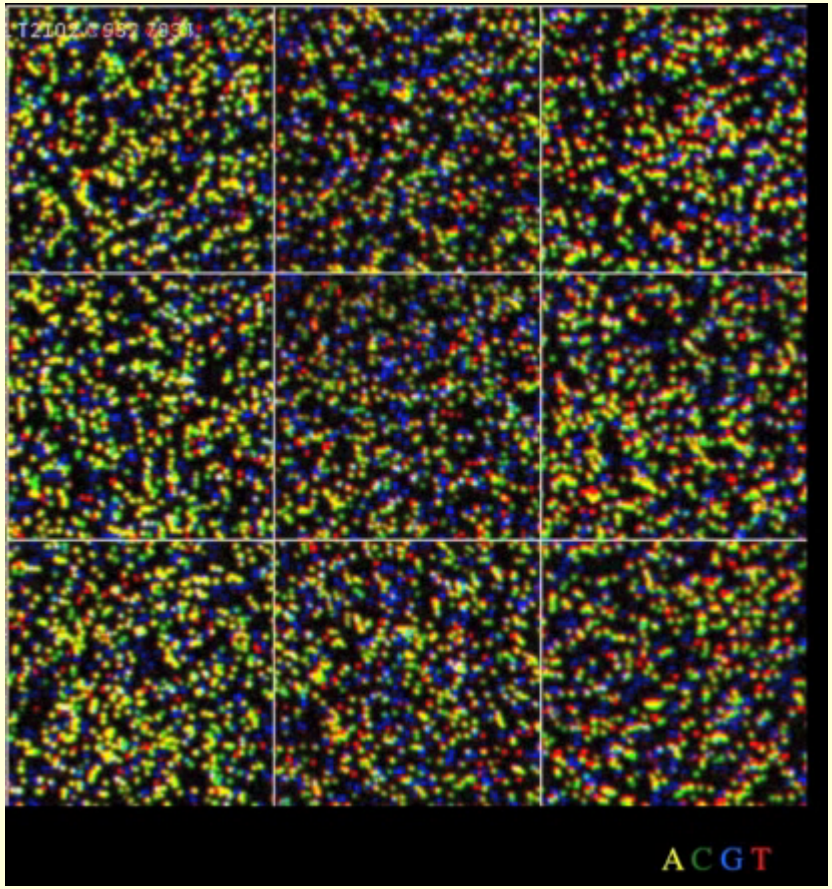
PLNT2530
2024
Unit 6e

DNA Sequencing



SOME RIGHTS RESERVED
creative commons

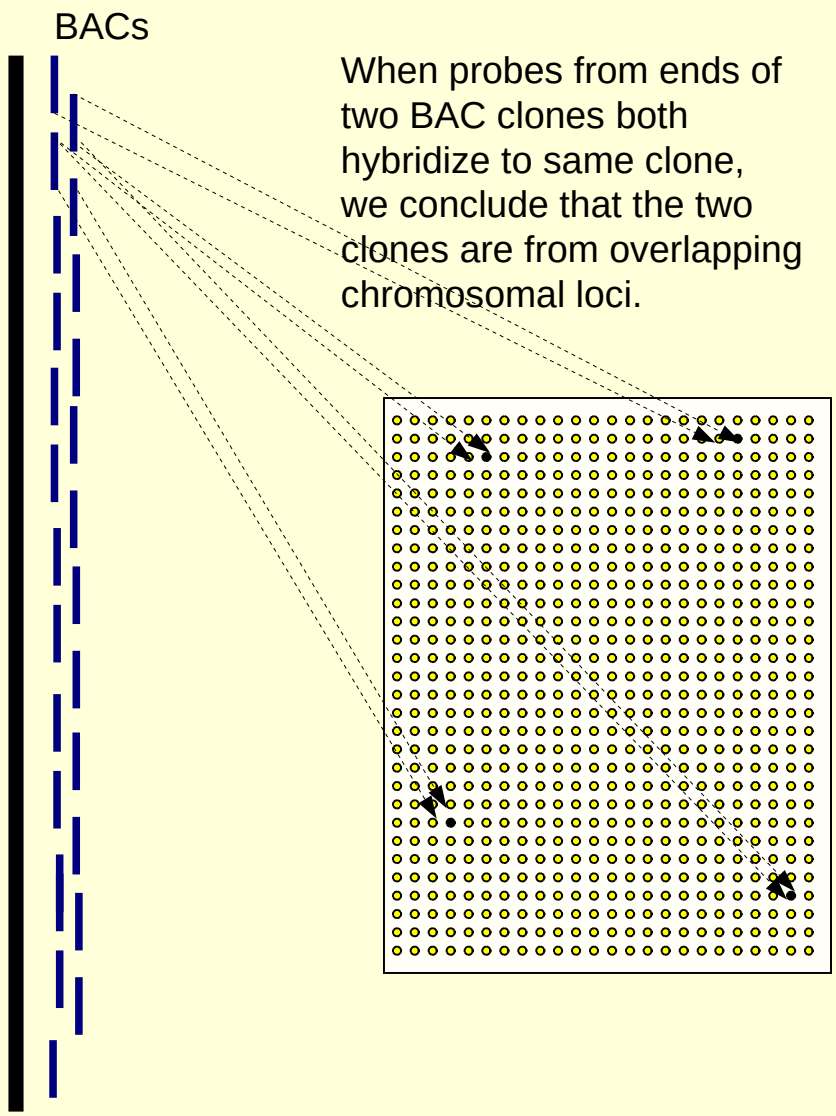
Unless otherwise cited or referenced, all content of this presenataion is licensed under the Creative Commons License Attribution Share-Alike 2.5 Canada

