The Abdus Salam
International Centre for Theoretical Physics

United Nations
Educational, Scientific and
Cultural Organization

IAEA
International Atomic Energy Agency

2255-1

2nd Conference on Systems Biology and New Sequencing Techniques" (2-4 November), preceded by Introductory Lectures on "Quantitative Approaches to Biological Problems" (31 October - 1 November)

31 October - 4 November, 2011

DNA Sequencing Methods Past, Present, and Future

Arjang Hassibi
University of Texas at Austin
USA

Strada Costiera 11, 34151 Trieste, Italy - Tel.+39 040 2240 111; Fax +39 040 224 163 - sci_info@ictp.it

DNA Sequencing Methods

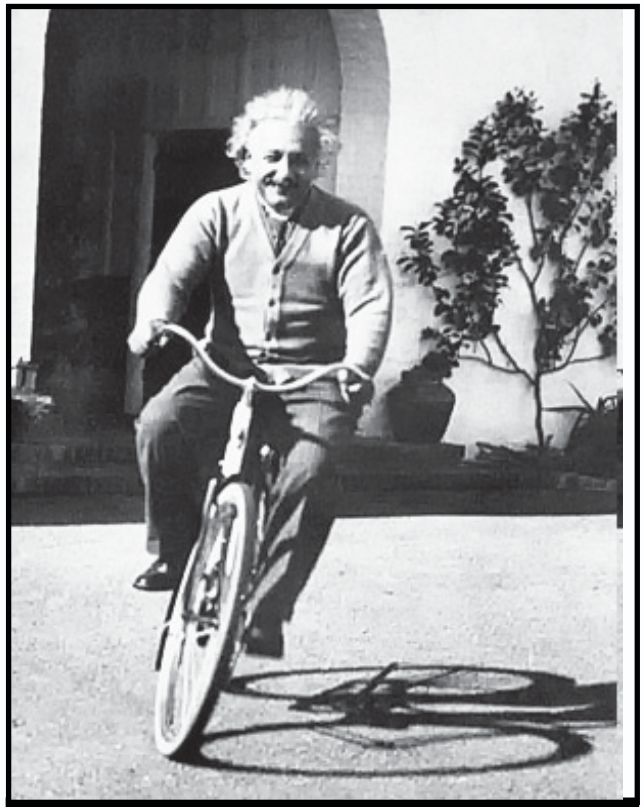
Past, Present, and Future

Arjang Hassibi
Electrical and Computer Engineering Department
Institute for Cellular and Molecular Biology
University of Texas at Austin, TX

November 1st, 2011

The Abdus Salam International Centre for Theoretical Physics

Why Sequence DNA?



Albert Einstein (1879-1955)
Greatest Achievement: Theory of Relativity

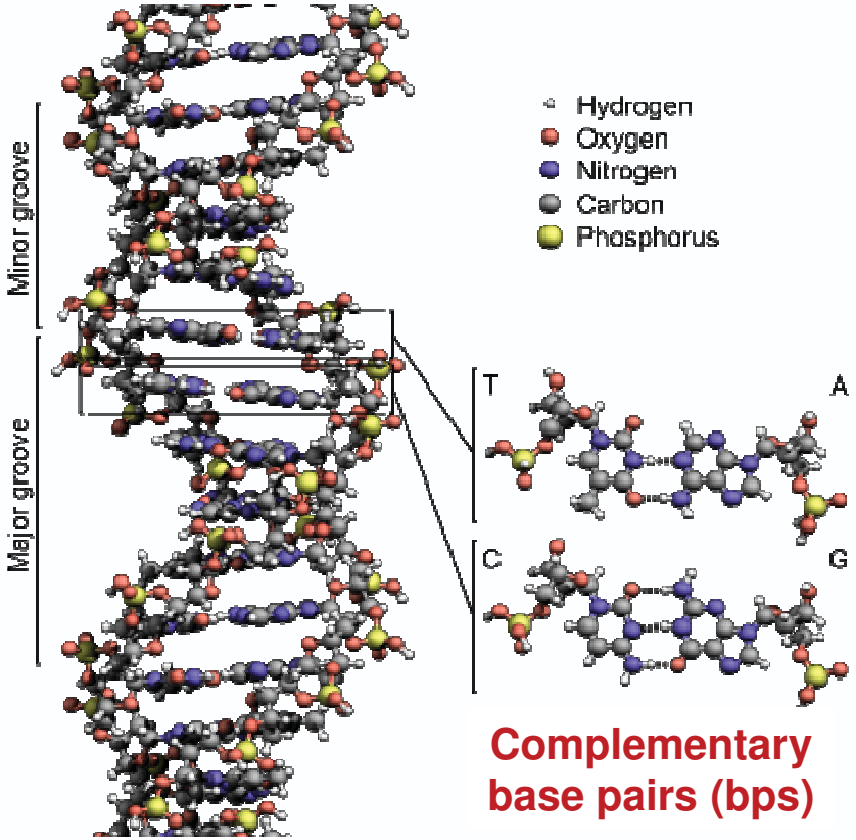
Bobo the Chimp (1995-Now)
Greatest Achievement: Shown Above

1.5% DNA Difference

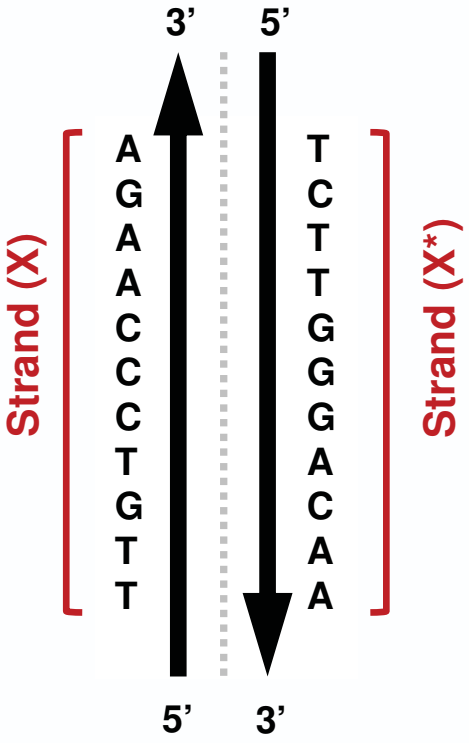
Small details in the DNA can make a huge difference!

Problem Definition

Identify the structure of a long polymer molecule that consists of only four building blocks (A, C, G, and T)

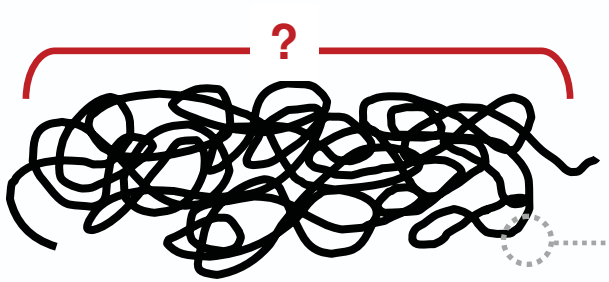
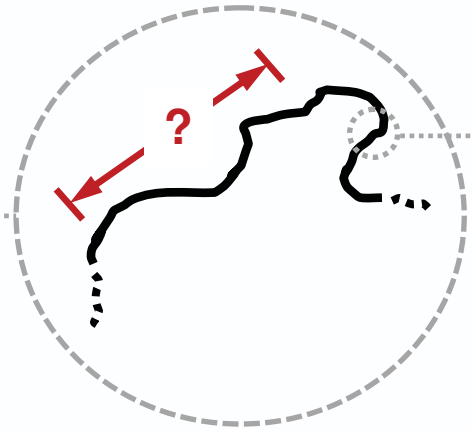


Complementary base pairs (bps)



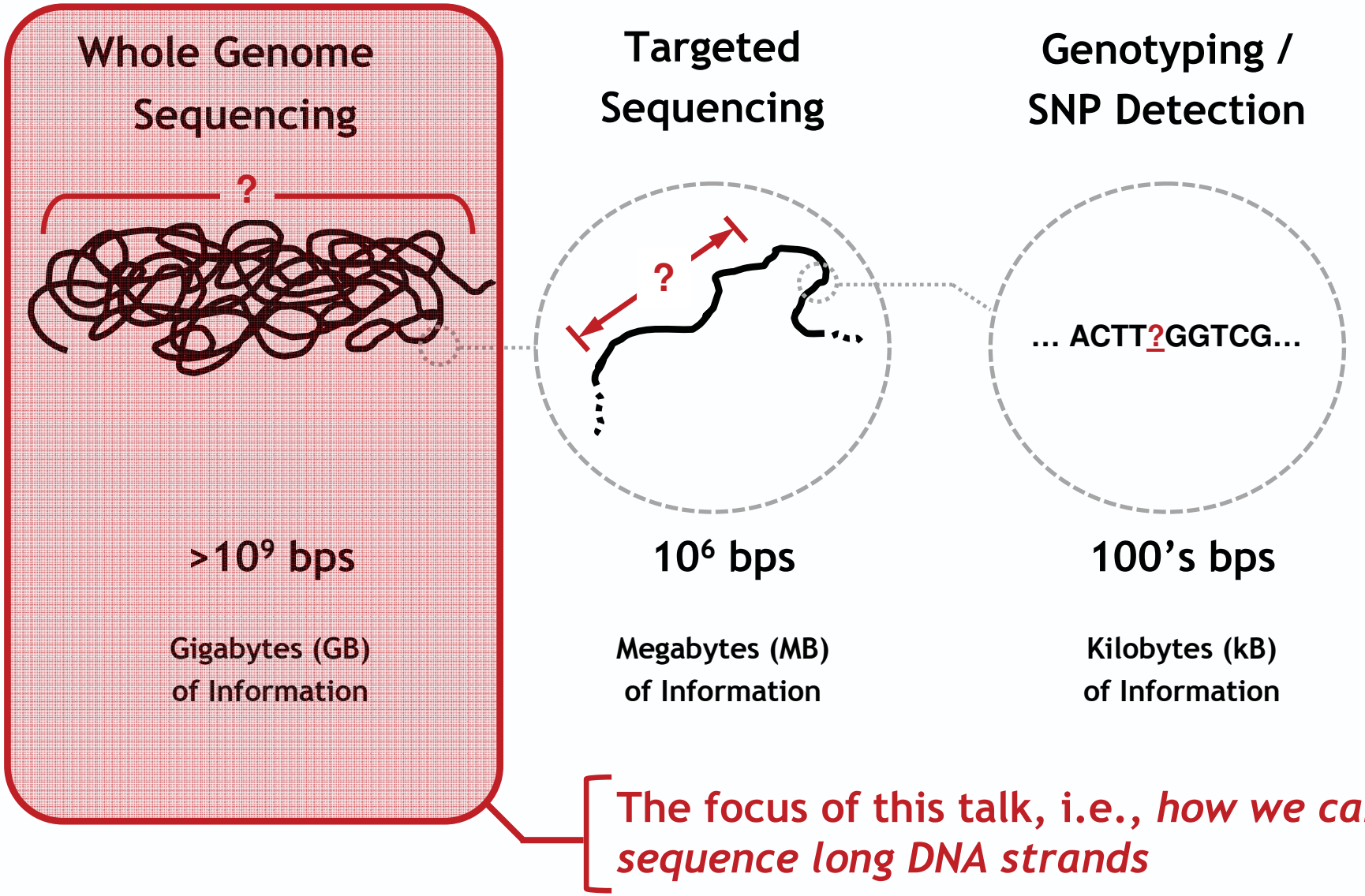
Strand (X) and (X*) are complementary

Scope of the Problem

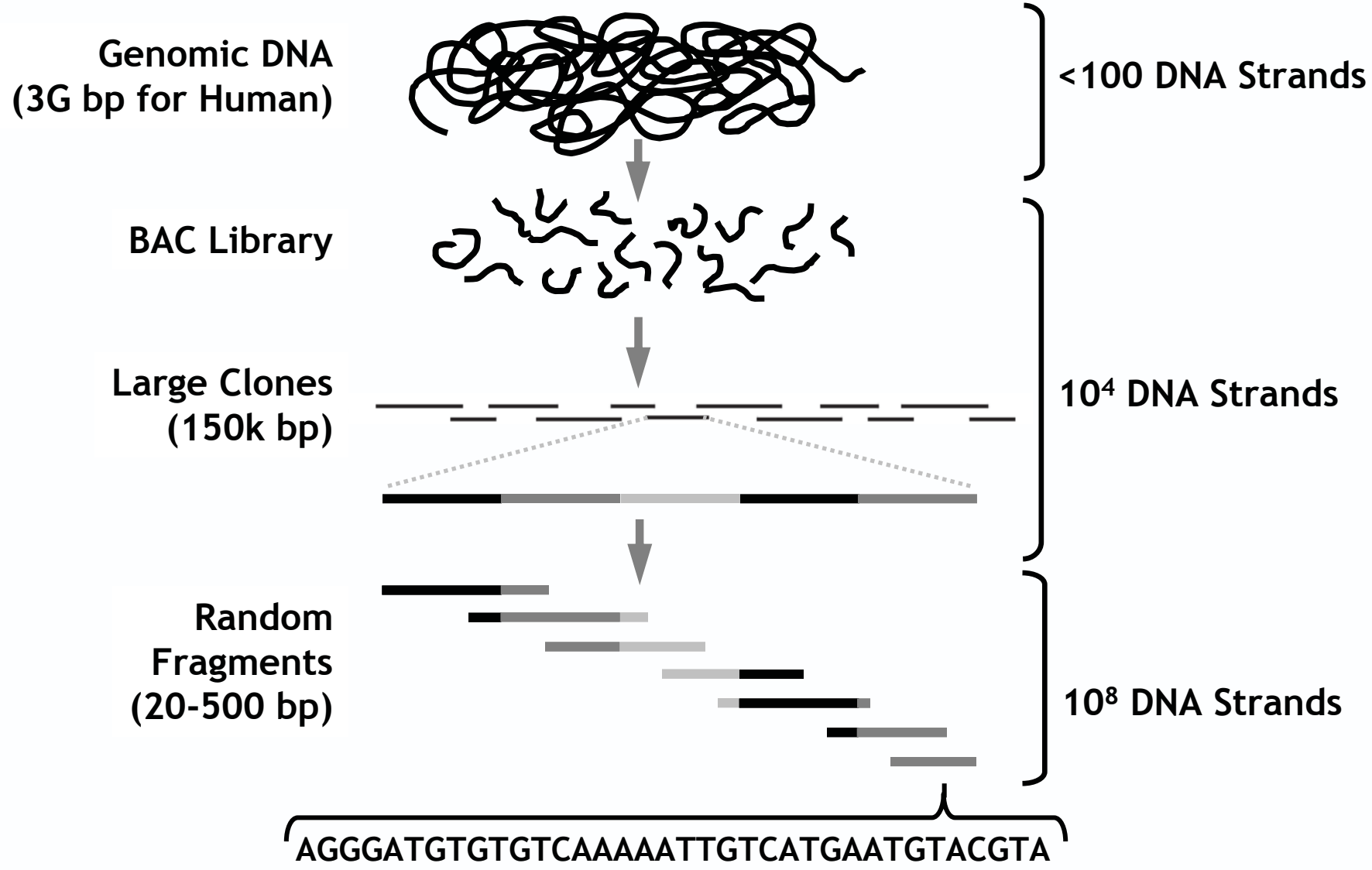
Whole Genome Sequencing	Targeted Sequencing	Genotyping / SNP Detection
		
>10^9 bps	10^6 bps	100's bps
Gigabytes (GB) of Information	Megabytes (MB) of Information	Kilobytes (kB) of Information

Research / Discovery → Diagnostics

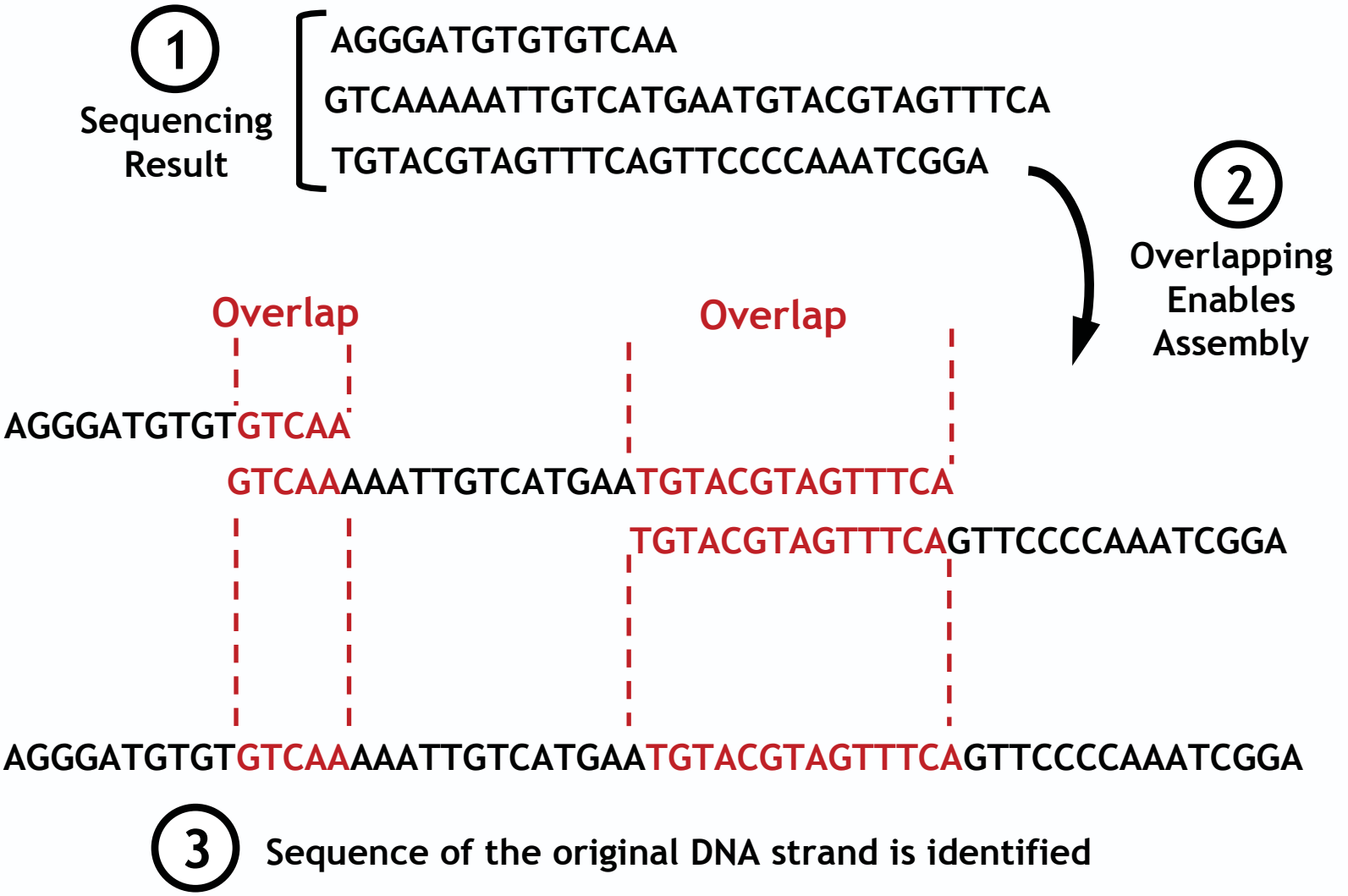
Scope of the Problem



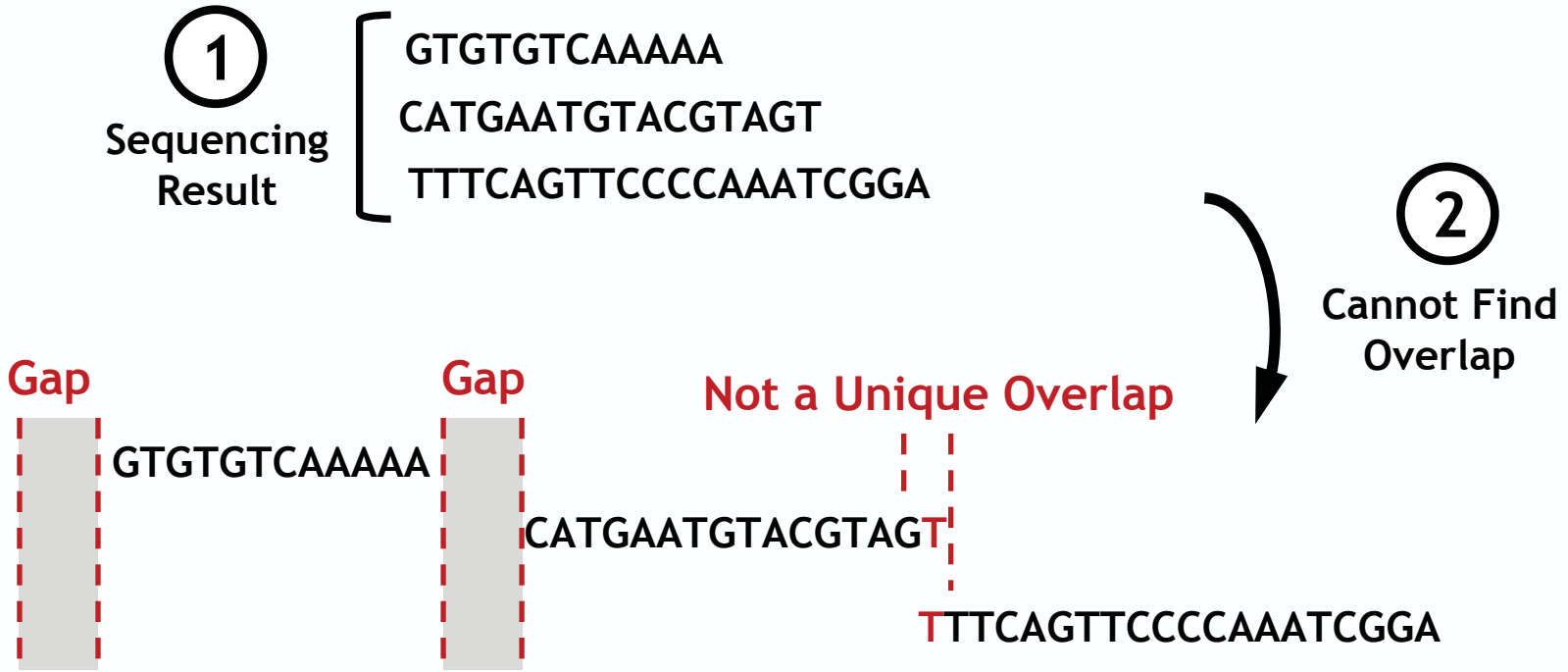
General Approach: Divide and Conquer



Redundancy vs. Accuracy (1)



