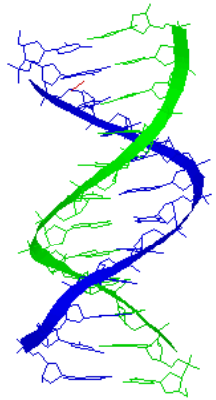
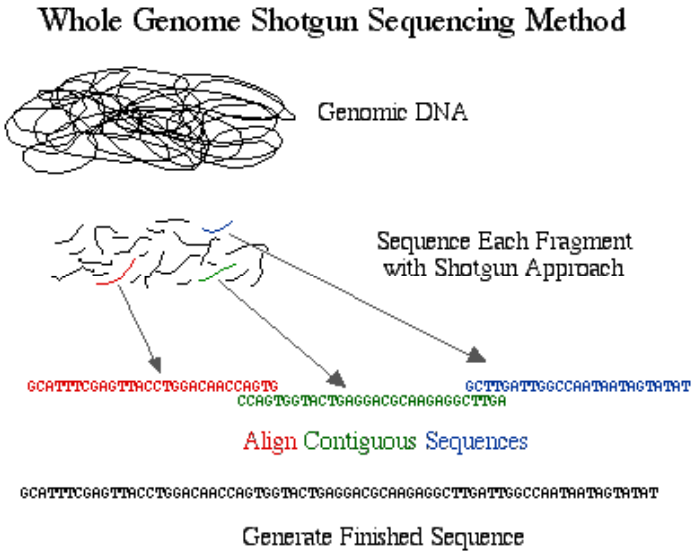
Sequencing among techniques of the molecular biology

DNA-sequencing



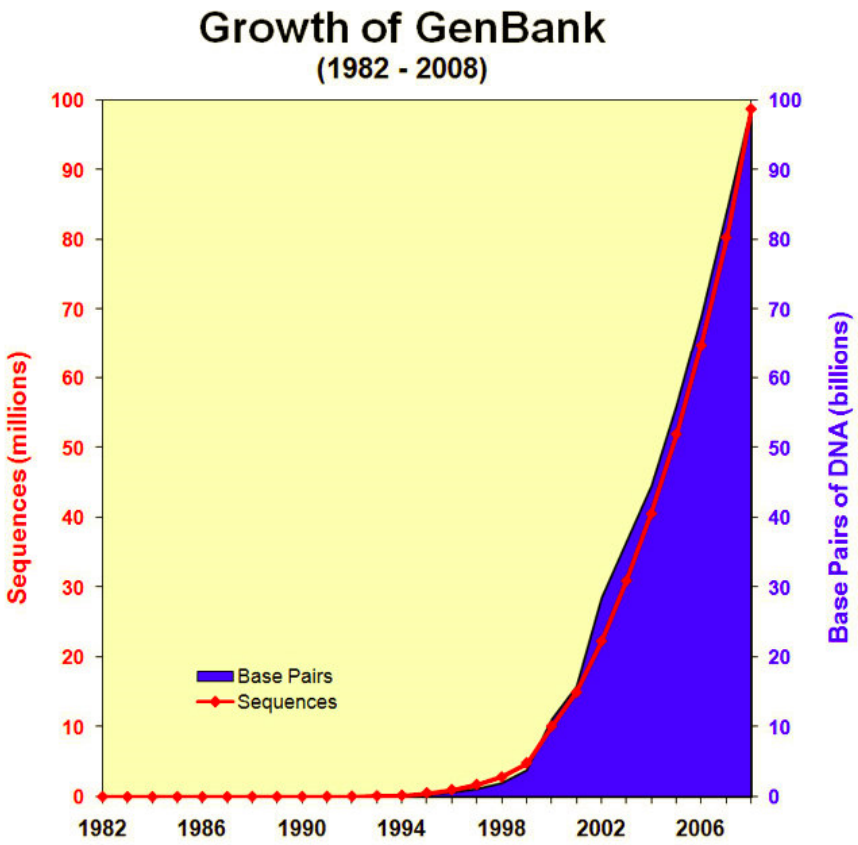
the determination of the precise sequence of nucleotides in a sample of DNA



The sequencing of the **reference human genome**

- 1990-2001 International Human Genome Sequencing Consortium; Strategy: hierarchial shotgun sequencing
- 1998- Celera (Craig Venter): whole genome random shotgun method (0.01$ vs 0.3$ cost of clone by clone)

Productivity of sequencing methods according to HGP



http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html

PSG Chain et al. Genome Project Standards in a New Era of Sequencing
Science 9 October 2009: Vol. 326 no. 5950 pp. 236-237

- reference sequence: „average" DNA sequence of a human
- Sequencing centers

Cost reduction:

  - 1990 10 $/base => 3 billion $
  - 2002 0.1 $/base => 3 million $

- Draft sequence, finished grade

(< 1 error per 100,000 bp, assembled into a single contiguous sequence with a minimal exceptions)

- New perspectives: birth of post-genomic era (to understand the huge amount of genomic data and using the understanding to solve real-world problems)
- Traditional genetics ➔ genomics (reduction ➔ expansion)

Increase in sequencing capability : bp/$

First human genome sequenced over 13 years at $2.7 billion.

May 09, 2011
Illumina Lowers Cost of Whole-Genome Sequencing Services

.. has lowered the cost of its human whole-genome sequencing services to $5,000 per genome for projects of 10 samples or more.
http://www.genomeweb.com/

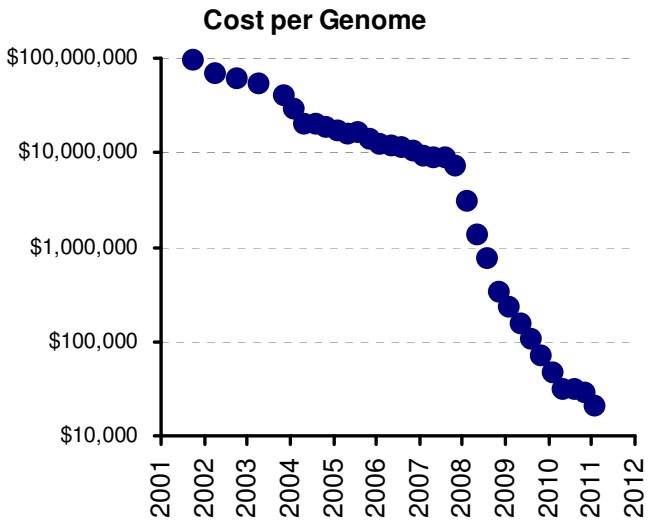
Table 2 | **Sequencing statistics on personal genome projects**

Personal Genome	Platform	Genomic template libraries	No. of reads (millions)	Read length (bases)	Base coverage (fold)	Assembly	Genome coverage (%)*	SNVs in millions (alignment tool)	No. of runs	Estimated cost (US$)
J. Craig Venter	Automated Sanger	MP from BACs, fosmids & plasmids	31.9	800	7.5	*De novo*	N/A	3.21	>340,000	70,000,000
James D. Watson	Roche/454	Frag: 500 bp	93.2‡	250§	7.4	Aligned*	95‖	3.32 (BLAT)	234	1,000,000¶
Yoruban male (NA18507)	Illumina/ Solexa	93% MP: 200 bp	3,410‡	35	40.6	Aligned*	99.9	3.83 (MAQ)	40	250,000¶
		7% MP: 1.8 kb	271	35				4.14 (ELAND)		
Han Chinese male	Illumina/ Solexa	66% Frag: 150–250 bp	1,921‡	35	36	Aligned*	99.9	3.07 (SOAP)	35	500,000¶
		34% MP: 135 bp & 440 bp	1,029	35						
Korean male (AK1)	Illumina/ Solexa	21% Frag: 130 bp & 440 bp	393‡	36	27.8	Aligned*	99.8	3.45 (GSNAP)	30	200,000¶
		79% MP: 130 bp, 390 bp & 2.7 kb	1,156	36, 88, 106						
Korean male (SJK)	Illumina/ Solexa	MP: 100 bp, 200 bp & 300 bp	1,647‡	35, 74	29.0	Aligned*	99.9	3.44 (MAQ)	15	250,000¶,#
Yoruban male (NA18507)	Life/APG	9% Frag: 100–500 bp	211‡	50	17.9	Aligned*	98.6	3.87 (Corona-lite)	9.5	60,000¶,**
		91% MP: 600–3,500 bp	2,075‡	25, 50						
Stephen R. Quake	Helicos BioSciences	Frag: 100–500 bp	2,725‡	32§	28	Aligned*	90	2.81 (IndexDP)	4	48,000¶
AML female	Illumina/ Solexa	Frag: 150–200 bp‡‡	2,730‡,‡‡	32	32.7	Aligned*	91	3.81‡‡ (MAQ)	98	1,600,000‖‖
		Frag: 150–200 bp§§	1,081‡,§§	35	13.9		83	2.92§§ (MAQ)	34	
AML male	Illumina/ Solexa	MP: 200–250 bp‡‡	1,620‡,‡‡	35	23.3	Aligned*	98.5	3.46‡‡ (MAQ)	16.5	500,000‖‖
		MP: 200–250 bp§§	1,351‡,§§	50	21.3		97.4	3.45§§ (MAQ)	13.1	
James R. Lupski CMT male	Life/APG	16% Frag: 100–500 bp	238‡	35	29.6	Aligned*	99.8	3.42 (Corona-lite)	3	75,000¶,¶¶
		84% MP: 600–3,500 bp	1,211‡	25, 50						

*A minimum of one read aligning to the National Center for Biotechnology Information build 36 reference genome. ‡Mappable reads for aligned assemblies. §Average read-length. ‖D. Wheeler, personal communication. ¶Reagent cost only. #S.-M. Ahn, personal communication. **K. McKernan, personal communication. ‡‡Tumour sample. §§Normal sample. ‖‖Tumour & normal samples: reagent, instrument, labour, bioinformatics and data storage cost. E. Mardis, personal communication. ¶¶R. Gibbs, personal communication. AML, acute myeloid leukaemia; BAC, bacterial artificial chromosome; CMT, Charcot–Marie–Tooth disease; Frag, fragment; MP, mate-pair; N/A, not available; SNV, single-nucleotide variant.

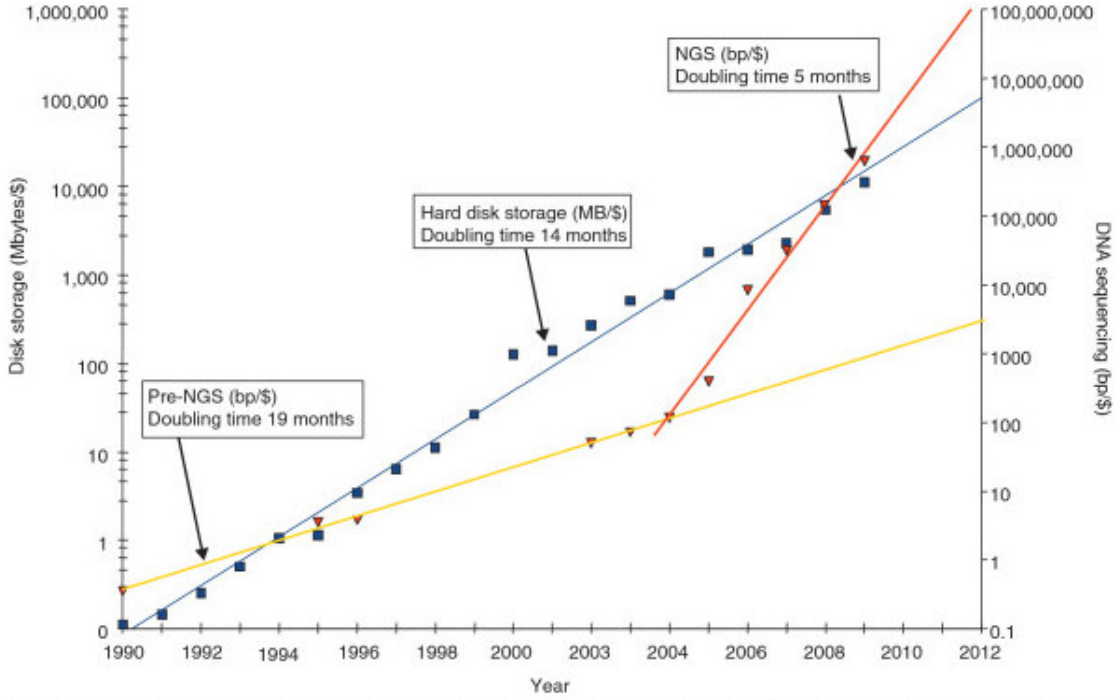
Metzker, M. L. (2010). "Sequencing technologies - the next generation." Nat Rev Genet **11**(1): 31-46.

The new Moore's law



Source:
National Human Genome Research Institute (NHGRI)

http://www.genome.gov/sequencingcosts/



Historical trends in storage prices versus DNA sequencing costs. The blue squares describe the historic cost of disk prices in megabytes per US dollar. The long-term trend (blue line, which is a straight line here because the plot is logarithmic) shows exponential growth in storage per dollar with a doubling time of roughly 1.5 years. The cost of DNA sequencing, expressed in base pairs per dollar, is shown by the red triangles. It follows an exponential curve (yellow line) with a doubling time slightly slower than disk storage until 2004, when next generation sequencing (NGS) causes an inflection in the curve to a doubling time of less than 6 months (red line). These curves are not corrected for inflation or for the 'fully loaded' cost of sequencing and disk storage, which would include personnel costs, depreciation and overhead.

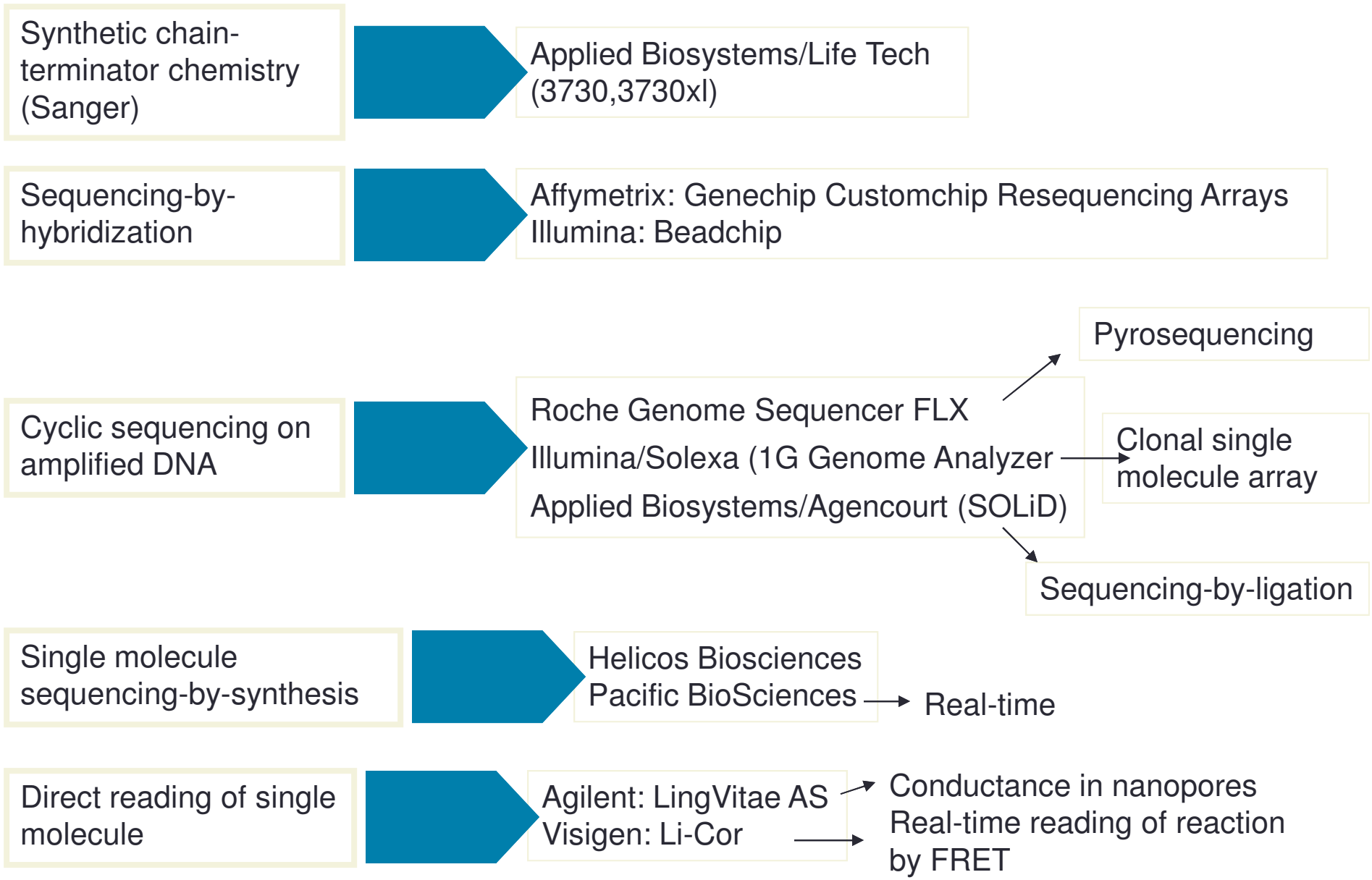
Stein *Genome Biology* 2010 **11**:207 doi:10.1186/gb-2010-11-5-207

Generations of sequencing technologies

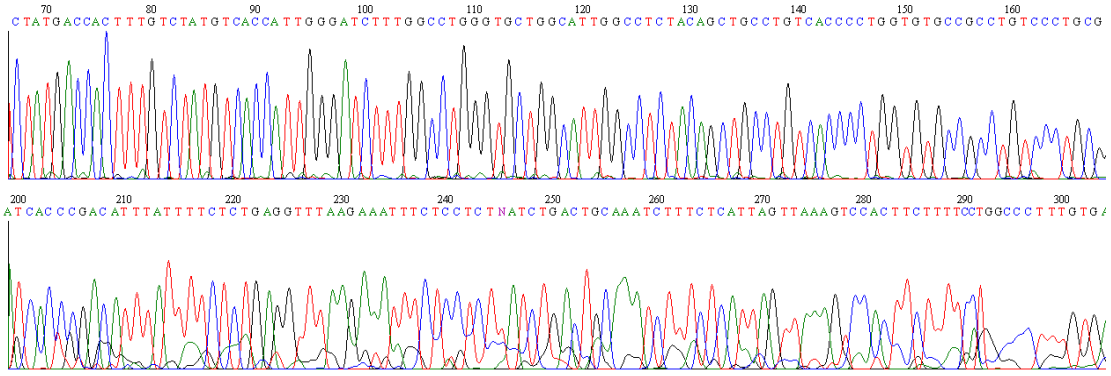
- **First** Generation sequencing technology: automated capillary sequencing machines
- **Second** Generation sequencing:
  - short reads: from something we sequence once, to something we sequence again and again.
  - it indicates a platform that requires amplification of the template molecules prior to sequencing. (454, Illumina, SOLID)
  - Second gen sequencing is not without its flaws:
    - cheap ($10-20k to sequence a human genome these days), but it still requires a lot of reagents, a lot of work and a lot of cost…
    - low read lengths are still a problem…
- Ion Torrent (**post-light sequencing**), and the Third Generation sequencers: Pacific Biosciences, Oxford Nanopore and Life Sciences Qdot technology ➔ to sequence **single molecules of DNA in real-time**. „**3rd generation**" indicates platforms that sequence directly individual DNA molecules

NGS

Sequencing approaches



Phred-score



- to characterize the quality of DNA sequences
- to compare the efficacy of different sequencing methods

$$q = -10\log_{10}(p)$$

p=error probability for the base

Phred quality scores are logarithmically linked to error probabilities

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90 %
20	1 in 100	99 %
30	1 in 1000	99.9 %
40	1 in 10000	99.99 %
50	1 in 100000	99.999 %

Phred creates a lookup table relating values of informative data metrics to empirical error rates

parameters (derived from the chromatogram):

- Peak spacing (largest/smallest, evenly spaced if R=1)
- Uncalled/called peak height ratio

(For each window of 3 or 7 bases the ratio of the height of the largest uncalled peak against the height of the shortest called peak)

- Distance to nearest unresolved position

Lookup table ➔ to assign qualities to new sequences

What is probability that a base having a phred quality score of 32 wasincorrectly called? q=32 p=10^(32/-10)=0.00063

$$p = 10^{\frac{q}{-10}}$$

Ewing B et al (1998): Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res. 8(3):175-185.
http://insilicase.co.uk/Web/PhredScores.aspx
http://noble.gs.washington.edu/~wnoble/genome373/lectures/kidd_phred_quality.ppt

Overview of leading commercial NGS platforms in 2011

Platform	Output / run (Megabase)	Read-length	Comment
ABI 3730xl	0.06	800 – 1000	Sanger
Roche 454 (GS Jr.T_FLX-T_FLX+)	50-500-900	400-400-700	Pyrosequencing
ABI SOLID (SOLID4_5500xl)	70.000-150.000	50+35 - 75+35	Colorspace
Illumina Solexa (Myseq_GAIIx_HiSeq2000)	1000-100.000-200.000	150+150-150+150- 100+100	Most used NGS
Helicos	28.000	35	Single molecule
PacBio RS	5-10	860-1100	SMRT

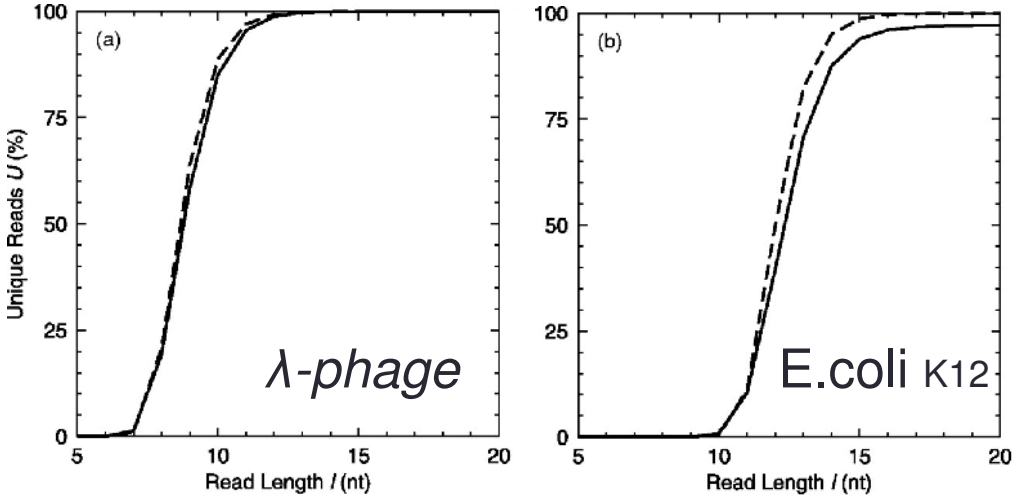
Platforms are quite diverse in sequencing biochemistry as well as in how the array is generated, their **work flows are conceptually similar**.

