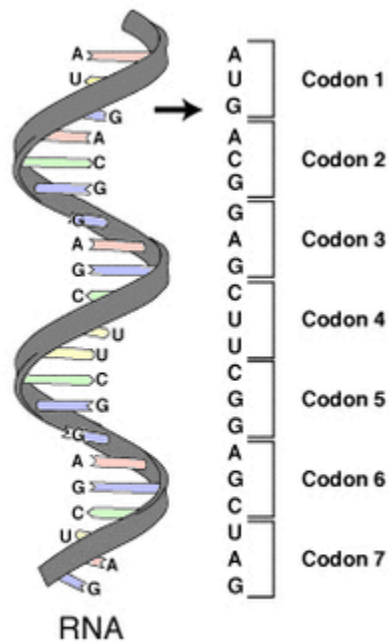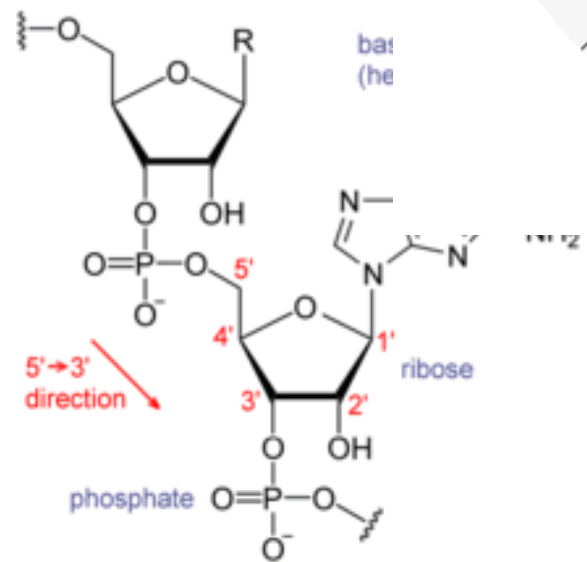# Sequencing and analysis of gene expression
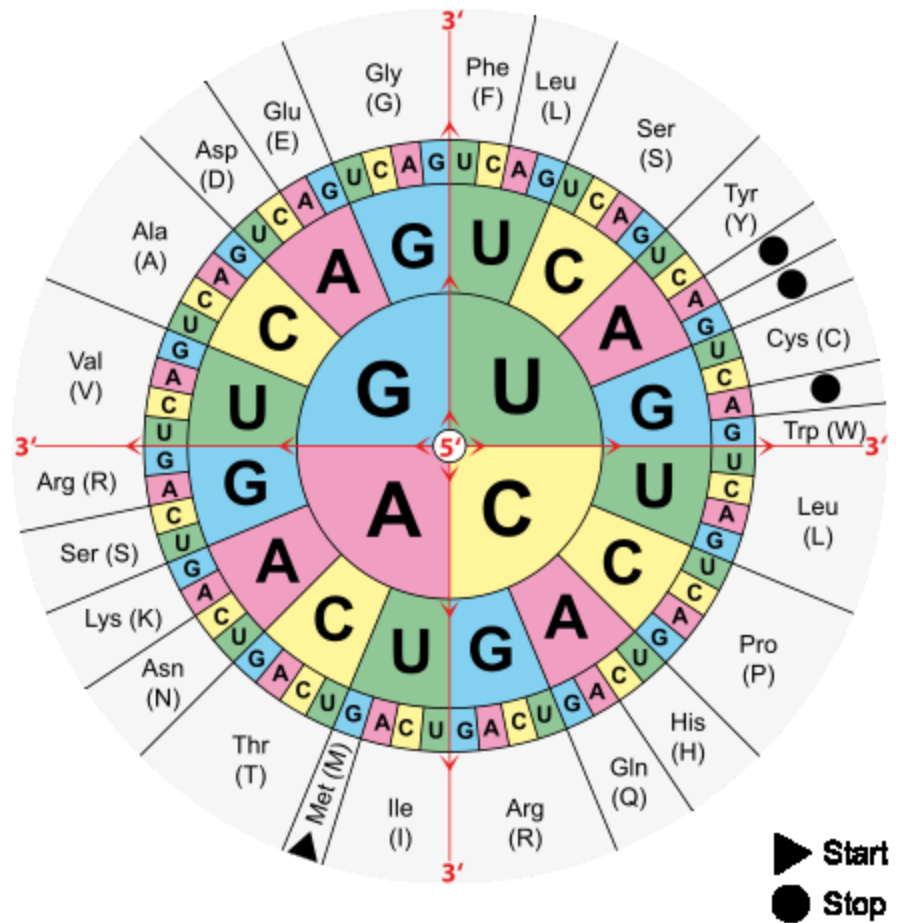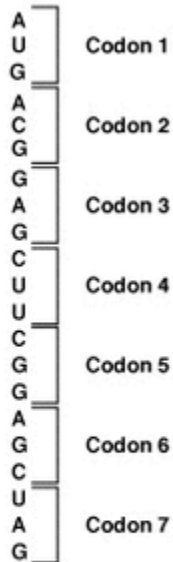
Jiří Fajkus

CG080 – Methods of Genomics and Proteomics

RNA

Ribonucleic acid

Codon 1
Codon 2
Codon 3
Codon 4
Codon 5
Codon 6
Codon 7

A
U
G
A
C
G
G
A
G
C
U
U
C
G
G
A
G
C
U
A
G

Gly (G)
Glu (E)
Asp (D)
Ala (A)
Val (V)
Arg (R)
Ser (S)
Lys (K)
Asn (N)
Thr (T)
Met (M)
Ile (I)
Arg (R)
Gln (Q)
His (H)
Pro (P)
Leu (L)
Trp (W)
Cys (C)
Tyr (Y)
Ser (S)
Leu (L)
Phe (F)

▶ Start
● Stop

5'→3' direction

phosphate

ribose

base (he

5'
4'
3'
2'
1'

OH
OH

A = adenine
C = cytosine
G = guanine
T = thymine
R = G A (purine)
Y = T C (pyrimidine)
K = G T (keto)
M = A C (amino)
S = G C (strong bonds)
W = A T (weak bonds)
B = G T C (all but A)
D = G A T (all but C)
H = A C T (all but G)
V = G C A (all but T)
N = A G C T (any)

http://www.wiley.com/college/pratt/0471393878/student/animations/dna_sequencing/index.html

Nobel Prize in Chemistry (1958) – amino acid sequence of insulin
Nobel Prize in Chemistry (1980) – dideoxy method of DNA sequencing
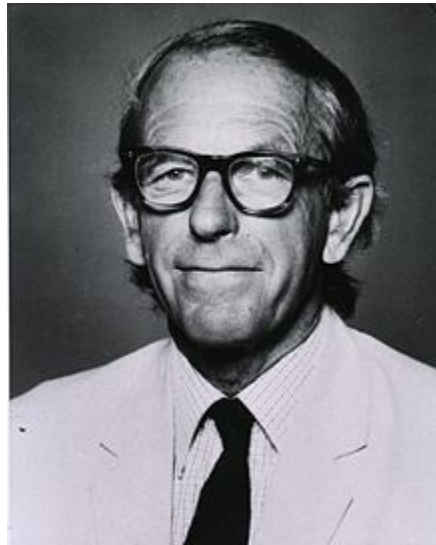
Frederick Sanger (13. 8. 1918 – 19. 11. 2013)

Sanger F, Nicklen S, Coulson AR (December 1977). "DNA sequencing with chain-terminating inhibitors". Proc. Natl. Acad. Sci. U.S.A. 74 (12): 5463–7.

Historical methods:

1975 - Sanger and Coulson – **plus-minus sequencing** (4 pairs of plus (1 dNTP) and minus (3 other dNTPs) reactions, restriction enzyme fragments as primers for elongation by DNA polymerase I.

– used for sequencing of 5,386 nucleotides of the single-stranded bacteriophage φX174

Chemical sequencing / Maxam-Gilbert Sequencing (Nobel Prize 1980)

1976–1977- Allan Maxam and Walter Gilbert (Maxam AM, Gilbert W (Feb 1977). "A new method for sequencing DNA". *Proc. Natl. Acad. Sci. U.S.A.* **74** (2): 560–4).

1) radioactive **labeling** at one 5' end of the DNA (typically by a kinase reaction using gamma-32P ATP) and purification of the DNA fragment to be sequenced.
2) **Chemical treatment** generates breaks at a small proportion of one or two of the four nucleotide bases in each of four reactions (G, A+G, C, C+T).
purines (**A+G**) are depurinated using formic acid,
guanines (**G**) (and to some extent the adenines) are methylated by dimethyl sulfate
pyrimidines (**C+T**) – bases are cleaved by hydrazine.
The addition of salt (NaCl) to the hydrazine reaction inhibits the methylation of thymine for the **C**-only reaction.
3) The modified DNAs are then **cleaved** by **hot piperidine** at the position of the modified base.
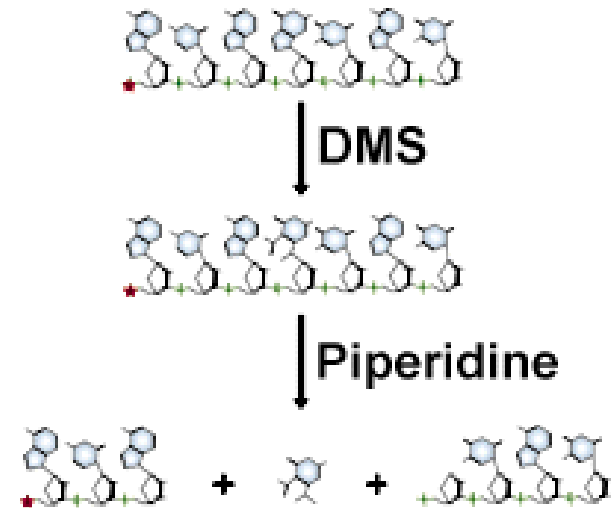Thus a series of labeled fragments is generated, from the radiolabeled end to the first "cut" site in each molecule.
4) The fragments in the four reactions are **electrophoresed** side by side in **denaturing PAGE for size separation**. To visualize the fragments, the gel is exposed to X-ray film for autoradiography, yielding a series of dark bands each corresponding to a radiolabeled DNA fragment, from which the sequence may be inferred.
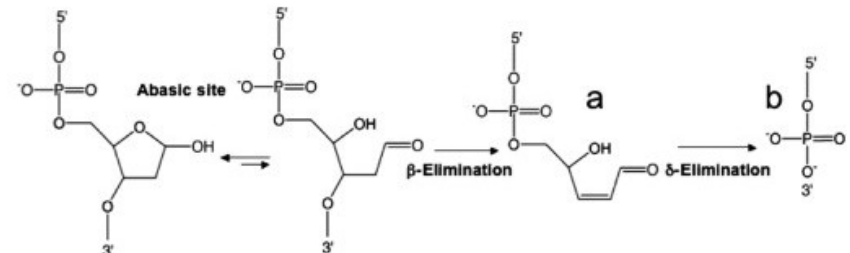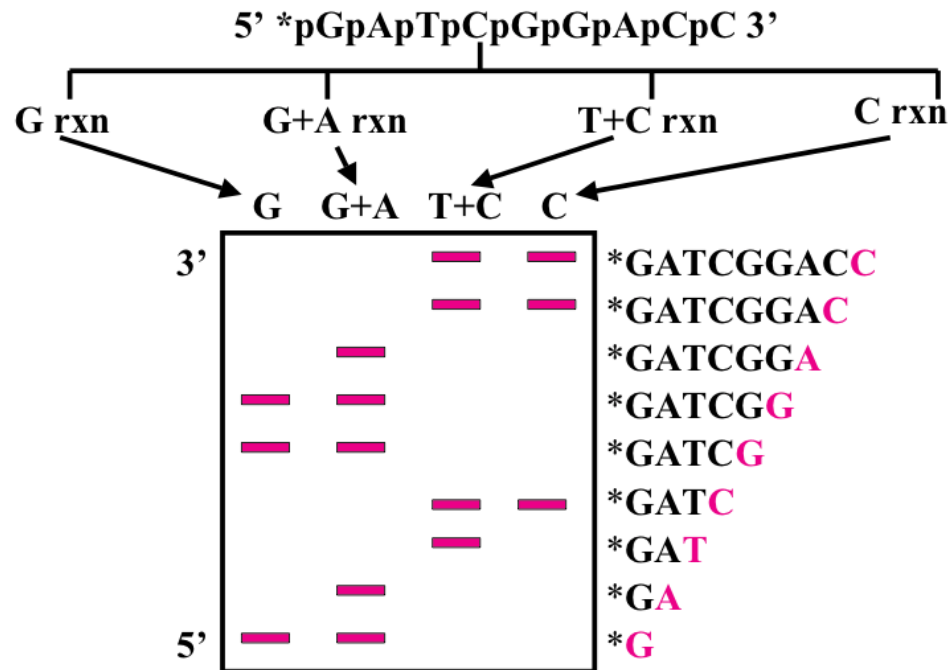
G-specific: DMS alkylation under alkaline conditions (pH 8)

(b)
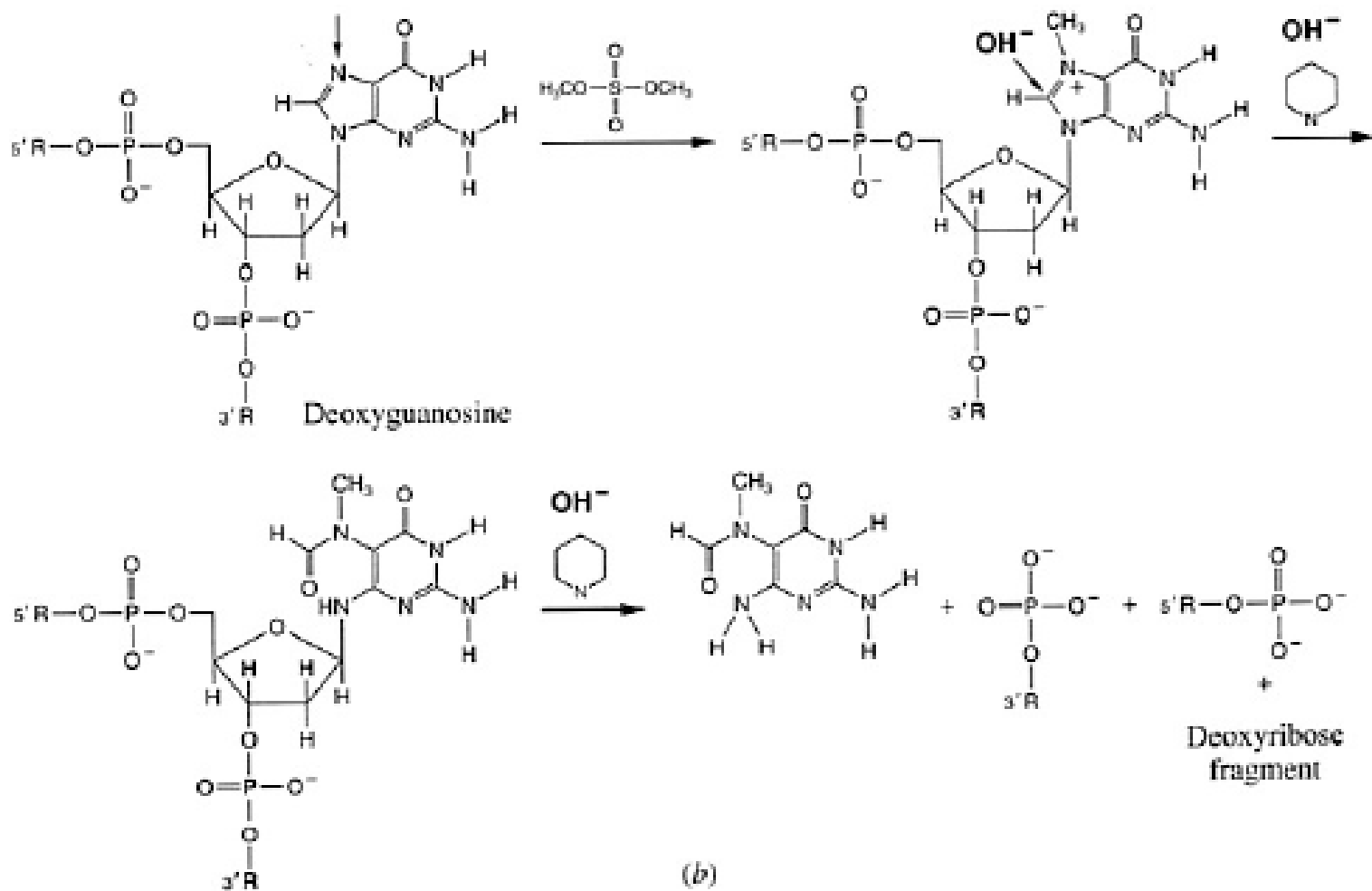
Reaction	Cleavage	Base Modification	Base Displacement	Strand Scission
R1	G>A	Dimethylsulfate	Heat at pH 7	NaOH
R2	A>G	Dimethylsulfate	acid	NaOH
* R3	C+T	Hydrazine	piperidine	piperidine
* R4	C	Hydrazine + Salt	piperidine	piperidine
* R5	G	Dimethylsulfate	piperidine	piperidine
* R6	G+A	Acid	acid	piperidine
R7	A>C	NaOH	piperidine	piperidine
R8	G	Methylene Blue	piperidine	piperidine
R9	T	Osmium Tetroxide	piperidine	piperidine

Maxam-Gilbert sequencing

5' *pGpApTpCpGpGpApCpC 3'

G rxn   G+A rxn   T+C rxn   C rxn

G   G+A   T+C   C

3'					
					*GATCGGACC
					*GATCGGAC
					*GATCGGA
					*GATCGG
					*GATCG
					*GATC
					*GAT
					*GA
5'					*G

Maxam-Gilbert (cont.)

- DNA is radiolabeled at one end (usually 5')

- Reaction conditions are selected such that on average each DNA strand in the reaction is cleaved once.

- Products are resolved on a DNA sequencing gel (PAGE)

- With 5' end-labeled DNA, the (5' to 3') sequence of the oligo is read from the bottom to the top of the gel.

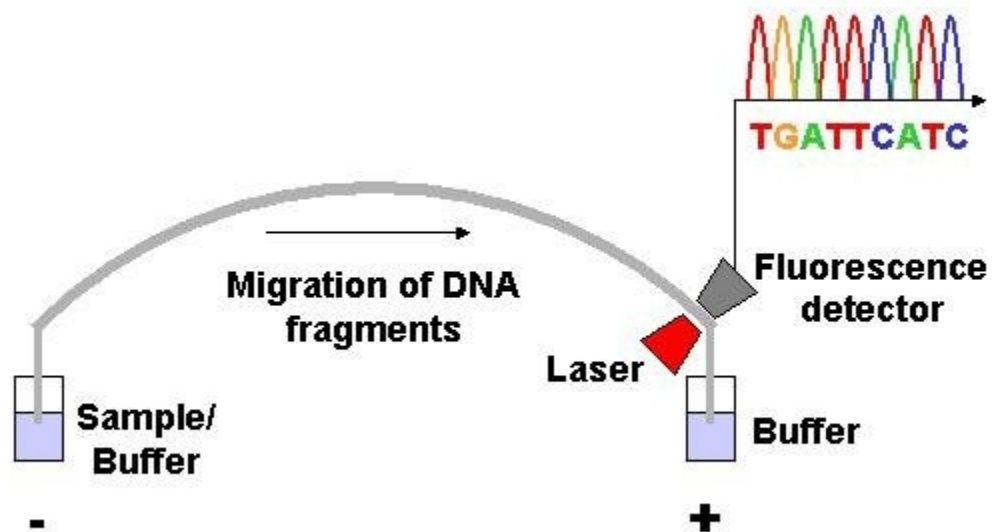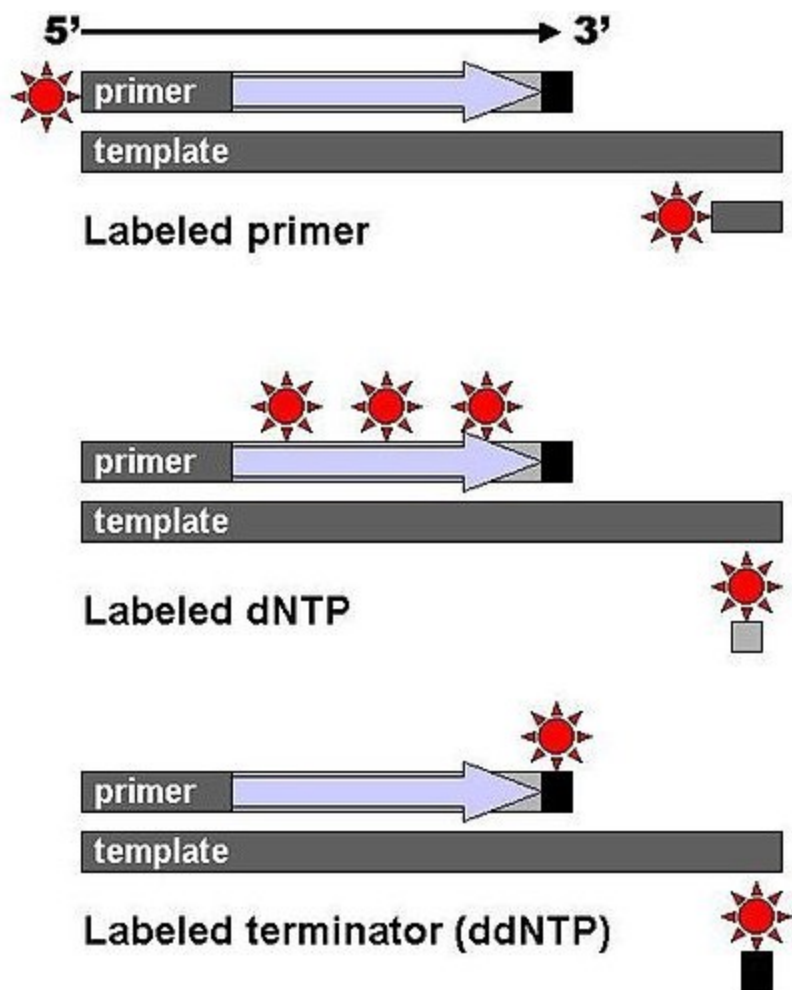**Chain-termination sequencing** (Sanger sequencing, dideoxy-sequencing)

- the use of dideoxynucleotide triphosphates (ddNTPs) as DNA chain terminators.

