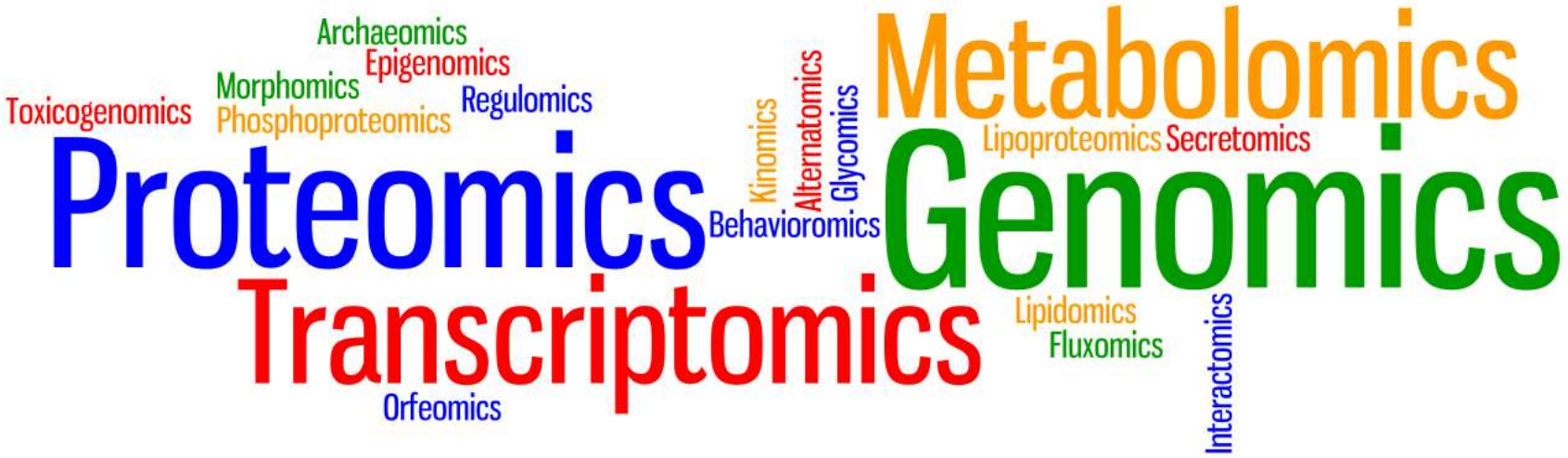


TOPIC 1. DNA-seq: techniques

Sanger sequencing. 2nd and 3rd generation sequencing techniques: 454, Illumina, SOLiD, IonTorrent, Oxford-Nanopore, PacBio.

Omics Techniques

Bachelor's Degree in Bioinformatics

Jaime Martinez-Urtaza, UAB

Why sequence a genome in the first place?

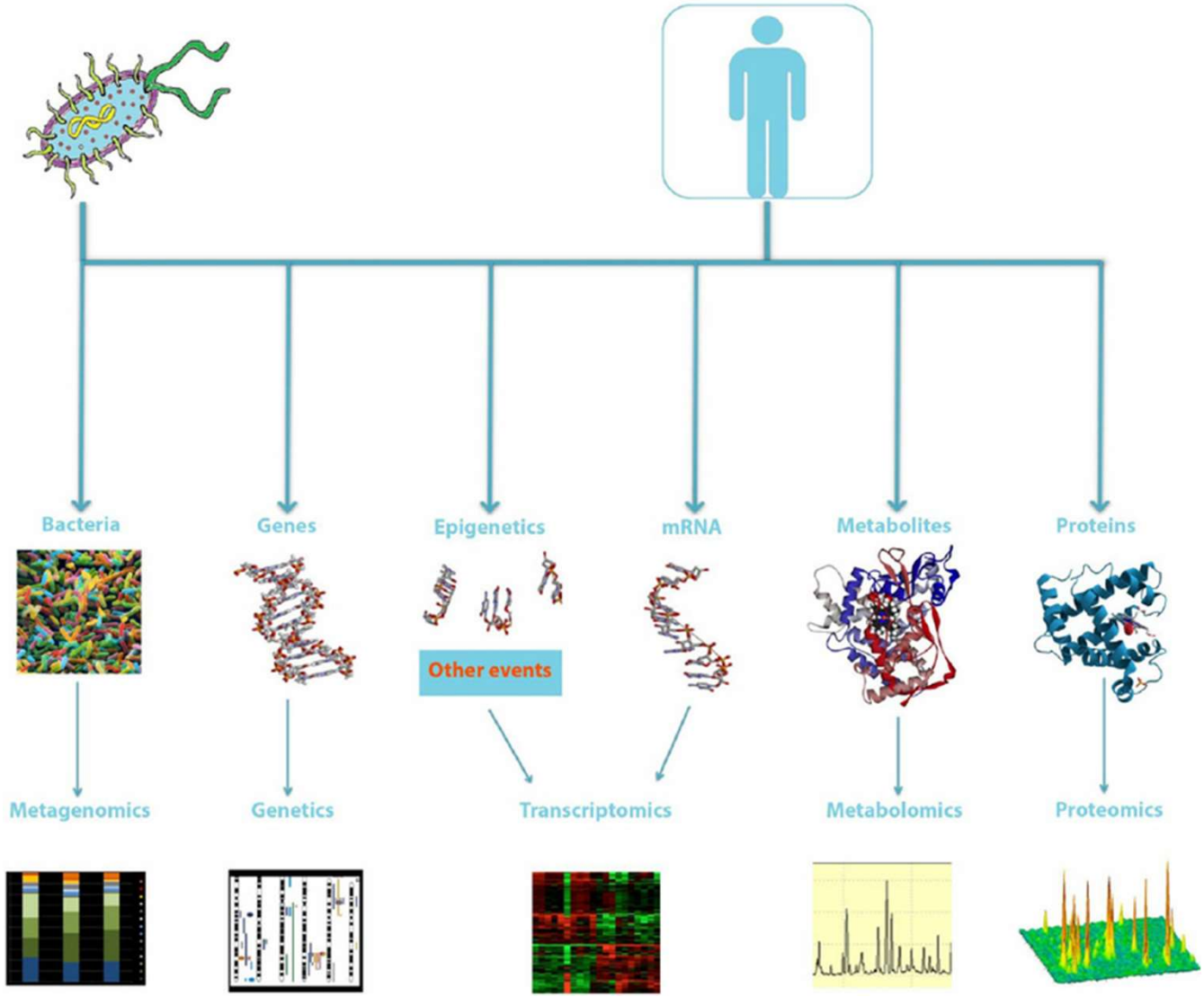
3 main reasons

- Description of sequence of every gene valuable. Includes regulatory regions which help in understanding not only the "biochemical" activities of a cell but also ways in which they are controlled.

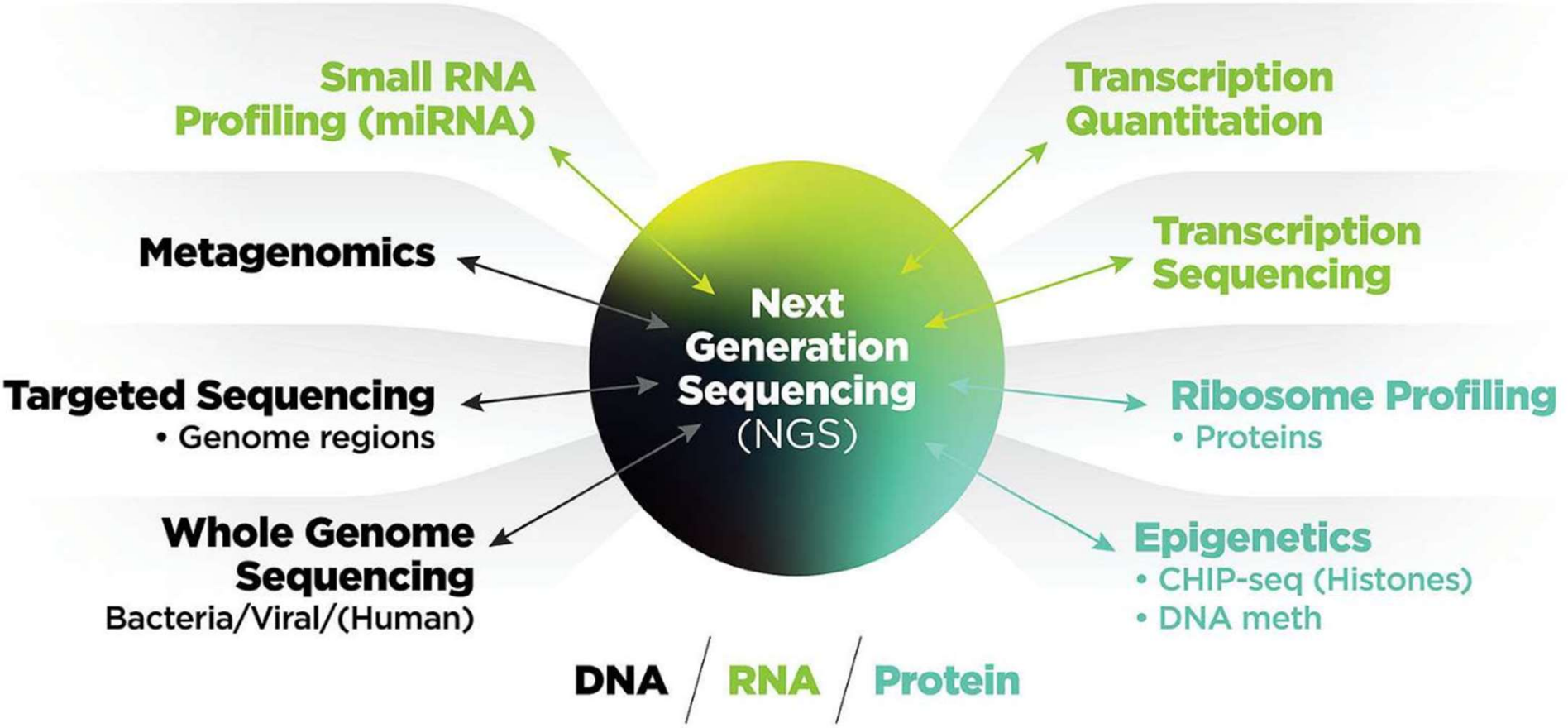
- Identify & characterise important inheritable disease genes or "useful" functional genes (e.g. bacterial genes for industrial use)

- To understand relationships between organisms and provide information on how they evolve.

Genetic and genomic technologies



Applications of Next Generation Sequencing Technologies



Source: Mehta NAL, Dow DJ, Battram AM. 2011. DNA sequencing technologies and emerging applications in drug discovery. European Pharmaceutical Review website. https://www.europeanpharmaceuticalreview.com/article/10409/dna-sequencing-technologies-and-emergingapplications-in-drug-discovery/. Accessed May 4, 2020.
Note: Colors used in the diagram are an adaptation of the original.

Glossary

Alignment	Similarity-based arrangement of DNA, RNA or protein sequences. In this context, subject and query sequence should be orthologous and reflect evolutionary, not functional or structural relationships
Annotation	Computational process of attaching biologically relevant information to genome sequence data
Assembly	Computational reconstruction of a longer sequence from smaller sequence reads
Barcode	Short-sequence identifier for individual labelling (barcoding) of sequencing libraries
BAC	(Bacterial artificial chromosome) DNA construct of various length (150–350 kb)
cDNA	Complementary DNA synthesized from an mRNA template
Contig	A contiguous linear stretch of DNA or RNA consensus sequence. Constructed from a number of smaller, partially overlapping, sequence fragments (reads)
Coverage	Also known as 'sequencing depth'. *Sequence coverage* refers to the average number of reads per locus and differs from *physical coverage*, a term often used in genome assembly referring to the cumulative length of reads or read pairs expressed as a multiple of genome size
De novo assembly	Refers to the reconstruction of contiguous sequences without making use of any reference sequence
EST library	Expressed sequence tag library. A short subsequence of cDNA transcript sequence
Fosmid	A vector for bacterial cloning of genomic DNA fragments that usually holds inserts of around 40 kb
GC content	The proportion of guanine and cytosine bases in a DNA/RNA sequence
Gene ontology (GO)	Structured, controlled vocabularies and classifications of gene function across species and research areas
InDel	Insertion/deletion polymorphism
Insert size	Length of randomly sheared fragments (from the genome or transcriptome) sequenced from both ends
K-mer	Short, unique element of DNA sequence of length k, used by many assembly algorithms
Library	Collection of DNA (or RNA) fragments modified in a way that is appropriate for downstream analyses, such as high-throughput sequencing in this case
Mapping	A term routinely used to describe alignment of short sequence reads to a longer reference sequence

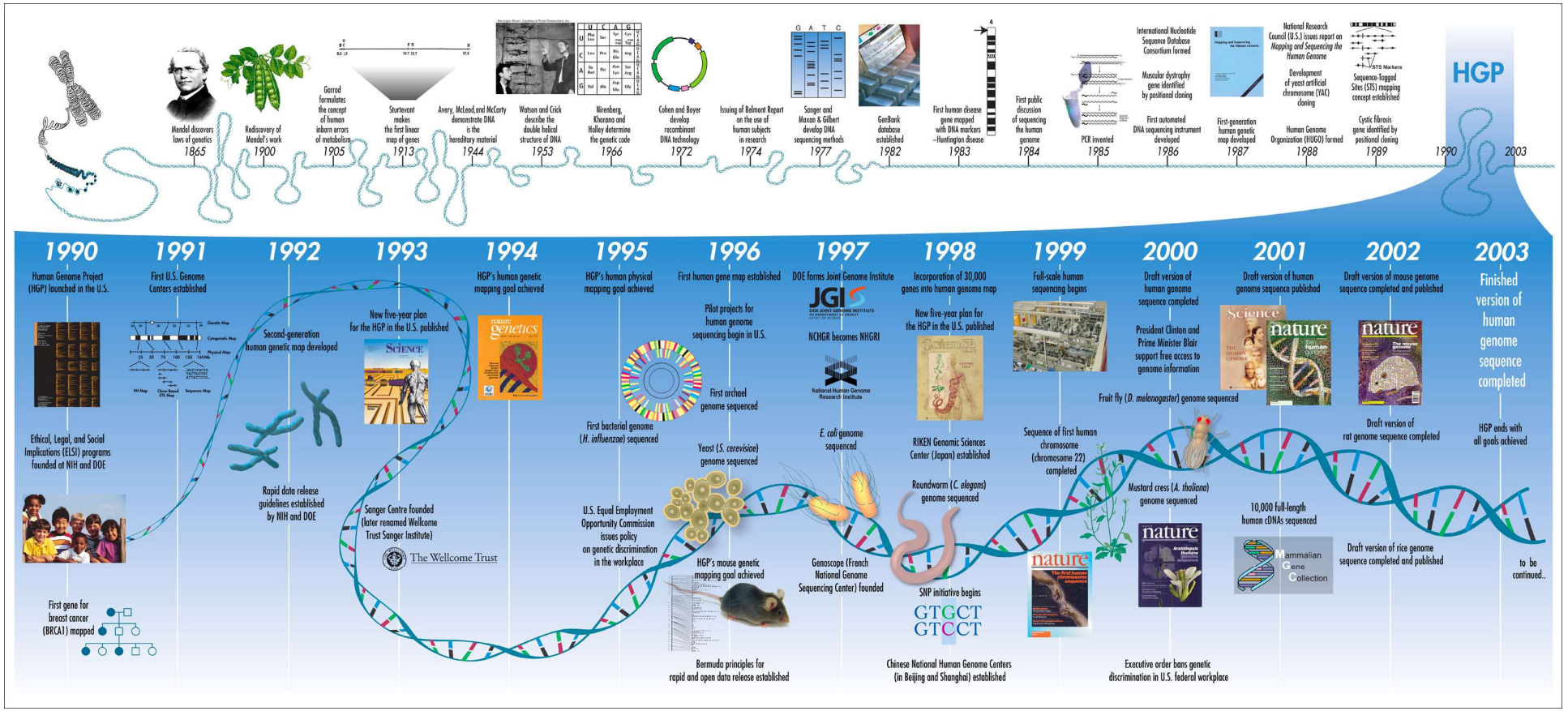
Evol Appl. 2014 Nov; 7(9): 1026–1042.