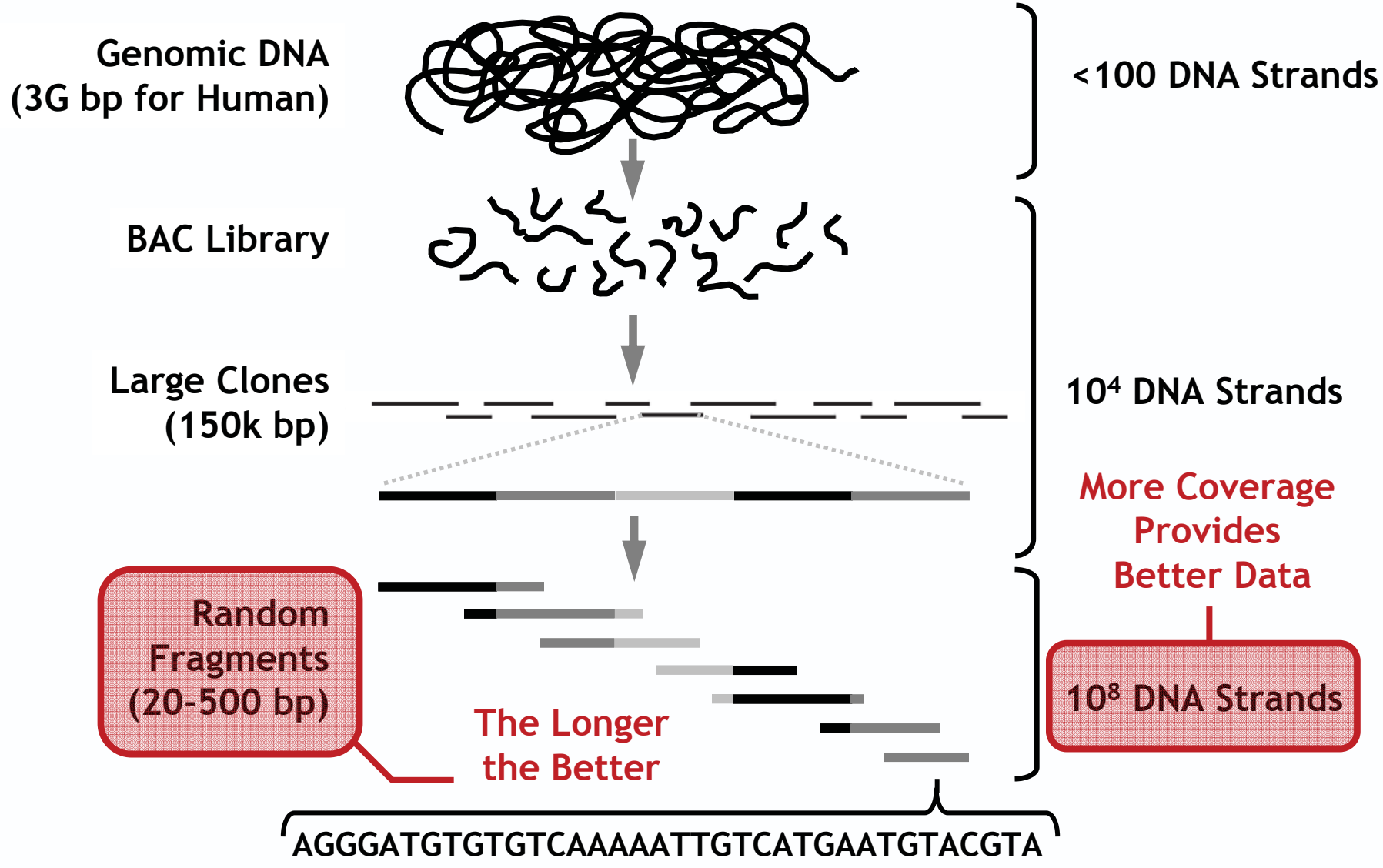
Redundancy vs. Accuracy (2)



AGGGATGTGTGTCAAAAATTGTCATGAATGTACGTAGTTTCAGTTCCCCAAATCGGA

FAILED! Not enough information to identify the sequence

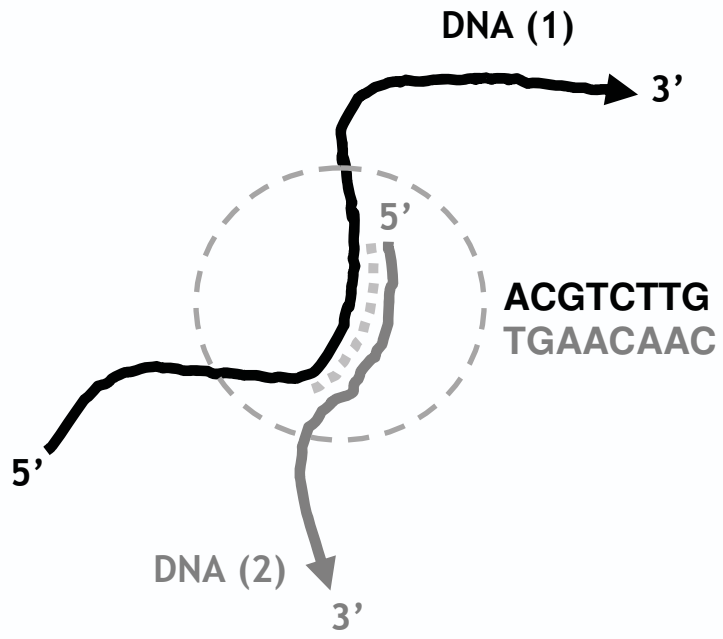
General Approach: Divide and Conquer



Our DNA Toolbox (1)

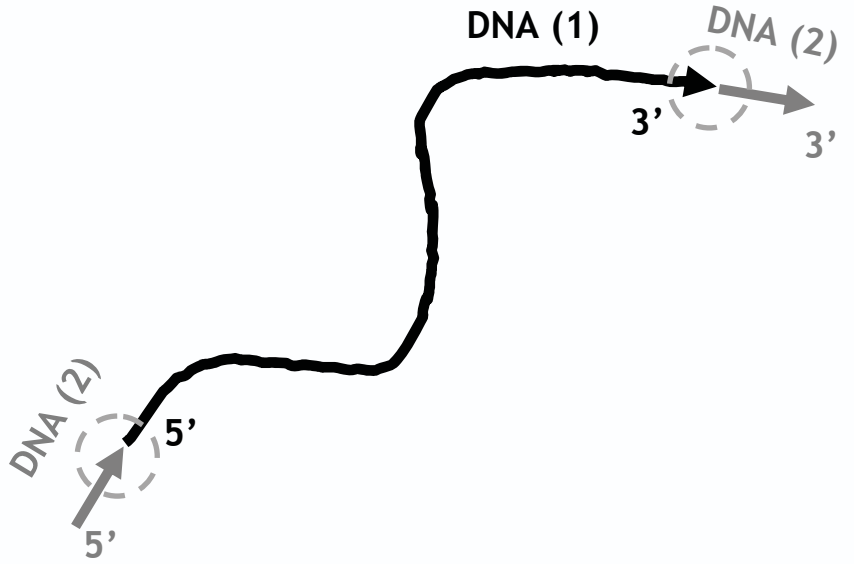
"Sticking" DNA fragments to one another

DNA Hybridization



Sequence-specific
(Energy ~0.2kT/bp at 300°K)

Covalent Attachment

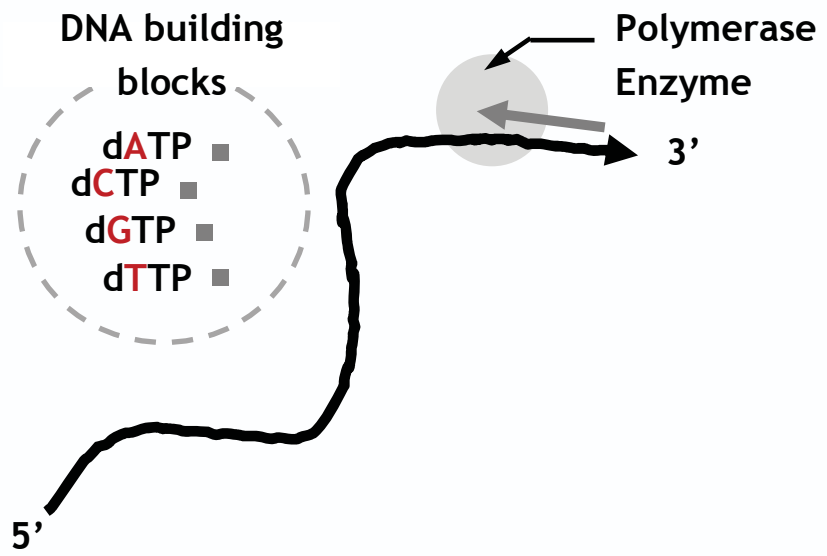


Non-specific and requires activation chemistries

Our DNA Toolbox (2)

Replicating/Copying DNA

DNA Polymerization



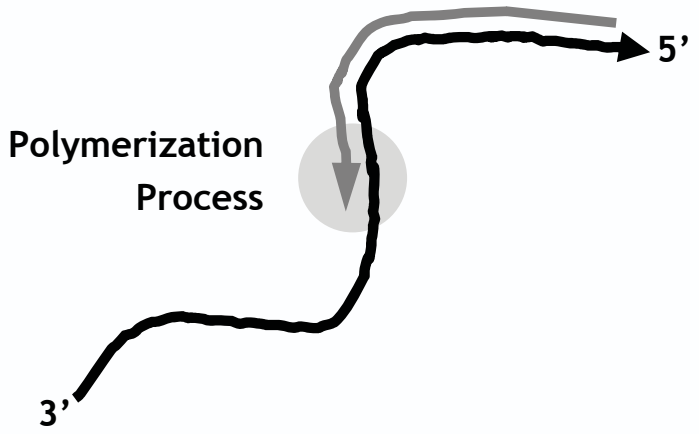
DNA can copy itself

Polymerase and nucleotides
(A, C, G, and T) should be available)

Our DNA Toolbox (2)

Replicating/Copying DNA

DNA Polymerization



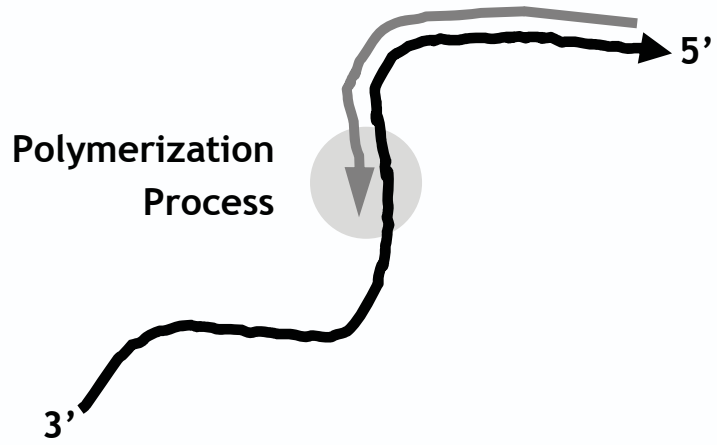
DNA can copy itself

Polymerase and nucleotides
(A, C, G, and T) should be available)

Our DNA Toolbox (2)

Replicating/Copying DNA

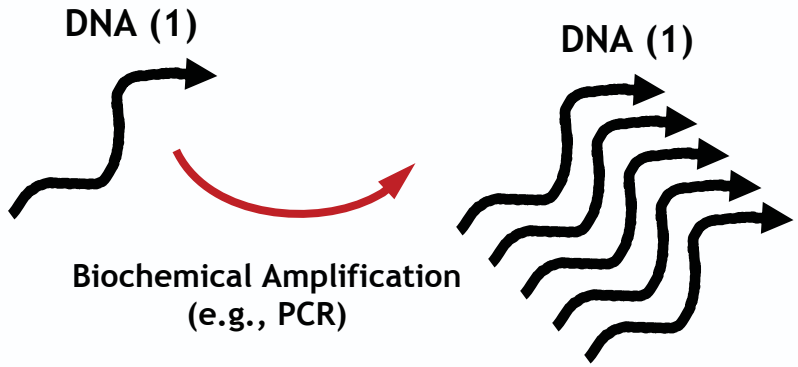
DNA Polymerization



DNA can copy itself

Polymerase and nucleotides
(A, C, G, and T) should be available)

DNA Amplification / Copying



Amplification gain can be >10^6

Sanger Sequencing (1977[1])

1958

1980

Frederick Sanger (1918-Present)

Nobel Prize in chemistry in 1958 (Structure of Proteins)

Nobel Prize in chemistry in 1980 (DNA Sequencing)

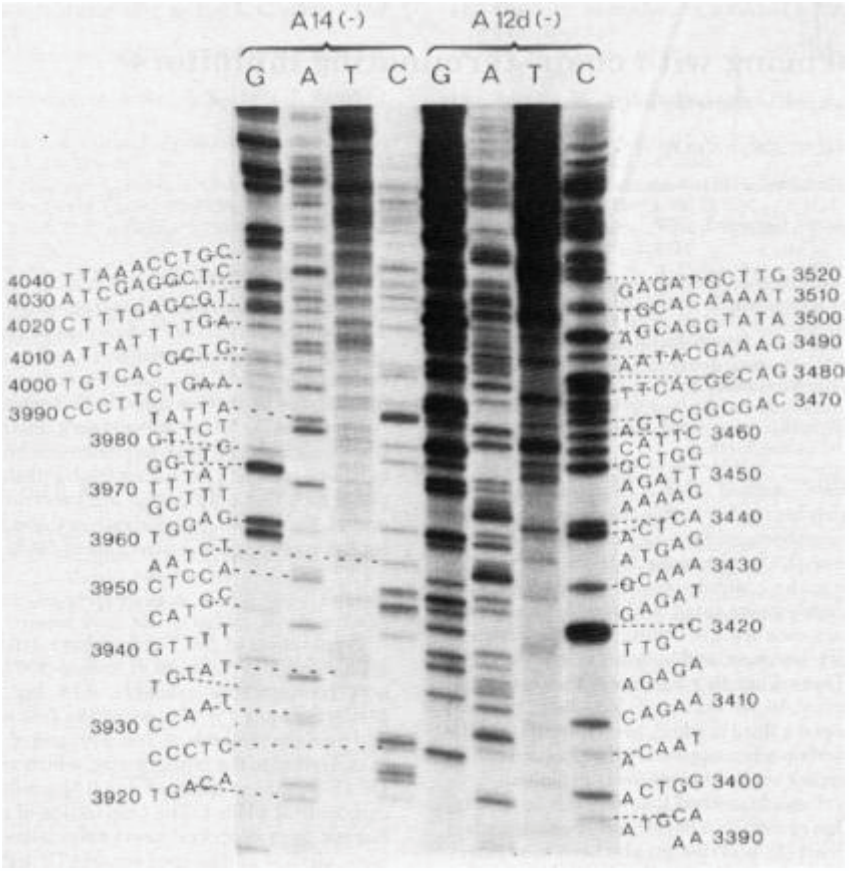
DNA sequencing with chain-terminating inhibitors

(DNA polymerase/nucleotide sequences/bacteriophage ϕX174)

F. SANGER, S. NICKLEN, AND A. R. COULSON

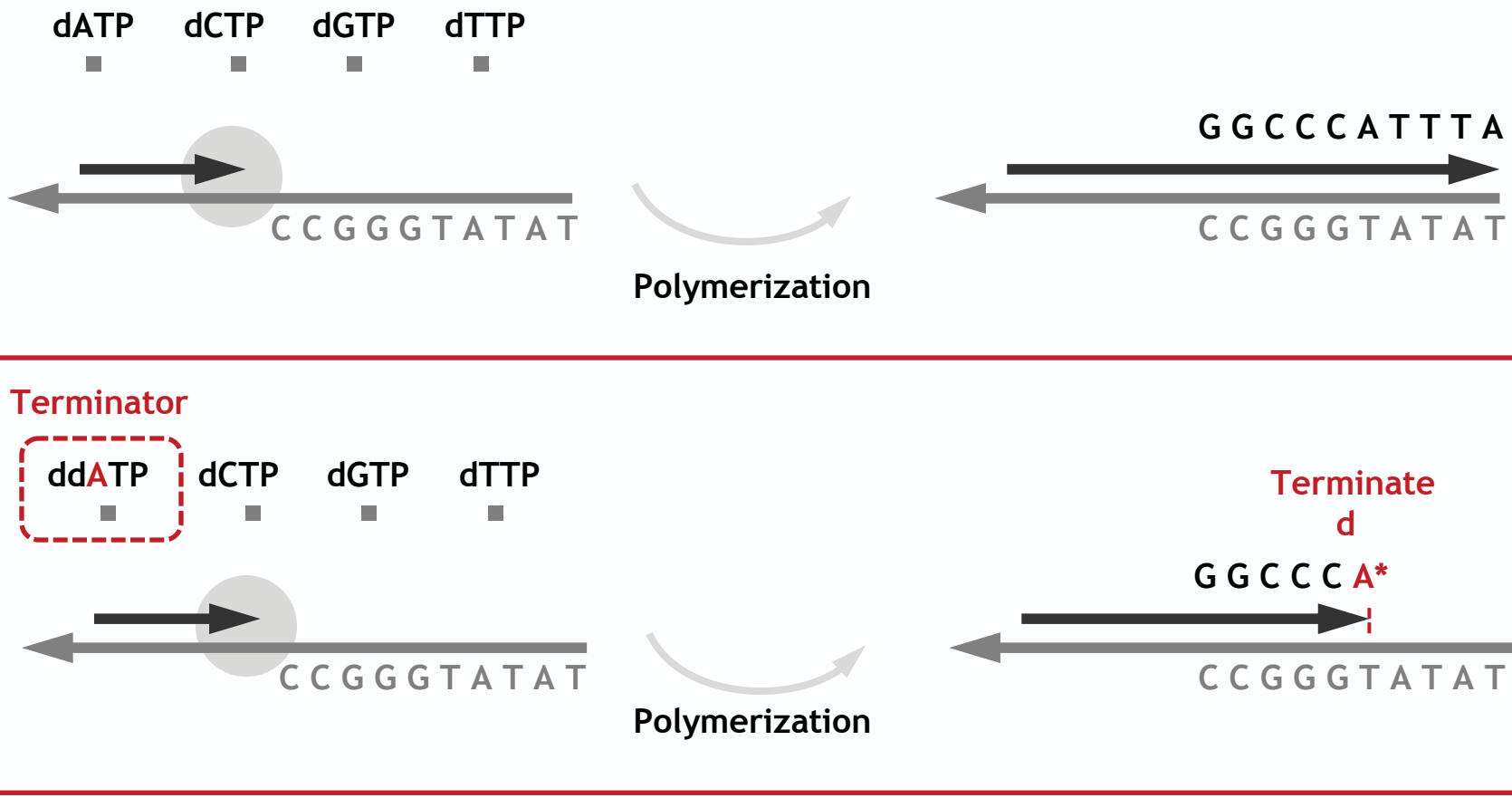
Medical Research Council Laboratory of Molecular Biology, Cambridge CB2 2QH, England

