Genome Sequencing

I: Methods

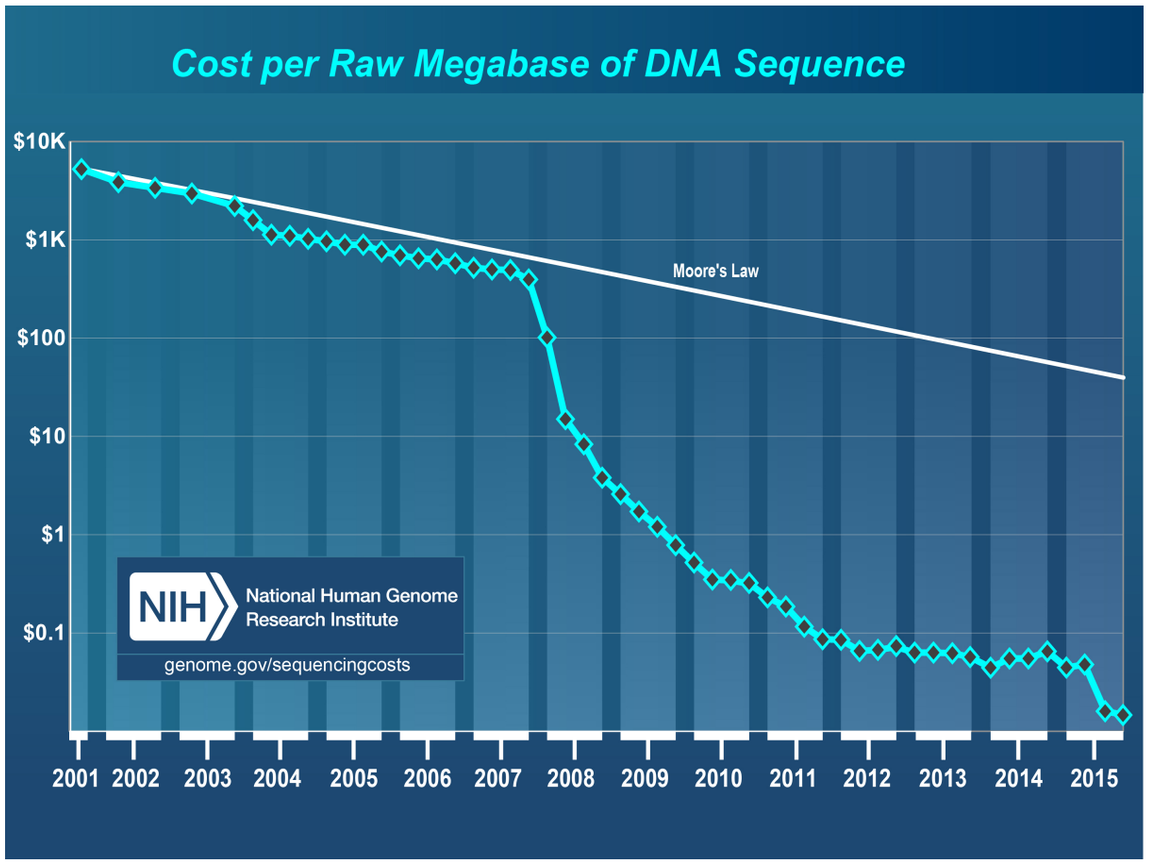
MMG 835, SPRING 2016
Eukaryotic Molecular Genetics

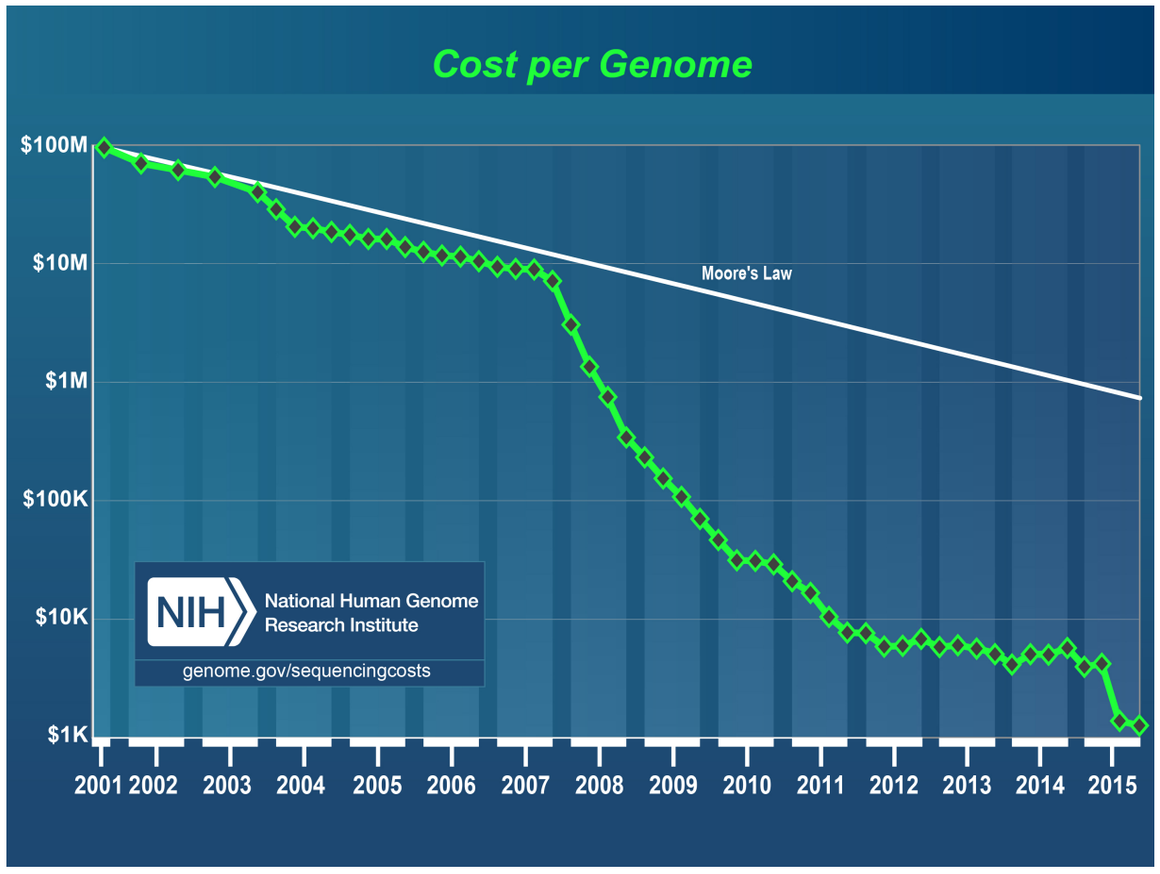
George I. Mias

Department of Biochemistry and Molecular Biology
gmias@msu.edu

Sequencing Methods

Cost of Sequencing





Wetterstrand KA. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP) Available at: www.genome.gov/sequencingcosts. Accessed 3/25/16.

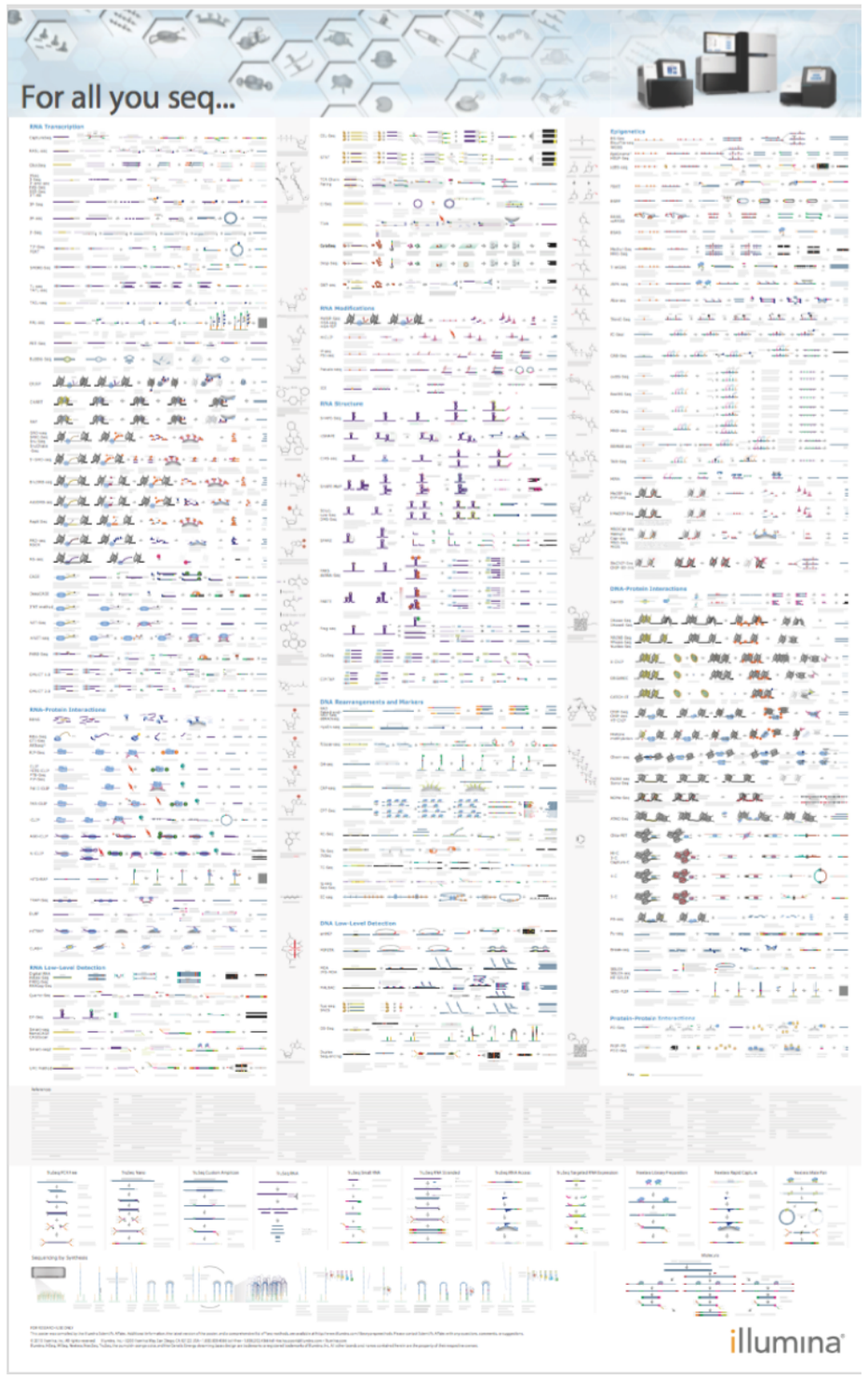
Overview [15838] | Eukaryotes [2867] | Prokaryotes [66600] | Viruses [5545] | Plasmids [6863] | Organelles [7827]

http://www.ncbi.nlm.nih.gov/genome/browse/ (3/26/2016)

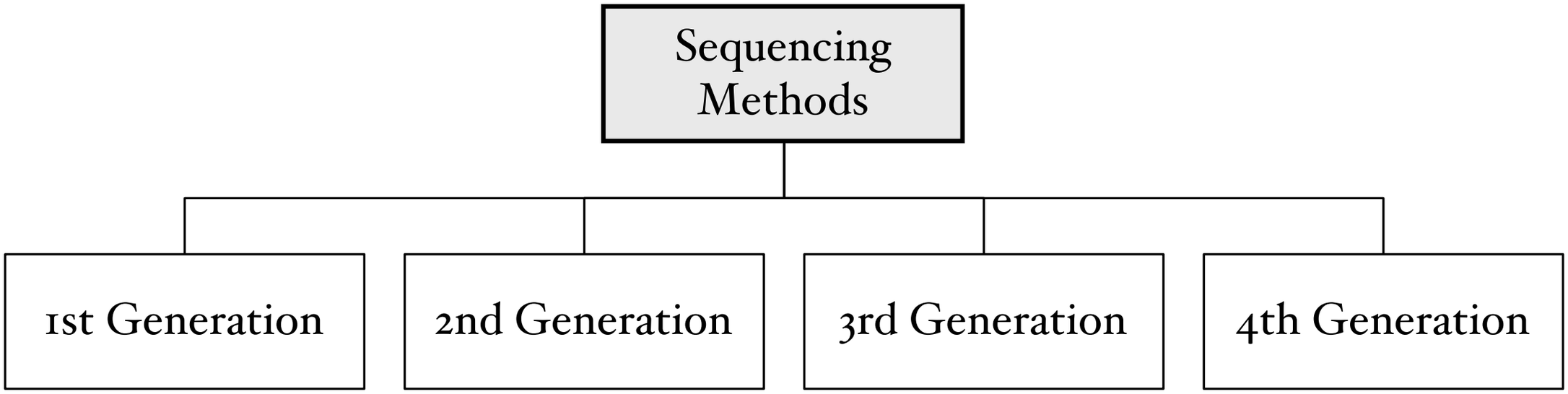
DNA Sequencing

Structure of Genomes
Function of Genomes

Polymorphisms
Whole Genome Sequencing
Exome Sequencing
RNA-Sequencing
Chip-Seq
all you -Seq



Generations of Sequencing Methods



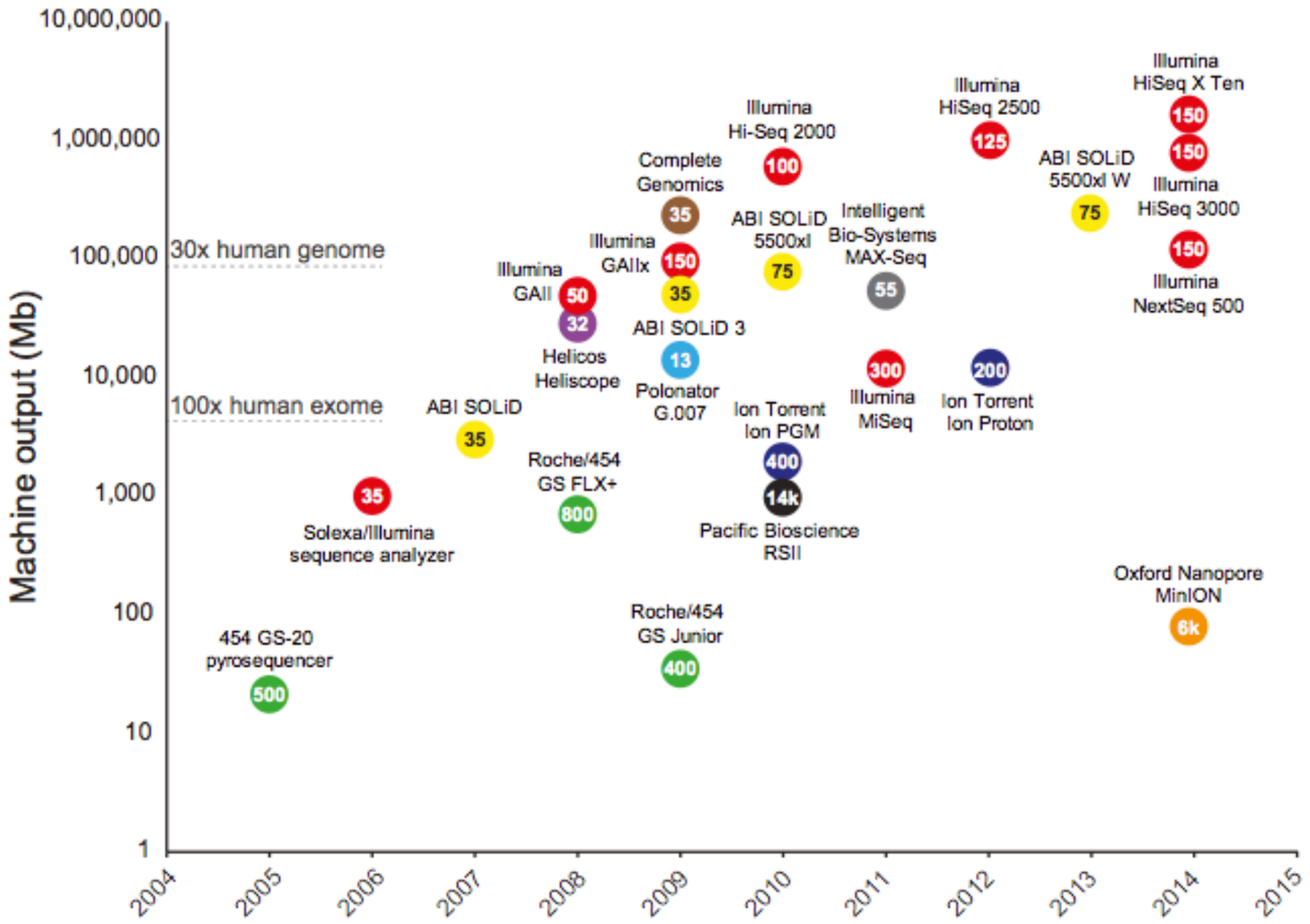
- Maxam-Gilbert (Chemical Degradation)
- Sanger (sequencing by synthesis [SBS])

- Solexa- Illumina (SBS)
- Pacific Biosciences (SBS)
- Ion Torrent (SBS)
- *454-Roche(SBS)*
- ABI SOLiD (ligation based)
- Complete Genomics ([CG] ligation based)
- *Helicos*

- Pacific Biosciences (SBS)

- Nanopores
  - Oxford Nanopore
  - Illumina
  - Genia
  - *NABsys*
  - *Noblege*

Next Gen Sequencing



Reuter et al., Molecular Cell 58, May 21, 2015

Next Gen Sequencing

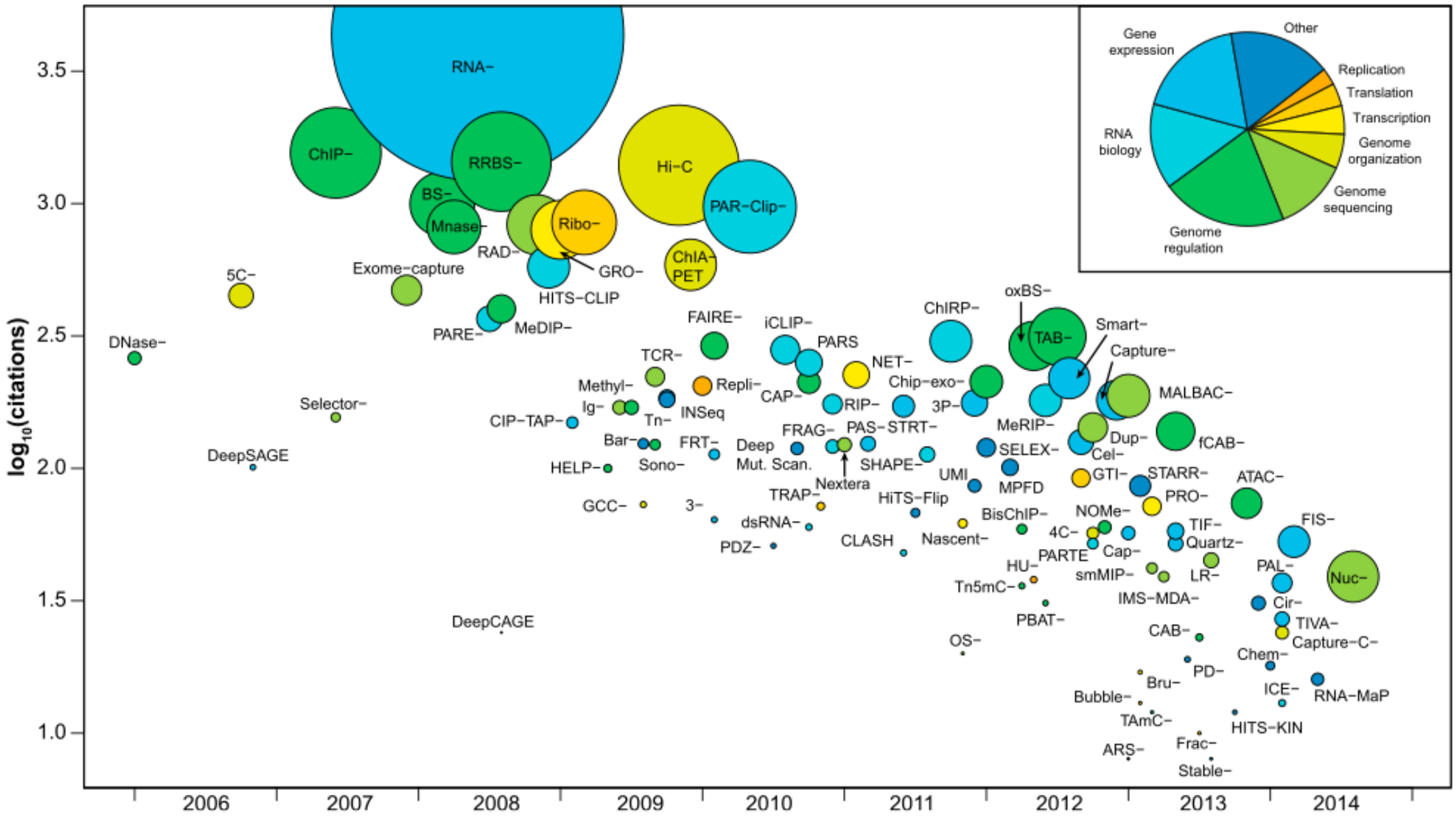
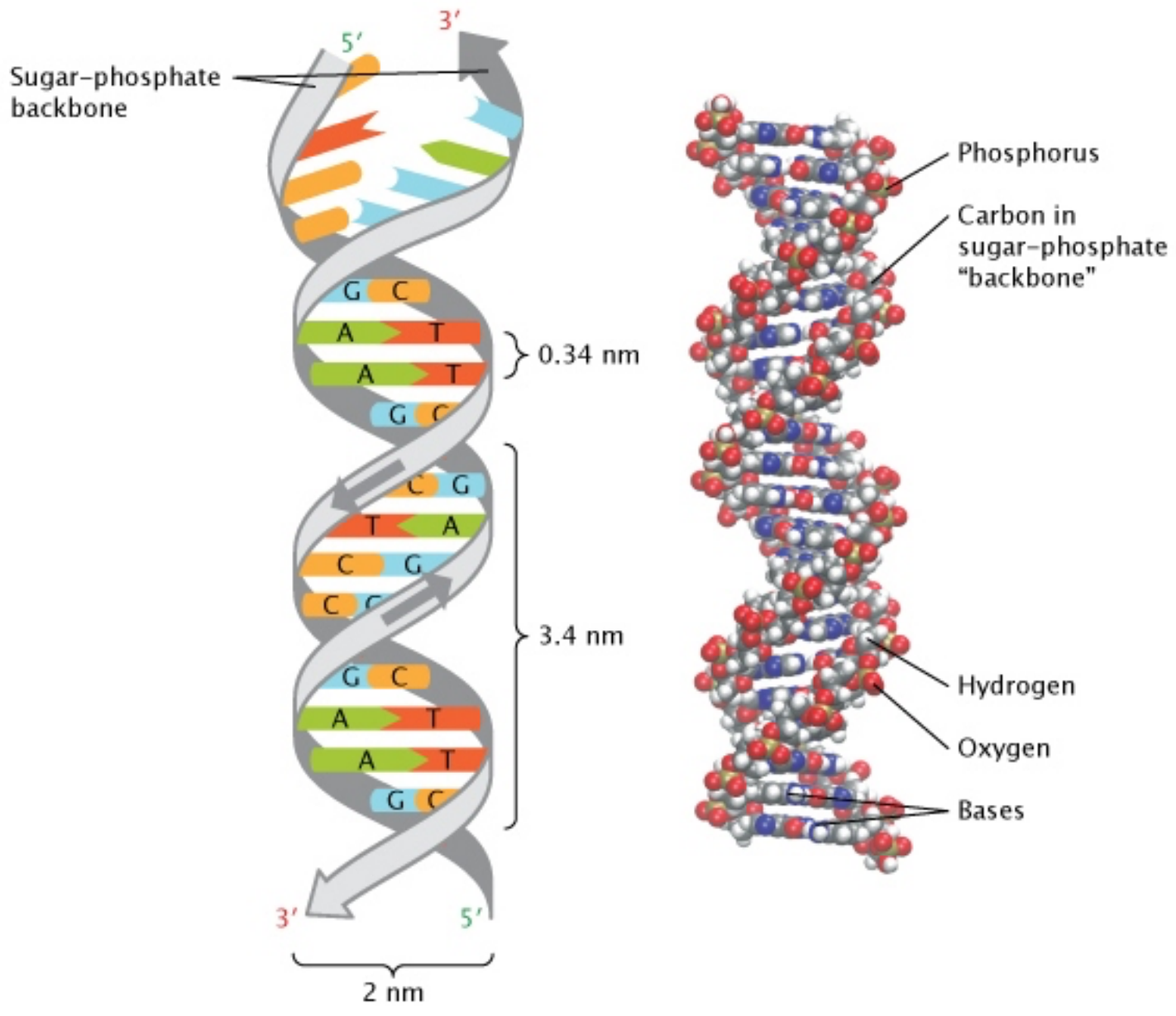


Figure 4. Overview of Selected HTS Applications

Publication date of a representative article describing a method versus the number of citations that the article received. Methods are colored by category, and the size of the data point is proportional to publication rate (citations/months). The inset indicates the color key as well the proportion of methods in each group. For clarity, seq has been omitted from the labels.

