-mer size affects the quality of sequence assembly. Shorter k-mers will be more repetitive in a sequence and contain less information, but increase chance of overlaps. Longer k-mers are less repetitive in a sequence and contain more information, but decrease chance of overlaps.

Metagenome Abundance

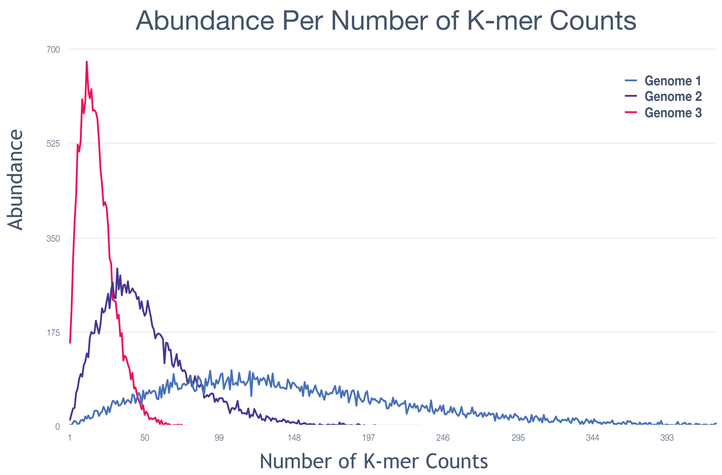


Figure 2. K-mer Abundance Profiles of 3 Different Genomes

K-mer abundance of three different algae genomes binned by MetaBAT. Here each distribution is represented by 7-mers. Ideally, each genome will require a different k-value depending on the abundance of each specie.

De Brujin Graph Assembly

Read AGATACAT

K-mers AGA, GAT, ATA, TAC, ACA, CAT, ATA, TAC

De Brujin Graph

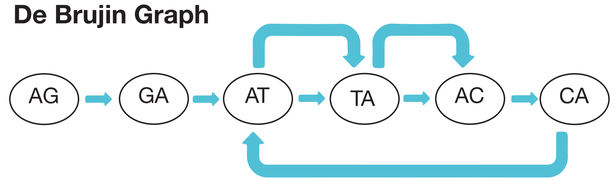


Figure 4. De Bruijn Graph Assembly with K = 3

Sequences are fragmented into k-mers, and overlaps represented by (k-1) are nodes in the graph. The arrows are directed edges which are used to reconstruct the the original genome.