CHARLES UNIVERSITY

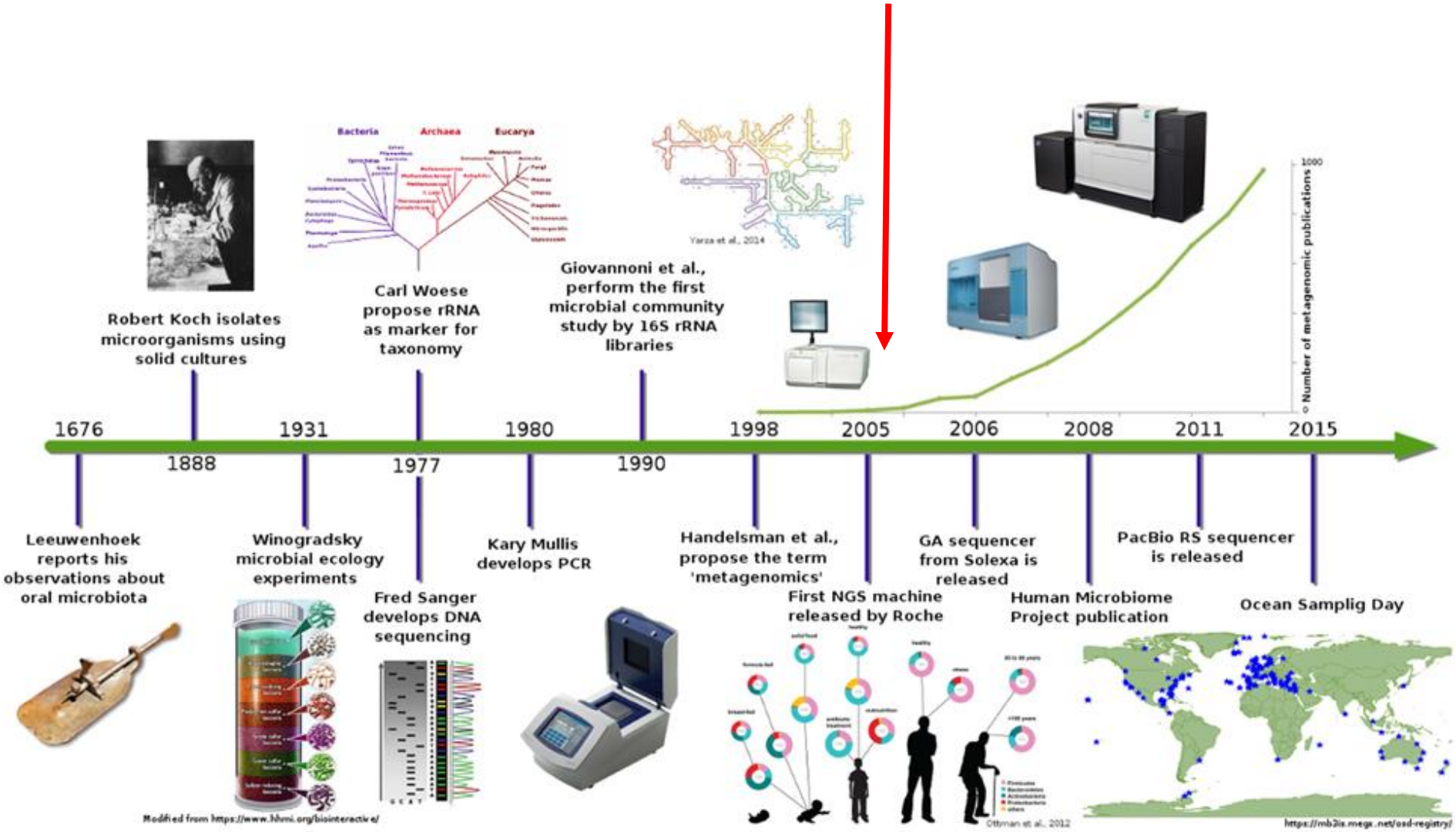
Bioinformatics and Microbiome Analysis MB140P94

