



# ГЕНОМИКА

## 18 марта 2020, Среда

---

### 2. Технология секвенирования ДНК:

- по Sanger
- next-generation sequencing (NGS) technology
- полногеномное *de novo* (whole genome *de novo* sequencing)
- ресеквенирование (resequencing)
- целевое (target)
- метагеномное (community (metagenomics) sequencing)

# Nitrogenous, Nucleoside and Nucleotide Bases

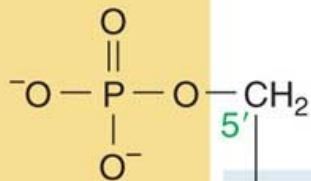
## Азотистое основание

(Азотсодержащее гетероциклическое соединение)

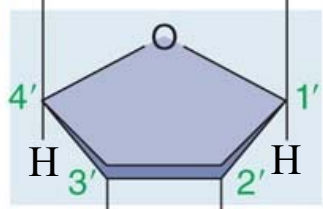
### 1 Structure of nucleotide

(nucleoside with a phosphate at 5' carbon)

Phosphate group



Nitrogenous base



OH in RNA

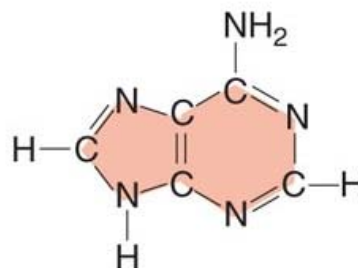
H in DNA

Sugar

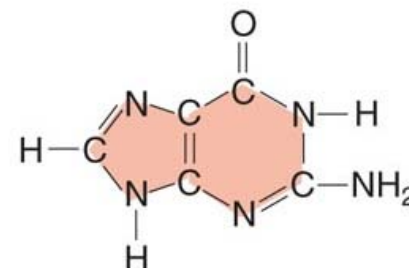
**Nucleoside**

(nitrogenous base linked to a 2-deoxy-D-ribose at 1' carbon)

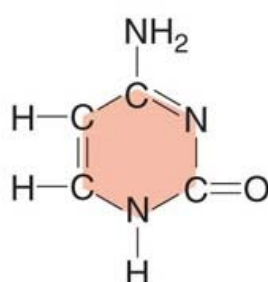
### 2 Nitrogenous bases



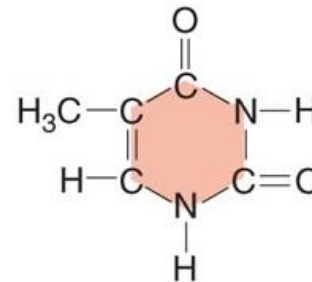
Adenine (**A**)



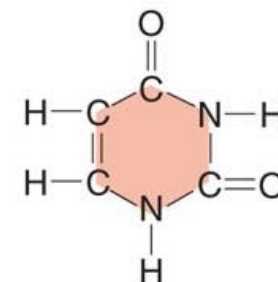
Guanine (**G**)



Cytosine (**C**)

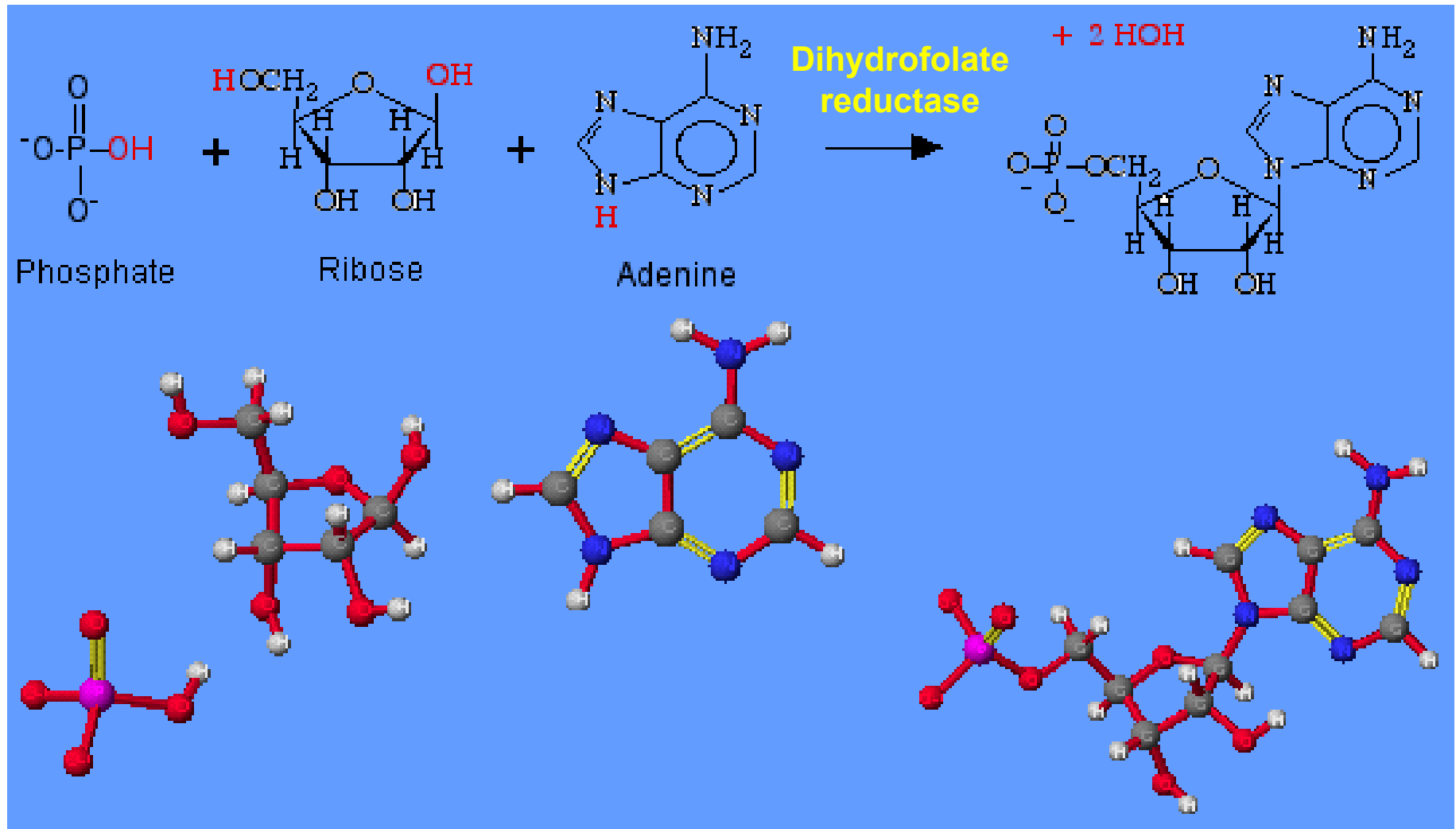


Thymine (DNA only) (**T**)



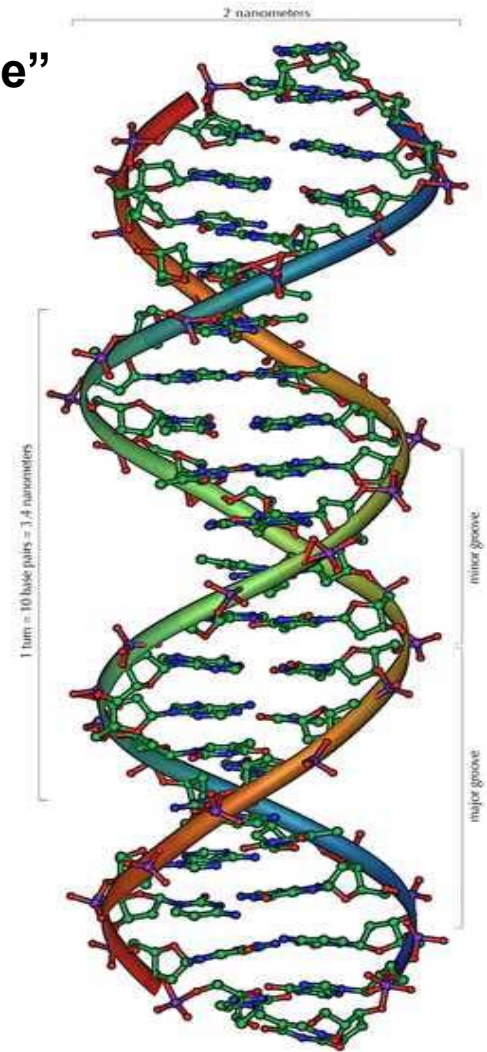
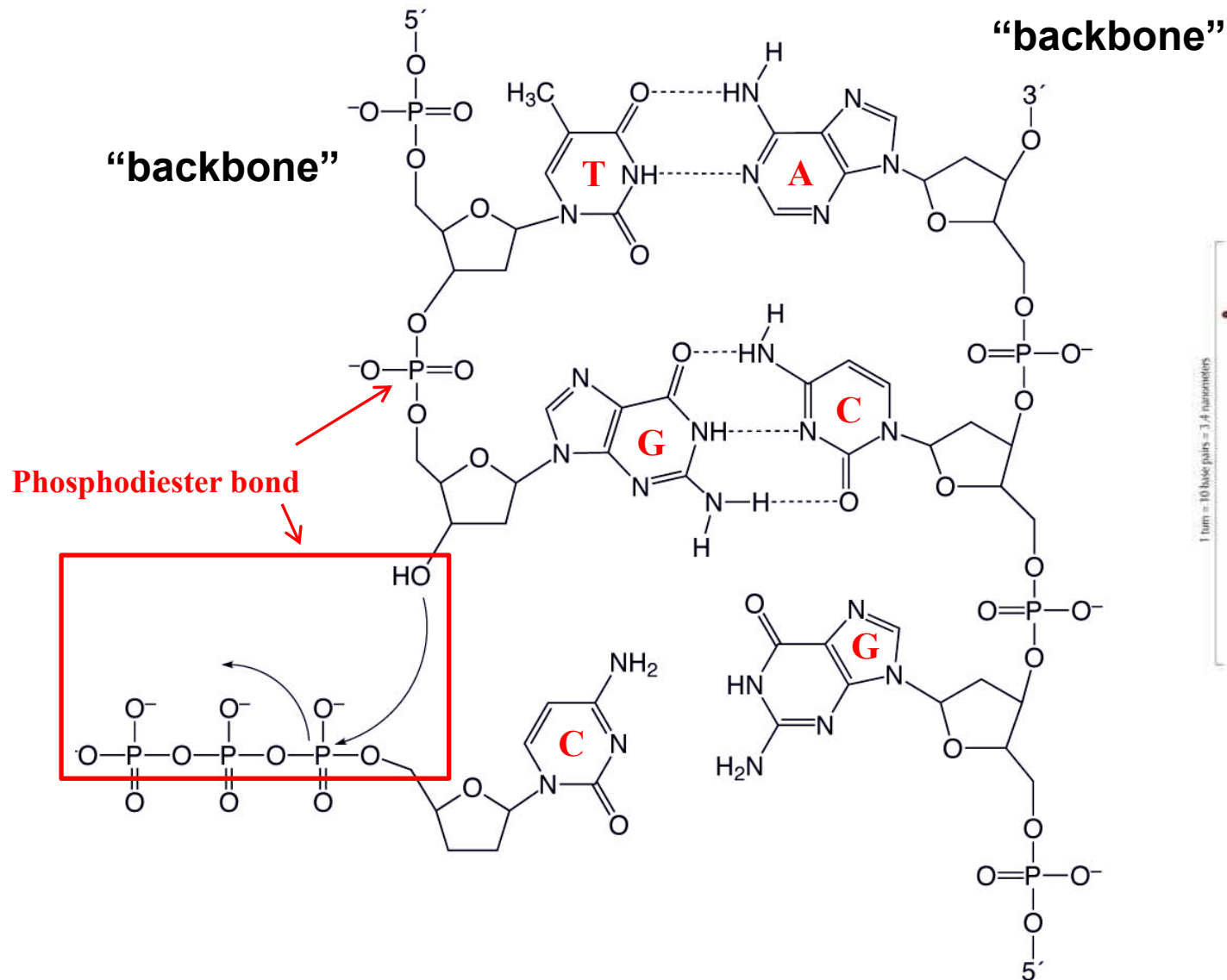
Uracil (RNA only) (**U**)

# Synthesis of a Nucleotide: Adenosine monophosphate (AMP)



# Nucleotide polymer

- DNA Polymerase



# DNA Sequencing Methods

---

## 1<sup>st</sup> generation sequencing methods:

- chemical degradation of nucleotides method (Allan Maxam and Walter Gilbert, 1977)
- chain-termination or dideoxy method (Frederick Sanger, 1977)

# Next-Generation Sequencing (NGS) Methods

---

## 2<sup>nd</sup> generation high-throughput massively parallel shotgun sequencing methods:

### – sequencing-by-synthesis methods:

- ❖ optical pyrosequencing by Jonathan Rothberg (Margulis et al. 2005) (originally **454 Corp.**, a subsidiary of **CuraGen Corp.**, then **454 Life Sciences**, a subsidiary of **Roche Diagnostics**) based on fixing fragmented (nebulized) and adapter-ligated DNA fragments to small DNA-capture beads PCR amplified in a water-in-oil emulsion in a PicoTiterPlate, and then sequenced using DNA polymerase, ATP sulfurylase, and luciferase (to generate light for detection of the individual nucleotides added to the template DNA) and adding sequentially 4 DNA nucleotides in a fixed order using the Genome Sequencer FLX Instrument (there was also a downscaled Junior version);
- ❖ optical based on fluorescently labelled reversible dye-terminators, the DNA are extended one nucleotide at a time followed by image acquisition - the camera takes images of the fluorescently labeled nucleotides, then the dye along with the terminal 3' blocker is chemically removed from the DNA, allowing the next cycle (**Solexa**, now **Illumina Inc.**) using Illumina GAIL, HiSeq, MiSeq and NextSeq instruments;
- ❖ nonoptical semiconductor sequencing developed by **Ion Torrent Systems Inc.** founded by Jonathan Rothberg (acquired by **Applied Biosystems Inc.** that became **Life Technologies Corp.** as a merger of **Invitrogen Corp.** and **Applied Biosystems Inc.** in 2008; now **Thermo Fisher Scientific Corp.** since 2014), based on the detection of hydrogen ions that are released during the DNA synthesis, as opposed to the optical methods used in other sequencing systems (Ion Torrent and Ion Proton Sequencers - Personal Genome Machines);

- sequencing-by-ligation method developed by **Applied Biosystems Inc.** (now **Thermo Fisher Scientific Corp.** since 2014), the DNA is amplified by emulsion PCR and then sequenced using the SOLiD Instrument.





# DNA Sequencing Methods

---

## 3<sup>rd</sup> generation single molecular (SM) based sequencing methods:

### – sequencing-by-synthesis methods:

- ❖ **optical SM** method by **Helicos Biosciences** (Cambridge, MA) uses bright fluorophores and laser excitation to detect base addition events from individual DNA molecules fixed to a surface, eliminating the need for molecular amplification. First, DNA fragments with added poly-A tail adapters are attached to the flow cell surface. The next steps involve extension-based sequencing with cyclic washes of the flow cell with fluorescently labeled nucleotides (one nucleotide type at a time, as with the Sanger method). The reads were short, up to 100 bp per run, and on November 15, 2012, Helicos BioSciences filed for Chapter 11 bankruptcy (purchased recently by SeqLL, LLC; [seqll.com](http://seqll.com));
- ❖ **optical SM real time (SMRT)** sequencing by **Pacific Biosciences** is based on the sequencing by synthesis of the DNA in zero-mode wave-guides (ZMWs) - small well-like containers with the signal capturing tools located at the bottom of the well using DNA polymerase attached to the ZMW bottom and producing reads of up to 15-20 Kbp, with mean read lengths of 2.5-2.9 Kbp;

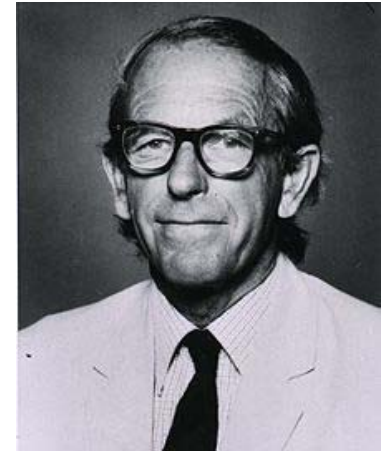
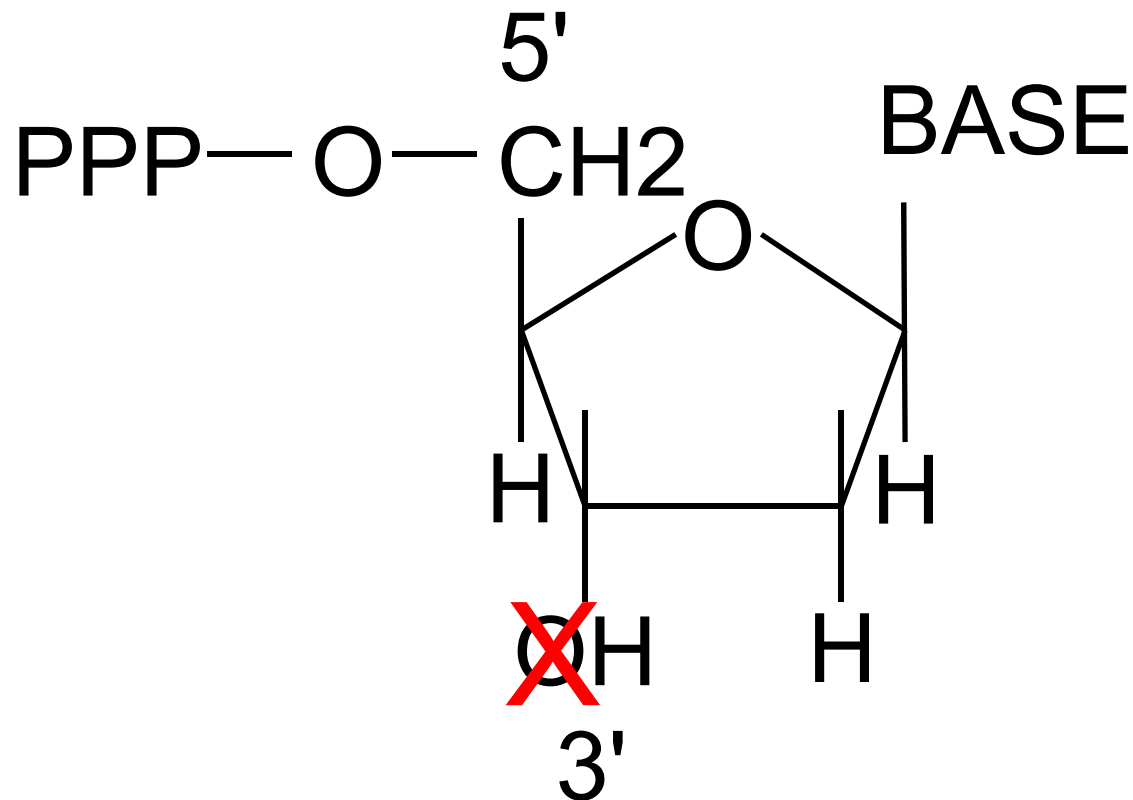
## 4<sup>th</sup> generation single molecular (SM) based sequencing methods:

**Roche Nanopore Sequencing** and **Oxford Nanopore Technologies** are based on the readout of electrical signals occurring at nucleotides passing through artificially manufactured pores in membranes, for example, in case of **Oxford Nanopore Technologies** the pores are created by the pore-forming protein  $\alpha$ -hemolysin covalently bound with cyclodextrin within the nanopore that will bind transiently to the DNA molecule being detected. The DNA passing through the nanopore changes its ion current. Each type of the nucleotide blocks the ion flow through the pore for a different period of time (GridION and PromethION systems and miniaturised MinION instrument).



# Dideoxy (Sanger) Method

**Dideoxynucleotides** without a hydroxyl group at 3'-end of ribose, which prevents strand extension, are used together with normal nucleotides in the sequencing reaction:



**Frederick Sanger**  
1918 –2013



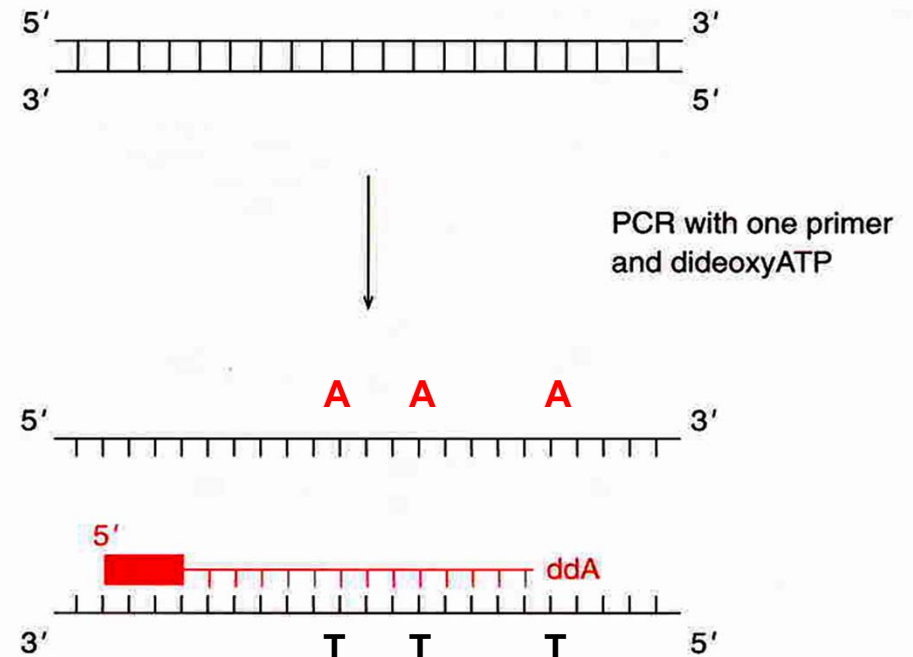
# Dideoxy (Sanger) Method

As a result there is a mixture of fragments of different size in a sequencing reaction:

## Sequencing cycle:

- 94°C: DNA denaturing
- 45°C: primer annealing
- 60-72°C: thermostable DNA polymerase primer extension

Repeat 25-35 times



After 4 cycles



**ddATP** in the reaction: a **ddA** will be occasionally added to the growing strand whenever there is a **T** in the template strand

# How to visualize DNA fragments?

---

- **Radioactivity**

- radioactive isotope labeled primers (kinase with  $^{32}\text{P}$ )
- radioactive isotope labeled dNTPs (gamma  $^{35}\text{S}$  or  $^{32}\text{P}$ )

