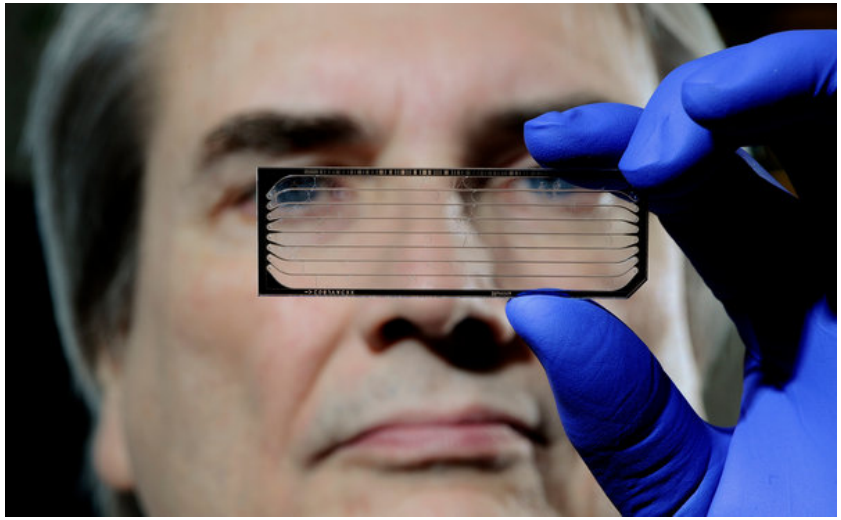
Lectures 7, 8: DNA Sequencing History and Methods

Spring 2020

February 20,27, 2020

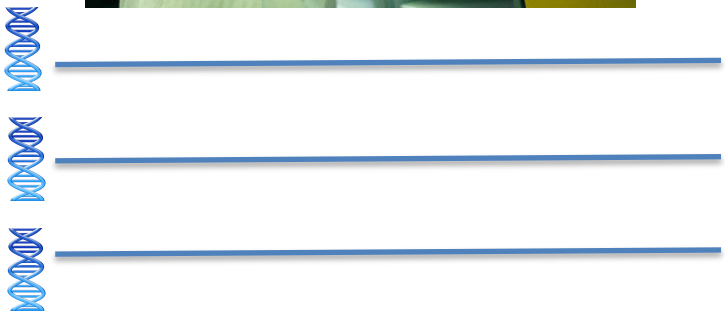
Introduction and History



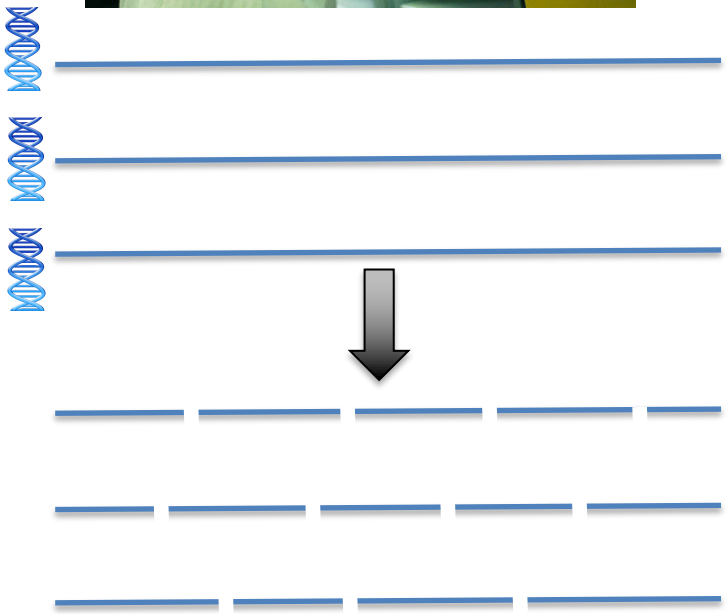
HMP


16 February 2001
Science
Vol. 291 No. 5507
Pages 1145–1434 $9
THE HUMAN GENOME
American Association for the Advancement of Science

Sample Preparation



Sample Preparation



Fragments

Sample Preparation

Fragments

Sequencing

Next Generation Sequencing (NGS)

