CONCEPT: SEQUENCING THE GENOME

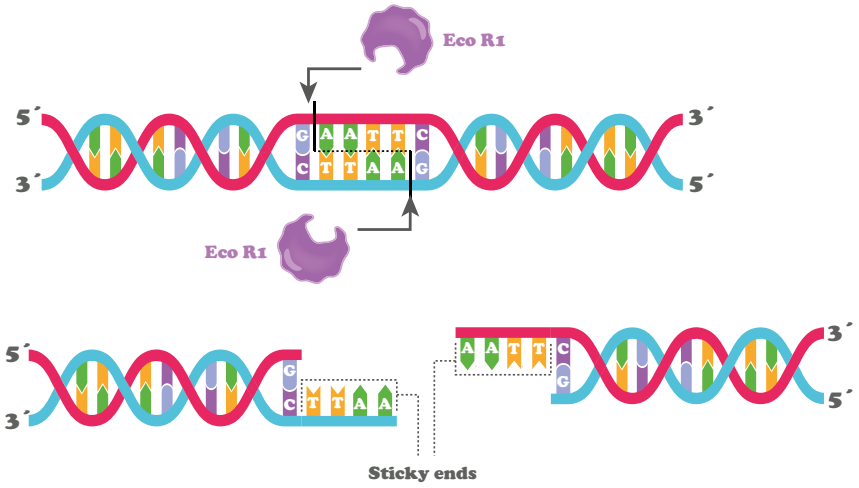
● Sequencing genomes uses a few main _____

1. DNA must be broken into millions of random, and overlapping segments

- **Restriction enzymes** are proteins that can chop the DNA at specific sequences

- **Reads** are what we call each of these fragments. They can range from 100-5000 bp long

EXAMPLE:



2. Sequence each read

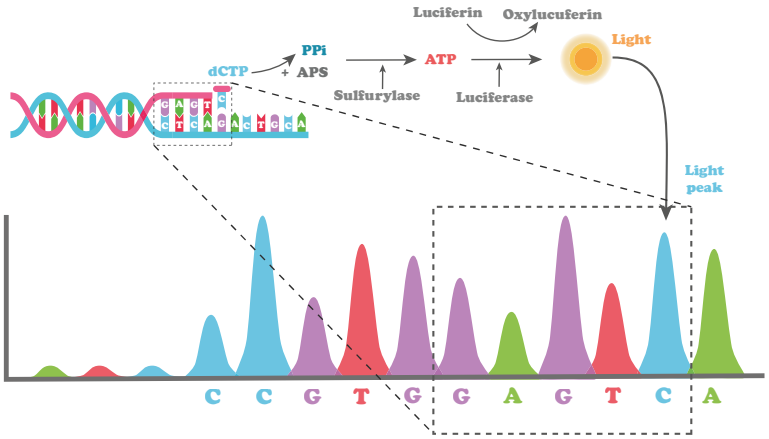
- **Pyrosequencing** takes each sequence, attaches it to bead, and then amplifies it

- A machine then runs each nucleotide across the sequence one at a time

- When a nucleotide binds it releases a pyrophosphate molecule, which can be converted into a light signal

- A camera detects the light signal and determines which nucleotide caused it by complementary binding

EXAMPLE:



3. Use _____ to find overlapping sequences and segments

4. Overlap each segment until all of the reads are linked

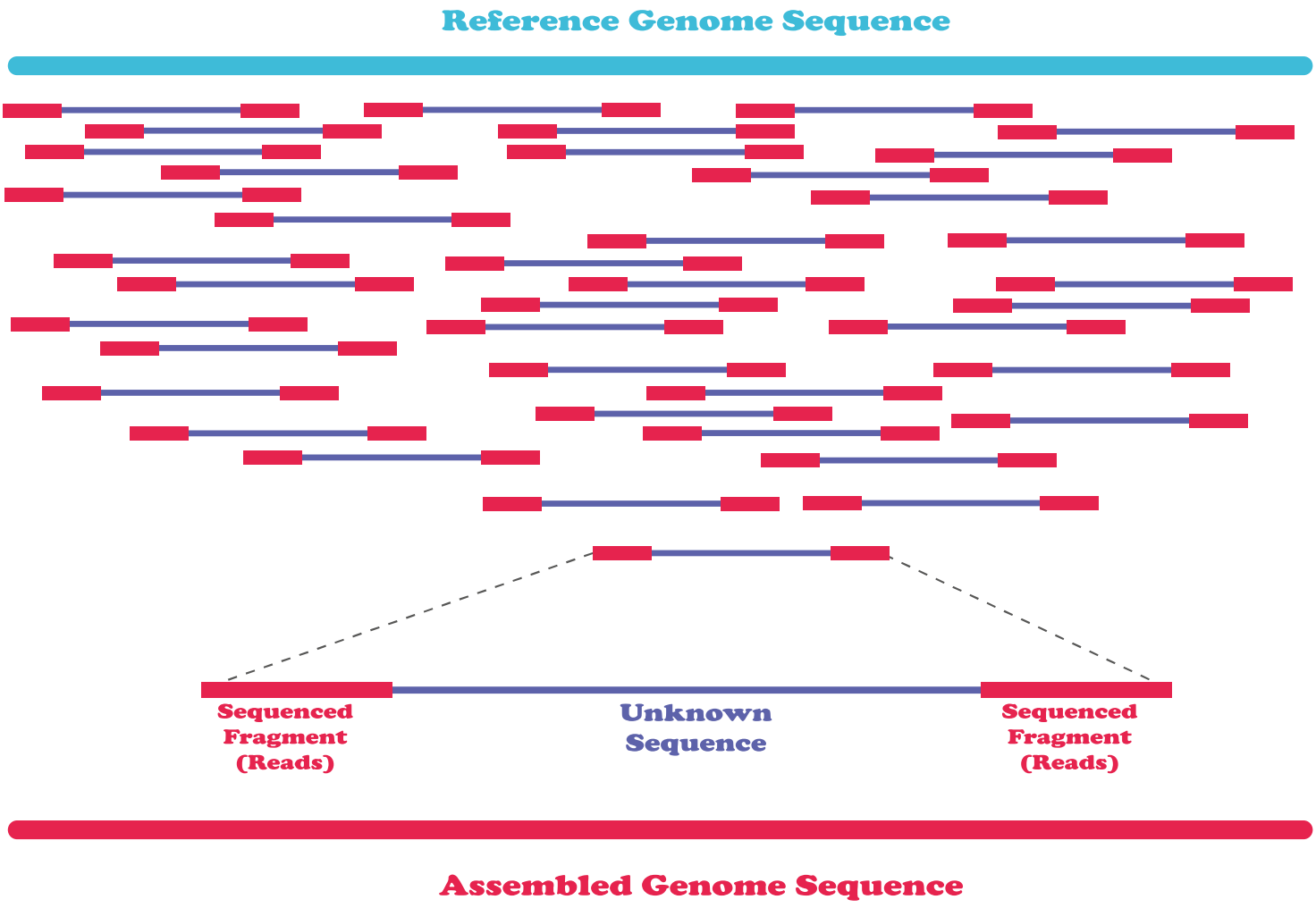
- **Sequence assembly** is building individual reads into a *consensus sequence*