- Requirements: ss (heat-denatured) DNA template, a DNA primer, a DNA polymerase, normal deoxynucleotidetriphosphates (dNTPs), and modified nucleotides (dideoxyNTPs, ddNTPs) that terminate DNA strand elongation.

-The DNA sample is divided into 4 separate sequencing reactions, containing all four of the standard deoxynucleotides (dATP, dGTP, dCTP and dTTP) and the DNA polymerase. To each reaction is added only **one** of the four **ddNTPs** (ddATP, ddGTP, ddCTP, or ddTTP) which are the chain-terminating nucleotides, lacking a 3'-OH group required for the formation of a phosphodiester bond between two nucleotides, thus terminating DNA strand extension and resulting in DNA fragments of varying length.

- The newly synthesized and labelled DNA fragments are heat denatured, and separated by size (with a resolution of just one nucleotide) by gel electrophoresis on a denaturing polyacrylamide-urea gel with each of the four reactions run in one of four individual lanes (lanes A, T, G, C);

- the DNA bands are then visualized by autoradiography or UV light, and the DNA sequence can be directly read off the X-ray film or gel image.
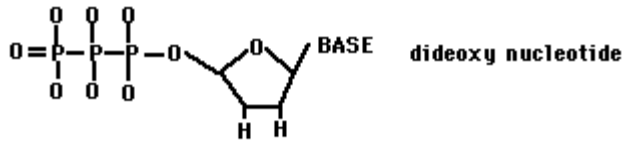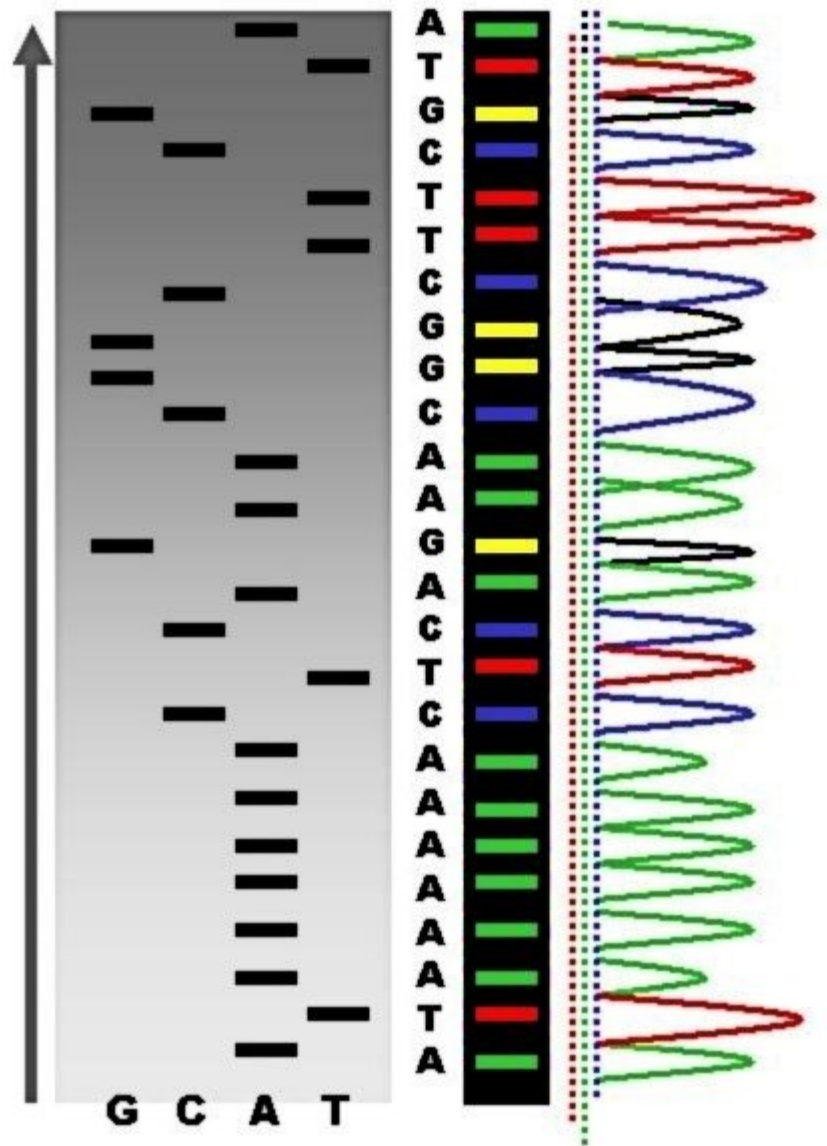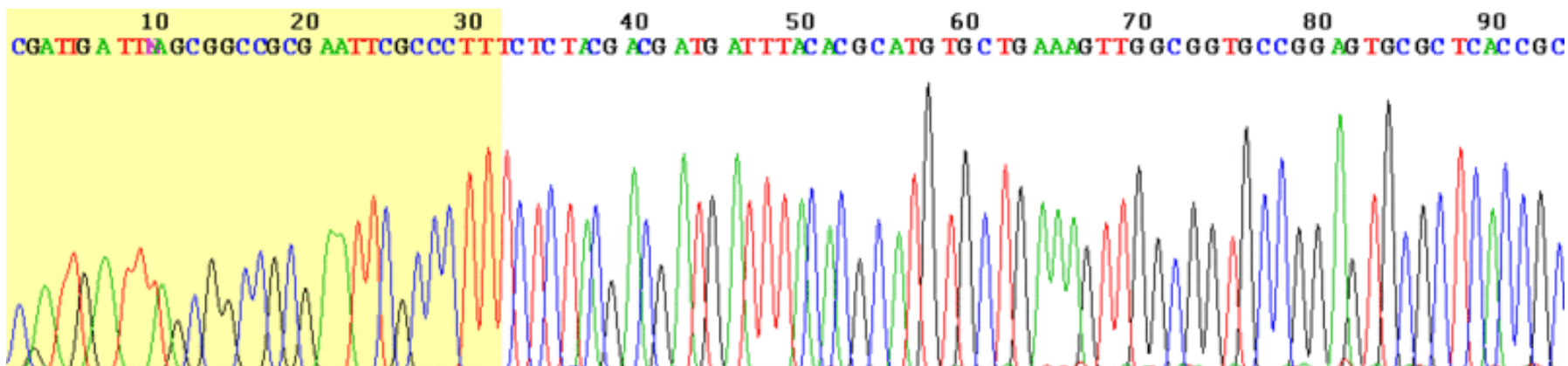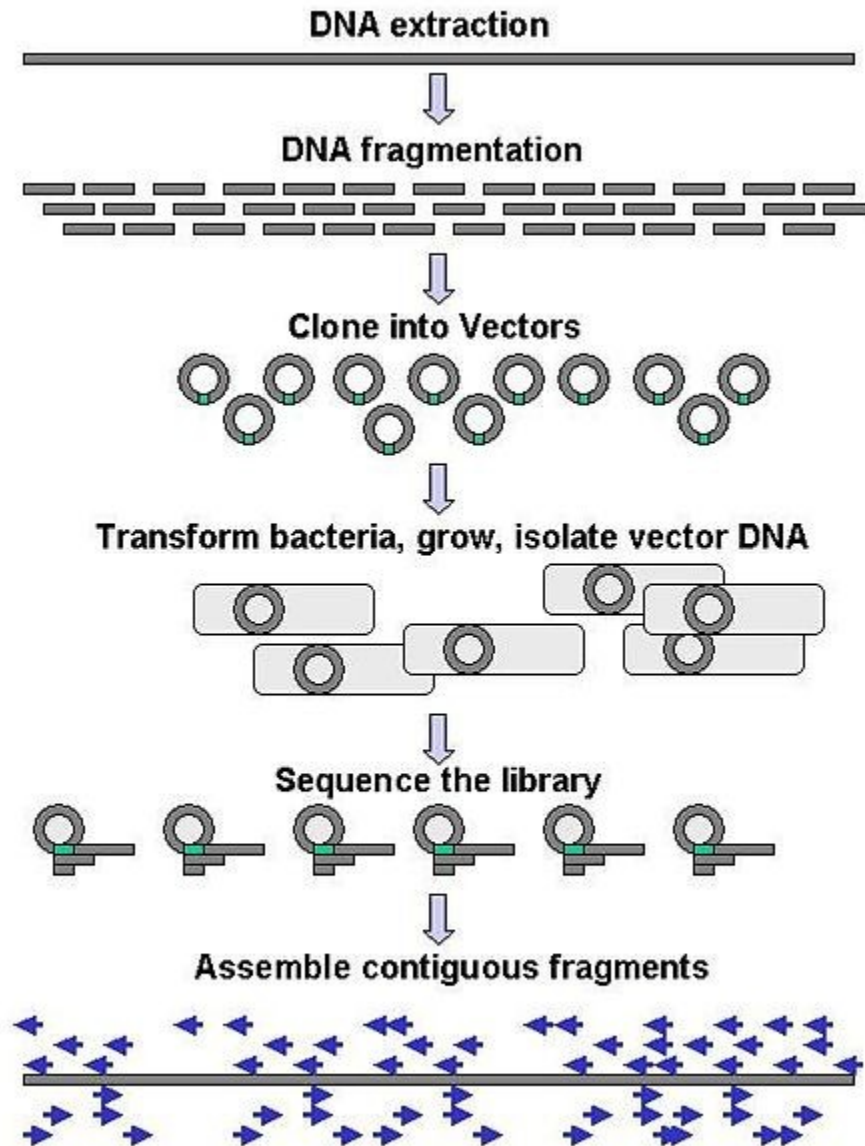
One of the four sequencing reactions
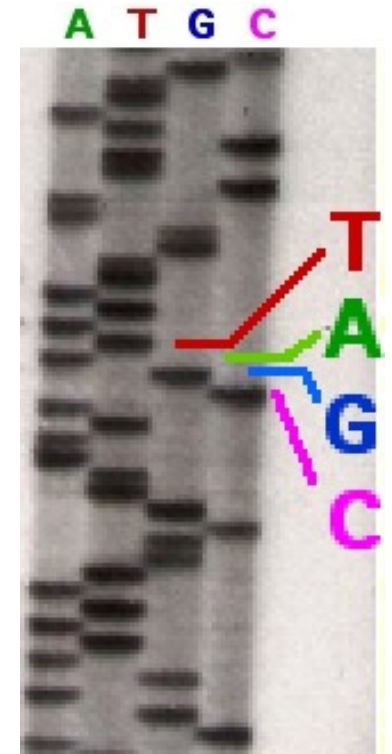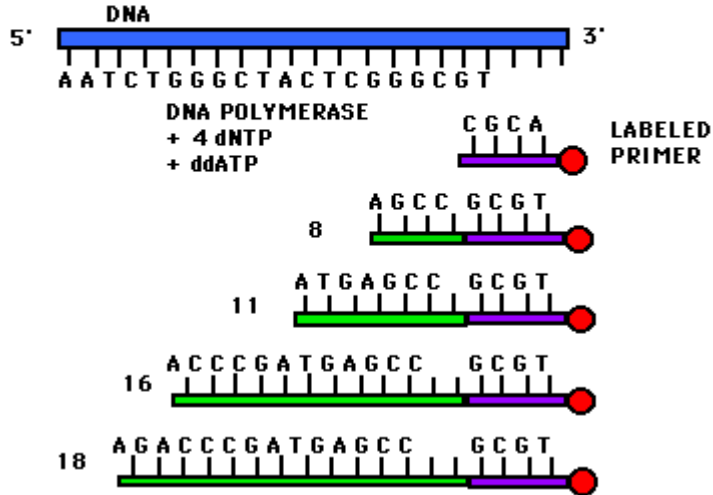
-In the image on the right, X-ray film was exposed to the gel, and the dark bands correspond to DNA fragments of different lengths. A dark band in a lane indicates a DNA fragment that is the result of chain termination after incorporation of a dideoxynucleotide (ddATP, ddGTP, ddCTP, or ddTTP). The relative positions of the different bands among the four lanes are then used to read (from bottom to top) the DNA sequence.

5' → 3'

Labeled primer

Labeled dNTP

Labeled terminator (ddNTP)

primer

template

Migration of DNA fragments

Sample/ Buffer

−

Laser

Buffer

+

Fluorescence detector

TGATTCATC

General sequencing strategy

DNA extraction

DNA fragmentation

Clone into Vectors

Transform bacteria, grow, isolate vector DNA

Sequence the library

Assemble contiguous fragments

Next generation
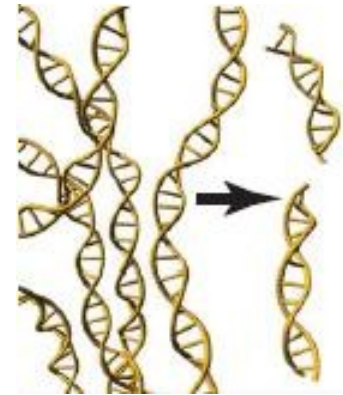# sequencing

Pyrosequencing *(Ronaghi and Nyrén, 1996)*

A parallelized version of pyrosequencing was developed by 454 Life Sciences, (*now part of Roche Diagnostics*) – **454 Pyrosequencing**.

Amplification of DNA inside water droplets in an oil solution (**emulsion PCR**), with each droplet containing a single DNA template attached to a single primer-coated bead that then forms a clonal colony. The sequencing machine contains many **picolitre-volume wells each containing a single bead** and sequencing enzymes. Pyrosequencing uses **luciferase** to generate light for detection of the individual nucleotides added to the nascent DNA, and the combined data are used to generate sequence read-outs. This technology provides **intermediate read length** and price per base compared to Sanger sequencing on one end and Solexa and SOLiD on the other.
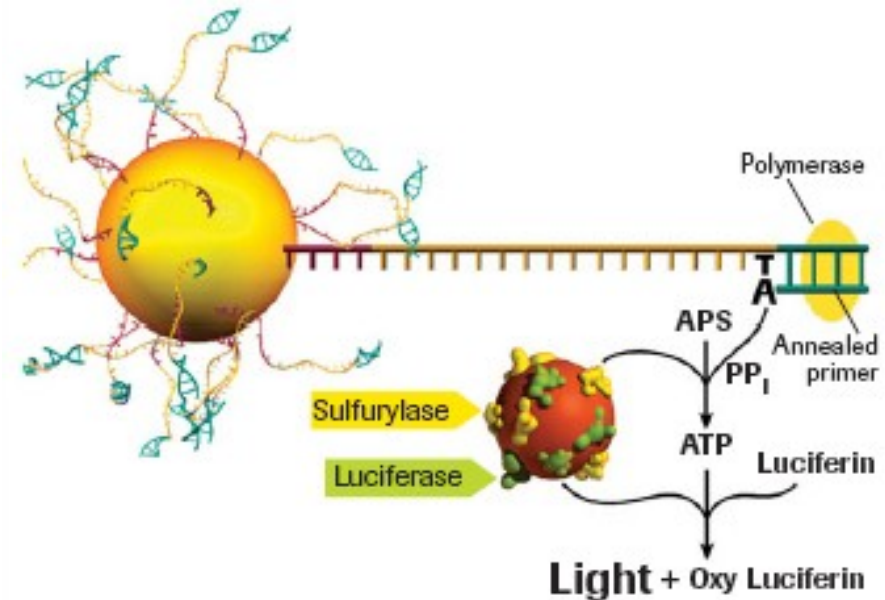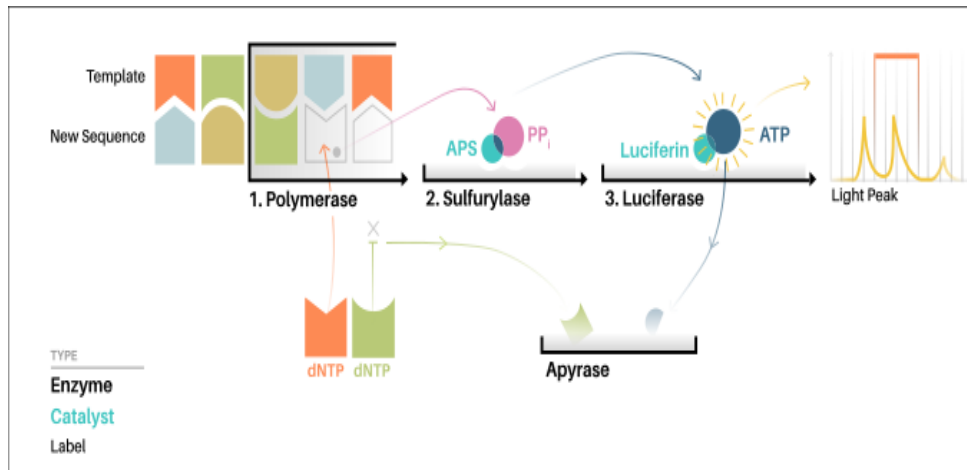
The Roche 454/GS FLX Sequencing Technology

The GS FLX sequencer supports sequencing of various different nucleic acid starting materials such as genomic DNA, PCR products, BACs and cDNA. Samples consisting of longer sequences are first **sheared into a random library of 300-800 base-pair long fragments.**
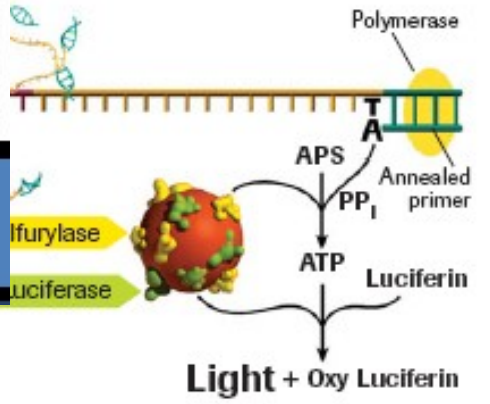

Adaptors essential for purification, amplification and sequencing are added to both ends of the fragments. If the sample is double stranded one strand is removed and the remaining single strandes are used in the following steps.
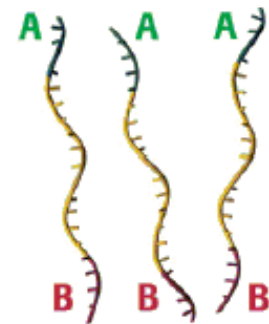
Aided by the adaptors **individual fragments are captured on their own unique beads**. A bead and the bound fragment together with a water-in-oil emulsion form a microreactor so that each fragment can be amplified without contamination via the so called **emulsion PCR** (emPCR). The entire fragment collection is amplified in parallel.
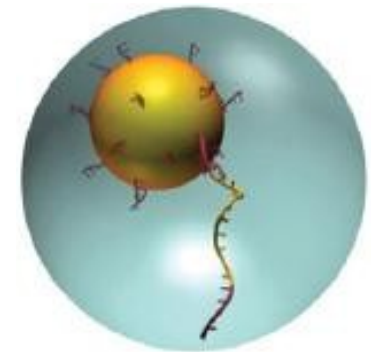
The emPCR amplifies each fragment several million times. After amplification the emulsion shell is broken and the clonally amplified beads are ready for loading onto the fibre-optic PicoTiterDevice for sequencing.