Glossary

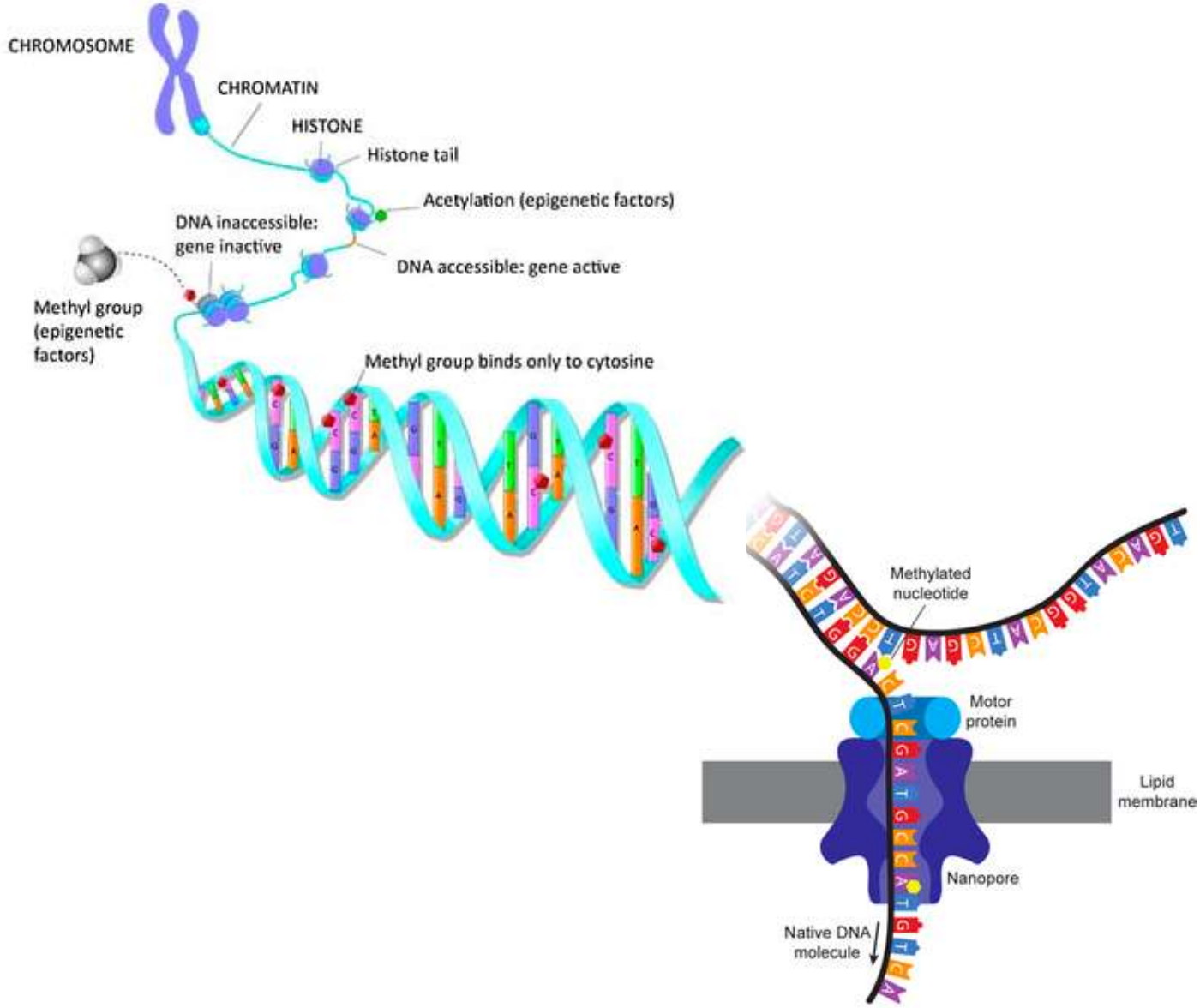
Masking	Converting a DNA sequence [A,C,G,T] (usually repetitive or of low quality) to the uninformative character state N or to lower case characters [a,c,g,t] (*soft masking*)
Massively parallel (or next generation) sequencing	High-throughput sequencing nano-technology used to determine the base-pair sequence of DNA/RNA molecules at much larger quantities than previous end-termination (e.g. Sanger sequencing) based sequencing techniques
Mate-pair	Sequence information from two ends of a DNA fragment, usually several thousand base-pairs long
N50	A statistic of a set of contigs (or scaffolds). It is defined as the length for which the collection of all contigs of that length or longer contains at least half of the total of the lengths of the contigs
N90	Equivalent to the N50 statistic describing the length for which the collection of all contigs of that length or longer contains at least 90% of the total of the lengths of the contigs
Optical map	Genomewide, ordered, high-resolution restriction map derived from single, stained DNA molecules. It can be used to improve a genome assembly by matching it to the genomewide pattern of expected restriction sites, as inferred from the genome sequence
Paired-end sequencing	Sequence information from two ends of a short DNA fragment, usually a few hundred base pairs long
Read	Short base-pair sequence inferred from the DNA/RNA template by sequencing
RNA-Seq	High-throughput shotgun transcriptome (cDNA) sequencing. Usually not used synonymous to RNA-sequencing which implies direct sequencing of RNA molecules skipping the cDNA generation step
Scaffold	Two or more contigs joined together using read-pair information
Transcriptome	Set of all RNA molecules transcribed from a DNA template

Evol Appl. 2014 Nov; 7(9): 1026–1042.

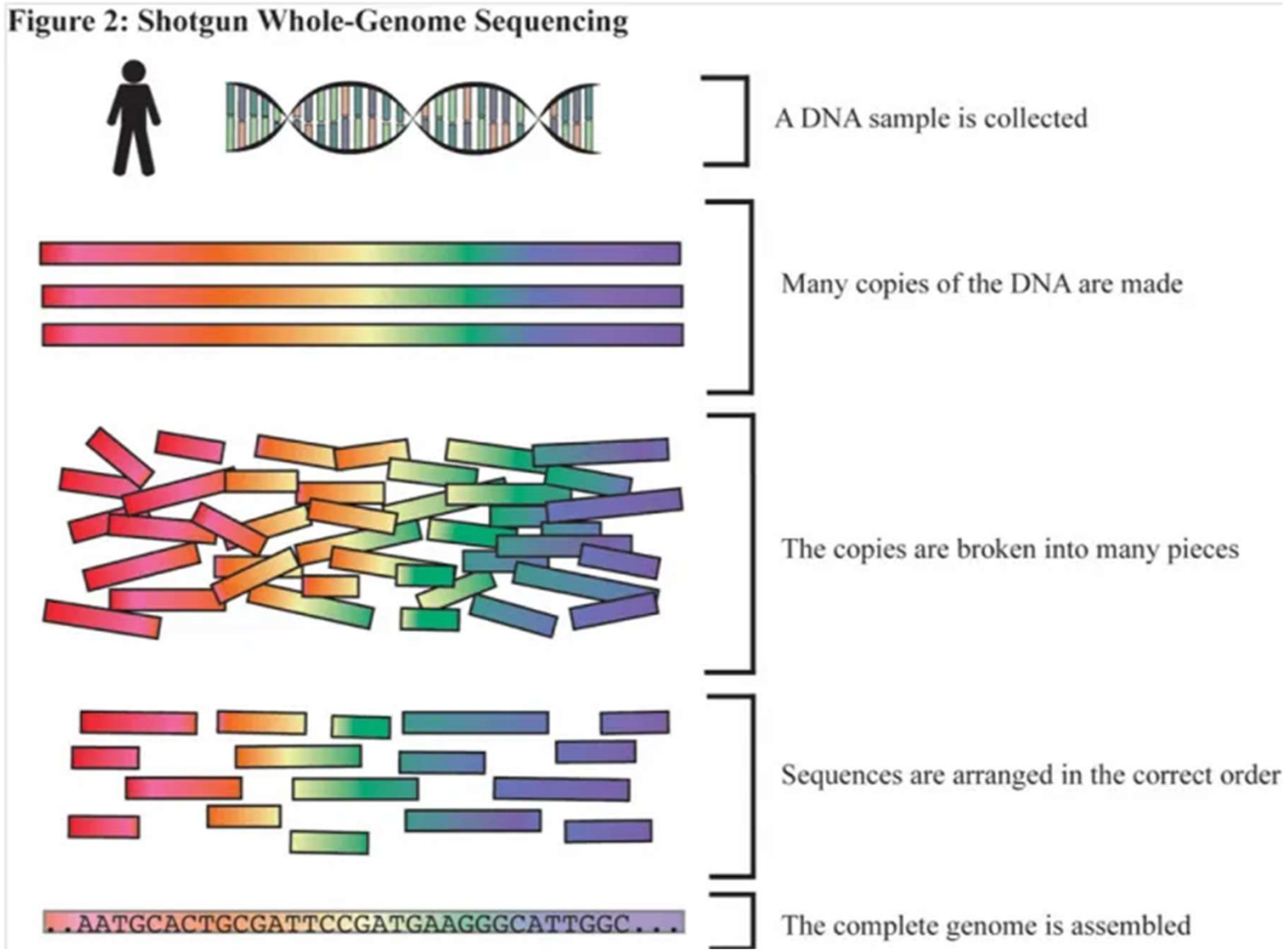
Timeline Sequencing Technologies



Sequencing Technologies



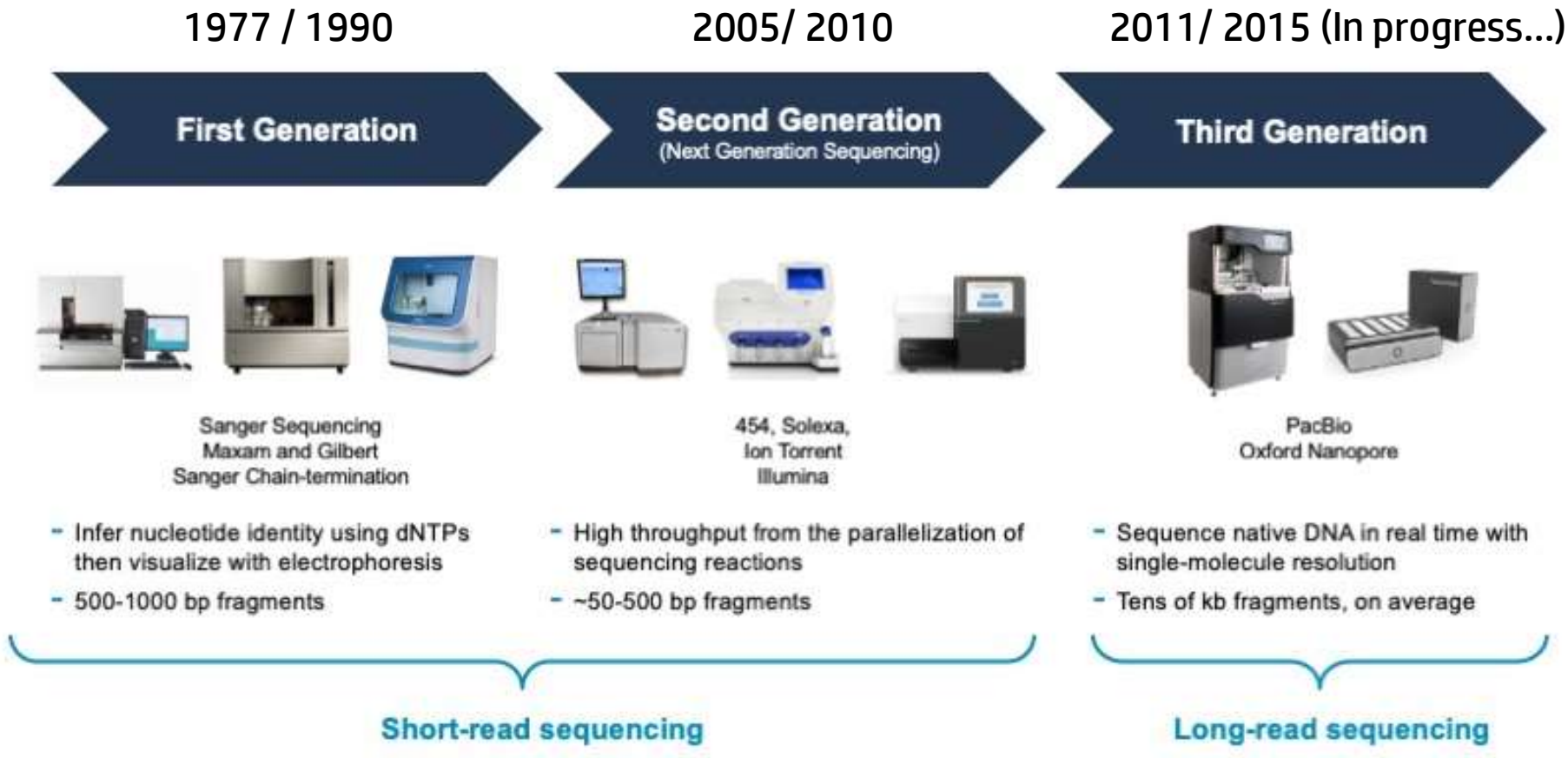
Sequencing Technologies



Sequencing Technologies

- First generation: Sanger (1977 / 1990)
- Second generation: 454 (2005)
  Illumina (2006)
  SOLiD (2007)
  Ion Torrent (2010)
- Third generation: PacBio (2011)
  Nanopore (2014)

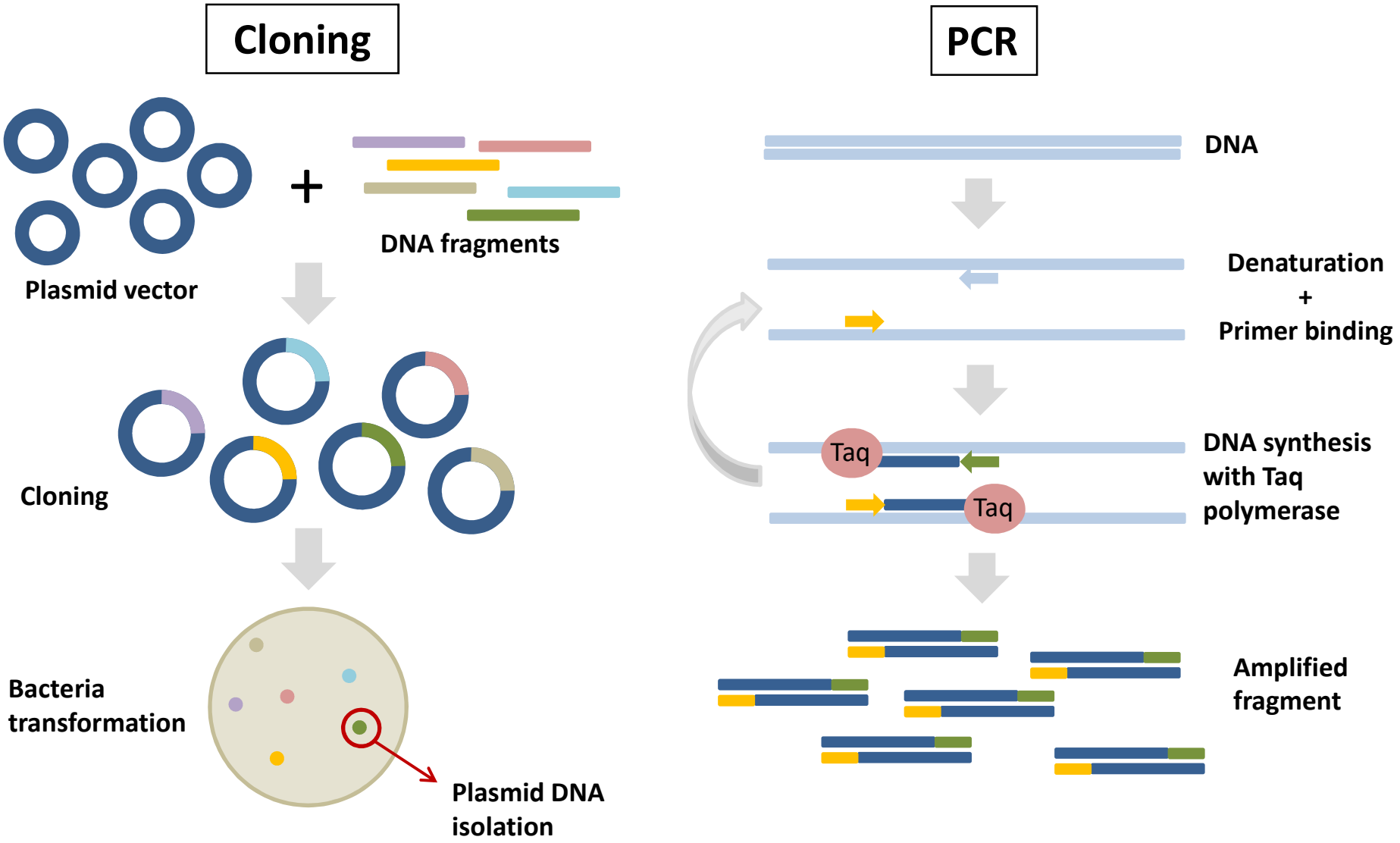
Sequencing Technologies



https://twitter.com/PacBio/status/1233091102800011266?s=20

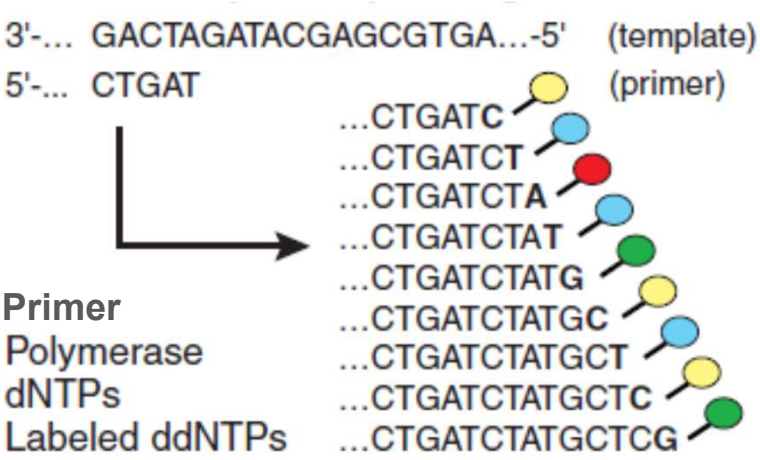
Sanger sequencing method

1. DNA amplification

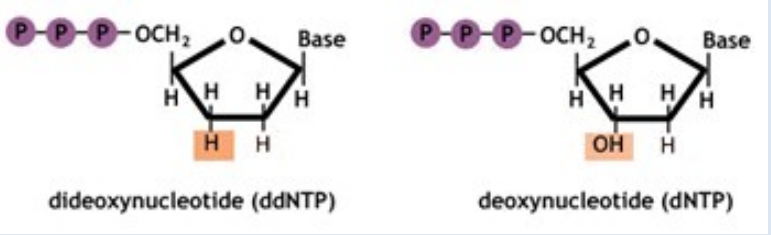


Sanger sequencing method

2. Sequencing reaction



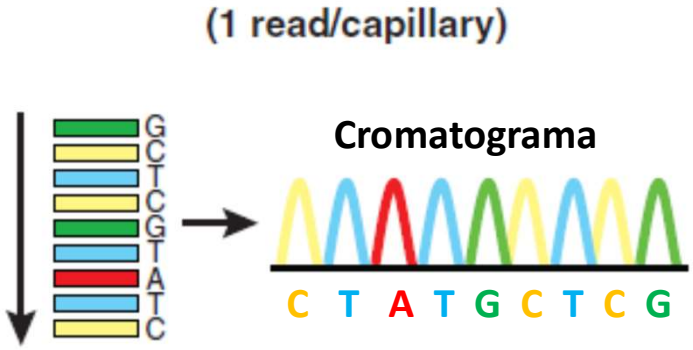
Dideoxynucleotides (ddNTPs)