- **Fluorescence**

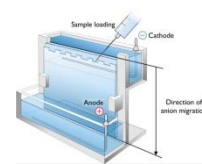
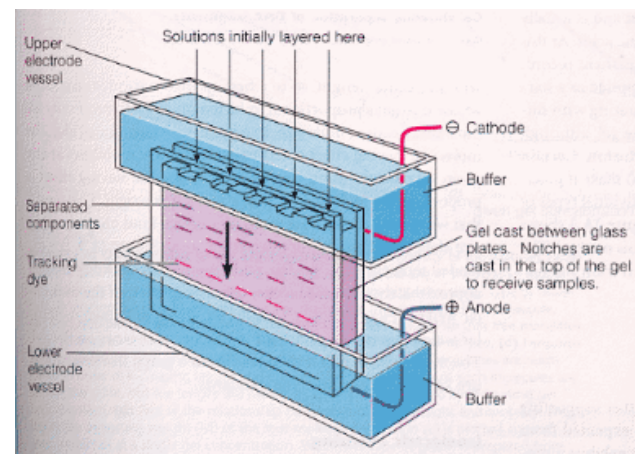
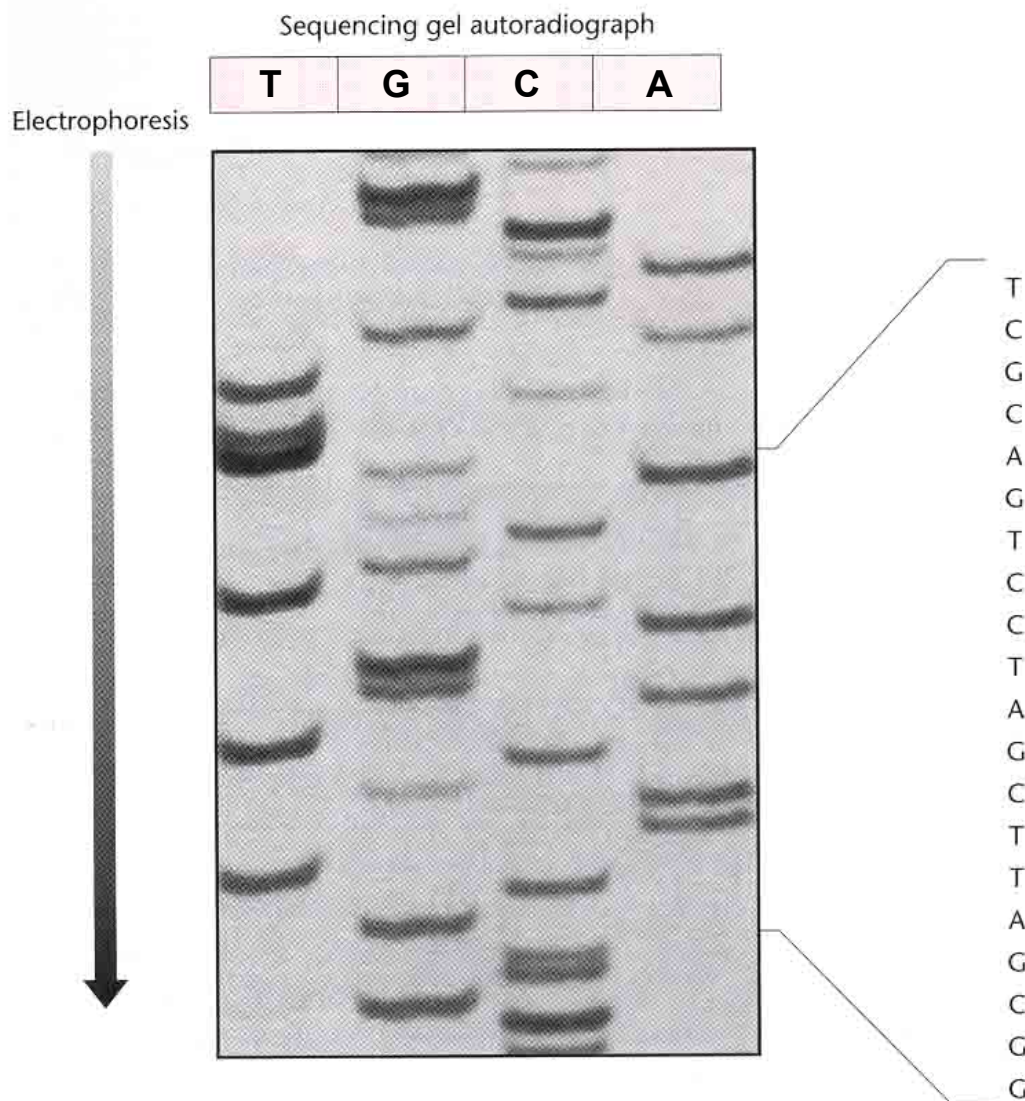
- ddNTPs chemically synthesized to contain fluorescent dye
- each ddNTP with a different fluorescent dye emits a light at a different wavelength allowing its identification

Then, fragments for both methods are separated in high resolution **polyacrylamide gel electrophoresis**

- slab gels
- capillary gels: require only a tiny amount of sample to be loaded, run much faster than slab gels, have higher resolutions, automated (best for high throughput sequencing)



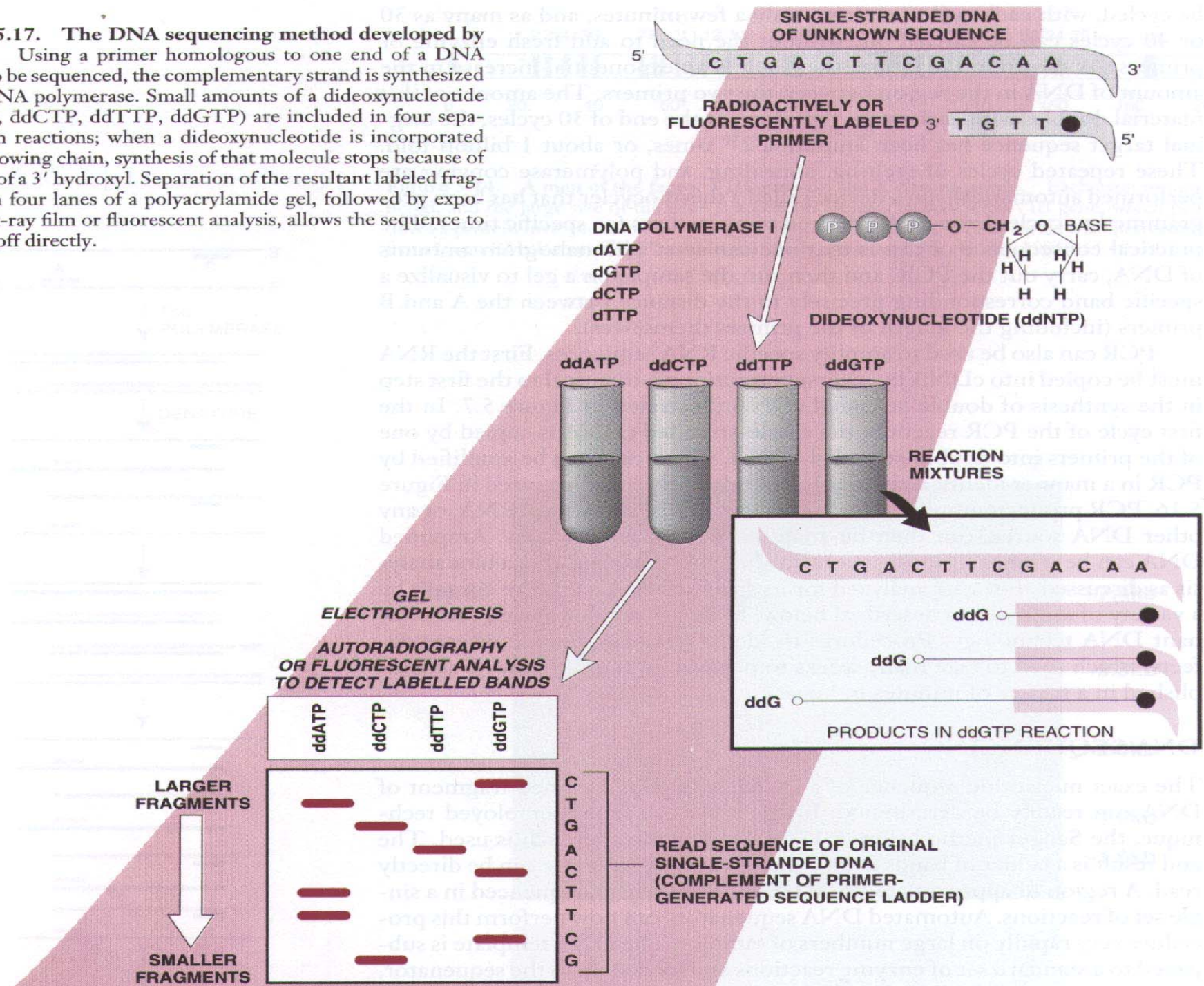
# Sequencing gel electrophoresis of DNA fragments amplified using radioactive isotope labeled primers or ddNTP



- Radioactively labelled primer or dNTPs in sequencing reaction
- 4 different ddNTPs used in 4 separate reactions, respectively
- Separation of sequencing products by slab gel electrophoresis
- Autoradiography

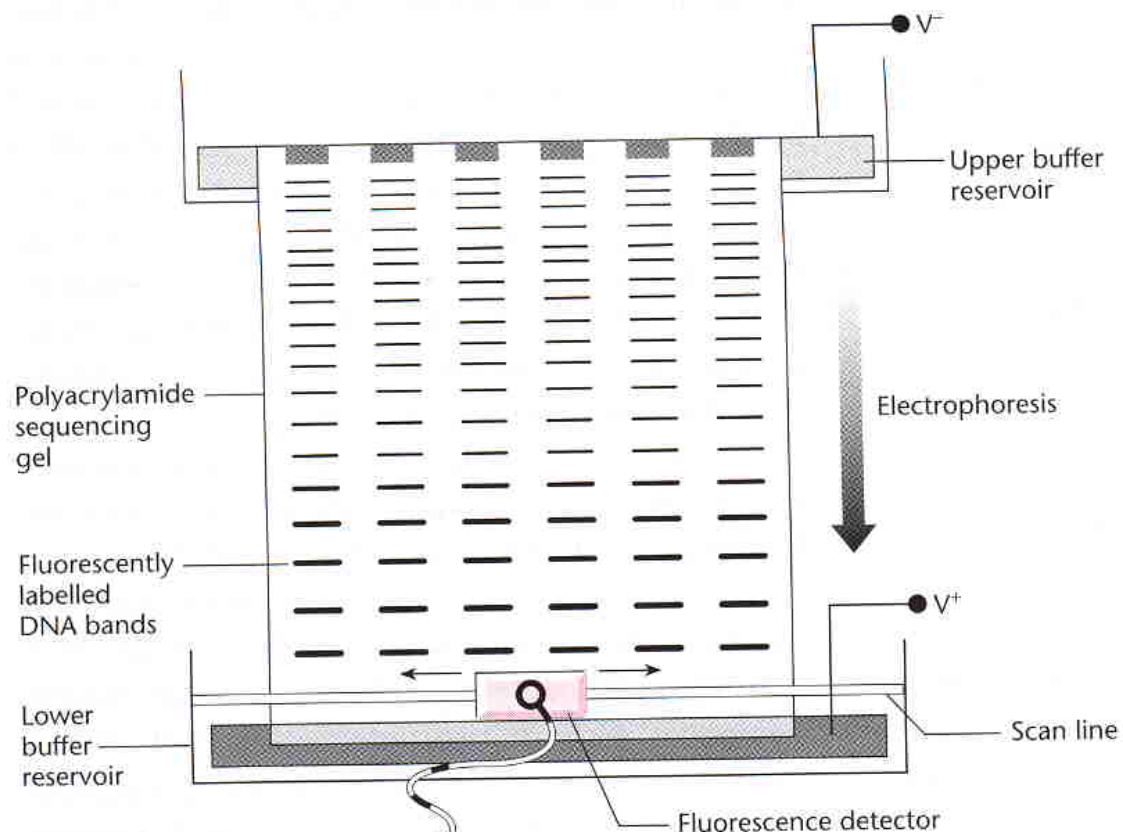


**Figure 5.17. The DNA sequencing method developed by Sanger.** Using a primer homologous to one end of a DNA strand to be sequenced, the complementary strand is synthesized using DNA polymerase. Small amounts of a dideoxynucleotide (ddATP, ddCTP, ddTTP, ddGTP) are included in four separate such reactions; when a dideoxynucleotide is incorporated into a growing chain, synthesis of that molecule stops because of the lack of a 3' hydroxyl. Separation of the resultant labeled fragments in four lanes of a polyacrylamide gel, followed by exposure to x-ray film or fluorescent analysis, allows the sequence to be read off directly.

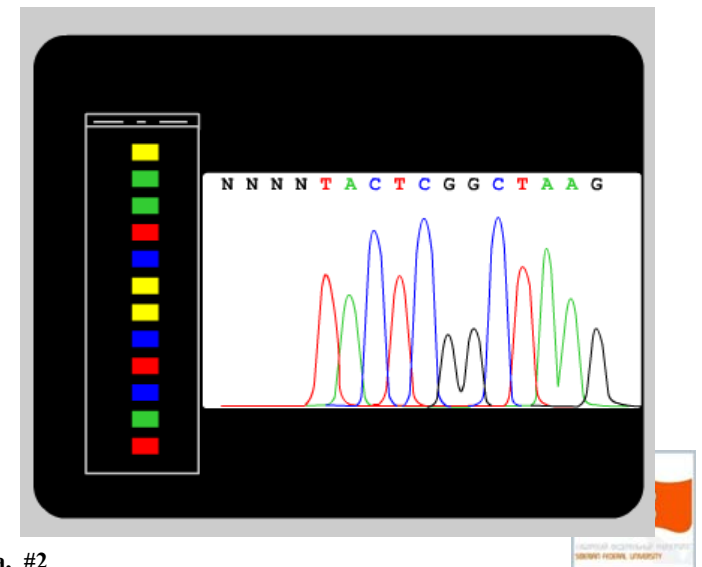
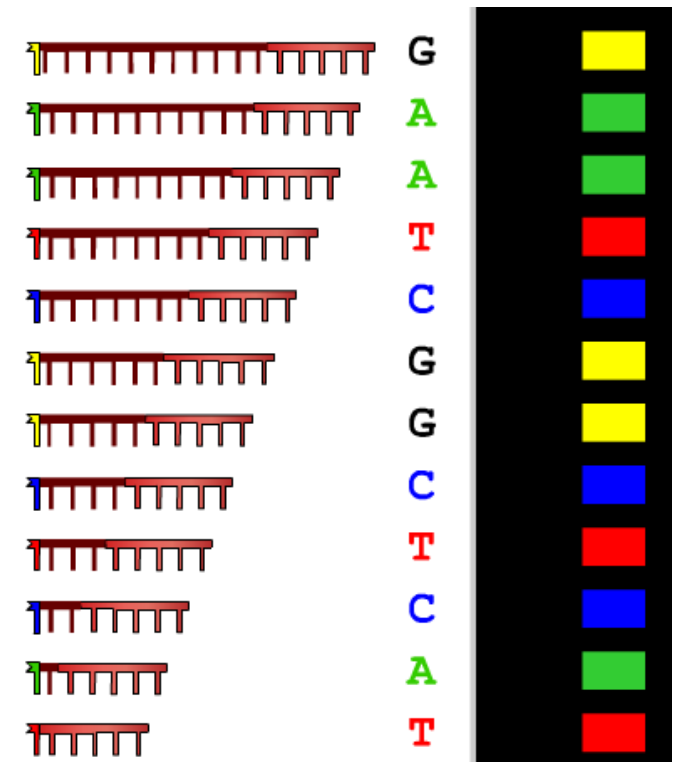


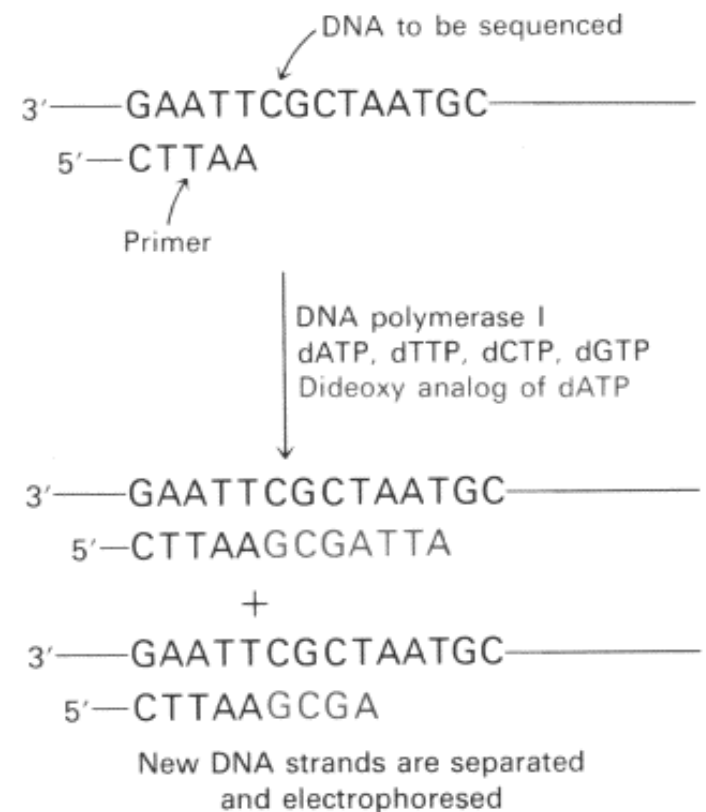
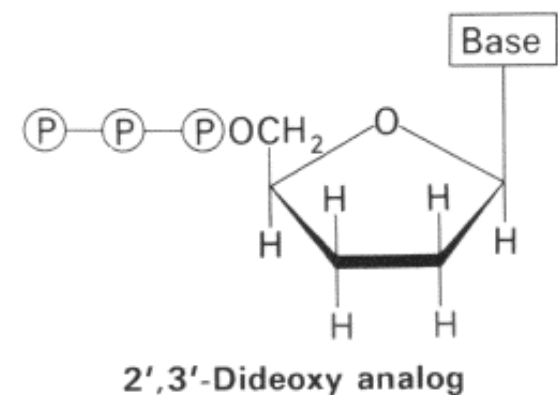
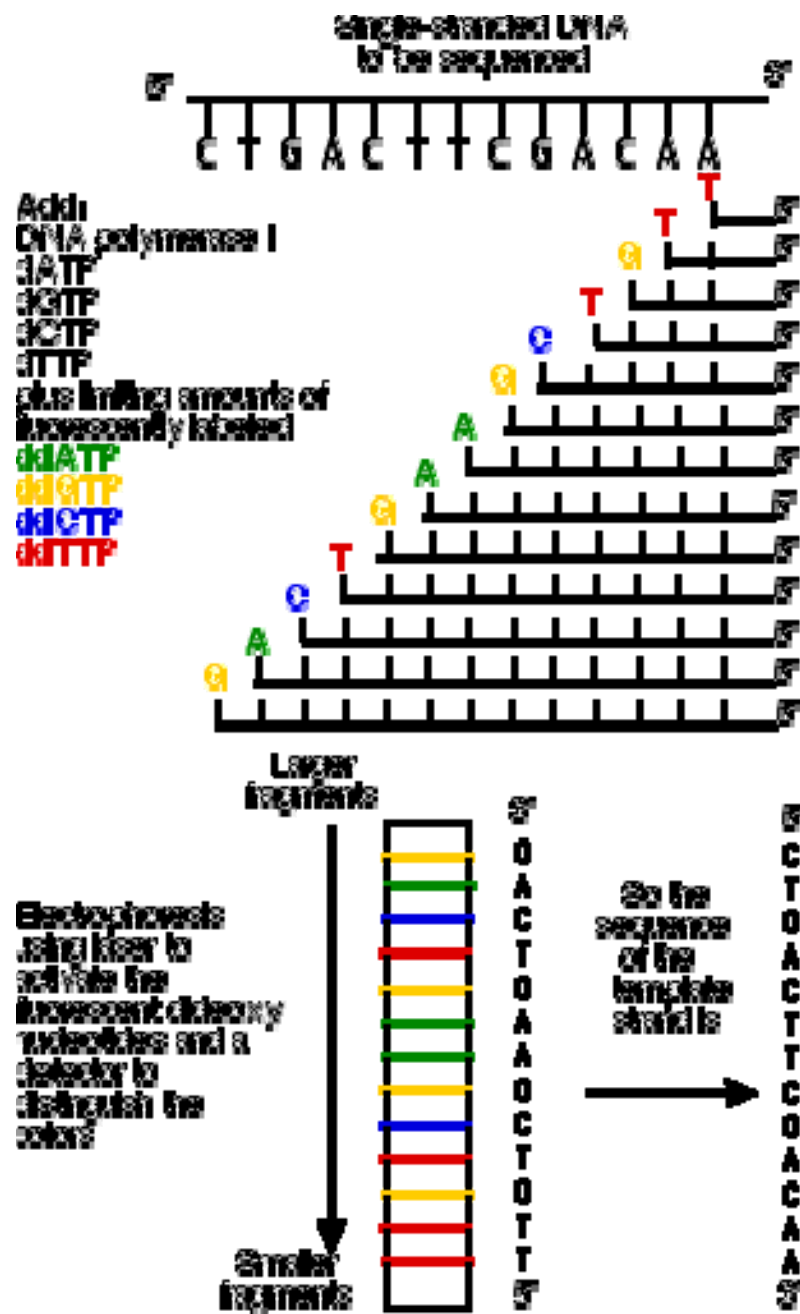
# An automated sequencer using mixed ddNTPs labelled by 4 different fluorescent dyes

## Slab or capillary gel electrophoresis:



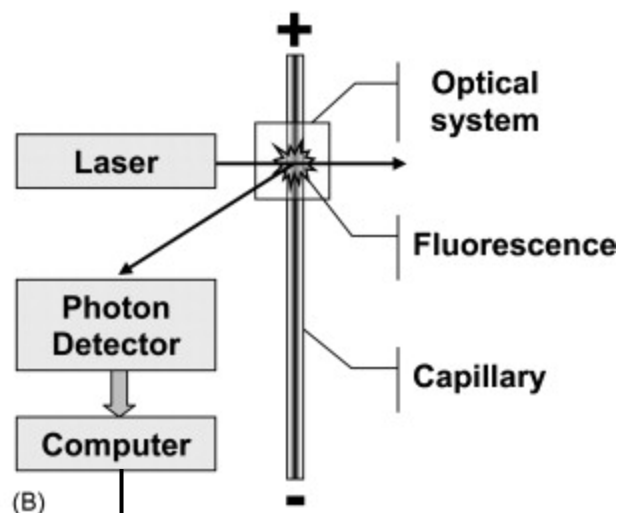
## Single lane or capillary output:



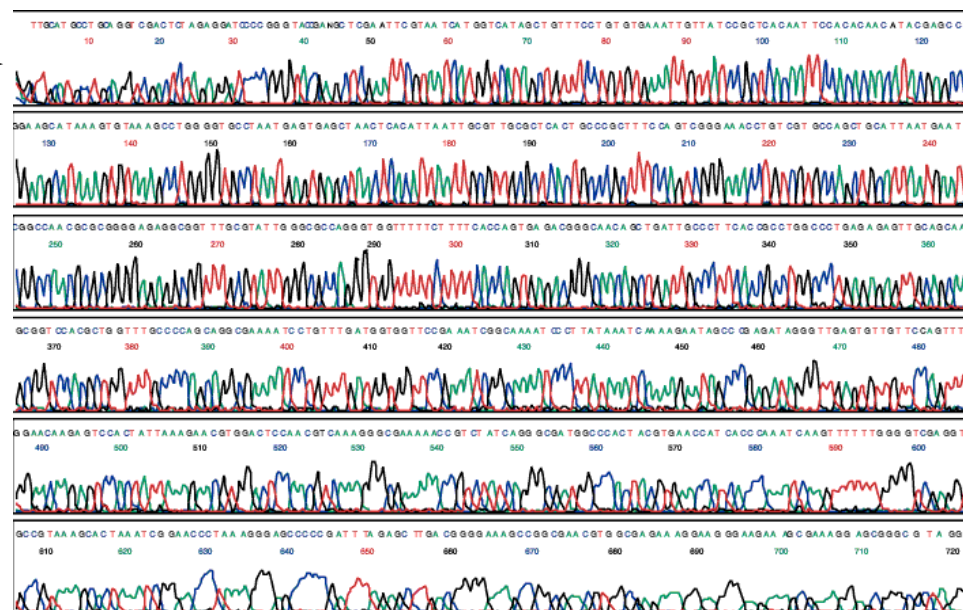
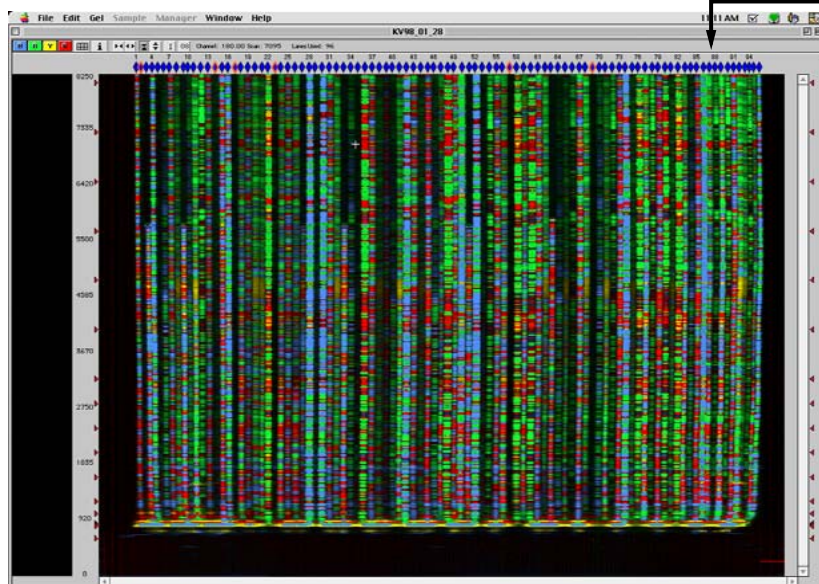