Contributed by F. Sanger, October 3, 1977



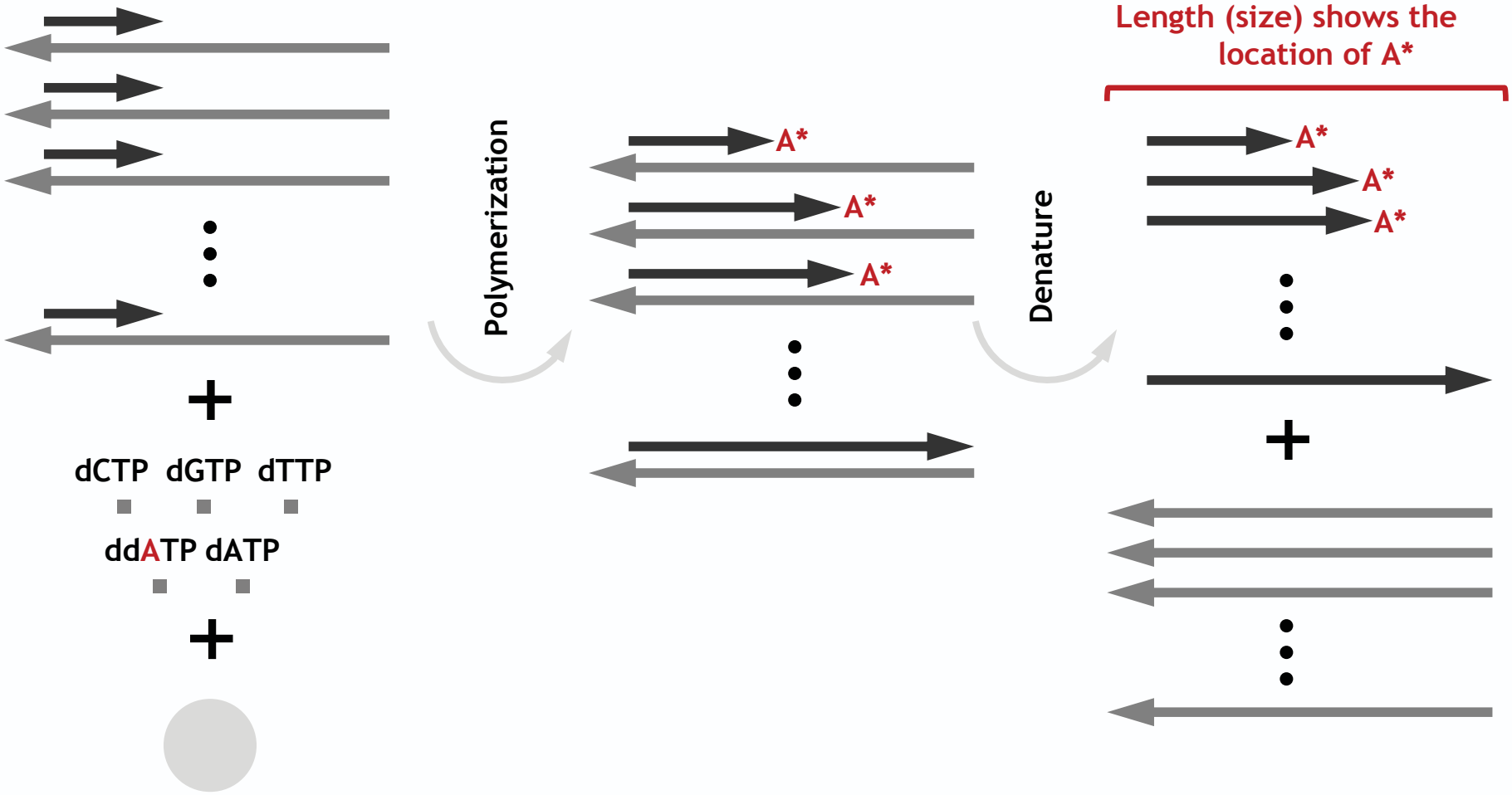
Chain Termination

By using dideoxynucleotides (ddNTPs) instead of deoxinucleotides (dNTPs), we can stop (terminate) polymerization

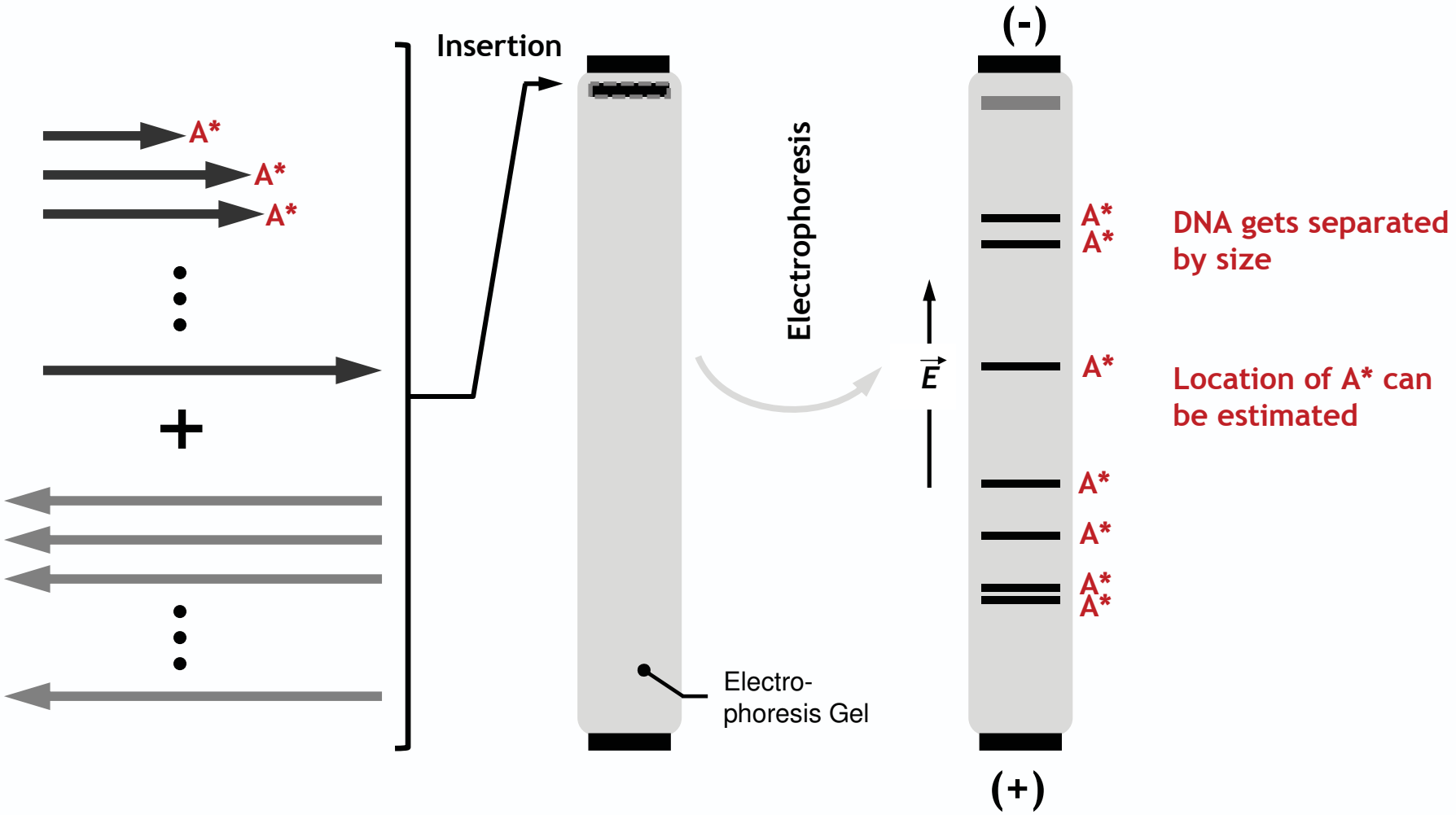


Sequence-Dependant Fragments

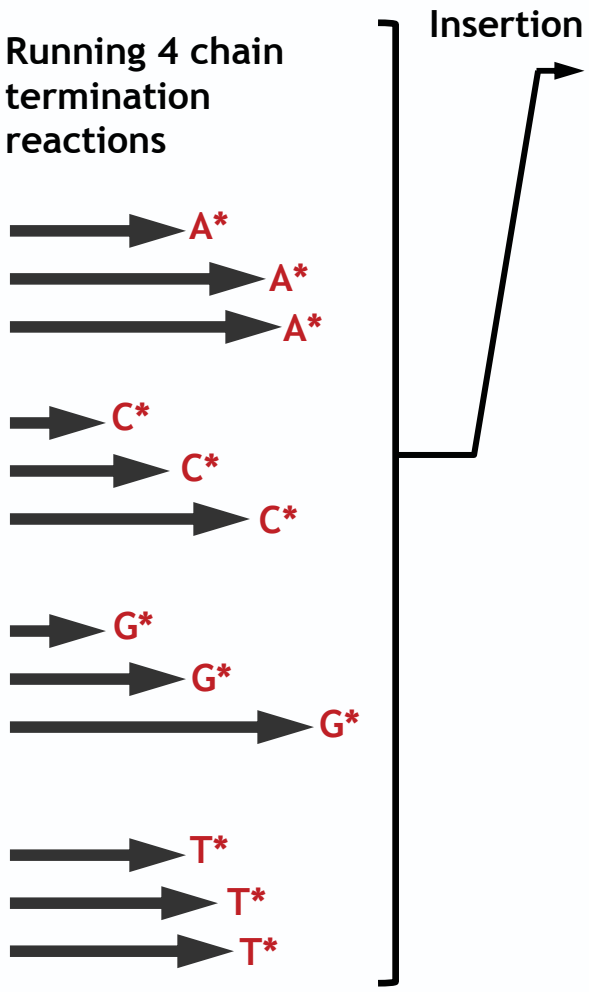
Polymerization with a mixture of dNTPs and ddTTPs

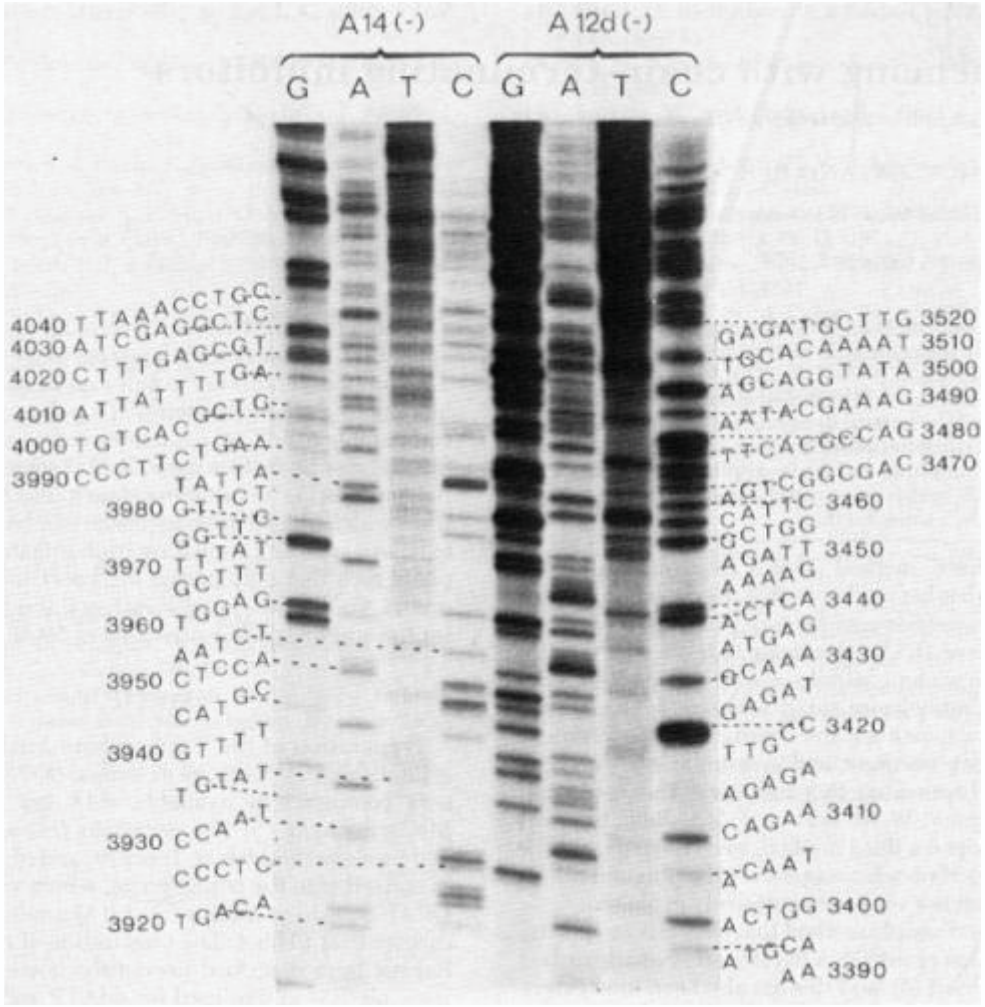


Finding the Length by Electrophoresis



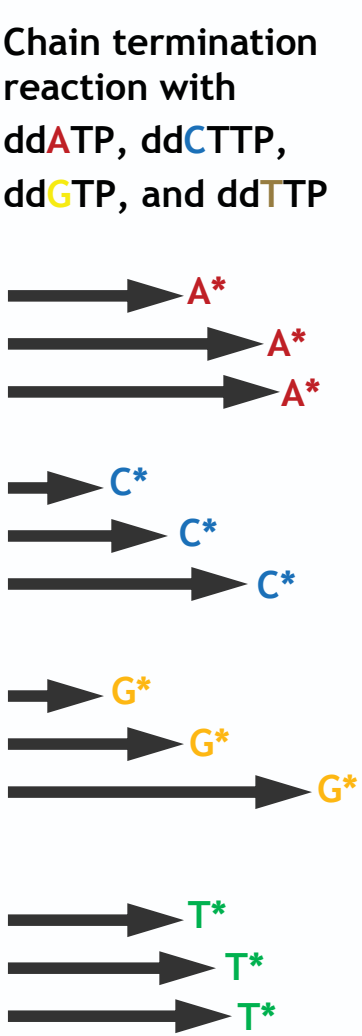
Imaging Using Radioactive Isotopes

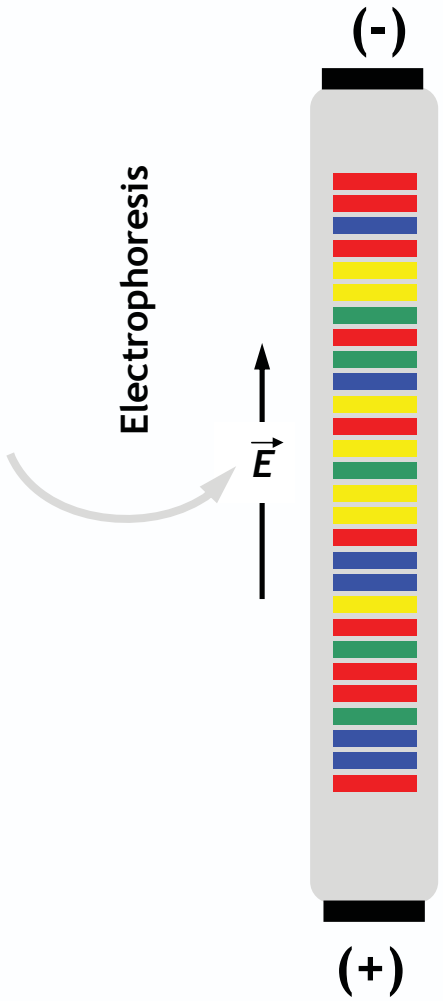




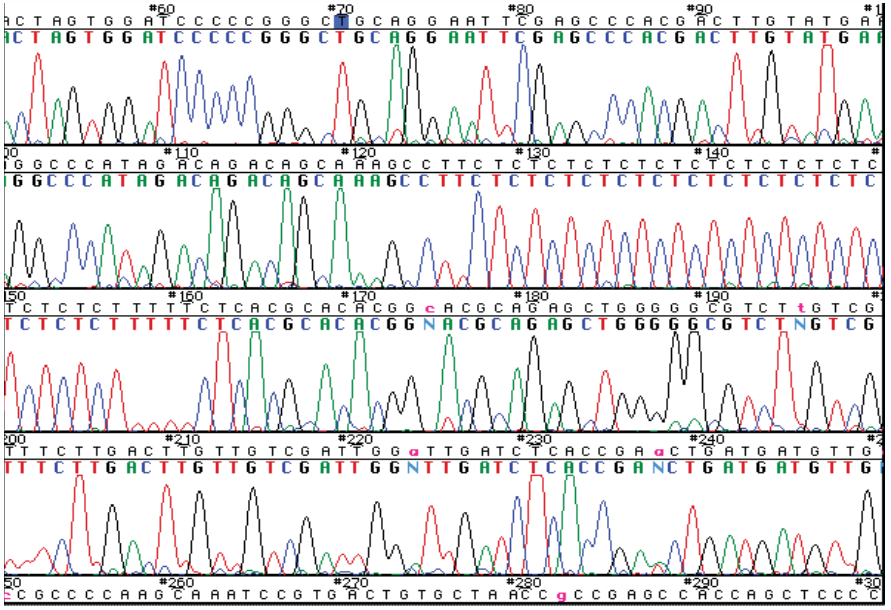
Sanger Using “Colored” Labels

Chain termination reaction with ddATP, ddCTTP, ddGTP, and ddTTP

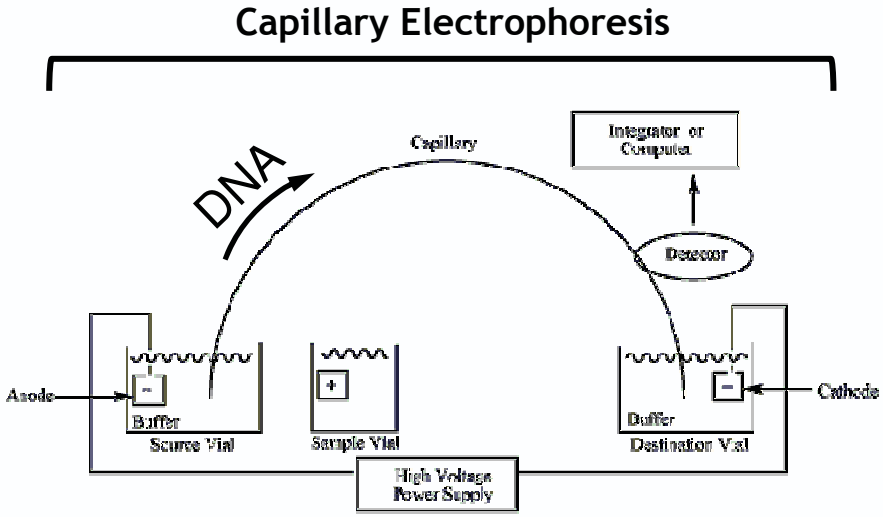




Typical Output Data of Sanger Sequencers



Sanger Sequencers in Human Genome Project (HGP[1])



Sanger DNA sequencer in 2010
ABI (Life Technologies) 3730

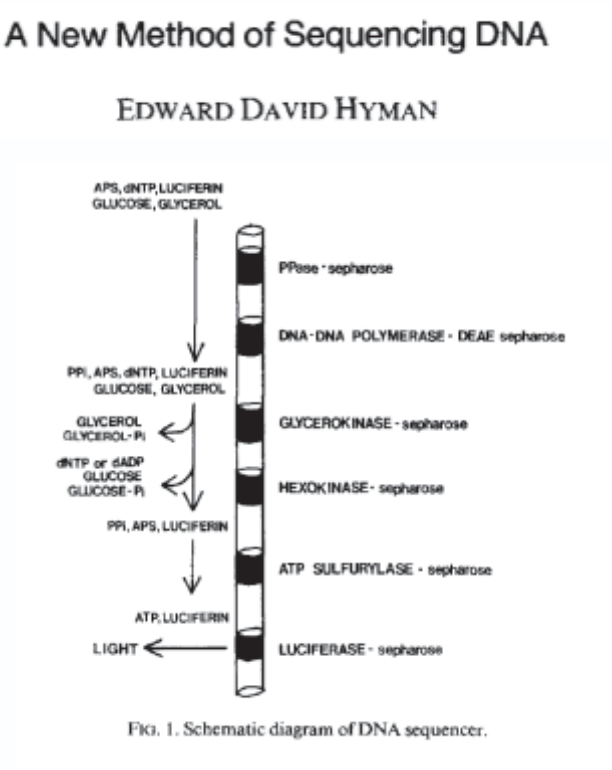
96 capillary electrophoresis
~1000 bps reads per channel
~100 kbps per day

It will take years to sequence a human genome!