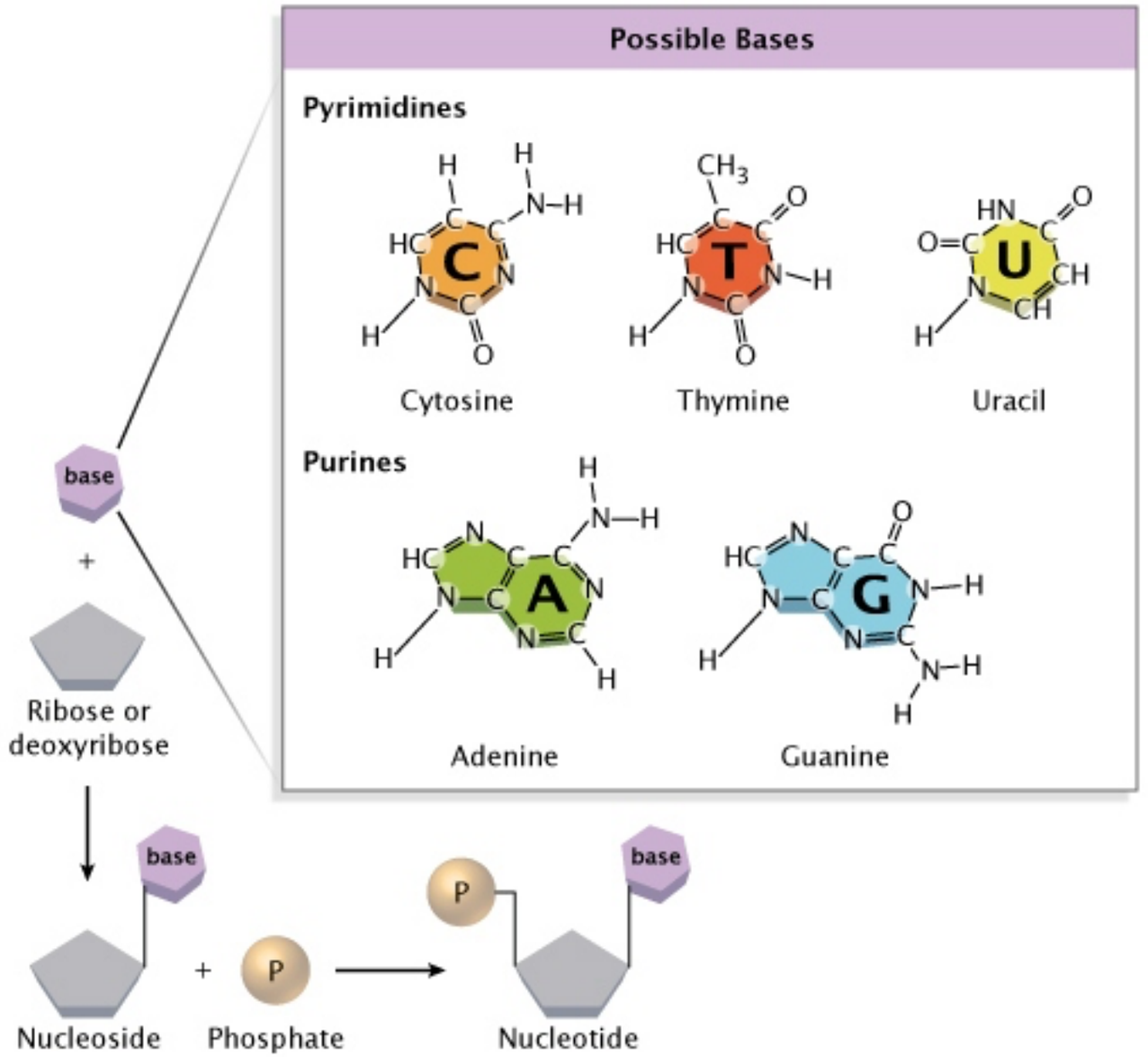
Reuter et al., Molecular Cell 58, May 21, 2015

Structure of DNA



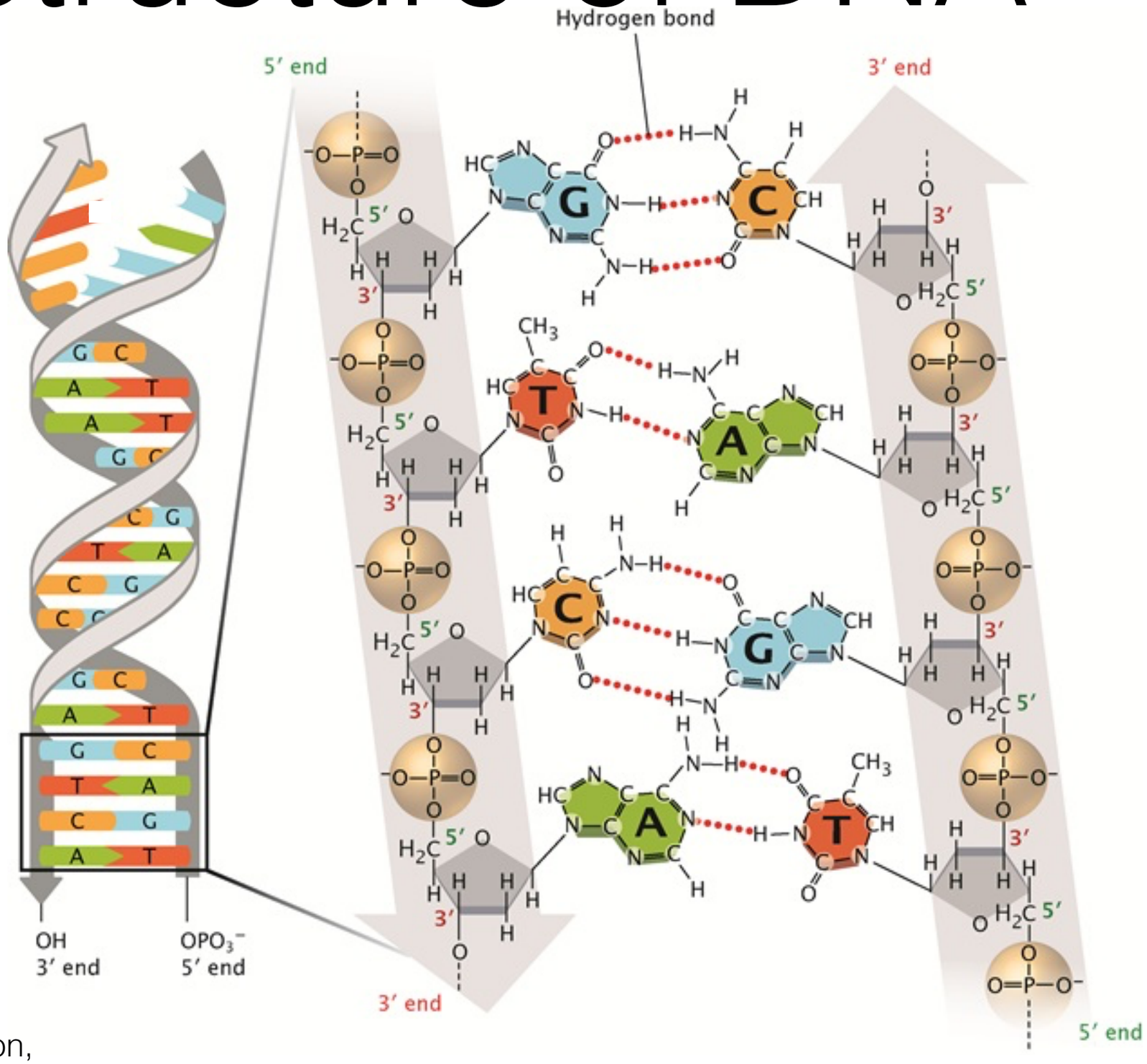
Pray, L. (2008) Discovery of DNA structure and function: Watson and Crick. Nature Education 1(1):100

Structure of DNA



Pray, L. (2008) Discovery of DNA structure and function: Watson and Crick. Nature Education 1(1):100

Structure of DNA



Pray, L. (2008) Discovery of DNA structure and function: Watson and Crick. Nature Education 1(1):100

Maxam-Gilbert

Maxam-Gilbert

A new method for sequencing DNA

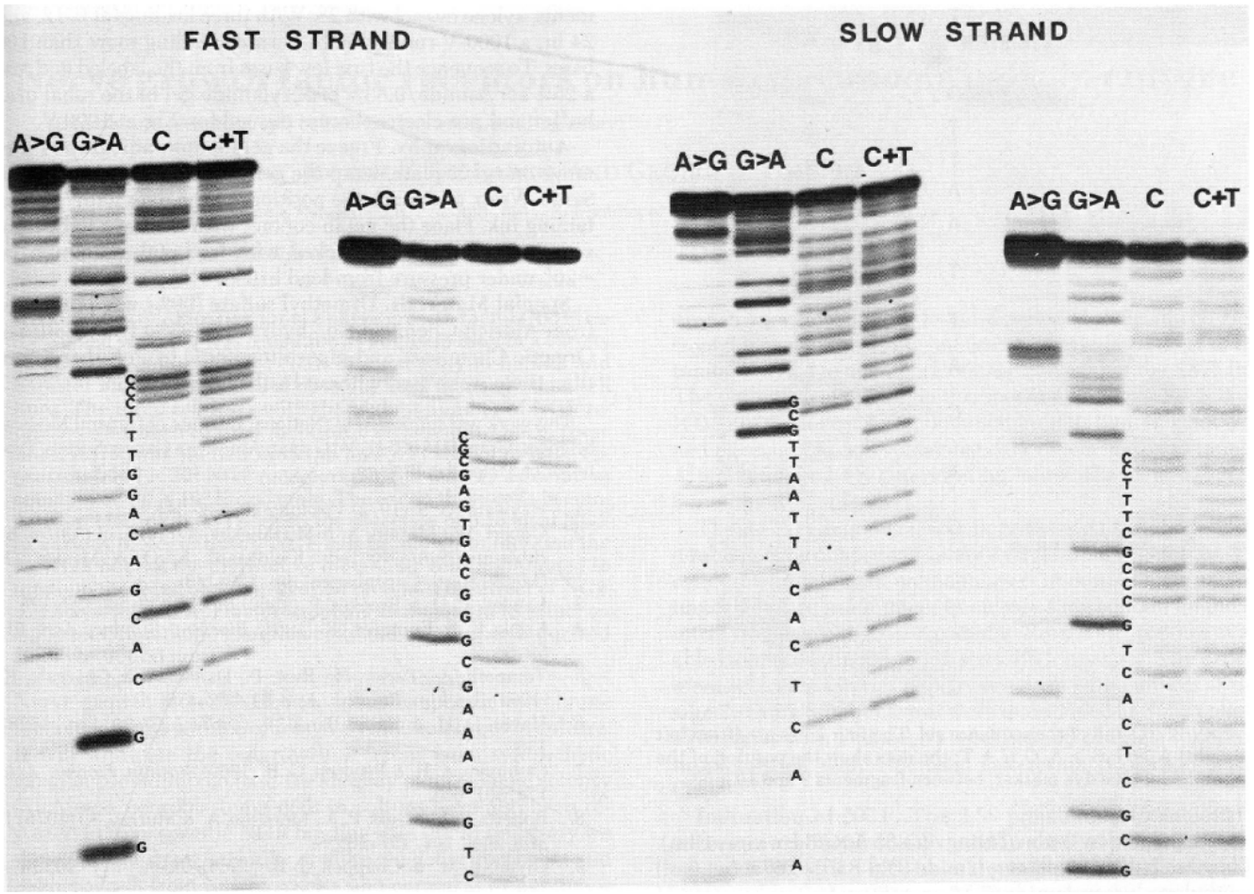
(DNA chemistry/dimethyl sulfate cleavage/hydrazine/piperidine)

ALLAN M. MAXAM AND WALTER GILBERT

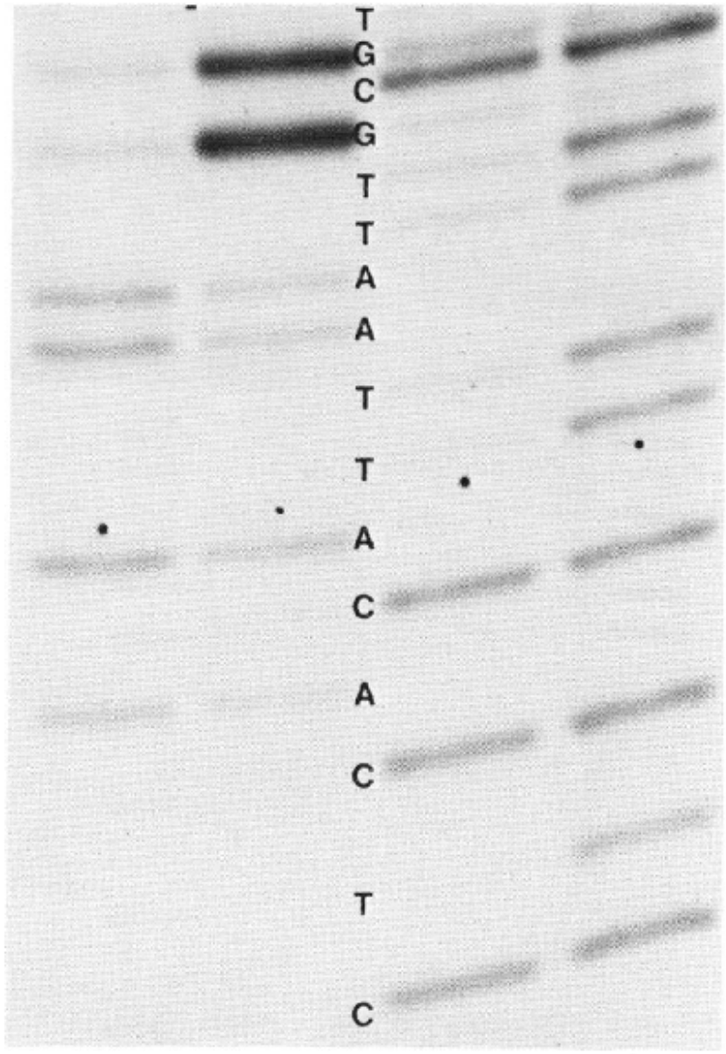
Department of Biochemistry and Molecular Biology, Harvard University, Cambridge, Massachus

Contributed by Walter Gilbert, December 9, 1976

ABSTRACT DNA can be sequenced by a chemical procedure that breaks a terminally labeled DNA molecule partially at each repetition of a base. The lengths of the labeled fragments then identify the positions of that base. We describe reactions that cleave DNA preferentially at guanines, at adenines, at cytosines and thymines equally, and at cytosines alone. When the products of these four reactions are resolved by size, by electrophoresis on a polyacrylamide gel, the DNA sequence can be read from the pattern of radioactive bands. The technique will permit sequencing of at least 100 bases from the point of labeling.