No additional nucleotide can be incorporated

Each ddNTP is labelled with a different fluorochrome

3. Capillary electrophoresis

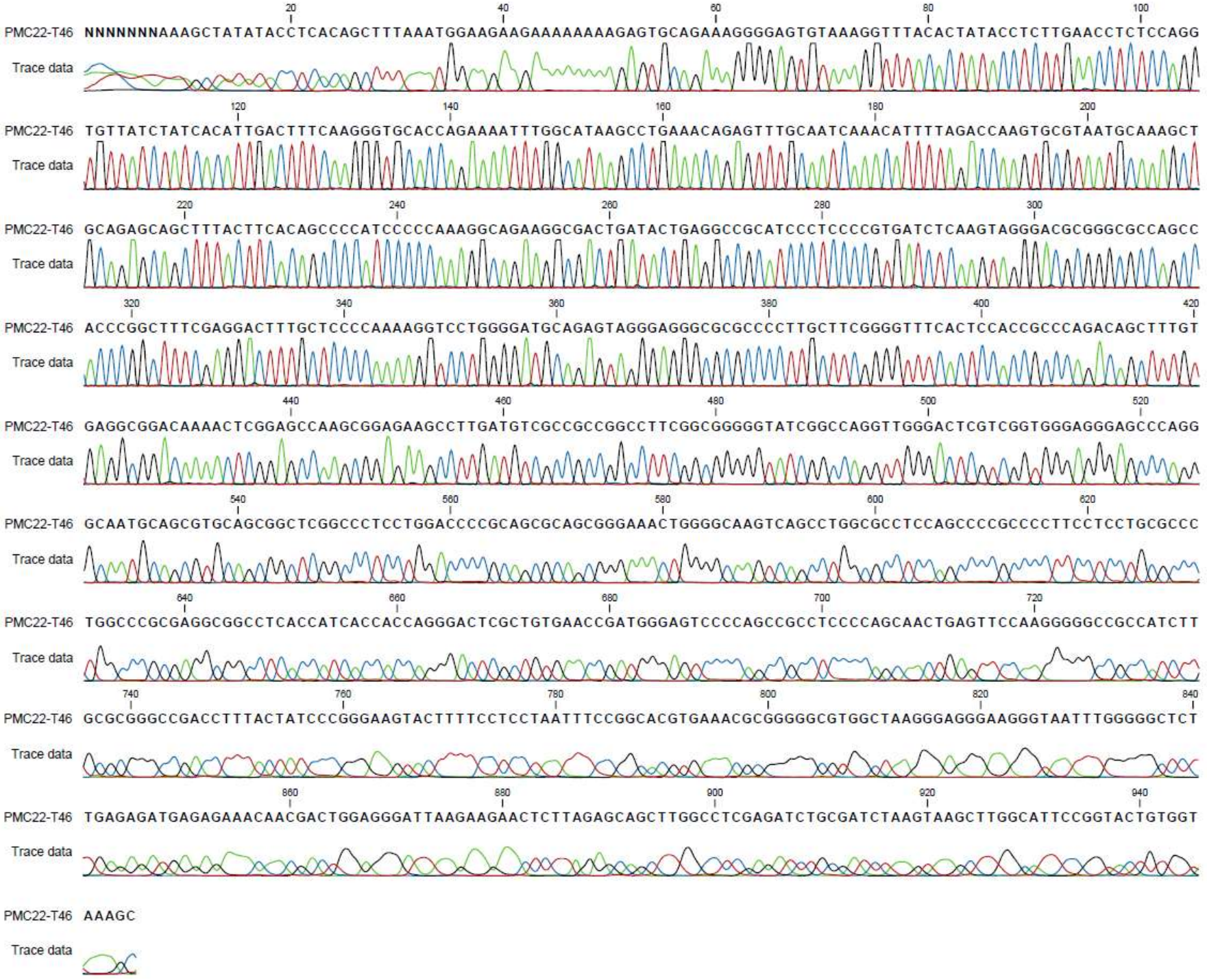


Size fragment separation in polyacrilamide gen + **Fluorescence detection**

RESULT OF SANGER SEQUENCING

- Long reads (500-1000 bp)
- Low throughput (96 reactions/run)

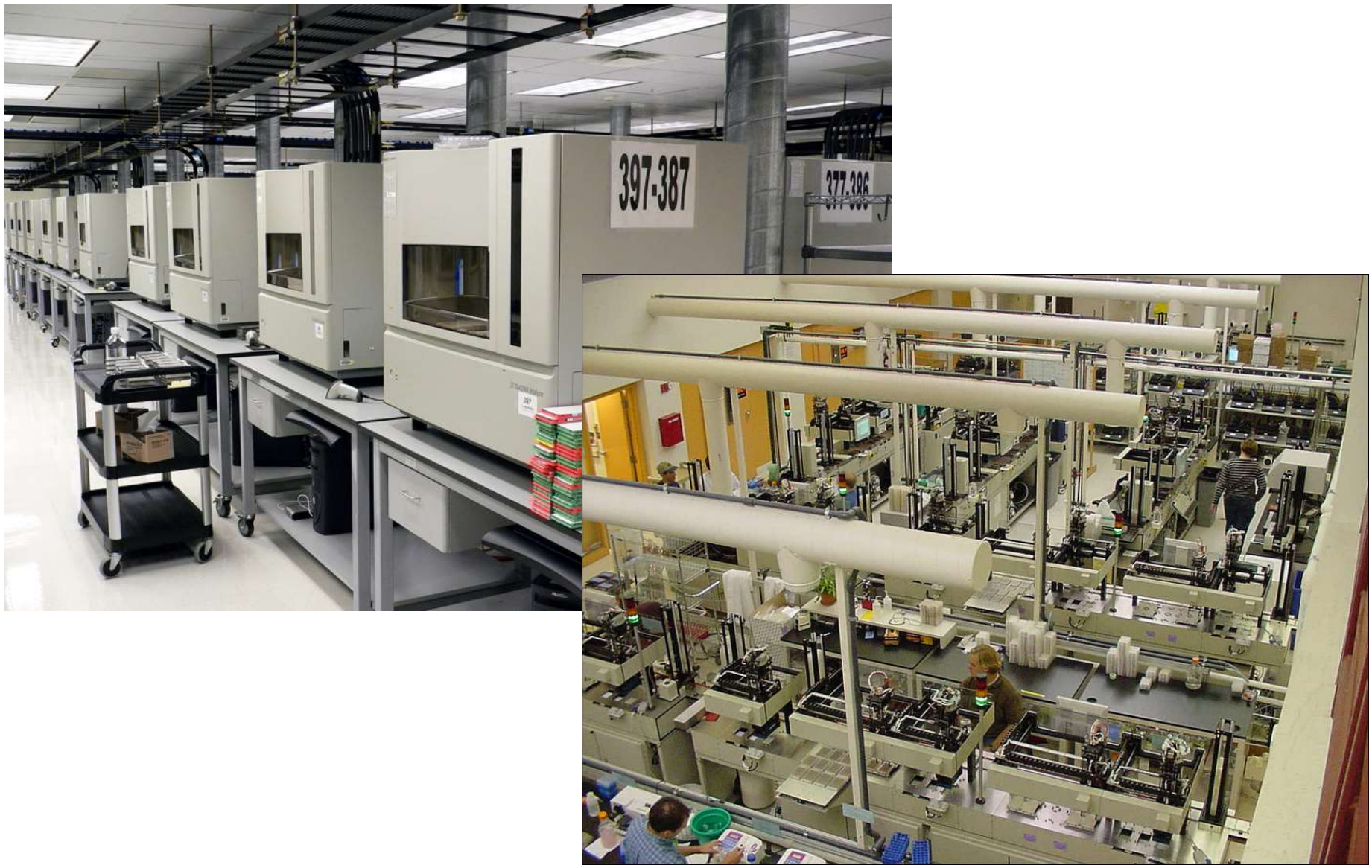
Figure 1. Shendure and Ji (2008) *Nature Biotechnology* 26: 1135-1145.


PMC22-T46 NNNNNNNAAAGCTATATACCTCACAGCTTTAAATGGAAGAAGAAAAAAAAGAGTGCAGAAAGGGGAGTGTAAAGGTTTACACTATACCTCTTGAACCTCTCCAGG
Trace data
PMC22-T46 TGTTATCTATCACATTGACTTTCAAGGGTGCACCAGAAAATTTGGCATAAGCCTGAAACAGAGTTTGCAATCAAACATTTTAGACCAAGTGCGTAATGCAAAGCT
Trace data
PMC22-T46 GCAGAGCAGCTTTACTTCACAGCCCCATCCCCCAAAGGCAGAAGGCGACTGATACTGAGGCCGCATCCCTCCCCGTGATCTCAAGTAGGGACGCGGGCGCCAGCC
Trace data
PMC22-T46 ACCCGGCTTTCGAGGACTTTGCTCCCCAAAAGGTCCTGGGGATGCAGAGTAGGGAGGGCGCGCCCCTTGCTTCGGGGTTTCACTCCACCGCCCAGACAGCTTTGT
Trace data
PMC22-T46 GAGGCGGACAAAACTCGGAGCCAAGCGGAGAAGCCTTGATGTCGCCGCCGGCCTTCGGCGGGGGTATCGGCCAGGTTGGGACTCGTCGGTGGGAGGGAGCCCAGG
Trace data
PMC22-T46 GCAATGCAGCGTGCAGCGGCTCGGCCCTCCTGGACCCCGCAGCGCAGCGGGAAACTGGGGCAAGTCAGCCTGGCGCCTCCAGCCCCGCCCCTTCCTCCTGCGCCC
Trace data
PMC22-T46 TGGCCCGCGAGGCGGCCTCACCATCACCACCAGGGACTCGCTGTGAACCGATGGGAGTCCCCAGCCGCCTCCCCAGCAACTGAGTTCCAAGGGGGCCGCCATCTT
Trace data
PMC22-T46 GCGCGGGCCGACCTTTACTATCCCGGGAAGTACTTTTCCTCCTAATTTCCGGCACGTGAAACGCGGGGGCGTGGCTAAGGGAGGGAAGGGTAATTTGGGGGCTCT
Trace data
PMC22-T46 TGAGAGATGAGAGAAACAACGACTGGAGGGATTAAGAAGAACTCTTAGAGCAGCTTGGCCTCGAGATCTGCGATCTAAGTAAGCTTGGCATTCCGGTACTGTGGT
Trace data
PMC22-T46 AAAGC
Trace data

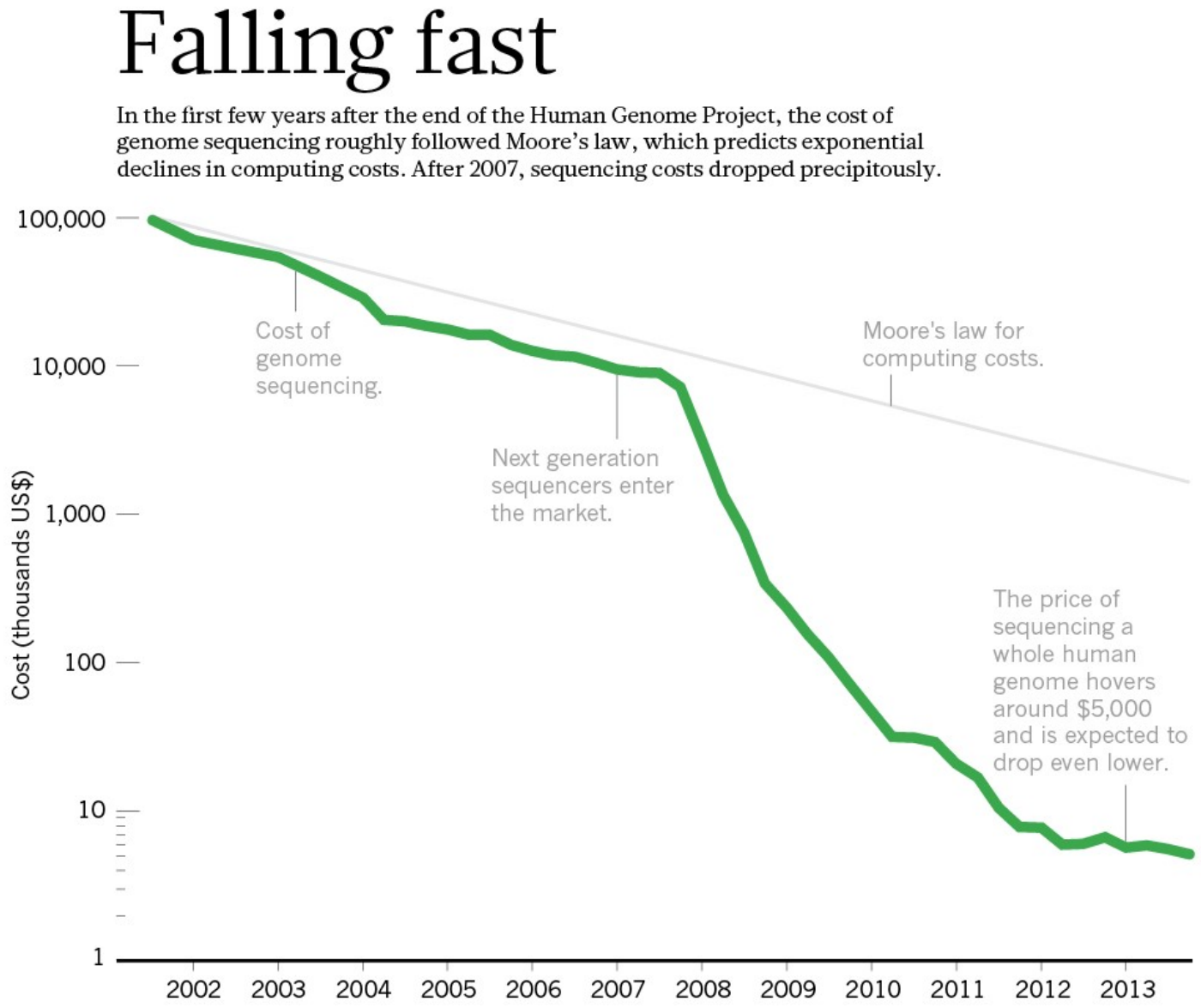
Limitations of classic sequencing techniques

- The main limitation of both these classical sequencing techniques is their low throughput, due to template preparation and in the case of Sanger Sequencing also to carrying out the enzymatic reaction.
- Each run in Sanger Sequencing can sequence up to 1000 bp, and with an automated sequencer 384 sequences can be run in parallel with a throughput of 80–100 kb per hour.
- Due its singleplex nature, Sanger Sequencing is not a hardly scalable process. In 1985, reading a single base cost $10, while in 2005, the various improvements reading 10,000 bases cost the same. However, large projects such as the Human Genome Project still required vast amounts of time and resources.
- Another limitation of First Generation Sequencing is that variants present at low frequency, such as mosaics, are difficult to detect due to high background levels.
- Finally, compared with modern technologies, the cost per base is still high.

