Next-generation sequencing

Lecture 1

NGS

- Introduction to the background
- NGS workflow
- Data format
- Assembly
  - Variation discovery
- RNA-seq
  - Aligner
  - Analysis tools
  - Applications, such as MiRNA
- Chip-seq
  - Applications

History: The Search for the Genetic Material

- T. H. Morgan -1910's
  - genes are located on chromosomes
  - composed of DNA and protein
- Frederick Griffith - 1928
  - Pathogenicity trait was heritable even from dead cells
- Erwin Chargaff – 1950
  - DNA is a polymer of nucleotides
  - Base composition varies among species
- Alfred Hershey & Martha Chase – 1952
  - DNA is the genetic material of a phage known as T2

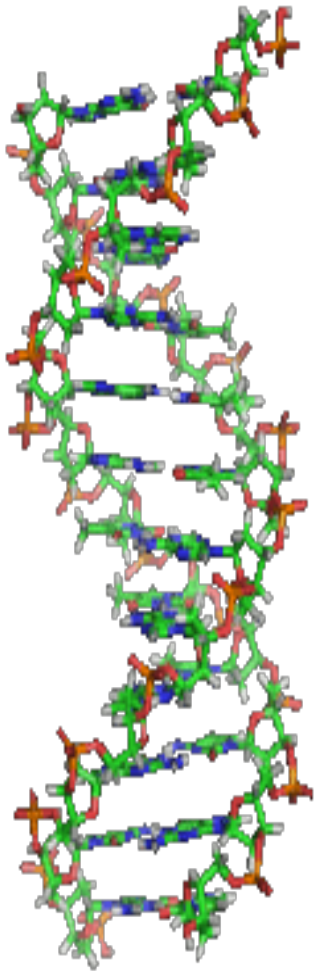
DNA: “Blueprints” for a cell

- Genetic information is stored in the sequence of bases along a nucleic acid chain, called DNA.

- DNA consists of only four types of nucleotides: Adenine (A), Cytosine (C), Guanine (G), Thymine (T)

- “A” pairs with “T”, “C” pairs with “G”

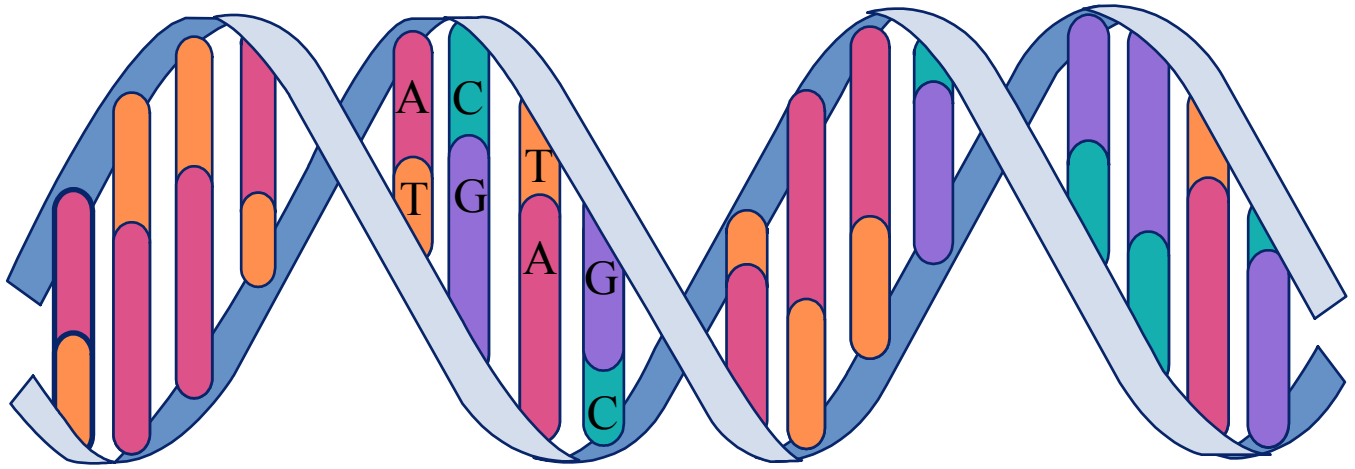
DNA Has Only Two Jobs



- It serves as a store of information
  - Ensuring that information is passed on to each new cell upon division, and the next generation.
- It directs the synthesis of proteins
  - Which are necessary to carry out the functions of a living organism.

DNA's Structure Explains How it Accomplishes Both Jobs

DNA Serves as a Store of Information



Adenine (A)

Always binds to

Thymine (T)

Cytosine (C)

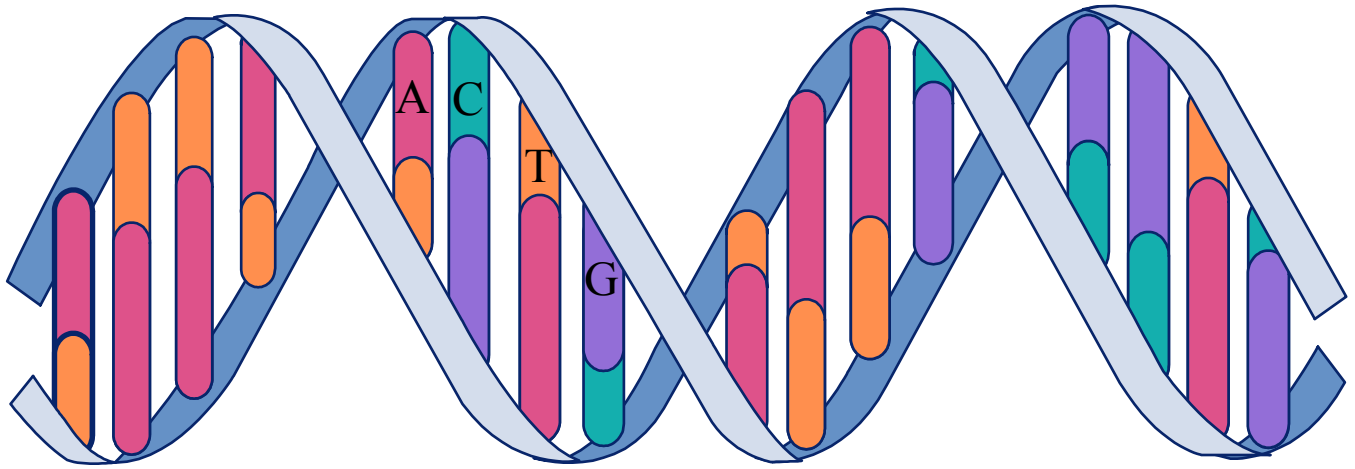
Always binds to

Guanine (G)

