

# **CH-413 Nanobiotechnology**

## **Next-Generation Sequencing (NGS)**

Angela Steinauer

April 3, 2025

# The Human Genome Project

---



Venter

Collins

*“Bill Clinton wanted biotech entrepreneur Craig Venter (left) and Francis Collins (centre) of the US National Institutes of Health to patch up their differences.”* Credit: Ron Sachs/Shutterstock

<https://www.nature.com/articles/d41586-020-01849-w>

**Start:** 1990

**Completed:** April 2003

## **Goal:**

Sequencing of all 3 billion ( $3 \times 10^9$ ) base pairs of human genome

Competition between private (Celera) and public projects

## **Recommended resources:**

<https://www.genome.gov/human-genome-project>

<https://www.nature.com/articles/464649a>

# The Human Genome Project

---

**What about the naysayers who asked, “Where are the cures for diseases that we were promised?”**

I became director of this institute three and a half years ago, and I remember when I first started going around and giving talks. Routinely I would hear: “You are seven years into this. Where are the wins? Where are the successes?”

I don’t hear that as much anymore. I think what’s happening, and it has happened in the last three years in particular, is just the sheer aggregate number of the success stories. The drumbeat of these successes is finally winning people over.

We are understanding cancer and rare genetic diseases. There are incredible stories now where we are able to draw blood from a pregnant woman and analyze the DNA of her unborn child.

Increasingly, we have more informed ways of prescribing medicine because we first do a genetic test. We can use microbial DNA to trace disease outbreaks in a matter of hours.

These are just game changers. It’s a wide field of accomplishment, and there is a logical story to be told.

A version of this article appears in print on April 16, 2013, Section D, Page 3 of the New York edition with the headline: Human Genome, Then and Now. [Order Reprints](#) | [Today's Paper](#) | [Subscribe](#)



Excerpt from an interview in the *New York Times* with **Eric Green**, then director of the National Human Genome Research Institute at the National Institutes of Health

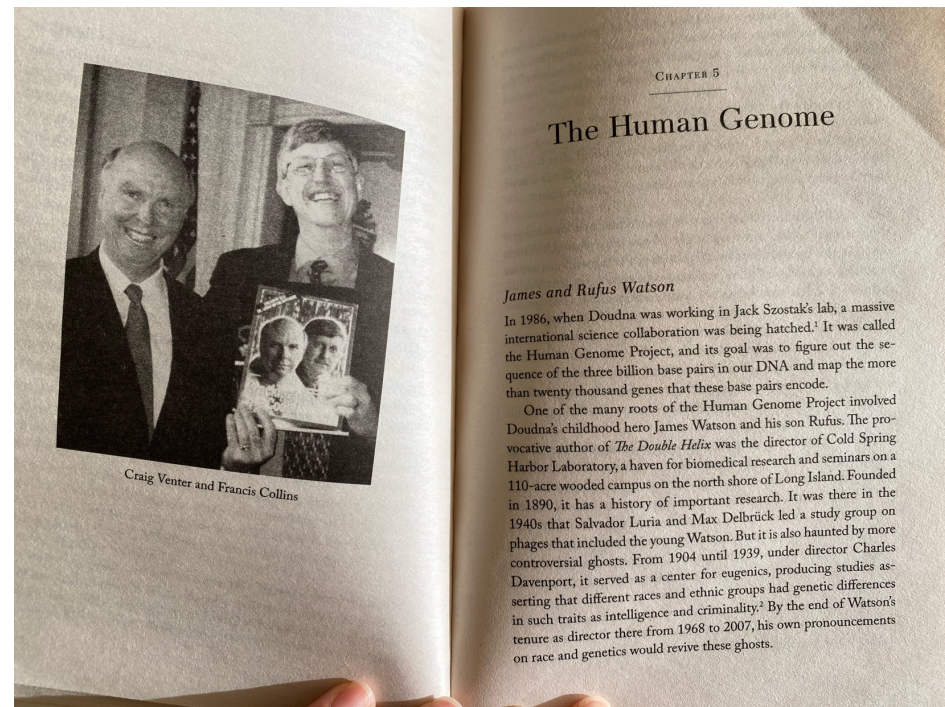
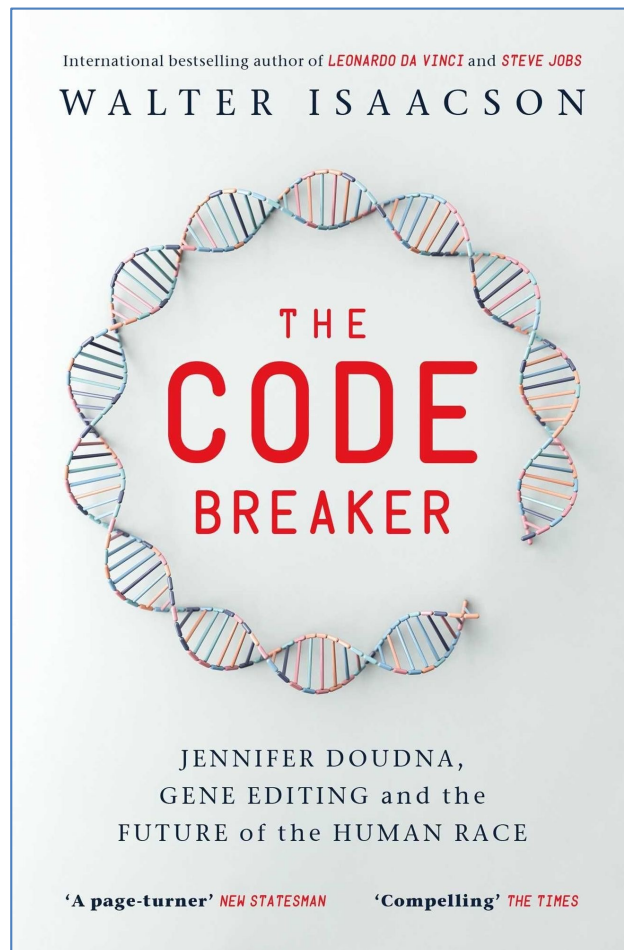
<https://www.nytimes.com/2013/04/16/science/the-human-genome-project-then-and-now.html>

# Impact of the Human Genome Project

---

Area	Impact
Genomics	Reference genome, gene catalog, understanding of variation
Technology	Rapid drop in sequencing cost, rise of NGS and bioinformatics
Medicine	Precision therapies, diagnostic tests, carrier screening
Science policy	Data-sharing norms, ELSI frameworks, global collaboration
Functional genomics	Mapping of non-coding regions, transcriptomics, epigenetics

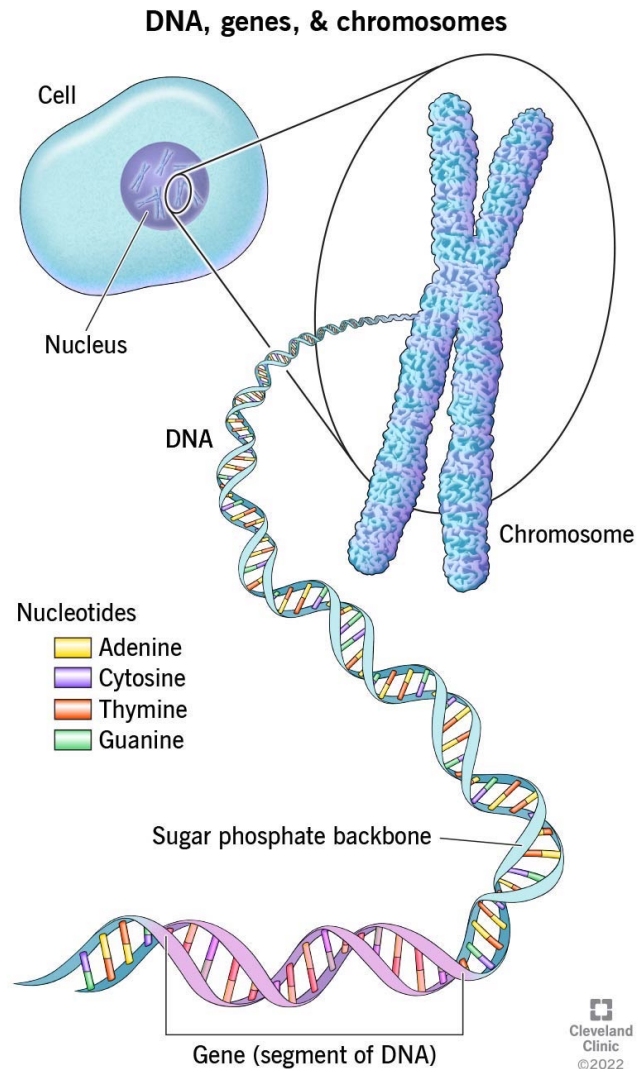
# Book recommendation





# Motivation: Read the genetic code to understand life itself

---



## DNA:

carrier of genetic information  
→ information storage

easily synthesized

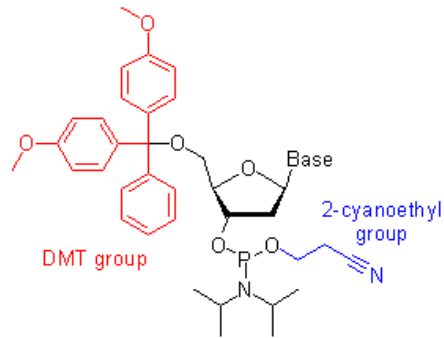
can be amplified

## Ideal information medium

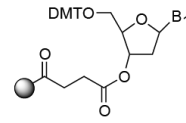
→ However, we need efficient methods to read the molecular information

# DNA synthesis

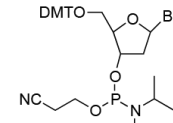
(DMT) dimethoxytrityl



Start here



deprotection



phosphoramidite  
building block

**Step 1**  
Activation and  
coupling

**Step 2**  
Capping

**Step 3**  
Oxidation

**Step 4**  
Deprotection

cleavage of  
finished  
oligonucleotide  
from support

deprotection

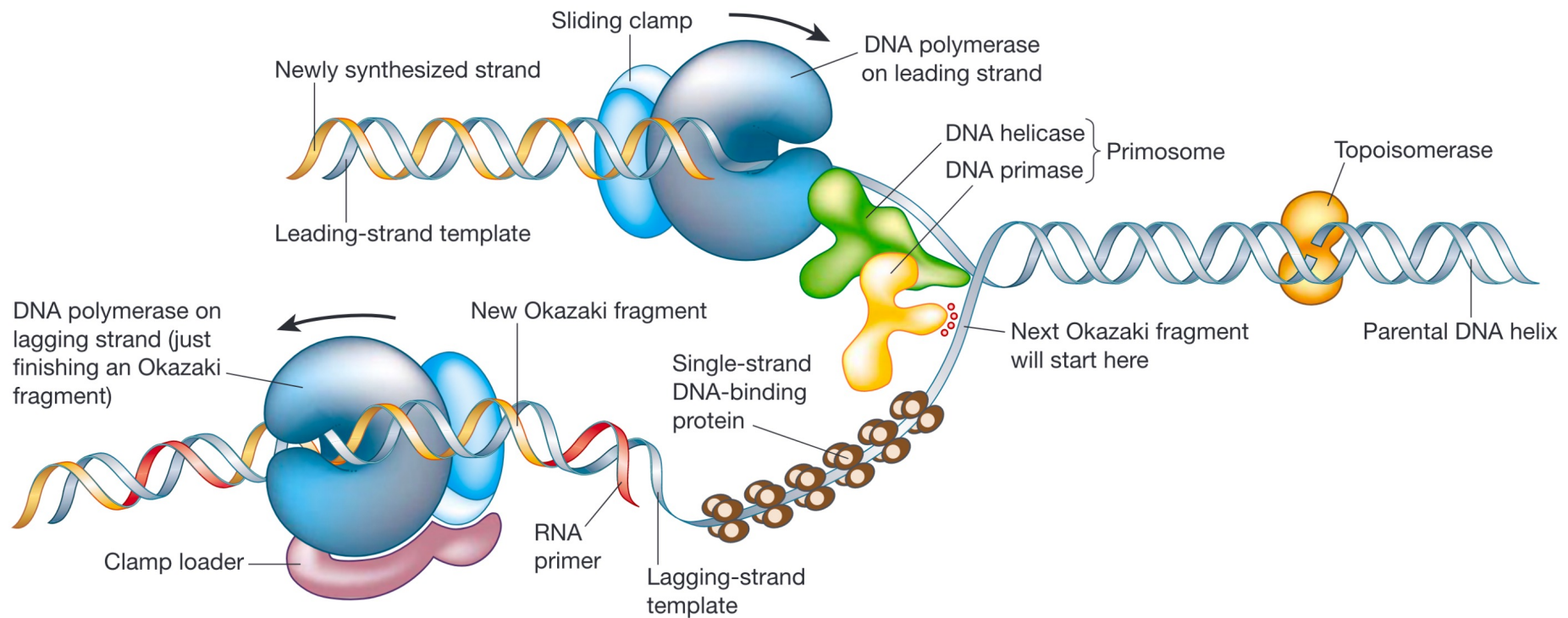
Highly automated,  
efficient and cheap  
Yield per cycle: 98-99.5%

