Welcome to the DNA Sequencing Revolution

A
T
C
G

Joel Sharbrough, PhD

Assistant Professor of Biology
New Mexico Tech
July 18th, 2022

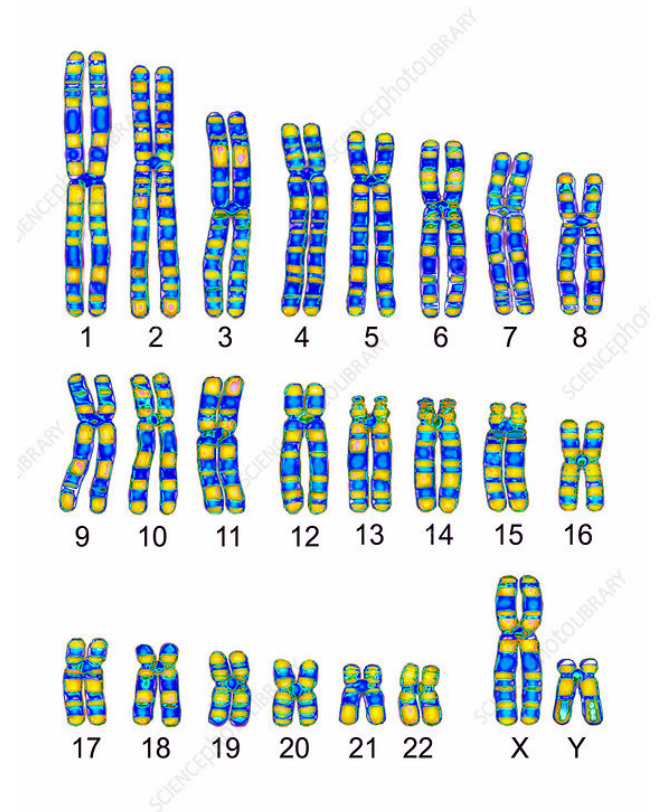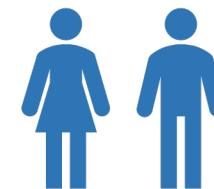# Current state of Biotechnology

A
T

C
G

- **Reading DNA**

- **Editing DNA**

- **Writing DNA**
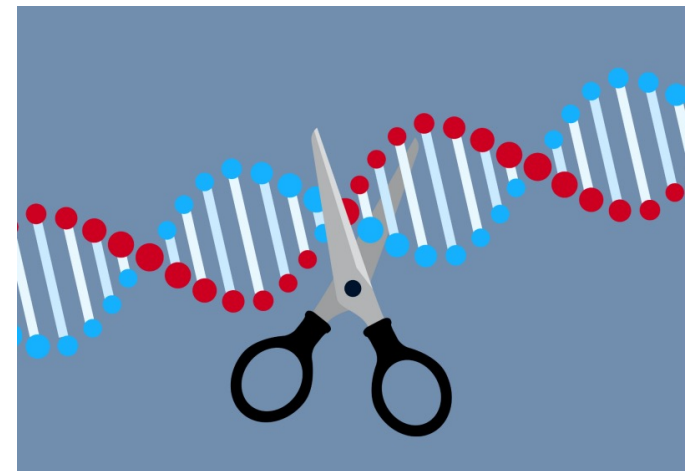
Current state of Biotechnology

- **Reading DNA –** Sequencing and decoding genomes

- **Editing DNA**

- **Writing DNA**

Current state of Biotechnology

- **Reading DNA** – Sequencing and decoding genomes

- **Editing DNA** – Making targeted modifications to a genome

- **Writing DNA**

Current state of Biotechnology

- **Reading DNA –** Sequencing and decoding genomes

- **Editing DNA –** Making targeted modifications to an existing genome

Corrected 28 June 2016; see full text.