As a double helix, each DNA molecule contains a copy of itself

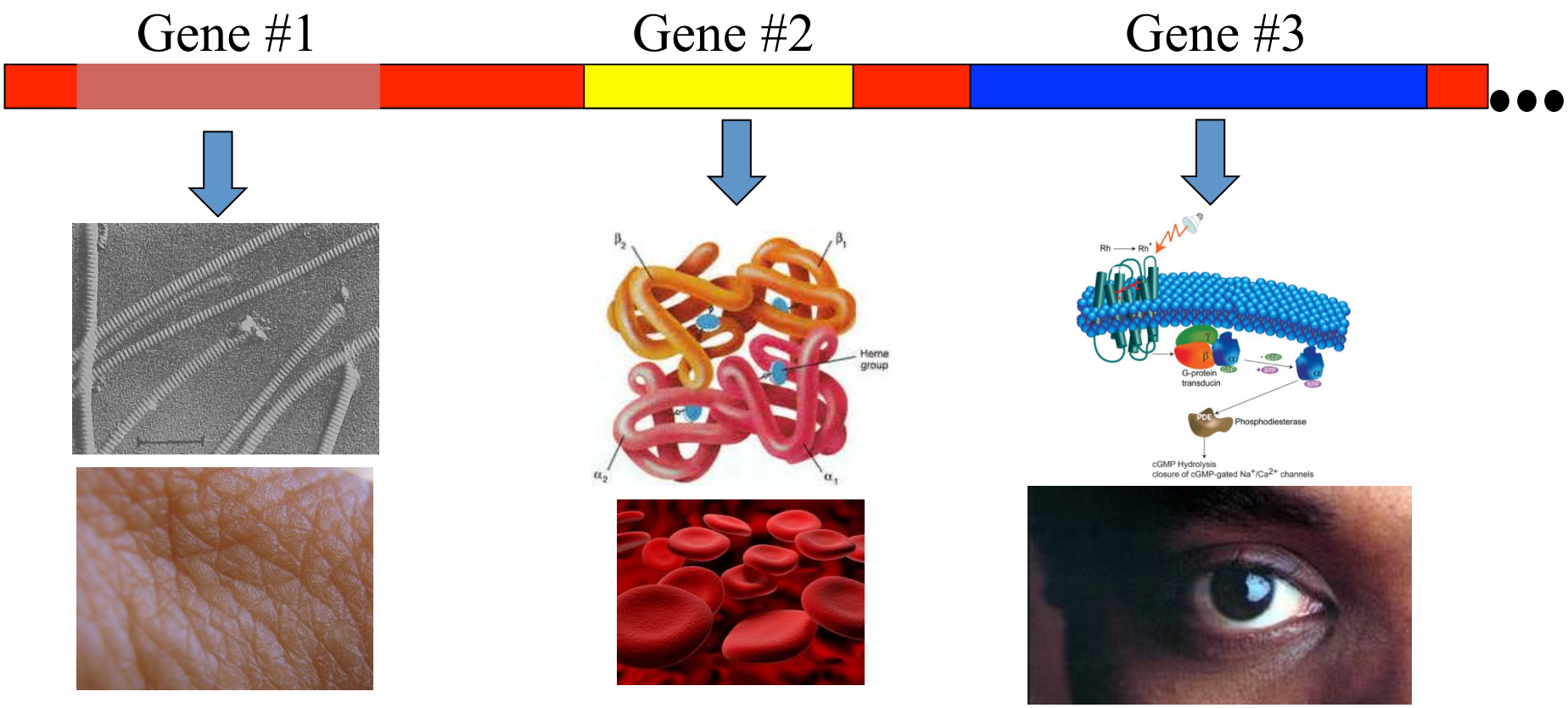
DNA's Structure Explains How it Accomplishes Both Jobs

DNA Directs the Synthesis of Proteins



- It is the *order of the bases* that provides the instructions for protein synthesis
- One stretch of DNA directs the synthesis of one type of protein, another stretch directs the synthesis of another type of protein

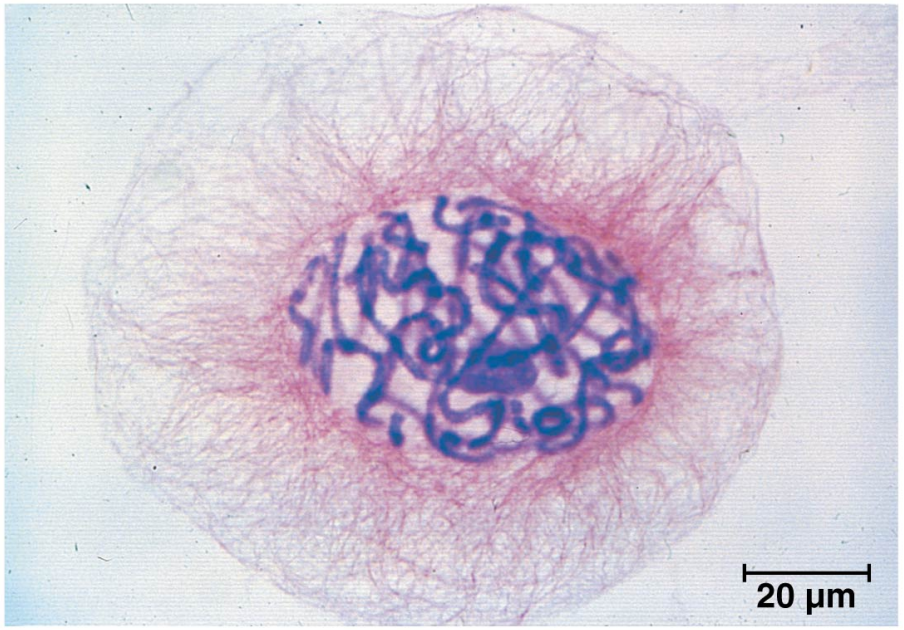
We call a stretch of DNA that directs the synthesis of one particular protein a *Gene*



DNA, Chromosome, Genome

- DNA molecules in a cell are packaged into **chromosomes**
- All the DNA in a cell constitutes the cell's **genome**

Chromosomes are stained in purple



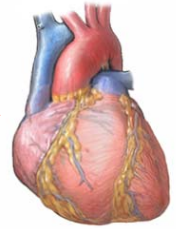
Eukaryotic chromosomes

Human Genome

- Human has 3 billion bases arrayed in a unique order, with ~20,000 genes that direct the synthesis of all the proteins that comprise the body
- The unique order of your bases greatly influences your health, e.g.
  - What disease you are more - or less - prone to
  - How you will react to different medications
- A human "genome" is about 6 feet of DNA
  - And each of your cells contains two copies of your genome

A Human DNA Sequence

- Obtain the sequence
- Store and display the sequence: 4000nt~1/1,000,000th of the Human Genome (need 11.5 days to show if one slide/sec)
- Analysis:
  - discover genes and their interactions
  - identify "polymorphisms" which differ between people because they are sometimes important, influencing traits or medically important characteristics



Sequencing

- Sequencing is the process by which you determine the exact order of the nucleotides in a given region of DNA.

- "Sequencing" DNA is simply the elucidation of the order of the bases in an organism's DNA strand.

Determining the Sequence of DNA

- Methods:
  1. Chain termination or dideoxy method
     - F. Sanger
  2. Next-generation sequence methods
     - Pyrosequencing

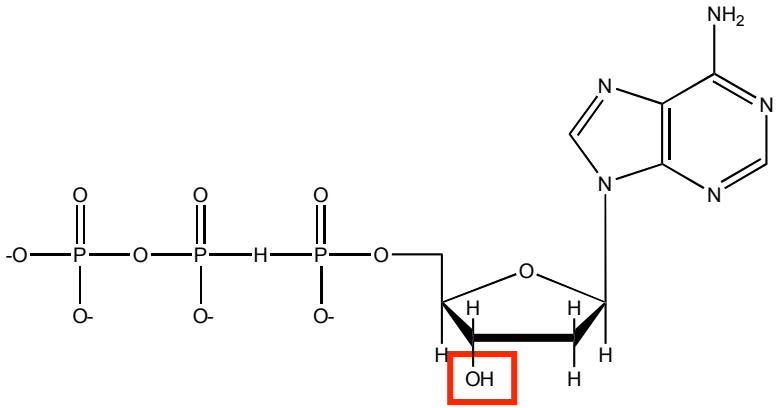
Sanger method

- The standard DNA sequencing technique is the Sanger method, named for its developer, Frederick Sanger, who shared the 1980 Nobel Prize in Chemistry.
- This method begins with the use of special enzymes to synthesize fragments of DNA that terminate when a selected base appears in the stretch of DNA being sequenced.
- These fragments are then sorted according to size by placing them in a slab of polymeric gel and applying an electric field -- a technique called electrophoresis.
- Because of DNA's negative charge, the fragments move across the gel toward the positive electrode. The shorter the fragment, the faster it moves. Typically, each of the terminating bases within the collection of fragments is tagged with a radioactive probe for identification.