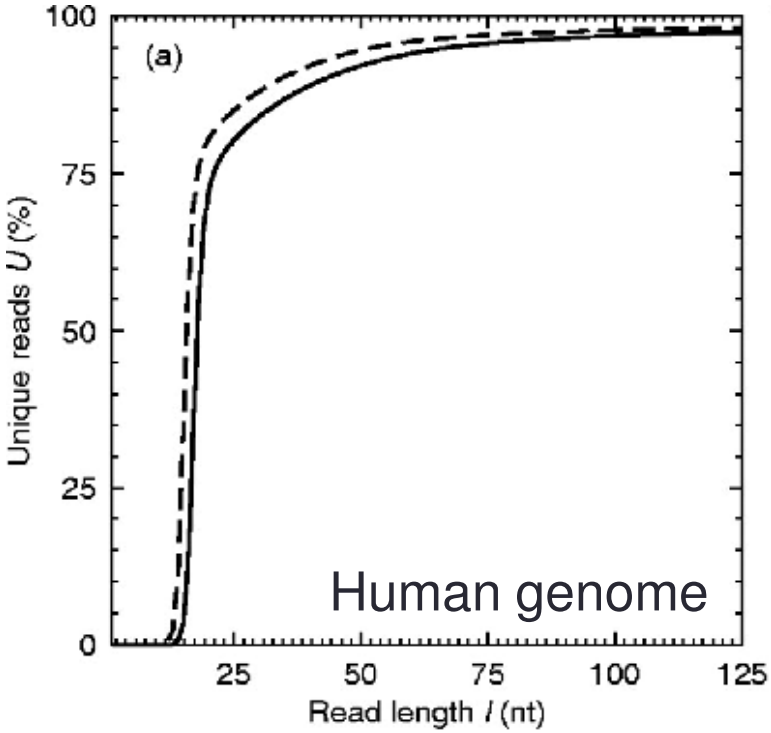
Mappable reads - read length

Mappable reads – very short DNA sequences that can be determined to originate from a single location in the genome (20–40 bases, length depends on genome complexity).

Percentage of unique reads as a function of read length.

The dashed curves show results for randomly generated sequences of the same size





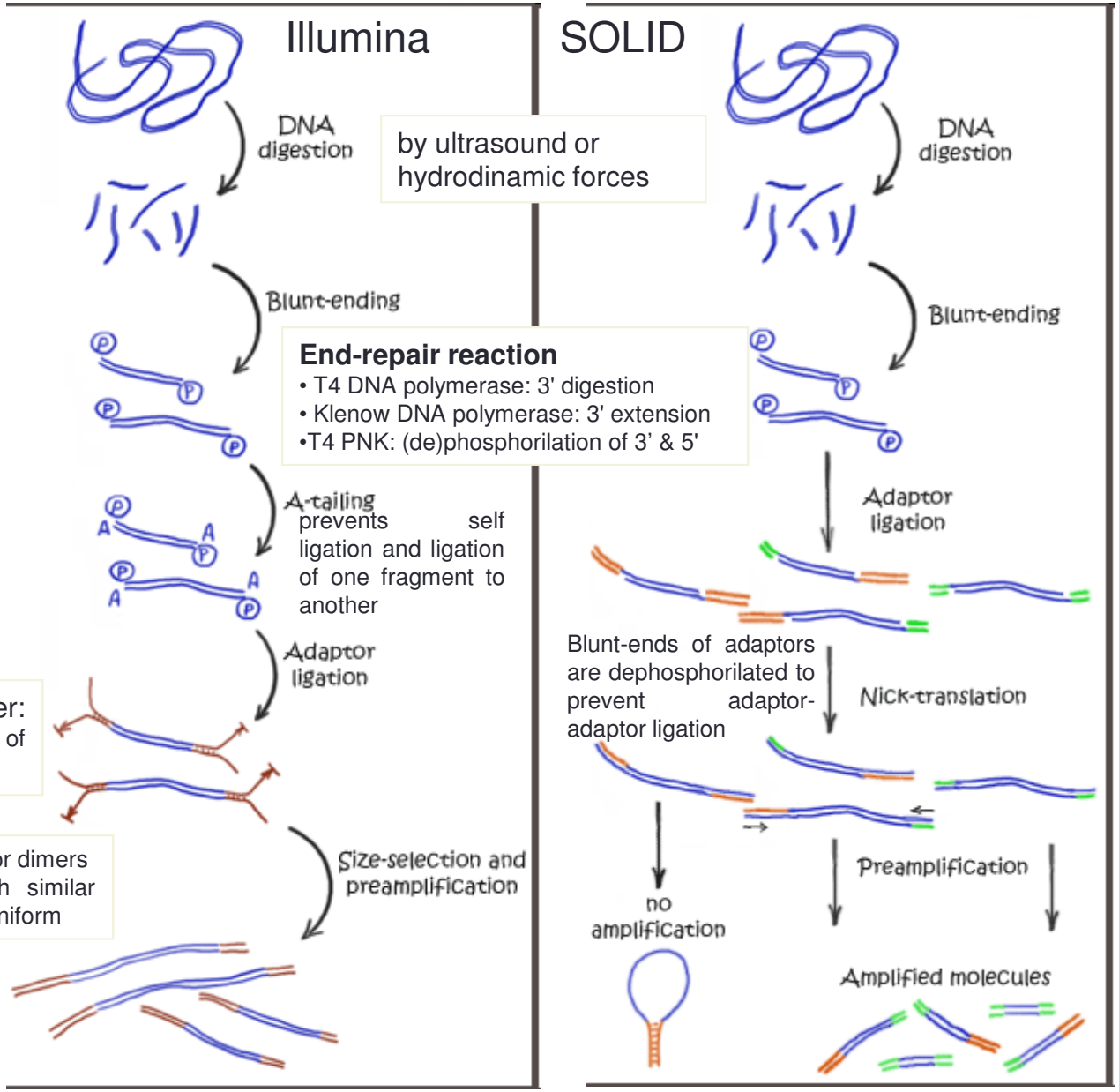
Percentage of unique sub-sequences for varying read length

the solid line shows uniqueness in the whole human genome, the dashed line shows uniqueness in human chromosome 1.

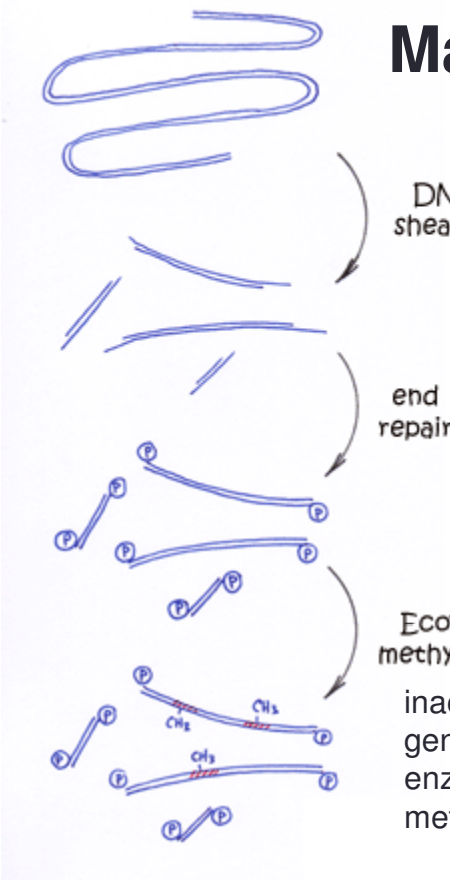
PL Roach et al.: An analysis of the feasibility of short read sequencing *Nucl. Acids Res. 33(19): e171*

Library preparation

- Template preparation: representative, non-biased source of nucleic acid material!
- random fragmentation of DNA, followed by in vitro ligation of common adaptor sequences (fragment vs mate-pair)
- clonally amplified vs. single molecule sequencing



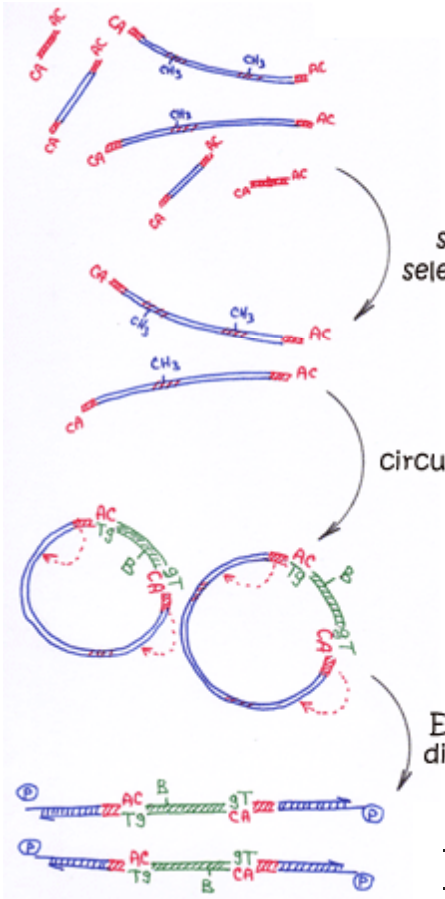
Mate-Paired library



inactivates EcoP15I sites in genomic DNA (restriction enzyme will not recognize methylated sites).

Eco CAP
adaptor ligation

ds adaptor contains EcoP15I recognition site



removes adaptor dimers

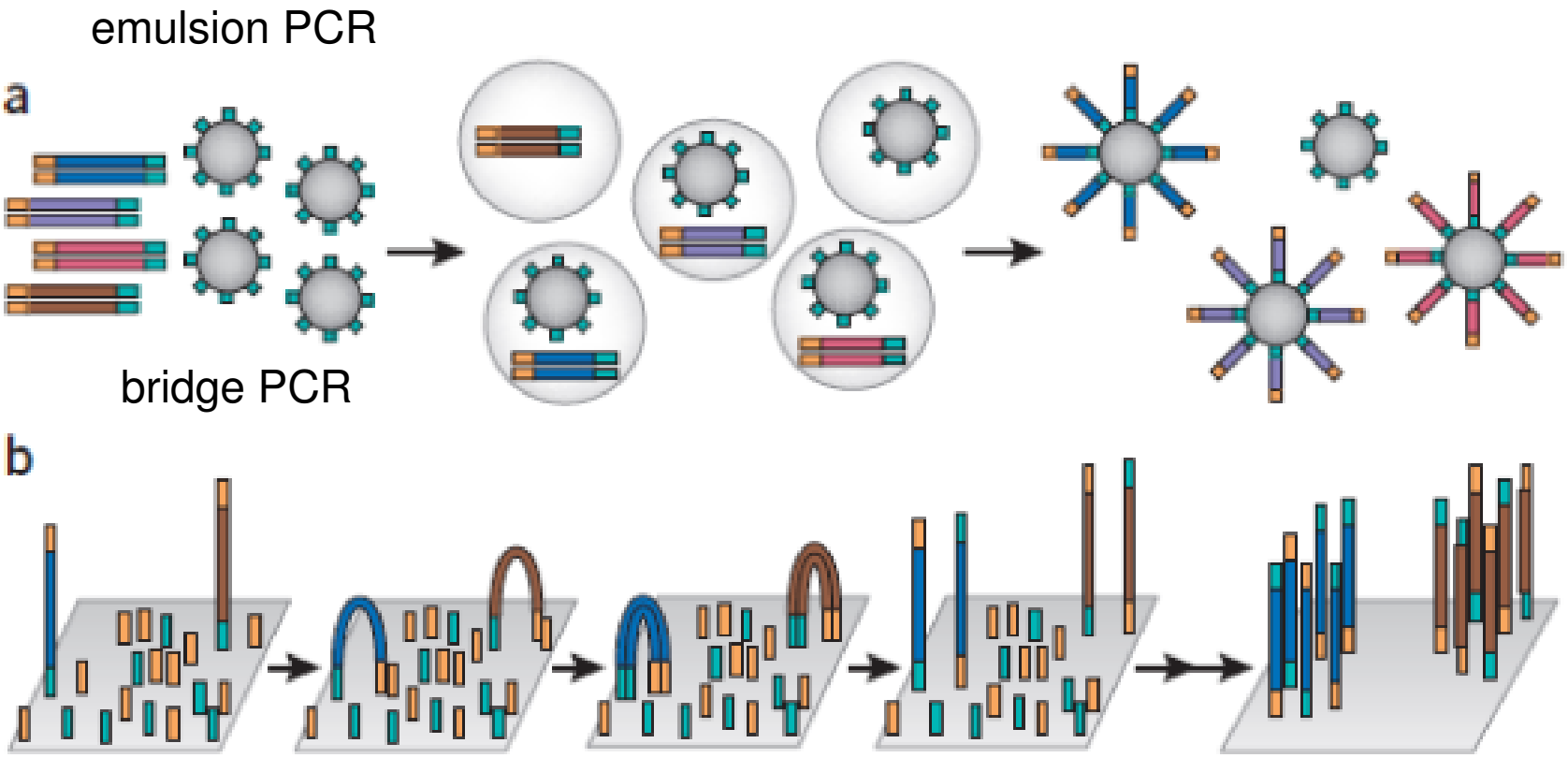
in presence of Internal adapter

$5'\cdots CAGCAG(N)_{25-27}\cdots 3'$

$3'\cdots GTCGTC(N)_{27-29}\cdots 5'$

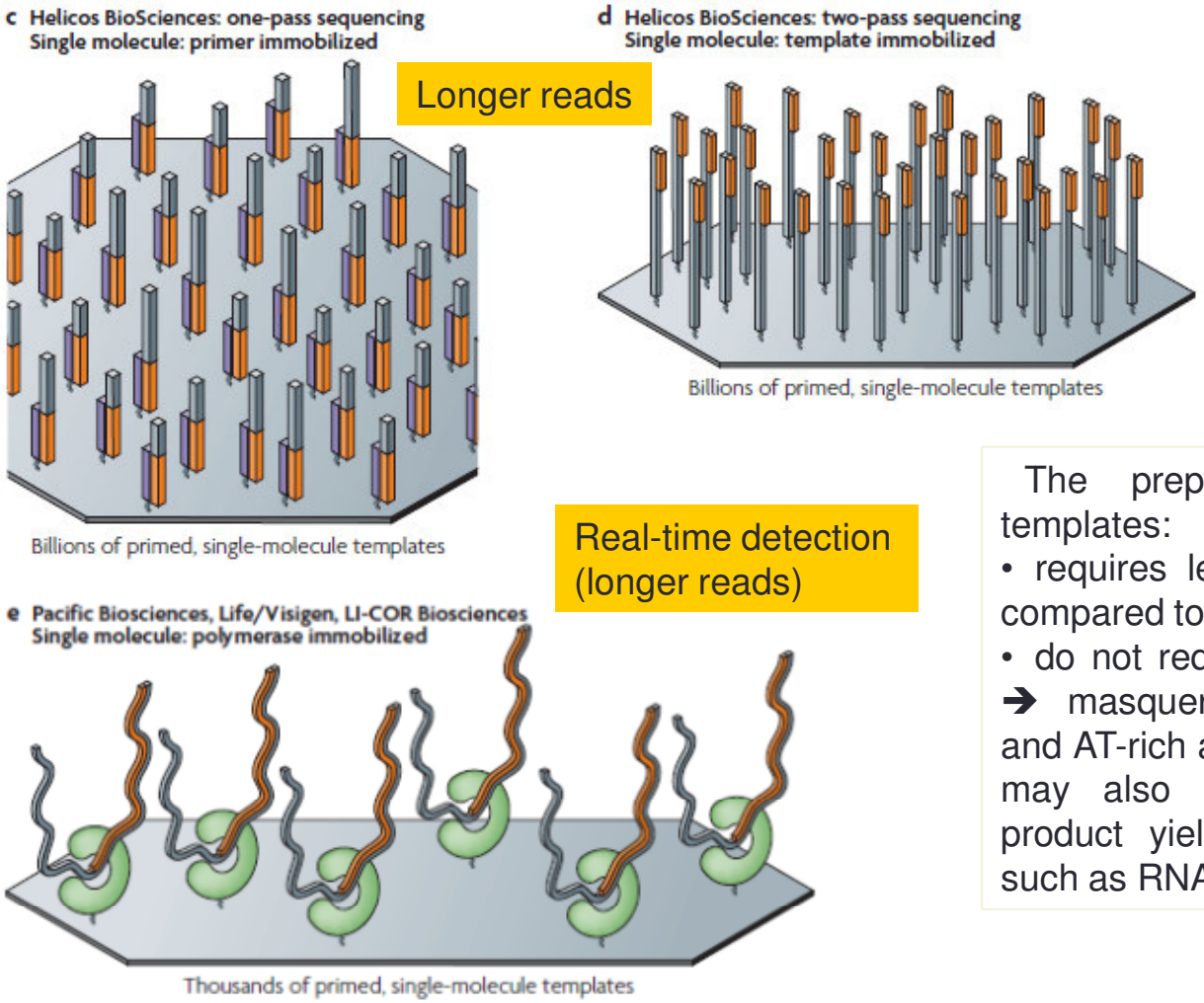
Clonally clustered amplicons

Common theme of different techniques: the template is attached or immobilized to a solid surface or support.



The immobilization of spatially separated template sites allows thousands to billions of sequencing reactions to be performed simultaneously.

Single-molecule templates:
template immobilization strategies



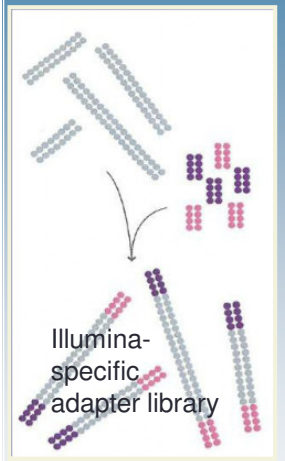
Longer reads

Real-time detection (longer reads)

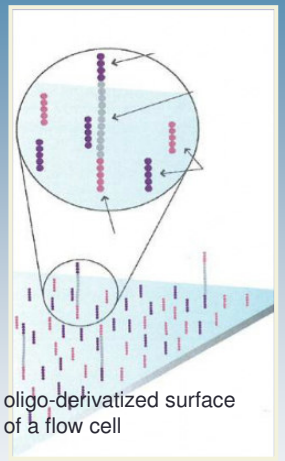
The preparation of single-molecule templates:

- requires less starting material (<1 µg) compared to clonally amplified methods
- do not require PCR (creates mutations ➔ masquerade as sequence variants, and AT-rich and GC-rich target sequences may also show amplification bias in product yield (Quantitative applications, such as RNA–seq!)

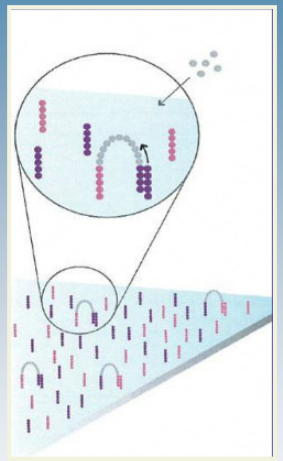
Illumina / Solexa



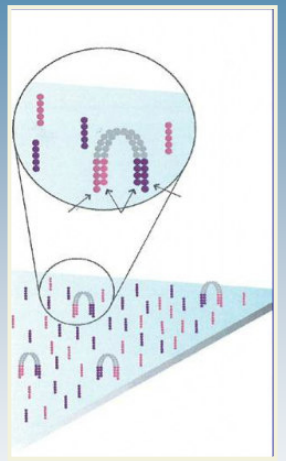
Randomly fragment genomic DNA and ligate adapters to both ends of the fragments.