[1] IHGSC "Finishing the euchromatic sequence of the human genome". *Nature* 431 (7011): 931–94, (2004).

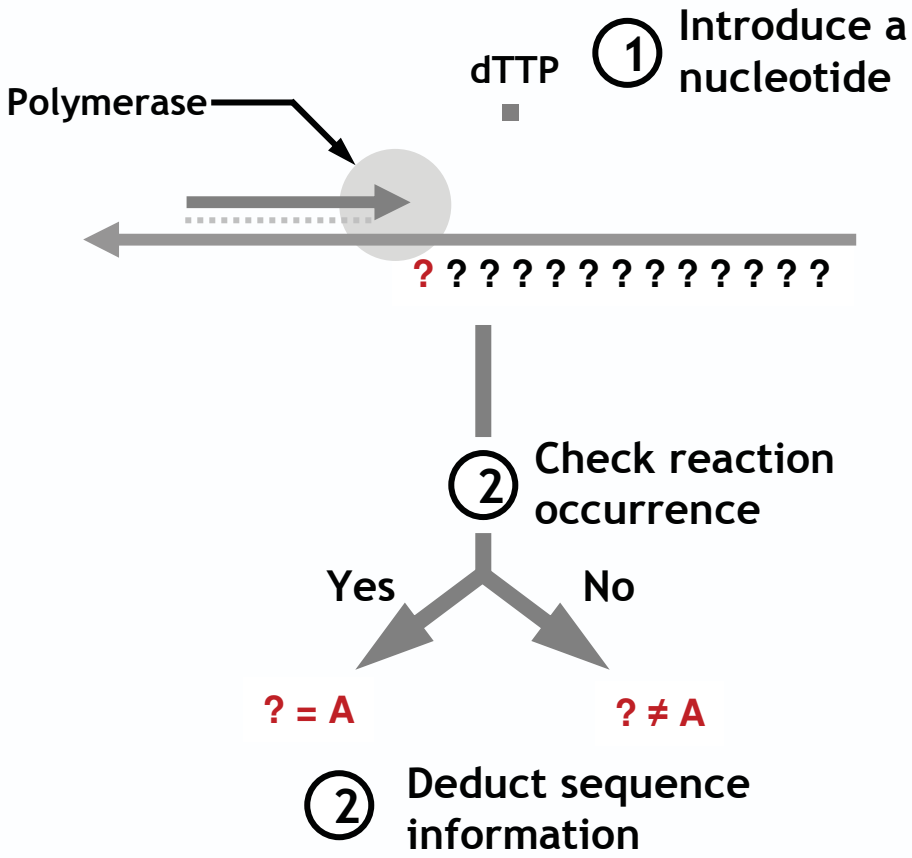
Sequence-by-Synthesis (SBS)[1]

Edward D. Hyman independently came up with this method in 1988. He made no money out of it and seldom gets any credit


A New Method of Sequencing DNA

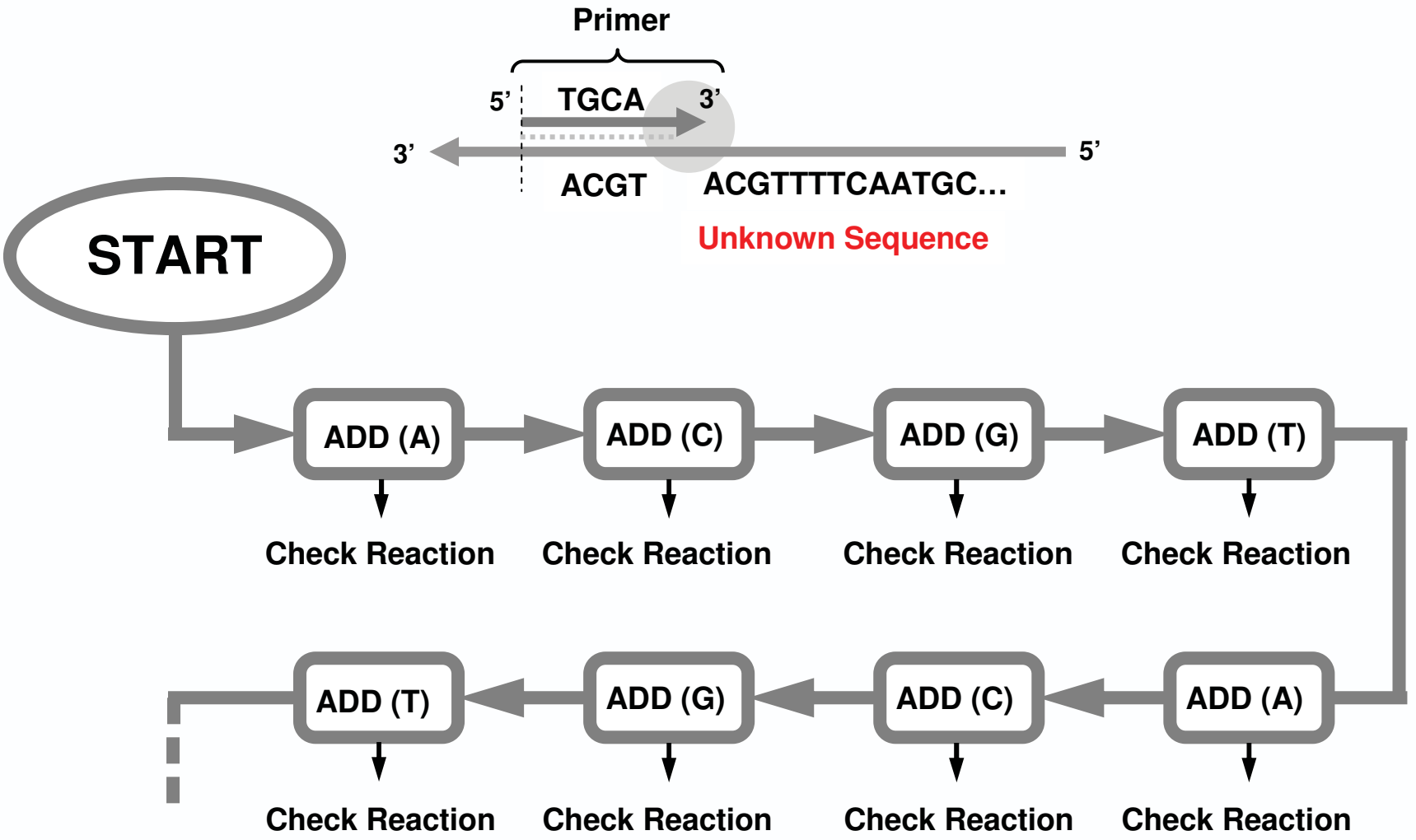
EDWARD DAVID HYMAN

FIG. 1. Schematic diagram of DNA sequencer.

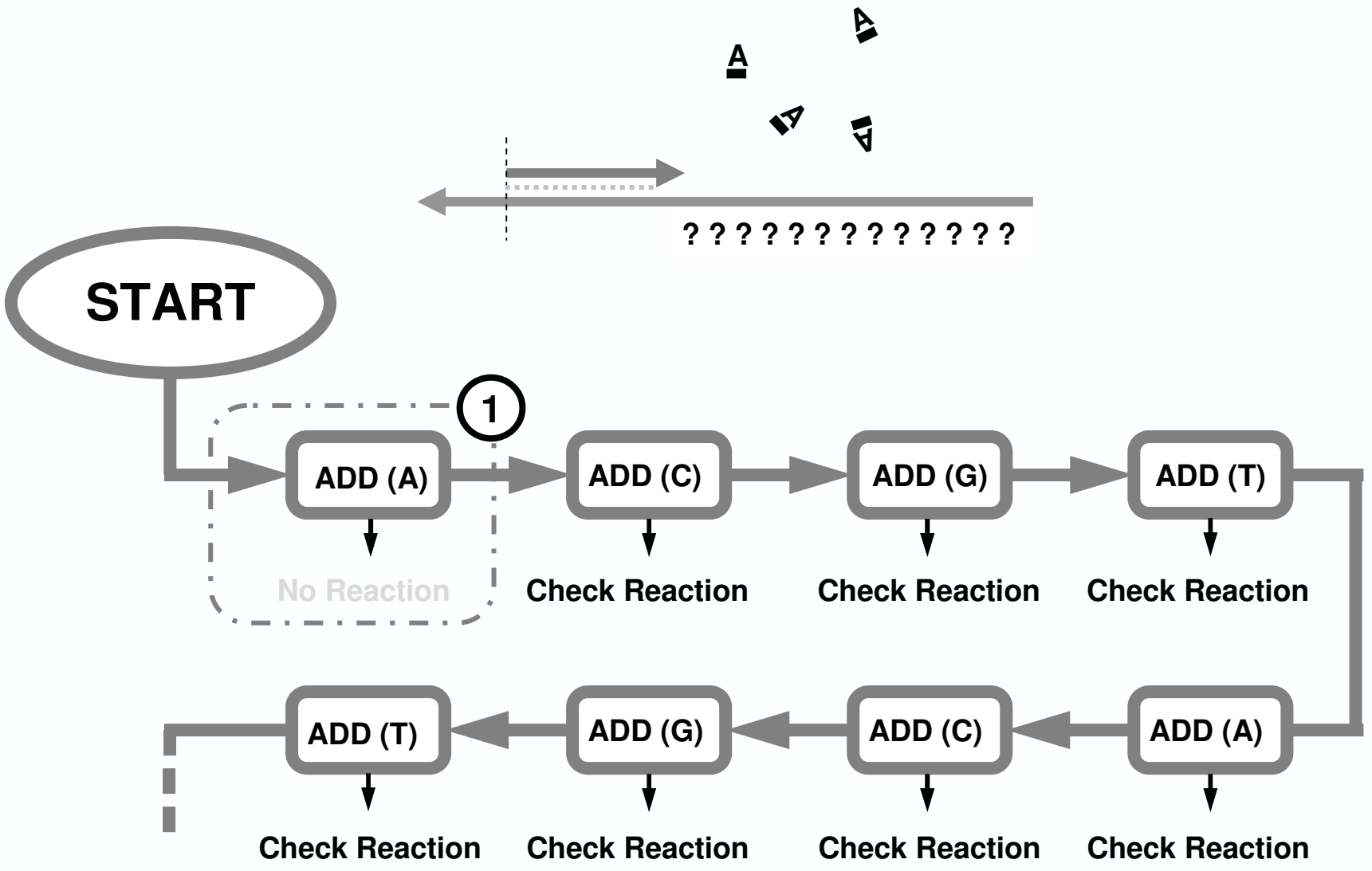


[1] ED Hyman "A new method of sequencing DNA," *Analytical Chemistry*, 174, 423-436, (1988).