Genomic Sequencing - Divide and Conquer Strategy

1. Create a BAC Library
2. Create an ordered library by finding out which BACs cross-hybridize. Essentially, this is a chromosome walk along the entire genome.
It may take hundreds or thousands of hybridizations to find at least one BAC
for every part of every chromosome.
3. Sequence at least one BAC at each chromosomal location.

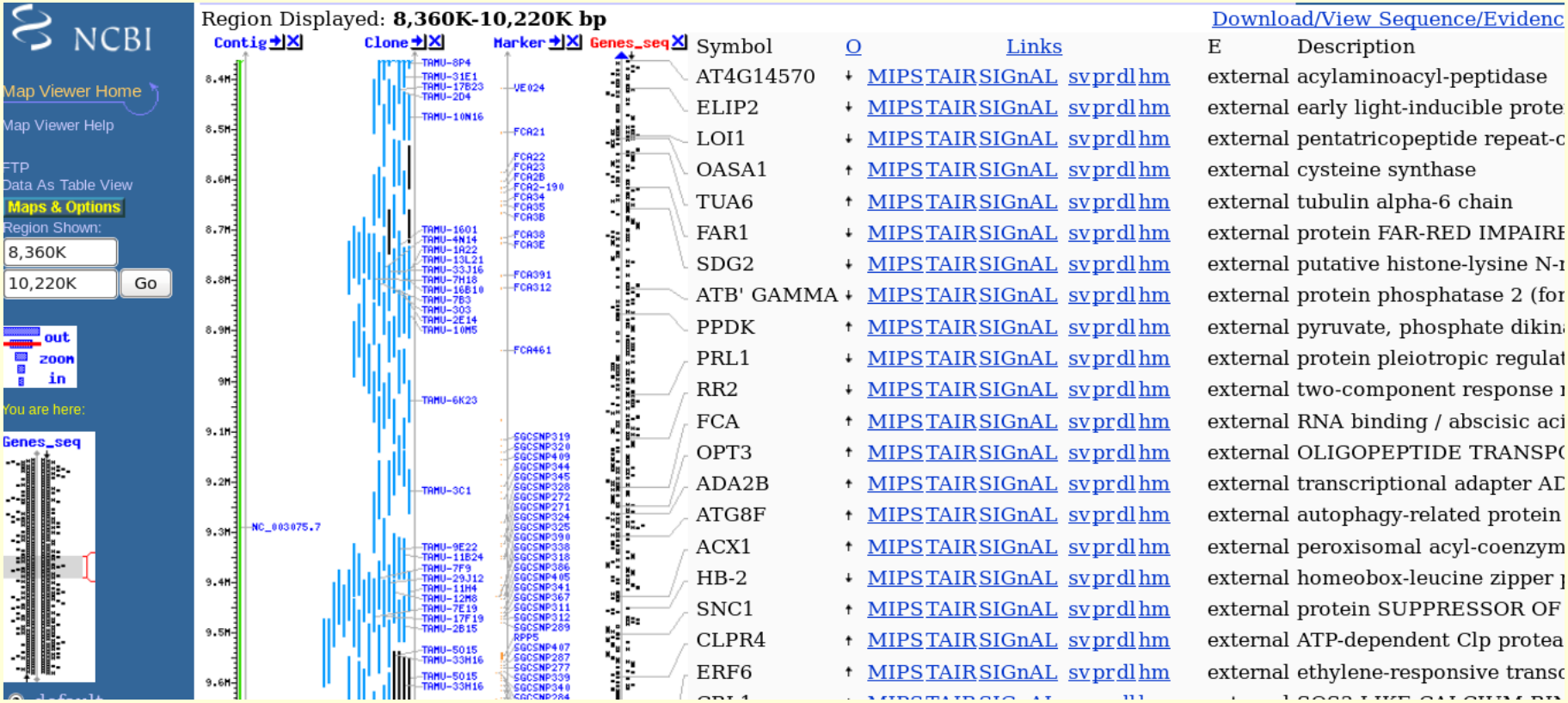
ADVANTAGE - This is the only way to be sure you have the entire genome.