ACGTAGAATCGACCATG

GGGACGTAGAATACGAC

ACGTAGAATACGTAGAA

Reads

Sample Preparation

Fragments

Sequencing

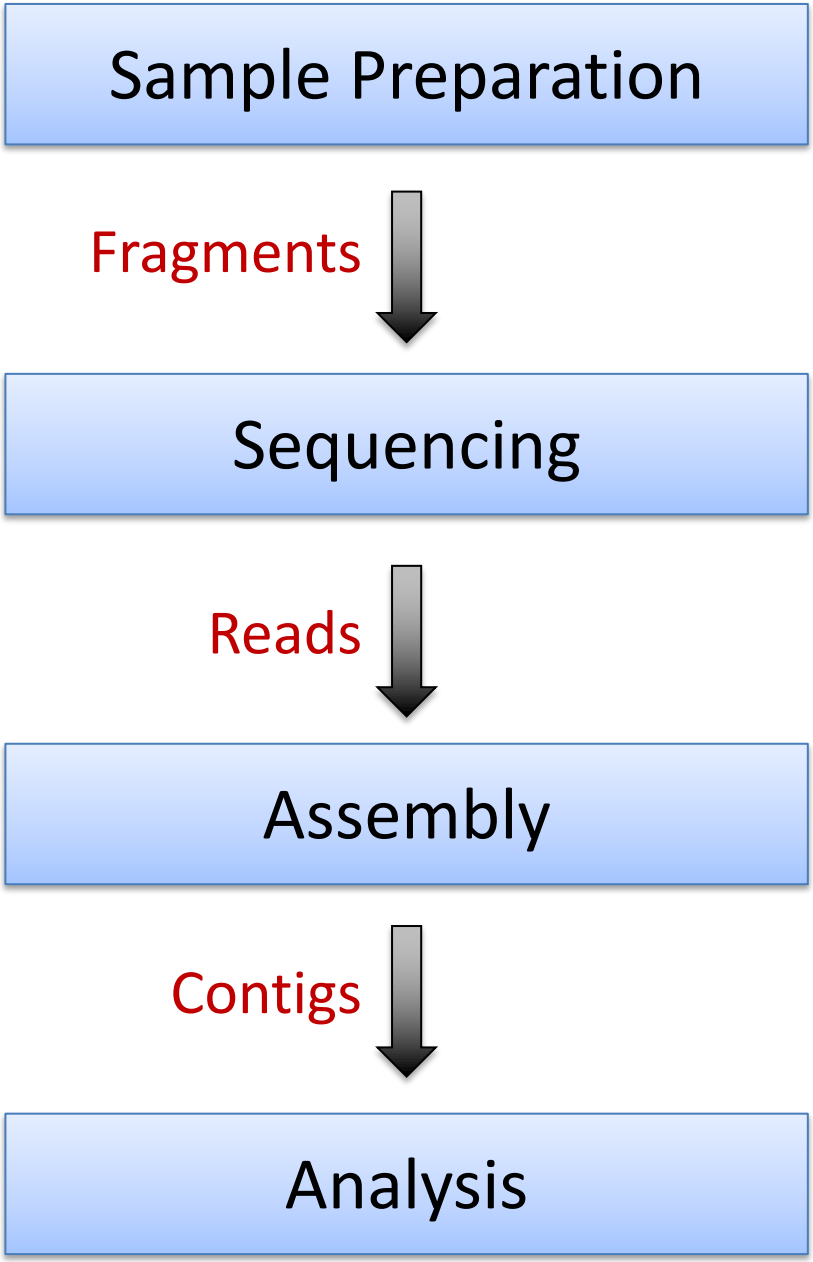
Reads

Assembly

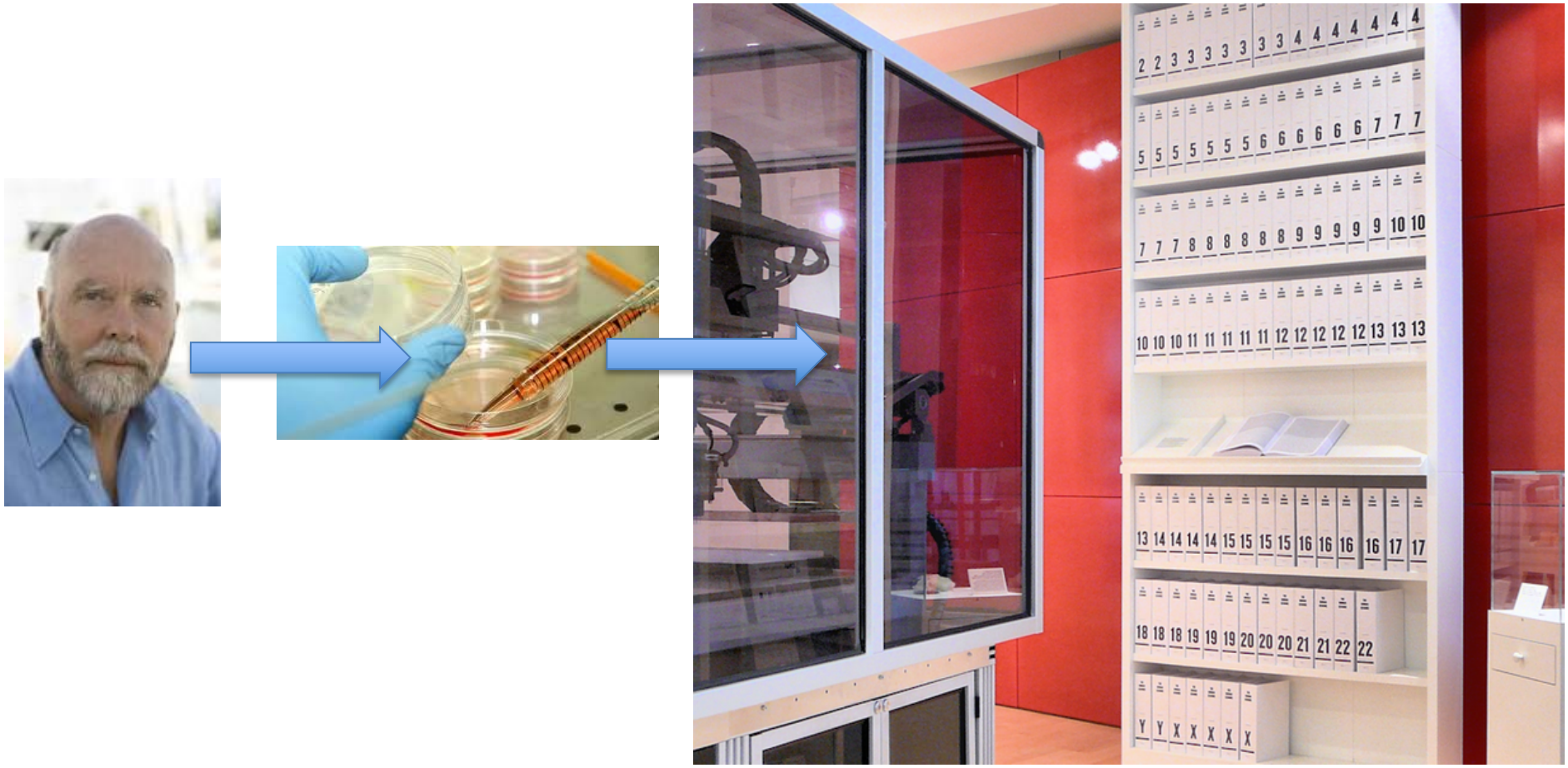
ACGTAGAATACGTAGAA
ACGTAGAATCGACCATG
GGGACGTAGAATACGAC

ACGTAGAATACGTAGAAACAGATTAGAGAG…

Contigs


Sample Preparation
Fragments
Sequencing
Reads
Assembly
Contigs
Analysis

Reference Genome

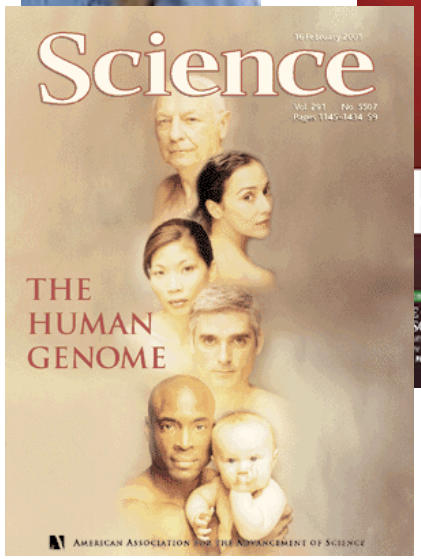


De novo vs. Re-sequencing

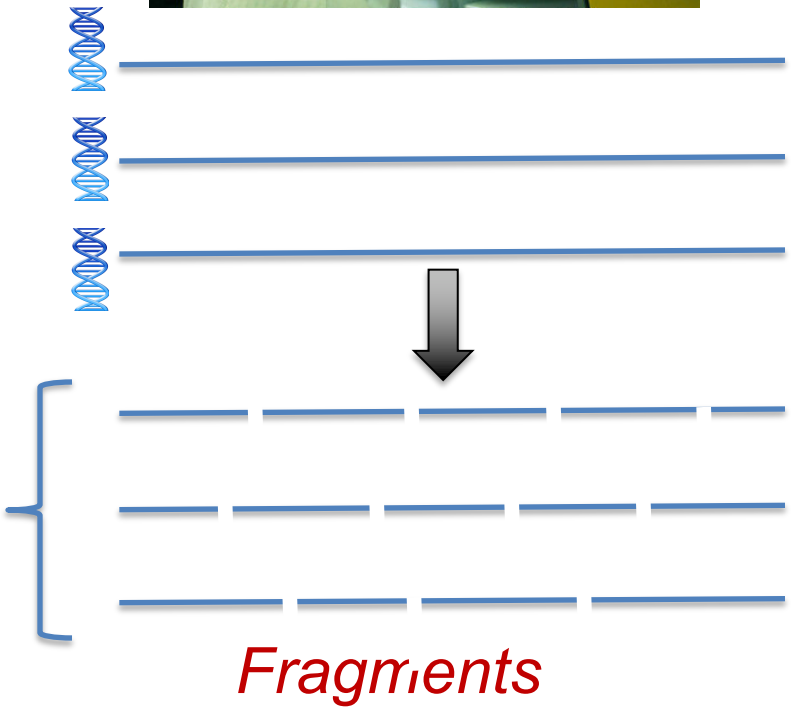
- ***De novo* assembly** ("from the beginning") implies that you have **no prior knowledge** of the genome.
- **Re-sequencing assembly** assumes you have a copy of the reference genome (that has been verified to a certain degree).
- The programs that work for re-sequencing will not work for *de novo*.

De novo vs. Re-sequencing

Giant Panda Database

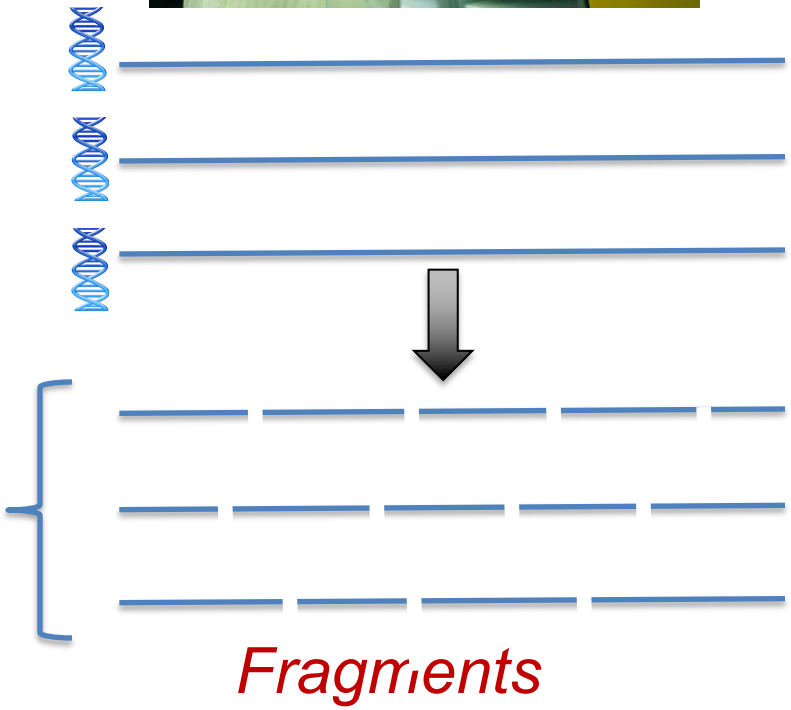


Sample Preparation

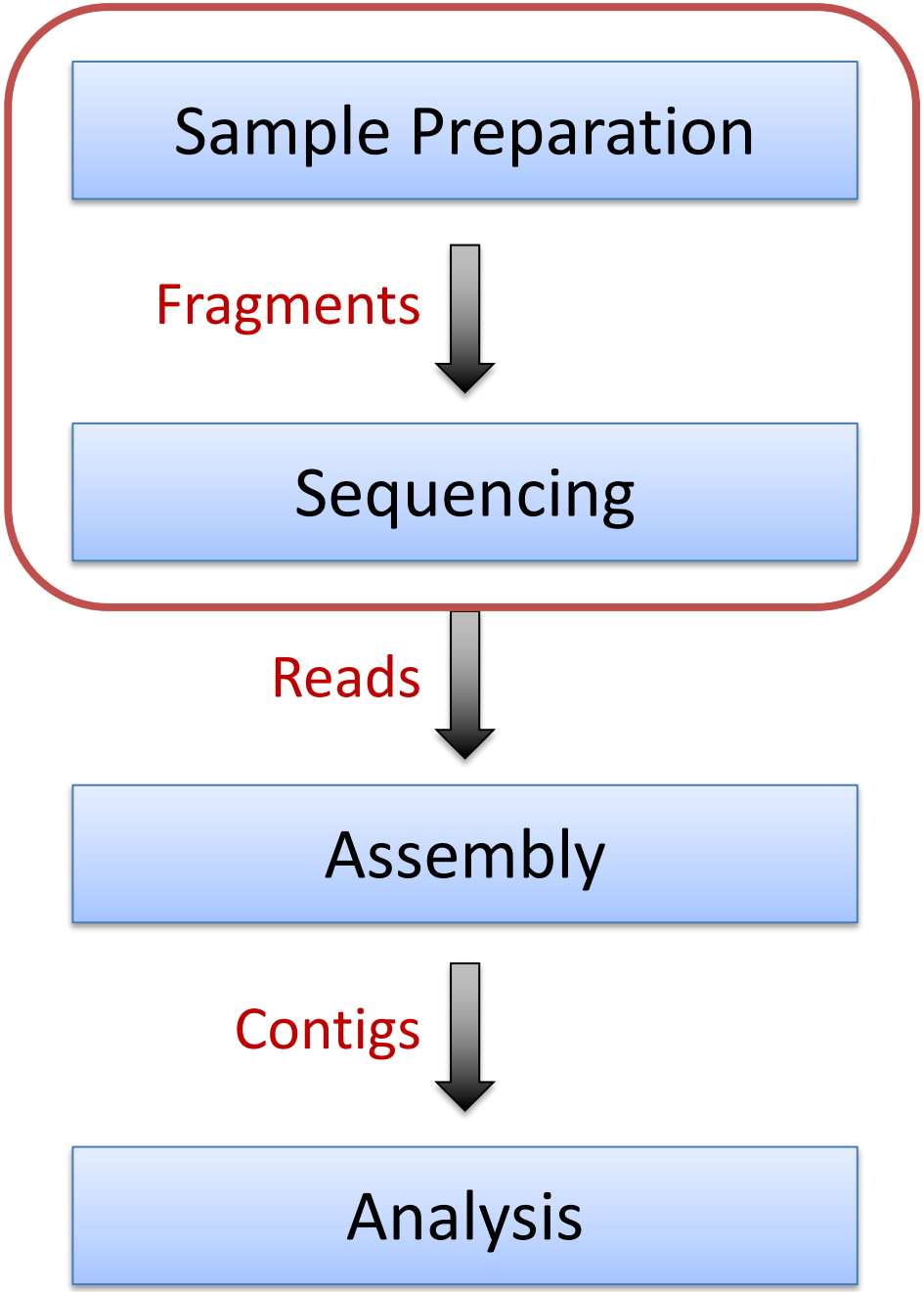


Re-sequencing (LOCAS, Shrimp) requires 15x to 30x coverage. Anything less and re-sequencing programs will not produce results or produce questionable results.

Sample Preparation



De-novo assembly requires higher coverage. At least 30x but upwards to 100x's coverage. Most de novo assemblers require paired-end data.



Our focus for today's lecture:

1. Comparison of sequencing platforms
2. Details of sample preparation
3. Definitions and terminologies concerning data and sequencing platforms

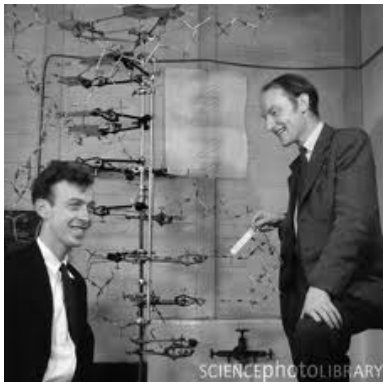
History and Background

Landmarks in Sequencing

Efficiency (bp/person/year)	Year	Event
	1870	Miescher: Discovers DNA
	1940	Avery: Proposes DNA as “Genetic Material”
	1953	Watson & Crick: Double Helix Structure of DNA
1	1965	Holley: transfer RNA from Yeast
1,500	1977	Maxam & Gilbert: "DNA sequencing by chemical degradation” Sanger: “DNA sequencing with chain-terminating inhibitors”
15,000	1981	Messing and his colleagues developed “shotgun sequencing” method
25,000	1987	ABI markets the first sequencing platform, ABI 370

Landmarks in Sequencing

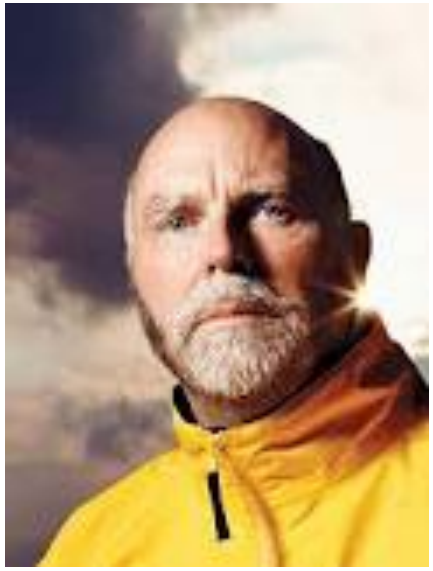
Efficiency (bp/person/year)	Year	Event
50,000	1990	NIH begins large-scale sequencing bacteria genomes.
200,000	1995	Craig Venter and Hamilton Smith at the Institute for Genomic Research (TIGR) published the first complete genome of a free-living organism in Science. This marks the first use of whole-genome shotgun sequencing, eliminating the need for initial mapping efforts.
	2001	A draft of the human genome was published in Science.
	2001	A draft of the human genome was published in Nature.
50,000,000	2002	454 Life Sciences comes out with a pyrosequencing machine.
100,000,000	2008	Next generation sequencing machines arrive.
Huge	2015+	Oxford Nanopore: 600 Million base pairs per hour.