Source:  
atdbio.com



# DNA synthesis in vivo: replication

---

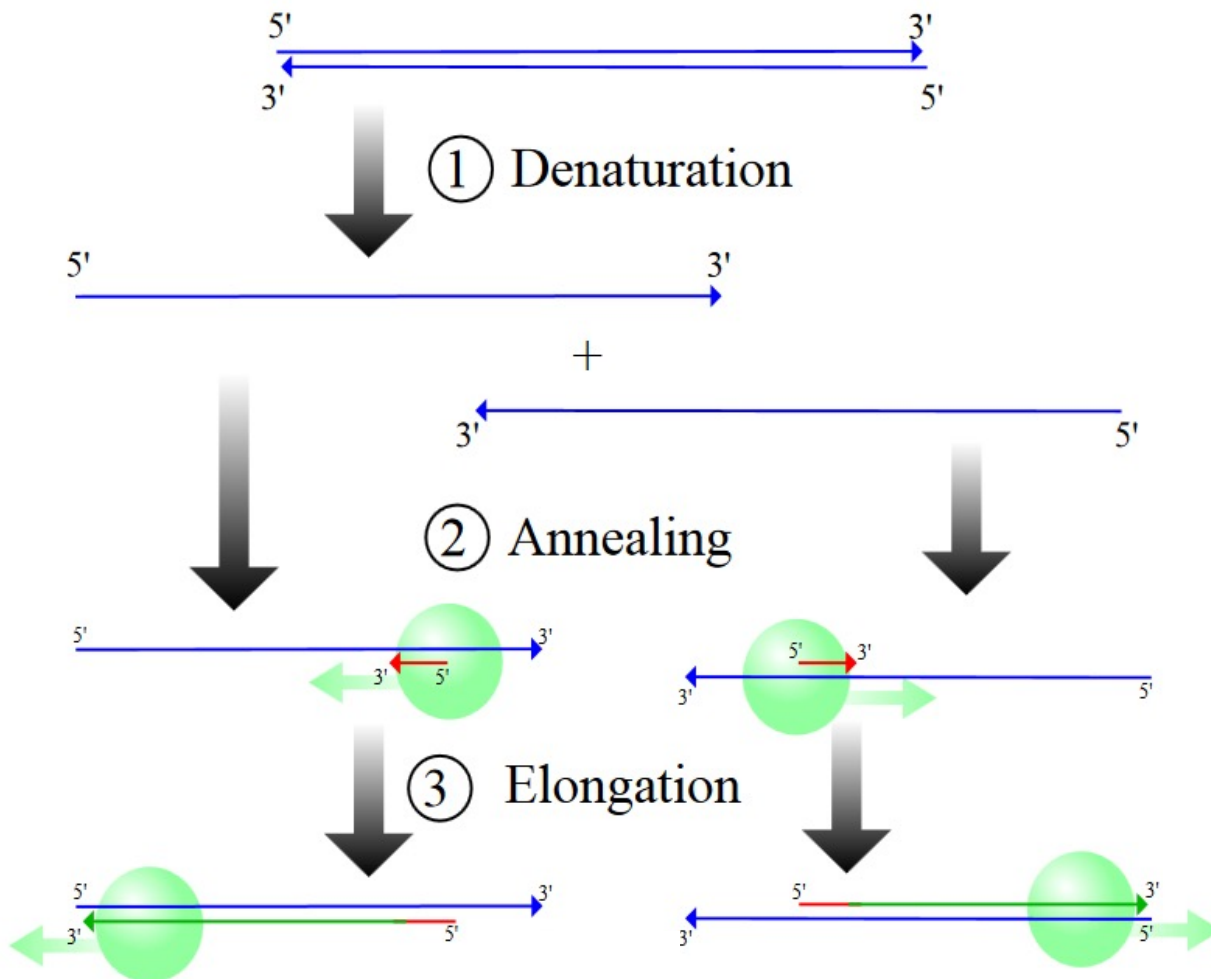


*Alberts, Nature 421, 431-453 (2003).*



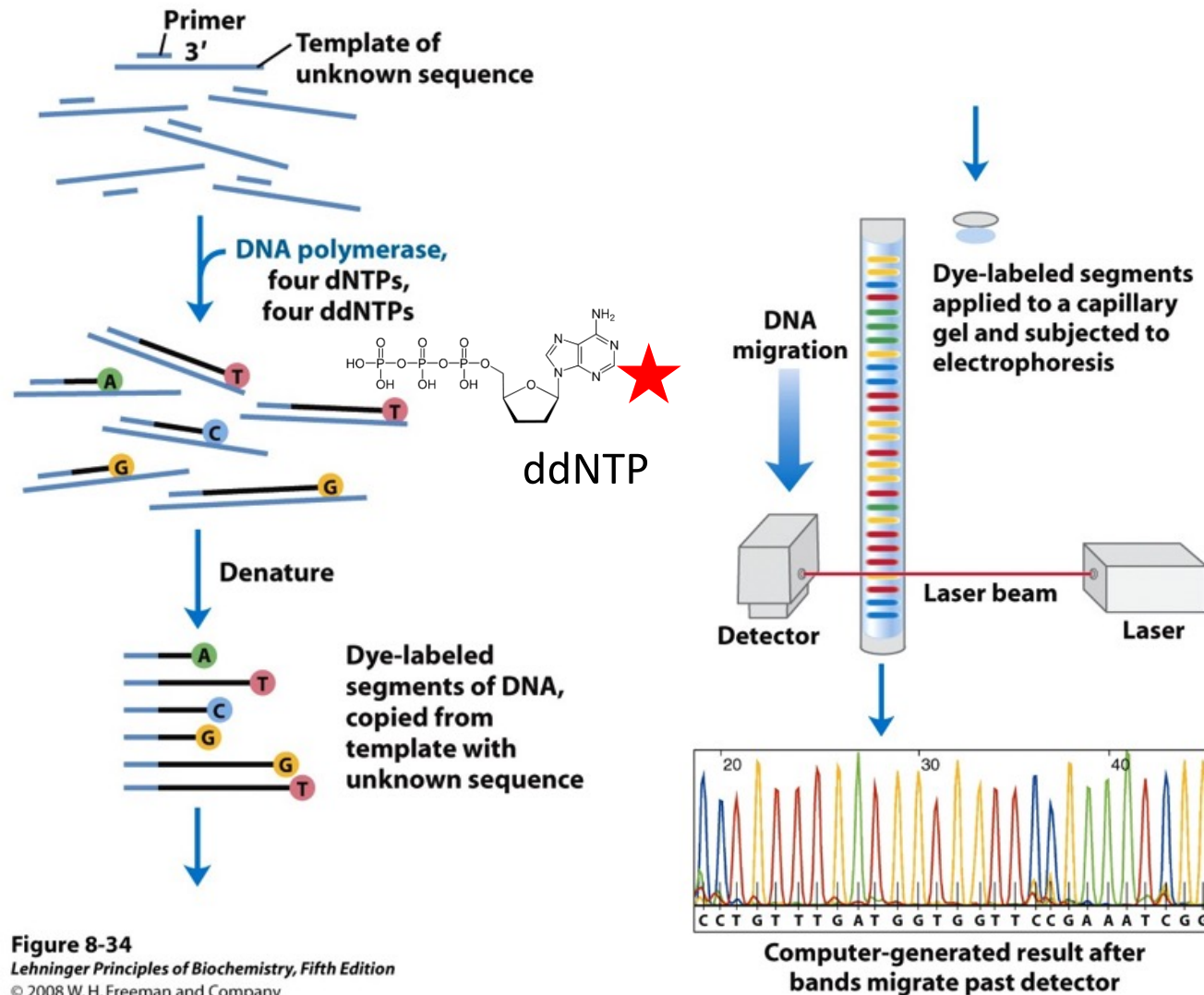
# PCR - Polymerase chain reaction

---



1. dsDNA is denatured by heating to 95 °C
2. Temperature is lowered to 50-65 °C: synthetic primers are annealed to the single-stranded DNA
3. Temperature is increased to 72 °C: DNA is elongated by heat-resistant polymerase
4. Cycle is repeated 20-35 times
5. **Exponential** increase in sequence of choice

