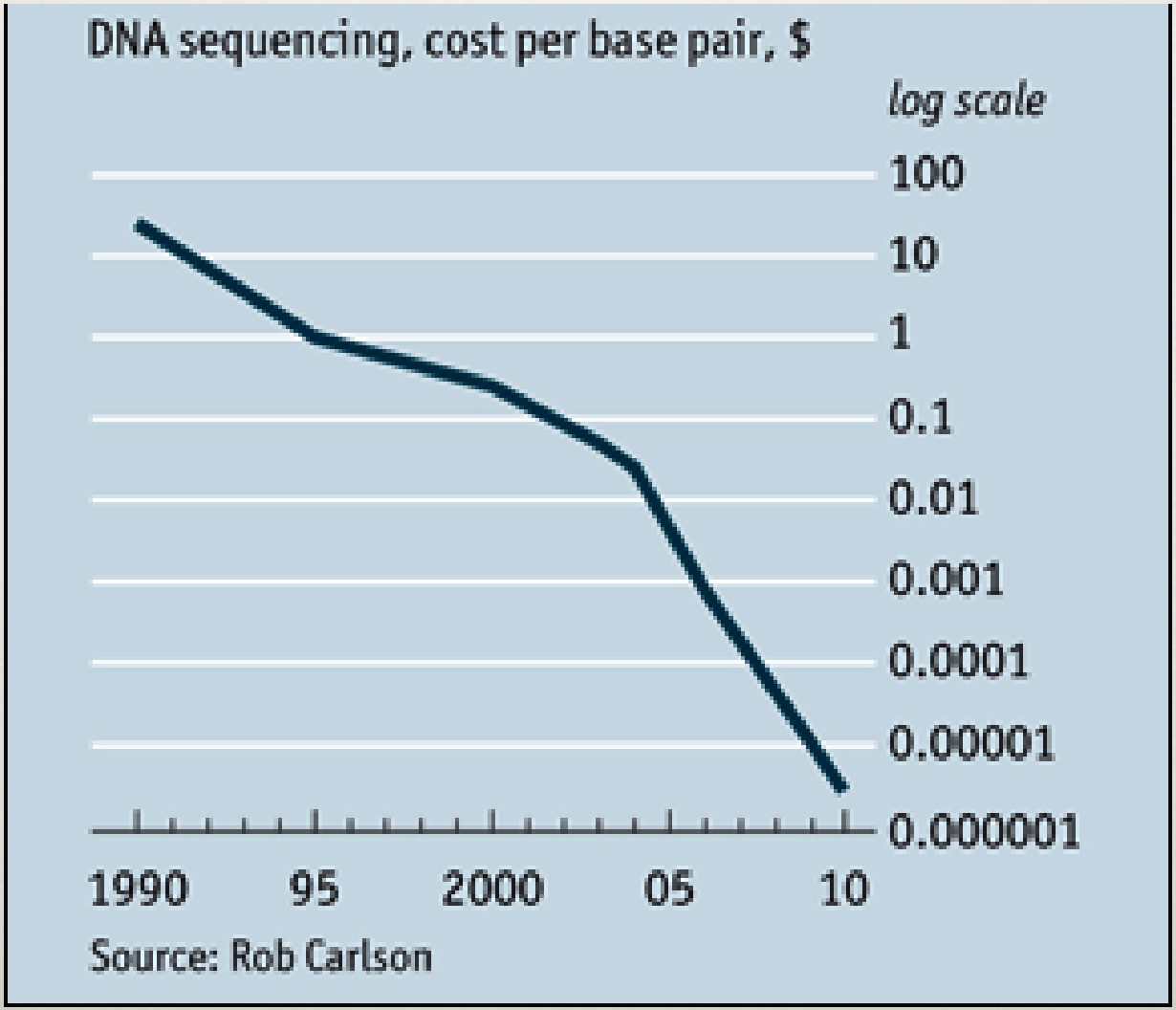
Next Generation DNA sequencing: why should we care

Shoudan Liang
(Keith Baggerly & Bradley Broom)
Microarray Analysis Class, Nov 23, 2010

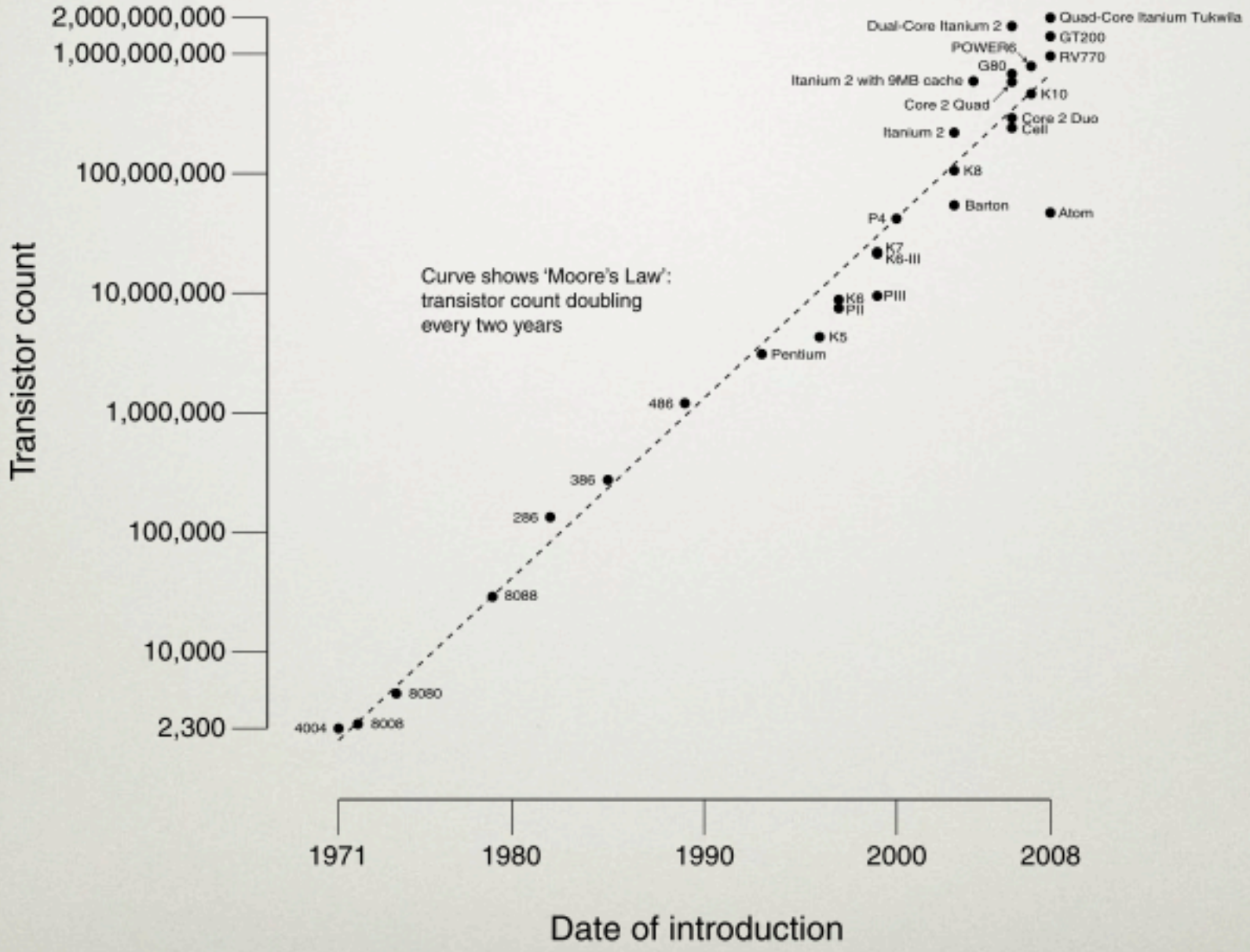
SEQUENCING CAPACITY IS GROWING EXPONENTIALLY

- first human genome sequenced over ten years at $3 billion.
- 2007, Watson's genome was sequenced in two months by 454 at $2 million.
- Last year, the cost (list price of reagent) of human genome re-sequencing using Solexa is $250,000.
- ABI SOLiD claim to be able to re-sequence at $10,000 this year.

The cost of sequencing DNA has dropped by more than a million folds in the last ten years



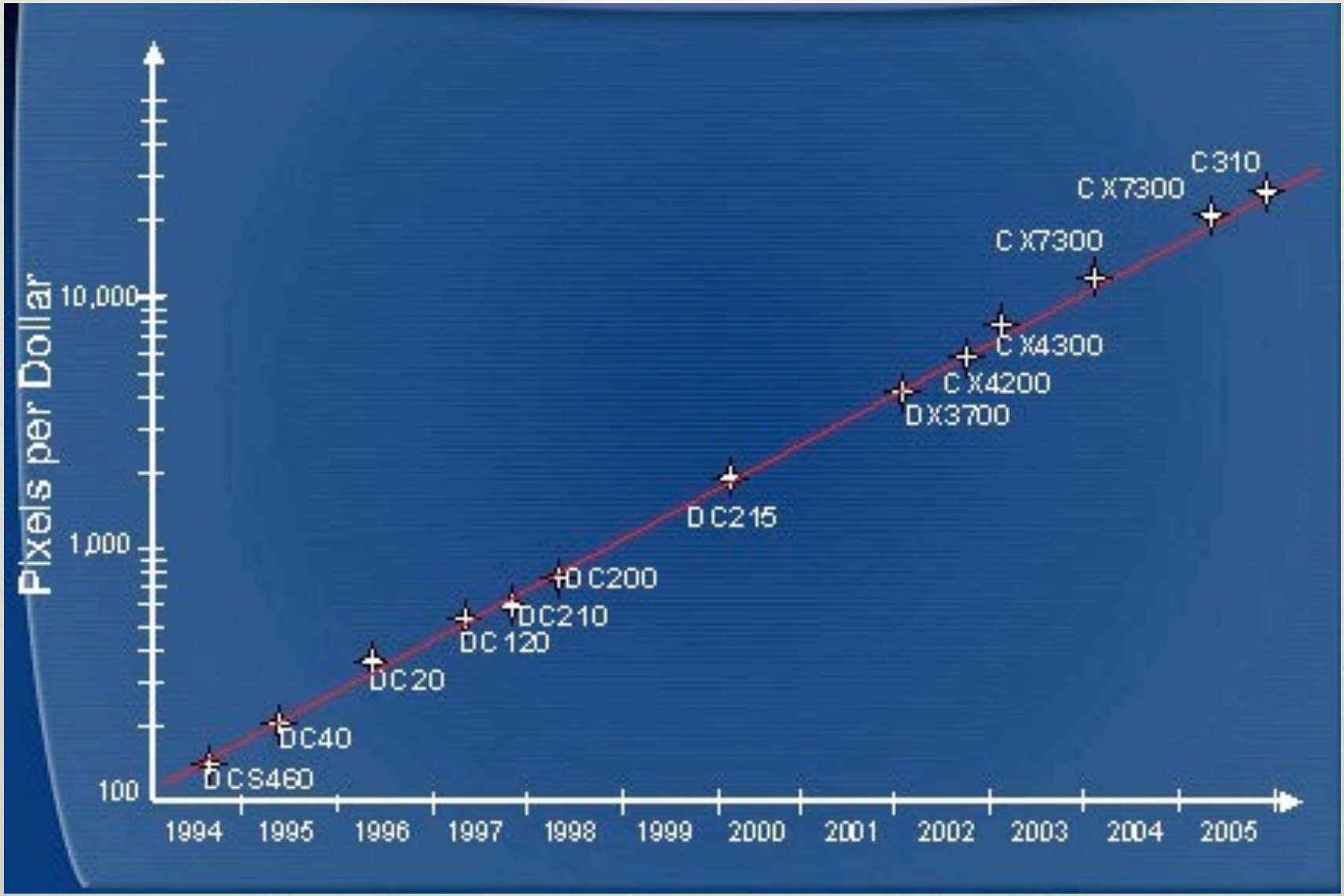
CPU Transistor Counts 1971-2008 & Moore's Law



sequencing is expected to follow Moore's-like law

- Moore's law: computing power a dollar can buy doubles every 18 months
- rate limiting step in NEXT GEN sequencing is imaging. CCD camera in sequencer will increase in capacity following Moore's law.
- DNA sequencing with semiconductor : merging of two technologies?

Pixels per dollar of Kodak digital cameras

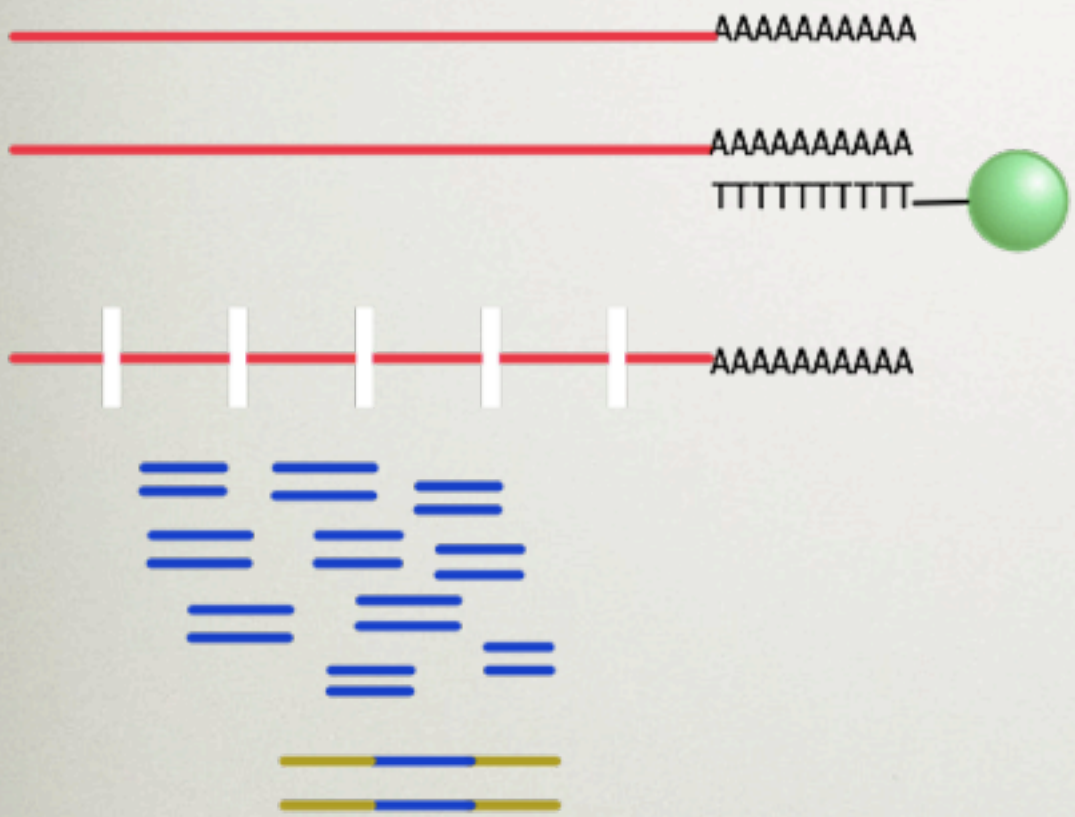


en.wikipedia.org

Applications in genome research

- a DNA sequence is a bar-code, and therefore an addressing system of a genome
- share similarities with microarray in measuring amount of DNA by genome locations

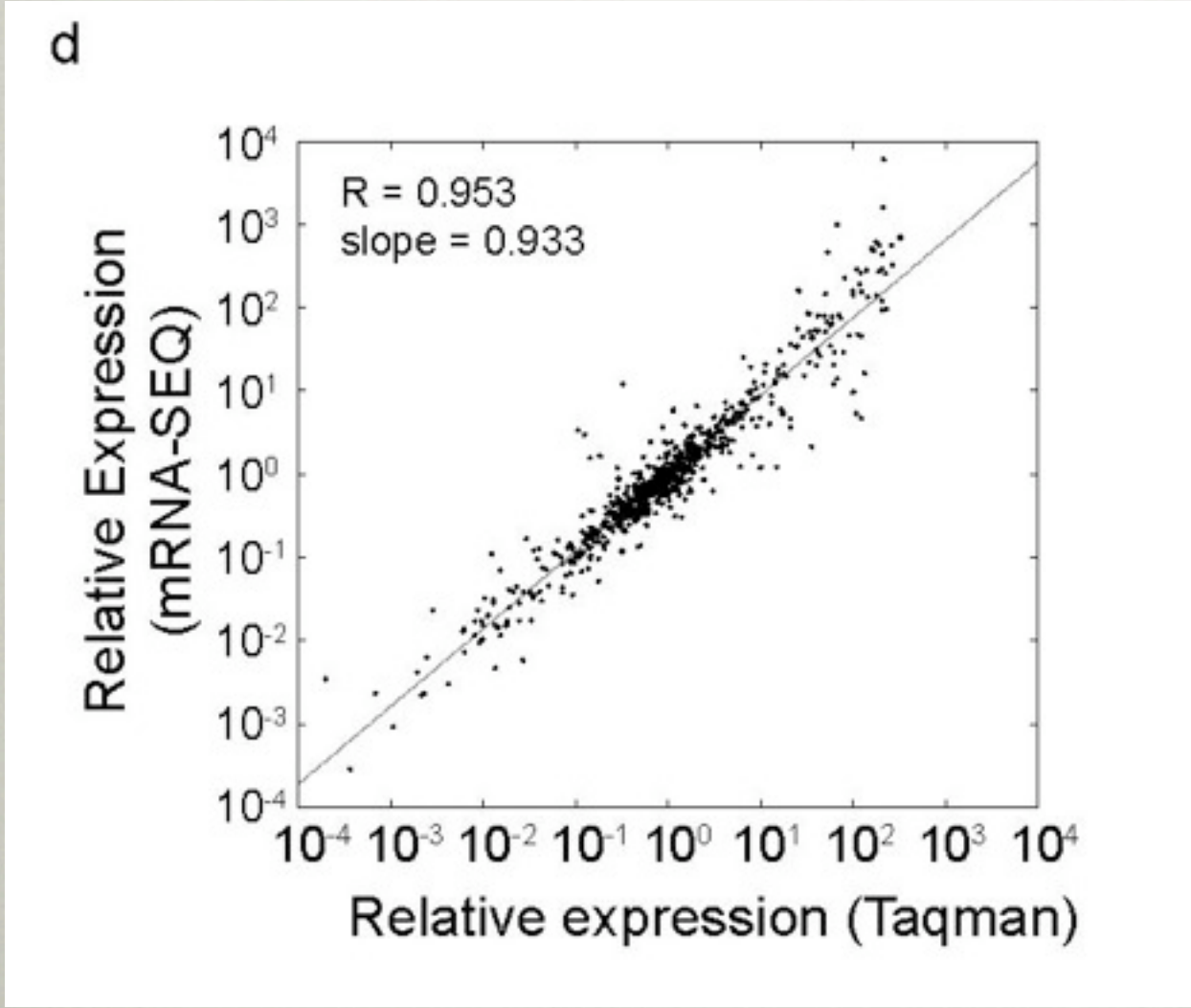
Steps in Preparing an RNA-Seq Library



1. Purify RNA
2. Bind polyA fraction (mRNA)
3. Fragment RNA (200 bp)
4. Convert to cDNA by random priming
5. Apply adaptors and sequence
6. Analyze millions of 25 bp reads

geospiza
FROM SAMPLE TO RESULTS™

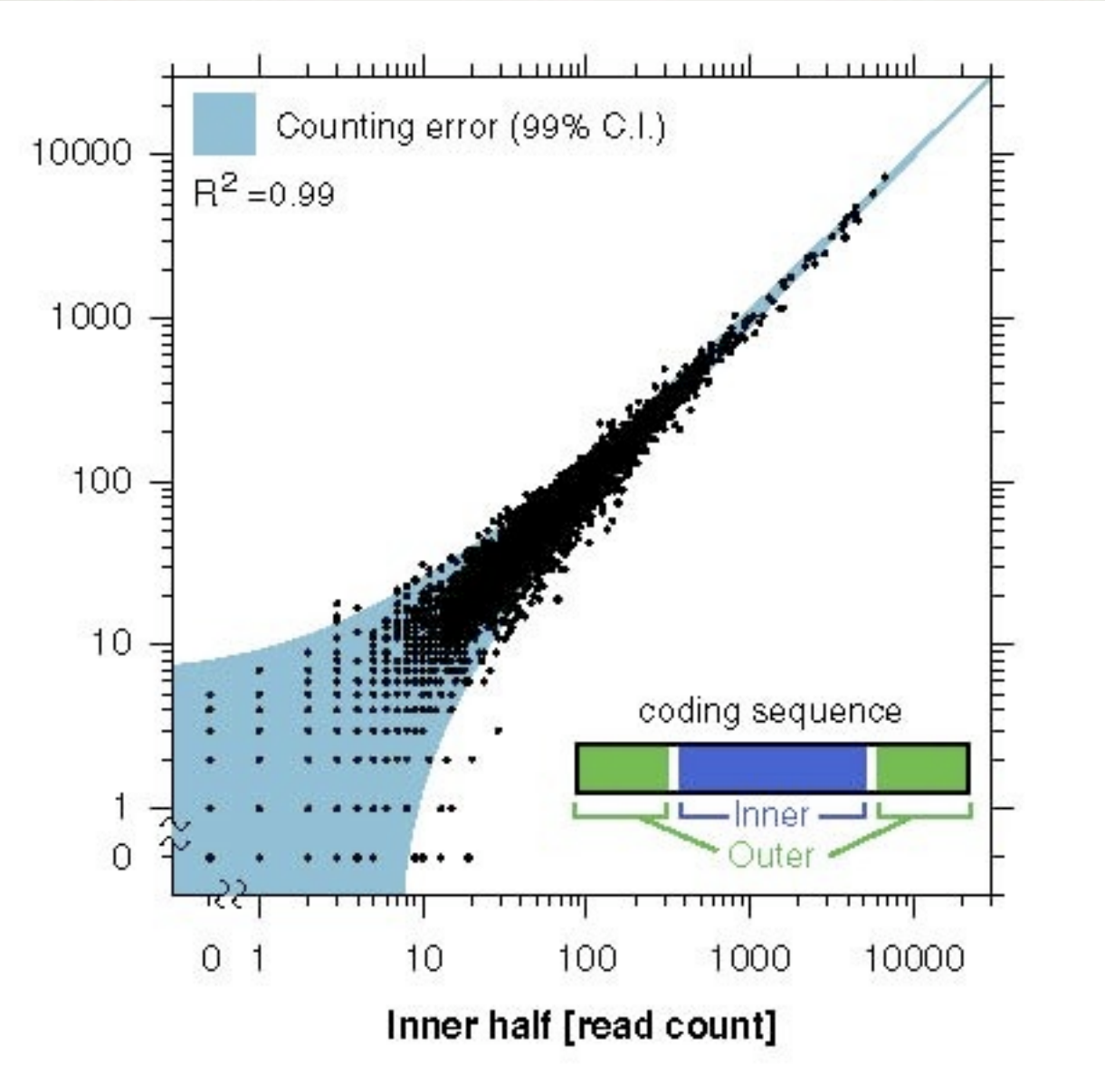
Digital gene expression Solexa vs Gold Standard



787 RefSeq human transcripts in brain and UHR

TaqMan is considered a gold standard

Wang et al Nature. 2008 Nov 27;456(7221):470-6.



Ingolia NT, et al Science. 2009 Feb 12.

Diversity of The Human Genome

ARTICLE

A map of human genome variation from population-scale sequencing

The 1000 Genomes Project Consortium*

Whole-genome low-coverage: 179 from 4 populations
Whole-genome high-coverage: 2 Mom-Dad-Child trios
Exon-targeted sequencing: 697 from 7 populations

Table 2 | Estimated numbers of potentially functional variants in genes

Class	Combined total	Combined novel	Low coverage		High-coverage trio		Exon capture		
			Total	Interquartile*	Total	Individual range	Total	Interquartile*	GENCODE extrapolation
Synonymous SNPs	60,157	23,498	55,217	10,572–12,126	21,410	9,193–12,500	5,708	461–532	11,553–13,333
Non-synonymous SNPs	68,300	34,161	61,284	9,966–10,819	19,824	8,299–10,866	7,063	396–441	9,924–11,052
Small in-frame indels	714	383	666	198–205	289	130–178	59	1–3	~25–75
Stop losses	77	40	71	9–11	22	4–14	6	0–0	~0–0
Stop-introducing SNPs	1,057	755	951	88–101	192	67–100	82	2–3	~50–75
Splice-site-disrupting SNPs	517	399	500	41–49	82	28–45	3	1–1	~50
Small frameshift indels	954	551	890	227–242	433	192–280	37	0–1	~0–25
Genes disrupted by large deletions	147	71	143	28–36	82	33–49	ND	ND	ND
Total genes containing LOF variants	2,304	NA	1,795	272–297	483	240–345	77	3–4	~75–100
HGMD 'damaging mutation' SNPs	671	NA	578	57–80	161	48–82	99	2–4	~50–100

NA, not applicable; ND, not determined.
*Interquartile range of the number of variants of specified type per individual.

{15 million SNPs; 1 million short indels; 20,000 structural variants were identified from total sequence data}

Amount of diversity observed *per individual genome*:

- 250-300 loss-of-function variants
- 50-100 variants implicated in inherited disorders
- ~10,000 non-synonymous cSNP differences compared to a published reference genome

TECHNOLOGY LOOKING FOR PROBLEMS

- currently, USA has 600 next-generation sequencers. The rest of the world another 500 or so.
- Number of human genomes to be sequenced by the end of next year is about 30,000.

platform	Feature generation	Cost per mega base	Cost per instrument	most common error	Read-length
Roche GS-FLX (454)	Emulsion PCR	$20	$500,000	Indel	400 bp
Illumina GA (Solexa)	Bridge PCR	$2	$430,000	Subst.	36 bp
ABI SOLiD	Emulsion PCR	$2	$591,000	Subst.	35 bp
HeliScope	Single molecule	$1	$1,350,000	Del	30 bp
Pacific Biosciences	Single molecule	-	-	Del/ Subst.	long
Complete Genomics	Nanoball	$0.01	NA	Subst.	35 bp

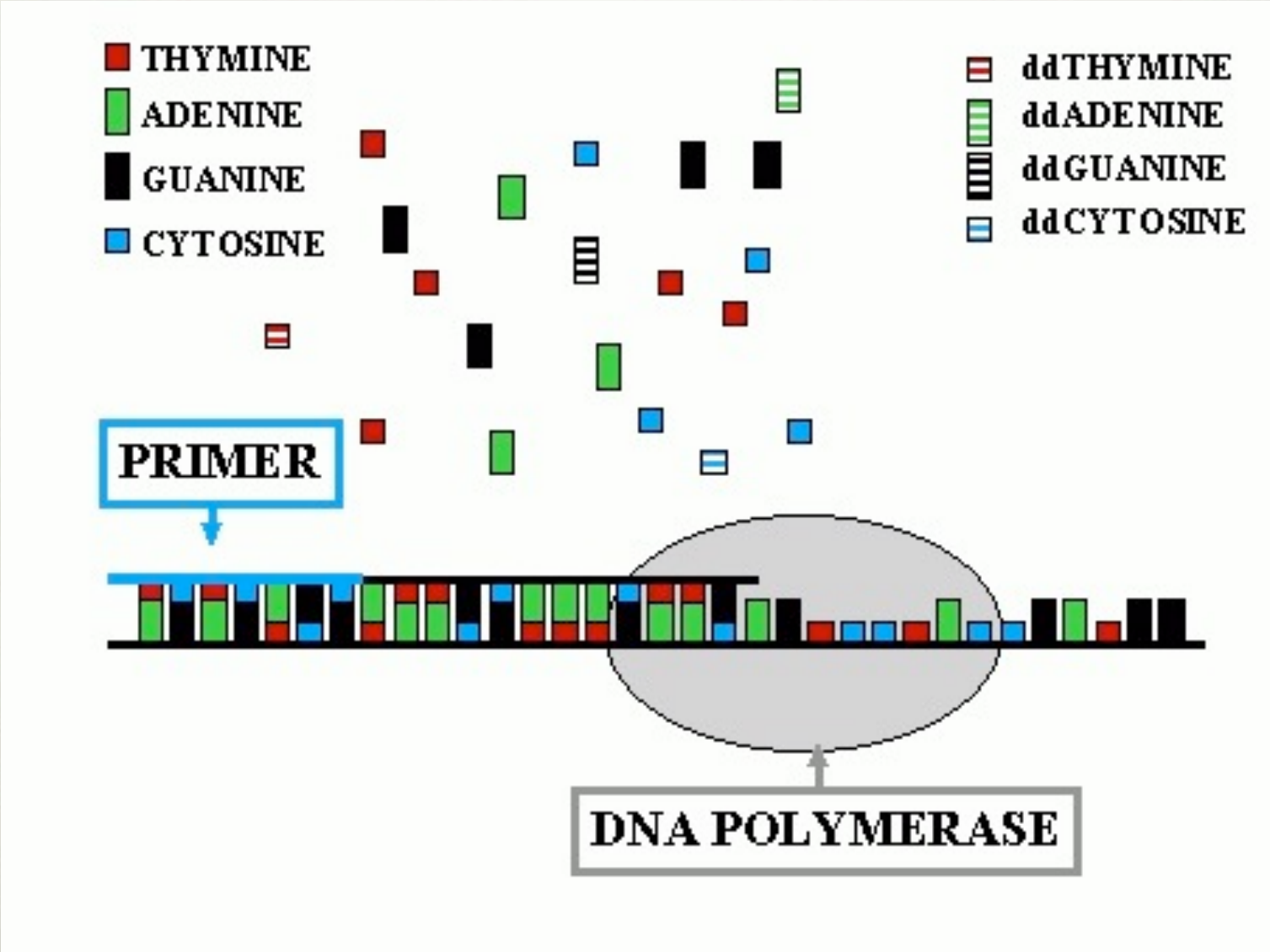
Modified from Shendure & Ji, *Nature Biotechnology* **26**, 1135 - 1145 (2008)

NEXT GEN SEQUENCING: HARDWARE

- Sanger
- 454
- ABI SOLiD
- Illumina Solexa
- Complete Genomics
- Pacific Biosciences

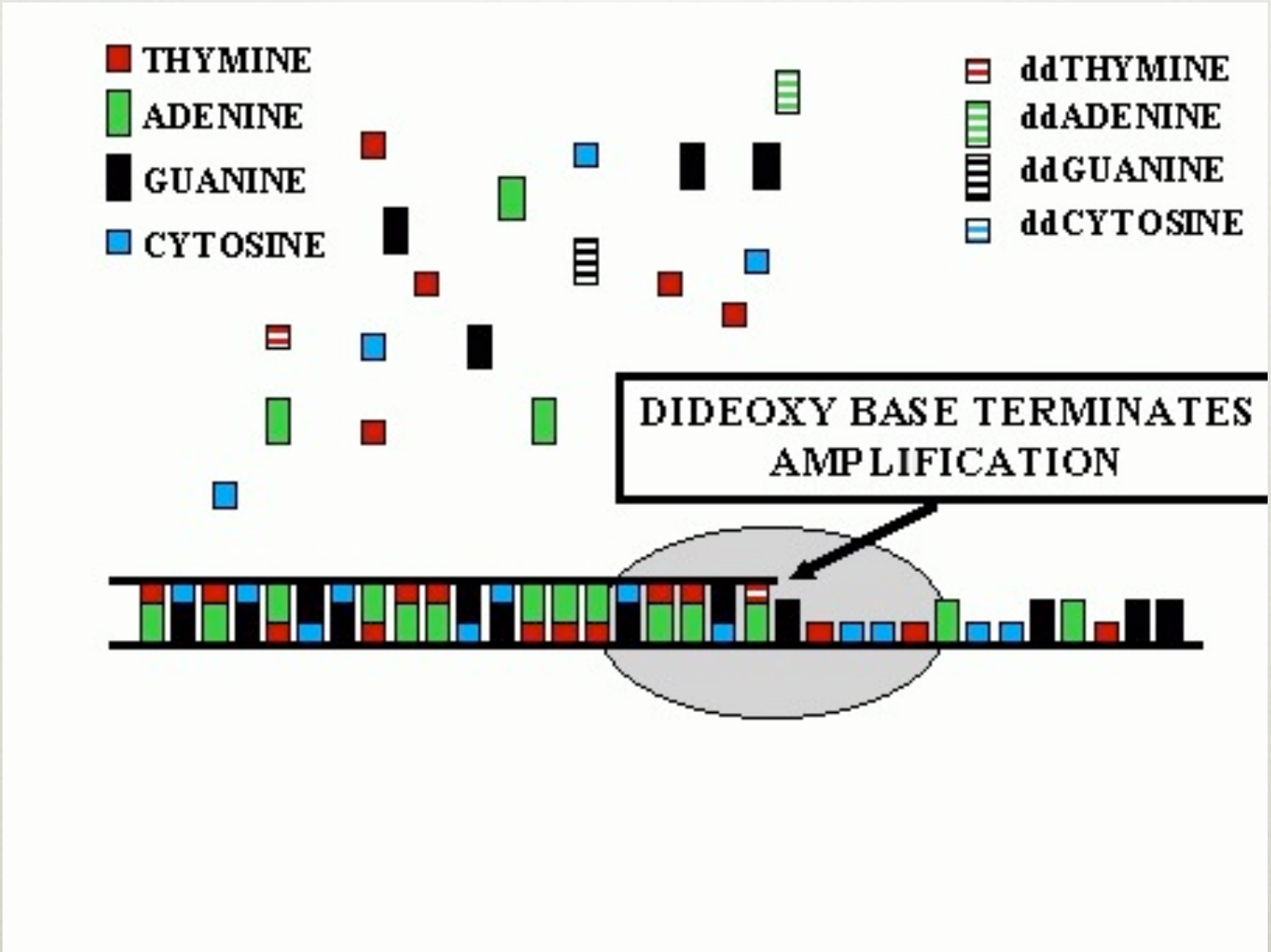
Sanger

ELONGATION AFTER PRIMER



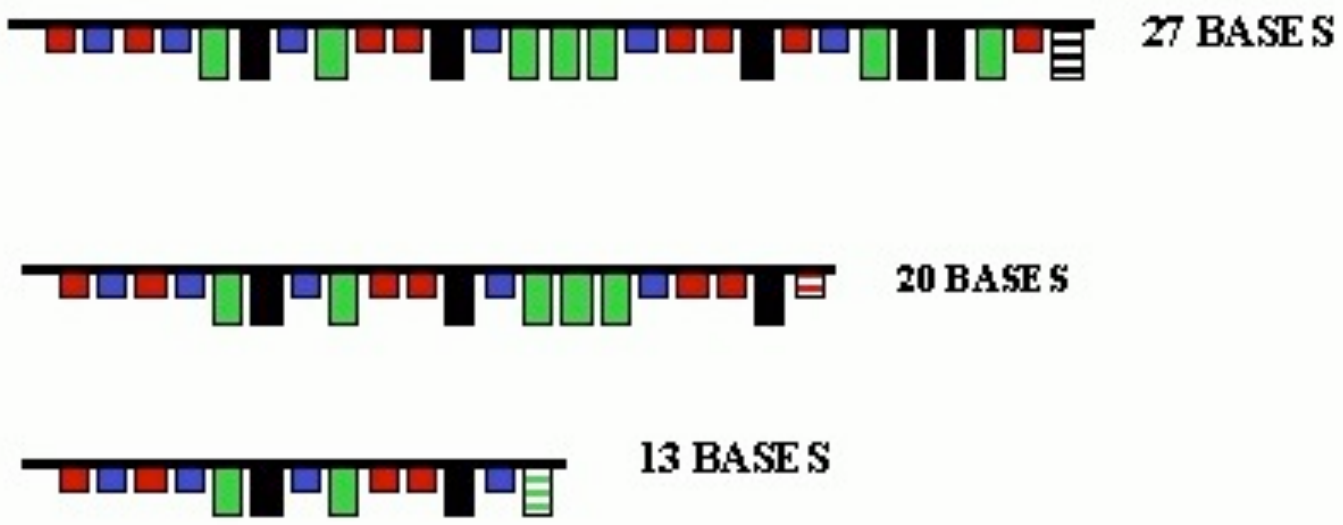
http://web.utk.edu/~khughes/SEQ

ENLONGATION STOPS WHEN DIDEXOY BASE IS ENCOUNTERED



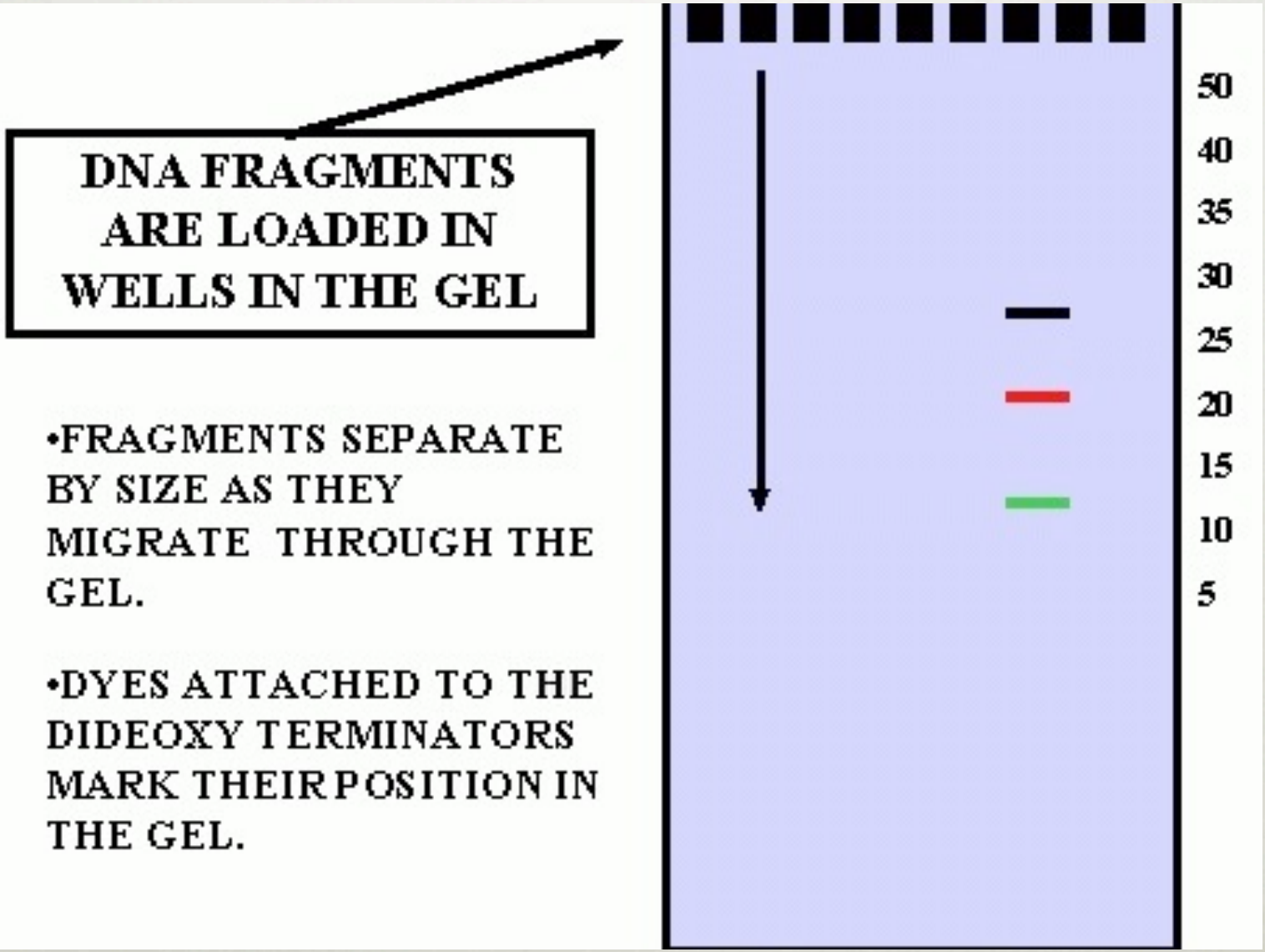
http://web.utk.edu/~khughes/SEQ

PRODUCING A LADDER



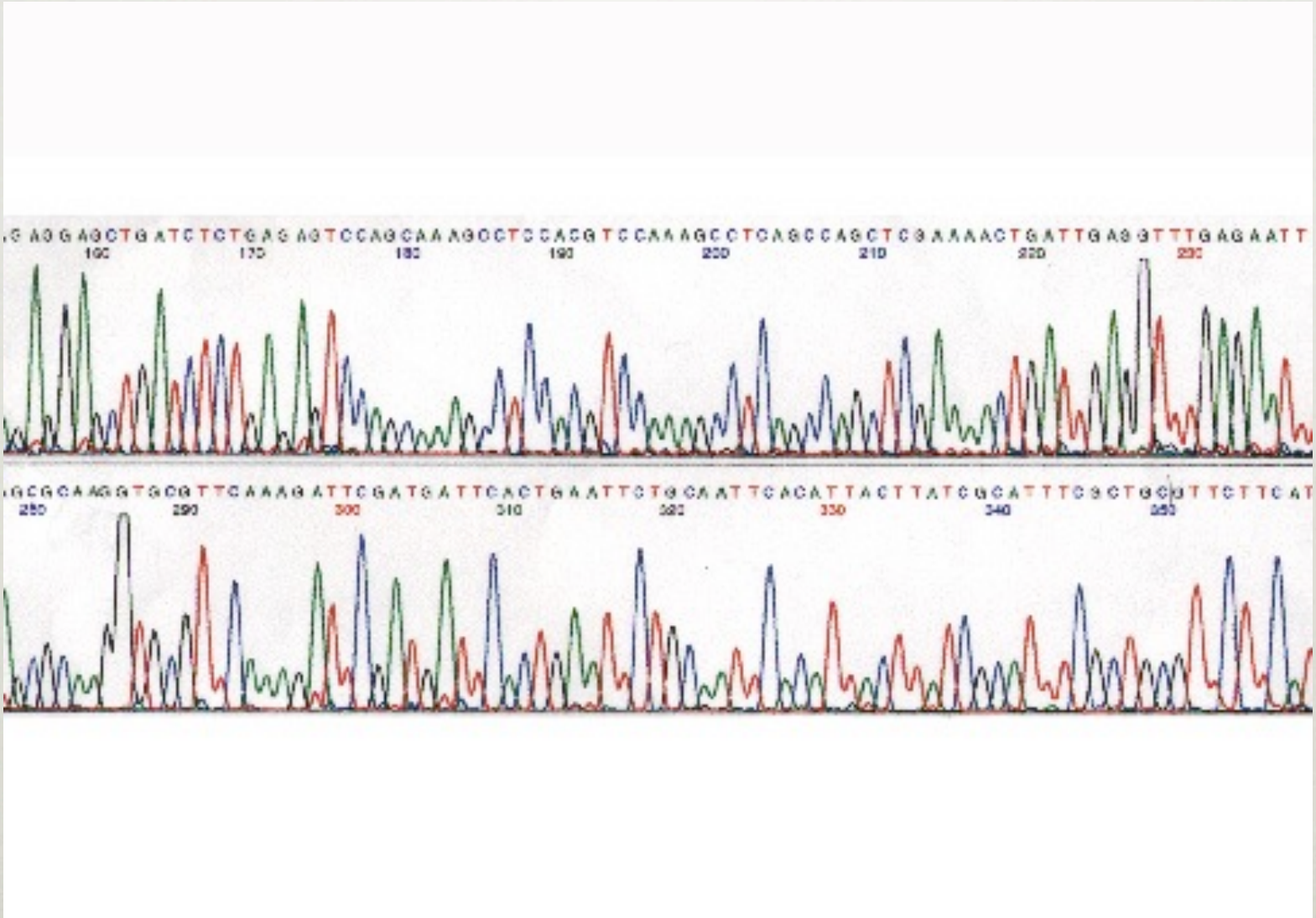
http://web.utk.edu/~khughes/SEQ

THAT CAN BE READ ON GEL



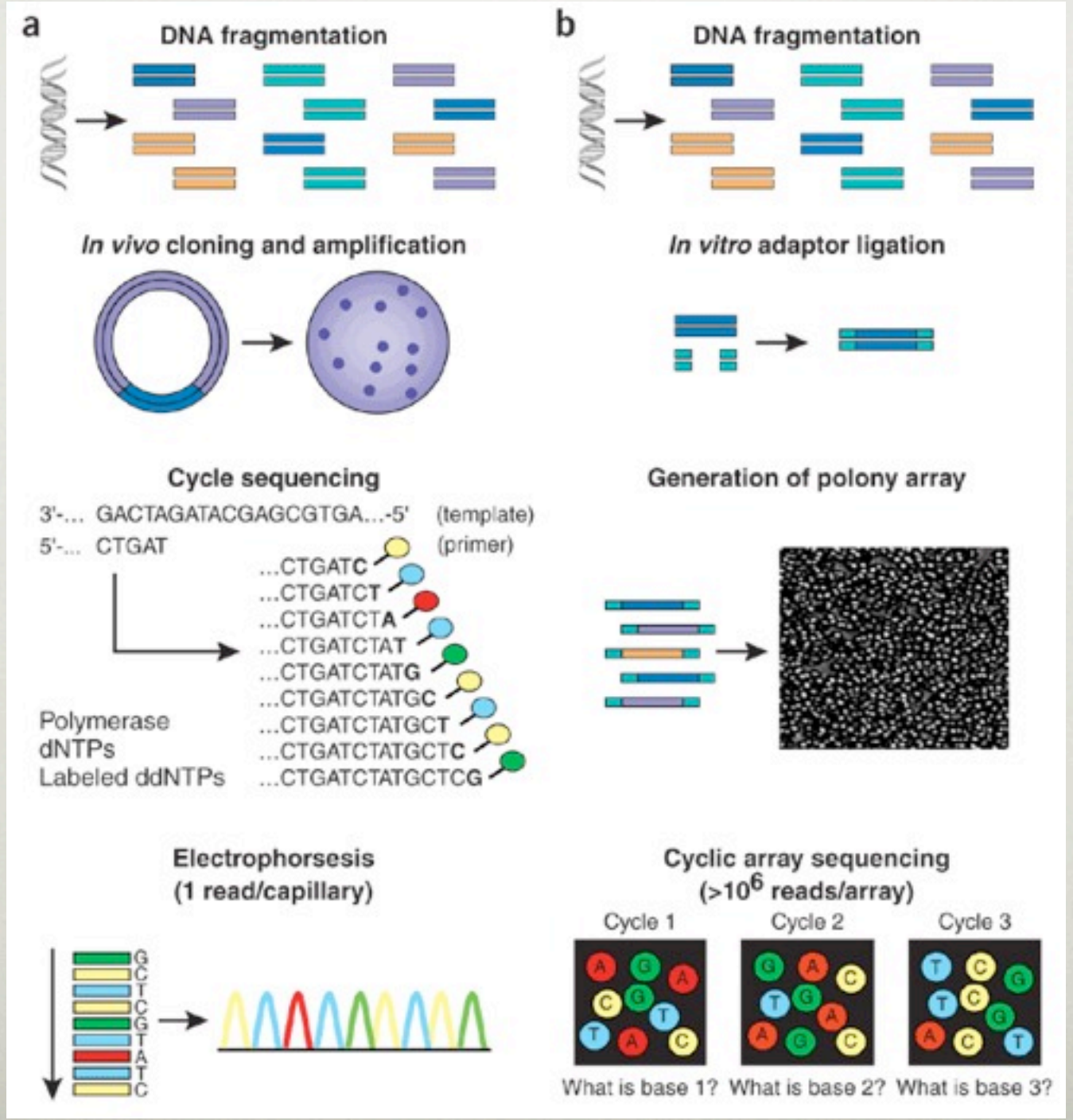
http://web.utk.edu/~khughes/SEQ

TRACING OF THE LADDER



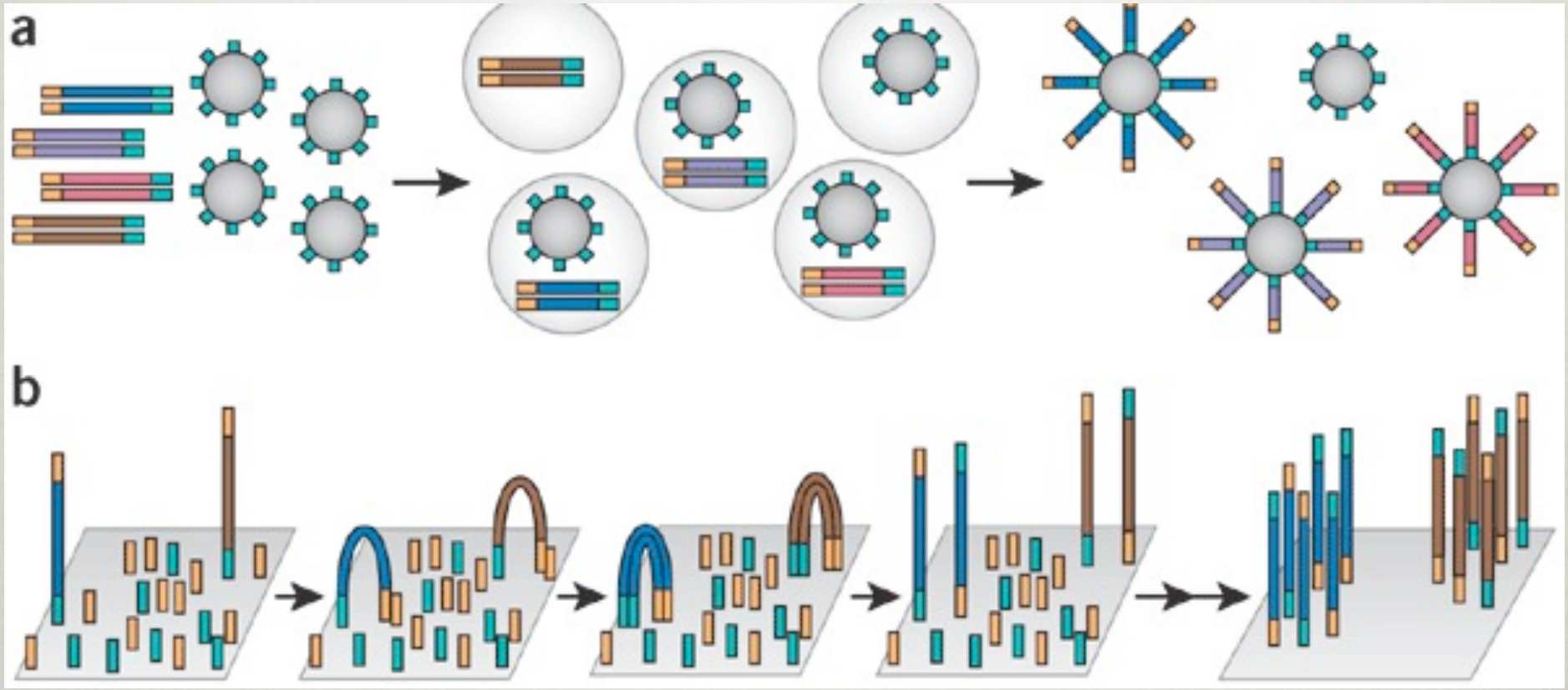
http://web.utk.edu/~khughes/SEQ

Sanger Next Gen



Nature Biotech. (2008) 26: 1135

TWO METHODS OF SINGLE MOLECULE PCR



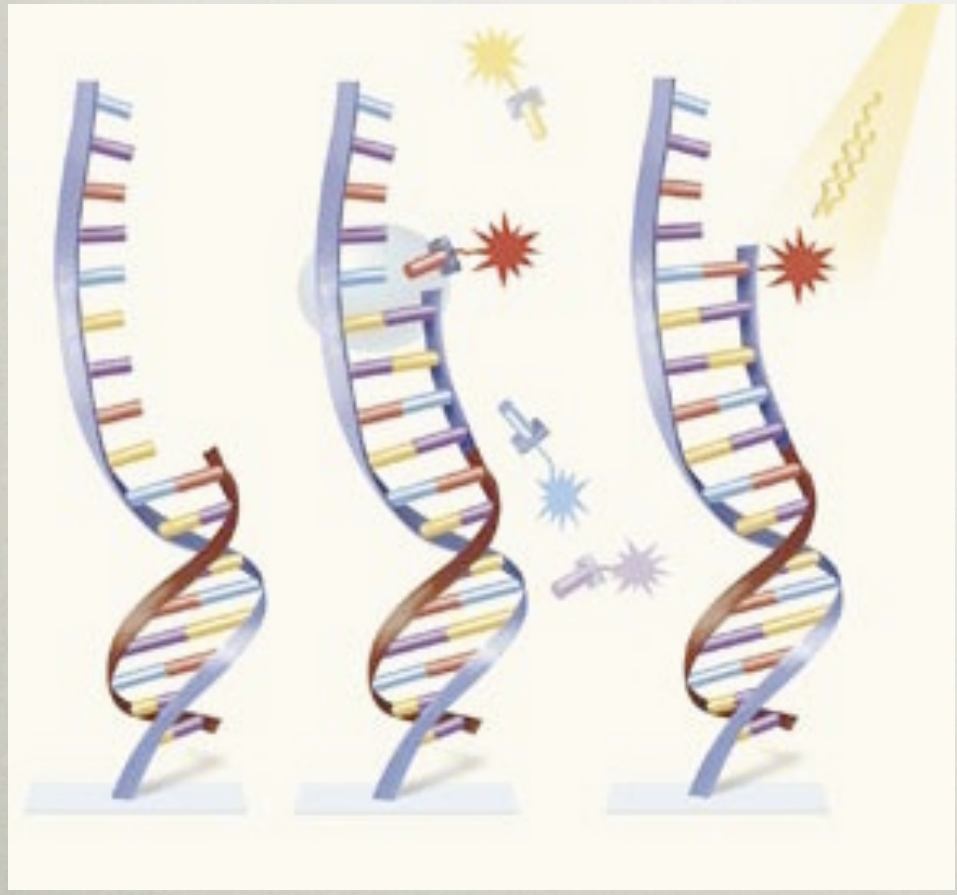
Nature Biotech. (2008) 26: 1135

a: emulsion PCR (454 & SOLiD)

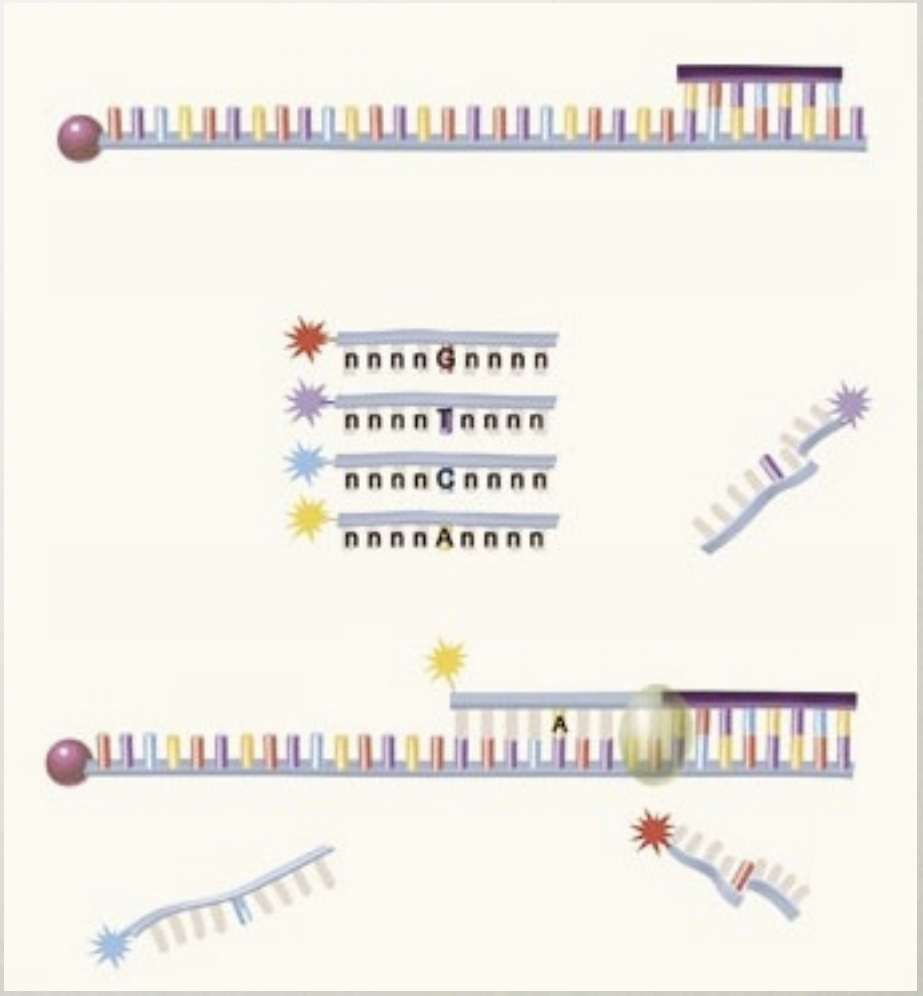
b: bridge PCR (Solexa)

SEQUENCING BY SYNTHESIS

base extension

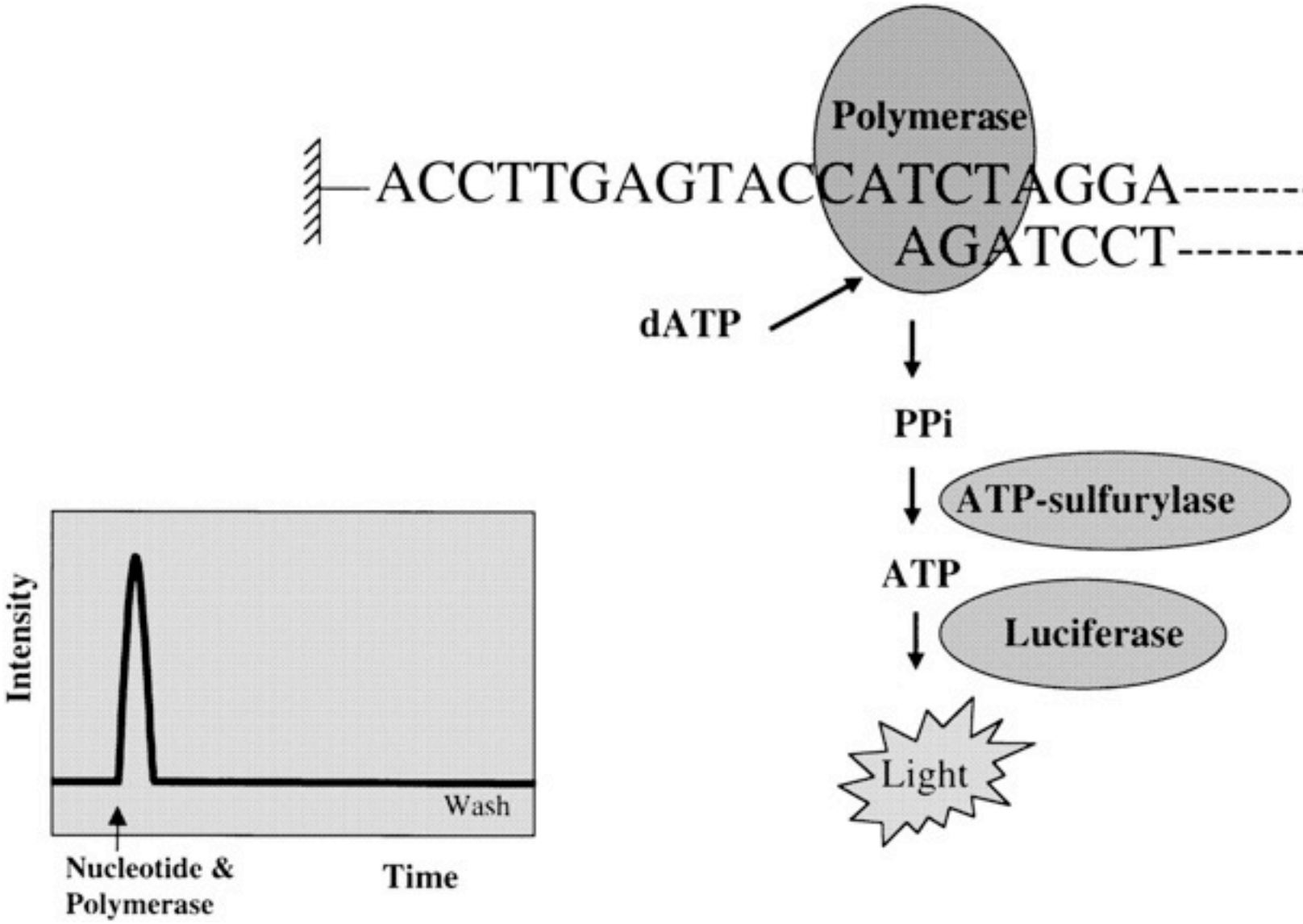


ligation



454

Schematic representation of the progress of the enzyme reaction in solid-phase pyrosequencing



Ronaghi M. Genome Res. 2001;11:3-11

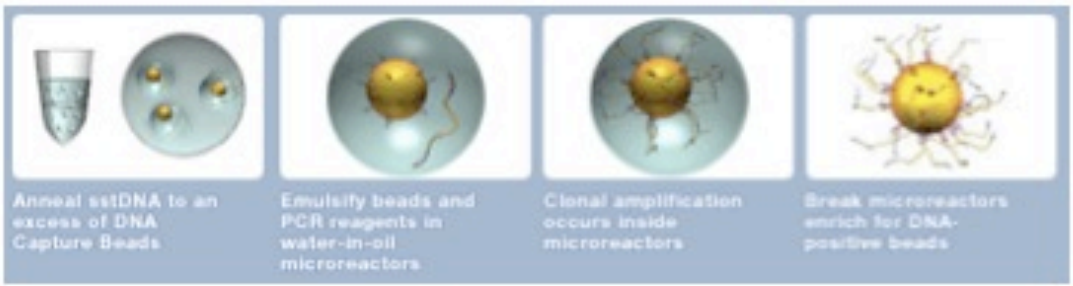
Pyrosequencing

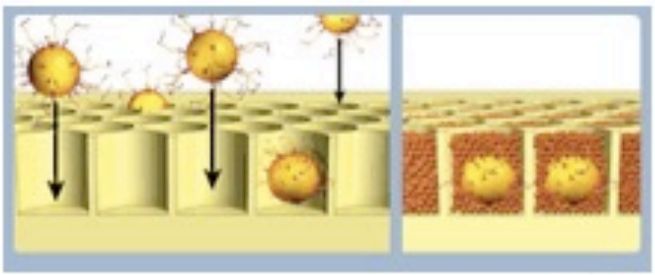
whole genome amplification

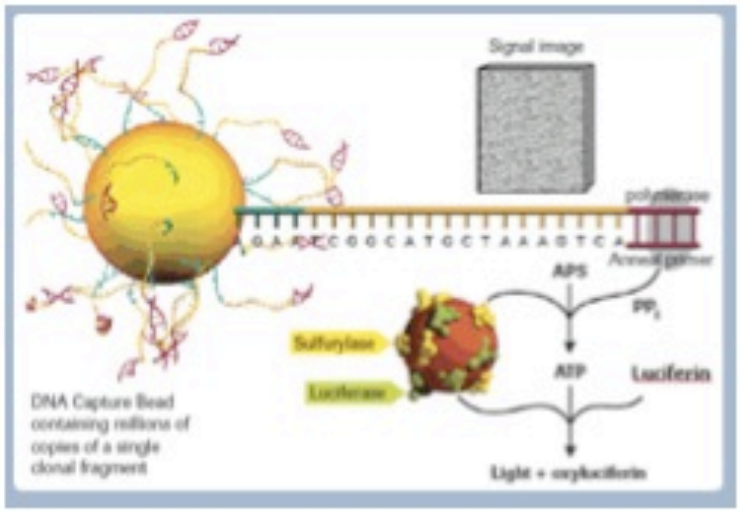
5-100ng DNA

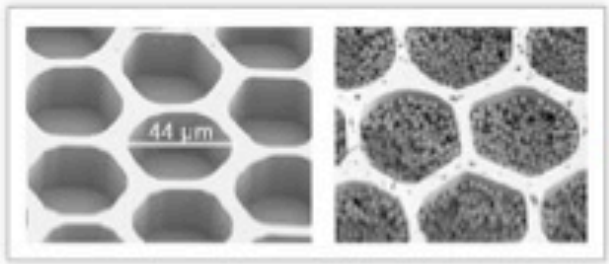
2-5 µg DNA

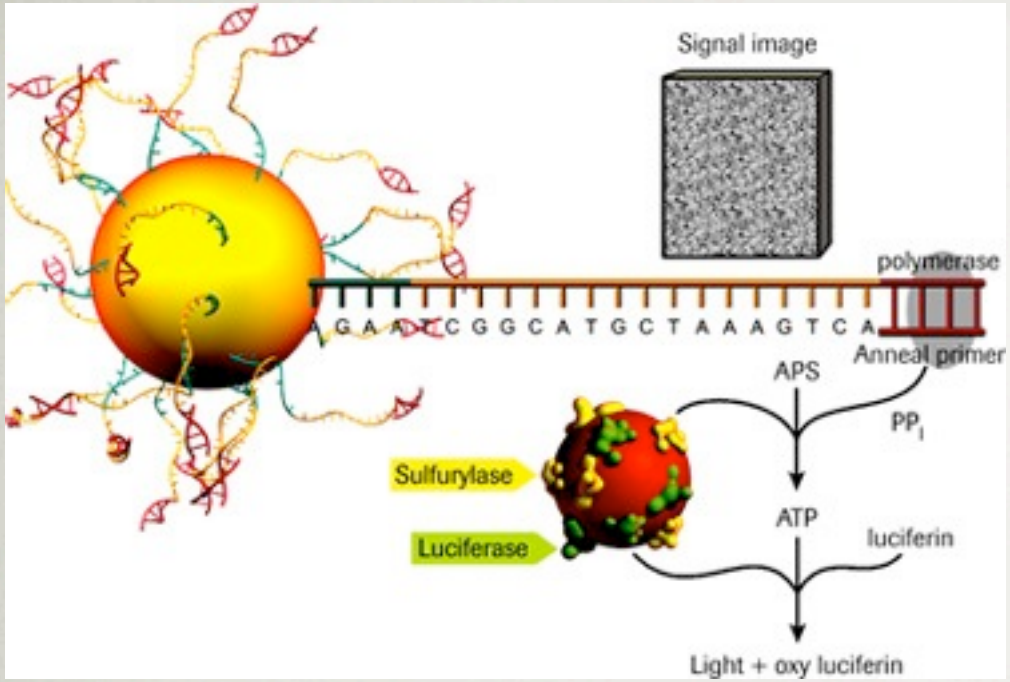
www.454.com

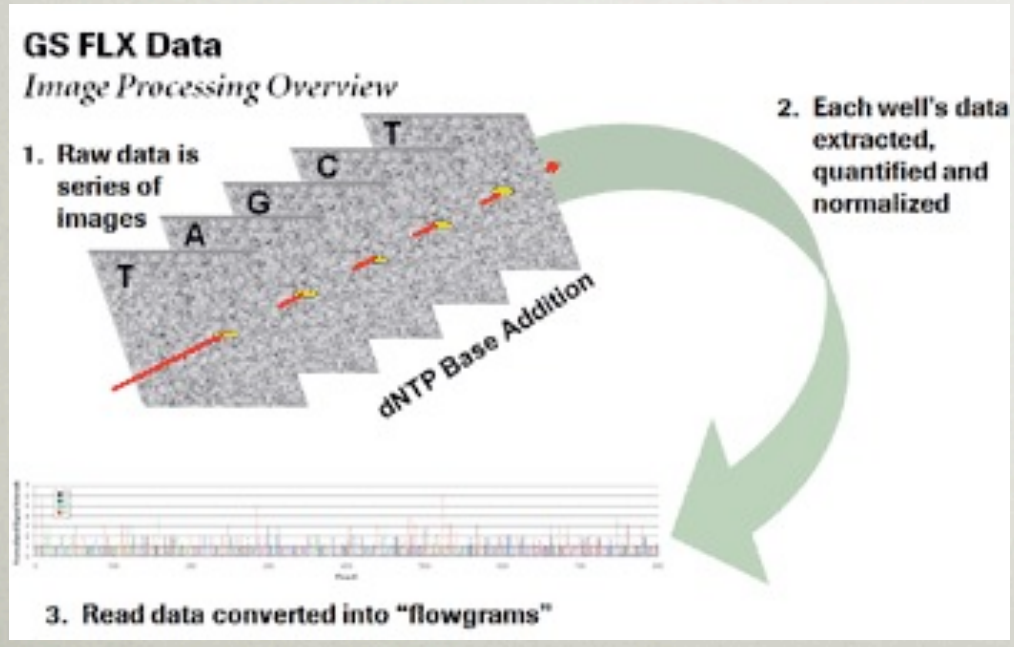




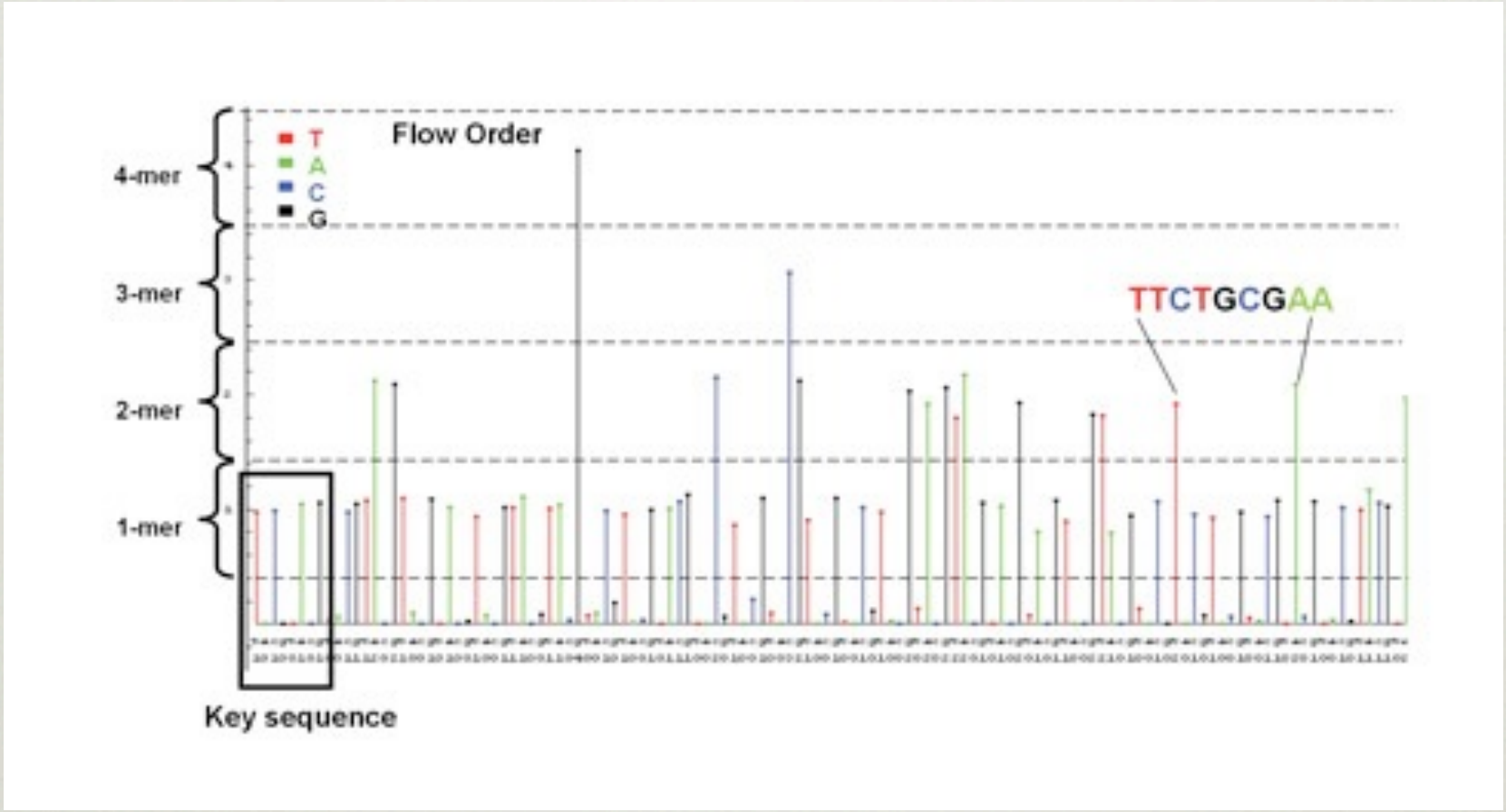





Signal image
polymerase
A G A T C G G C A T G C T A A A G T C A
Anneal primer
APS
PPi
Sulfurylase
Luciferase
ATP
luciferin
Light + oxy luciferin

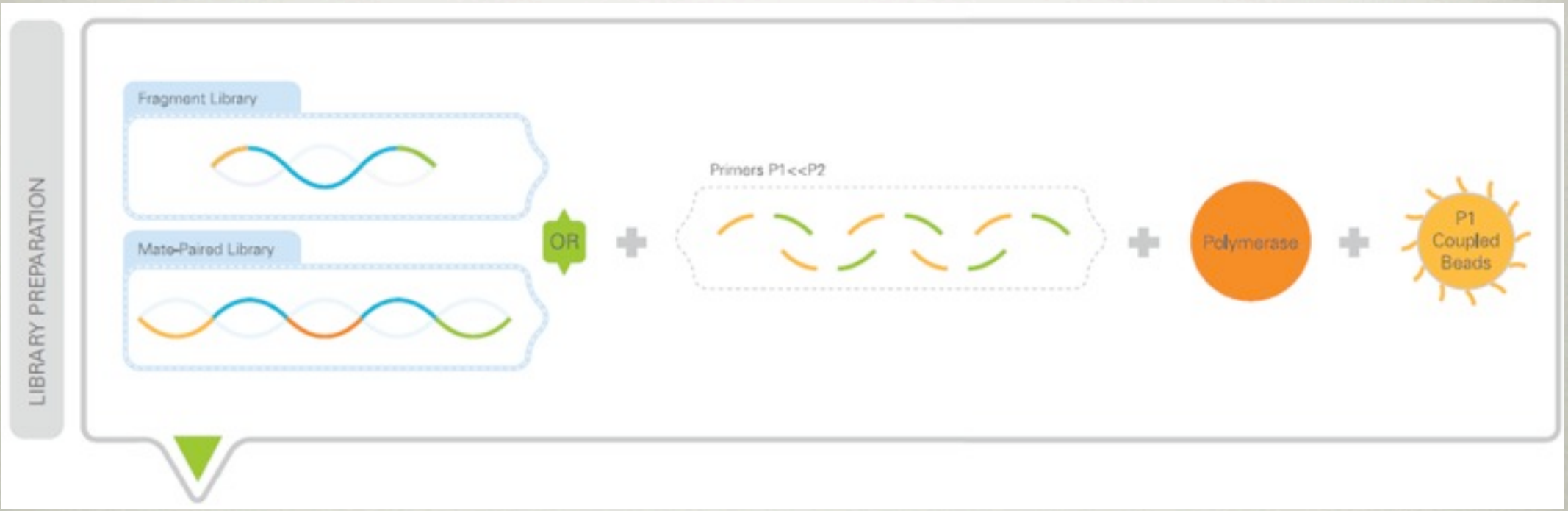

GS FLX Data
Image Processing Overview
1. Raw data is series of images
T
C
G
A
T
dNTP Base Addition
2. Each well's data extracted, quantified and normalized
3. Read data converted into "flowgrams"