Watson and Crick

Robert Holley and team in 1965

Messing: World's most-cited scientist

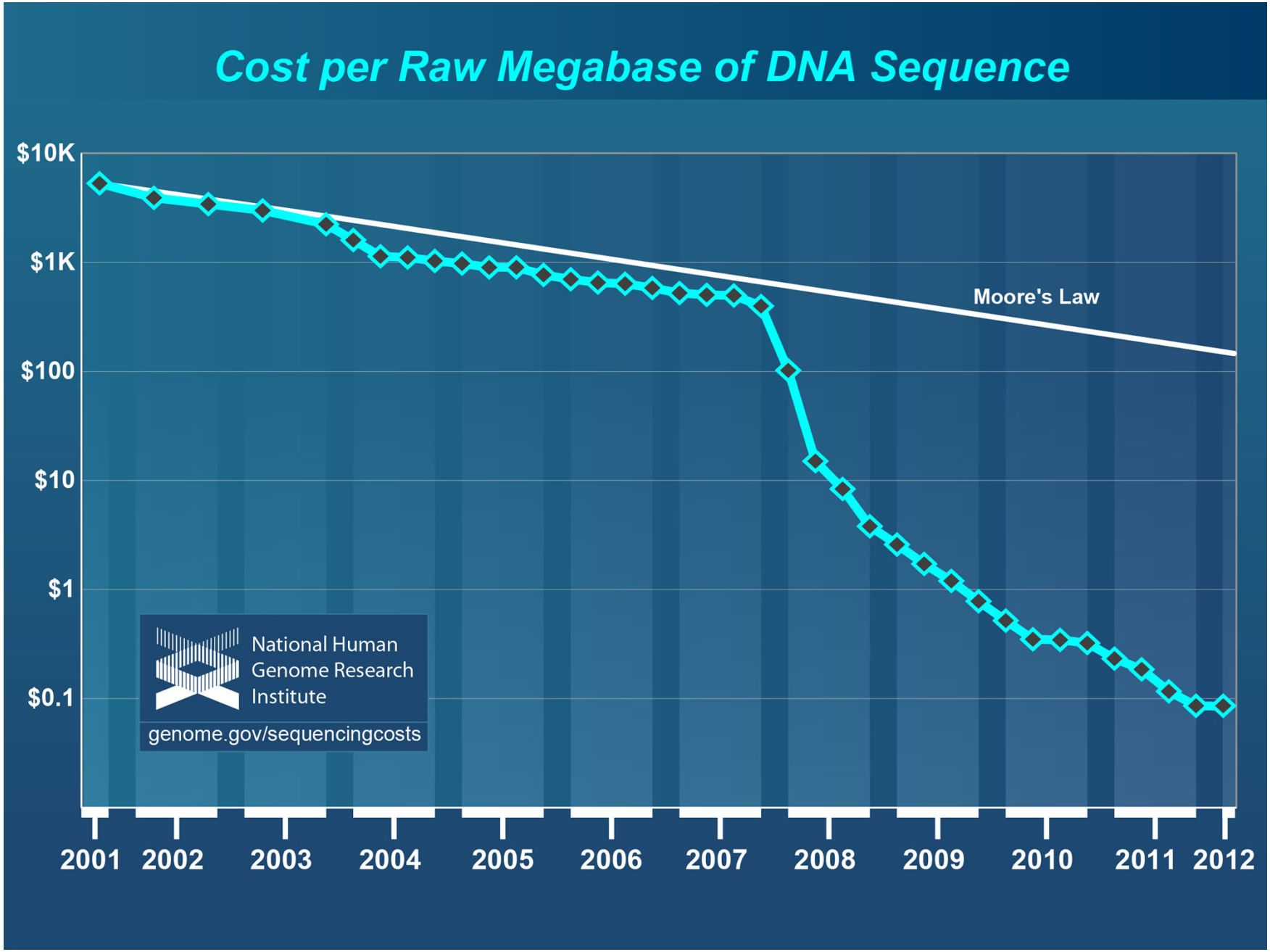


President Clinton and geneticists J. Craig Venter (left) and Francis Collins (right) celebrate

Francis Collins: Private Human Genome project.


Science
16 February 2001
Vol. 291 No. 5507
Pages 1145–1434 59
THE HUMAN GENOME
American Association for the Advancement of Science


nature
the human genome
Nuclear fission
Seafloor spreading
Career prospects
naturejobs


Cost per Raw Megabase of DNA Sequence
$10K
$1K
$100
$10
$1
$0.1
Moore's Law
National Human Genome Research Institute
genome.gov/sequencingcosts
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012

Next-Gen Sequencing Platforms

454/Roche GS-20/FLX
(2005)

PacBio RS (2009-2010)
3rd generation?

Illumina HiSeq
(2007)

ion torrent

by life technologies™

Comparison of Platforms

Technology	Reads per run	Average Read Length	bp per run	Types of errors
454 (Roche)	400,000	250-1000bp	70 Million	Indels
SoLiD (ABI)	88-132 Million	35bp	1 Billion	Indels
Illumina HiSeq	2.5 Billion	100 – 250bp	600 Billion	Substitution
PacBio	45,000	2000-10,000bp	45 Million	Insertions and deletions

\

Sequencing Methods and Terminology

Sanger Sequencing

Sanger method (1977):
labeled ddNTPs terminate DNA copying at random points.

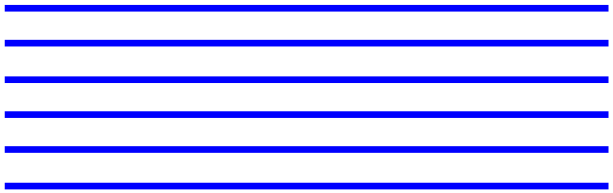
Gilbert method (1977):
chemical method to cleave DNA at specific points (G, G+A, T+C, C).

Both methods generate labeled fragments of varying lengths that are further electrophoresed.