Load beads into PicoTiter™Plate

Load Enzyme Beads

Centrifuge Step

44 μm
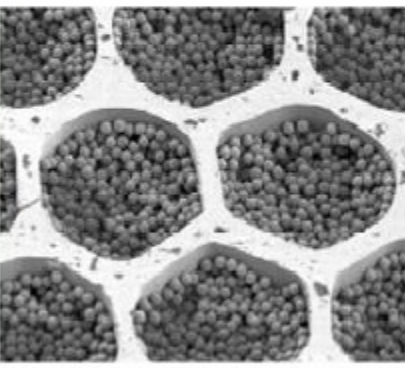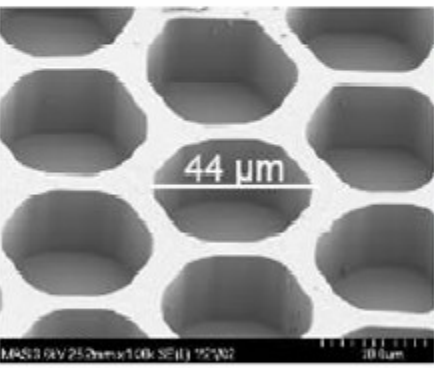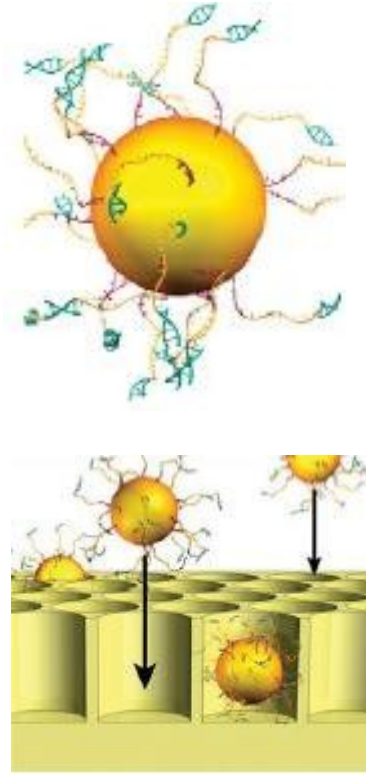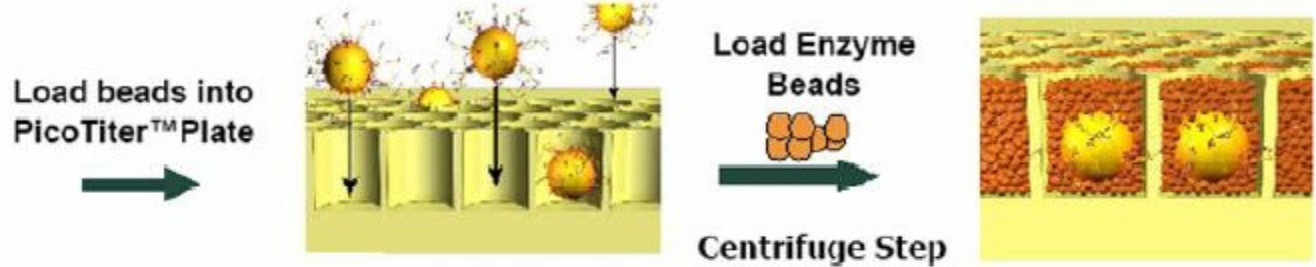
150 micron ~ Tip of a human hair

Each plate has 1.6 Million wells

Polymerase

APS

Annealed primer

PP$_I$

ATP

Luciferin

lfurylase

uciferase

Light + Oxy Luciferin

cascade. This results in the **generation of a light signal** which is captured by a CCD camera.

ATP sulfurylase converts PPi to ATP in the presence of adenosine 5' phosphosulfate (APS). This ATP drives the luciferase-mediated **conversion of luciferin to oxyluciferin** that **generates visible light** in amounts that are proportional to the amount of ATP. The light produced in the luciferase-catalyzed reaction is detected by a charge coupled device (CCD) chip and seen as a peak in the raw data output (Pyrogram). The height of each peak (light signal) is proportional to the number of nucleotides incorporated.

Nucleotide incorporation generates light seen as a peak in the Pyrogram trace
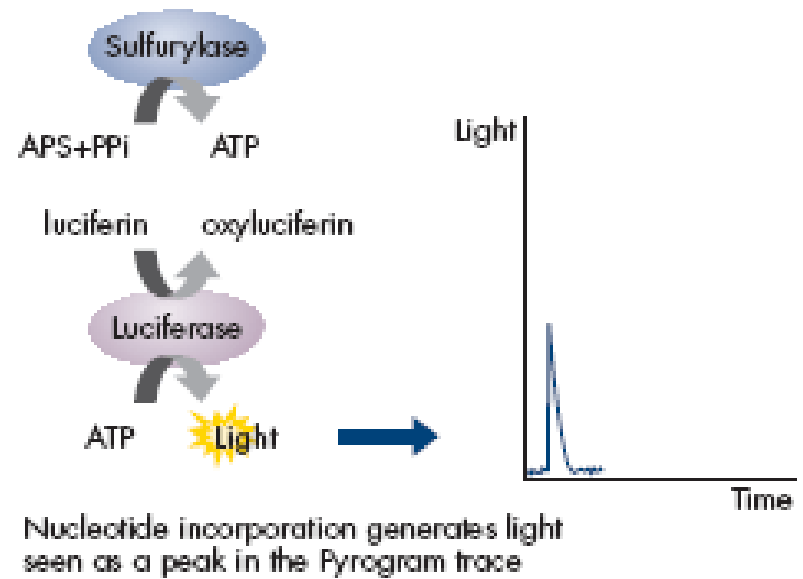
Addition of dNTPs is performed sequentially. It should be noted that deoxyadenosine **alfa-thio triphosphate (dATP·S) is used as a substitute for the natural dATP** since it is efficiently used by the DNA polymerase, but **not recognized by the luciferase**. As the process continues, the complementary DNA strand is built up and the nucleotide sequence is determined from the signal peaks in the Pyrogram trace.
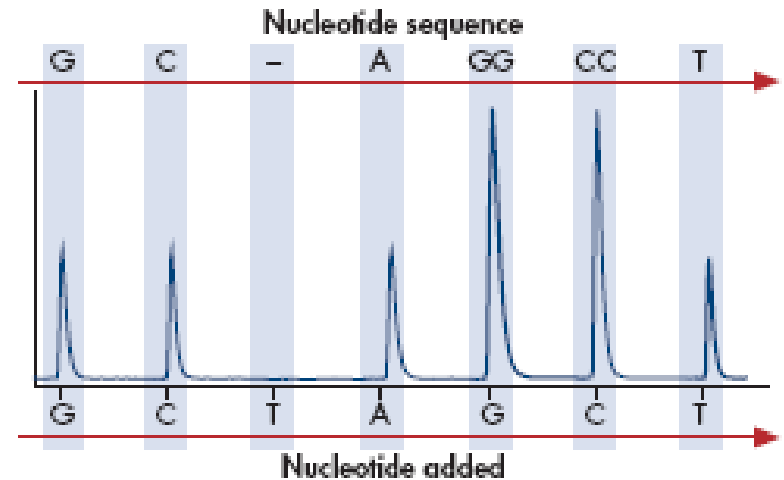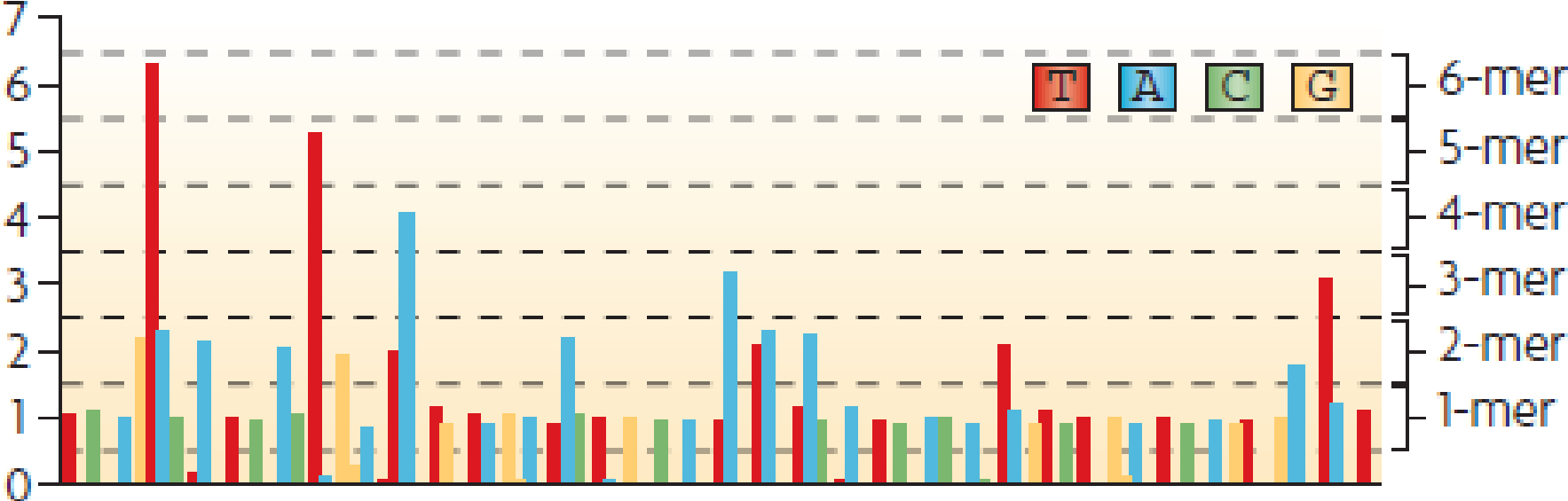
Ilumina (Solexa) sequencing

Solexa, now part of Illumina, developed a sequencing technology based on **reversible dye-terminators**.
DNA molecules are first attached to primers on a slide and amplified so that local clonal colonies are formed (isothermal bridge amplification). Four types of reversible terminator bases (RT-bases) are added, and non-incorporated nucleotides are washed away. Unlike pyrosequencing, the **DNA can only be extended one nucleotide at a time**. A camera takes images of the fluorescently labelled nucleotides, then the dye along with the terminal 3' blocker is chemically removed from the DNA, allowing the next cycle.

Ilumina

1. PREPARE GENOMIC DNA SAMPLE
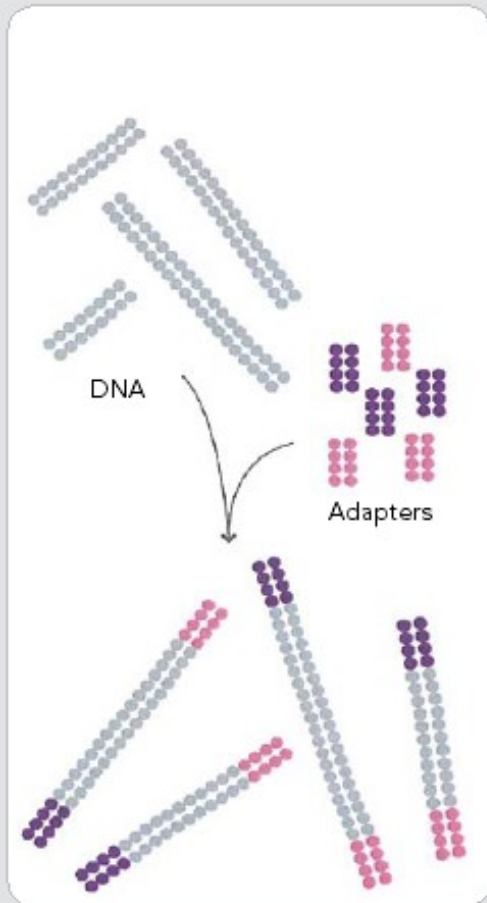
Randomly fragment genomic DNA and ligate adapters to both ends of the fragments.

2. ATTACH DNA TO SURFACE
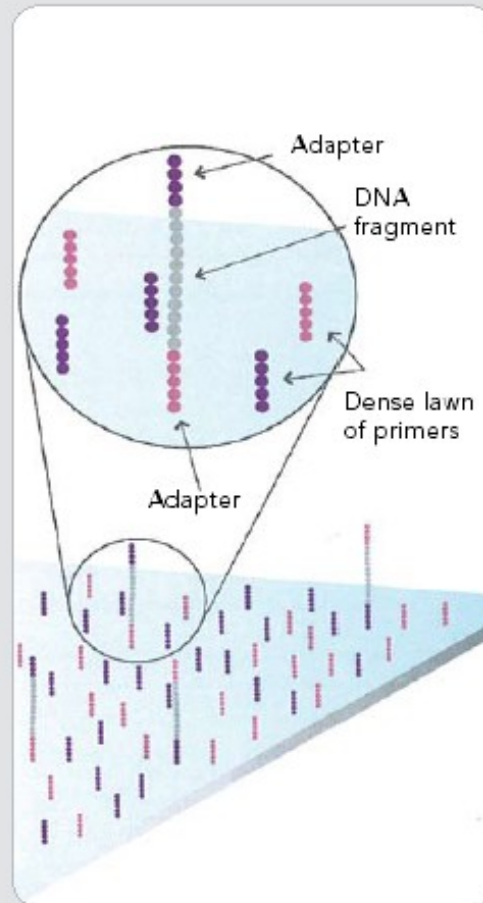
Bind single-stranded fragments randomly to the inside surface of the flow cell channels.

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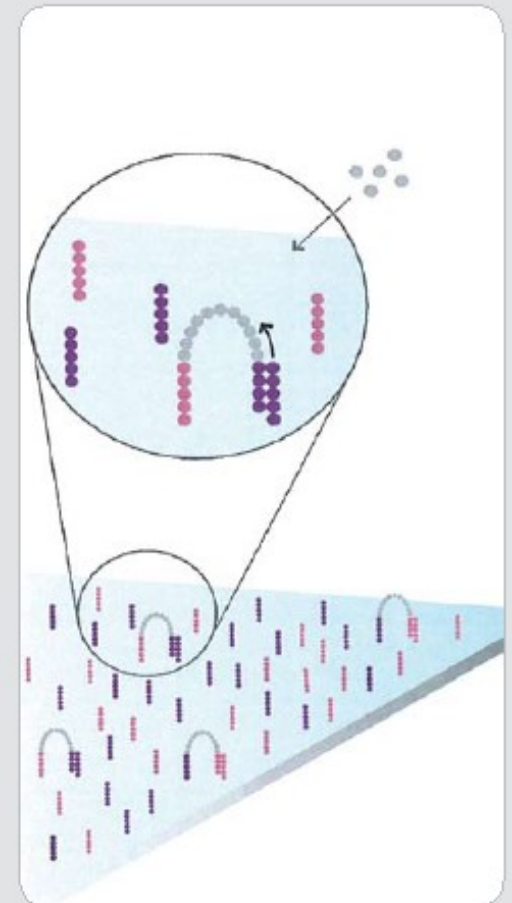
Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification.

Ilumina

4. FRAGMENTS BECOME DOUBLE-STRANDED

Attached terminus

Free terminus

Attached terminus

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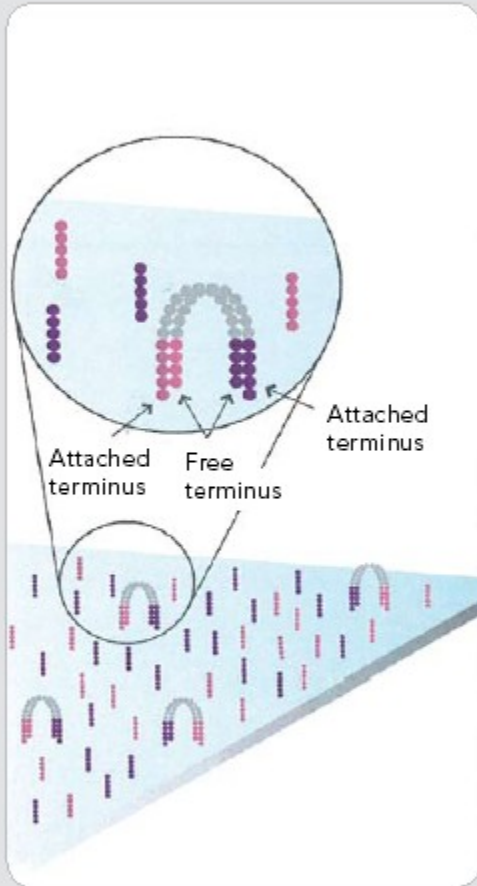
The enzyme incorporates nucleotides to build double-stranded bridges on the solid-phase substrate.

5. DENATURE THE DOUBLE-STRANDED MOLECULES

Attached

Attached
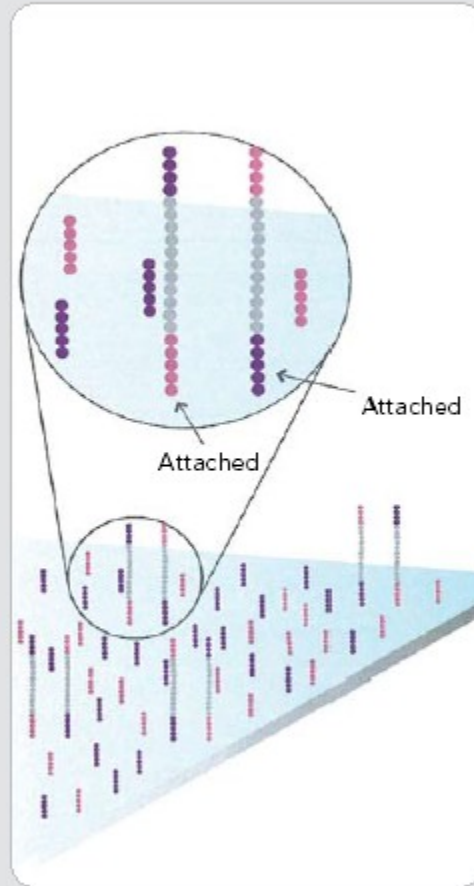
Denaturation leaves single-stranded templates anchored to the substrate.

6. COMPLETE AMPLIFICATION
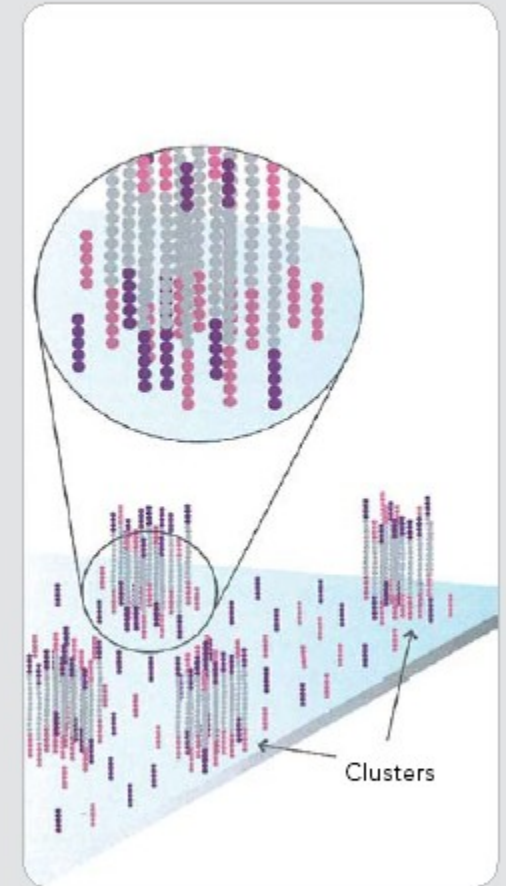
Clusters

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

Ilumina

7. DETERMINE FIRST BASE
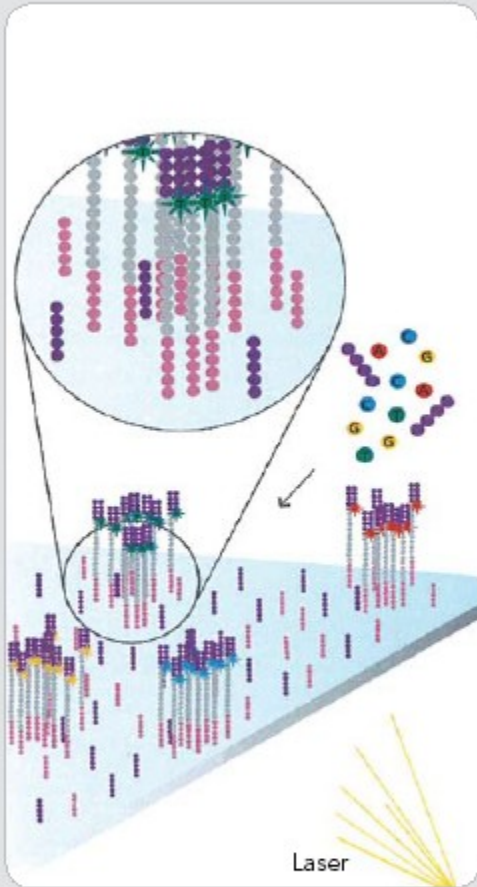
The first sequencing cycle begins by adding four labeled reversible terminators, primers, and DNA polymerase.

8. IMAGE FIRST BASE
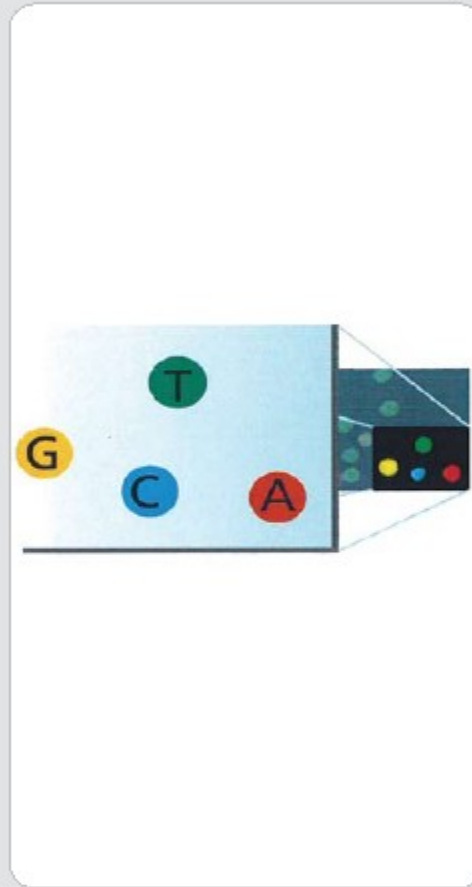
After laser excitation, the emitted fluorescence from each cluster is captured and the first base is identified.