# Automated Version of the Dideoxy Method



96 capillaries





# Dideoxy (Sanger) Method

---

## Advantages:

- relatively long fragments (500-750 bp)
- low frequency of sequencing errors (“gold standard”)

## Disadvantages:

- expensive
- laborious
- low productivity



# Next Generation Sequencing (NGS)

---

- Overview of the **Next Generation Sequencing (NGS)** technologies: **second, third and fourth generations**
- **Sequencing-by-synthesis** using reversible nucleotide terminators and luciferase and luminescence signal detection - **Pyrosequencing (454 / Roche)**
- **Sequencing-by-synthesis** using reversible nucleotide terminators and fluorescent signal detection - **Illumina** sequencing technology
- Application of NGS for marker discovery and genotyping
  - whole genome *de novo* sequencing
  - whole genome resequencing
  - whole transcriptome sequencing
  - genotyping by sequencing (GBS)
  - genomic complexity reduction and target resequencing
  - sequence data are used to develop high density SNP genotyping assays (such as Illumina Infinium)



# Available 2nd-Generation Sequencing Technologies

Company	Platform name	Method of sequencing	Method of detection	Read length	Advantages	Relative limitation
Roche/454	GS FLX Titanium	Pyrosequencing	Optical	500-800	Longest read lengths among second generation; high-throughput with respect to first generation (Sanger) sequencing	Challenging sample prep; difficulty reading through repetitive/homopolymer regions; sequential reagent washing gives steady accumulation of errors; expensive instrument (\$500K)
<b>However, the 454 sequencers was phased out in 2016!</b>						
Illumina	HiSeq/(MiSeq) (MiniSeq NextSeq NovaSeq)	Reversible terminator Sequencing by Synthesis	Fluorescence/Optical	2x150 (HiSeq)/2x300 (MiSeq) (MiniSeq NextSeq NovaSeq)	Very high-throughput (HiSeq)/Desktop Sequencer (MiSeq)	Expensive instrument (HiSeq); significant cost of data managing and analysis (HiSeq)
ABI/SOLiD Thermo Fischer Scientific	5500xl SOLiD System	Sequencing by Ligation	Fluorescence/Optical	25-35	Very high throughput; lowest reagent cost needed to reassemble a human genome among the widely accepted second generation platforms (Illumina, 454, SOLiD)	Long sequencing runs; short reads increase cost and difficulty of data analysis and genome assembly; high instrument cost (~\$700K)
Helicos (SeqLL now)	Heliscope	Single-molecule sequencing by synthesis	Fluorescence/Optical	11-100	High throughput; single-molecule nature of technology unique among second-gen platforms	Short reads increasing the costs and reduce quality of genome assembly; very costly instrument (~\$1M)

Niedringhaus et al. (2011) Landscape of Next-Generation Sequencing Technologies. *Anal. Chem.* 83: 4327–4341

According to a new report published by Allied Market Research, titled, "**DNA Sequencing Market by Product, Application, Technology, and End User: Global Opportunity Analysis and Industry Forecast, 2017-2023**," the global DNA sequencing market was valued at \$5,156 million in 2016, and is projected to reach \$18,284 million by 2023

**Illumina** continue dominates the sequencing market (\$ 4.5 bln) with 70% share, while **ABI Life Technologies/Thermo Fischer Scientific** and **Roche** split nearly all of the remaining market at 14% each.



# Available 2nd-Generation Sequencing Technologies

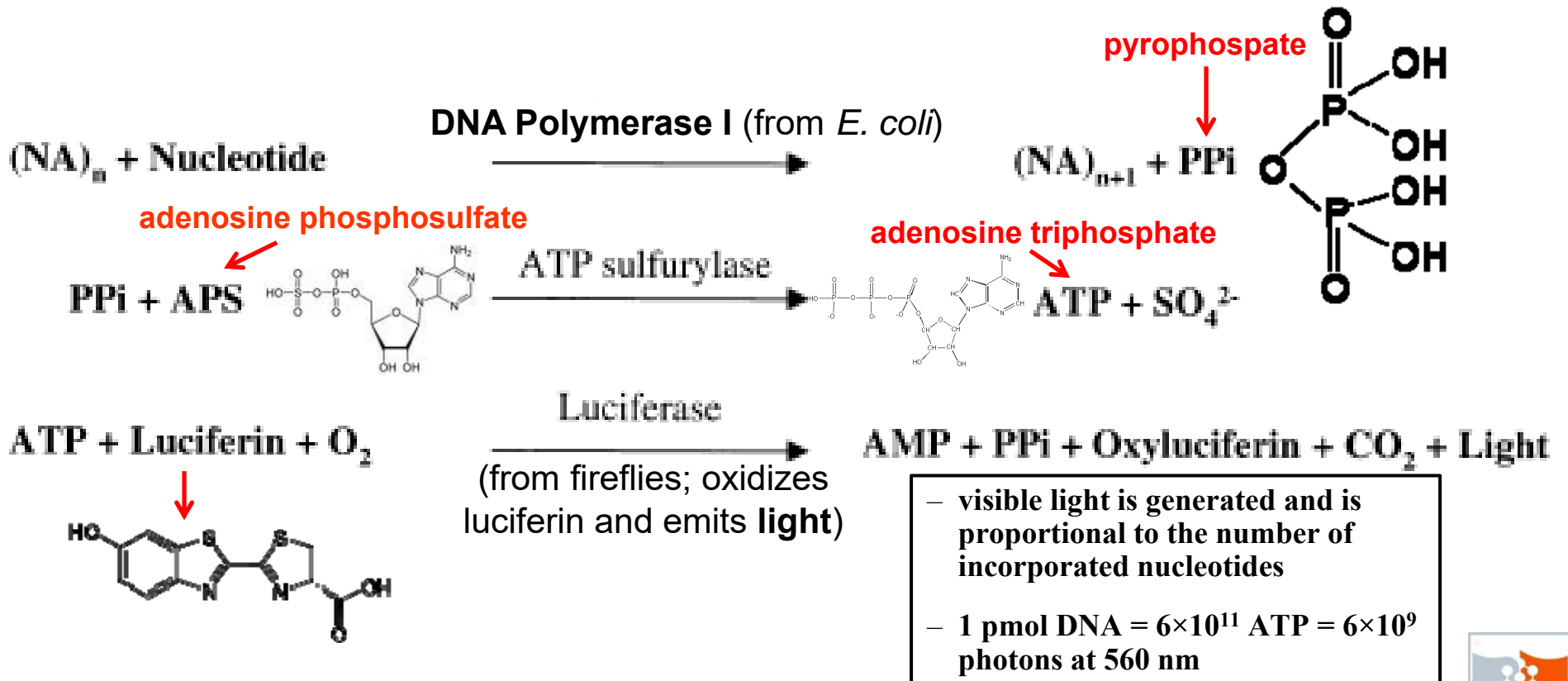
---

- **454/Roche** ([www.454.com](http://www.454.com)):  
**GS FLX Titanium (phased out in 2016)**
- **Illumina** ([www.illumina.com](http://www.illumina.com)):  
**HiSeq1000/2000/2500/3000/4000/X, NovaSeq 6000, MiSeq, MiniSeq, NextSeq**
- **Thermo Fischer Scientific/Life/ABI** ([www.appliedbiosystems.com](http://www.appliedbiosystems.com)): **Solid 5500xl**
- **SeqLL** (former Helicos BioSciences)  
(<http://seqll.com/>): **HeliScope**
- **Beijing Genomics Institute** (<https://www.bgi.com>):  
**BGISEQ-500**



# 2<sup>nd</sup> Generation Sequencing: Pyrosequencing method – Basic Principle

- Pyrosequencing is based on the generation of **light** signal through release of **pyrophosphate (PPi)** on nucleotide addition:  $(NA)_n + dNTP \rightarrow (NA)_{n+1} + PP_i$
- PPi** is used to generate ATP from **adenosine phosphosulfate (APS)**:  $APS + PP_i \rightarrow ATP$
- ATP and **luciferase** generate **light** by conversion of **luciferin** to **oxyluciferin**.



# Pyrosequencing: Preparation of DNA library

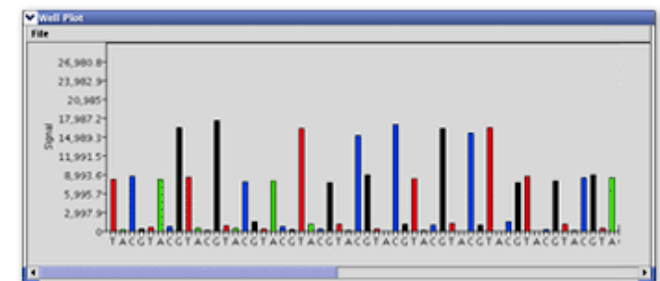
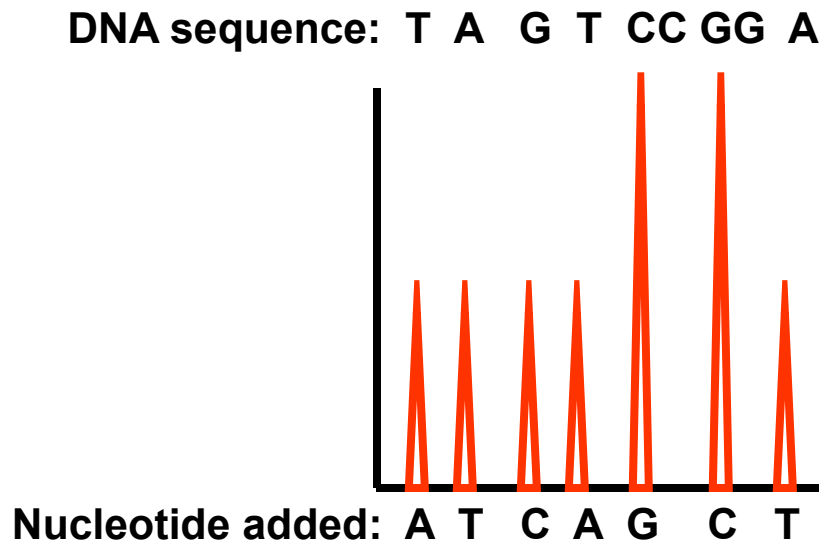
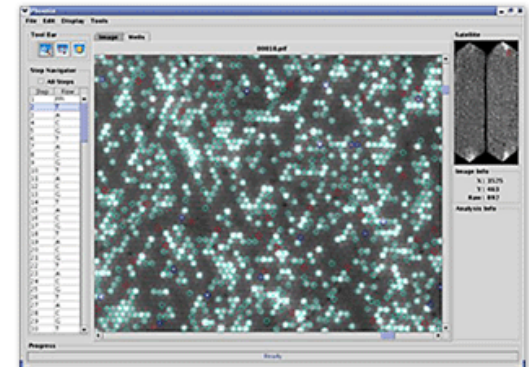
---

- shearing genomic DNA to small DNA fragments 500-800 bp long
- attaching single DNA fragments to very small plastic beads (one fragment per bead)
- emulsion-based clonal PCR amplification (emPCR) of the DNA on each bead to cover each bead with a cluster of identical fragments to enhance the light signal
- placing each bead in a separate well on a PicoTiterPlate, a fiber optic chip with up to 1.6 million wells (A mix of enzymes such as DNA polymerase, ATP sulfurylase, and luciferase are also packed into the well.)
- The PicoTiterPlate is then placed into the GS FLX System for sequencing.



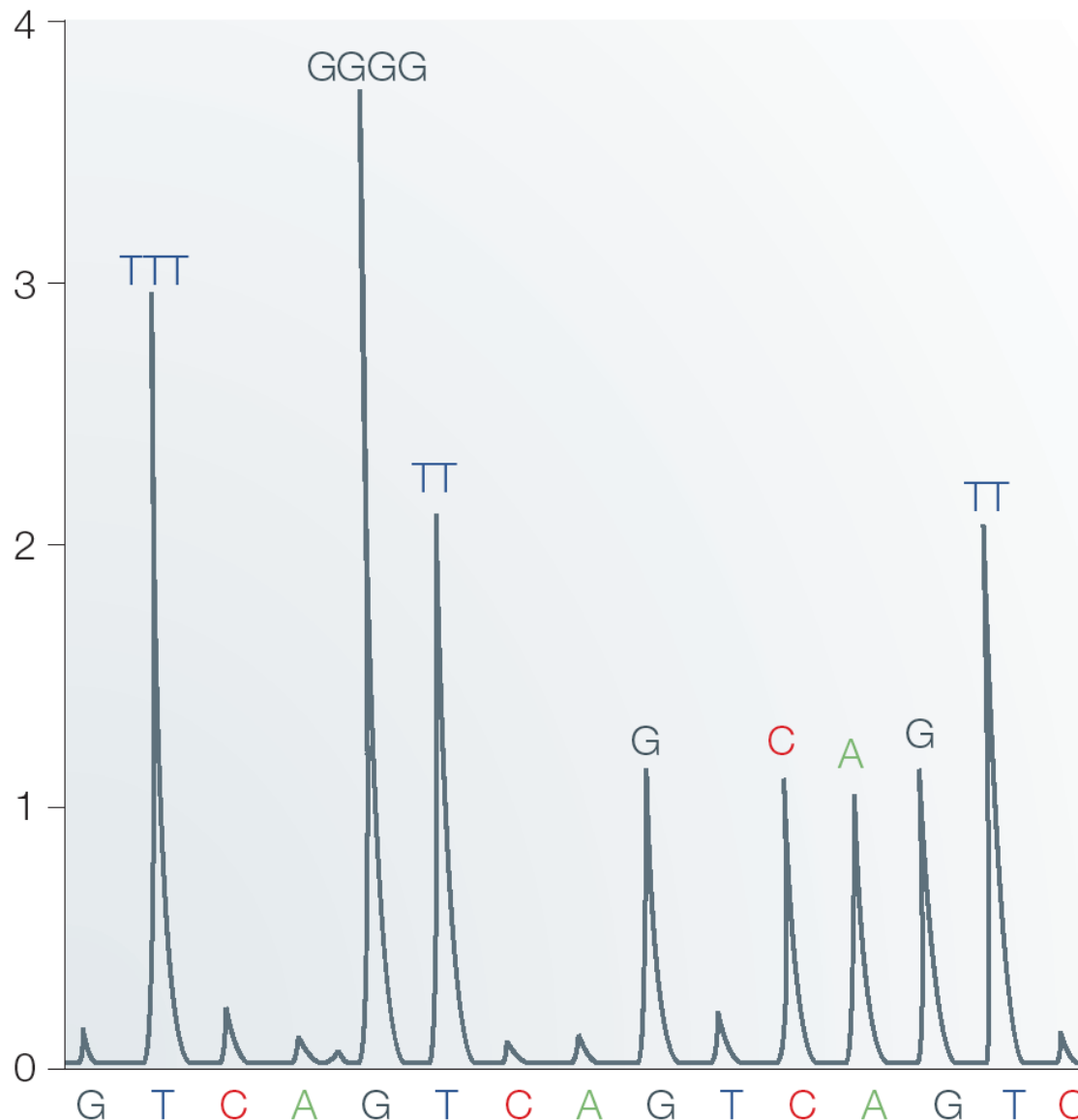
# Pyrosequencing – Basic Principle

- Sequence-by-synthesis via DNA polymerase directed chain extension, one base at a time in the presence of a reporter (luciferase). Each nucleotide is added separately in a separate cycle.
- Only one of four will generate a light signal. Luciferase will emit a photon of light in response to the pyrophosphate (PPi) released upon nucleotide addition by DNA polymerase
- The remaining nucleotides are removed enzymatically.
- The light signal is recorded on a **pyrogram**:





# Pyrosequencing Results:

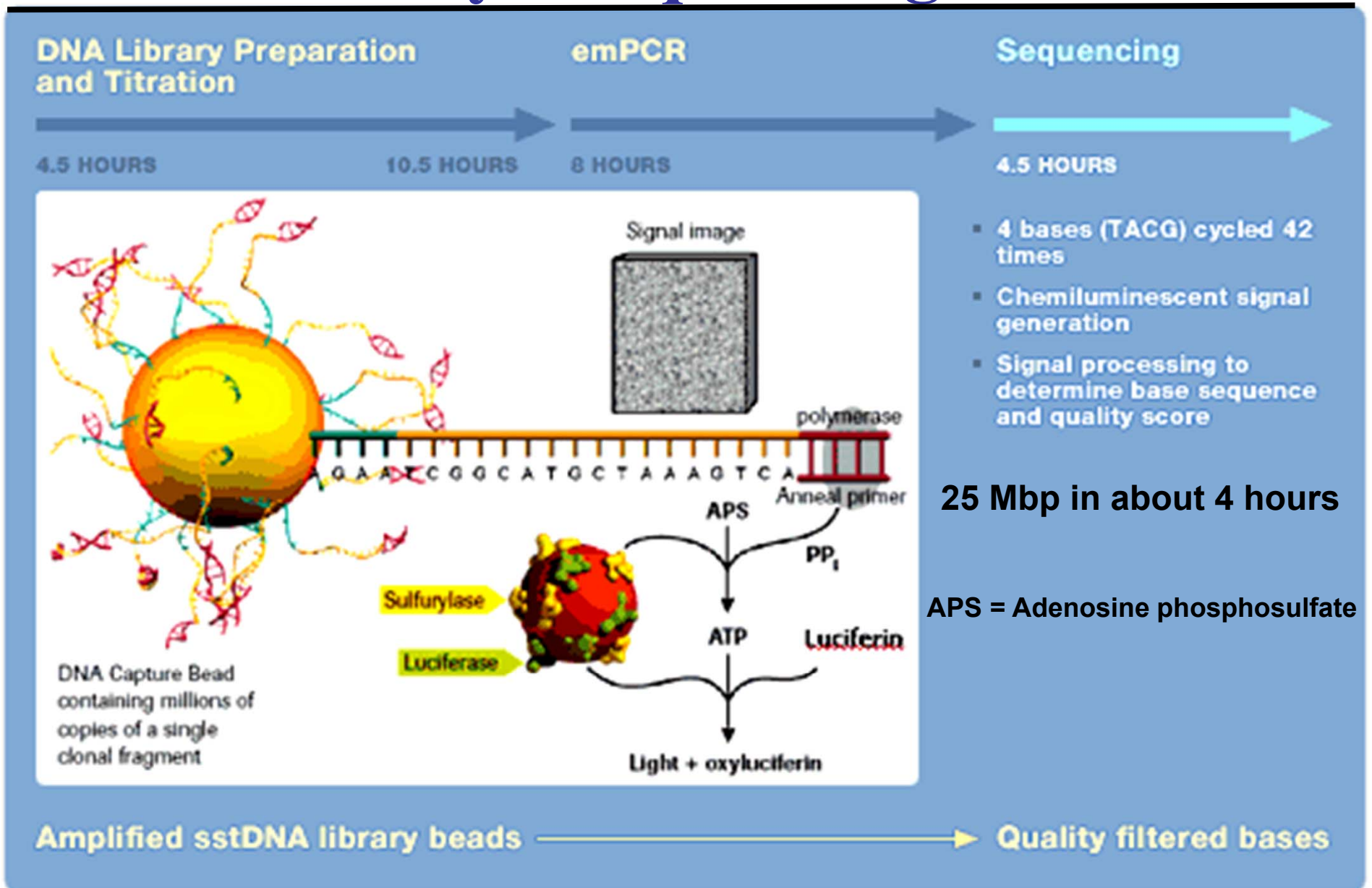


Height of peak indicates the number of dNTPs added

This sequence: AAACCCCAACGTCAA



# Pyrosequencing



<http://www.youtube.com/watch?v=bFNjxKHP8Jc>

# Pyrosequencing

---

- Sequencing by synthesis
- **Advantages:**
  - accurate - low frequency of sequencing errors
  - relatively long fragments (500-750 bp)
  - parallel processing
  - automated
  - no need for labeled primers and nucleotides
  - no need for gel electrophoresis
- **Disadvantages:**
  - expensive
  - laborious
  - low productivity
  - nonlinear light response after more than 5-6 identical nucleotides



# References

---

Applied Biosystems Automated DNA Sequence Chemistry Guide. (2000)

Garrett & Grisham. (2007) Biochemistry. Thomson and Brooks/Cole. 3<sup>rd</sup> ed. Pgs 337-340.

Maxam, A. & Gilbert, W. (1977) A new method for sequencing DNA. *Proc. Natl. Acad. Sci.* **74**, 560-564.

Ronaghi, M. (2001) Pyrosequencing sheds light on DNA sequencing. *Genome Res.* **11**, 3-11.

Sanger, F., Nicklen, S., & Coulson, A.R. (1977) DNA Sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci.* **94**, 5463-5467.

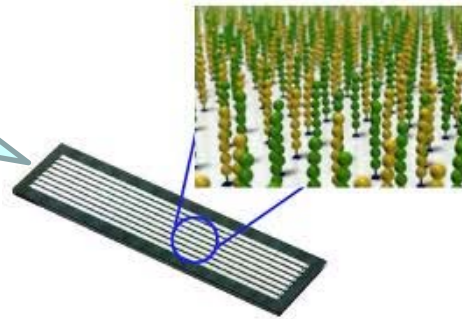
Shendure, J. & Ji, H. (2008) Next-generation DNA Sequencing. *Nature Biotech.* **26**, 1135-1145

Venter, C, et al. (2001) The sequence of the human genome. *Science.* **291**, 1304.



# Illumina: Preparation a DNA library for sequencing - Cluster formation

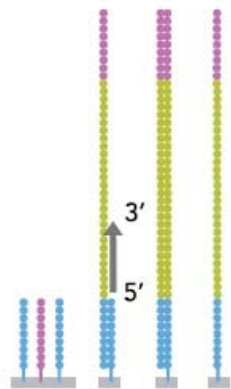
Samples are loaded in a flowcell with 8 lanes and clusters attached to the cell surface are generated and prepared for sequencing using cBot instrument



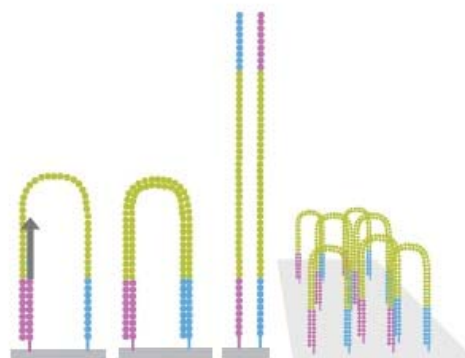
Flowcell with 8 separate lanes and primers complementary to adapters immobilized on the flowcell surface.