DISADVANTAGE - This method takes years for most genomes.



2 Mb region of A. thaliana chromosome 4

This output from the genome map viewer at NCBI gives you some idea of how many BAC clones (light blue) were sequenced in a 2 million base region.



http://www.ncbi.nlm.nih.gov/projects/mapview/maps.cgi?TAXID=3702&CHR=4&MAPS=cntg-r,clone,tair_marker,genes[1.00%3A18585056.00]&zoom=10

Whole Genome Shotgun Sequencing (WGS)

A single BAC clone is usually 100 kb - 200 kb.

A plant genome may be 10^7 to 10^{12} bp.

How do we sequence such large genomes?

In principle, the answer is: sequence millions of very short DNA fragments and use sophisticated software to find the overlaps to assemble complete chromosomal sequences.

In practice, this ideal is not possible for lots of reasons. Nonetheless, we can assemble many thousands of runs to create long contigs that span several thousand to several hundred thousands of bases each.

High-throughput “Next Generation Sequencing” (NGS)

NGS technolologies (eg. Illumina, IonTorrent) can generate several million short sequences in a single
run. The short sequences can be overlapped, to assemble a sequence thousands, or even hundreds of thousands of bases long.

NGS technologies now make it possible to:

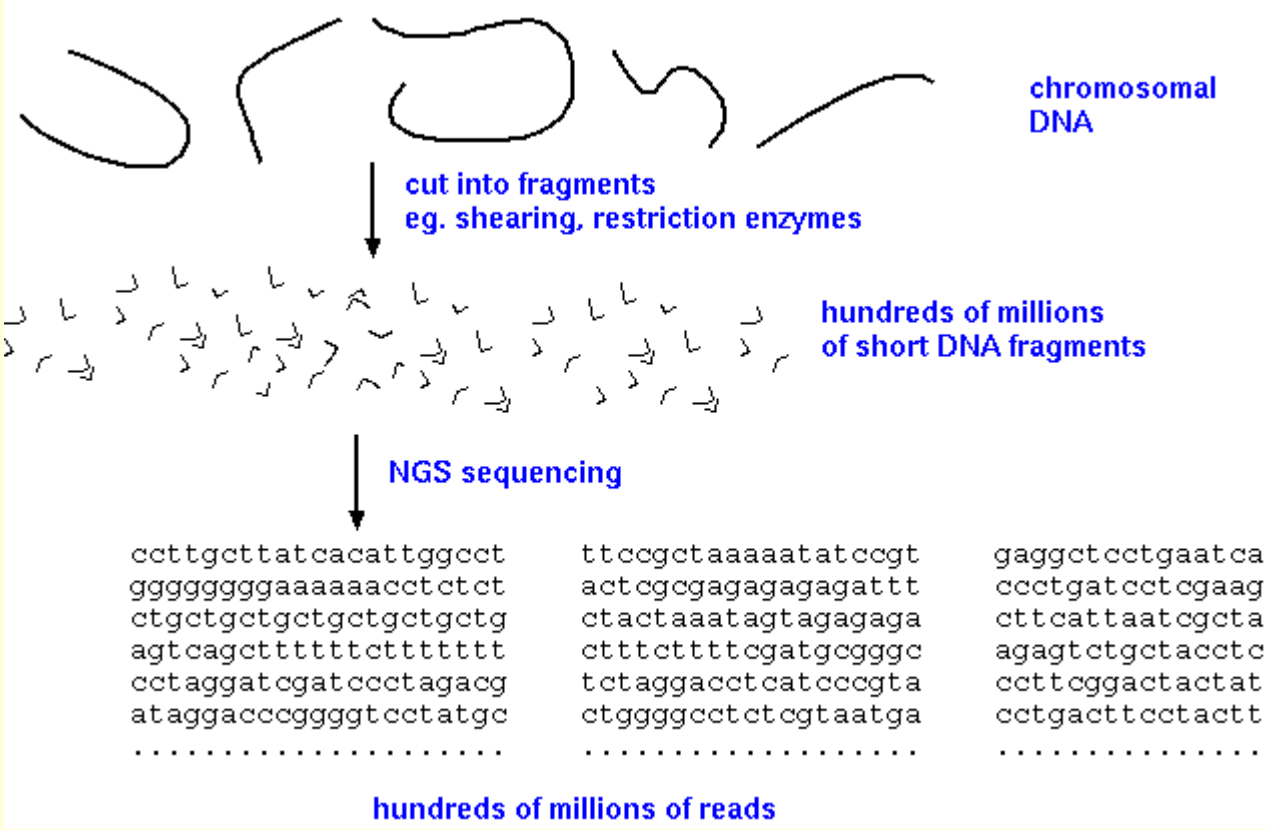
- generate almost complete genomic sequences for several thousand dollars
- generate sequences for DNA samples from mixed populations of microbes, allowing measurement of biodiversity (Metagenomics)
- generate sequences for RNA populations, making it possible to directly measure gene expression levels for thousands of genes simultaneously.

Next Generation DNA sequencing

There are numerous NGS sequencing methods. Essentially all of them follow the same steps for sequencing a genome:

1) **fragment** genomic DNA into small fragments of a few hundred bases
2) **immobilize** individual DNA molecules onto a solid surface
3) **amplify** each molecule by PCR many thousands of times
4) perform **DNA synthesis** using nucleotides that emit a characteristic wavelength of light each time a base is added
5) **read** the sequence by imaging the emission of light in real time
6) from the millions of reads, **assemble** overlapping reads into long contiguous segments known as "contigs"

Whole Genome Shotgun Sequencing (WGS)



WGS by Illumina Sequencing

Principal

sequencing by synthesis

as each base is added, a photon is emitted at a characteristic wavelength for A,G,C and T

Method

Illumina video (5:03)