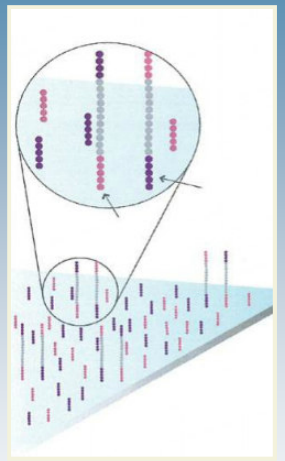
Attach DNA to Surface: Bind single-stranded fragments **randomly** to the inside surface of the flow cell channels.



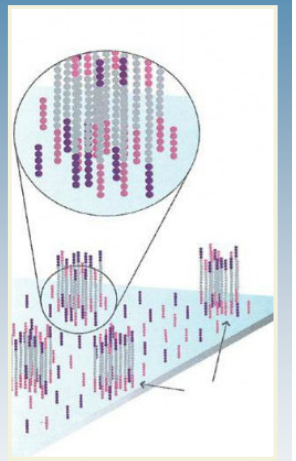
Bridge Amplification: Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification.



Synthesis completed

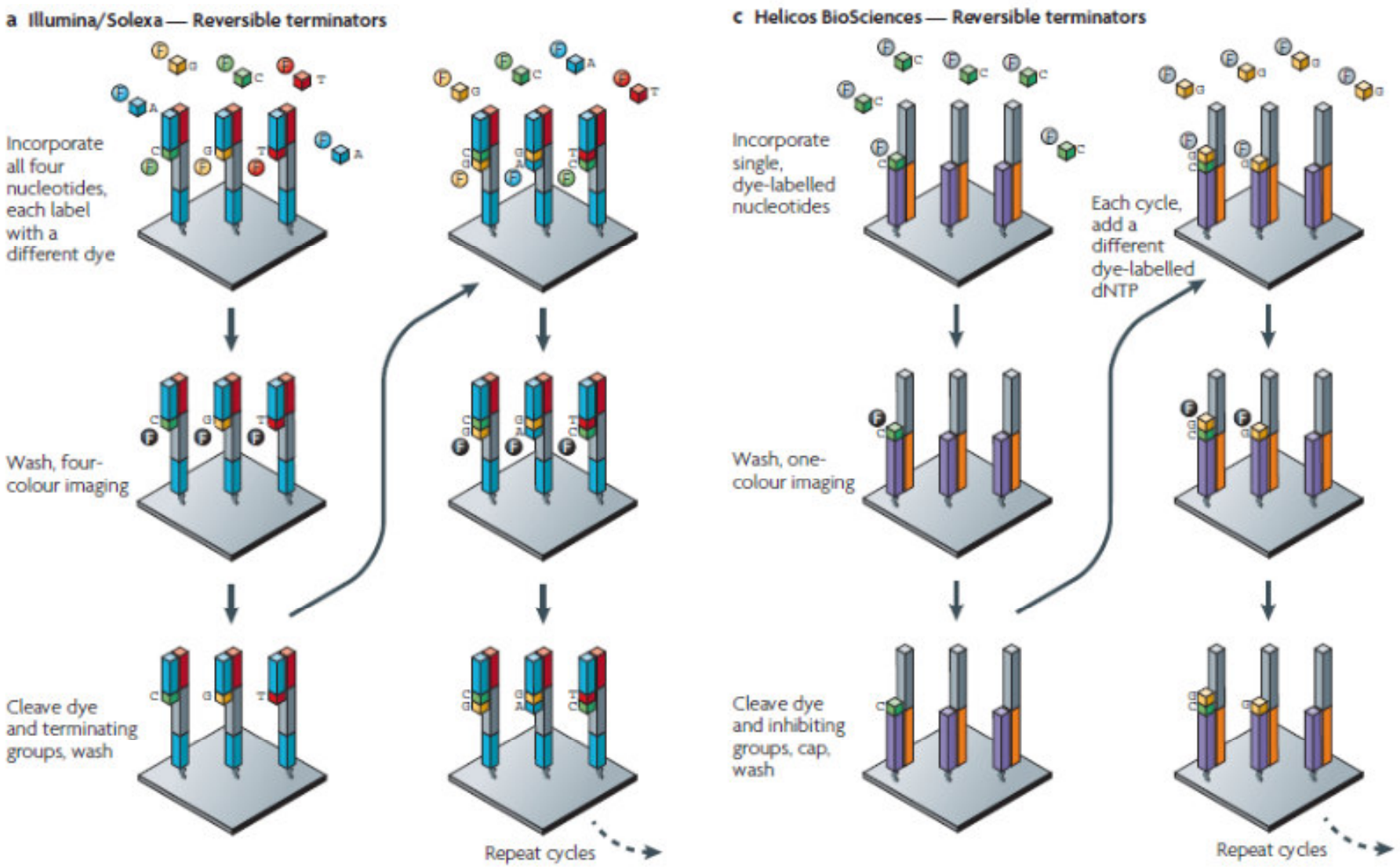


Denaturation

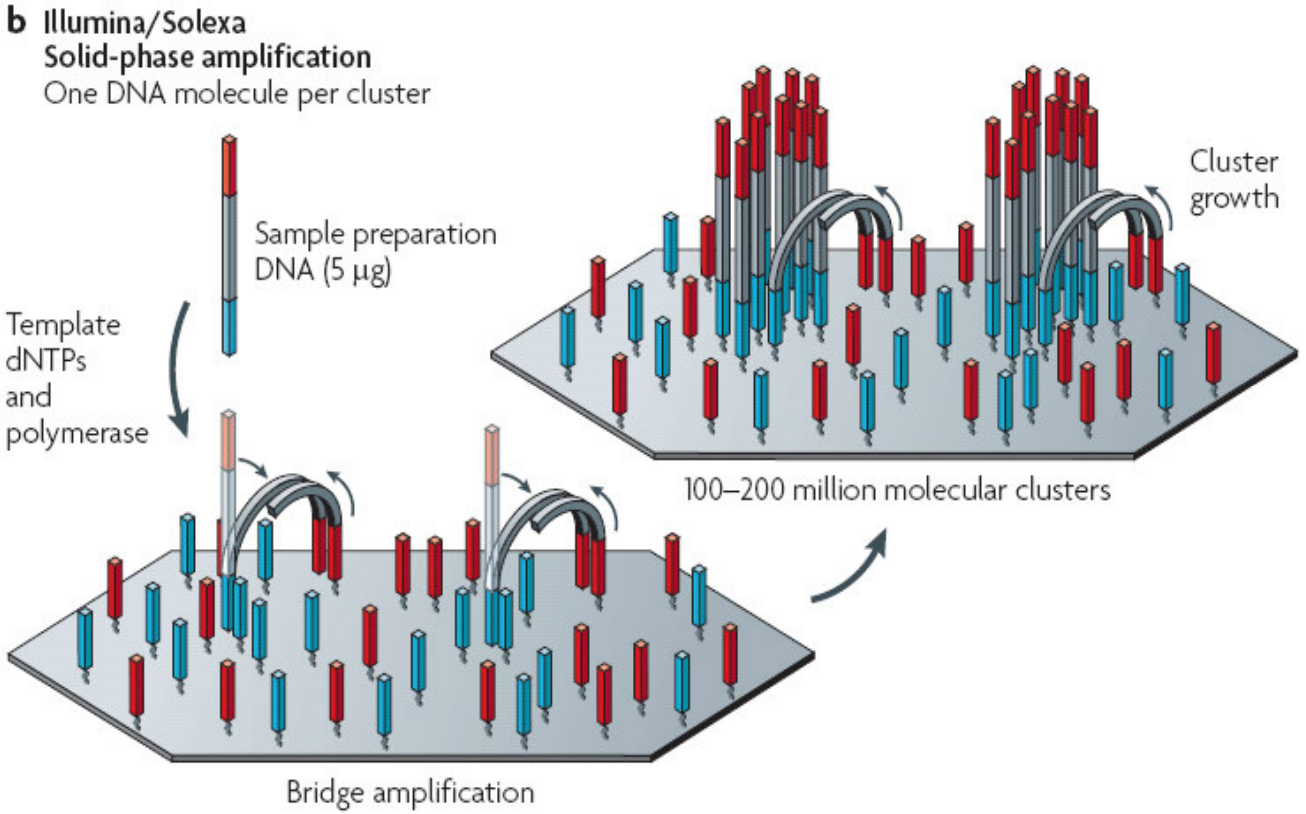


Complete Amplification clusters, that each represent the single molecule that initiated the cluster amplification.

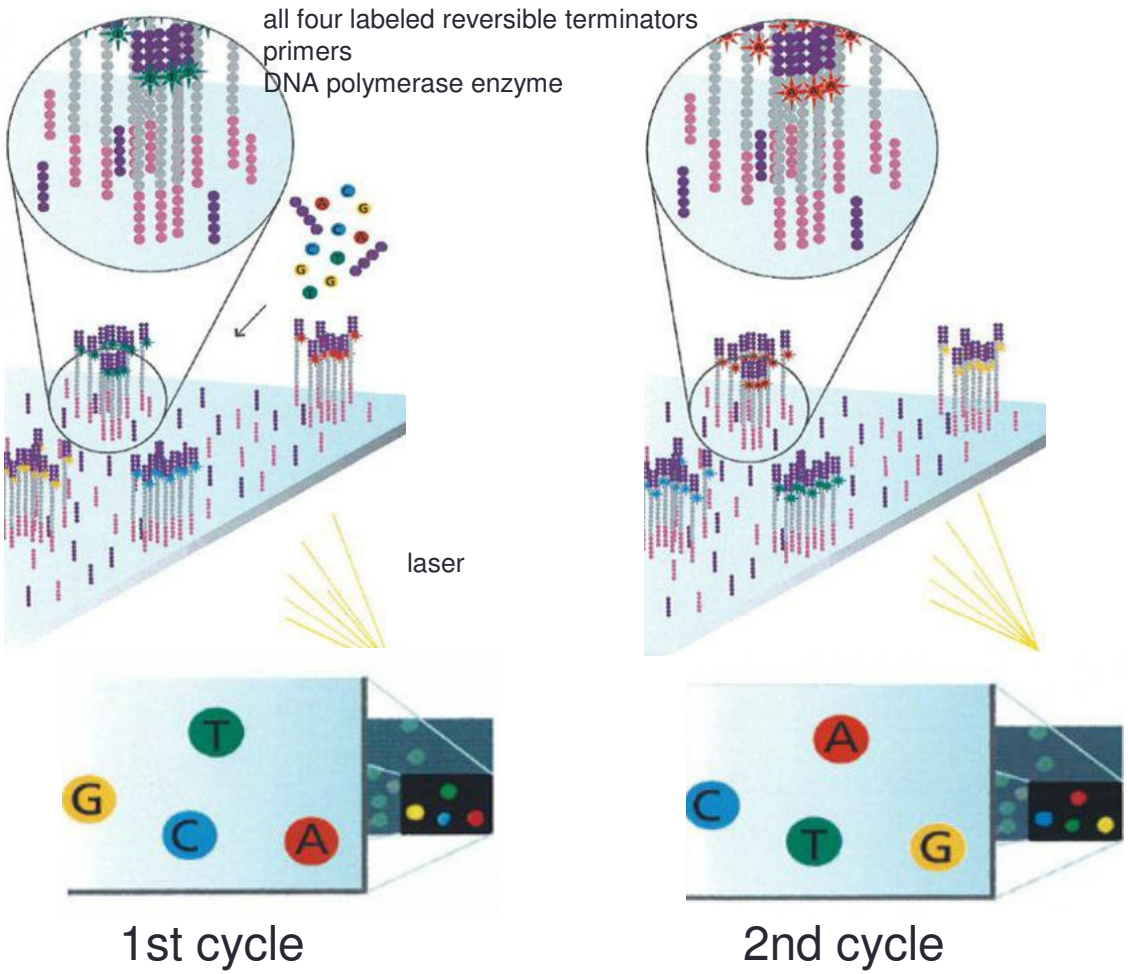
- Cluster Station: automated device
- Bridge amplification: fragments on its surface, DNA polymerase, multiple DNA copies, clusters
- cBot – a required accessory instrument for many Illumina sequencers in which Bridge PCR is completed
- **Flow cell** : 8-channel sealed glass microfabricated device. A separate library can be added to each of the eight channels, or the same library can be used in all eight, or combinations thereof. **> 10 million clusters**
- Clusters: each represent the single molecule that initiated the cluster amplification. **Each cluster** contains approximately **one million copies of the original fragment**, which is sufficient for reporting incorporated bases at the required signal intensity for detection during sequencing.



Metzker, M. L. (2010). "Sequencing technologies - the next generation." Nat Rev Genet **11**(1): 31-46.

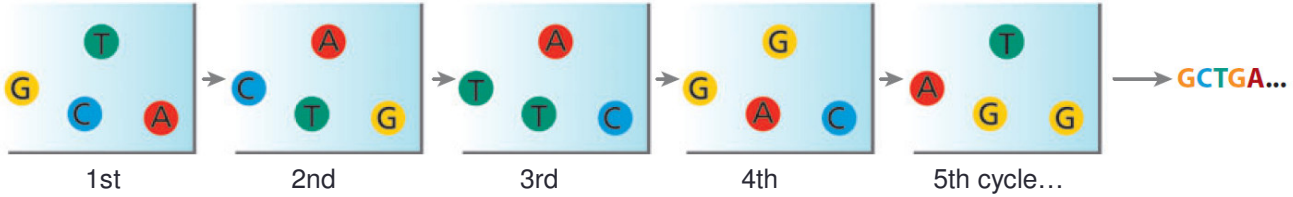


Metzker, M. L. (2010). "Sequencing technologies - the next generation." Nat Rev Genet **11**(1): 31-46.



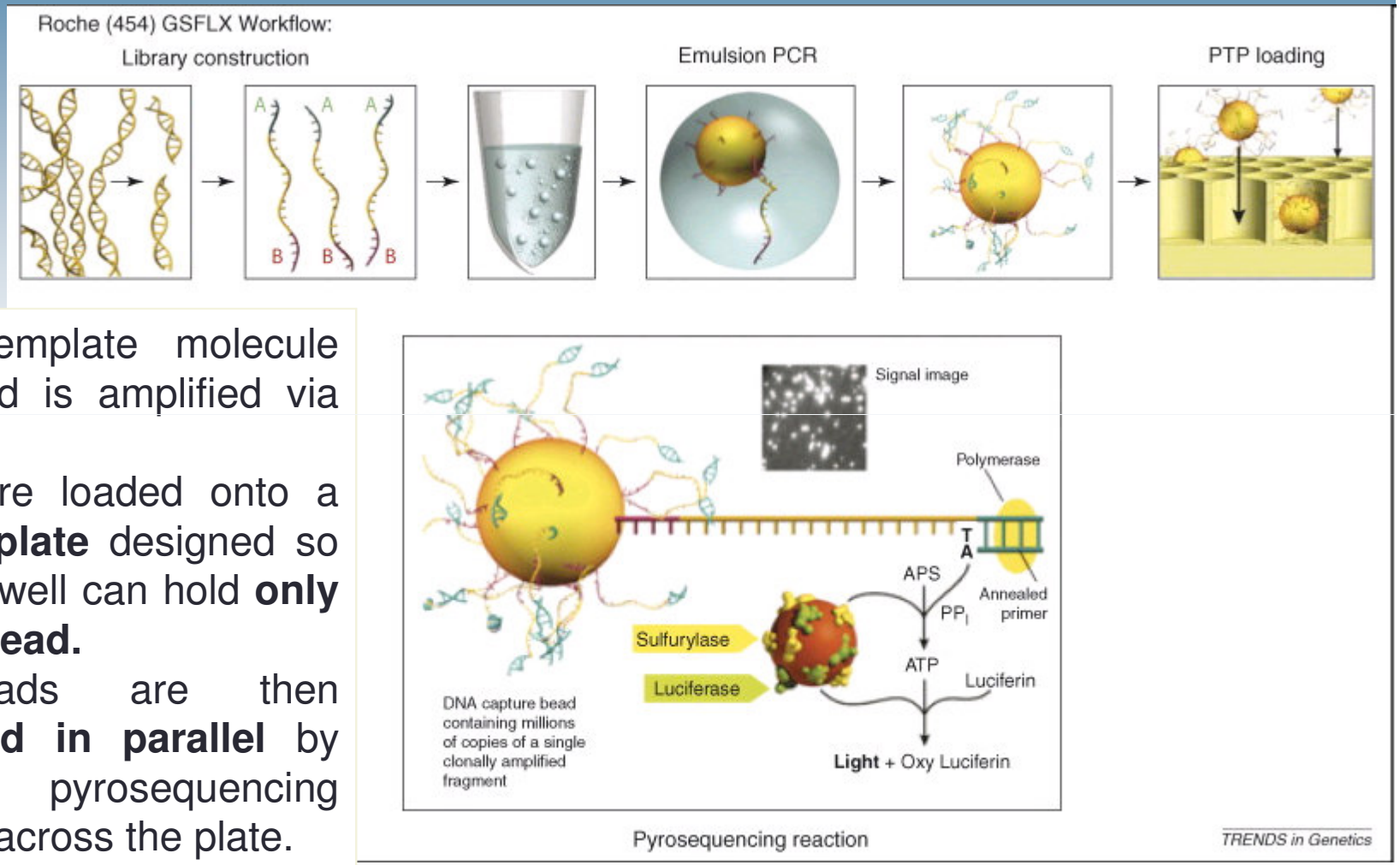
Clonal Single Molecule Array technology

- the strategy that is similar to Sanger sequencing
- Sequencing-by-synthesis: reversible terminator-based
- **All four nucleotides are added simultaneously** to the flow cell channels (base-unique fluorescent label and the 3-OH group is chemically blocked)
- An **imaging** step follows each base incorporation step.
- After each imaging step, the 3 **blocking group** is chemically **removed**
- This series of steps ➔ read lengths of 25–35 bases.



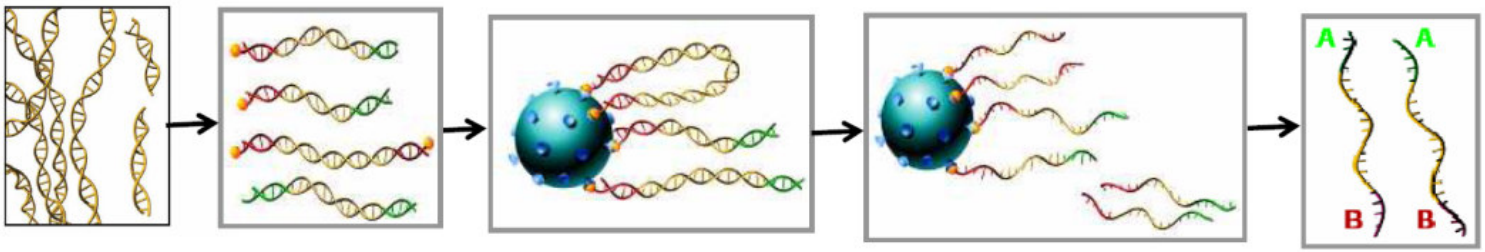
454/Roche

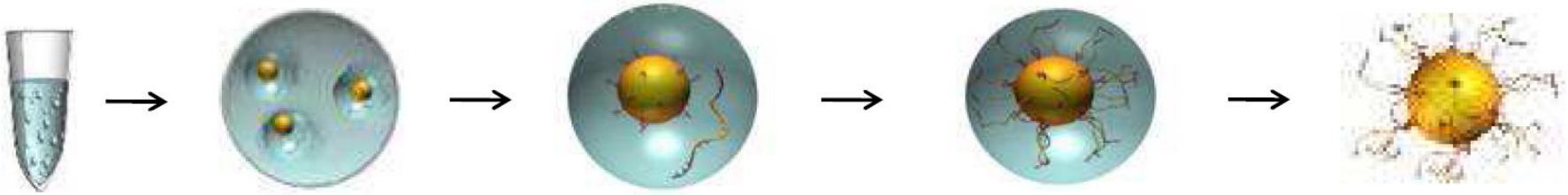
the 1st commercial NGS platform



- single template molecule on a bead is amplified via **emPCR**
- beads are loaded onto a **picotitre plate** designed so that each well can hold **only a single bead.**
- all beads are then **sequenced in parallel** by **flowing** pyrosequencing **reagents** across the plate.