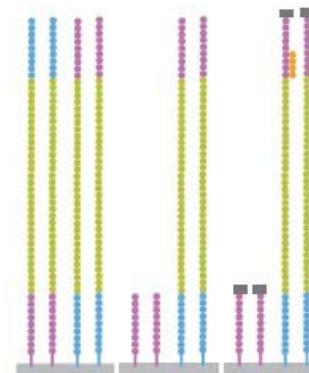
**cBOT**



**Hybridization and second strand synthesis**



**Bridge PCR**



**Linearization and primer annealing**



**HiSeq2000**



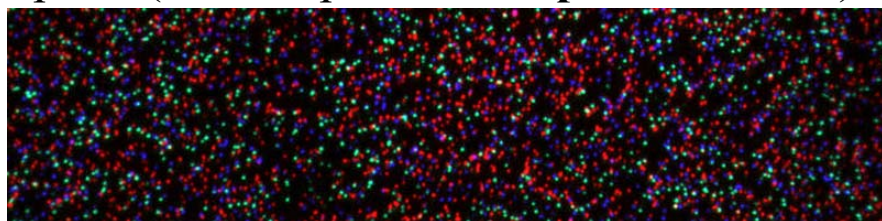
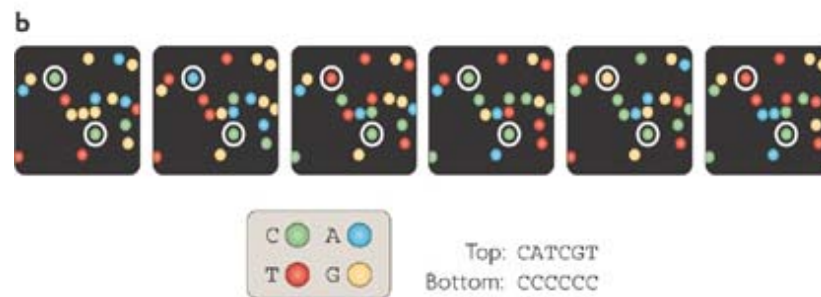
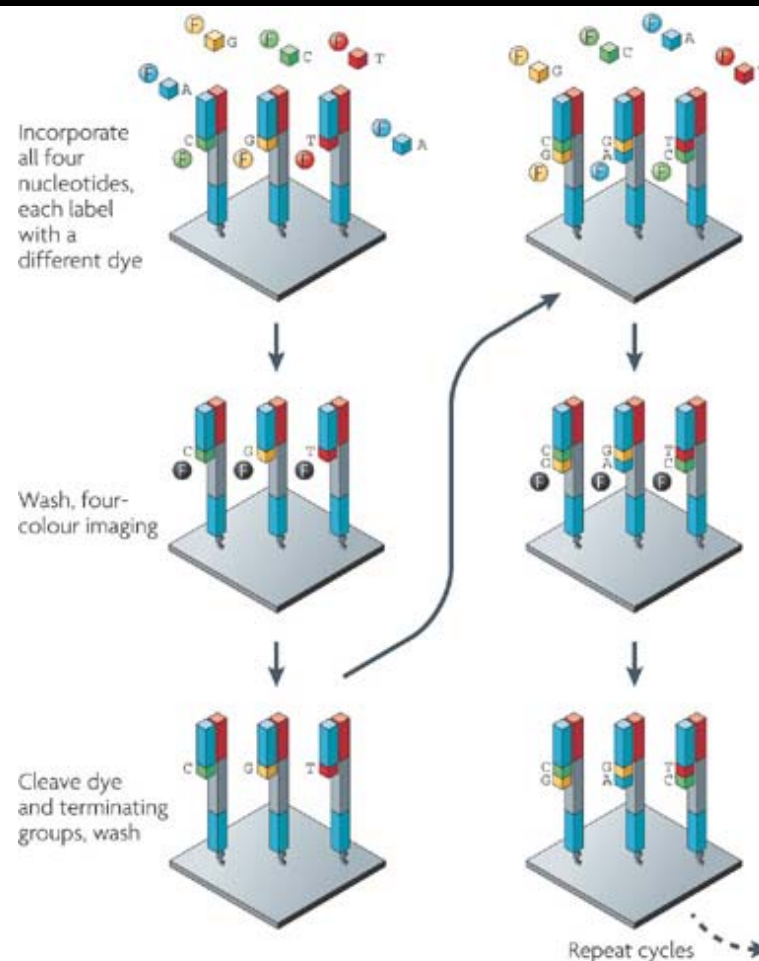
# Illumina: Sequencing by Synthesis (SBS)

- Sequencing is performed by addition of 4 labeled reversible terminators (1 base is added per cycle to a growing chain in all the fragments).
- The flowcell is then scanned
- The block and phluorophore is then removed, and another cycle starts up to the desired read length.

## Production-scale sequencer NovaSeq 6000:

**1-1.25 billions of clusters per lane** (8 lanes per flowcell) – **16-20 billions** for dual flowcell; total output up to **4.8-6 Tbp** for a full run with 2 flowcells for paired-end sequences 150 bp × 2

**Benchtop sequencer MiSeq:** 25 Millions of clusters per flowcell with a single lane; up to 300 bp × 2 (total output = **15 Gbp** for a full run)

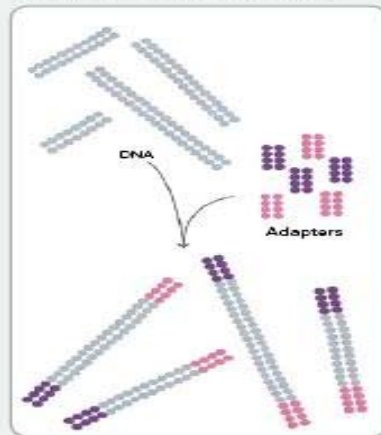




# Illumina sequencing technology in 15 steps

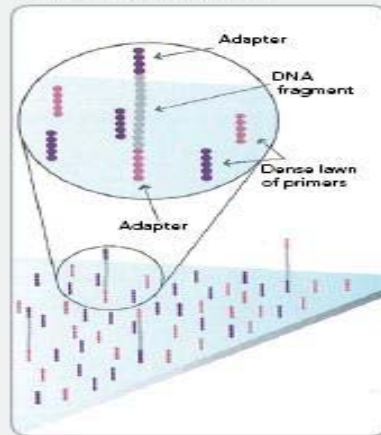
FIGURE 2: SEQUENCING TECHNOLOGY OVERVIEW

1. PREPARE GENOMIC DNA SAMPLE



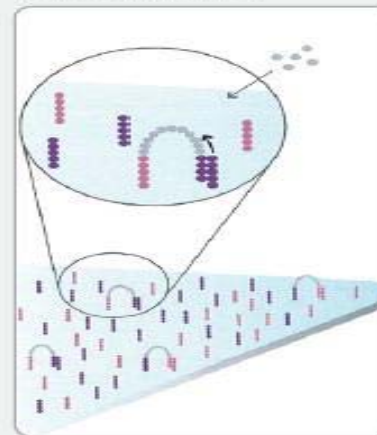
Randomly fragment genomic DNA and ligate adapters to both ends of the fragments.

2. ATTACH DNA TO SURFACE



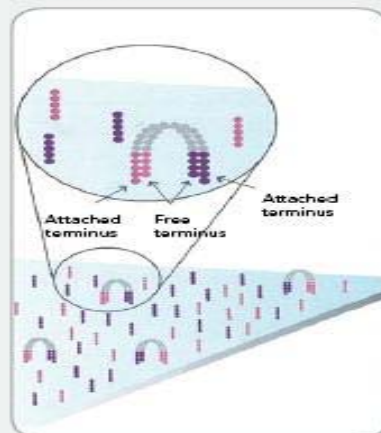
Bind single-stranded fragments randomly to the inside surface of the flow cell channels.

3. BRIDGE AMPLIFICATION



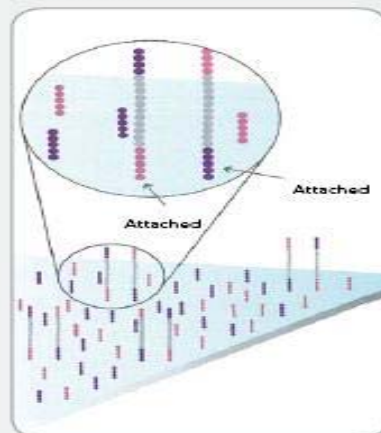
Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification.

4. FRAGMENTS BECOME DOUBLE STRANDED



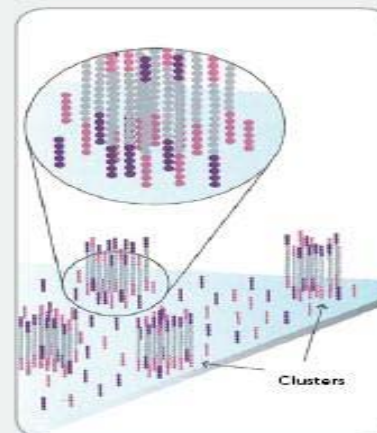
The enzyme incorporates nucleotides to build double-stranded bridges on the solid-phase substrate.

5. DENATURE THE DOUBLE-STRANDED MOLECULES



Denaturation leaves single-stranded templates anchored to the substrate.

6. COMPLETE AMPLIFICATION



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

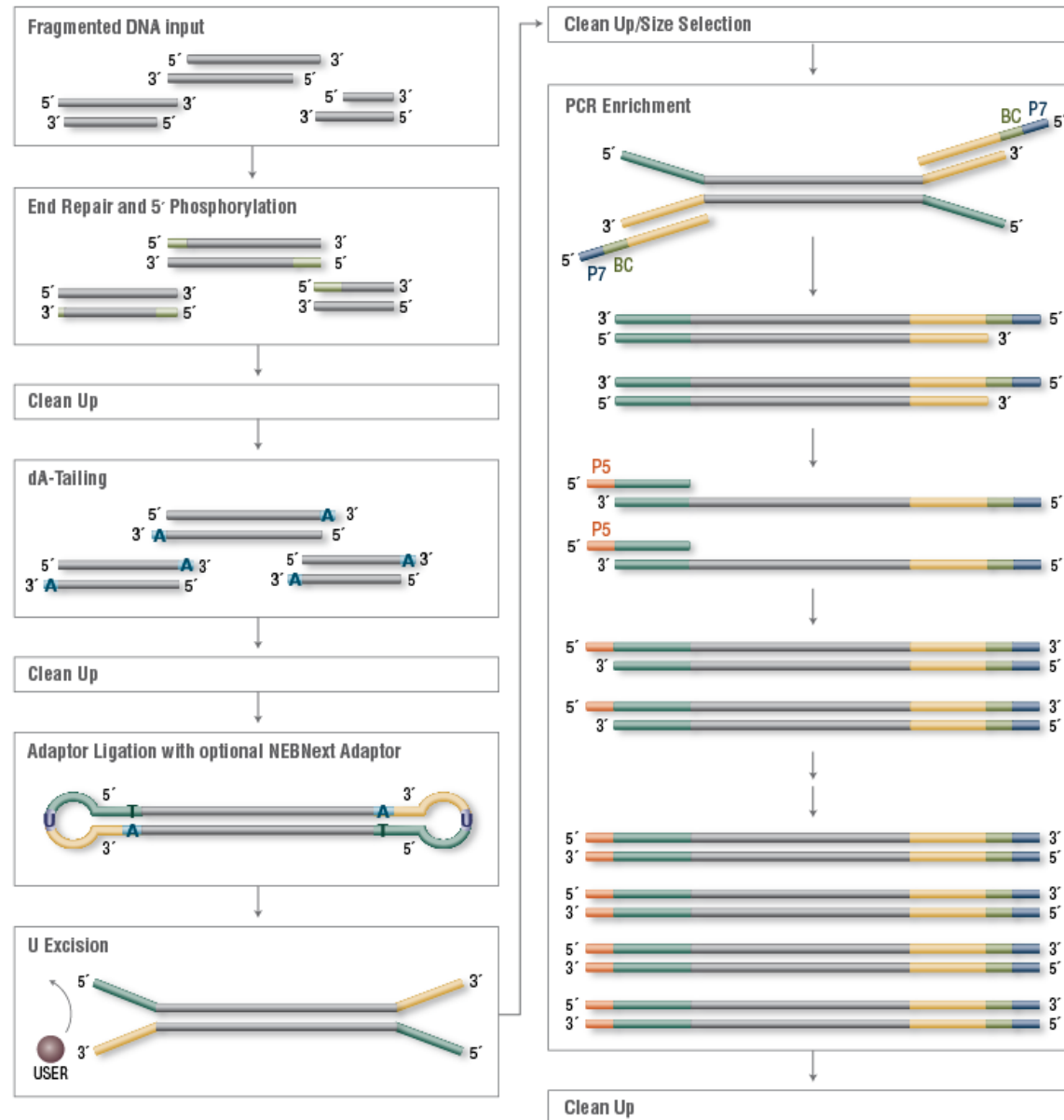
## Important note:

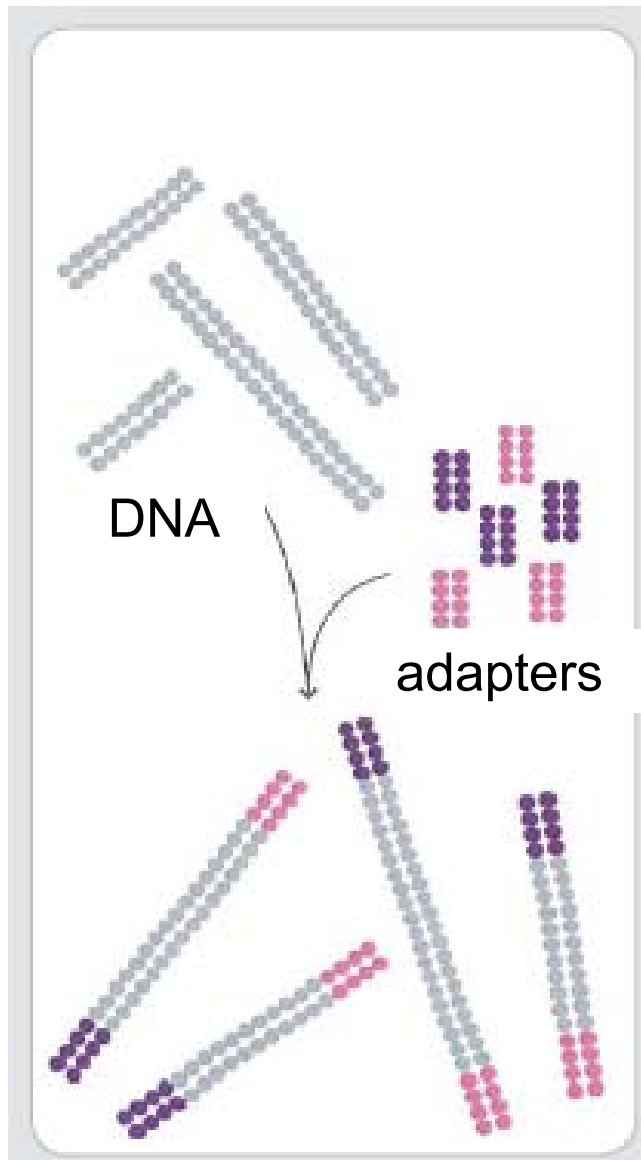
Between 1. and 2. steps PCR is performed to selectively enrich those DNA fragments that have adapters on both ends and to amplify the amount of DNA in the library. The PCR is performed with a PCR primer cocktail that anneals to the ends of the adapters.

[https://www.illumina.com/documents/products/techspotlights/techspotlight\\_sequencing.pdf](https://www.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf)



