Sequencing Technologies - Part I

Sonja J. Prohaska

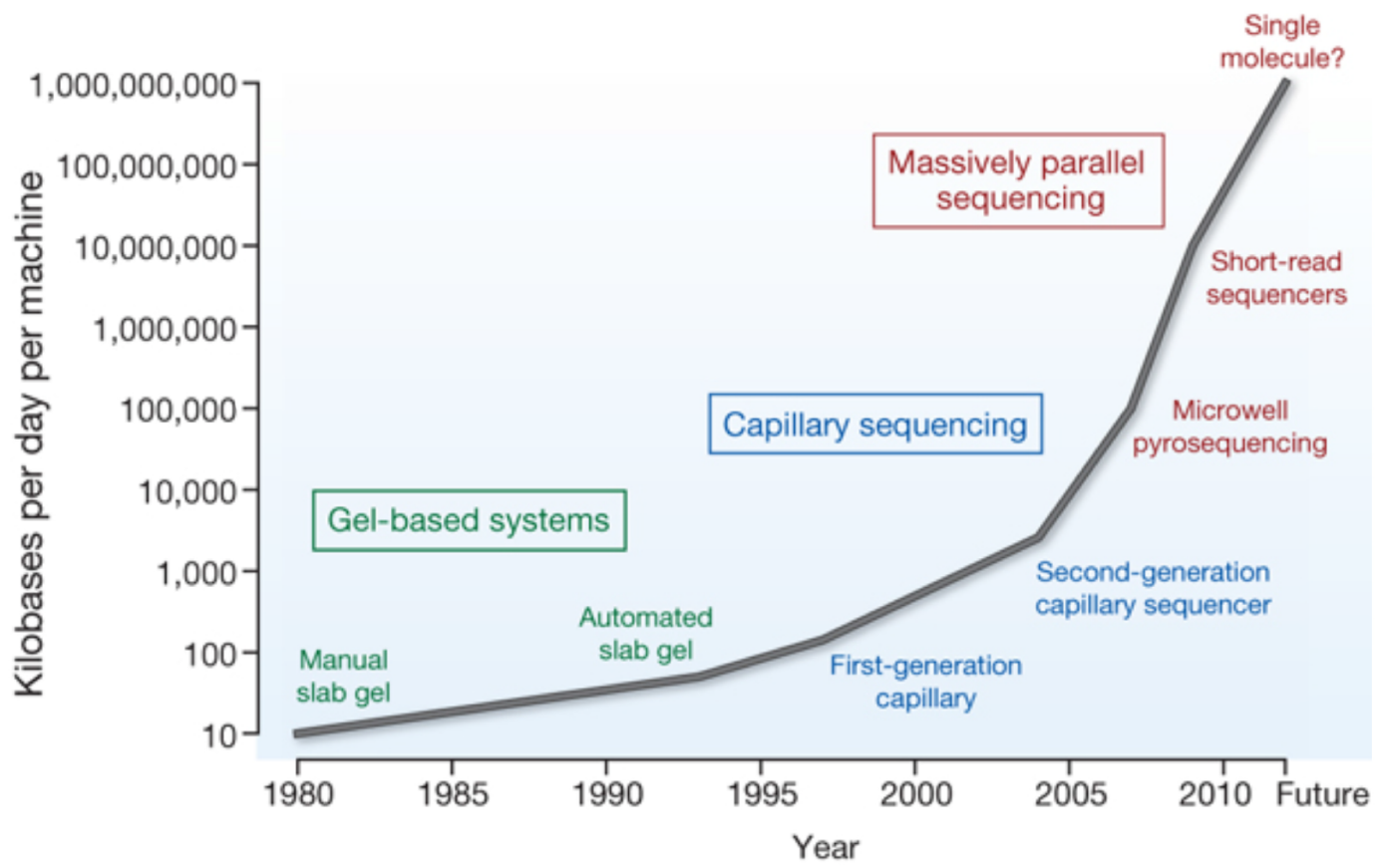
Computational EvoDevo, University of Leipzig
Santa Fe Institute, Santa Fe, NM, U.S.

WS 2025/26

High-Throughput Sequencing: What is sequenced and why?

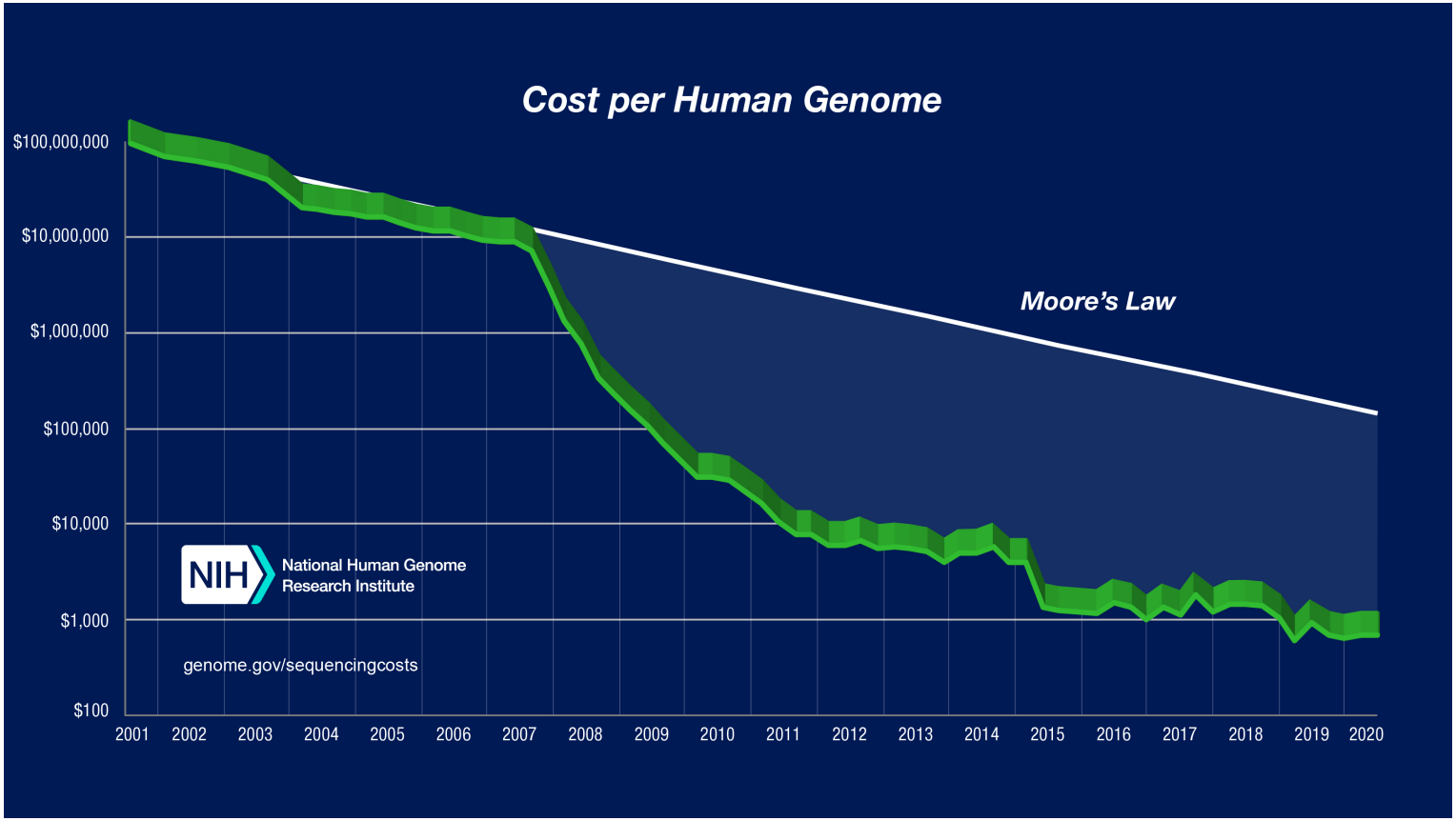
- High-Throughput sequencing (HTS) allows us to determine the nucleic acid sequence of the whole **genome**, i.e. all genetic material (DNA), or the whole **transcriptome**, i.e. the full set of transcripts (RNA), of a cell or organism.
- The "genome" can give us an insight in a cells repertoire of genes and regulatory units.
- The "transcriptome" sheds light on those genes that are currently expressed.
- clever exertion of lab, sequencing, and bioinformatic techiques makes it possible to even examine specific **regulatory interactions**