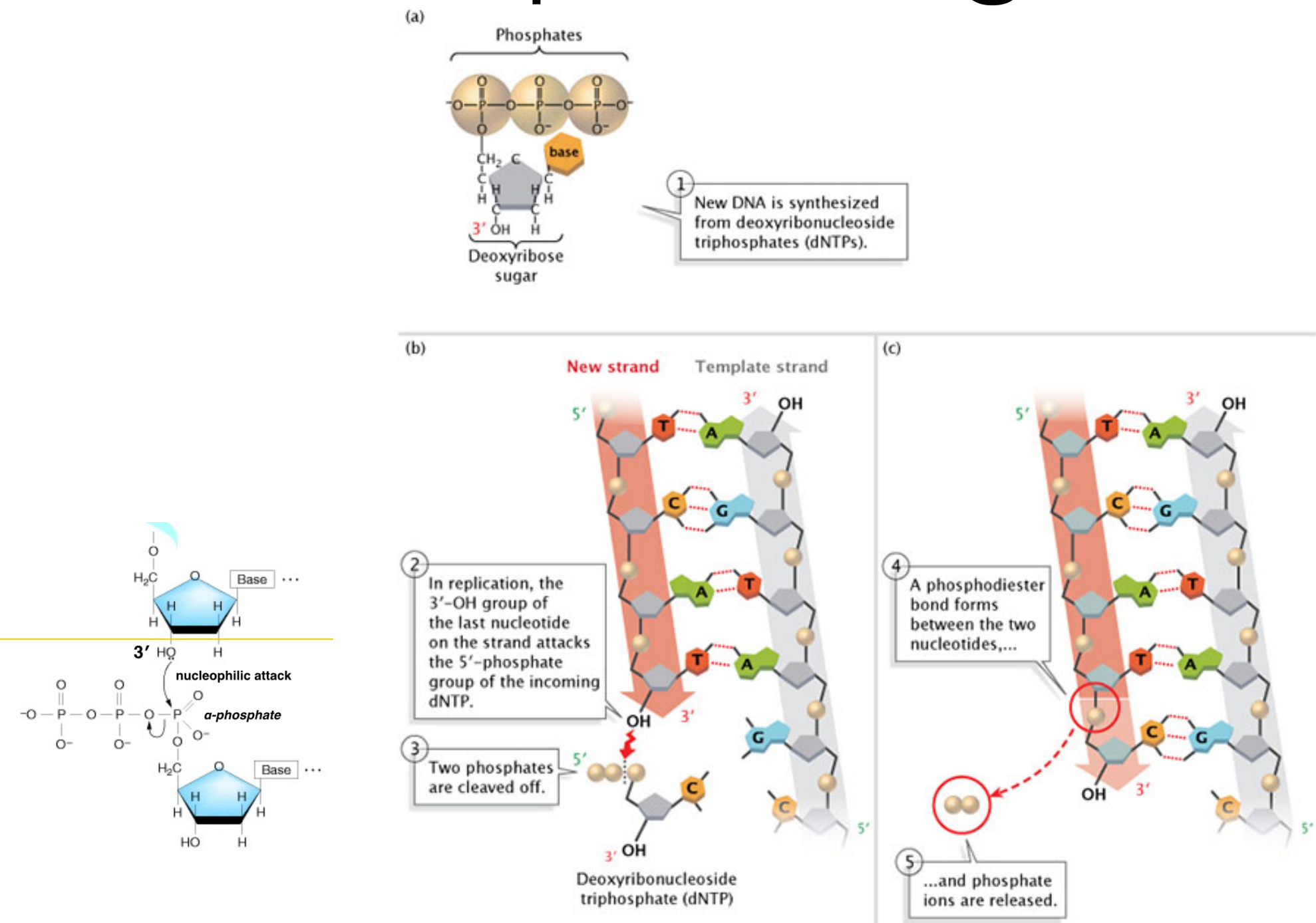


Maxam Gilbert, PNAS Vol.74,No.2,pp.560-564,February1977

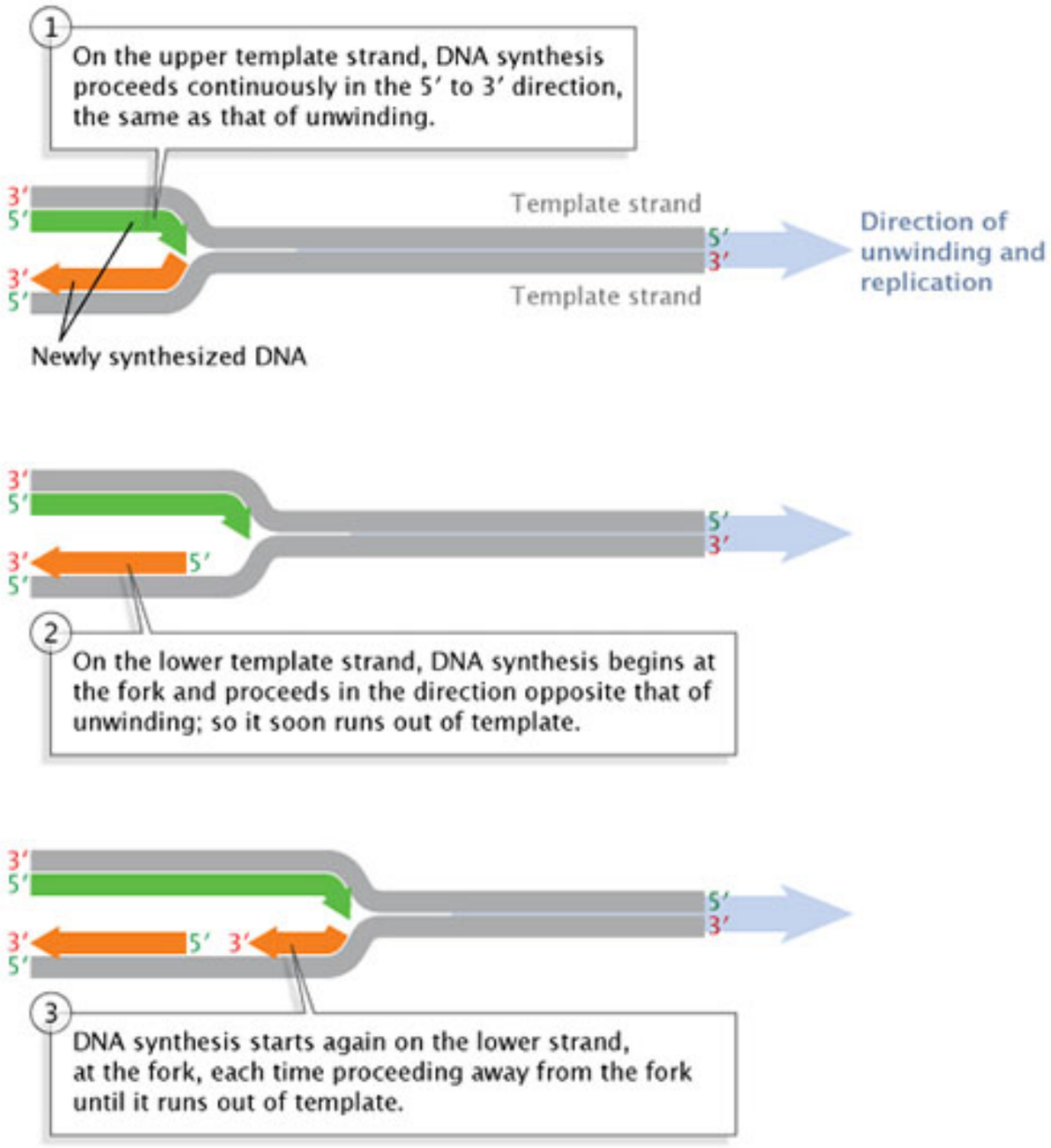


Sanger Sequencing

Replicating DNA



Replicating DNA



An Introduction to Genetic Analysis. 7th edition. Griffiths AJF, Miller JH, Suzuki DT, et al. New York: W. H. Freeman; 2000.

Sanger Sequencing

DNA sequencing with chain-terminating inhibitors

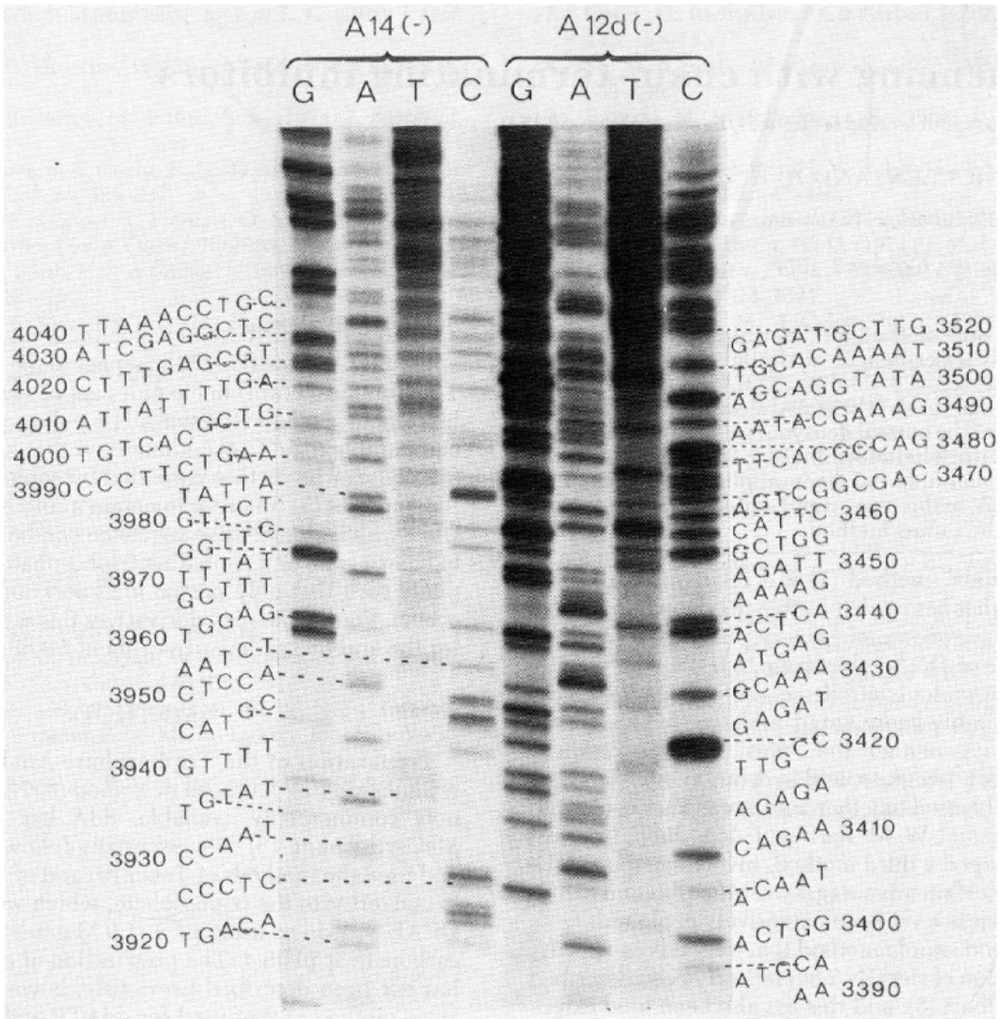
(DNA polymerase/nucleotide sequences/bacteriophage ϕX174)

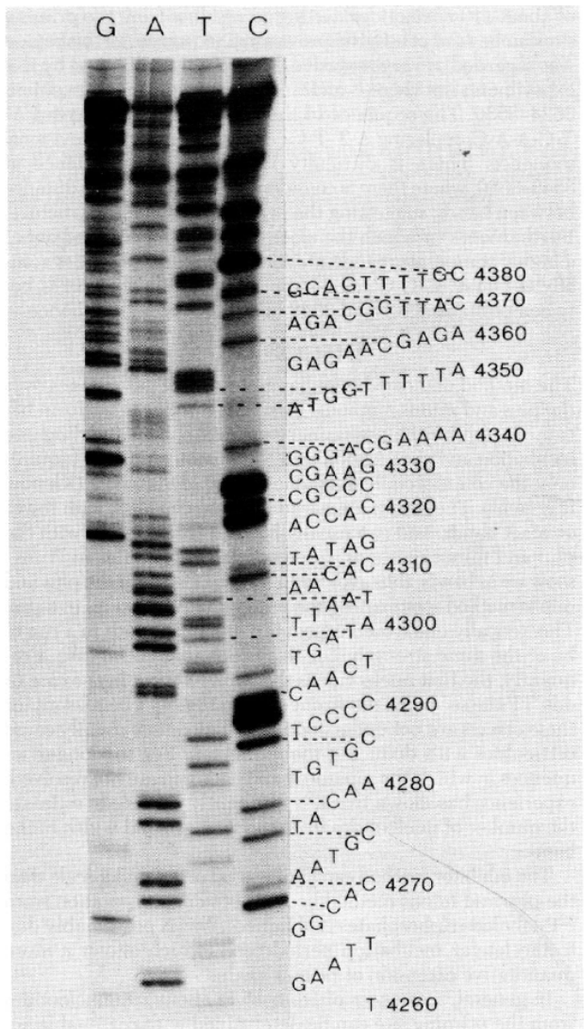
F. SANGER, S. NICKLEN, AND A. R. COULSON

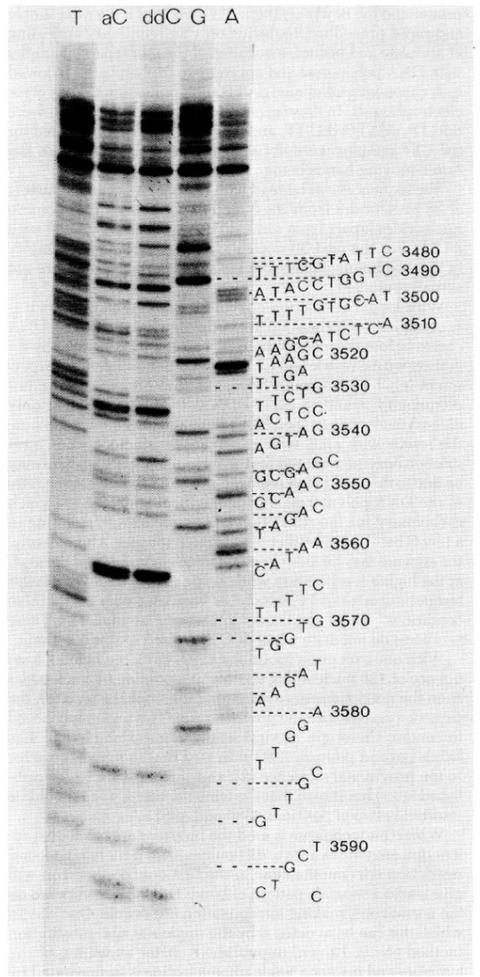
Medical Research Council Laboratory of Molecular Biology, Cambridge CB2 2QH, England

Contributed by F. Sanger, October 3, 1977

ABSTRACT A new method for determining nucleotide sequences in DNA is described. It is similar to the "plus and minus" method [Sanger, F. & Coulson, A. R. (1975) *J. Mol. Biol.* 94, 441–448] but makes use of the 2′,3′-dideoxy and arabinonucleoside analogues of the normal deoxynucleoside triphosphates, which act as specific chain-terminating inhibitors of DNA polymerase. The technique has been applied to the DNA of bacteriophage ϕX174 and is more rapid and more accurate than either the plus or the minus method.

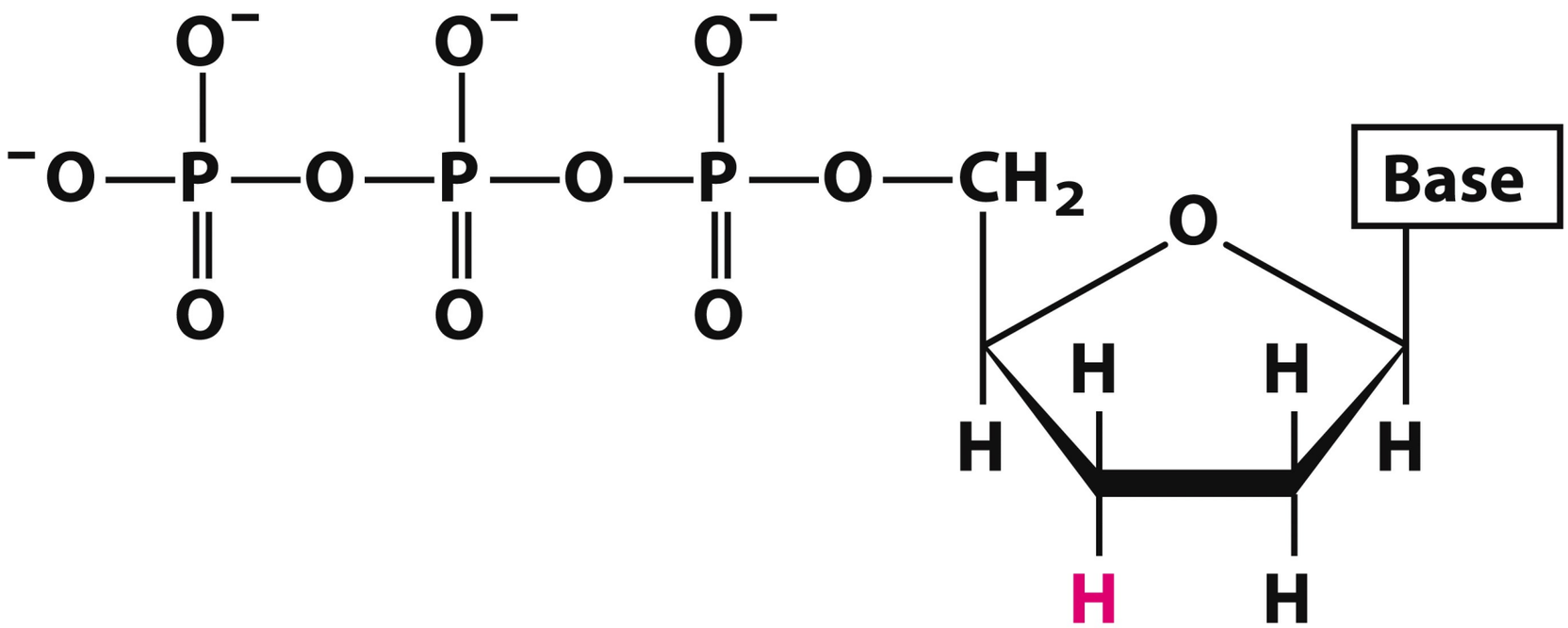






Sanger, Nickten and Coulson, PNAS Vol. 74, No. 12, pp. 5463-5467, December 1977

Chain Termination by ddNTP

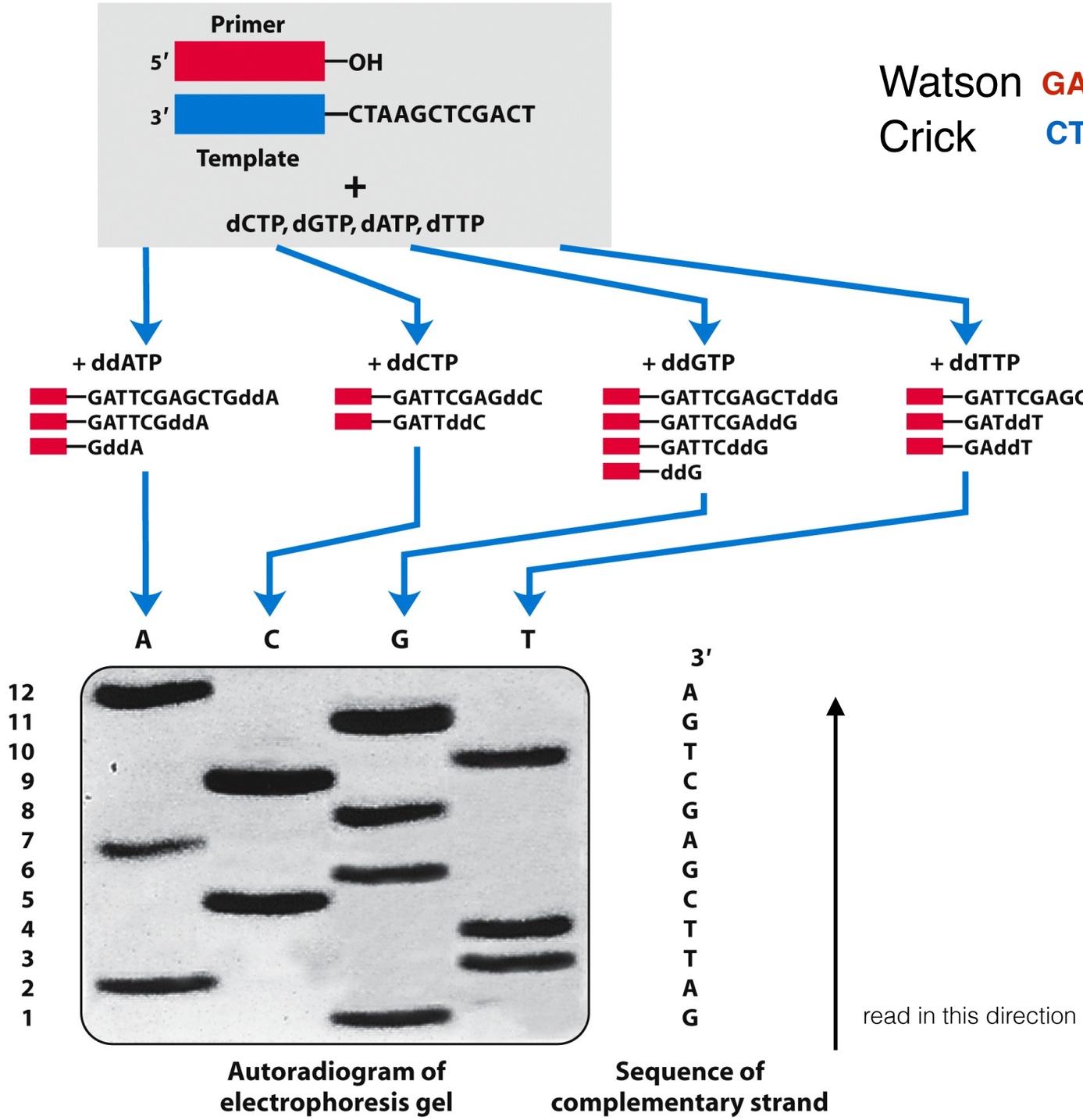


ddNTP analog

Figure 8-33b
Lehninger Principles of Biochemistry, Sixth Edition

Will terminate sequence as dideoxy cannot be extended

Sanger Sequencing



Watson GATTCGAGCTGA

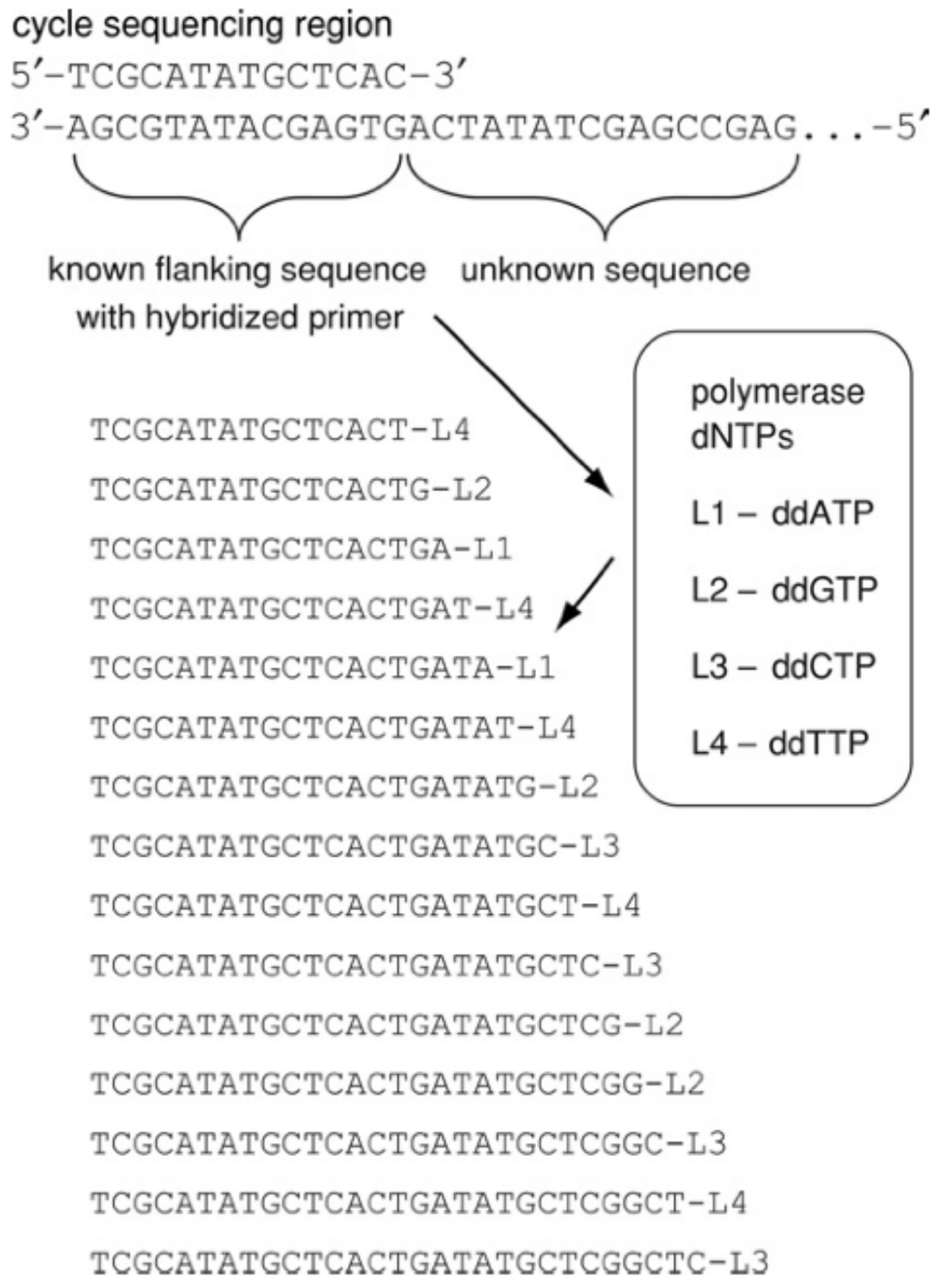
Crick CTAAGCTCGACT

radioactively marked nucleotides

Figure 8-33c
Lehninger Principles of Biochemistry, Sixth Edition

Automated Sanger Sequencing

Watson
Crick



- Cycle sequencing (multiple cycles of primer annealing, primer extension, and denaturation) are performed with polymerase, dNTPs, and **fluorescently** labeled ddNTPs (where a different label is present on each species of ddNTP).