https://youtu.be/fCd6B5HRaZ8

Argonne Lab Video showing real Illumina images

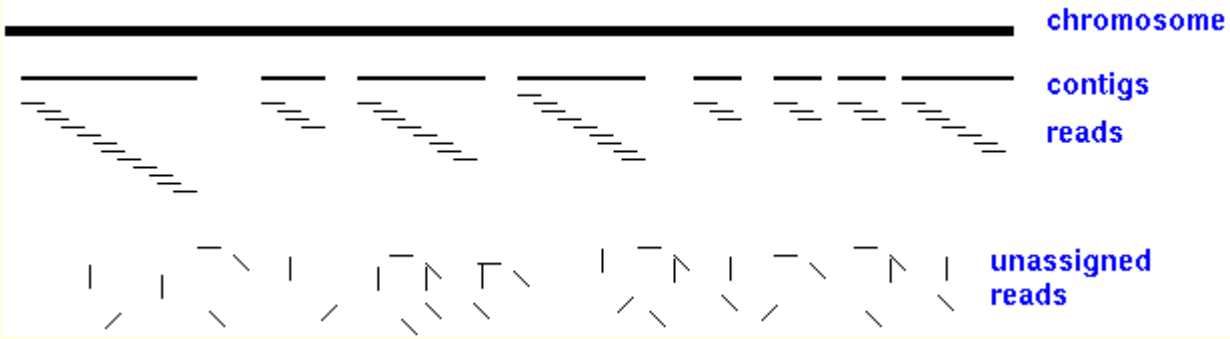
http://youtu.be/tuD-ST5B3QA

Sequence Assembly

http://mira-assembler.sourceforge.net/docs/bookfigures/454_autoedit2.png

Whole Genome Shotgun Sequencing (WGS)

High performance computers and software assemble millions of reads into contigs



- Contigs are built by overlapping reads.
- There are always gaps between contigs where the software cannot extend contigs any further.
- There are always large numbers of reads left over that cannot be assigned to a contig.
- The main reason: when reads contain parts of repetitive sequences, they may overlap thousands of other reads, making it impossible to uniquely determine overlaps. Consequently: Very few eukaryotic genomes are ever completely sequenced.

Illumina Sequencing

Advantages

- Very large number of reads
- Low cost
- Hundredfold coverage or greater

Disadvantages

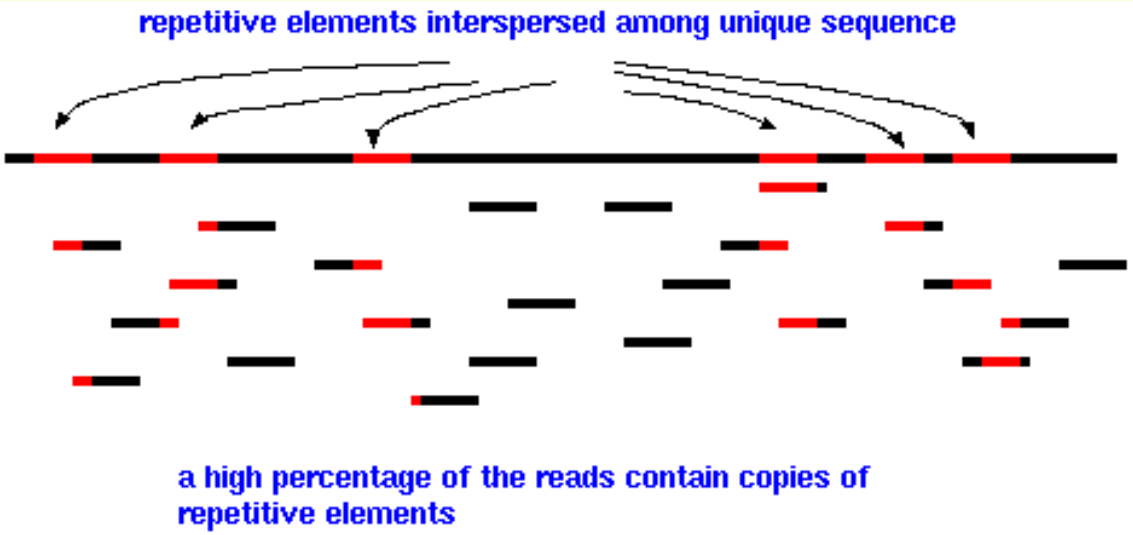
- Reads are short, 150 - 250 bp
- Error rate < 0.4 per 100 bases
- Short reads present many problems during sequence assembly

Example: 100 kb BAC clone: (100,000 bp/150 bp per read) x 100 fold coverage = 40,000 reads.

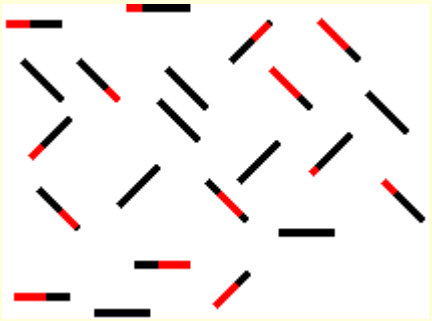
Example: Soybean Genome (1.1×10^9 bp/150 bp per read) x 100 fold coverage = 440 million reads

Sequencing Very Large Genomes using Long Read Sequencing

Problem: Plant Genomes are mostly repetitive DNA.