Rate of Sequencing over time



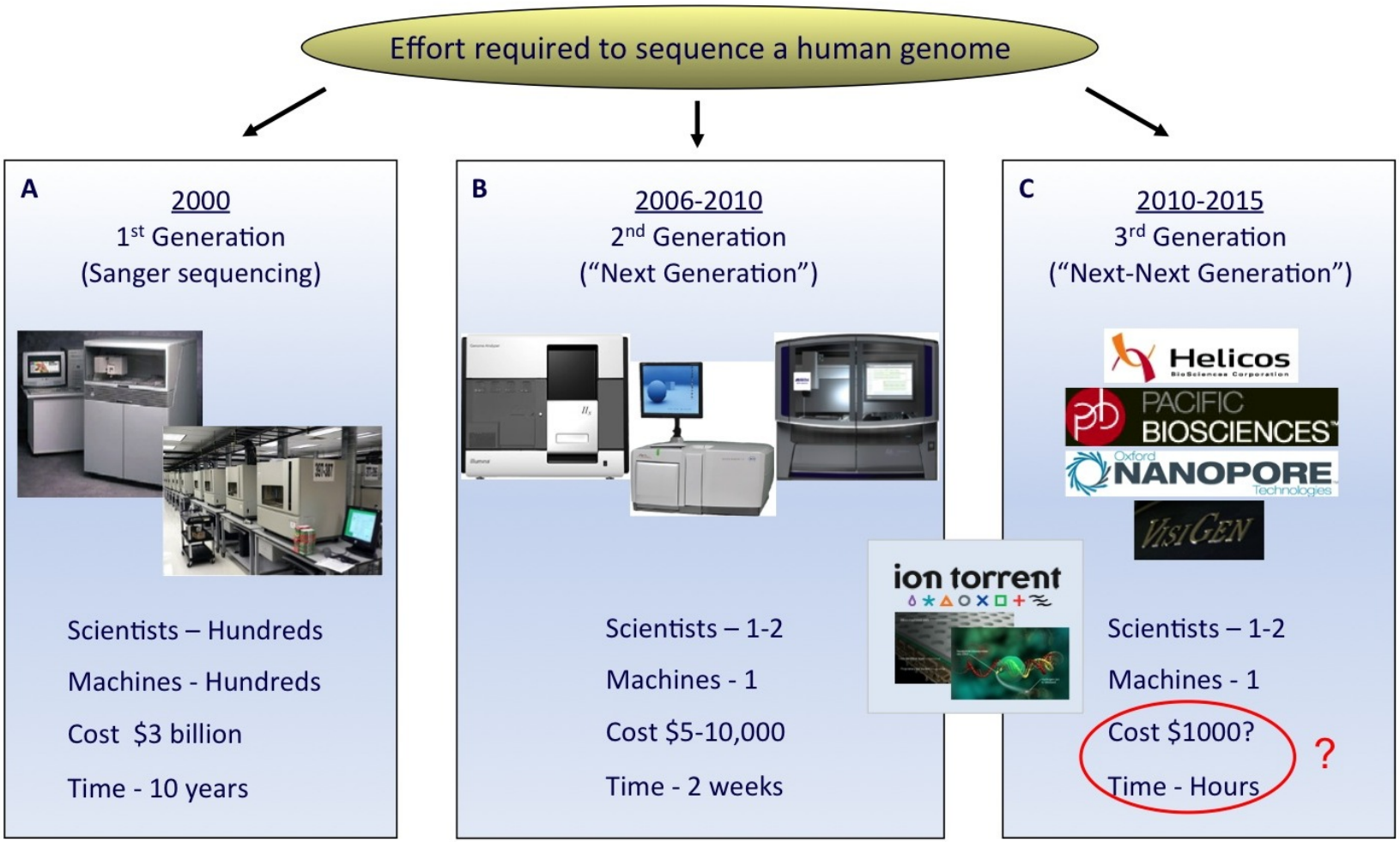
Reprinted by permission from Macmillan Publishers Ltd: Nature 458, 719-724 (2009)

Sequencing Cost Over Time



Wetterstrand KA. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP) Available at: www.genome.gov/sequencingcostsdata. Accessed 12/11/2020.