# Adaptor ligation:





## 1. Fragment genomic DNA and ligate adapters

2. Attach DNA to surface

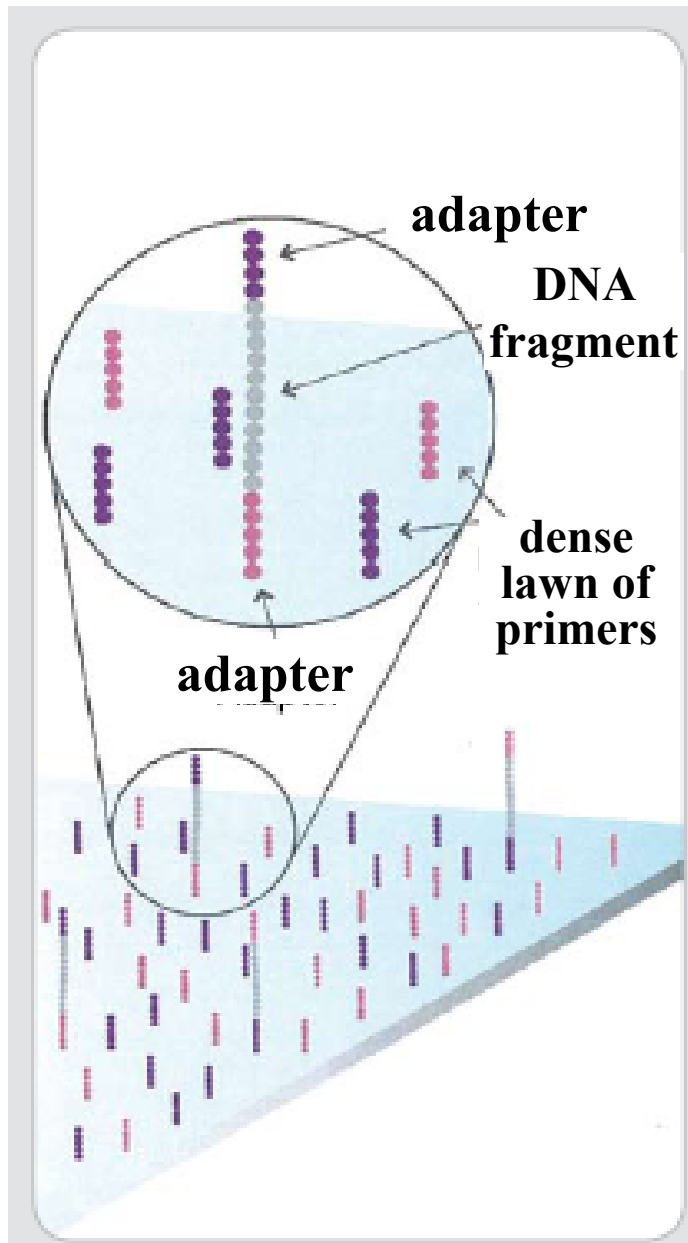
3. Bridge amplification

4. Fragments become double stranded

5. Denature the double-stranded molecules

6. Complete amplification

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



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

1. Fragment genomic DNA and ligate adapters

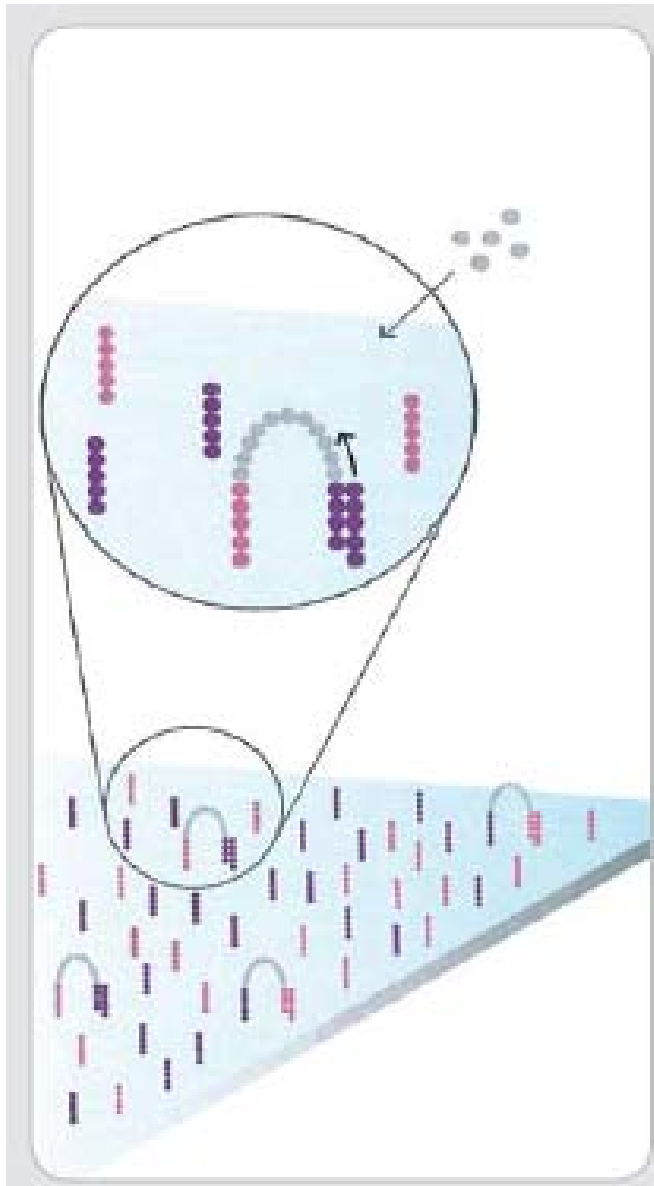
**2. Attach DNA fragments to surface**

3. Bridge amplification

4. Fragments become double stranded

5. Denature the double-stranded molecules

6. Complete amplification



1. Fragment genomic DNA and ligate adapters

2. Attach DNA fragments to surface

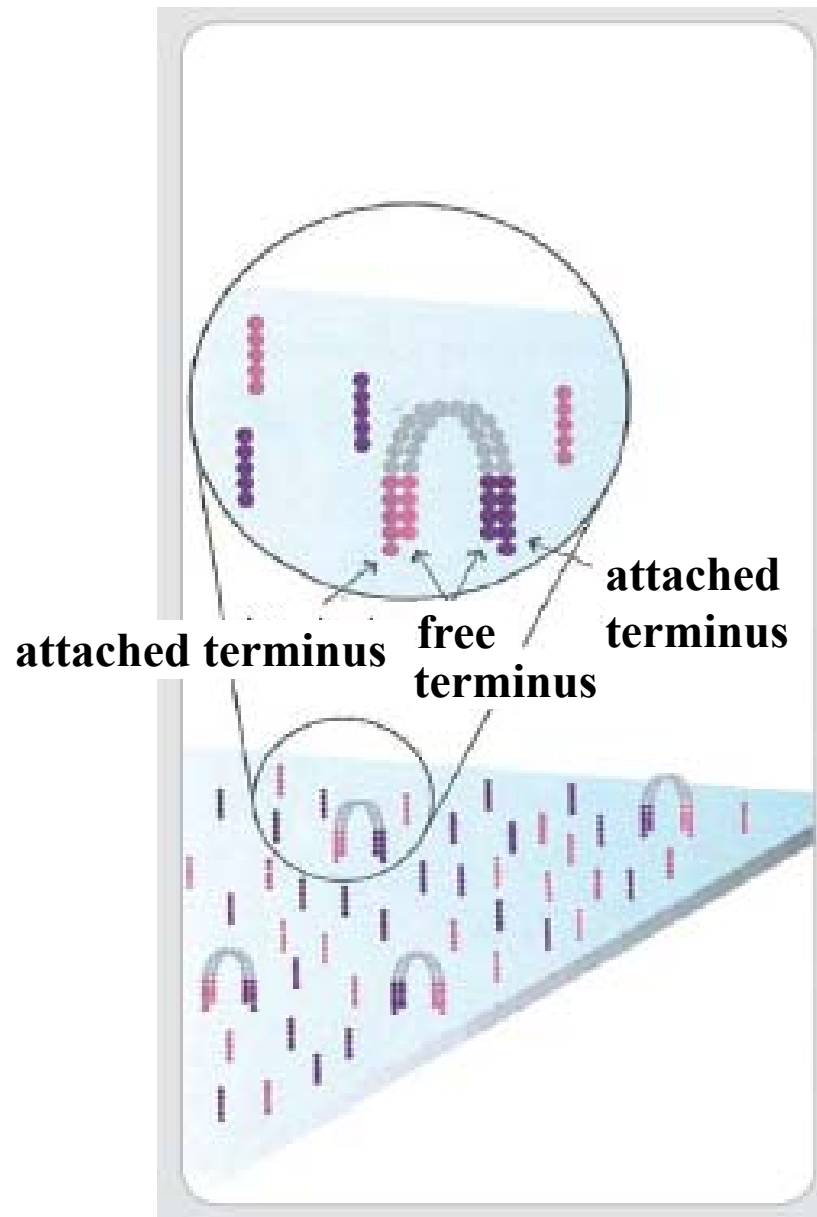
### 3. Bridge amplification

4. Fragments become double stranded

5. Denature the double-stranded molecules

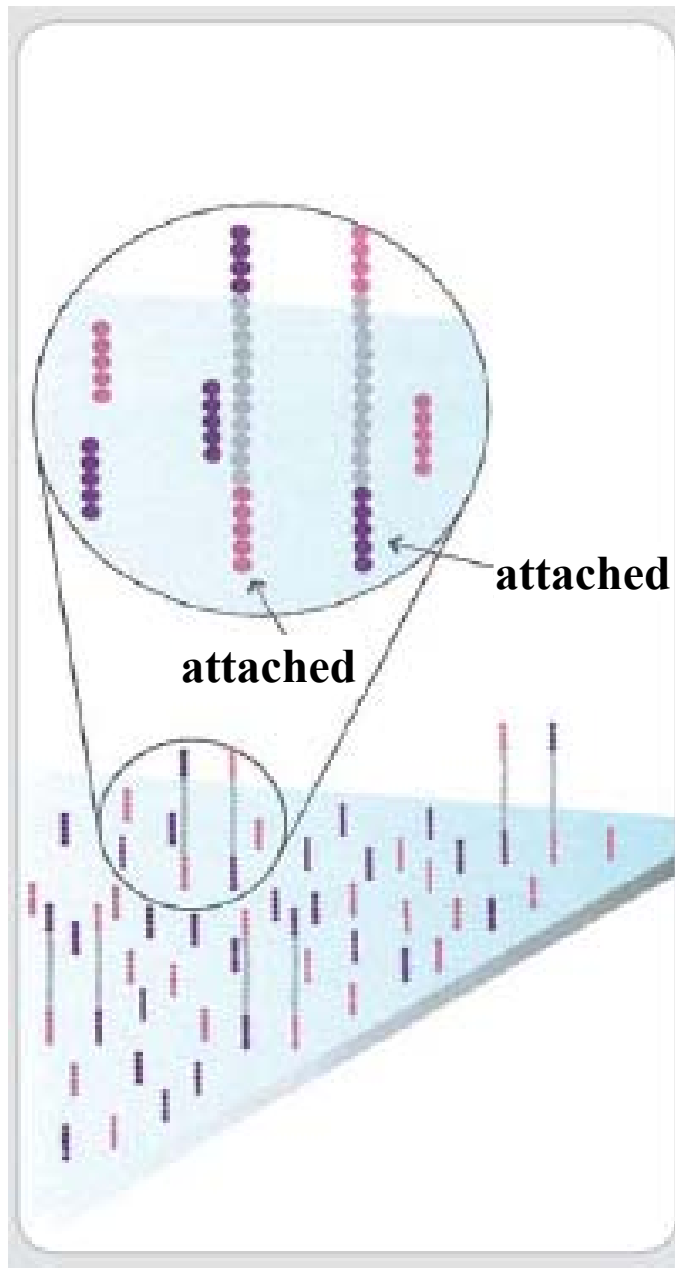
6. Complete amplification

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



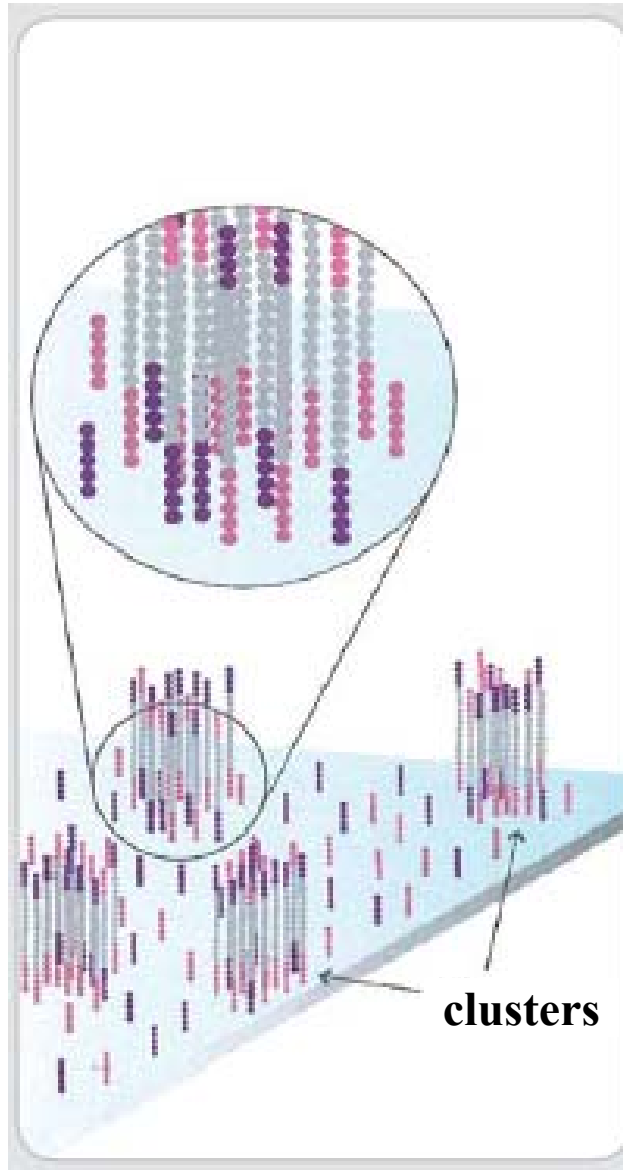
1. Fragment genomic DNA and ligate adapters
2. Attach DNA fragments to surface
3. Bridge amplification
- 4. Fragments become double stranded**
5. Denature the double- stranded molecules
6. Complete amplification

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



1. Fragment genomic DNA and ligate adapters
2. Attach DNA fragments to surface
3. Bridge amplification
4. Fragments become double stranded
- 5. Denature the double-stranded molecules**
6. Complete amplification

**Denaturation leaves single-stranded templates anchored to the substrate**



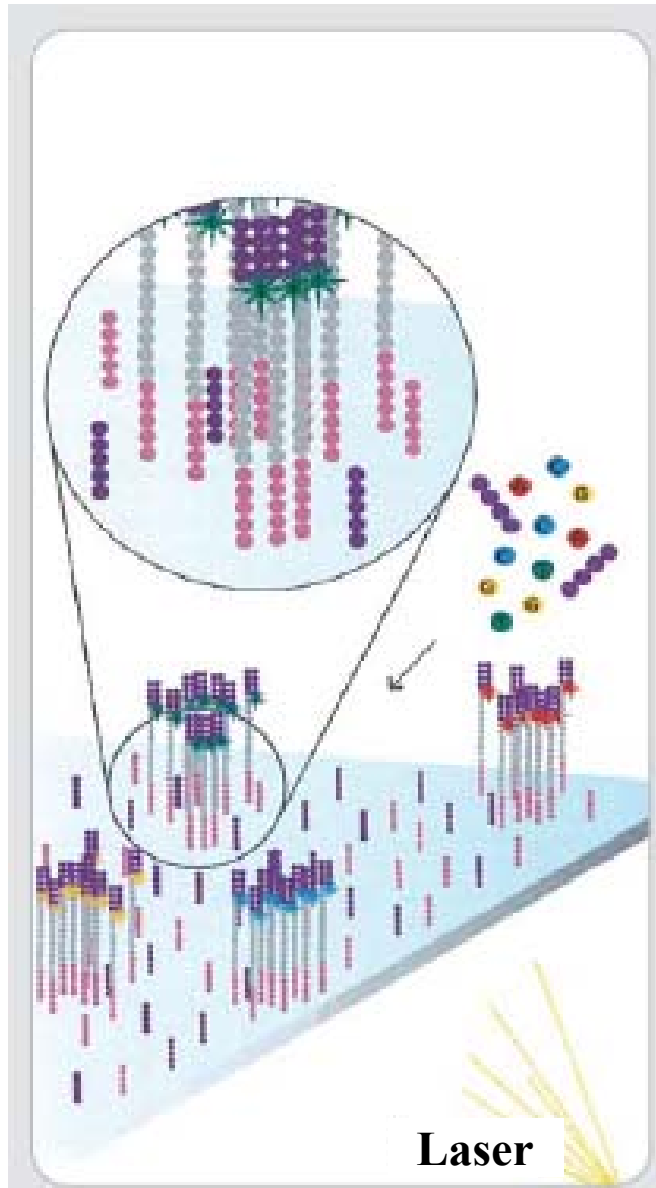
**Several million dense clusters of double-stranded DNA are generated in each channel of the flow cell**

1. Fragment genomic DNA and ligate adapters
2. Attach DNA fragments to surface
3. Bridge amplification
4. Fragments become double stranded
5. Denature the double- stranded molecules

**6. Complete amplification**

**7. Reversed complement strands are cleaved and washed away**





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

**8. First four labeled reversible nucleotide bases (terminators) adding and the first nucleotide base synthesis, and unincorporated nucleotides are washed away**

9. Image first base

10. The block and phluorophore are then removed

11. Second four nucleotide base adding and the second nucleotide base synthesis followed by washing

12. Image second base

13. The block and phluorophore are then removed

14. Sequencing over multiple chemistry cycles

15. Process sequence reads, etc.



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

8. First four labeled reversible nucleotide bases (terminators) adding and the first nucleotide base synthesis, and unincorporated nucleotides are washed away

### **9. Image first base**

10. The block and phluorophore are then removed

11. Second four nucleotide base adding and the second nucleotide base synthesis followed by washing

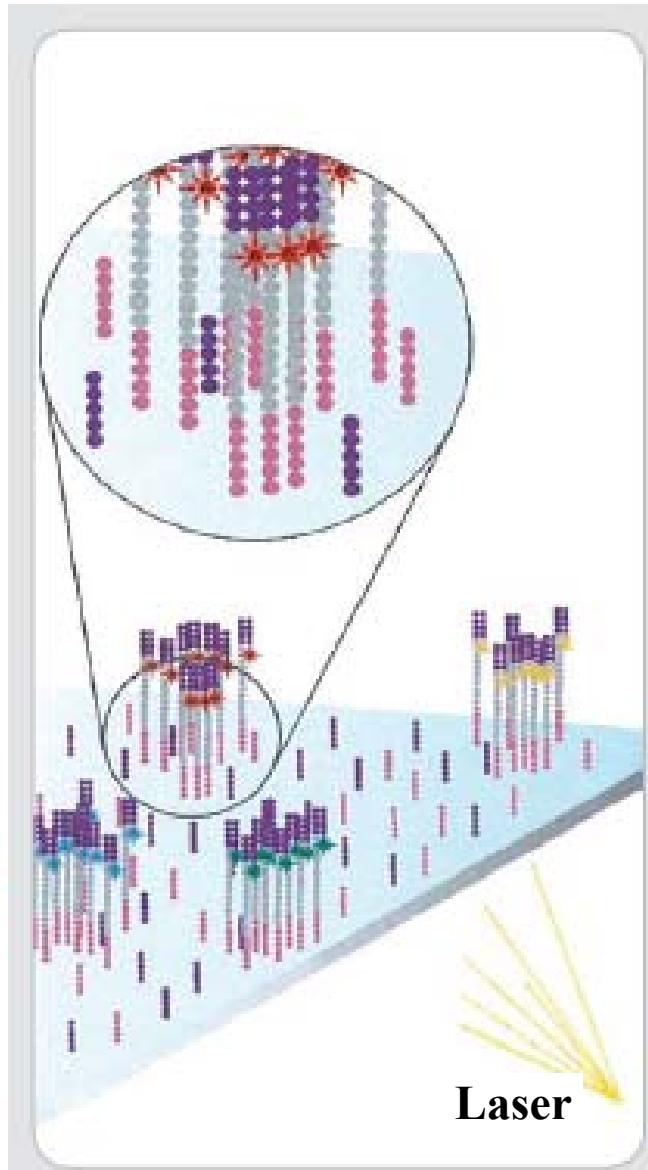
12. Image second base

13. The block and phluorophore are then removed

14. Sequencing over multiple chemistry cycles

15. Process sequence reads, generate contigs, map (align) reads or contigs to the reference sequence, if it is available





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

8. First four labeled reversible nucleotide bases (terminators) adding and the first nucleotide base synthesis, and unincorporated nucleotides are washed away

9. Image first base

10. The block and phluorophore are then removed

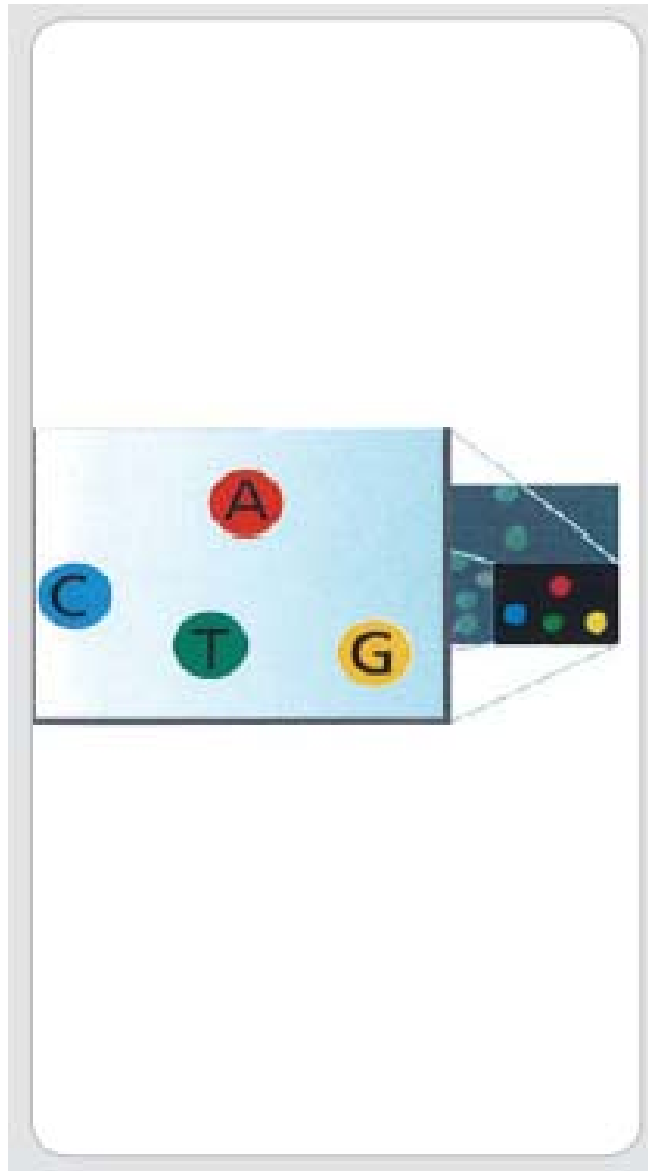
**11. Second four nucleotide base adding and the second nucleotide base synthesis followed by washing**

12. Image second base

13. The block and phluorophore are then removed

14. Sequencing over multiple chemistry cycles

15. Process sequence reads, generate contigs, map (align) reads or contigs to the reference sequence, if it is available



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

8. First four labeled reversible nucleotide bases (terminators) adding and the first nucleotide base synthesis, and unincorporated nucleotides are washed away

9. Image first base

10. The block and phluorophore are then removed

11. Second four nucleotide base adding and the second nucleotide base synthesis followed by washing

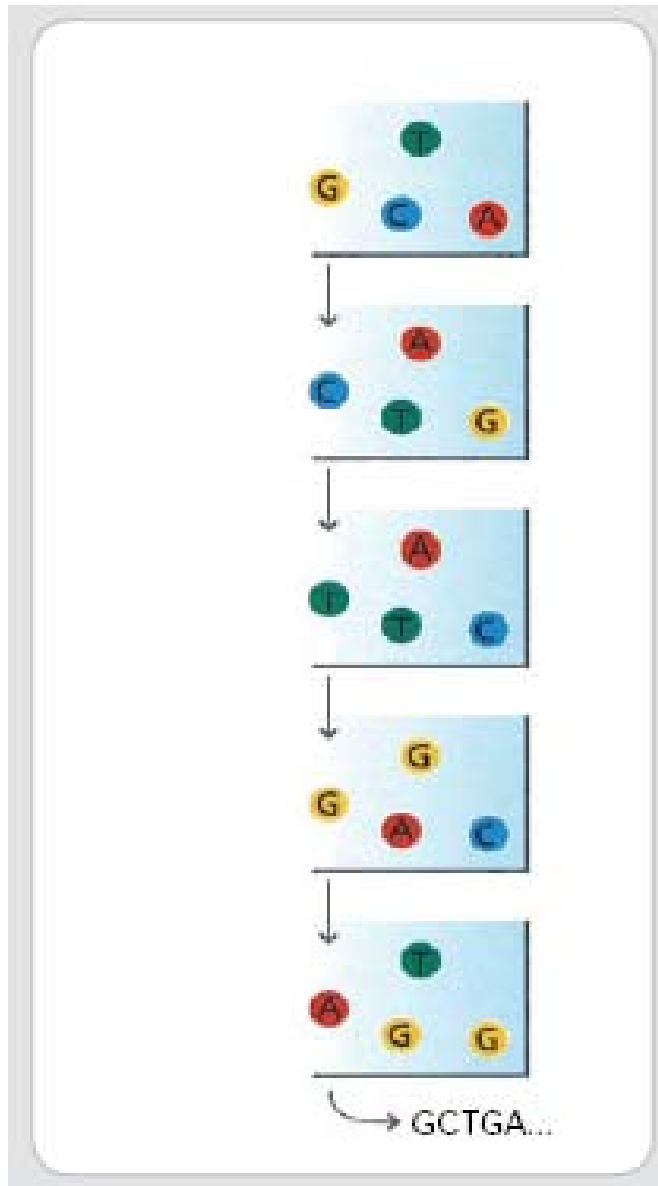
**12. Image second base**

13. The block and phluorophore are then removed

14. Sequencing over multiple chemistry cycles

15. Process sequence reads, generate contigs, map (align) reads or contigs to the reference sequence, if it is available





8. First four labeled reversible nucleotide bases (terminators) adding and the first nucleotide base synthesis, and unincorporated nucleotides are washed away

9. Image first base

10. The block and phluorophore are then removed

11. Second four nucleotide base adding and the second nucleotide base synthesis followed by washing

12. Image second base

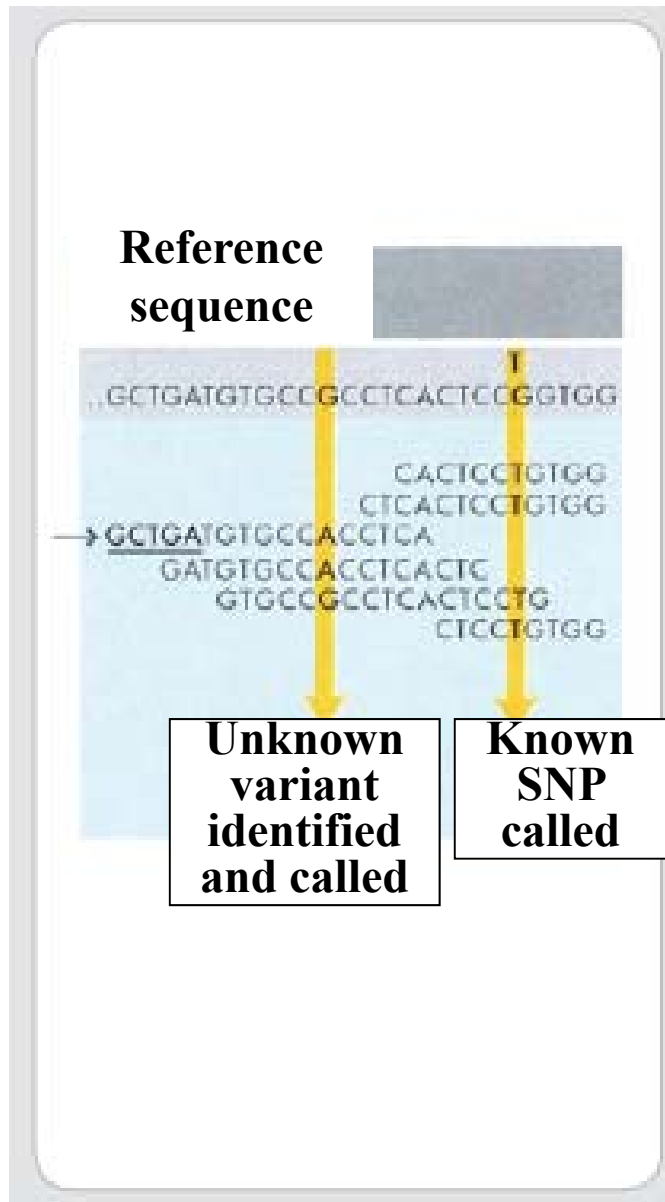
13. The block and phluorophore are then removed

**14. Sequencing over multiple chemistry cycles**

The sequencing cycles are repeated to determine the sequence of bases in a fragment, one base at a time.

15. Process sequence reads, generate contigs, map (align) reads or contigs to the reference sequence, if it is available

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



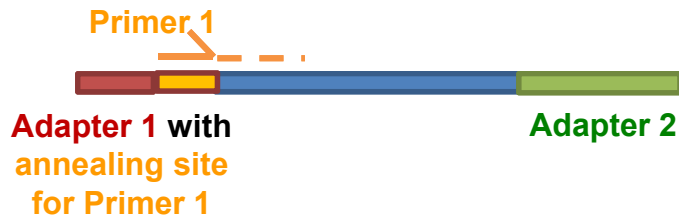
The data are aligned and compared to a reference, and sequencing differences are identified.

8. First four labeled reversible nucleotide bases (terminators) adding and the first nucleotide base synthesis, and unincorporated nucleotides are washed away
9. Image first base
10. The block and phluorophore are then removed
11. Second four nucleotide base adding and the second nucleotide base synthesis followed by washing
12. Image second base
13. The block and phluorophore are then removed
14. Sequencing over multiple chemistry cycles
15. Process sequence reads, generate contigs, **map (align) reads or contigs to the reference sequence**, if it is available



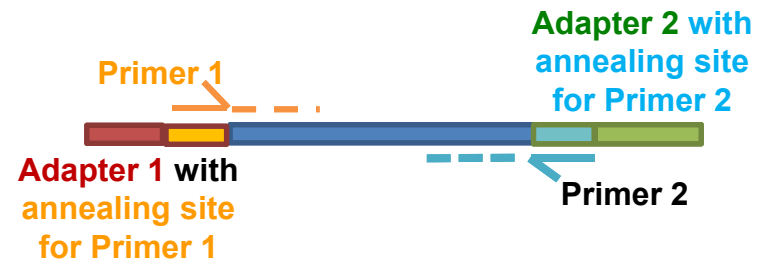
# Single, paired-end and multiplexed sequencing

## Single read sequencing:



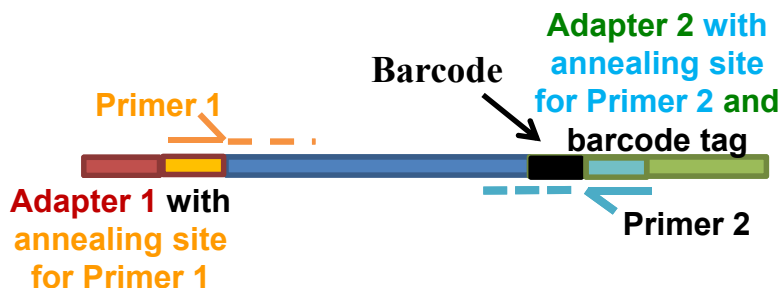
- resequencing
- expression quantification

## Pair-end sequencing:




- *de novo* sequencing for better assembly
- allow to resolve better transcript isoforms
- can be used to detect gene fusion

## Multiplexed sequencing:



- The sequencing of a 6-nucleotides barcode on one of the adapters allows to identify samples pooled together and sequenced in a single lane.
- Up to 96 samples per lane.