Large sequencing facilities



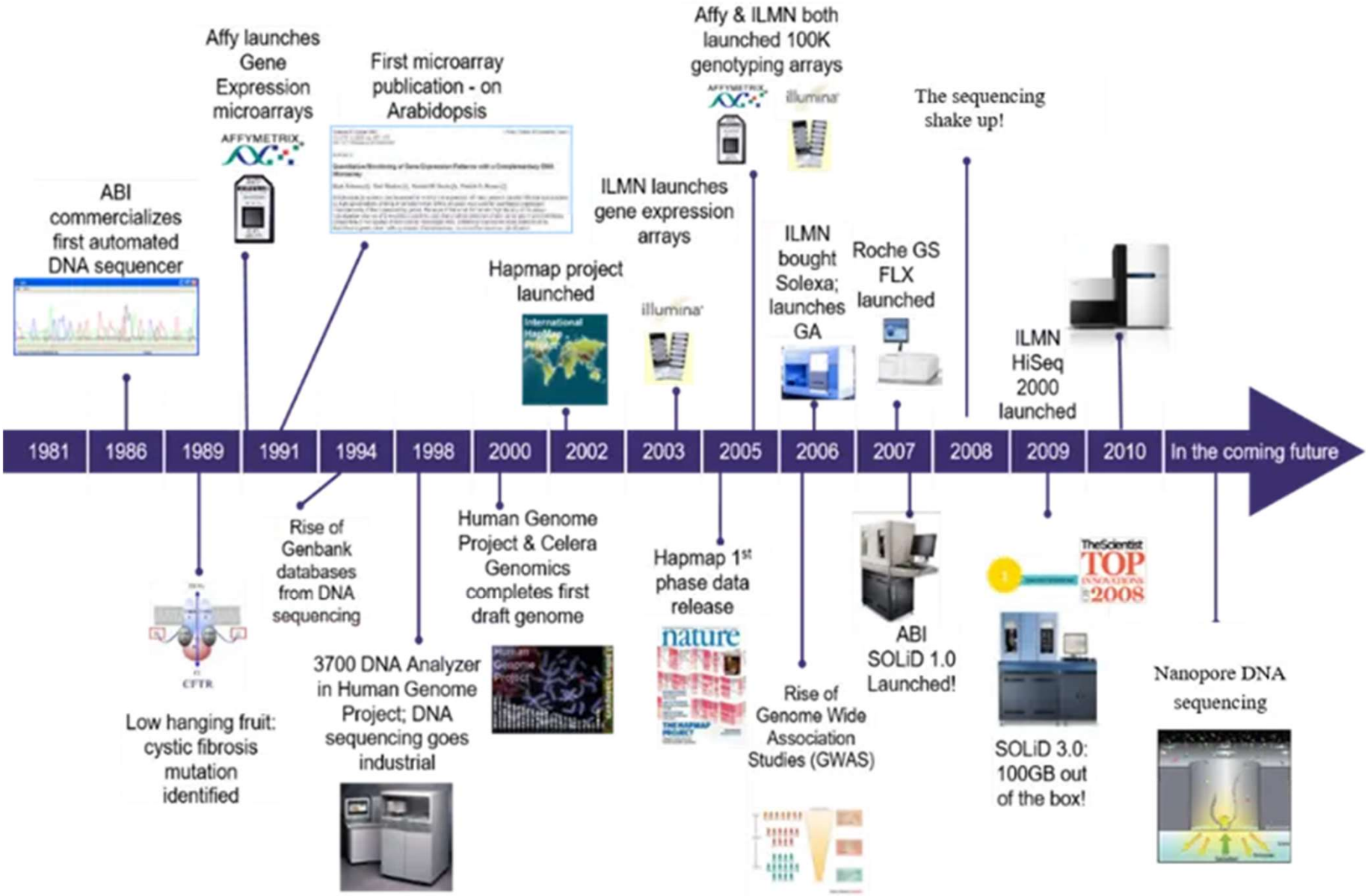
The quest for the $1,000 genome



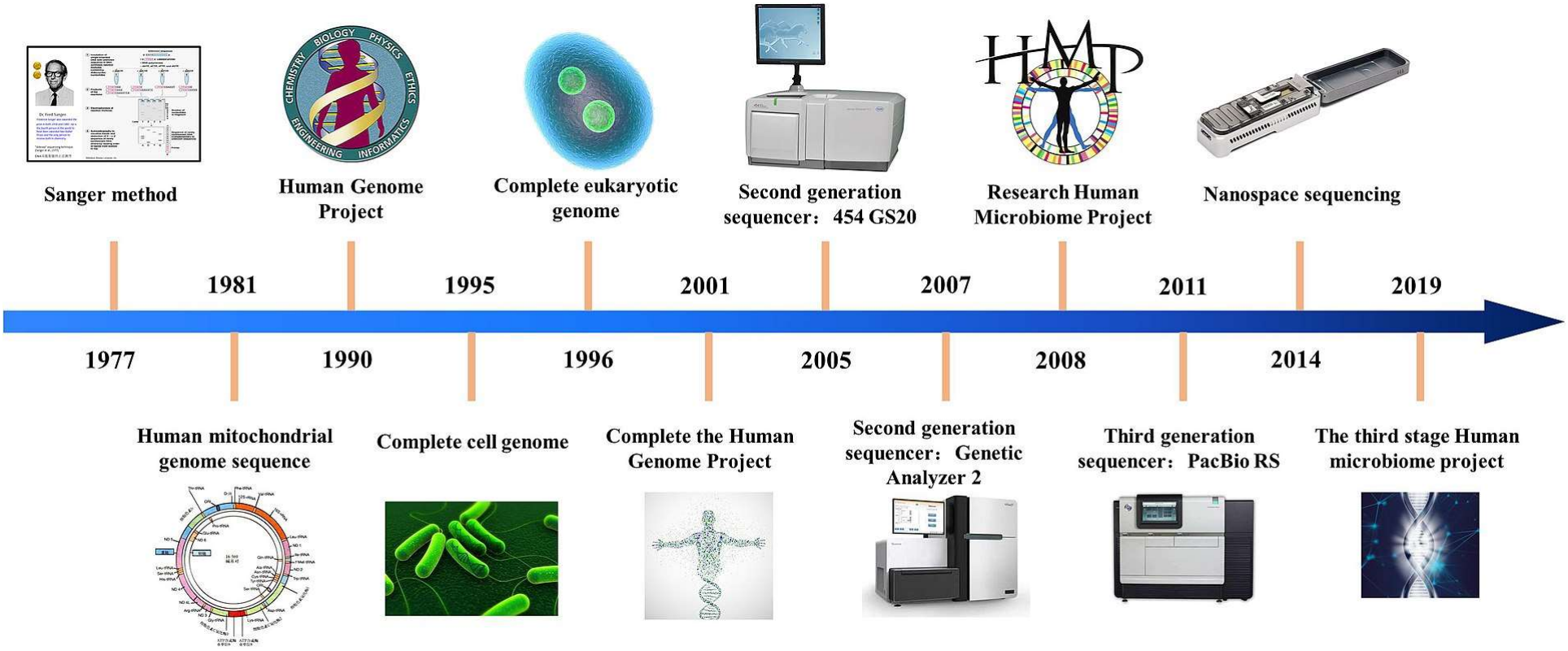
Next generation sequencing (NGS) methods

Next generation sequencing instruments can generate as much data in one day as several hundred Sanger DNA capillary sequencers!

Next generation sequencing systems



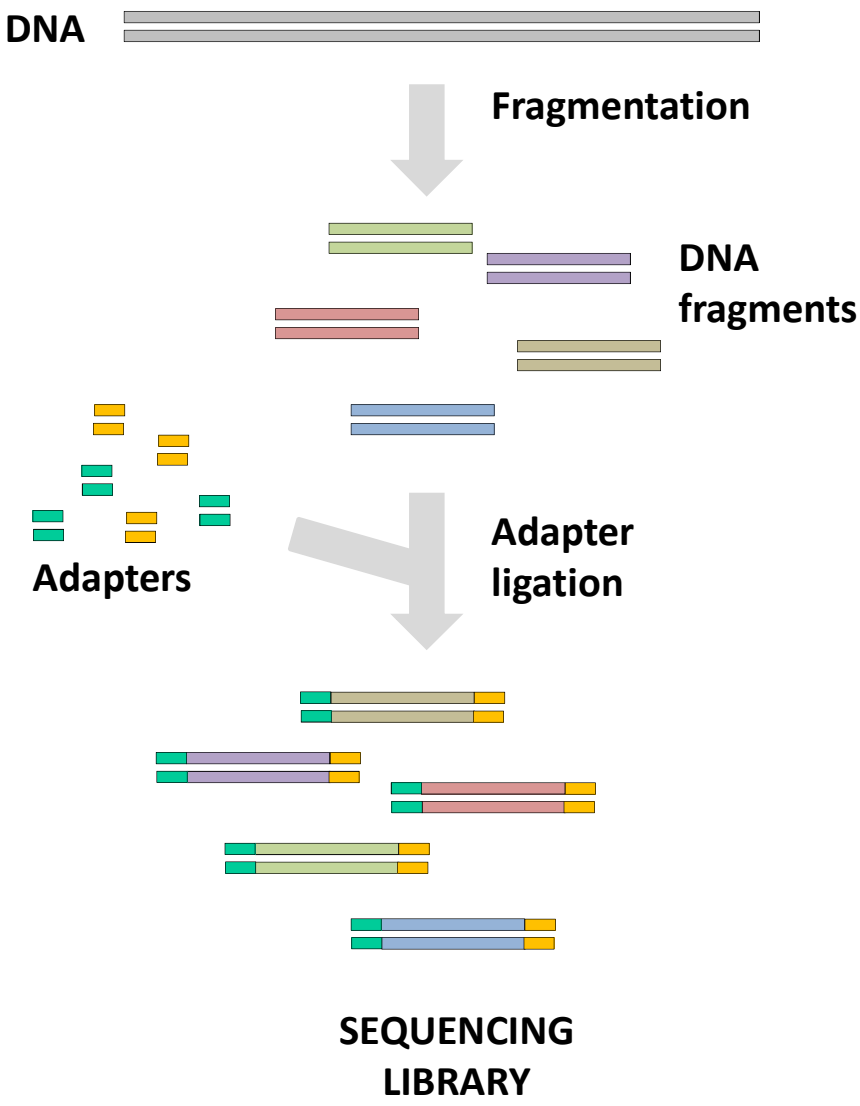
Next generation sequencing systems



Common characteristics of NGS methods

1. Cell-free preparation of sequencing library
2. Solid-phase amplification
3. Massively parallel sequencing reaction of each DNA fragment independently
4. Direct sequencing without need of electrophoresis

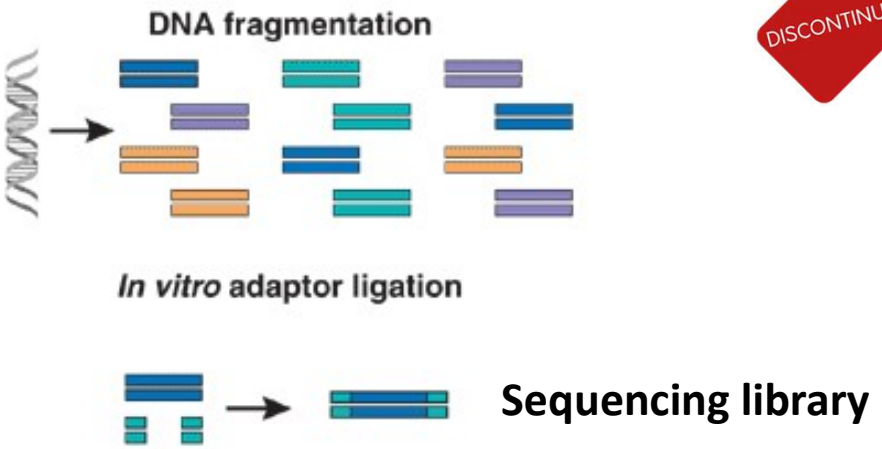
MASSIVELY PARALLEL SEQUENCING



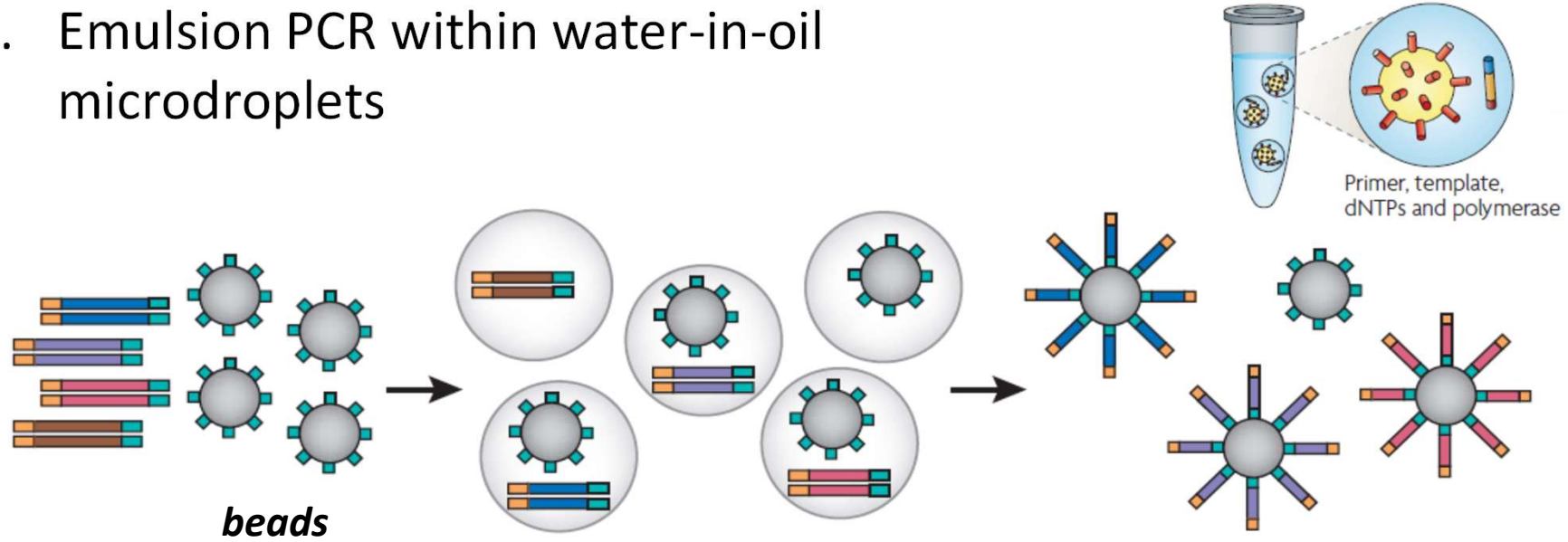
454/Roche – Pyrosequencing

DISCONTINUED

1. DNA fragmentation and adapter ligation



2. Emulsion PCR within water-in-oil microdroplets



Figures 1 and 2. Shendure and Ji (2008) *Nature Biotechnology* 26: 1135-1145.

454/Roche – Pirosequencing

DISCONTINUED

3. Distribution in individual wells

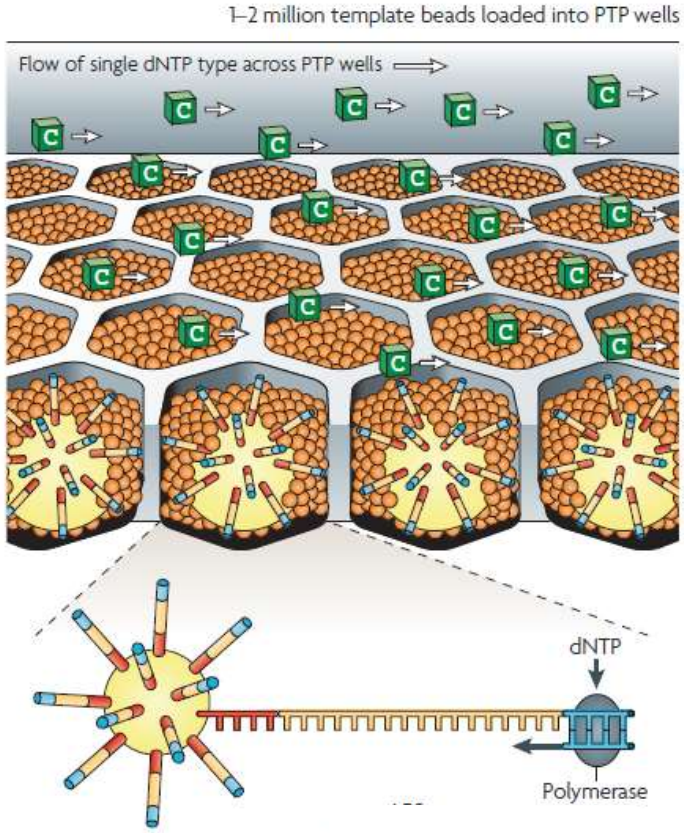
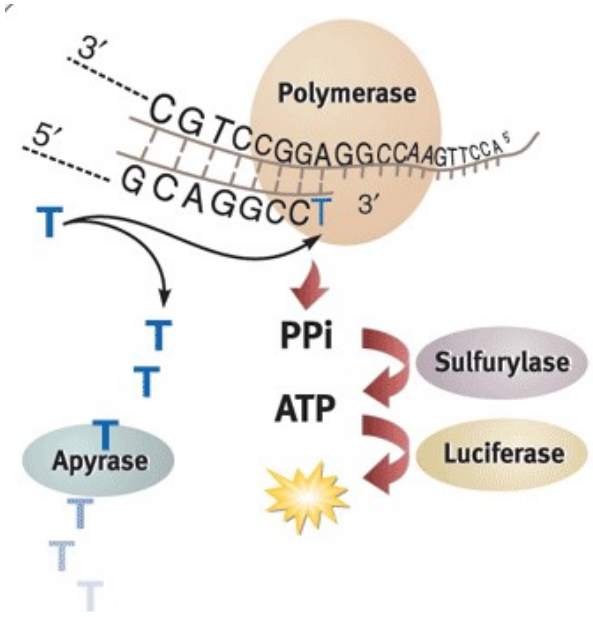


Figure 3. Metzker (2010) *Nature Reviews Genetics* 11: 31-46.