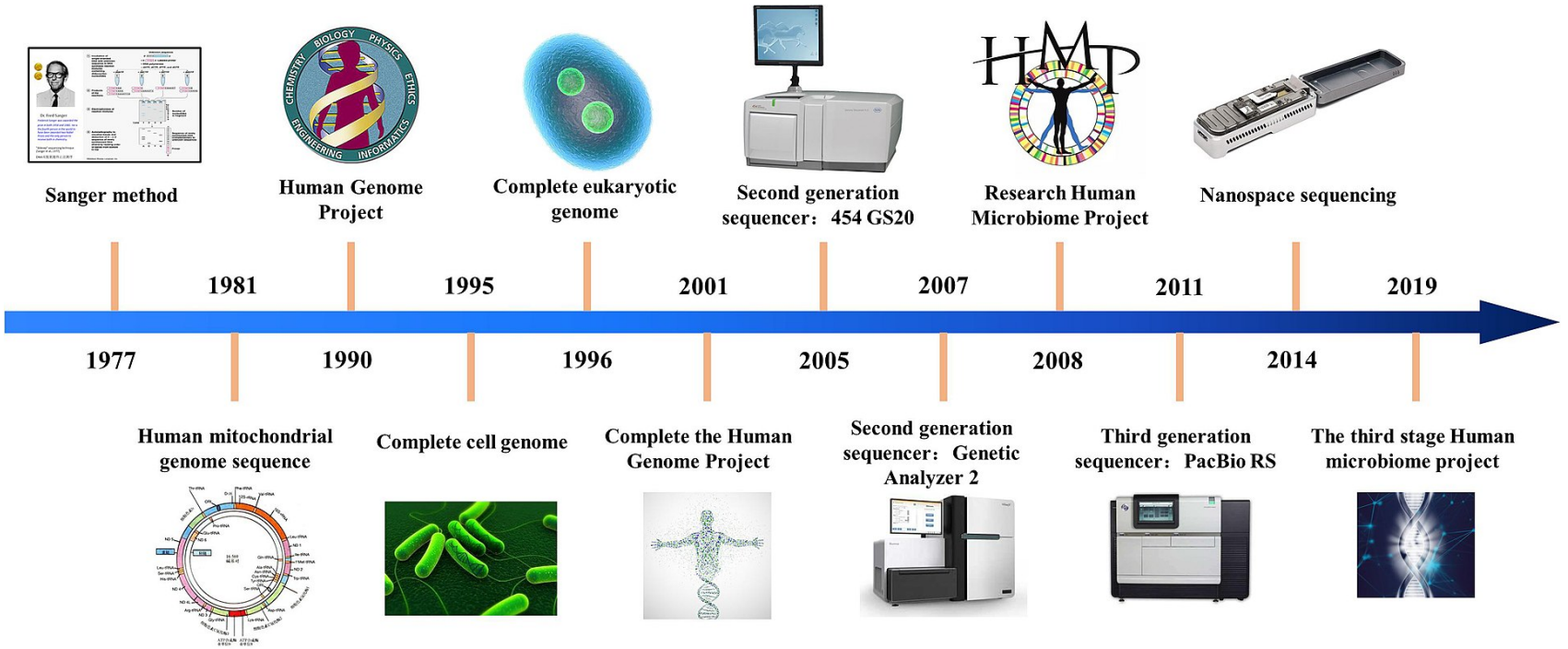
Currently Available Sequencing Technologies



Sequencing Today

- the throughput of sequencing technologies has increased drastically over the last two decades
- meanwhile the cost for sequencing was decreasing quickly
- the throughput is still going up
- the cost per human genome stays around 1,000$ since some time
- sequencing is “cheap enough” for medical purposes at the moment
- focus is put on improving the quality
- most expensive item after the sequencing itself is the bioinformatics analysis (followed by consumable or lab personel)

Development of Sequencing Technologies



- **1st generation**: since 1977, Sanger sequencing
- **2nd generation**: since 2005, short read sequencing

  e.g. Illumina
- **3rd generation**: since 2009, single molecule, long read sequencing

  e.g. Nanopore, PacBio

Heather, James M., and Benjamin Chain. "The sequence of sequencers: the history of sequencing DNA." Genomics 107.1 (2016): 1-8.

Sequencing is Based on a "Simple" Concept

DNA synthesis in vitro (meaning "im Glas")

Components

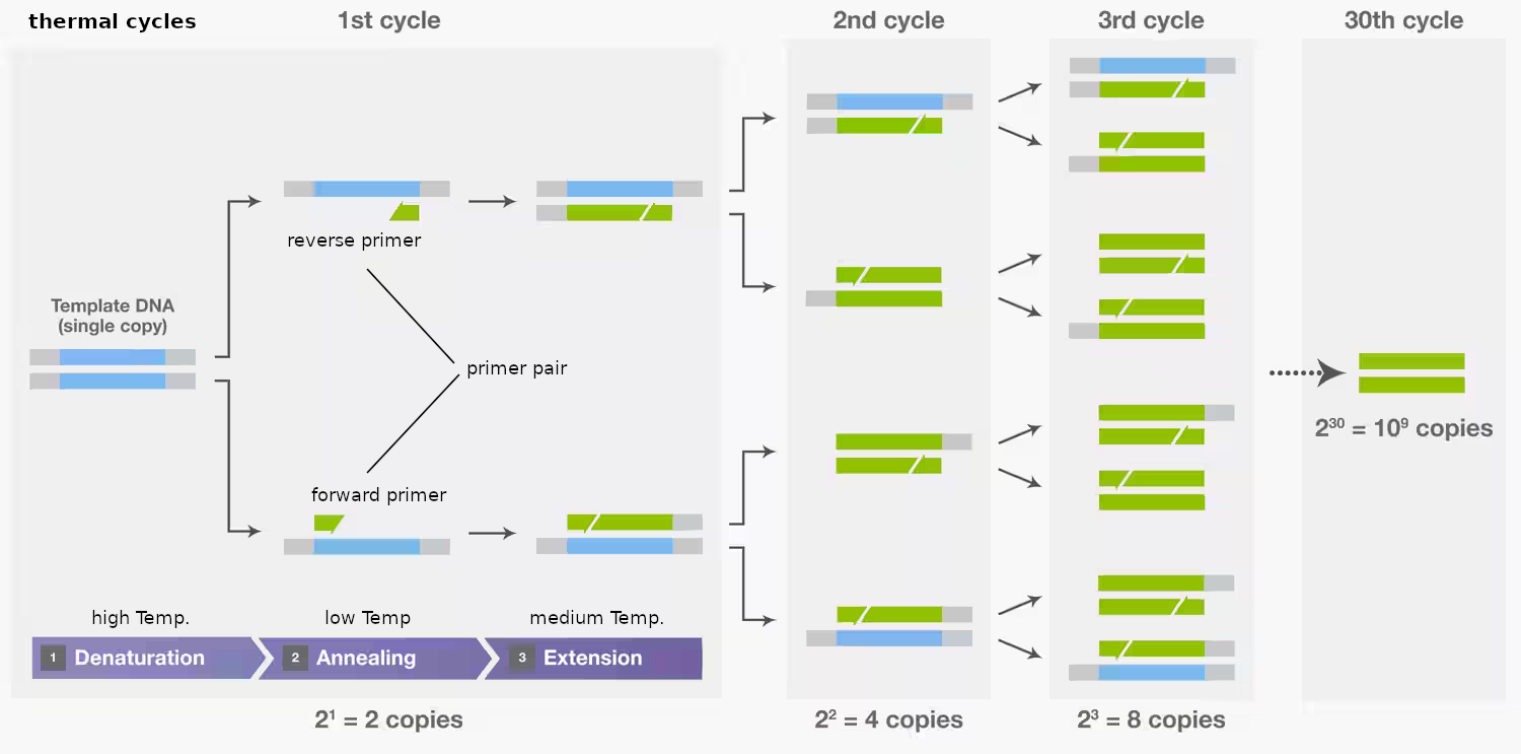
- DNA template (a single copy)
- regular deoxynucleotides: dATPs, dCTPs, dGTPs, dTTPs
- synthetic location-specific RNA primer pair
- heat-resistent DNA polymerase (Taq polymerase)

The (cyclic) Process

- **Denaturation**, i.e. separation, of double stranded DNA
  at high temperature (94-98 deg. C)
- **Primer annealing** of the RNA primer to the 3'-end of the template DNA
  at low temperature (50-65 deg. C)
- **Primer extension** and insertion of complementary bases by DNA polymerase (until the end of the template is reached)
  at medium temperature (75-80 deg. C)

This concept is used for sequence amplification and known as Polymerase Chain Reaction (PCR).