- **Writing DNA –** Building a new genome from the ground up

A
T
C
G

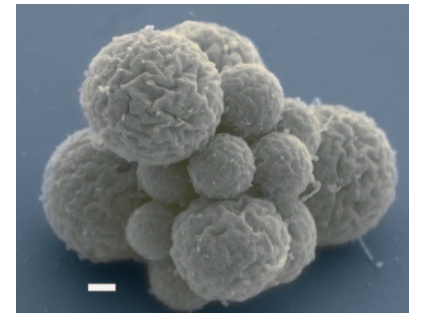**Design and synthesis of a minimal bacterial genome**
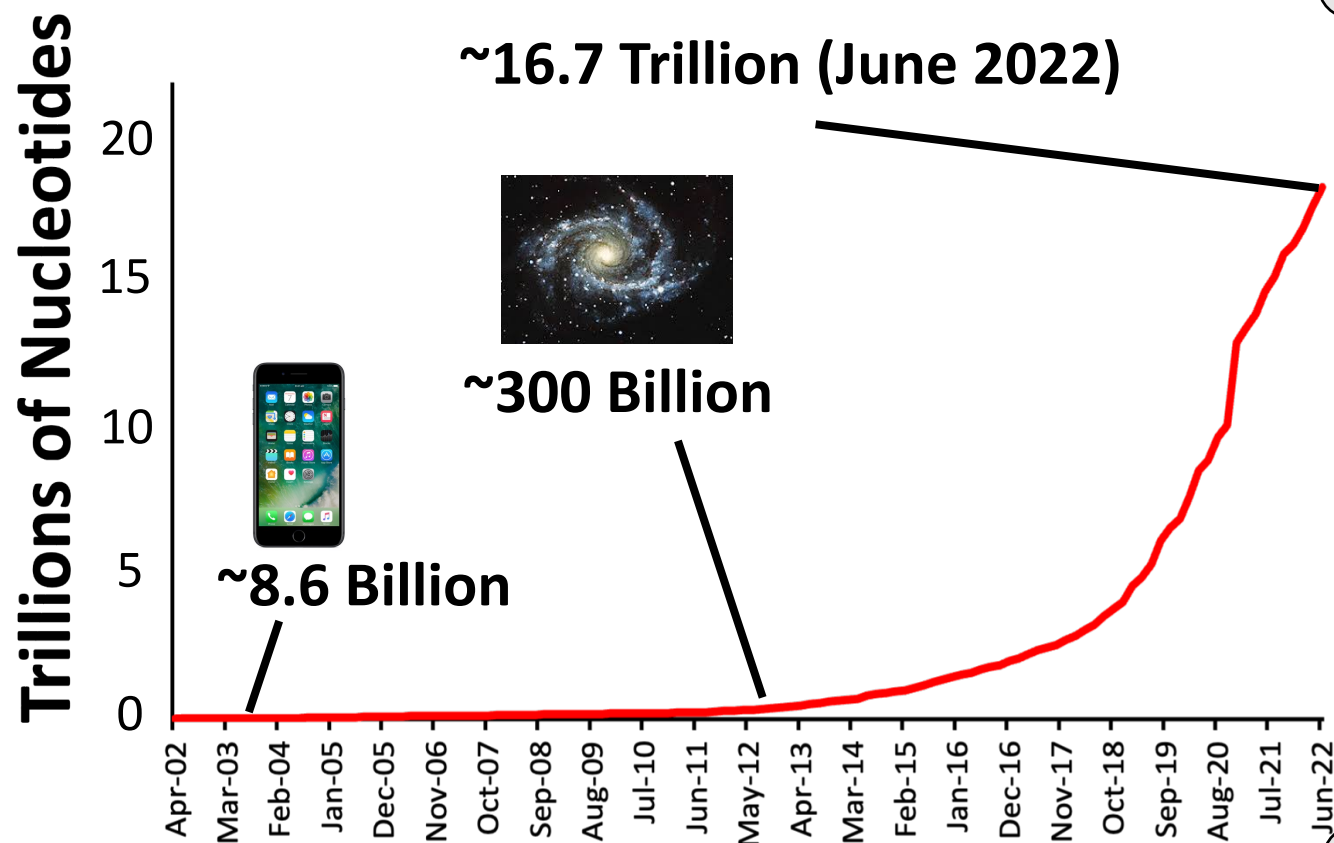
Current state of Biotechnology

- **Reading DNA –** Sequencing and decoding genomes

- **Editing DNA –** Making targeted modifications to an existing genome

- **Writing DNA –** Building a new genome from the ground up

DNA Sequencing – a biological revolution

~16.7 Trillion (June 2022)

~300 Billion

~8.6 Billion

Trillions of Nucleotides

20
15
10
5
0

Apr-02 Mar-03 Feb-04 Jan-05 Dec-05 Nov-06 Oct-07 Sep-08 Aug-09 Jul-10 Jun-11 May-12 Apr-13 Mar-14 Feb-15 Jan-16 Dec-16 Nov-17 Oct-18 Sep-19 Aug-20 Jul-21 Jun-22