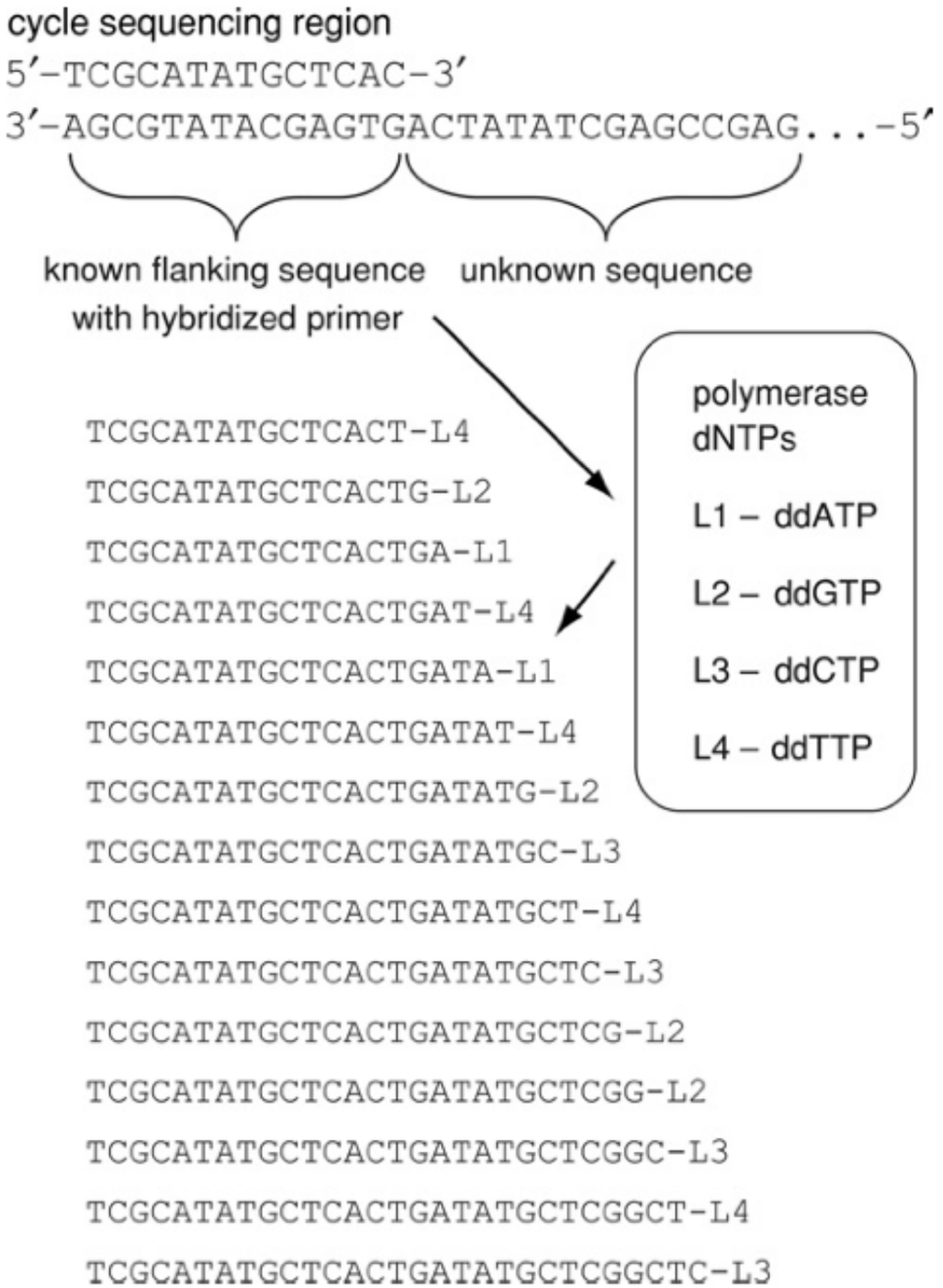
- Products of the cycle sequencing reaction are run into a capillary containing a denaturing polymer. This yields size-based separation with single-base-pair resolution, with the shortest fragments running the fastest.

- Observation of the emission spectra in four channels (corresponding to the fluorescent labels for the four ddNTP species) over time, as fragments emerge from capillary electrophoresis, can be used to infer the primary sequence of the unknown template.

Shendure, Porreca, and Church, Curr. Protoc. Mol. Biol. 81:7.1.1-7.1.11.

Automated Sanger Sequencing

Watson
Crick



Shendure, Porreca, and Church, Curr. Protoc. Mol. Biol. 81:7.1.1-7.1.11.

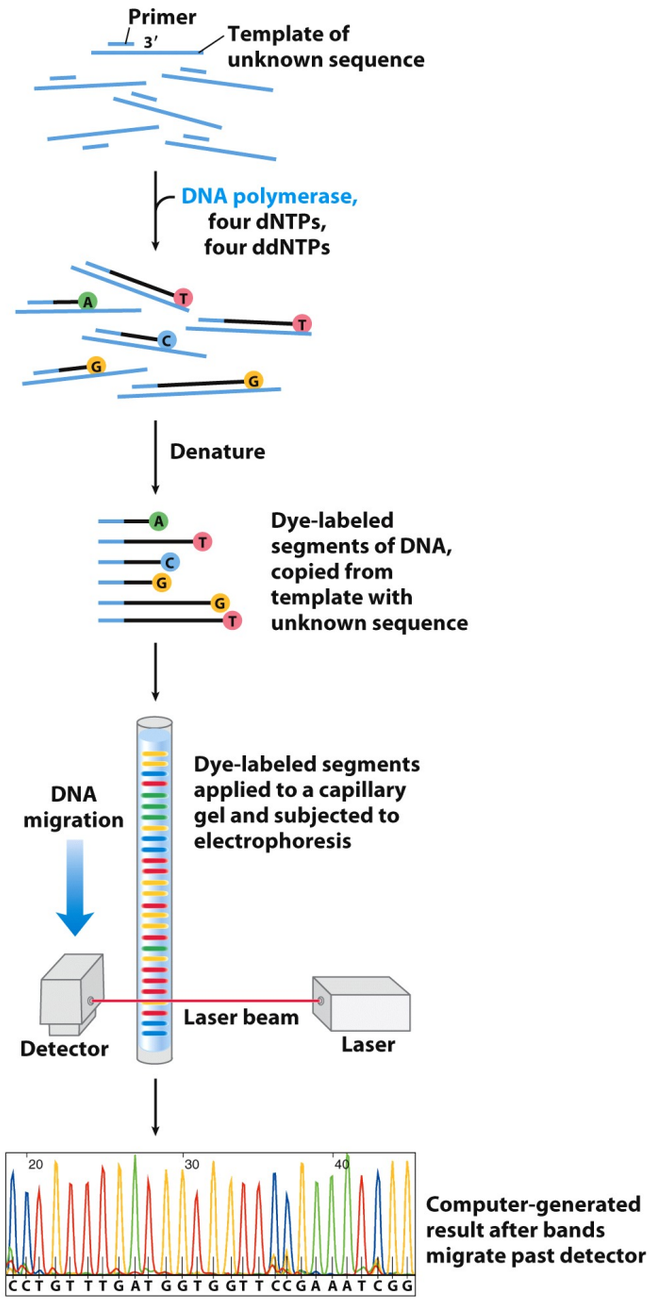
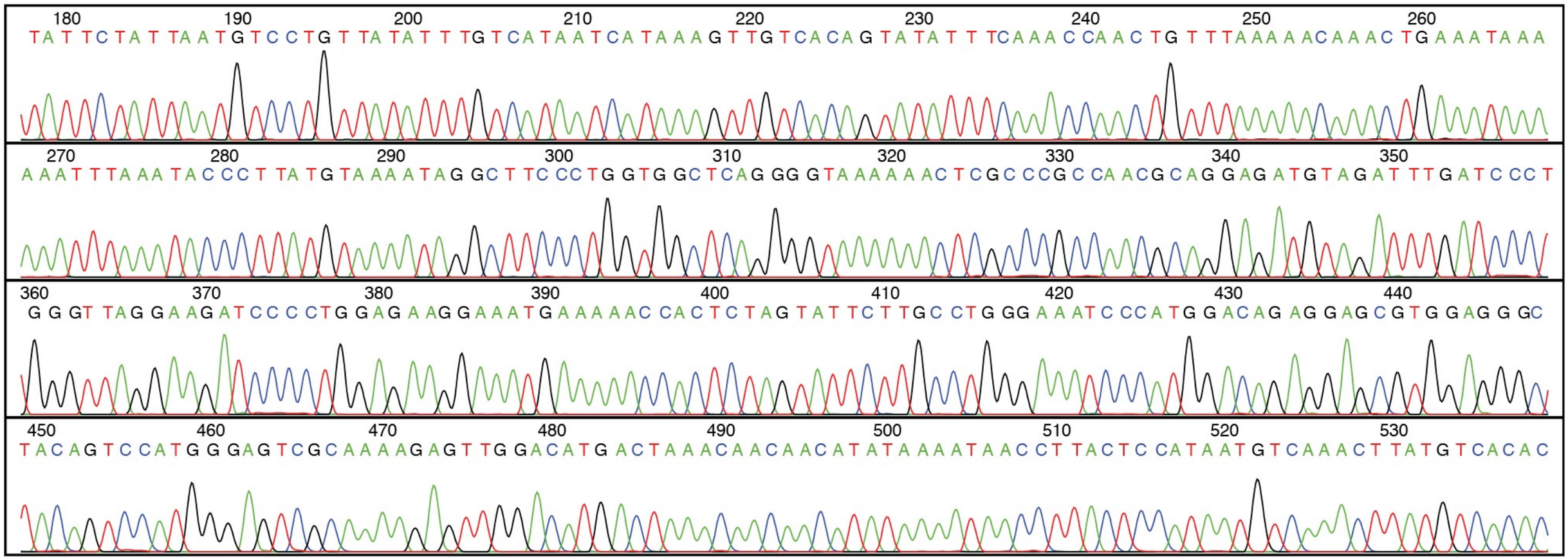


Figure 8-34
Lehninger Principles of Biochemistry, Sixth Edition

Automated Sanger Sequencing



DNA sequence data from an automated sequencing machine

http://www.genome.gov/dmd/img.cfm?node=Photos/Graphics&id=85152 Credit: Darryl Leja, NHGRI

Automated Sanger Sequencing

	310	3130/3130*xl*	3500/3500xL	3730/3730*xl*
Sequencing				
Sequencing read length (bp)	up to 600	up to 950	up to 850	up to 900
Minimum run time	38 minutes	35 minutes	30 minutes	20 minutes
Maximum sequencing throughput (bases pair reads/day)	15 k	78 k (3130), 315 k (3130*xl*)	138 k (3500), 414 k (3500xL)	1.3 M (3130), 2.6 M (3130*xl*)
Sequencing reagents	BigDye® Terminator v1.1 BigDye® Terminator v3.1	BigDye® Direct BigDye® Terminator v1.1 BigDye® Terminator v3.1	BigDye® Direct BigDye® Terminator v1.1 BigDye® Terminator v3.1	BigDye® Direct BigDye® Terminator v1.1 BigDye® Terminator v3.1
Fragment analysis				
Minimum run time	30 minutes	20 minutes	30 minutes	20 minutes
Fragment throughput (bp)	up to 920	up to 3 k (3130), 12 k (3130*xl*)	up to 5 k (3500), 16 k (3500xL)	up to 10 k (3730), 21 k (3730*xl*)
Fragment analysis reagents	SNaPshot® Multiplex Kit GeneScan™ Size Standards	SNaPshot® Multiplex Kit GeneScan™ Size Standards	SNaPshot® Multiplex Kit GeneScan™ Size Standards	SNaPshot® Multiplex Kit GeneScan™ Size Standards

AB Capillary Sequencers
Life Technologies
Thermo Scientific

~ Million bp/day

8-h work day 1920 samples 1.2 megabases
24-h work day 4600 samples 2.8 megabases

MegaBACE Systems
Amersham/GE

LI-COR

limited new instrument availability

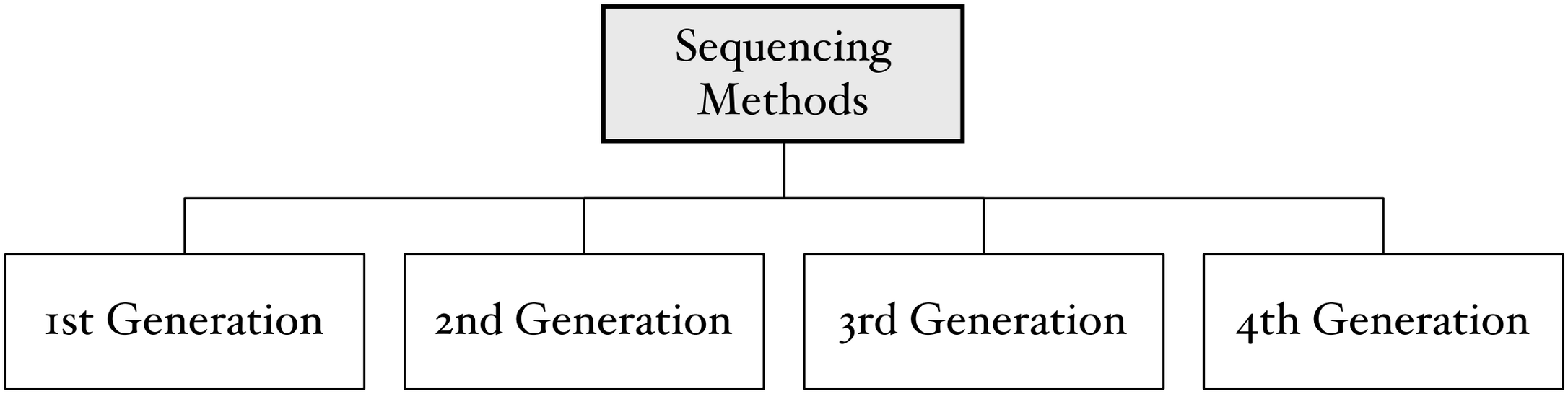
Next Generation
Whole Genome Sequencing

- *Additional Coverage*
- *More and longer reads needed for assembly*
- *Depends on complexity and genome size*

- No more cloning - nanotechnologies
- No more electrophoresis
- Initially short reads - now longer as well
- Massively Parallel
- Digital readout

Next Generation sequencing methods (2nd gen) use PCR amplification. Possible PCR Artifacts:
Exactly duplicated reads
Preferential amplification of low complexity samples
Introduces polymerase substitution errors
CG Bias in bridge amplification

Generations of Sequencing Methods



- Maxam-Gilbert (Chemical Degradation)
- Sanger (sequencing by synthesis [SBS])

- Solexa- Illumina (SBS)
- Pacific Biosciences (SBS)
- Ion Torrent (SBS)
- *454-Roche(SBS)*
- ABI SOLiD (ligation based)
- Complete Genomics ([CG] ligation based)
- *Helicos*

- Pacific Biosciences (SBS)

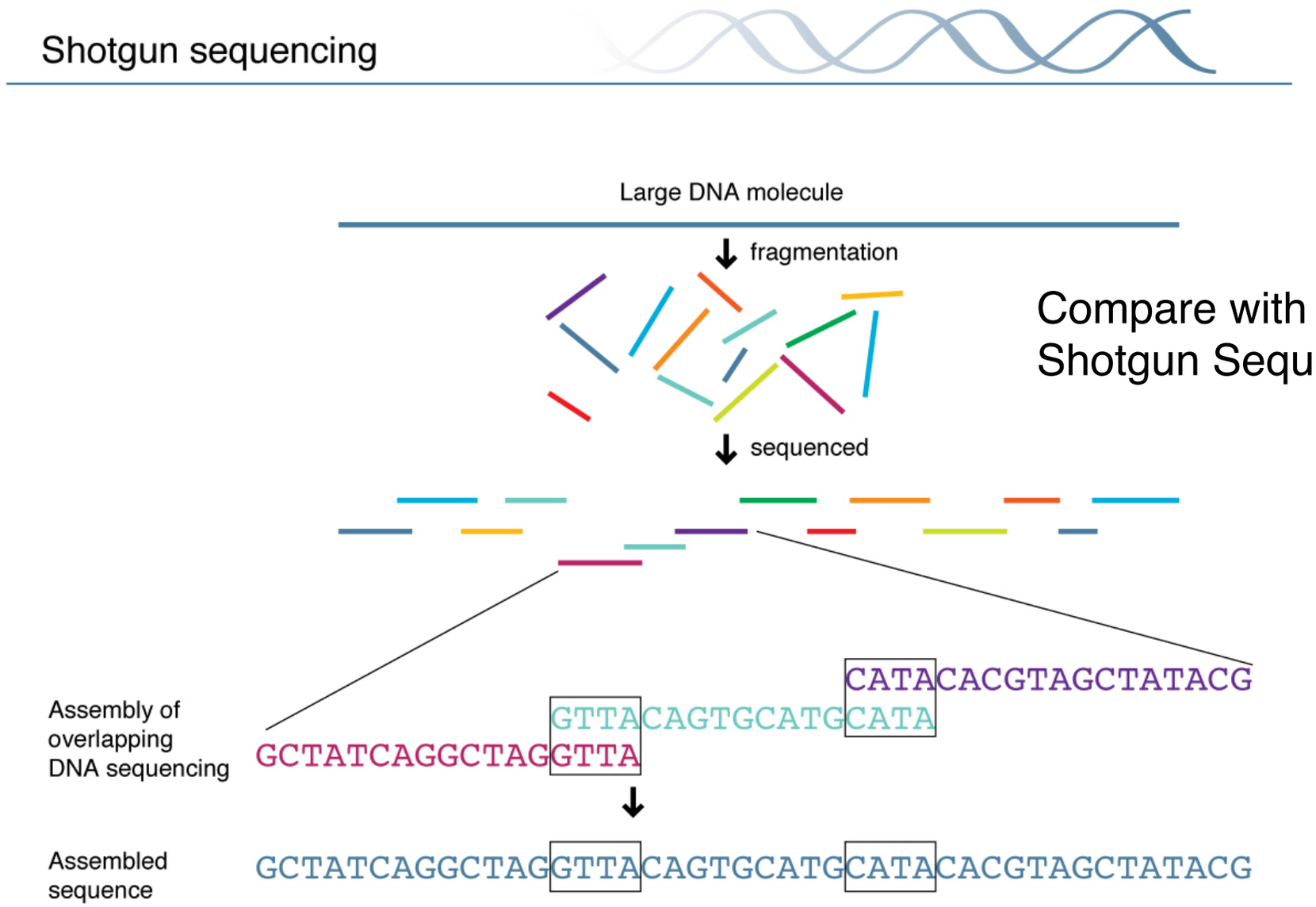
- Nanopores
  - Oxford Nanopore
  - Illumina
  - Genia
  - *NABsys*
  - *Noblege*

Some Glossary Terms

- **Read.** A data string of A,T, C, and G bases corresponding to the sample DNA.
- **Contigs.** A stretch of continuous sequence, in silico, generated by aligning overlapping sequencing reads.
- **Coverage.** The average number of sequenced bases that align to each base of the reference DNA. For example, a whole genome sequenced at 30× coverage means that, on average, each base in the genome was sequenced 30 times.
- **Gb**. Gigabase, or one billion nucleotides. The Gb required for a given application can vary by the size of the genome (e.g. human genome versus a microbe).
- **Library**. A collection of DNA fragments with adapters ligated to each end.
- **Long reads**. Sequence reads that generally are at least 400 base pairs long in a single direction. Long reads make nucleic acid scaffold construction from subsequences an easier bioinformatic task.
- **Mapped read depth**. The total number of bases sequenced and aligned at a given reference base position.
- **Reference genome.** A reference genome is a fully sequenced and assembled genome that acts as a scaffold against which new sequence reads are aligned and compared. Typically, reads generated from a sequencing run are aligned to a reference genome as a rst step in data analysis. In the absence of a reference genome, the newly sequenced reads must be constructed by contig assembly (de novo sequencing).

www.illumina.com/technology/next-generation-sequencing.html

Shotgun Sequencing



Compare with Hierarchical Shotgun Sequencing with BAC

Talking Glossary of Genetic Terms
NATIONAL HUMAN GENOME RESEARCH INSTITUTE
NATIONAL INSTITUTES OF HEALTH | genome.gov
Illustration by Darryl Leja, NHGRI

Illumina

Illumina

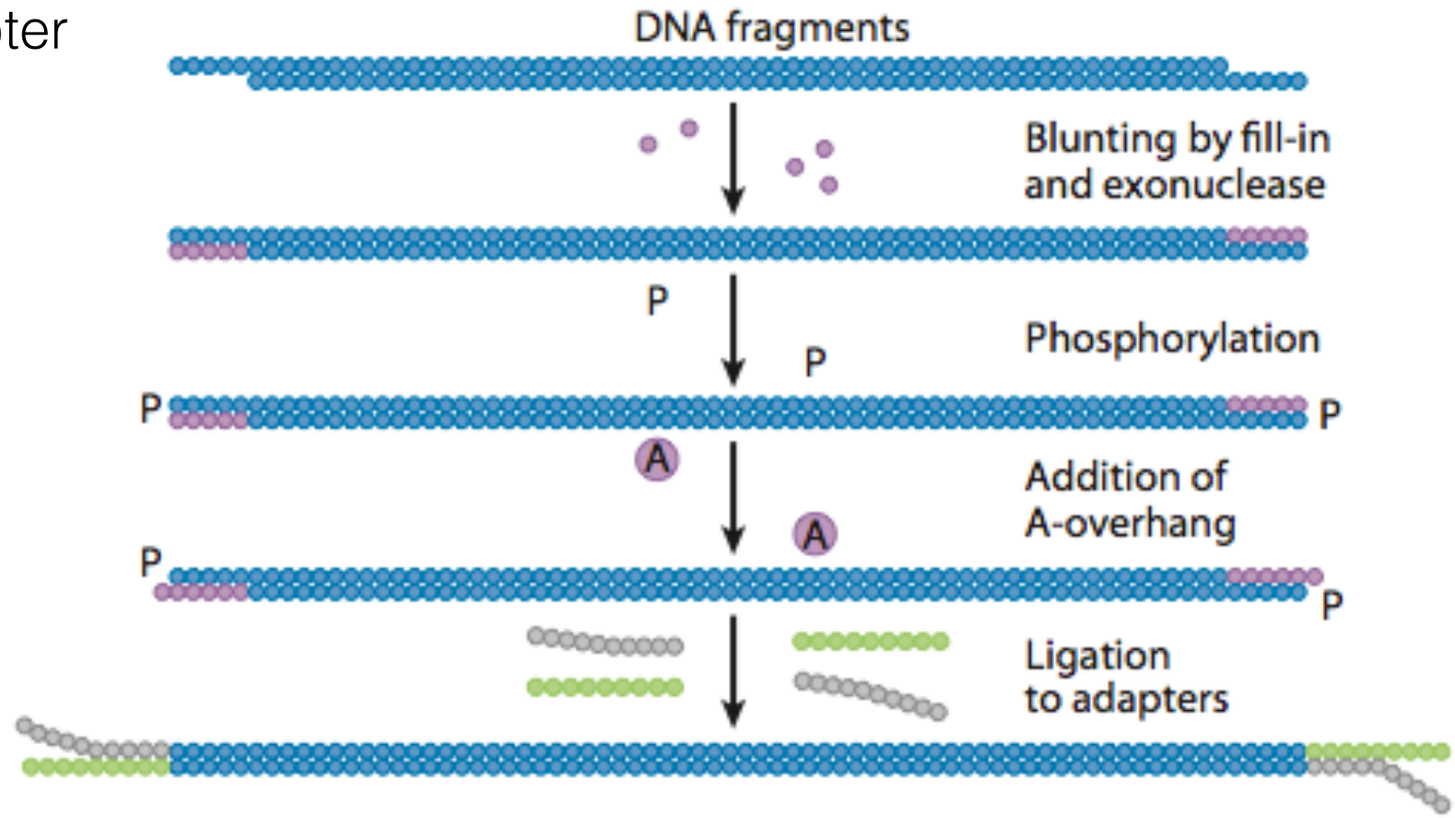
	MiniSeq	MiSeq	NextSeq	HiSeq		HiSeq X	
				2500	*4000*	*X Five*	*X Ten*
Output per Run:	7.5 Gb	15 Gb	120 Gb	1 Tb	1.5 Tb	1.8 Tb	1.8 Tb

illumina.com

Step 1 of 4

Illumina

1. Library Preparation—The sequencing library is prepared by random fragmentation of the DNA or cDNA sample, followed by 5’ and 3’ adapter ligation.