9. DETERMINE SECOND BASE
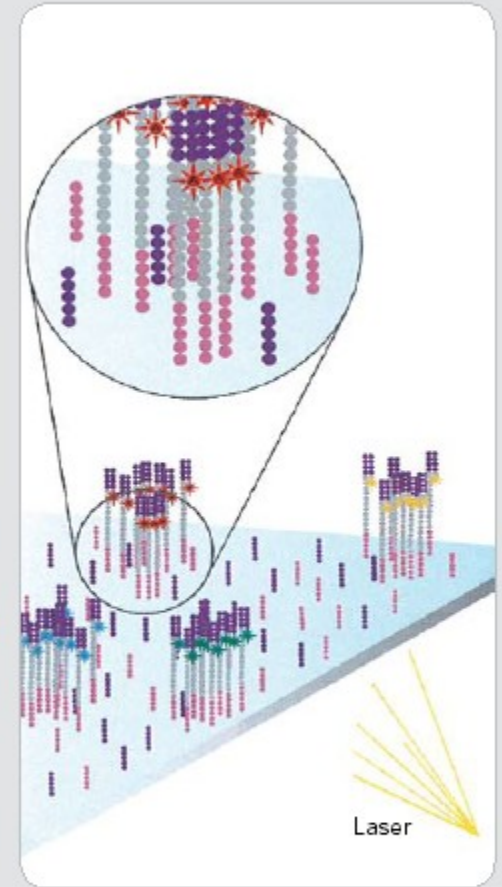
The next cycle repeats the incorporation of four labeled reversible terminators, primers, and DNA polymerase.

Ilumina

10. IMAGE SECOND CHEMISTRY CYCLE
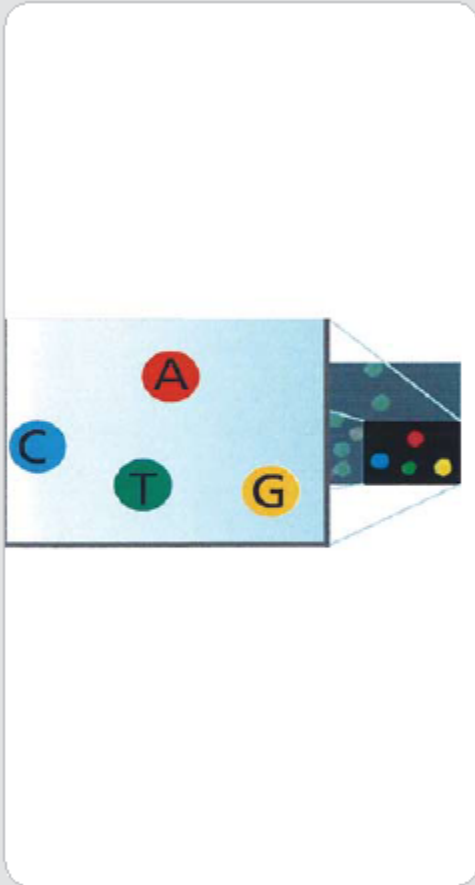
After laser excitation, the image is captured as before, and the identity of the second base is recorded.

11. SEQUENCING OVER MULTIPLE CHEMISTRY CYCLES

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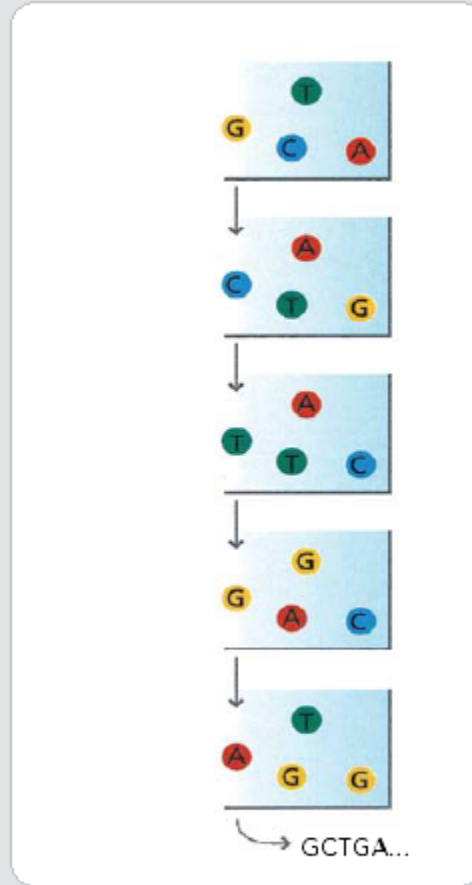
The sequencing cycles are repeated to determine the sequence of bases in a fragment, one base at a time.

12. ALIGN DATA

Reference sequence

..GCTGATGTGCCGCCTCACTCCGGTGG

CACTCCTGTGG
CTCACTCCTGTGG
GCTGATGTGCCACCTCA
GATGTGCCACCTCACTC
GTGCCGCCTCACTCCTG
CTCCTGTGG

Unknown variant identified and called

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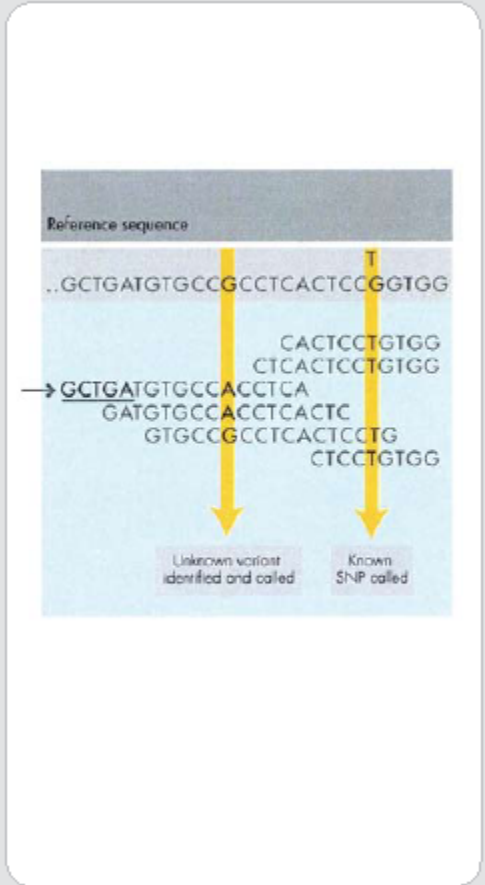
The data are aligned and compared to a reference, and sequencing differences are identified.

SOLiD sequencing

Applied Biosystems' SOLiD technology employs sequencing by ligation. Here, a pool of all possible oligonucleotides of a fixed length are labeled according to the sequenced position. Oligonucleotides are annealed and ligated; the preferential ligation by DNA ligase for matching sequences results in a signal informative of the nucleotide at that position. Before sequencing, the DNA is amplified by emulsion PCR. The resulting bead, each containing only copies of the same DNA molecule, are deposited on a glass slide. The result is sequences of quantities and lengths comparable to Illumina sequencing.

SEE ANIMATION: http://www.youtube.com/watch?v=nlvyF8bFDwM

SOLiD sequencing

Possible Dinucleotides Encoded By Each Color

2nd Base

1st Base

Template Sequence

	AT	AC	AA	GA
	CG	CA	C C	TC
	GC	GT	GG	AG
	TA	TG	T T	CT
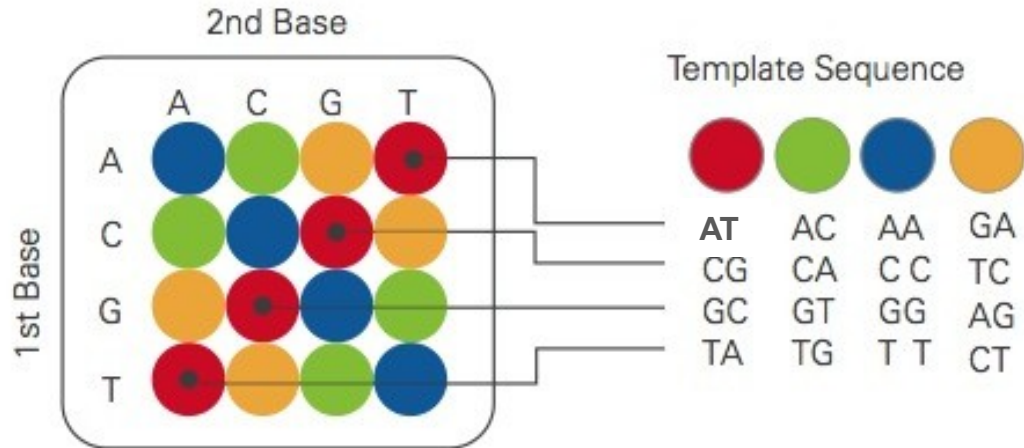
Double Interrogation

With 2 base encoding each
base is defined twice

A ←⟶ T ←⟶ G ←⟶ G ←⟶ A
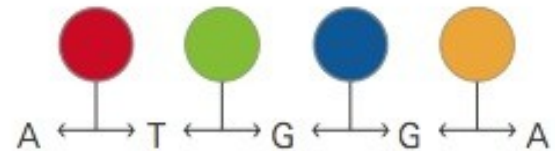
Library Preparation
Prepare one of the two types of libraries for SOLiD™ System sequencing- fragment or mate-paired. Your choice of library depends on the application you're performing and the information you desire from your experiments.

SOLiD sequencing
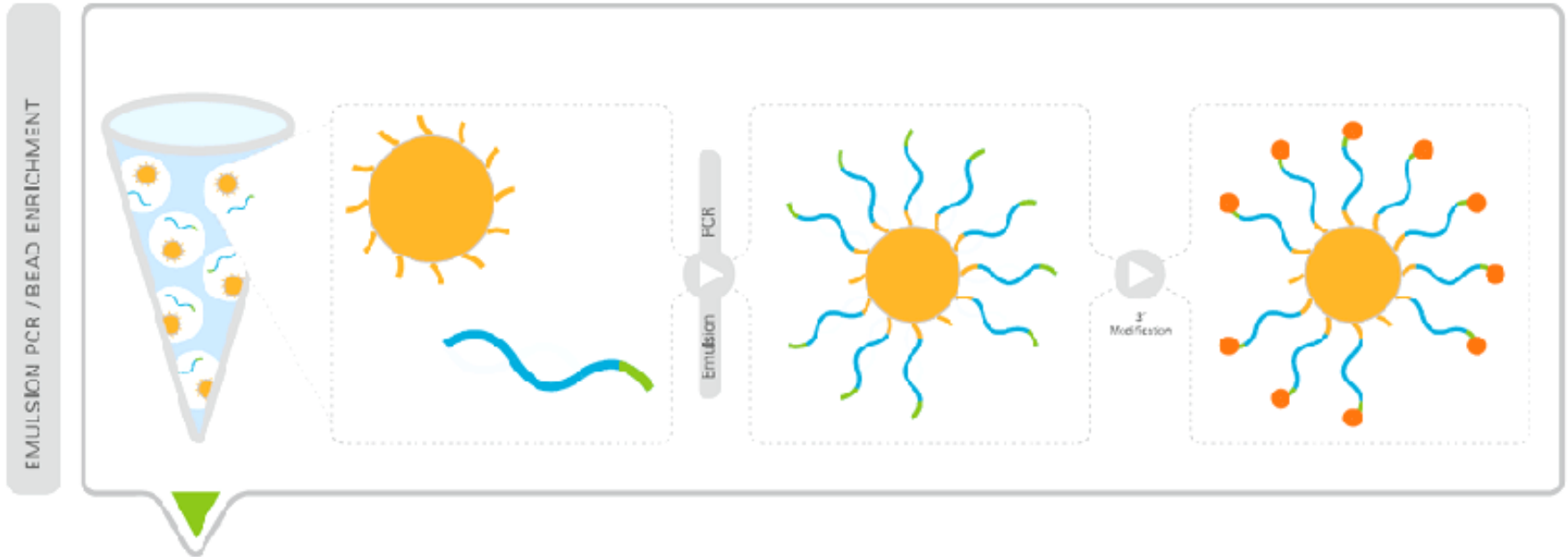
Emulsion PCR/Bead Enrichment

Prepare clonal bead populations in microreactors containing template, PCR reaction
   components, beads, and primers.
After PCR, denature the templates and perform bead enrichment to separate beads
   with extended templates from undesired beads. The template on the selected
   beads undergoes a 3' modification to allow covalent attachment to the slide.

SOLiD sequencing

BEAD DEPOSITION
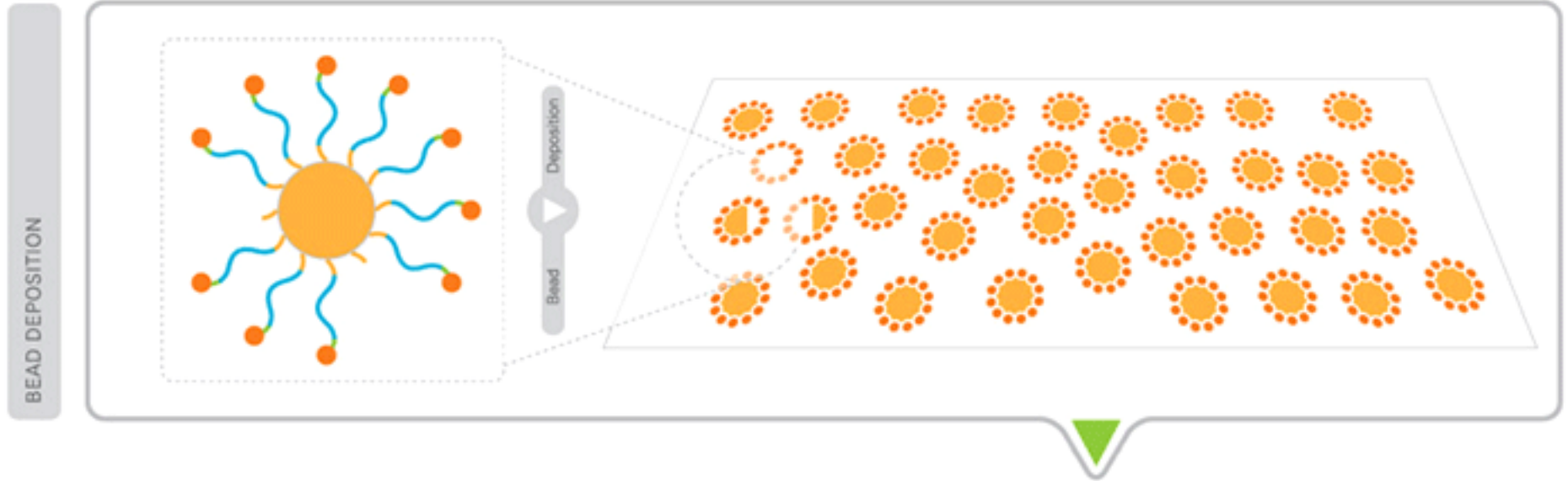
Bead · Deposition

Bead Deposition

Deposit 3' modified beads onto a glass slide. During bead loading, deposition chambers enable you to segment a slide into one, four, or eight sections. A key advantage of the system is the ability to accommodate increasing densities of beads per slide, resulting in a higher level of throughput from the same system.
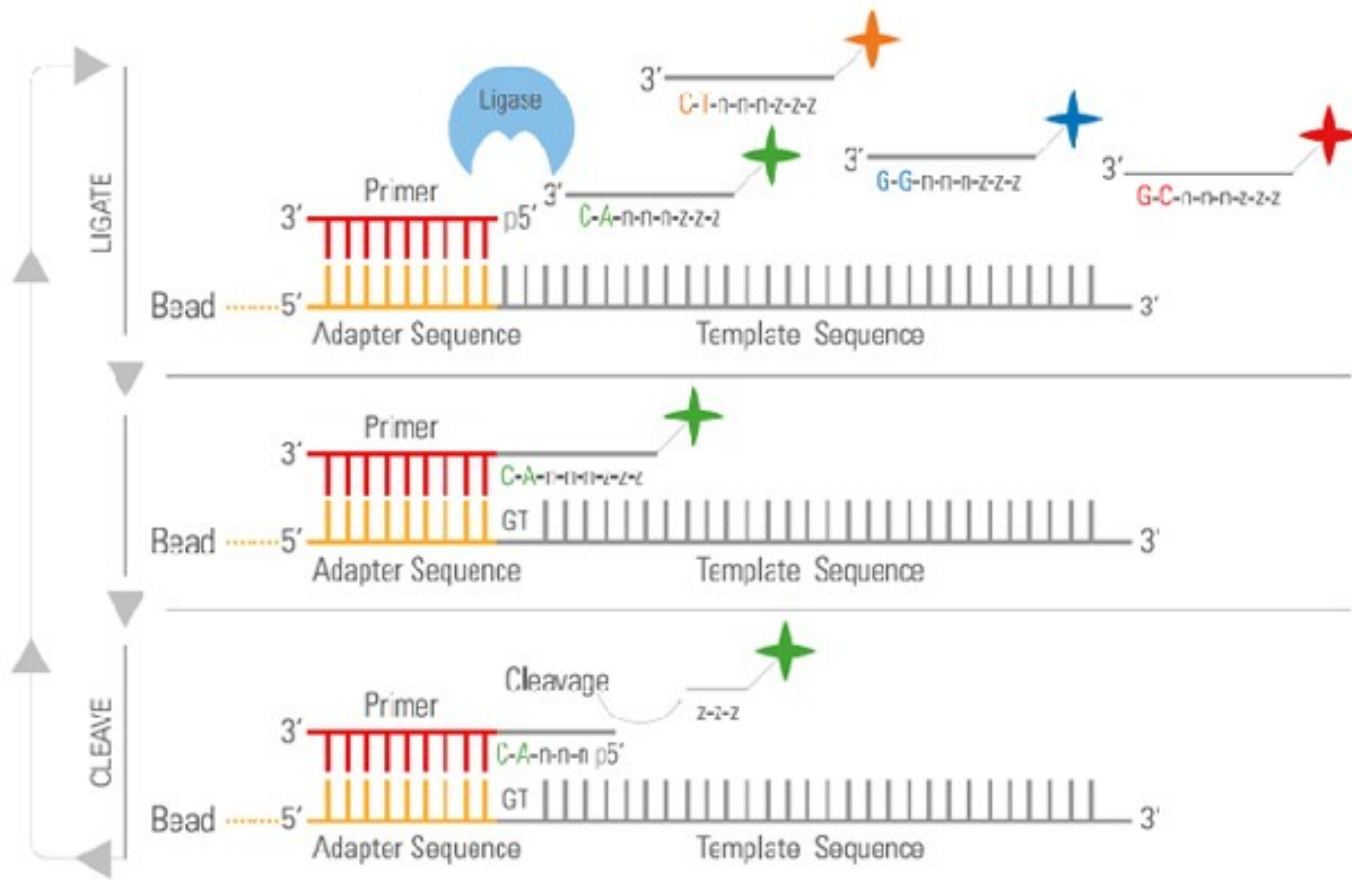
Sequencing by Ligation

1. Primers hybridize to the P1 adapter sequence on the templated beads.

2. A set of four fluorescently labeled di-base probes compete for ligation to the sequencing primer. Specificity of the di-base probe is achieved by interrogating every 1st and 2nd base in each ligation reaction.

3. Multiple cycles of ligation, detection and cleavage are performed with the number of cycles determining the eventual read length.

4. Following a series of ligation cycles, the extension product is removed and the template is reset with a primer complementary to the n-1 position for a second round of ligation cycles.