Polymerase Chain Reaction (PCR)



PCR was invented 1983 by Kary Mullis who was awarded the Nobel Prize in Chemistry in 1993.
Sequencing methods based on **sequencing by synthesis** are still in use.

Polymerase Chain Reaction (PCR)

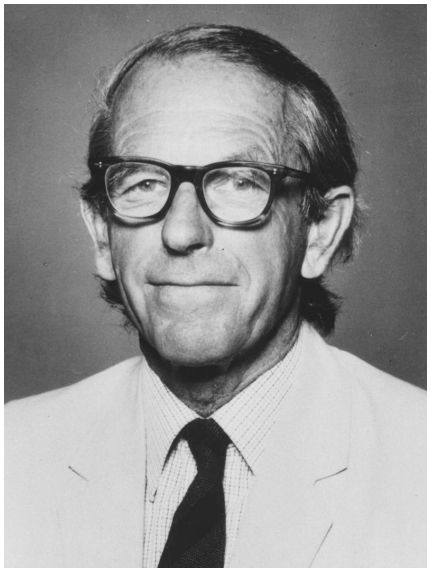
Videos for a better understanding

Polymerase Chain Reaction (PCR):
`https://www.youtube.com/watch?v=2KoLnIwoZKU`

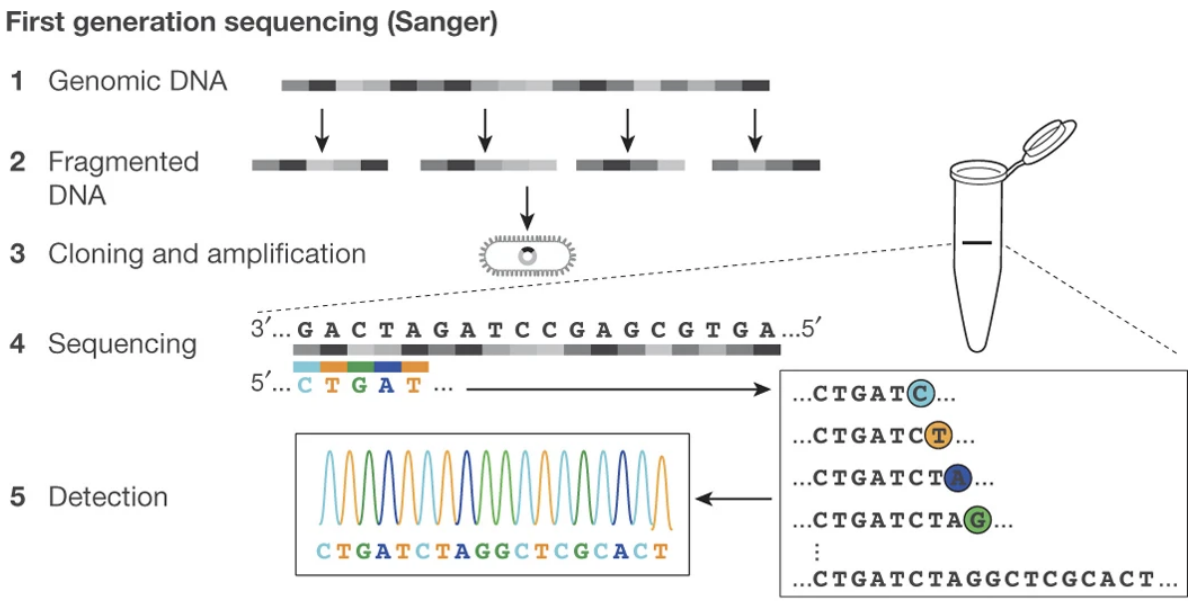
A bit longer but with nice cartoons and Coronavirus test as example application:
`https://www.youtube.com/watch?v=a5jmdh9AnS4`

The FIRST Generation of High-Throughput Sequencing

The concept behind PCR is the basis of sequencing.

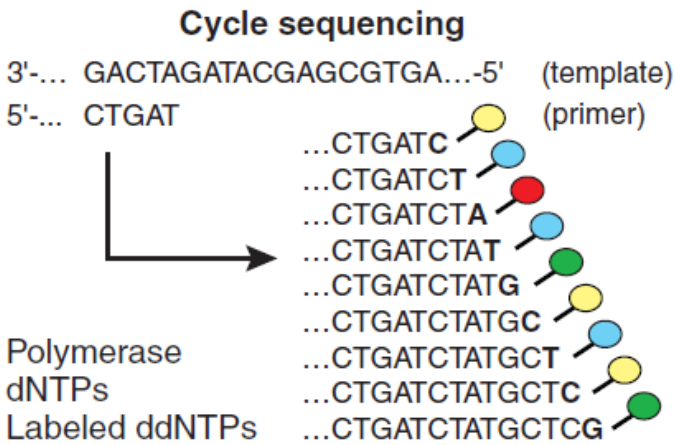


Frederick Sanger

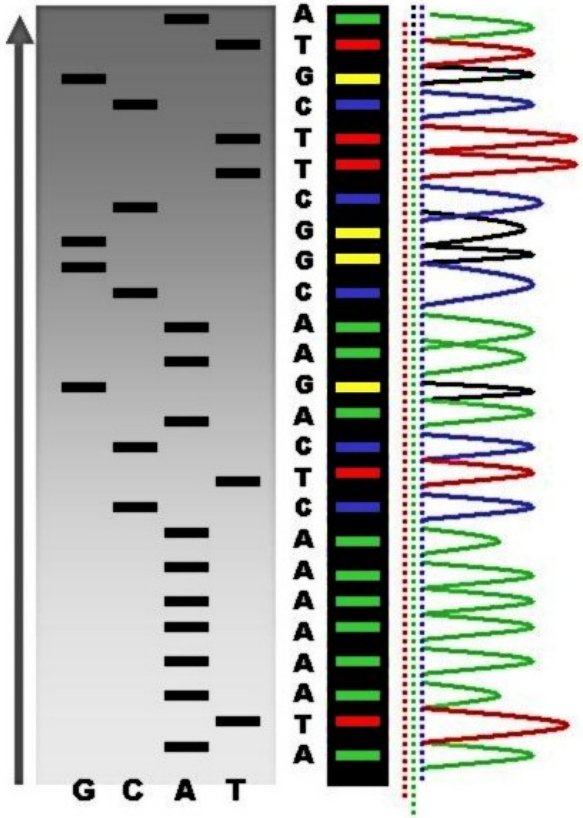


PCR + labeled nucleotides $\approx$ sequencing.