4. Pirosequencing



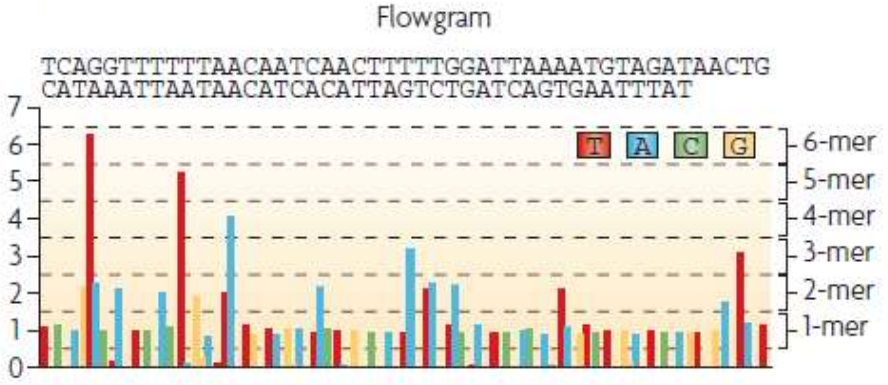
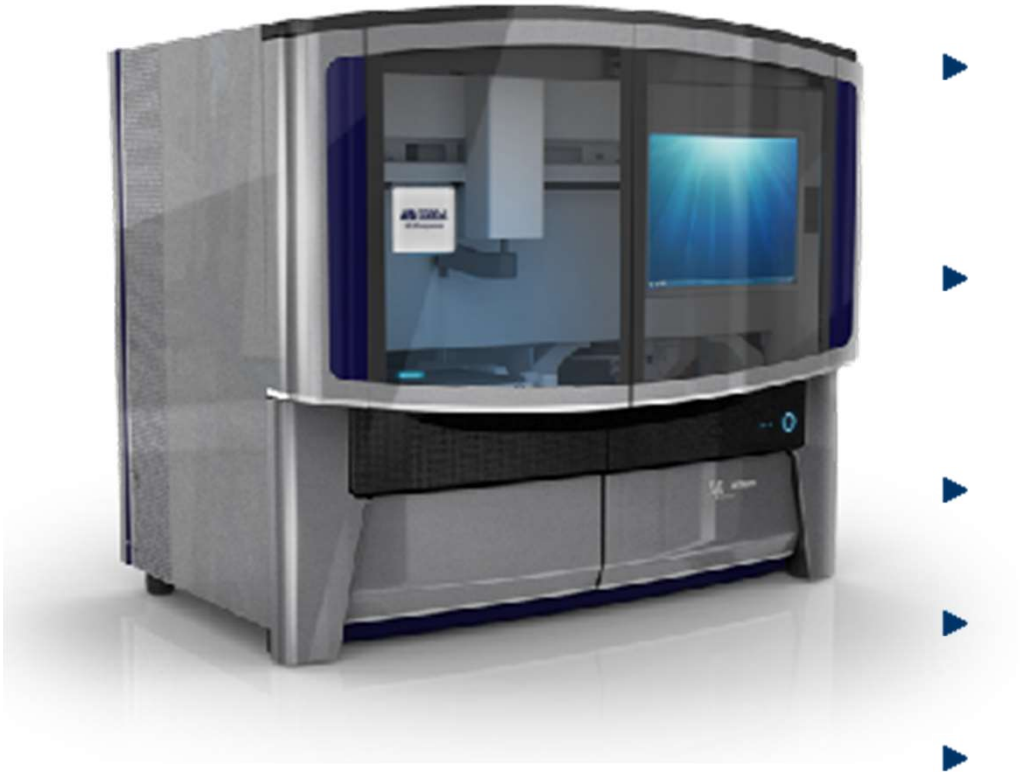


Figure 1. England and Pettersson (2005) *Nature Methods 2*: Application Note

SOLiD™ SOLiD (Sequencing by Oligonucleotide Ligation and Detection)



5500 SOLiD

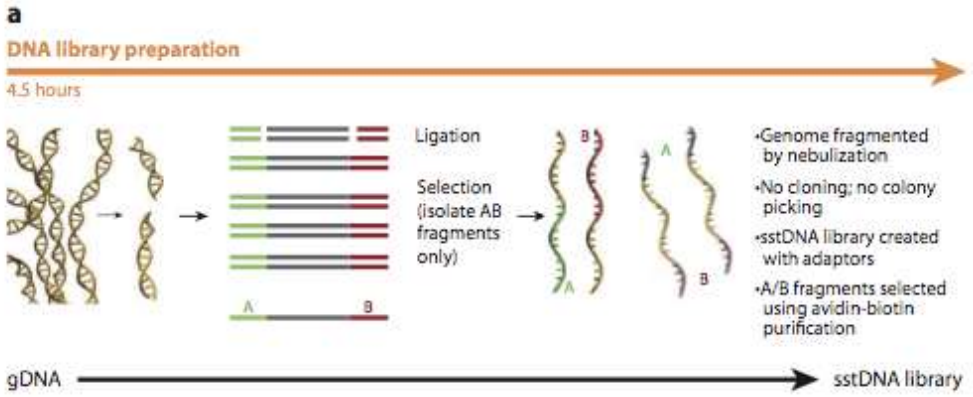
- Chemistry based on sequencing by ligation
- Sample amplified by emulsion PCR
- Read length 35-75 bp
- 100-500 million reads per run
- 90-300 Gb of sequence
- 1-7 days run

Applied Biosystems

http://www3.appliedbiosystems.com

SOLiD™ SOLiD (Sequencing by Oligonucleotide Ligation and Detection)

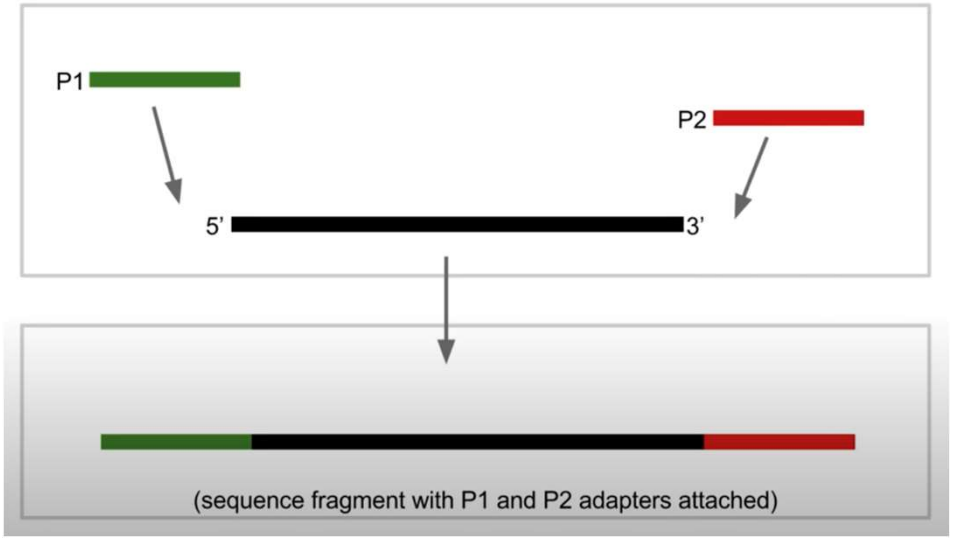
1. DNA fragmentation and adapters ligation



Fragmentation of gDNA

NEBULIZATION	SONICATION	DIGESTION
Compressed nitrogen is used to force DNA through a small hole, creating mechanically sheared fragments	Ultrasonic waves used to create gas bubbles in sample, and shear DNA by resonance vibration	Restriction enzymes used to cleave DNA, reaction kits with enzymes commercially available

Ligation of Adapter Sequences



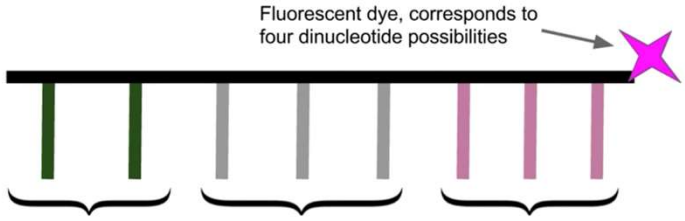
SOLiD™ (Sequencing by Oligonucleotide Ligation and Detection)

Ligation Chemistry Process (Overview)

1) Primer binds to template strands
2) Probe hybridization and ligation
3) Fluorescence measured
4) Dye-end (3) nucleotides cleaved
5) Steps 1-4 repeated 6+ times

Process completed 5 times, each time primer is offset by 1 base

Probe Anatomy

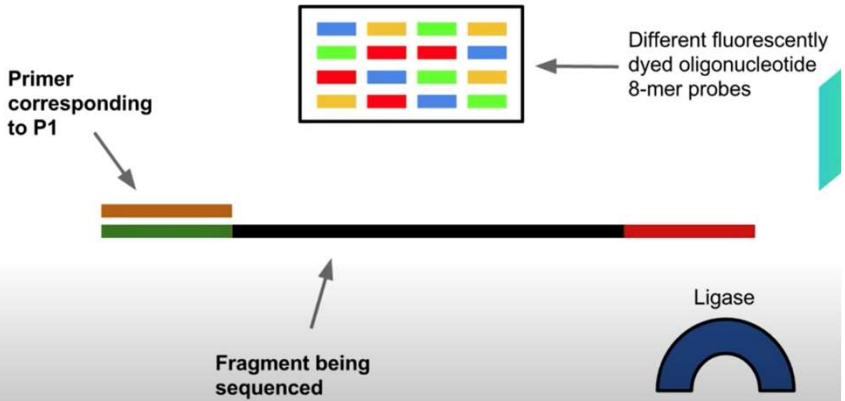


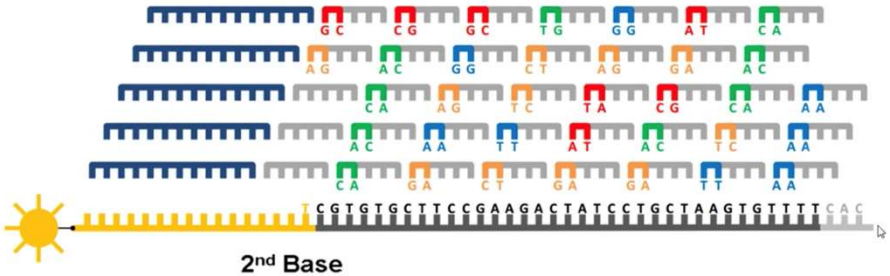
Two actual bases, each dinucleotide permutation corresponds to a dye color (red, green, blue, or yellow/orange) 16 possible dinucleotide permutations

Universal bases, bind to any of the 4 nucleotides

Universal bases with fluorescent dye, measure for fluorescence and cleaved in each cycle, so attached probe is only 5 nucleotides long

Primer Binds

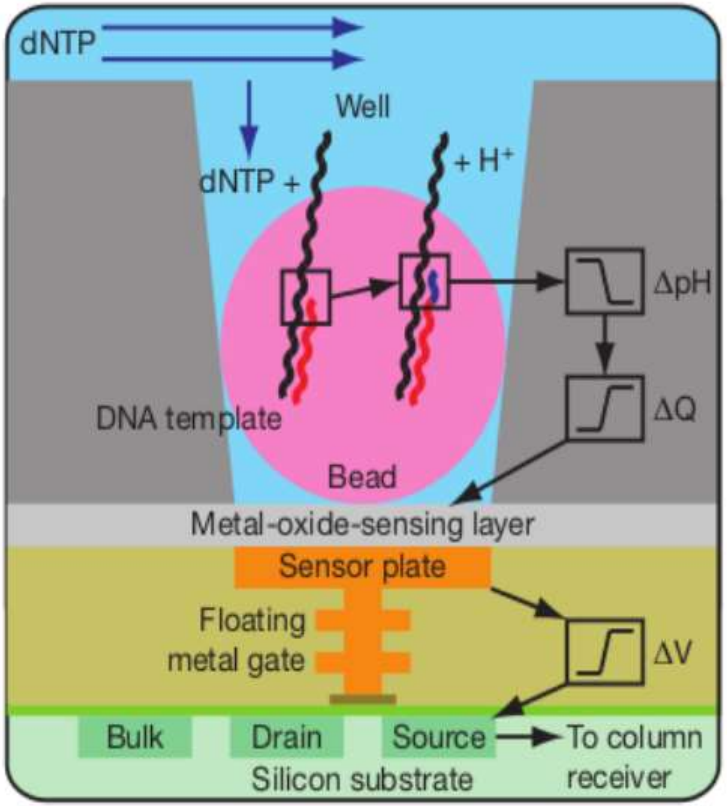




A **base** and a **color** define the next base in the sequence

Ion Torrent – Proton detection

- Emulsion PCR within water-in-oil microdroplets
- Real-time sequencing by using a semiconductor plate to count proton release during DNA synthesis



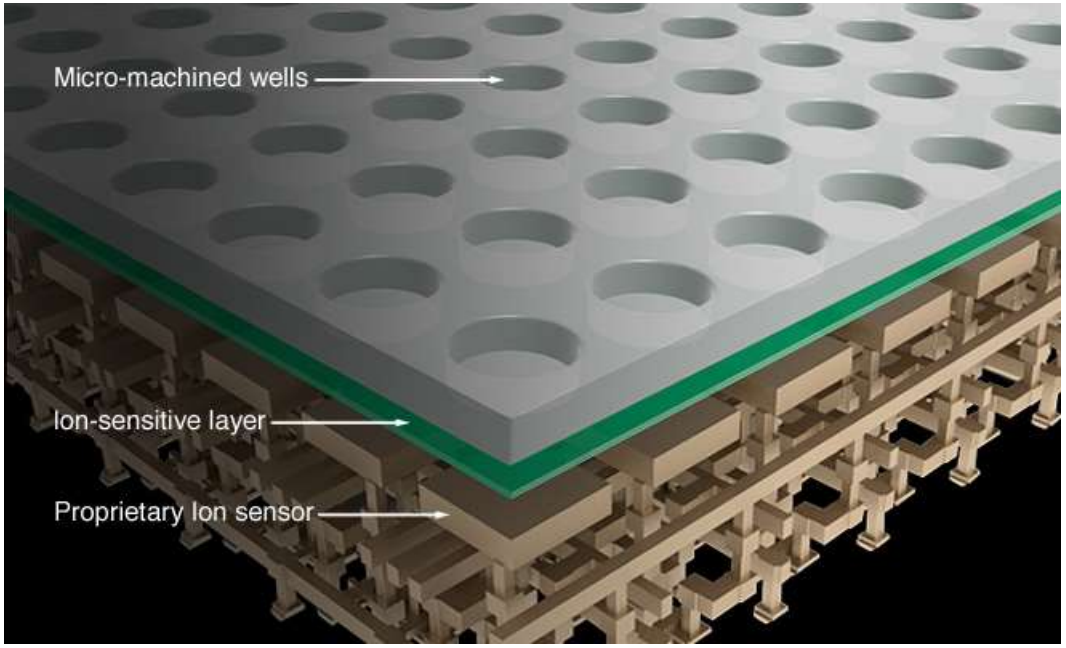
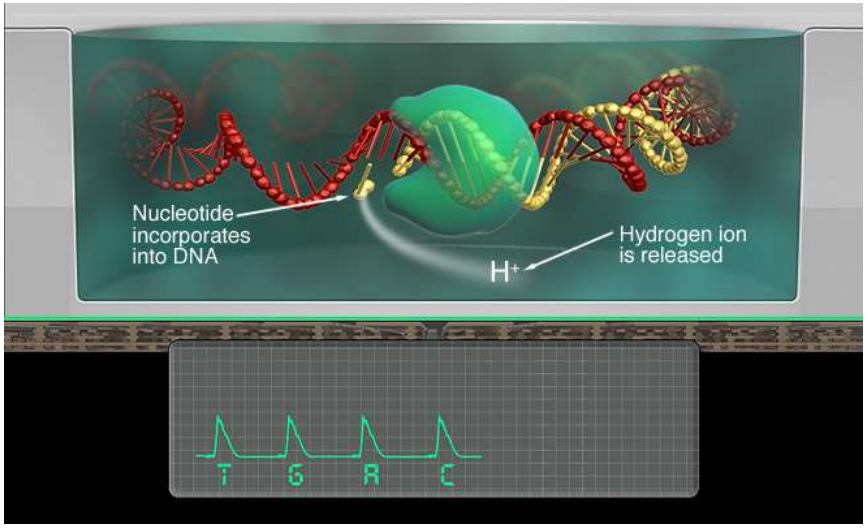
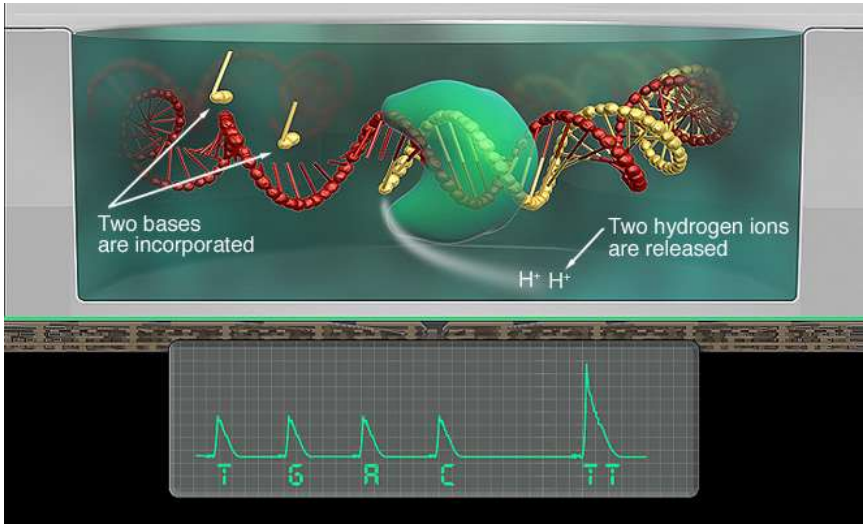


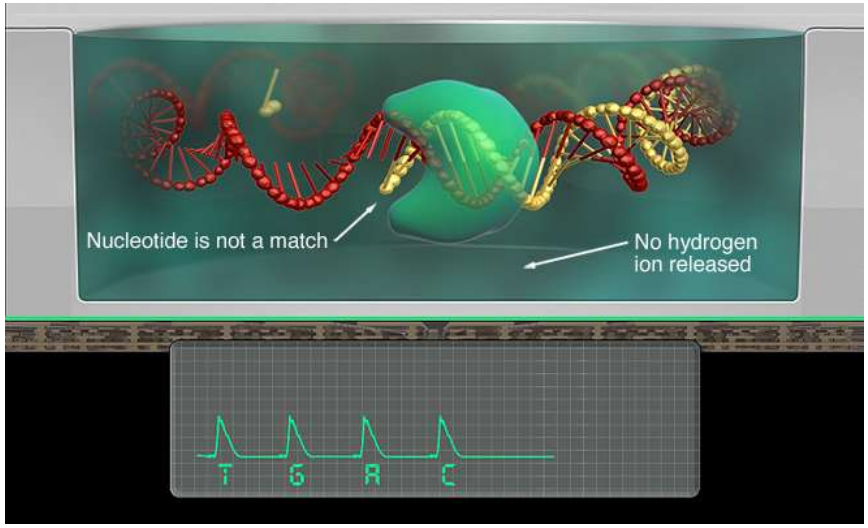
Figure 1. Rothberg *et al*. (2011) *Nature* 475:348-52.

http://www.iontorrent.com/technology-how-does-it-work-more/

Ion Torrent – Proton detection







- Normal nucleotides (not labelled) flow sequentially through the chip.
- The incorporation of one nucleotide released one proton (pH change). This is detected by the semi-conductor plate, which converts the chemical information into digital information.
- No optical machines are needed (no scanning, fluorescence, laser excitation, …)

Illumina – Reversible terminators

Illumina sequencing method

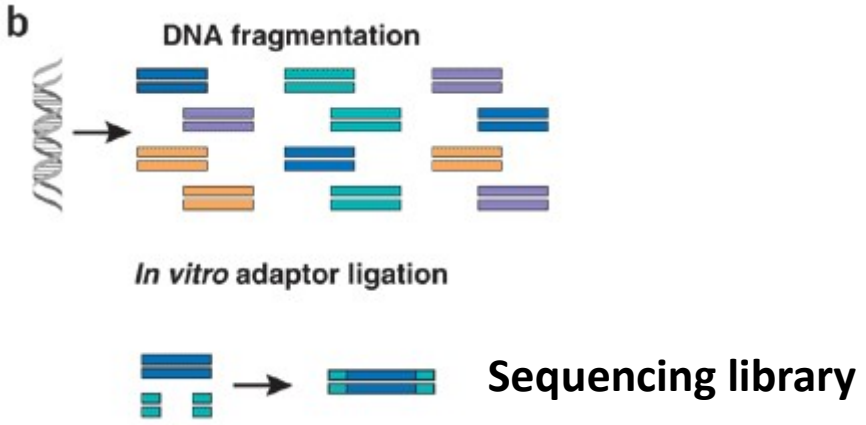
Illumina HiSeq

- Chemistry based on reversible terminators
- Sample amplified by solid-phase amplification
- Read length 36-100-150-300 bp
- 2-5 billion reads per run
- ~200-1500 Gb of sequence
- 2.5-6 days run

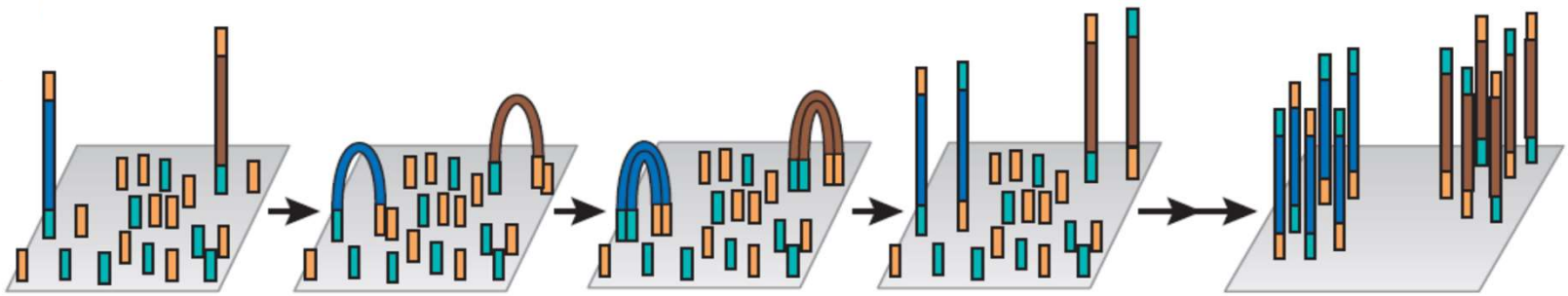
http://www.illumina.com/

Illumina – Reversible terminator

1. DNA fragmentation and adapter ligation

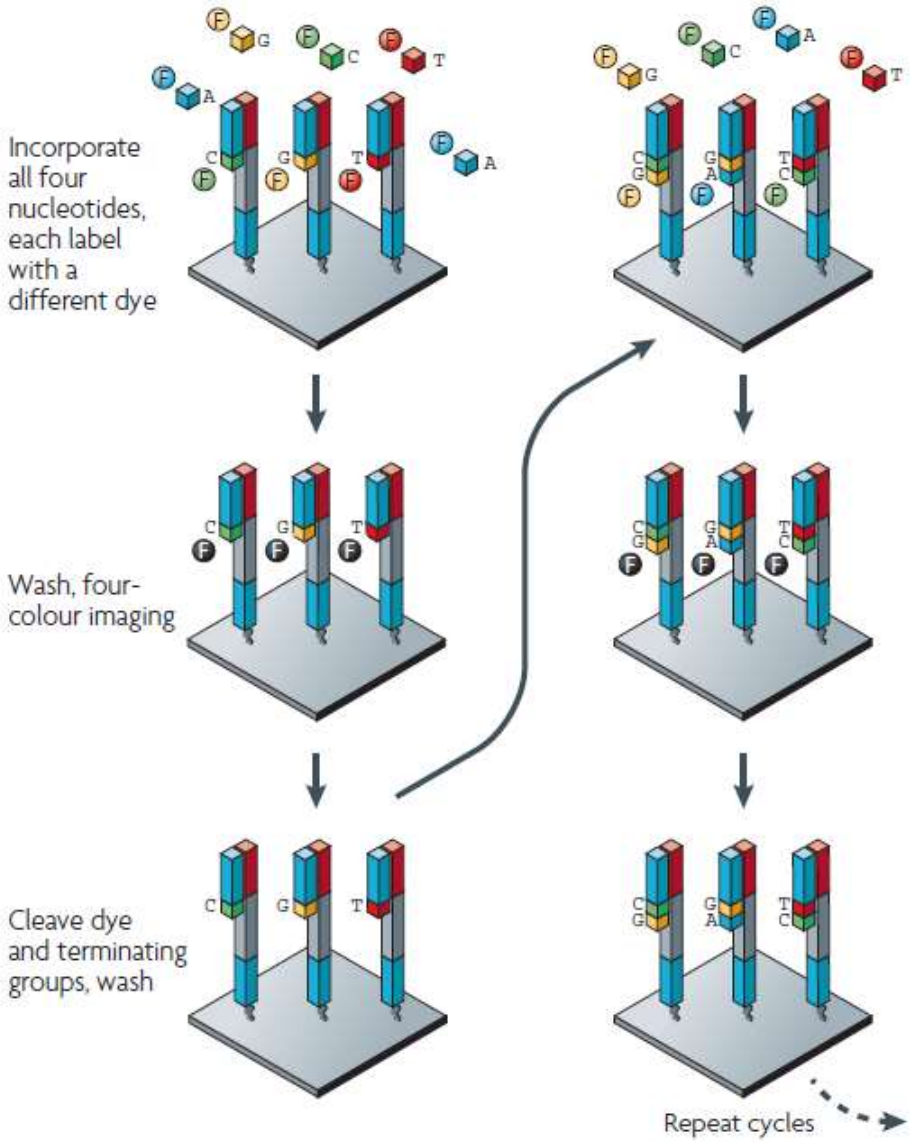


2. Solid-phase amplification and cluster generation by bridge PCR



Figures 1 and 2. Shendure and Ji (2008) *Nature Biotechnology* 26: 1135-1145.

Illumina – Reversible terminator



3. Flowing of fluorescent reversible terminator dNTPs and incorporation of a single base per cycle

4. Read identity of each base of a cluster from sequential images

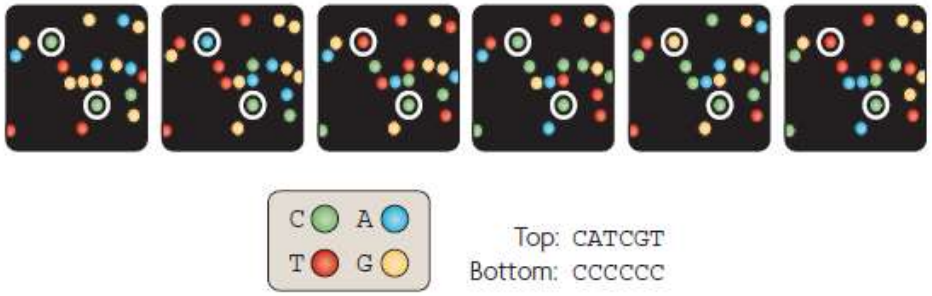


Figure 2. Metzker (2010) *Nature Reviews Genetics* 11: 31-46.

Illumina – Reversible terminator

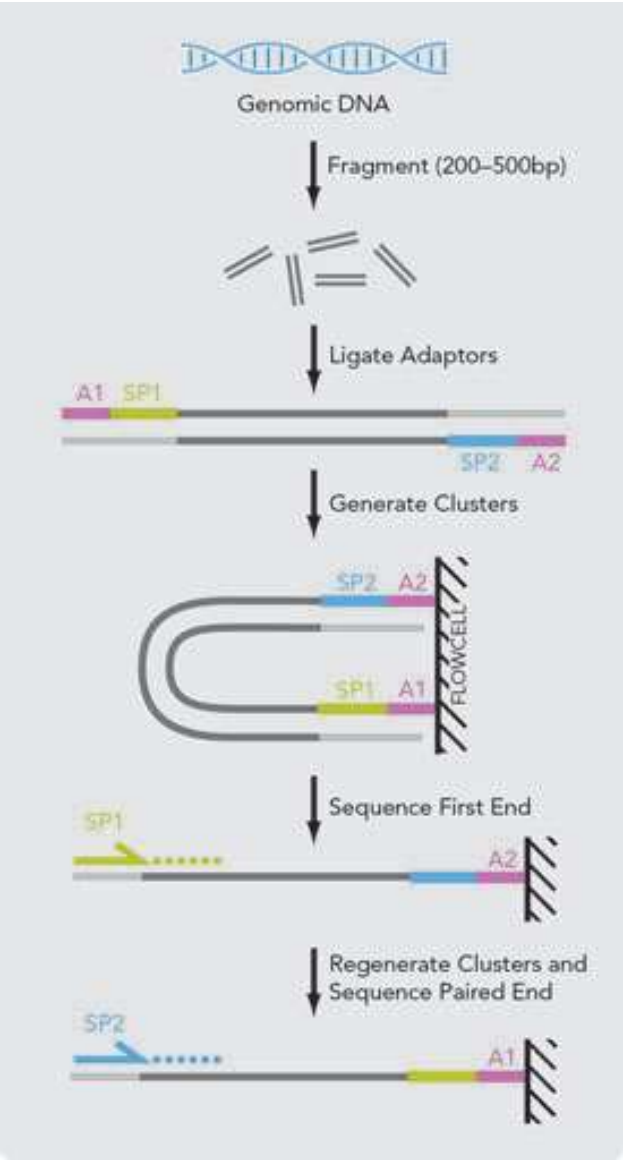
Illumina flow cell

- One (1) flow cell
- Eight (8) channels/lanes (1.4-mm wide)
- Input requirement: 0.1–1.0 µg (single- and paired-end reads), 10 µg (Mate Pair reads)
- 96-120 million reads (clusters) per flow cell, each containing ~1,000 copies of the same template
- Each lane can run 12-96 differently tagged libraries (Multiplexed Sequencing)

Illumina – Reversible terminator

Paired-end sequencing

- After completion of the first read, the templates can be regenerated *in situ* to enable a second >50 bp read from the opposite end of the fragments
- Double amount of sequence can be generated from the same amount of DNA
- Up to 2 x 150 bp
- Longer run times
- Very important for some applications (transcript analysis, structural variants,…)



Sequencing power for every scale

	 MiniSeq System	 MiSeq Series	 NextSeq Series	 HiSeq Series	 HiSeq X Series*
Key Methods	Amplicon, targeted RNA, small RNA, and targeted gene panel sequencing.	Small genome, amplicon, and targeted gene panel sequencing.	Everyday exome, transcriptome, and targeted resequencing.	Production-scale genome, exome, transcriptome sequencing, and more.	Population- and production-scale whole-genome sequencing.
Maximum Output	7.5 Gb	15 Gb	120 Gb	1500 Gb	1800 Gb
Maximum Reads per Run	25 million	25 million†	400 million	5 billion	6 billion
Maximum Read Length	2 × 150 bp	2 × 300 bp	2 × 150 bp	2 × 150 bp	2 × 150 bp
Run Time	4–24 hours	4–55 hours	12–30 hours	<1–3.5 days (HiSeq 3000/HiSeq 4000) 7 hours–6 days (HiSeq 2500)	<3 days
Benchtop Sequencer	Yes	Yes	Yes	No	No
System Versions	• MiniSeq System for low-throughput targeted DNA and RNA sequencing	• MiSeq System for targeted and small genome sequencing • MiSeq FGx System for forensic genomics • MiSeqDx System for molecular diagnostics	• NextSeq 500 System for everyday genomics • NextSeq 550 System for both sequencing and cytogenomic arrays	• HiSeq 3000/HiSeq 4000 Systems for production-scale genomics • HiSeq 2500 Systems for large-scale genomics	• HiSeq X Five System for production-scale whole-genome sequencing • HiSeq X Ten System for population-scale whole-genome sequencing

* First instrument to generate a 30 x genome for $1000

- Denser clustering
- Faster camera
- Faster enzymes

Comparison of different sequencing methods

Table 1. Comparison of high-throughput sequencing technologies available

	Throughput	Length	Quality	Costs	Applications	Main sources of errors
Sanger	6 Mb/day	800 nt	10^{-4}–10^{-5}	~500$/Mb	Small sample sizes, genomes/ scaffolds, InDels/SNPs, long haplotypes, low complexity regions, etc.	Polymerase/amplification, low intensities/missing termination variants, contaminant sequences
454/Roche	750 Mb/day	400 nt	10^{-3}–10^{-4}	~20$/Mb	Complex genomes, SNPs, structural variation, indexed samples, small RNA^{+}, mRNAs^{+}, etc.	Amplification, mixed beads, intensity thresholding, homopolymers, phasing, neighbor interference
Illumina	5,000 Mb/day	100 nt	10^{-2}–10^{-3}	~0.50$/Mb	Complex genomes, counting (SAGE, CNV ChIP, small RNA), mRNAs, InDels/homopolymers, structural variation, bisulfite data, indexing, SNPs^{+}, etc.	Amplification, mixed clusters/neighbor interference, phasing, base labeling
SOLiD	5,000 Mb/day	50 nt	10^{-2}–10^{-3}	~0.50$/Mb	Complex small genomes, counting (SAGE, ChIP, small RNA, CNV), SNPs, mRNAs, structural variation, indexing, etc.	Amplification, mixed beads, phasing, signal decline, neighbor interference
Helicos	5,000 Mb/day	32 nt	10^{-2}	<0.50$/Mb	Non-amplifiable samples, counting (SAGE, ChIP, small RNA), etc.	Polymerase, low intensities/thresholding, molecule loss/termination

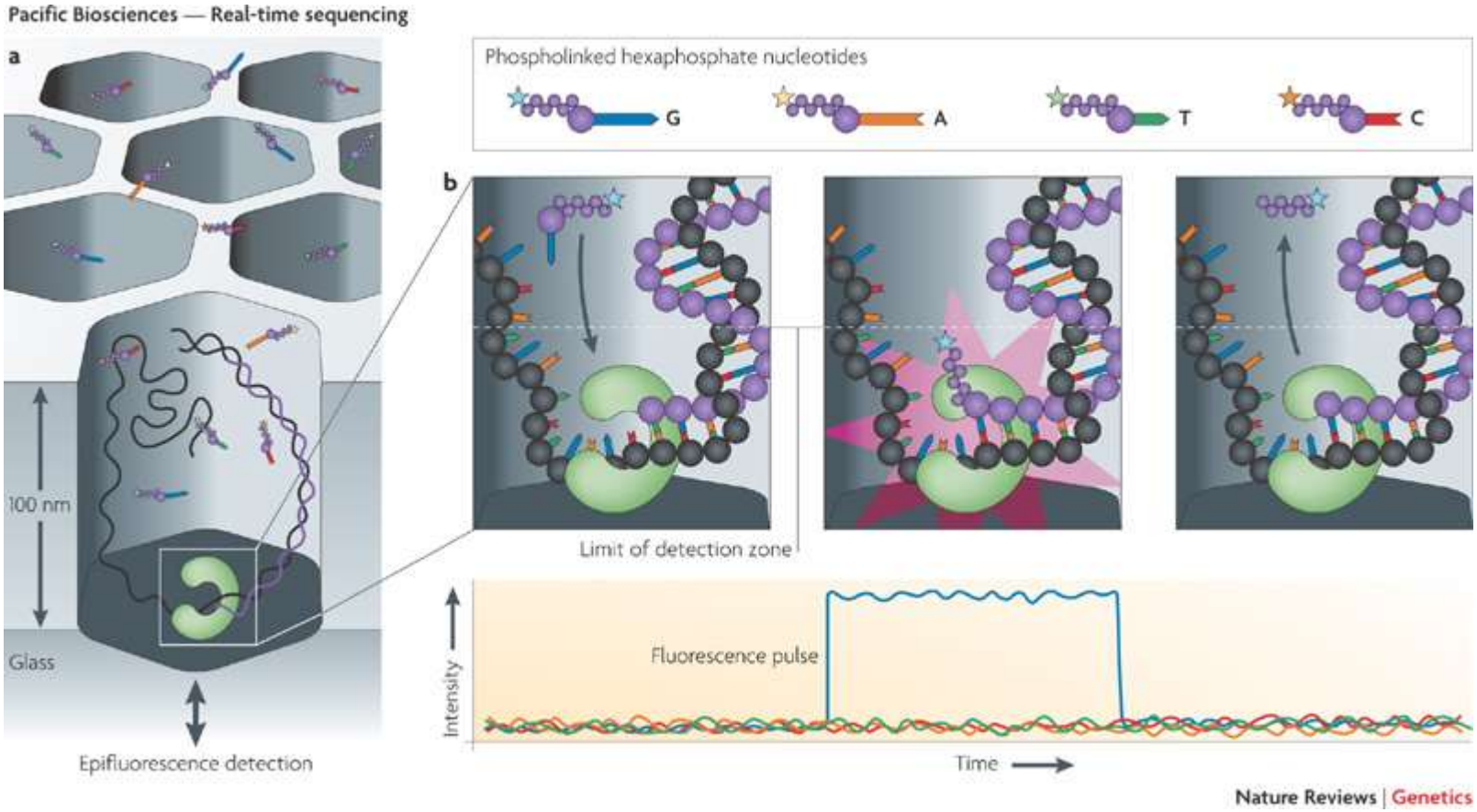
Figure 1. Kircher and Kelso (2010) *Bioessays* 32: 524-536.