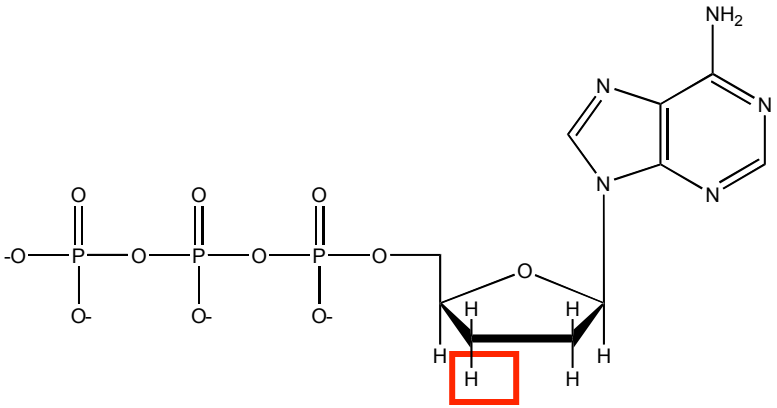
Dideoxynucleotides

Here is an example comparing dATP and ddATP:

dATP

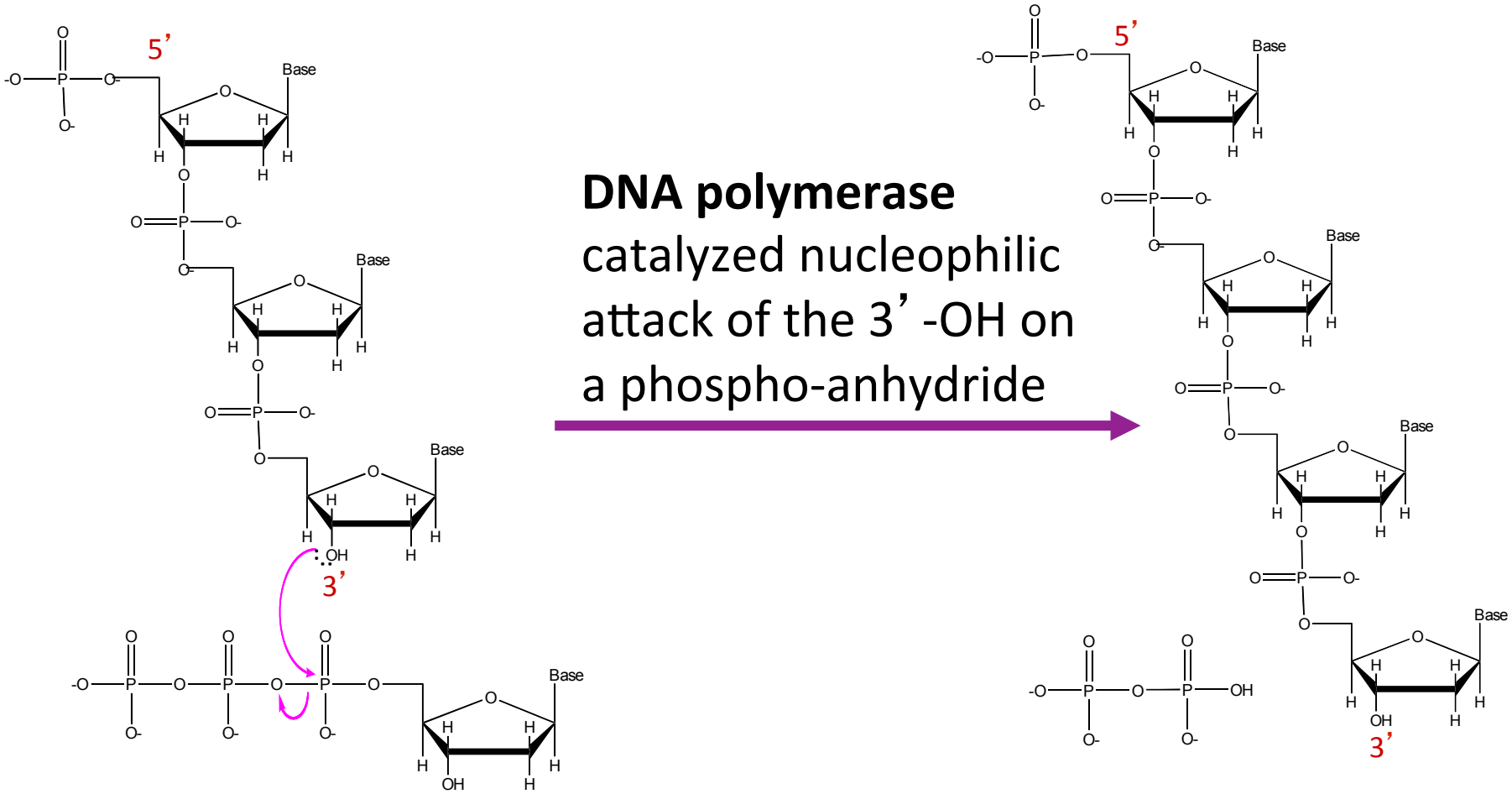


ddATP



The 3’ hydroxyl has been changed to a hydrogen in ddNTP’s, which terminates a DNA chain because a phosphodiester bond cannot form at this 3’ location

Mechanism of DNA polymerization



**** Since the 3' –OH is changed to a –H in ddNTPs, it is unable to form a phosphodiester bond by nucleophilic attack on the phosphate, and it will cause a termination in the DNA chain**

Sanger Method

- 4 Steps:
  1. Denaturation of DNA
  2. Primer attachment and extension of bases. With addition of enzyme (DNA polymerase), the primer is extended until a ddNTP is encountered.
  3. Termination. With the proper dNTP:ddNTP ratio, the chain will terminate throughout the length of the template. All terminated chains will end in the ddNTP added to that reaction.
  4. Gel electrophoresis

Fluorescent Dideoxynucleotide Labeled Sequencing

1. Here we have one reaction vessel, with four copies of our Unknown fragment. The reaction vessels have millions of our unknown fragment.

2. A region of known sequence

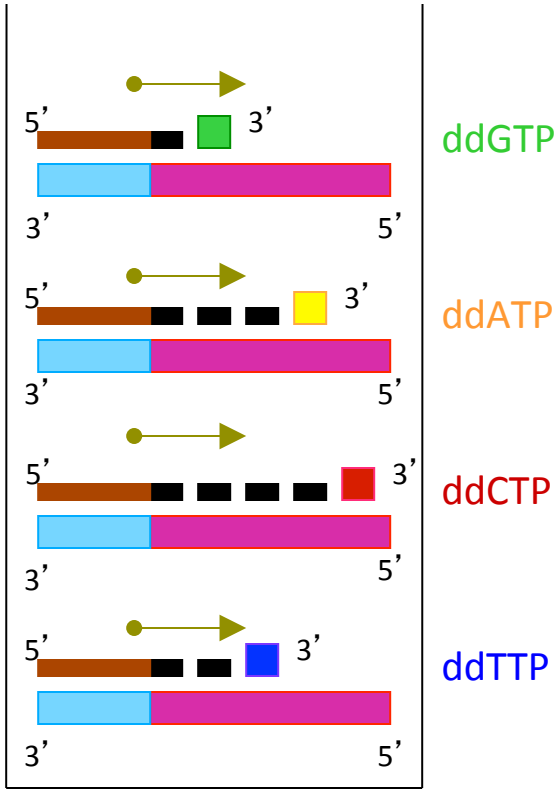
3. Complementary primer

4. dNTP's (dATP, dGTP, dCTP, and dTTP)

5. Fluorescent labeled ddNTP's. Each labeled with a different fluorescent dye

6. DNA Polymerase. Four separate reactions. One type of ddNTP per reaction.

7. Again ddNTP incorporation stops chain synthesis. Add enough so each ddNTP is randomly and completely incorporated at each base.