Human Genome (2003):
13 years, $3 Billion

Personal Genomics (2009):
1 month, $48,000

1 Million bp (2012):
2-3 weeks, $199

Whole Genome (2022):
$350

AB Applied Biosystems

illumina®

23andMe

ancestry

PACBIO®

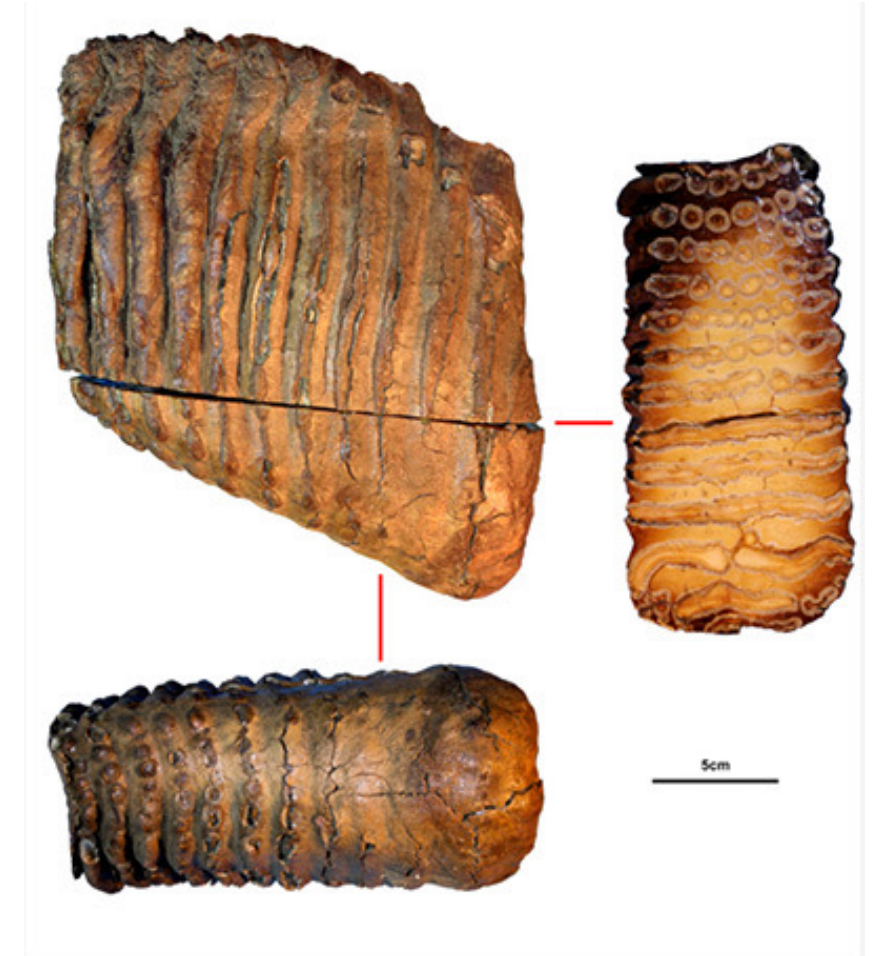NANOPORE

10X GENOMICS®

Pfizer

moderna®

Sequencing genomes using ancient DNA

2021: 1.2 Million year old DNA sample!

(Beth Shapiro's lab at UCSC)

# Ancient Neanderthal DNA sequencing as a tool to understand our past

A Complete Neandertal Mitochondrial Genome Sequence Determined by High-Throughput Sequencing

Richard E. Green,[1,*] Anna-Sapfo Malaspinas,[2] Johannes Krause,[1] Adrian W. Briggs,[1] Philip L.F. Johnson,[3] Caroline Uhler,[4] Matthias Meyer,[1] Jeffrey M. Good,[1] Tomislav Maricic,[1] Udo Stenzel,[1] Kay Prüfer,[1] Michael Siebauer,[1] Hernán A. Burbano,[1] Michael Ronan,[5] Jonathan M. Rothberg,[6] Michael Egholm,[5] Pavao Rudan,[7] Dejana Brajković,[8] Željko Kućan,[7] Ivan Gušić,[7] Mårten Wikström,[9] Liisa Laakkonen,[10] Janet Kelso,[1] Montgomery Slatkin,[2] and Svante Pääbo[1]