Challenges of NGS methods

- Increase read length
- Improve sequence accuracy
- Single-molecule sequencing (no amplification)
- De-novo assembly of complex genomes
- Sequencing of complex regions

Pacific Biosciences – Real-time sequencing

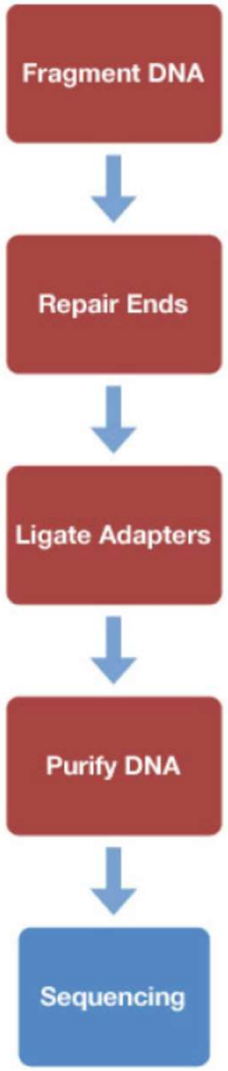
Pacific Biosciences – Real-time sequencing

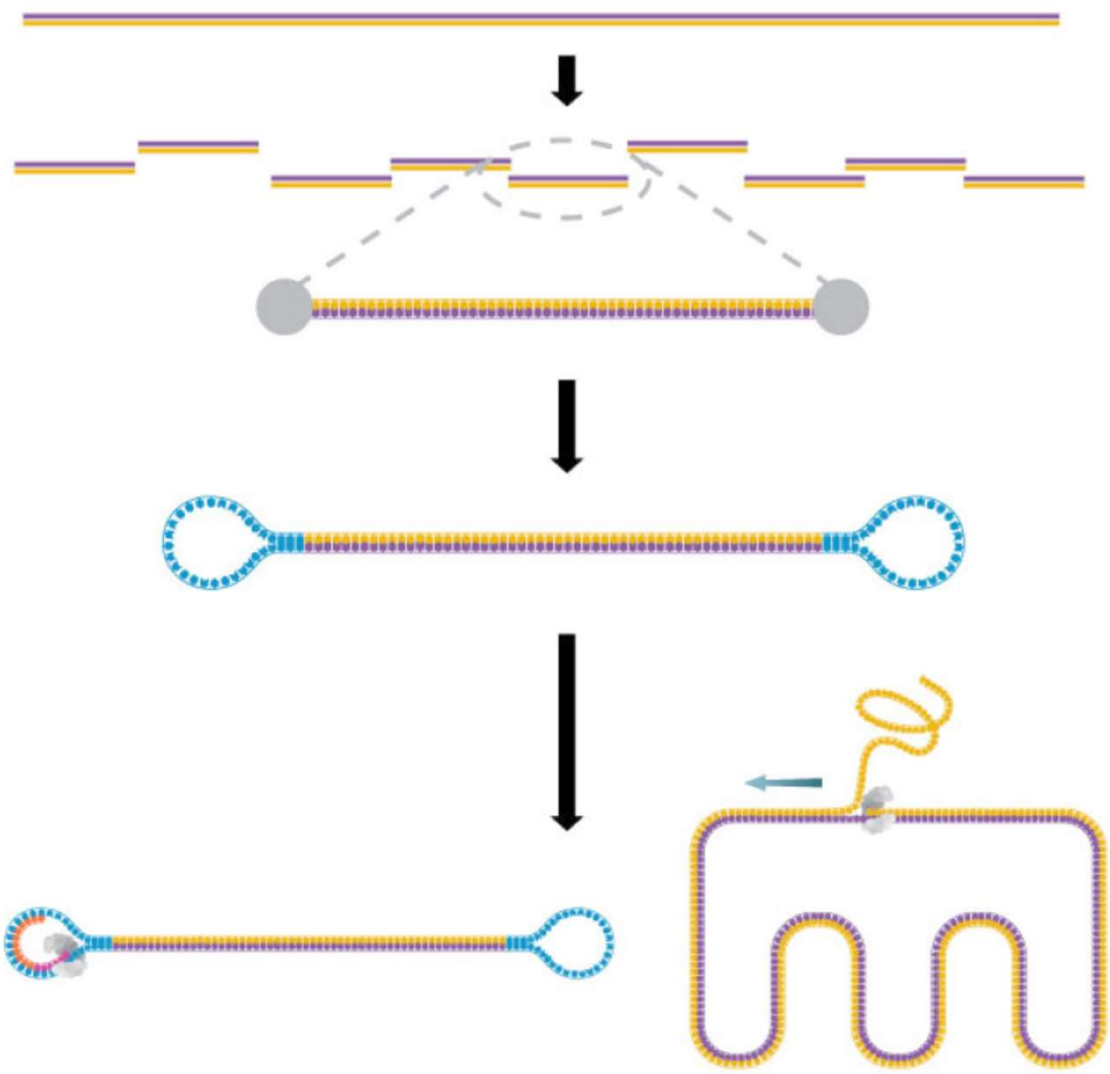


- **Real-time** monitoring of **single-molecule** sequencing as it occurs with an immobilized DNA polymerase (*Single Molecule Real Time* –SMRT-technology)
- Read length avg. ~**10-15 kb**, up to ~**50 kb**, but error rate is ~**15%**

Figure 4. Metzker (2010) *Nature Reviews Genetics* 11: 31-46.

PacBio workflow





Sample genome

DNA fragments

SMRT bell library

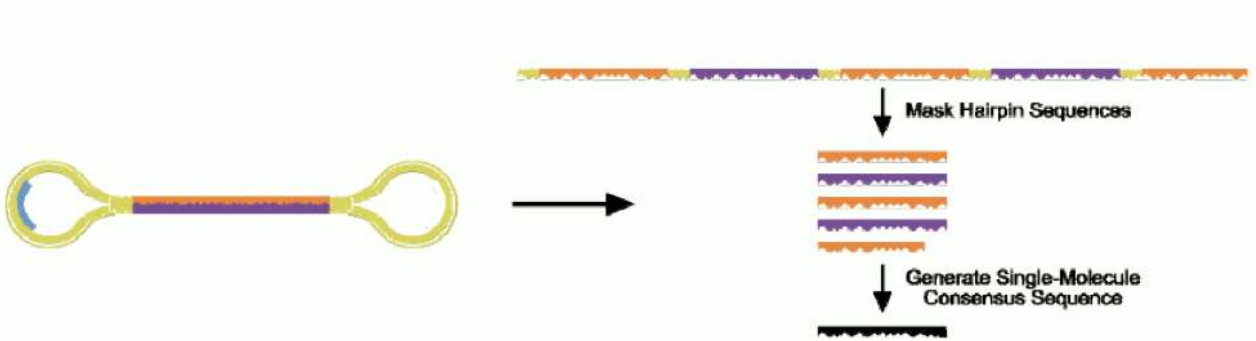
Strand displacement and DNA synthesis
(Fwd+Rev)

PacBio workflow

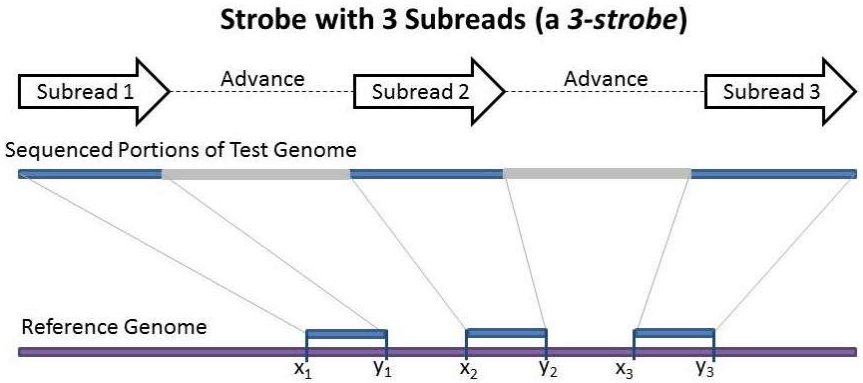
Multiple sequencing protocols are possible:

- Standard sequencing (long read lengths)

- Circular consensus sequencing (high accuracy)

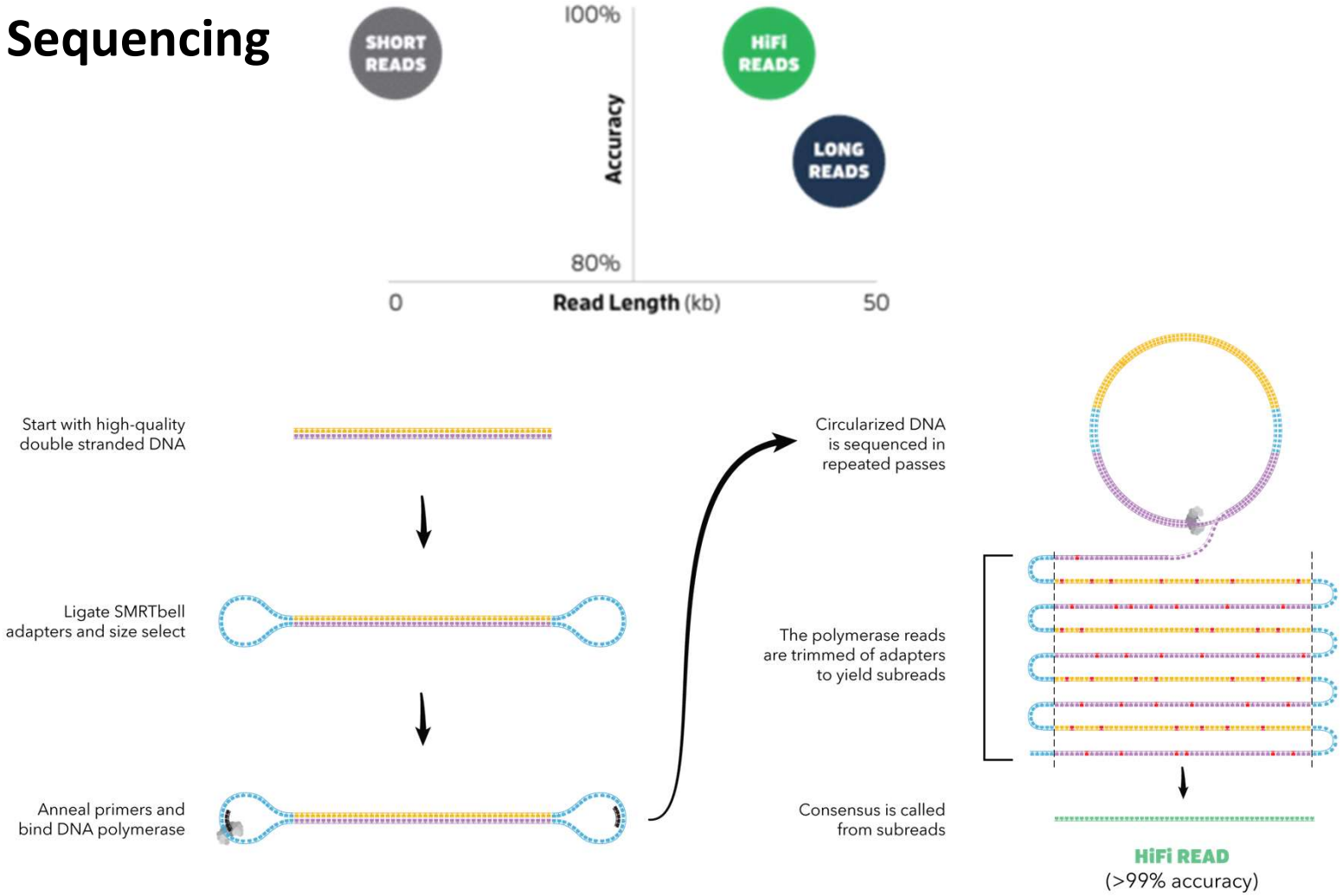


- Strobe sequencing (structural variation detection)



Pacific Biosciences – Real-time sequencing

HiFi Sequencing

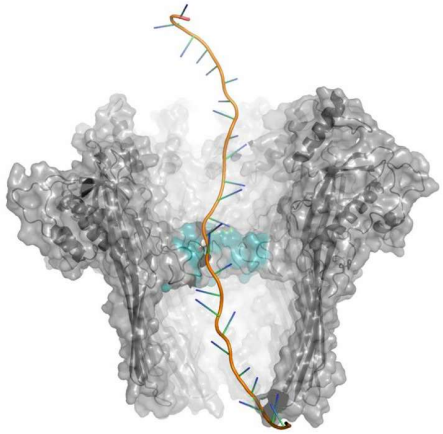


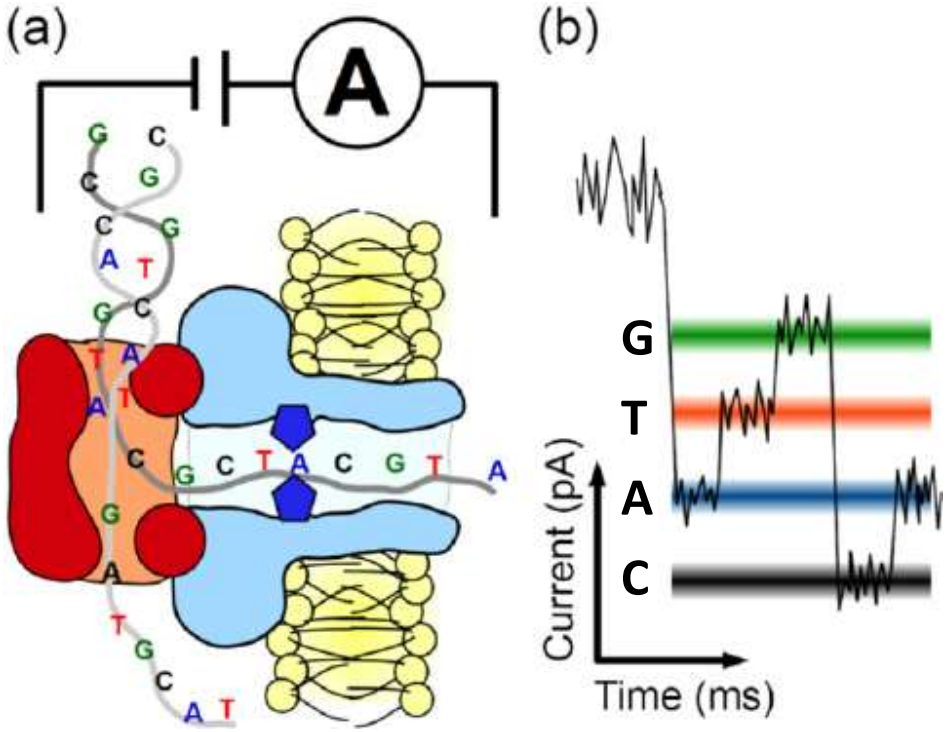
HiFi reads are produced by calling consensus from subreads generated by multiple passes of the enzyme around a circularized template. This results in a HiFi read that is both long and accurate.