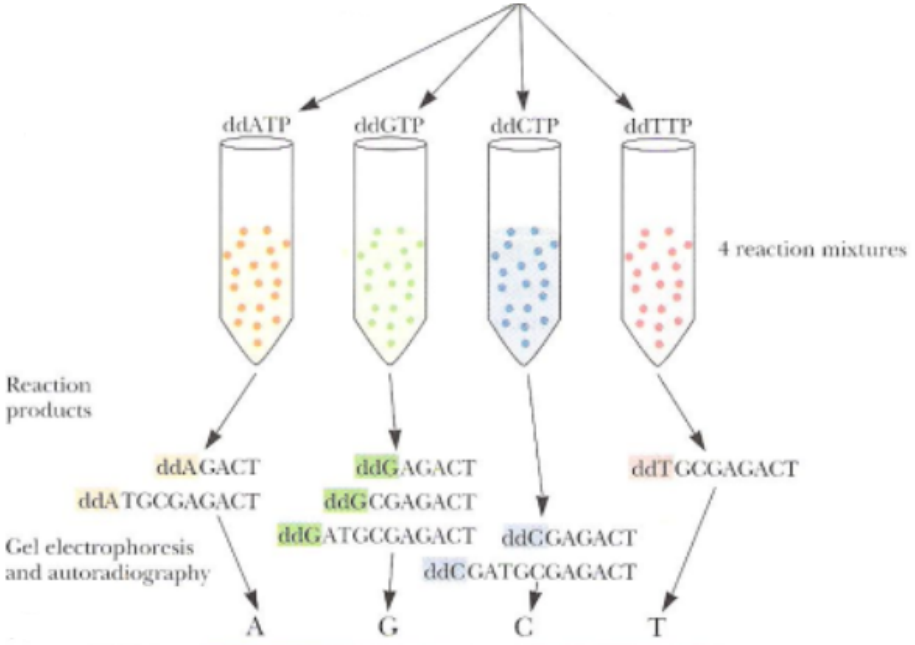


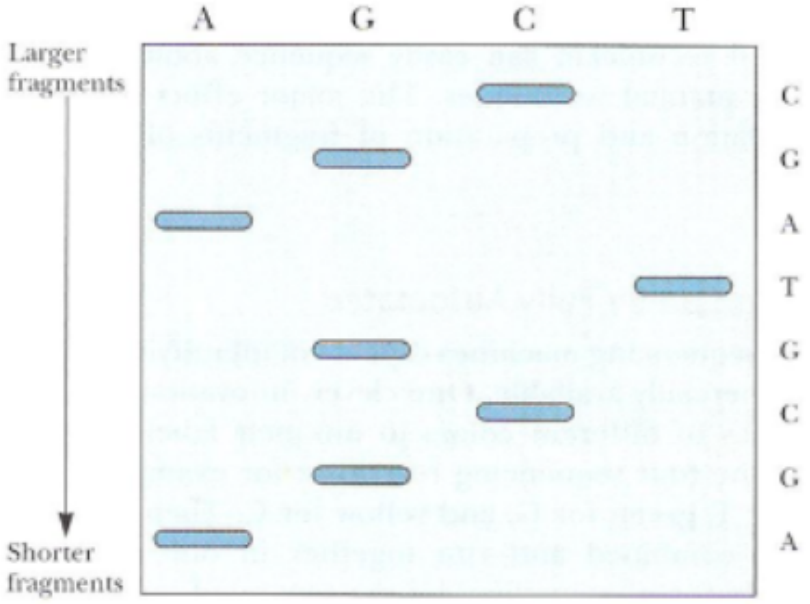
One reaction vessel

Remember each reaction has many molecules each one incorporating its respective ddNTP and stopping at a different length.

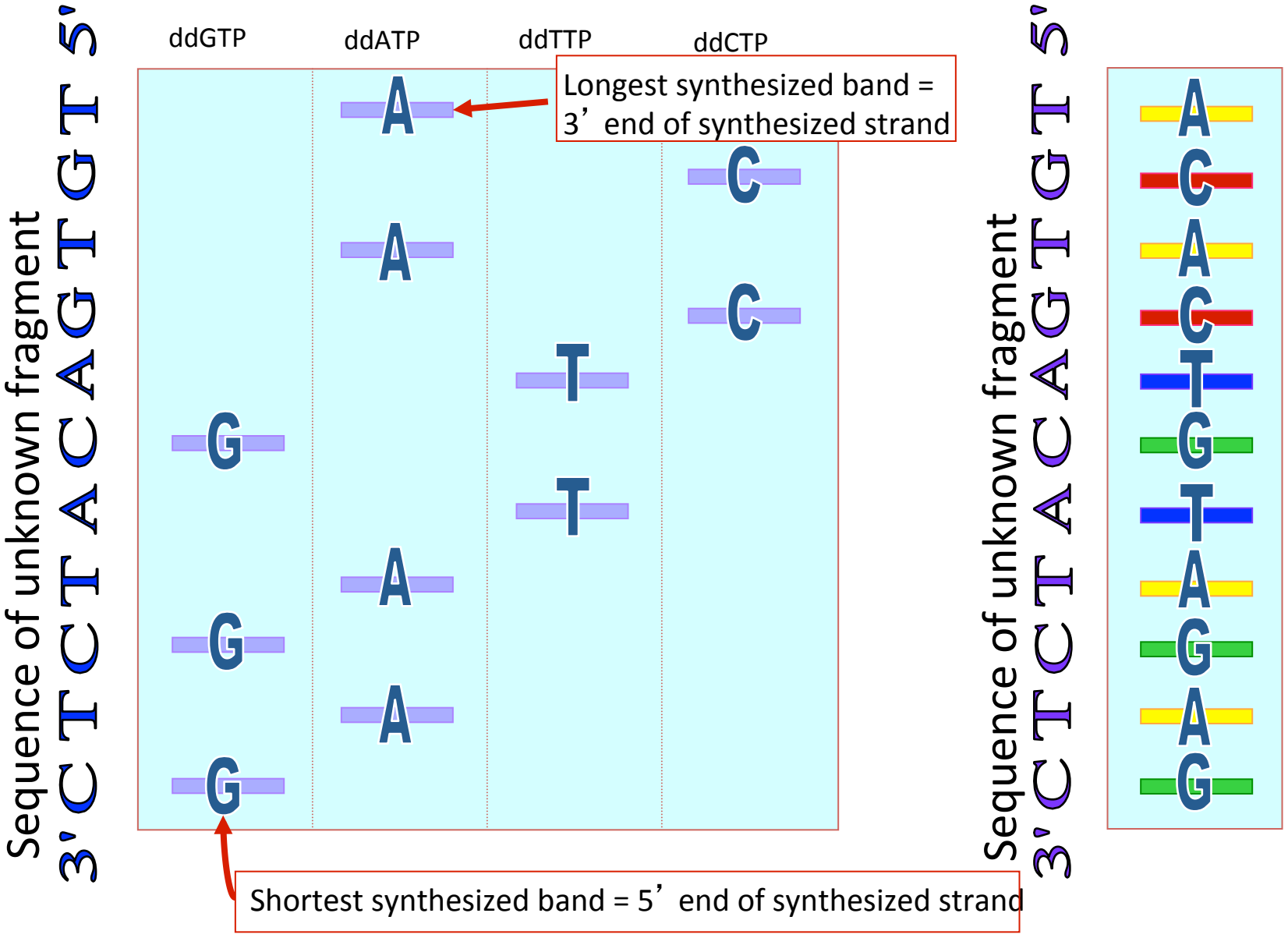
Sanger Method

- Run four separate reactions each with different ddNTPs.
- Run on a gel in four separate lanes.
- Read the gel from the bottom up.