Modern sequencing methods

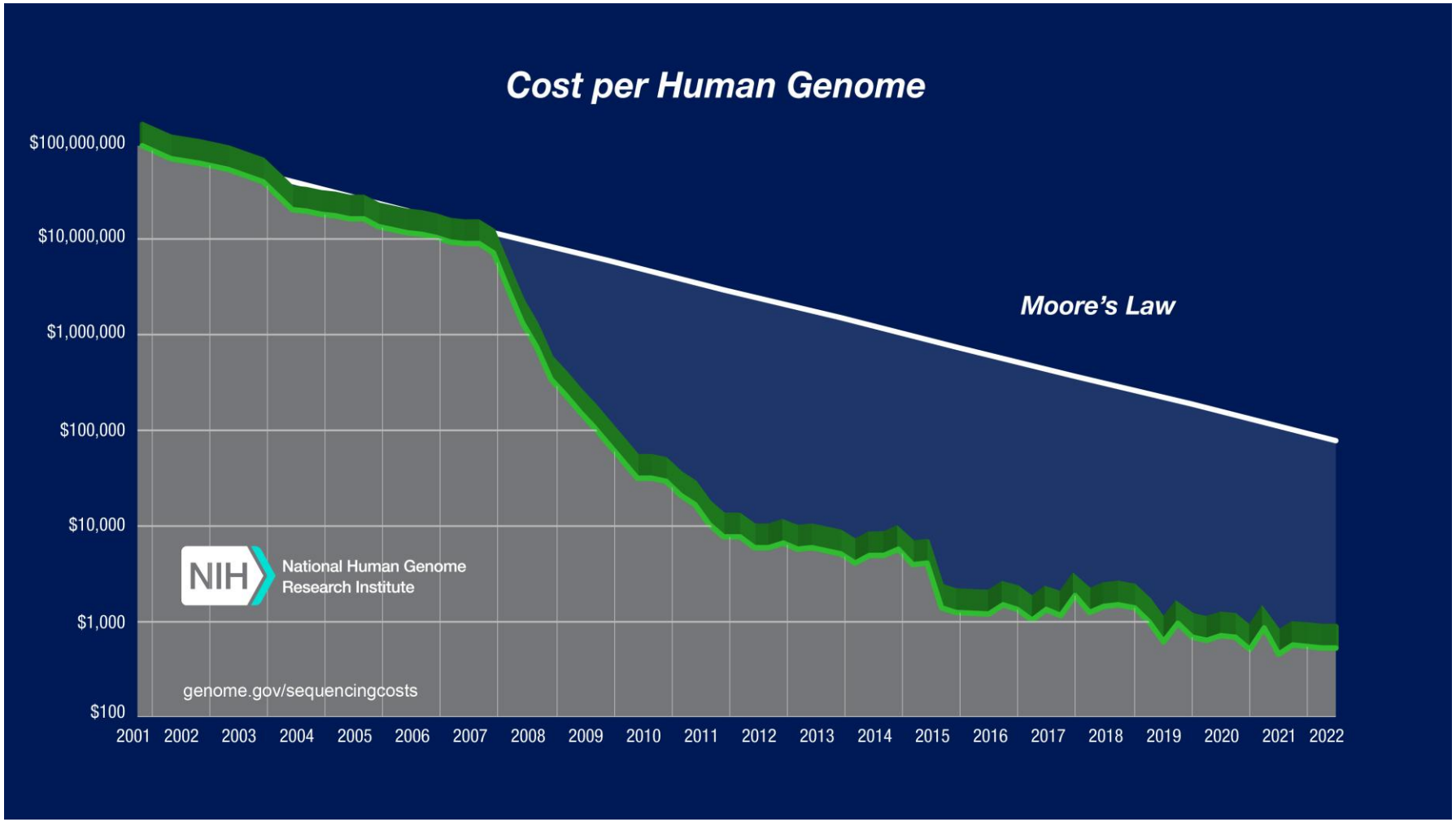
Tomáš Větrovský, Priscila Thiago Dobbler
Laboratory of Environmental Microbiology
Institute of Microbiology of the CAS

Genomics and DNA sequencing

high throughput sequencing (HTS) techniques were introduced by 454 Pyrosequencing