- Individual differences prevents any one sequence from truly representing the genome

- Requires multiple reads of each base pair

- Ex: 10-fold coverage means that each base pair is found in at least 10 reads

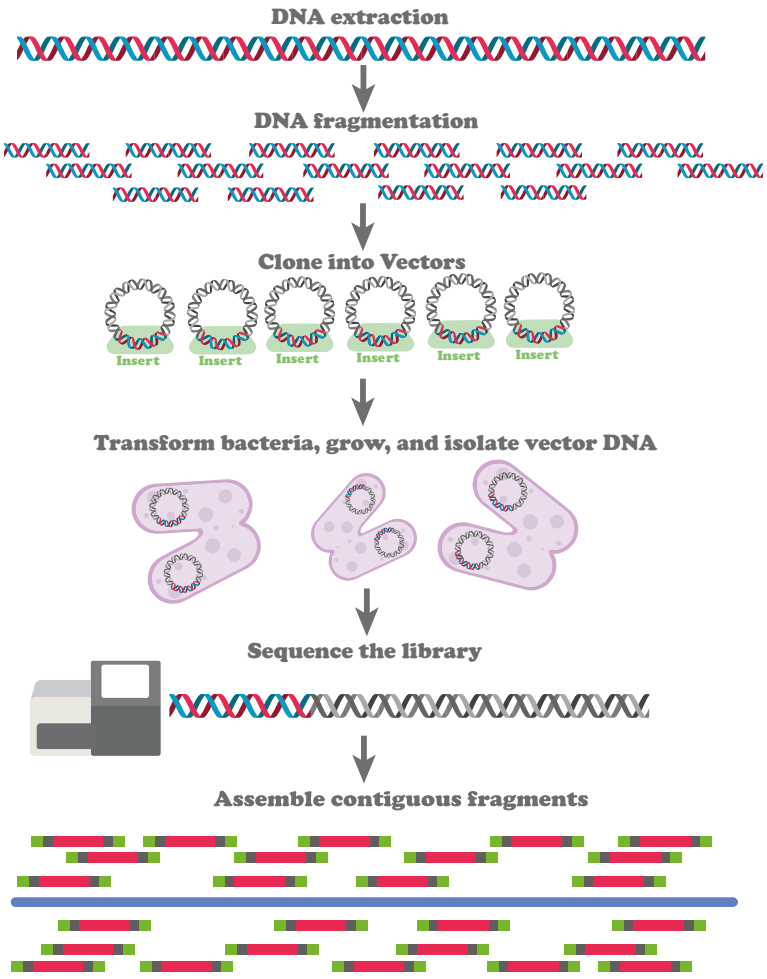
EXAMPLE:



Traditional vs. Next Generation Sequencing

- There are many different types of sequencing that have _____ over the years
  - □ **Traditional whole genome sequencing** (WGS) require the use of cell
    - DNA fragments are placed into **vectors** (plasmid) and grown in bacteria
    - Sequence reads are obtained through isolating these sequences from vectors
    - Sequence **contigs** is the final contiguous sequence each overlapping read is arranged into
  - □ **Next Generation WGS** does not require the use of cells
    - DNA is prepared using cell-free reactions
    - DNA fragments are isolated and sequenced using software and sequencing machines
    - Very small reaction volumes

EXAMPLE:



Difficulties of Whole Genome Assembly

● Certain genome characteristics make sequence assembly _____

□ One difficulty is repetitive DNA sequences that are longer than the reads

- Scientists are unable to determine where the overlap starts

- **Paired-end reads** are pairs of sequences that are read from opposite ends of genomic inserts

- Pair-end reads may span the gap and help determine the sequence between two contigs

EXAMPLE: