INTRODUCTION TO NEXT-GENERATION SEQUENCING

Dr Ben Dickins

Useful Reading

- Start here for a basic overview of the 454, SOLiD and Illumina technologies.
- For more advanced guides:
  - Quail, M. A., Smith, M., ... & Gu, Y. (2012). A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC genomics*, *13*(1), 341.
  - Chapter 2 of Kircher, M. (2011). Understanding and improving high-throughput sequencing data production and analysis. PhD Thesis, Universität Leipzig
- For a review of the key business decisions read this article.
- Up-to-date information on sequencing costs is here and look here for current global usage of NGS platforms.

The Development of Sequencing

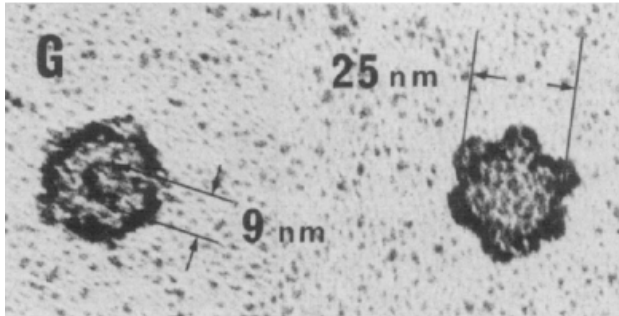
- 1970s – Sanger sequencing
- 1980s – Automated Sanger sequencing
- Continuous updates: e.g., ABI 3500 series
- Then:

Company	Technology	Platform
Illumina/Solexa	cyclic reversible termination	MiSeq, NextSeq, HiSeq
Roche/454	single nucleotide addition/pyrosequencing	GS Junior+, GS FLX+
Applied Biosystems (ABI)/SOLiD	sequencing by ligation	5500W Series
BGI-Shenzhen/Complete Genomics	sequencing by ligation	Revolocity
Ion Torrent	real-time sequencng (pH)	Ion S5, Ion PGM, Ion Proton
Pacific Biosciences (PacBio)	real time sequencing (SMRT)	RSII, Sequel
Oxford Nanopore Technologies	real time sequencing (pore-based)	MinION, PromethION, GridION

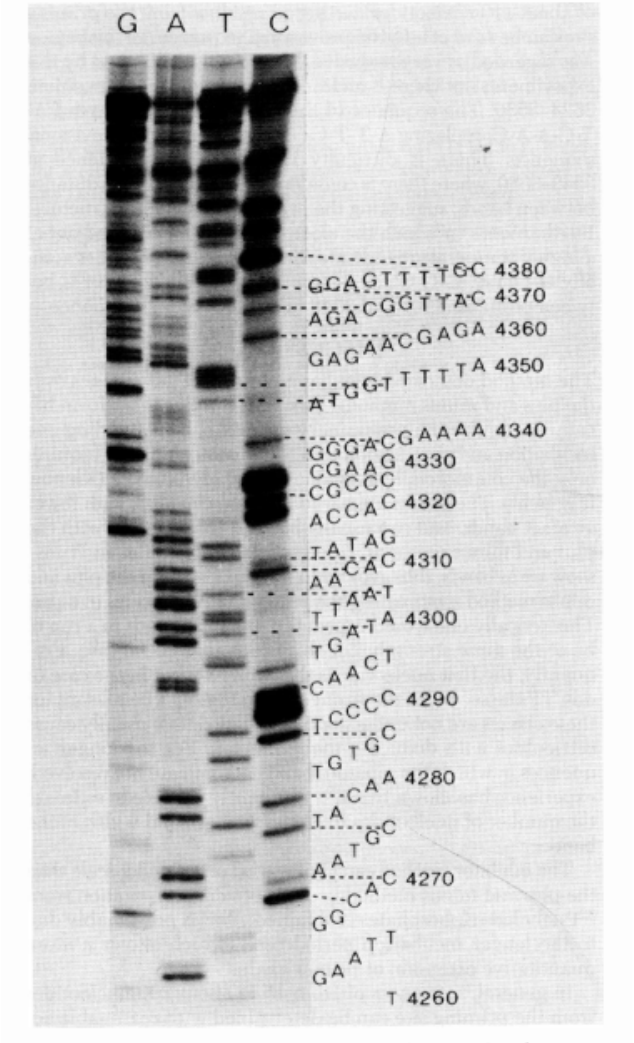
Sequencing Firsts

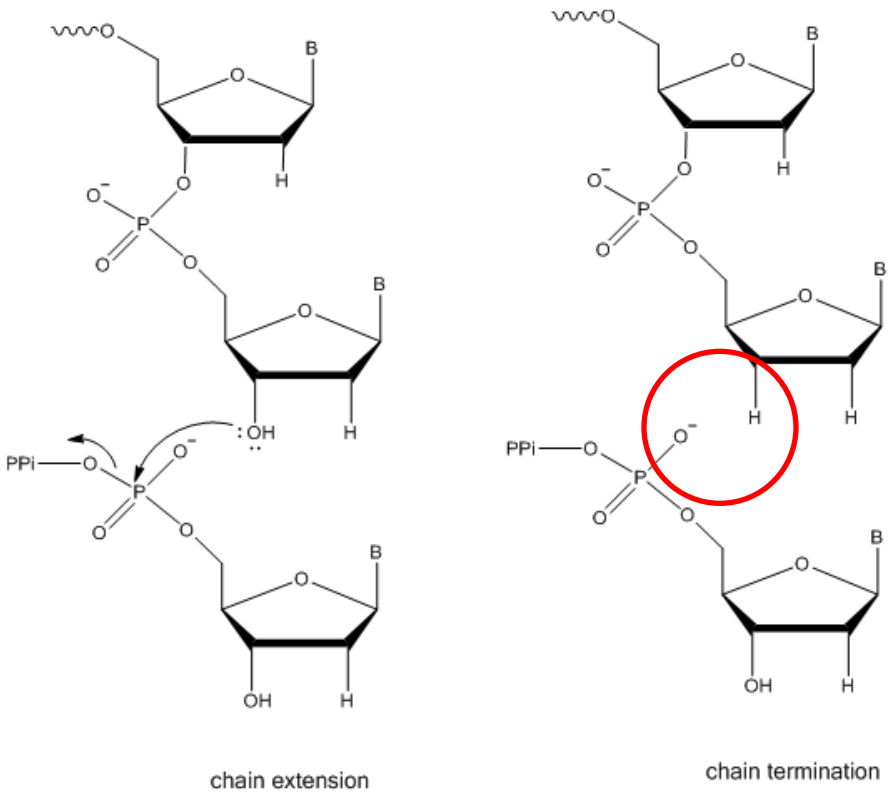
- Early sequencing was painstaking, e.g., 24 bp by "wandering spot" analysis (Gilbert and Maxam, 1973).
- 1st genome sequenced = MS2: ~3.6kb RNA (Fiers et al., 1976).

- Then an enzymatic method was developed by Fred Sanger and colleagues in the 1970s now known as "dideoxy", "chain termination" or "Sanger" sequencing
- 1st DNA genome sequenced = ΦX174: ~5.4kb ssDNA (Sanger et al., 1977, 1978).



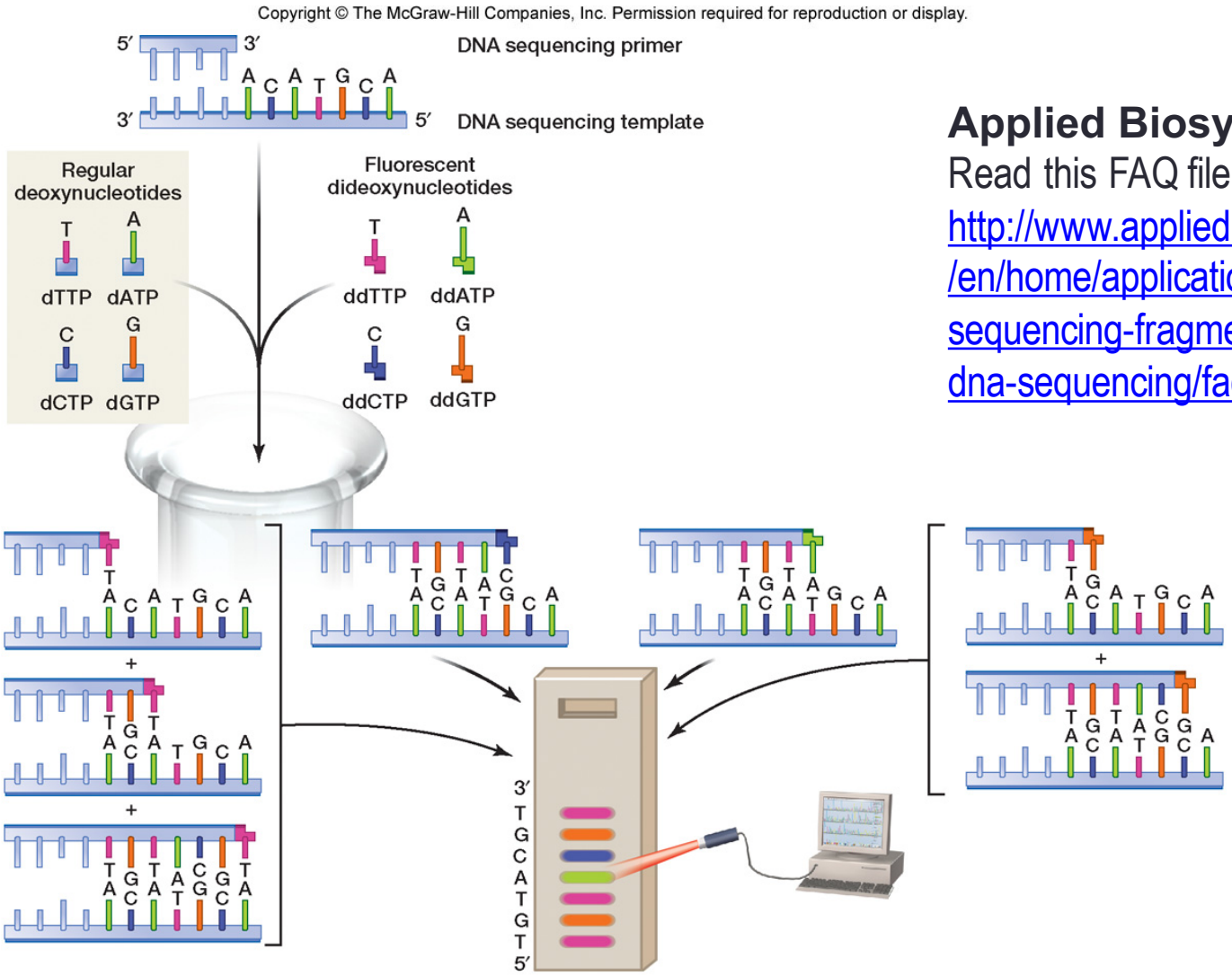
Chain termination





Source of image and good place to read a brief description: http://biologicalchemistry.tumblr.com/post/86350473870/sanger-dideoxy-dna-sequencing-the-first-enzymatic

Automated DNA sequencing



Applied Biosystems, 1986

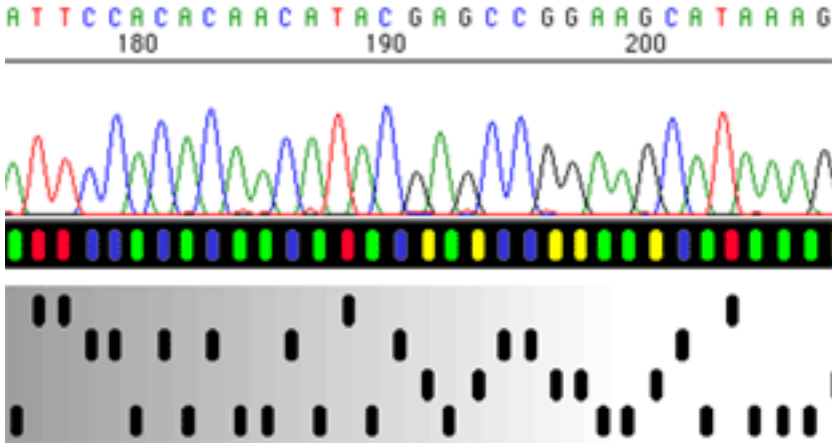
Read this FAQ file #1-6:

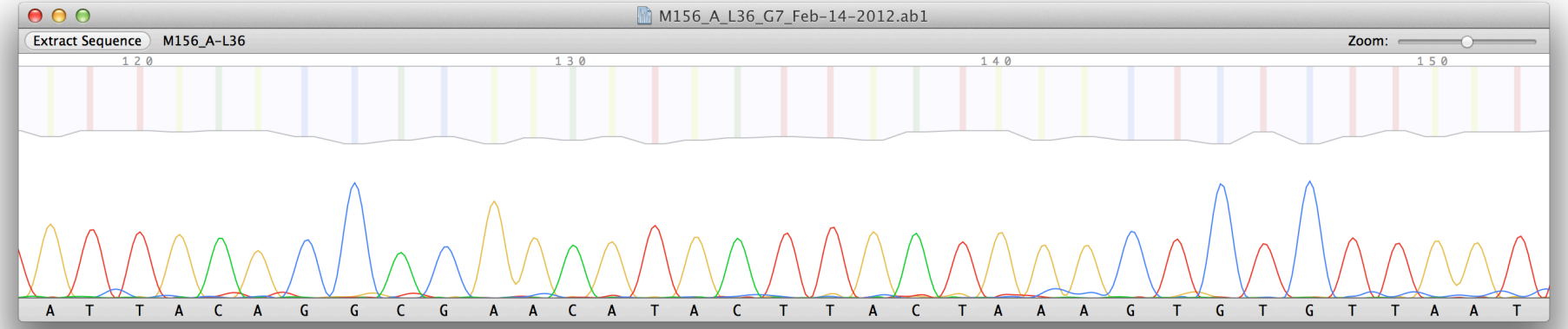
http://www.appliedbiosystems.com/absite/us/en/home/applications-technologies/dna-sequencing-fragment-analysis/overview-of-dna-sequencing/faq.html

Automated DNA sequencing

The benefits of cycle sequencing:

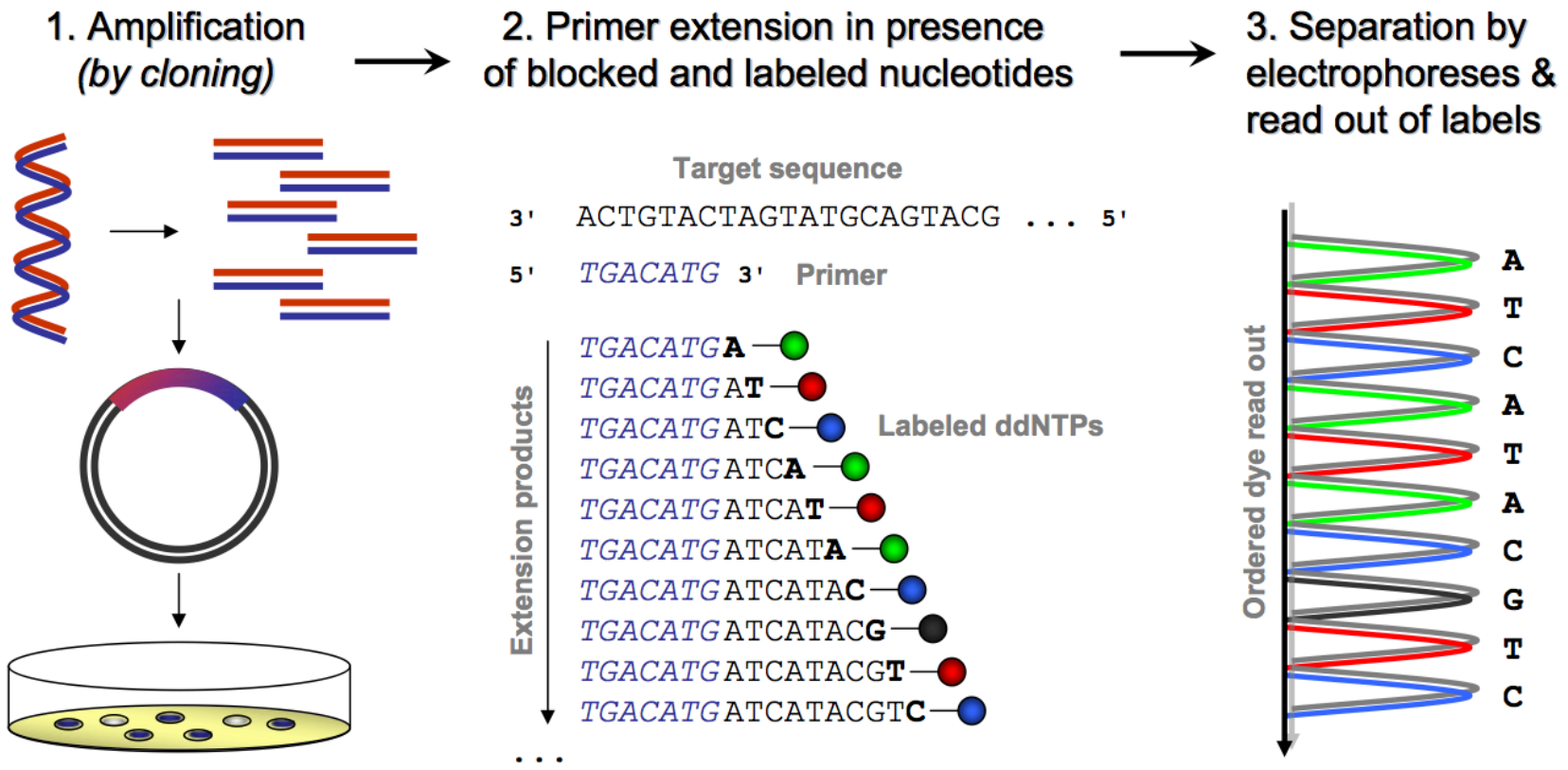
(and use of capillary)





Some DNA I sequenced!

Sanger Workflow



Sanger Developments

- The costs of Sanger sequencing have been declining (see this reference although note this is slightly dated).
- And throughput has been increasing: current capillary technology can process 96 well plates of samples (and Applied Biosystem's 3500xL analyzer has 24 capillaries).
- But to sequence we need:
  - to know some sequence (for the primer),
  - to clone the DNA fragments first (BAC clones for genome).

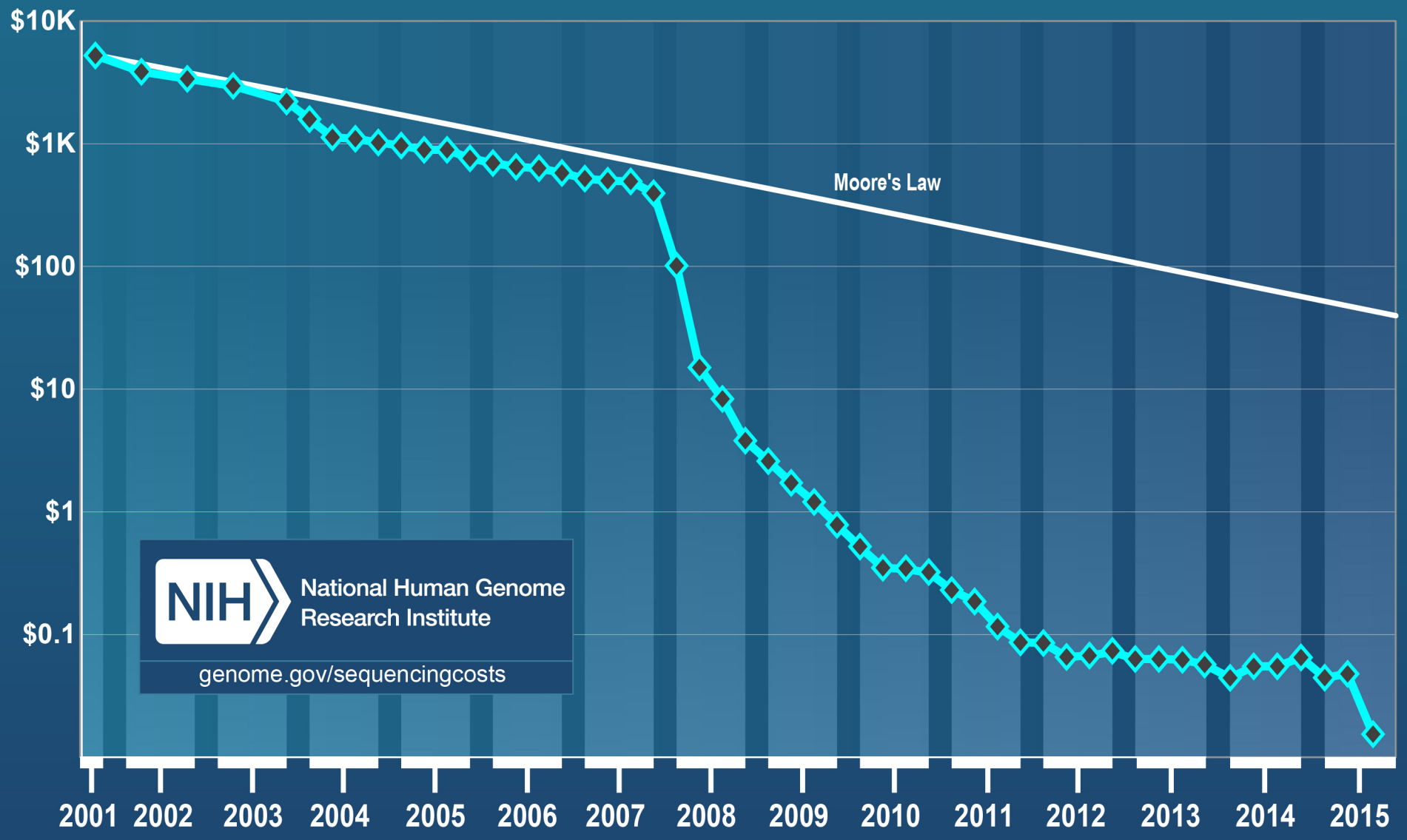
Advances in DNA sequencing

- Throughput is limited. More machines are required for large genomes.
- Several months/years to sequence a genome (bacterial/eukaryotic)
- US$10 million to US$25 million to sequence a single human genome and $20,000–$50,000 to sequence a microbial genome.
- Next-Generation Sequencing (NGS) or Massively Parallel Sequencing Technologies (MPST) have since been developed.
- There are two major methods:
  - short read sequencing
  - single-molecule sequencing

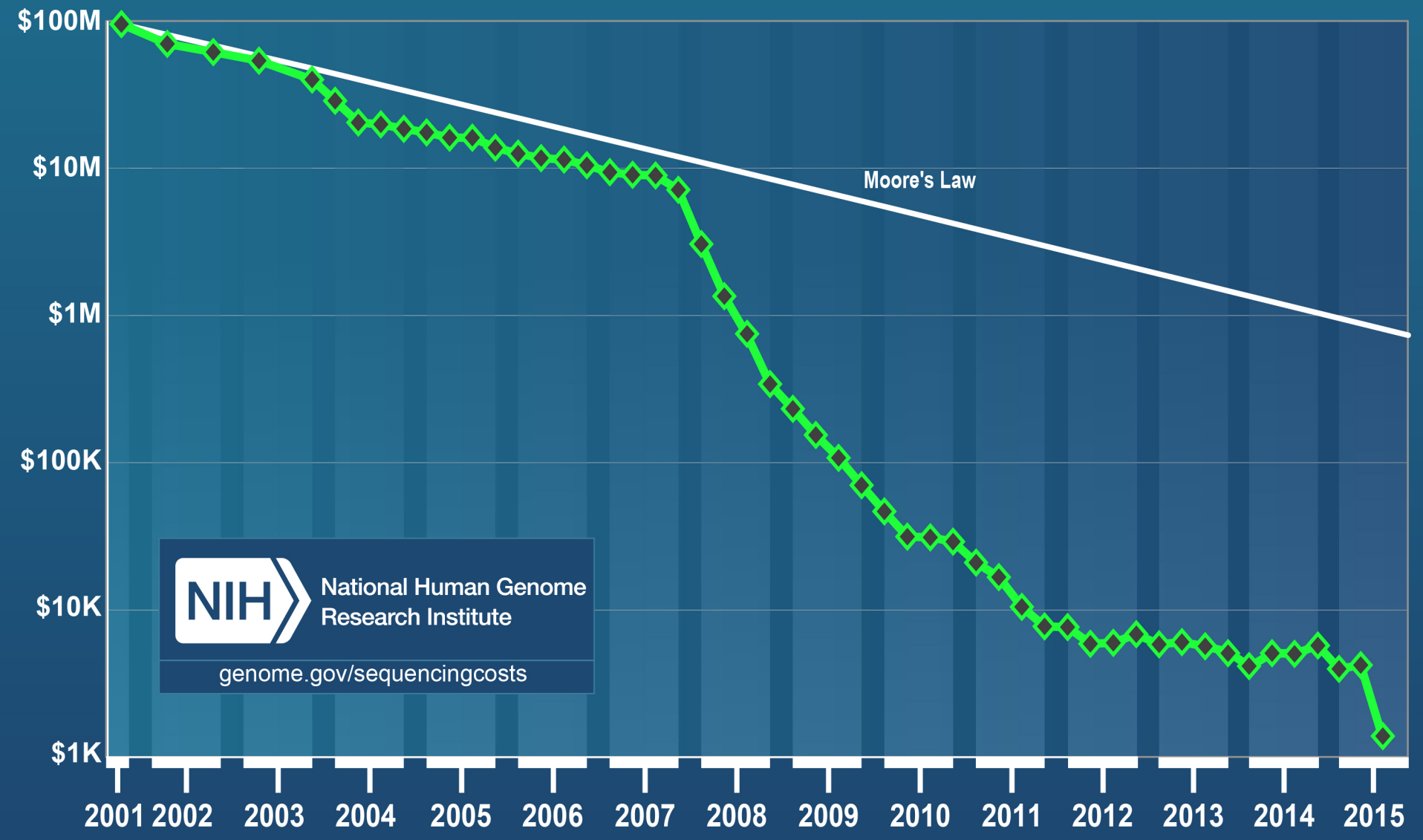
Uses of NGS

- Genome sequencing
- Exome sequencing
- Metagenomics
- Strain re-sequencing/detection
- Transcriptomics (RNA-Seq)
- Methylation detection
- Chromatin analysis (ChIP-Seq)

Cost per Raw Megabase of DNA Sequence



Cost per Genome



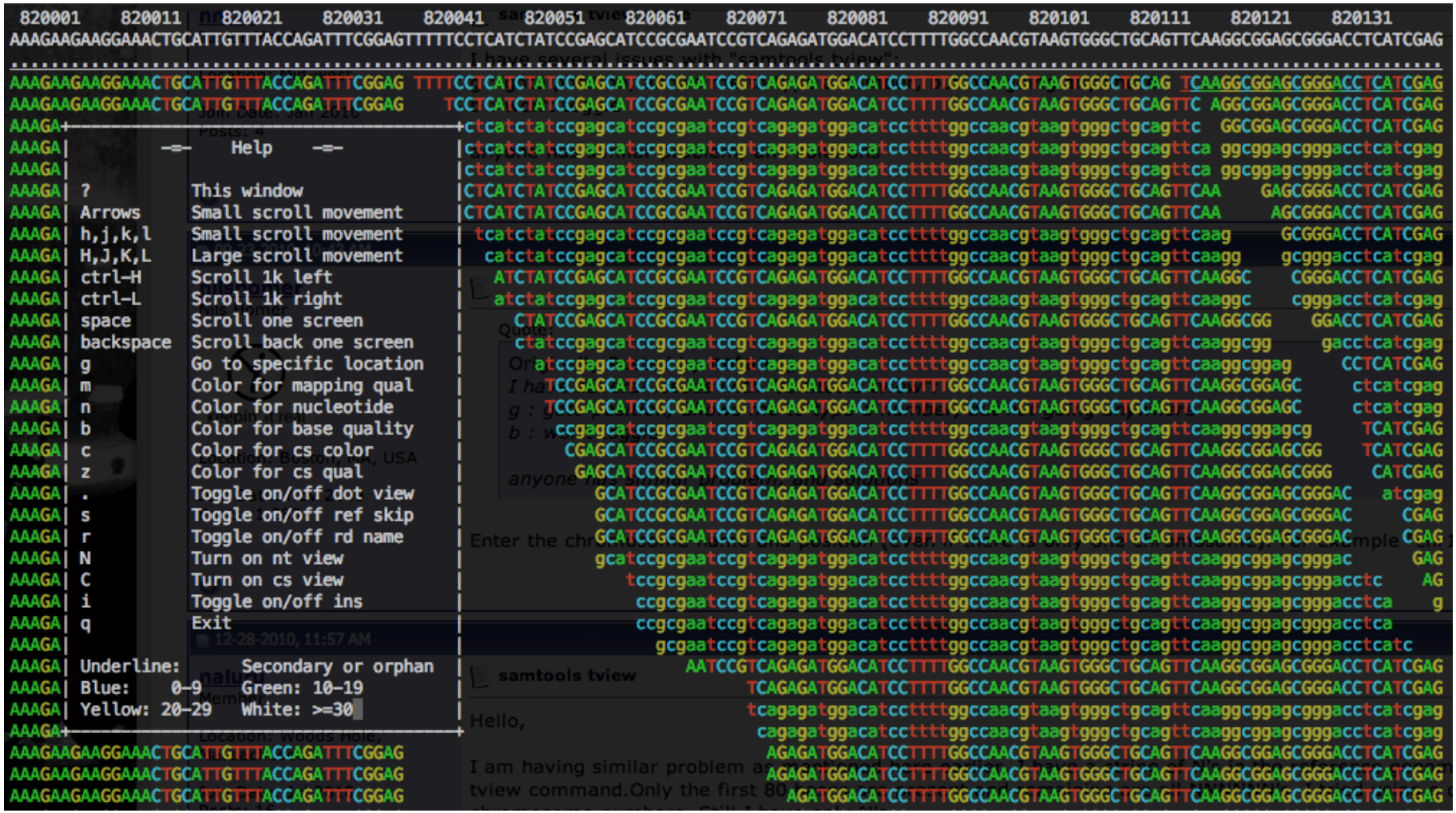
Developments in this field are rapid and new technologies are emerging constantly.

Illumina = ILMN is a large company on the NASDAQ. Market cap: 22.99B [checked: 19/10/2014].

NGS Overview

- Number of steps involved in all NGS:
  - template preparation
  - sequencing and imaging
  - data analysis
- The last step has been enabled by rapid improvements in computer technology.
- Short sequence "reads" must be assembled to infer original sequence or aligned/mapped against a reference.
- This will be discussed in later lectures.

Short Reads



SAMtools tview

Next Generation Sequencing

- Single nucleotide addition (SNA)
- Real time sequencing
- Cyclic reversible termination (CRT)
- Sequencing by ligation (SBL)

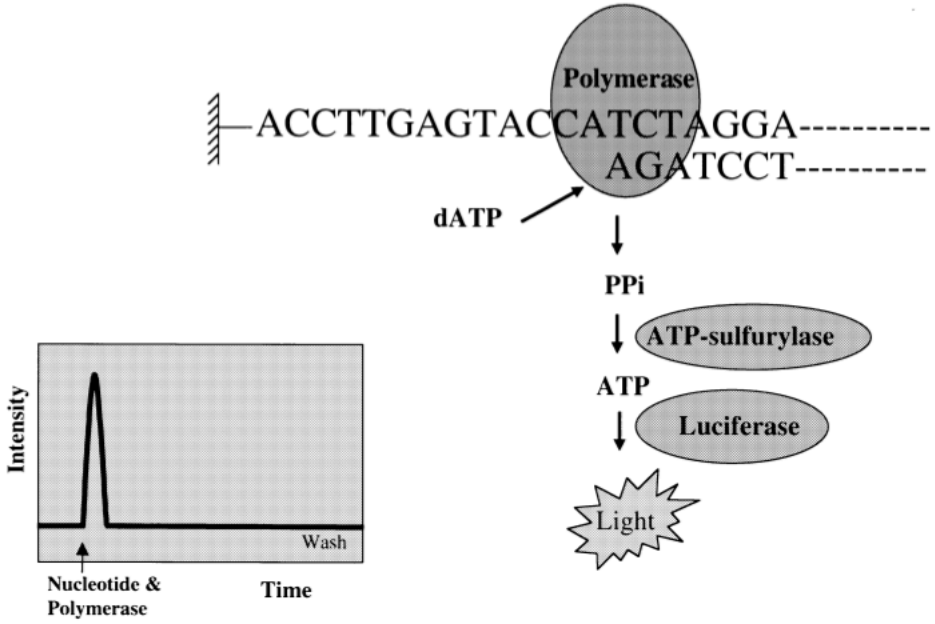
Next Generation Sequencing

- Single nucleotide addition (SNA)

➔Pyrosequencing

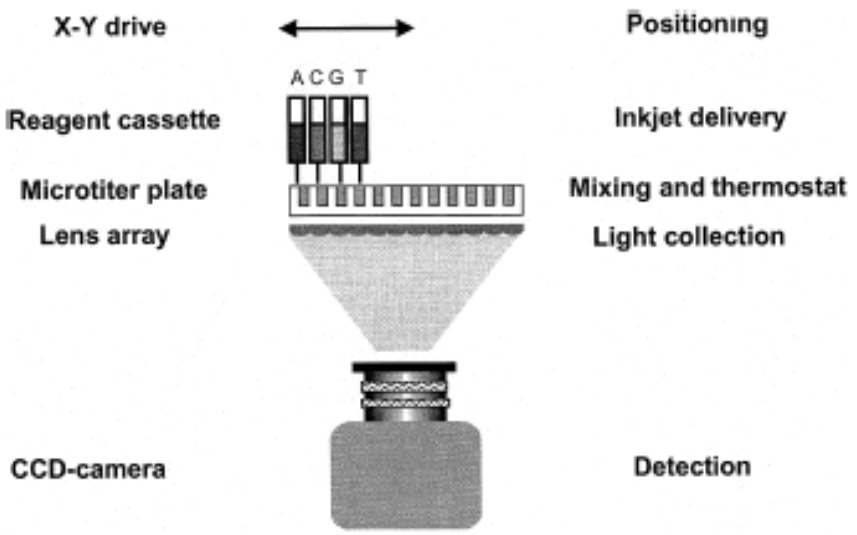
Pyrosequencing

- Theory (Melamede, 1985), practice (Hyman, 1988).
- Developed by Nyrén and Ronaghi (1996)
- DNA sequencing technique based on the detection of released pyrophosphate (PPi) during DNA synthesis.
- Cascade of enzymatic reactions, visible light is generated that is proportional to the number of incorporated nucleotides.

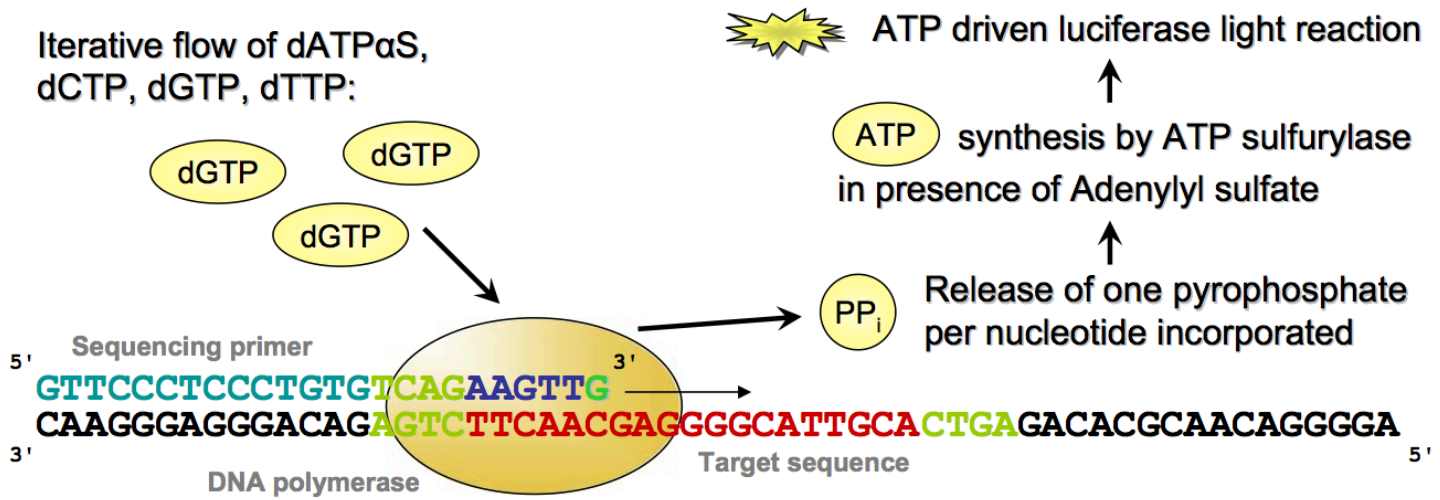


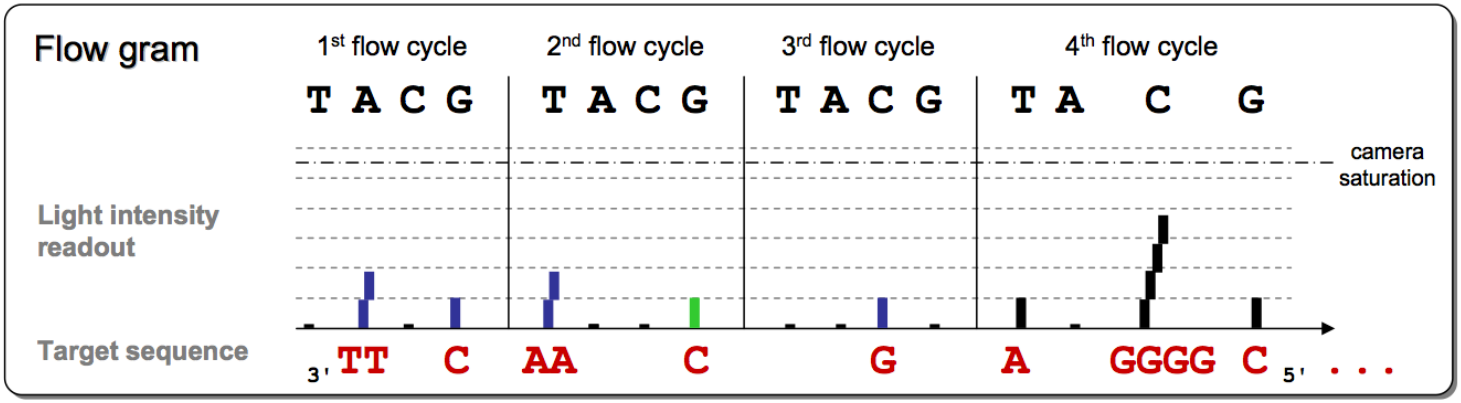
Pyrosequencing – some details

- Overall reaction from polymerization to light detection takes place within 3–4 sec at room temperature.
- One pmol of DNA in a pyrosequencing reaction yields $6x10^{11}$ ATP molecules which, in turn, generate more than $6x10^{9}$ photons at a wavelength of 560 nm
- This amount of light is easily detected by a photodiode, photomultiplier tube, or a charge-coupled device camera (CCD) camera
- Liquid phase or solid phase pyrosequencing is possible.

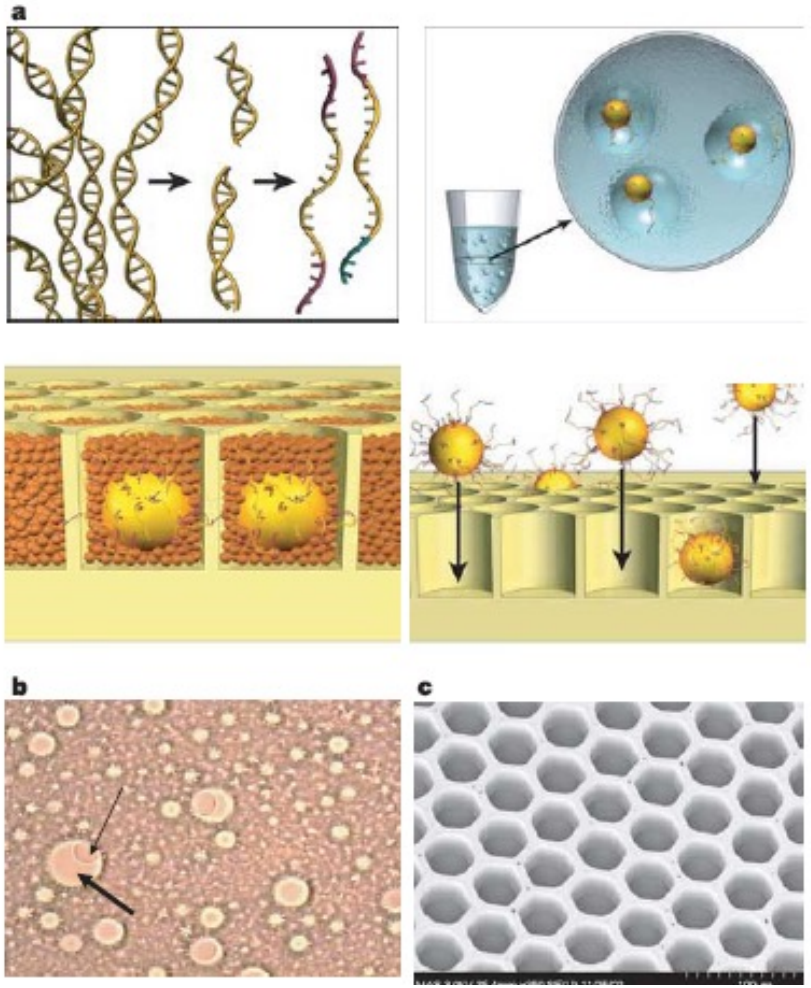


Pyrosequencing Workflow



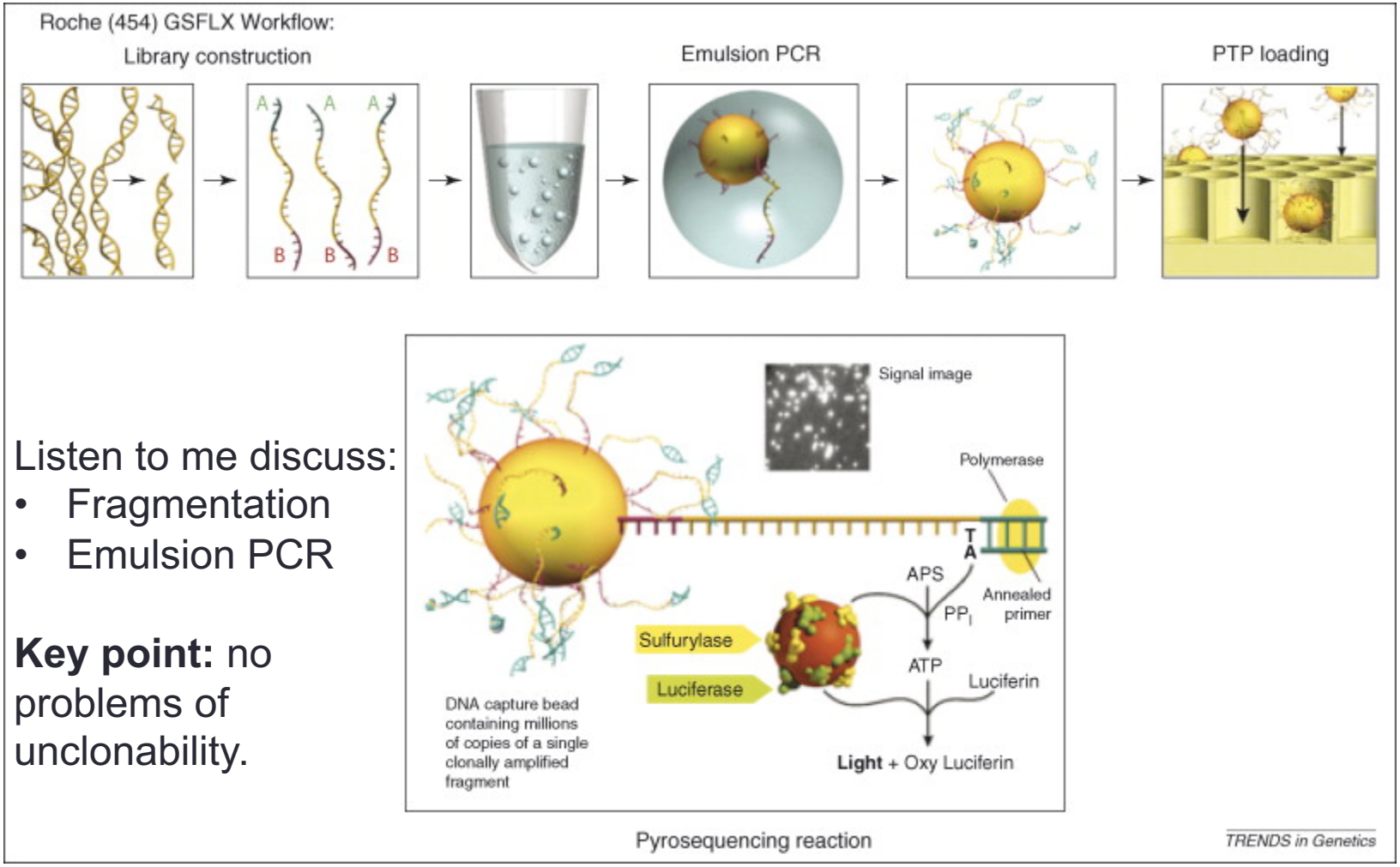


454: Genome sequencing in picolitre plates



- 454 later purchased by Roche Diagnostics
- Marguiles et al (2005): emulsion method for DNA amplification and an instrument for sequencing by synthesis.
- Pyrosequencing protocol optimized for solid support and picolitre-scale volumes.
- Novel fibre-optic slide of individual wells.

454 Workflow

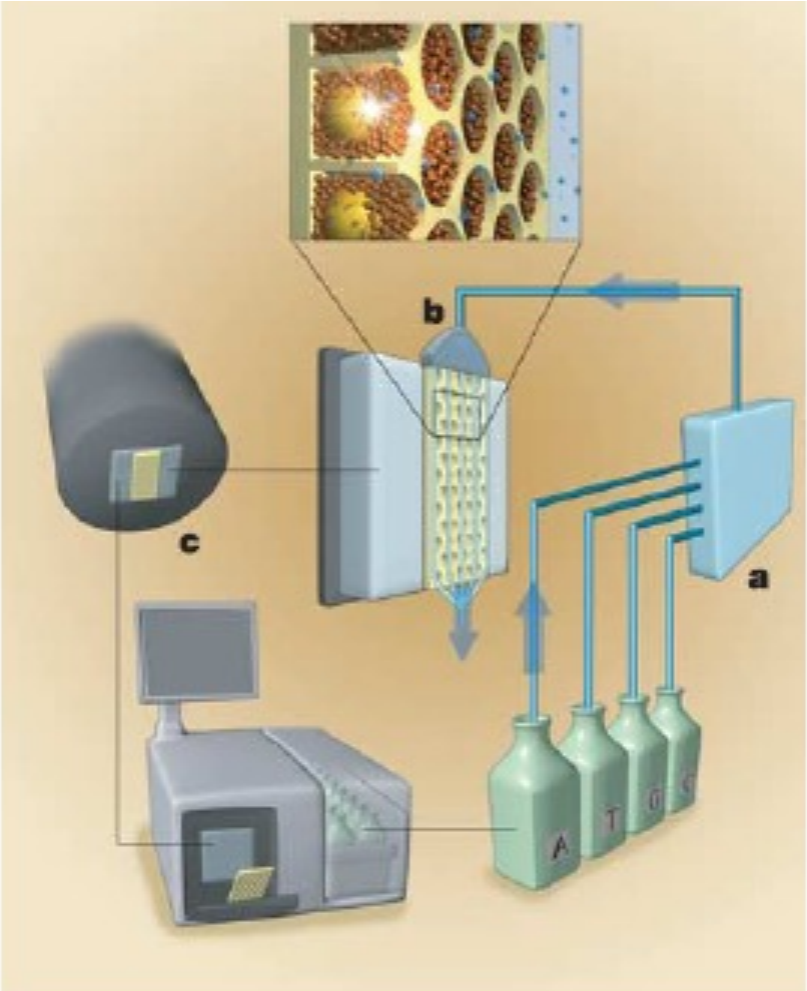


Listen to me discuss:

- Fragmentation
- Emulsion PCR

Key point: no problems of unclonability.

454 Instrumentation

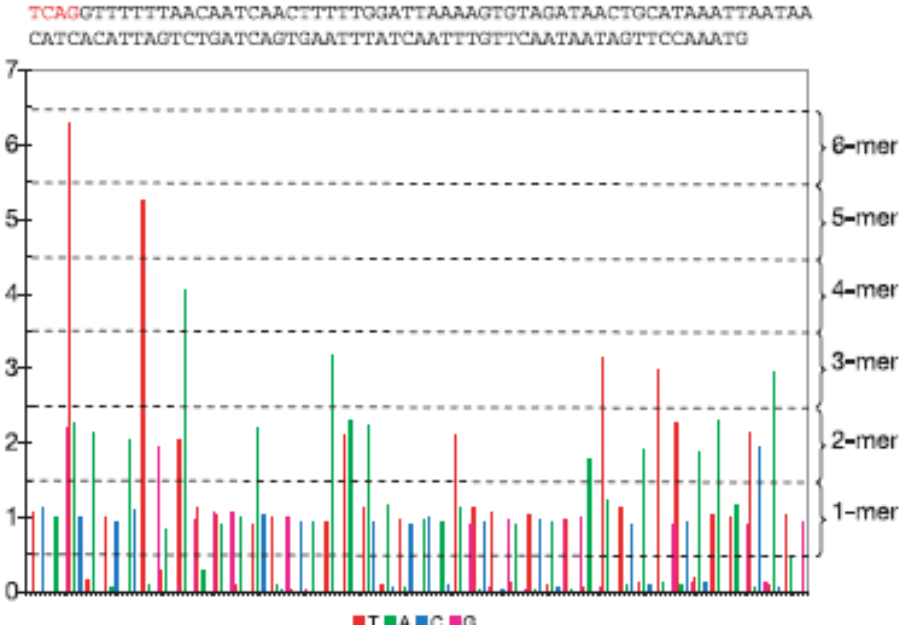


Sequencing instrument consists of following major subsystems:

- a fluidic assembly (a), a flow chamber that includes the well-containing fibre-optic slide
- (b) a CCD camera-based imaging assembly
- (c) a computer that provides the necessary user interface and instrument control

454 Applications

- Margulies et al (2005): Shotgun sequencing and *de novo* assembly of *Mycoplasma genitalium* genome with 96% coverage at 99.96% accuracy in one run of a machine. Machine could sequence 25M bases in 4 hours.
- Previously: "A total of 9846 sequencing reactions were performed by five individuals using an average of eight AB 373 DNA sequencers per day for a total of 8 weeks." (Fraser et al., 1995).



Current Systems

- 454 reads are relatively long (~700bp). Able to sequence 700M bases in a day.
- Among the lowest error rate ~10^{-3} per base.
- Homopolymer (repeats of same base) problems.
- Current platforms from Roche:
  - GS Junior+ (benchtop)
  - GS FLX+.

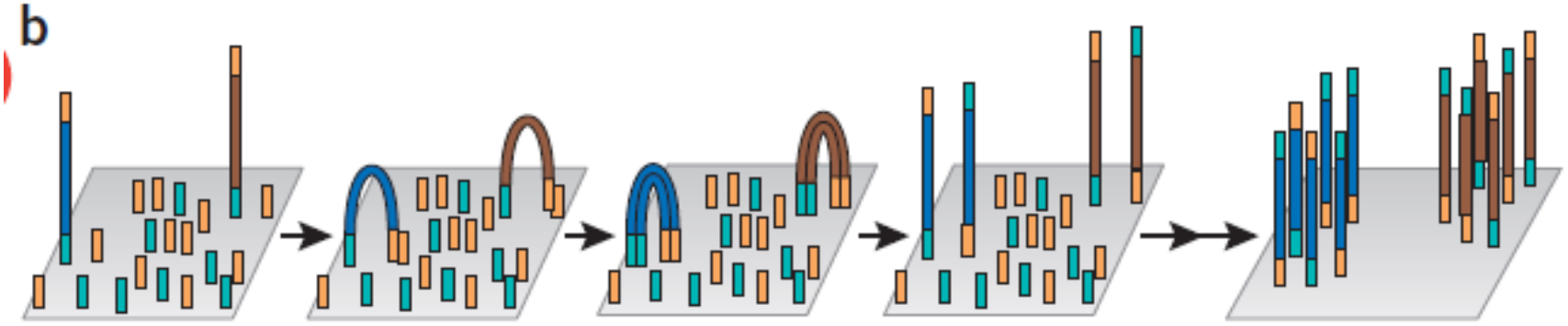
- More up to date information: http://www.454.com

Next Generation sequencing

- Cyclic reversible termination (CRT)
- nucleotide incorporation, fluorescence imaging and cleavage

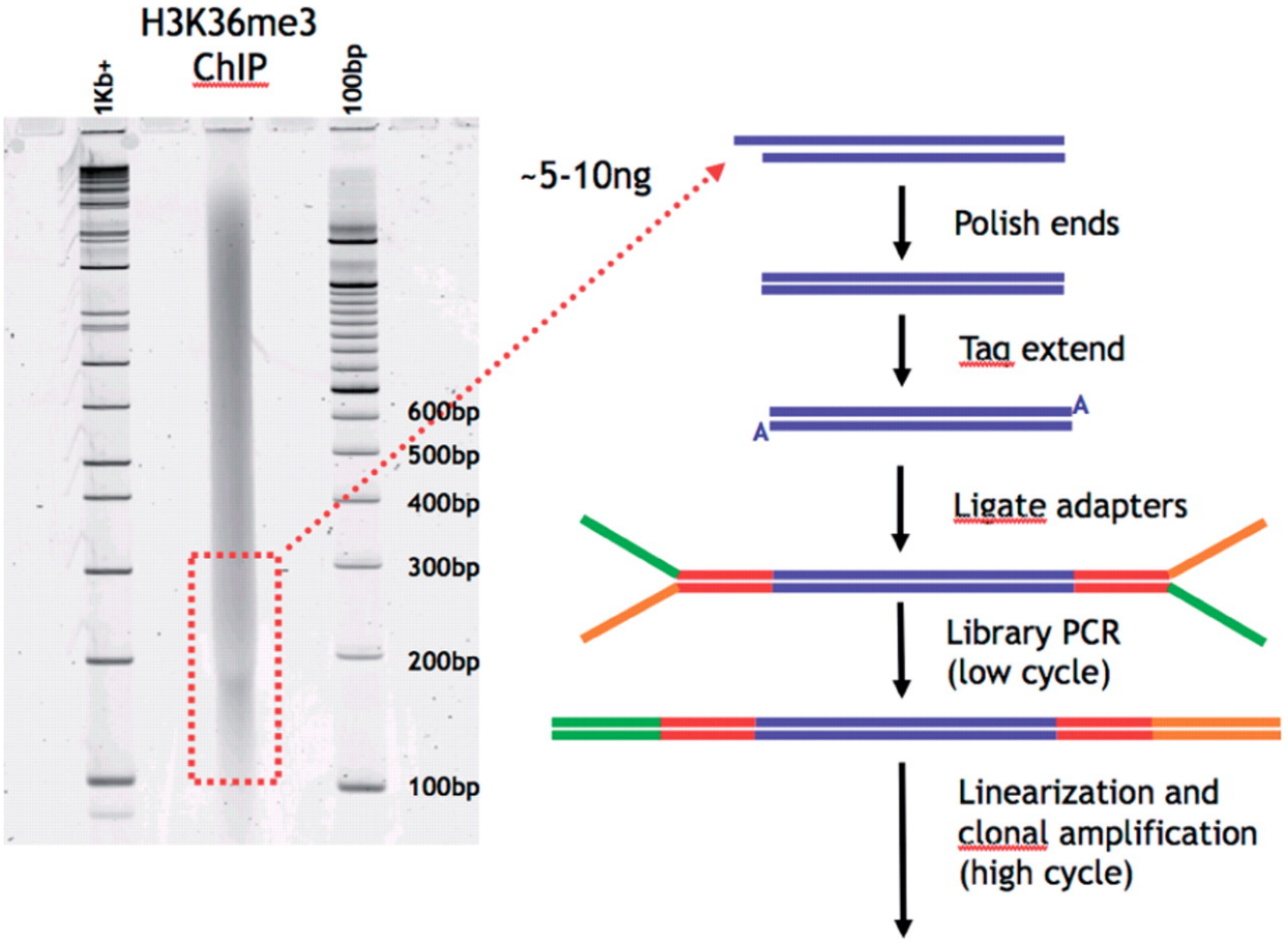
Illumina

- Sequencing by synthesis. Solexa introduced in 2007. Taken over by Illumina Inc.
- Some amplification during library preparation (usually). Then on a plate by "bridge" or "fold-back" PCR.
- Both primers densely coat the surface of a solid substrate, attached at their 5′ ends by a flexible linker.
- Amplification products originating from any given member of the template library remain locally tethered near the point of origin.
- Clonal cluster containing ~1,000 copies of a single member of the template library



Overview of library construction.

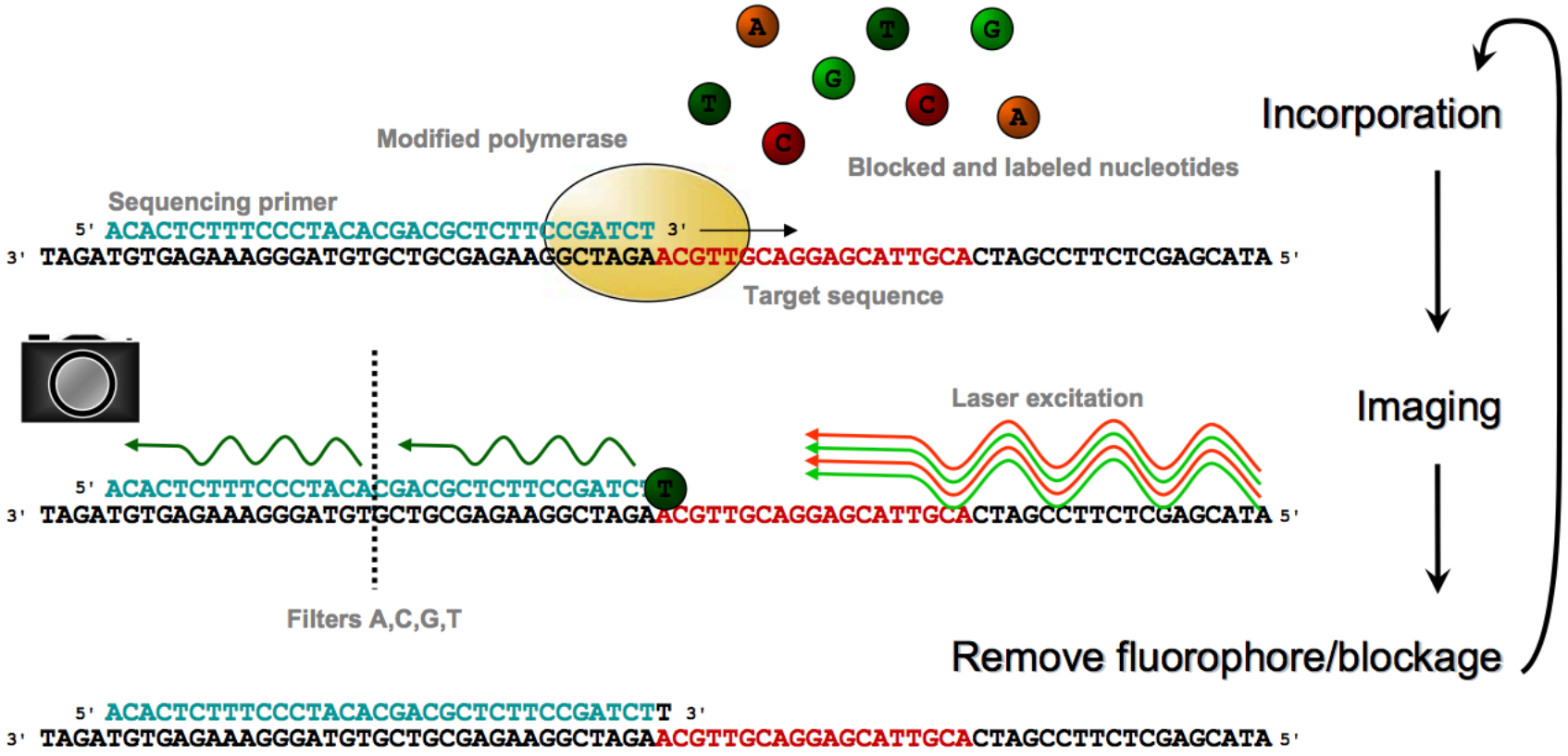
Fragmentation* followed by distinct 5'/3' (forked) adaptors



Hirst M , and Marra M A Briefings in Functional Genomics 2010;9:455-465

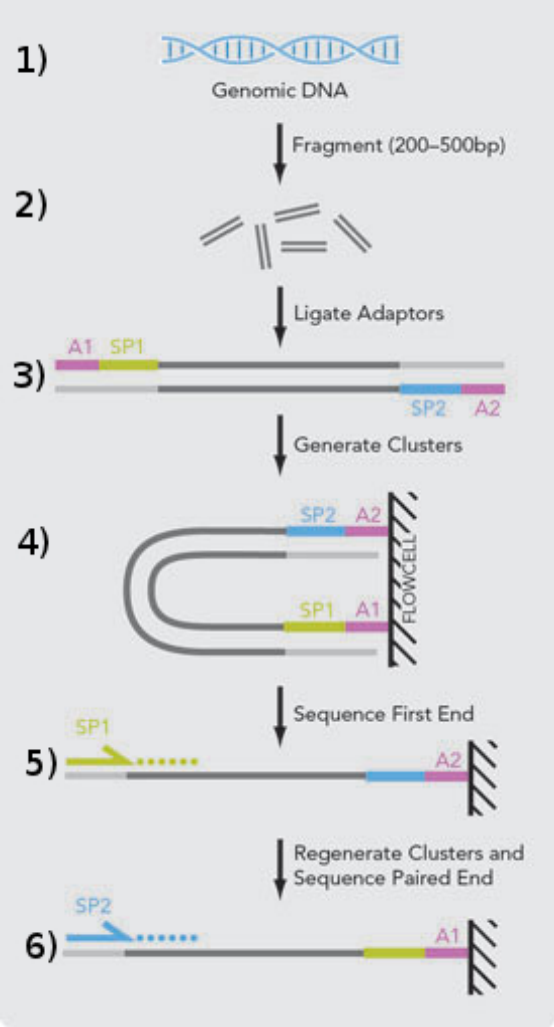
Briefings in Functional Genomics

Illumina Workflow



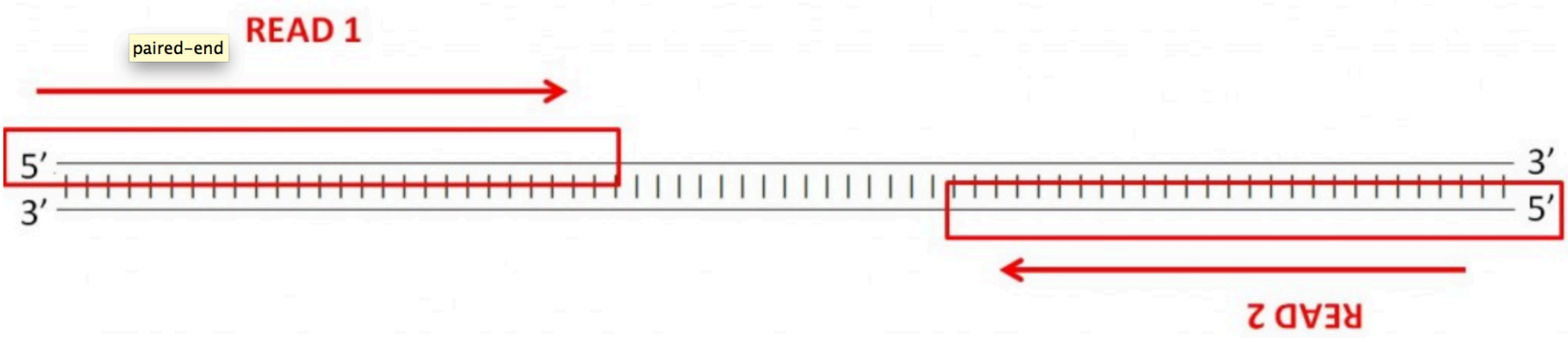
Paired-end (PE) and Mate-pair (MP) Sequencing

- PE sequencing is now common>
- Process:
  - remove synthesized strand (NaOH)
  - repeat bridge amplification
  - target base modifications of flow cells oligos
- MP a little different – circularize fragments during library prep.



PE Sequencing

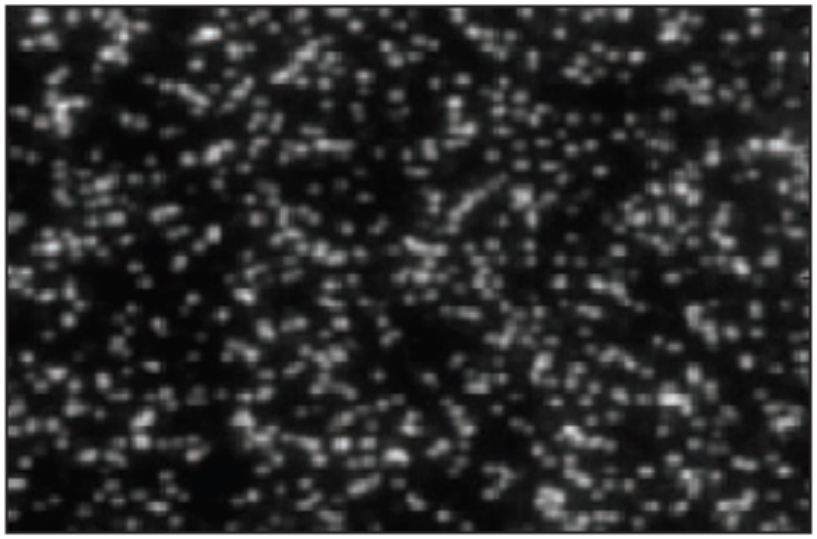
- Critical function: gives information on sequence at either end of fragment = can help in assembly.

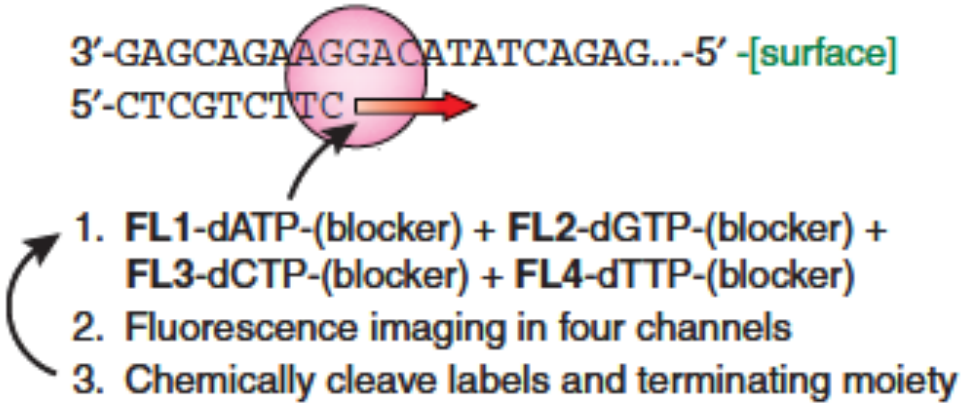


- In MP sequencing the fragments have opposite orientations.