SOLiD sequencing

• Indicates positions of interrogation   Ligation Cycle 1 2 3 4 5 6 7

T ⇒ known nucleotide from adapter

TACggAATCAATCCgCCACTACgTTCAAgTgAgATg
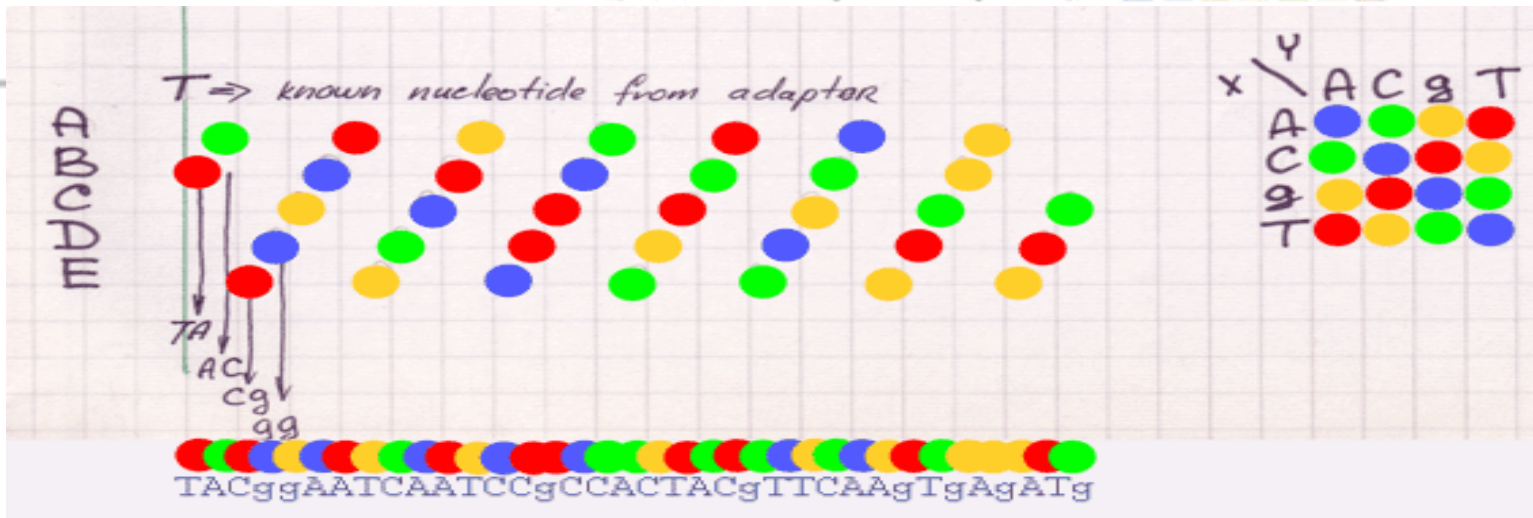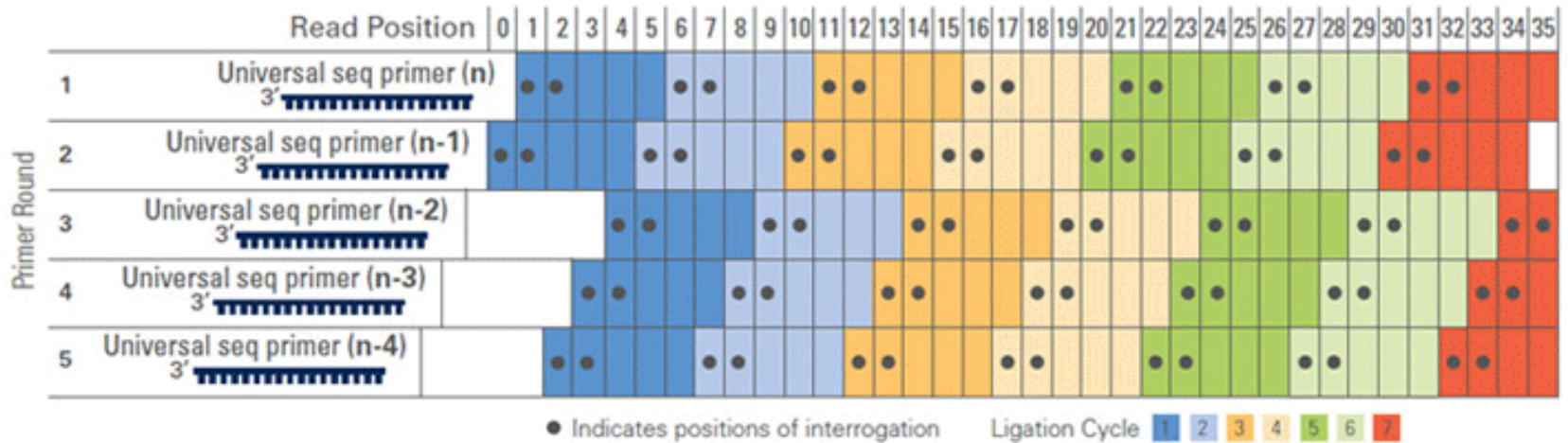
Primer Reset

Five rounds of primer reset are completed for each sequence tag. Through the primer reset process, virtually every base is interrogated in two independent ligation reactions by two different primers.

For example, the base at read position 5 is assayed by primer number 2 in ligation cycle 2 and by primer number 3 in ligation cycle 1. This dual interrogation is fundamental to the unmatched accuracy characterized by the SOLiD™ System.

a Life/APG — Sequencing by ligation

Primer round 1

Universal seq primer (*n*)

P1 adapter   Target sequence

1,2-probes
x, y Interrogation bases
n   Degenerate bases
z   Universal bases

3′ xynnnzzz 5′
3′ xynnnzzz 5′
3′ xynnnzzz 5′
3′ xynnnzzz 5′

Excite   Fluorescence, four-colour imaging

Cleavage agent

Repeat ligation cycles   Ligation cycle 1 2 3 4 5 6 Z.. (*n* cycles)

AT TT CT GT TT CA GC
TA AA GA CA AA GT CG

Reset primer (*n* − 1), repeat ligation cycles

Primer round 2   1 base shift

Universal seq primer (*n* − 1)

−1

AA CT GC TG AT CC CG
T   GA CG AC TA GG GC

Reset primer three more times

b
Two-base encoding: each target nucleotide is interrogated twice

2nd base   Template sequence

A C G T
A   ATACAAGA
C   CGCACCTC
G   GCGTGGAG
T   TATGTTCT

Alignment of colour-space reads to colour-space reference genome

SNP

TCGGATTCAGCCTGCTGCTCTATCA
                A

c Roche/454 — Pyrosequencing

1–2 million template beads loaded into PTP wells

Flow of single dNTP type across PTP wells

dNTP

Polymerase

APS   PP_i

Sulphurylase

Luciferase

ATP

Luciferin

Light and oxyluciferin

d Flowgram

TCAGGTTTTTTAACAATCAACTTTTTGGATTAAAATGTAGATAACTG
CATAAATTAATAACATCACATTAGTCTGATCAGTGAATTTAT

Nature Reviews | Genetics
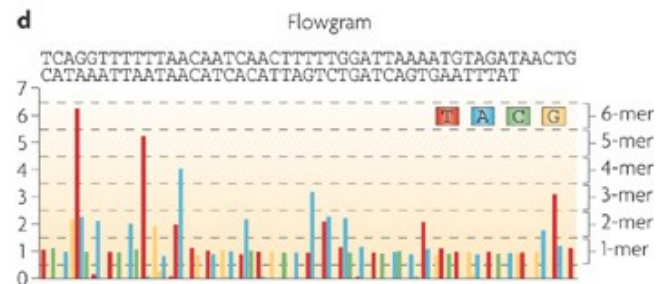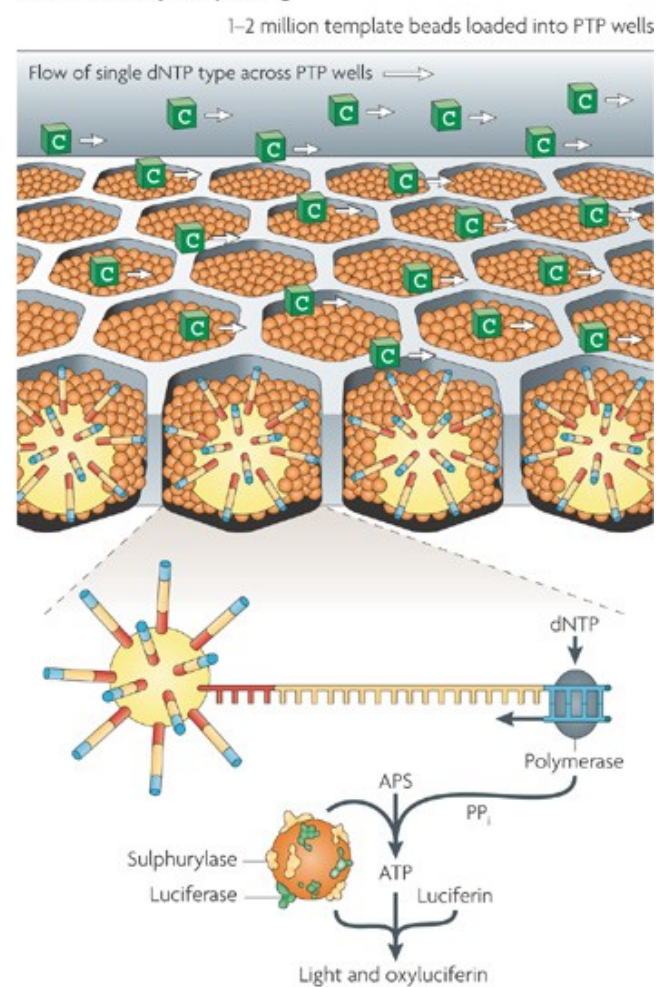
Applications

De novo sequencing: 454 (longer reads and high coverage - preferred for genome assembly)

Resequencing: Solexa, SOLiD (shorter, but reliable reads) – identification of differences compared to reference sequence – e.g. molecular diagnostics,...

Transcriptome sequencing – substitutes for cDNA/RNA microarrays,

- representation of reads transcripts (normalized to reference transcript) reflects abundance of the transcript

- more precise than microarrays

-possibility of detection of unknown transcripts

Whole genome sequencing, **1000 genome project (human genome variation)**

Metagenomics – analysis of complete genetic material from environmental samples **without cultivation –** real diversity (e.g. microbial composition of certain habitats, human organs - only about 1 out of 10 cells in the human body is actually a human cell: most of the cells that make up our bodies are microbes!)

Epigenetics / epigenomics - DNA methylation, ChIP-Seq

Ion semiconductor sequencing

Ion Torrent Systems Inc. (now owned by Life Technologies) developed a system based on using standard sequencing chemistry, but with a novel, **semiconductor based detection system**. This method of sequencing is based on the **detection of hydrogen ions** that are released during the polymerisation of DNA, as opposed to the optical methods used in other sequencing systems. A microwell containing a template DNA strand to be sequenced is flooded with a single type of nucleotide. If the introduced nucleotide is complementary to the leading template nucleotide it is incorporated into the growing complementary strand. This causes the release of a hydrogen ion that triggers a hypersensitive ion sensor, which indicates that a reaction has occurred. If homopolymer repeats are present in the template sequence multiple nucleotides will be incorporated in a single cycle. This leads to a corresponding number of released hydrogens and a proportionally higher electronic signal.

http://www.iontorrent.com/

Ion Torrent

Step One
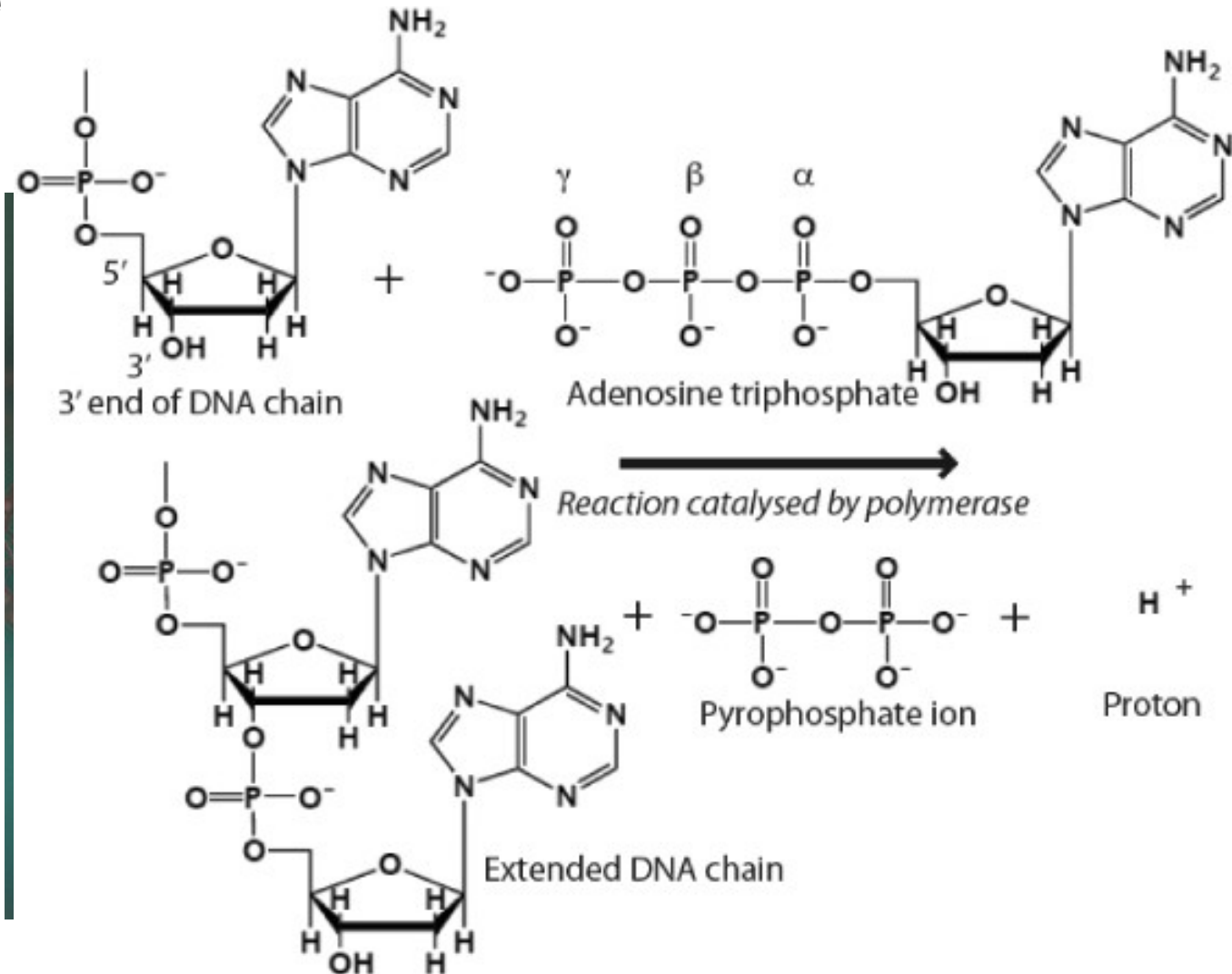
In nature, when a nucleotide is incorporated into a strand of DNA by a polymerase, a **hydrogen ion is released** as a byproduct.

Step Two
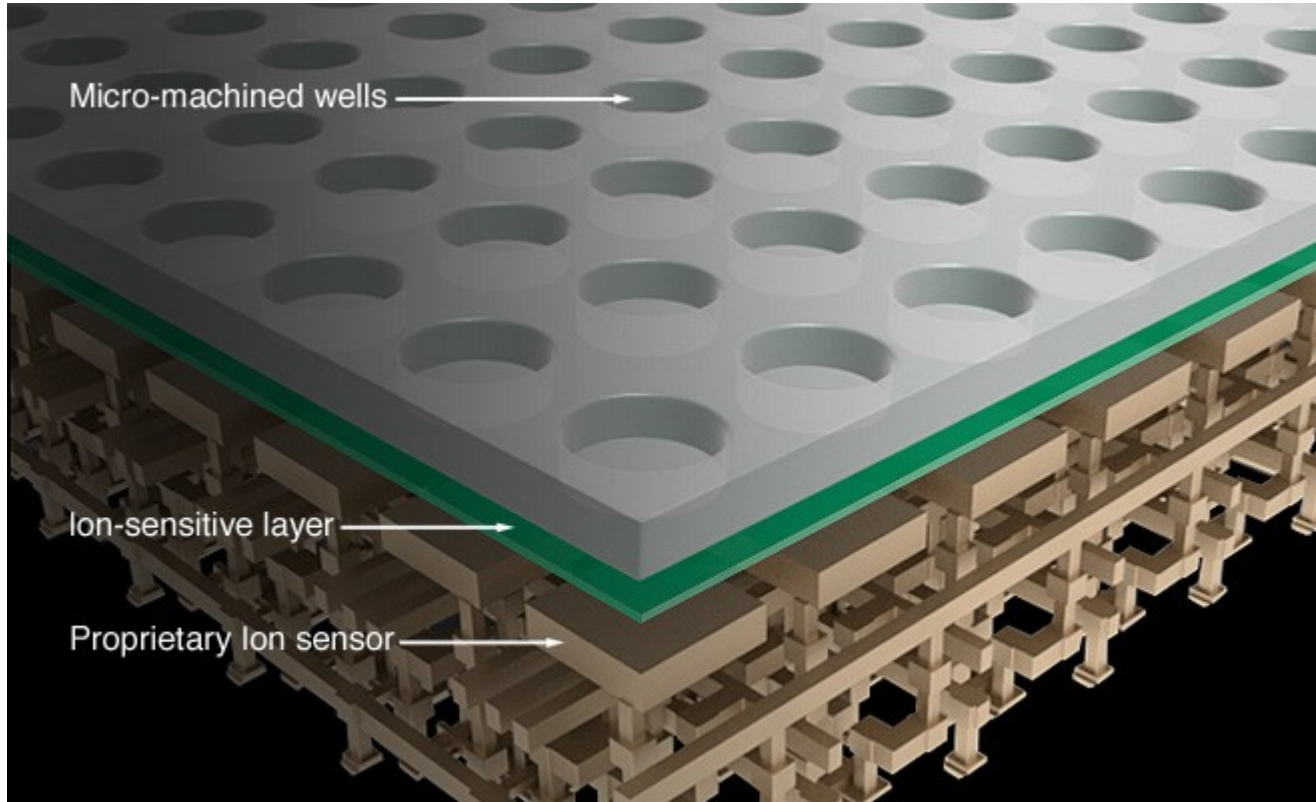
Ion Torrent™ uses a high-density array of micro-machined wells to perform this biochemical process in a massively parallel way. Each well holds a different DNA template. Beneath the wells is an ion-sensitive layer and beneath that a proprietary Ion sensor.

Ion Torrent

Step Three
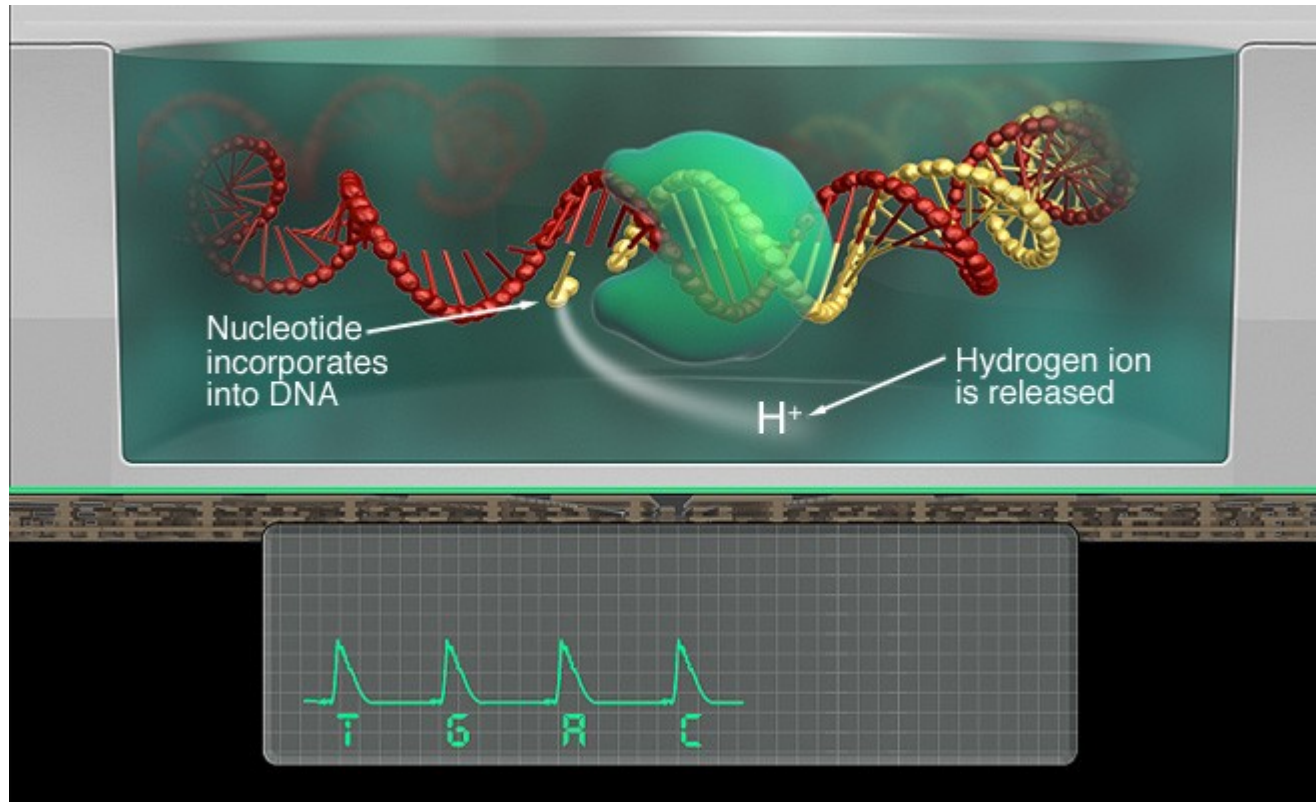
If a nucleotide, for example a C, is added to a DNA template and is then incorporated into a strand of DNA, a hydrogen ion will be released. The charge from that ion will change the pH of the solution, which can be detected by our proprietary ion sensor. The sequencer - essentially the world's smallest solid-state pH meter - will call the base, going directly from chemical information to digital information.