# Illumina Sequencers and Genotypers








From Genome-Wide Discovery to Targeting Validation and Screening

	Sequencing					Arrays	
Instrument	HiSeq X™ Ten*	HiSeq® 2500	NextSeq™ 500	MiSeqDx™	MiSeq®	HiScan®	iScan
Technologies	Sequencing by Synthesis (SBS) Powered by TruSeq Chemistry					BeadArray, Infinium, GoldenGate, DASL	
Applications	Population-Scale Whole Human Genome Sequencing	Production-Scale Genome, Exome, Transcriptome Sequencing and More	Exome, Transcriptome, Whole-Genome, Sequencing and More	FDA-Cleared <i>in vitro</i> Diagnostic System Cystic Fibrosis Screening and User-Defined Assays	Small Genome, Amplicons, Targeted Gene Panel Sequencing	SNP and Whole-Genome Genotyping, CNV Analysis, Gene Regulation and Epigenetic Analysis, Gene Expression Analysis, Cytogenetic Analysis	

\* The HiSeq X Ten consists of 10 sequencing systems.



# Illumina Sequencers

	 <b>MiniSeq System</b>	 <b>MiSeq Series</b>	 <b>NextSeq Series</b>	 <b>HiSeq Series</b>	 <b>HiSeq X Series<sup>*</sup></b>
<b>Key Methods</b>	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.
<b>Maximum Output</b>	7.5 Gb	15 Gb	120 Gb	1500 Gb	1800 Gb
<b>Maximum Reads per Run</b>	25 million	25 million <sup>†</sup>	400 million	5 billion	6 billion
<b>Maximum Read Length</b>	2 × 150 bp	2 × 300 bp	2 × 150 bp	2 × 150 bp	2 × 150 bp
<b>Run Time</b>	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
<b>Benchtop Sequencer</b>	Yes	Yes	Yes	No	No
<b>System Versions</b>	<ul style="list-style-type: none"> <li>• <b>MiniSeq System</b> for low-throughput targeted DNA and RNA sequencing</li> </ul>	<ul style="list-style-type: none"> <li>• <b>MiSeq System</b> for targeted and small genome sequencing</li> <li>• <b>MiSeq FGx System</b> for forensic genomics</li> <li>• <b>MiSeqDx System</b> for molecular diagnostics</li> </ul>	<ul style="list-style-type: none"> <li>• <b>NextSeq 500 System</b> for everyday genomics</li> <li>• <b>NextSeq 550 System</b> for both sequencing and cytogenomic arrays</li> </ul>	<ul style="list-style-type: none"> <li>• <b>HiSeq 3000/HiSeq 4000 Systems</b> for production-scale genomics</li> <li>• <b>HiSeq 2500 Systems</b> for large-scale genomics</li> </ul>	<ul style="list-style-type: none"> <li>• <b>HiSeq X Five System</b> for production-scale whole-genome sequencing</li> <li>• <b>HiSeq X Ten System</b> for population-scale whole-genome sequencing</li> </ul>

<https://www.illumina.com/systems.html>

# Illumina HiSeq 2500 or X, or NovaSeq

Read length	Dual Flow Cell <b>HiSeq 2500</b>	Single Flow Cell	Dual Flow Cell Run Time
1×36	128-144 Gb	64-72 Gb	29 hrs
2×50	360-400 Gb	180-200 Gb	2.5 days
2×100	720-800 Gb	360-400 Gb	5 days
2×125	900-1 Tb	450-500 Gb	6 days
Reads Passing Filter	Up to 4 billion single read or 8 billion paired-end reads	Up to 2 billion single read or 4 billion paired-end reads	
Quality	Greater than 85% of bases above Q30 at 2×50 bp Greater than 80% of bases above Q30 at 2×100 bp Greater than 80% of bases above Q30 at 2×125 bp		



Table 1: HiSeq X Ten Performance Parameters\*

	Dual Flow Cell	Single Flow Cell
Output/Run	1.6–1.8 Tb	800–900 Gb
Reads Passing Filter†	5.3–6 billion	2.6–3 billion
Supported Read Length	2 × 150	
Run Time	< 3 days	
Quality	≥ 75% of bases above Q30 at 2 × 150 bp	

## Sequencing Output per Flow Cell

	NovaSeq 6000 System		
Flow Cell Type	S1*	S2	S4
2 × 50 bp	134-167 Gb	280-333 Gb	NA†
2 × 100 bp	266-333 Gb	560-667 Gb	NA†
2 × 150 bp	400-500 Gb	850-1000 Gb	2400-3000 Gb

# Illumina MiSeq & MiniSeq Personal Sequencers

## MISEQ REAGENT KIT V2

READ LENGTH	TOTAL TIME*	OUTPUT
1 × 36 bp	~4 hrs	540-610 Mb
2 × 25 bp	~5.5 hrs	750-850 Mb
2 × 150 bp	~24 hrs	4.5-5.1 Gb
2 × 250 bp	~39 hrs	7.5-8.5 Gb



## MISEQ REAGENT KIT V3

READ LENGTH	TOTAL TIME*	OUTPUT
2 × 75 bp	~20 hrs	3.3-3.8 Gb
2 × 300 bp	~55 hrs	13.2-15 Gb



### **MiniSeq:**

**Read length**

**Output**

**2 x 300 bp**

**15 Gbp (25 mln reads)**

# Third-Fourth Generation Sequencing Technologies

---

- Pacific Biosciences (<http://www.pacificbiosciences.com/>): PacBioRS
- BGI-Shenzhen (former Complete Genomics ) (<http://www.completegenomics.com/>) (only for human genome)
- Ion Systems (Life/ABI/Thermo Scientific Inc.) (<http://www.iontorrent.com> <http://www.lifetechnologies.com>): Ion Torrent, Ion Proton, Ion S5, Ion Personal Genome Machine (PGM™) sequencers

## Fourth generation:

- Oxford Nanopore (<http://www.nanoporetech.com/>): GridION, PromethION and miniaturised MinION
- Roche Nanopore Sequencing (<http://sequencing.roche.com/>)





# PacBio Single Molecule Real Time (SMRT) Sequencing

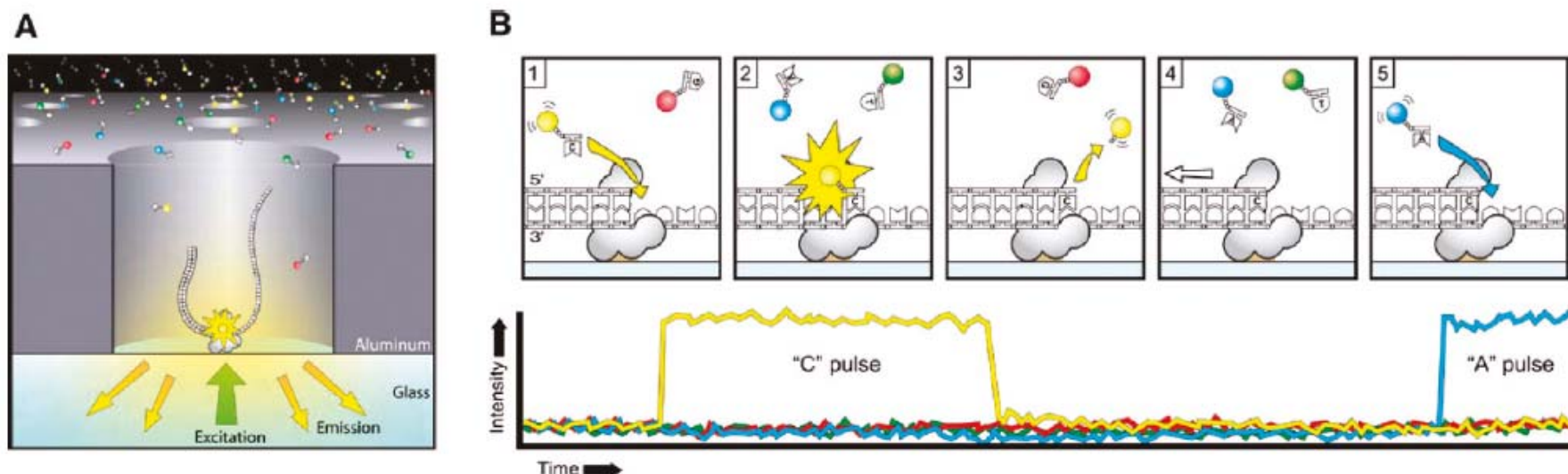
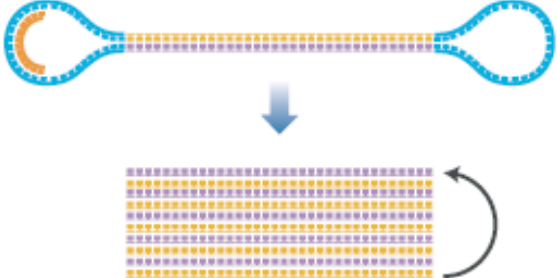
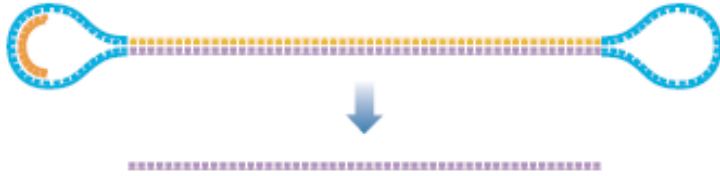


Figure 2. Schematic of PacBio's real-time single molecule sequencing. (A) The side view of a single ZMW nanostructure containing a single DNA polymerase ( $\Phi 29$ ) bound to the bottom glass surface. The ZMW and the confocal imaging system allow fluorescence detection only at the bottom surface of each ZMW. (B) Representation of fluorescently labeled nucleotide substrate incorporation on to a sequencing template. The corresponding temporal fluorescence detection with respect to each of the five incorporation steps is shown below. Reprinted with permission from ref 39. Copyright 2009 American Association for the Advancement of Science.

**ZMW = Zero Mode Waveguide**

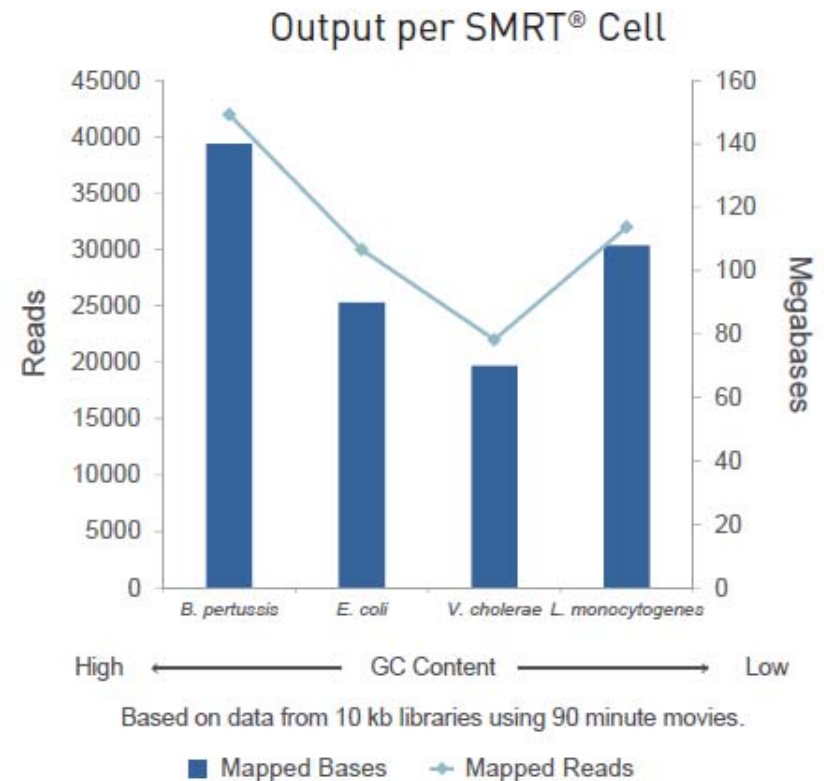
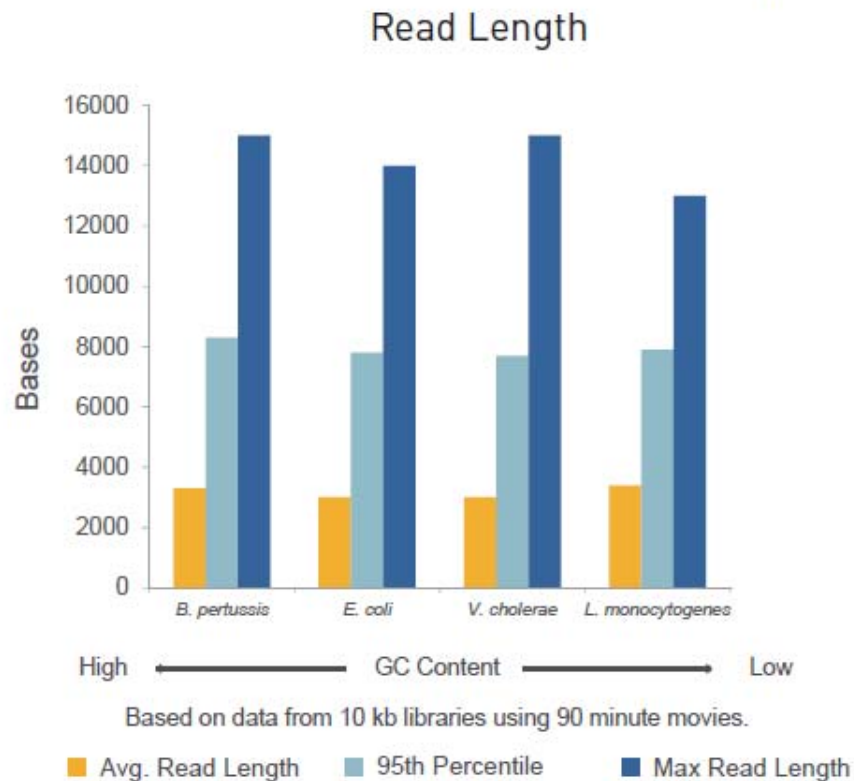
<http://www.youtube.com/watch?v=v8p4ph2MAvI>

# PacBio Sample prep

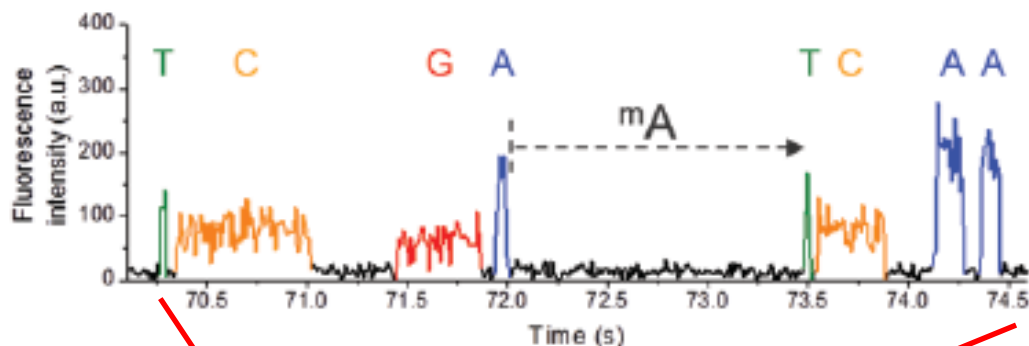
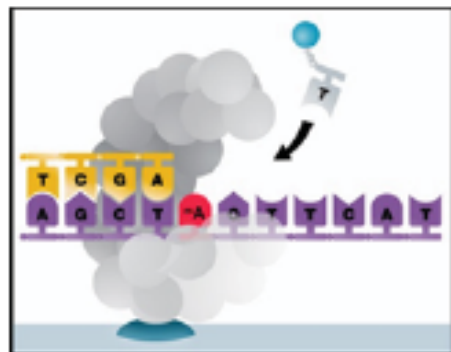
Insert Size (bp)	Input DNA required per library prep (ng)	Sequencing protocols per insert size
250	250	 <p>Circular consensus provides multiple subreads on shorter insert sizes.</p>
500	250	
1,000	500	
2,000	500	 <p>Standard sequencing provides a single pass read on longer insert sizes.</p>
5,000	2,000	
10,000	5,000	

# PacBio SMRT Sequencing

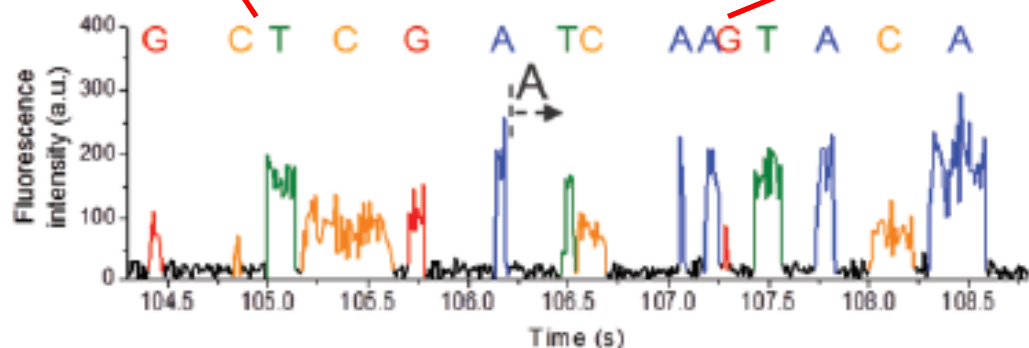
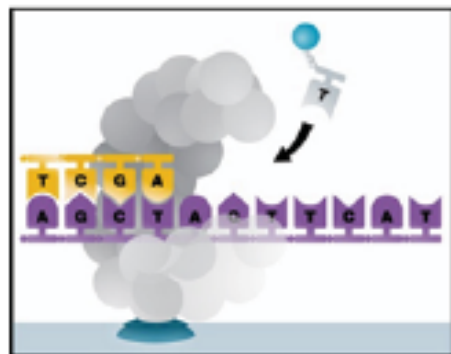
## Typical Results



# PacBio Base Modification Detection (Application in development)



**methylated**

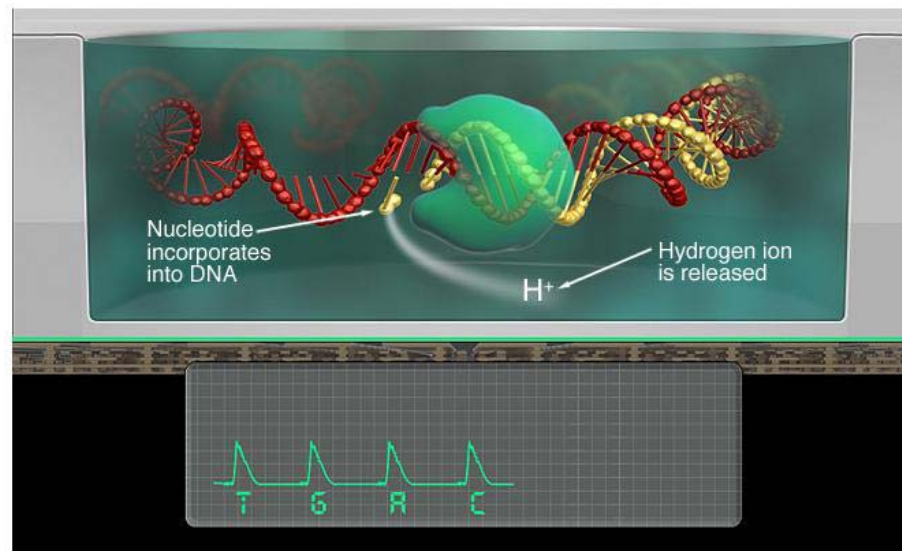
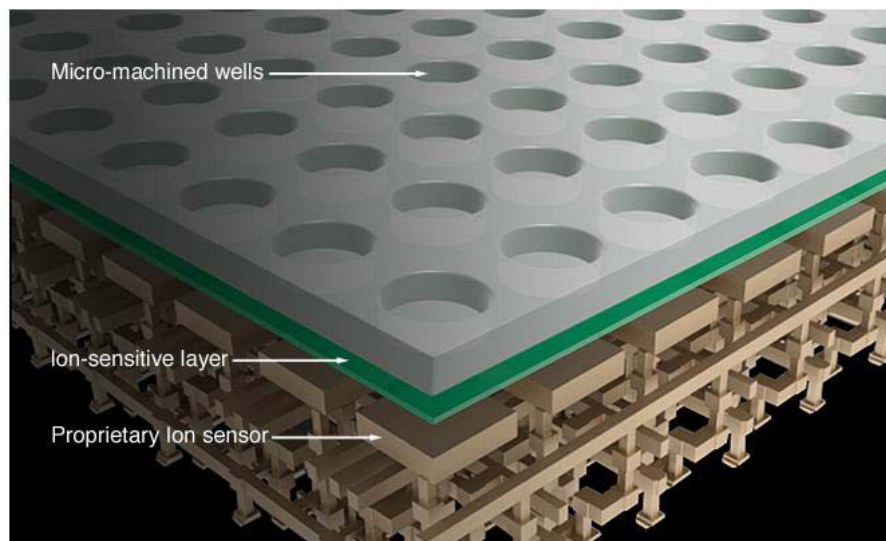


**unmethylated**

A methylated adenine in the template (top) slows the incorporation of a thymine in the replicating strand of DNA. The rate of incorporation can be compared to an unmodified version of the same template (bottom) which has a much faster thymine addition. Differences between the modified and unmodified incorporation rates indicate potential sites of modified bases. These differences often span multiple bases, creating a distinctive signature.