Illumina Raw Data

- Each sequencing cycle includes the simultaneous addition of a mixture of four modified deoxynucleotide species, each bearing one of four fluorescent labels and a reversibly terminating moiety at the 3′ hydroxyl position.
- A modified DNA polymerase drives synchronous extension of primed sequencing features. This is followed by imaging in four channels and then cleavage of both the fluorescent labels and the terminating moiety
- Illumina





What you get back from most sequencers

FASTQ = common to most NGS platforms:

```
@HWI-EAS209_0006_FC706VJ:5:58:5894:21141#ATCACG/1
TTAATTGGTAAATAAATCTCCTAATAGCTTAGATNTTACCTTNNNNNNNNNNTAGTTTCTTGAGATT
TGTTGGGGGAGACATTTTTGTGATTGCCTTGAT
+HWI-EAS209_0006_FC706VJ:5:58:5894:21141#ATCACG/1
efcfffffcfeefffcffffffddf`feed]`]_Ba_^__[YBBBBBBBBBBRTT\]][]dddd`dd
d^dddadd^BBBBBBBBBBBBBBBBBBBBBB
```

$$Q = -10 \log_{10} P$$

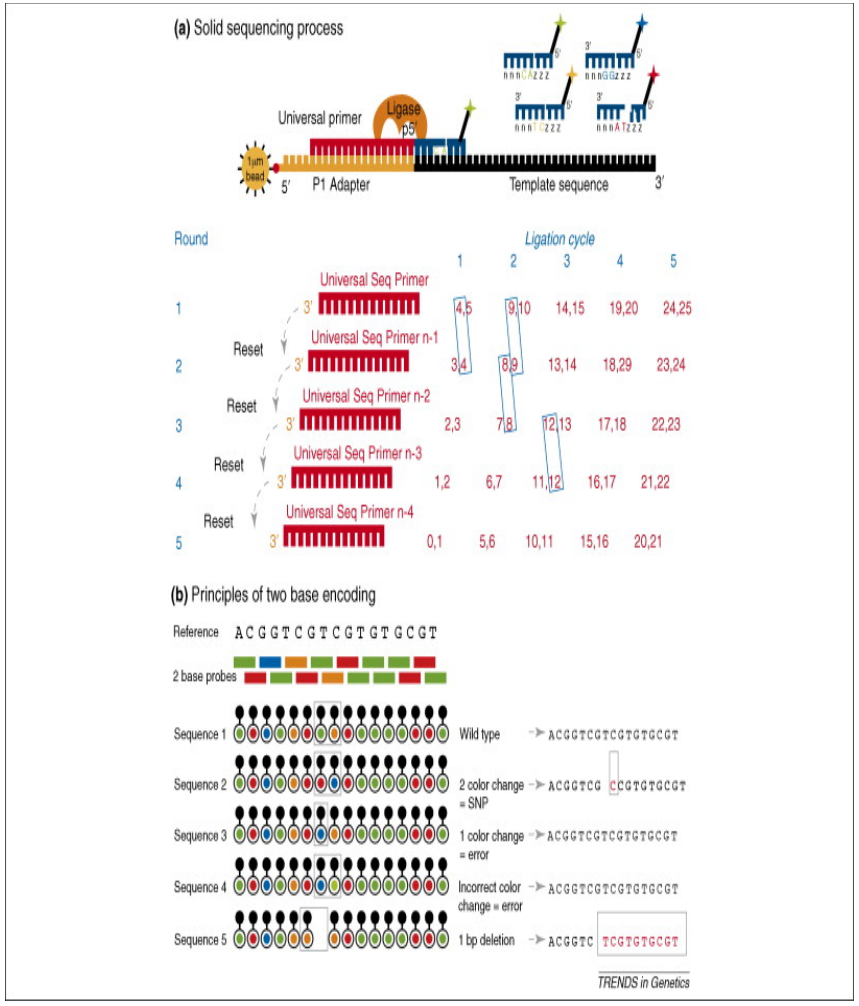
The first line is the sequence (including ambiguous characters), the second line is the quality (Q, hex encoded). P = probability that a base is miscalled. In PE sequencing you get two files or an interspersed file (/2 after read ID). Check out Wikipedia for description: http://en.wikipedia.org/wiki/FASTQ_format (accessed 19/10/2014]

Current Systems

- Reasonably accurate: ~0.1% per base.
- Phasing problems.
- Current platforms from Illumina:
  - MiSeq (benchtop)
  - HiSeq 2500
- Up to date information: http://www.illumina.com

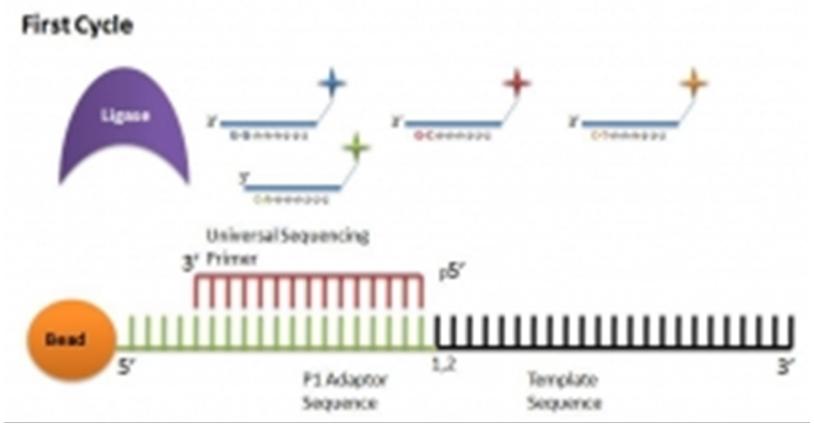
SOLiD sequencer

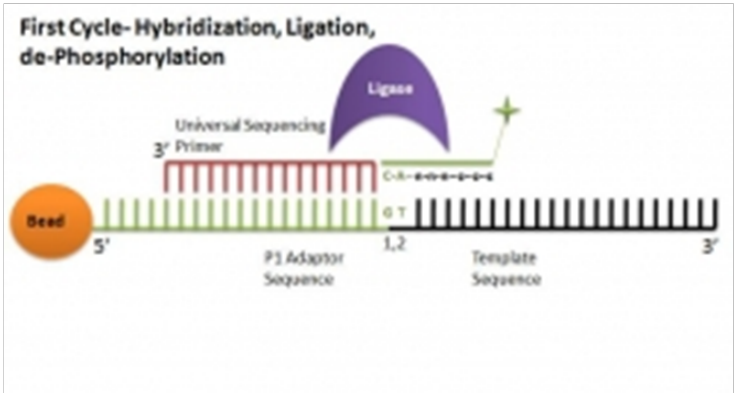
- George Church's lab (2005).
- Now AB/Life Sciences
- Sequencing by ligation.
- Library prepared by emulsion PCR.
- Beads deposited on glass slide.



SOLiD sequencer

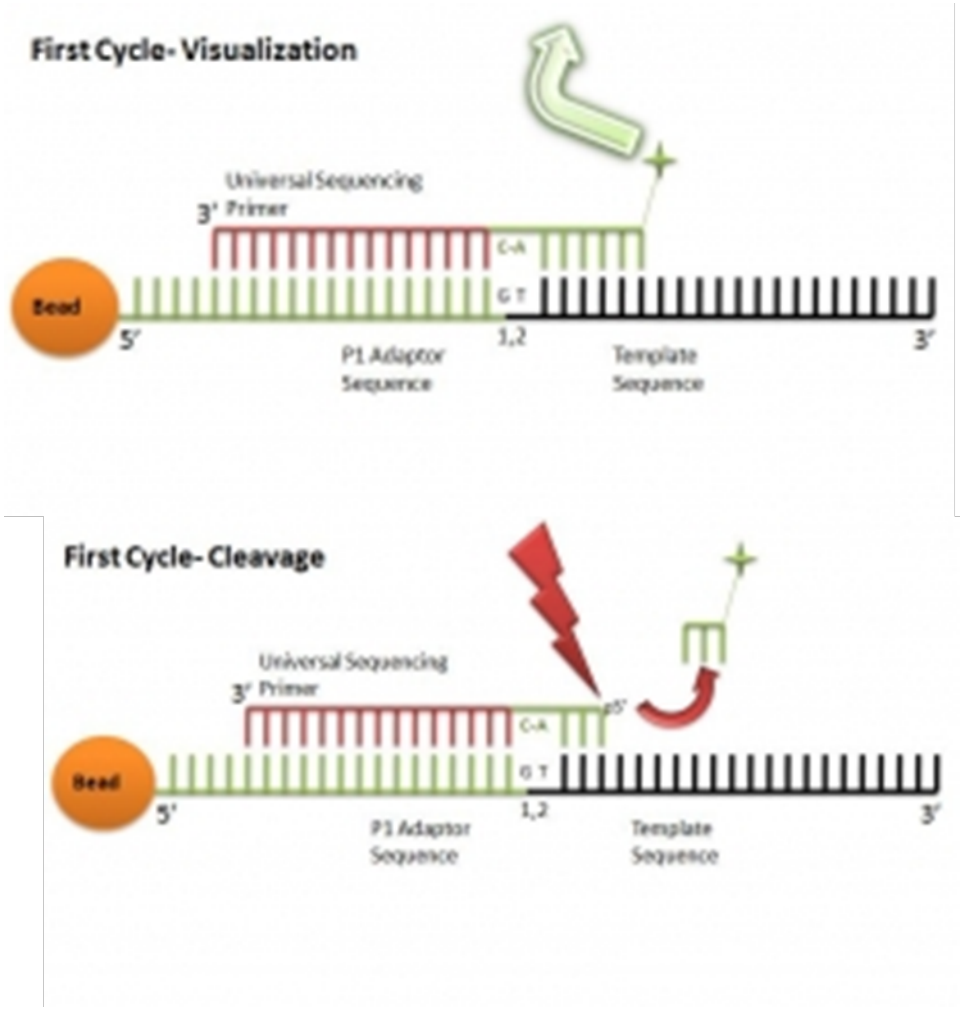
- A sequencing primer is annealed to single-stranded copies of sequences to be determined.
- Octamer probes are hybridized, ligated to the sequencing primer, and a fluorescent dye at the 5′ end of the ligated 8-mer probes, encoding the two 3′-most nucleotides of the probe, is read out.





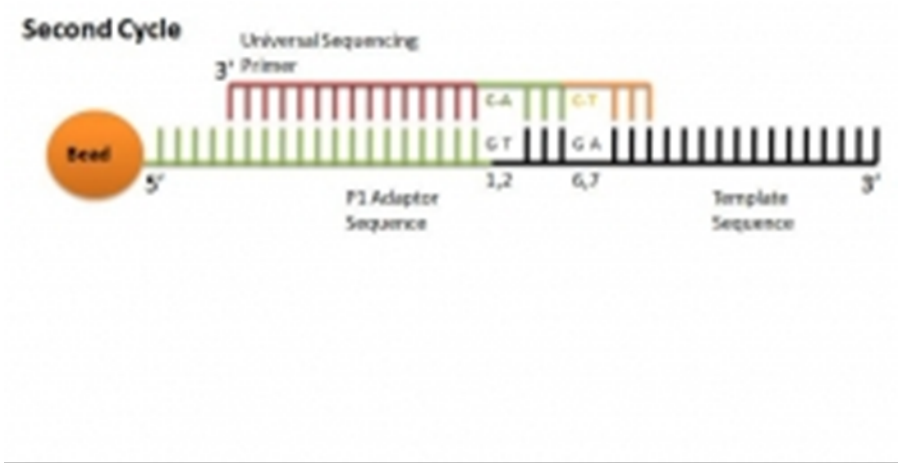
SOLiD sequencer

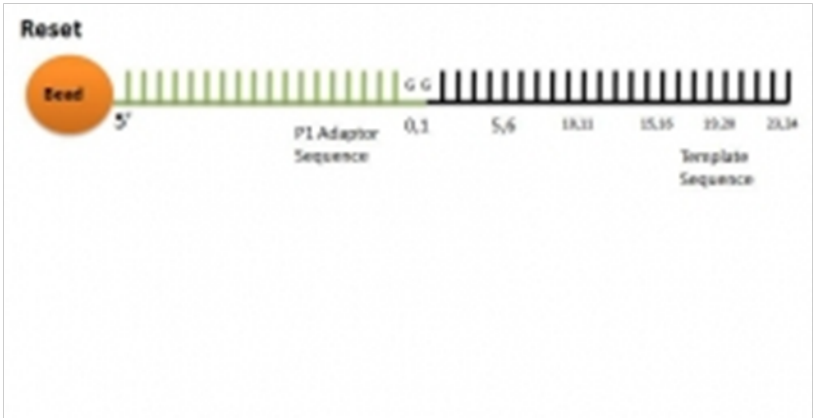
- Non-extended primers are dephosphorylated.
- Three nucleotides of the probe including the dye are cleaved, creating a free 5′ phosphate for further ligations.



SOLiD sequencer

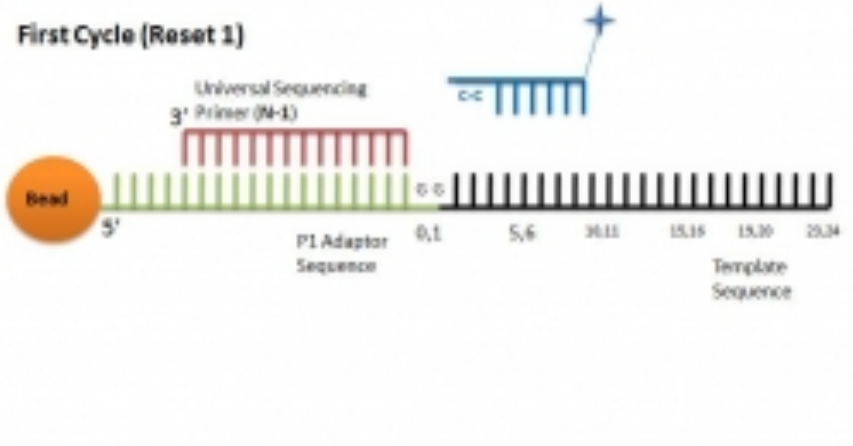
- Starting from the new sequencing primer the ligation reaction is repeated.
- After multiple ligations, the synthesized strands are melted and the ligation product is washed away before a new, by-one-nucleotide-shifted sequencing primer is annealed.



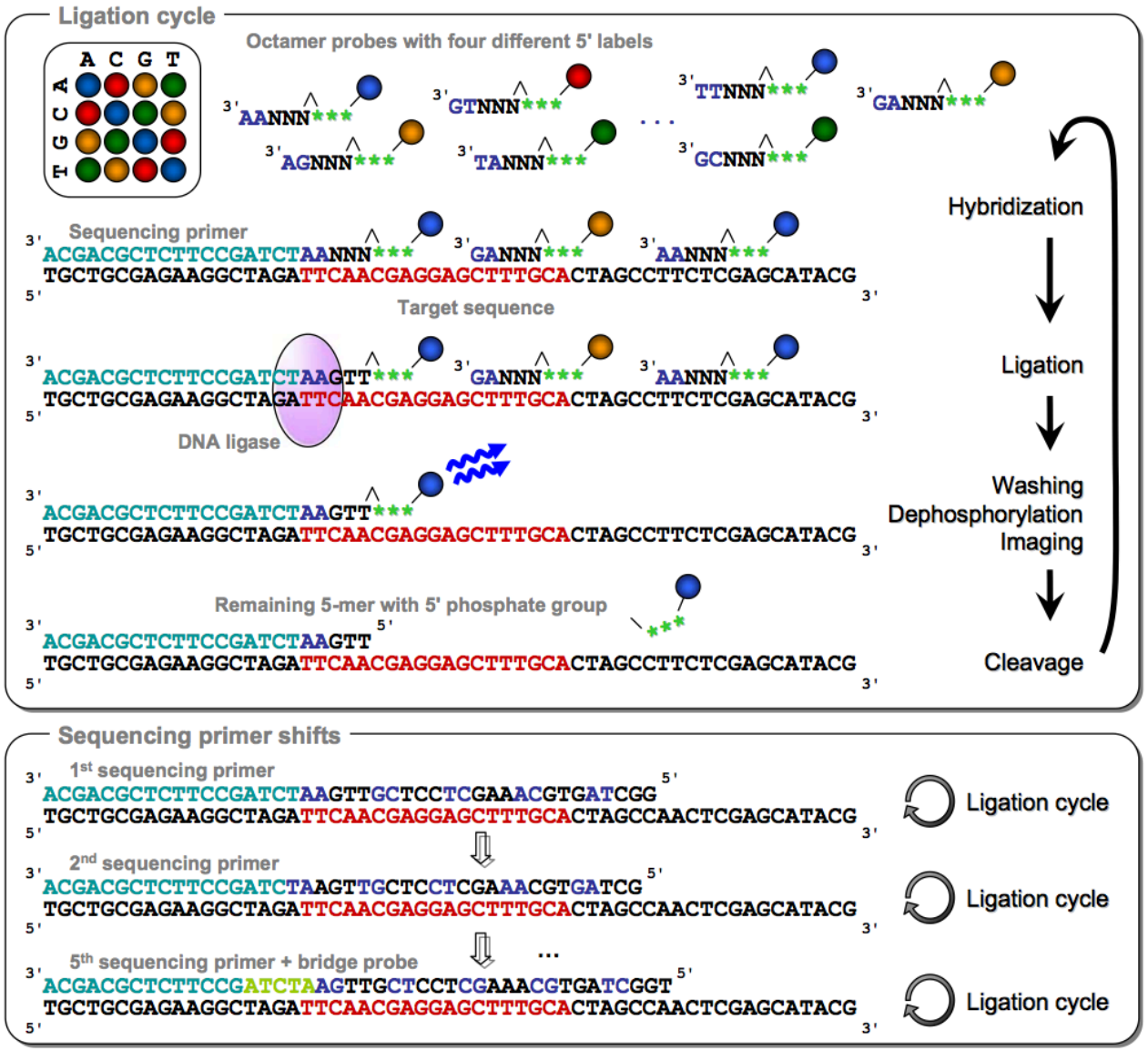


SOLiD sequencer

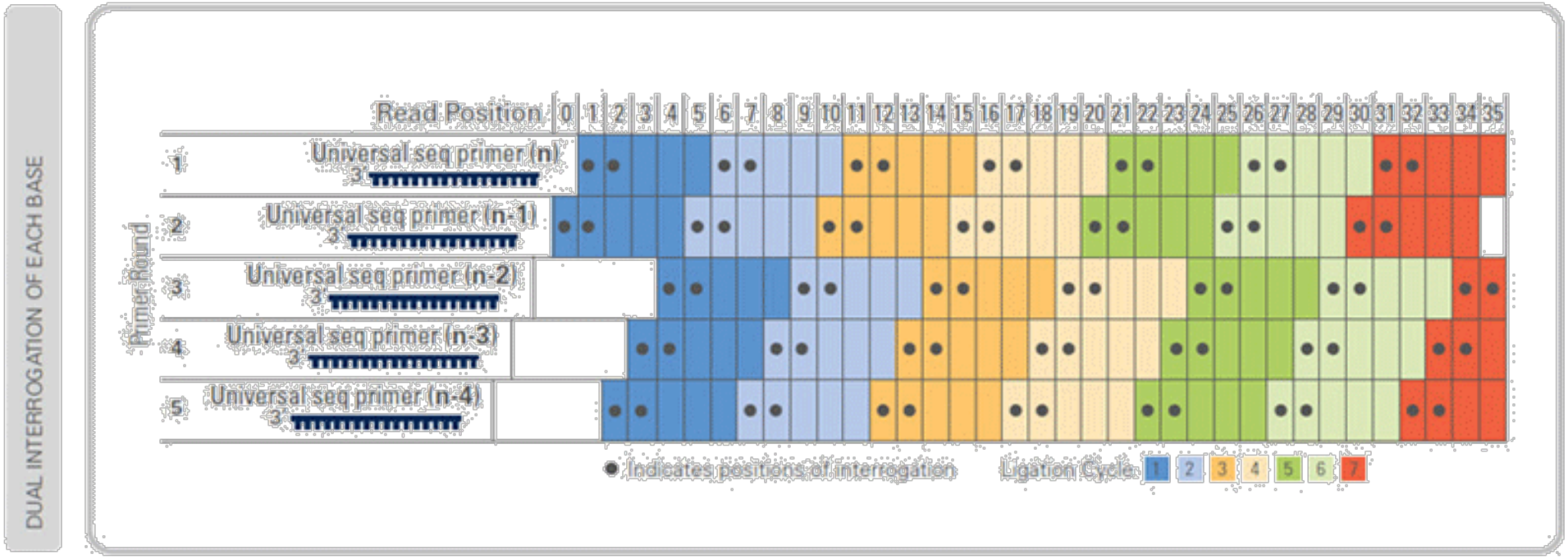
- The procedure is repeated with the three other primers, allowing the read out of the dinucleotide label for every position in the sequence.