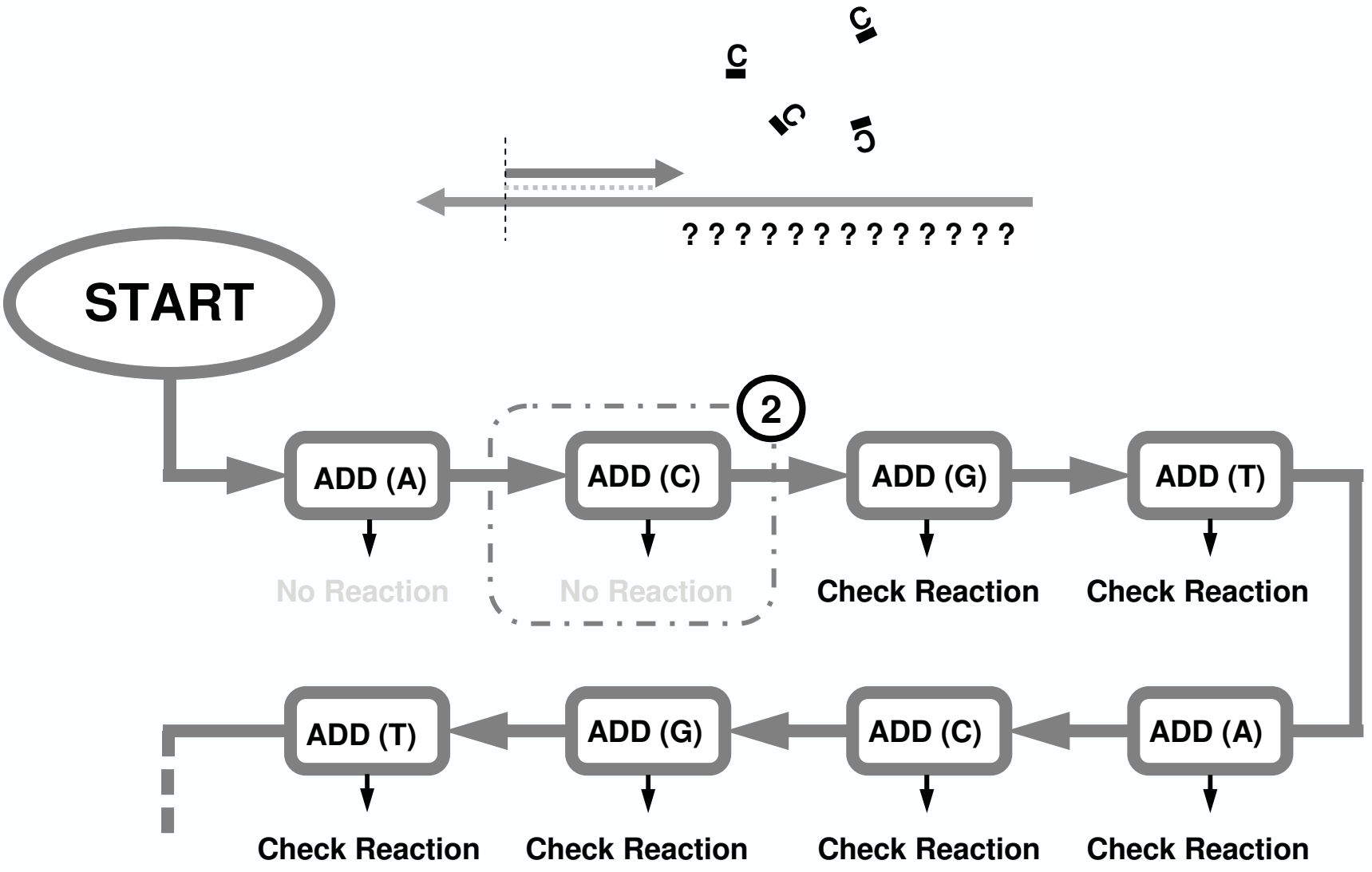
SBS Detection Algorithm



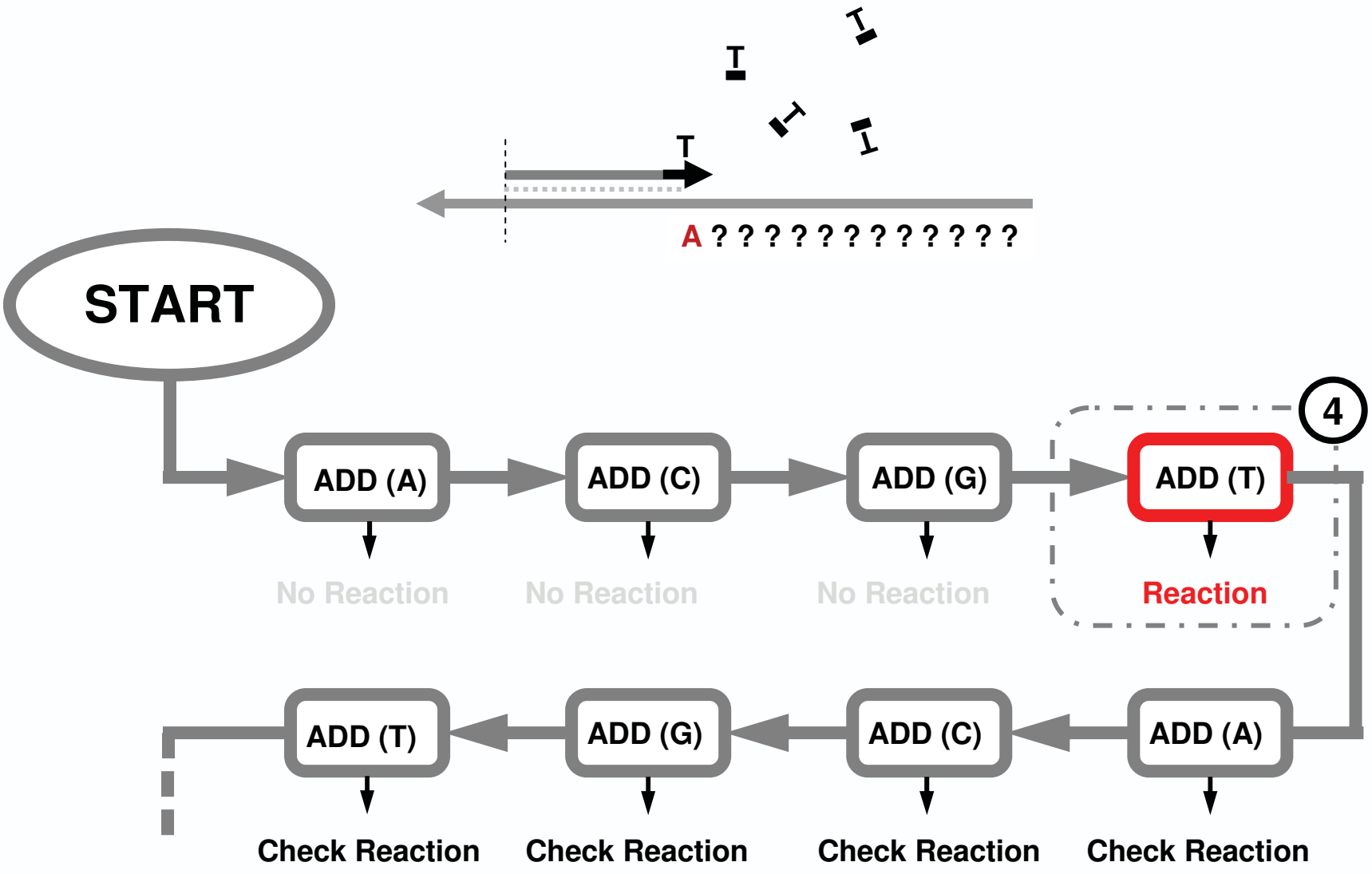
SBS Detection Algorithm



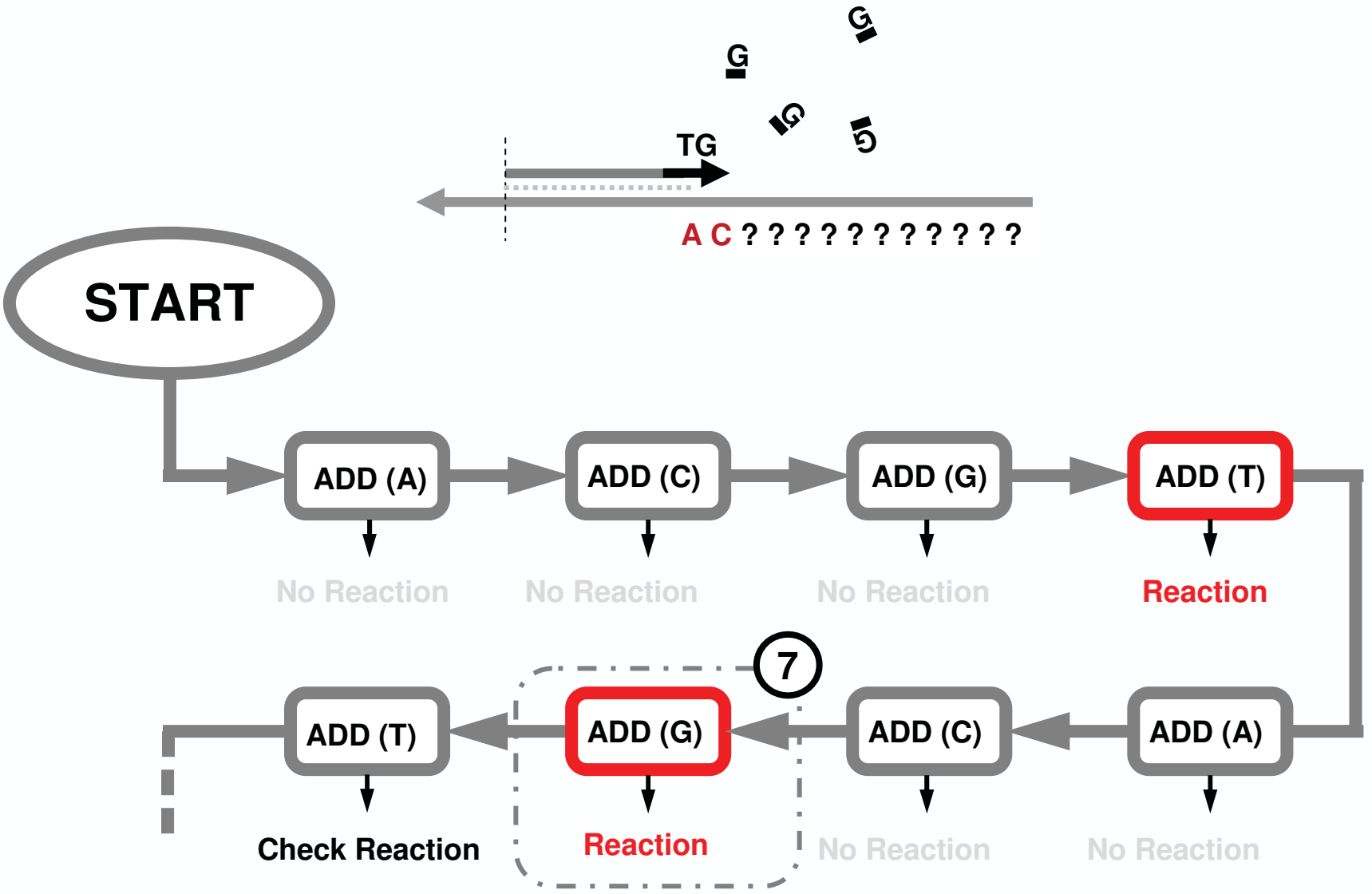
SBS Detection Algorithm



SBS Detection Algorithm

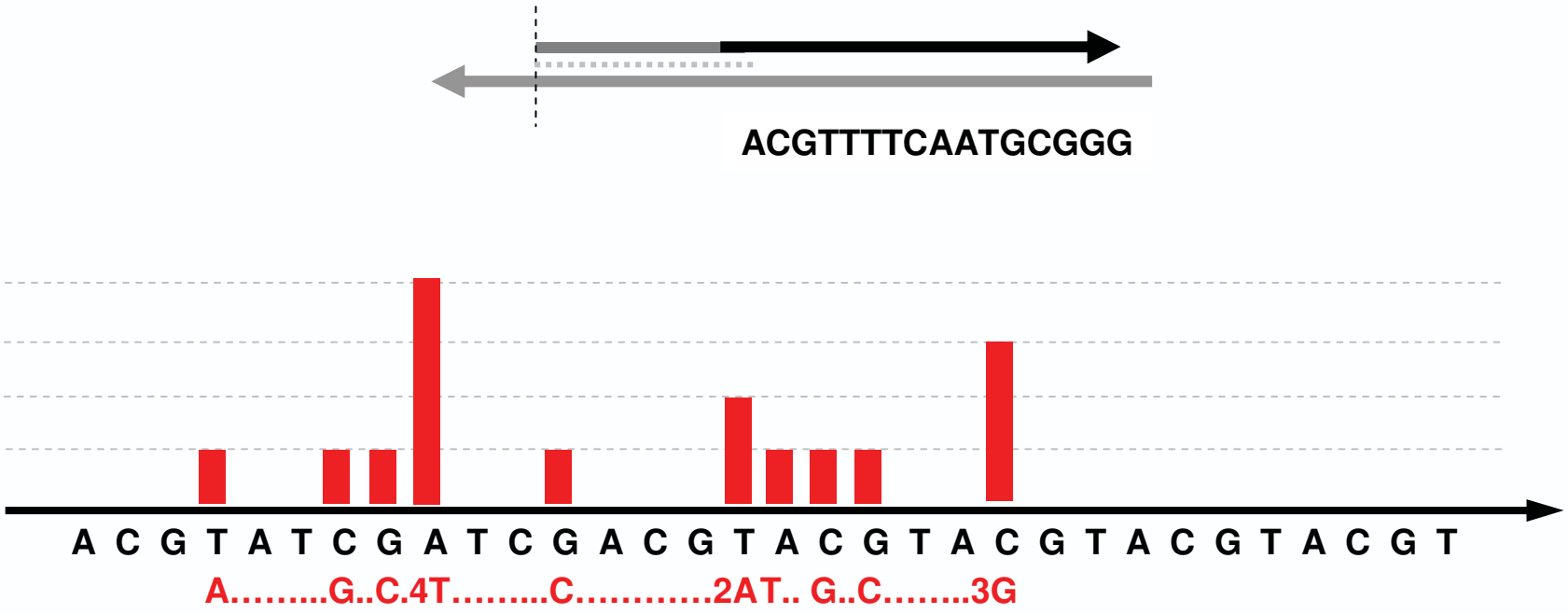


SBS Detection Algorithm



SBS Reaction Diagram

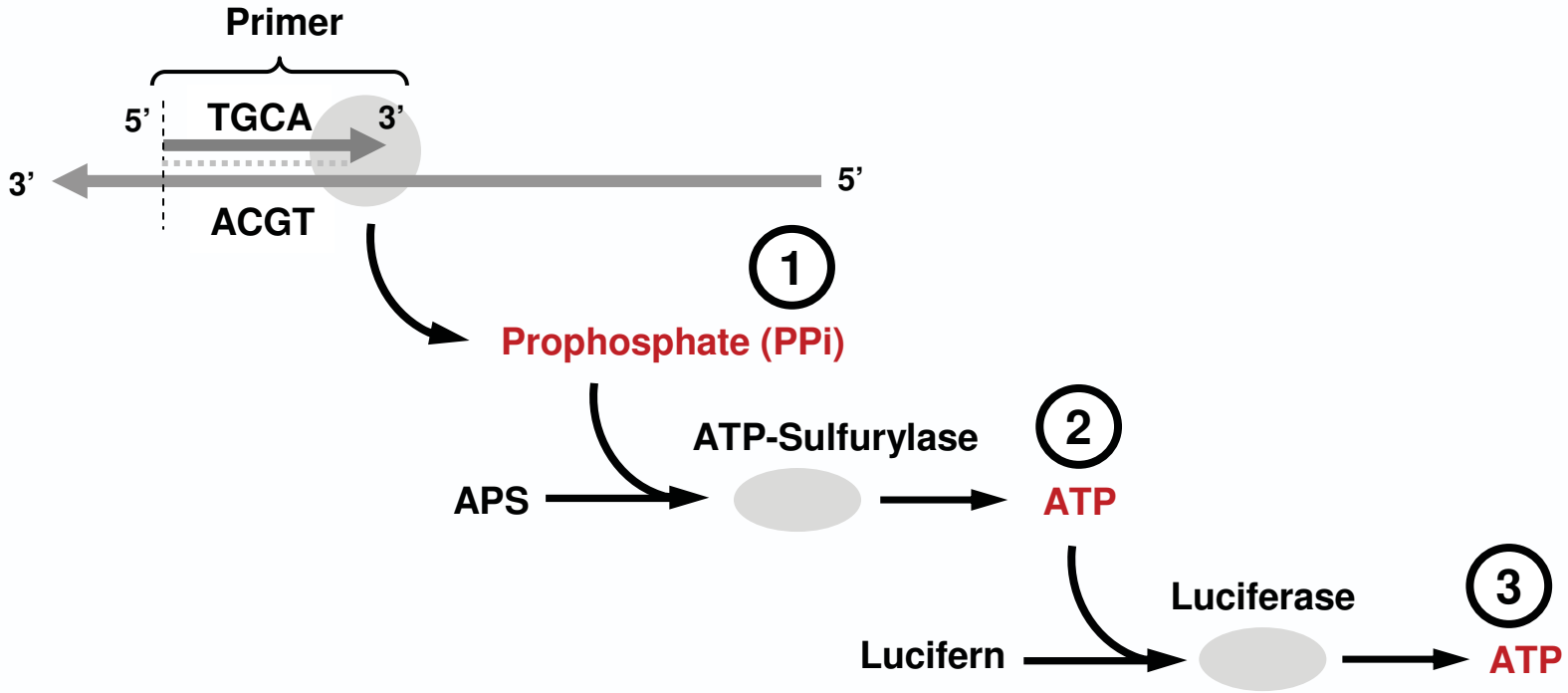
Presence of a reaction and its magnitude is used to sequence the unknown DNA fragment



Sequence: ACGTTTTCAATGCGGG

Pyrosequencing[1]

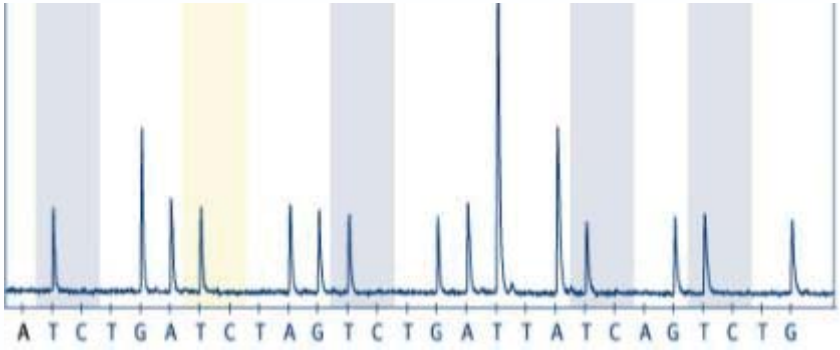
In 1998, Mostafa Ronaghi and Pål Nyrén improved Hyman's SBS method, created a robust assay and coined the term Pyrosequencing

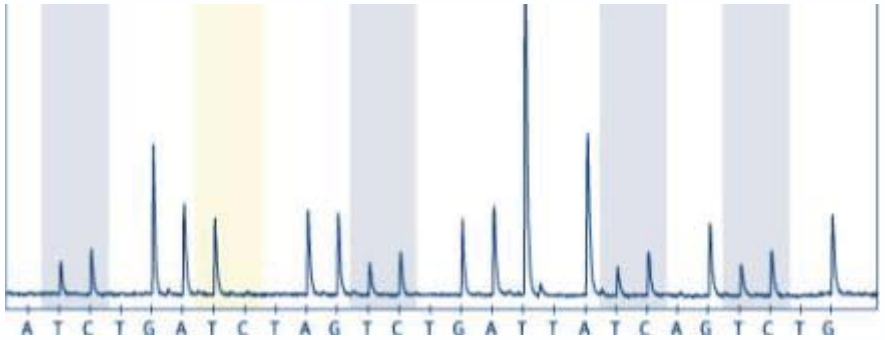


In Pyrosequencing successful reaction trigger an enzymatic cascade to generate photons with λ=562nm

[1] M. Ronaghi, M. Uhlén, and P. Nyrén, "A sequencing method based on real-time pyrophosphate,". *Science*, 1998.

Pyrosequencing Data





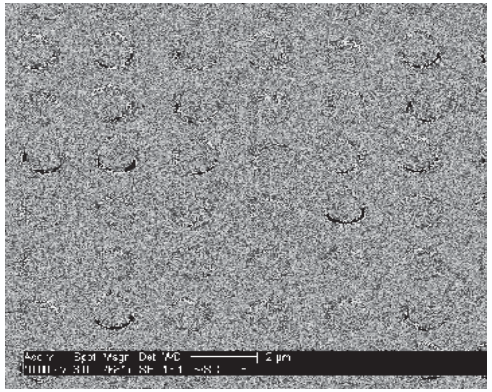
Qiagen's PyroMark (2011)

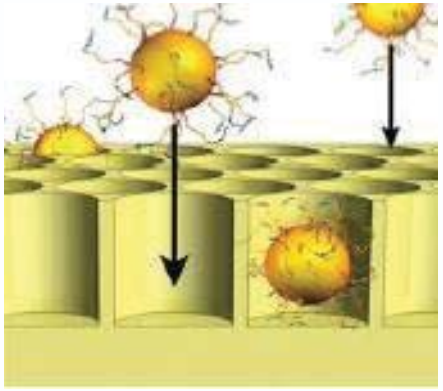
24 microtitre well format
~100 bps read length
~50 kbps per day

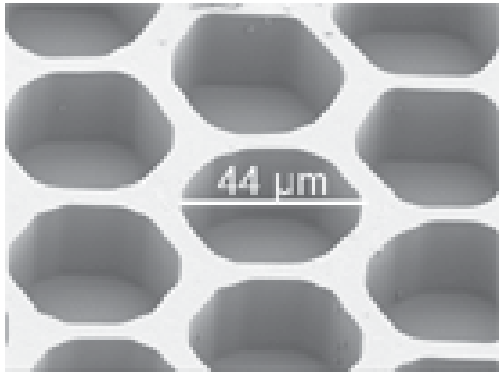
It will take years to sequence a human genome!

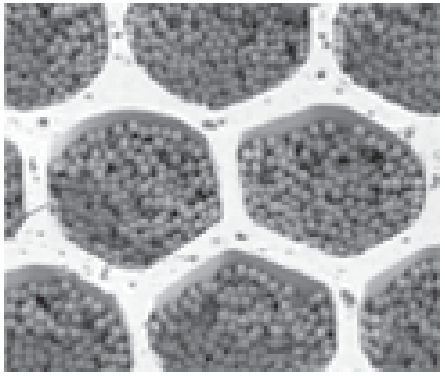
Random Bead Arrays + Pyrosequencing

In 2005, the company 454, a subsidiary of Curagen created a high-throughput Pyrosequencing system enabled by random bead arrays









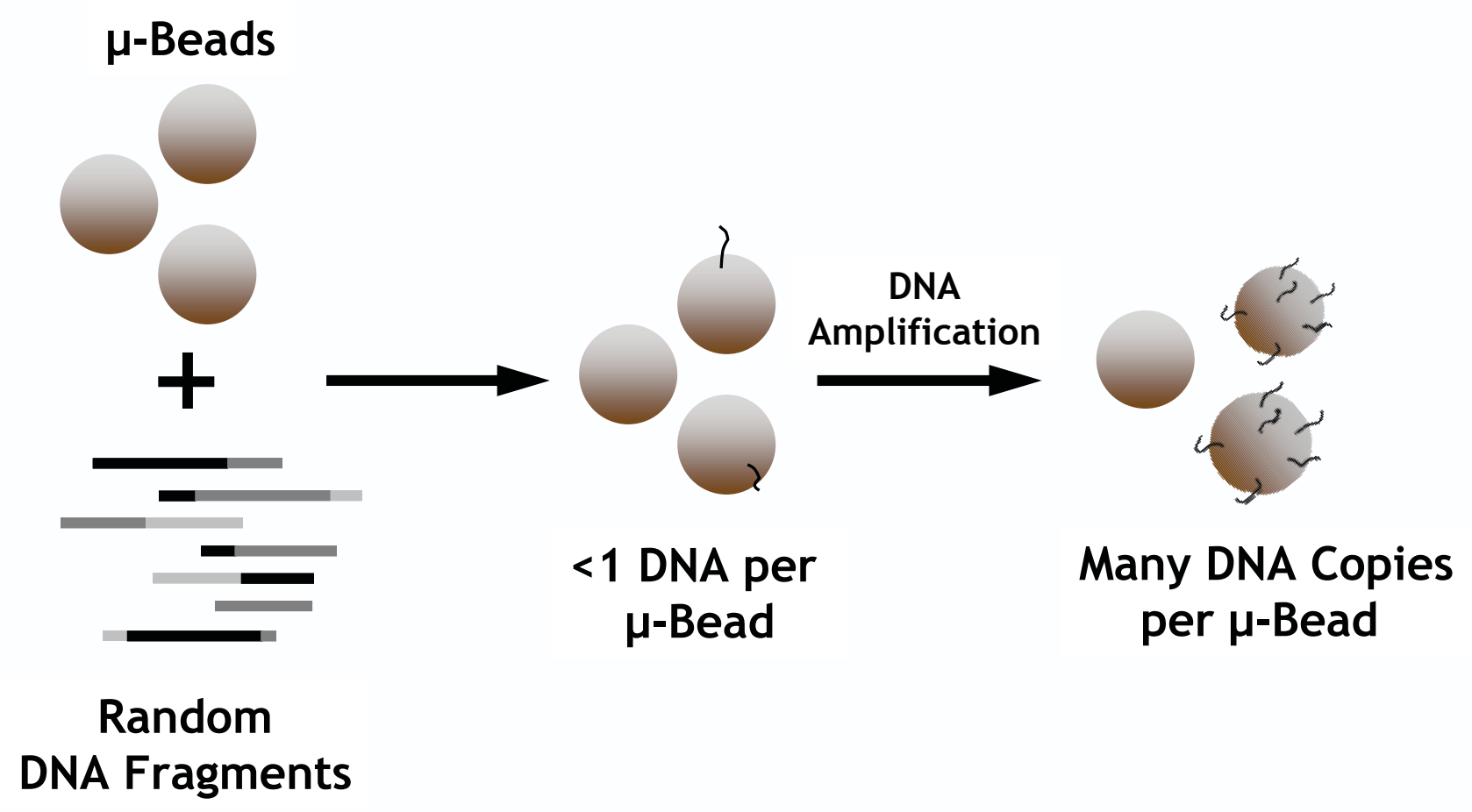
454 (Roche) GS FLX+ (2011)

>1 M micro-wells
~700 bps read length
~700 Mbs per day

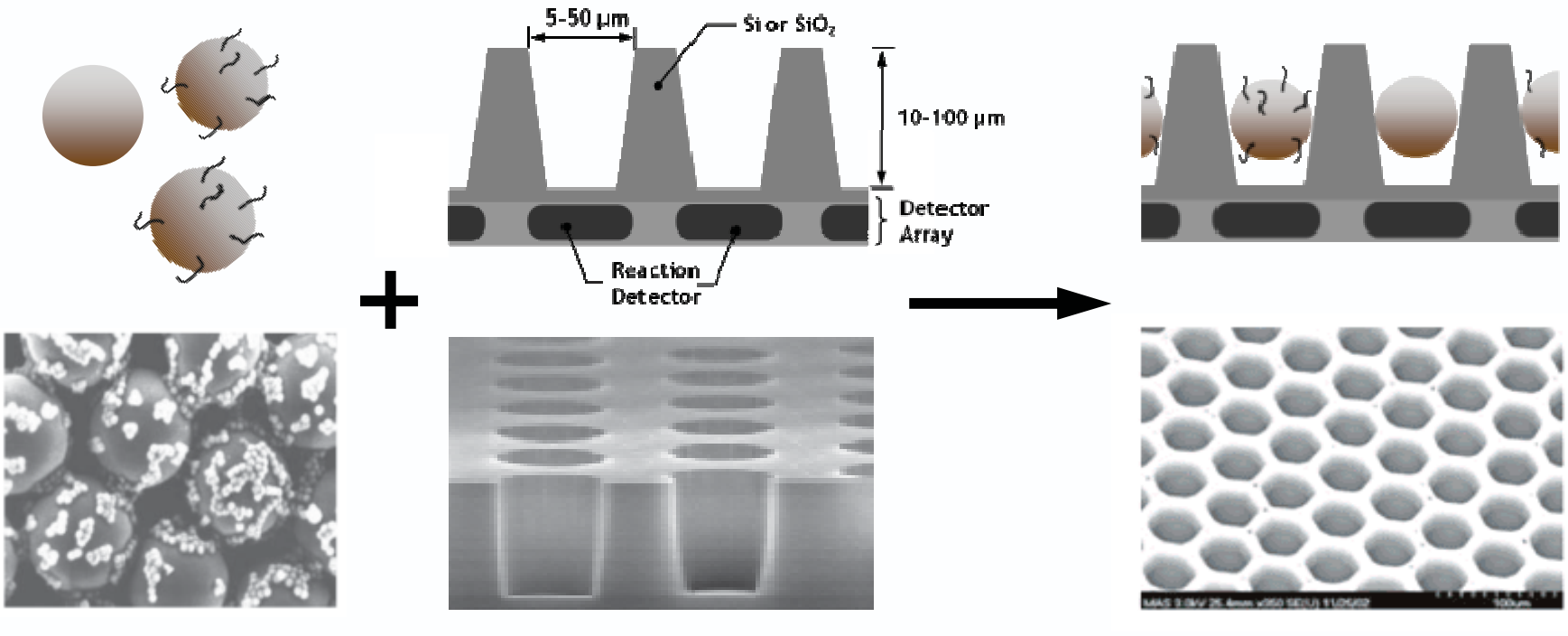
It will take 2 weeks to sequence a human genome

[1] M. Margulies + 54 additional coauthors "Genome sequencing in open microfabricated high density picoliter reactors," *Nature* (2005).