Green et al., 2008

Genetic history of an archaic hominin group from Denisova Cave in Siberia

David Reich[1,2]*, Richard E. Green[3,4]*, Martin Kircher[3]*, Johannes Krause[3,5]*, Nick Patterson[2]*, Eric Y. Durand[6]*, Bence Viola[4,7]*, Adrian W. Briggs[1,3], Udo Stenzel[3], Philip L. F. Johnson[4], Tomislav Maricic[3], Jeffrey M. Good[3], Tomas Marques-Bonet[10,11], Can Alkan[10], Qiaomei Fu[3,12], Swapan Mallick[1,2], Heng Li[2], Matthias Meyer[3], Evan E. Eichler[10], Mark Stoneking[3], Michael Richards[7,13], Sahra Talamo[7], Michael V. Shunkov[14], Anatoli P. Derevianko[14], Jean-Jacques Hublin[7], Janet Kelso[3], Montgomery Slatkin[6] & Svante Pääbo[3]
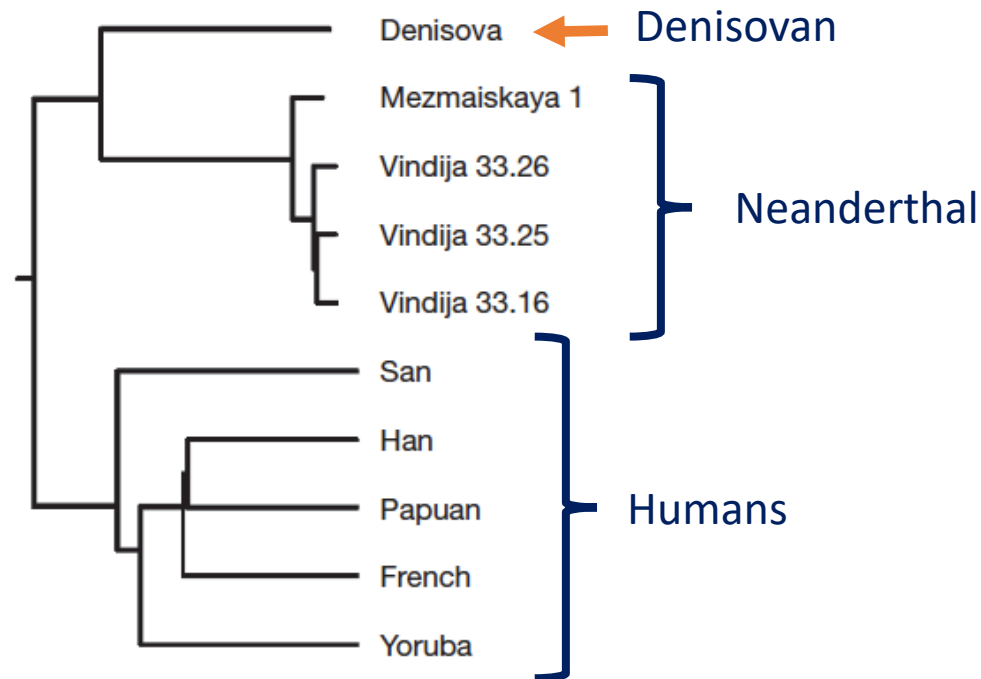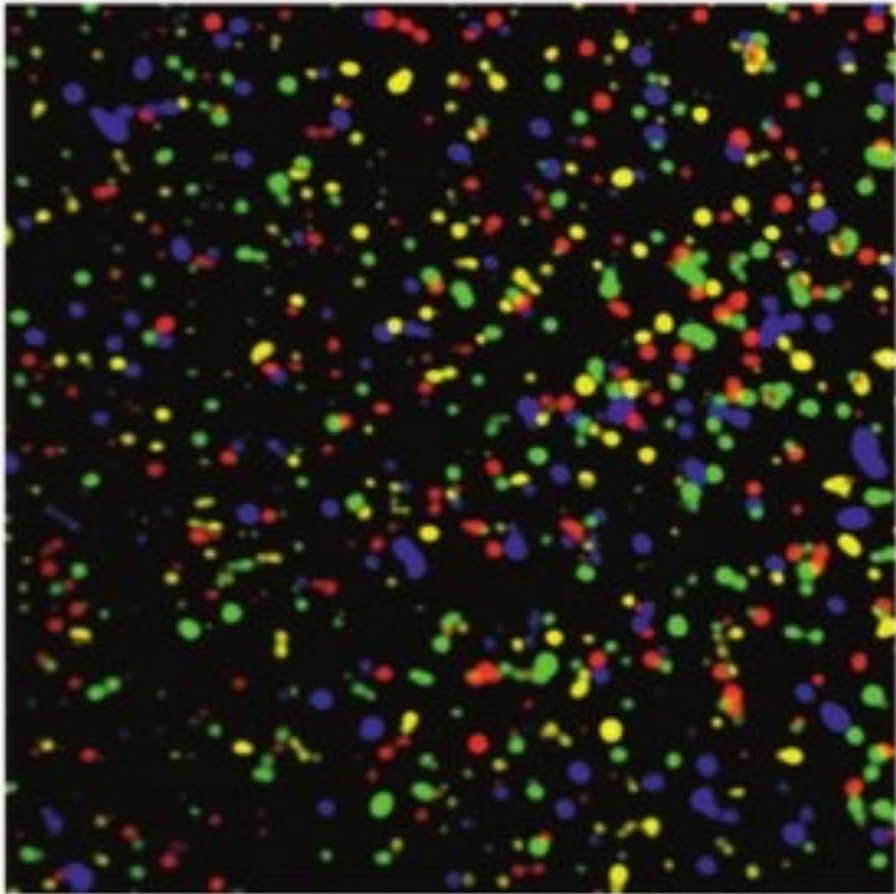
DNA Sequencing Methods: Sanger sequencing