Flowgram

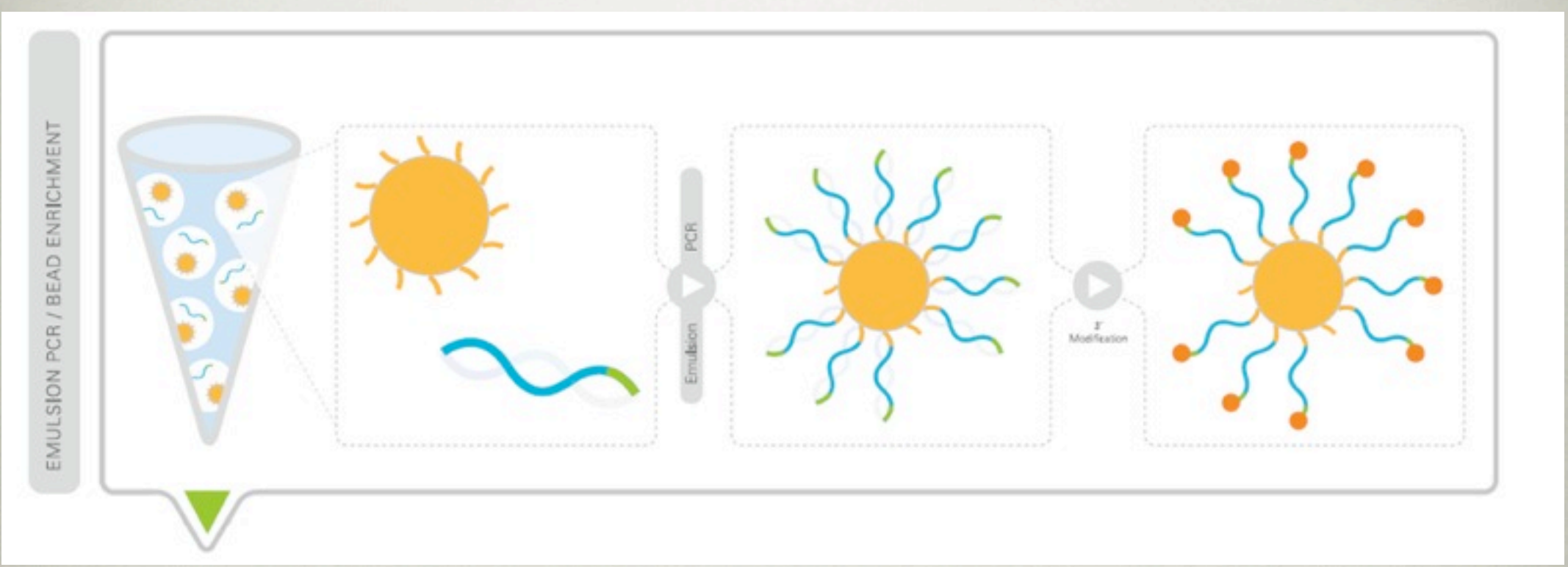


ABI SOLiD

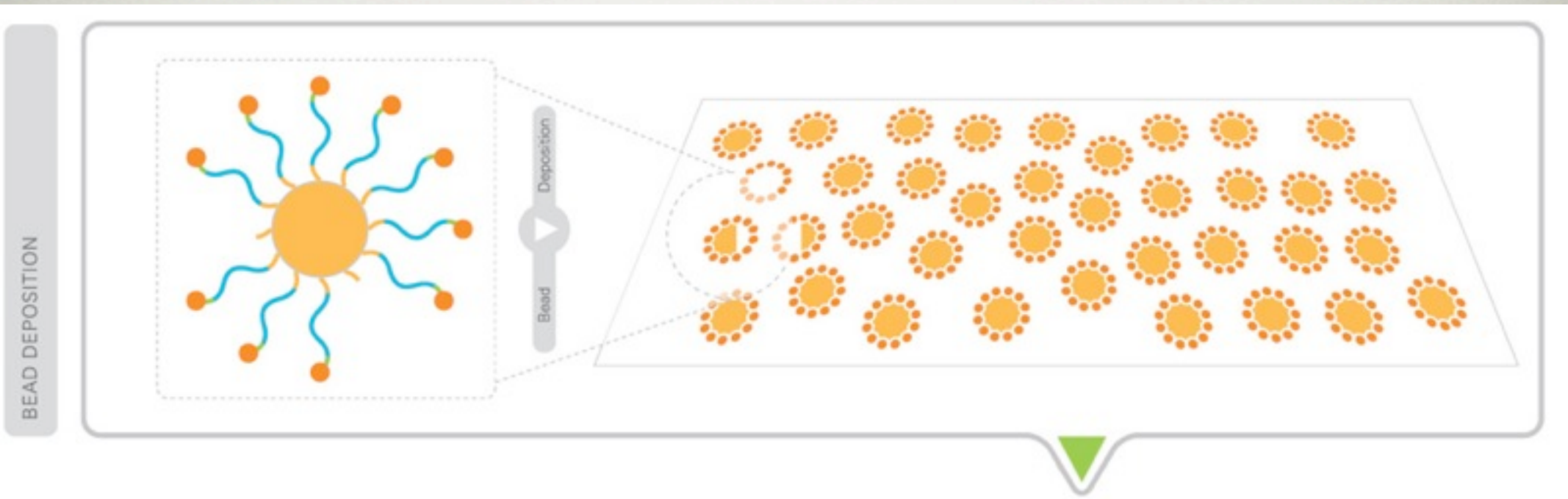
PREPARE LIBRARY OF SINGLE STRANDED DNA



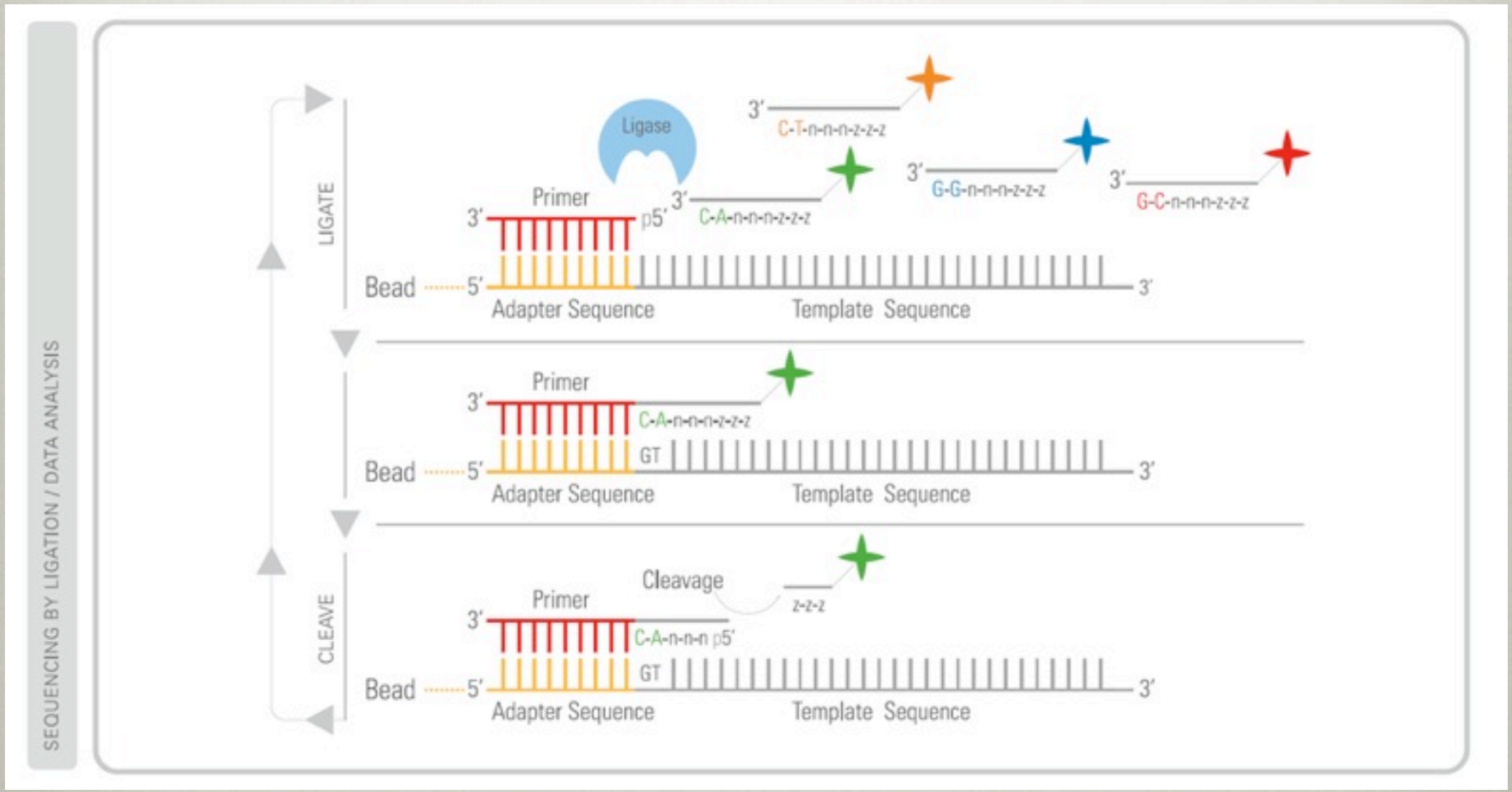
SINGLE MOLECULE PCR



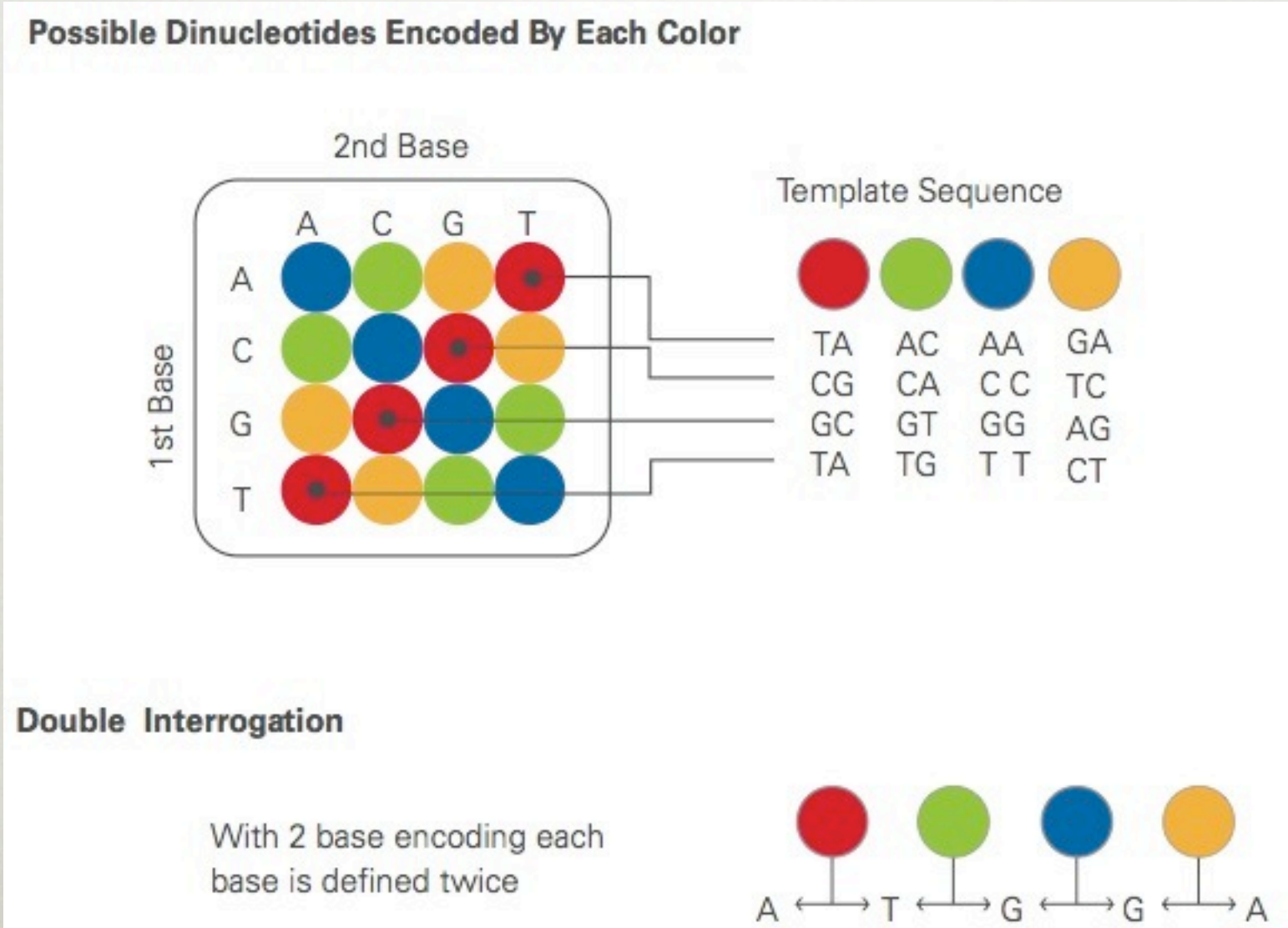
Beads on surface



ABI SOLiD Sequencing

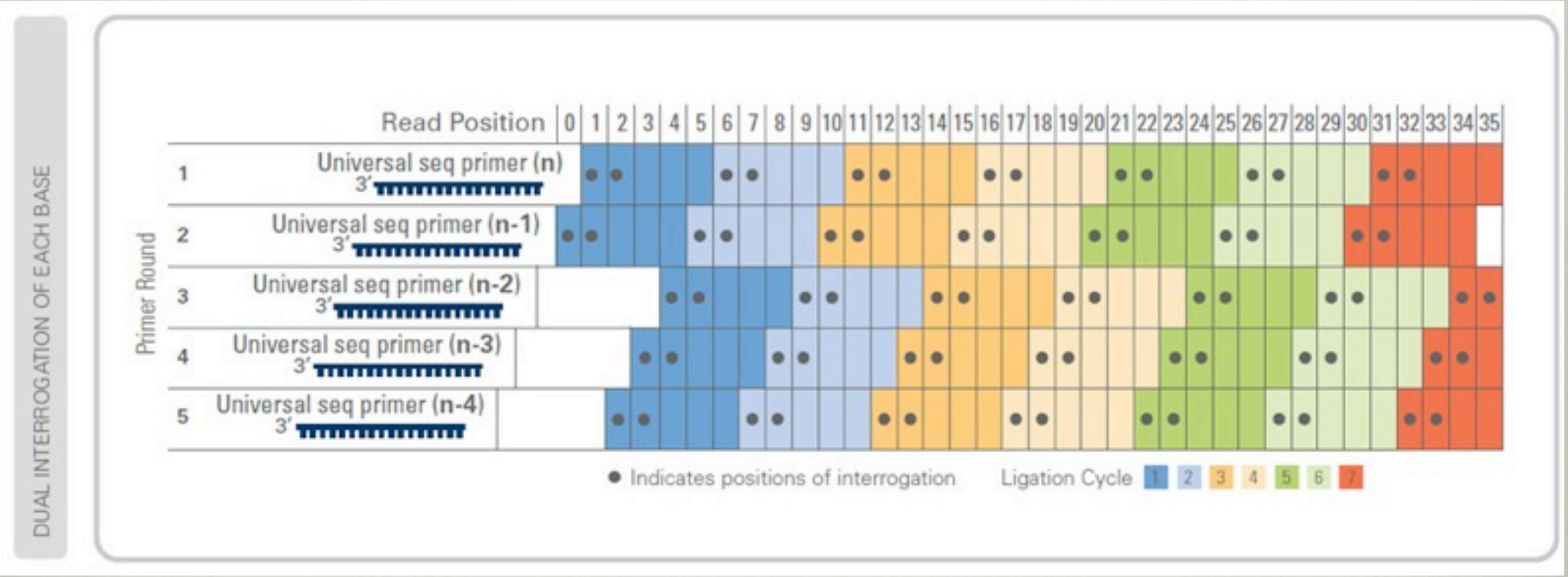


16 di-nucleotides probes in 4 steps



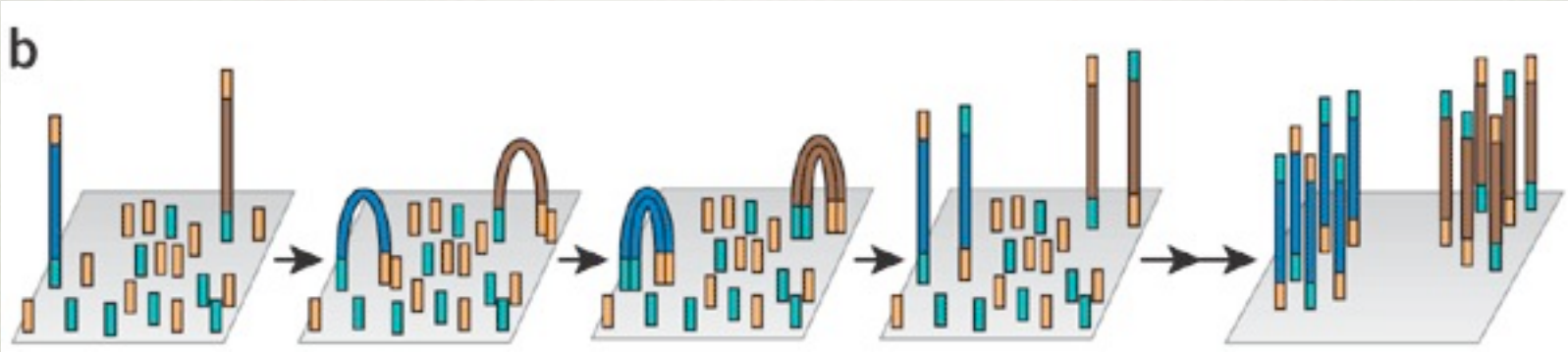
ABI SOLiD

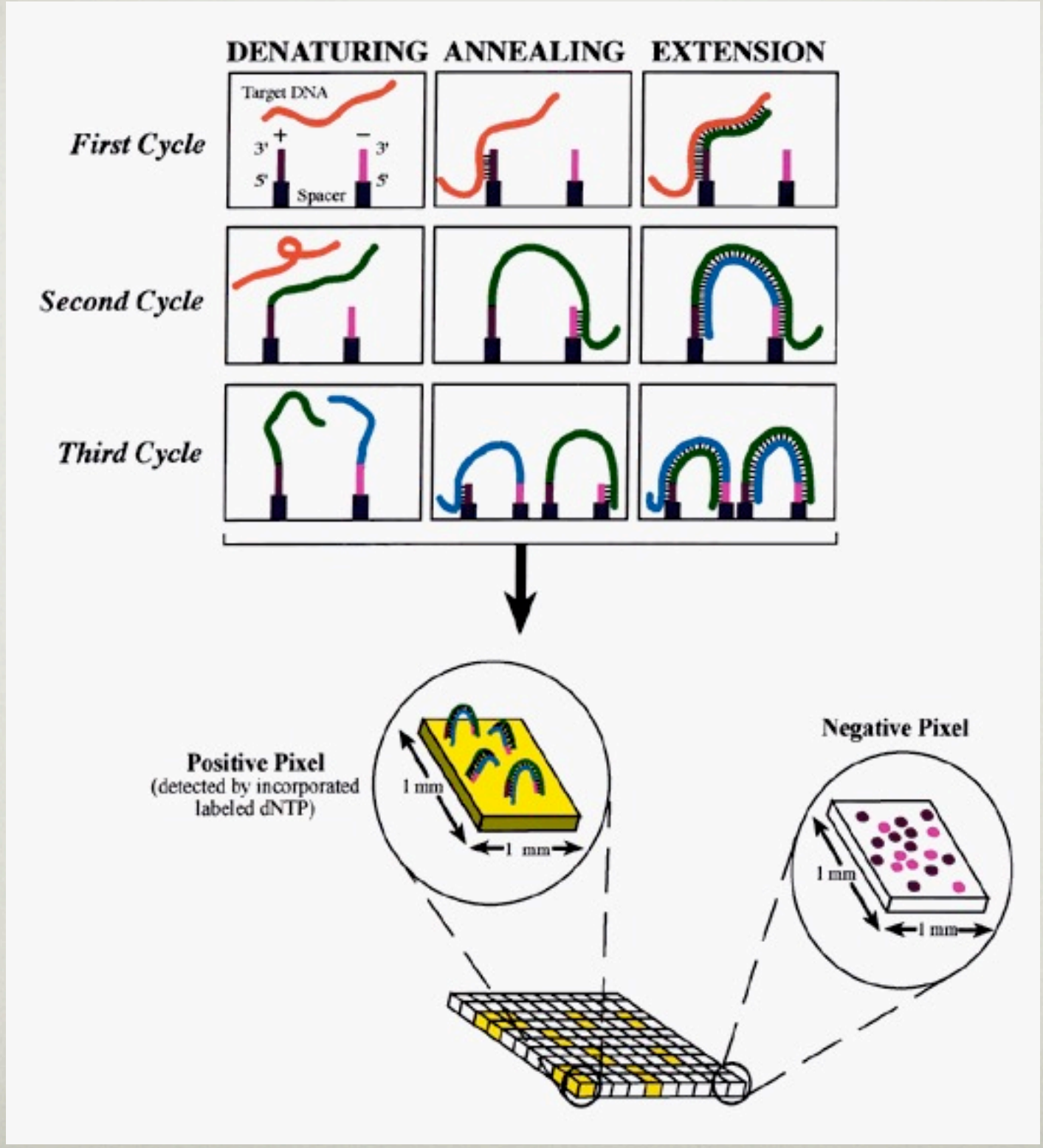
CYCLING



Illumina Solexa

BRIDGE PCR




DENATURING
ANNEALING
EXTENSION
First Cycle
Second Cycle
Third Cycle
Target DNA
+
−
3'
3'
5'
5'
Spacer
Positive Pixel
(detected by incorporated
labeled dNTP)
1 mm
1 mm
Negative Pixel
1 mm
1 mm