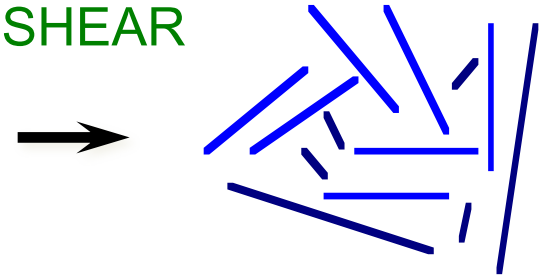
Sanger Sequencing Video

Sanger Sequencing

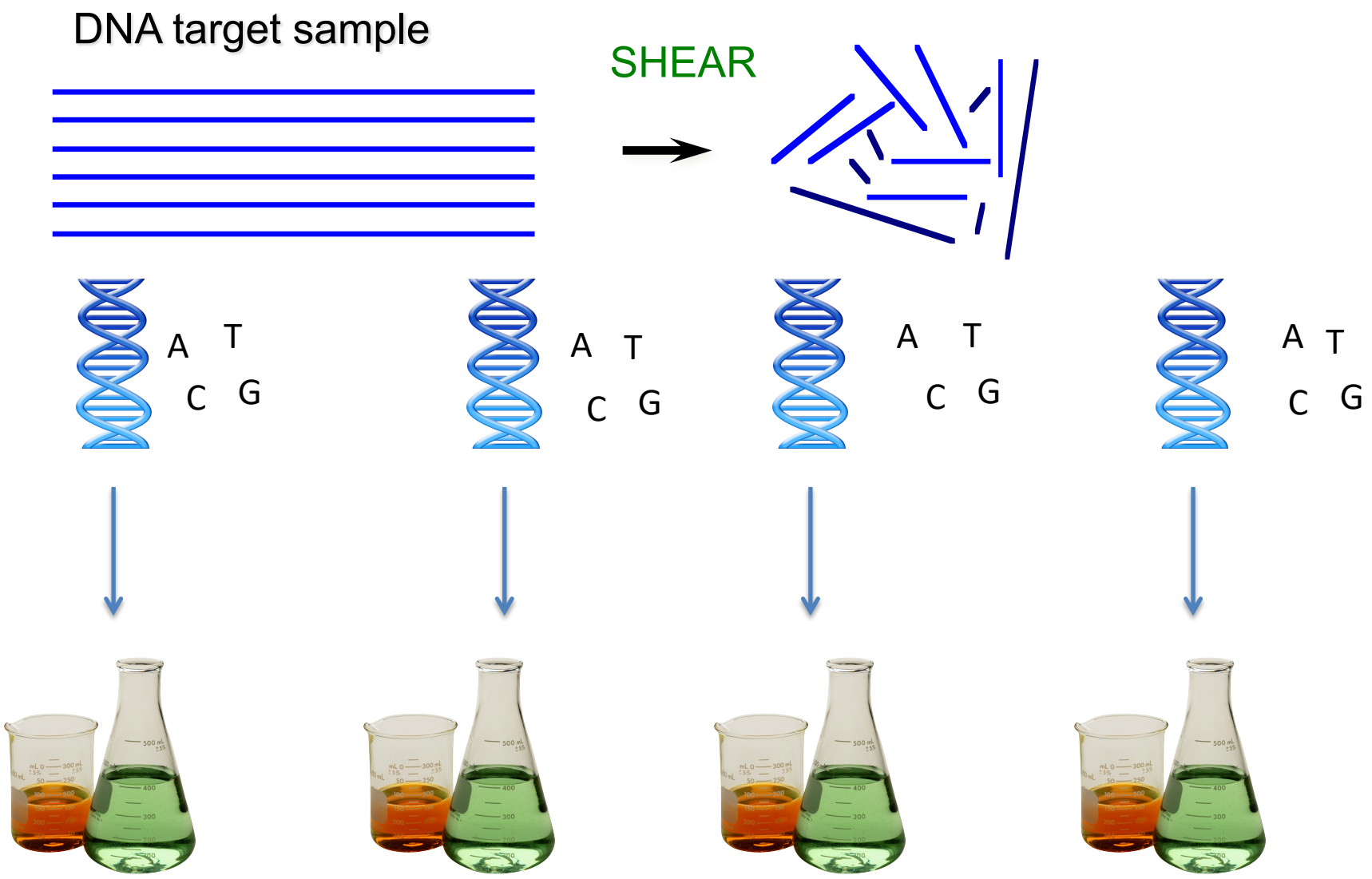
DNA target sample



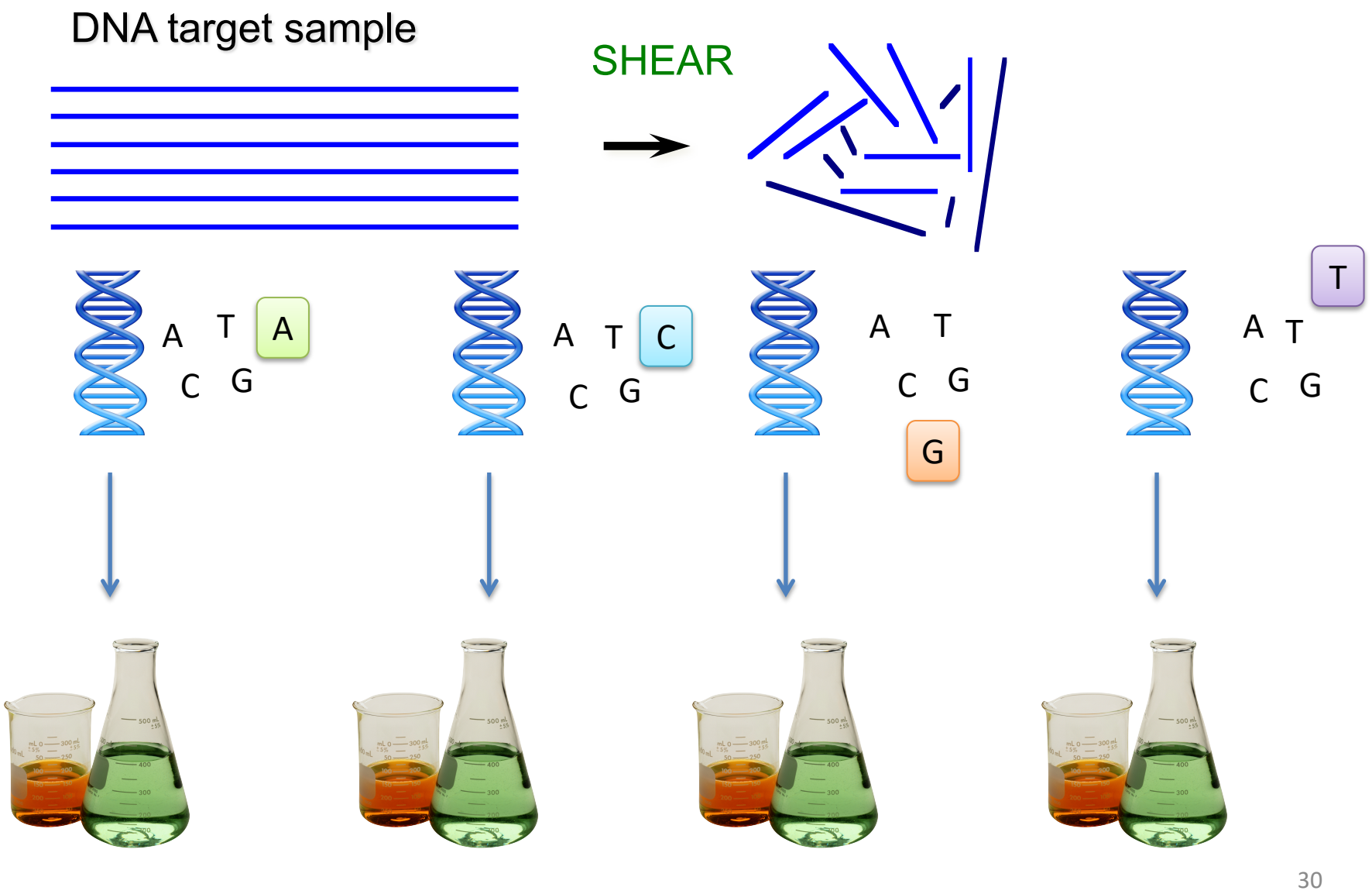
SHEAR



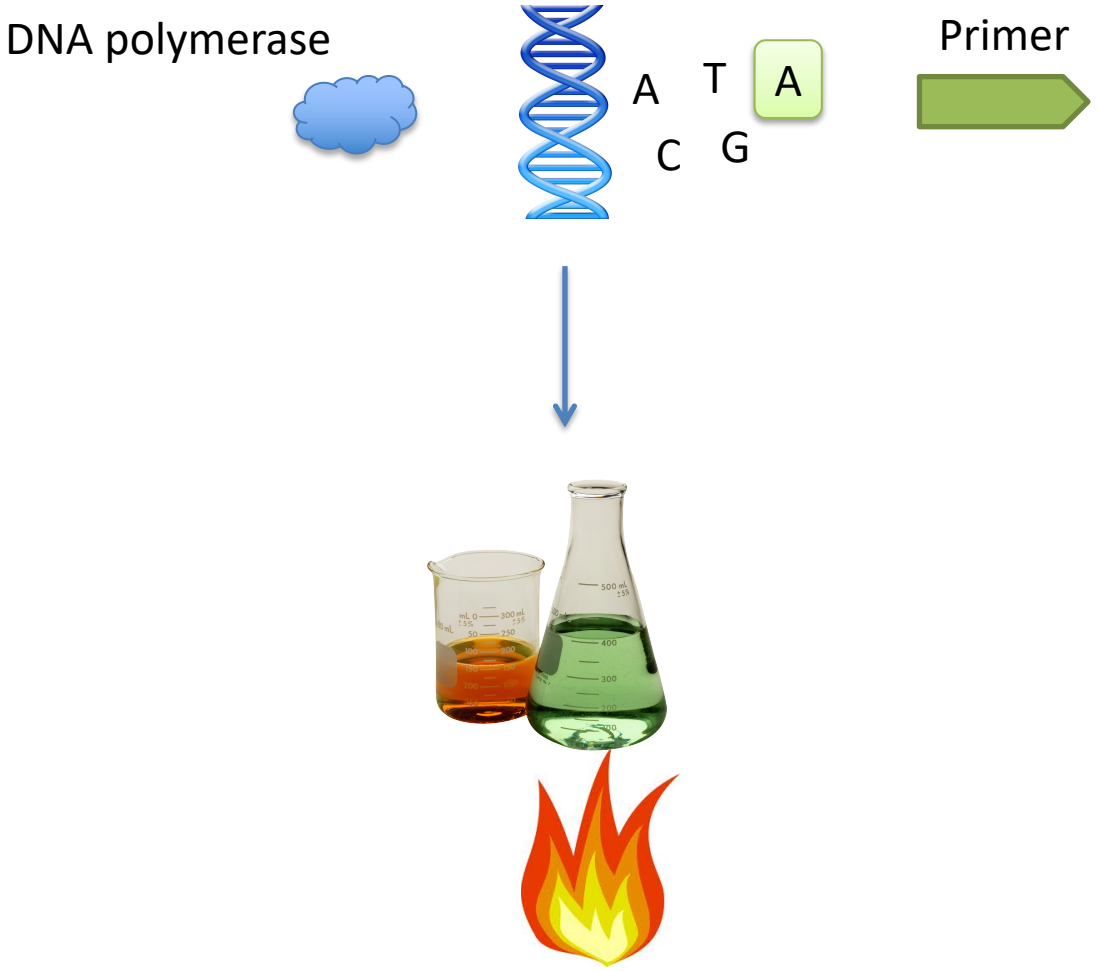
Sanger Sequencing



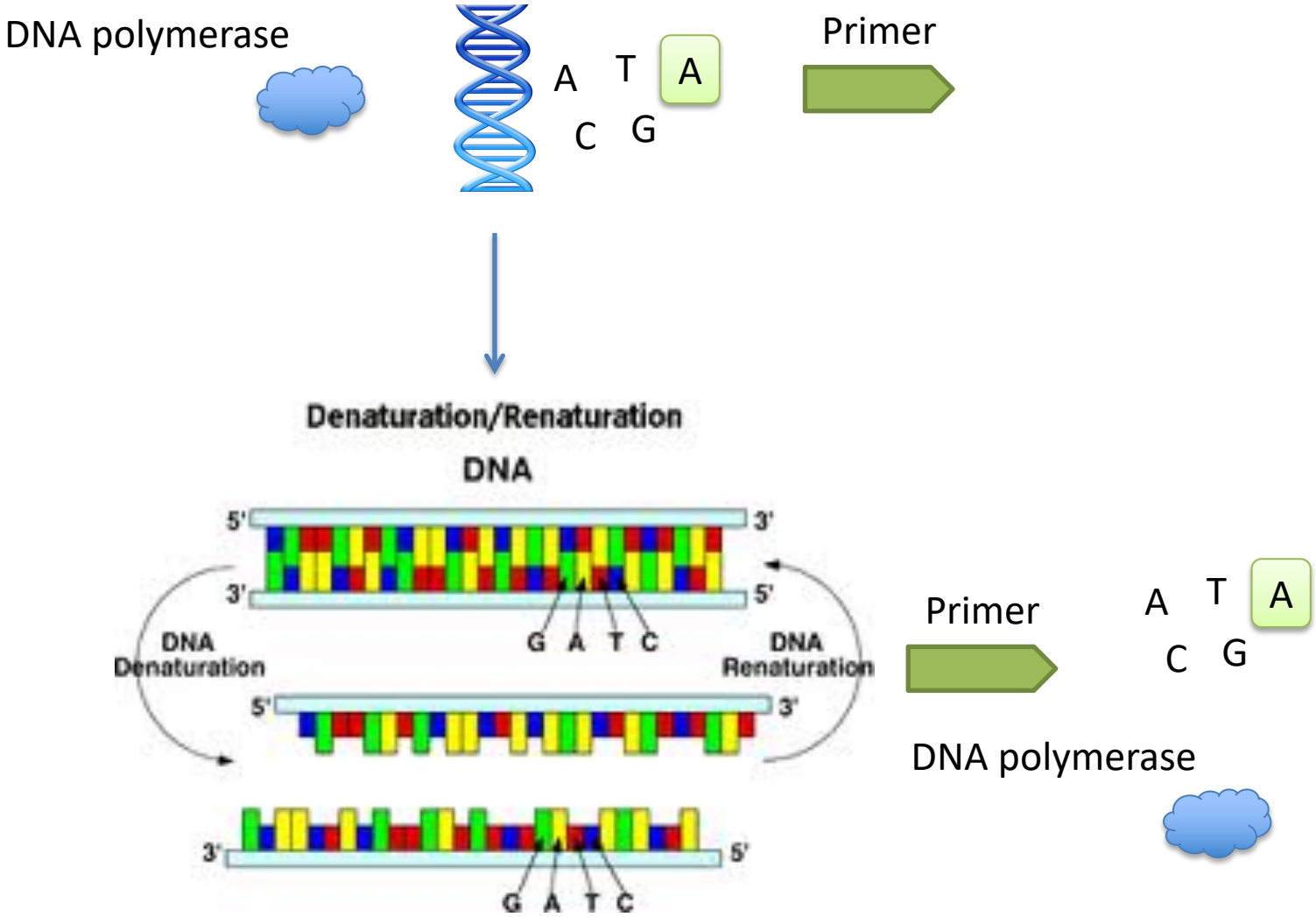
Sanger Sequencing



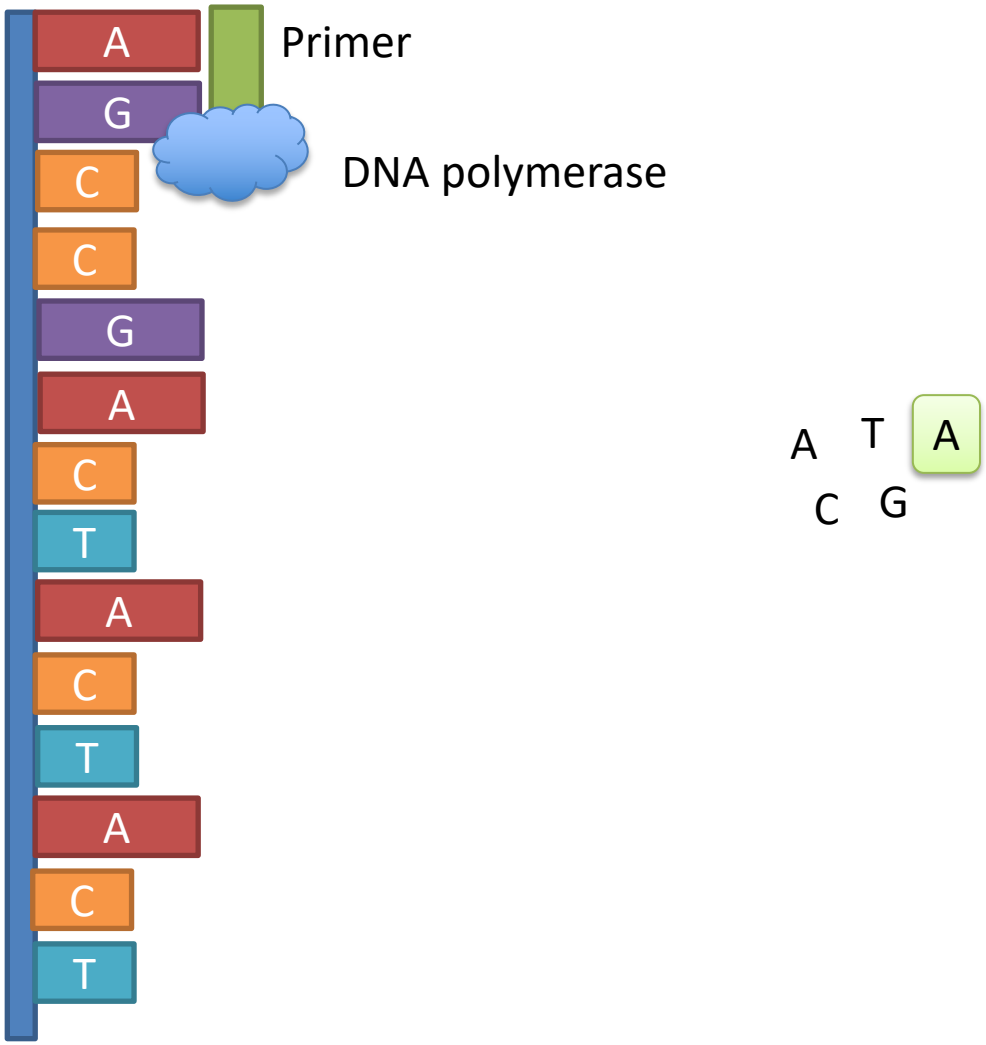
Sanger Sequencing



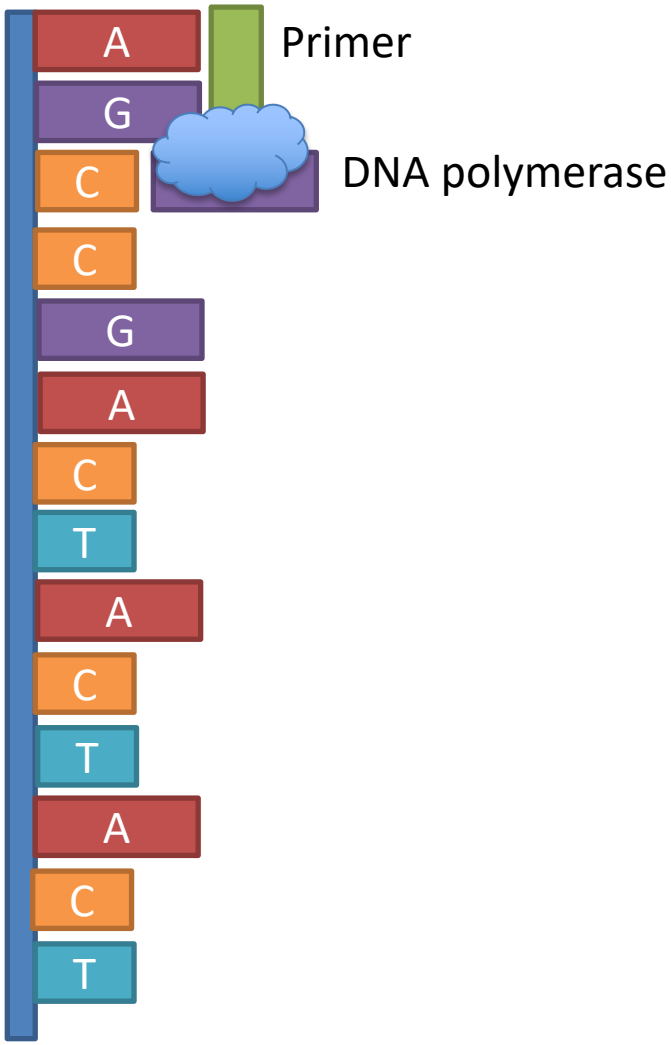
Sanger Sequencing



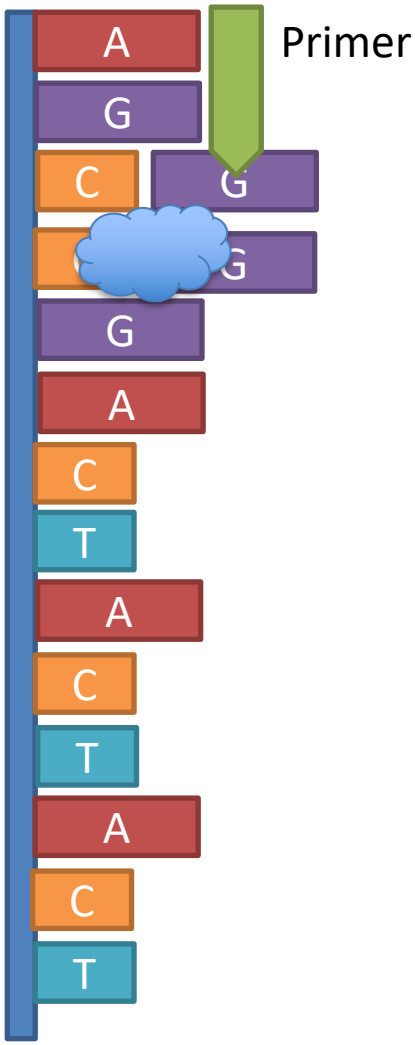
Sanger Sequencing



Sanger Sequencing

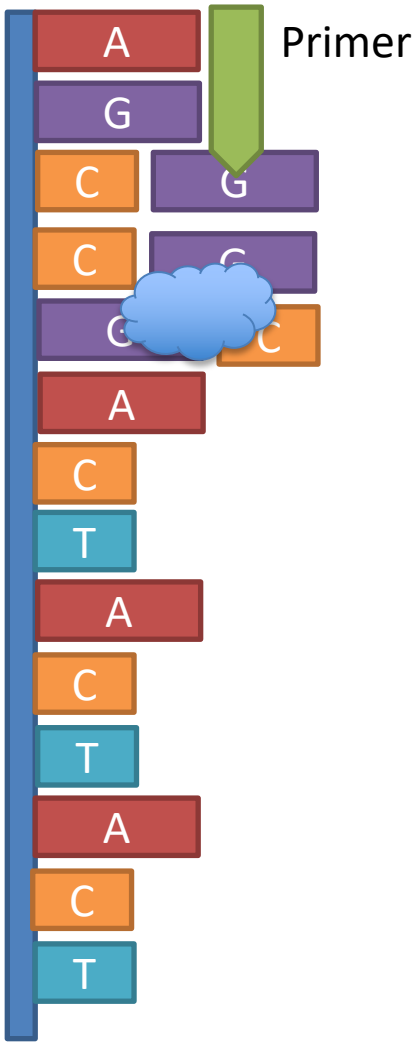


Sanger Sequencing

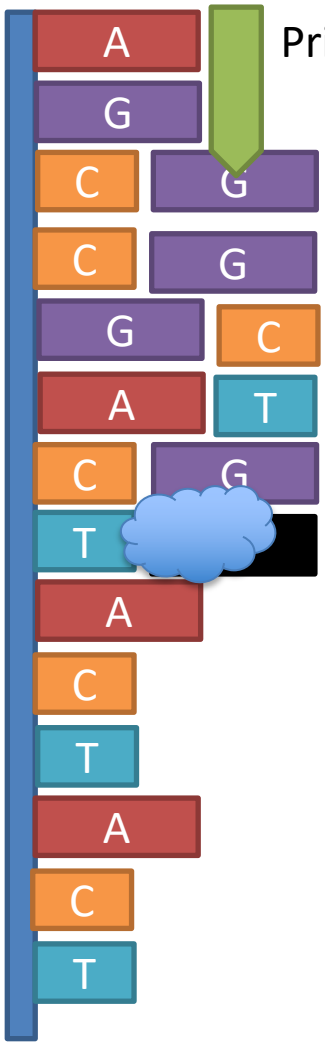


A T A
C G

Sanger Sequencing



Sanger Sequencing

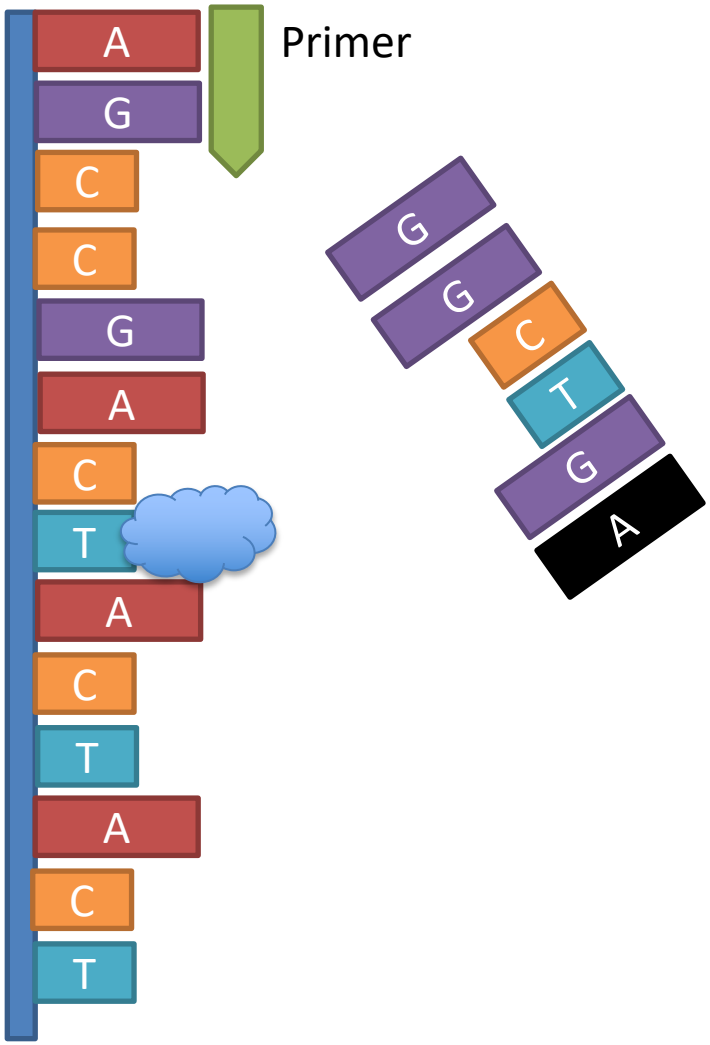


Primer

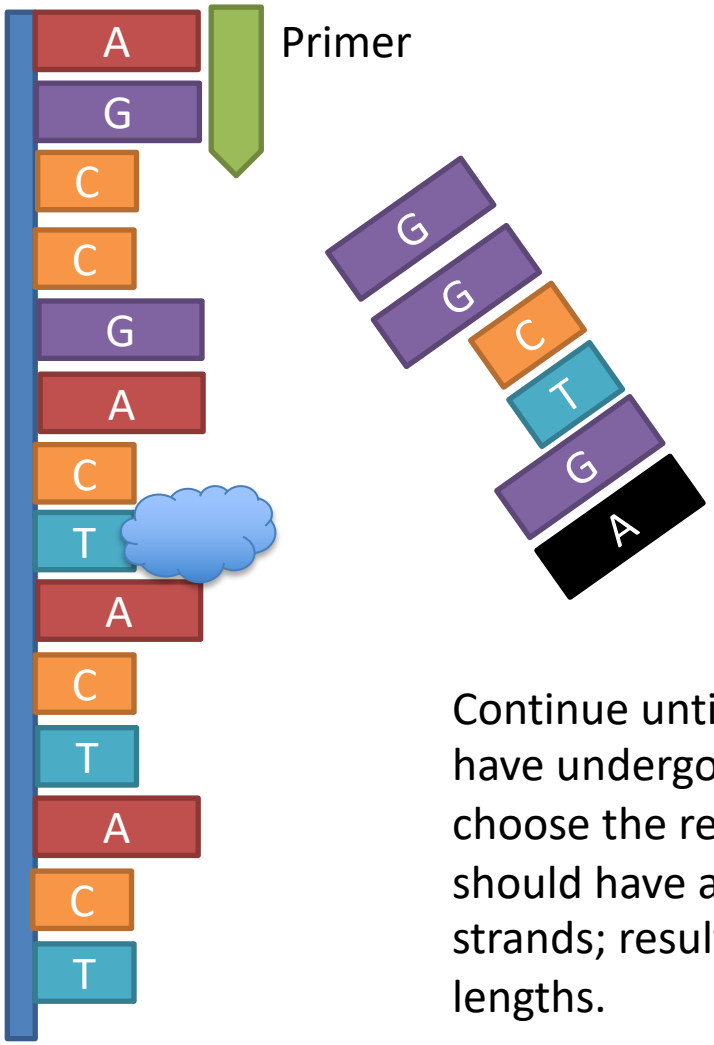
A T A

C G

Sanger Sequencing

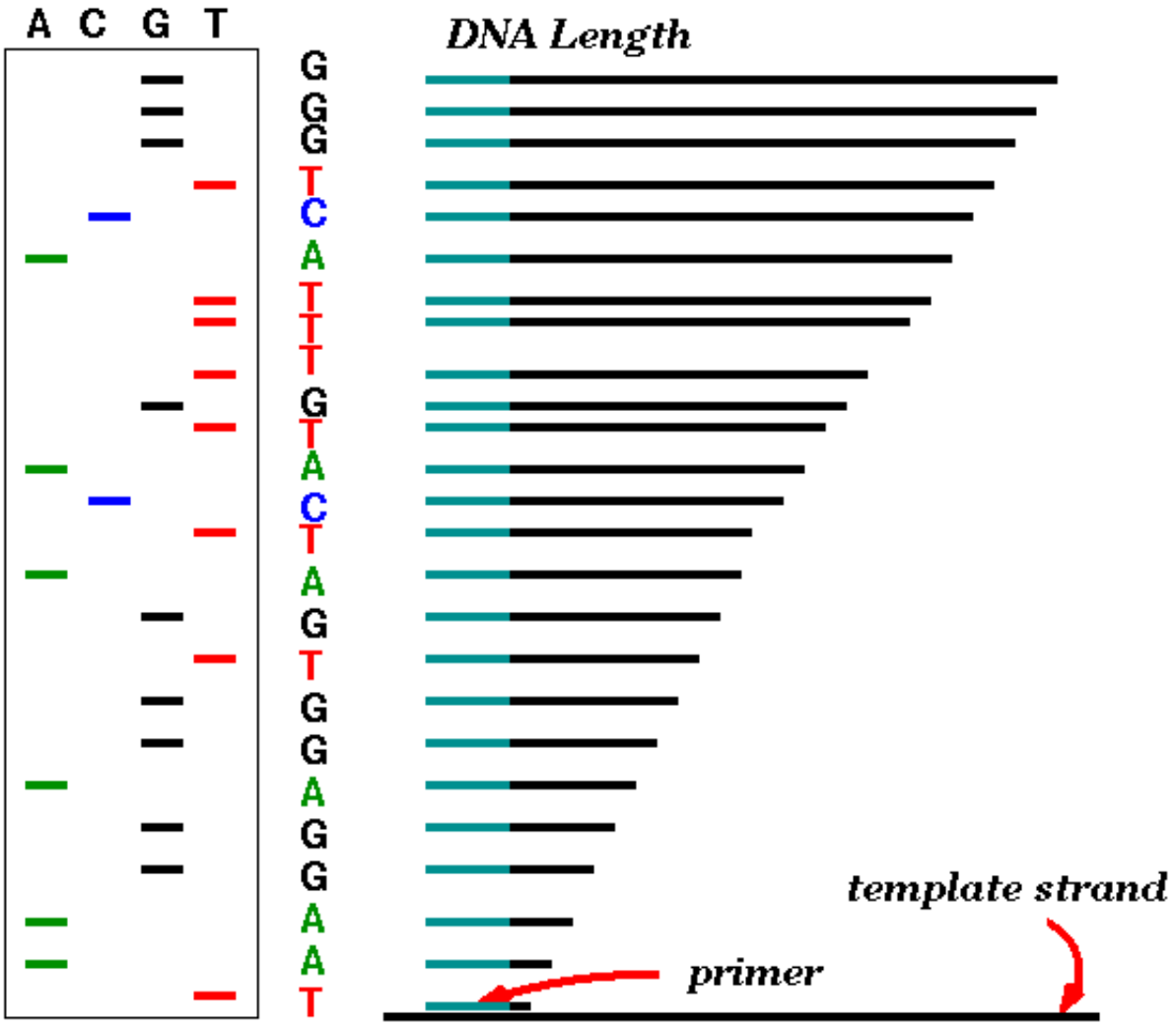


Sanger Sequencing

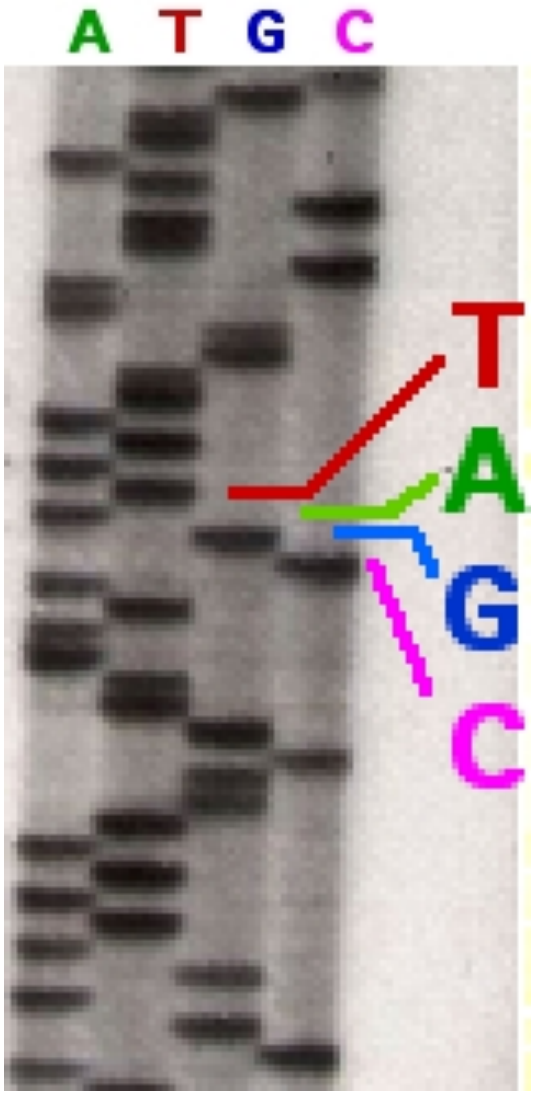


Continue until all strands of DNA have undergone this reaction. If you choose the reagents correctly then you should have all possible A-terminated strands; resulting in sequences of varying lengths.