Sanger Sequencing: a Chain Termination Method



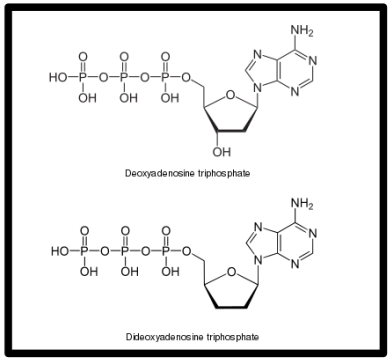
- amplified target sequence
- a pool of dNTPs
- some chain terminating, **labeled** **d**dNTPs: **d**dATPs (green), **d**dCTPs (blue), **d**dGTPs (yellow), **d**dTTPs (red)
- chain terminations at **d**dNTPs but random positions

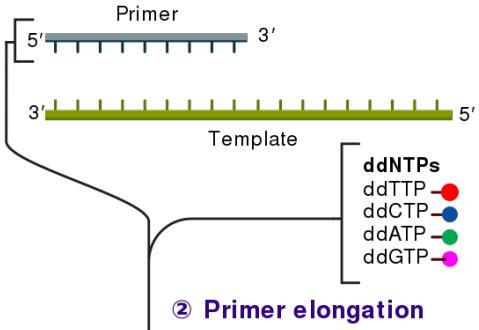


- separation by length
  e.g. gel-electrophoresis

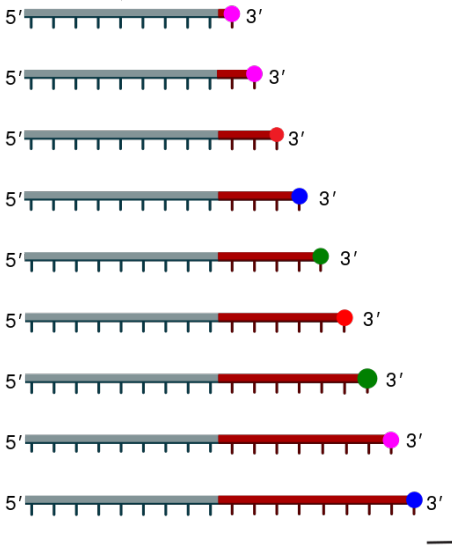
① **Reaction mixture**

- Primer and DNA template
- DNA polymerase
- ddNTPs with flourochromes
- dNTPs (dATP, dCTP, dGTP, and dTTP)

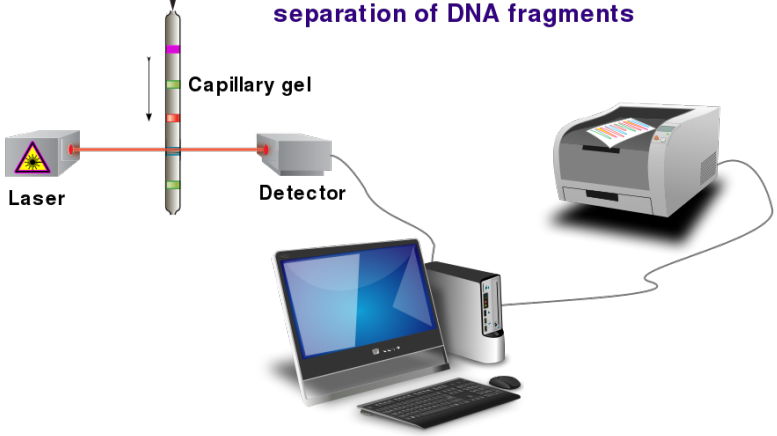


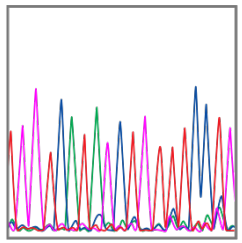


② **Primer elongation and chain termination**



③ **Capillary gel electrophoresis separation of DNA fragments**





Chromatograph

④ **Laser detection of flourochromes and computational sequence analysis**

Sanger Sequencing: a Chain Termination Method

Problems and Limitation

- **separation by length**:
  discriminating longer sequences on nucleotide resolution
  - scales very bad with sequence length
  - the number of errors increases at the end of the sequence
  - typical errors: order confunsion, substitution, deletion

Sequencing length is limited!

Sanger Sequencing: a Chain Termination Method

Actually, Sanger sequencing is easier to understand from a video. So feel free to take a look at one or all of them.

A very short one:
`https://www.youtube.com/watch?v=6ldtdWjDwes`

A more detailed video: `https://dnalc.cshl.edu/resources/animations/cycseq-video.html`

Alternative explanation from Promega Corp.:
`https://www.youtube.com/watch?v=dVRB4CaLizc`

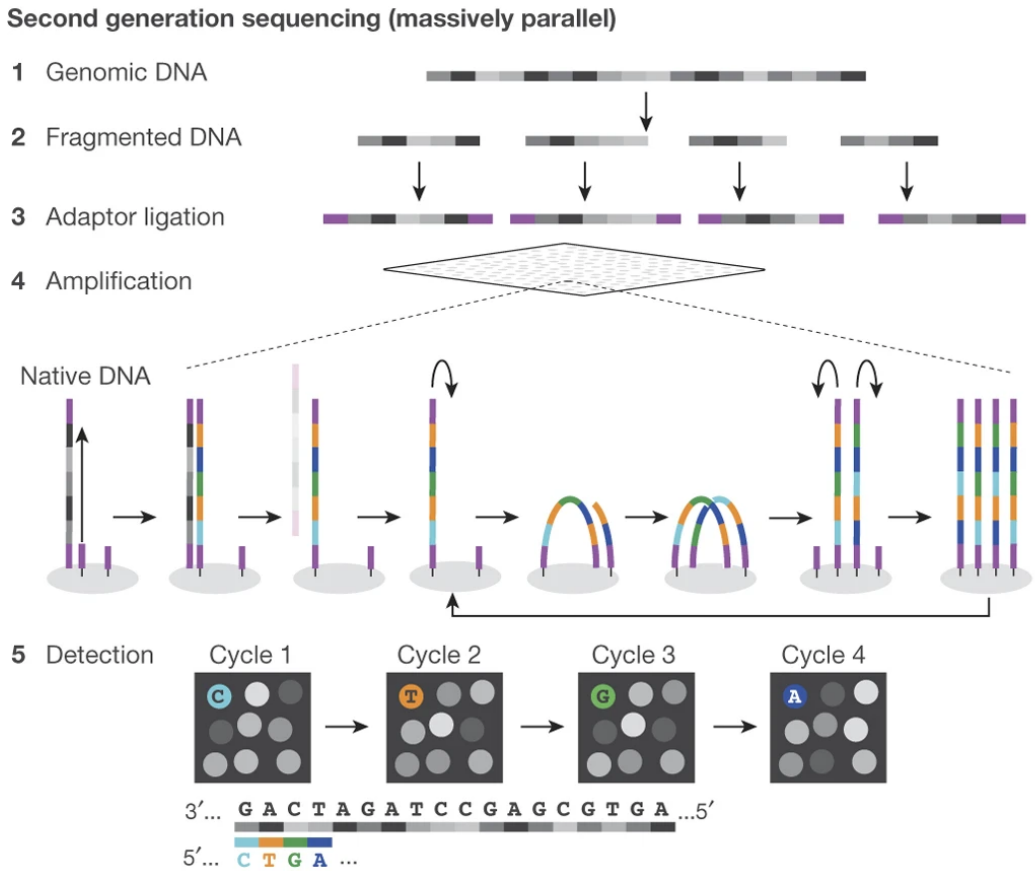
Alternative explanation from Thermo Fisher Scientific:
`https://www.youtube.com/watch?v=e2G5zx-OJIw`