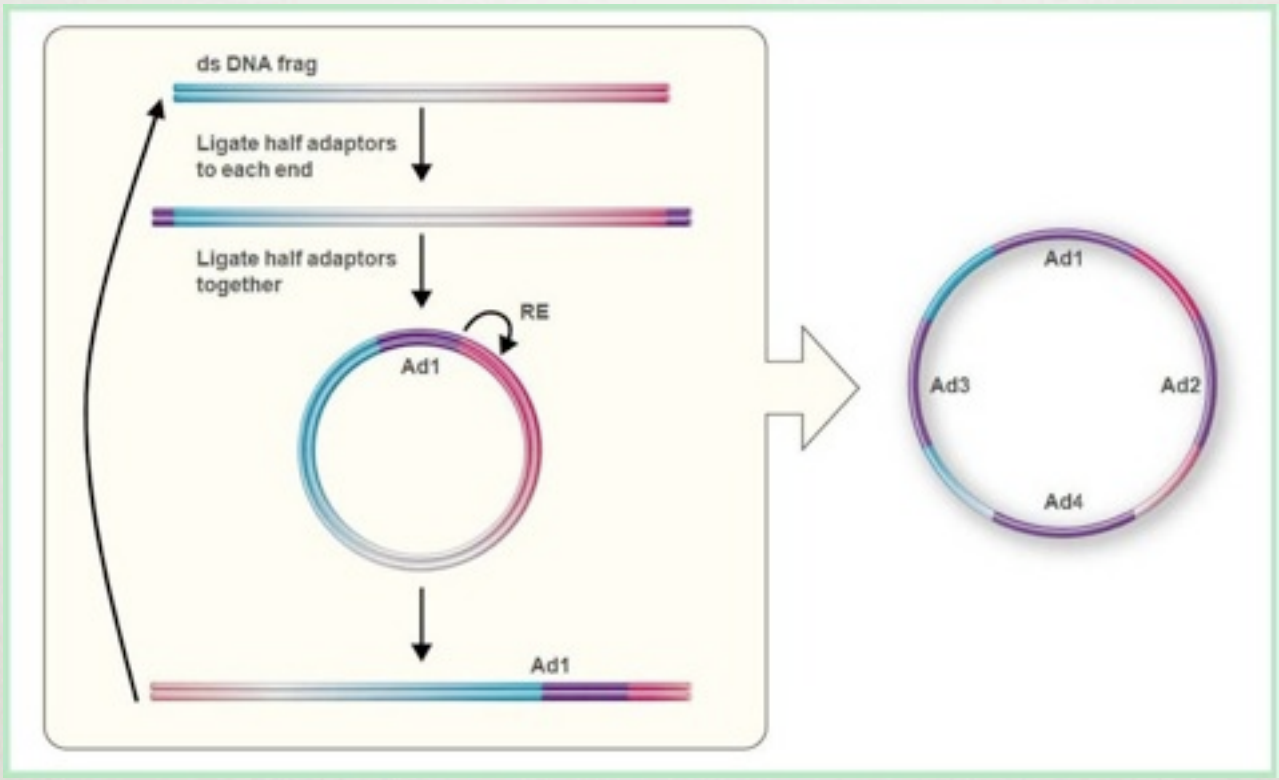
- 8 lanes per run
- 200 pictures per lane
- 4X36 pictures for 36-mer
- 1/4 million pictures per 3-day run -> 0.5TB of data

new HiSeq 2000

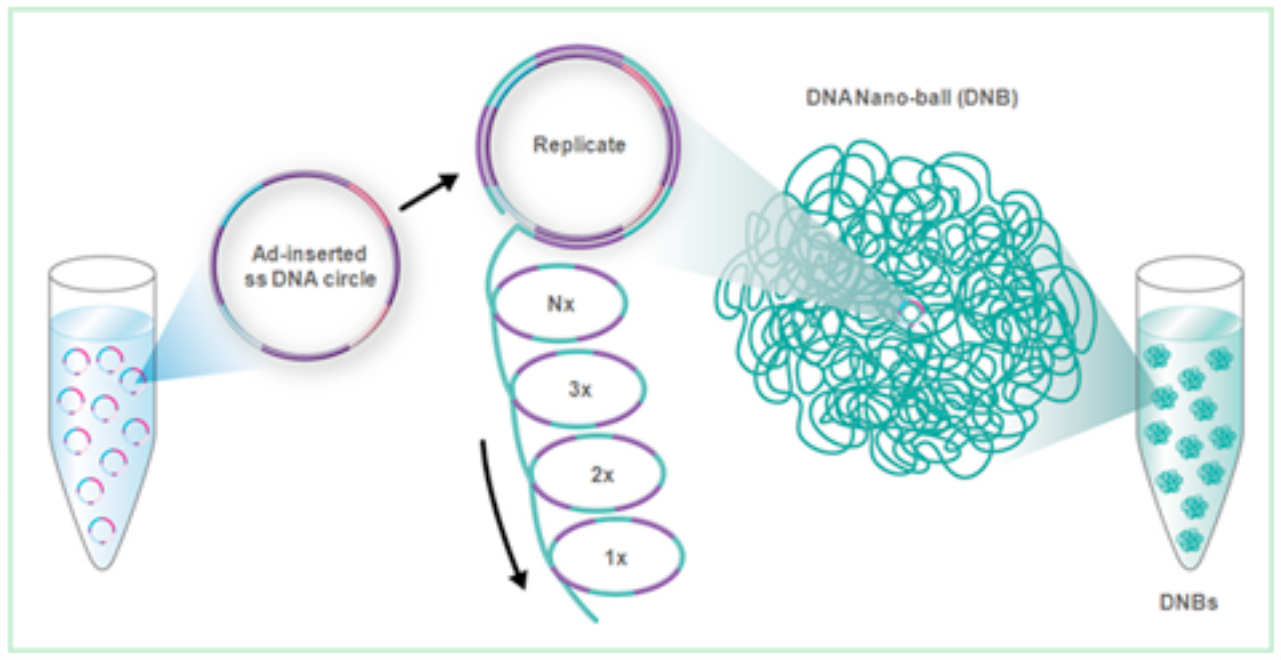
- sequence up to 100bp
- 1 billion tags per experiment
- 25Gbase per day
- reagent cost is about 10 times cheaper than the current product

Complete Genomics

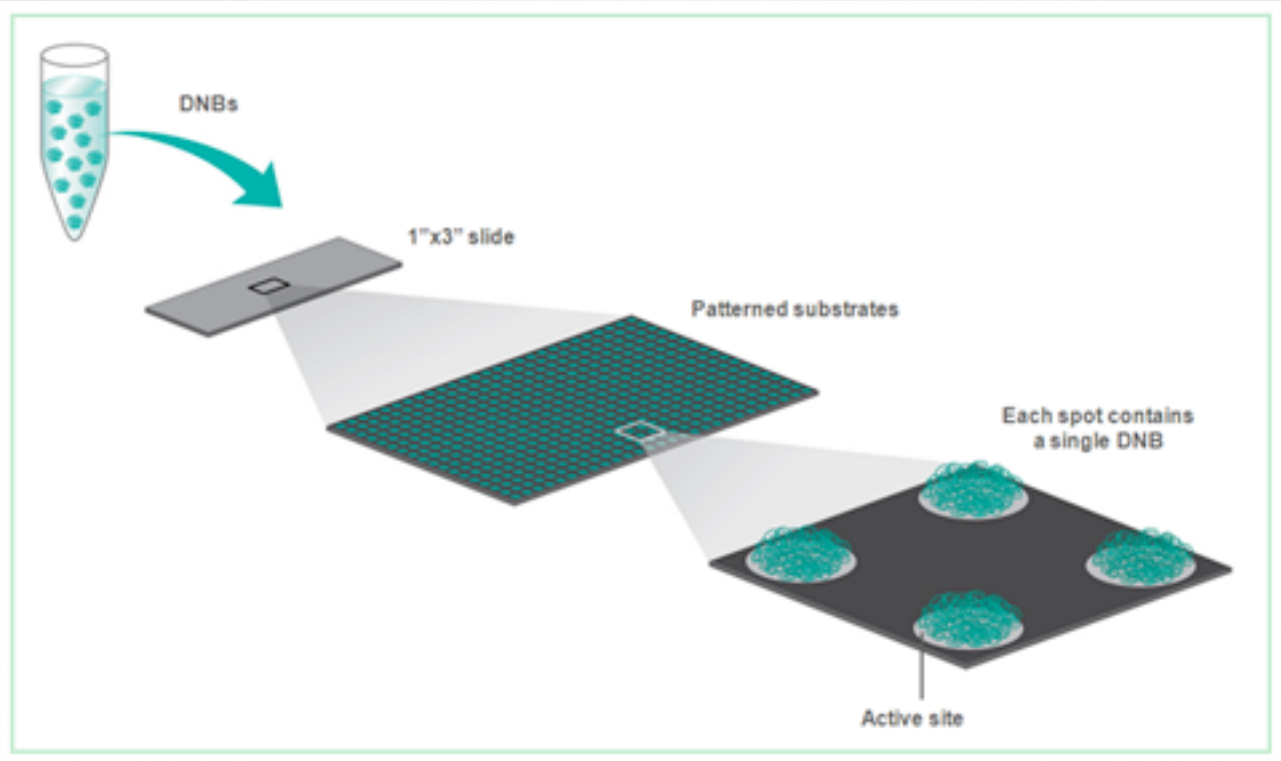
CONSTRUCT A CIRCULAR DNA



ROLLING CIRCLE AMP



HIGH DENSITY PACKING

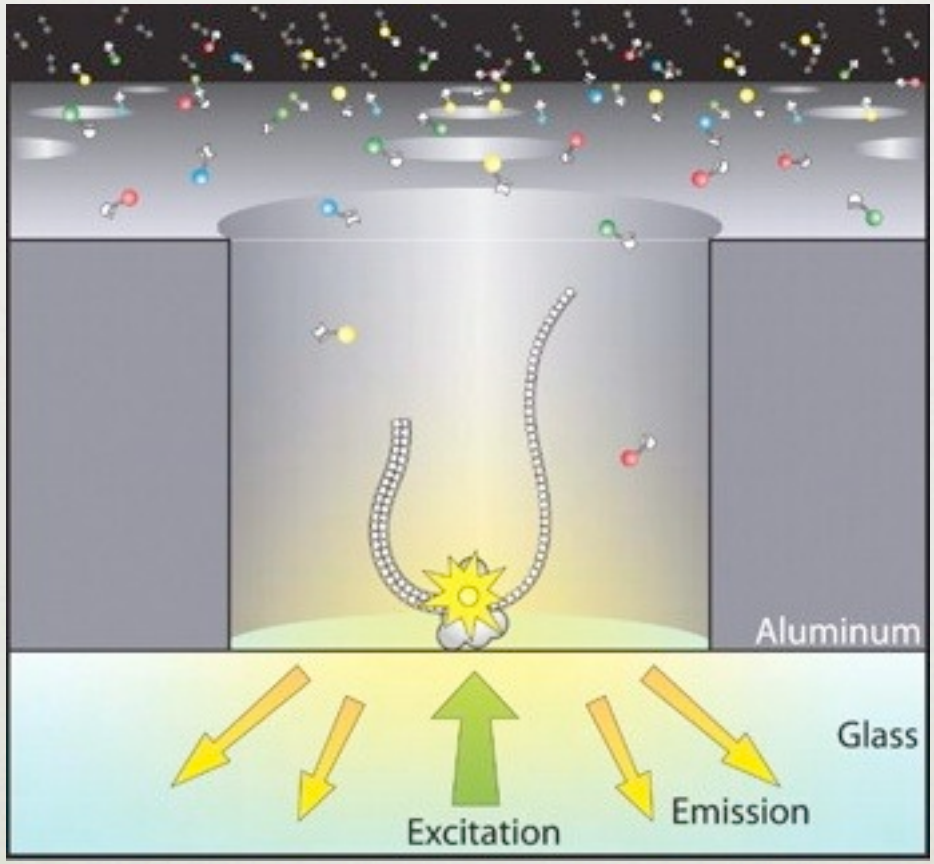


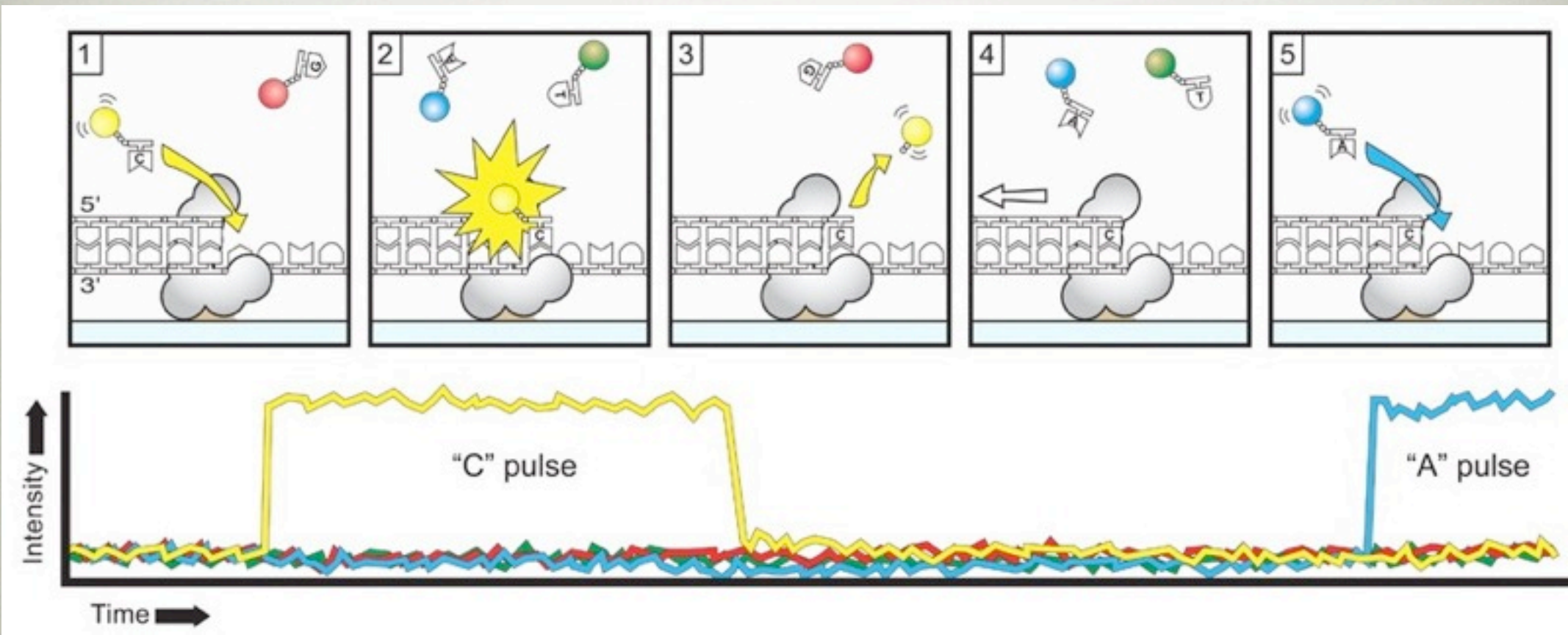
- each DNA nano-ball is 80 bp genomic DNA (plus 4 adaptors) repeated 200 times.
- reads are equivalent to two 35-bp on paired ends of 500bp DNA.
- rolling circle amplification replaces emulsion or bridge PCR
- 1″X3″ silicon slide holds one billion DNA nano-balls

- reagent cost is 1/1000 of Solexa
- demonstrated 8.8 Gb per machine run per day.
- a completed genome sequence on company's web site
- June 2009, launch of commercial run: 200Gb per machine run lasting 8 days.
- data center: 60,000 processors and 30 petabytes storage.