Fluorescent dyes (fluorophores) attached to nucleotides made this human genome project possible
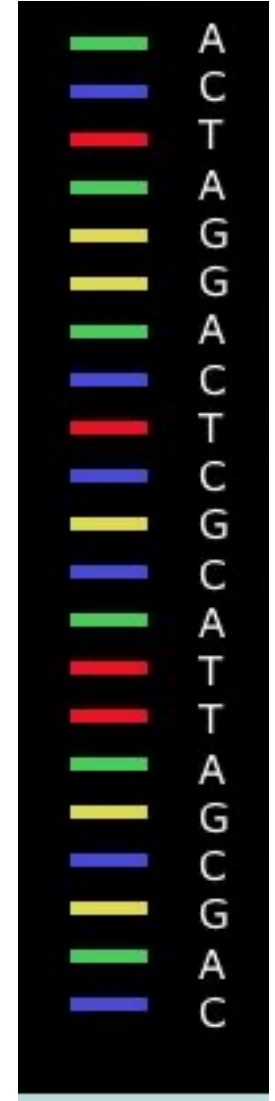
DNA Sequencing Methods: Illumina "cluster sequencing"

Voelkerding et al 2009. Clin Chem.

- Microscope images of slide after each round of nucleotide additions

- Massively parallel sequencing

DNA Sequencing Methods: PacBio "single-molecule real-time"

Each well has a little video camera placed above it

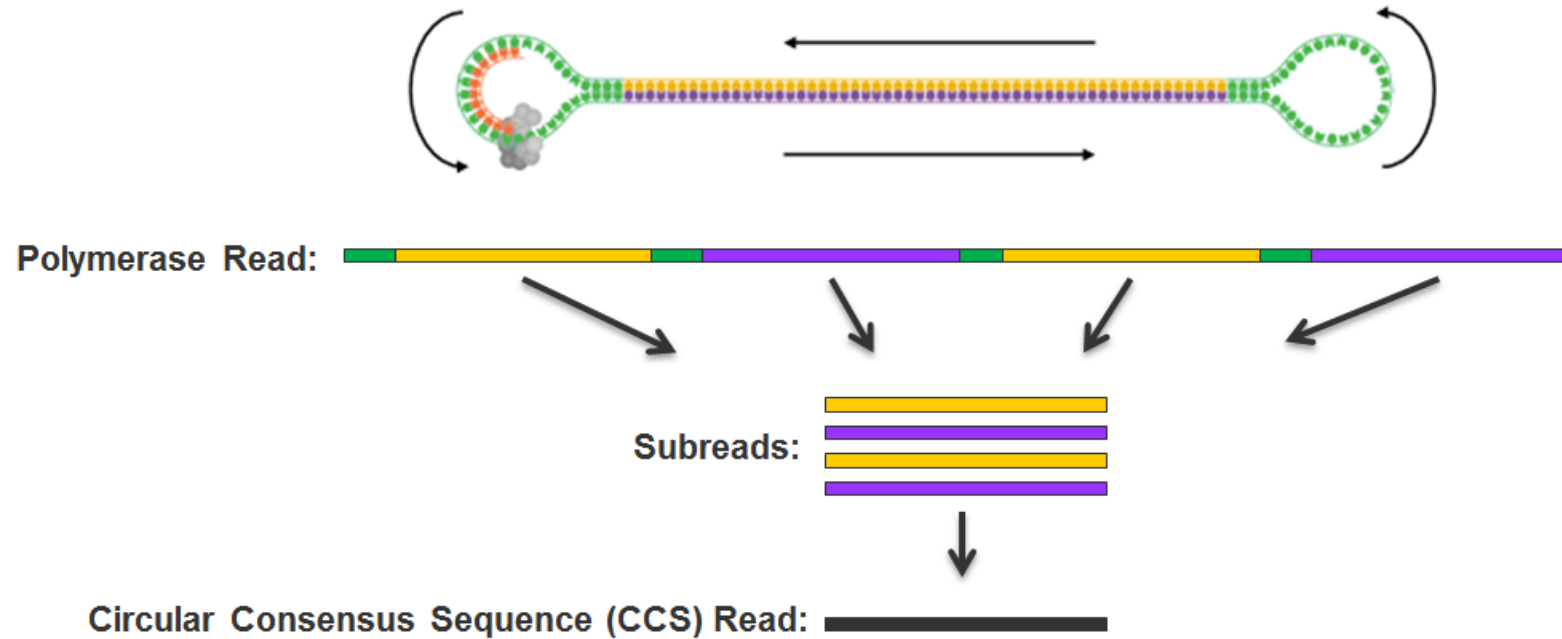Fluorophore emits light at nucleotide incorporation

As nucleotides are added, video of the fluorescent nucleotide being added is captured