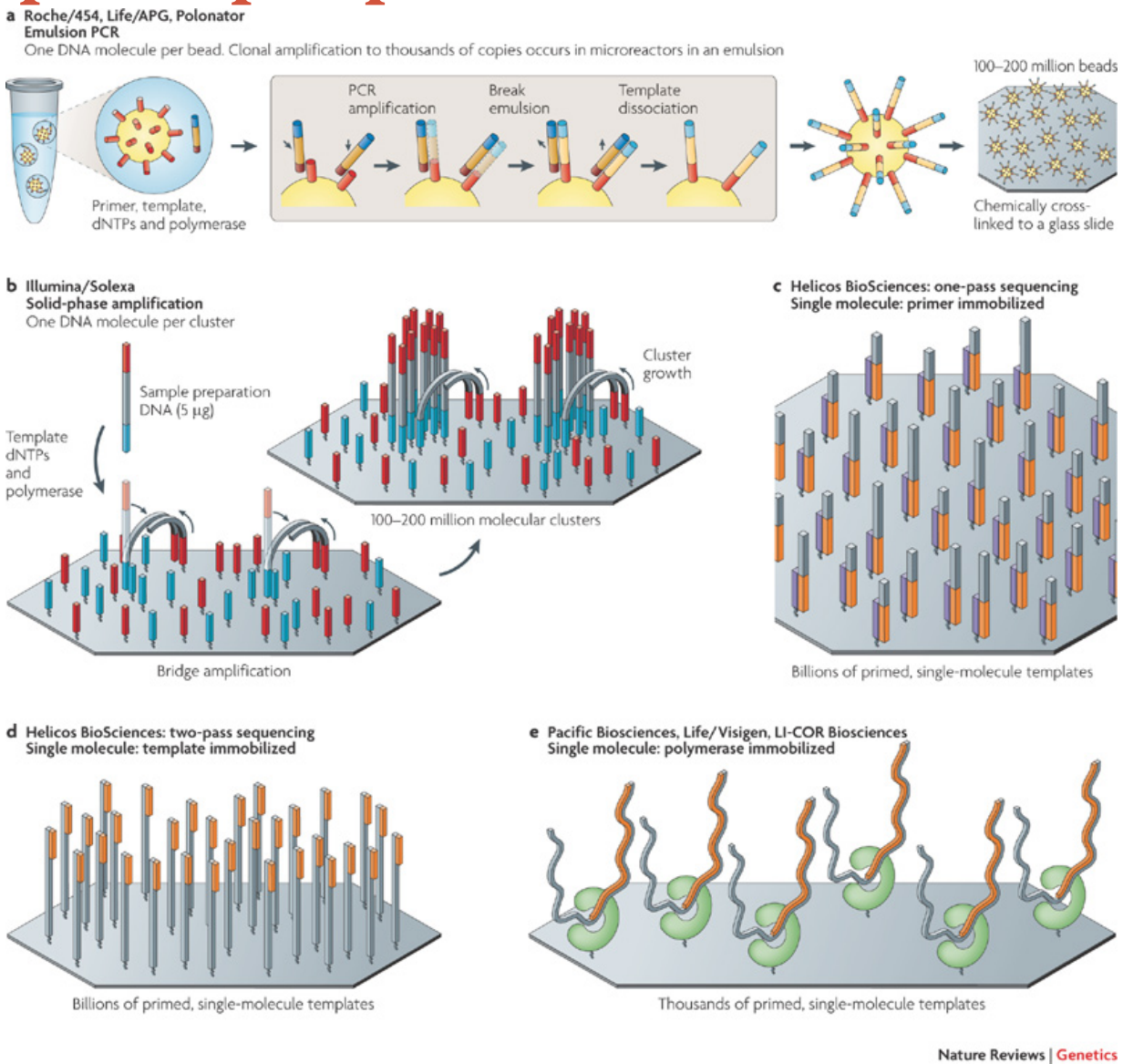
SOLiD sequencer



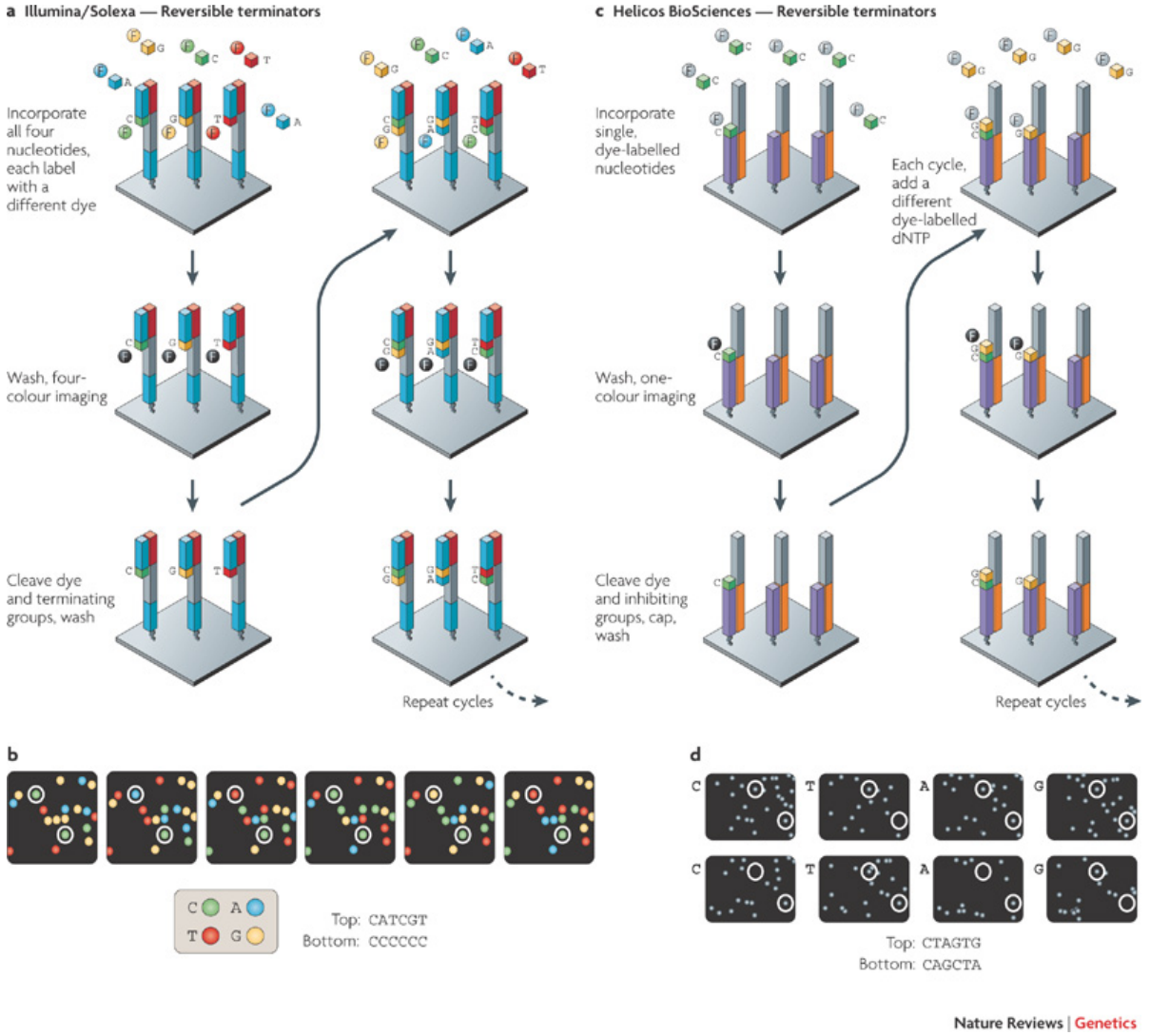
SOLiD "Colorspace"