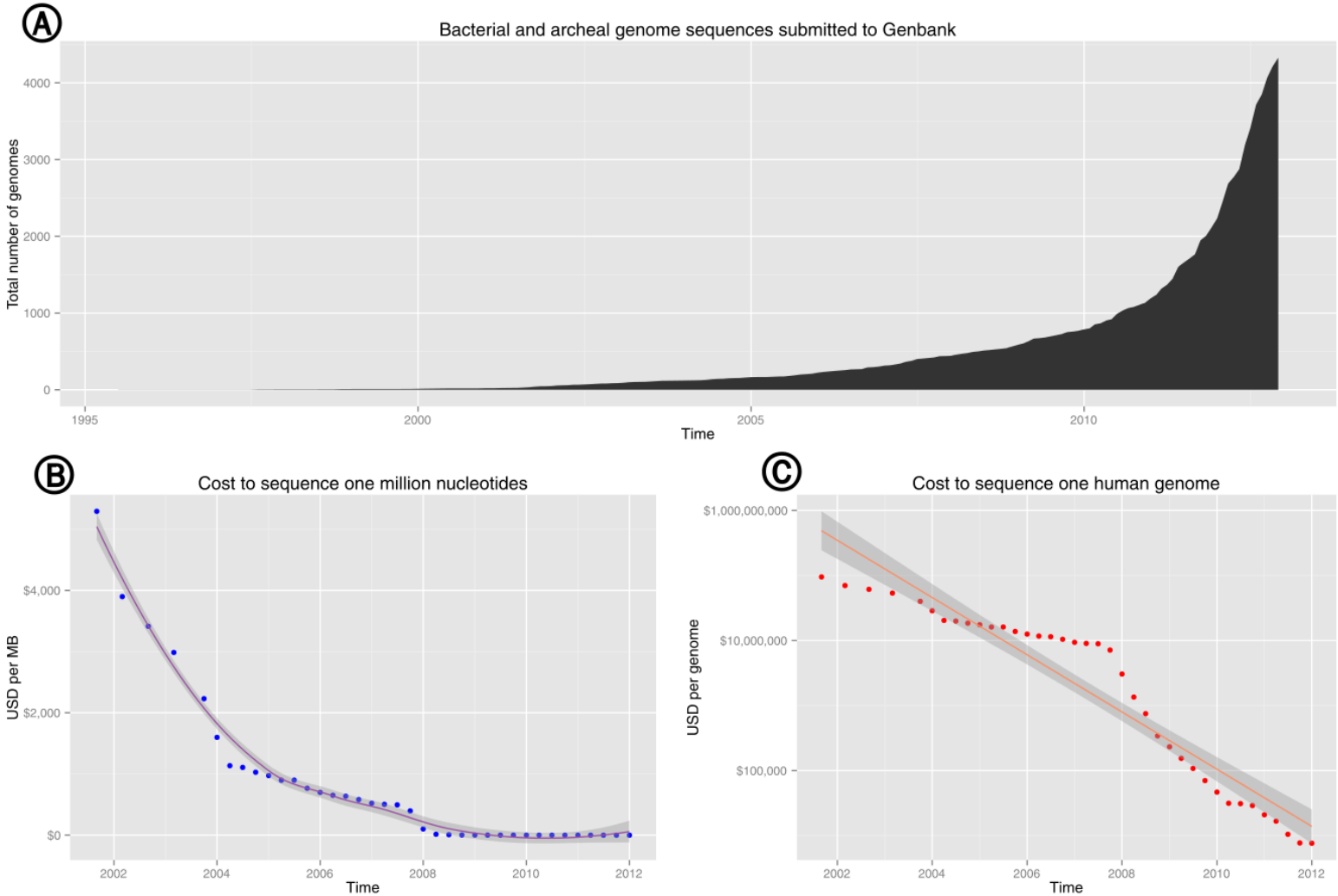


Fall of sequencing costs



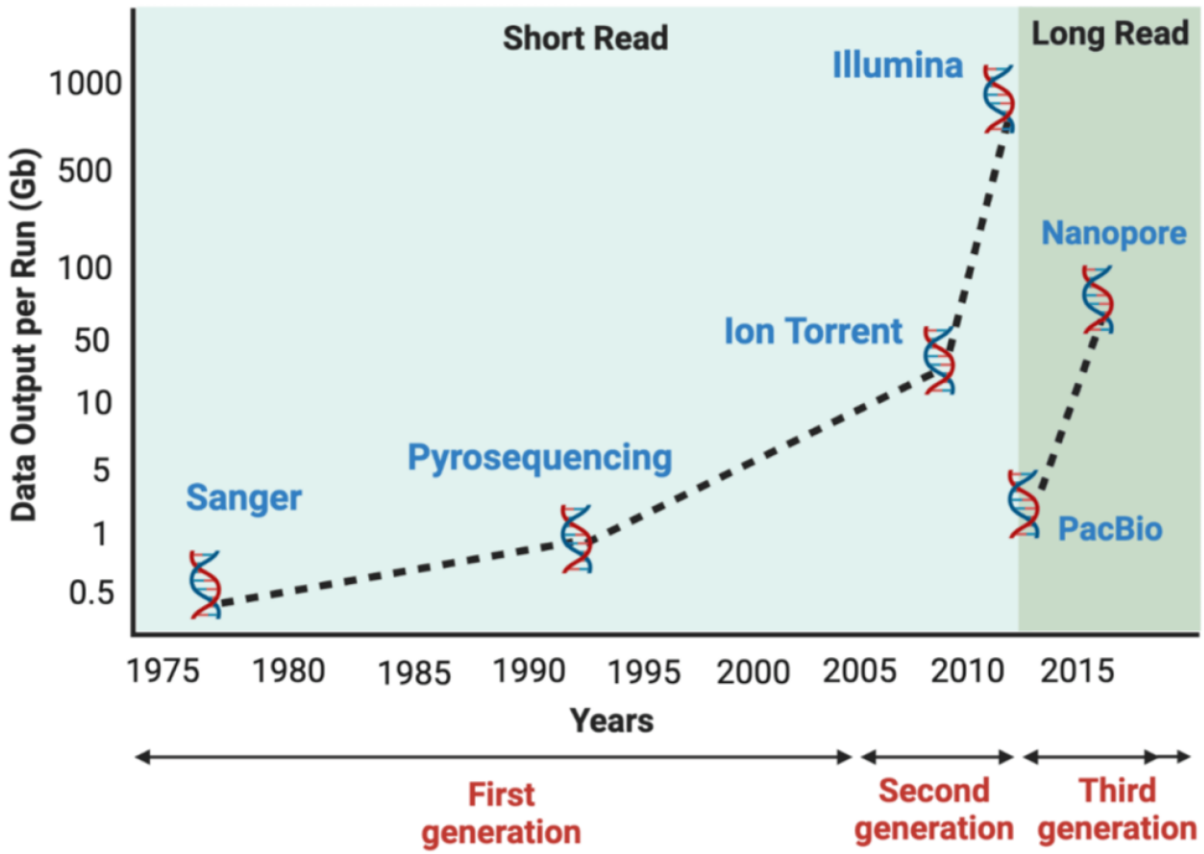
$1,000 genome

High throughput sequencing technologies have become essential in studies on genomics