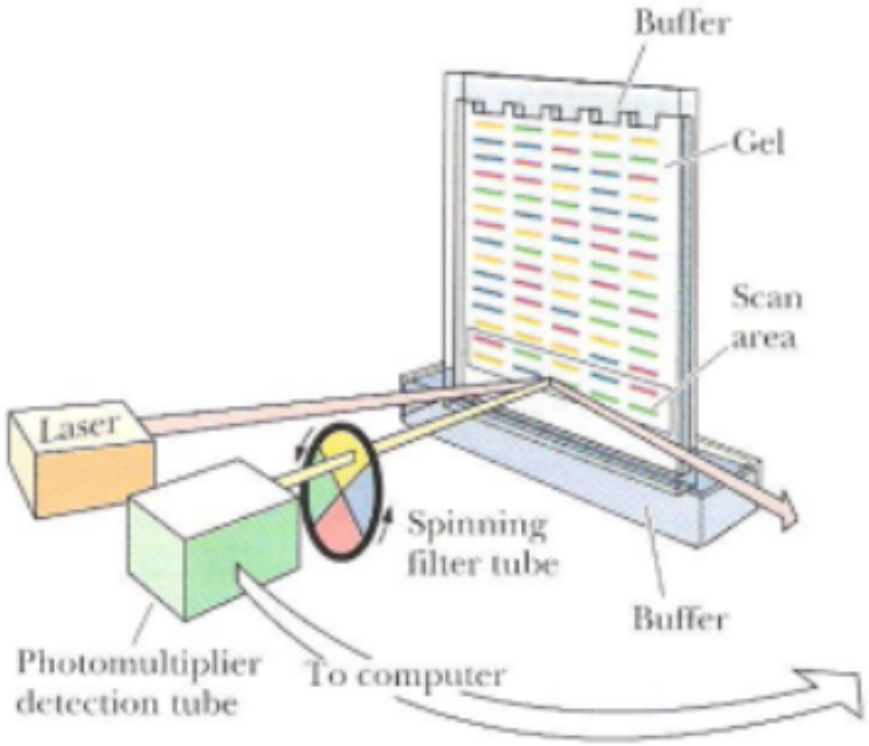


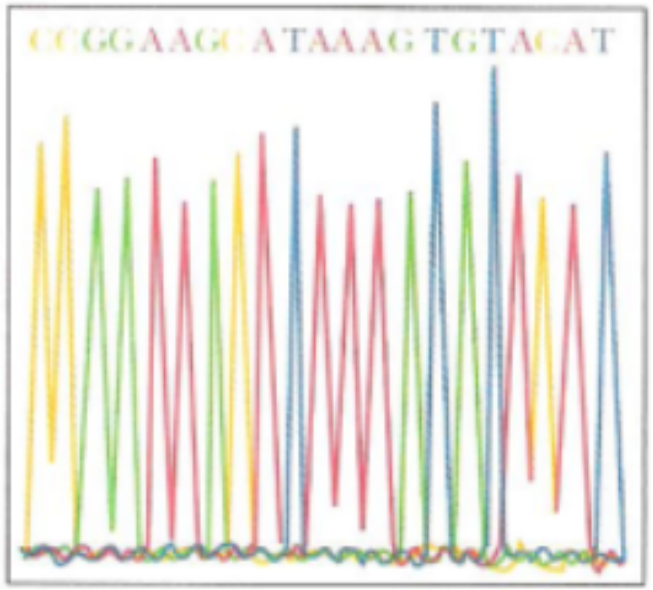


Gel Electrophoresis and Readout of Reaction Products:



Automated Version of the Sanger Method





Commercial Sanger Sequencer

- Uses capillary electrophoresis instead of gels. It can tolerate higher temperature and higher voltage, and is Higher throughput
- Performance:
  - 1920 samples/day
  - Average read length: 650 base pairs

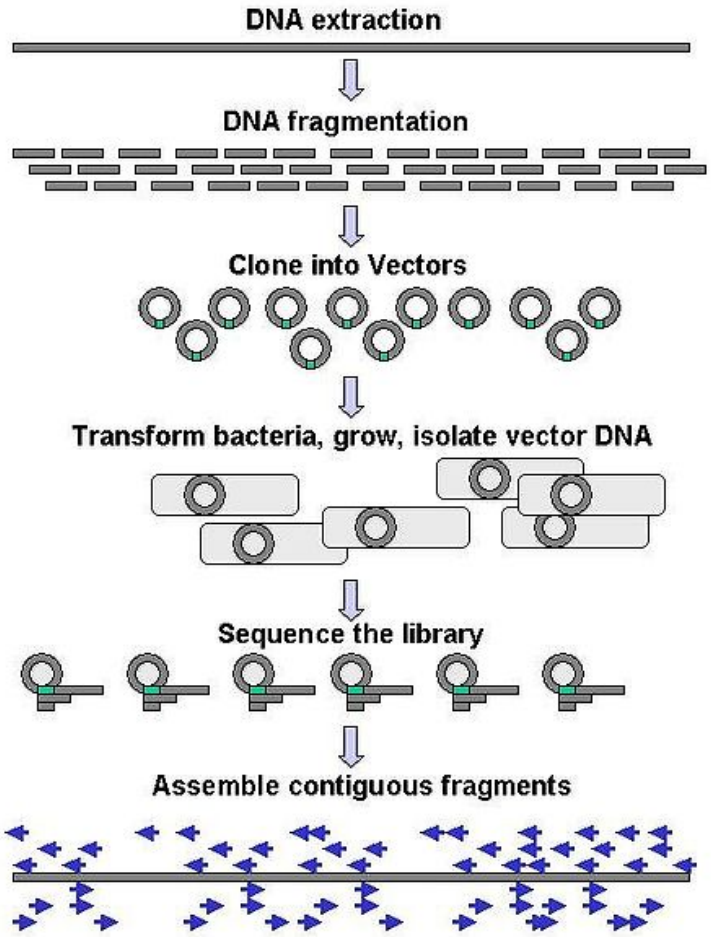
Applied Biosystems (ABI) 3730

So, What's Wrong With It?

- The Sanger method is good only for 500-750bp reactions.
- The biggest limitation to sequencing is that the genome is big
  - So carrying out these reactions for an entire genome is slow and expensive

Shotgun Sequencing

- Since the Sanger method can only be used for fairly short strands.
- Used to sequence whole genomes
- Steps:
  - DNA is broken up randomly into smaller fragments
  - Dideoxy method produces reads
  - Look for overlap of reads



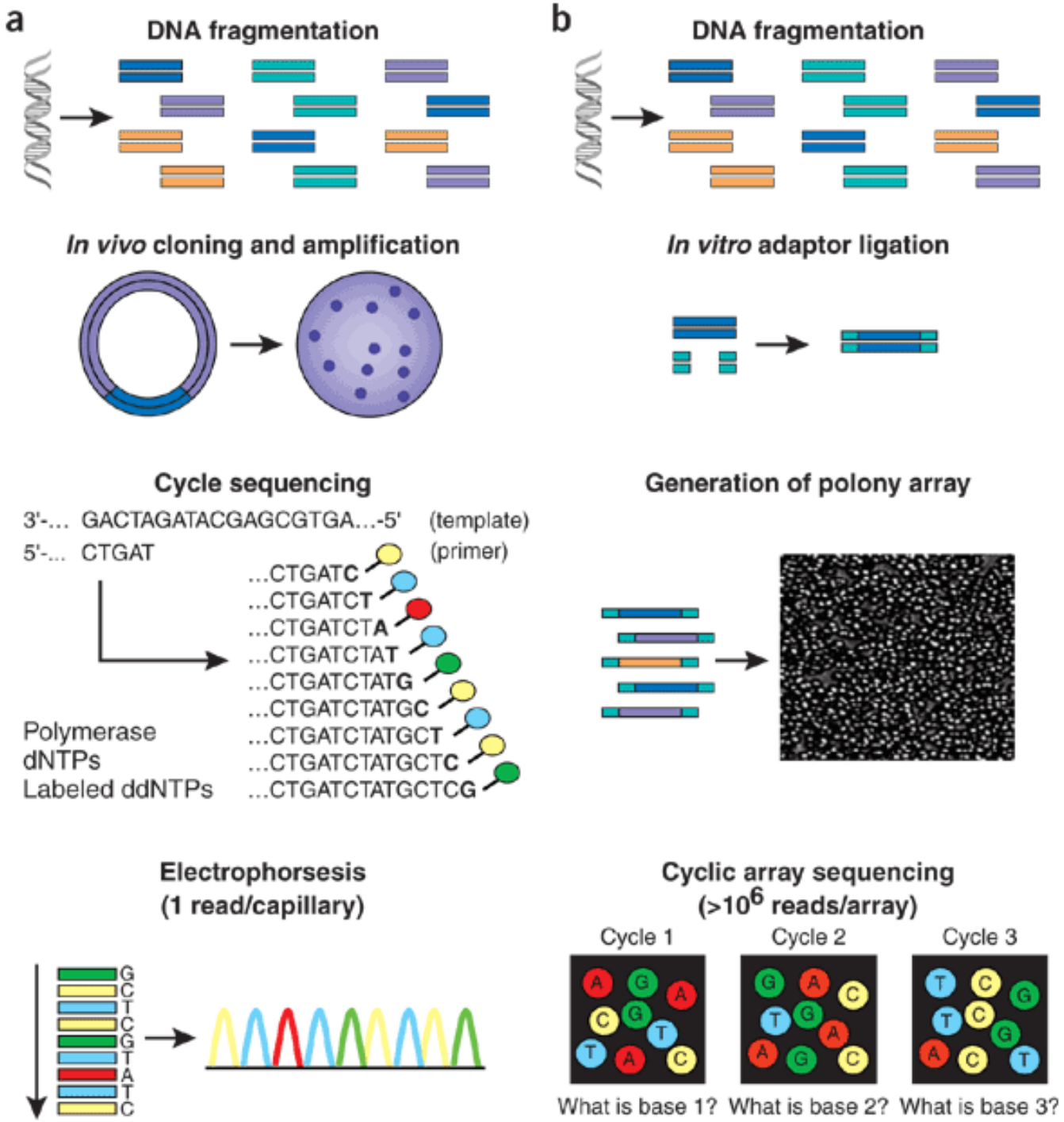
Strand	Sequence
First Shotgun Sequence	AGCATGCTGCAGTCATGCT------- --------------------TAGGCTA
Second Shotgun Sequence	AGCATG-------------------- ------CTGCAGTCATGCTTAGGCTA
Reconstruction	AGCATGCTGCAGTCATGCTTAGGCTA