SMRTbell
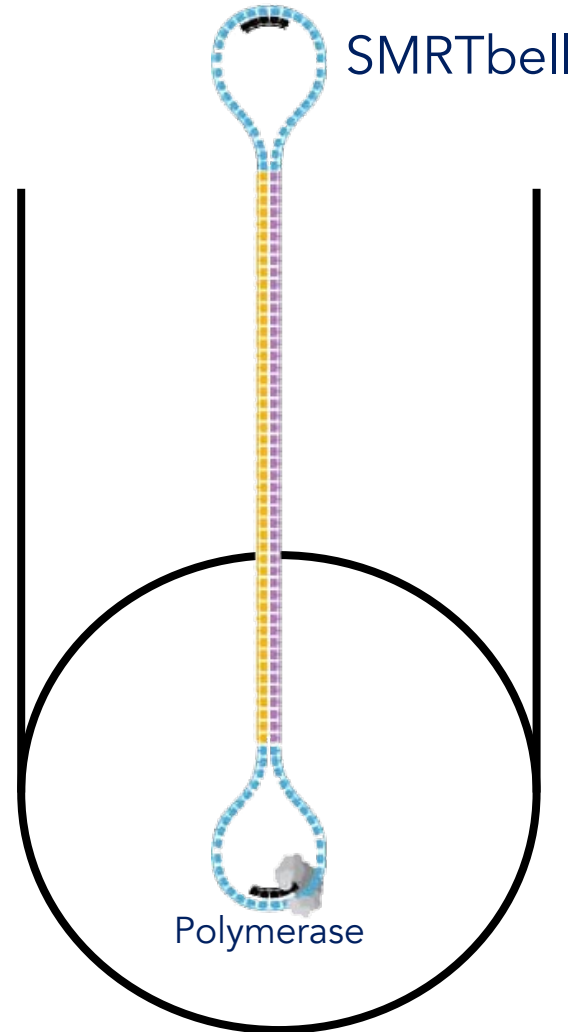
Polymerase

DNA Sequencing Methods: PacBio "single-molecule real-time"

Circular Consensus Sequencing Reads (CCS Reads) are produced when polymerase goes around the circle ≥3 times
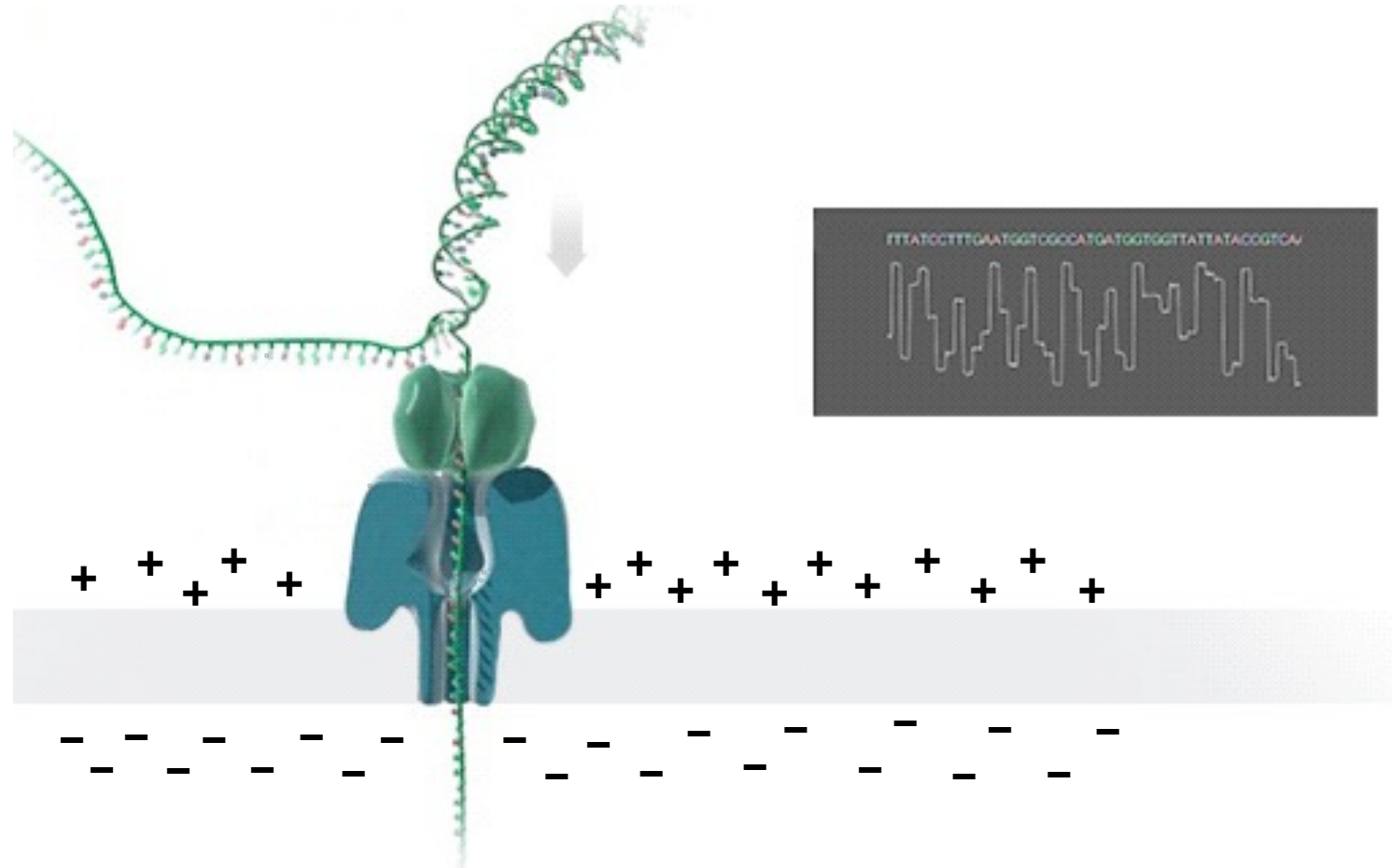
DNA Sequencing Methods: Nanopore sequencing