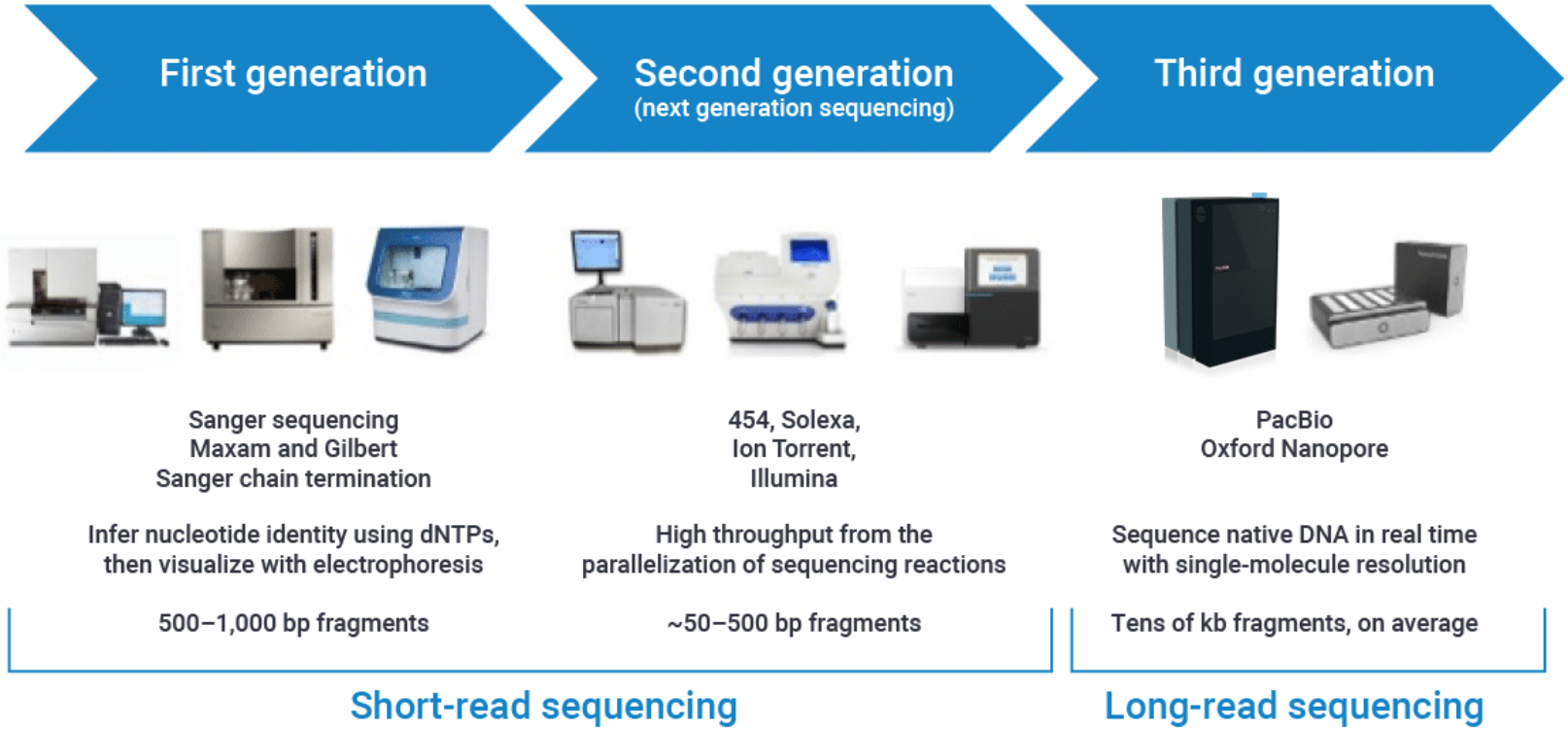


The number of genome projects has increased as technological improvements continue to lower the cost of sequencing. **(A)** Exponential growth of genome sequence databases since 1995. **(B)** The cost in US Dollars (USD) to sequence one million bases. **(C)** The cost in USD to sequence a 3,000 Mb (human-sized) genome on a log-transformed scale.

Evolution of sequencing technologies



The development of sequencing technologies over the past four decades can be categorized into three generations. The first generation was represented by **Sanger sequencing**, providing the foundation for DNA sequencing. The second generation introduced **massively parallel sequencing** with platforms such as Illumina and Ion Torrent, enabling high-throughput sequencing. The current third generation includes PacBio and Nanopore, offering **long-read and single-molecule sequencing** capabilities.



2nd generation sequencing

- exponentially increased sequence throughput and accuracy
- allows simultaneous sequencing of millions of different DNA fragments (cDNA, RNA) of approx. 30-1000 bp length (depending on the chosen platform and sequencing kit)
- fragment amplification (emulsion PCR, bridge amplification) - higher signal during nucleotide incorporation during sequencing, allowing detection

3nd generation sequencing

- it does not use amplification to increase the signal (should be higher accuracy)
- produces long reads