Ion Torrent

Step Four
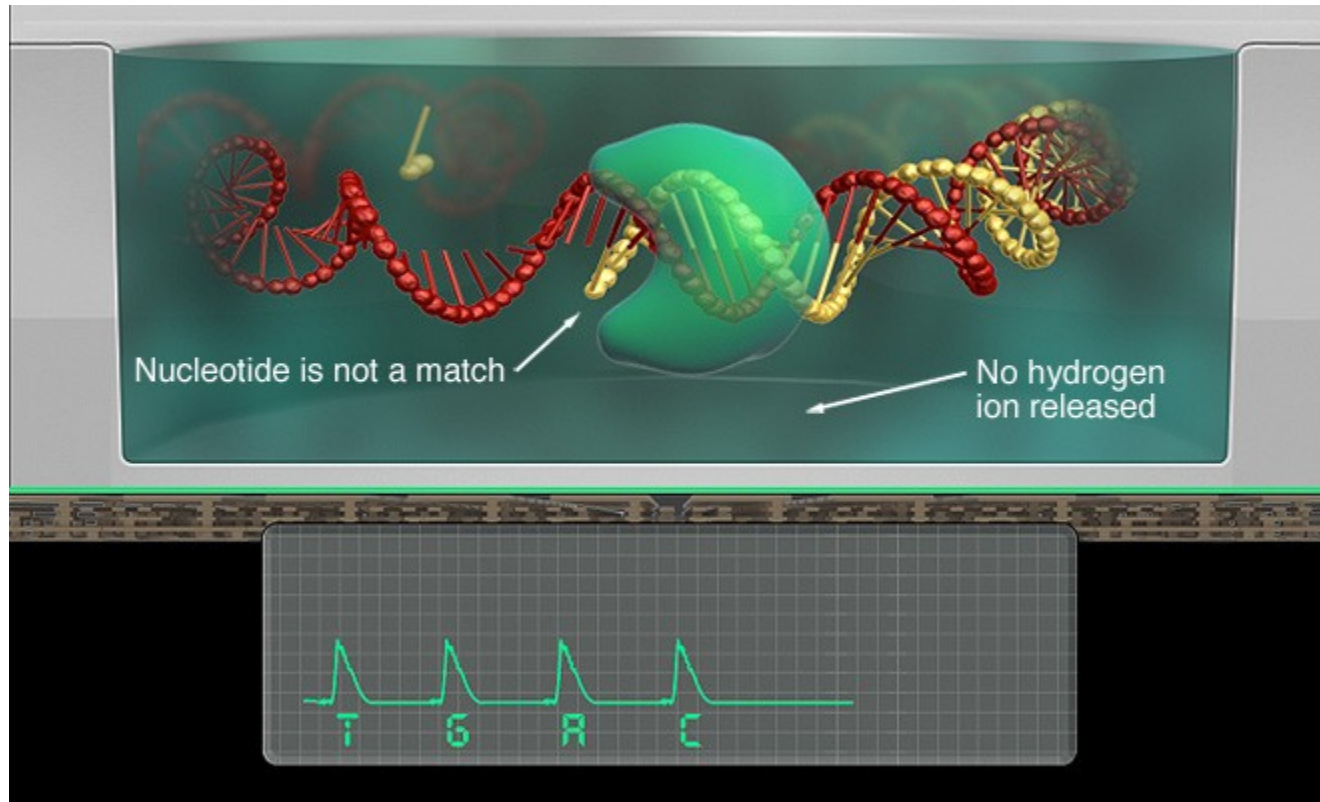
The Ion Personal Genome Machine™ (PGM™) sequencer then sequentially floods the chip with one nucleotide after another. If the next nucleotide that floods the chip is not a match, no voltage change will be recorded and no base will be called.

Ion Torrent

Two bases are incorporated

Two hydrogen ions are released

H⁺ H⁺

T   G   A   C   T T
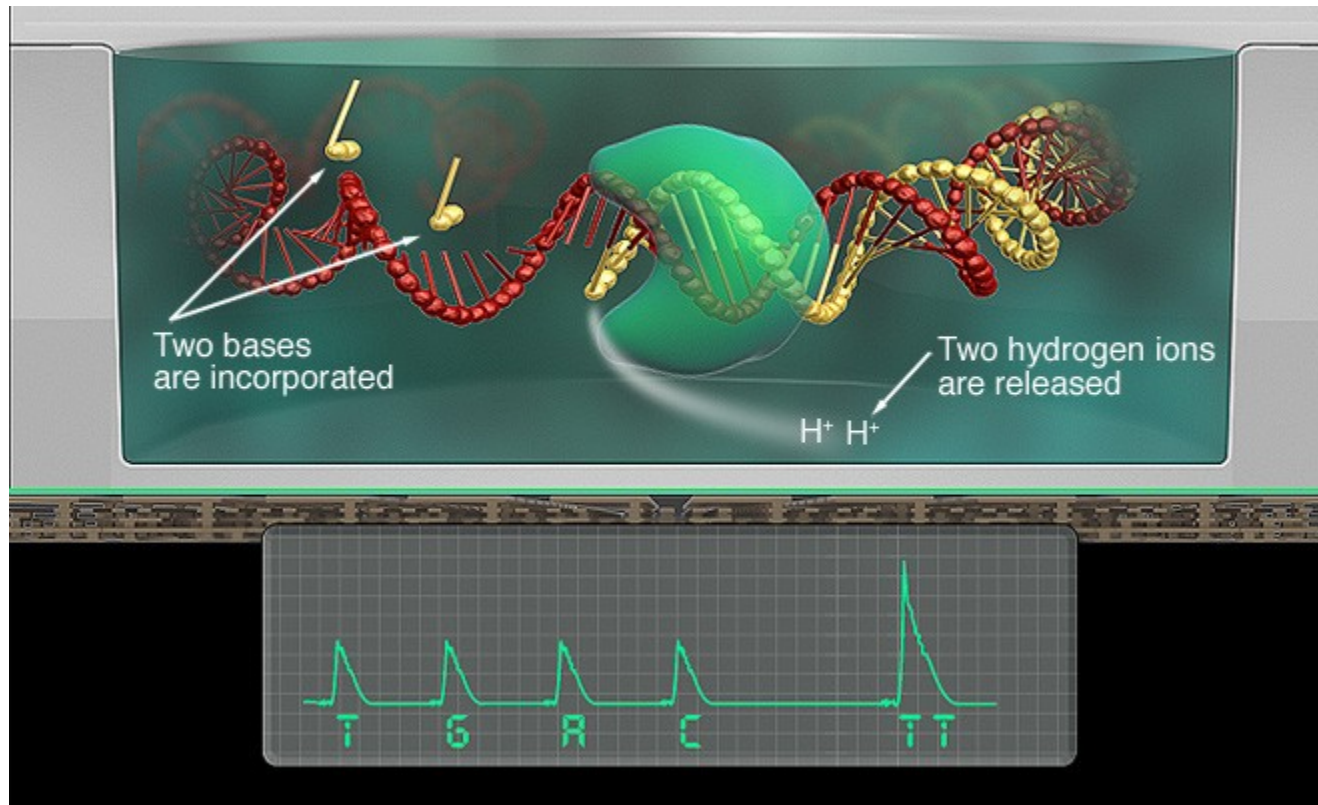
If there are two identical bases on the DNA strand, the voltage will be double, and the chip will record two identical bases called. Because this is direct detection - no scanning, no cameras, no light - each nucleotide incorporation is recorded in seconds. The semiconductor will inevitably transform the life sciences, just as it has transformed every other industry it has touched. By creating a direct connection between chemical and digital information, Ion Torrent™ will democratize research, providing a fast, simple, scalable sequencing solution that every lab can afford. Eventually, Ion Torrent Technology™ will also be able to provide diagnostics that are less expensive and more reliable, improving human health around the world.

Helioscope^{TM} single molecule sequencing

Based on **"true single molecule sequencing"** technology, Helioscope sequencing uses DNA fragments with added polyA tail adapters, which 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 are performed by the Helioscope sequencer. The reads are short, up to 55 bases per run, but recent improvement of the methodology allows more accurate reads of homopolymers (stretches of one type of nucleotides) and RNA sequencing.

Advantages: High number of reads, no sequencing bias (no amplification step)

a Illumina/Solexa — Reversible terminators

Incorporate all four nucleotides, each label with a different dye

Wash, four-colour imaging

Cleave dye and terminating groups, wash

Repeat cycles

b

C	A
T	G

Top: CATCGT
Bottom: CCCCCC
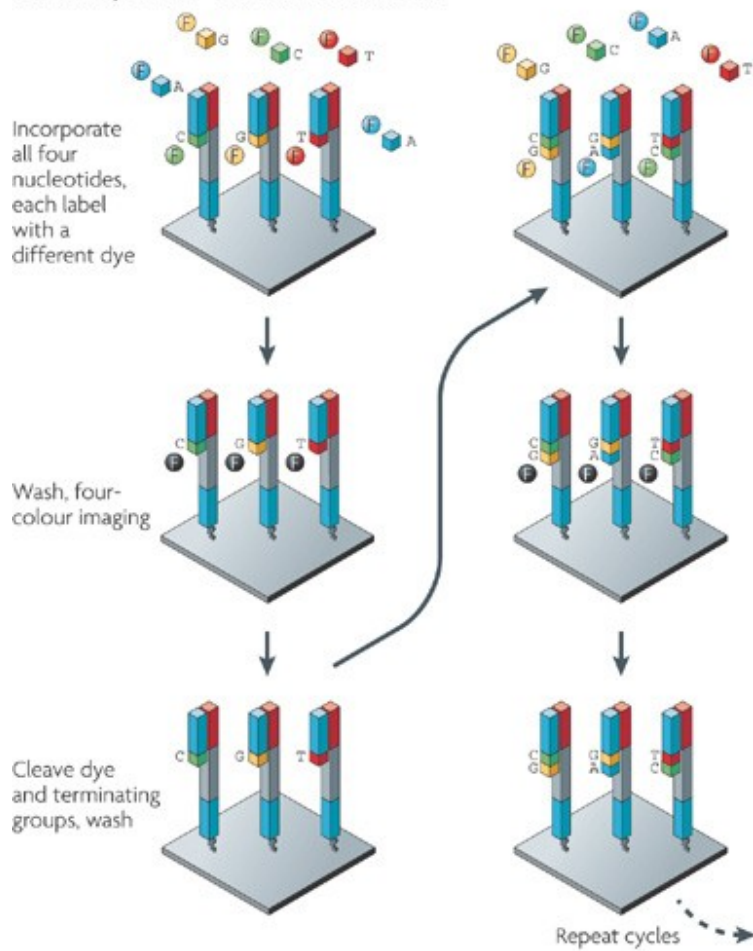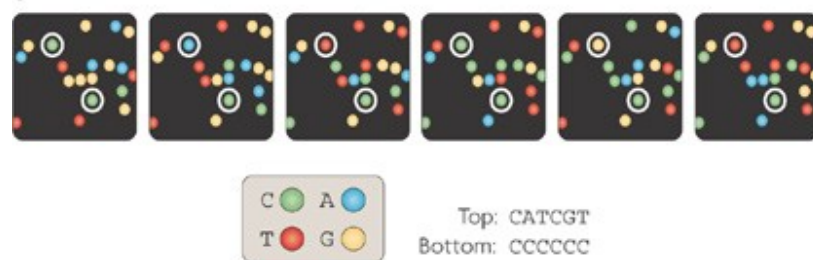
c Helicos BioSciences — Reversible terminators

Incorporate single, dye-labelled nucleotides

Each cycle, add a different dye-labelled dNTP

Wash, one-colour imaging

Cleave dye and inhibiting groups, cap, wash

Repeat cycles

d

C   T   A   G

C   T   A   G

Top: CTAGTG
Bottom: CAGCTA
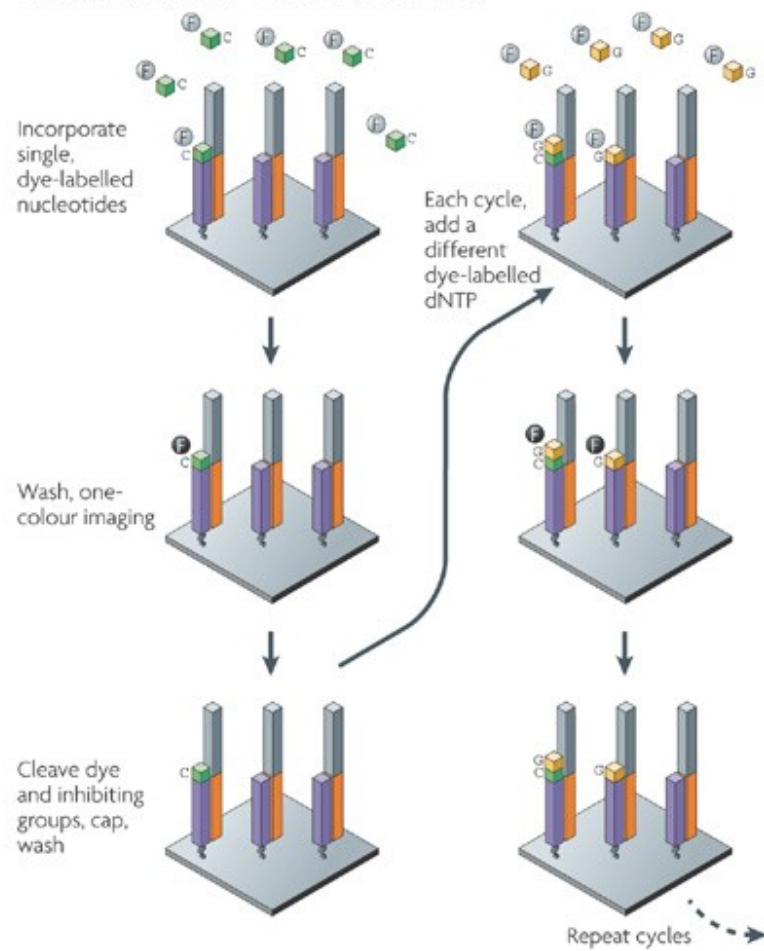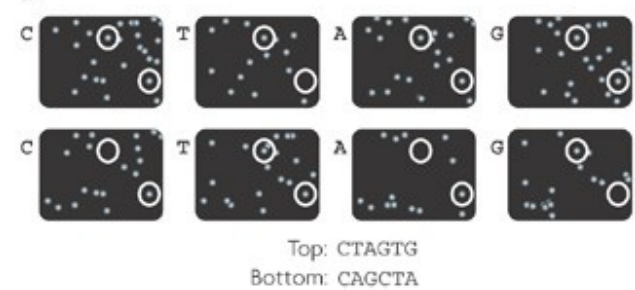
How tSMS Works

Within two flow cells, billions of single molecules of sample DNA are captured on an application-specific proprietary surface. These **captured strands** serve as templates for the sequencing-by-synthesis process:

1. Polymerase and **one** fluorescently labeled nucleotide (C, G, A or T) are added.
2. The polymerase catalyzes the sequence-specific incorporation of fluorescent nucleotides into nascent complementary strands on all the templates.
3. After a wash step, which removes all free nucleotides, the incorporated nucleotides are imaged and their positions recorded.
4. The fluorescent group is removed in a highly efficient cleavage process, leaving behind the incorporated nucleotide.
5. The process continues through each of the other three bases.

Multiple four-base cycles result in complementary strands greater than 25 bases in length synthesized on billions of templates—providing a greater than 25-base read from each of those individual templates.

	Position		
	1	2	3
Cycle 1	-	-	G
Cycle 2	C	C	-
Cycle 3	A	A	A
Cycle 4	-	-	T
Cycle 5	C	-	-
Cycle X	G	G	-
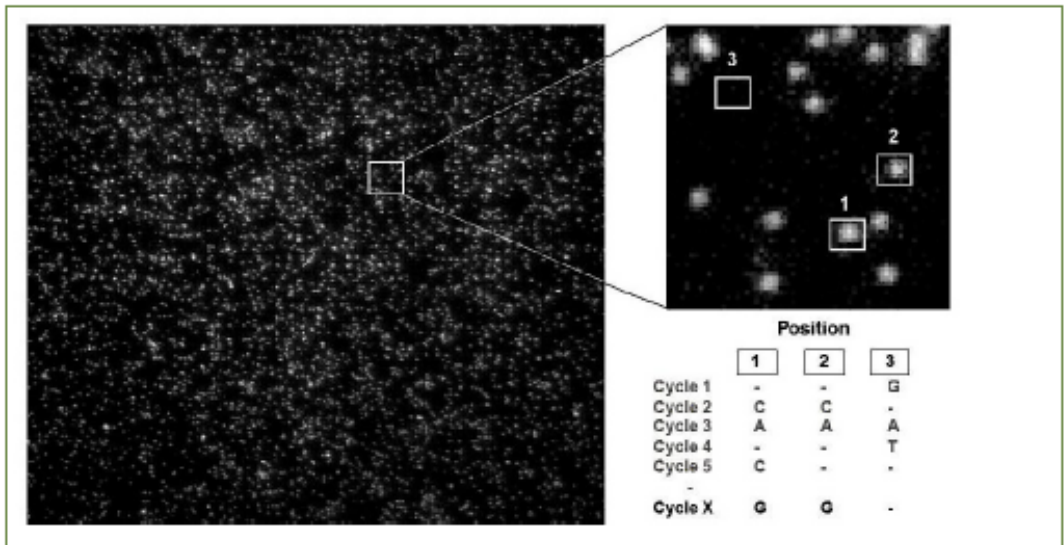
An image taken by the HeliScope Single Molecule Sequencer. Inset shows a close-up view of individual single molecules that incorporated a fluorescent "G" nucleotide in this cycle.
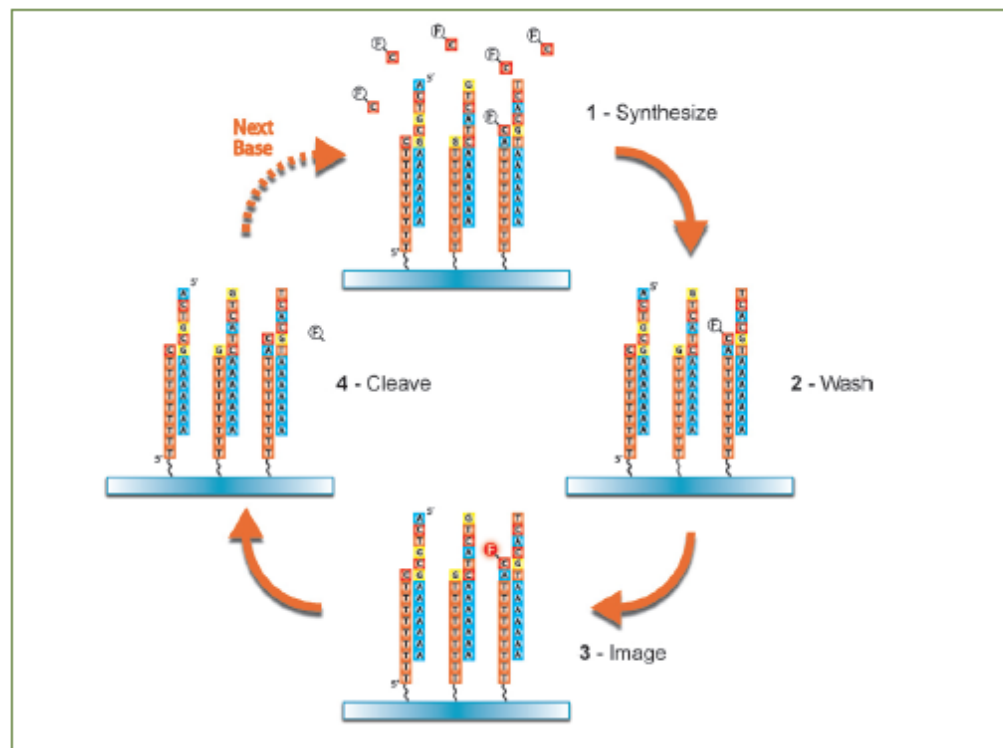
After hybridization of modified genomic DNA to oligo dT capture primers inside the flow cell, acquisition of template positions, and cleaving the fluorescent templated lablel, the tSMS process is a cyclical process involving multiple rounds of (1) synthesis using labeled nucleotides, (2) washing, (3) imaging, and (4) cleaving the fluorescent label until the desired read length is achieved.

	Sanger Sequencing	"Next Generation" Sequencing	Helicos True Single Molecule Sequencing
Information Capacity:	**100's** of reads per experiment	**100,000,000's** of reads per experiment	**1,000,000,000's of reads per experiment**
Scalability of Sample Preparation:	A few at a time	A few at a time	**Hundreds at a time Easily automated**
Amplification:	Required	Required	**No amplification True direct DNA measurement**
Accuracy:	*Analog base calls* Sequencing biases Not quantitative	*Analog base calls* Sequencing biases Not quantitative	*Digital base call* **No amplification biases Digital quantitation**

Single molecule SMRT™ sequencing

SMRT sequencing is based on the sequencing by synthesis approach. The DNA is synthesized in **zero-mode wave-guides** (ZMWs) - small well-like containers with the capturing tools located at the bottom of the well. The sequencing is performed with use of unmodified **polymerase** (**attached to the ZMW bottom**) and fluorescently labelled nucleotides flowing freely in the solution. The wells are constructed in a way that **only the fluorescence occurring by the bottom of the well is detected**. The fluorescent label is detached from the nucleotide at its incorporation into the DNA strand, leaving an unmodified DNA strand. According to *Pacific Biosciences*, the SMRT technology developer, this methodology allows detection of nucleotide modifications (such as cytosine methylation). This happens through the observation of polymerase kinetics. This approach allows reads of up to **15,000 nucleotides**, with mean read lengths of **2.5 to 2.9 kilobases**

Boundary Surface

Diffracted Wave

Evanescent Wave (totally reflected part not shown)

The SMRT Cell

One of the fundamental challenges with observing a **DNA polymerase working in real time** is the ability to detect the incorporation of a single nucleotide, taken from a large pool of potential nucleotides, during DNA synthesis. We applied the same principle that operates in the metallic screen of a microwave oven door. In a microwave oven, the screen is perforated with holes that are much smaller than the wavelength of the microwaves. Because of their relative size, the holes prevent the much longer microwaves from passing through and penetrating the glass. However, the much smaller wavelength visible light is able to pass through the holes in the screen, allowing food to be visible. We have reduced this same principle to the nanoscale and we call our innovation a **zero-mode waveguide**, or ZMW.