Sanger Sequencing



Sanger Sequencing



In the gel, the longer DNA fragments move faster to the bottom and the shorter ones move slower and remain at the top.

The sequence can be read off by going from top to bottom.

Challenges

- Requires **a lot of space and time**: you need a place to run the reaction, and then you need a gel to determine the length of the DNA
  - You could only run perhaps a hundred of these reactions at any one time.
  - There are 3 billion base pairs of DNA in the human genome, meaning about 6 million 500-base pair fragments of DNA.
- Nonetheless it was still used to come up with the first copy of the human genome

Celera Sequencing (2001)

- 300 ABI DNA sequencing platforms
- 50 production staff
- 20,000 square feet of wet lab space
- 1 million dollars / year for electrical service
- 10 million dollars in reagents

Total cost of human genome: 2.7 Billion dollars

Celera Sequencing (2001)

- 300 ABI DNA sequencing platforms
- 50 production staff
- 20,000 square feet of wet lab space
- 1 million dollars / year for electrical service
- 10 million dollars in reagents

Current cost of human genome: < 1,000 $

Second/Next Generation Sequencing

- Second generation sequencing techniques overcome the restrictions by finding ways to sequence the DNA without having to move it around.
- You stick the bit of DNA you want to sequence in a little dot, called a **cluster**, and you do the sequencing there; as a result, you can pack many millions of clusters into one machine.

Sequencing a strand of DNA while keeping it held in place is tricky, and requires a lot of cleverness.

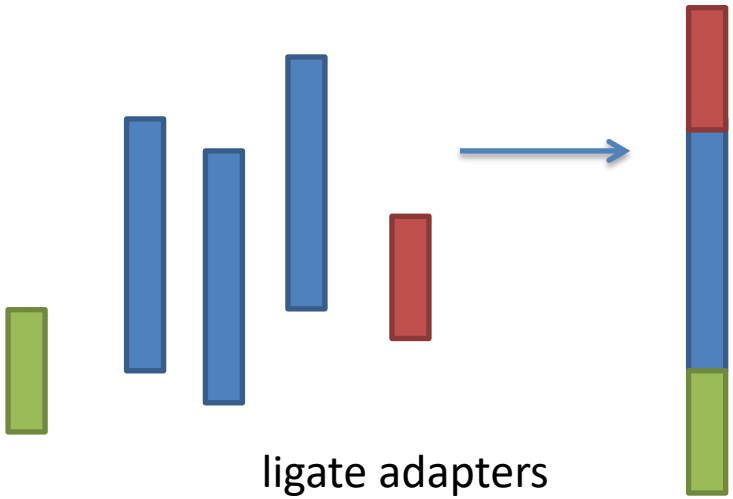
Illumina Sequencing: Video

Steps in Illumina sequencing

- Turn on the sequencing machine and wait (1 week)…

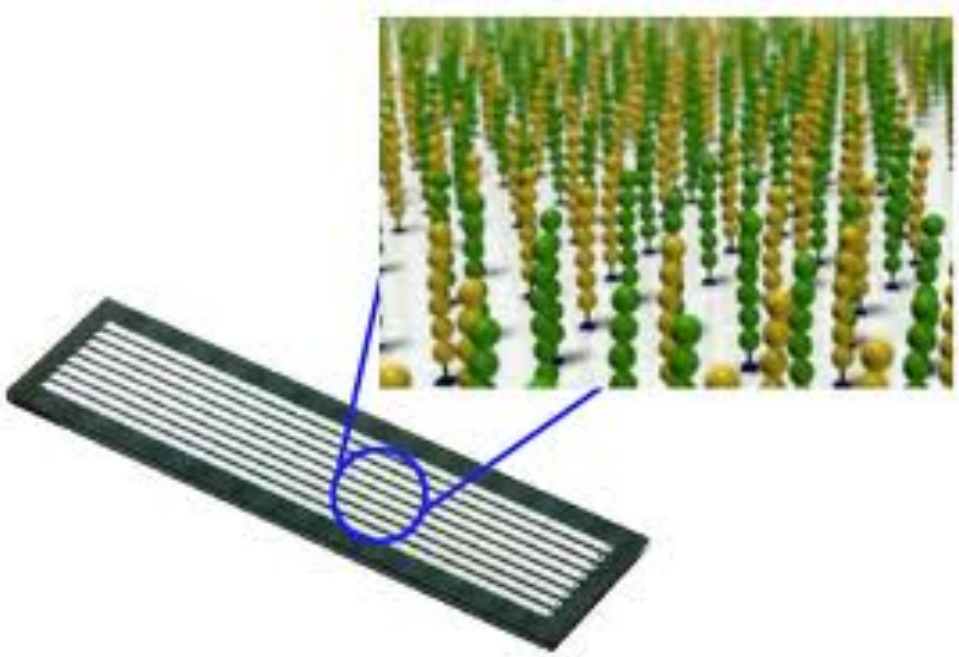
Steps in Illumina sequencing

- Sample prep: size select fragments, add adapters to ensure the fragments ligate to the flow cell (1 to 5 days)



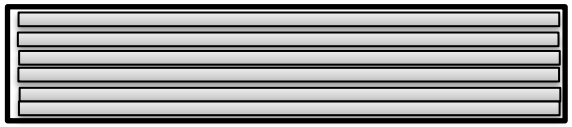
Steps in Illumina sequencing

- Cluster generation on flow cell

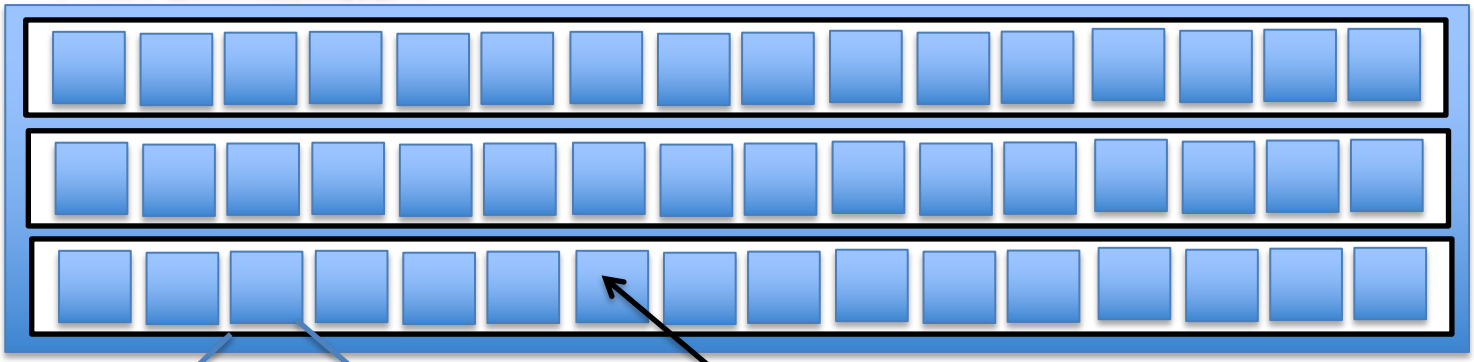


Why do we need clusters?