Mardis ER. Next-generation sequencing platforms. Annu Rev Anal Chem 2013;6:287-303.

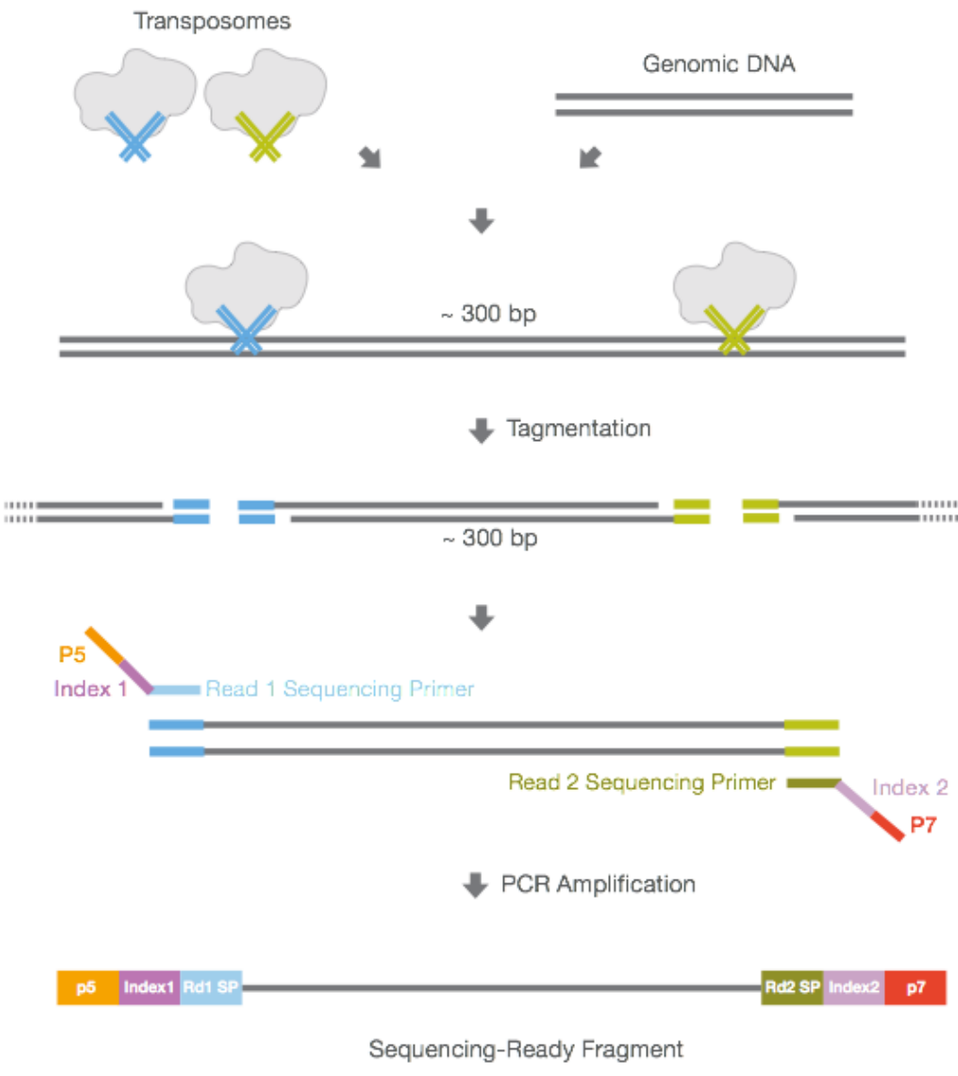
Step 1 of 4

Illumina

Alternatively, “tagmentation” combines the fragmentation and ligation reactions into a single step that greatly increases the efficiency of the library preparation process.

Adapter-ligated fragments are then PCR amplified and gel purified.

Figure 2: Nextera Sample Preparation Biochemistry



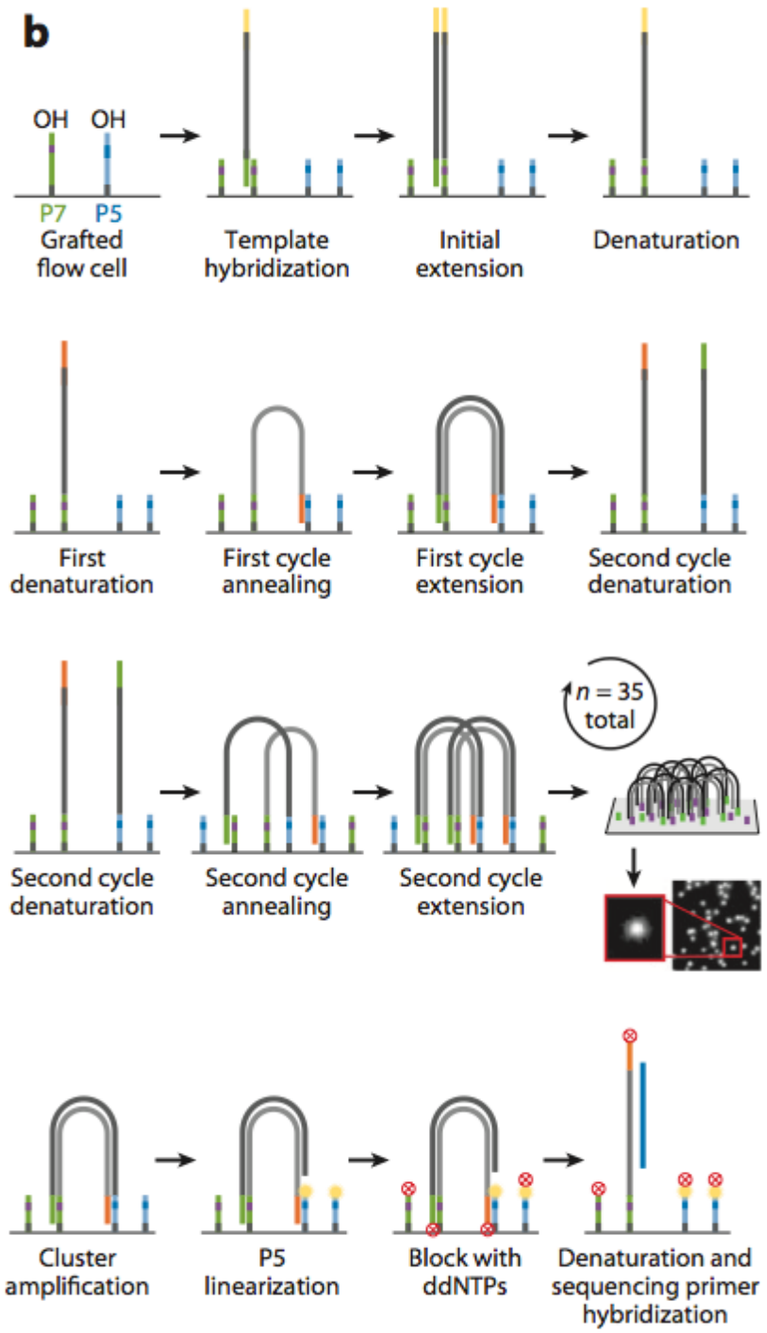
Nextera chemistry simultaneously fragments and tags DNA in a single step. A simple PCR amplification then appends sequencing adapters and sample indices to each fragment.

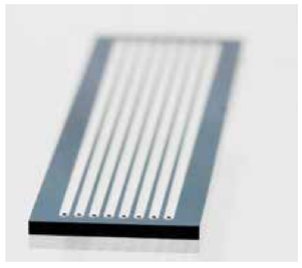
Step 2 of 4

Illumina

2. Cluster Generation—For cluster generation, the library is loaded into a flow cell where fragments are captured on a lawn of surface-bound oligos complementary to the library adapters.

Each fragment is then amplified into distinct, clonal clusters through bridge amplification. When cluster generation is complete, the templates are ready for sequencing.





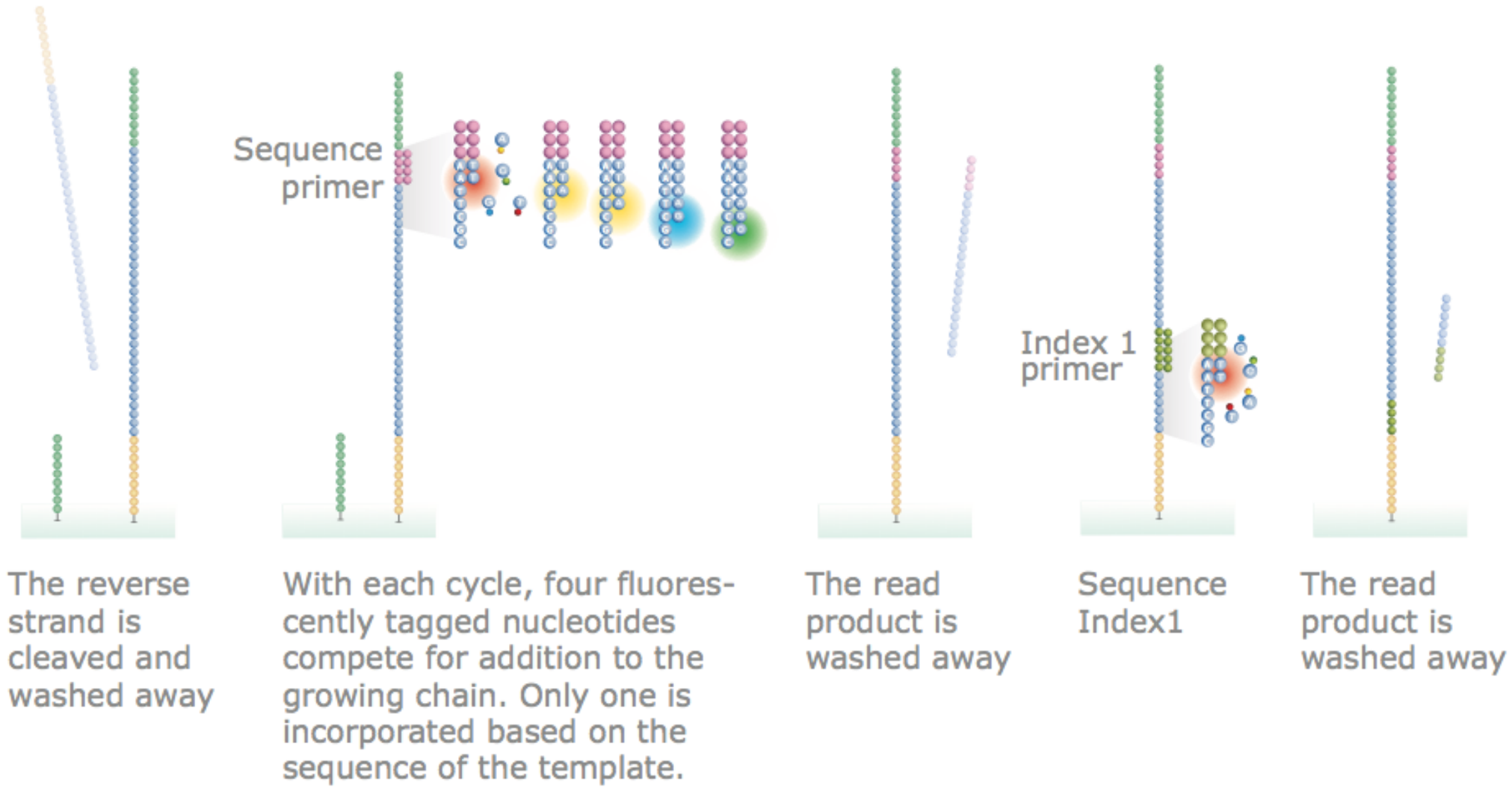
- **Flow cell** A glass slide with 1, 2, or 8 physically separated lanes, depending on instrument platform.
- Each lane is coated with a lawn of surface bound, adapter-complimentary oligos.
- A single library or a pool of up to 96 multiplexed libraries can be run per lane depending on application parameters.

Mardis ER. Next-generation sequencing platforms. Annu Rev Anal Chem 2013;6:287-303.

Step 3 of 4

Illumina

3. Sequencing—Illumina SBS technology utilizes a proprietary reversible terminator–based method that detects single bases as they are incorporated into DNA template strands. All 4 reversible terminator-bound dNTPs are present during each sequencing cycle.

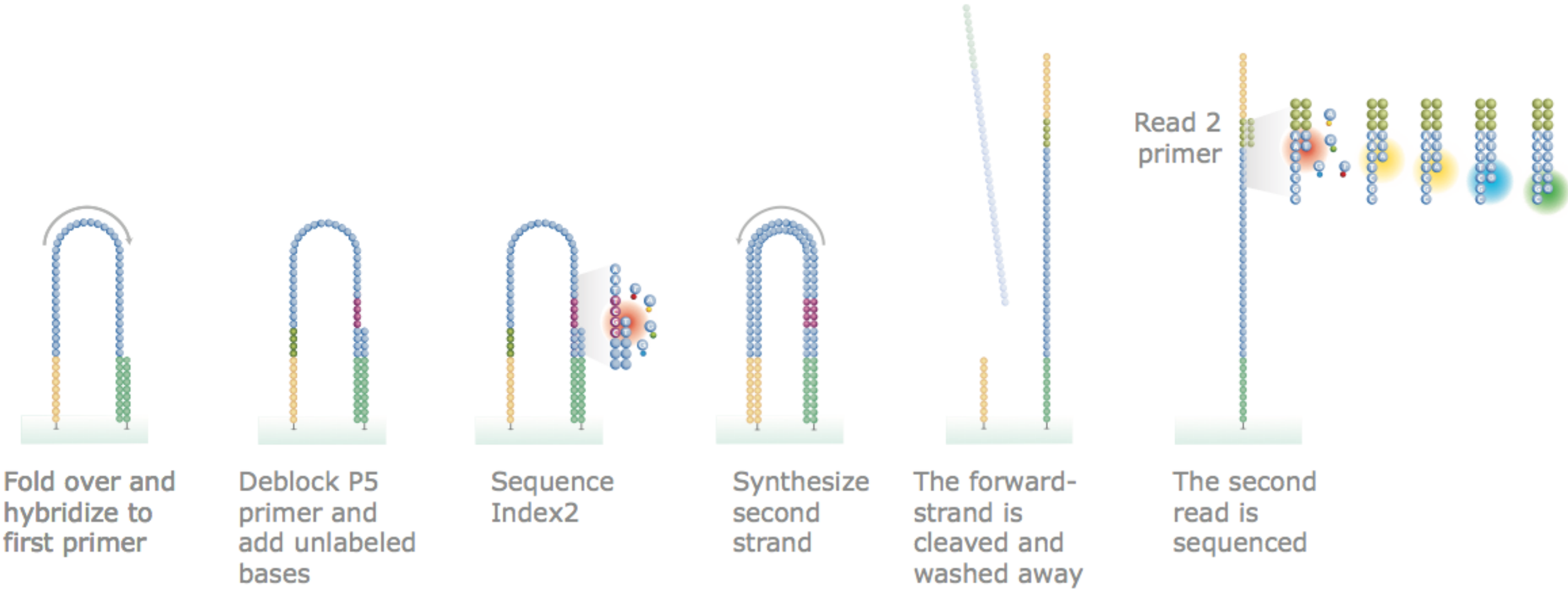


http://www.illumina.com/libraryprepmethods

Step 3 of 4

Illumina

3. Sequencing—Illumina SBS technology utilizes a proprietary reversible terminator–based method that detects single bases as they are incorporated into DNA template strands. All 4 reversible terminator-bound dNTPs are present during each sequencing cycle.



http://www.illumina.com/libraryprepmethods

Step 3 of 4

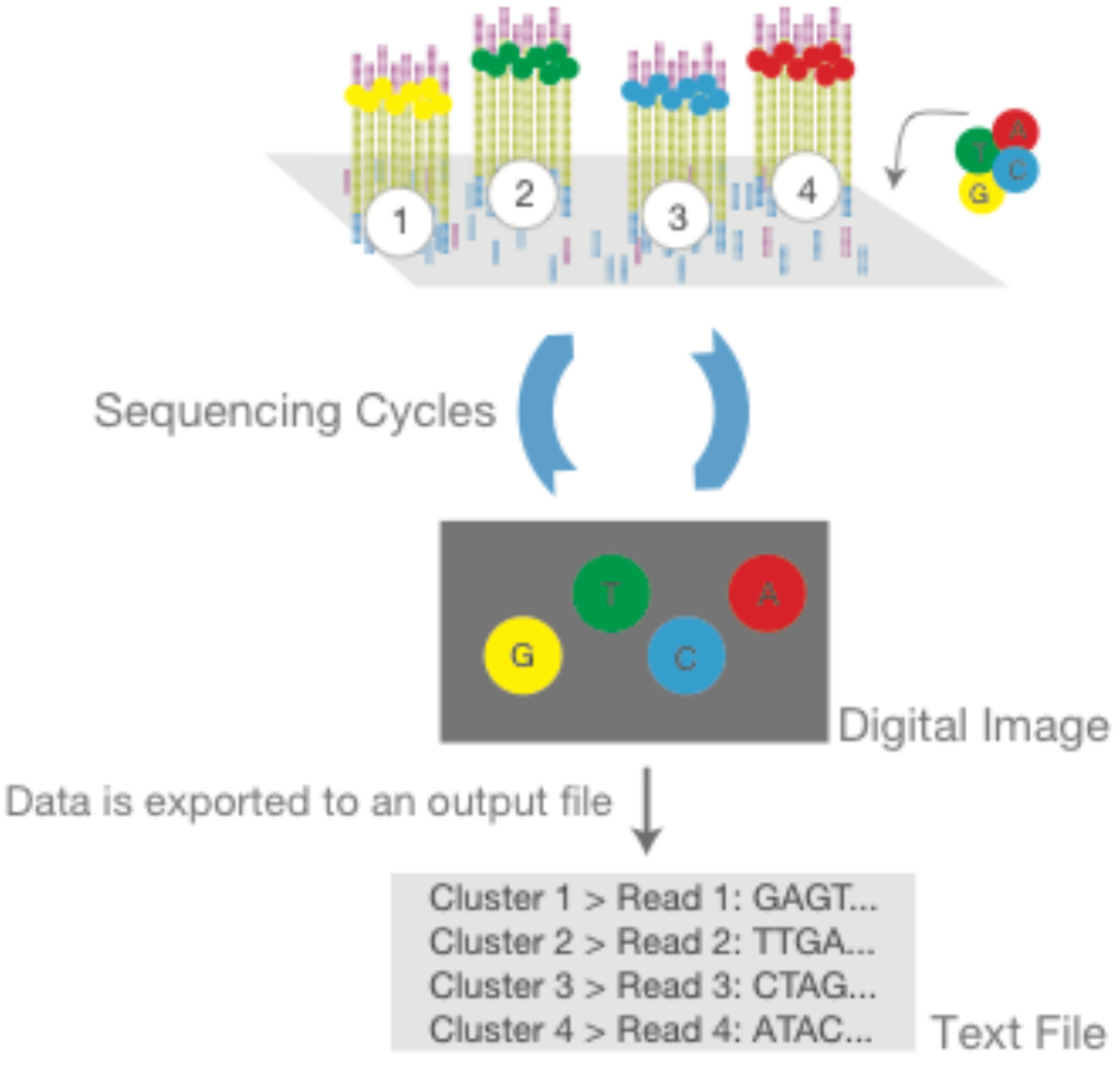
Illumina

Sequencing reagents, including fluuorescently labeled nucleotides, are added and the first base is incorporated.