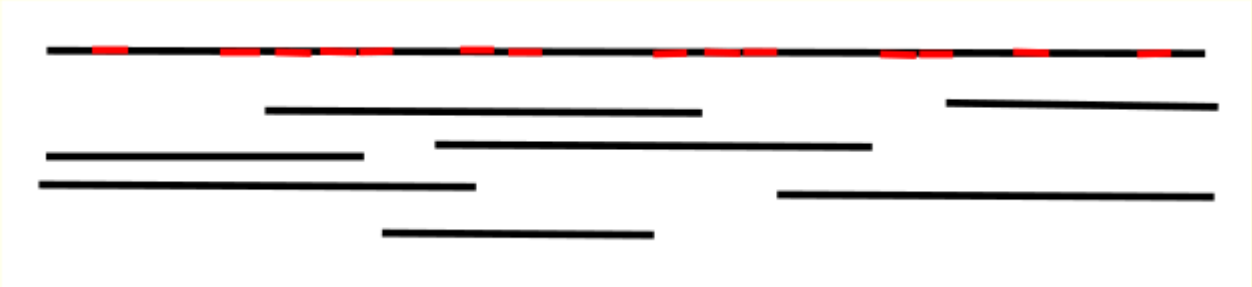


When the software tries to assemble from short reads, one copy of a repetitive sequence (red) will match every other copy of that sequence. There is now way to decide which reads overlap.



Sequencing Very Large Genomes using Long Read Sequencing

Solution: Sequence longer reads!



Reads that are several kb in length will always have substantial amounts of unique DNA on either side of repetitive DNA sequences. The unique DNA gives sequence assemblers enough information to find a unique path to assemble overlapping reads into contigs.

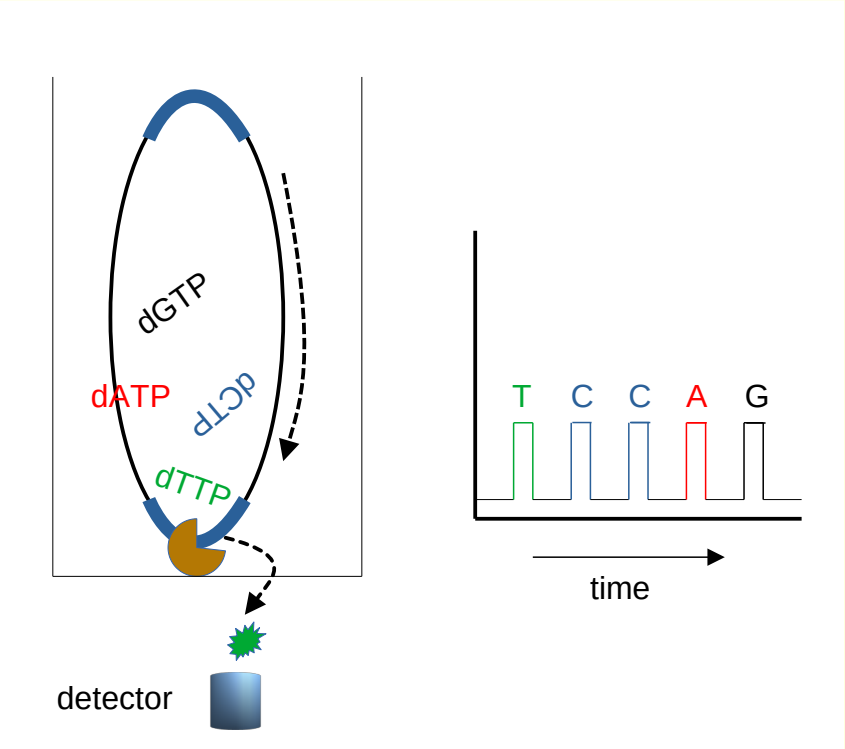
Long Read Sequencing

Long read methods sequence single DNA molecules. Instead of immobilizing the DNA fragments, long read methods pull long single-stranded DNAs through an enzyme. As the strand is pulled through, sequence is read. There are currently two competing technologies: PacBio SMRT sequencing and Oxford Nanopore.