- according to Dr Drmanac, CSO of Complete Genomics, Inc
- next generation of machine will have
  - $10 per genome reagent cost
  - $20 per genome of instrument cost

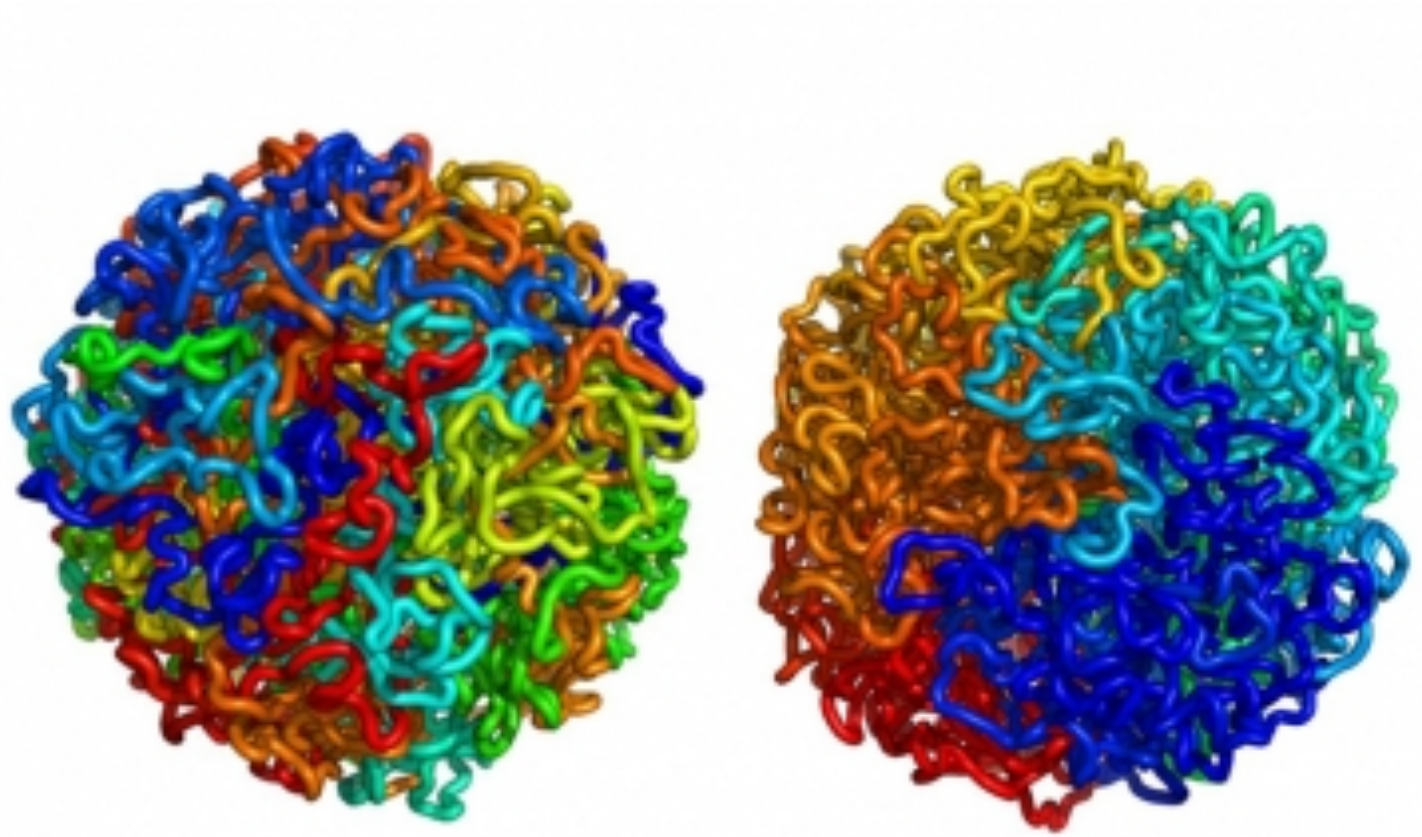
Pacific Biosciences

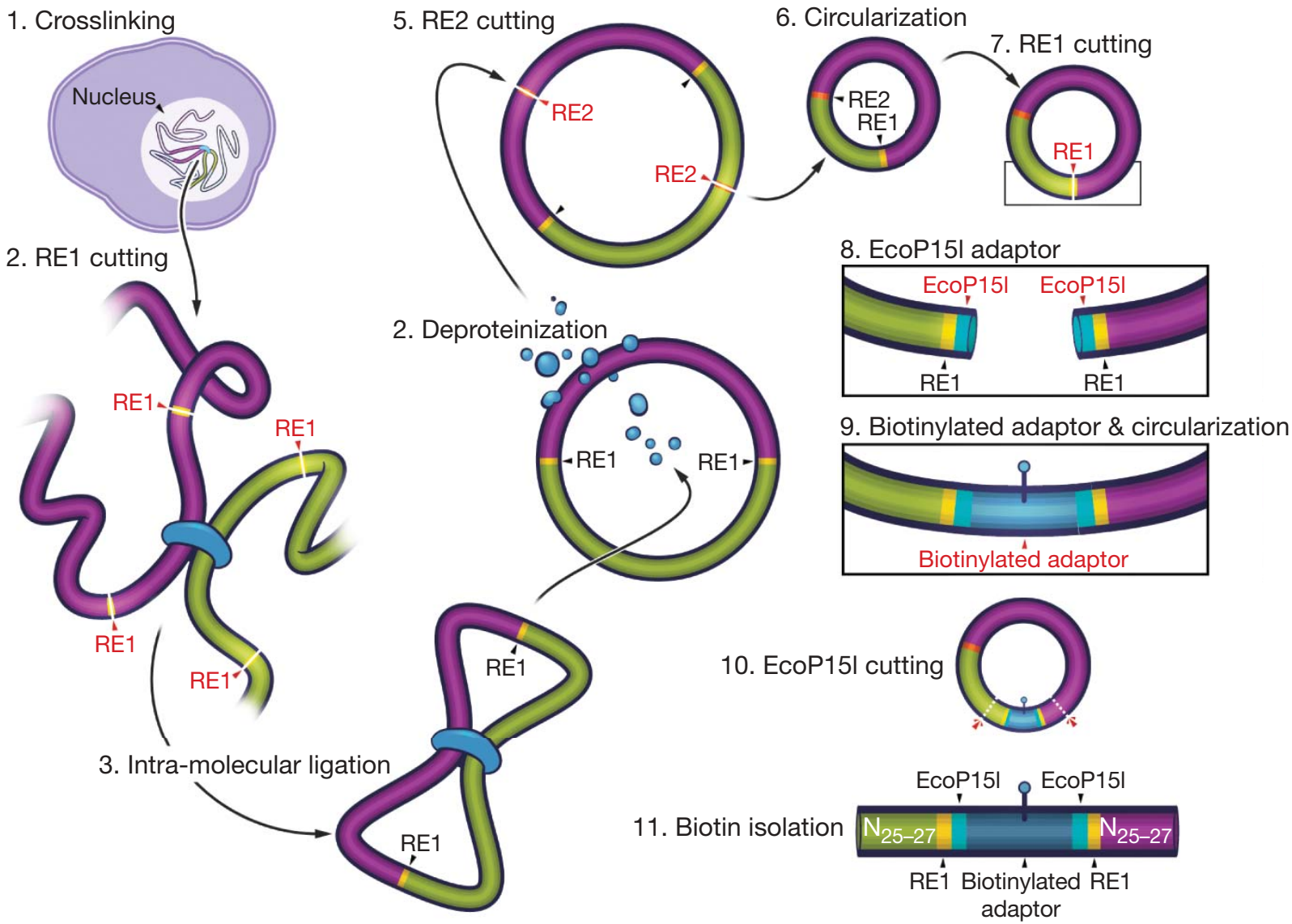

Aluminum
Glass
Excitation
Emission


1
2
3
4
5
5'
3'
Intensity
Time
"C" pulse
"A" pulse

Unlike Sanger sequencing, which average over many molecules, in next gen sequencing PCR errors do not average away

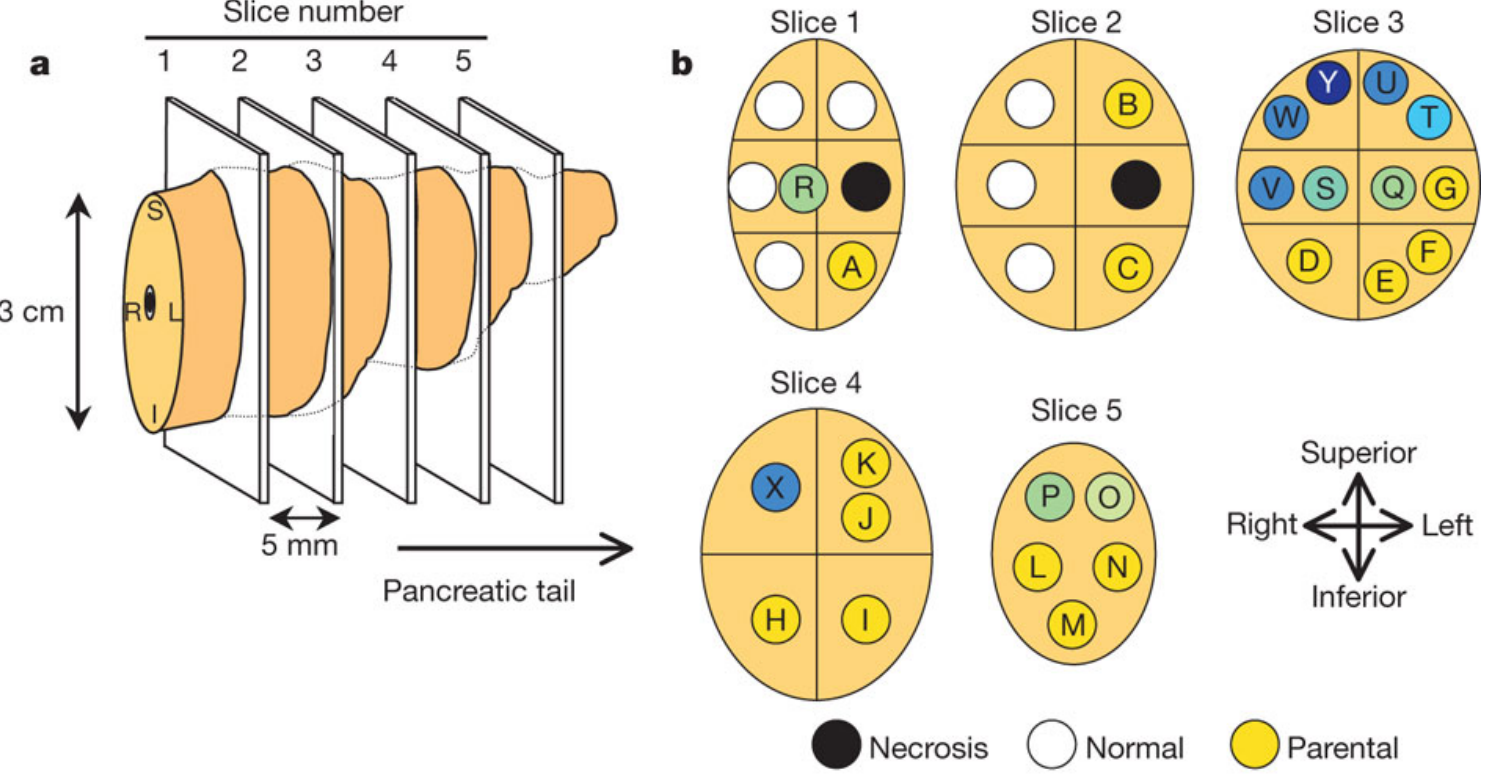
Application: 3D genome

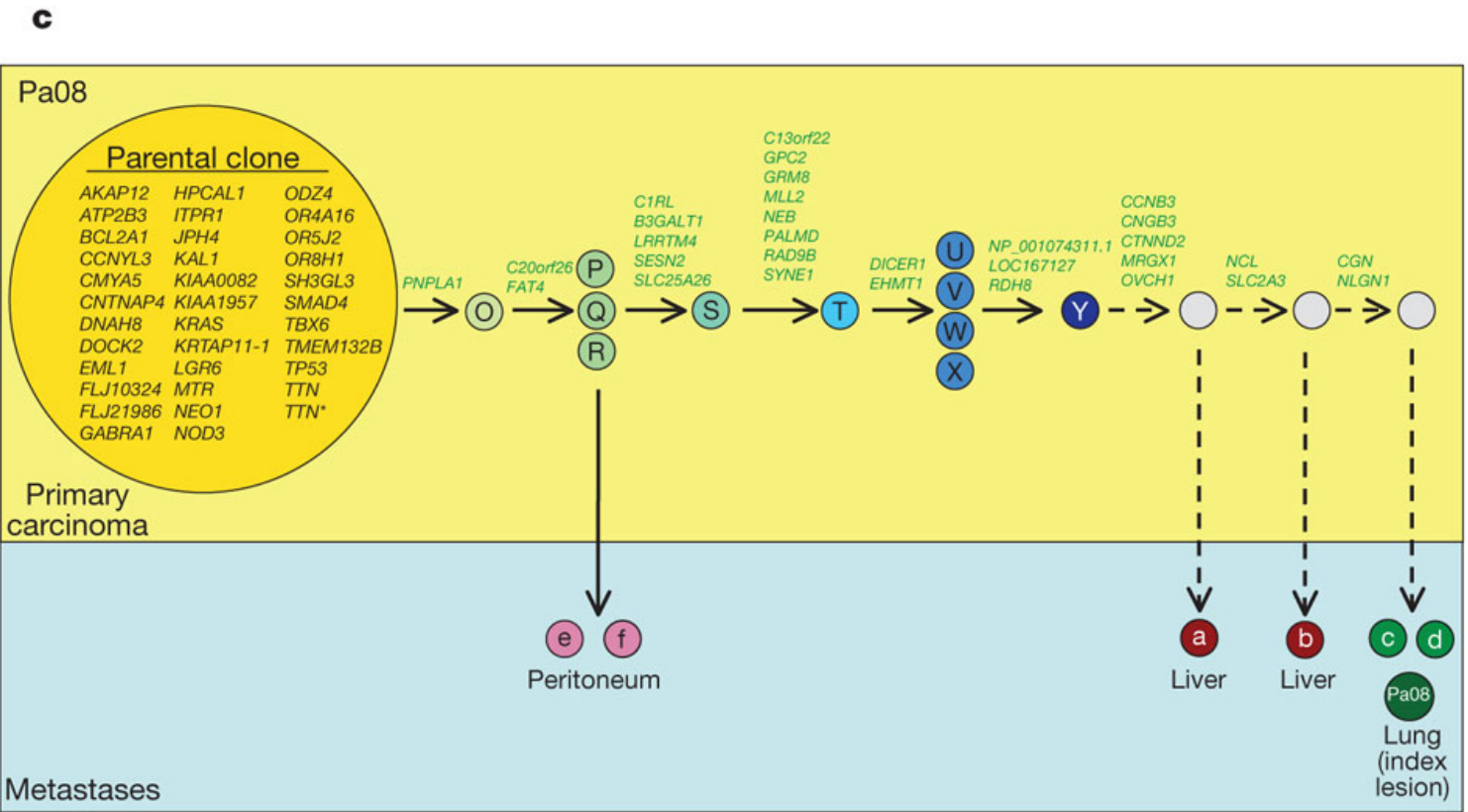




Nature **465** 363 (2010)

Application: tumorigenesis





Geographic mapping of metastatic clones within the primary carcinoma and proposed clonal evolution of Pa08.

S Yachida *et al. Nature* **467**, 1114-1117 (2010) doi:10.1038/nature09515

- happy thanksgiving