The flow cell is imaged and the emission from each cluster is recorded.

The emission wavelength and intensity are used to identify the base.

This cycle is repeated “n” times to create a read length of “n” bases.



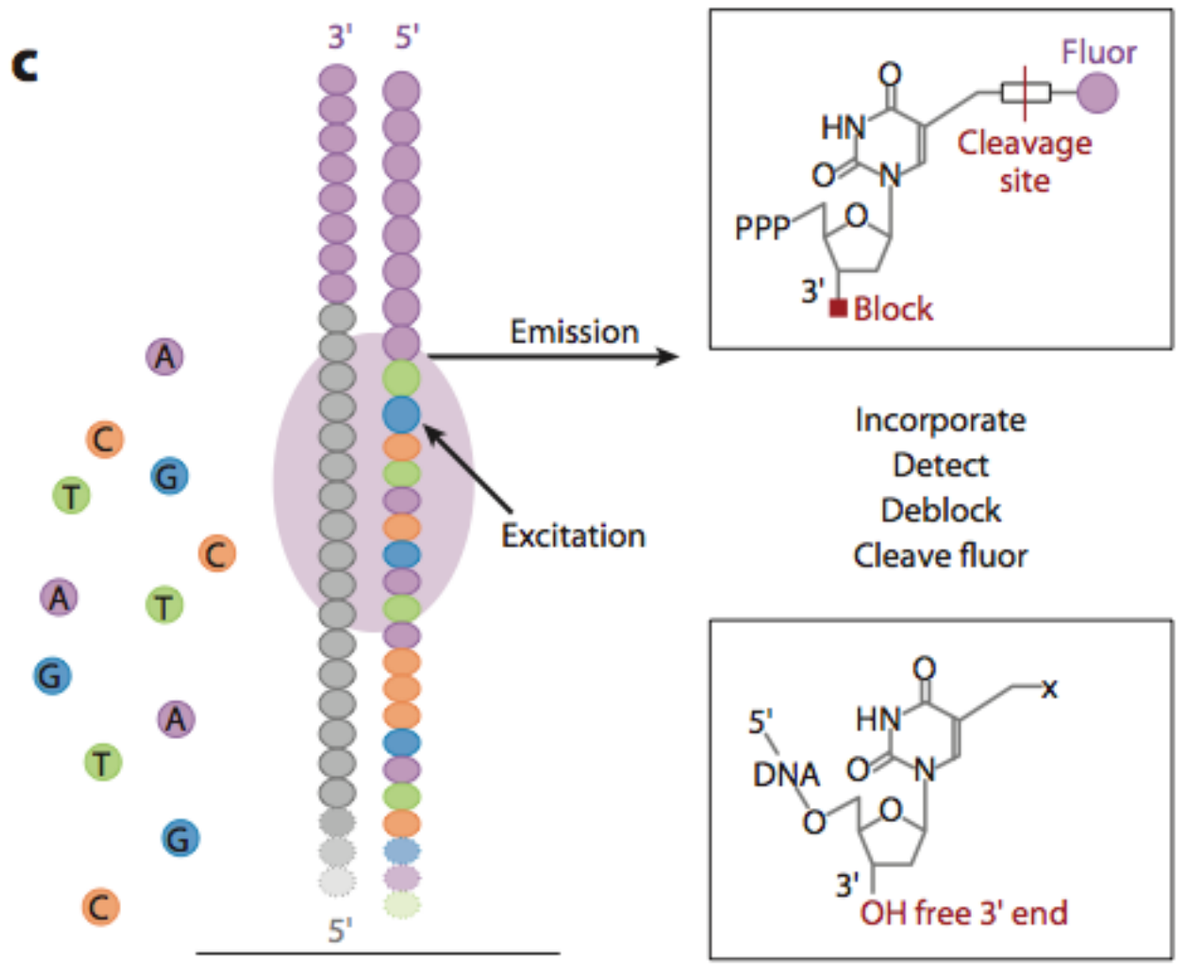
www.illumina.com/technology/next-generation-sequencing.html

Step 3 of 4

Illumina

The sequencing occurs as single-nucleotide addition reactions because a blocking group exists at the 3′-OH position of the ribose sugar, preventing additional base incorporation reactions by the polymerase.

In each step:
(a) The nucleotide is added by polymerase,
(b) unincorporated nucleotides are washed away
(c) the flow cell is imaged on both inner surfaces to identify each cluster that is reporting a fluorescent signal
(d) the fluorescent groups are chemically cleaved, and (e) the 3′-OH is chemically deblocked.



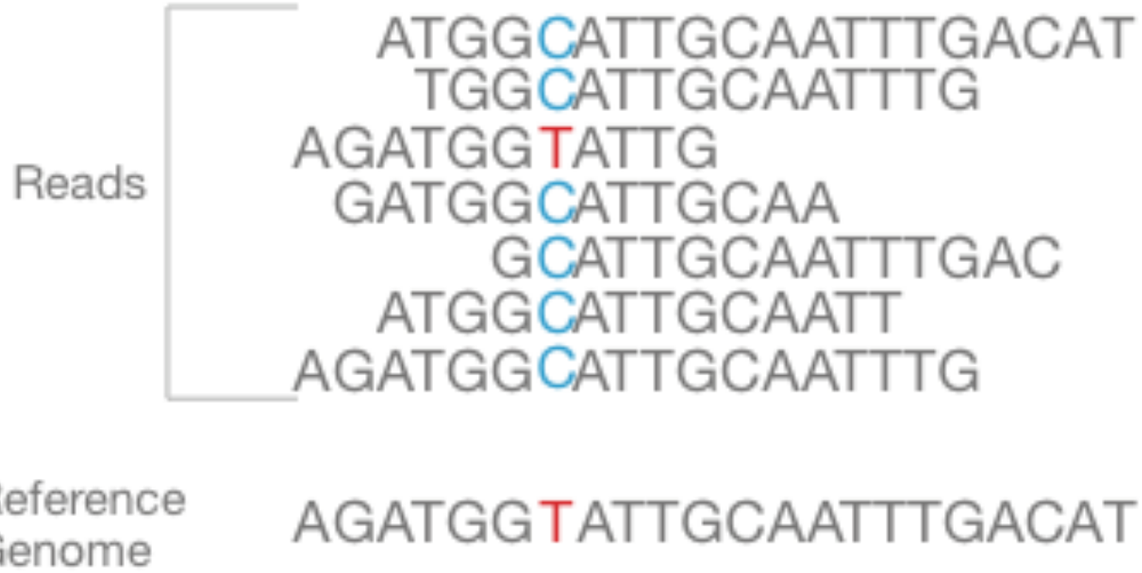
Sequencing by synthesis with reversible dye terminators.

Mardis ER. Next-generation sequencing platforms. Annu Rev Anal Chem 2013;6:287-303.

Step 4 of 4

Illumina

4. Data Analysis — During data analysis and alignment, the newly identified sequence reads are then aligned to a reference genome. Following alignment, many variations of analysis are possible such as single nucleotide polymorphism (SNP) or insertion-deletion (indel) identification, read counting for RNA methods, phylogenetic or metagenomic analysis, and more.



www.illumina.com/technology/next-generation-sequencing.html

Additional Considerations

Illumina

Multiple Kinds of Libraries

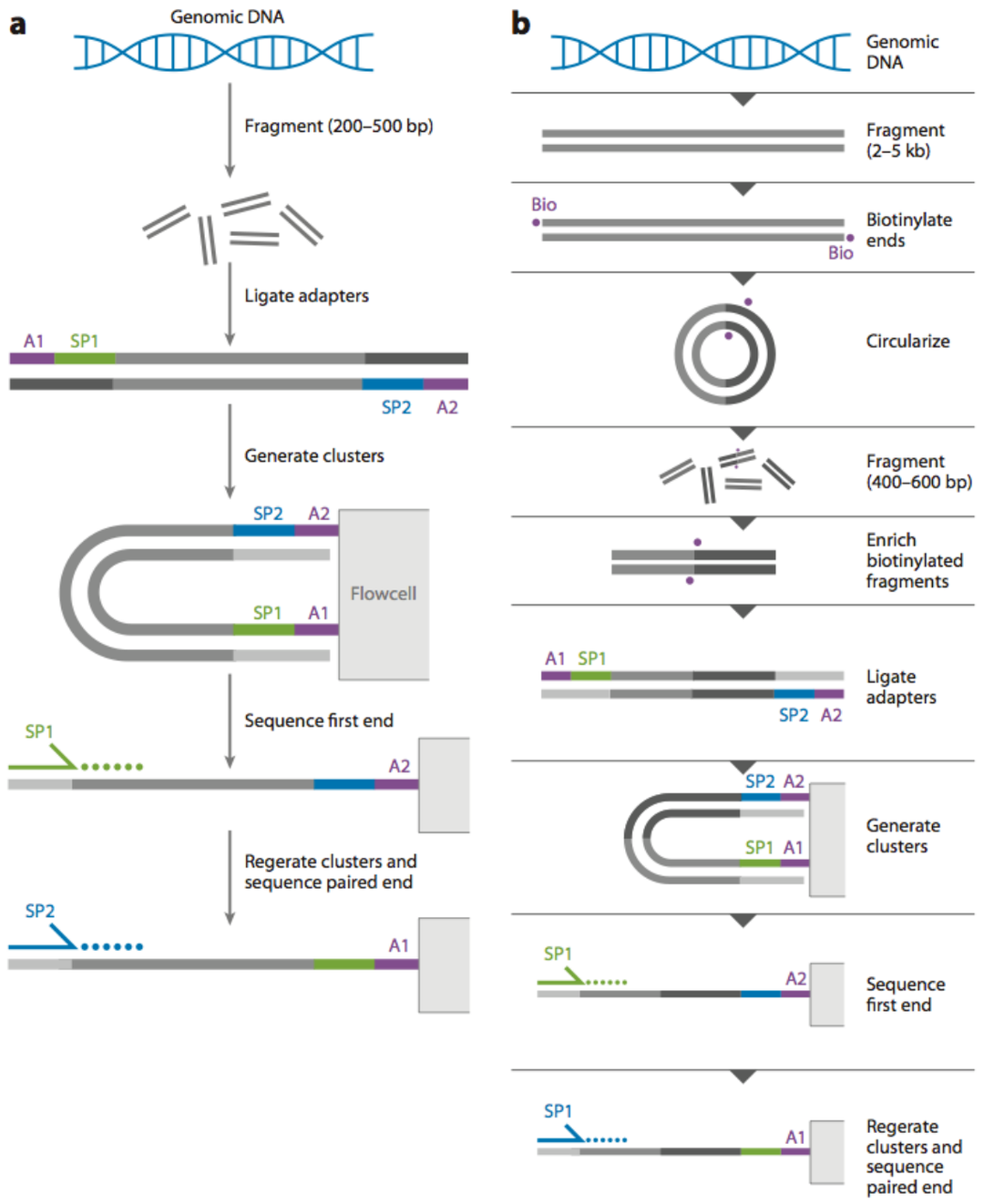


Figure 2

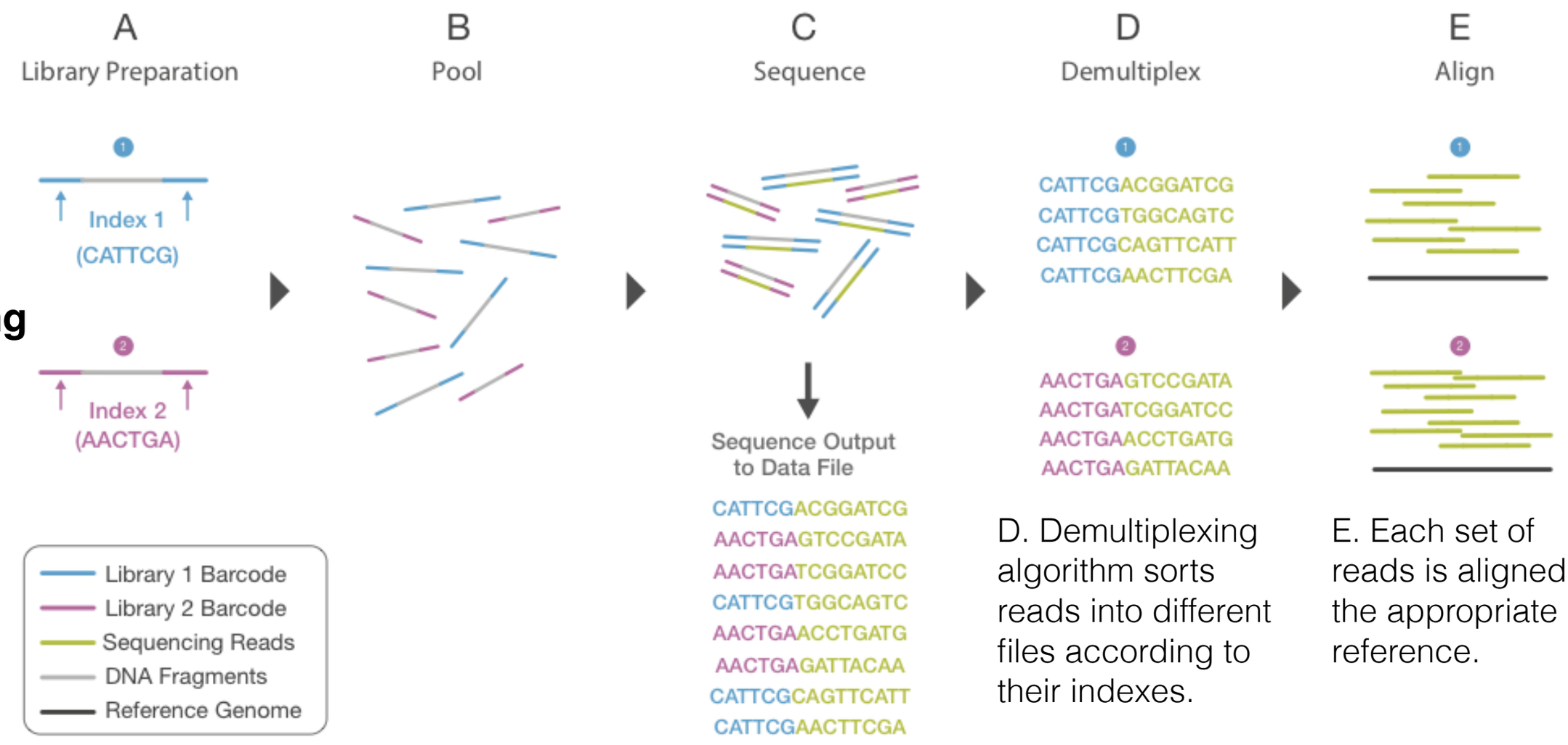
Comparison between (*a*) paired-end and (*b*) mate-pair sequencing library-construction processes.

Mardis ER. Next-generation sequencing platforms. Annu Rev Anal Chem 2013;6:287-303.

Additional Considerations

Illumina

Multiplexing



A. Two distinct libraries attached to unique index sequences during library preparation.

B. Libraries pooled together & loaded into same flow cell lane.

C. Libraries sequenced together during single instrument run.
All sequences exported to single output file.

D. Demultiplexing algorithm sorts reads into different files according to their indexes.

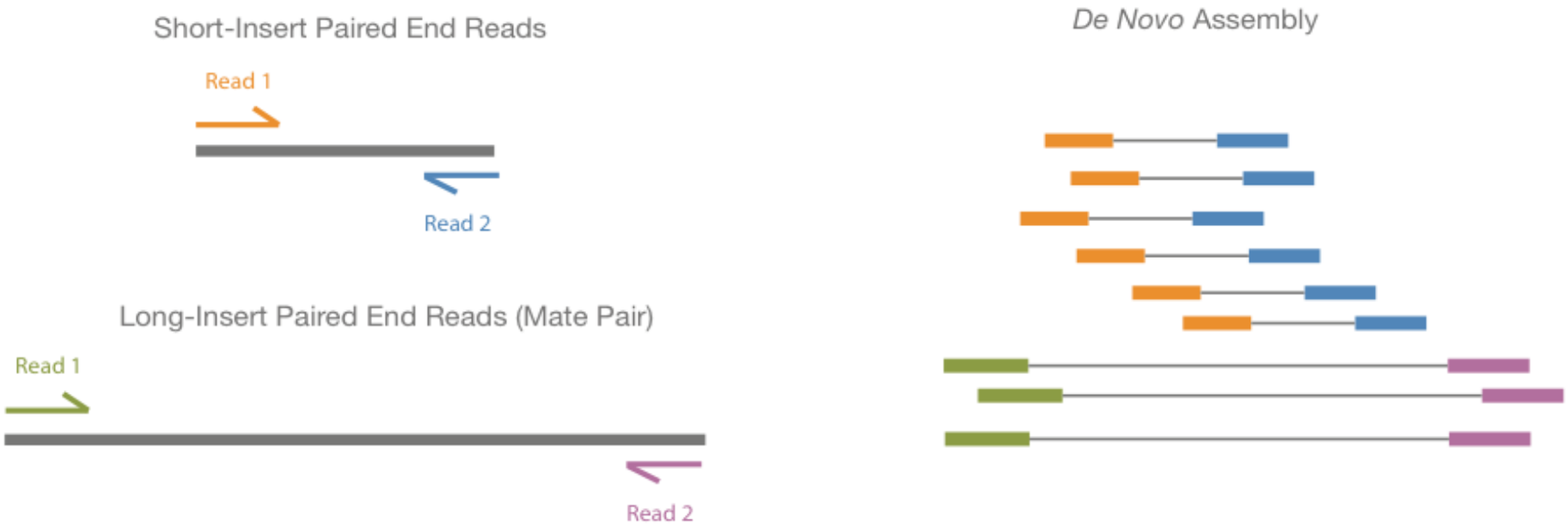
E. Each set of reads is aligned to the appropriate reference.

www.illumina.com/technology/next-generation-sequencing.html

Additional Considerations

Illumina

Paired End Reads



Mate Pairs and De Novo Assembly—combination of short and long insert sizes with paired-end sequencing for in maximal coverage the genome for *de novo* assembly.

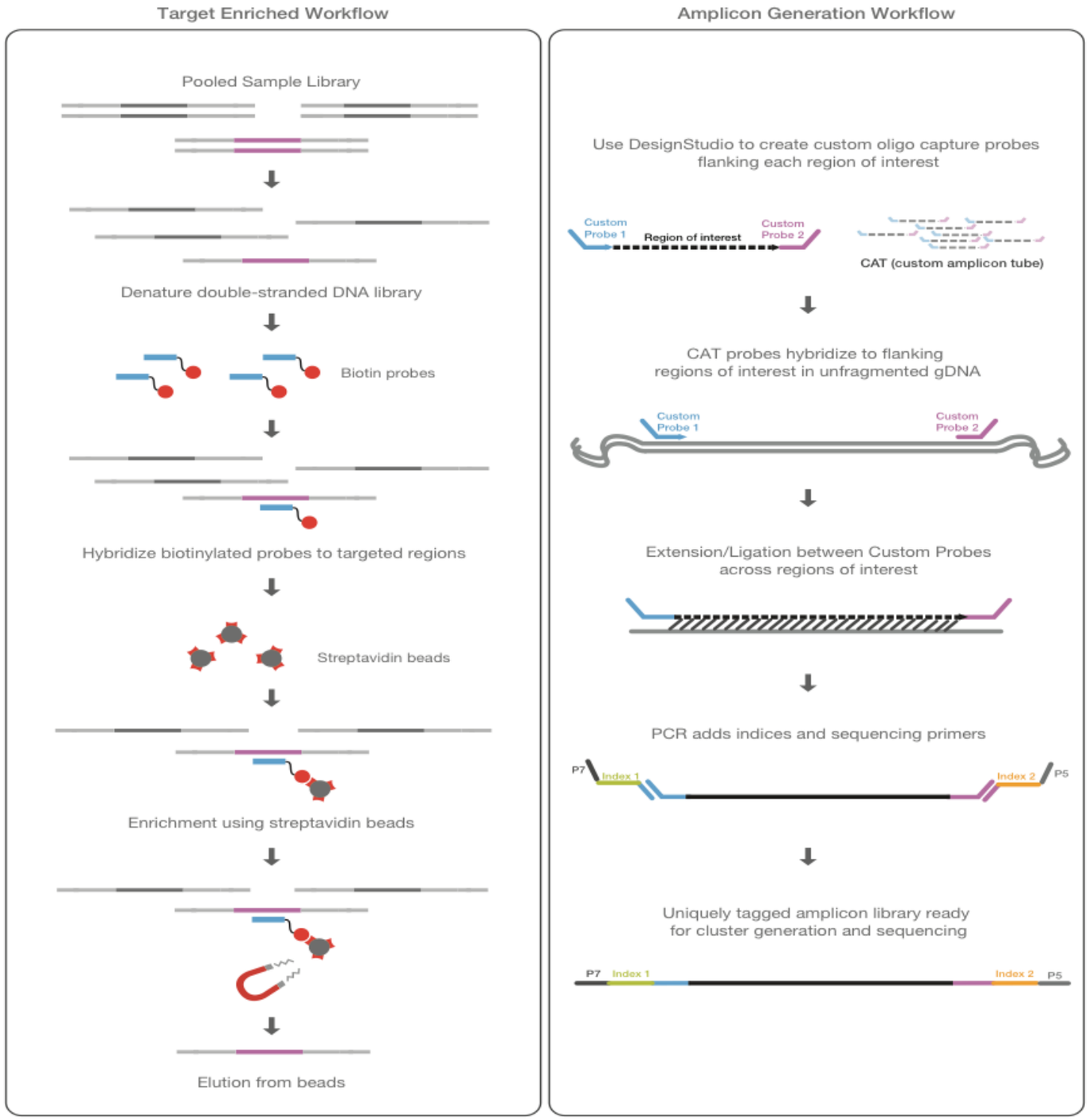
Longer inserts - better ability for repetitive sequences and regions.
Shorter inserts can fill in gaps at high depths.