- good sequencing of GC rich areas
- Epigenetics
- PacBio and Nanopore (MinION)

Third generation

PacBio
Oxford Nanopore

Sequence native DNA in real time with single-molecule resolution

Tens of kb fragments, on average

Long-read sequencing

DNA sequencing platforms

Technology	Number of lanes	Injection volume (nL)	Analysis time	Average read length	Throughput (including analysis; Mb/h)
Slab gel	96	500–1000	6–8 hours	700 bp	0.0672
Capillary array electrophoresis	96	1–5	1–3 hours	700 bp	0.166
Microchip	96	0.1–0.5	6–30 minutes	430 bp	0.660
454/Roche FLX (2008)		< 0.001	4 hours	200–300 bp	20–30
Illumina/Solexa (2008)			2–3 days	30–100 bp	20
ABI/SOLiD (2008)			8 days	35 bp	5–15
Illumina MiSeq (2019)			1–3 days	2x75–2x300 bp	170–250
Illumina NovaSeq (2019)			1–2 days	2x50–2x150 bp	22,000–67,000
Ion Torrent Ion 530 (2019)			2.5–4 hours	200–600 bp	110–920
BGI MGISEQ-T7 (2019)			1 day	2x150 bp	250,000
Pacific Biosciences SMRT (2019)			10–20 hours	10–30 kb	1,300
Oxford Nanopore MinIon (2019)			3 days	13–20 kb[15]	700

Sanger

performance values for genome sequencing technologies including Sanger methods and next-generation methods

Sequencing platforms





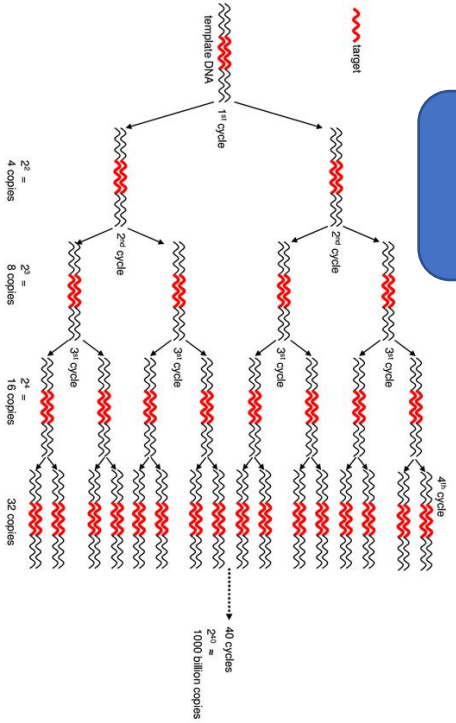
- 454 (Roche)
- SOLiD (Life Technologies)
- Ion Torrent (Life Technologies)
- Illumina (Illumina)
- **PACBIO (PacBio)**
- **Oxford Nanopore Technologies**
- BGISEQ-500 (BGI)
- PyroMark (QIAGEN)
- MGI (DNB SEQ)
- Element AVITI System (Element Biosciences)
- Onso (PacBio)