E. coli channel protein embedded in membrane nanopore

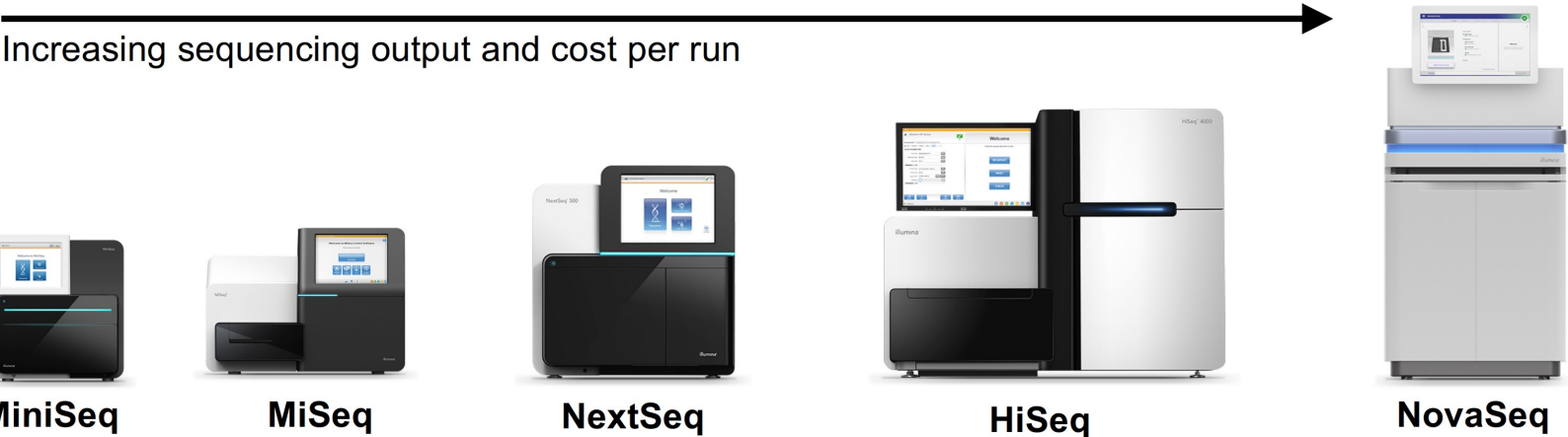Double-stranded DNA is unwound and fed through a channel

Change in voltage across membrane measured by flow of ions through channel

The extent to which **DNA blocks the flow of ions** is the output signal

Illumina sequencing platforms, costs, and outputs

	Clusters (millions)	Max Read-Length	Max Output (Gb)	Cost	Bacterial Genomes	Eukaryotic Transcriptomes	Human Genomes
MiniSeq	25	150 bp	7.5	$1,500	15	1.5	0.08
MiSeq	25	300 bp	15	$1,530	30	3	0.17
NextSeq 500 (mid)	130	150 bp	40	$1,650	80	8	0.43
NextSeq 500 (high)	400	150 bp	120	$4,240	240	24	1.33
HiSeq 4000 Lane	300	150 bp	90	$1,925	180	18	1
NovaSeq S4 Lane	2500	150 bp	750	$6,000	1500	150	8.33

Increasing sequencing output and cost per run

MiniSeq   **MiSeq**   **NextSeq**   **HiSeq**   **NovaSeq**

Slide courtesy of Dan Sloan

Oxford Nanopore Sequencing Platforms

SmidgION

-Will fill up your
phone in seconds

Flongle

-Low-throughput
(126 channels)
-Cheap
-Long queue

MinION

-Mid-throughput
30 Gb per flow cell
7-12 million reads
~$1000 starter kit
~$900 per flow cell after