# First-generation method: Automated Sanger sequencing



## Sequencing is based on PCR reaction

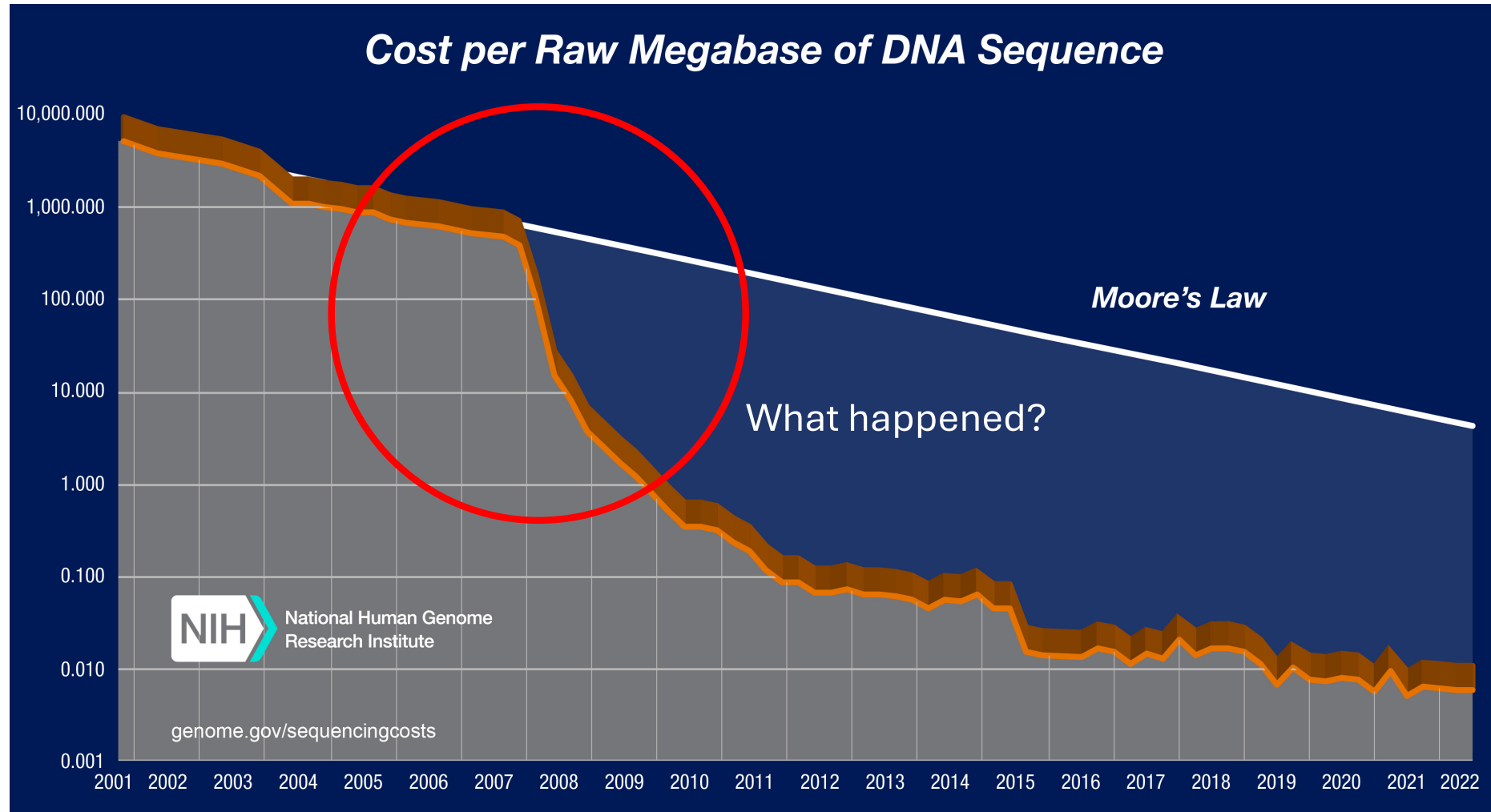
- PCR with mixture dNTPs and **fluorescent dideoxynucleotides (ddNTPs)**
- DNA synthesis stops at ddNTP
- Sequences are separated by capillary electrophoresis (shortest fragments elute first)
- Last nucleotide added is identified by its fluorescence signal
- Sequence (up to a 1000 bp) is reconstructed

## Disadvantage

- Slow
- Max 1000 bp read length
- **One DNA fragment per sequencing run**
- Material intensive

**Figure 8-34**  
*Lehninger Principles of Biochemistry, Fifth Edition*  
© 2008 W.H. Freeman and Company

# Sequencing costs



# Next-generation sequencing (NGS) methods

---

## A revolution of DNA analysis:

- Cheap generation of large amounts of data
  - one billion of short reads per instrument run
- Low amount of material required (single-molecule sequencing possible)
- Evolution of bioinformatics methods to assemble genome
- Methods very versatile, not limited to genome sequencing
  - Replacing microarray methods
  - Readout of experiments
  - barcoding

# NGS workflow

---

Library preparation

Amplification

Sequencing

Data analysis



# Next generation sequencing

---

## Two fundamentally different approaches:

1. **Clonally amplified templates** originating from a single DNA molecule
2. Directly from a **single DNA molecule**

## Range of templates:

- Genomic DNA (randomly broken into fragments of suitable size)
- DNA/RNA fragments from biological assays (ChIP, RNA from cells, etc.)
- DNA from sensor application

# Next-generation sequencing methods

---

- Roche/454
- Illumina/Solexa
- Life/APG
- Helicos BioSciences
- Polonator
- Pacific BioSciences
- Ion torrent sequencing
- Qiagen GeneReader
- Oxford Nanopore sequencing
- 10X Genomics
- ...



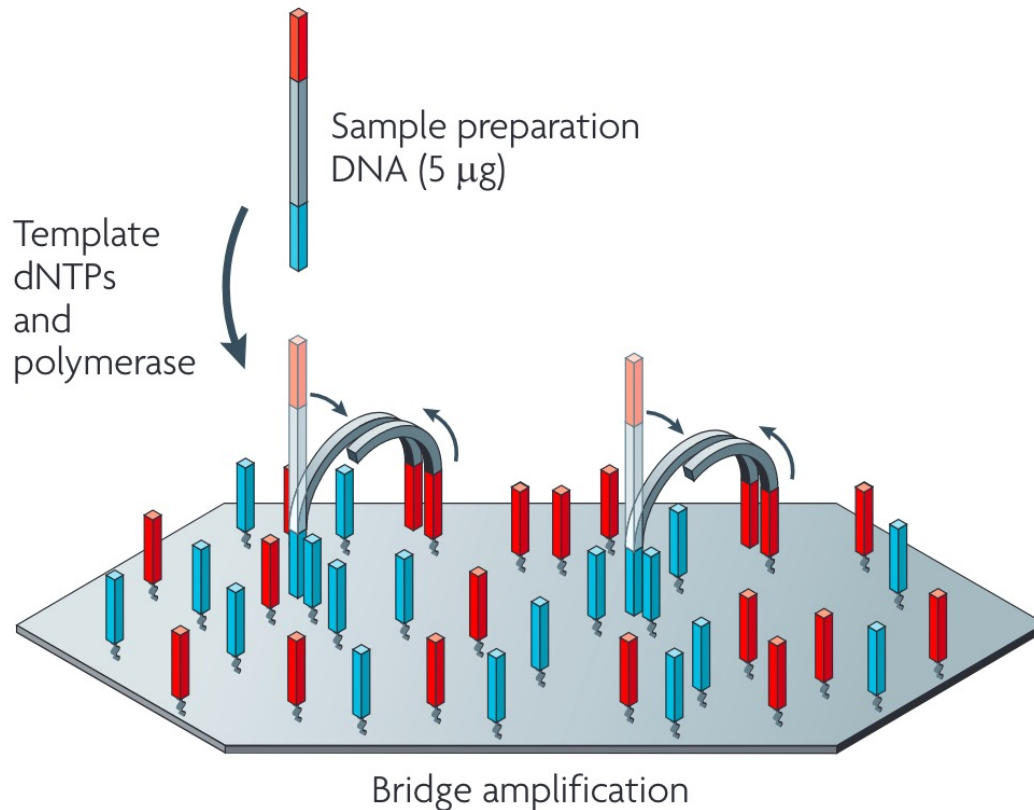
# Next-generation DNA sequencing

---

- Introduction
- **Illumina: short-read, sequencing-by-synthesis**
- Helicos: short-read, single-molecule, sequencing-by-synthesis
- PacBio: long-read, single-molecule, real-time sequencing
- Nanopore: long-read, nanopore-based electrical sensing (see last week!)

# Illumina: Solid-phase amplification

## **b** Illumina/Solexa Solid-phase amplification One DNA molecule per cluster

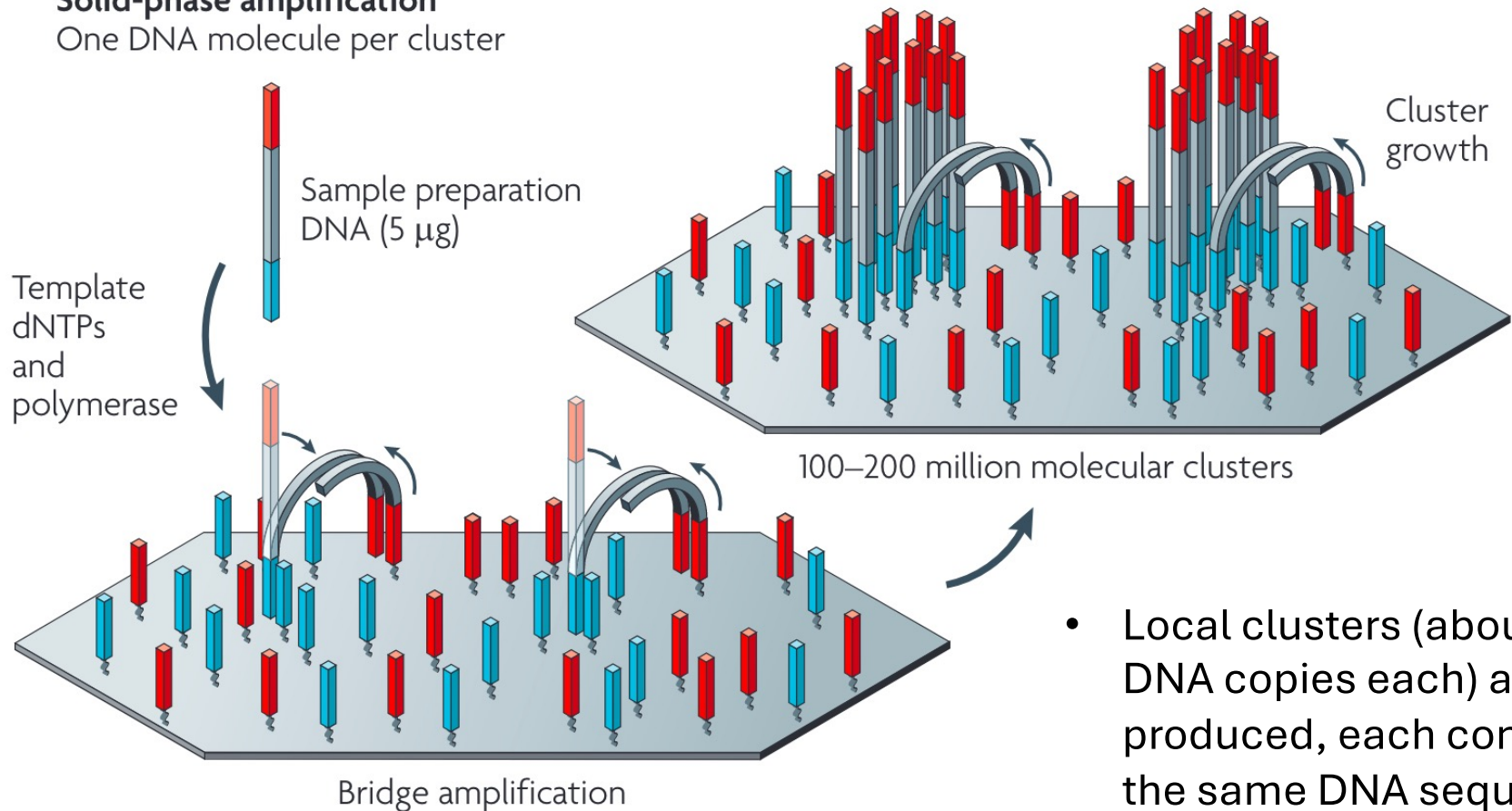


- Short DNA fragments are produced by enzymatic digestion (about 300 bp max) and standardized adapters are ligated
- Adapters contain primer binding sites, flow cell attachment sequence and indices/barcodes
- DNA strands are annealed to solid support
- High-density forward and reverse primers covalently attached (=attachment sequences)
- Ratio of the primers to template defines surface density
- Bridge amplification is employed to produce local enrichment of the same sequence