www.illumina.com/technology/next-generation-sequencing.html

Additional Considerations

Targeted Sequencing

Illumina



www.illumina.com/technology/next-generation-sequencing.html

Additional Considerations

Illumina

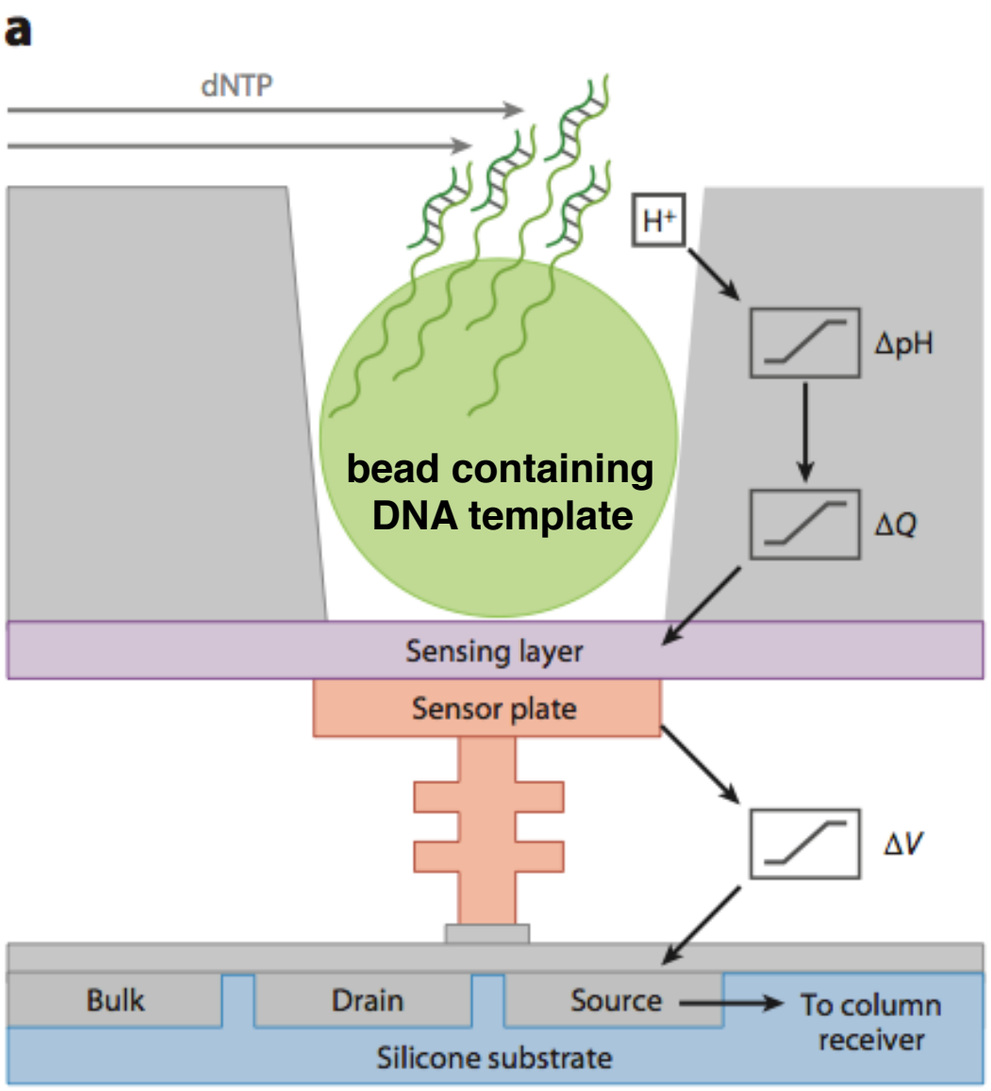
Sequencers

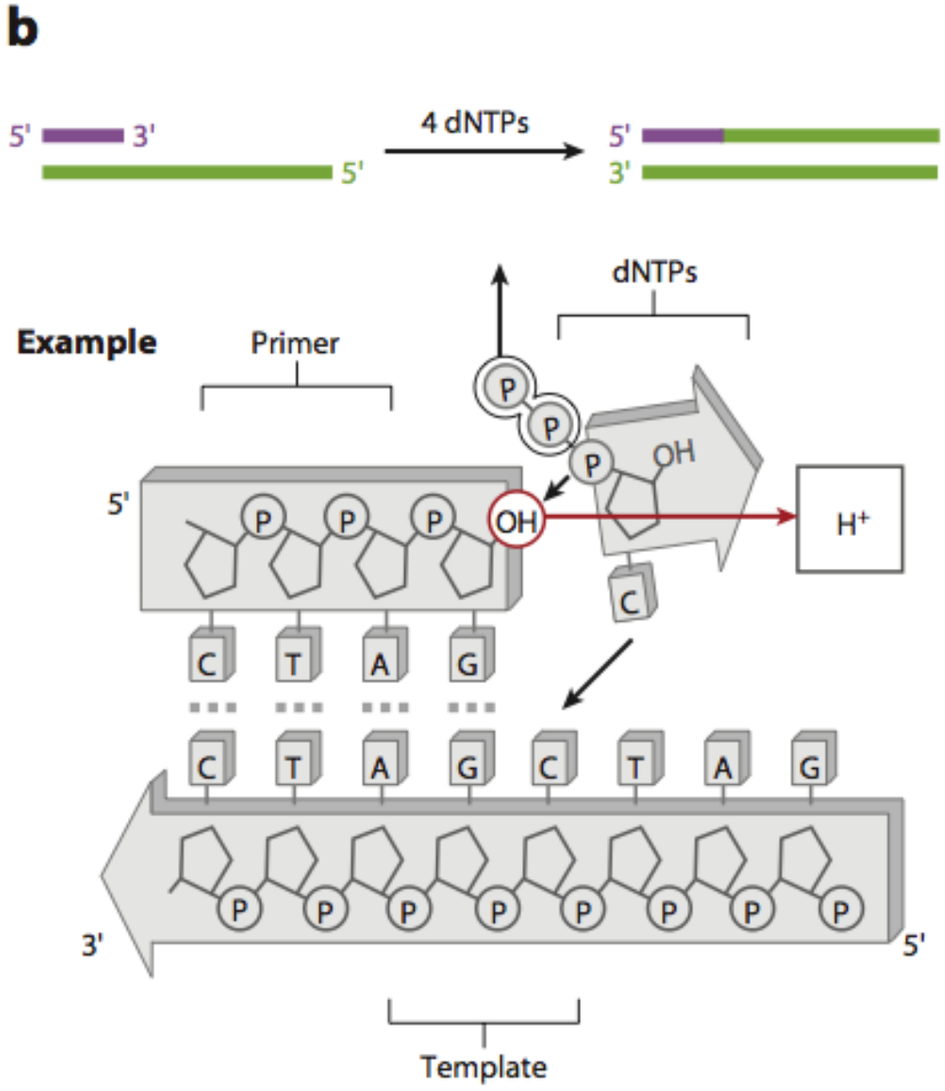
	MiniSeq System	MiSeq Series	NextSeq Series	HiSeq Series	HiSeq X Series*
Key Methods	Amplicon, targeted RNA, small RNA, and targeted gene panel sequencing.	Small genome, amplicon, and targeted gene panel sequencing.	Everyday exome, transcriptome, and targeted resequencing.	Production-scale genome, exome, transcriptome sequencing, and more.	Population- and production-scale whole-genome sequencing.
Maximum Output	7.5 Gb	15 Gb	120 Gb	1500 Gb	1800 Gb
Maximum Reads per Run	25 million	25 million†	400 million	5 billion	6 billion
Maximum Read Length	2 × 150 bp	2 × 300 bp	2 × 150 bp	2 × 150 bp	2 × 150 bp
Run Time	4–24 hours	4–55 hours	12–30 hours	<1–3.5 days (HiSeq 3000/HiSeq 4000) 7 hours–6 days (HiSeq 2500)	<3 days
Benchtop Sequencer	Yes	Yes	Yes	No	No
System Versions	• MiniSeq System for low-throughput targeted DNA and RNA sequencing	• MiSeq System for targeted and small genome sequencing • MiSeq FGx System for forensic genomics • MiSeqDx System for molecular diagnostics	• NextSeq 500 System for everyday genomics • NextSeq 550 System for both sequencing and cytogenomic arrays	• HiSeq 3000/HiSeq 4000 Systems for production-scale genomics • HiSeq 2500 Systems for large-scale genomics	• HiSeq X Five System for production-scale whole-genome sequencing • HiSeq X Ten System for population-scale whole-genome sequencing

www.illumina.com

Ion Torrent

Ion Torrent and pH sensing of nucleotide incorporation





Protons released when nucleotides (dNTP) incorporated on the growing DNA strands, changing the pH of the well (ΔpH)

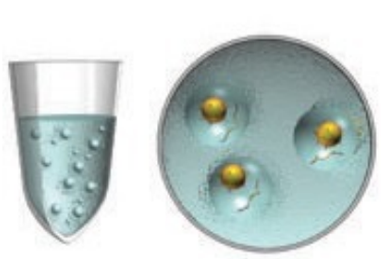
pH change induces a change in surface potential of the metal-oxide-sensing layer, and a change in potential (ΔV) of the source terminal of the underlying field-effect transistor

Mardis ER. Next-generation sequencing platforms. Annu Rev Anal Chem 2013;6:287-303
Rothberg JM et al. W, Rearick TM, Schultz J, Mileski W, et al. Nature 475:348–52 (2011)

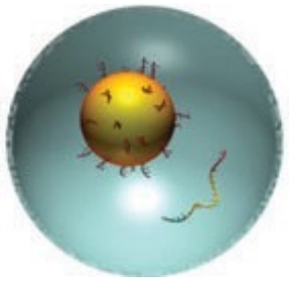
Ion Torrent and pH sensing of nucleotide incorporation

Emulsion PCR

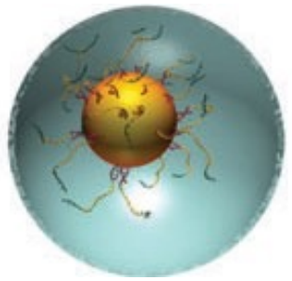
8 hours



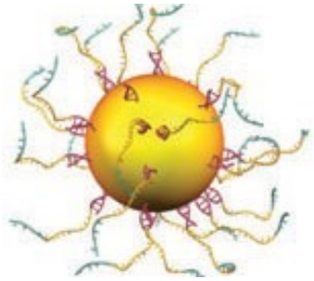
Anneal sstDNA to an excess of DNA capture beads



Emulsify beads and PCR reagents in water-in-oil microreactors



Clonal amplification occurs inside microreactors

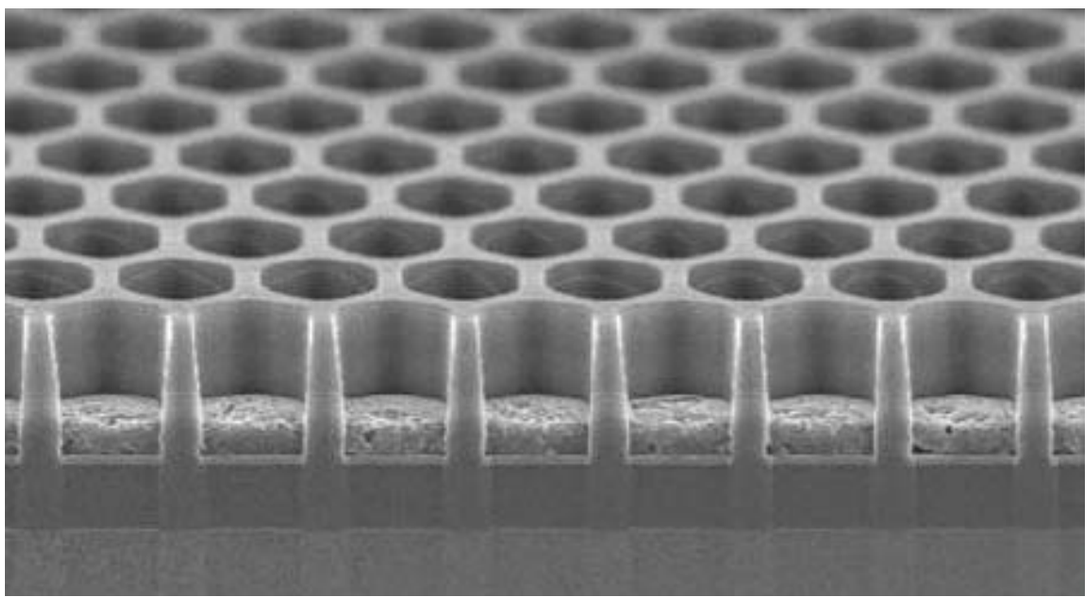


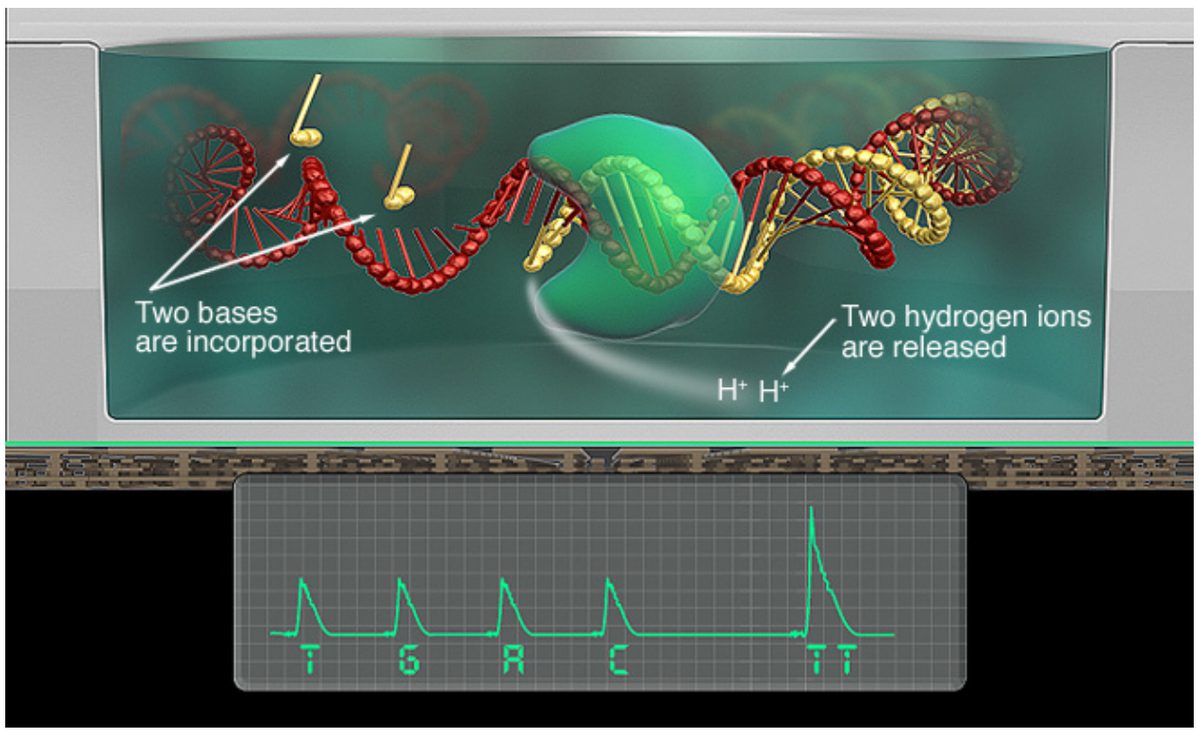
Break microreactors and enrich for DNA-positive beads

sstDNA library → Bead-amplified sstDNA library

Similar to 454

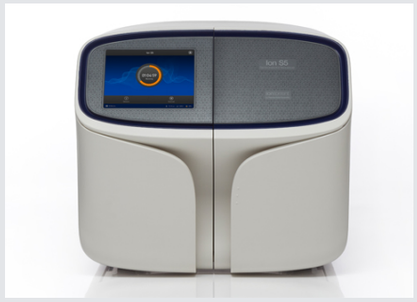
Mardis 2008. Annual Rev. Genetics 9: 387.





Rothberg JM et al. W, Rearick TM, Schultz J, Mileski W, et al. Nature 475:348–52 (2011)
https://www.thermofisher.com/us/en/home/life-science/sequencing/next-generation-sequencing/ion-torrent-next-generatio sequencing-technology.html (Accessed 3/2016)

Ion Torrent Instruments and Chips

Instrument	Ion Chip	Read Length	Key Research Applications
Ion S5 System 	Ion 520™ Chip	Up to 400 bp	Targeted DNA sequencing, targeted RNA sequencing, small RNA sequencing, *de novo* microbial sequencing, bacterial typing, viral typing, metagenomics, sequencing by genotyping
	Ion 530™ Chip		
	Ion 540™ Chip	Up to 200 bp	Exome sequencing, transcriptome sequencing, copy number analysis
Ion PGM System 	Ion 314™ Chip	Up to 400 bp	Targeted DNA sequencing, targeted RNA sequencing, small RNA sequencing, *de novo* microbial sequencing, bacterial typing, viral typing, metagenomics, sequencing by genotyping
	Ion 316™ Chip		
	Ion 318™ Chip		
Ion Proton System 	Ion PI™ Chip	Up to 200 bp	Exome sequencing, transcriptome sequencing, copy number analysis

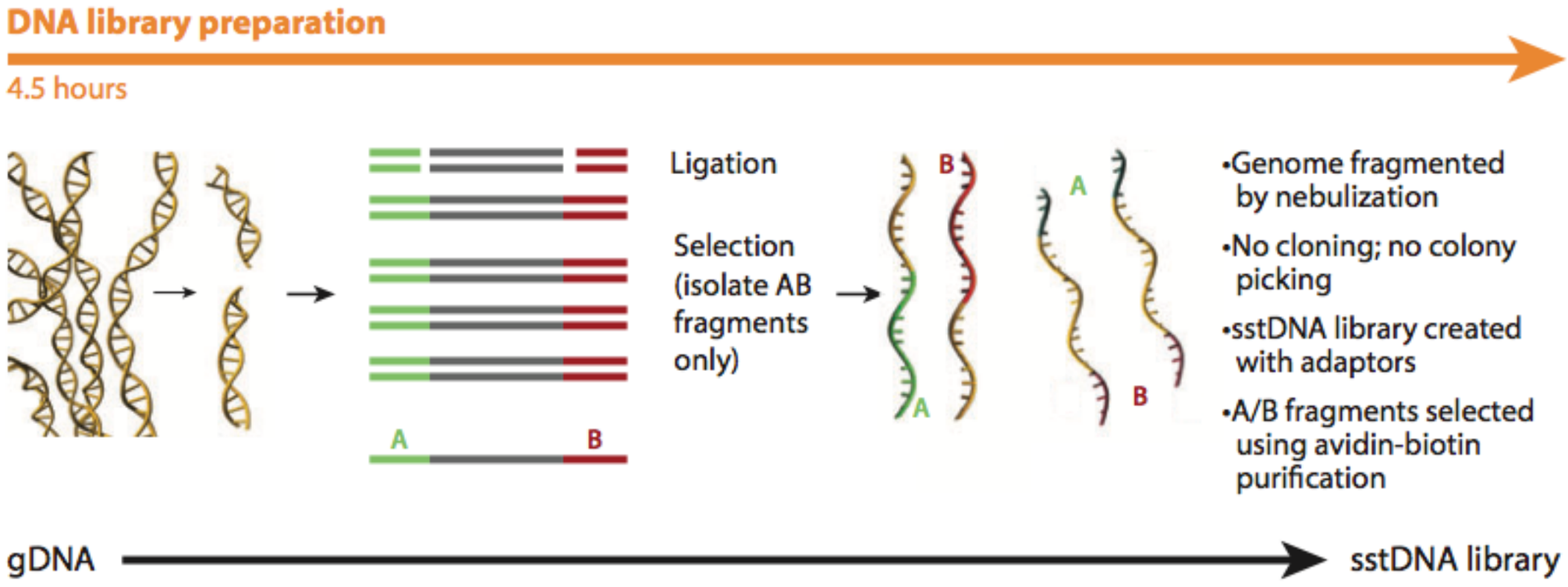
https://www.thermofisher.com/us/en/home/life-science/sequencing/next-generation-sequencing/ion-torrent-next-generatio
sequencing-workflow/prepare-template/ion-torrent-next-generation-sequencing-ion-chef-system.html.html (3/2016)

Roche-454

Phased Out

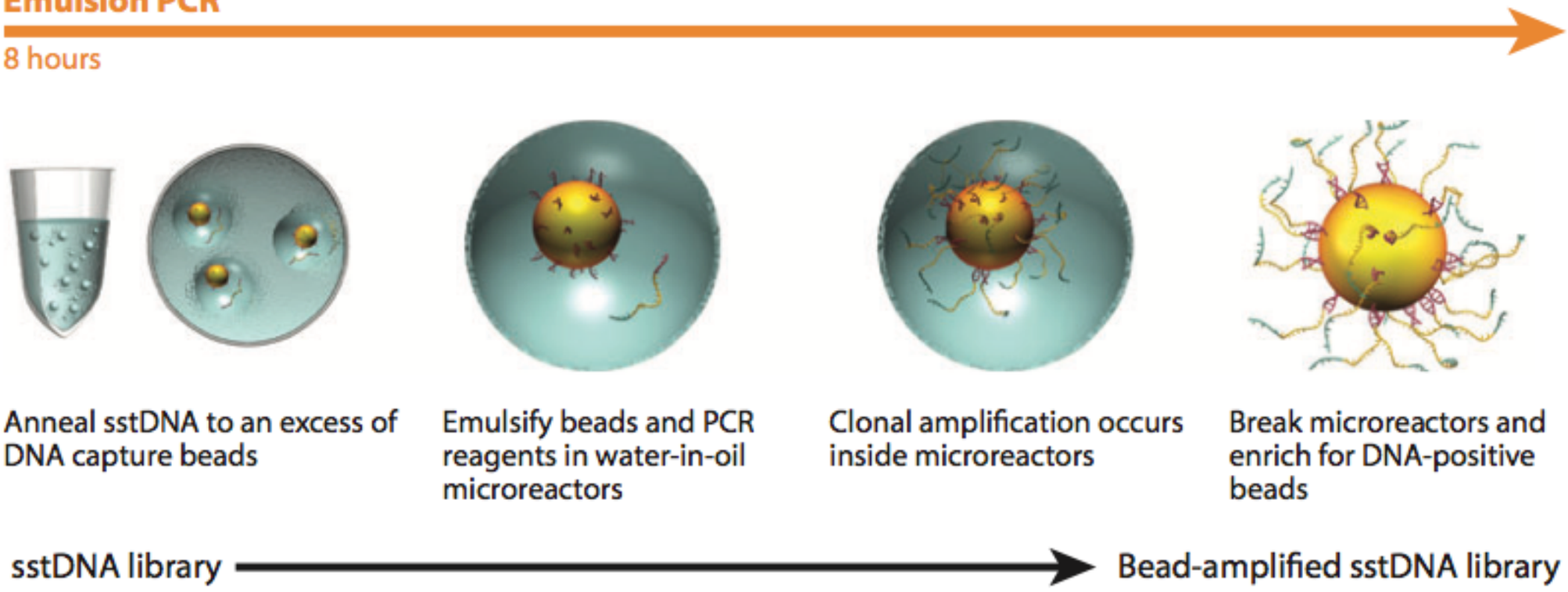
Roche-454

First commercial next gen sequencing system (2004)