template preparation

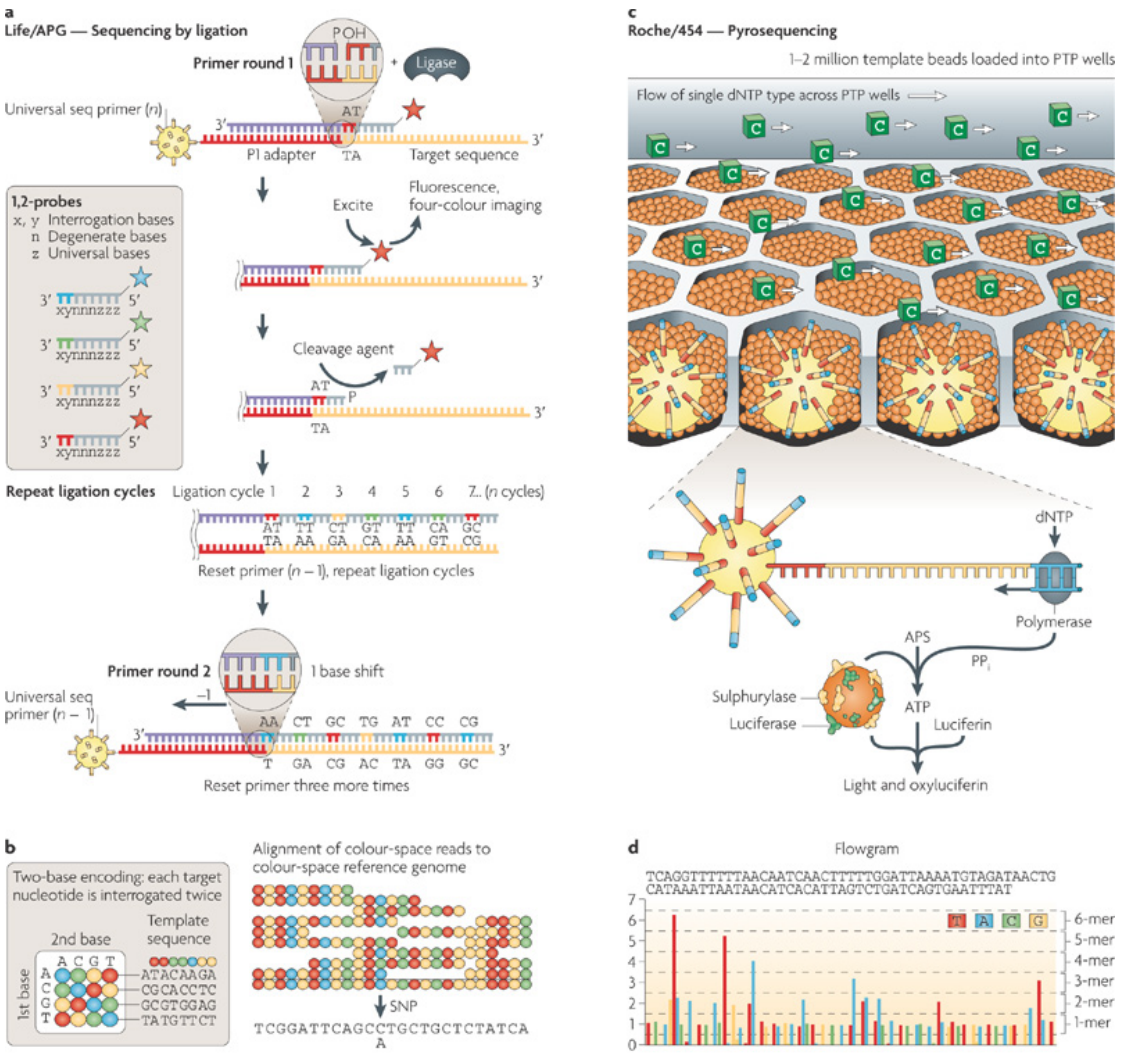


Four colour and one colour cyclic reversible termination methods

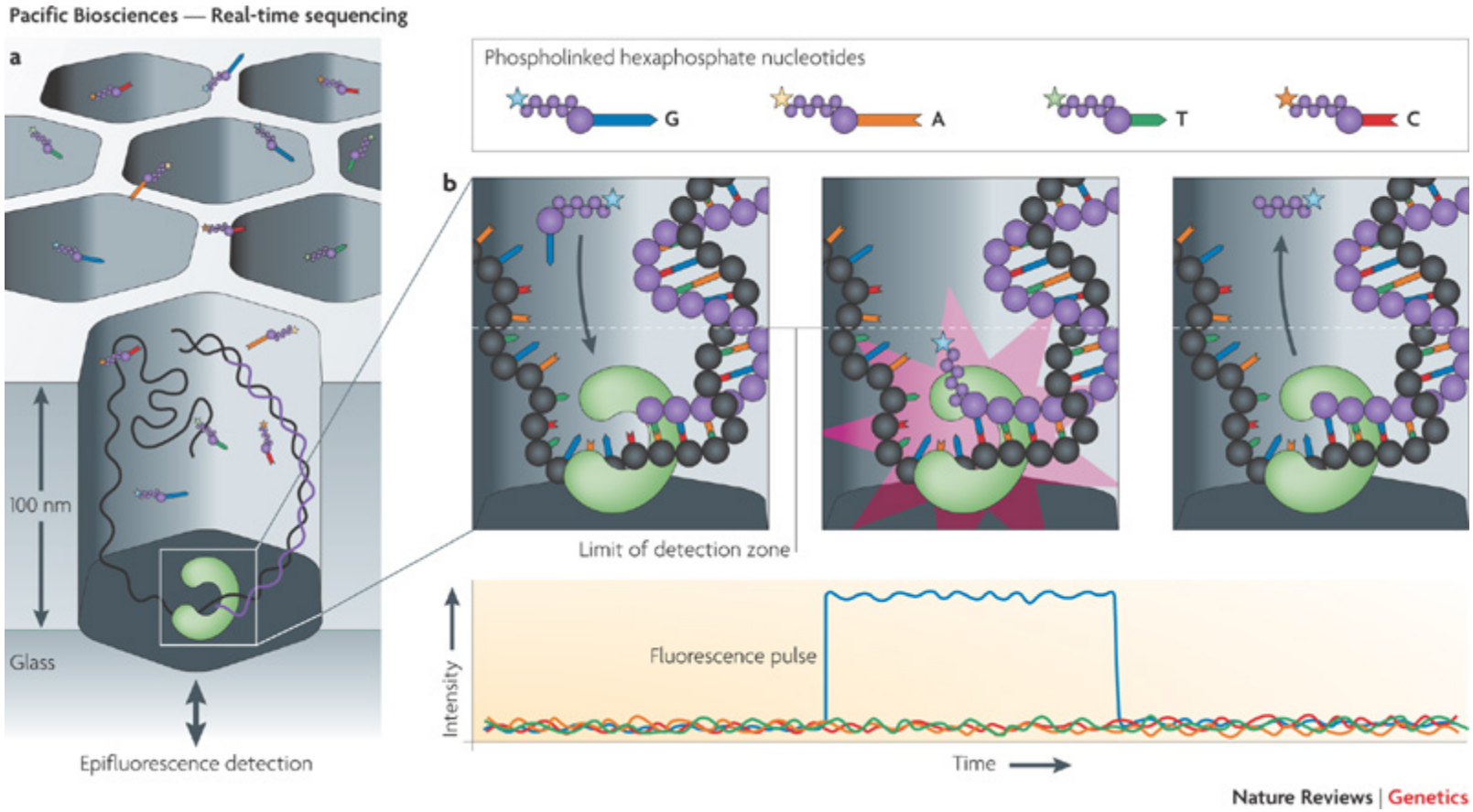


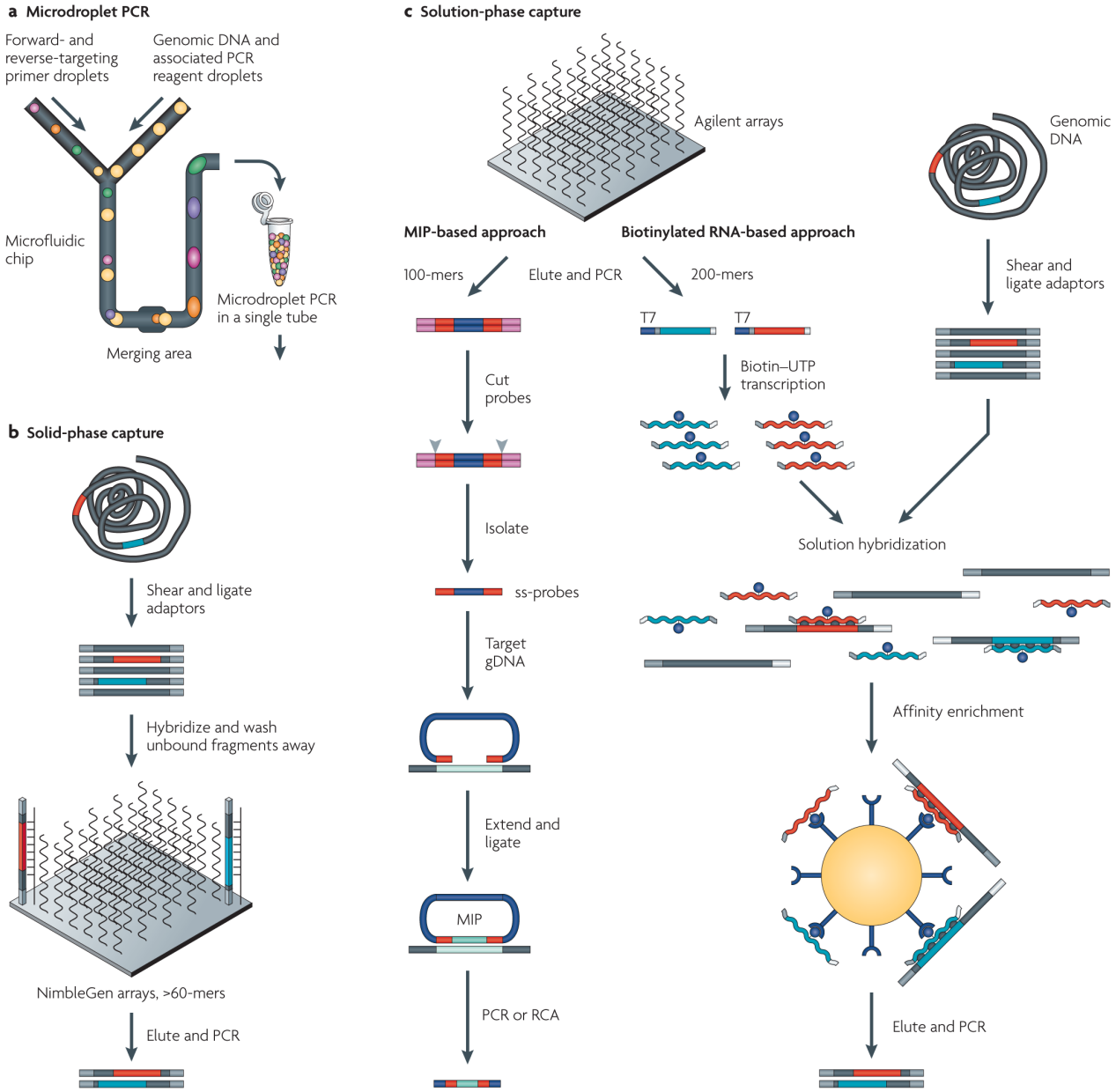
Nature Reviews | Genetics

NGS that use ePCR



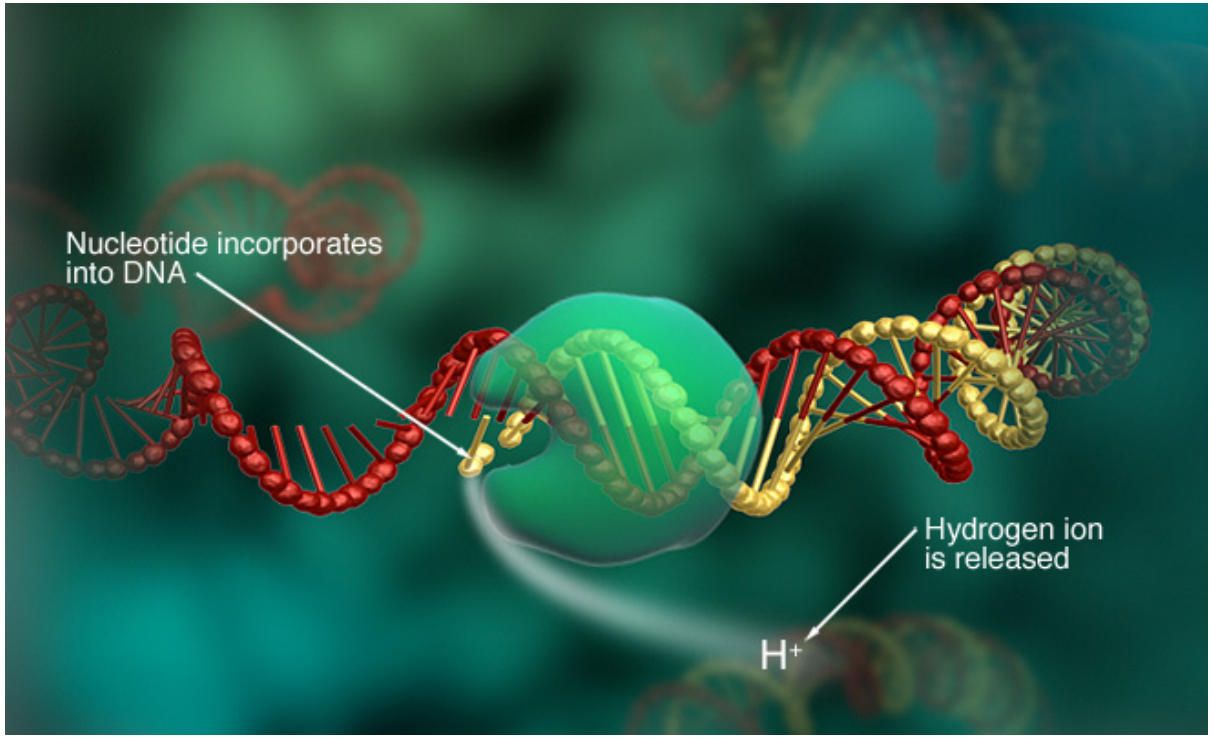
Real-time sequencing




a Microdroplet PCR
Forward- and reverse-targeting primer droplets
Genomic DNA and associated PCR reagent droplets
Microfluidic chip
Microdroplet PCR in a single tube
Merging area
b Solid-phase capture
Shear and ligate adaptors
Hybridize and wash unbound fragments away
NimbleGen arrays, >60-mers
Elute and PCR
c Solution-phase capture
Agilent arrays
MIP-based approach
Biotinylated RNA-based approach
100-mers
Elute and PCR
200-mers
T7
T7
Cut probes
Biotin–UTP transcription
Isolate
ss-probes
Target gDNA
Extend and ligate
MIP
PCR or RCA
Genomic DNA
Shear and ligate adaptors
Solution hybridization
Affinity enrichment
Elute and PCR

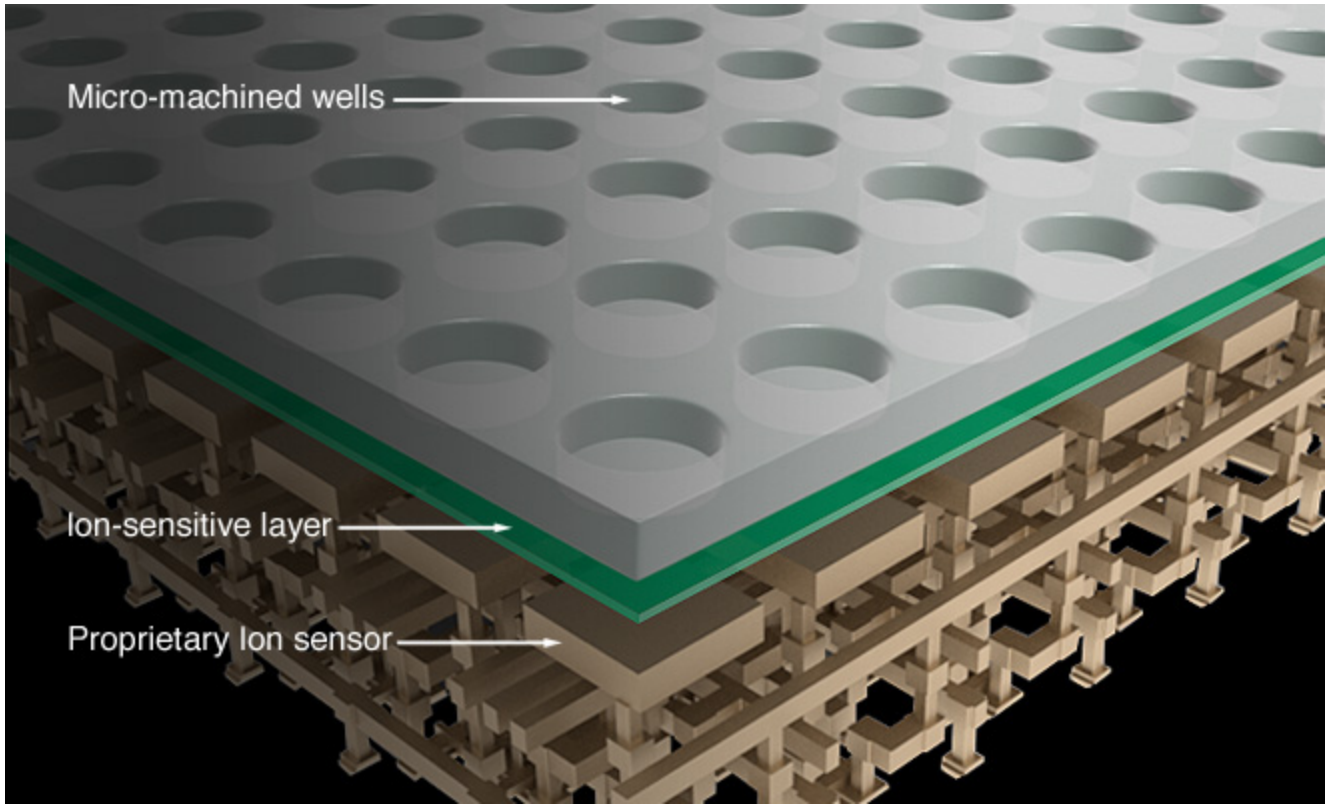
Ion Torrent Sequencing

- Non-optical system
- Semi-conductor sequencing chips
- **Nice video here:** http://ioncommunity.lifetechnologies.com/community/intro