*Metzker et al. Nature Rev  
Genetics 2010*

# Illumina/Solexa: Solid-phase amplification

## b Illumina/Solexa Solid-phase amplification One DNA molecule per cluster



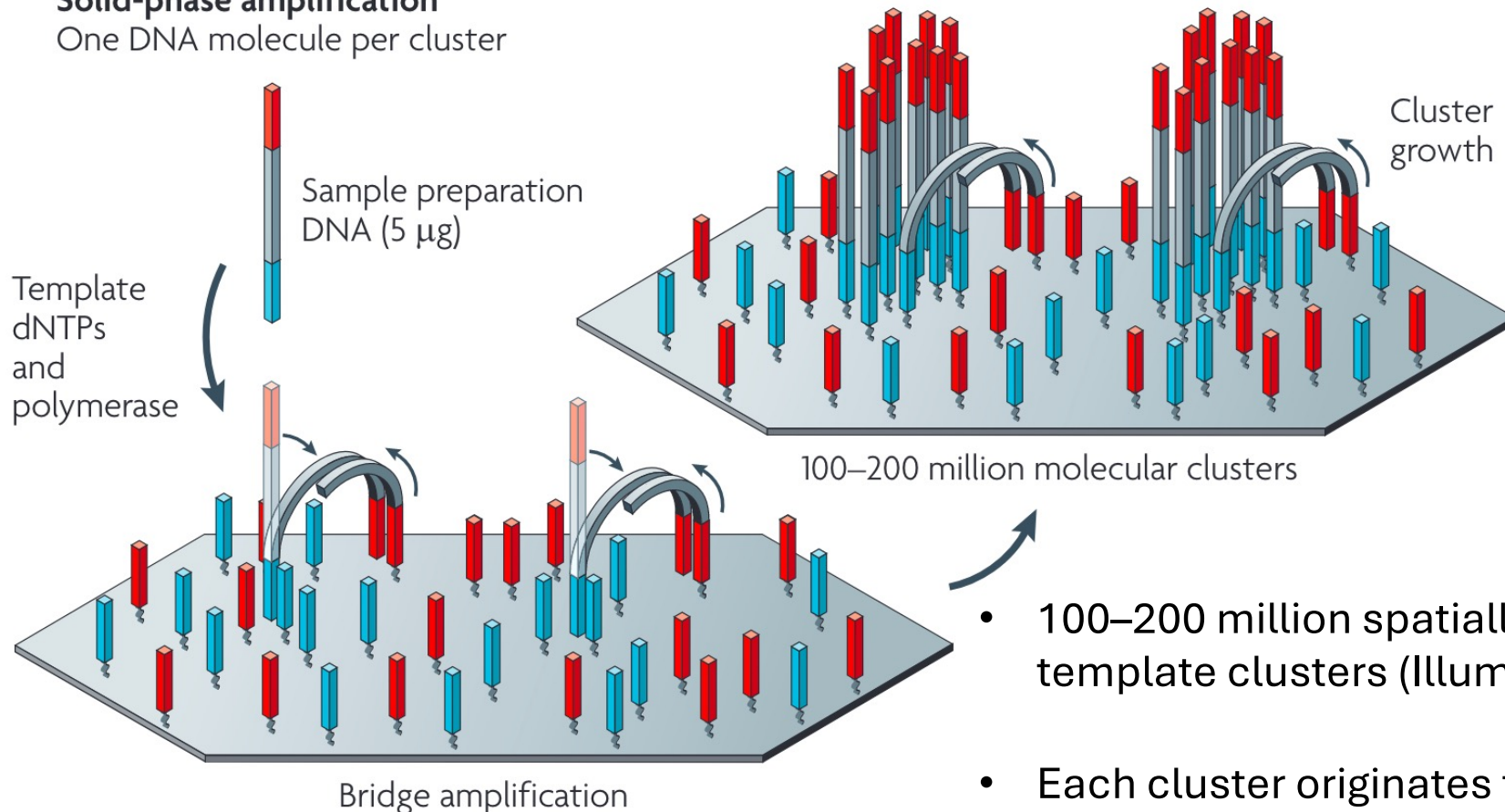
- Local clusters (about 1000 DNA copies each) are produced, each containing the same DNA sequence

→ **Clonal amplification**



# Illumina/Solexa: Solid-phase amplification

## b Illumina/Solexa Solid-phase amplification One DNA molecule per cluster



- 100–200 million spatially separated template clusters (Illumina/Solexa),
- Each cluster originates from a single molecule, and will produce a single, strong signal during sequencing.

Metzker et al. *Nature Rev Genetics* 2010

# Illumina sequencing

---

- [https://www.youtube.com/watch?v=fCd6B5HRaZ8&ab\\_channel=Illumina](https://www.youtube.com/watch?v=fCd6B5HRaZ8&ab_channel=Illumina)

# What are index reads? (see video above)

---

- Short sequences used as **unique barcodes**
- Added to each DNA sample before sequencing
- Not part of the DNA samples studied, synthetic sequences

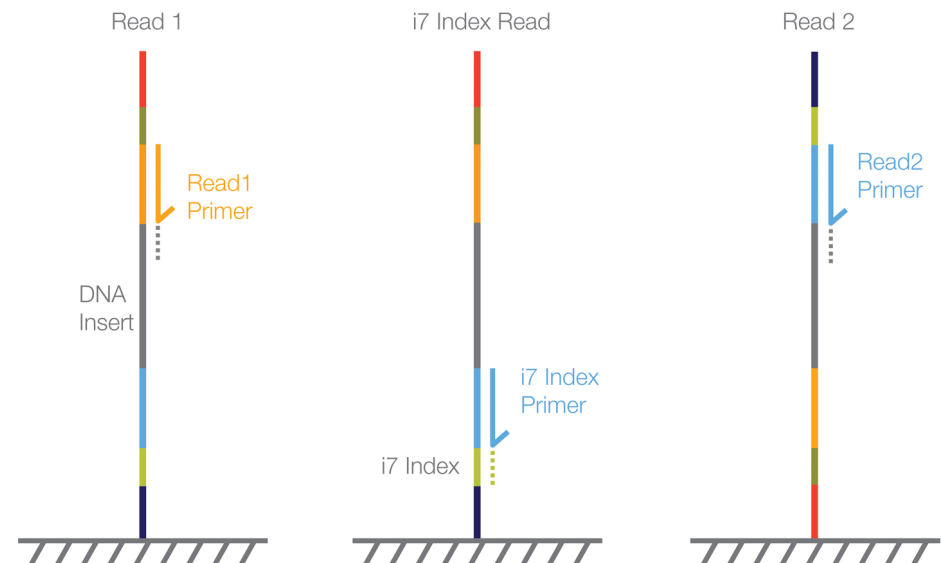
## Index 1:

- First barcode sequence added to samples during the library preparation phase
- Read after the first sequencing read (Read 1) for paired-end sequencing

## Index 2:

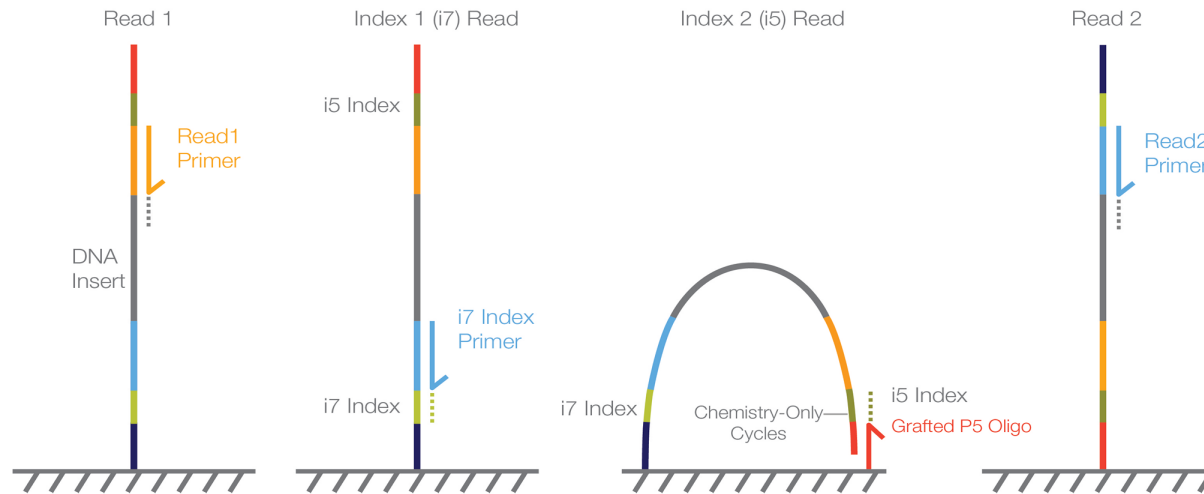
- Second barcode, also added during library preparation
- Used in a different configuration:
- For paired-end sequencing, this index is read after Read 2.

Figure 1 Single-Indexed Sequencing



# Dual-indexed sequencing on a paired-end flow cell

Figure 2 Dual-Indexed Sequencing on a Paired-End Flow Cell (Forward Strand)



1. **Read 1**—Read1 follows the standard Read 1 sequencing protocol using SBS reagents. The Read 1 sequencing primer is annealed to the template strand during the cluster generation step.
2. **Read preparation**—The Read 1 product is removed and the Index 1(i7) sequencing primer is annealed to the same template strand
3. **Index 1 (i7) Read**—Following Index Read preparation, the Index 1(i7) Read performs up to 20 cycles of sequencing. The maximum number of cycles in each Index Read depends on the system and run parameters.
4. **Index 2 (i5) Read**—The Index1(i7) Read product is removed and the template anneals to the grafted P5 primer on the surface of the flow cell. The run proceeds through an additional seven chemistry-only cycles (no imaging occurs), followed by up to 20 cycles of sequencing.
5. **Read 2 resynthesis**—The Index Read product is removed and the original template strand is used to regenerate the complementary strand. The original template strand is then removed to allow hybridization of the Read 2 sequencing primer.
6. **Read 2**—Read 2 follows the standard paired-end sequencing protocol using SBS reagents.

# Activity: Index reads

---

- Take a look at the figure in the previous slide
- Go through the process by yourself, do you have questions about the index read workflow? (3 min)
- Find someone in the room you haven't talked to yet: Help each other answer your remaining questions.
- Discuss: what are the benefits of using 1 vs. 2 index reads?



# Clonal amplification sequencing:

## Advantages and disadvantages

---

### Advantages

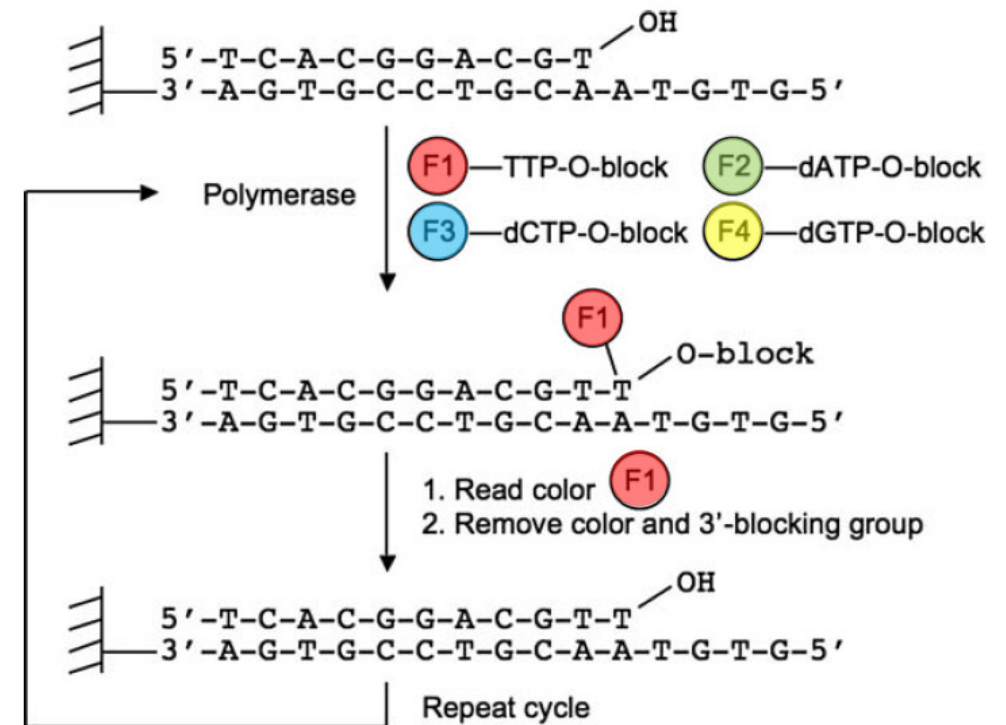
- Improved signal strengths (exponential signal increase)
- Higher accuracy and base-calling confidence
- Error correction via redundancy
- Multiplexing and high throughput
- No lost sequences