The SECOND Generation of High-Throughput Sequencing

also called **NEXT Generation Sequencing**

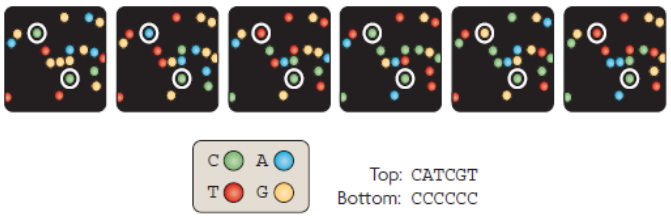


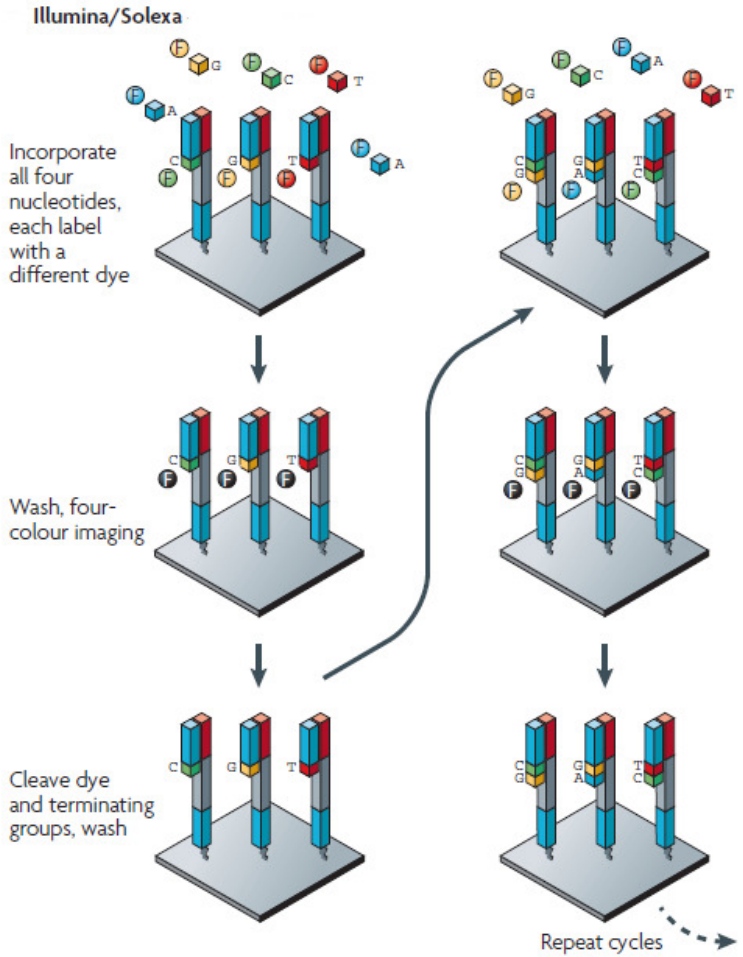
Larry Bock
(Illumina)



Illumina sequencing – the Main Concept

- fix DNA fragments to a surface
- supply only chain terminating, **labeled** dNTPs
- cycle through
  - let addition of exactly one nucleotide occur
  - take image
  - remove terminator moiety (and label)





The Trick: Using Modified Nucleotides

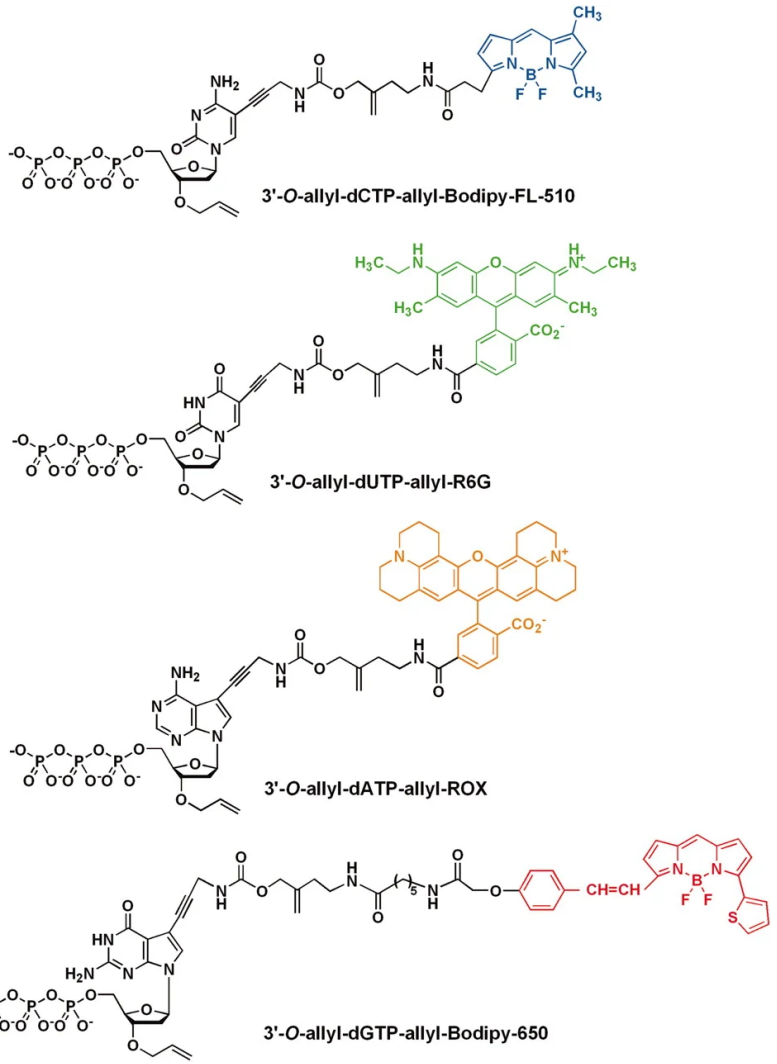
Illumina nucleotides

Reversible terminator:
capping the 3'-OH
with a reversible moiety
here, an allyl group
(-CH_2-CH=CH_2)