With an **active polymerase immobilized at the bottom of each ZMW**, nucleotides diffuse into the ZMW chamber. In order to detect incorporation events and identify the base, each of the four nucleotides A, C, G and T are labeled with a different fluorescent dye having a distinct emission spectrum. Since the excitation illumination is directed to the bottom of the ZMW, nucleotides held by the polymerase prior to incorporation emit an extended signal that identifies the base being incorporated.

A **ZMW** is a hole, tens of nanometers in diameter, fabricated in a 100nm metal film deposited on a glass substrate. The small size of the ZMW prevents visible laser light, which has a wavelength of approximately 600nm, from passing entirely through the ZMW. Rather than passing through, the light **exponentially decays** as it enters the ZMW. Therefore, by shining a laser through the glass into the ZMW, **only the bottom 30nm of the ZMW becomes illuminated**. **Within each ZMW, a single DNA polymerase molecule is anchored to the bottom glass surface** using a proprietary technique. Nucleotides, each type labeled with a different colored fluorophore, are then flooded above an array of ZMWs at the required concentration. Diffusion at the nanoscale is incredibly fast. Within microseconds, labeled nucleotides travel down into the ZMW, surround the DNA polymerase, then diffuse back up and exit the hole. **As no laser light penetrates up through the holes to excite the fluorescent labels, the labeled nucleotides above the ZMWs are dark**. Only when they diffuse through the bottom 30nm of the ZMW do they fluoresce. **When the correct nucleotide is detected by the polymerase, it is incorporated into the growing DNA strand in a process that takes milliseconds in contrast to simple diffusion which takes microseconds**. This difference in time results in higher signal intensity for incorporated versus unincorporated nucleotides, which creates a high signal-to-noise ratio. Thus, the **ZMW has the ability to detect a single incorporation event against the background of fluorescently labeled nucleotides** at biologically relevant concentrations.

DNA sequencing is performed on proprietary SMRT Cells, each having an array of approximately 75,000 ZMWs. Each ZMW is capable of containing a DNA polymerase loaded with a different strand of DNA sample. As a result, the SMRT Cell enables the potential detection of approximately 75,000 single molecule sequencing reactions in parallel.
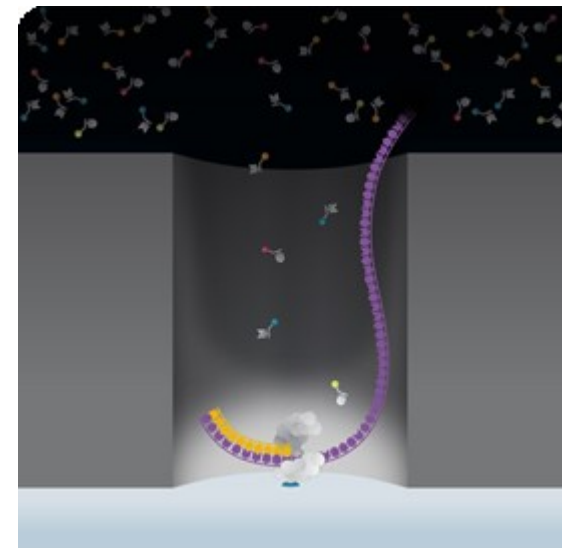
Phospholinked Nucleotides

Previous labeling technologies attach a fluorescent label to the base of the nucleotide, which is incorporated into the DNA strand. This is problematic for any system to observe DNA synthesis in real time because the dye's large size can interfere with the activity of the DNA polymerase. Typically, a DNA polymerase can incorporate only a few base-labeled nucleotides before it halts. Our proprietary phospholinked nucleotides have a **fluorescent dye attached to the phosphate chain of the nucleotide** rather than to the base. The phosphate chain is cleaved when the nucleotide is incorporated into the DNA. Thus, upon incorporation of a phospholinked nucleotide, the DNA polymerase naturally frees the dye molecule from the nucleotide when it cleaves the phosphate chain. Upon cleaving, the label quickly diffuses away, with no evidence of labeling remaining.

Pacific Biosciences — Real-time sequencing
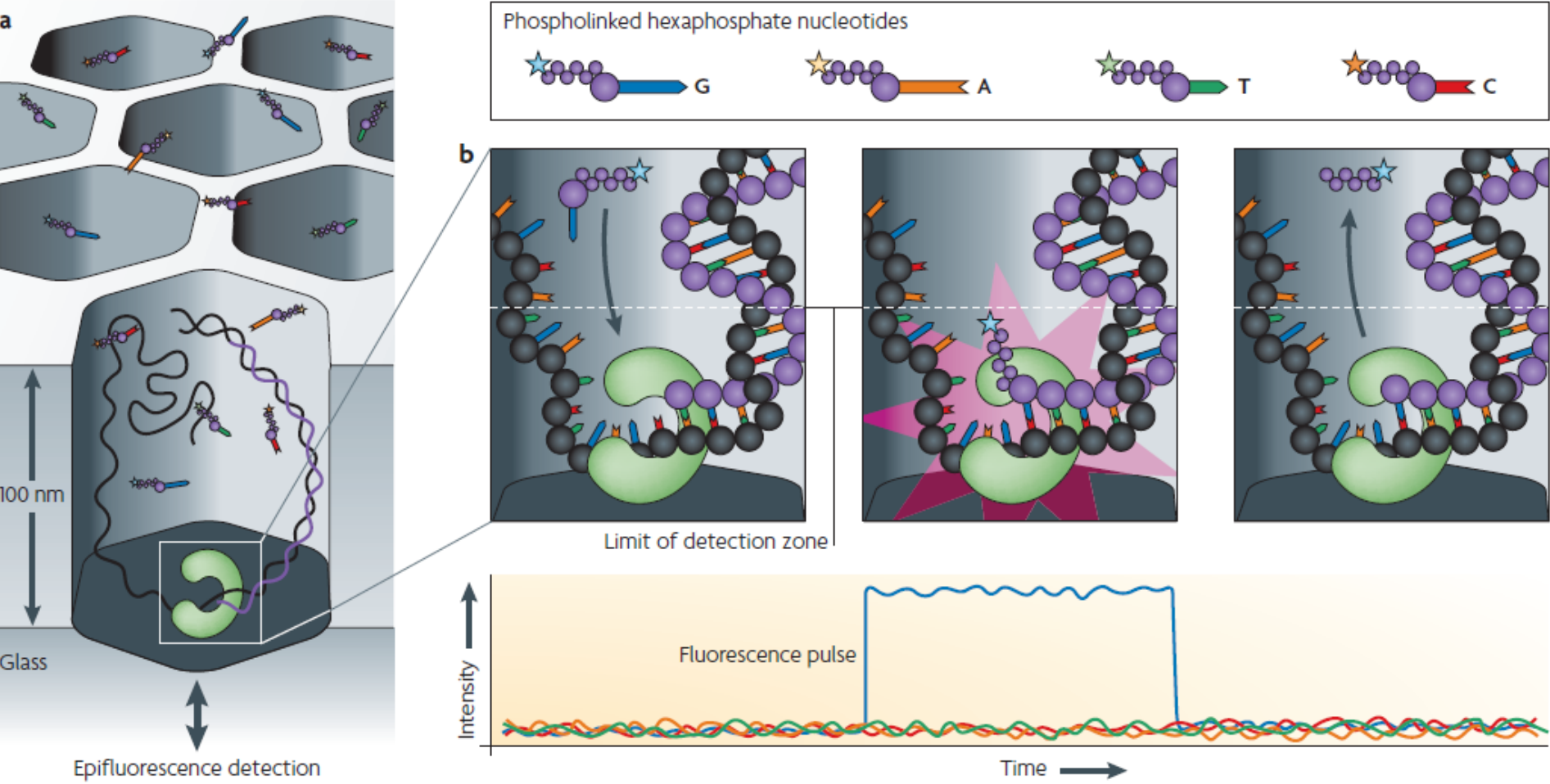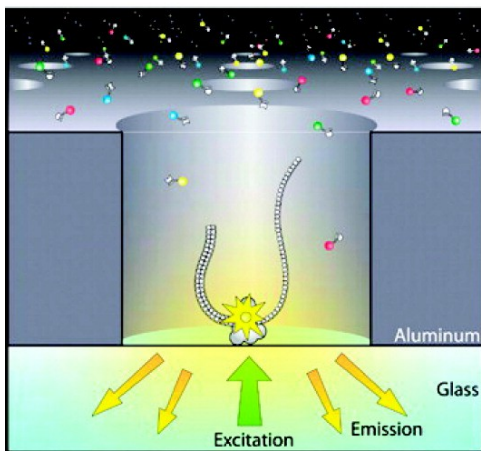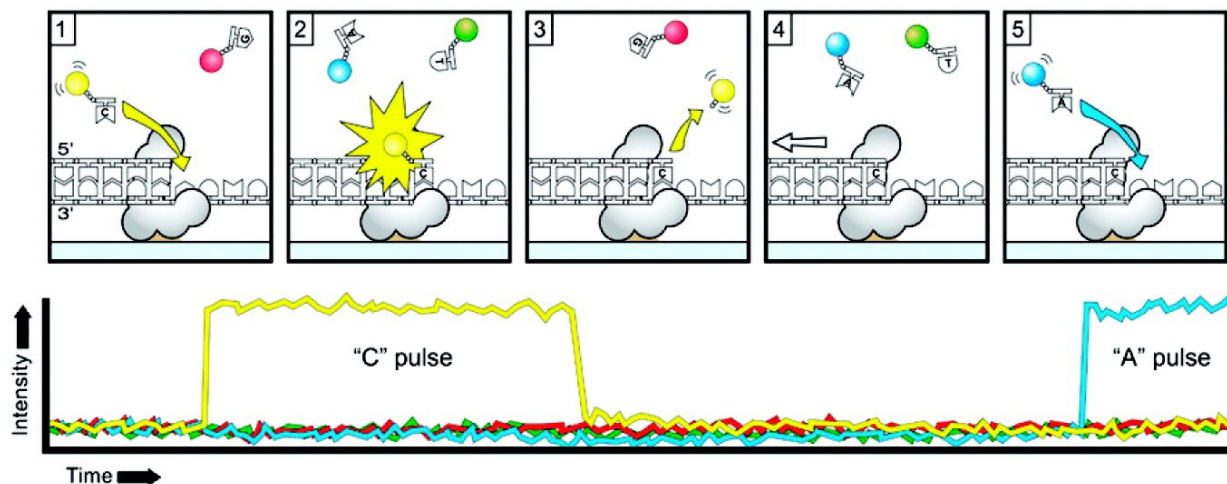
Schematic of PacBio's real-time single molecule sequencing. (A) The side view of a single ZMW nanostructure containing a single DNA polymerase (Φ29) bound to the bottom glass surface. The ZMW and the confocal imaging system allow fluorescence detection only at the bottom surface of each ZMW. (B) Representation 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.

21.3.2017

Optical mapping: bridging between chromosome and the sequence

Individual DNA molecules hundreds of kbp long may be stretched and visualized by optical microscopy.

► An optical barcode is generated by fluorescent labeling of short sequence motifs along the stretched DNA.

► Optical maps complement DNA sequencing for gap closing, finishing, validation and *de novo* assembly of genomes.

► Genome structural variations not accessible to sequencing or DNA arrays may be directly visualized.

► Epigenetic marks such as DNA methylation and DNA binding proteins may also be mapped on single genomic fragments.

Michal Levy-Sakin , Yuval Ebenstein

Beyond sequencing: optical mapping of DNA in the age of nanotechnology and nanoscopy
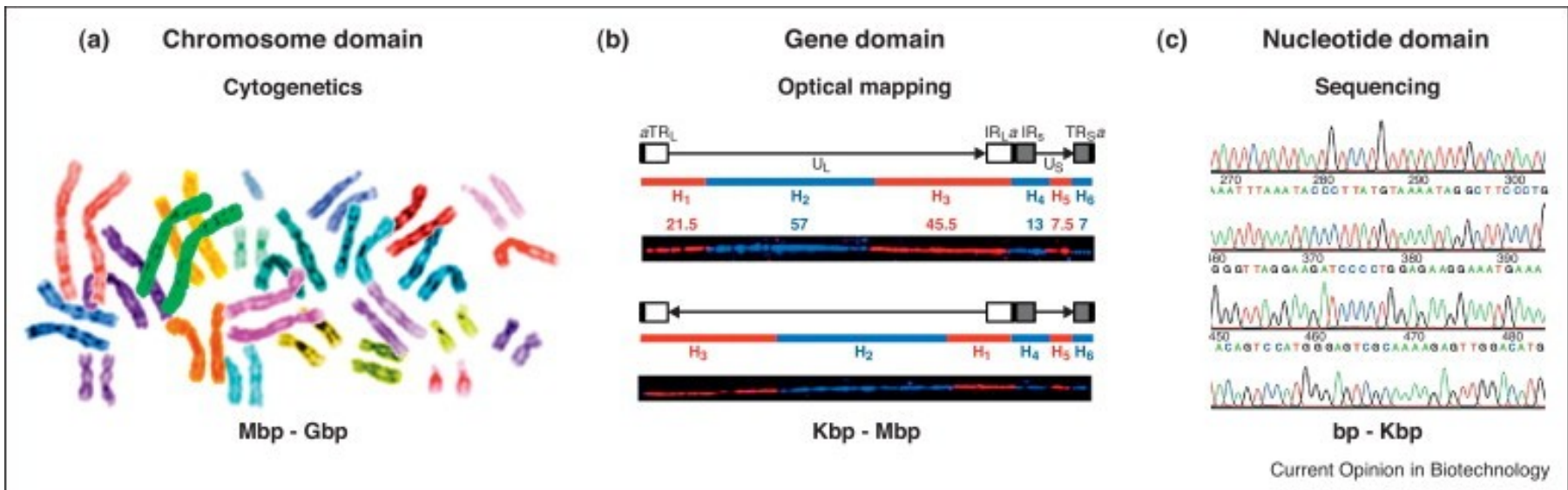
Genomic data domains — from whole chormosome to single nucleotide. **(a)** Spectral human karyotype illustration (taken by Jane Ades, NHGRI, adapted from www.genome.gov). Spectral Karyotyping (SKY) and FISH analysis allow the detection of whole or parts of chromosomes in resolution ranging between a few Mbp to 1 Gbp. **(b)** Optical mapping allows detection of DNA with resolution of 1 kbp to several Mbp, here as an example for optical mapping, genomic morse code of *Herpes simplex* virus-1 (HSV-1) isomers (modified from [44], with permission, American Society for Microbiology). Fluorescent hybridization arrays illustrating the HSV isomers P and IL are shown below the scheme. The sizes of the fragments shown are given in kbp according to the reference HSV-1 genome sequence (Genbank accession number: NC_001806). **(c)** The highest resolution of DNA mapping is achieved by DNA sequencing which allows single nucleotide analysis.

Michal Levy-Sakin , Yuval Ebenstein
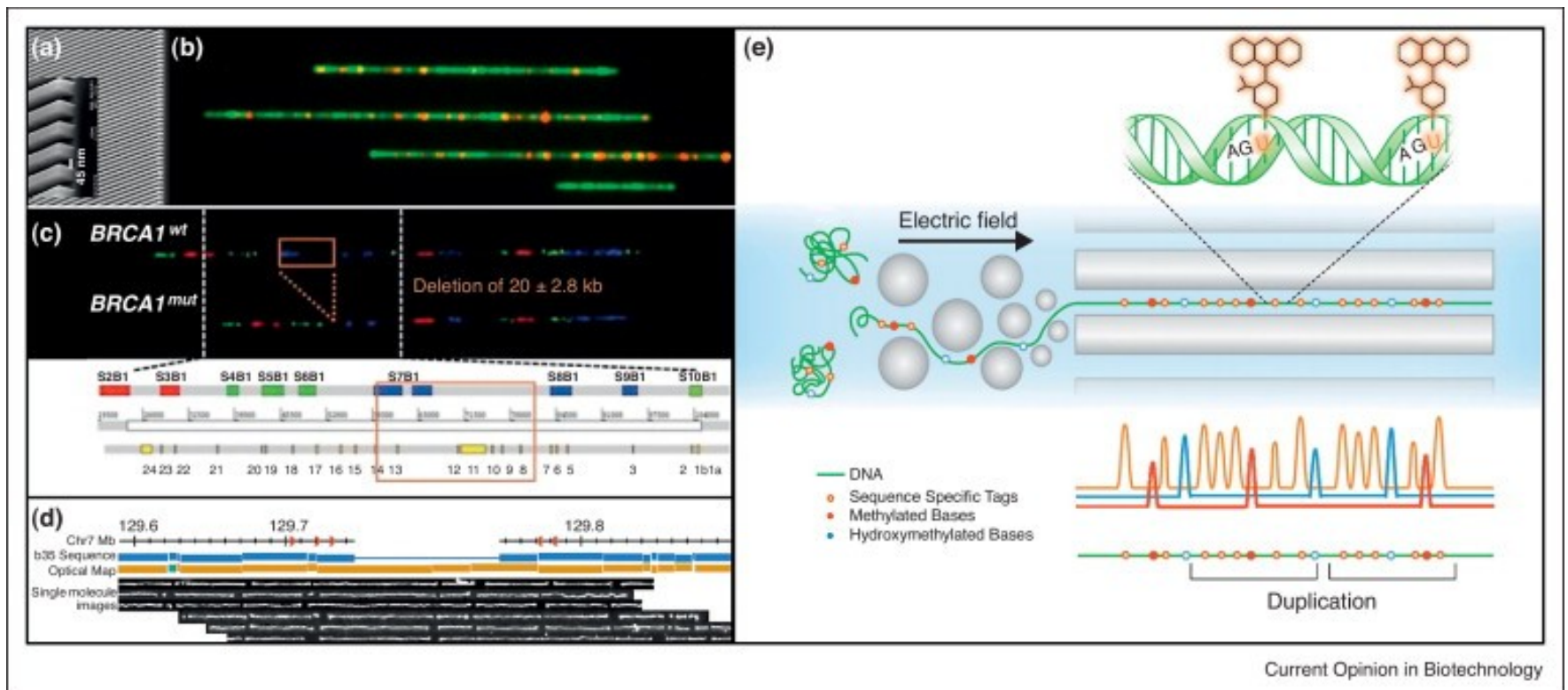
Optical mapping schemes. **(a)** Scanning electron microscope (SEM) image of a highly parallel nanochannel-array chip (*source*: BioNano Genomics, with permission). **(b) BAC DNA molecules (green) are nicked by an enzyme at specific sequence motifs and repaired by a polymerase that incorporates fluorescently labeled nucleotides** (orange dots) and imaged in nanochannel-array. **(c)** DNA molecular combing. Large rearrangements in BRCA1 are detected in breast cancer patients. Del ex 8–13 is visible as a deletion of the blue signal S7B1. **(d)** Restriction-optical mapping. A 90 kbp insertion is detected (red arrows). Included below the map is a montage of several of the single-molecule images that give evidence to support this insertion. **(e)** Future perspective for optical mapping in nanochannels: DNA molecules (green) are specifically labeled with different colored probes revealing different genomic/epigenomic information such as sequence motifs and methylation sites (colored dots). An applied electric field drives the molecules through a series of progressively smaller nanoscale obstacles (gray circles) that funnel the molecules into nanochannels. Once DNA is stretched and confined within the channels, the distances between labels can be accurately measured using a fluorescence microscope. Software is used to generate a consensus map of the sequence motifs overlaid with the epigenetic patterns of individual molecules. The maps facilitate the analysis of structural and epigenetic variation, such as sequence duplications and methylation patterns.