### Disadvantages

- PCR can introduce mutations, multiple coverage required
- Amplification bias due to sequence (AT and GC rich)
- Could lead to dropout of rare species
- Not compatible with long-read sequencing

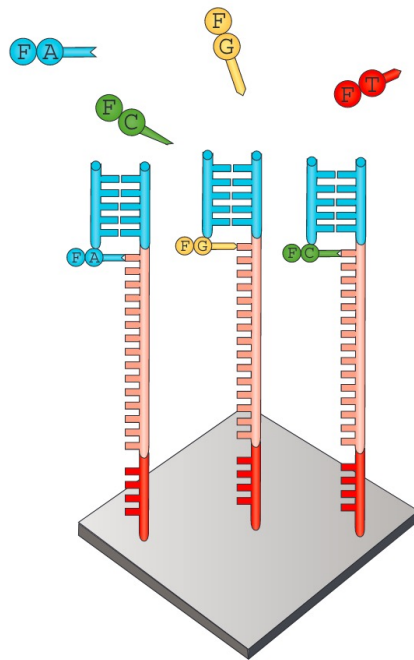
# Sequencing chemistry (Illumina/Solexa): Cyclic reversible termination (CRT)

- Addition of a nucleotide
- Wash
- Readout
- Deprotection / cleavage



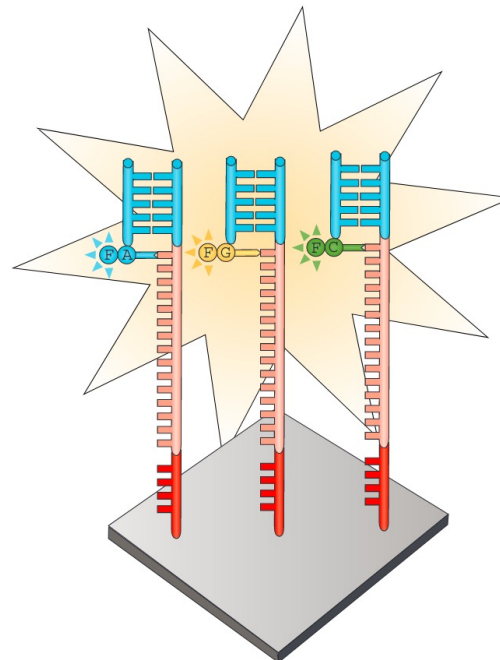
*Hutter et al., Nucleosides, Nucleotides and Nucleic Acids, 29:879–895, 2010*

# Illumina/Solexa: Sequencing by synthesis



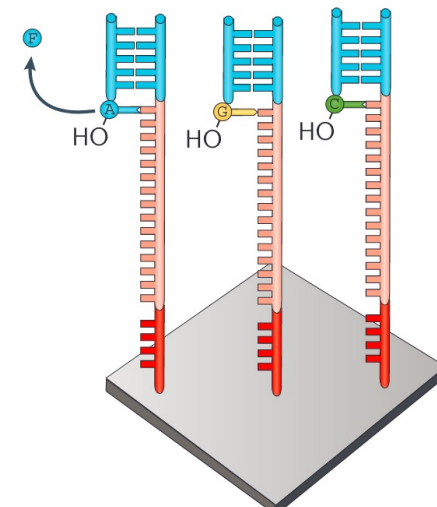
## Nucleotide addition

Fluorophore-labelled, terminally blocked nucleotides hybridize to complementary base. Each cluster on a slide can incorporate a different base.



## Imaging

Slides are imaged with either two or four laser channels. Each cluster emits a colour corresponding to the base incorporated during this cycle.

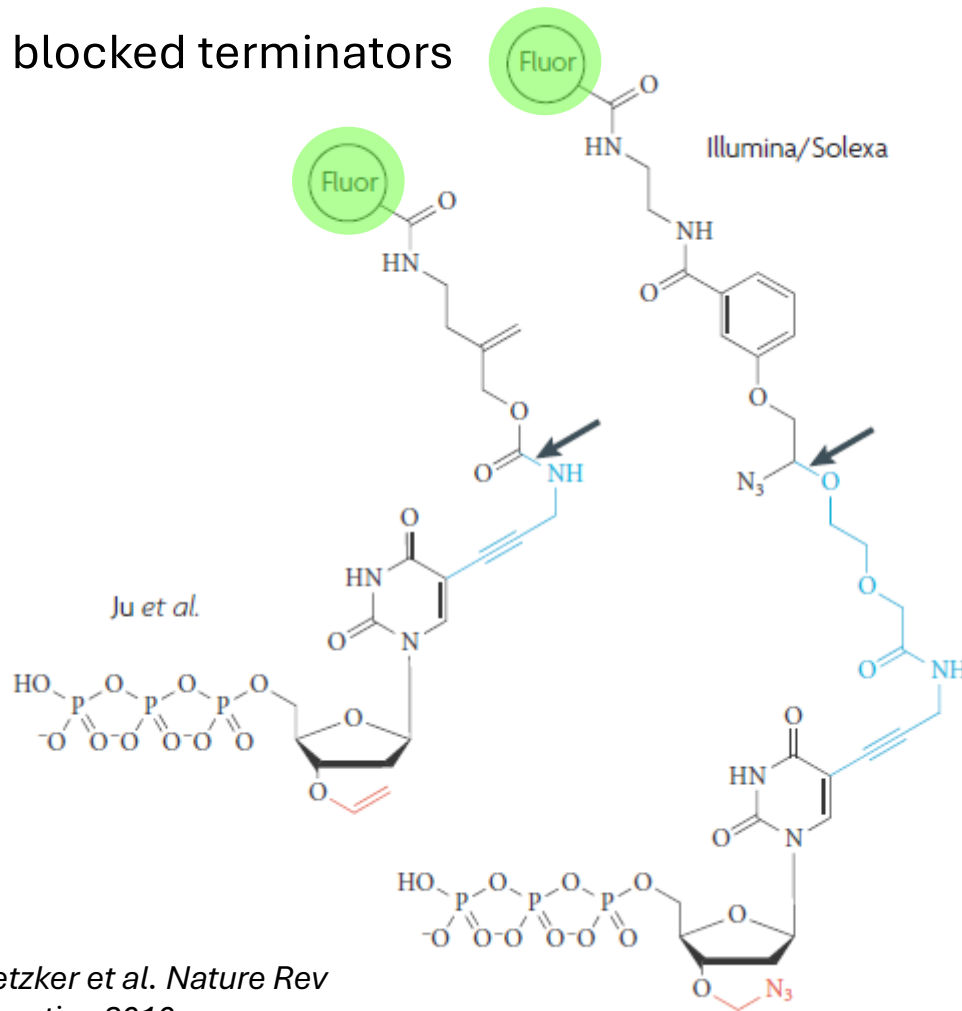


## Cleavage

Fluorophores are cleaved and washed from flow cells and the 3'-OH group is regenerated. A new cycle begins with the addition of new nucleotides.

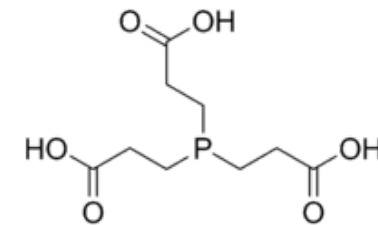
# Nucleotides used in CRT

3' blocked terminators



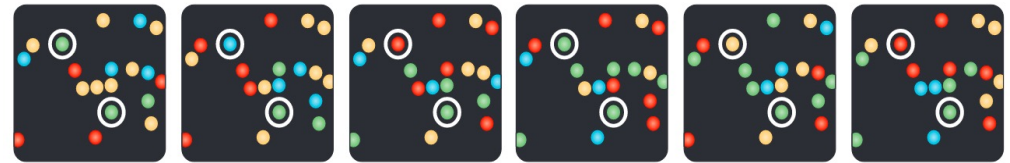
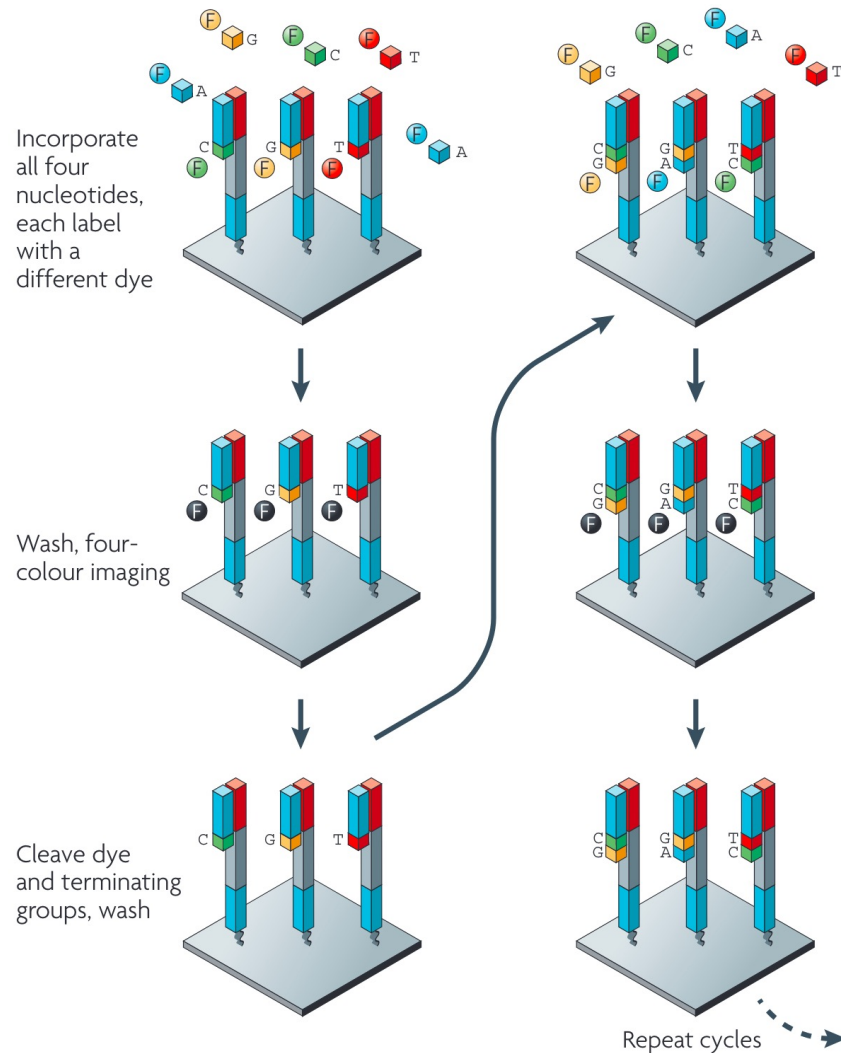
Metzker et al. Nature Rev  
Genetics 2010

- All protecting groups can be simultaneously cleaved using a reducing agent. Usually, tris(2-carboxyethyl) phosphine **TCEP**, is used



- Cleavable **fluorophore** is used for nucleotide identification
- **Blue**: residual linker/molecular scar
- **Red**: cleavable terminating functional groups
  - O-allyl group
  - 3'-azidomethyl group