Amarashinge SL et al. (2020) Opportunities and challenges in long-read sequencing data analysis. BMC Genome Biology 21:30 https://doi.org/10.1186/13059-020-1935-5

PAC Bio - SMRT sequencing

By adding adapters to both ends of a double-stranded fragment, a single-stranded circular molecule is created. The circular molecule is captured by a DNA polymerase which is anchored to the transparent floor of the cell. As each nucleotide is added, a photon is produced at a wavelength characteristic for A, G, C or T. Each nucleotide is therefore detected as a peak of fluorescence over time.



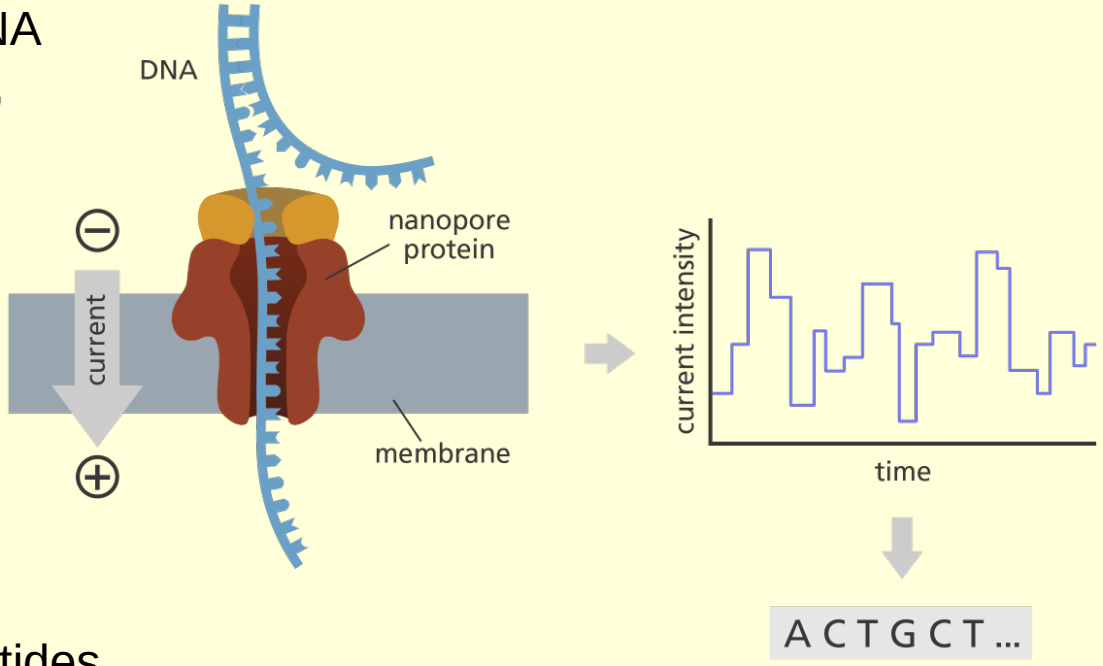
High accuracy is achieved by cycling the circular DNA through the DNA polymerase many times. The redundant sequence serves to confirm the nucleotide read at each position.

2021 - Average read lengths are around 30 kb.

Oxford nanopore sequencing

Nanopore utilizes a membrane bound motor protein, α-hemolysin, which has been genetically engineered to pull a single strand of DNA through a membrane pore.

As the single-stranded DNA enters the lower chamber, the conductivity of the solution at the inner surface of the membrane changes. Because each of the four nucleotides changes the conductivity to a different degree, a plot of current intensity over time gives a reading of the sequence of nucleotides as they emerge from the pore.



As of 2021, the average read lengths range between 15 - 20 kb. The longest reads often exceed 1000 - 2000 kb.

Image from https://www.yourgenome.org/facts/what-is-oxford-nanopore-technology-ont-sequencing