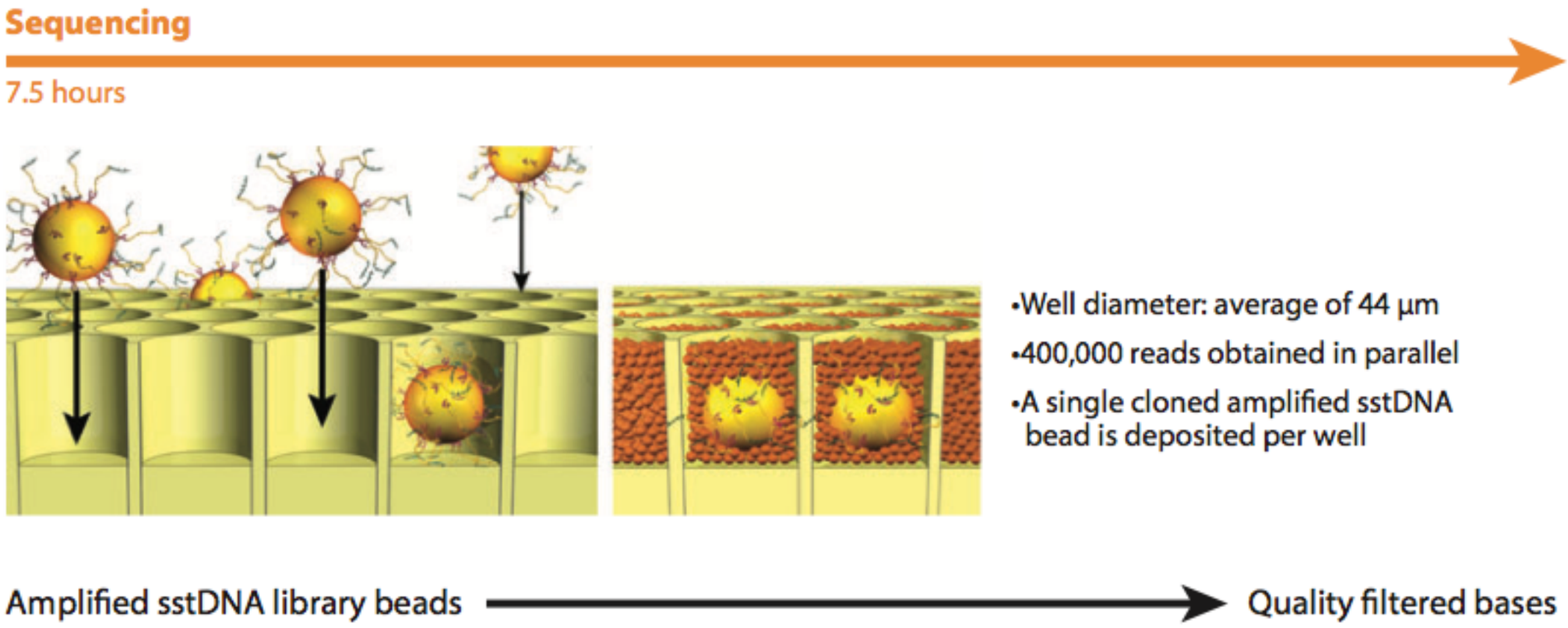
Mardis 2008. Annual Rev. Genetics 9: 387.

Roche-454

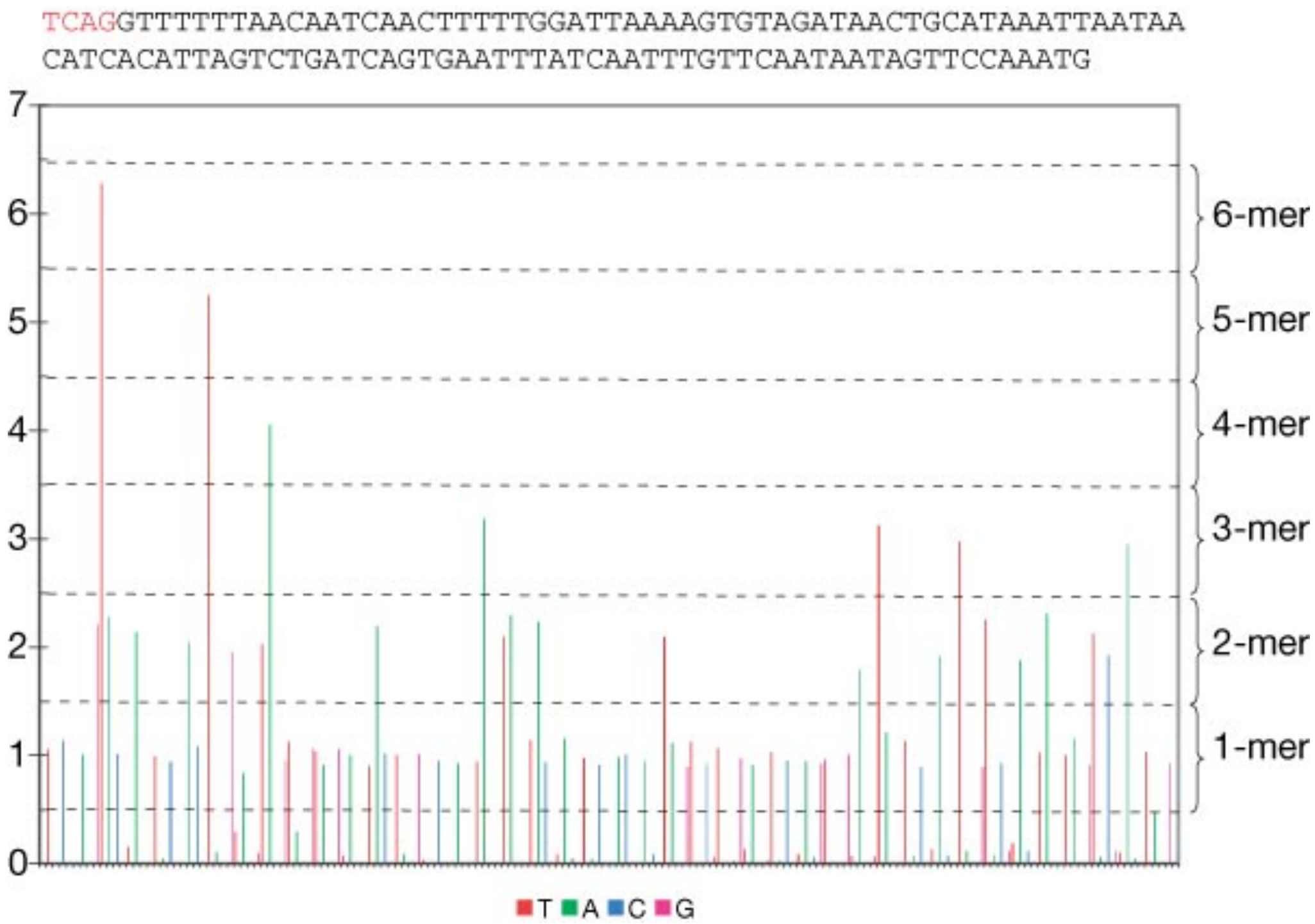


Mardis 2008. Annual Rev. Genetics 9: 387.

Roche-454



Mardis 2008. Annual Rev. Genetics 9: 387.



Nucleotides are flowed in the order T, A, C, G. The sequence is shown above the flowgram. The signal value intervals corresponding to the various homopolymers are indicated on the right. The first four bases (in red, above the flowgram) constitute the 'key' sequence, used to identify wells containing a DNA-carrying bead.

Margulies et al., Nature 437:376-380 (2005)

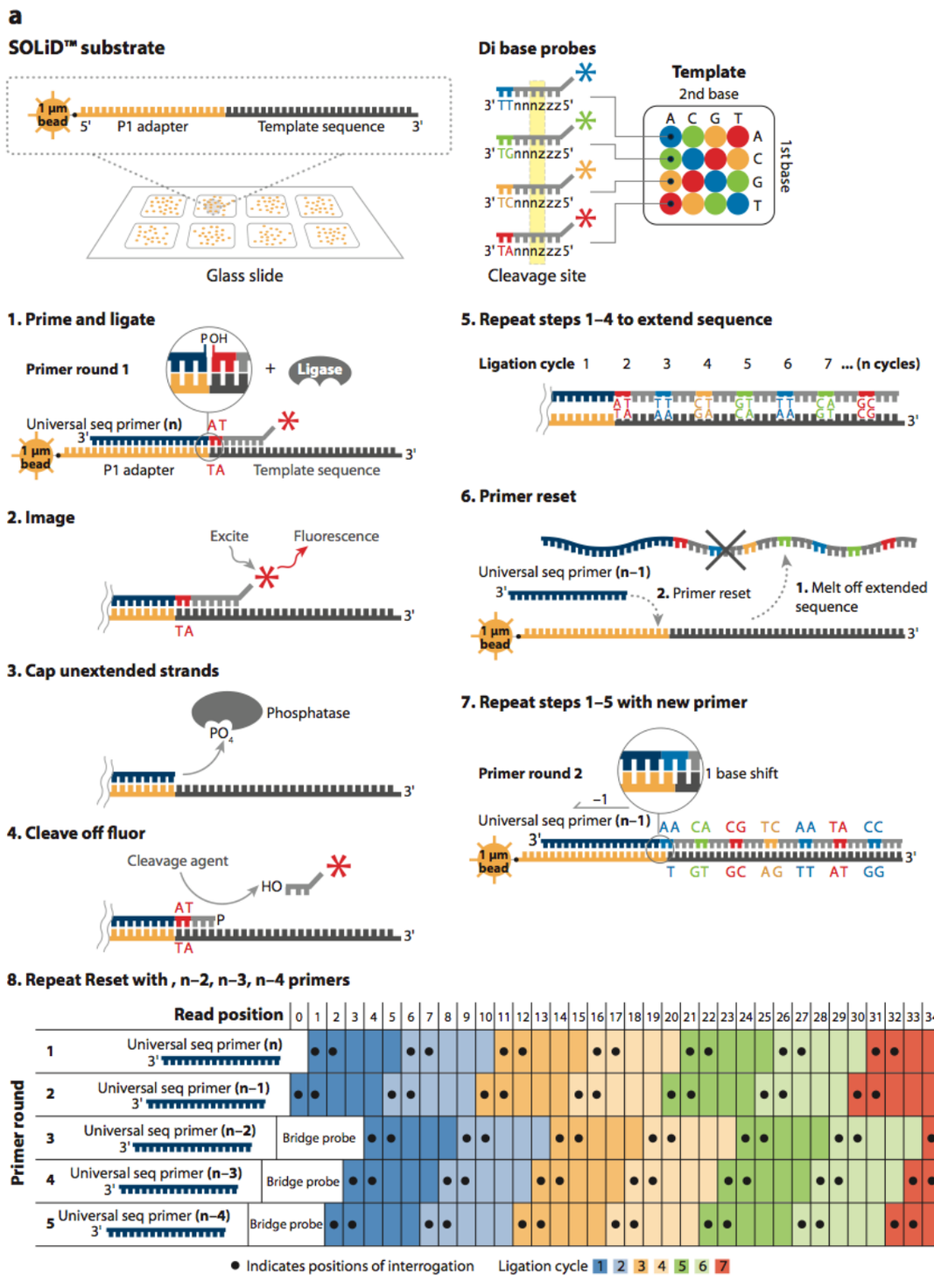
ABI SOLiD

Phased Out

ABI SOLiD

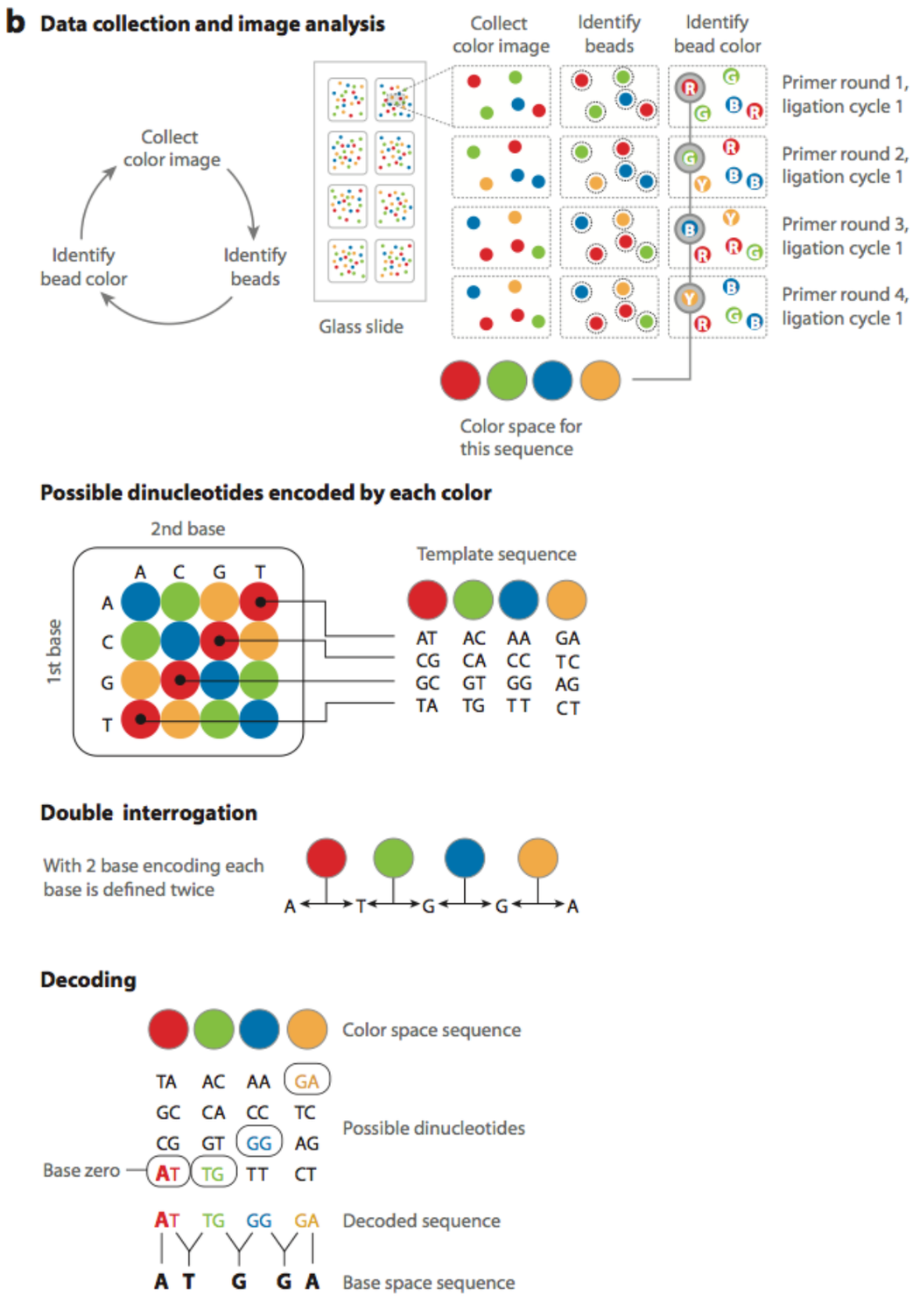
- The ligase-mediated sequencing approach of the Applied Biosystems SOLiD sequencer. In a manner similar to Roche/454 emulsion PCR amplification, DNA fragments for SOLiD sequencing are amplified on the surfaces of 1-µm magnetic beads to provide sufficient signal during the sequencing reactions, and are then deposited onto a flow cell slide.
- Ligase-mediated sequencing begins by annealing a primer to the shared adapter sequences on each amplified fragment, and then DNA ligase is provided along with specific fluorescent- labeled 8mers, whose 4th and 5th bases are encoded by the attached fluorescent group.
- Each ligation step is followed by fluorescence detection, after which a regeneration step removes bases from the ligated 8mer (including the fluorescent group) and concomitantly prepares the extended primer for another round of ligation.

Mardis 2008. Annual Rev. Genetics 9: 387.



ABI SOLiD

- Because each fluorescent group on a ligated 8mer identifies a two-base combination, the resulting sequence reads can be screened for base-calling errors versus true polymorphisms versus single base deletions by aligning the individual reads to a known high-quality reference sequence.



Mardis 2008. Annual Rev. Genetics 9: 387.

Complete Genomics

Complete Genomics

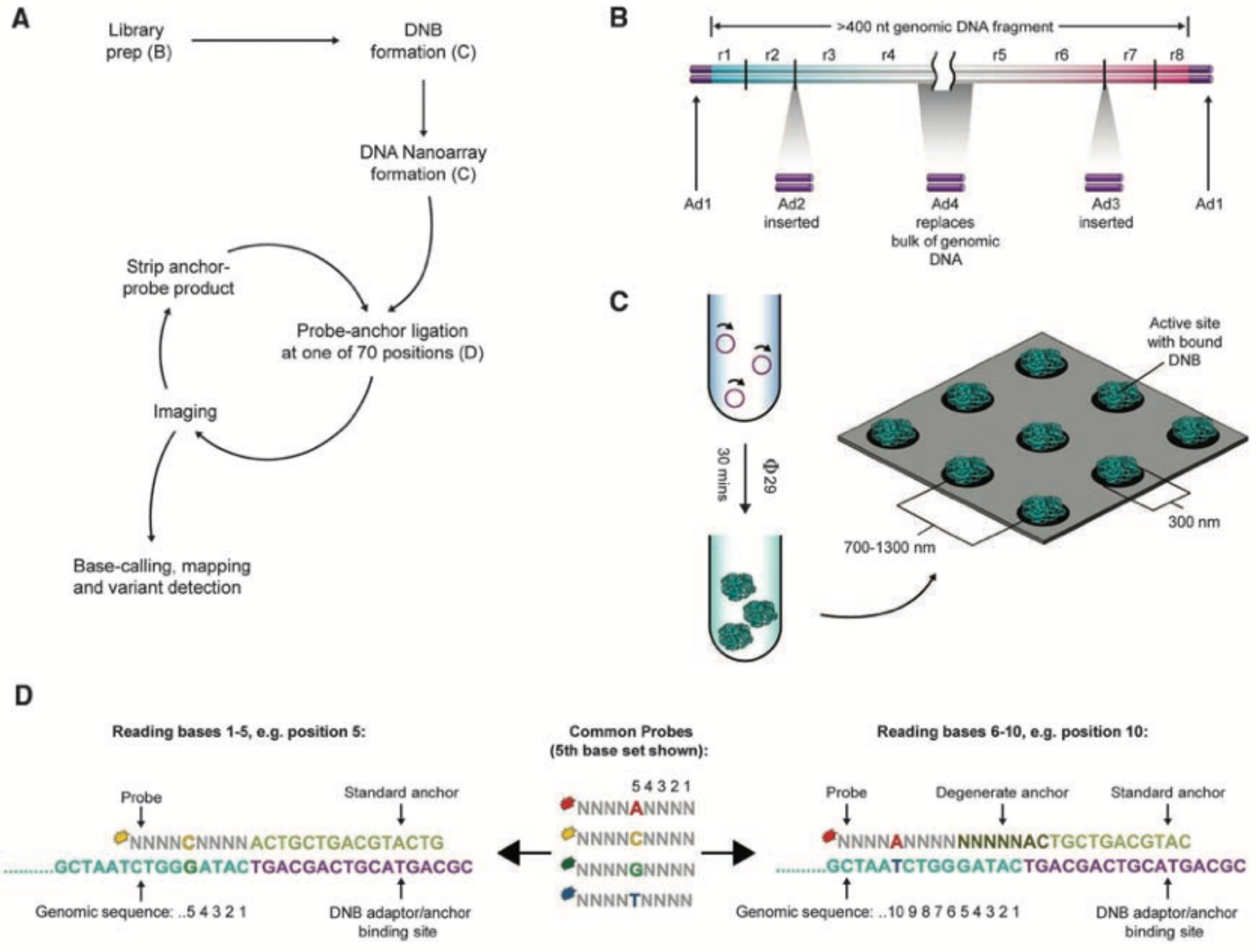


Fig. 1. Amplified DNA nanoarray platform. (**A**) Schematic flow diagram of the process used. (**B**) Library construction schematic (fig. S1). r1 to r8 are gDNA regions adjacent to distinct adapter ends; Ad1 to AD4 indicate adapters 1 to 4. (**C**) DNB production using *Phi*29 DNA polymerase (fig. S11) and nanoarray formation (SOM) schematics. (**D**) Schematic of cPAL products (SOM).

Drmanac et al., Science 327(5961) pp. 78-81 (2010)

Video Illustrations

Video Illustrations

Illumina

https://youtu.be/HMyCqWhwB8E
https://youtu.be/pfZp5Vgsbw0

Ion Torrent

https://youtu.be/WYBzbxIfuKs

Roche - 454

https://youtu.be/rsJoG-AulNE