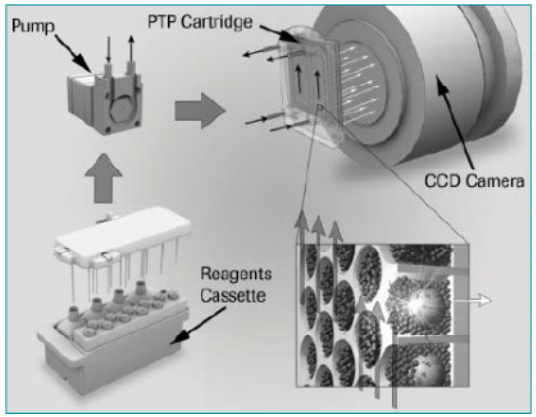
emPCR

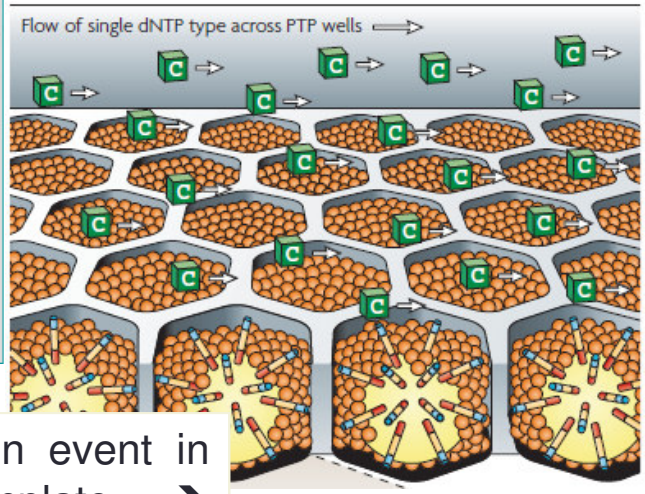


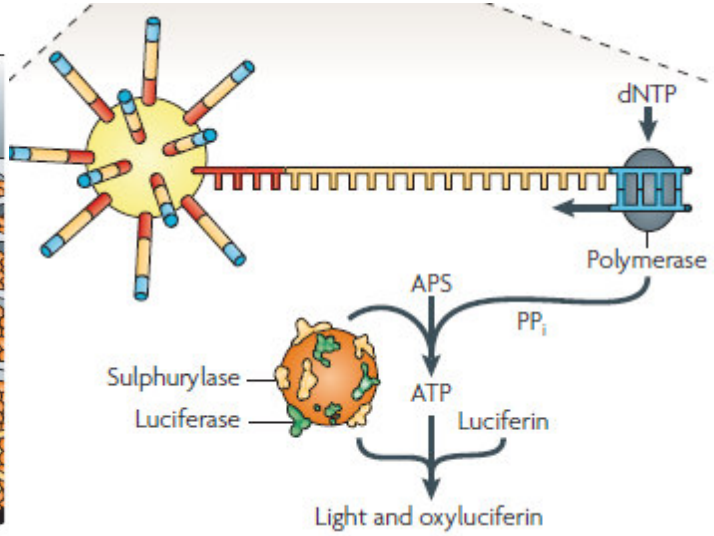


• PCR that occurs within **aqueous microdroplets** separated by oil (up to 1000 of independent reactions /ul)
• one **primer is usually covalently linked to a bead** ➔ PCR only occurs in microdroplets with beads,
• a **single template molecule per bead / microdroplet** is needed, ➔ each bead having a homogeneous set of template molecules,
• used in 454, Ion Torrent, and SOLiD sequencers

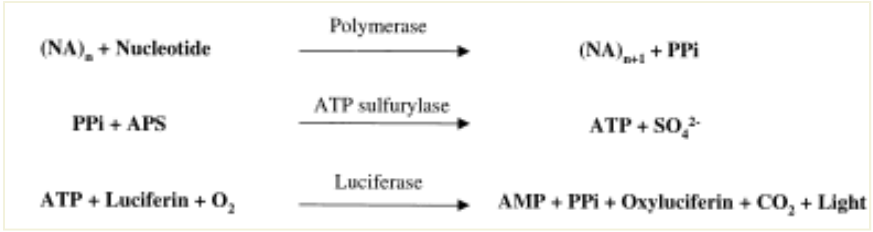
Pyrosequencing by 454

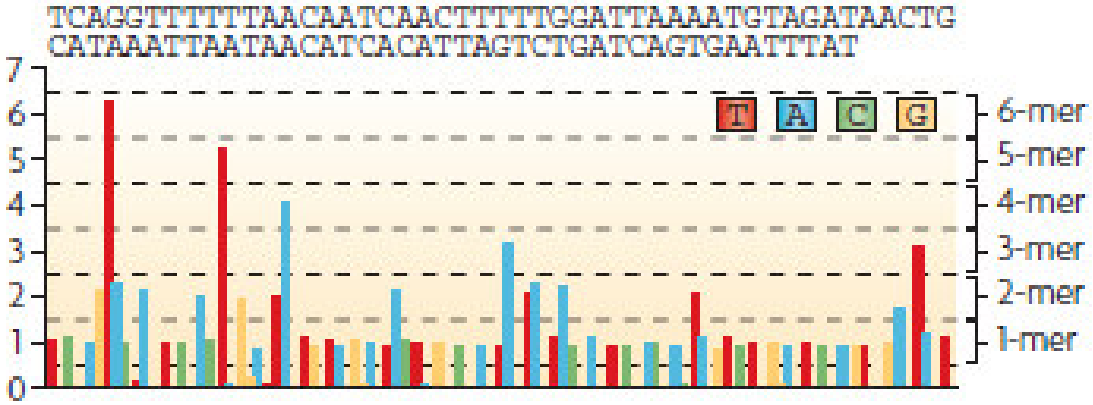




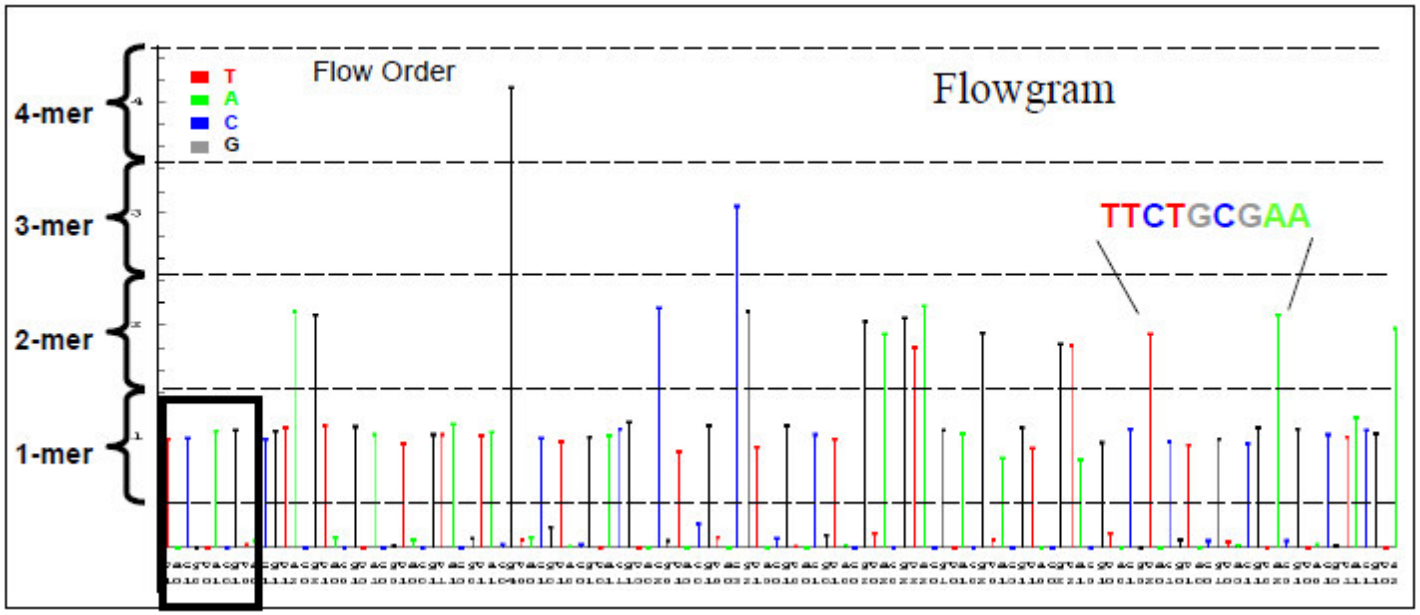


- A nucleotide-incorporation event in a well containing template ➔ **pyrophosphate release** ➔ picotiter-plate well–localized luminescence ➔ transmitted through the **fiber-optic plate** ➔ charge-coupled device camera
- strength: the **longer read length**, which facilitates de novo assembly of genomes
- estimates of **homopolymer** length (>3–4 bases) are less accurate with increasing length ➔ metal coating of the walls of picotiter wells





Pyrosequencing by 454

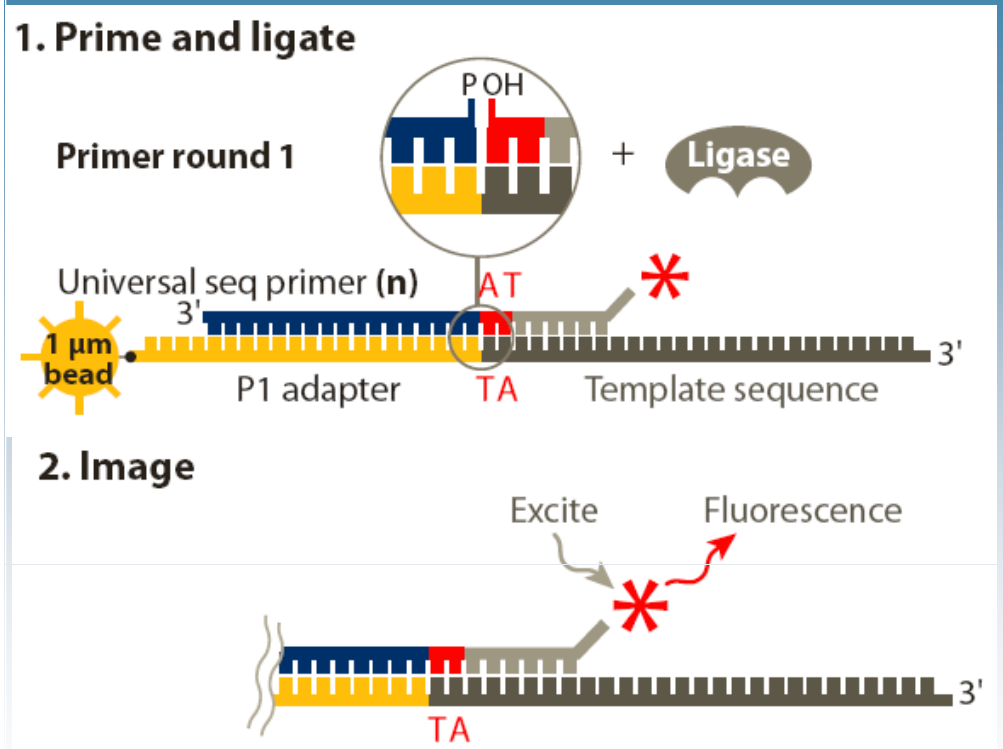


Key sequence = TCAG for signal calibration

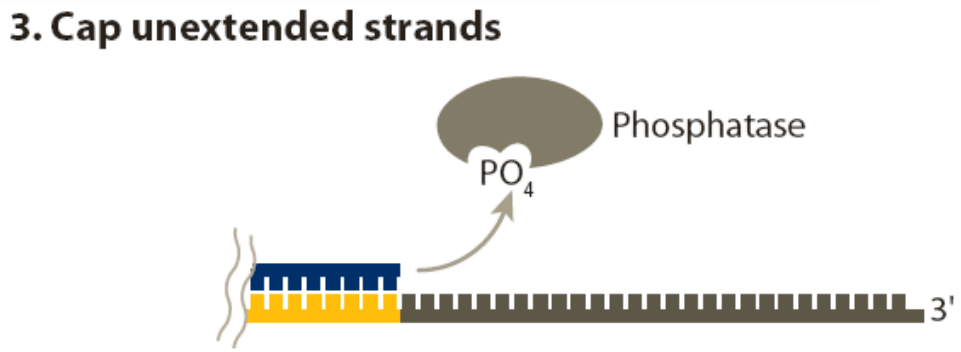
- chemi-luminescent signal ➔ bar graph of light intensities: “flowgram” for each well contained on the PicoTiterPlate
- The signal strength is proportional to the number of nucleotide incorporated
- conversion raw data into basecalls and quality scores
- FASTA and Standard Flowgram Format (SFF) files

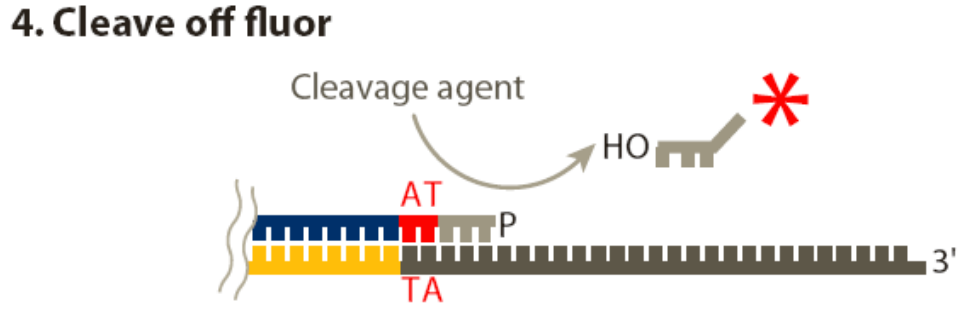
http://www.454.com/downloads/news-events/how-genome-sequencing-is-done_FINAL.pdf

Sequencing by Oligonucleotide Ligation and Detection

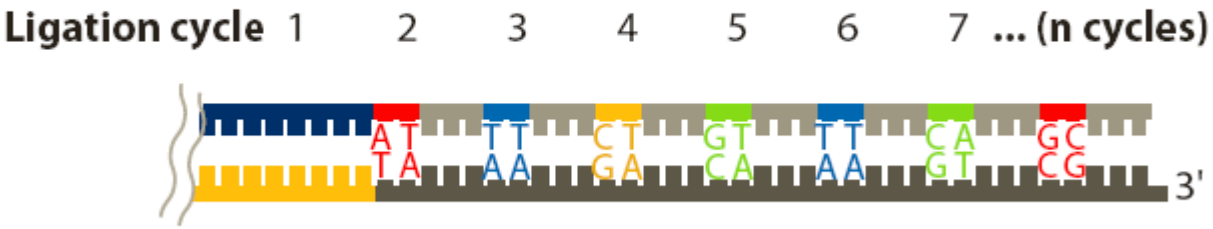


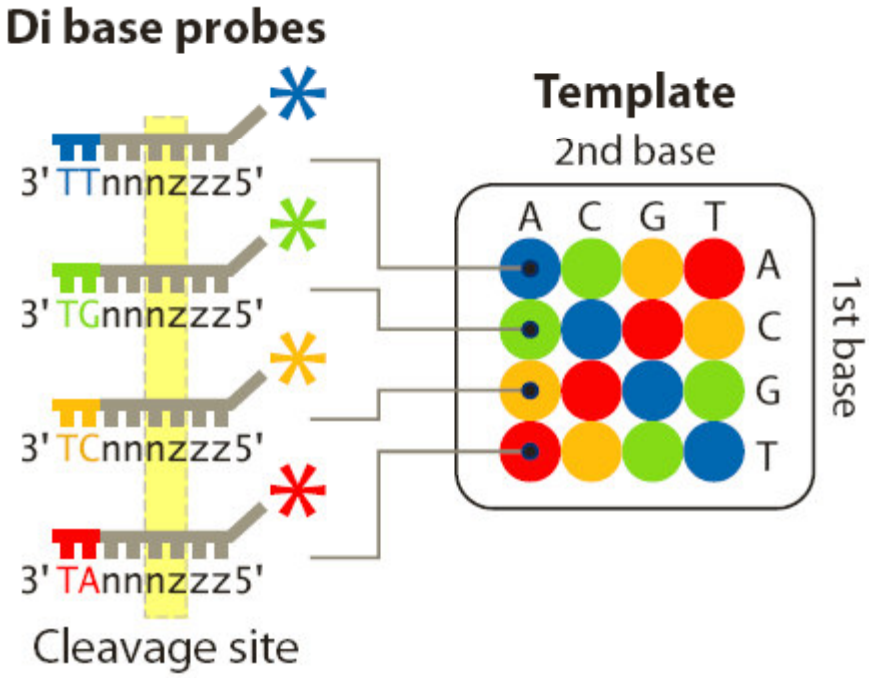
- adapter-ligated fragment library
- to amplify the fragments for sequencing: emulsion PCR approach with magnetic beads (1um)
- an unique approach to sequence: SOLiD uses DNA ligase specific fluorescent labeled 8mers, whose 4th and 5th bases are encoded by the attached fluorescent group
- after detection, a regeneration step removes bases from the ligated 8mer (including the fluorescent group) and concomitantly prepares the extended primer for another round of ligation





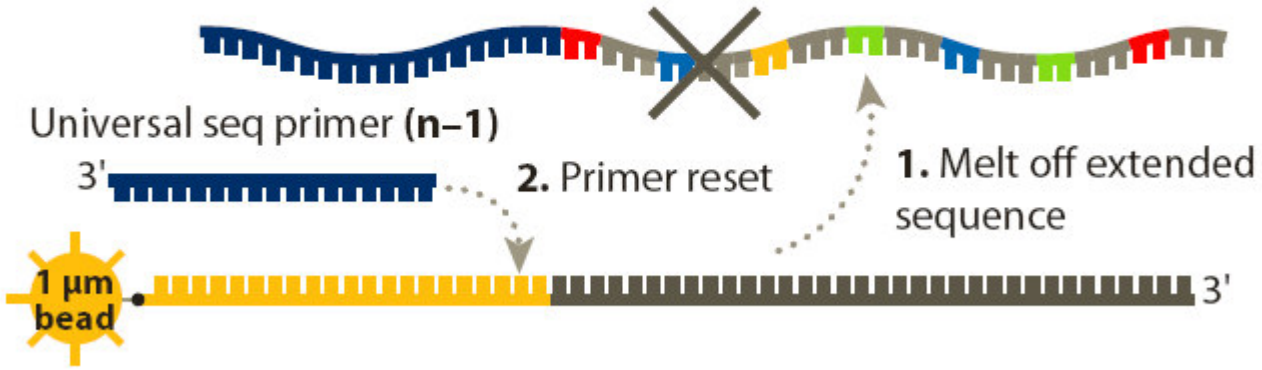
5. Repeat steps 1–4 to extend sequence



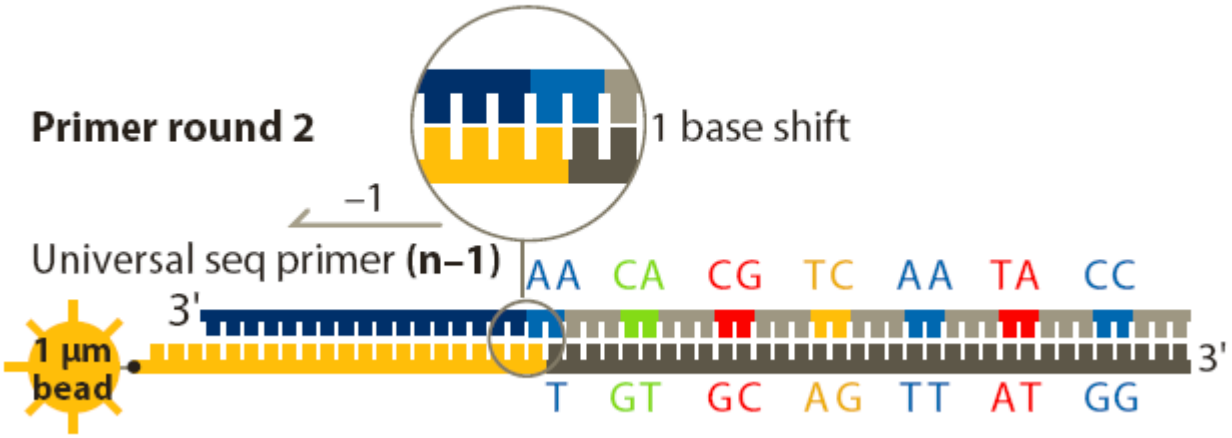


- Each fluorescent group on a ligated 8mer identifies a two-base combination
- 1024 different detection oligos:
  (dinucleotide + 3 degenerate)^4=5^4
- First two nucleotides determine the colour of the fluorophore
- Three next positions — degenerate nucleotides (64 different for each dinucleotide).
- Three last positions: universal bases, they are the same for all detector
- Dark oligonucleotides have the same internal structure, but have no fluorophores oligonucleotides.