# Illumina/Solexa: Cyclic reversible termination (CRT)



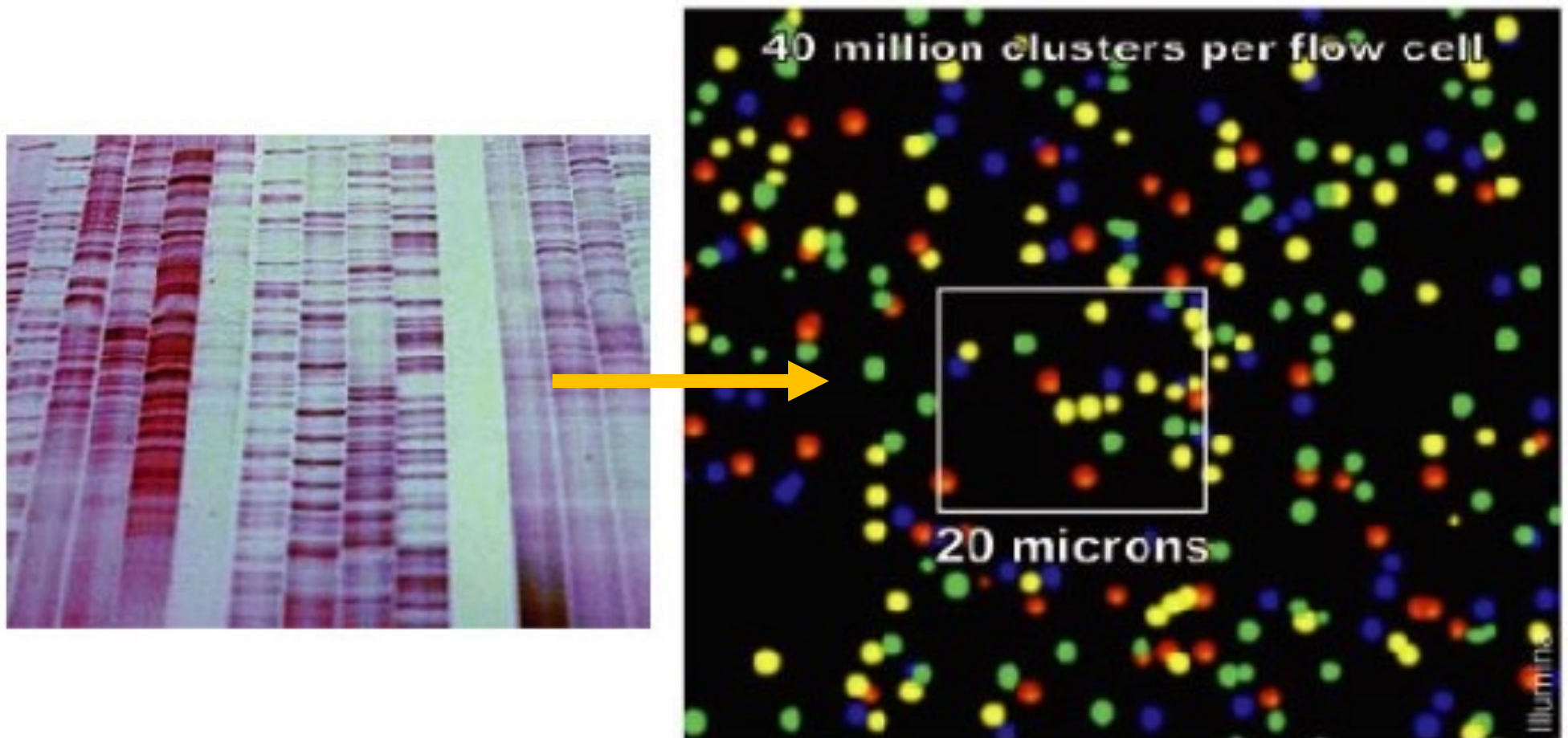
- **Elongation:** Requires modified polymerase (tolerant to modified bases)
- **Deprotection:** Reducing agent, TCEP
- **Readout / imaging:** two-laser **TIRF** imaging of fluorescence emission
- Cycle is repeated until the sequences are determined (max length: 300 bp)

Metzker et al. Nature  
Rev Genetics 2010



# From bands (Sanger) to colorful dots (Illumina)

---

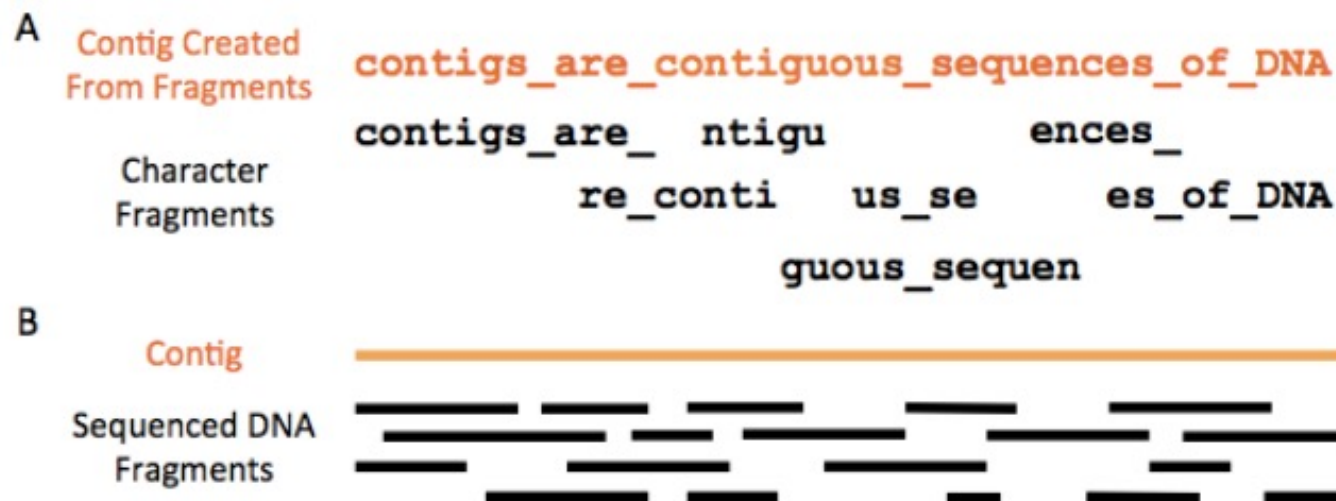


Color dots replace bands as a sequence is read with reversible terminators on a Solexa (Illumina) sequencer.

# Data analysis for NGS

---

- Alignment to reference genome
- De-novo assembly



<http://gcat.davidson.edu/phast/>

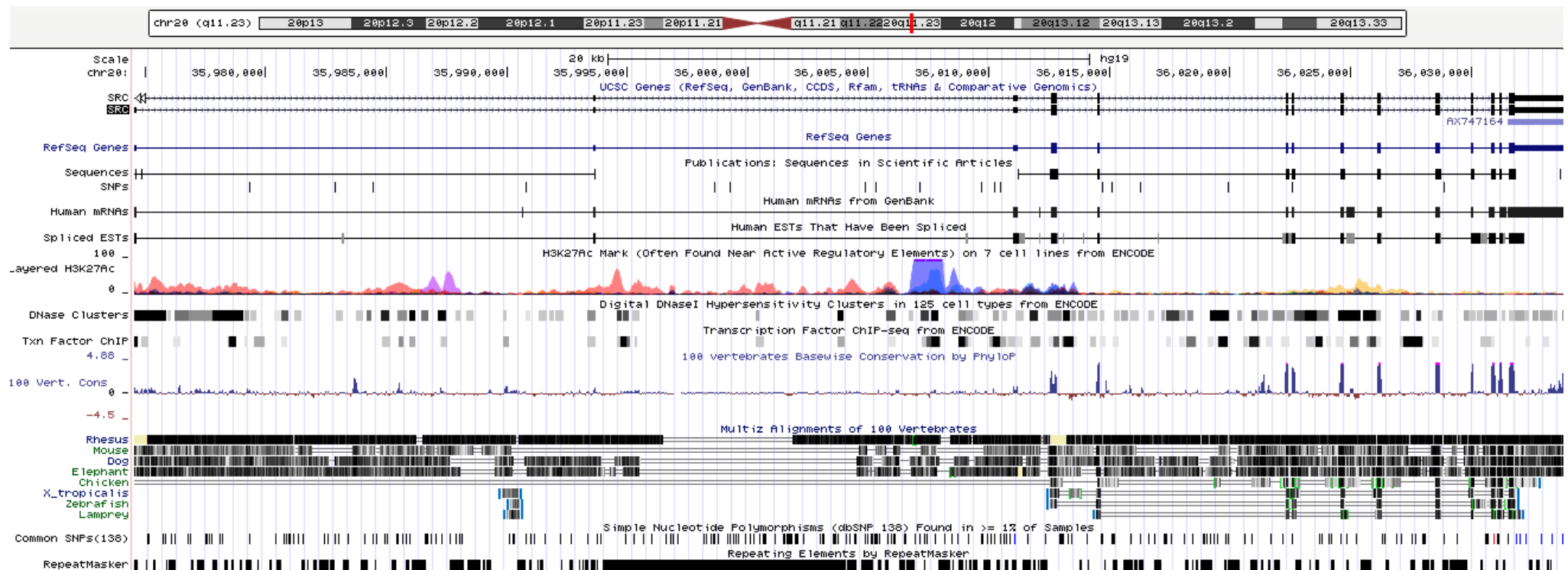
# Computational challenge



- Huge amounts of complex data
  - SNPs and structural variants
  - Sequencing errors
  - Complex search algorithms required
- Bioinformatic challenge
- From NGS reads (short and high volume) high oversampling is required to construct a genome:
  - 30-40x oversampling is required

# Applications of next gen DNA sequencing

- Whole genome sequencing
- Sequence variants
- Single cell sequencing
- cDNA, RNA, micro-RNA sequencing



# Next-generation DNA sequencing

---

- Introduction
- Illumina: short-read, sequencing-by-synthesis
- Helicos: short-read, single-molecule, sequencing-by-synthesis
- PacBio: long-read, single-molecule, real-time sequencing
- Nanopore: long-read, nanopore-based electrical sensing (see last week!)

# Next-generation DNA sequencing

---

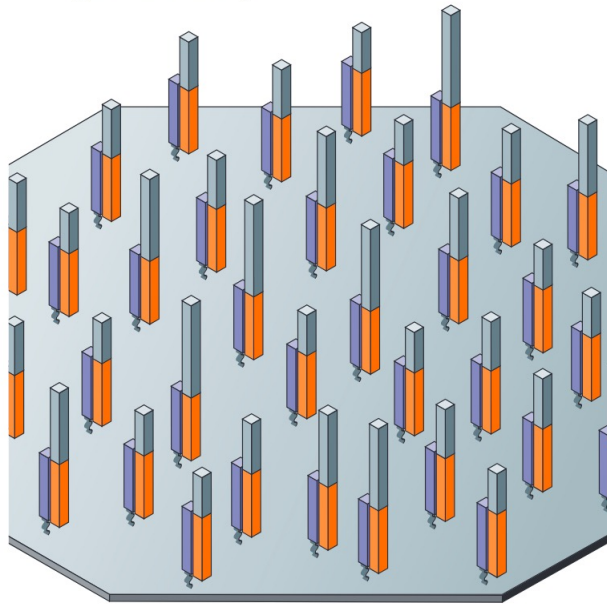
- Introduction
- Illumina: short-read, sequencing-by-synthesis
- **Helicos: short-read, single-molecule, sequencing-by-synthesis**
- PacBio: long-read, single-molecule, real-time sequencing
- Nanopore: long-read, nanopore-based electrical sensing (see last week!)