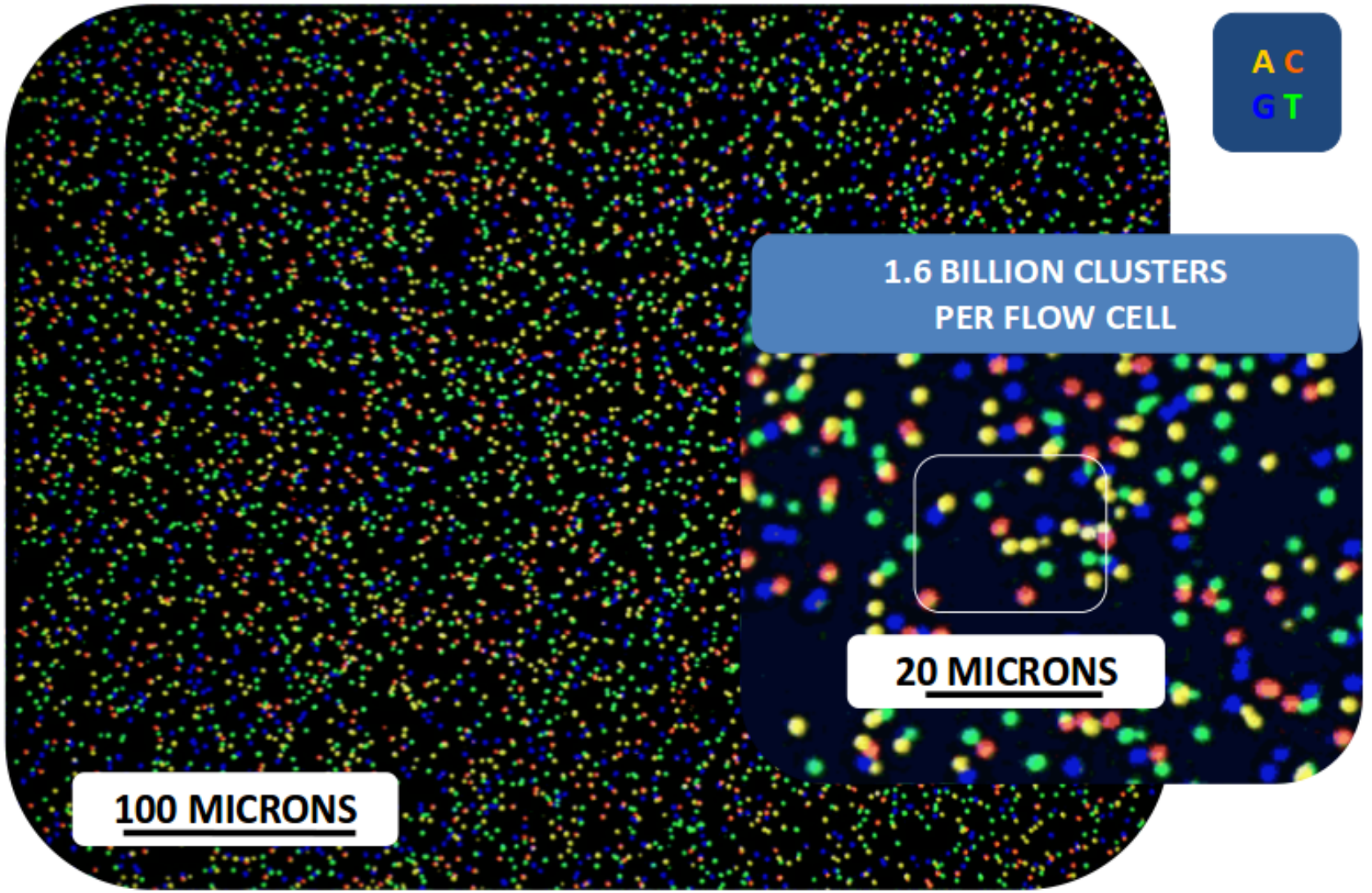
Cleavable fluorophores:
let nucleotides light up

Wavelength – absorbtion, emission

- dCTP – 502, 510nm (blue)
- dUTP – 525, 550nm (green)
- dATP – 585, 602nm (yellow)
- dGTP – 630, 650nm (red)



Upon activation (by light) the fluorophors emit light of different wavelength (color).

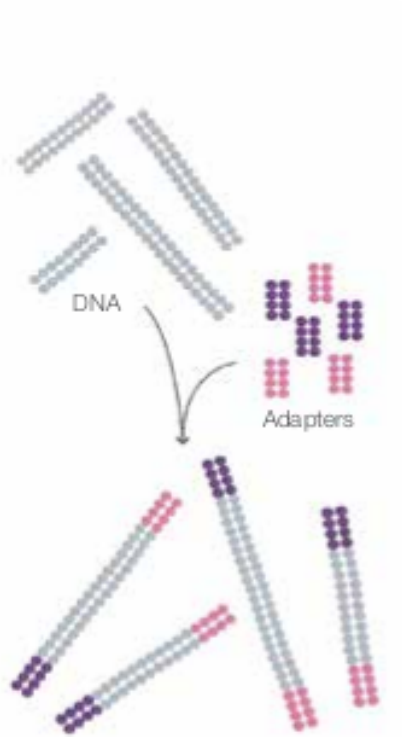

A C
G T
1.6 BILLION CLUSTERS
PER FLOW CELL
20 MICRONS
100 MICRONS

Illumina sequencing

The video gives a very good overview about Illumina sequencing. The next five slides recapture the main steps of an Illumina sequencing. If you had problems to follow them in the video please go through them again and try to understand what is happening in each step.

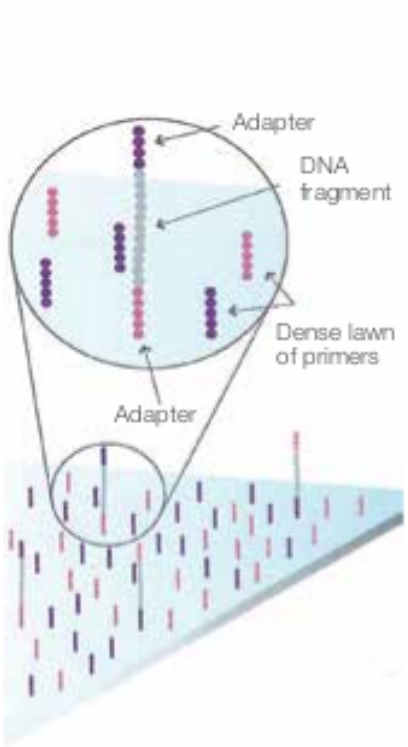
`https://www.youtube.com/watch?v=fCd6B5HRaZ8`

Figure 2: Prepare Genomic DNA Sample



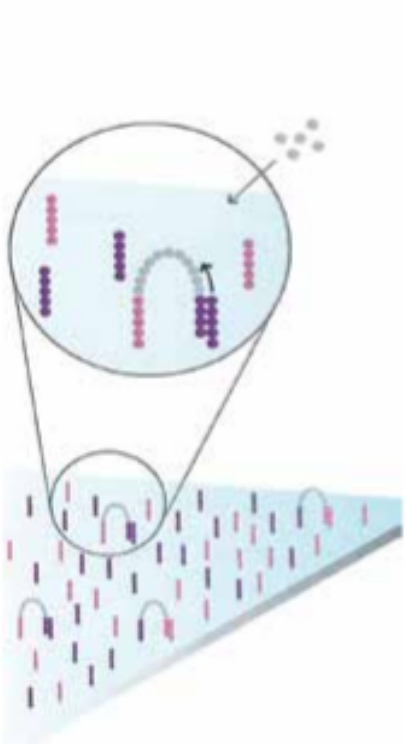
Randomly fragment genomic DNA and ligate adapters to both ends of the fragments.

Figure 3: Attach DNA to Surface



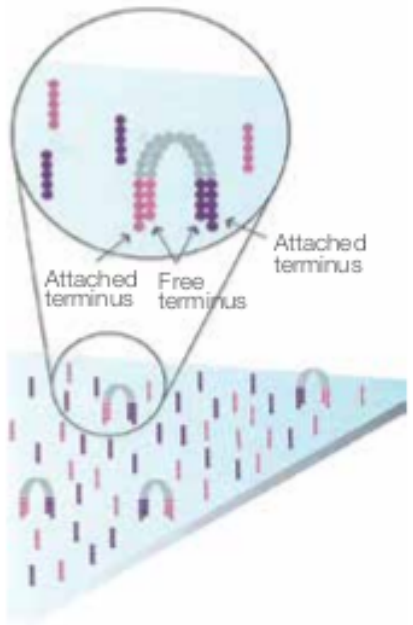
Bind single-stranded fragments randomly to the inside surface of the flow cell channels.