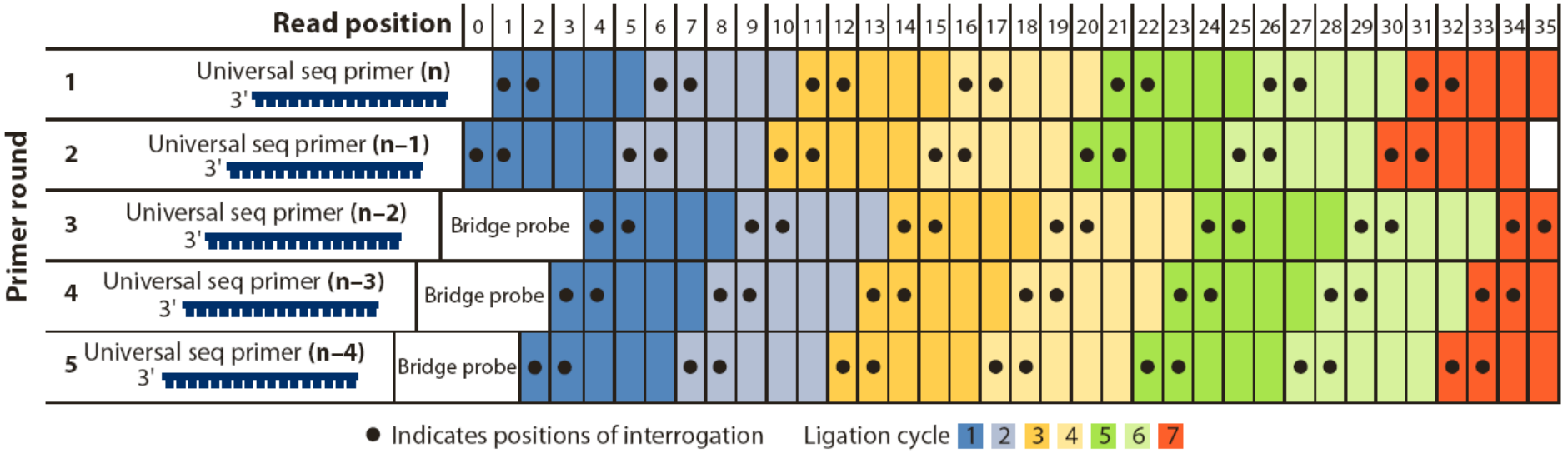
6. Primer reset

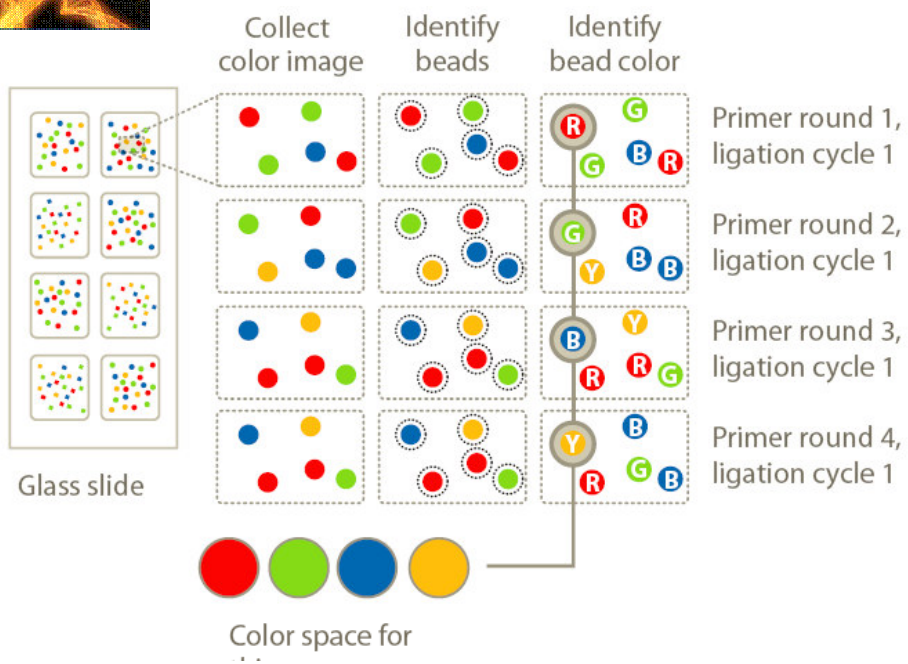


7. Repeat steps 1–5 with new primer



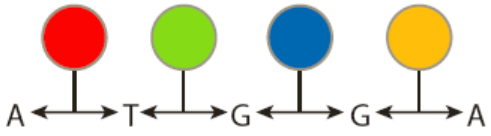
8. Repeat Reset with , n–2, n–3, n–4 primers


Read position
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35
Primer round
1 Universal seq primer (n) 3'
2 Universal seq primer (n–1) 3'
3 Universal seq primer (n–2) 3'
Bridge probe
4 Universal seq primer (n–3) 3'
Bridge probe
5 Universal seq primer (n–4) 3'
Bridge probe
● Indicates positions of interrogation
Ligation cycle 1 2 3 4 5 6 7

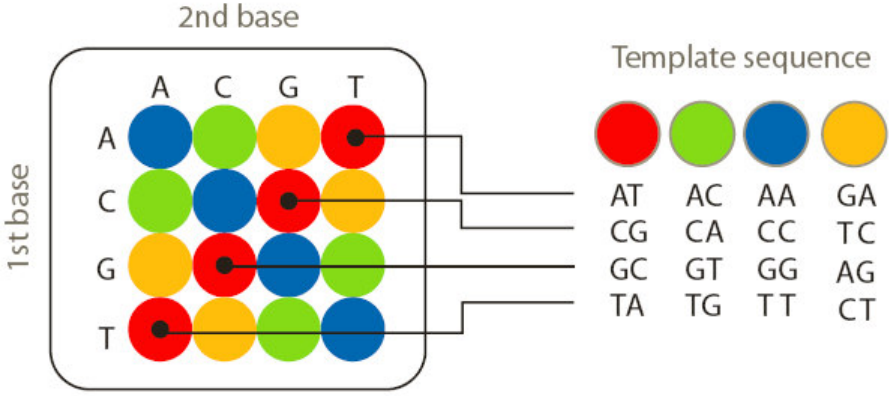


Double interrogation

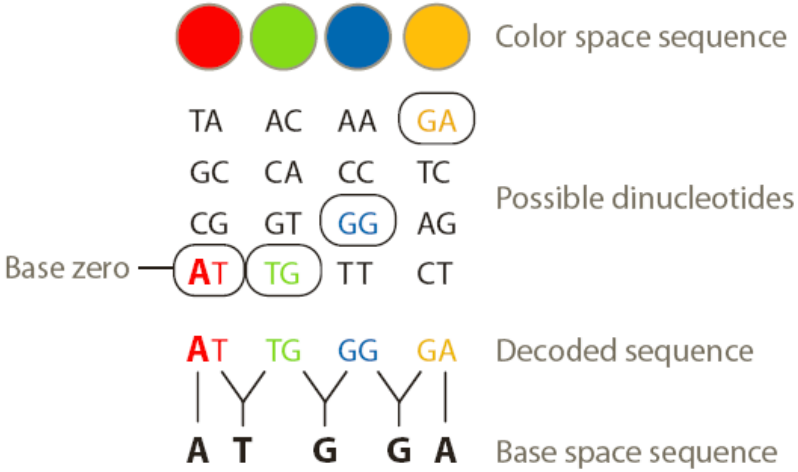
With 2 base encoding each base is defined twice

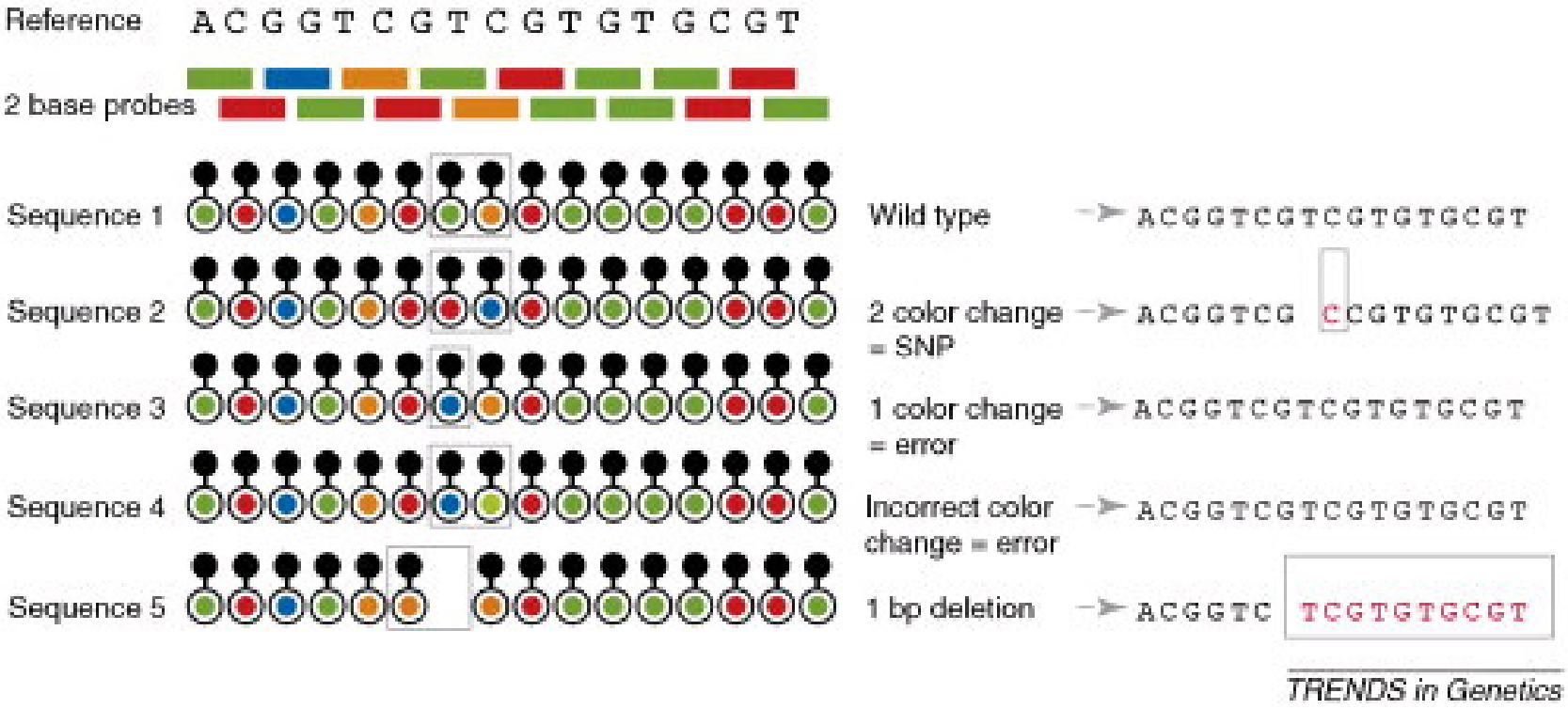


Possible dinucleotides encoded by each color



Decoding

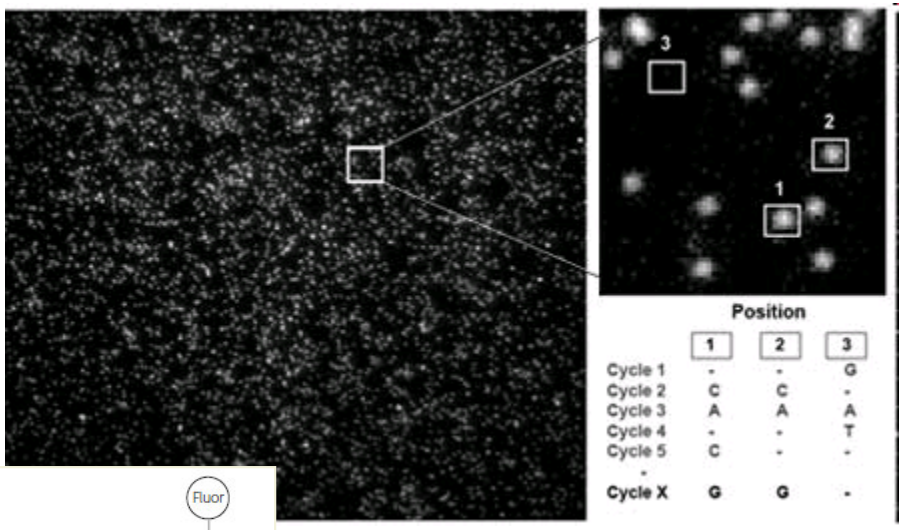



Reference
A C G G T C G T C G T G T G C G T
2 base probes
Sequence 1
Wild type
A C G G T C G T C G T G T G C G T
Sequence 2
2 color change
= SNP
A C G G T C G C C G T G T G C G T
Sequence 3
1 color change
= error
A C G G T C G T C G T G T G C G T
Sequence 4
Incorrect color
change = error
A C G G T C G T C G T G T G C G T
Sequence 5
1 bp deletion
A C G G T C T C G T G T G C G T
TRENDS in Genetics

HeliScope

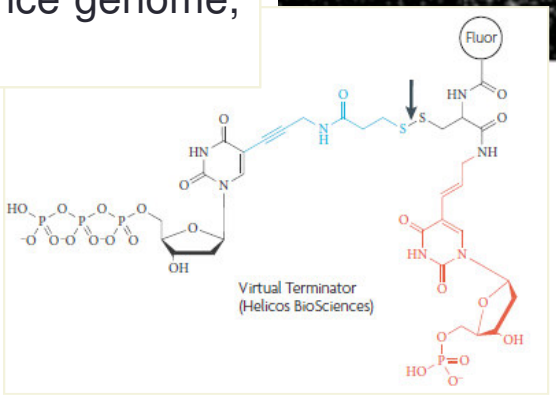
• True **single molecule sequencing** (1st, Helicos BioSciences): No amplification step is needed

• Nature Biotechnology (proof of concept): „single-molecule methods to sequence an individual human genome": 24- to 70-bp reads (32 bp average) to ~90% of the NCBI reference genome, with 28× average coverage



http://www.helicosbio.com/

• sequence output of 1 Gb/day
• 2-pass sequencing: accuracy was improved when template molecules were sequenced twice.



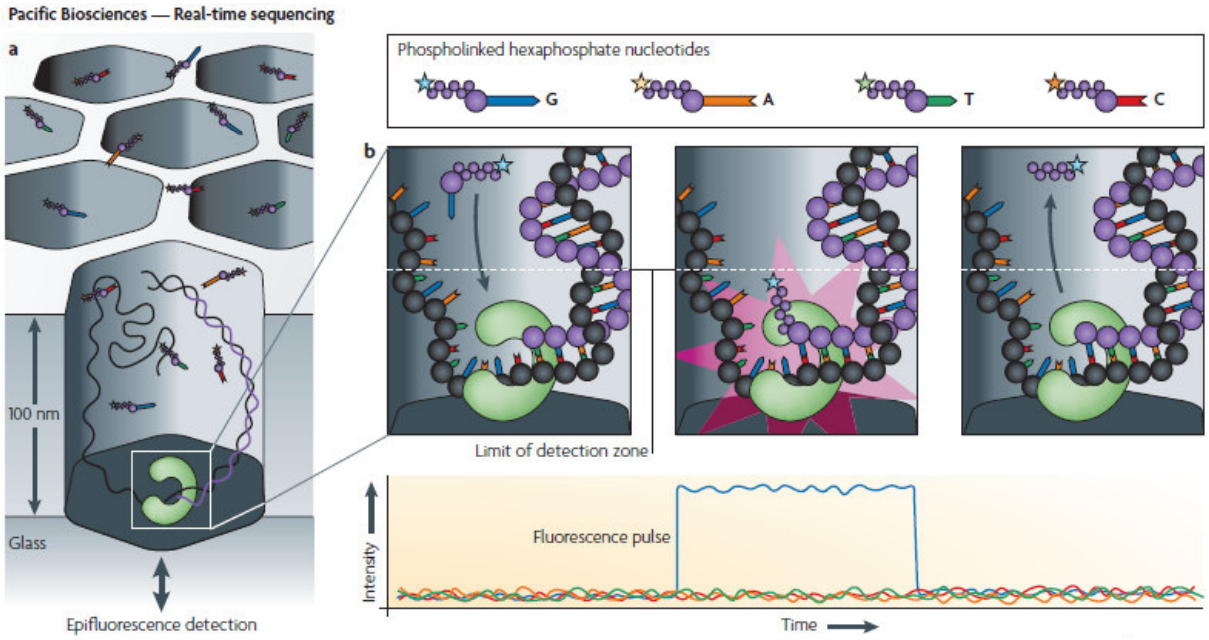
• accuracy of homopolymer sequencing ! (**3' unblocked terminators**): "virtual terminators," reduce polymerase processivity so that only single bases are added.
• **high cost of the instruments** and short read lengths limited adoption of this platform.
• Helicos no longer sells instruments, but conducts sequencing via a service centre model.

➔ fragmented sample DNA polyadenylated at the 3' end with the final adenosine fluorescently labeled.
➔ hybridized to poly(dT) oligonucleotides immobilized on a flow-cell surface at a capture density of up to 100 millions template strands /cm2. After the positional coordinates of the captured strands are recorded by a CCD-camera, the label is cleaved and washed away before sequencing.
➔ For sequencing, polymerase and one of 4 Cy5-labeled dNTPs are added to the flow cell, which is imaged to determine incorporation into individual strands

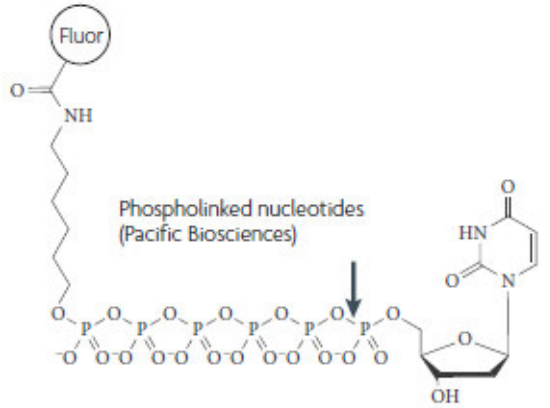
Pushkarev, D., N. F. Neff, et al. (2009). "Single-molecule sequencing of an individual human genome." Nat Biotech **27**(9): 847-850.

PacBio RS

- SMRT= Single Molecule Real-Time technology
- Each chip with waveguides: 100nm hole to watch DNA-polymerase perform sequencing by synthesis
- Phospholinked nucleotides labeled with fluorophore
- Long reads, short run times, high error rate



Individual DNA polymerases are attached to the surface



zero-mode waveguide (ZMW)

Light cannot propagate through these small **waveguides** ➔ **evanescent wave** = excitation wave with an intensity that decreases exponentially away from the surface. (<200 nm) ➔ it **reduces** the **observation volume** at the surface of the polymerase reaction down to the zeptolitre range (10^{-21} l) ➔ polymerization reaction can be performed at **higher dye-labelled nucleotide concentrations**

Ion Torrent

- When nucleotide incorporated into a strand of DNA by a polymerase, a hydrogen ion is released
- High density array of wells:
  - → made by semiconductor technology,
  - → with each well with different template
  - → beneath the well is an ion-sensitive layer and a sensor
- Sequentially floods the chip with one nucleotide after another
- If the given dNTP matches, a hydrogen ion is relelased and the change in the pH of the solution is detected

Ion Personal Genome Machine

Ion Semiconductor Sequencing Chip	Output	Read Length		Total Sequencing Time
314	> 10Mb	2011	2012	< 2 hours
316	> 100Mb	> 200bp	> 400bp	
318	> 1Gb			
Accuracy:	>99.99% consensus accuracy and >99.5% raw accuracy.			

Sequencing strategy **similar to the 454**, except that
(i) H+ are detected instead of a pyrophosphatase cascade and
(ii) sequencing chips conform to common design and standards ➔ low-cost manufacturing

- no lasers, cameras or fluorescent dyes are needed
- Ion Torrent was purchased by Life Technologies in 2010

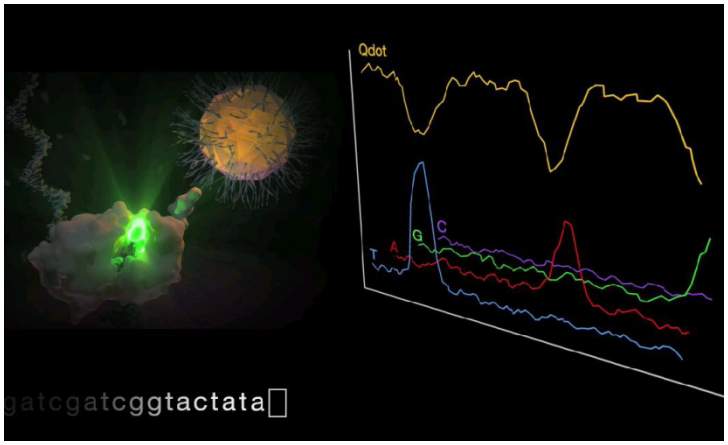
http://www.iontorrent.com/

Starlight

• **single-molecule sequencing**: **quantum dot nanocrystal tethered to a DNA polymerase** as a photon source for **four-color terminally labeled nucleotides** and measures light emitted from both the Qdot and the labels **in real time**

• DNA is attached to the surface of a microscope slide (~PacBio, but DNA **polymerase can be replaced** after it has lost activity. Thus, sequencing can continue along the entire length of a template.)
• about 50,000 single-stranded DNA strands at a time are monitored
• When the correct **nucleotide binds to the polymerase**, the **Qdot transfers some of its light** to the nucleotide dye in a process called Förster resonance energy transfer (**FRET**).