Step 1: Create DNA Clusters on µ-Beads



Step 2: Loading the Array

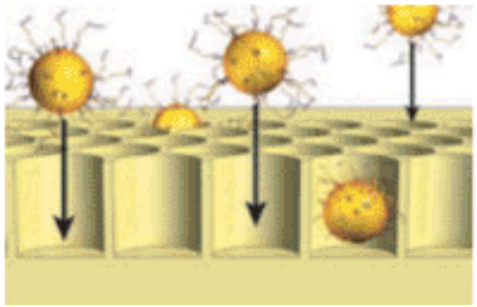


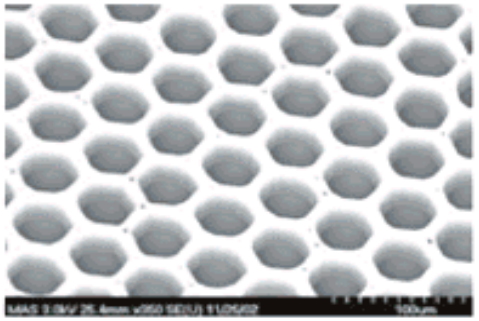
Bead + DNA

µ-Well

Random Bead Array

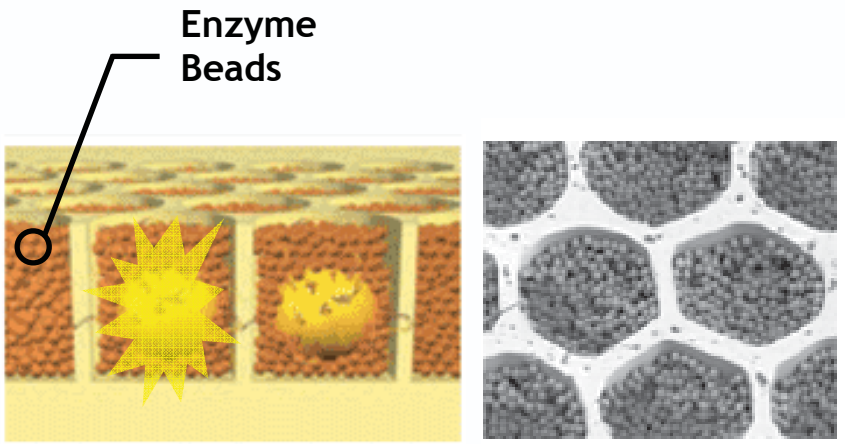
Step 3: Pyrosequencing

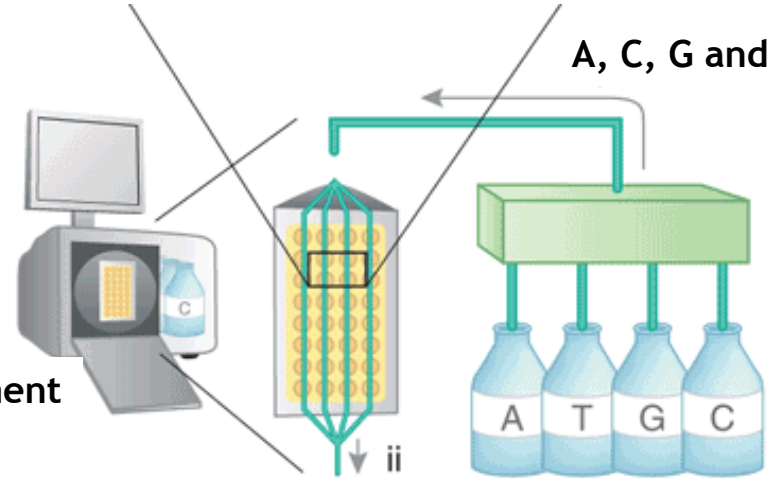




Random Bead Array

Running Reactions



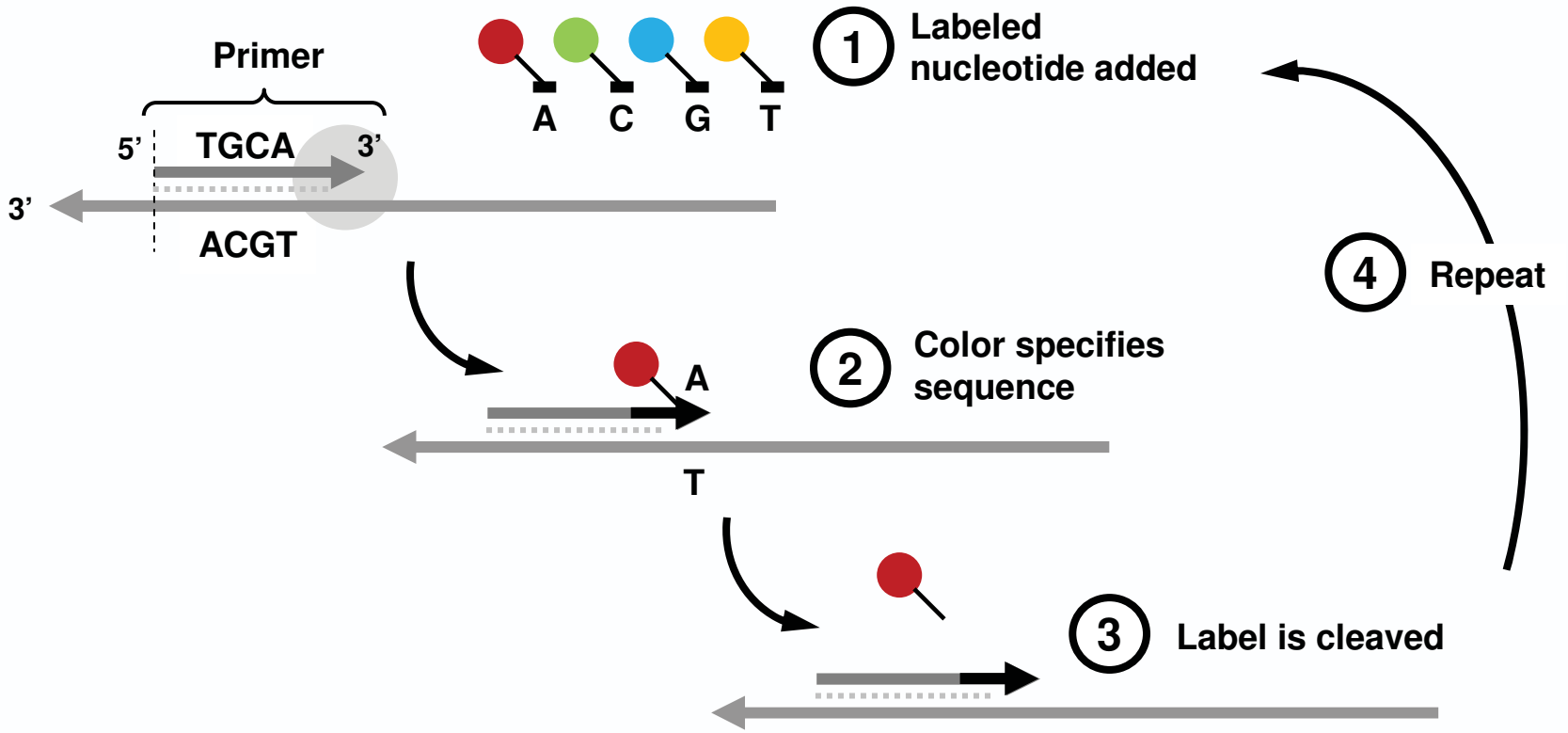


Flow-Through Chip

[1] M. Margulies + 54 additional coauthors "Genome sequencing in open microfabricated high density picoliter reactors," *Nature* (2005).

Illumina (Solexa) DNA Sequencing

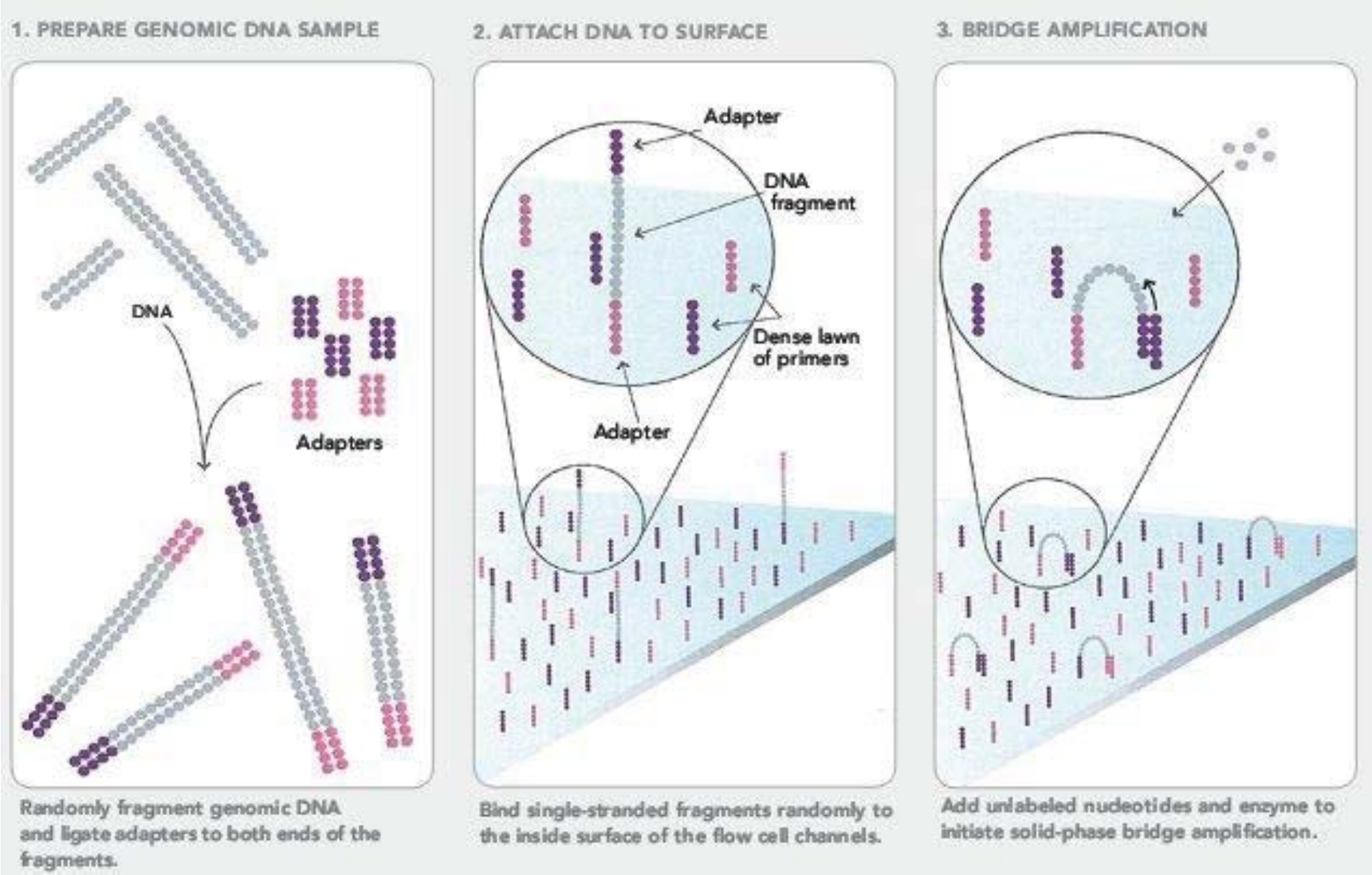
Merges the technologies from Solexa[1] (UK), Manteia[2] (Switzerland) to create a high-throughput SBS system



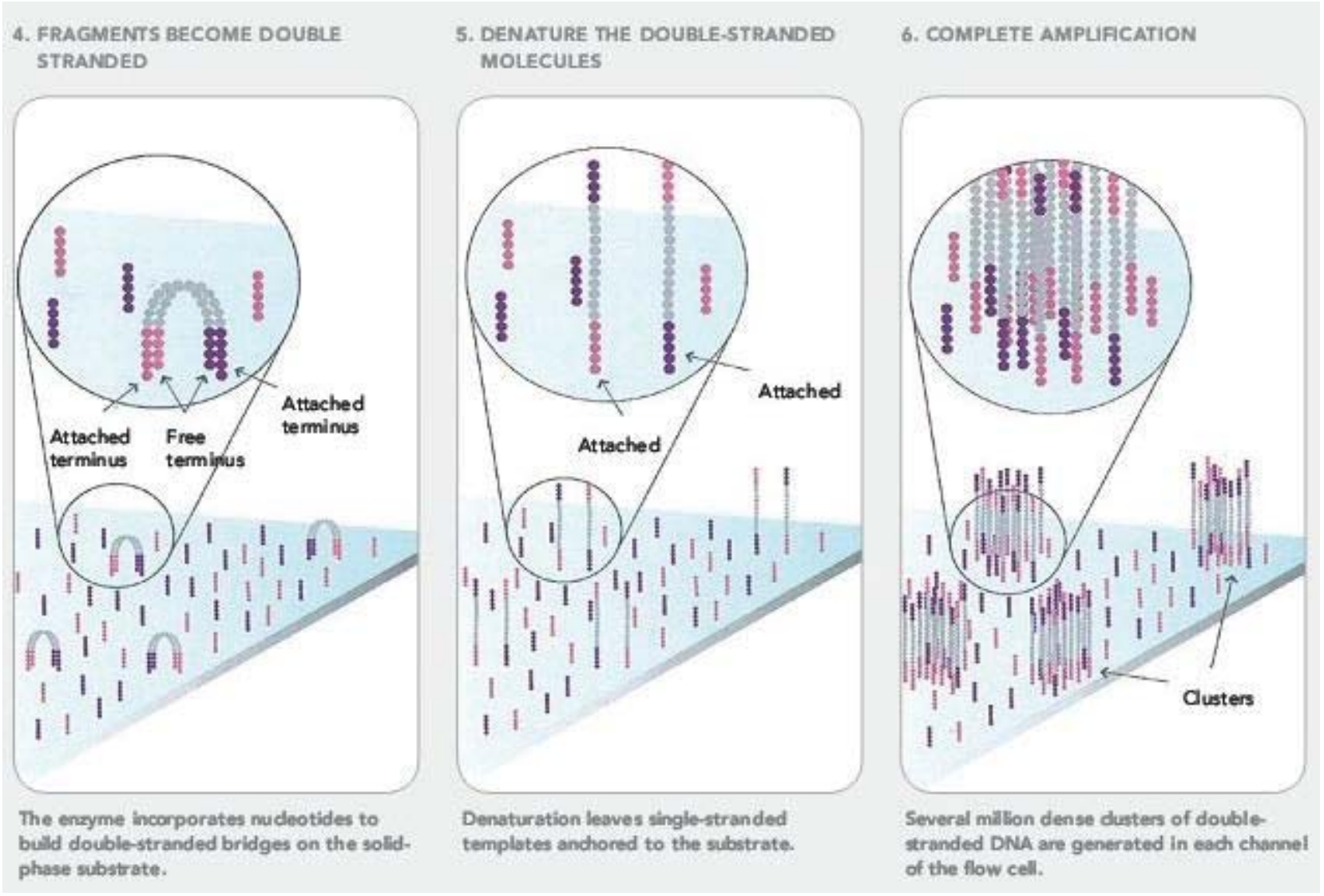
[1] Turcatti *et al.*, "A new class of cleavable fluorescent nucleotides: synthesis and optimization as reversible terminators for DNA sequencing by synthesis," *Nucleic Acids Res.*, (2008).

[2] Adessi *et al.*, "Solid phase DNA amplification: characterization of primer attachment and amplification mechanisms," *Nucleic Acids Res.*, (2000).