Human Genome Project

- Why?
  - Human evolution
  - Nature versus nurture
  - Causes of disease
- Began in 1990. The Human Genome Project was declared complete in April 2003. An initial rough draft of the human genome was available in June 2000 and by February 2001.
- Cost = $3 billion
- The Human Genome Project fostered development of faster, less expensive sequencing techniques.

Next-Generation Sequencing

- Such as Sequencing by synthesis (SBS), pyrosequencing, sequencing by ligation
- Advantages:
  - Accurate
  - Parallel processing
  - Easily automated
  - Eliminates the need for labeled primers and nucleotides
  - No need for gel electrophoresis

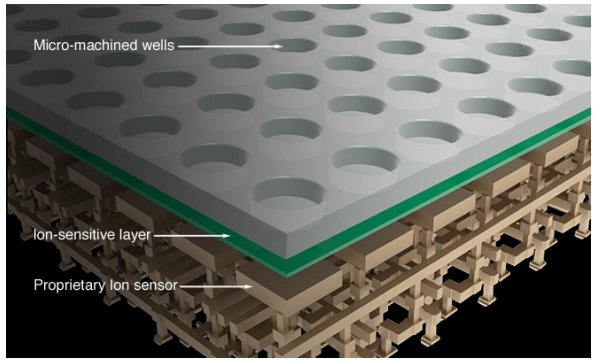


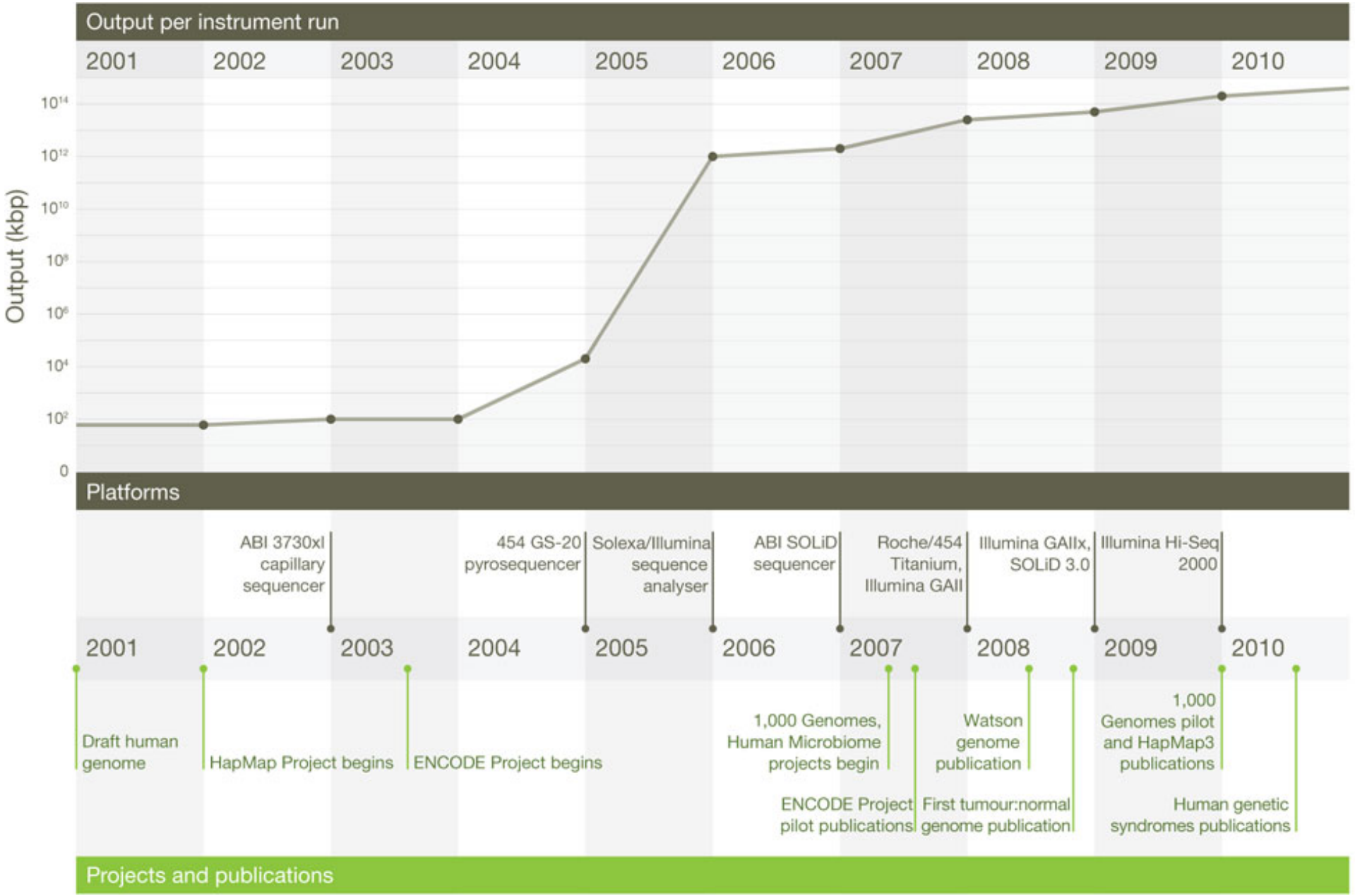
Work flow of conventional versus second-generation sequencing.

Shndure and Ji, 2008

Next Generation Sequencing

- Takes advantage of miniaturization to engage in massively parallel analysis
  - Essentially carrying out millions of sequencing reactions simultaneously in each of 10 million tiny wells

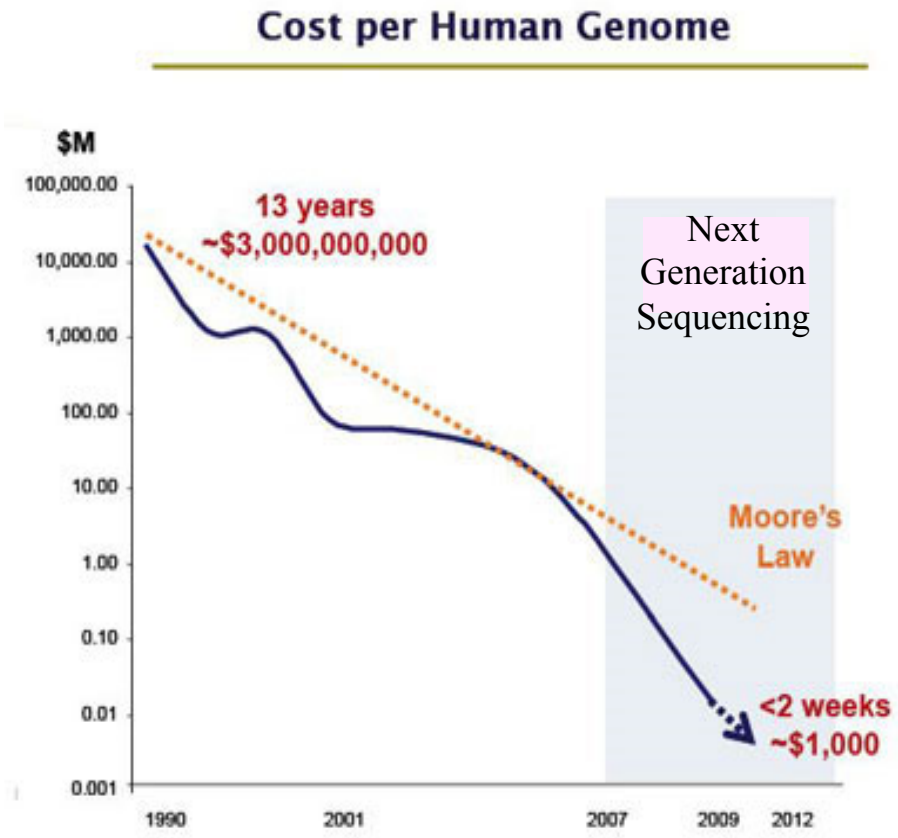




Changes in instrument capacity over the past decade, and the timing of major sequencing projects