• use a weaker laser than other platforms, conferring **less damage to the polymerase**
• FRET➔the signal is spatially confined (only in the vicinity of the Qdot): **low background noise**.
• A second signal for each base incorporation comes from the Qdot itself: as it transfers photons to the nucleotide label, it becomes dimmer for a short time: "**extra sequencing signal**"



genomeweb.com
http://www3.appliedbiosystems.com/cms/groups/global_marketing_group/documents/generaldocuments/cms_091831.pdf

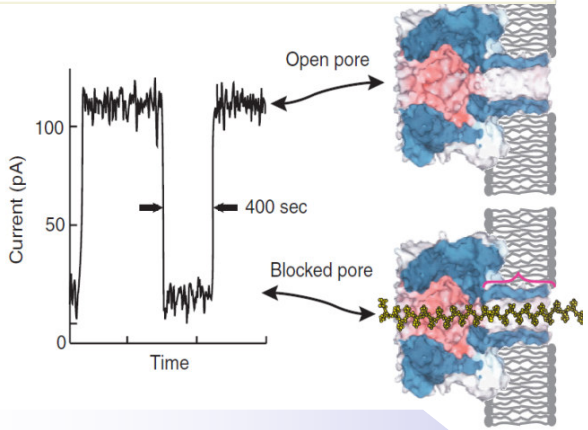
Comparison of sequencing instruments

sorted by cost/Mb, with expected performance by mid 2011

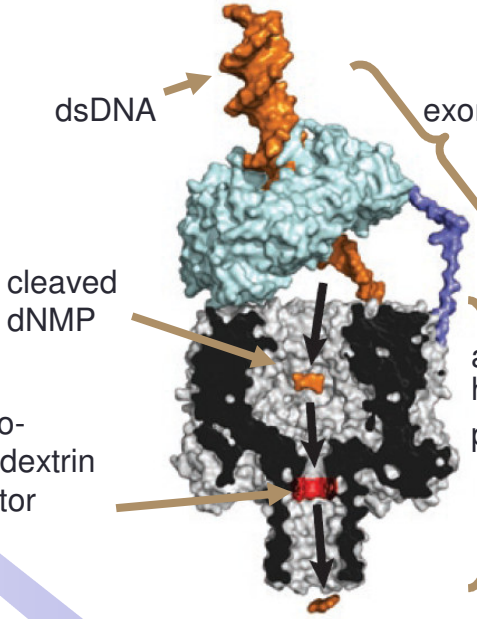
Instrument	Run time[a]	Millions of reads/run	Bases/read[b]	Yield Mb/run	Reagent cost/run[c]	Reagent cost/Mb	Minimum unit cost (% run)[d]
3730xl (capillary)	2 h	0.000096	650	0.06	$96	$1500	$6 (1%)
Ion Torrent – '314' chip	2 h	0.10	100	>10	$500	<$50	~$750 (100%)
454 GS Jr. Titanium	10 h	0.10	400	50	$1100	$22	$1500 (100%)
Starlight*	†	~0.01	>1000	†	†	†	†
PacBio RS	0.5–2 h	0.01	860–1100	5–10	$110–900	$11–180	†
454 FLX Titanium	10 h	1	400	500	$6200	$12.4	$2000 (10%)
454 FLX+[e]	18–20 h	1	700	900	$6200	$7	$2000 (10%)
Ion Torrent – '316' chip*	2 h	1	>100	>100	$750	<$7.5	~$1000 (100%)
Helicos[f]	N/A	800	35	28 000	N/A	NA	$1100 (2%)
Ion Torrent – '318' chip*	2 h	4–8	>100	>1000	~$925	~$0.93	~$1200 (100%)
Illumina MiSeq*	26 h	3.4	150 + 150	1020	$750	$0.74	~$1000 (100%)
Illumina iScanSQ	8 days	250	100 + 100	50 000	$10 220	$0.20	$3000 (14%)
Illumina GAIIx	14 days	320	150 + 150	96 000	$11 524	$0.12	$3200 (14%)
SOLiD – 4	12 days	>840[g]	50 + 35	71 400	$8128	<$0.11	$2500 (12%)
Illumina HiSeq 1000	8 days	500	100 + 100	100 000	$10 220	$0.10	$3000 (12%)
Illumina HiSeq 2000	8 days	1000	100 + 100	200 000	$20 120[h]	$0.10	$3000 (6%)
SOLiD – 5500 (PI)*	8 days	>700[g]	75 + 35	77 000	$6101	<$0.08	$2000 (12%)
SOLiD – 5500xl (4hq)*	8 days	>1410[g]	75 + 35	155 100	$10 503[h]	<$0.07	$2000 (12%)
Illumina HiSeq 2000 – v3[i]*	10 days	≤3000	100 + 100	≤600 000	$23 470[h]	≥$0.04	~$3500 (6%)

Nanopore sequencing

Strand-sequencing using ionic current blockage

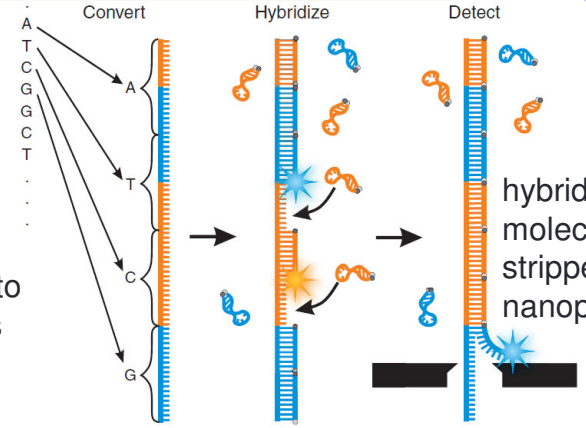


Exonuclease-sequencing by modulation of the ionic current



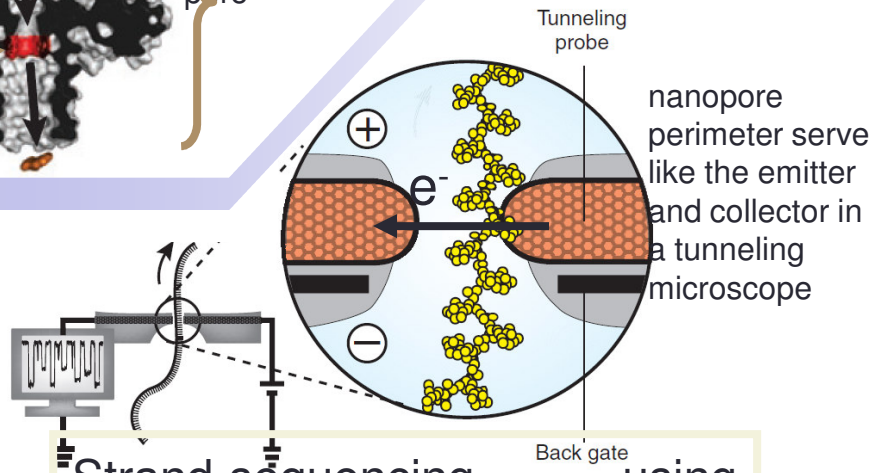
level of the current blockade ~ identity of dNMPs

Nanopore sequencing using synthetic DNA and optical readout



target ssDNA: converted to code- units (12mer oligo)

hybridized molecular beacons: stripped off using a nanopore



nanopore perimeter serve like the emitter and collector in a tunneling microscope

Strand-sequencing using transverse electron currents

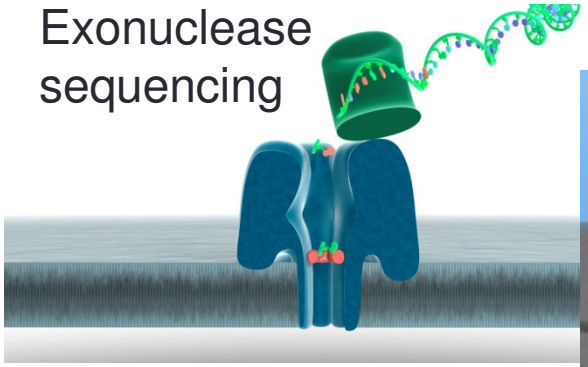
Branton et al.: The potential and challenges of nanopore sequencing Nature Biotechnology 2008.
http://www.mcb.harvard.edu/branton/projects-NanoporeSequencing.htm

mv1

Oxford NANOPORE Technologies®

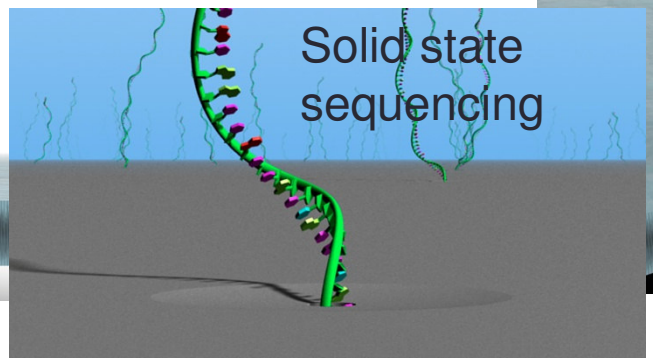
a label-free, electrical, single-molecule DNA sequencing method

Exonuclease sequencing

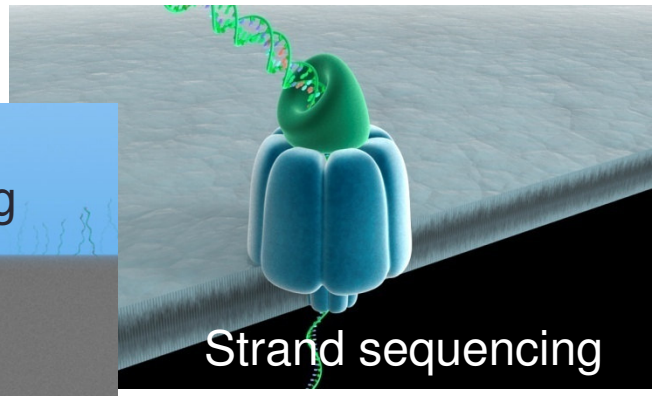


Using a processive enzyme to cleave individual nucleotides from a DNA strand and pass them through a protein nanopore.

Solid state sequencing



Using synthetic materials, rather than protein pores, to create nanopores.



Identifying individual nucleotides on a DNA strand as it passes intact through a protein nanopore.

http://www.nanoporetech.com/sequences
http://www.nanoporetech.com/sections/index/80

Slide 35

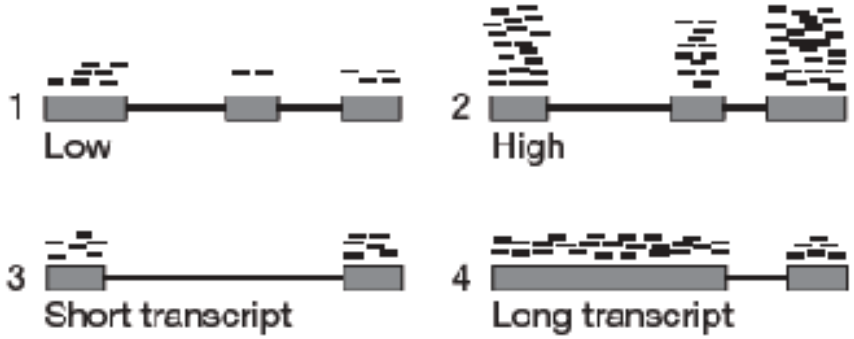
mv1

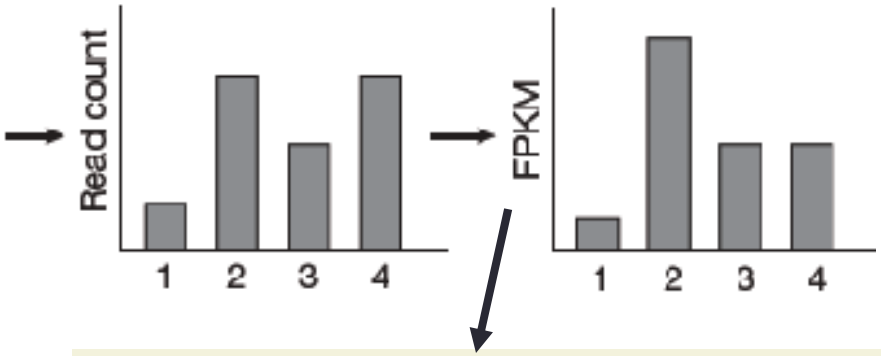
David Stoddart
Single-nucleotide discrimination in immobilized DNA oligonucleotides with a biological nanopore
http://www.pnas.org/content/106/19/7702.full

S. Garaj
Graphene as a subnanometre trans-electrode membrane
http://www.nature.com/nature/journal/v467/n7312/abs/nature09379.html
viktor.molnar, 6/7/2011

RNA-seq

Sensitivity of RNA-Seq ~
- molar concentration
- fragment length



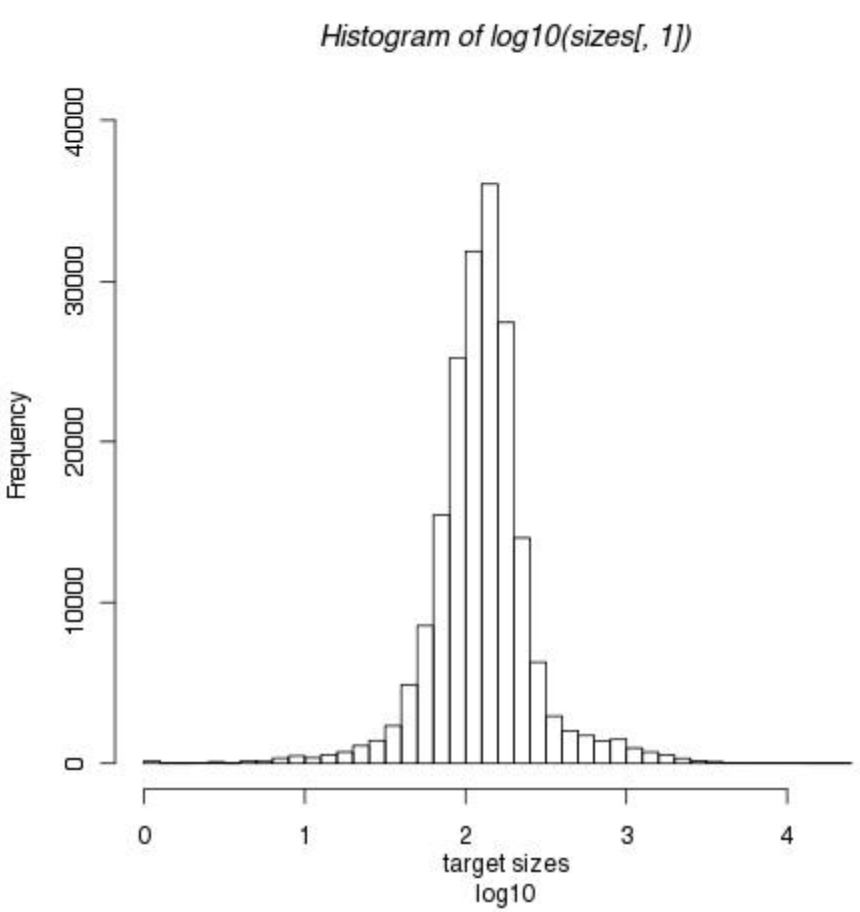


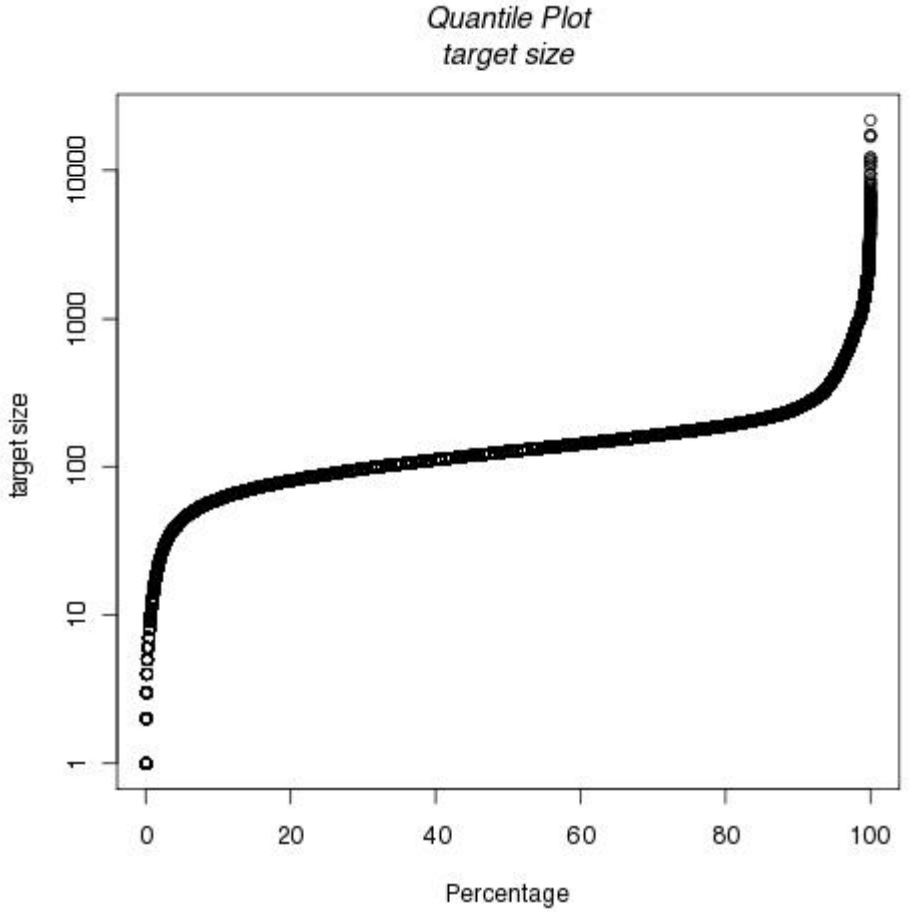
experimental procedures that generate DNA sequence reads derived from the entire RNA molecule

FPKM (fragments per kilobase of exon per million fragments mapped): **read density** reflects the molar concentration of the starting sample by **normalizing for RNA lenth** and the **total read number** in the measurement.

M Garber: Computational methods for transcriptome annotation and quantification using RNA-seq, Nature Methods, 2011 May

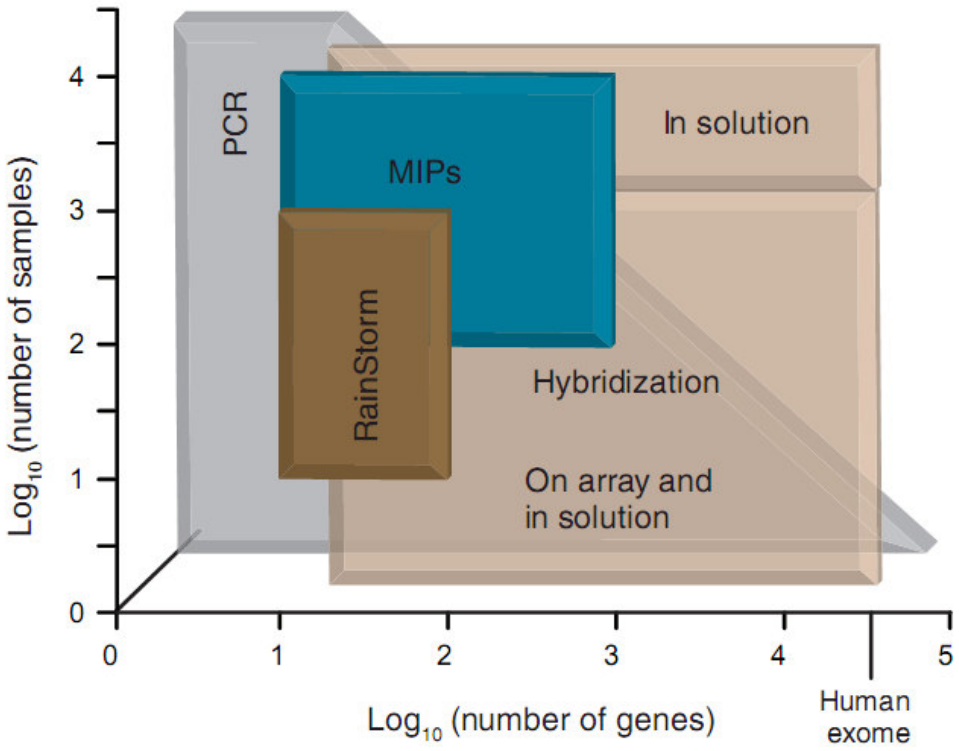
Exome length distribution





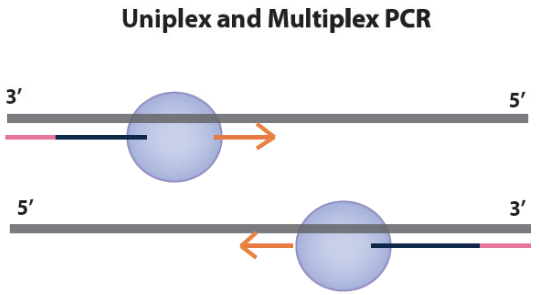
Target-enrichment strategies

Target-enrichment methods allow to selectively capture genomic regions of interest from a DNA sample prior to sequencing. Several target-enrichment strategies have been developed.



PCR

- Most widely used enrichment strategies for over 20 years.
- In classical Sanger sequencing
- Uniplex PCR used to generate a single DNA sequence is comparable in read length to a typical amplicon.
- Multiplex PCR reactions which require several primers are challenging
- Large amount of genomic target needed due to workload and quantity of DNA required.
- Highly effective
  - Not feasible to target > several megabases in size
  - Large quantity of DNA required and high cost.