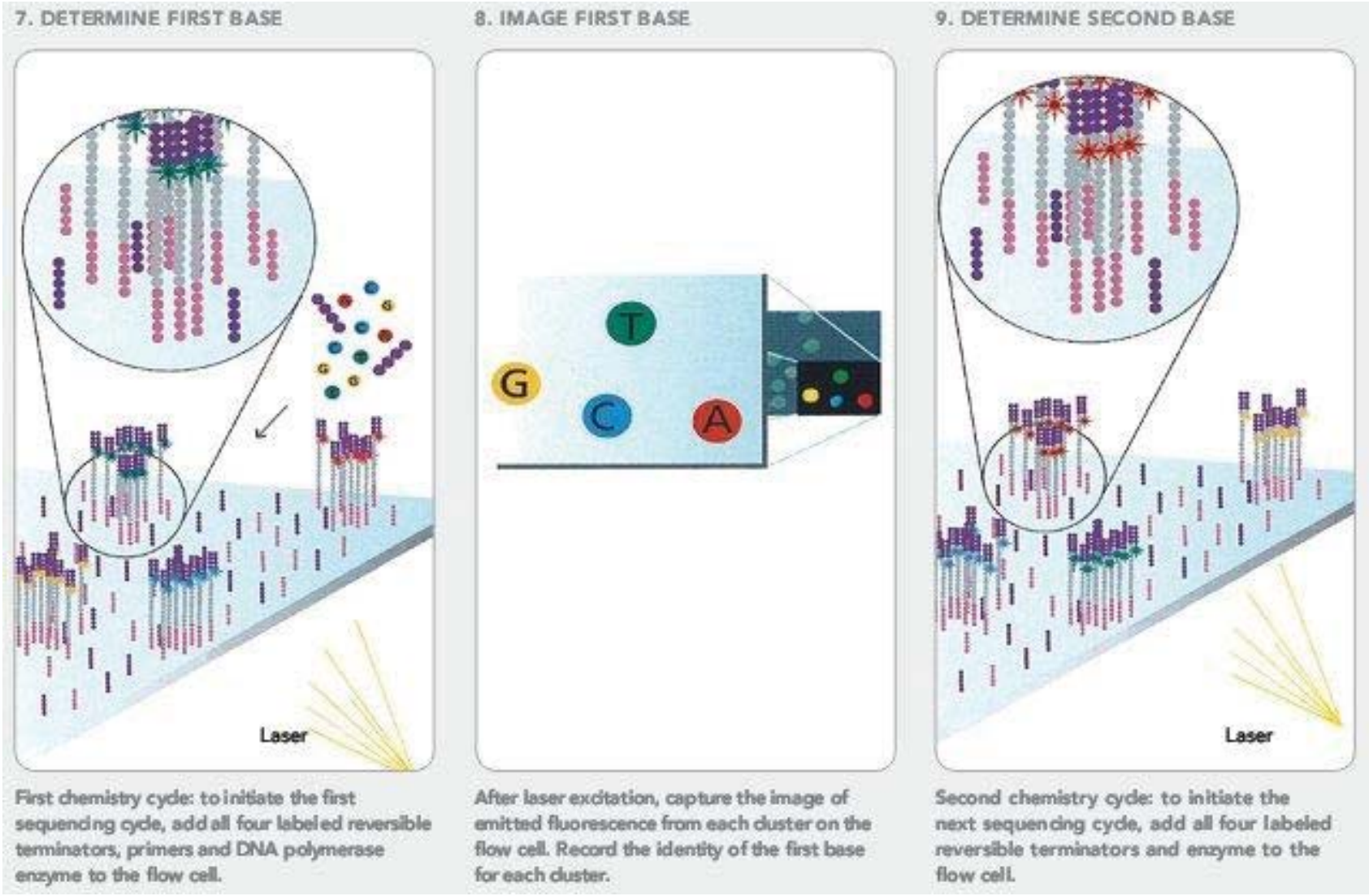
Step 1: Creating Fragments



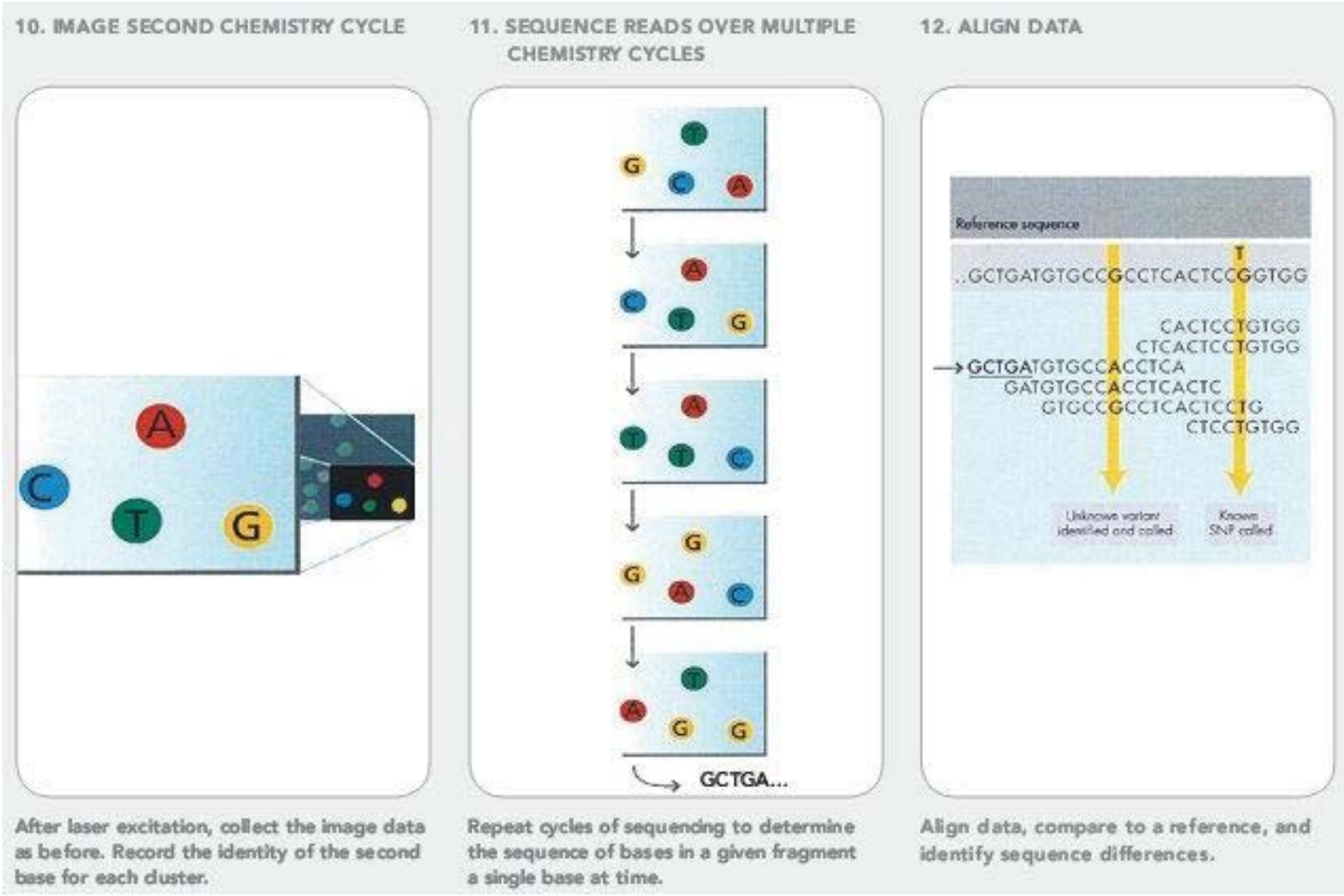
Step 2: Bridge Amplification



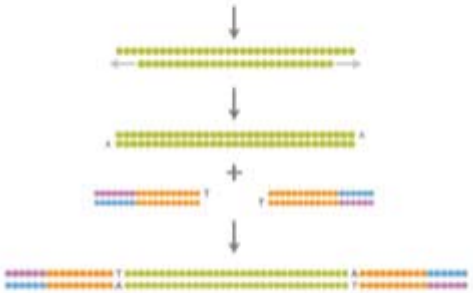
Step 3: SBS Using Labels



Step 4: Data Analysis



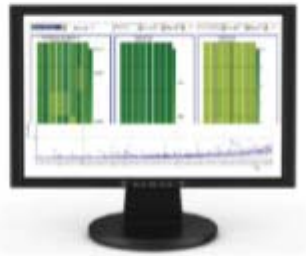
Illumina (Solexa) DNA Sequencing



Library Preparation
~2 h [15 min hands-on (Nextera)]
< 6 h [< 3 h hands-on (TruSeq)]

Cluster Generation
~5 h (<10 min hands-on)

Sequencing by Synthesis
~1.5 to 11 days



CASAVA
2 days (30 min hands-on)

HiSeq 2000 (2011)

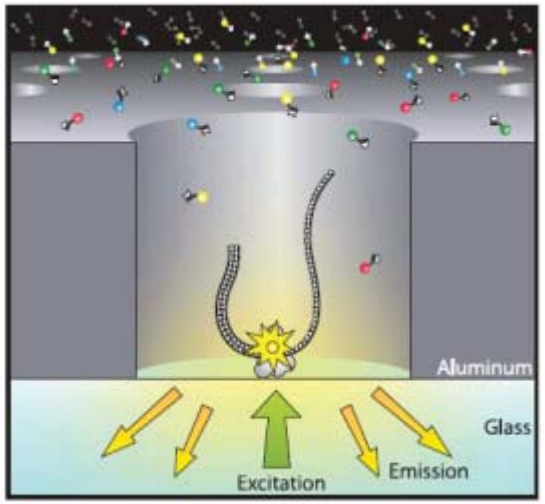
~100 bps read length
~100 Gbps per 2 days

It will take 2 days to sequence a human genome!

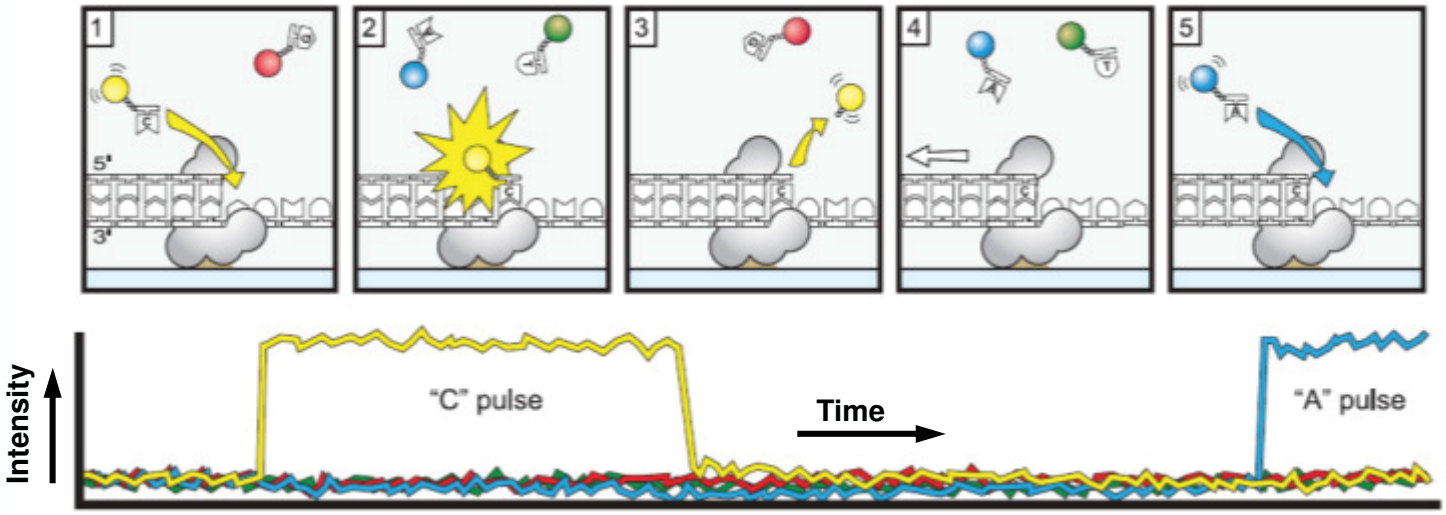
Upcoming Sequencers

***Disclaimer*:** The presenter cannot verify the performance of these systems in October 2011

Pacific Biosciences[1] (PacBio)



Polymerization of a single molecule is detected in real-time using a labeled nucleotides

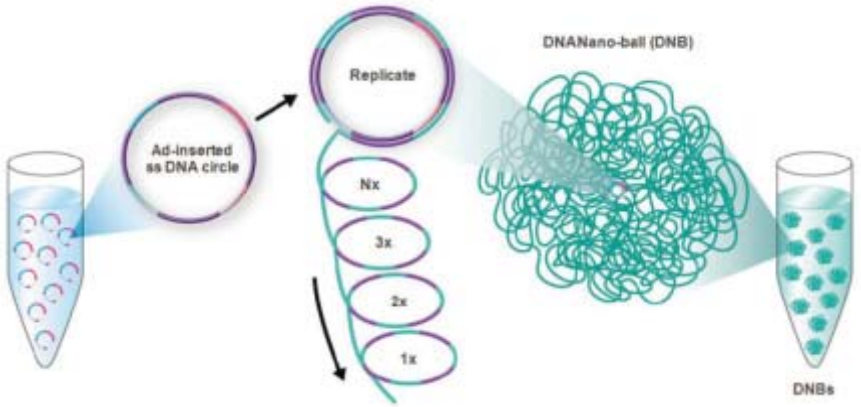


[1] J. Eid *et al.*, "Real-time DNA sequencing from single polymerase molecules," *Science*, 2009.

Pacific Biosciences (PacBio)

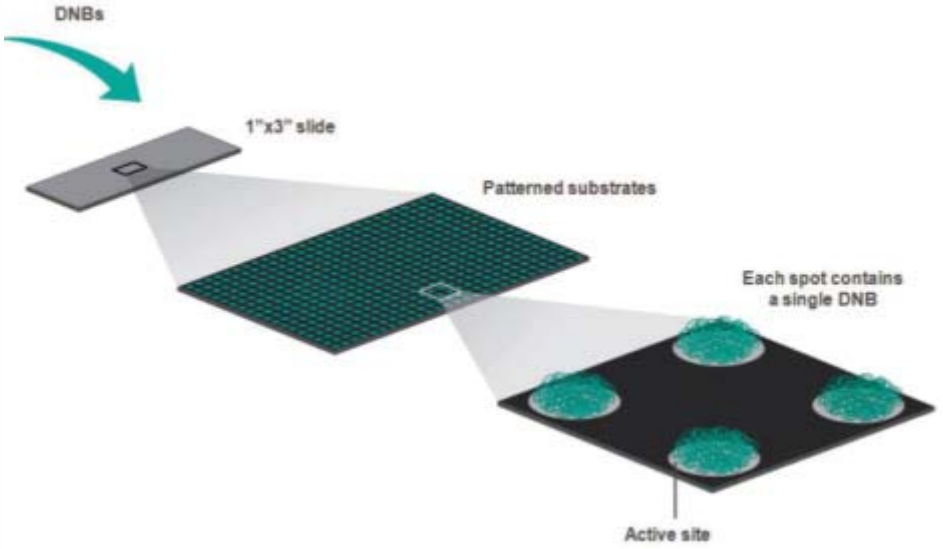
[1] J. Eid *et al.*, "Real-time DNA sequencing from single polymerase molecules," *Science*, 2009.

DNA Nanoball Sequencing[1] (Complete Genomics)



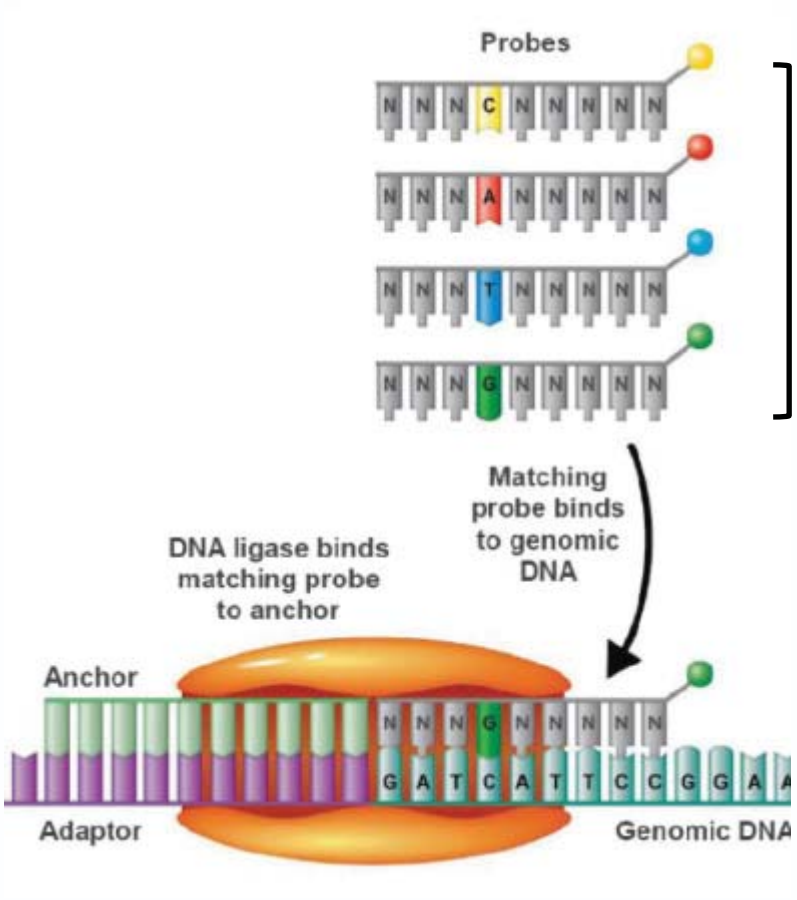
① **Unknown DNA is amplified into repetitive long sequence to form a "nanoball"**

② **The "nanoballs" get randomly attached on a substrate with ~2.8B spots**



[1] R. Drmanac *et al.*, "Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays," *Science*, 2010.

DNA Nanoball Sequencing (Complete Genomics)

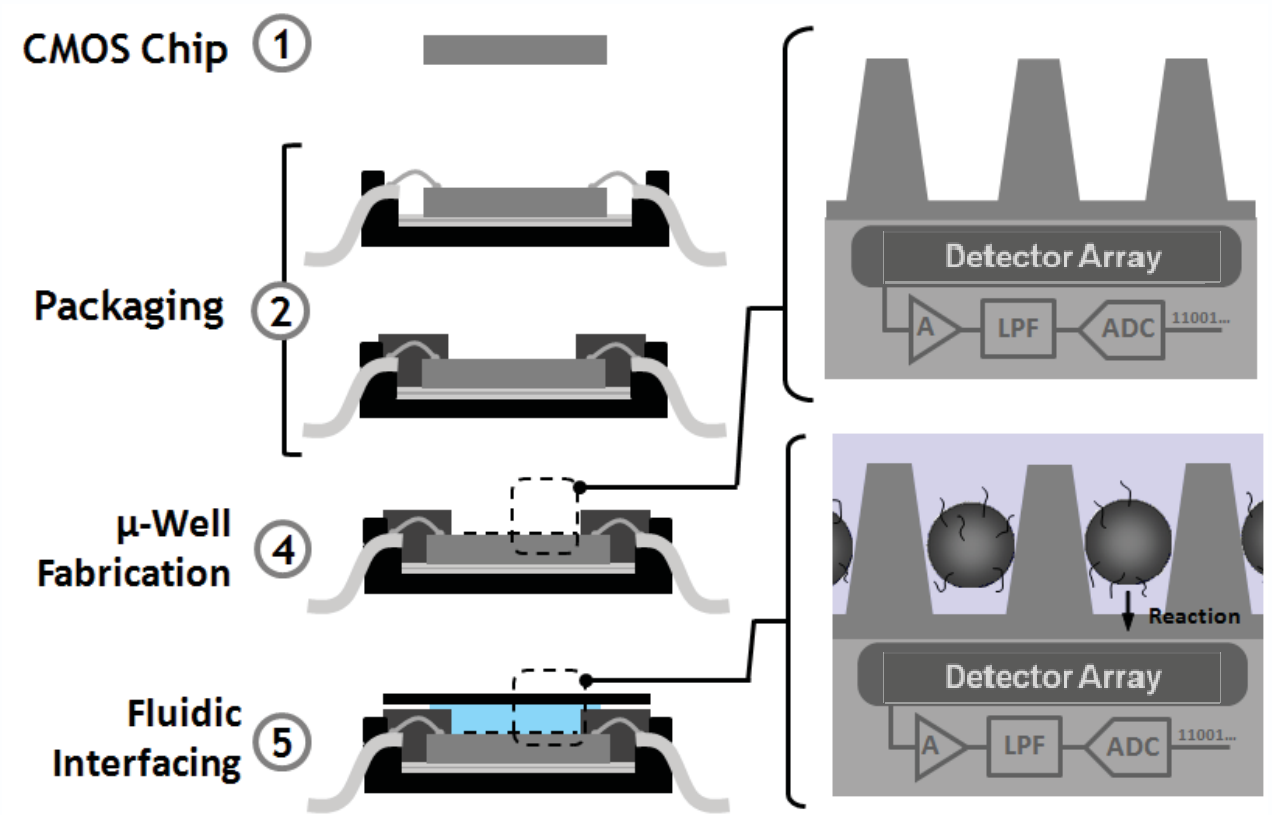


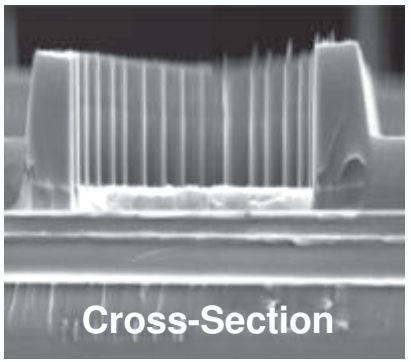
"N" is degenerate nucleotide, i.e., A, C, G, or T

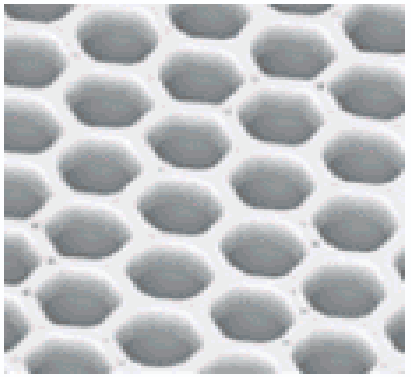
Hybridization and ligation are to used to find the sequence of specific locations on the unknown strand

Semiconductor Sequencing[1] (Ion Torrent)

Detecting polymerization electronically



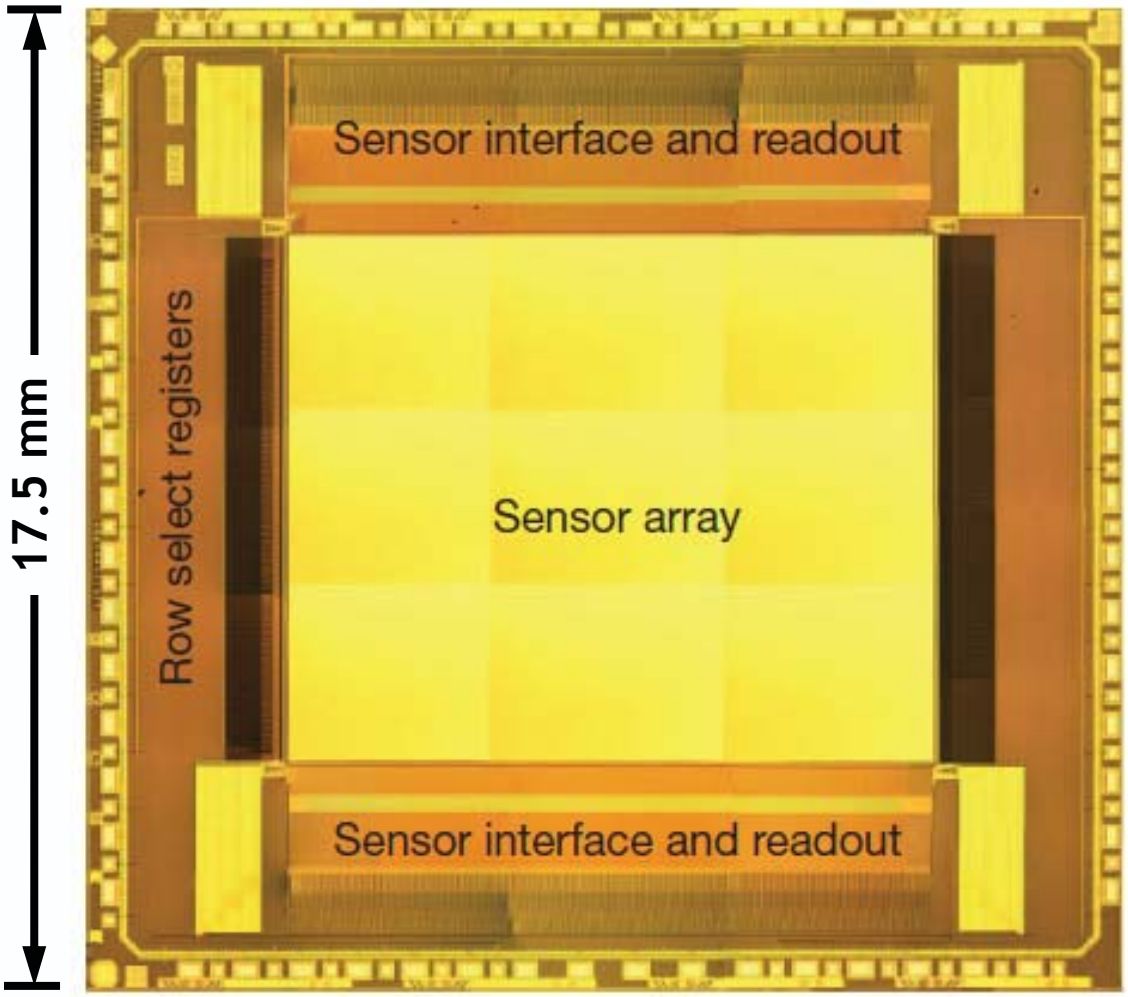


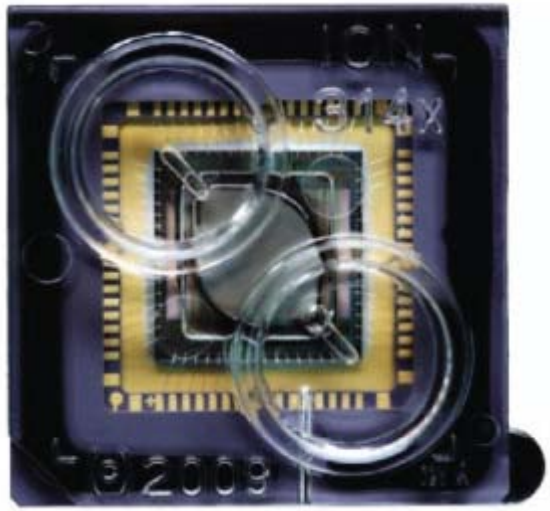


Top View

[1] Rothberg *et al.*, *Nature* (2011)