Michal Levy-Sakin, Yuval Ebenstein
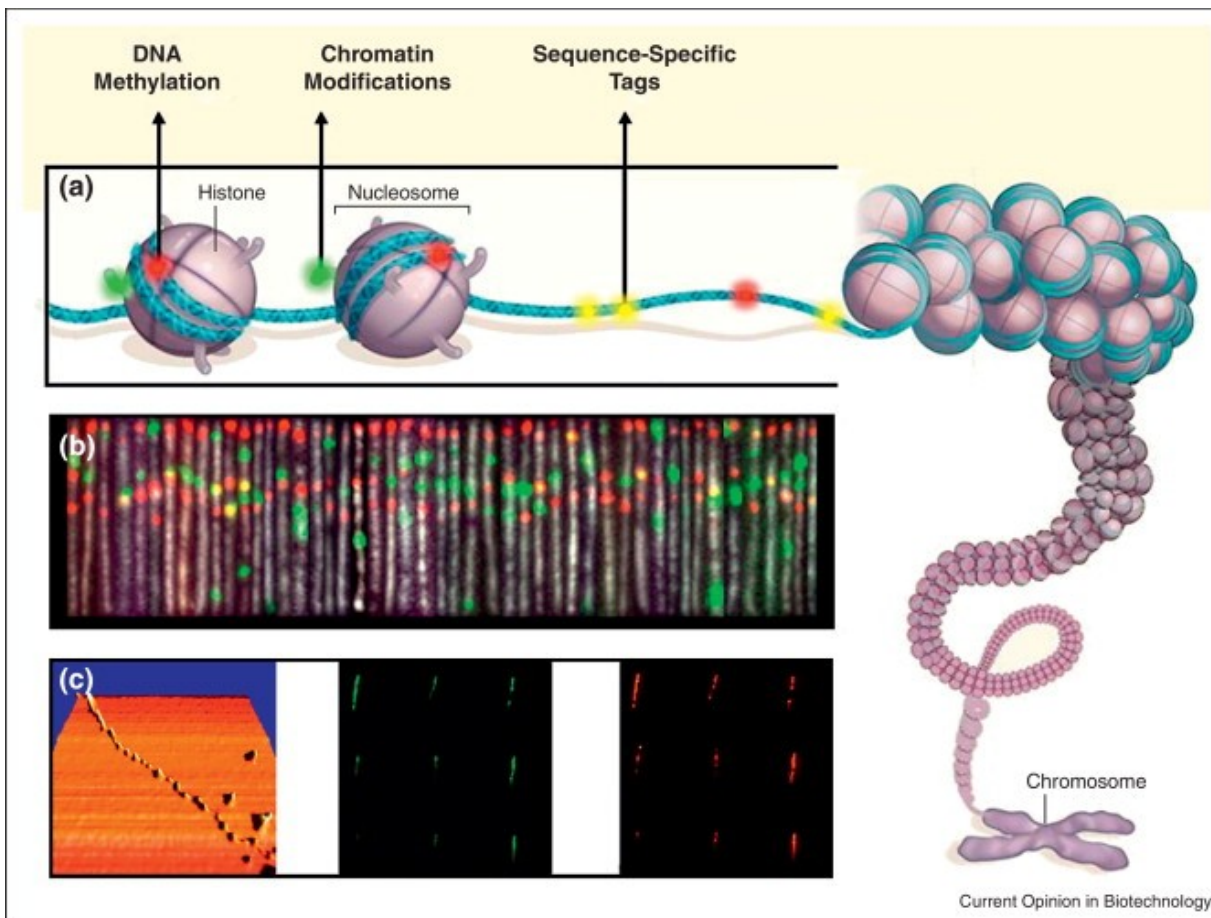**Beyond sequencing: optical mapping of DNA in the age of nanotechnology and nanoscopy**

Beyond the sequence — optical mapping of chromatin. **(a)** Schematic illustration of DNA labeling. DNA sequences, chromatin modifications and epigenetic data such as methylation can be detected by optical mapping. **(b)** Cropped color overlay images of T7 genomes (white) carrying sequence specific tags (red) and fluorescently labeled RNAP (green). Scale bar 3 µm. **(c)** Stretching of single chromatin fragments characterized by atomic force microscopy (AFM) (first panel) and fluorescence microscopy (second and third panels). Three-dimensional representation of the chromatin molecules reveals nucleosome distribution and positioning along the stretched chromatin molecule. Chromatin fragments were stained with YOYO-1 (green) and labeled with Alexa647-histone H3 probes (red).

Michal Levy-Sakin , Yuval Ebenstein

Beyond sequencing: optical mapping of DNA in the age of nanotechnology and nanoscopy

Nanopore DNA sequencing

This method is based on the readout of electrical signal occurring at nucleotides passing by alpha-hemolysin pores covalently bound with cyclodextrin. The DNA passing through the nanopore changes its ion current. This change is dependent on the shape, size and length of the DNA sequence. Each type of the nucleotide blocks the ion flow through the pore for a different period of time.
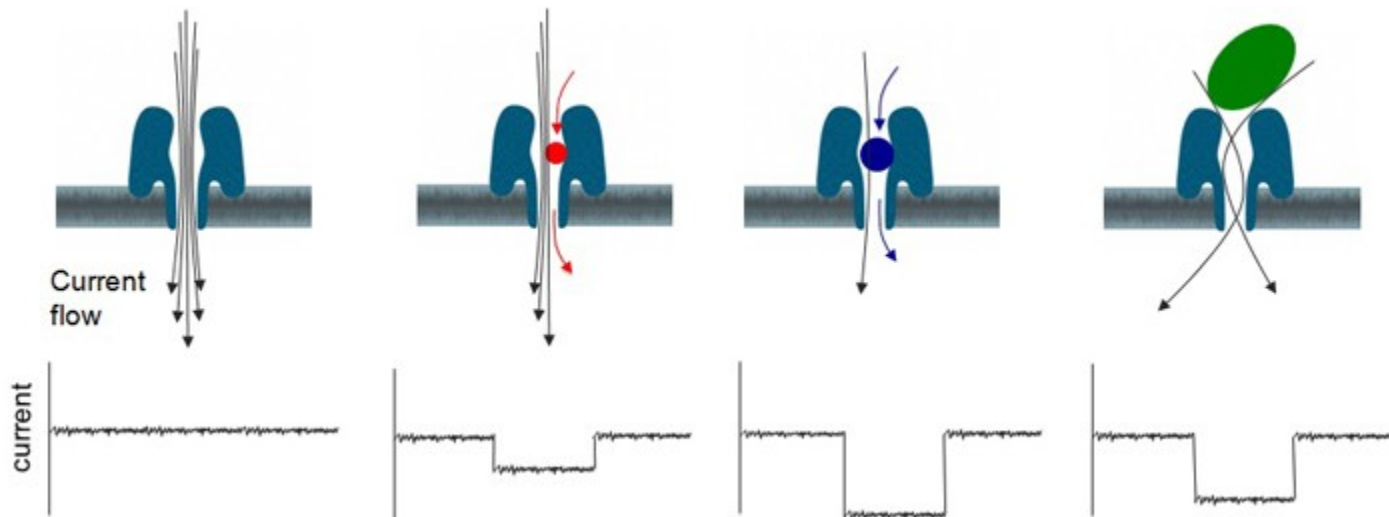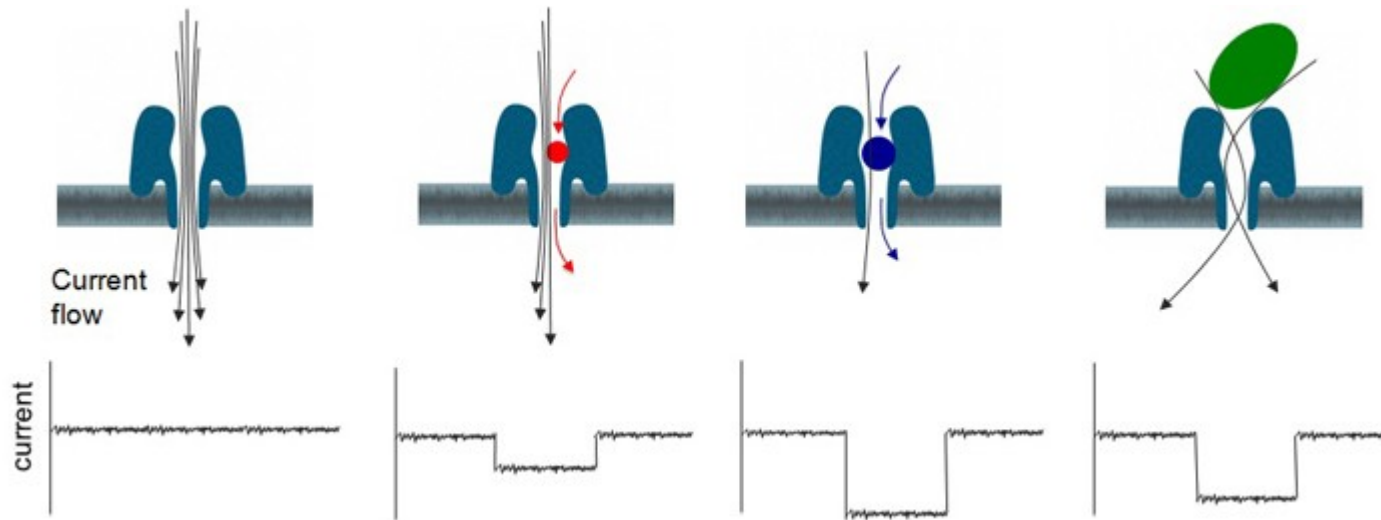
This diagram shows a protein nanopore set in an electrically resistant membrane bilayer. An ionic current is passed through the nanopore by setting a voltage across this membrane.

If an analyte passes through the pore or near its aperture, this event creates a **characteristic disruption in current**. By measuring that current it is possible to identify the molecule in question. For example, this system can be used to distinguish the four standard DNA bases and G, A, T and C, and also modified bases. It can be used to identify target proteins, small molecules, or to gain rich molecular information for example to distinguish the enantiomers of ibuprofen or molecular binding dynamics.

Strand sequencing

Exonuclease sequencing

A. Electronic Measurements

Ionic current   Tunneling current   Capacitive change

$I_{(ion)}, I_{(tunnel)}, V$

t

B. Optical Readout

Convert   Hybridize   Detect

G
A
T
C
T

G
A
T
C
T

Bulk (massively parallel)   Single Molecule
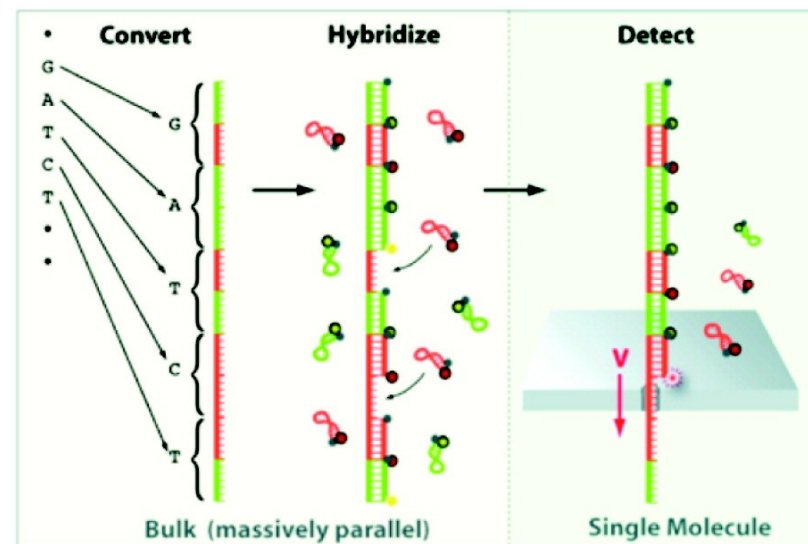
Nanopore DNA sequencing using electronic measurements and optical readout as detection methods. (A) In electronic nanopore schemes, signal is obtained through ionic current, tunneling current, and voltage difference measurements. Each method must produce a characteristic signal to differentiate the four DNA bases. (B) In the optical readout nanopore design, each nucleotide is converted to a preset oligonucleotide sequence and hybridized with labeled markers that are detected during translocation of the DNA fragment through the nanopore.
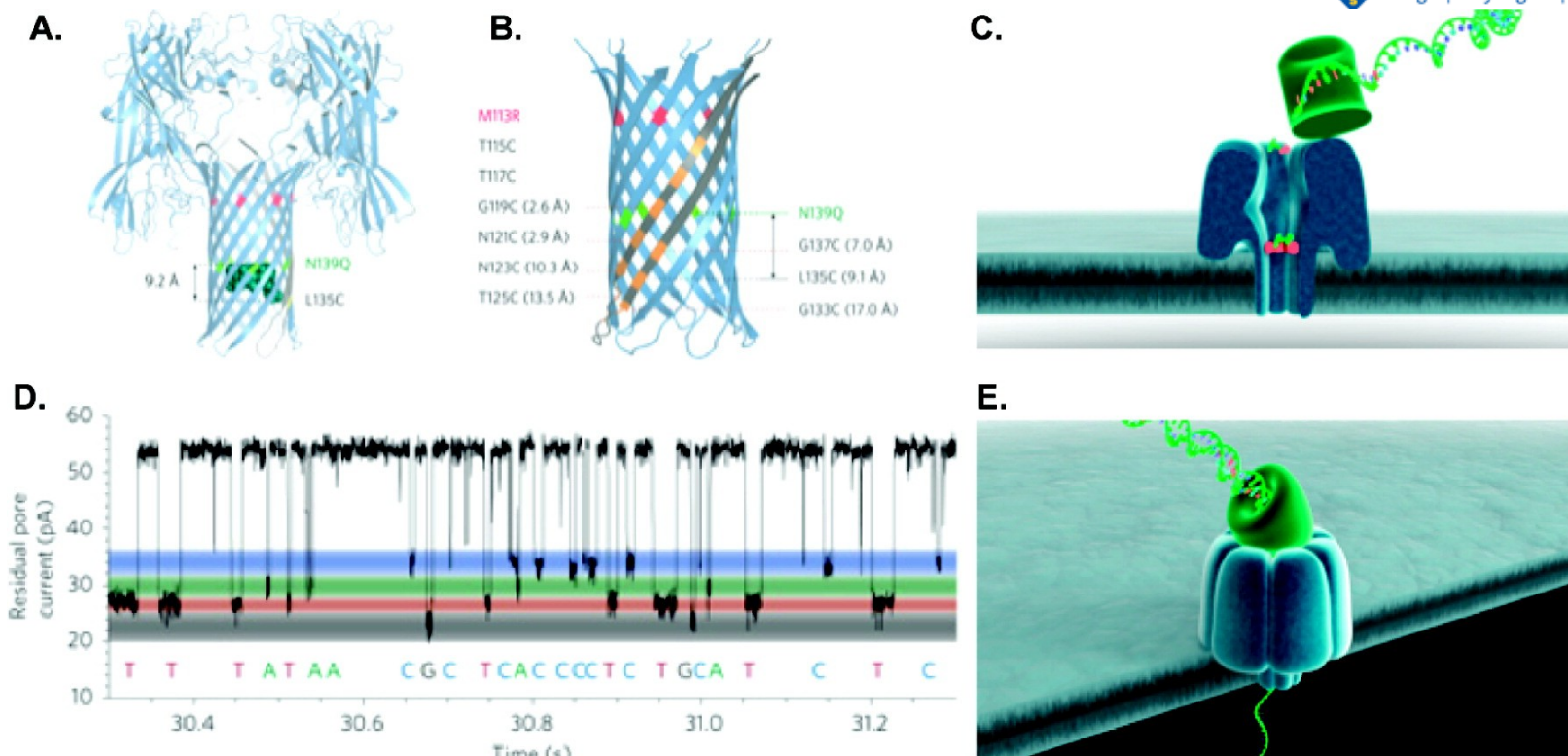
Biological nanopore scheme employed by Oxford Nanopore. (A) Schematic of αHL protein nanopore mutant depicting the positions of the cyclodextrin (at residue 135) and glutamines (at residue 139). (B) A detailed view of the β barrel of the mutant nanopore shows the locations of the arginines (at residue 113) and the cysteines. (C) **Exonuclease sequencing**: A processive enzyme is attached to the top of the nanopore to cleave single nucleotides from the target DNA strand and pass them through the nanopore. (D) A residual current-vs-time signal trace from an αHL protein nanopore that shows a clear discrimination between single bases (dGMP, dTMP, dAMP, and dCMP). (E) **Strand sequencing**: ssDNA is threaded through a protein nanopore and individual bases are identified, as the strand remains intact.

VisiGen Biotechnologies approach

VisiGen Biotechnologies introduced a specially **engineered DNA polymerase** for use in their sequencing. This **polymerase acts as a sensor** - **having incorporated a donor fluorescent dye by its active centre**. This donor dye acts by **FRET** (fluorescent resonance energy transfer), **inducing fluorescence of differently labeled nucleotides**. This approach allows reads performed at the speed at which polymerase incorporates nucleotides into the sequence (several hundred per second). The nucleotide fluorochrome is released after the incorporation into the DNA strand. The expected read lengths in this approach should reach 1000 nucleotides.

Future methods

Sequencing by hybridization is a non-enzymatic method that uses a DNA microarray. A single pool of DNA whose sequence is to be determined is fluorescently labeled and hybridized to an array containing known sequences. Strong hybridization signals from a given spot on the array identifies its sequence in the DNA being sequenced. Mass spectrometry may be used to determine mass differences between DNA fragments produced in chain-termination reactions. DNA sequencing methods currently under development include labeling the DNA polymerase, reading the sequence as a DNA strand transits through nanopores,and microscopy-based techniques, such as AFM or transmission electron microscopy that are used to identify the positions of individual nucleotides within long DNA fragments (>5,000 bp) by nucleotide labeling with heavier elements (e.g., halogens) for visual detection and recording. Third generation technologies aim to increase throughput and decrease the time to result and cost by eliminating the need for excessive reagents and harnessing the processivity of DNA polymerase.

In microfluidic Sanger sequencing the entire thermocycling amplification of DNA fragments as well as their separation by electrophoresis is done on a single glass wafer (approximately 10 cm in diameter) thus reducing the reagent usage as well as cost. In some instances researchers have shown that they can increase the throughput of conventional sequencing through the use of microchips. Research will still need to be done in order to make this use of technology effective. In October 2006, the X Prize Foundation established an initiative to promote the development of full genome sequencing technologies, called the Archon X Prize, intending to award $10 million to "the first Team that can build a device and use it to sequence 100 human genomes within 10 days or less, with an accuracy of no more than one error in every 100,000 bases sequenced, with sequences accurately covering at least 98% of the genome, and at a recurring cost of no more than $10,000 (US) per genome."

Each year NHGRI promotes grants for new research and developments in genomics. 2010 grants and 2011 candidates include continuing work in microfluidic, polony and base-heavy sequencing methodologies.

Major landmarks in DNA sequencing

1953 Discovery of the structure of the DNA double helix.

1972 Development of recombinant DNA technology, which permits isolation of defined fragments of DNA; prior to this, the only accessible samples for sequencing were from bacteriophage or virus DNA.

1977 The first complete DNA genome to be sequenced is that of bacteriophage φX174.

1977 Allan Maxam and Walter Gilbert publish "DNA sequencing by chemical degradation".
Frederick Sanger, independently, publishes "DNA sequencing with chain-terminating inhibitors".

1984 Medical Research Council scientists decipher the complete DNA sequence of the Epstein-Barr virus, 170 kb.

1986 Leroy E. Hood's laboratory at the California Institute of Technology and Smith announce the first semi-automated DNA sequencing machine.

1987 Applied Biosystems markets first automated sequencing machine, the model ABI 370.

1990 The U.S. National Institutes of Health (NIH) begins large-scale sequencing trials on *Mycoplasma capricolum*, *Escherichia coli*, *Caenorhabditis elegans*, and *Saccharomyces cerevisiae* (at US$0.75/base).

1991 Sequencing of human expressed sequence tags begins in Craig Venter's lab, an attempt to capture the coding fraction of the human genome.

1995 Craig Venter, Hamilton Smith, and colleagues at The Institute for Genomic Research (TIGR) publish the first complete genome of a free-living organism, the bacterium *Haemophilus influenzae*. The circular chromosome contains 1,830,137 bases and its publication in the journal *Science* marks the first use of whole-genome shotgun sequencing, eliminating the need for initial mapping efforts.

1996 Pål Nyrén and his student Mostafa Ronaghi at the Royal Institute of Technology in Stockholm publish their method of pyrosequencing

1998 Phil Green and Brent Ewing of the University of Washington publish "pyrex" for sequencer data analysis.

2000 Lynx Therapeutics publishes and markets "MPSS" - a parallelized, adapter/ligation-mediated, bead-based sequencing technology, launching "next-generation" sequencing.

2001 A draft sequence of the human genome is published.

2004 454 Life Science markets a parallelized version of pyrosequencing. The first version of their machine reduced sequencing costs 6-fold compared to automated Sanger sequencing, and was the second of a new generation of sequencing technologies, after MPSS.

Platform	Library size	Data output
GS FLX	450bp Shot-gun sequencing	1 million reads
GS FLX	2 X 175bp paired-end library(3, 8, 20Kb insert)	1.2 million reads
Illumina HiSeq2000	50bp Single-end sequencing	280-300 million reads
Illumina HiSeq2000	100bp Single-end sequencing	280-300 million reads
Illumina HiSeq2000	2 X 50 pair-end sequencing	300 million reads
Illumina HiSeq2000	2 X 100 pair-end sequencing	300 million reads

An overview of NGS platforms

Platform	Year	Sequencing Method	Amplification	Detection	Features
454	2005	Pyro-sequencing	Emulsion PCR	Light	First NGS
Illumina	2007	Synthesis	Bridge PCR	Light	90% of Market
SOLiD	2008	Ligation	Emulsion PCR	Light	Lowest Error Rate
Ion Torrent	2010	Synthesis	Emulsion PCR	Hydrogen Ion	Semiconductor Chip
Pacific Biosciences	2010	Synthesis	None = Single Molecule	Light	Anchored Polymerases
Oxford Nanopore	2012	Nanopore	None = Single Molecule	Electrical Conductivity	"Run Until" Sequencing

Modified from T. C. Glenn. 2011. Field guide to next-generation DNA sequencers. Molecular Ecology Resources 11: 759-769.