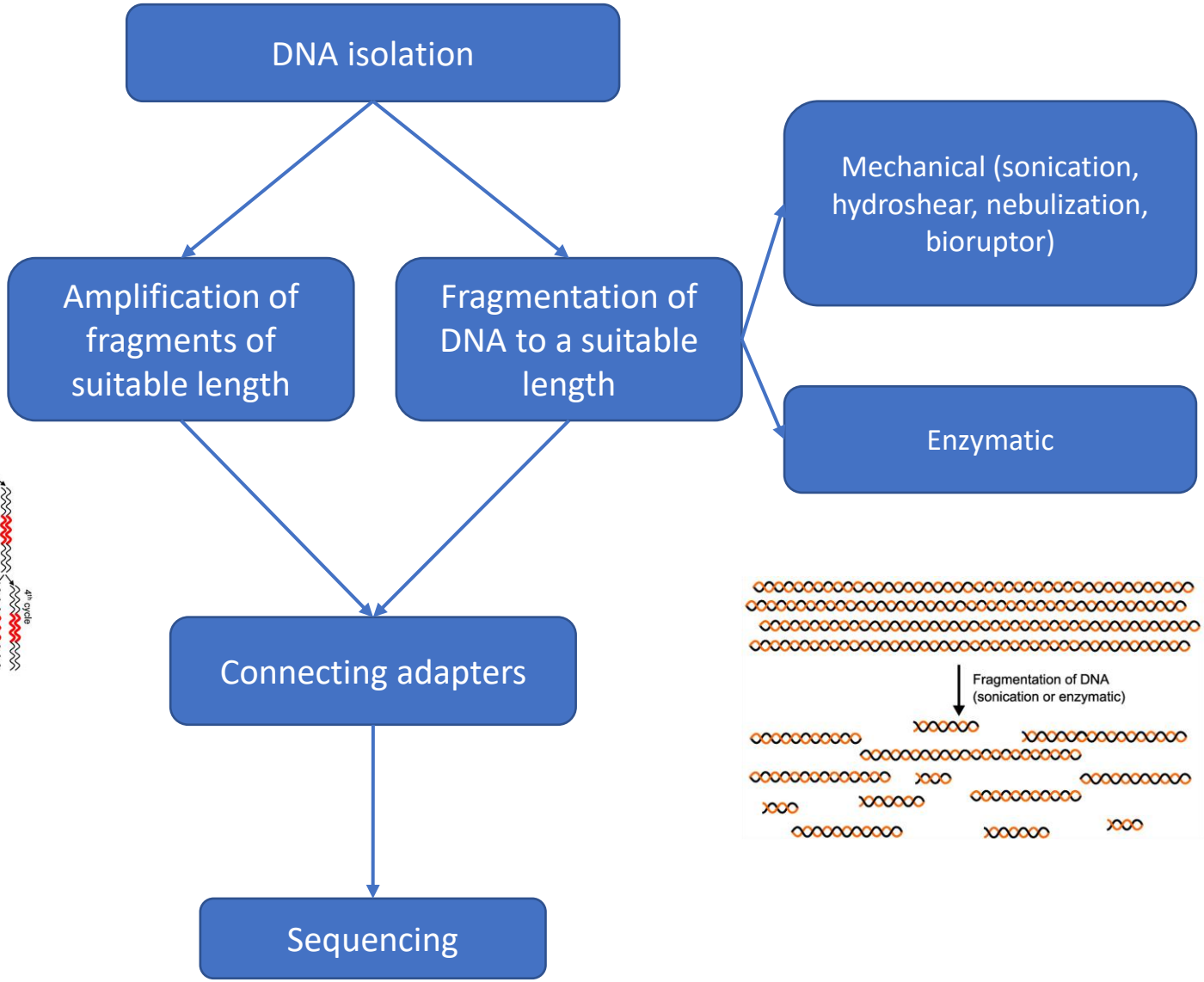




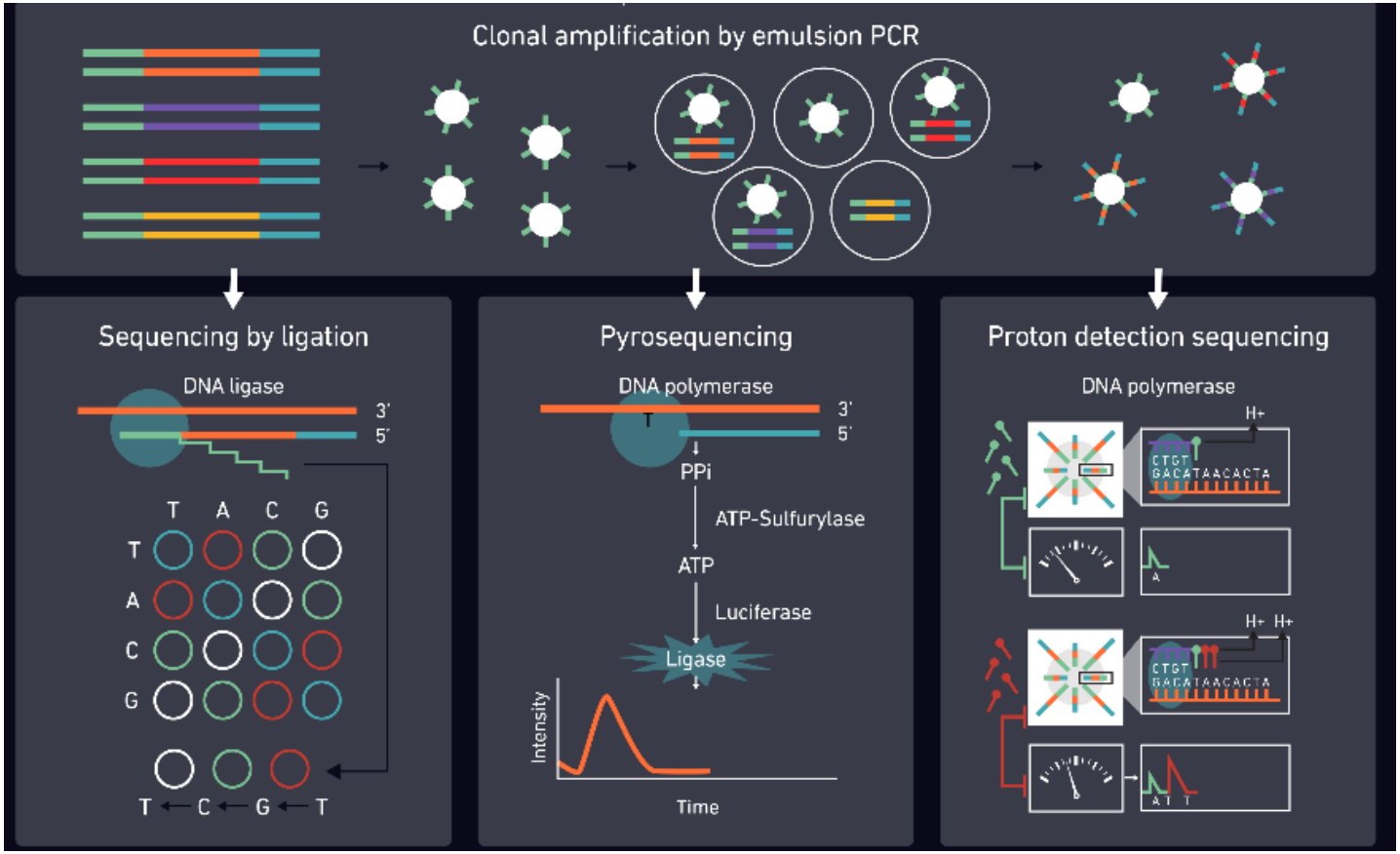