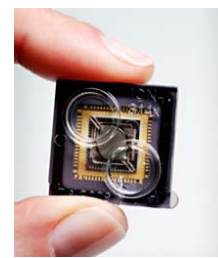
Flusberg et al. (2010) Nature Methods 7: 461-465

# Ion Torrent & Proton: Personal Genome Machine (PGM)



**When a nucleotide is added to a DNA template and is then incorporated into a strand of DNA, a hydrogen ion is released. The charge from that ion changes the pH of the solution, which can be detected by a ion sensor.**

- reads 200 or 400 bp long (S5, Torrent, Proton, PGM)
- from 0.6 to 15 Gb per run (different chips)
- max 3-80 M reads (2.5-16.5 hours)
- useful for Amplicon-seq, small RNA-seq, small genomes sequencing (i.e. bacterial, virus)



(<https://www.thermofisher.com/de/en/home/life-science/sequencing/next-generation-sequencing.html>)

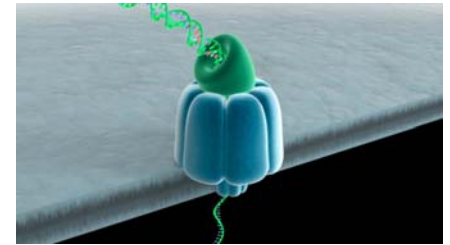
<http://www.youtube.com/watch?v=yVf2295JqUg>



# Oxford Nanopore: The GridION system

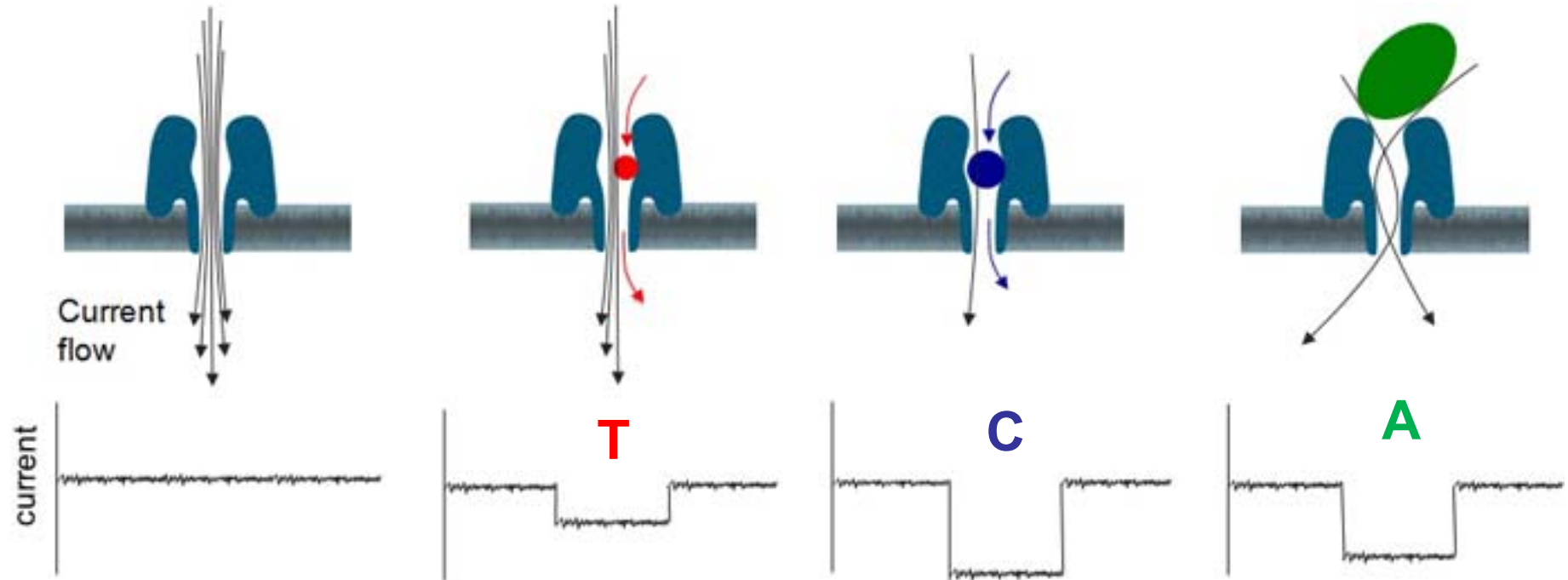
---

- Electrical single-molecule sequencing
- Protein nanopores used as biosensor
- Exonuclease sequencing: combining a protein nanopore and processive enzyme for the sequential identification of DNA bases as they pass through the pore
- Oxford Nanopore signed a commercialisation agreement with Illumina for this technology



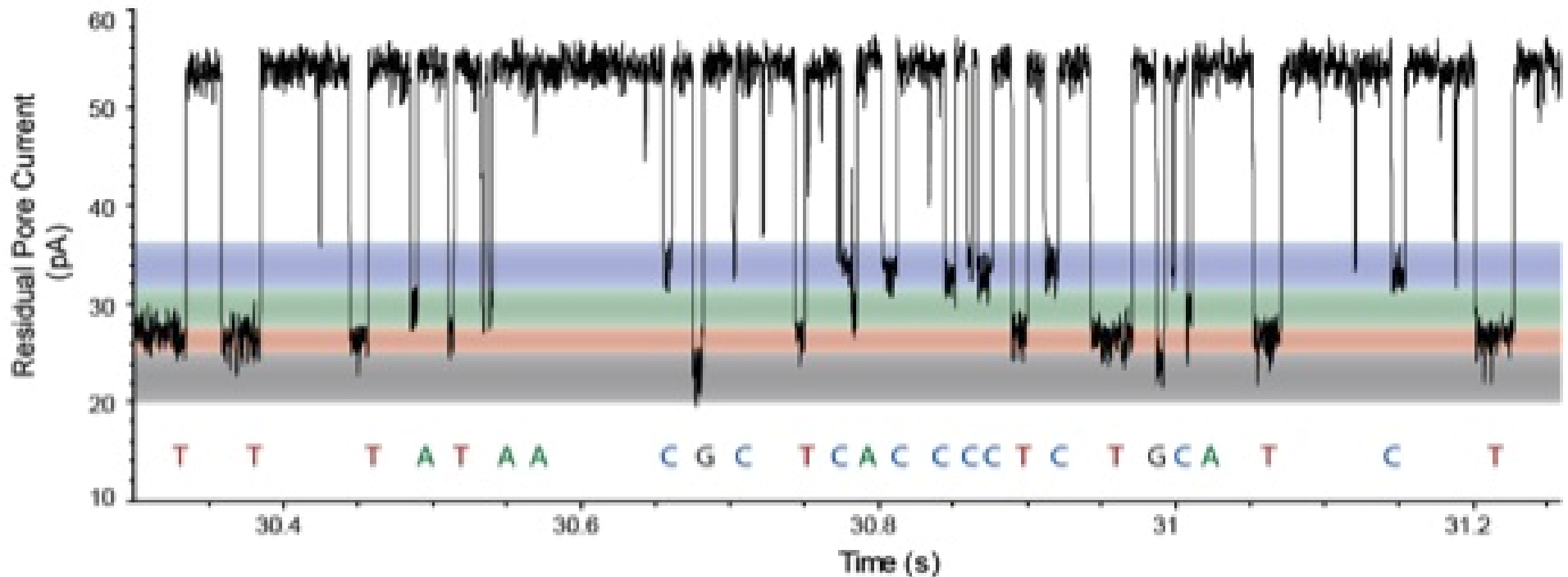


# The GridION system: Nanopore sensing



**Exonuclease sequencing:** combining a protein nanopore and processive enzyme (e.g., a exonuclease) for the sequential identification of DNA bases as they pass through the pore.

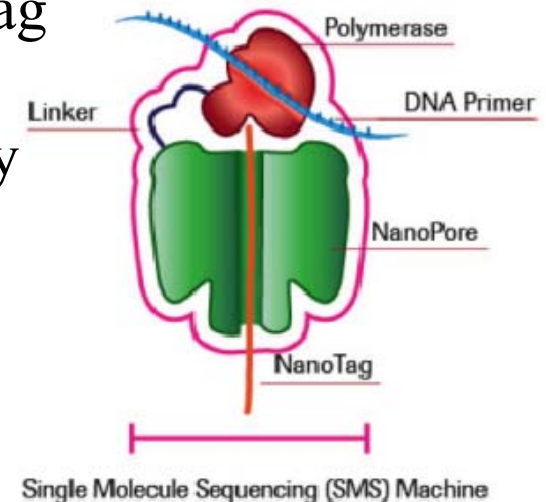
# The GridION system: Electrical sequencing trace



- The system is designed to give ultra-high read length (tens of Kb)
- First generation commercial system is designed to achieve tens of Gb per day

# Roche Nanopore Sequencing: Genia Technologies

- electrical single-molecule sequencing
- biological nanopores used as biosensor (a protein pore embedded in a lipid bilayer membrane)
- based on a proprietary integrated circuit and NanoTag chemistry from Genia Technologies developed in collaboration with Columbia and Harvard University
- uses a DNA replication enzyme to sequence a template strand with single base precision as base-specific engineered NanoTags are captured by the nanopore
- as the phosphate tagged nucleotides (NanoTags) enter the pore, they attenuate the current flow through the pore in an identity-dependent manner



<http://sequencing.roche.com/research---development/nanopore-sequencing.html>

# Third-Forth Generation Sequencing Technologies

Company	Platform name	Method of sequencing	Method of detection	Read length	Advantages	Relative limitation
Pacific Biosciences	PacBio RS	Real-time, singlemolecule DNA sequencing	Fluorescence/Optical	variable: 200 bp to few thousand Kbp	Long average read lengths; decreased sequencing time compared to seggen platforms; no amplification of sequencing fragments; longest individual reads approach 3,000 bases	Inefficient loading of DNA polymerase in ZMWs; low single-pass accuracy (81-83%); degradation of the polymerase in ZMWs; overall, high cost per base (expensive instrument)
Complete Genomics / BGI	In-house lab-built instrumentation BGISEQ-50 & -500	Combinatorial probe anchor hybridization and ligation (cPAL)	Fluorescence/Optical	35-300 bp	Highest (claimed) throughput of thirdgen platforms; lowest reagent cost for reassembling a human genome of all sequencing technologies; each sequencing step is independent, minimizing accumulation of errors	Short read lengths; template preparation prevents sequencing through long repetitive regions; labor intensive sample preparation; no commercially available instrument
Ion Torrent/Life Technologies	Personal Genome Machine (PGM) sequencer	Sequencing by synthesis	Change in pH detected by Ion-Sensitive Field Effect Transistors (ISFETs)	175, 200 or 400 bp	Direct measurement of nucleobase incorporation events; DNA synthesis reaction operates under natural conditions (no need for modified DNA bases)	Sequential washing steps can lead to accumulation of errors; potential difficulties in reading through highly repetitive or homopolymer regions of the genome
Oxford Nanopore	gridION	Nanopore exonuclease sequencing	Current	9-10 Kb	Potential for long read lengths; low cost of RHLnanopore production; no fluorescent labeling or optics necessary	Cleaved nucleotides may be read in the wrong order; difficult to fabricate a device with multiple parallel pores

## 2016 NGS Field Guide: Overview

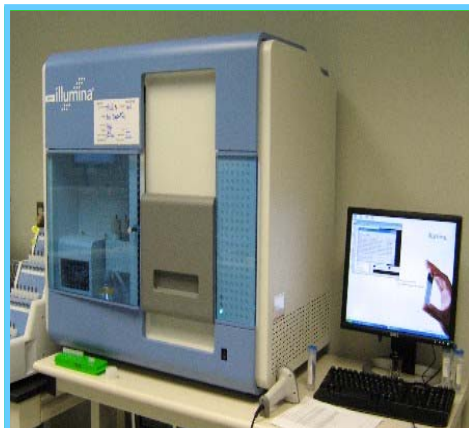
<http://www.molecular ecologist.com/next-gen-fieldguide-2016/>

## Single-molecule Label-free Electronic DNA Sequencer

<https://gamma-dna.com>



# NGS Platform statistics



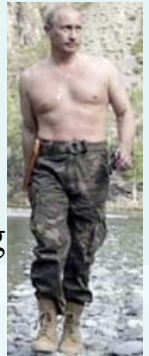
Instrument	Amplification	Run time	M of Read / run	Bases / read	Reagent Cost / run, \$	Reagent Cost / Gb, \$	Reagent Cost /M reads, \$	Gbp / run	cost / Gb, \$
Illumina HiSeq X (2 flow cells)	BridgePCR	3 days	6000	300	12750	7	2	1800	7
Illumina HiSeq 2500 - high output v4	BridgePCR	6 days	2000	250	14950	30	7	500	30
Life Technologies SOLiD – 5500xl	emPCR	8 days	1410	110	10503	68	7	155.1	68
Illumina NextSeq 500	BridgePCR	30 hrs.	400	300	4000	33	10	120	33
Oxford Nanopore GridION 8000	None - SMS	varies	10	10000	1000	10	100	100	10
Ion Torrent - Proton III	emPCR	6 hrs.	500	175	1000	11	2	87.5	11
Illumina MiSeq v3	bridgePCR	55 hrs.	22	600	1442	109	66	13.2	109
Ion Torrent – PGM 318 chip	emPCR	7.3 hrs.	4.75	400	874	460	184	1.9	460
Ion Torrent – PGM 316 chip	emPCR	4.9 hrs.	2.5	400	674	674	270	1	674
Oxford Nanopore MinION	None - SMS	≤6 hrs.	0.1	9000	900	1000	9000	0.9	1000
454 FLX+	emPCR	20 hrs.	1	650	6200	9538	6200	0.65	9538
Ion Torrent – PGM 314 chip	emPCR	3.7 hrs.	0.475	400	474	2495	998	0.19	2495
Pacific Biosciences RS II	None - SMS	2 hrs.	0.03	3000	100	1111	3333	0.09	1111

**‘Benchtop’ NGS technology:** will substitute the capillary electrophoresis (CE) sequencers for common experiments, such as Illumina libraries verification, amplicon sequencing and small genome sequencing.

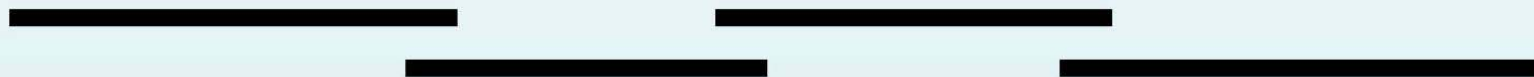
# Major approaches for *de novo* complete genome sequencing

## 1. Traditional based on BAC library – very costly for large genomes:

- 1) developing BAC library (arranged and microplatted) (costly for large genomes, for instance, loblolly pine (*Pinus taeda*) library with ~1.8M BACs (~8X genome coverage) costs ~\$1.8M (<http://www.pine.msstate.edu/bac.htm>);
- 2) fingerprinting each clone - ~\$ 5 mln;
- 3) selection of unique overlapping BACs, building a “**minimum tiling path**” and sequencing

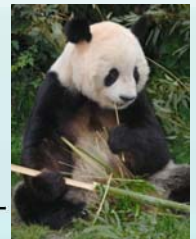


### Minimum path (1.3X)



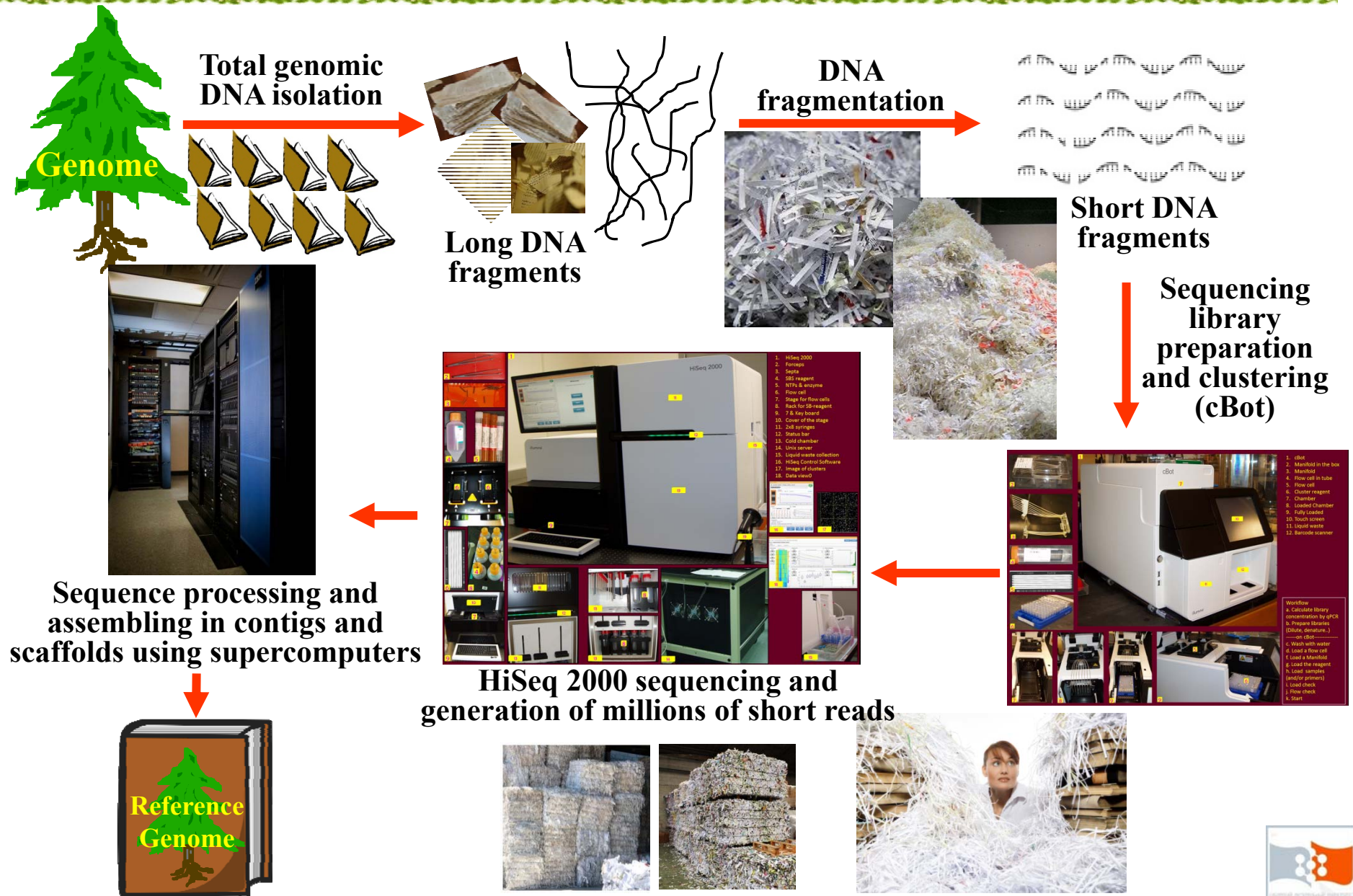
## 2. Whole Genome Shot-Gun Sequencing (WSGS) using NGS platforms:

### Shotgun sequencing (30X or more)

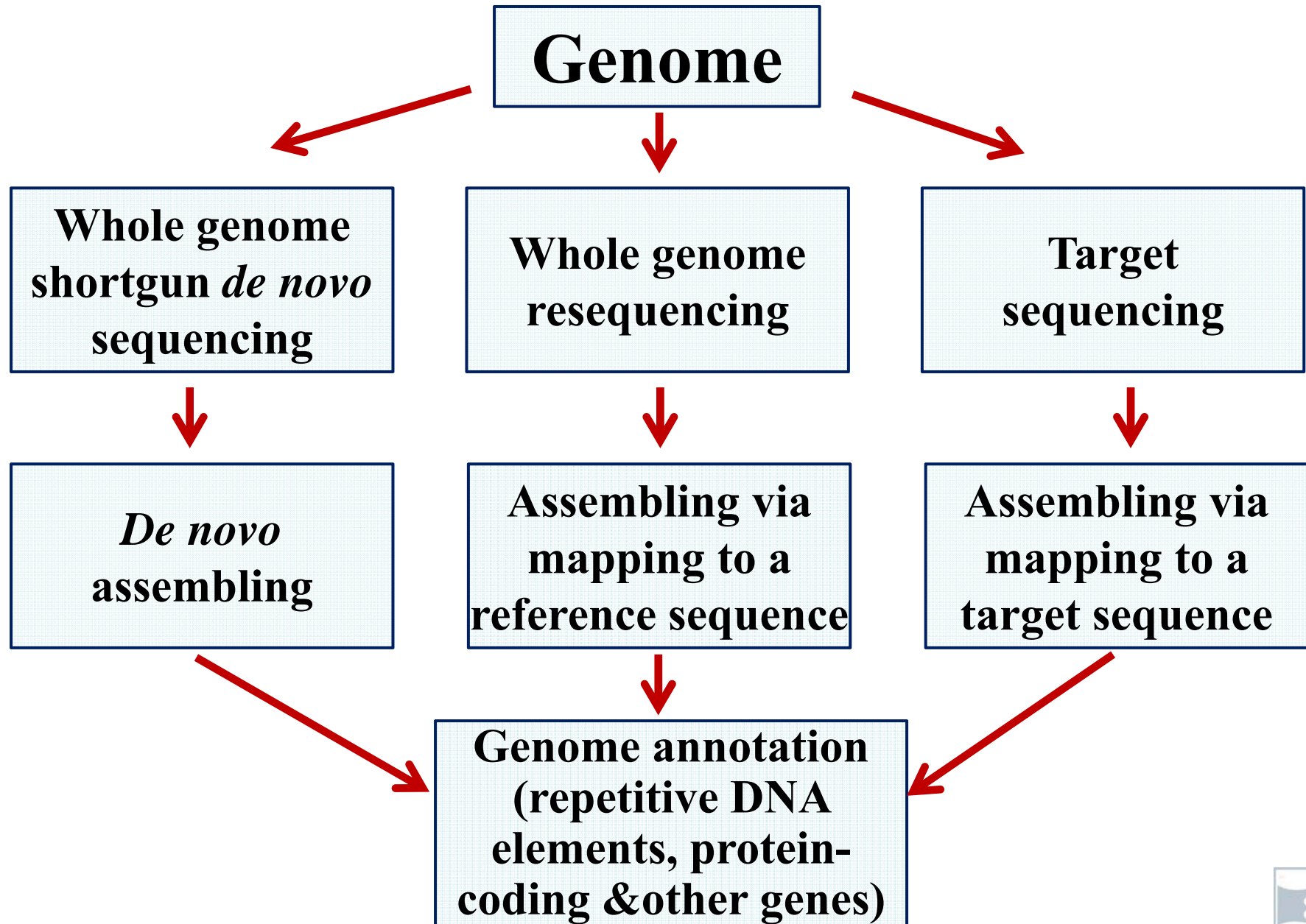




# Whole genome *de novo* sequencing and assembling using NGS short reads and supercomputers



# Overview of genome sequencing analysis



<b>Applications of Genome Sequencing</b>		
<b>Purpose</b>	<b>Template</b>	<b>Example</b>
<i>De novo</i> sequencing	genome sequencing	<b>The Genome 10K project</b> (sequencing 10,000 vertebrate species genomes, approximately one for every vertebrate genus); <b>1K Plant Genomes Project</b> , <b>10K fungi genomes</b>
	ancient DNA	Extinct Neanderthal genome
	metagenomics	Human guts
Resequencing	whole genomes	Sequencing <b>1000 individual human genomes</b> project
	genomic regions	Assessment of genomic rearrangements or disease-associated regions
	somatic mutations	Sequencing mutations in cancer
Transcriptome	full-length transcripts	Defining regulated messenger RNA transcripts
	Noncoding RNAs	Identifying and quantifying microRNAs in samples
Epigenetics	Methylation changes	Measuring methylation changes in cancer

Table 13.15 in Bioinformatics and Functional Genomics by J. Pevsner (2<sup>nd</sup> ed., Wiley-Blackwell, 2009) p. 538



# Ancient genome DNA sequencing projects

---

## Special challenges:

- fragmented & degraded by nucleases
- deamination of cytosine to uracil
- the majority of DNA from unrelated organisms such as bacteria that invaded after death
- The majority of DNA in samples is contaminated by human DNA
- Determination of authenticity requires special controls, and analysis of multiple independent extracts



Green, R. E. *et al.* (2010) A draft sequence of the Neandertal genome. *Science* 328, 710–722