# Single-molecule templates: Helicos (defunct)

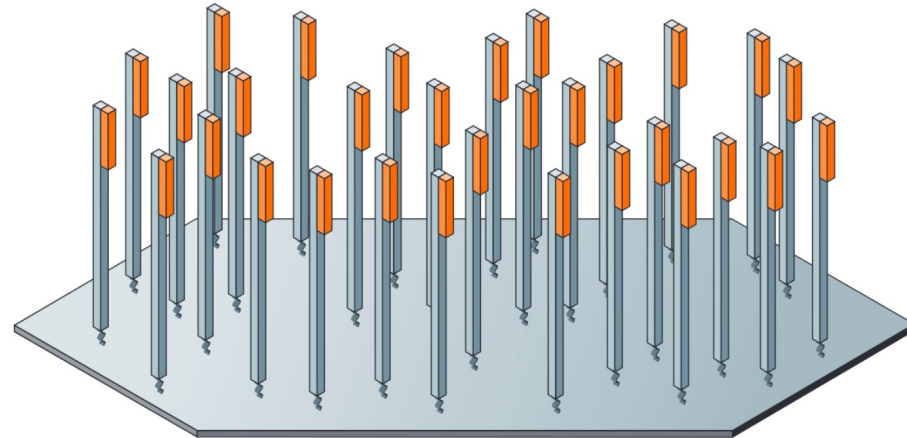
---

c Helicos BioSciences: one-pass sequencing  
Single molecule: primer immobilized



Billions of primed, single-molecule templates

d Helicos BioSciences: two-pass sequencing  
Single molecule: template immobilized



Billions of primed, single-molecule templates

## Why important?

First example of:

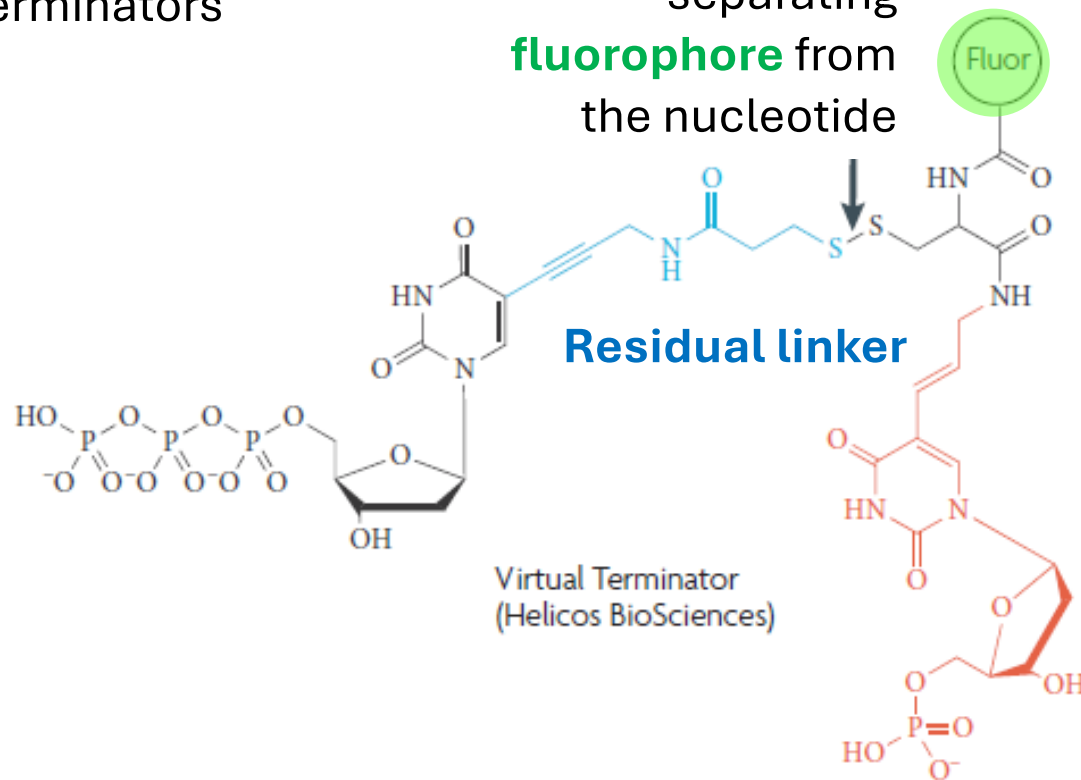
- Amplification-free sequencing
- Direct sequencing of RNA molecules
- Single-molecule sensitivity



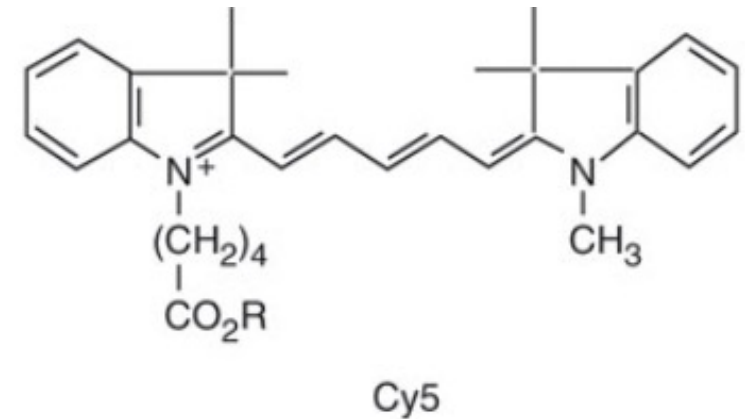
# Single-molecule CRT

3' unblocked terminators

Site of cleavage:  
separating  
**fluorophore** from  
the nucleotide



fluorophore



Deprotection by  
reduction

**Steric hindrance:** inhibitory function

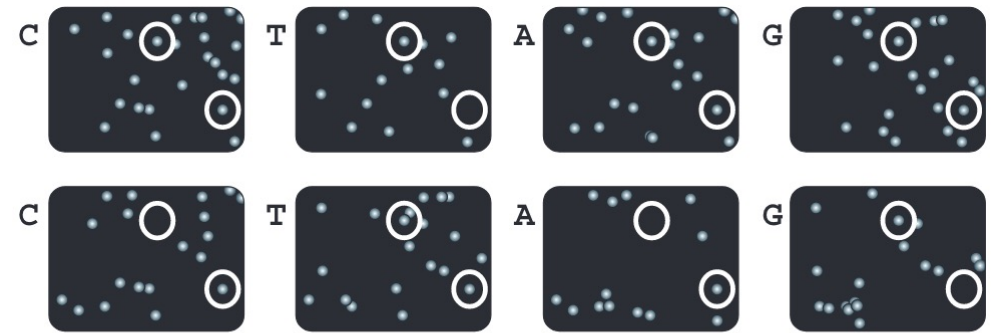
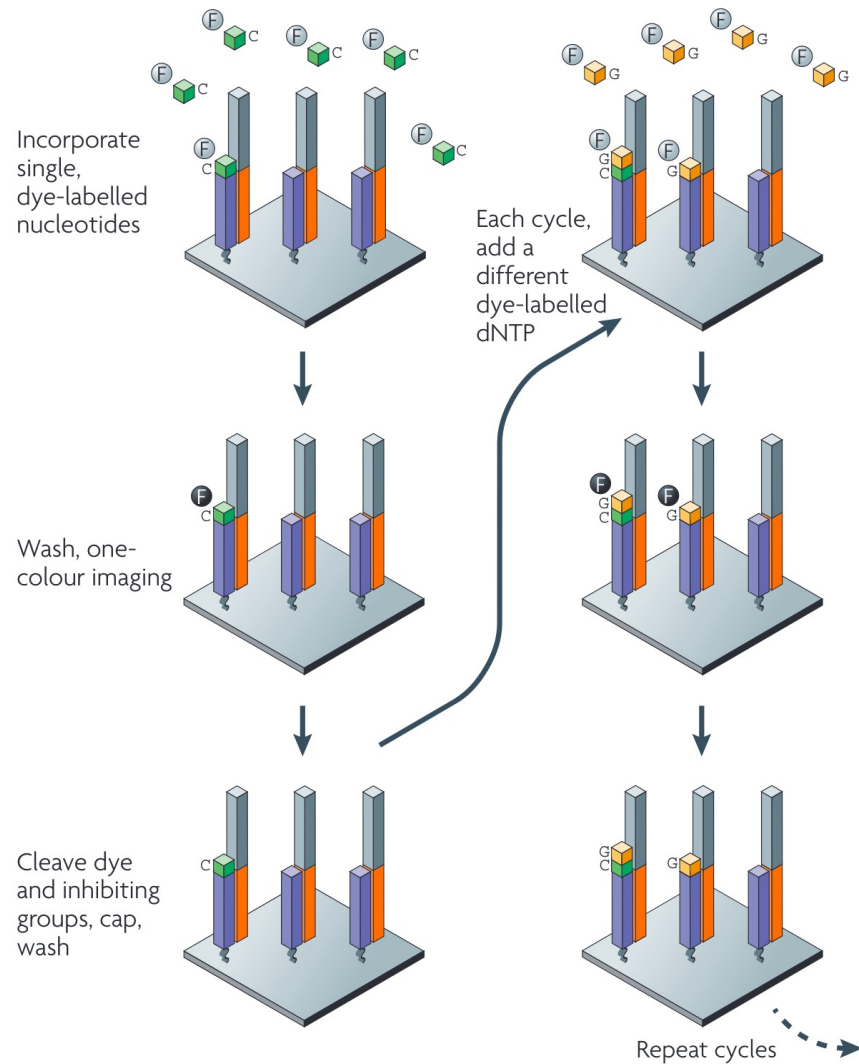
*Metzker et al. Nature  
Rev Genetics 2010*

# Activity: Comparing Illumina and Helicos sequencing methods

---

- Take a look at the figure on the next slide (1-color CRT, Helicos)
- What is the same, what is different compared to Illumina/Solexa CRT sequencing (see previous slides)?
- Why is only one dye needed for Helicos sequencing?
- What are the implications of using only one color vs. multiple colors?
- Given what you've learned, why might Helicos sequencing have struggled commercially, despite its innovative design?

# 1-color CRT, Helicos



Top: CTAGTG  
Bottom: CAGCTA

- **Elongation:** Less stringent for the polymerase
- **Deprotection:** Only one step
- **Readout / imaging:** single molecule Cy5 TIRF imaging

Metzker et al. *Nature Rev Genetics* 2010

# Next-generation DNA sequencing

---

- Introduction
- Illumina: short-read, sequencing-by-synthesis
- Helicos: short-read, single-molecule, sequencing-by-synthesis
- **PacBio: long-read, single-molecule, real-time sequencing**
- Nanopore: long-read, nanopore-based electrical sensing (see last week!)