General (DNA) sequencing procedure



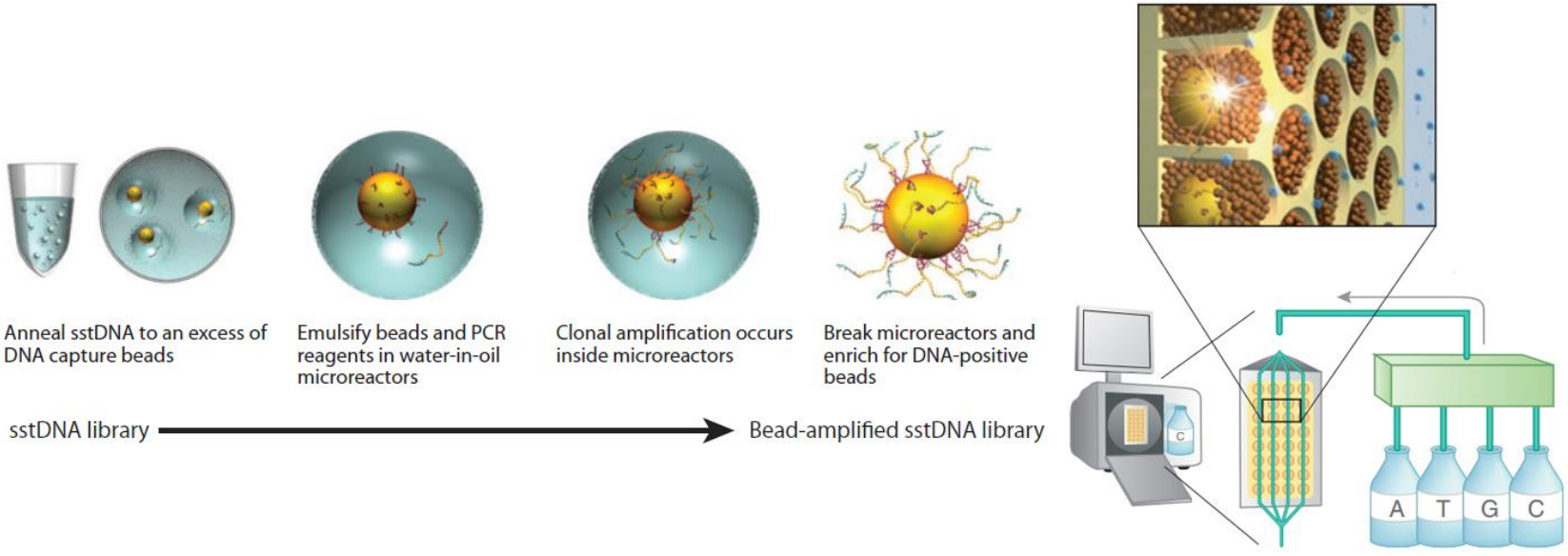