Short and Long PCR

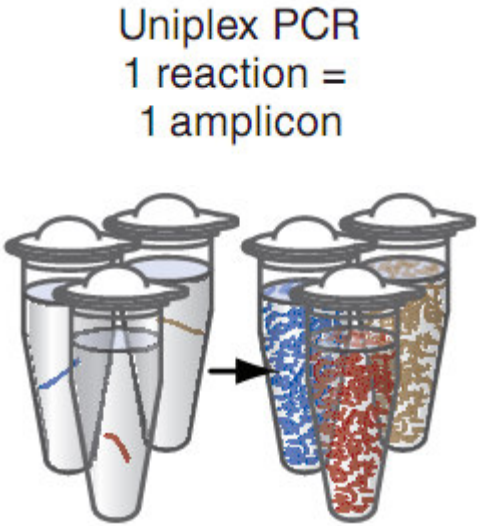
- Short
  - Length < 500 bp
  - Common, less specific, higher success rate
  - More overlap loss
  - Multiple reactions needed for large coverage
- Long
  - Length up to 50 kb (Taq polymerase)
  - Less overlap
  - Higher failure rate
  - Higher sensitivity to primer design (hairpins, GC content)

Long PCR

- Long range PCR allows the amplification of PCR products,
- Much larger than with conventional *Taq* polymerases.
- Up to 27 kb fragments are possible from good quality genomic DNA
- 10 - 20 kb fragments are routinely achievable
- Uses a mixture of thermostable DNA polymerases,
  - *Taq* DNA polymerase
    - high processivity (i.e. 5'-3' polymerase activity
  - PWO polymerase.
    - 3'-5' proofreading abilities
- Longer primer extension than can be achieved with *Taq* alone.
- Polymerase detachment results in uneven fragment lengths
  - Uneven coverage
  - Overrepresentation of regions close to primer

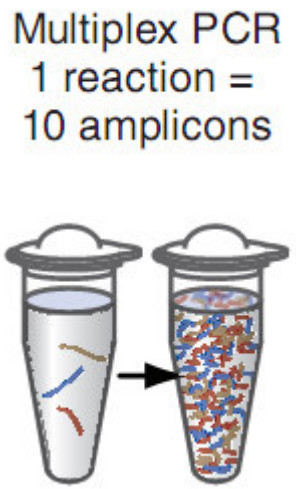
Uniplex PCR

- 1 reaction – 1 amplicon
- Compatible with all NGS platforms
- Straightforward
- 10 kb maximum practical length
  - Longer loses robustness – need multiple overlapping regions
- Validation and optimization required
  - Minimize needed total DNA mass
- Normalization needed



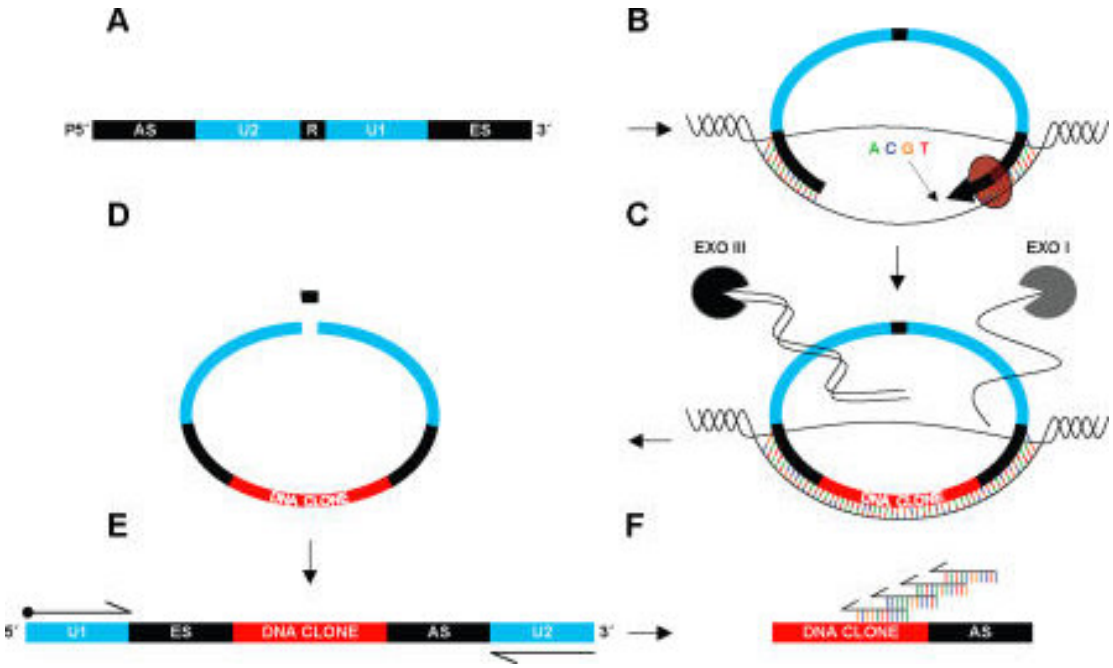
Multiplex PCR

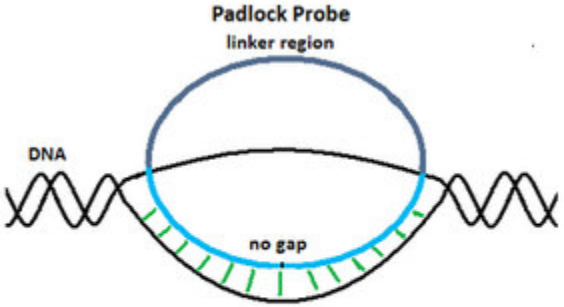
- 1 reaction ~ 10 amplicons
- Compatible with all NGS platforms
- 100 kb maximum practical length
  - Up to several hundred KB
- Multiple primers interact
  - Nonspecific amplification
  - Failure to amplify
  - Uneven amplification
- Lower specific cost
- Visual inspection of intensity bands on agarose gel

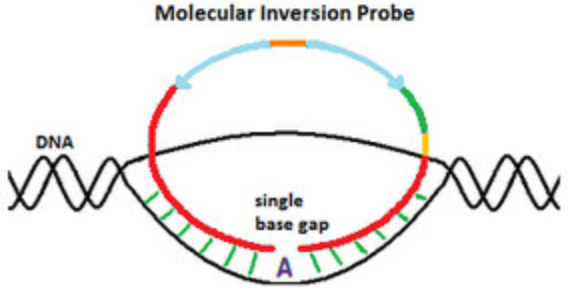


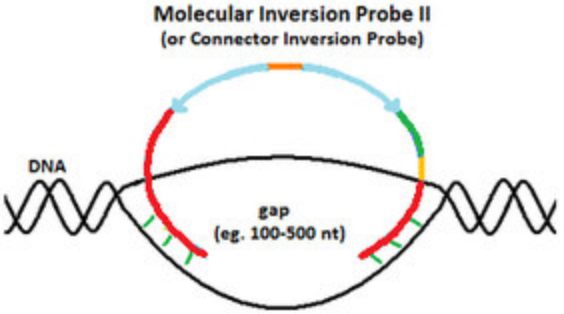
Molecular Inversion Probes (MIP)

- This is an enzymatic technique that targets the amplification of genomic regions by multiplexing based on target circularization.

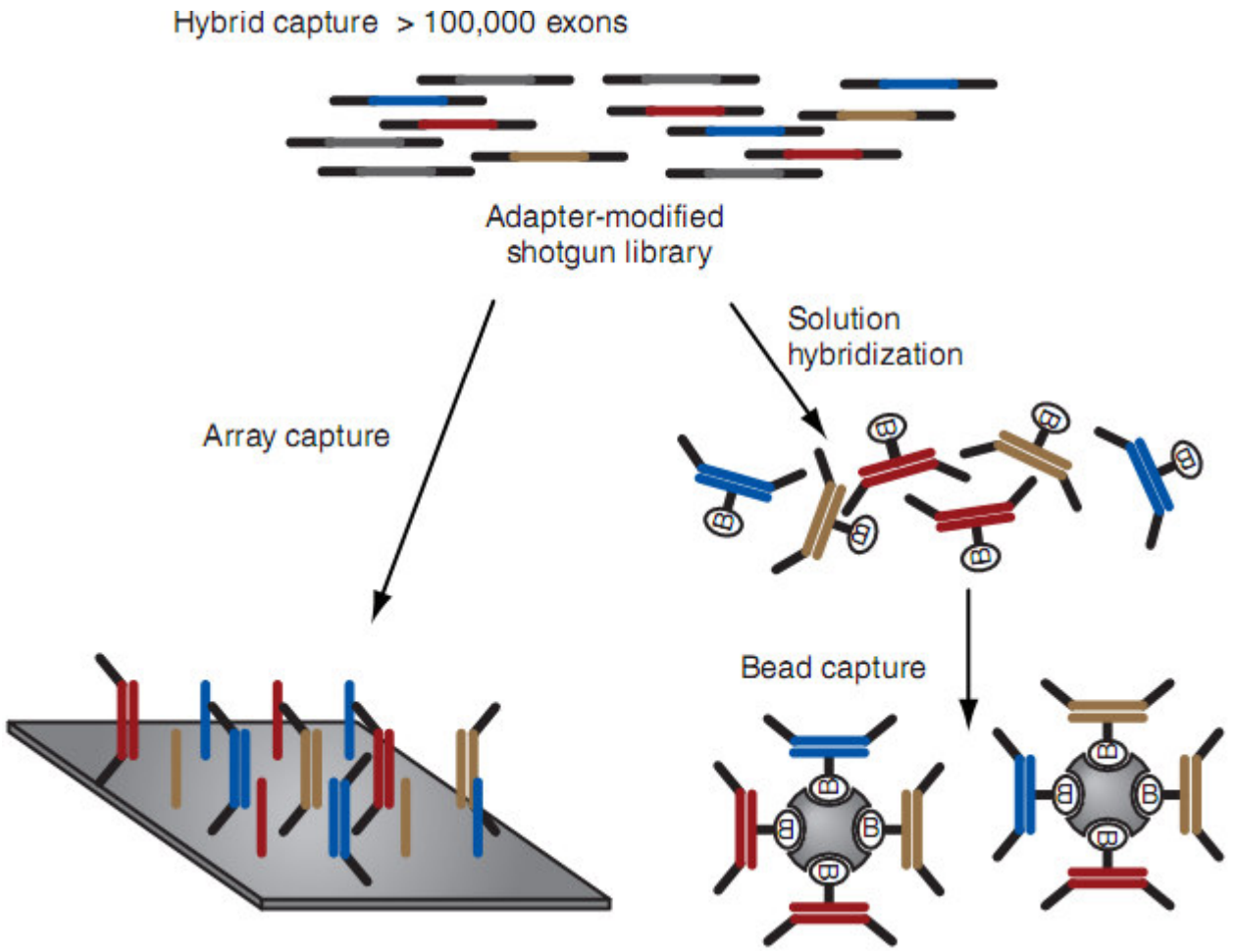




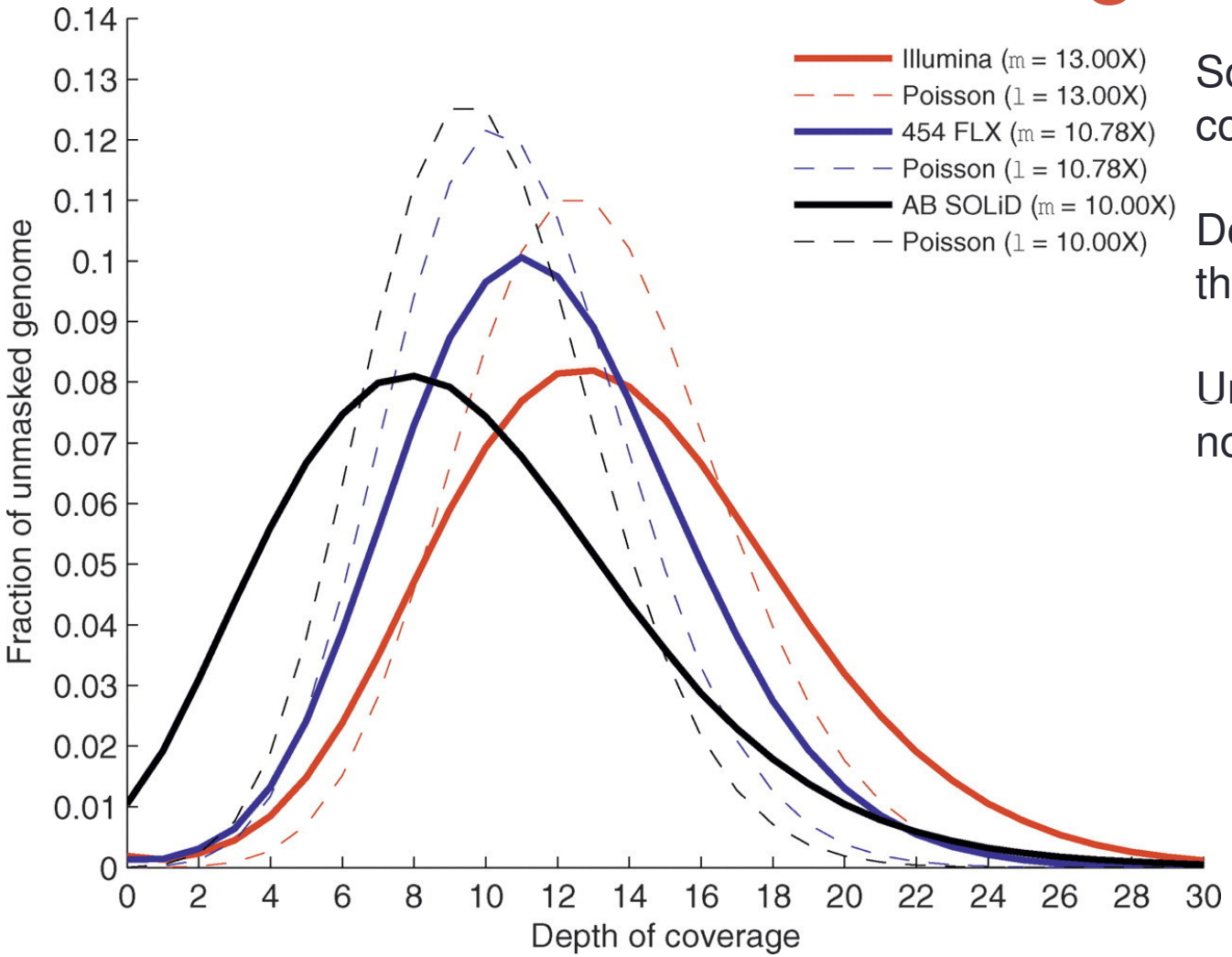




Hybrid Capture



Theoretical vs true coverage



Solid lines are actual coverage

Dotted lines show theoretical limits

Unmasked genome – no capture bias

Reproducibility

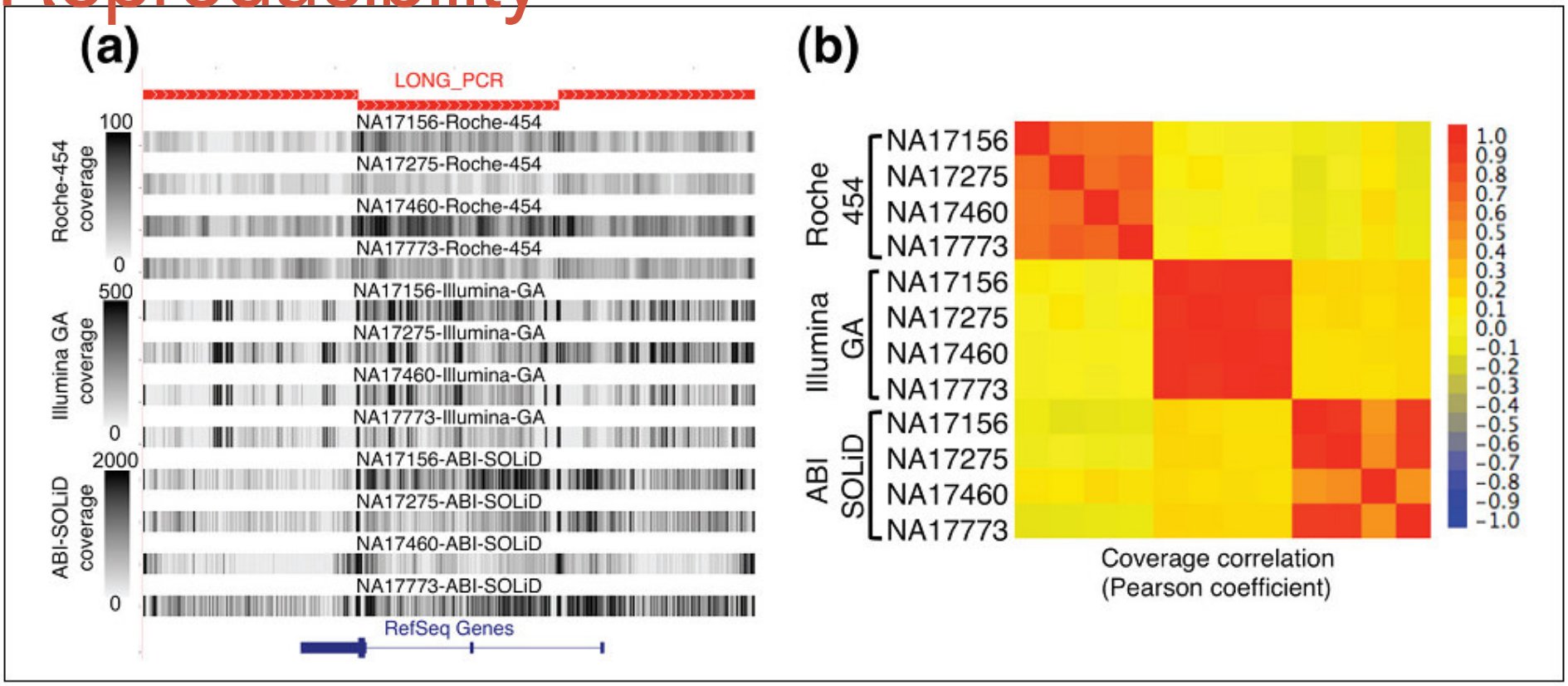


Figure 3

Each NGS technology generates a consistent pattern of non-uniform sequence coverage. **(a)** Sequence coverage depth is displayed as a gray-scale (0-100× for Roche 454; 0-500× for Illumina GA and ABI SOLiD) along an approximately 25-kb region of chromosome 11 amplified by three long-range PCR products (red rectangles). **(b)** A heat-map colored matrix displays the coefficient of correlation of coverage across the entire 260 kb of analyzed sequence between each of the 72 possible pair-wise comparisons (four samples by three technologies). The apparent lower correlation of the Roche-454 sequence coverage is more reflective of the smaller amplitude in the coverage variability (lower average coefficient of variance) than a lack of coverage correlation from sample to sample. The correlation of NA17460 with the other three samples on the ABI SOLiD platform is slightly lower due to technological issues (Additional data file 2) and was therefore excluded from the coefficient of correlation calculation reported in the text.

Methods I.

Sequencing device subdivision

- Achieves physical separation
- No additional steps required
- Limited subdivision level
- No additional associated cost
- Full read length retained
- Limited post-processing required

Methods II.

Bar code library ligation

- Expensive and time consuming
- Bar codes must be designed and synthesized
- Large numbers of samples can be multiplexed
- Reduces read length
- Requires post sequencing separation prior to mapping

Methods III.