Mardis, 2011

Accelerating Technology & Plummeting Cost



Commodore PET 2001 with 4KB memory is $795 in 1977 (=$2,800 in current $)

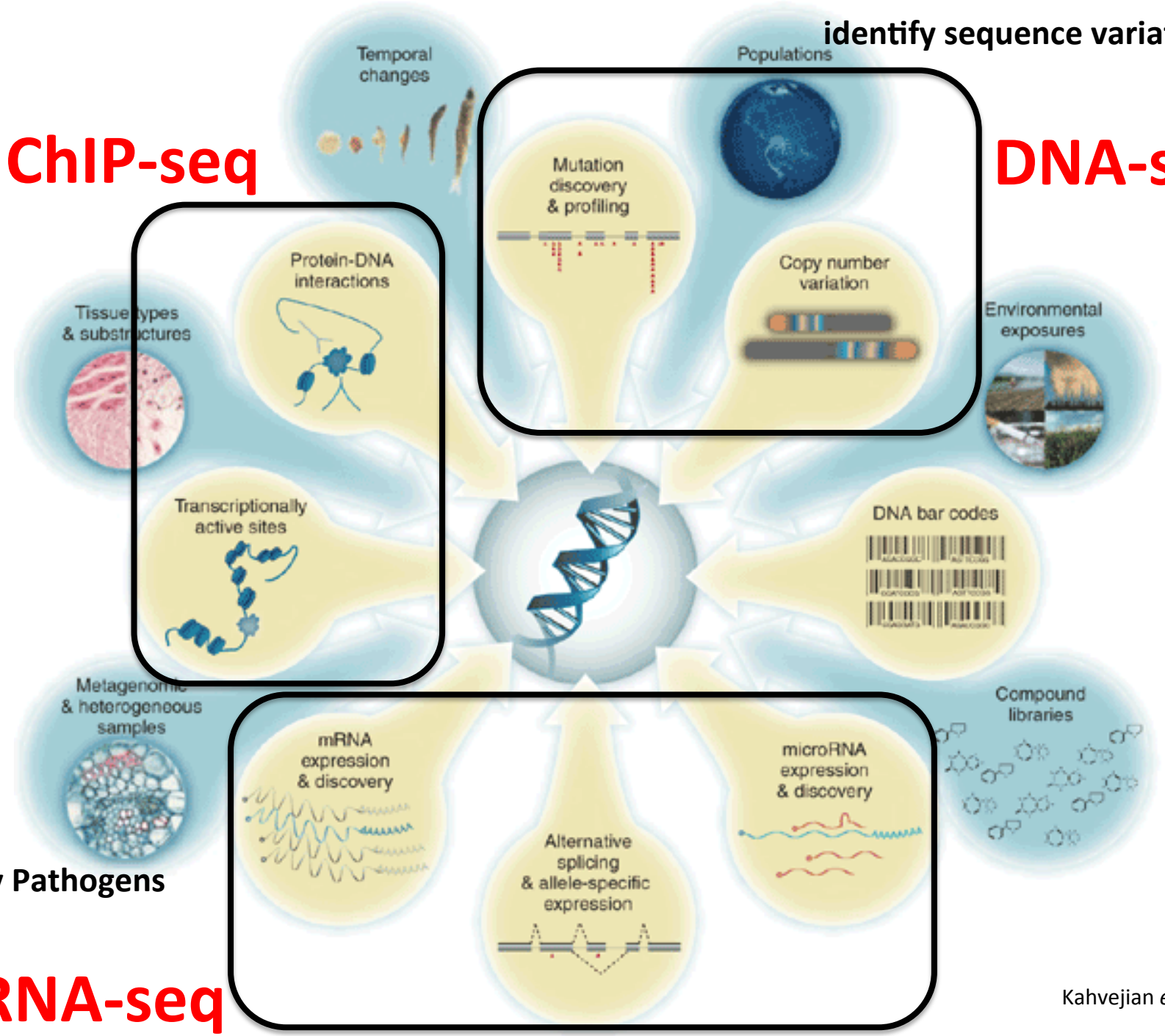
iPad has 16GB memory and is $499.

What can we do with NGS?

Category	Examples of applications
Complete genome resequencing	Comprehensive polymorphism and mutation discovery in individual human genomes
Reduced representation sequencing	Large-scale polymorphism discovery
Targeted genomic resequencing	Targeted polymorphism and mutation discovery
Paired end sequencing	Discovery of inherited and acquired structural variation
Metagenomic sequencing	Discovery of infectious and commensal flora
Transcriptome sequencing	Quantification of gene expression and alternative splicing; transcript annotation; discovery of transcribed SNPs or somatic mutations
Small RNA sequencing	microRNA profiling
Sequencing of bisulfite-treated DNA	Determining patterns of cytosine methylation in genomic DNA
Chromatin immunoprecipitation– sequencing (ChIP-Seq)	Genome-wide mapping of protein-DNA interactions
Nuclease fragmentation and sequencing	Nucleosome positioning
Molecular barcoding	Multiplex sequencing of samples from multipleindividuals