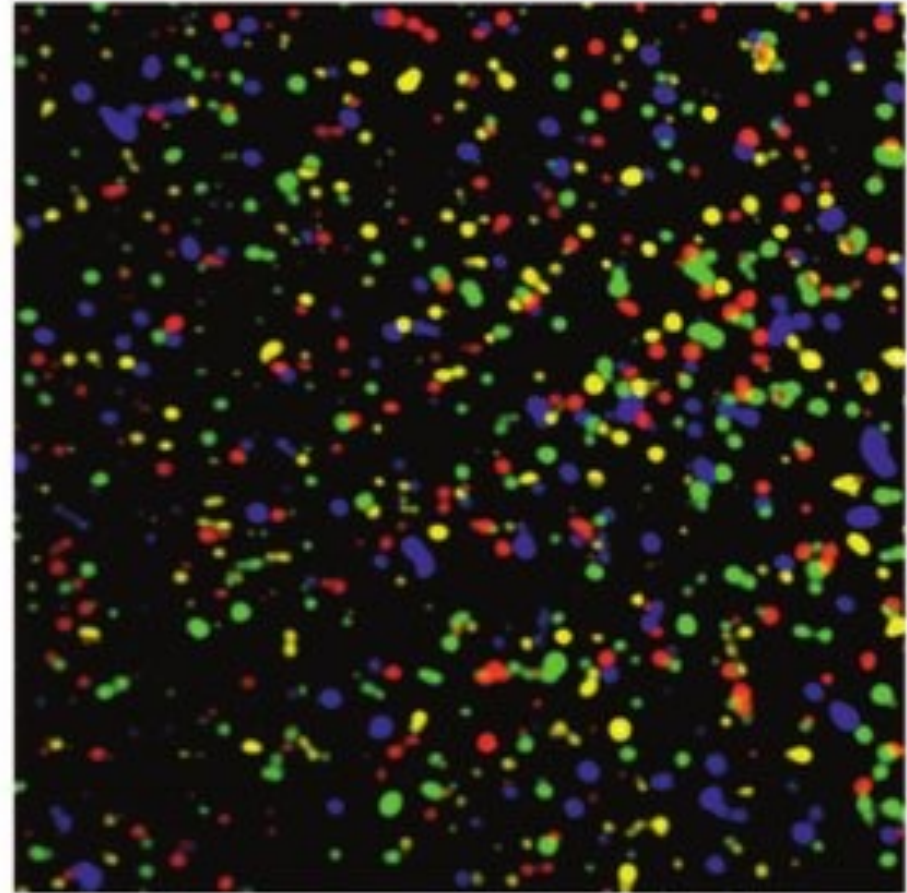**GridION**

-Mid/high-throughput
5 x Flow Cells

PromethION

-High-throughput
24/48 x Flow Cells

What does DNA sequence data look like?

FASTQ Format

@sequence_name

sequence

+

base_quality_values

Voelkerding et al 2009. *Clin Chem.*

FASTQ Format

```
@ERR2000666.1 1 length=60
CCCTTCGCTGGGCAGCGCAGTACTGGCTGCCGATGGTGCAGGTCACGCCGCGCGGGATGT
+ERR2000666.1 1 length=60
I]]]]]]]]]]GIBHI]ICI]]]]]HIDHIIICDIIDIIDI]]ICIHIHFIFGEFD?@C>
@ERR2000666.2 2 length=152
NCGGCGTGCGTNNNNNNNNNGCGANNNNNNNNNNNNNNNNNTNNNNNNNNNNNNNNNNNNNNNNNNN
+ERR2000666.2 2 length=152
!DHGFH?FFG@!!!!!!!!!IIIC!!!!!!!!!!!!!!!!!8!!!!!!!!!!!!!!!!!!!!!!!!!
@ERR2000666.3 3 length=152
CNCCGGTGNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
+ERR2000666.3 3 length=152
3!D=?415!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
@ERR2000666.4 4 length=43
TCGTGGTGCTGGATCATCTTGGTGAGGAAGGCCTGGTCGAACG
+ERR2000666.4 4 length=43
]]]]]]]S]]]]]]]BI]]]]]]]]]]]]>]B]]]A?]FA?@:>]W]
```

The National Center for Bioinformatic Information (NCBI)

- GenBank – NCBI database search tool

- ftp – Easy download for smaller items (e.g., assemblies/annotations)

- sra – Bulk download for sequencing read datasets

Other Data repositories

- [EBI](#) – Ensembl, European Bioinformatics Institute
- [DDBJ](#) – DNA Data Bank of Japan
- [Phytozome](#) – JGI Genome Portal for plants
- [EnsemblPlants](#) – Database containing. plant genomes
- [Dryad](#) – Digital Object Identifier permanent data storage
- [FigShare](#) – Digital Object Identifier permanent data storage
- [GitHub](#) – Limited permanent data storage (mostly for version control)