Nature Biotechnology volume 37, pages1155–1162(2019)

Oxford Nanopore technology

Oxford Nanopore technology



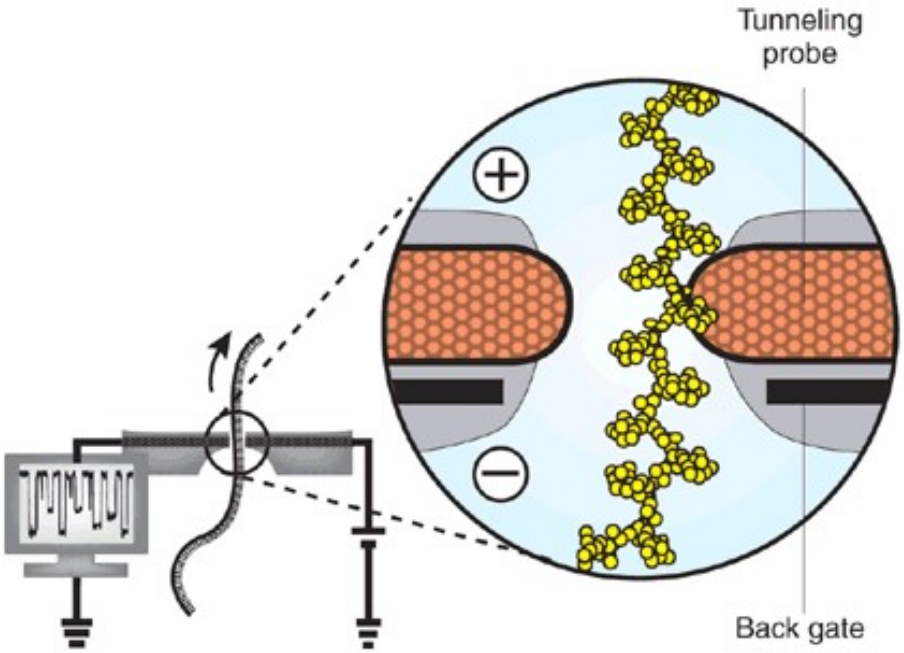


Base identification through differences in conductance of nucleic acid molecules driven through a nanopore

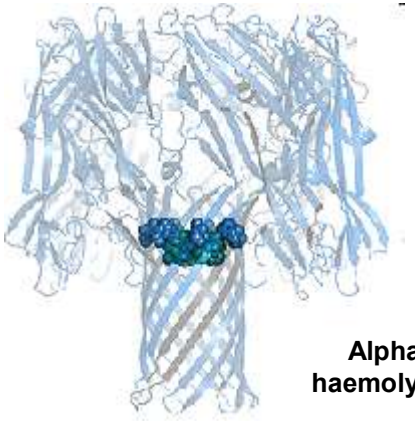
- Single-molecule sequencing (no amplification needed)
- Very long read lengths
- High throughput
- Electrical base detection (no optics)

MinION Mobile Sequencing

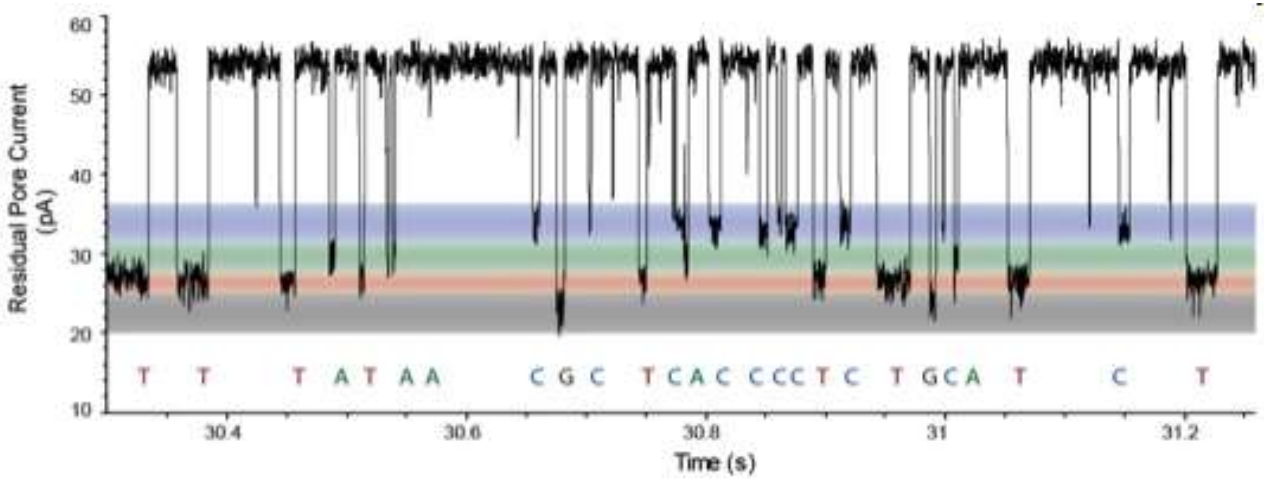
Oxford Nanopore technology



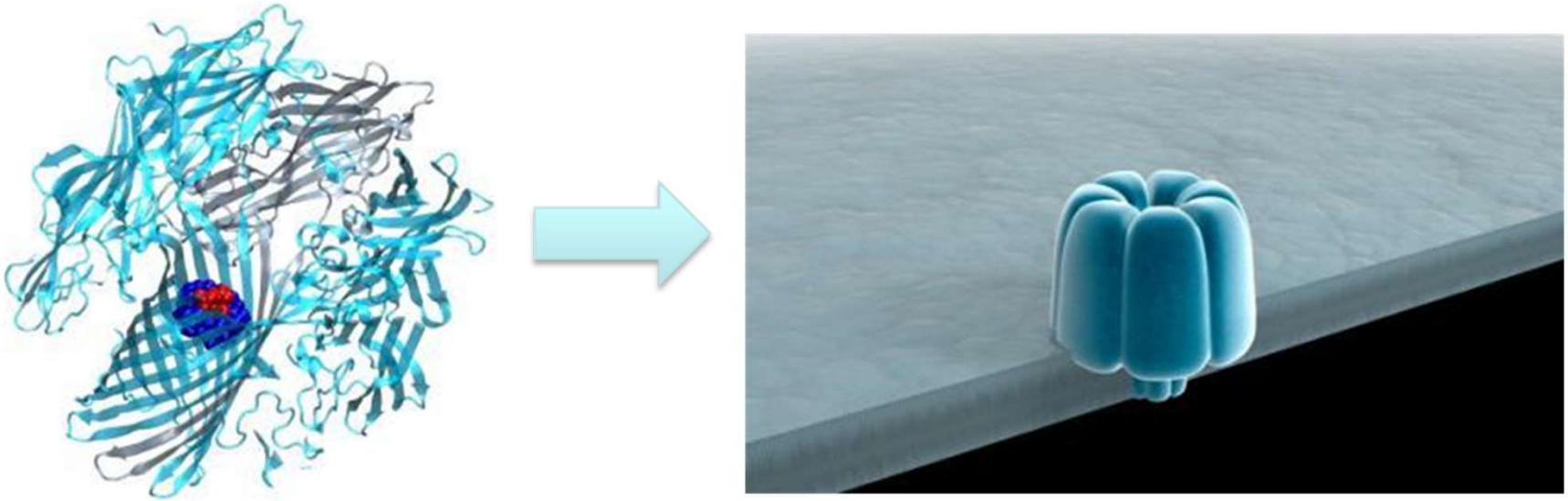
Base identification through differences in conductance of nucleic acid molecules driven through a nanopore



Alpha haemolysin



Engineering nature's nanopores



Oxford Nanopore technology

Nanopore Accuracy

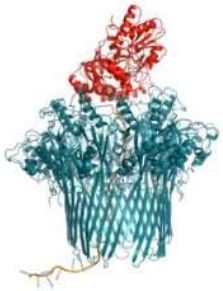
Q20+ raw read

Sequencing chemistry tuned with latest base callers

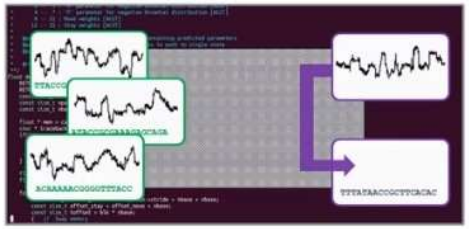
- New "Q20+", kit 12 release upgrades enzyme motor
- Refined motor "E8.1" – better movement quality, 250 bps
- Combine with latest base callers for high accuracy
- Works with standard R9.4.1 flowcells and R10.4
- Hitting > Q20 raw read, single pass accuracy (Simplex)

9 Kit 10 = R9.4.1 nanopore E8.0 motor

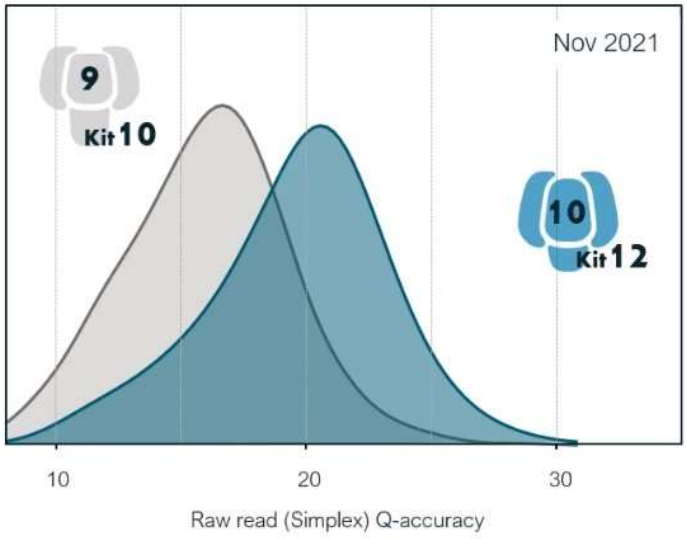
10 Kit 12 = R10.4 nanopore E8.1 motor



+



=



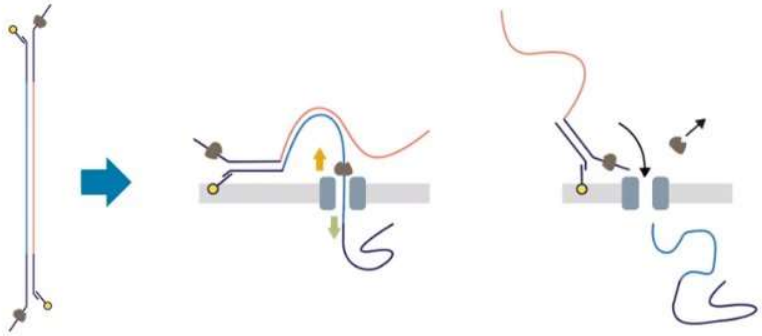
R9.4.1 and R10.4 chemistries > Q20 Simplex

Reading Both Strands

Duplex

Combining data from both strands

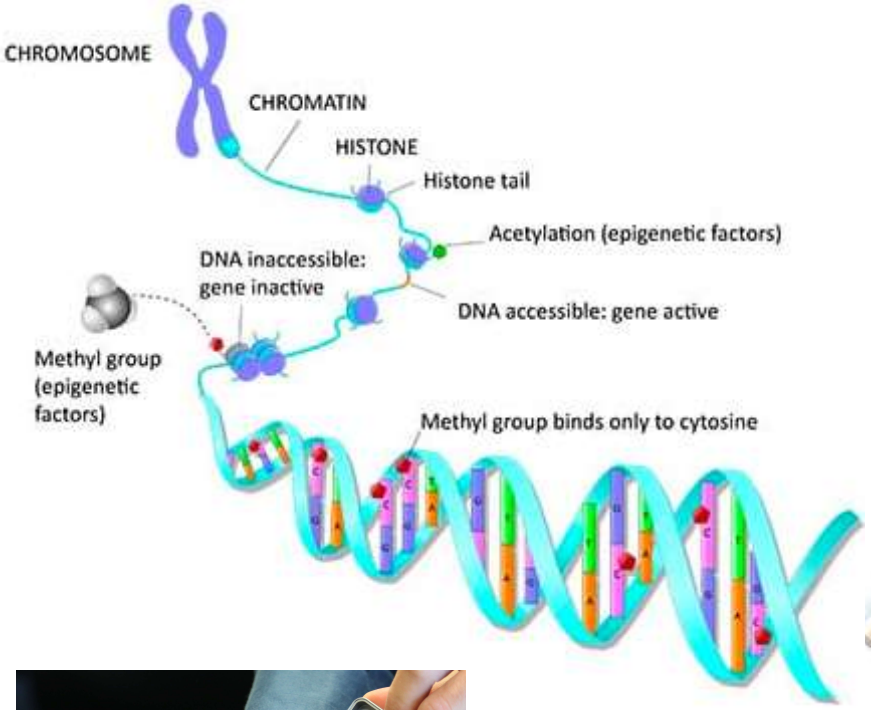
- Reading both strands has always been a strong feature
  - Two looks at the same sequence pairs
  - Different error profiles on reverse complement
  - Can help resolve modifications
- Previous versions of the chemistry had low natural "follow on"
- Typical rates of ~ 5% of the data collected was from duplex pairs

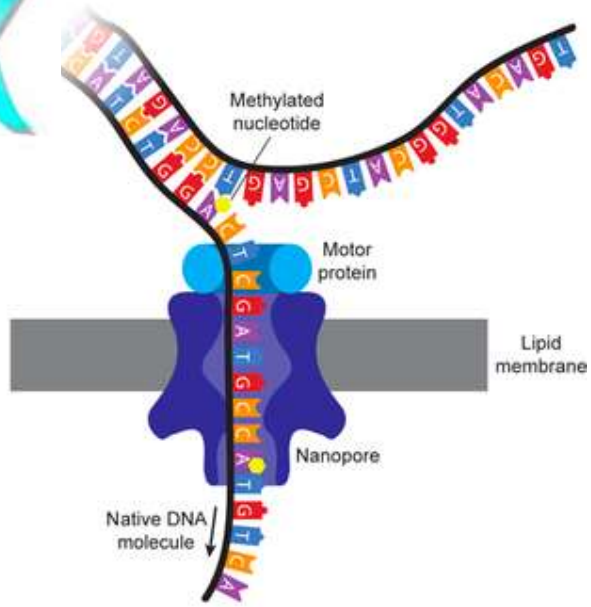


Combining data from both strands

- Chance of capturing complement after a template can be high
  - Need to optimise the setup to increase your odds
- Things that help:
  - DNA must have adapters at both ends
  - Fully adapted, repair any breaks or nicks
  - Lowering competition from other stands (dilute loading)

Sequencing Technologies





Oxford Nanopore technology

Comparable costs

NANOPORE PRODUCT FAMILY

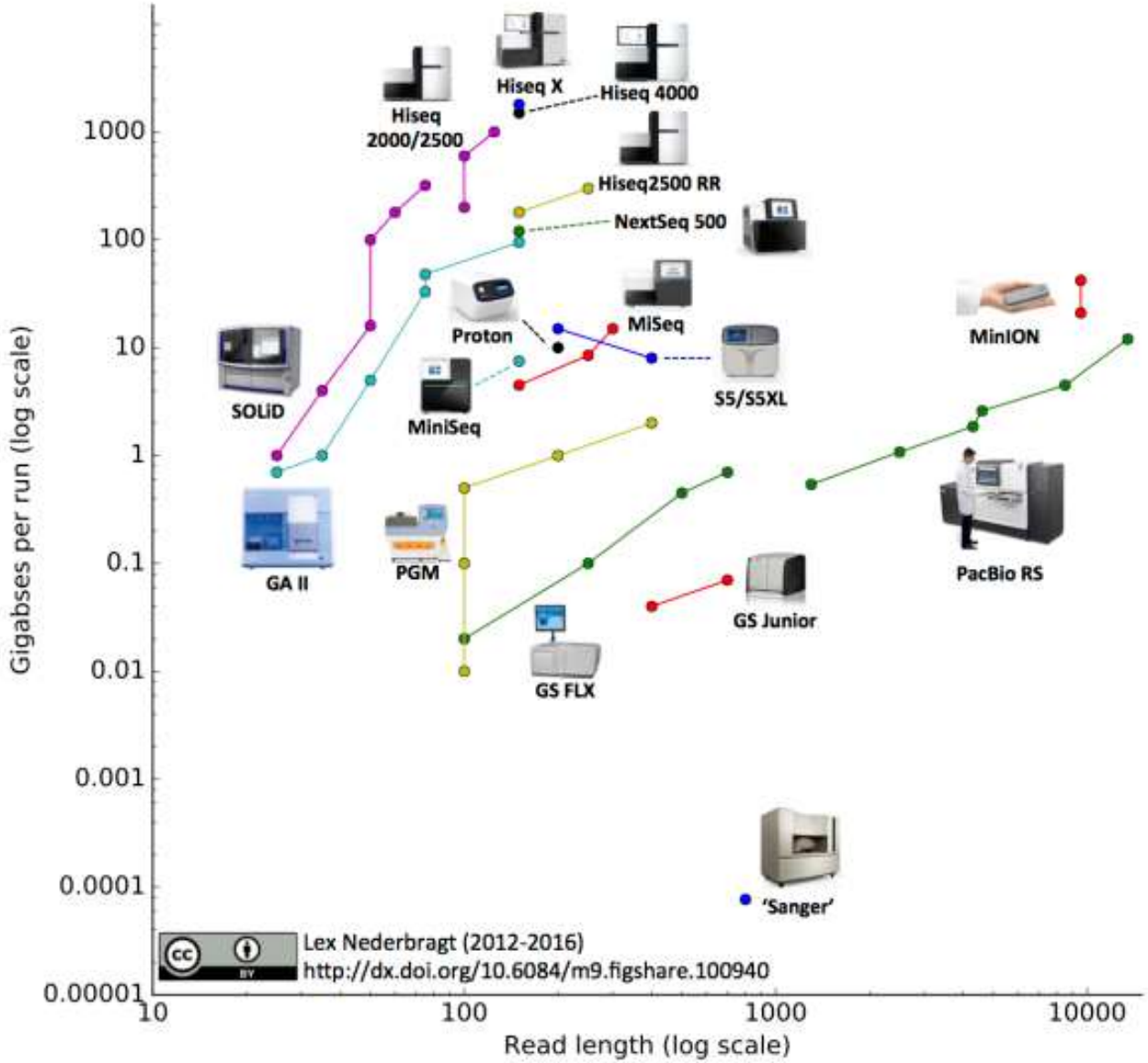
One core technology, real-time, on-demand





Key	SmidgION	Flongle	MinION	GridION	PromethION
System Price	TBC	Included in $5K Starter Pack	Included in $1K Starter Pack	Included in $50K Starter Pack	Included in $135K Starter Pack
Number of channels	200 channels	128 channels	512 channels	5 x 512 = 2,560*	48 x 3,000* = 144,000
Per flow cell Current Data – Max Data	TBC	1 - 3.3 Gb	17 - 40 Gb	17 - 40 Gb	125 - 311 Gb
Per Device Current Data – Max Data				85 - 200 Gb	3/6 - 20 Tb
Price per Gb Current Data – Max Data	**TBC**	**$90 - $30**	**$30 - $12.5**	**$17.5 - $7.5**	**$5 - $2**

Comparison of different NGS platforms



Applications of Next Generation Sequencing Technologies

Key messages

- Not all sequencing technologies are equivalent and equally applicable. They are a toolkit with different levels of outputs and qualities, and need to be selected according to the study design.

- Frequently selection of technology relies on budget availability or accessibility rather than on quality criteria.

- Today it is increasingly more frequent the combination of technologies to reach better results (hybrid approaches).