Name: _____________
BIO/CMPSC 300
Chapter 8 Introduction
Fall 2019

As you know, the genome is the instruction manual that contains the information to create an organism. Everything from viruses to humans have genomes. In the third week of the course, you evaluated and edited DNA sequence data in the form of chromatograms (Figure 1). These data were generated using dye terminator sequencing and the length of a sequence of this type is typically 600-800 bp. Recent advances in DNA sequencing technology have significantly improved our ability to quickly and efficiently collect DNA sequence data using methods collectively referred to as Next Generation sequencing. However, the length of a given sequence generated using Next Generation sequencing is drastically smaller than the length of a sequence generated using dye terminator sequencing – just 20-500 bp.

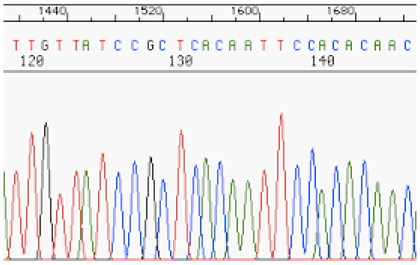

Figure 1 - DNA Sequence Chromatogram

The smallest known genome is that of the Porcine circovirus and is just 1759 bp while the largest known genome belongs to the marbled lungfish with 130,000,000,000 bp. Even the smallest known genome, and clearly the largest known genome, is too big to sequence with a single DNA sequencing read. The solution to this challenge is shown in Figure 2 below. The first step of a genome sequencing project is to isolate many copies of the target genome and then break the large genome into smaller pieces randomly. These smaller pieces are sequenced and then the genome is put back together based on overlapping nucleotides present in each sequence.

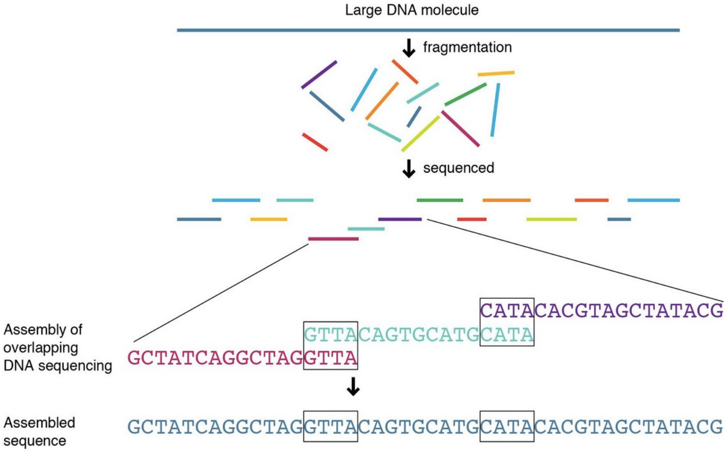

Figure 2 – Genome Sequencing and Assembly

This process has been compared to putting together a jigsaw puzzle or trying to read the New York Times after it has been blown apart into millions of pieces (Figure 3).

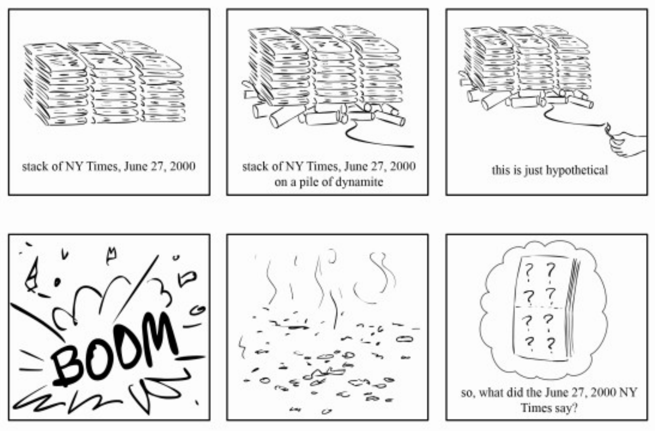
stack of NY Times, June 27, 2000
stack of NY Times, June 27, 2000 on a pile of dynamite
this is just hypothetical
BOOM
so, what did the June 27, 2000 NY Times say?

Figure 3 – Exploding newspaper analogy

For today's class activity, you will use **EGassembler** to assemble the genome sequence of a virus. The raw data you will assemble is contained in a file called **reads.txt**. This file contains ~2,500 sequences in FASTA format representing the fragmented genome of of an unknown virus. These sequences range in length from 100 to 500 bases and contain between 1 and 10 random substitutions or single-nucleotide deletions each, representing the errors inherent in sequencing data.

1. Locate the *reads.fasta* file in the same directory as this activity sheet.

2. Navigate to the EGassembler website (https://www.genome.jp/tools/egassembler/) and either upload or copy and paste the sequences from the reads.txt file into the input field.

EGassembler includes software to scan for low-quality sequence (e.g. sequences containing many Ns, representing unreadable nucleotides) and remove sequences matching databases of other DNA sources (e.g. organelles (mitochondrial DNA in a human nuclear genome project) as well as highly repetitive sequences.

3. For our purposes, turn off the options other than sequencing cleaning and the assembly step itself by unchecking the box next to the word “enable” for repeat masking, vector masking, and organelle masking. Run the program. You should immediately see the results of sequencing cleaning; You can view a .cln file to identify reads that were discarded and then examine these reads in the original sequence file.

In a few minutes, the results should become functional. From the results page, you can view (1) the contig or contigs that resulted from the assembly of your sequence reads; (2) any “singletons”, which are reads that could not be assembled into the contigs or that were not used in creating the contig; and (3) an alignment of the individual sequence reads showing how they led to the generation of the consensus contig sequence.

4. Process 1 describes the Sequence Cleaning. Looking at the text on the results page, how many sequence reads were rejected in the sequence cleaning process? Why they were rejected?

5. Download the .contigs file. Use BLAST to compare your contig sequence(s) with known

sequences in Genbank. The assembled sequence should match one known sequence with a high degree of similarity. What genome have you just assembled.

6. Because next-generation sequencing produces random short reads, there is no guarantee that even 2,500 reads would be sufficient to completely sequence a particular genome. Did the sequence reads you assembled cover the entire genome or do gaps remain? (Hint – you'll need to click on the genome accession number to determine the length of the previously published genome)

7. You used the default parameters in your EGassembler run. In a real sequencing project, however, you might want to change variables such as the overlap percent identity cut-off (the minimum percentage of nucleotides that must be identical in the overlapping region of two fragments) By default, the assembler is quite tolerate of sequencing errors (and in fact automatically compensates for some of the common problems of high-throughput sequencing such as as low-quality sequence at the beginnings and ends of fragments). To see how these parameters affect the assembly, in the "Enable Sequence Assembly Process" options, try setting the overlap percent identity cut-off to 100%. What happens to your contig? Does the quality of your alignment change?