Figure 4: Bridge Amplification



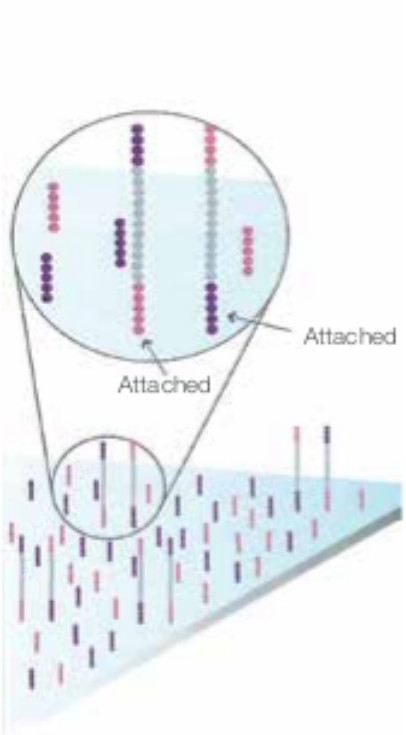
Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification.

Figure 5: Fragments Become Double Stranded



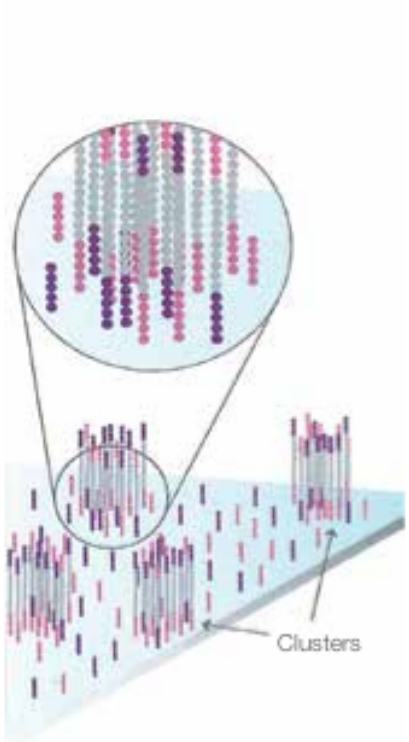
The enzyme incorporates nucleotides to build double-stranded bridges on the solid-phase substrate.

Figure 6: Denature the Double-Standed Molecules



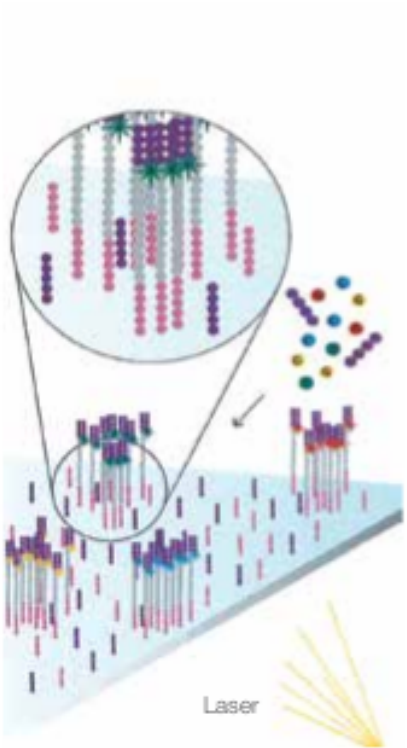
Denaturation leaves single-stranded templates anchored to the substrate.

Figure 7: Complete Amplification



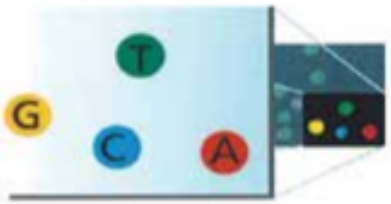
Several million dense clusters of double-stranded DNA are generated in each channel of the flow cell.

Figure 8: Determine First Base



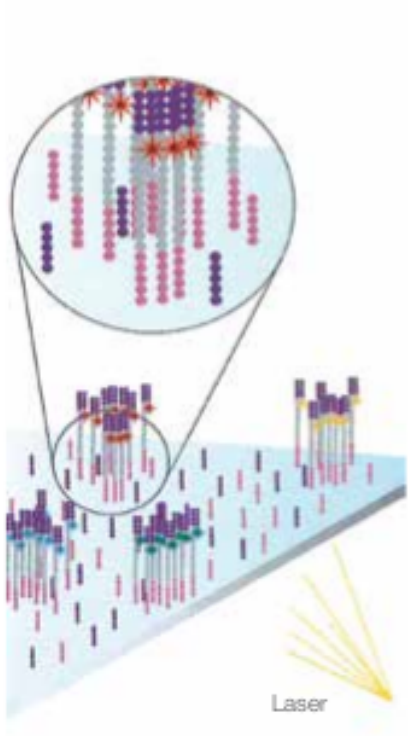
The first sequencing cycle begins by adding four labeled reversible terminators, primers, and DNA polymerase.

Figure 9: Image First Base



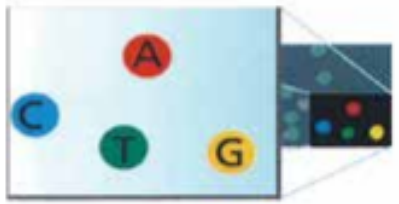
After laser excitation, the emitted fluorescence from each cluster is captured and the first base is identified.

Figure 10: Determine Second Base



The next cycle repeats the incorporation of four labeled reversible terminators, primers, and DNA polymerase.

Figure 11: Image Second Chemistry Cycle



After laser excitation, the image is captured as before, and the identity of the second base is recorded.

Illumina sequencing

Problems and Limitations

- **synchrony of sequences in a cluster**:
  incomplete removal of fluorophor and/or chain terminator can causes asynchrony in the cluster and misinterpretation of mixed light signal
  - scales bad with sequence length
  - the number of errors increases at the end of the sequence
  - typical error: substitution

Sequencing length is limited!

Next Generation High-Throughput Sequencing (HTS)

- HTS made (whole genome) sequencing and resequencing available to many more researchers and projects
- cost and time have been greatly reduced
- error profiles and limitations of new platforms differ significantly from previous technologies
- selection of appropriate sequencing platform for particular experiment is important
- requires detailed understanding of technologies available; that includes sources of error, error rate, throughput and cost

Further reading

In addition to these slides, please read the following paper:

(The pdf is in moodle)

"DNA sequencing at 40: past, present and future"
`http://www.nature.com/doifinder/10.1038/nature24286`

In the next lecture we will look at some aspects of Illumina again and discuss third generation sequencing methods, as well.

Shendure, Jay, et al. "DNA sequencing at 40: past, present and future." Nature 550.7676 (2017): 345-353.