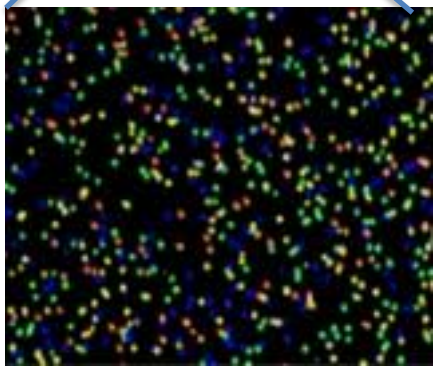
A flow cell
contains 8 lanes



Each lane contains three columns of tiles



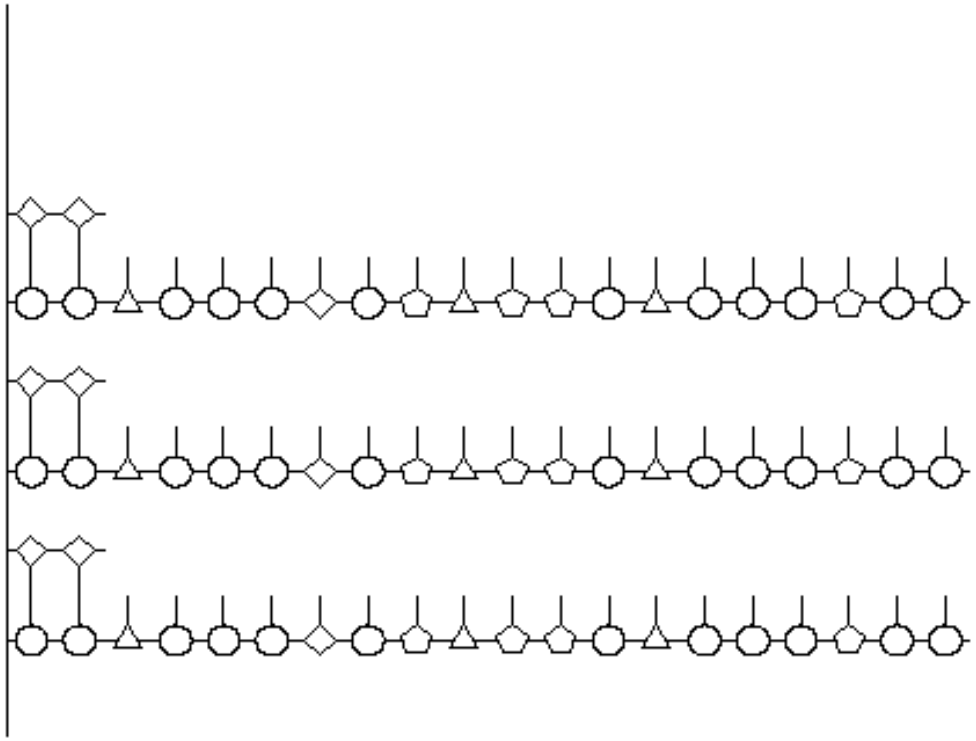
Each column contains 100 tiles



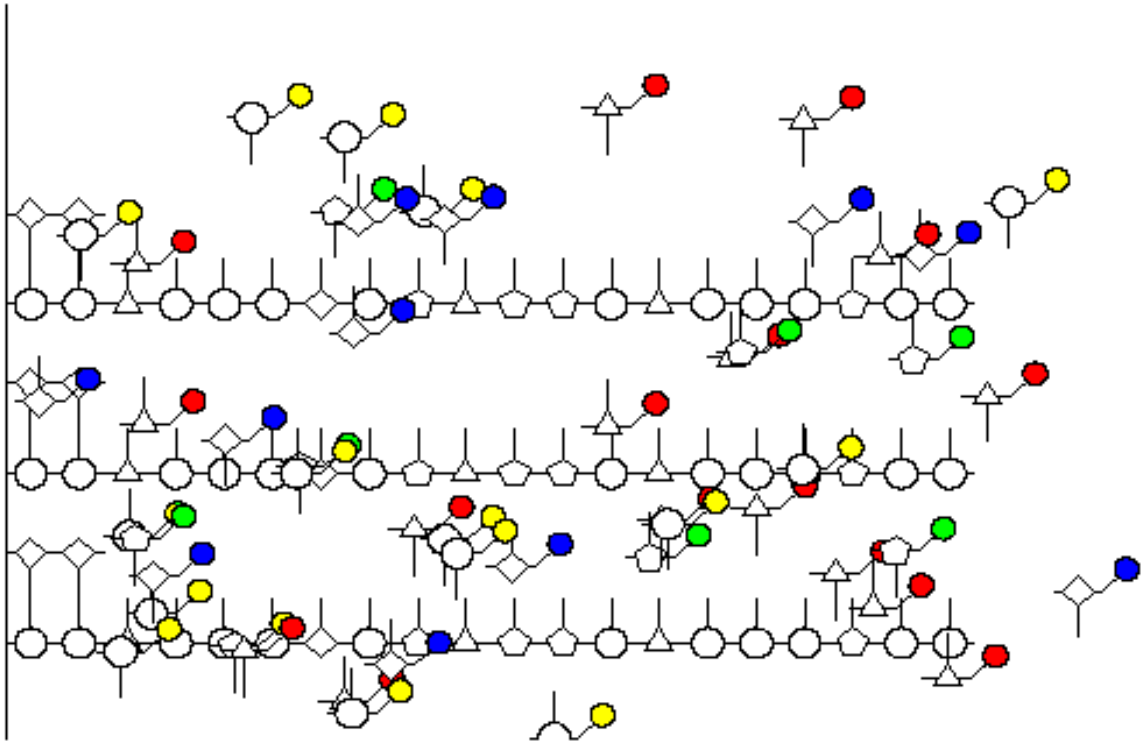
20K to 30K clusters

Each tile is imaged four times per cycle,
which is one image per base

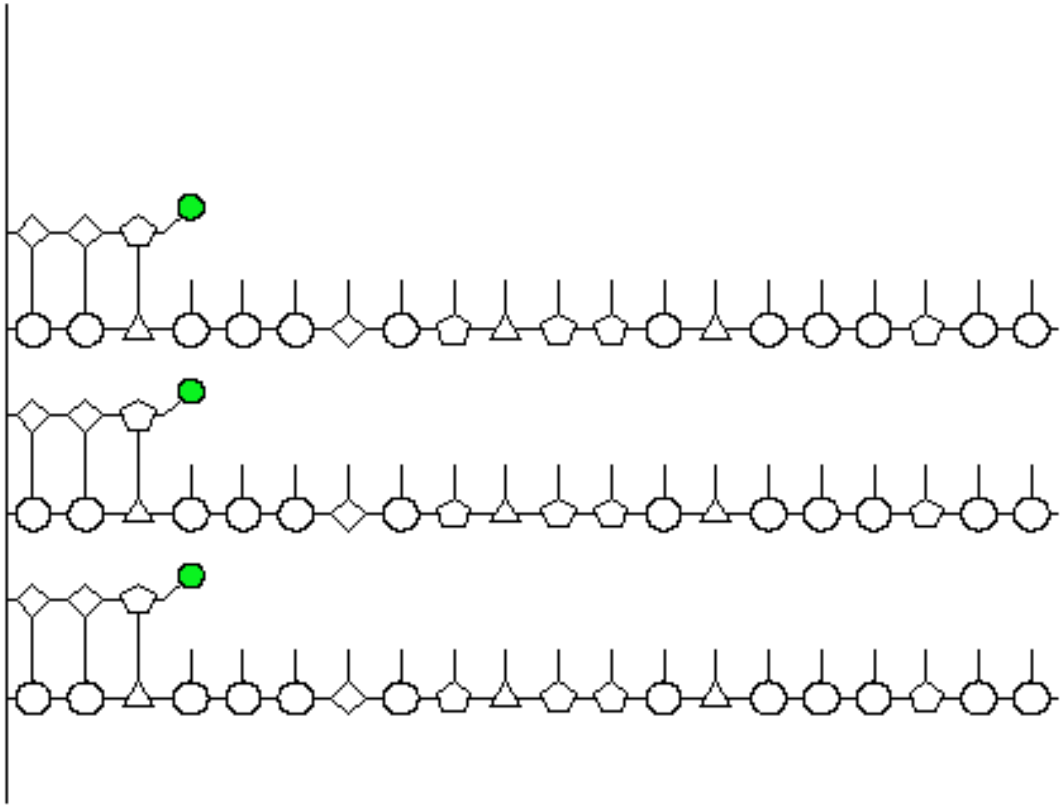
We multiply up the template stand, i.e. the bit of DNA that we are sequencing, and stick on a few bases of ‘adaptor sequence’; this sequence sticks on to complementary bits of DNA stuck to a surface, which holds the DNA in place while we sequence it:



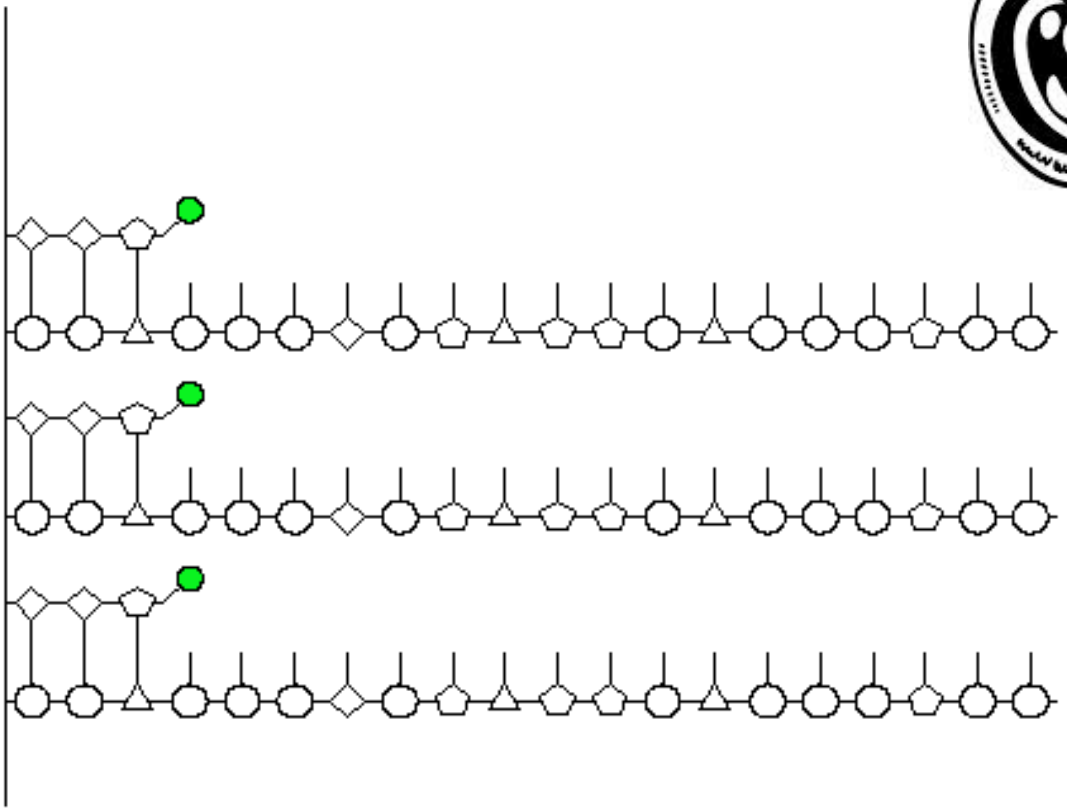
We then flood the DNA with Reversible Terminator (RT)-bases. We also add a polymerase enzyme, which incorporates the RT-base into the new strand that is complementary to the template strand:

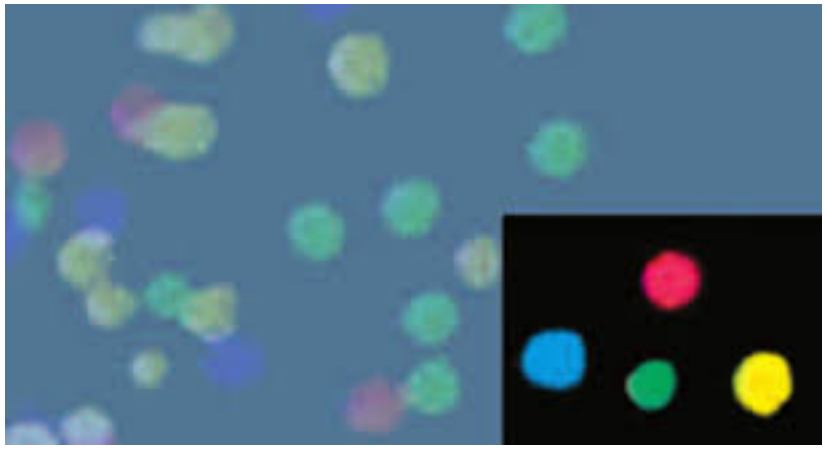


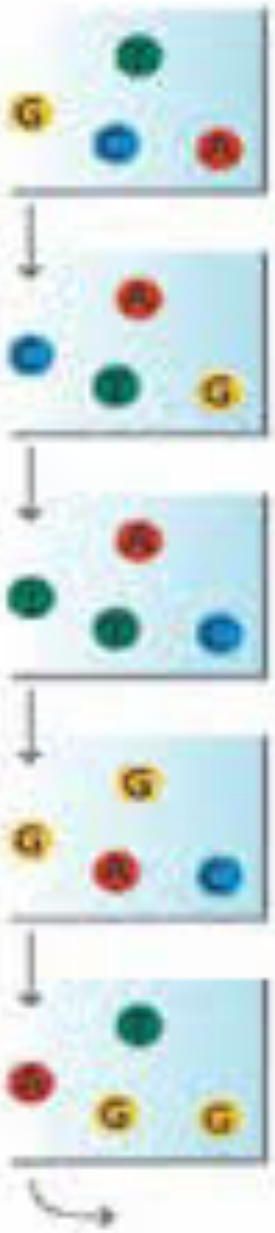
We then wash away all the RT-bases, leaving just those that were incorporated into the new strand; we can read off what base this is by looking at the color of the dye:

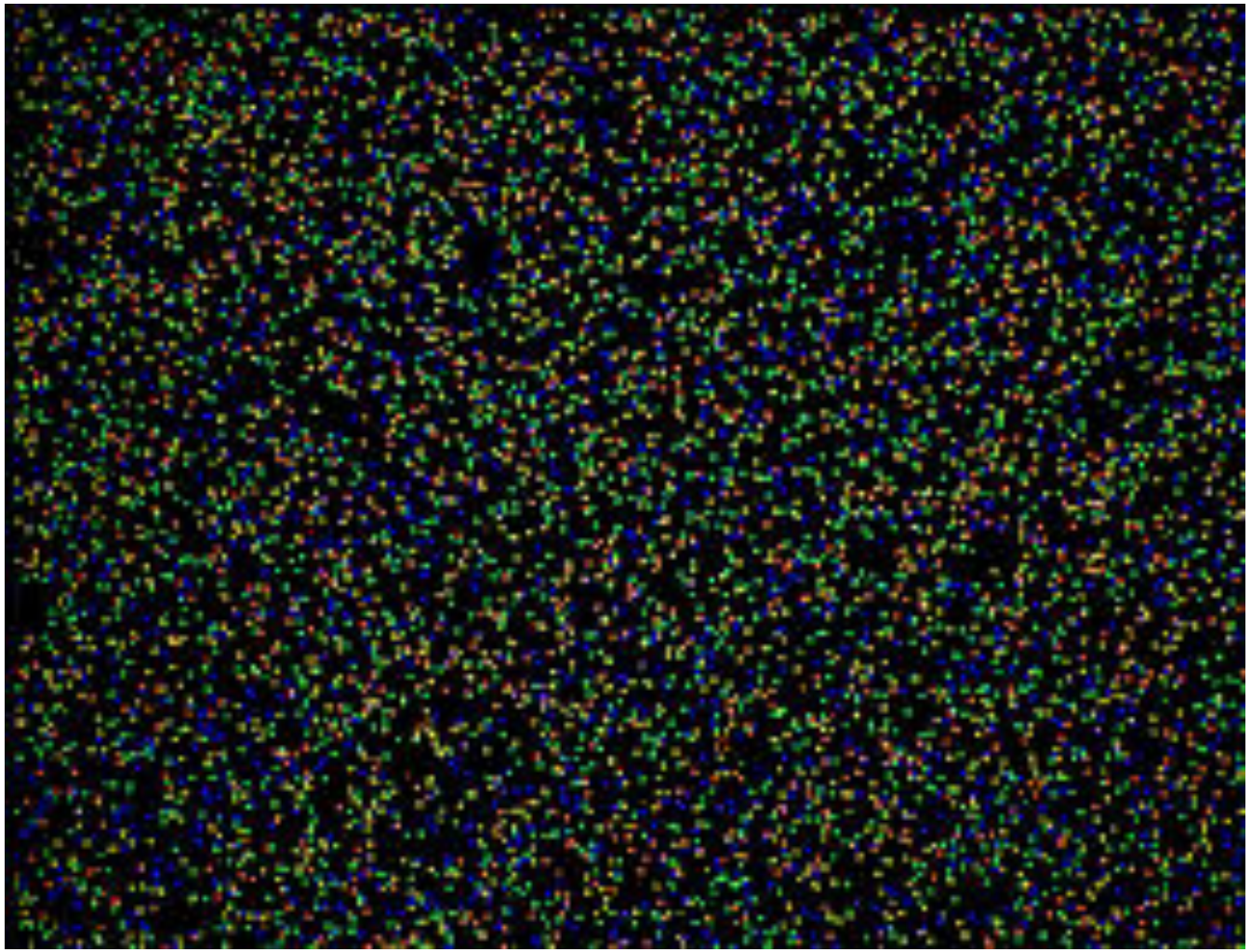


There exists a cleavage enzyme that chops all the extra molecules off, and turns the RT-base into a normally functioning nucleotide.









Illumina Characteristics

- Illumina uses the modified version of Sanger sequencing called *reversible terminator* method.

- The dye is washed after imaging and the last nucleotide is extended in the next round.

- In a single Illumina machine we have hundreds of millions of these clusters; cameras look at all of these dots and record how they change color over time, allowing you to determine the sequence of bases of millions of bits of DNA at once.

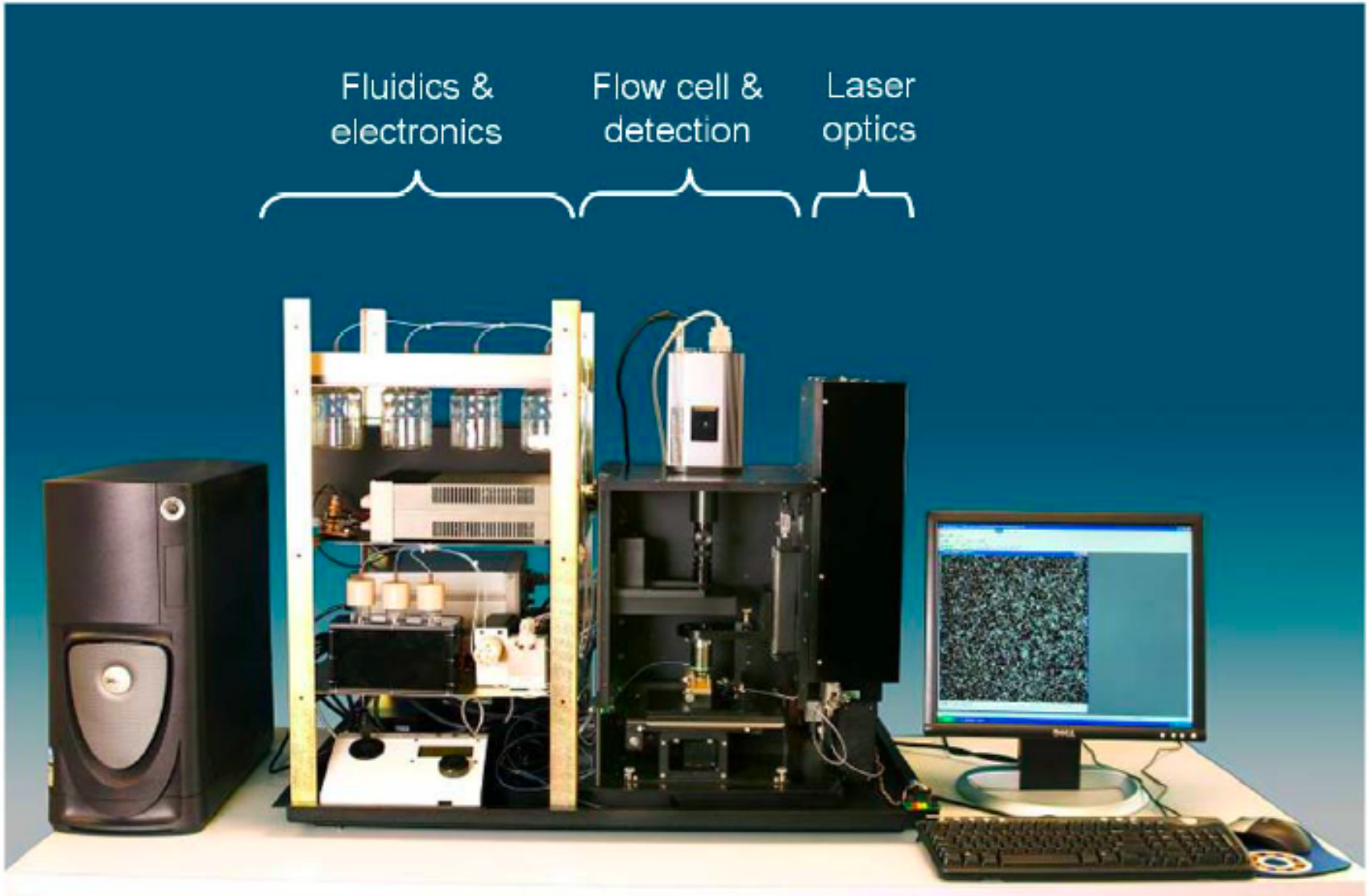
Illumina Characteristics

- Sequencing method is actually pretty inefficient, however, the machine is capable of sequencing millions of fragments of DNA at once.

- Due to controlled sequence of termination, washing, and chemical deactivation/activation events, Illumina reads have (almost) only substitution errors.

- Paired reads with small insert size (< 800 bp) can be reliably generated. Large insert mate pairs can be made using unreliable, difficult, time-consuming, and expensive chemical hacks.

Inside the Illumina Machine



Pyrosequencing: Video 454 Roche System

https://www.youtube.com/watch?v=nFfgWGFe0aA

Pyrosequencing Characteristics

- Pyrosequencing differs from Sanger sequencing, in that it relies on the detection of pyrophosphate release on nucleotide incorporation, rather than chain termination with dideoxynucleotides.

- Since there is no chain termination in pyrosequencing other than by designed unavailability of the other 3 nucleotides, pyrosequencing reads have insertion/deletion errors particularly in or next to runs of homopolymers:

  hard to distinguish between AAAAA and AAAAAA

Pyrosequencing Characteristics

- Relatively long reads: 800-1000 bp.
- Reliable paired read protocol with large insert sizes: 3 kbp, 8 kbp, 20 kbp.
- For instance, a pair of 1000 bp reads back to back (insert size = 2 kbp) essentially gives a 2000 bp read.
- Dealing with 2%-3% indels in 454 reads is the main challenge beside higher sequencing costs in comparison with Illumina.

Single Molecule Sequencing: Video Pacific Biosciences System

https://www.youtube.com/watch?v=v8p4ph2MAvI
https://www.youtube.com/watch?v=NHCJ8PtYCFc

PacBio Characteristics

- PacBio reads are long, e.g. on average a few kilobases.
- Since PacBio relies on the signal from a single molecule, the signal to noise ratio is small, and PacBio reads have lots of *uniformly random* errors, up to 15%.
- PacBio errors are primarily indels, which makes efficacious computational error correction currently intractable.
- PacBio reads are currently used for limited validation of contiguity information or helping datasets generated with other technologies.

Nanopore Sequencing: Video Oxford Nanopore System

https://www.youtube.com/watch?v=3UHw22hBpAk

Oxford Nanopore Characteristics

- Nanopore reads are pretty long, up to 100+ kbp.
- They have lots of errors, 10%-40%.
- Errors are primarily indels like PacBio's but the Nanopore error model is not clear yet [PacBio errors are pretty much uniformly random].
- Nanopore has just started a world-wide benchmarking project which is still going on.