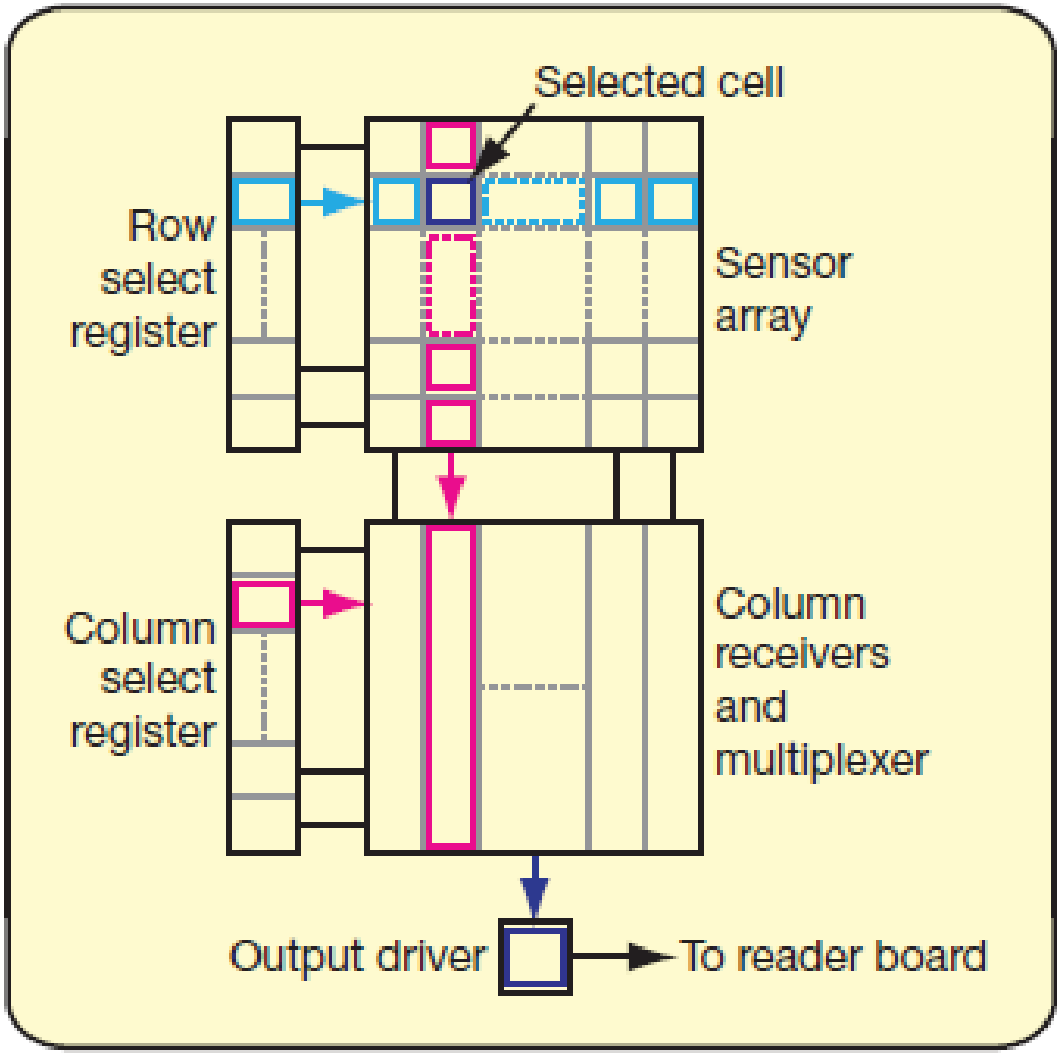
Ion Torrent Sequencing Chip[1]

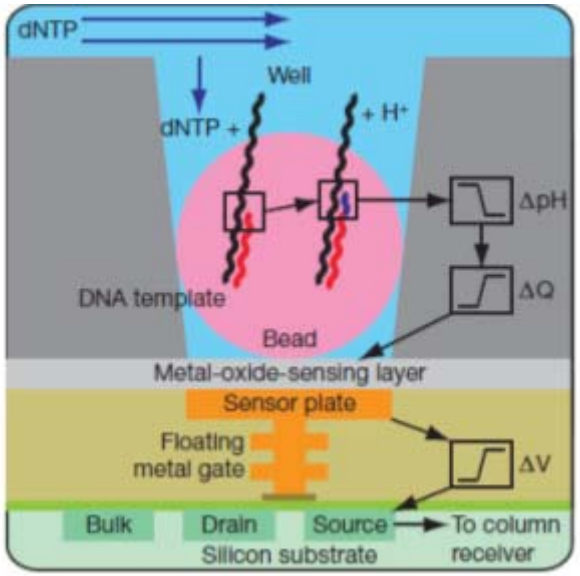


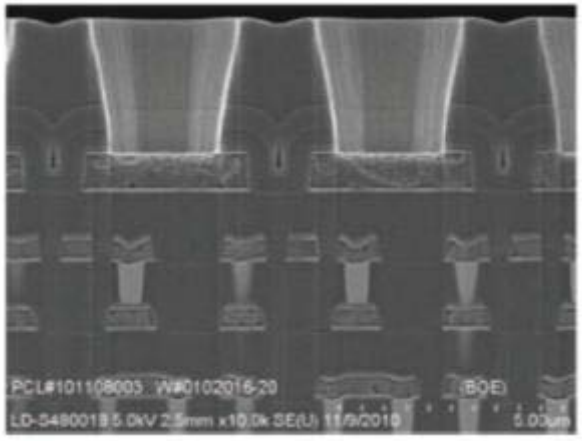


[1] Rothberg *et al.*, *Nature* (2011)

Architecture[1]

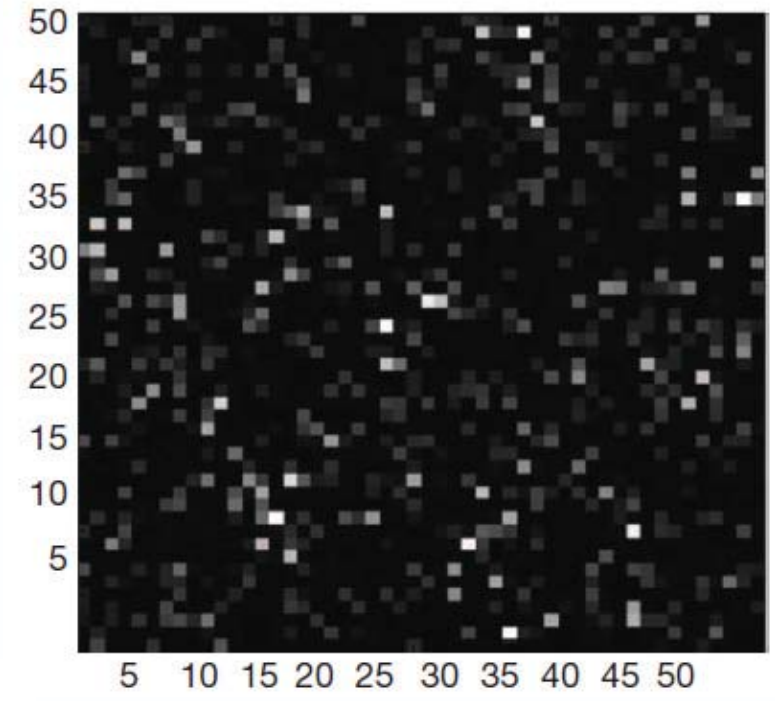


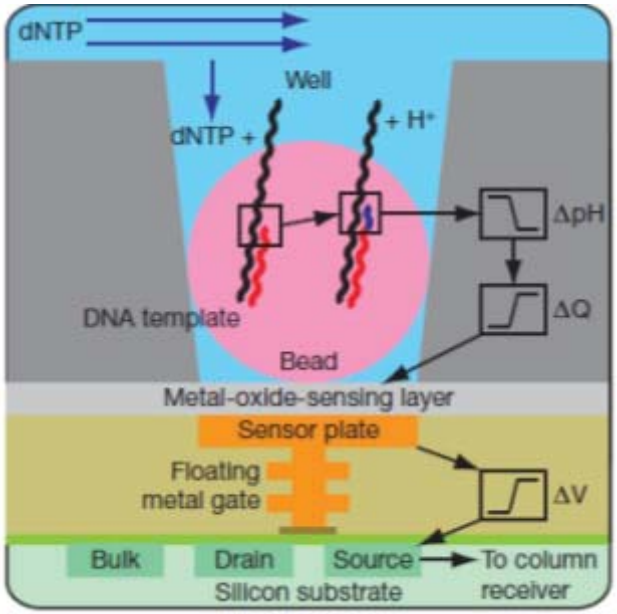


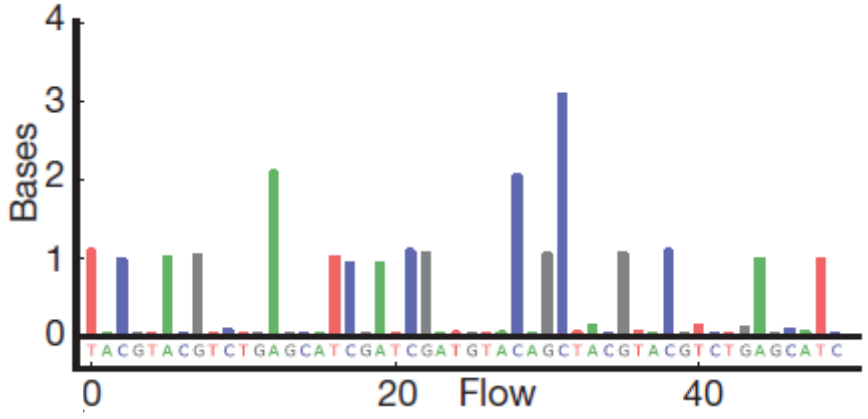


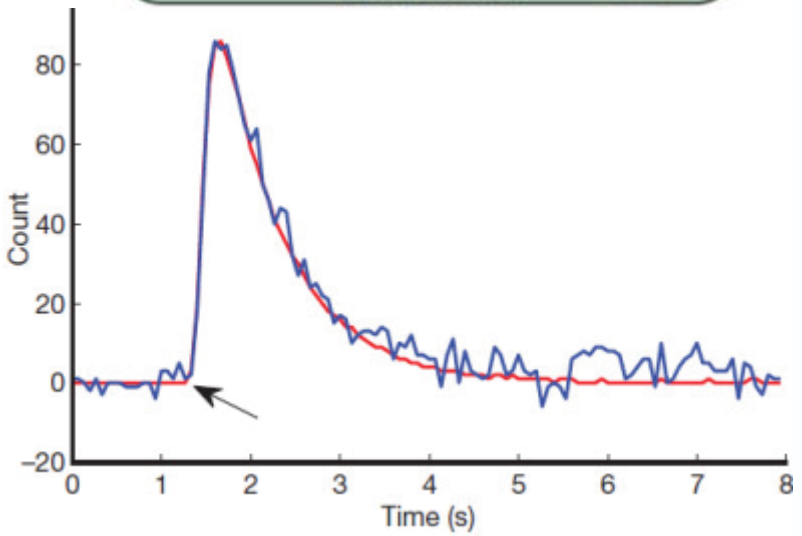
[1] Rothberg *et al.*, *Nature* (2011)

Electronic SBS Signals[1]



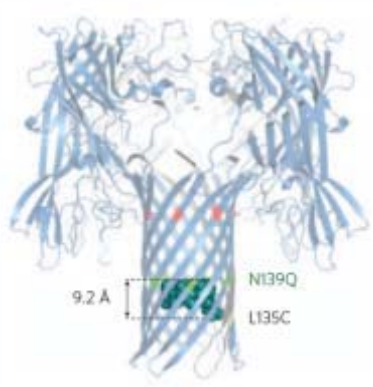


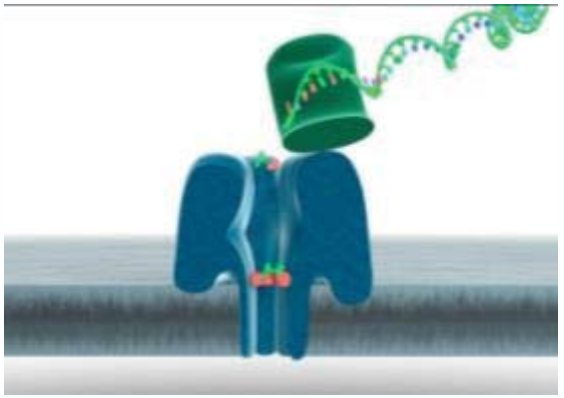


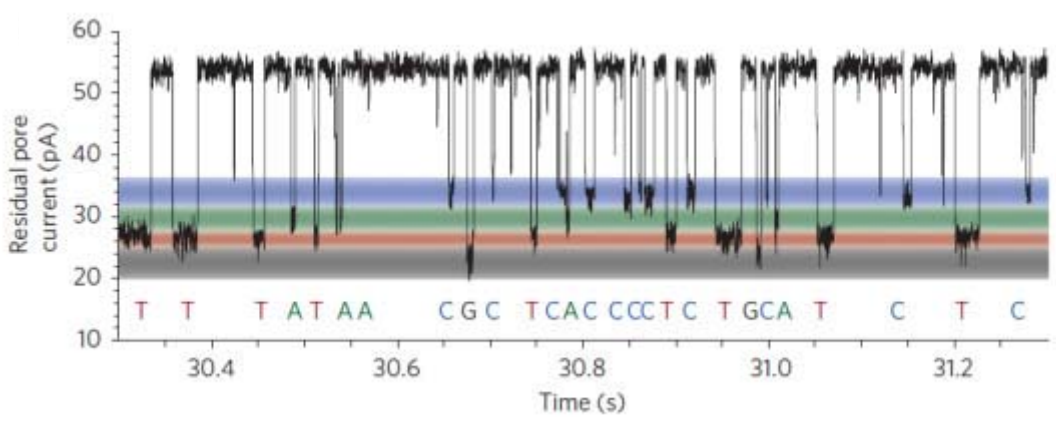


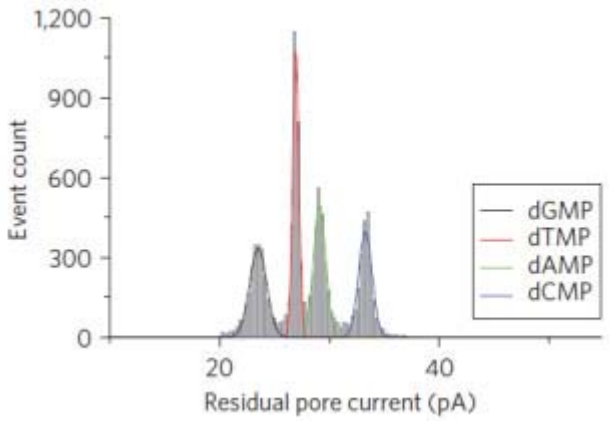
Nanopore Sequencing[1] (Oxford Nanopore)

Digested DNA bases can create dissimilar blockage currents in a organic nanopore structure





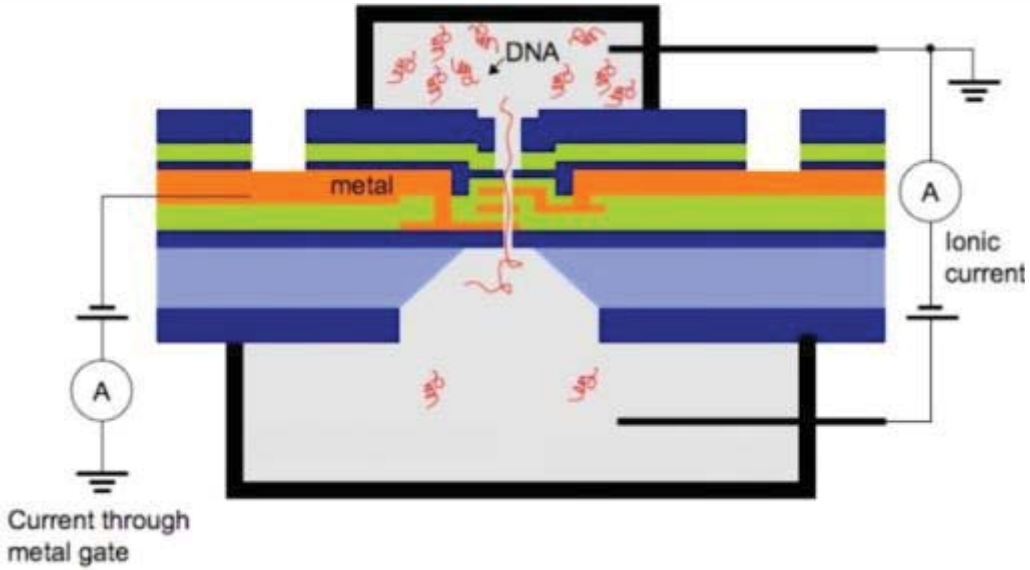


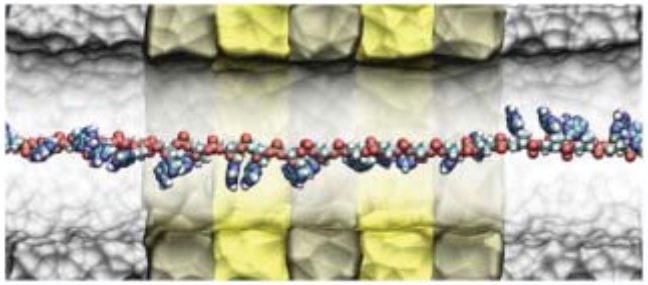


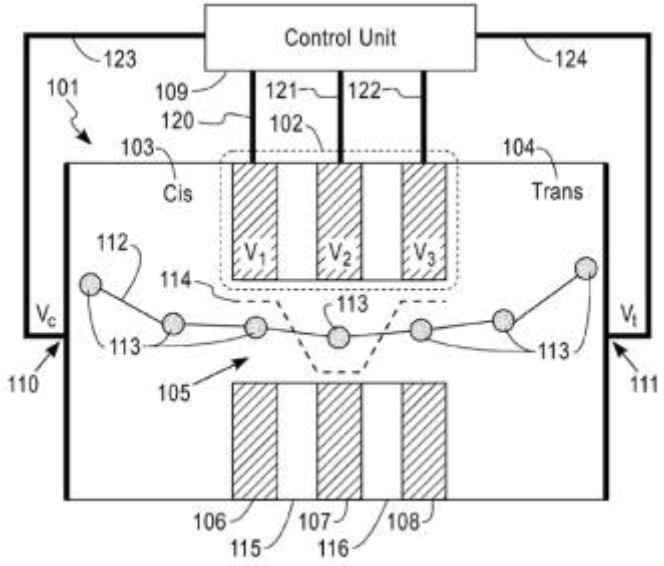
[1] Clarke *et al.*, "Continuous base identification for single-molecule nanopore DNA sequencing," *Nature Nano*, 2009.

Nanopore Sequencing[1] (IBM/Roche)

Use solid-state probes with embedded electrodes to simultaneously ratchet and sequence DNA







[1] Polonsky *et al.*, "Nanopore in metal-dielectric sandwich for DNA position control," *Applied Physics Letters*, 2007.

Conclusion

- Sequencing matters
- It is a complicated problem
- It requires much more than biology and perhaps deep understanding of physics and engineering
- There are many crazy methods (and nebulous ideas) out there
- You should try to come up with your own method … seriously!