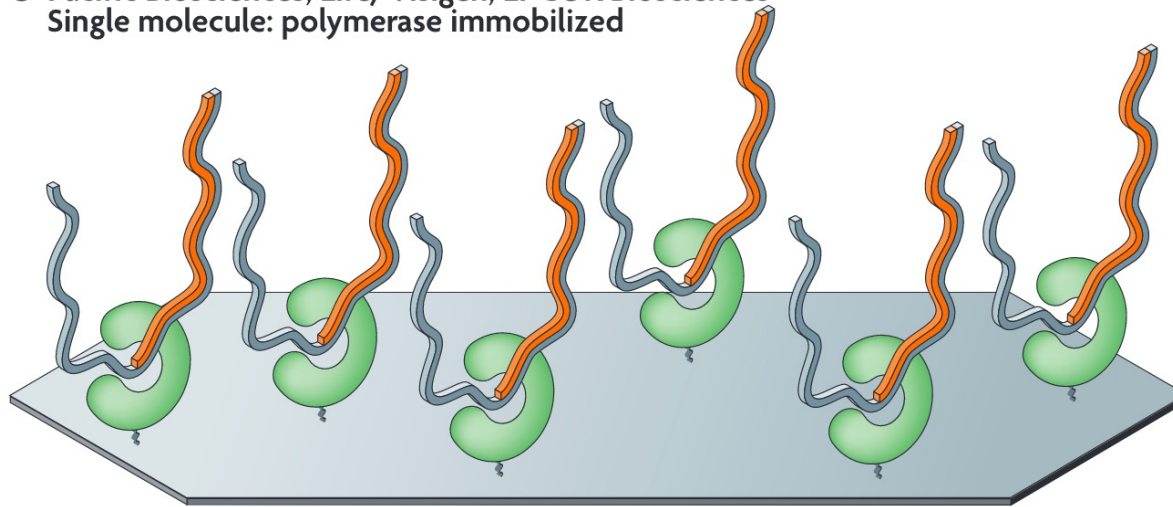
# PacBio sequencing

---

- [https://www.youtube.com/watch?v=\\_lD8JyAbwEo&ab\\_channel=PacBio](https://www.youtube.com/watch?v=_lD8JyAbwEo&ab_channel=PacBio)  
O

# Single-molecule real-time sequencing: PacBio

e Pacific Biosciences, Life/Visigen, LI-COR Biosciences  
Single molecule: polymerase immobilized



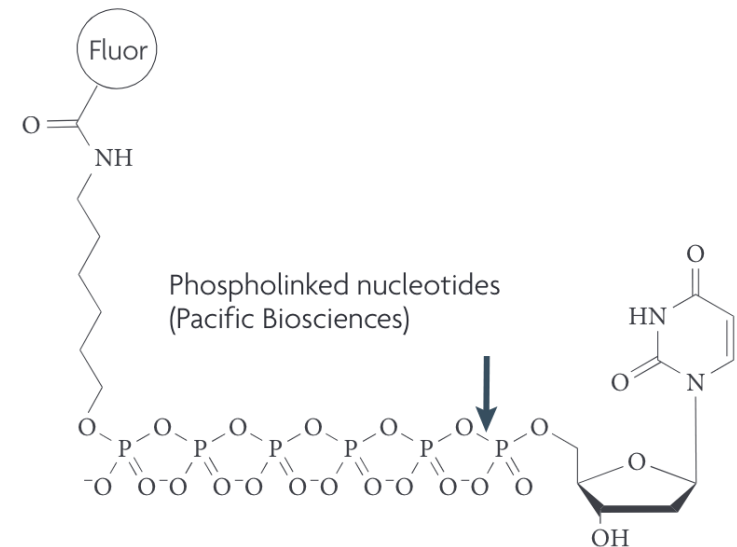
Thousands of primed, single-molecule templates

## Phospholinked nucleotide (hexaphosphate linker)

- Hexaphosphate slows down incorporation process slightly
- PacBio DNA pol was engineered to recognize and incorporate these modified nucleotides
- Hexaphosphate group serves as a cleavable linker

*Metzker et al. Nature  
Rev Genetics 2010*

- Immobilization of DNA polymerase molecules
- Observation of the **incorporation of single, fluorescently labeled nucleotides** (no reversible terminators)



# Single-molecule real-time sequencing: PacBio

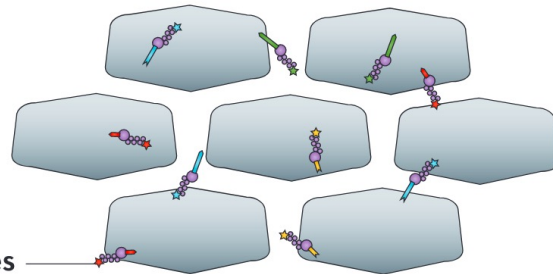
## SMRTbell template

Two hairpin adapters allow continuous circular sequencing



## ZMW wells

Sites where sequencing takes place

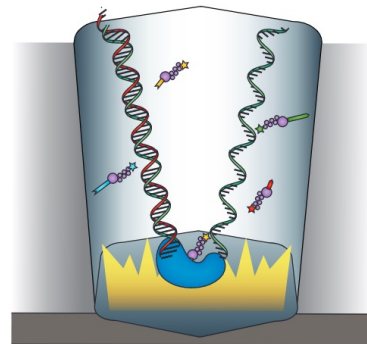


## Labelled nucleotides

All four dNTPs are labelled and available for incorporation

## Modified polymerase

As a nucleotide is incorporated by the polymerase, a camera records the emitted light



## PacBio output

A camera records the changing colours from all ZMWs; each colour change corresponds to one base



## Single-stranded circular DNA:

- Template fragments processed and ligated to hairpin adapters at each end
- Primers and  $\phi$ 29 DNA polymerase are attached to ssDNA regions on SMRTbell template

## Zero mode waveguides (ZMW):

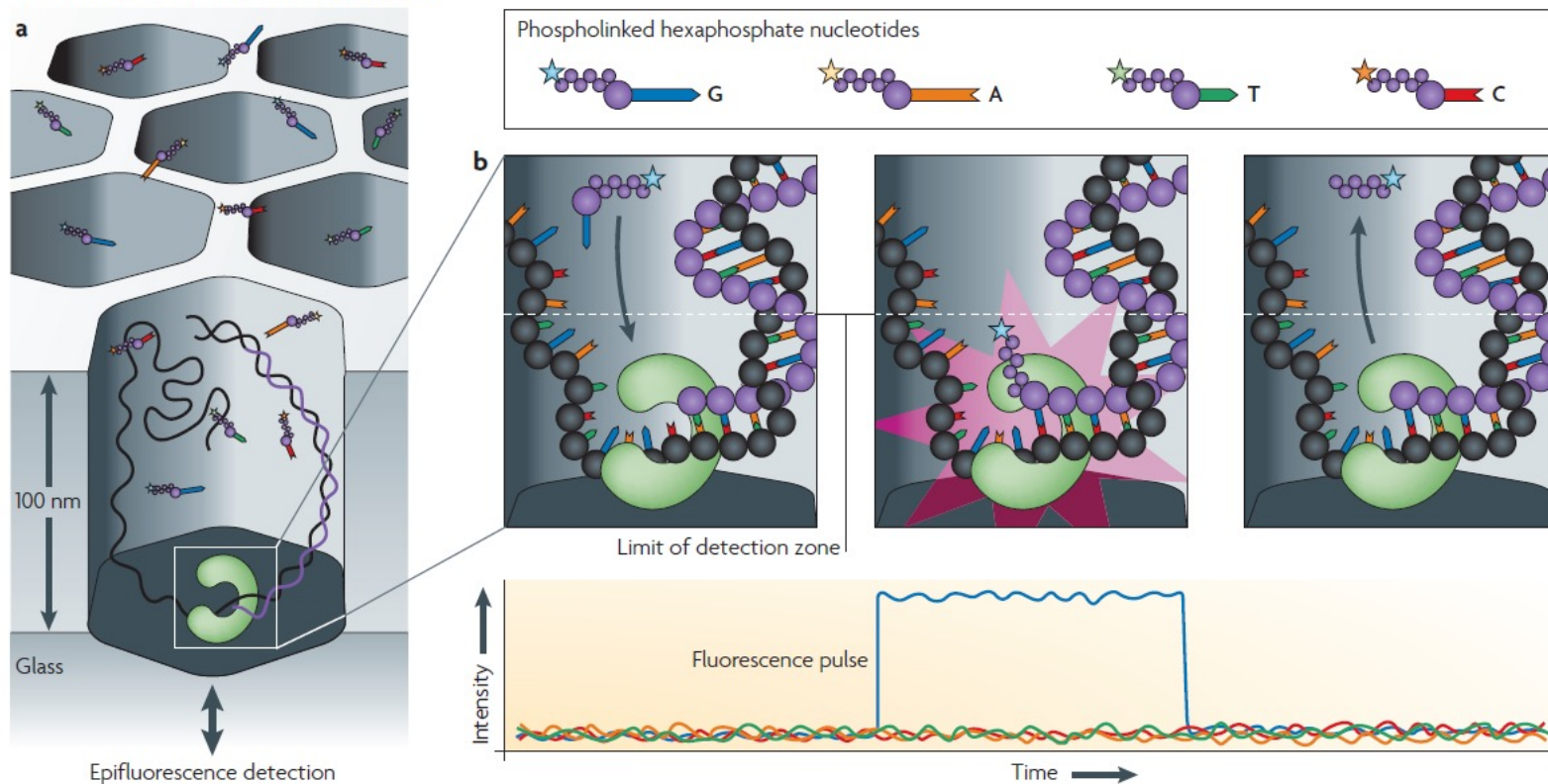
- Cavity in thin metal film with
- Reduction of illumination volume to zeptoliter size
- Allows to observe molecules at high concentration, **one at a time**

*Goodwin et al. Nature Rev Genetics 2016*



# Real-time sequencing

Pacific Biosciences — Real-time sequencing



- Residence time of phospholinked nucleotides in the active site: millisecond scale
- Long enough to record a **fluorescent pulse**
- Released, dye-labeled pentaphosphate by-product quickly diffuses away

# Performance of real-time sequencing

---

- Average read-length: several thousands of nucleotides
- Error sources:
  - Deletions
  - Insertions
  - Mismatches
- Mostly due to very fast reaction times
- **Accuracy** per run: 83%
  - Requires multiple sequencing runs for the template

# Next-generation DNA sequencing

---

- Introduction
- Illumina: short-read, sequencing-by-synthesis
- Ion Torrent (Thermo Fisher): short-read, semiconductor sequencing
- Helicos: short-read, single-molecule, sequencing-by-synthesis
- PacBio: long-read, single-molecule, real-time sequencing
- **Nanopore: long-read, nanopore-based electrical sensing (see last week!)**

# Summary of NGS methods

---

Feature	Illumina	Helicos	Ion Torrent	PacBio (HiFi)	Oxford Nanopore
Detection method	Fluorescence (SBS)	Fluorescence (single-molecule SBS)	pH sensing (semiconductor)	Fluorescence (real-time SMRT)	Ionic current through nanopore
Read length	75–300 bp	25–55 bp	100–400 bp	10–25 kb (HiFi reads)	10–100+ kb
Accuracy (raw)	>99.9%	~98%	~98–99%	>99.9% (HiFi)	~95–98% (raw); improving
Throughput	Very high (up to terabases/run)	Moderate (low density vs. modern SBS)	Moderate	Moderate	Variable (scalable)
Run time	Hours to days	~1-2 days	Hours	~24 hrs	Real-time (minutes–hours)
Library amplification	Yes	No (single molecule)	Yes	No (SMRTbell prep)	No
Molecule type	Amplified DNA	Native DNA	Amplified DNA	Native DNA or RNA	Native DNA or RNA
Strengths	Accuracy, throughput, cost per base	Amplification-free, early SMS platform	Fast, inexpensive instruments	Long reads + high accuracy	Ultra-long reads, portability
Limitations	Short reads, PCR bias	Short reads, discontinued platform	Lower accuracy, pH drift	Higher cost per base, lower throughput	Higher error rate, context bias
Best for	WGS, RNA-Seq, panels, clinical use	Single-molecule transcription profiling	Targeted panels, quick runs	Isoform sequencing, genome assembly	Structural variation, field use
Manufacturer	Illumina	Helicos BioSciences (defunct)	Thermo Fisher	PacBio	Oxford Nanopore Technologies

---

# Connections

---

- [https://connections.swellgarfo.com/game/-NvBOf2\\_nOaAEYOnwTU5](https://connections.swellgarfo.com/game/-NvBOf2_nOaAEYOnwTU5)