Shndure and Ji, 2008

identify sequence variations



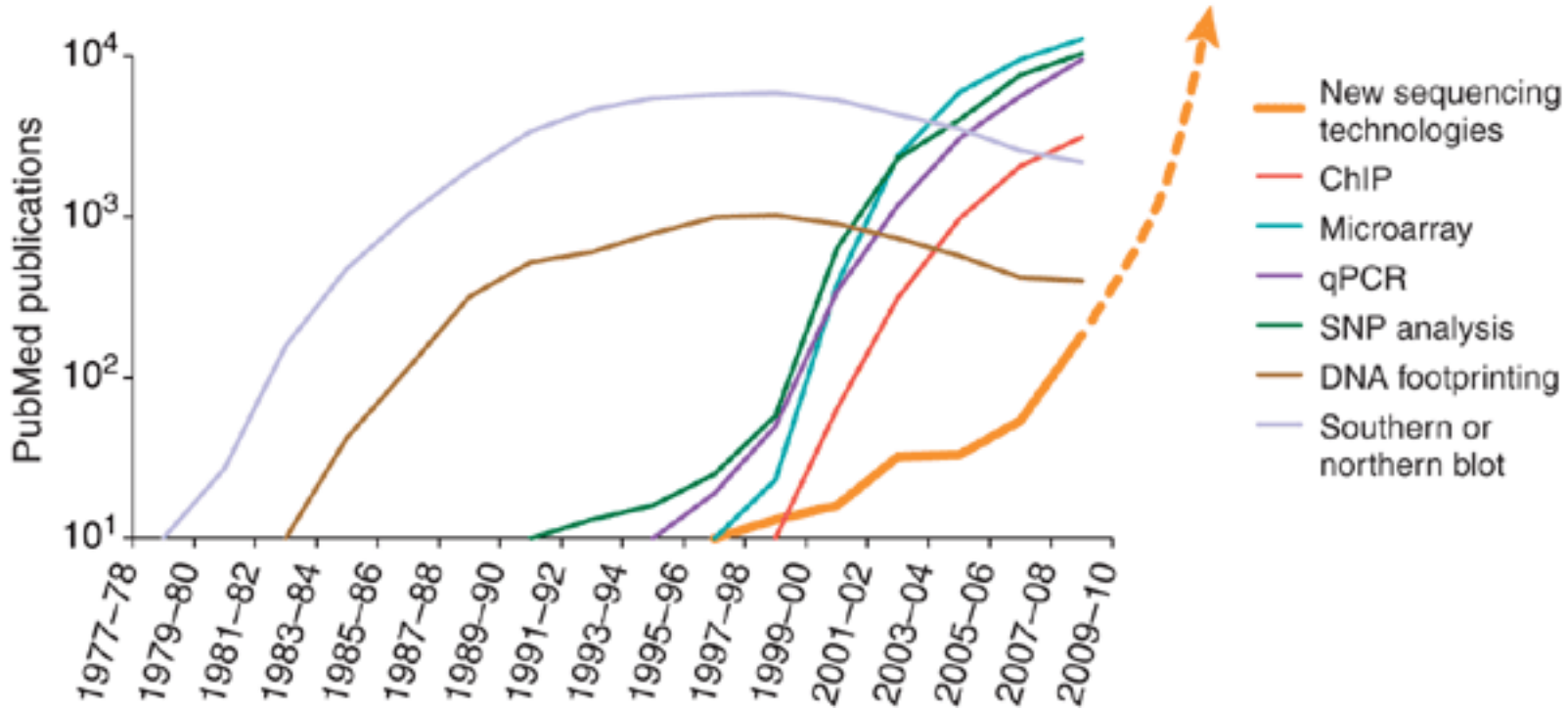
Identify Pathogens

Kahvejian *et al*, 2008

Achievements of NGS

- New genomes can be mapped easily and rapidly. Sophisticated computer analysis of huge amounts of information allows “assembly" of a given sequence.
- 1000 Cancer Genomes project: disease associations can be mapped in hope for a cure.
- Discover or quantitate rare sequence variants.
  - Cancer related mutants within a single patient
- Genomic diagnostics.
- Personalized medicine. The goal of the Archon X prize in Genomics is to enable a $1000 geneome. (currenlty at $5000-$50,000).
- Human microbiomes (an astronomic number of them), the bacterial flora of our bodies, can be sampled.

Number of Publications



The number of publications with keywords for nucleic acid detection and sequencing technologies.

Kahvejian et al. Nature Biotechnology 2008

Platforms

- Illumina
- SOLiD (Life Technology)
- 454 (Roche)
- Helicos
- Pacific Viosciences
- Ion Torrent (Life Technology)

Illumina

Run Mode	High Output	Rapid Run*
Output (2 × 100 bp)	600 Gb	120 Gb
Run Time (2 × 100 bp)	~11 days	~27 hours
Cluster Generation	cBot	On board
Paired-end Reads	6 Billion	1.2 Billion
Single Reads	3 Billion	600 Million
Maximum Read Length**	2 × 100 bp	2 × 150 bp

HiSeq 2500

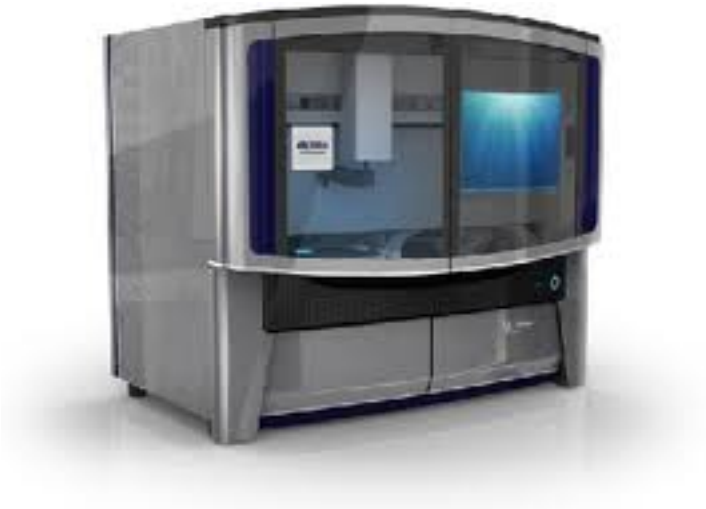
454

Sequencing Kit	GS FLX Titanium XL+
Read Length	Up to 1,000 bp
Mode Read Length	700 bp
Throughput Profile	- 85% of total bases from reads >500 bp - 45% of total bases from reads >700 bp
Typical Throughput	700 Mb
Reads per Run	~1,000,000 shotgun
Consensus Accuracy*	99.997%
Run Time	23 hours
Sample Input	gDNA or cDNA

Roche GS FLX Titanuim XL+

SOLiD

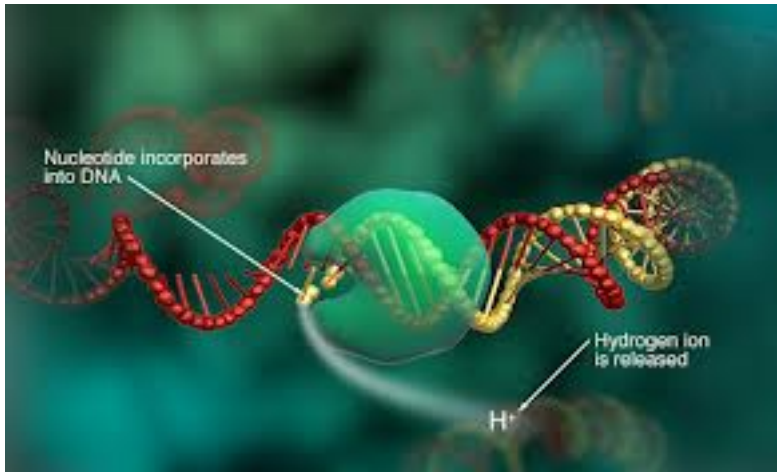
System and features	5500xl System (1.0 µm microbeads)
Independent lanes	1–12 (2 FlowChips)
Throughput	10–15 Gb/day
Human genome/run	Up to 2 genomes (30X average coverage)
Maximum read lengths	Mate-paired: 2 x 60 bp Paired-end: 75 bp x 35 bp Fragment: 75 bp

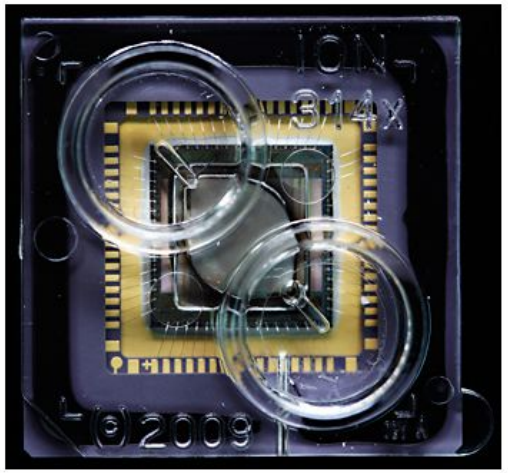


ABI 5500xl SOLiD System

Ion Torrent

- **First Ion semiconductor sequencing**
- **Ion Semiconductor Sequencing** is a method of DNA sequencing based on the detection of hydrogen ions that are released during the polymerization of DNA.





Some issues

- Certain platforms are better suited for certain tasks
  - Counting applications (ChIP-Seq, RNA-Seq) need more reads
  - De novo assembly work needs longer reads
  - Whole genome re-sequencing requires lower errors rate and high processivity (long reads).

References

- Bentley DR *et al*. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**: 53-59 (2008)
- Kahvejian A, Quackenbush J & Thompson JF. What would you do if you could sequence everything? *Nature Biotechnology* **26**: 1125-1133 (2008)
- Korbel JO *et al*. Paired-end mapping reveals extensive structural variation in the human genome. *Science* **318**: 420-426 (2007)
- Mardis ER. A decade's perspective on DNA sequencing technology. *Nature* **470**: 198-203 (2011)
- Metzker ML. Sequencing technologies - the next generation. *Nature Reviews Genetics* **11**:31-46 (2010)
- Shendure J & Ji H. Next-generation DNA sequencing. *Nature Biotechnology* **26**:1135-1145 (2008)
- Turner EH *et al*. Methods for genomic partitioning. *Annual Review of Genomics and Human Genetics* **10** (2009)