# Microbial Community Sequencing (Metagenomics) Projects

---

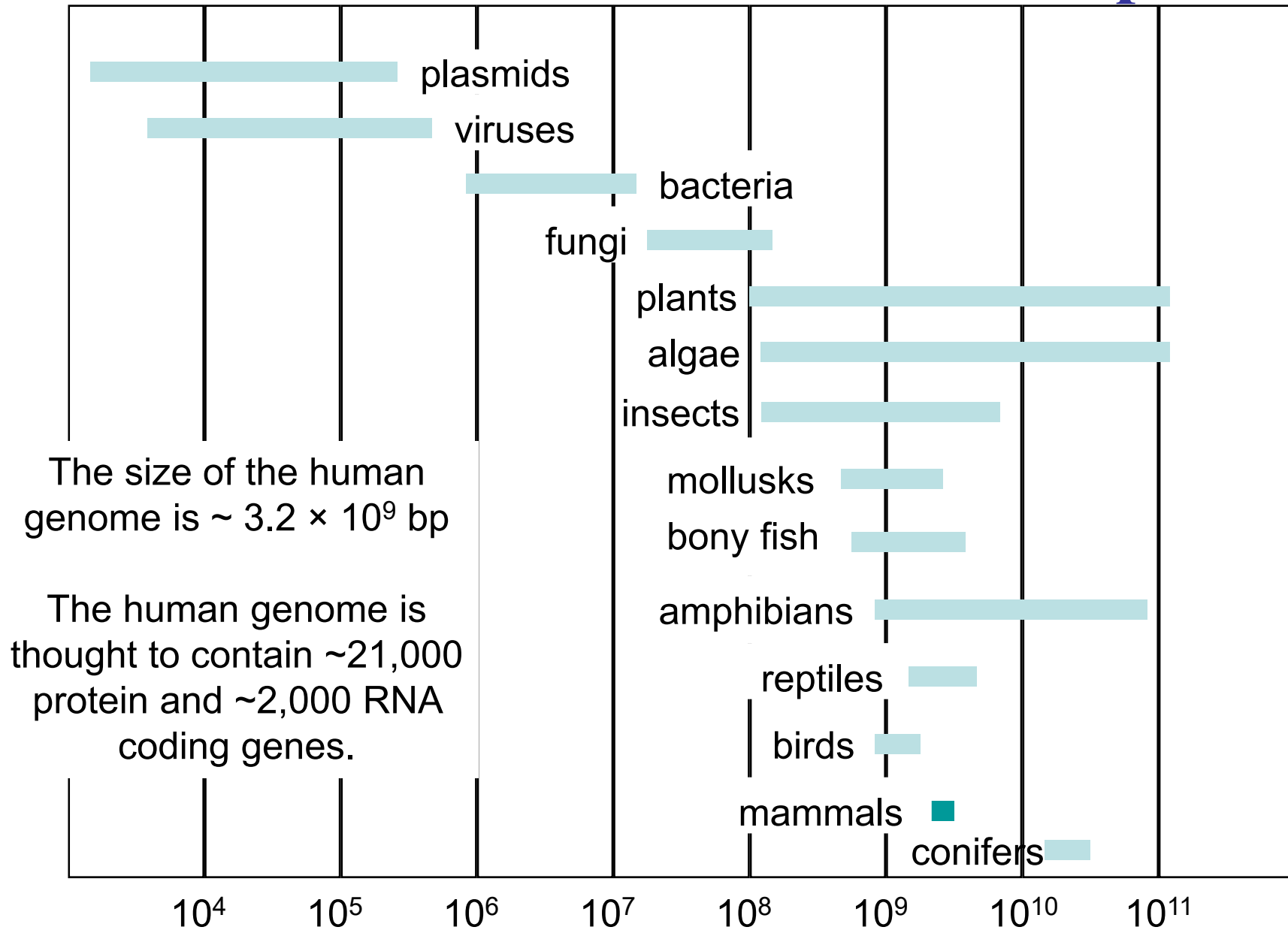
## Two broad areas:

- Environmental (ecological)  
e.g. hot spring, ocean, sludge, soil
- Organismal  
e.g. human gut, feces, lung





# Genome sizes in nucleotide base pairs



<http://www3.kumc.edu/jcalvet/PowerPoint/bioc801b.ppt>





## Eukaryotic completed genome projects > 2 Gb

Genus, species	Subgroup	Size (Mb)	#chr	Common name
<i>Pinus taeda &amp; lambertiana</i>	Land Plants	20150 & 28900	12	Loblolly & sugar pines
<i>Picea abies</i>	Land Plants	20000	12	Norway spruce
<i>Triticum urartu</i>	Land Plants	4940	7	wheat A-genome progenitor
<i>Aegilops tauschii</i>	Land Plants	4360	7	Tausch's goatgrass
<i>Macropus eugenii</i>	Mammals	3800	8	tammar wallaby
<i>Oryctolagus cuniculus</i>	Mammals	3500	22	rabbit
<i>Capsicum annuum</i>	Land Plants	3480	12	pepper
<i>Cavia porcellus</i>	Mammals	3400	31	guinea pig
<i>Homo sapiens</i>	Mammals	3200	23	human
<i>Pan troglodytes</i>	Mammals	3100	24	chimpanzee
<i>Bos taurus</i>	Mammals	3000	30	cow
<i>Dasypus novemcinctus</i>	Mammals	3000	32	nine-banded armadillo
<i>Loxodonta africana</i>	Mammals	3000	28	African savanna elephant
<i>Sorex araneus</i>	Mammals	3000	20	European shrew
<i>Rattus norvegicus</i>	Mammals	2750	21	rat
<i>Nicotiana glauca</i>	Land Plants	2636	12	tobacco
<i>Canis familiaris</i>	Mammals	2400	39	dog
<i>Zea mays</i>	Land Plants	2365	10	corn



# GC content varies across genomes

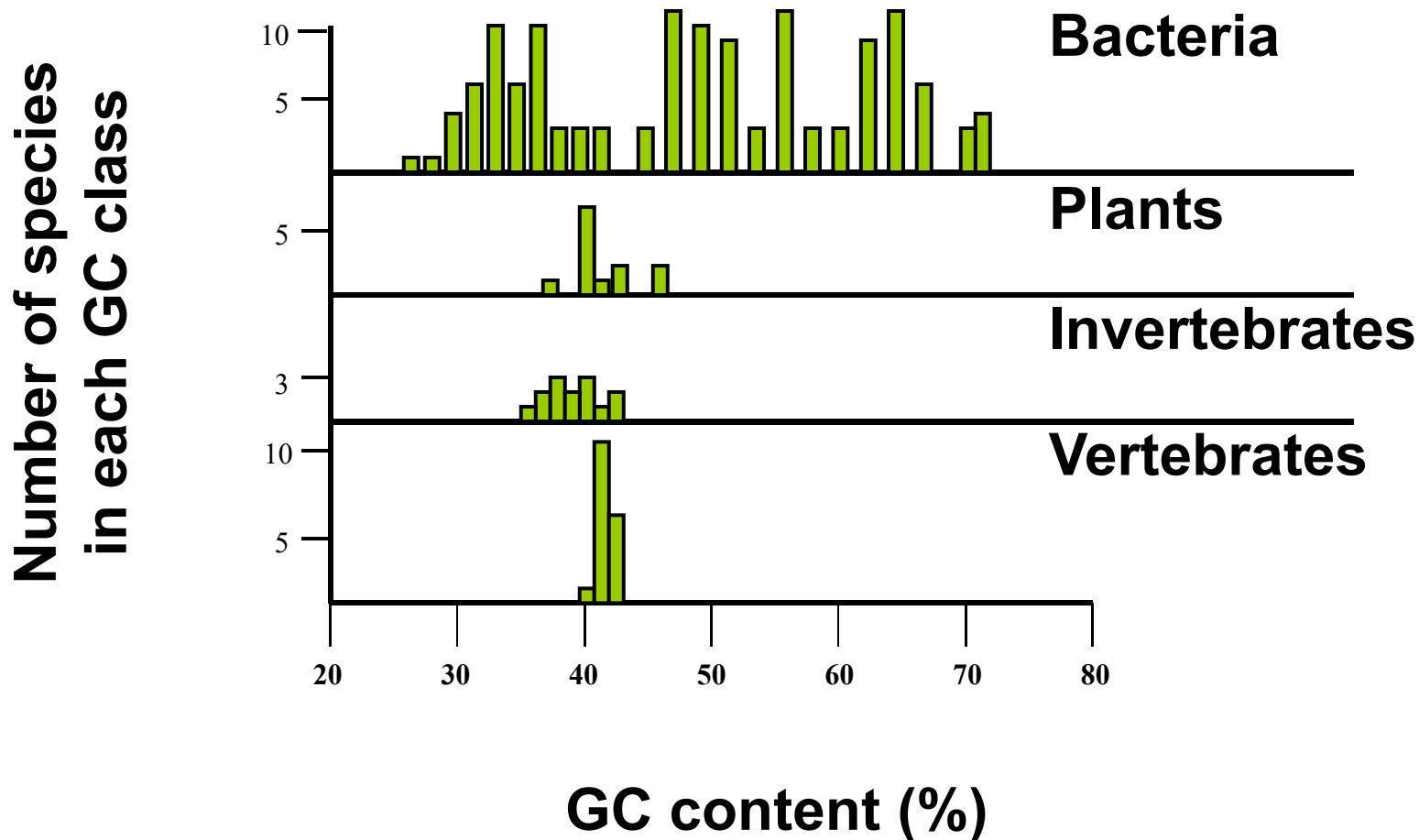
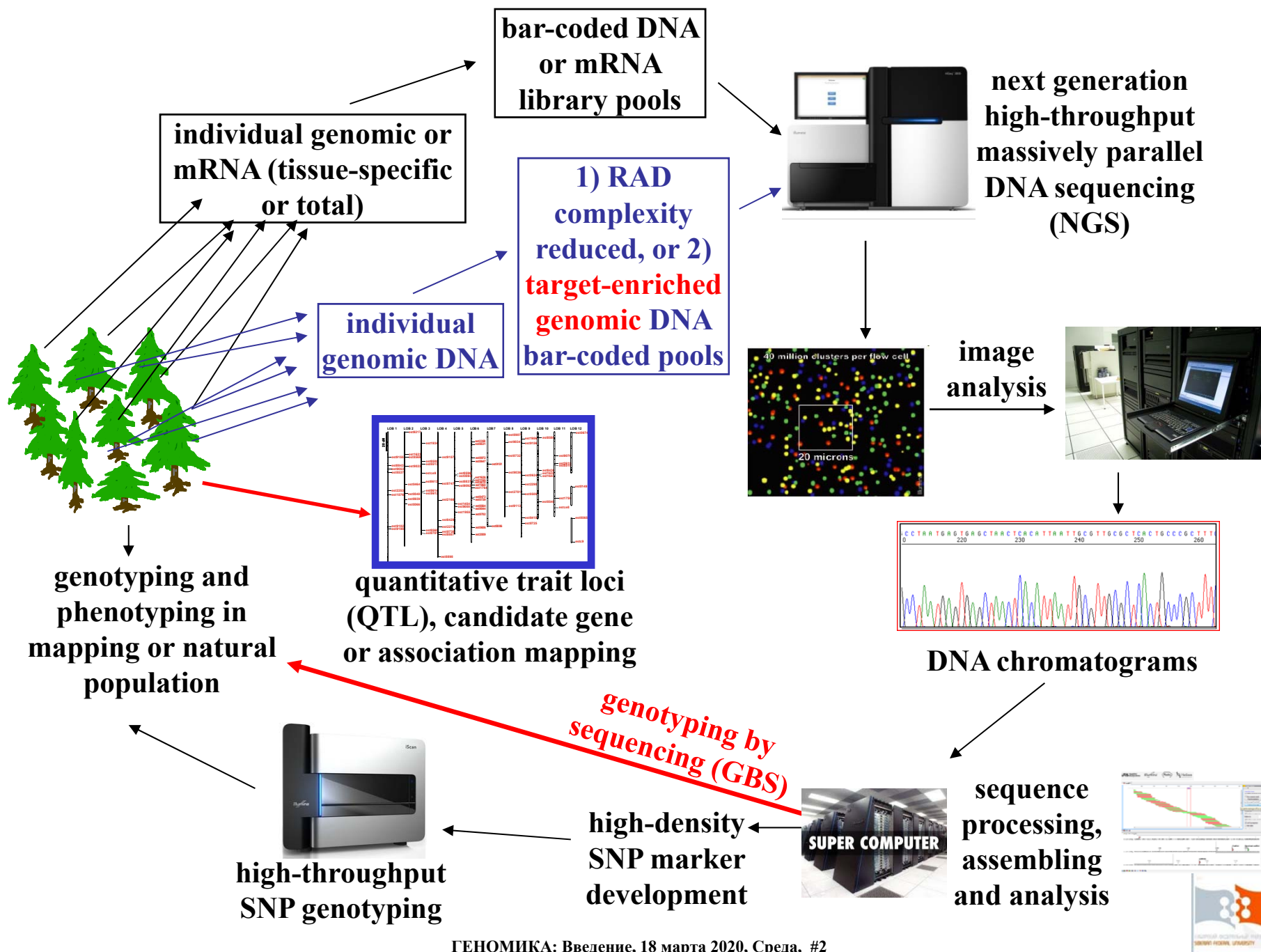


Fig. 13.15 in Bioinformatics and Functional Genomics by J. Pevsner (2<sup>nd</sup> ed., Wiley-Blackwell, 2009) p. 556

## Genomic markers development and genotyping using next generation sequencing



# Genomic DNA target enrichment for high-throughput massively parallel sequencing using the Agilent's SureSelect Target Enrichment System

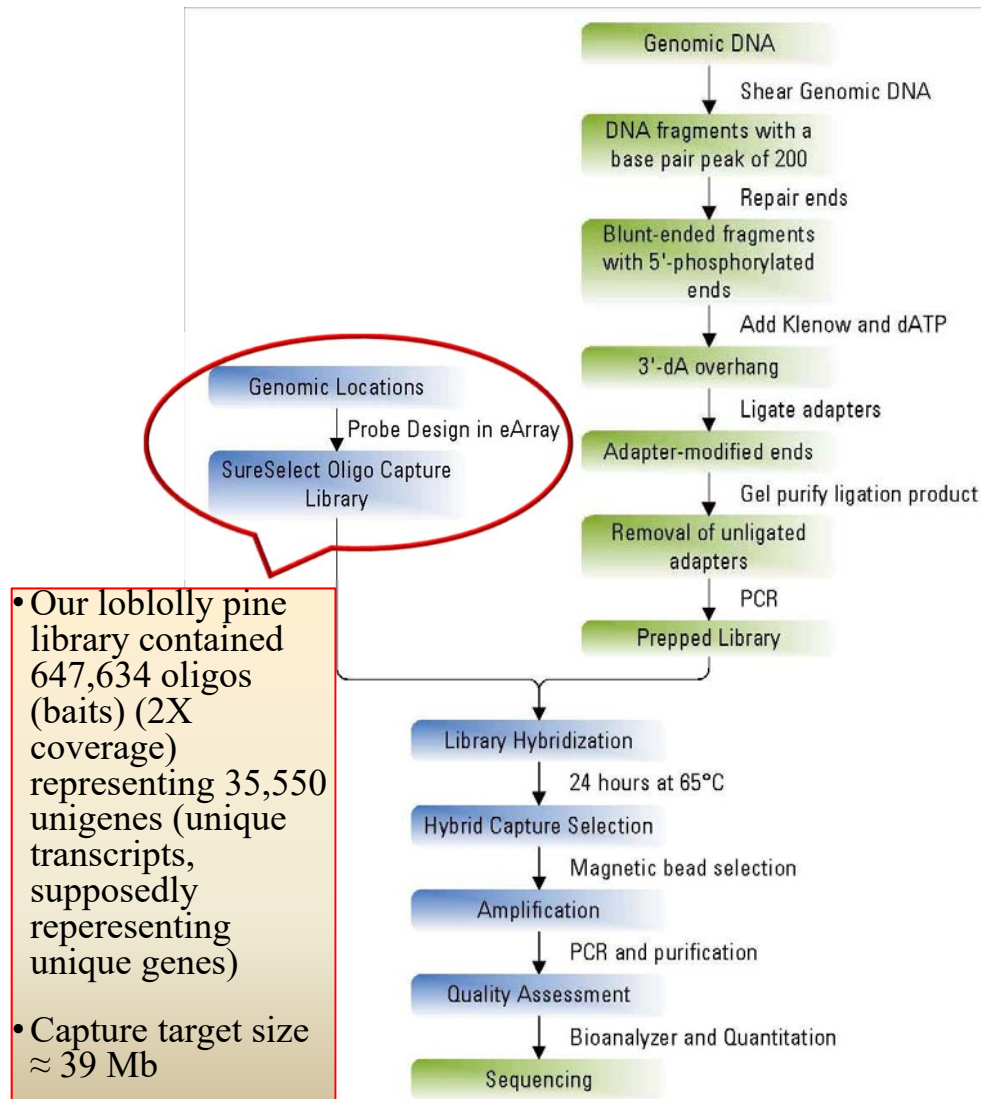


Figure 1 Overall sequencing sample preparation workflow.

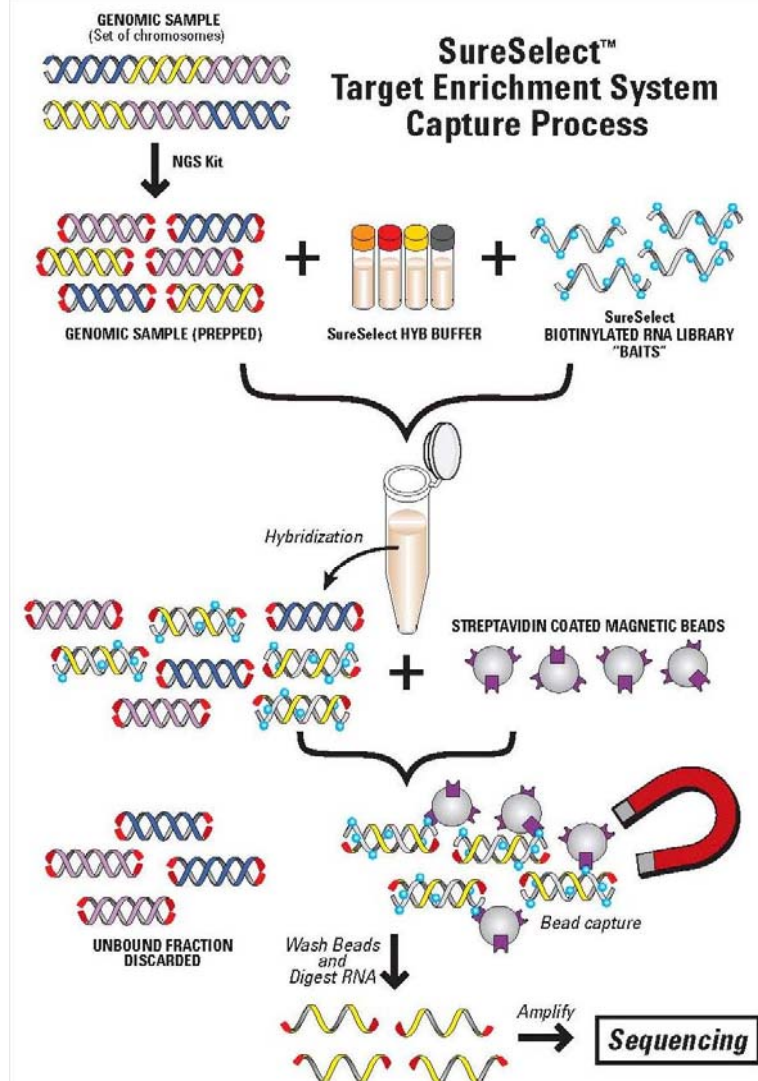
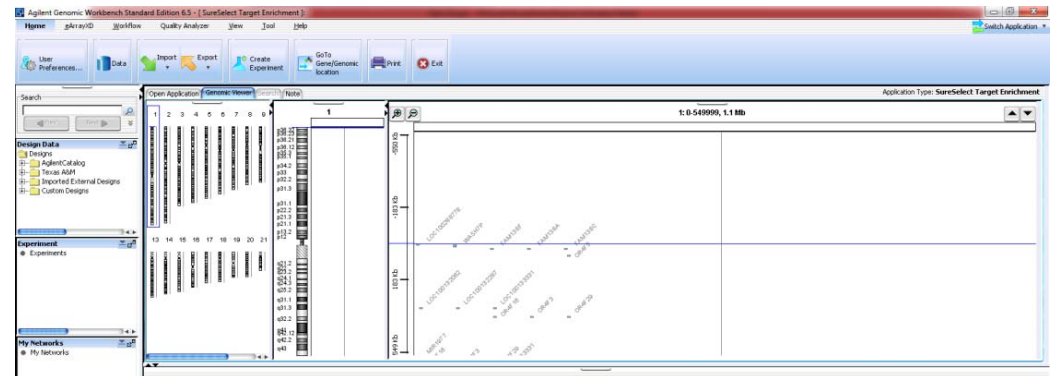


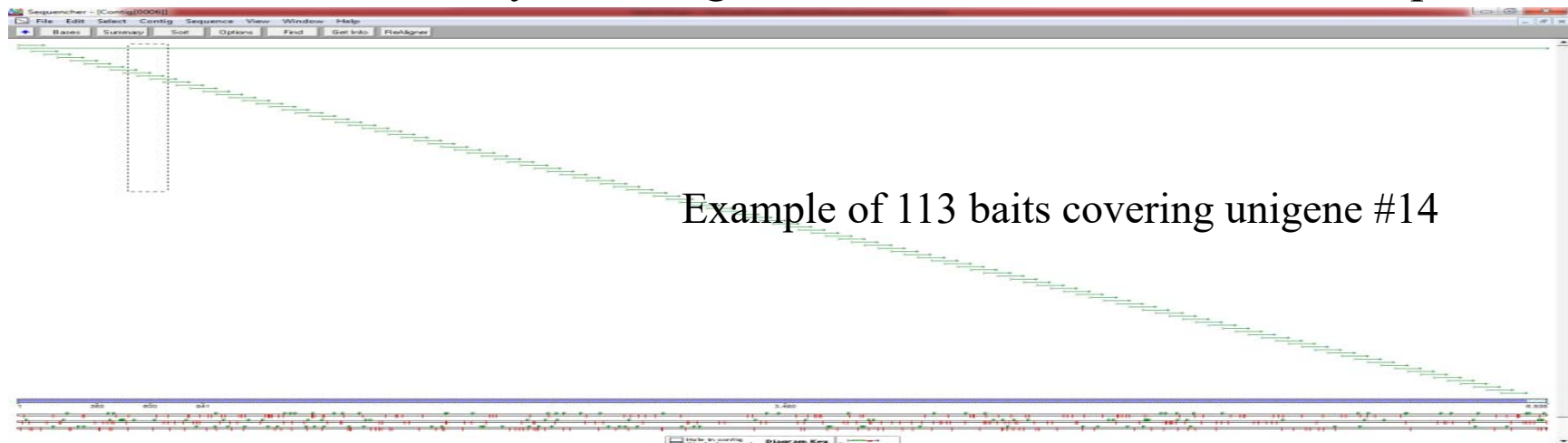
Figure 2 SureSelect Target Enrichment System Capture Process

[http://www.opengenomics.com/SureSelect\\_Target\\_Enrichment\\_System](http://www.opengenomics.com/SureSelect_Target_Enrichment_System)

# Genomic DNA enrichment for the entire exome (~48,000 genes) in loblolly pine for high-throughput massively parallel sequencing using bar-coding and the Roche NimbleGen Inc. Target Enrichment System



Several million oligonucleotide hybridization 50-100 bp long probes (baits) based on 196,068 exons were designed to target 49 Mbp of gene space using Roche NimbleGen Inc. system to gene enrich DNA libraries for sequencing



Example of 113 baits covering unigene #14

Lu, M., K. V. Krutovsky, C.D. Nelson, T. E. Koralewski, T. D. Byram, and C. A. Loopstra, 2016 Exome genotyping, linkage disequilibrium and population structure in loblolly pine (*Pinus taeda* L.). BMC Genomics 17:730 (<http://www.biomedcentral.com/content/pdf/s12864-016-3081-8.pdf>)