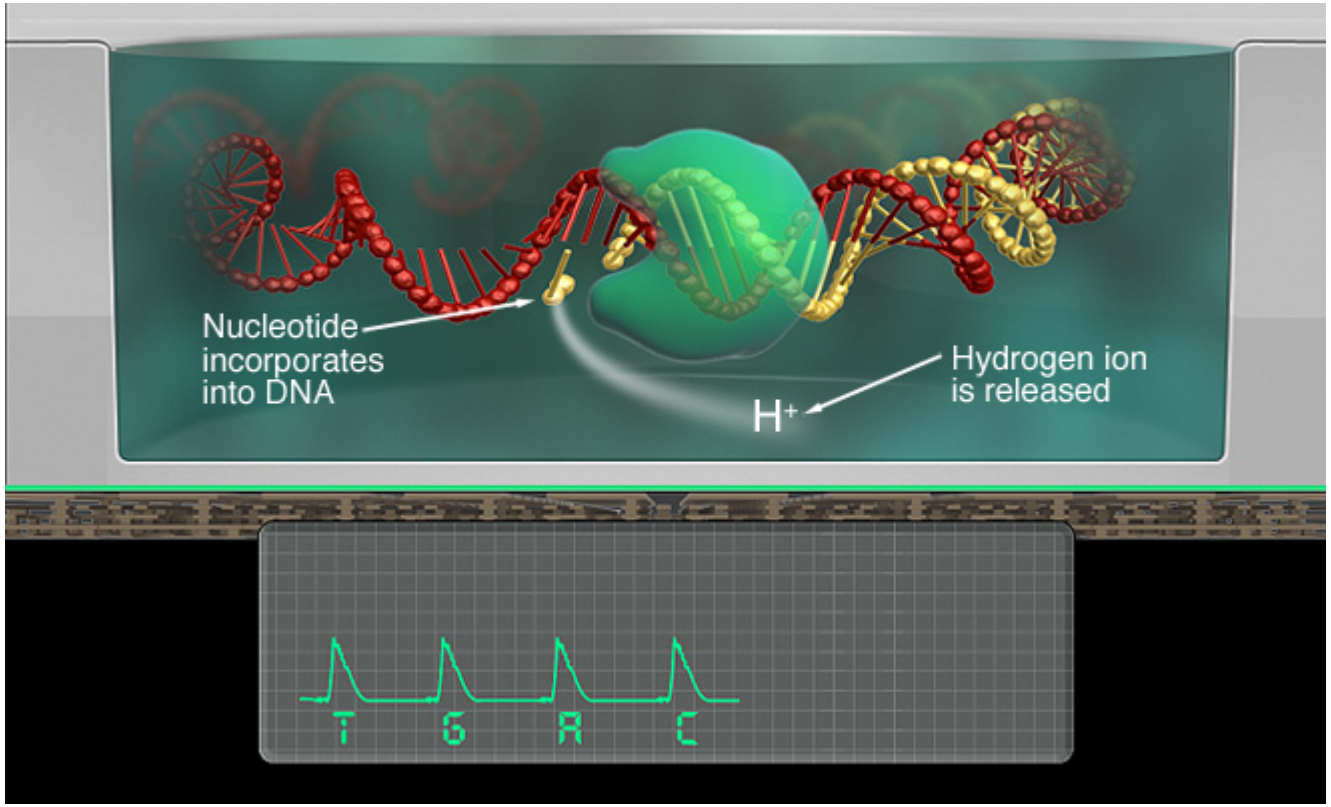
Ion Torrent Sequencing

- Non-optical system
- Semi-conductor sequencing chips



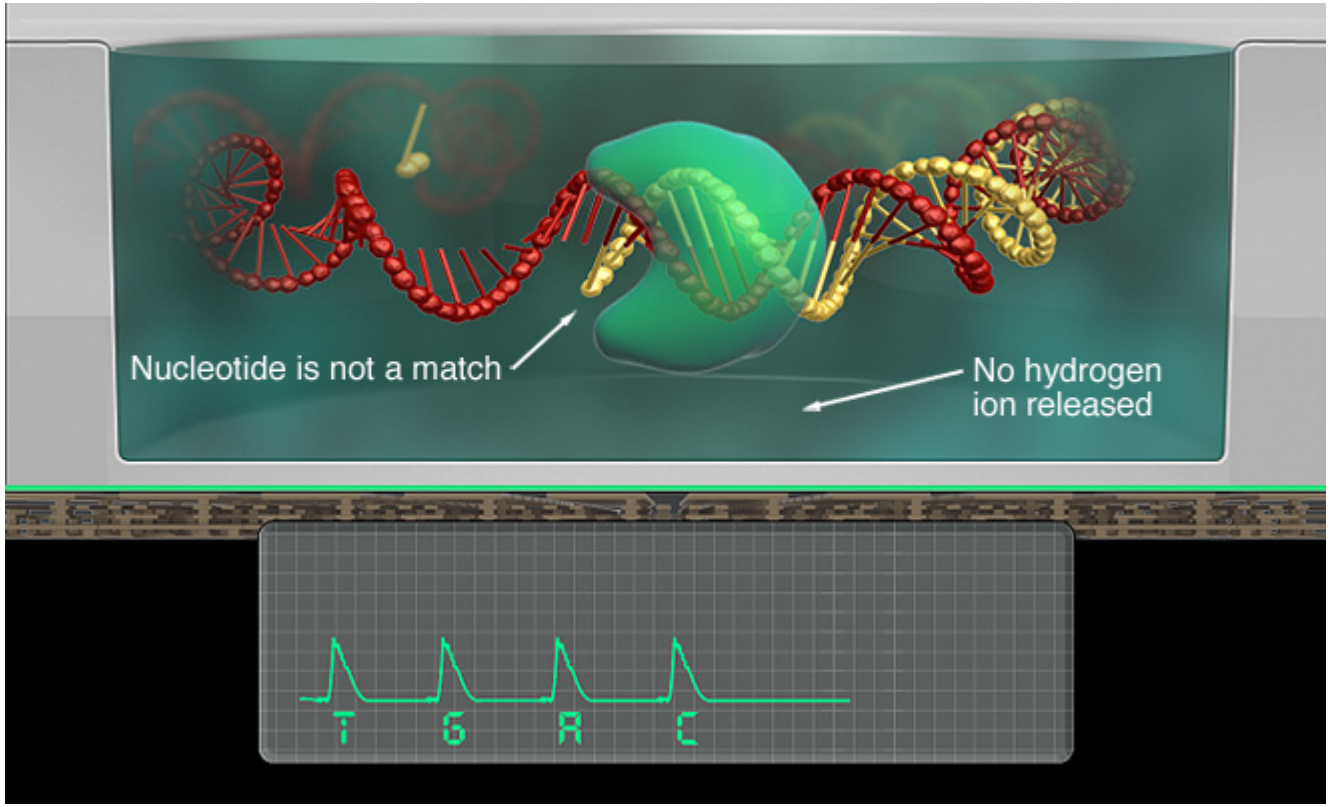
Ion Torrent Sequencing

- Non-optical system
- Semi-conductor sequencing chips



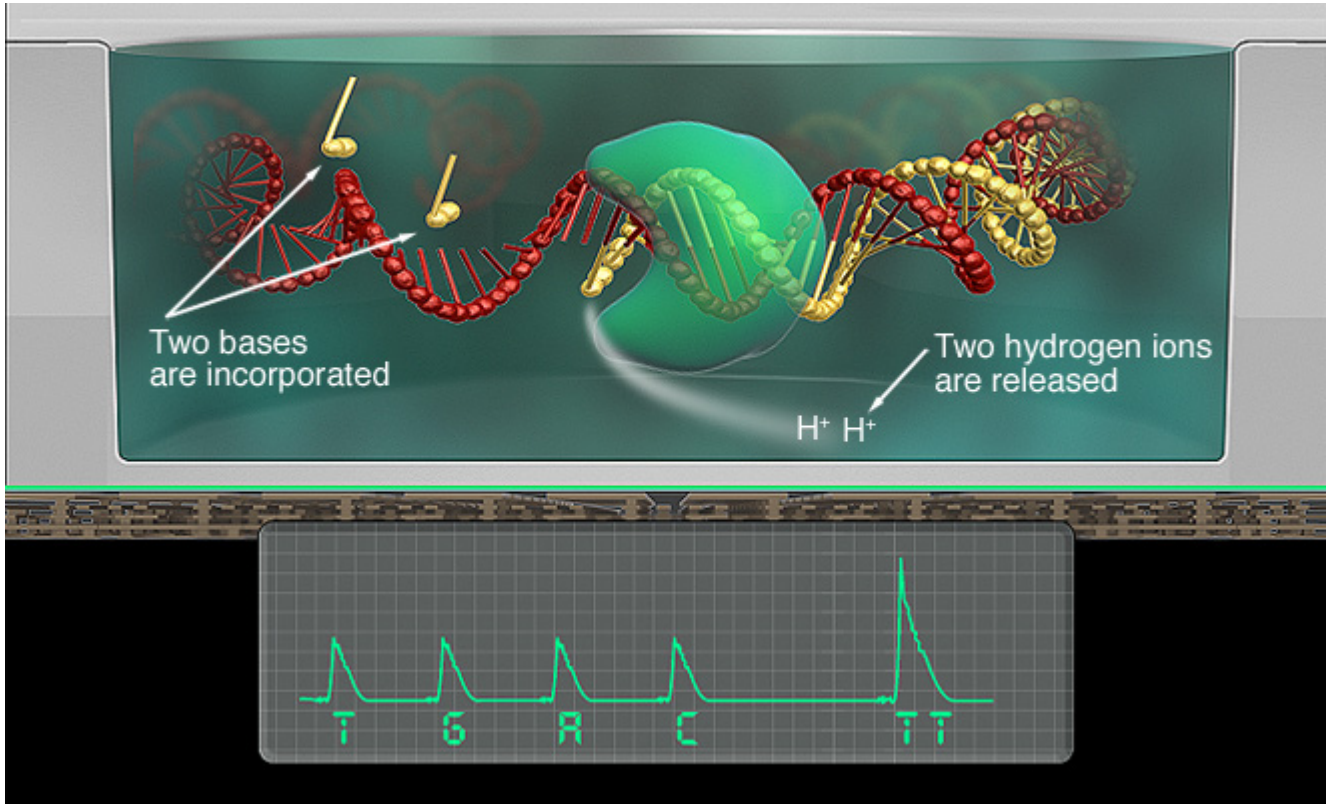
Ion Torrent Sequencing

- Non-optical system
- Semi-conductor sequencing chips



Ion Torrent Sequencing

- Non-optical system
- Semi-conductor sequencing chips



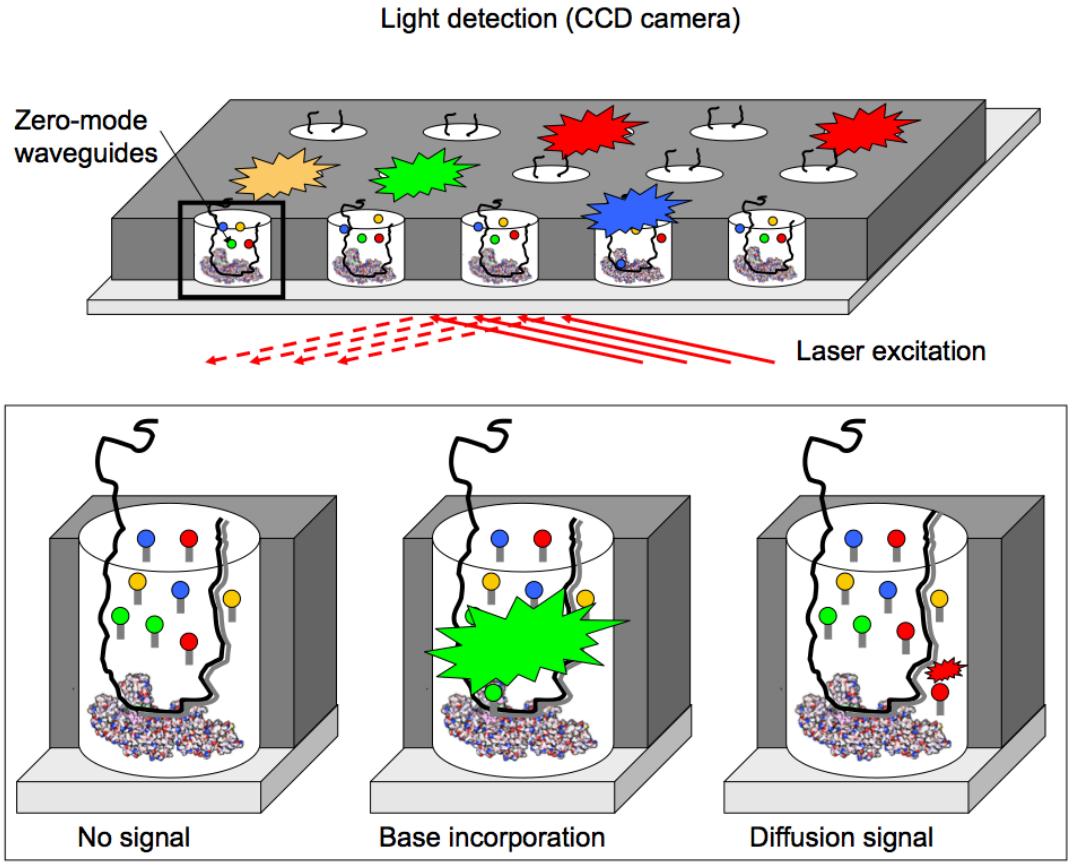
SINGLE MOLECULE SEQUENCING

A major breakthrough?

Single-Molecule sequencing

- Library preparation often entails PCR which introduces error and bias as some sequences are over-represented.
- If we are interested in minor allele frequencies (I am!) then this is a problem.
- Most sequencing approaches require amplification of template for detection, as majority of imaging systems not designed to detect single fluorescent events
- It is possible, however, to sequence individual DNA molecules.
- This brings a second benefit: much longer reads!
- At the moment these methods are limited in accuracy.

Pacific Biosciences/SMRT Sequencing



SMRT Video

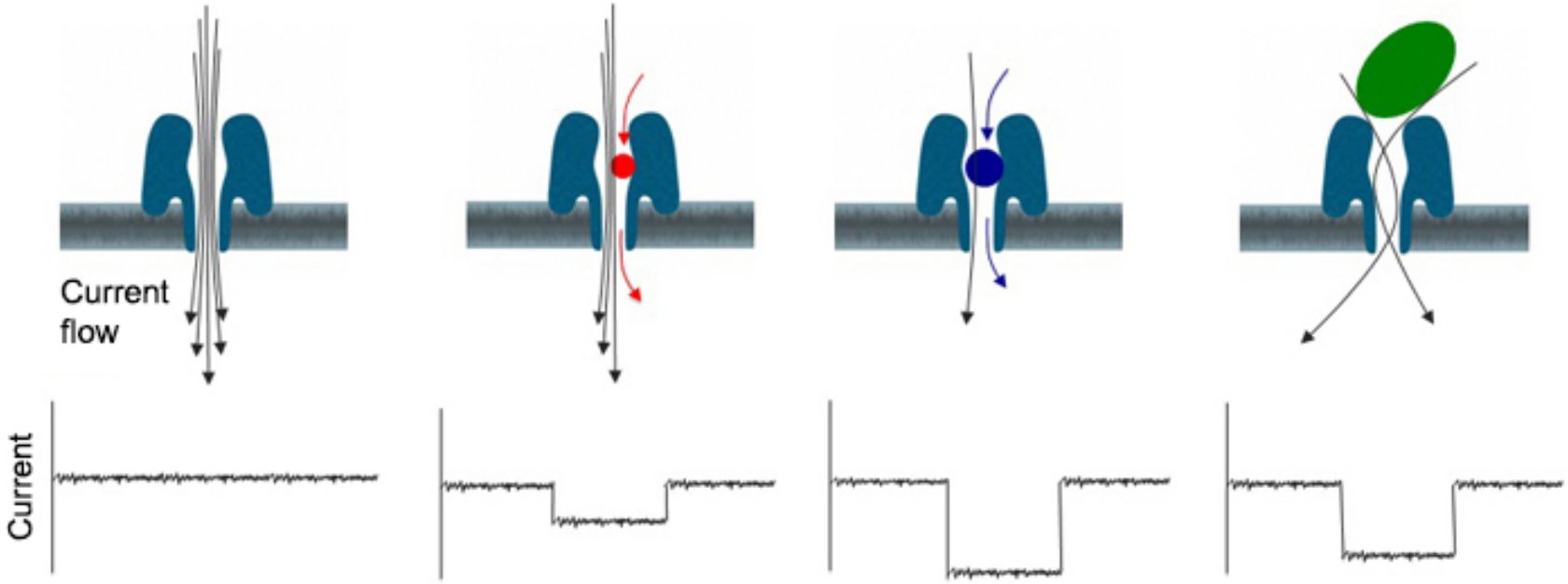
- http://www.pacb.com/smrt-science/smrt-sequencing/

SMRT/Pacific

- First publication: 2009 in Science.
- Zero-mode waveguides are holes 10^{-21} litres in size!
- During the time it takes for a new nucleotide to be incorporated its presence can be detected (10s of milliseconds) = real-time sequencing.
- Read length ~10-15kb with PacBio® RSII. Pacbio blog: http://blog.pacificbiosciences.com
- Single-pass error rate of ~14% according to this source: http://allseq.com/knowledgebank/sequencing-platforms/pacific-biosciences
- Like many companies Pacific focus on consensus accuracy – in this case this is reasonable as errors are stochastic.

Oxford Nanopore

- British company – extrusion of molecule through pore.
- Biological and solid state nanopores – sequencing by exclusion of current flow:

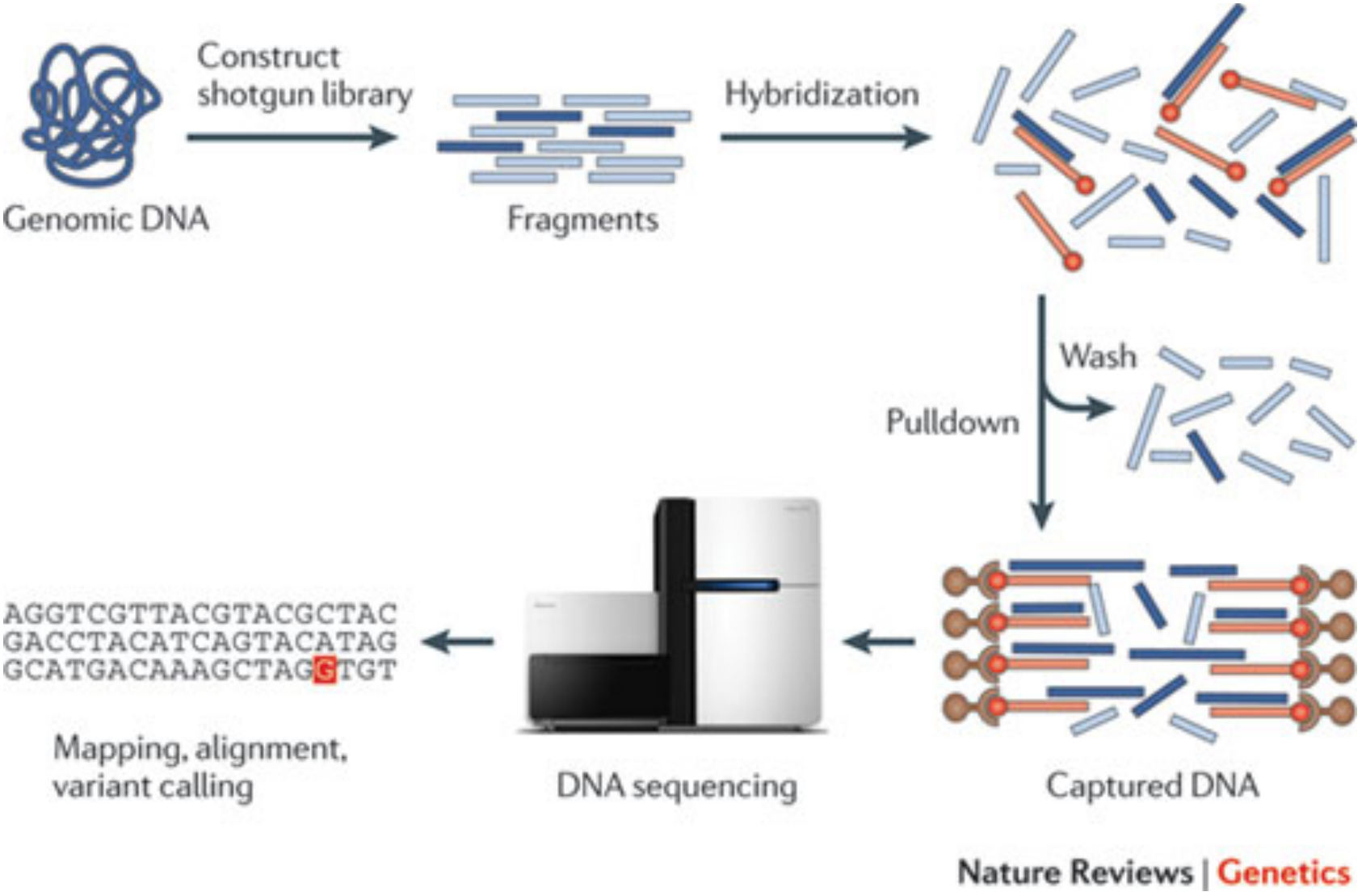


- Or by exonuclease activity on the other side of the pore.
- Library prep – fragment only (can take place in a tube).

MinION

- Oxford Nanopore has PromethION, GridION and MinION platforms.
- MinION is the size of a USB key.
- It is being tested in beta today by researchers.
- And open-source software is being developed now: http://biomickwatson.wordpress.com/2014/08/10/pore-an-r-package-for-the-visualization-and-analysis-of-nanopore-sequencing-data/

Example: exome sequencing



Platform comparisons

Platform	Read length (bp)	Per base error rate (excl. indels)
Sanger (capillary)	~800	~10^{-4}
454 (GS FLX +)	700 (mode)	~10^{-3}
Illumina (HiSeq 2500)	2 × 150	~0.1%
Ion Torrent	200	~1%
SOLiD (5500xl)	2 × 60	<=0.1%
Pacific SMRT (RS II)	15,000	~15%
Oxford Nanopore MinION		

- Above comparisons are very approximate – please do not cite!
  Caution: most companies give consensus accuracy.
- Future technologies: Single-molecule technologies of great interest.
- Obsolescence: Helicos' Heliscope is no more.

Types of error

Platform	Base-specific errors	Progressive (read) errors
454	*In vivo* amplification, >1 sequence per bead, single-base indels, homopolymer runs (esp. > 10nt), ghost wells	phasing (lagging), intensity loss: fewer strands, enzyme decay/loss?
Illumina	*In vivo* amplification, >1 sequence per cluster, cross-talk (A/C, G/T pairs excited by same laser), dust/lint	phasing (bi-directional), intensity loss: fewer strands, dimming fluorophores, background noise
SOLiD	*In vivo* amplification, >1 sequence per bead, dust/lint	Minimal phasing (phosphatase efficiency), intensity loss: incomplete dye removal

- Single molecule sequencing generally has a low signal-to-noise ratio. This also affects Ion Torrent (voltage ceiling also affects homopolymer accuracy).
- 454 appears to be the most accurate technology, but it is expensive.

Bleeding Edge Technologies

- Here is a good reference for sequencing technologies (with a market/applications focus). This shows that Illumina is dominant at present.

New Methods of Note

- Emulsion-based library prep has enabled short-read sequencing to trace back to source molecules:
  - http://10xgenomics.com
- BioNano for visualising genome maps:
  - http://www.bionanogenomics.com

More guides

- Ion Torrent: http://www.thermofisher.com/uk/en/home/life-science/sequencing/next-generation-sequencing/ion-torrent-next-generation-sequencing-technology.html#
- SOLiD Overview:
- http://www.appliedbiosystems.com/absite/us/en/home/applications-technologies/solid-next-generation-sequencing/next-generation-systems/solid-sequencing-chemistry.html
- Nanopore Overview:
- https://nanoporetech.com/science-technology/how-it-works

New Trend: Direct to Consumer

- No recommendation of these products or services is intended by the information on this slide.

- Bead-based microarrays from Illumina are offered as an individual service by 23andme (£125):
  - https://www.23andme.com/en-gb/
- One startup is even now offering full genome sequencing (*estimate* for full genome ~$745):
  - https://www.fullgenomes.com

- I expect this to be attractive to "Quantified Self" movement:
  - http://quantifiedself.com