Primer bar coding

- Used in amplicon sequencing
- Bar code is unique to primer, not sample
- Lower cost

Actual codes:

```
GSMIDs{   mid = "MID1", "ACGAGTGCGT", 2;
          mid = "MID2", "ACGCTCGACA", 2;
          mid = "MID3", "AGACGCACTC", 2;
          mid = "MID4", "AGCACTGTAG", 2;
          mid = "MID5", "ATCAGACACG", 2;
          mid = "MID6", "ATATCGCGAG", 2;
          mid = "MID7", "CGTGTCTCTA", 2;
          mid = "MID8", "CTCGCGTGTC", 2;
          mid = "MID9", "TAGTATCAGC", 2;
          mid = "MID10", "TCTCTATGCG", 2;
          mid = "MID11", "TGATACGTCT", 2;
          mid = "MID12", "TACTGAGCTA", 2;
          mid = "MID13", "CATAGTAGTG", 2;
          mid = "MID14", "CGAGAGATAC", 2;}
```

- Hamming distance of 6
- Detection of 5 errors
- Correction of 2 errors

```
RLMIDs{   mid = "RL1", "ACACGACGACT", 1, "AGTCGTGGTGT";
          mid = "RL2", "ACACGTAGTAT", 1, "ATACTAGGTGT";
          mid = "RL3", "ACACTACTCGT", 1, "ACGAGTGGTGT";
          mid = "RL4", "ACGACACGTAT", 1, "ATACGTGGCGT";
          mid = "RL5", "ACGAGTAGACT", 1, "AGTCTACGCGT";
          mid = "RL6", "ACGCGTCTAGT", 1, "ACTAGAGGCGT";
          mid = "RL7", "ACGTACACACT", 1, "AGTGTGTGCGT";
          mid = "RL8", "ACGTACTGTGT", 1, "ACACAGTGCGT";
          mid = "RL9", "ACGTAGATCGT", 1, "ACGATCTGCGT";
          mid = "RL10", "ACTACGTCTCT", 1, "AGAGACGGAGT";
          mid = "RL11", "ACTATACGAGT", 1, "ACTCGTAGAGT";
          mid = "RL12", "ACTCGCGTCGT", 1, "ACGACGGGAGT";}
```

- Hamming distance of 4
- Detection of 3 errors
- Correction of 1 error

Raw error rate : Roche 454

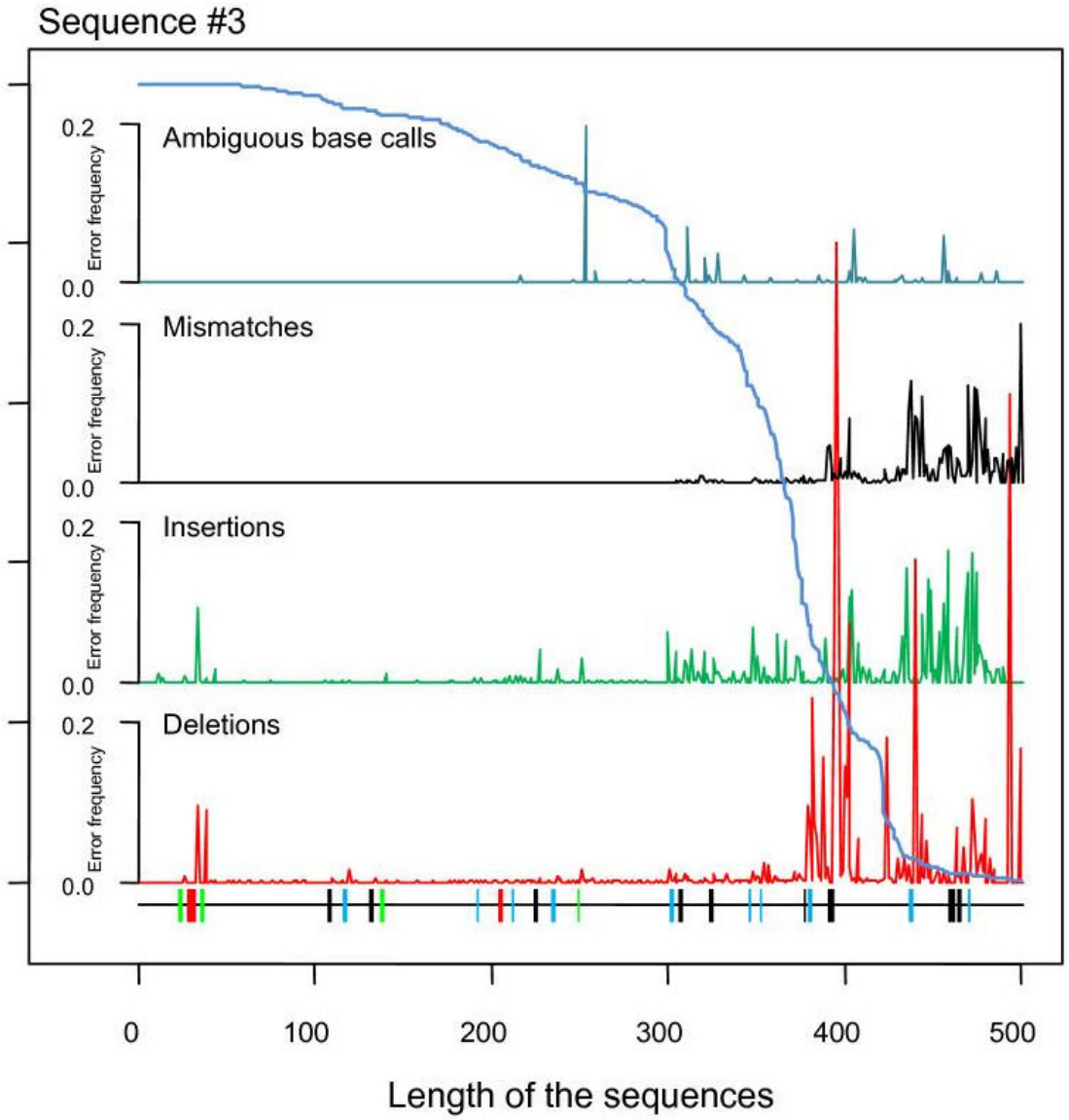
- Raw error rate
  - 0.1% (reported substitution error)
  - 1.07% (experimental, not randomly distributed)
    - Rose to 50% - sequence dependant
- Sequence position
- Sequence size
- Physical localization on PicoTiter plates
  - For indels
- Homopolymer streches are difficult to sequence
  - 10% error rate with 6bp
  - 50% error rate with 8bp

Errors vs sequence length

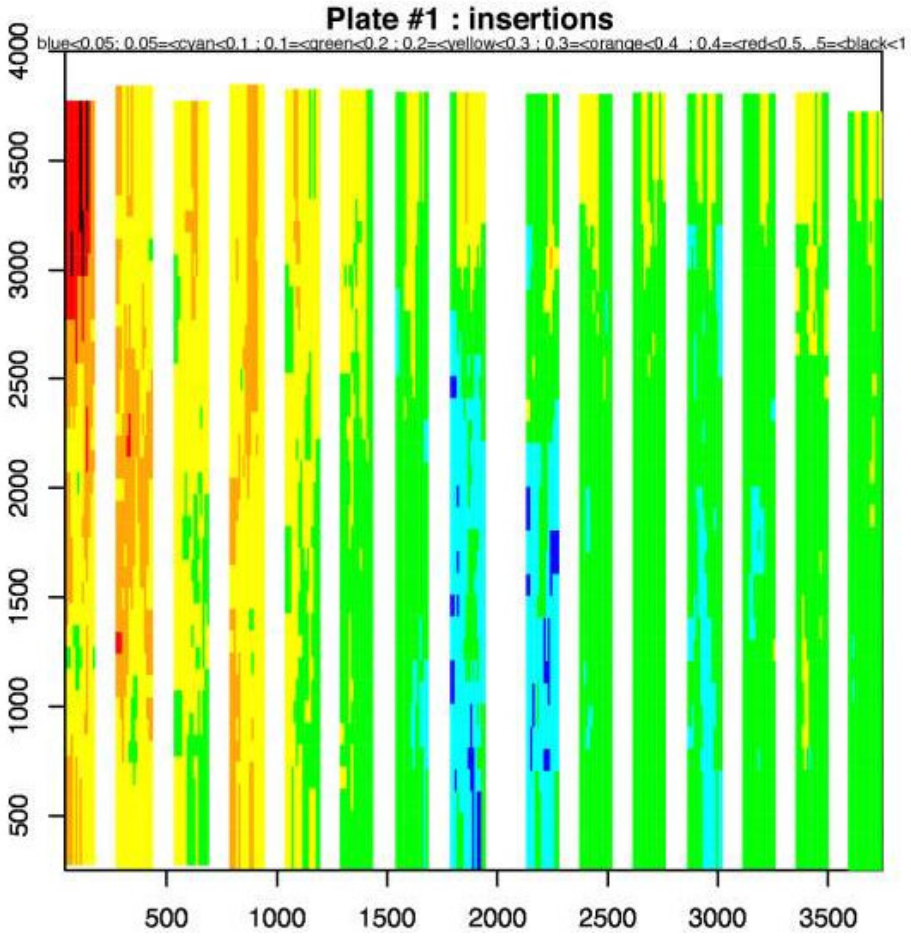
Blue line is the proportion of generated sequences

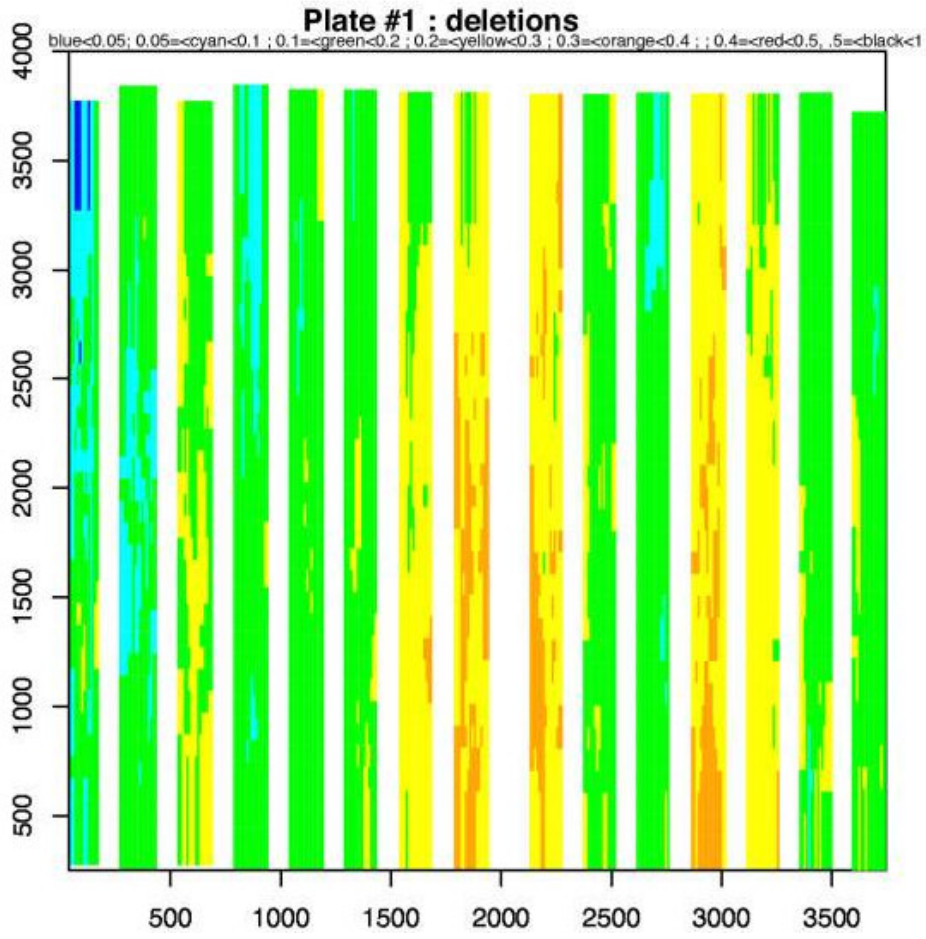
The position and length of homopolymers for each base are given on the x-axis

Green: A, red: T, black: G, blue: C



Error localization on PT

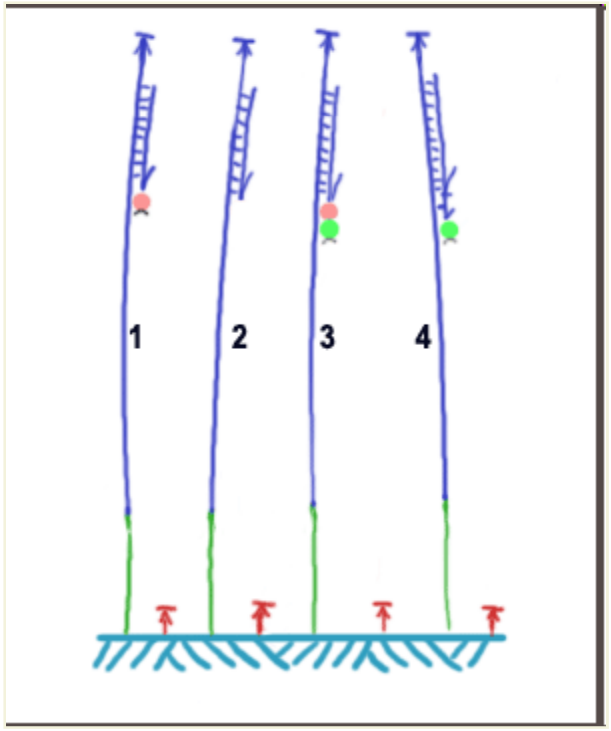




Dephasing: why are the read not longer?

A technical concern of Illumina sequencing is that base-call **accuracy decreases with increasing read length**

During a given sequencing cycle, nucleotides can be under- or overincorporated, or block removal can fail. ➔ these aberrations accumulate to produce a heterogeneous population in a cluster of strands of varying lengths ➔ decreases signal purity and reduces precision in base calling, especially at the 3' ends of reads.



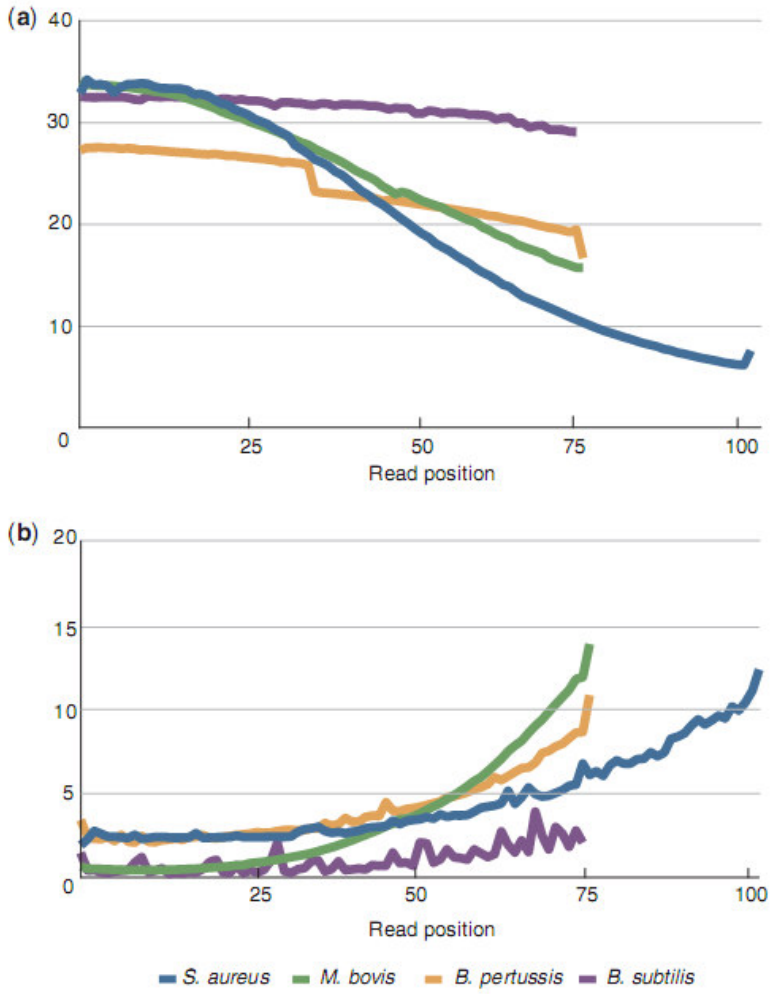
Dephasing: the main reason of a limited readlength.
possible reasons :

1. normal extension // no dephasing;
2. nucleotide does not incorporated // negative dephasing on next cycle;
3. incorporated nucleotide have no terminator // positive dephasing;
4. incorporated nucleotide have no fluorophore and terminator // positive dephasing;

it is also possible, that terminator was not removed during cleavage, it will result in negative dephasing on a next cycle

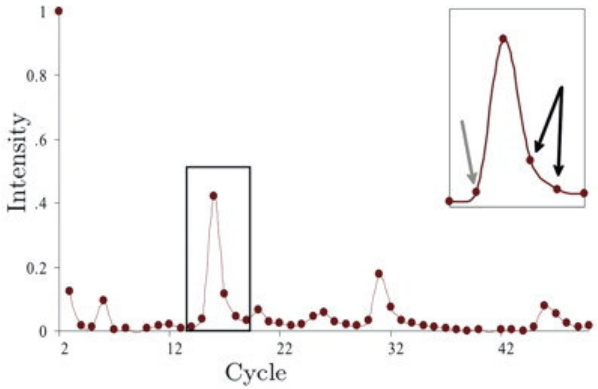
Illumina Dephasing

- Increasingly higher error rate towards the tail (3')
- As seqencing by synthesis progresses, individual amplicons in the clonal cluster progress at different rates,
  - A: Phred score
  - B: Mismatch ratio
- Discard 3' ends of fragments

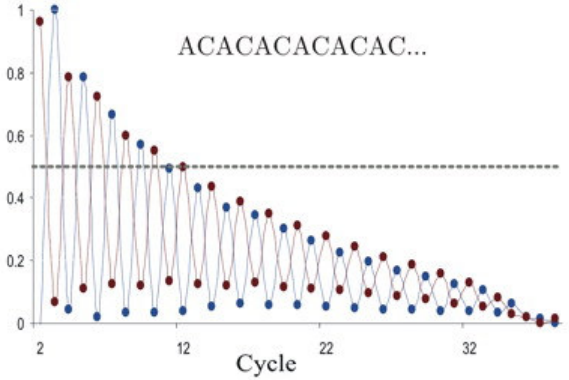


Illumina noise patterns

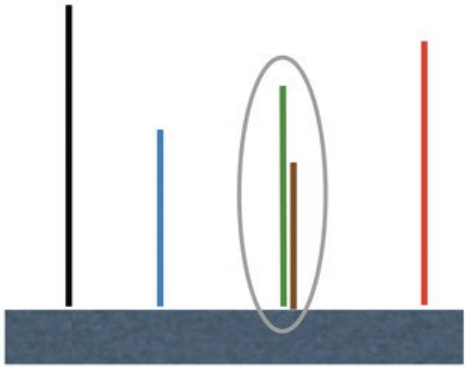
Phasing noise ϕ



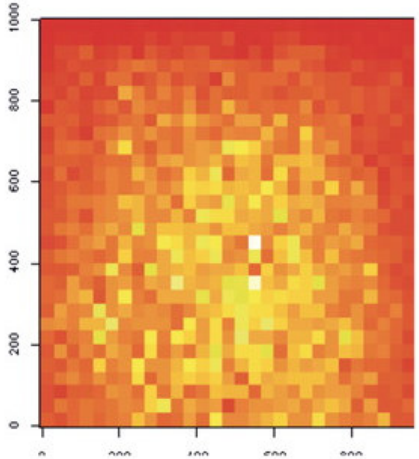
Signal Decay δ



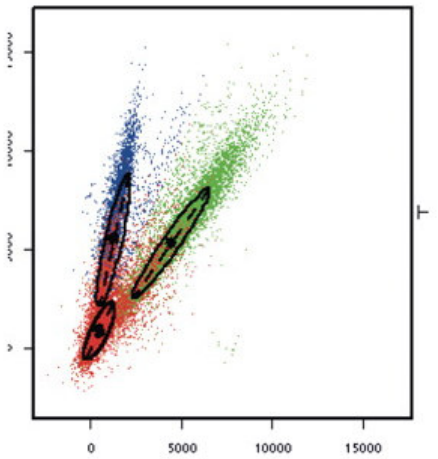
Mixed Cluster μ



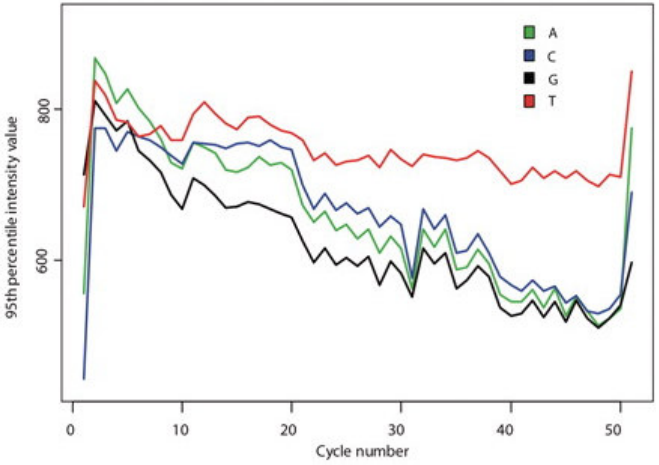
Boundary effects ω



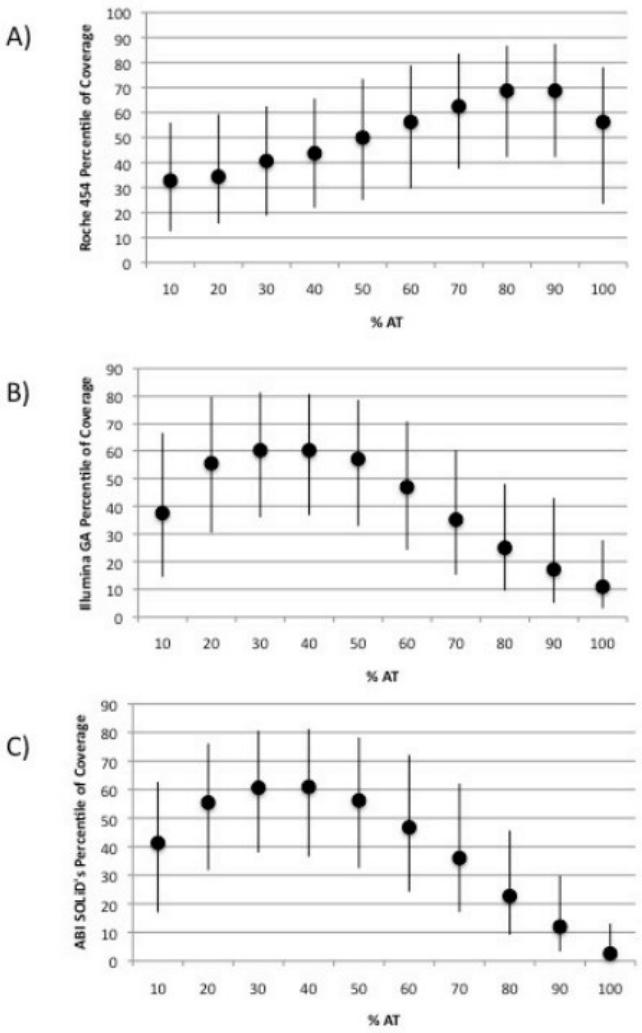
Cross-talk $\sum$



T fluophore accumulation τ



AT rich regions



- 454/Illumina/ABI
- Nonoverlapping 10-bp windows
  - windows with a particular AT % content grouped together (x-axis)
- Percentile coverage relative to the entire sample
- Illumina: coverage decreases with AT%
- 454: tolerates a wider range of AT%
- ABI: very sharp decline in coverage of AT rich regions