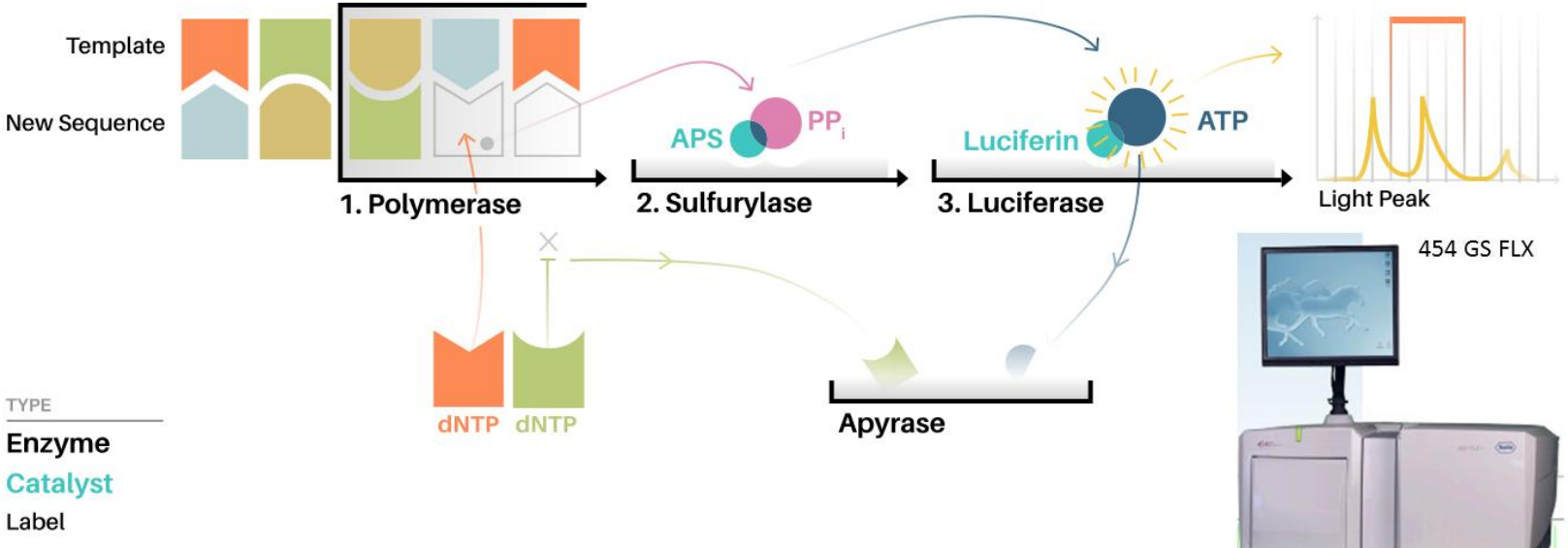
Extinct (and almost extinct) platforms



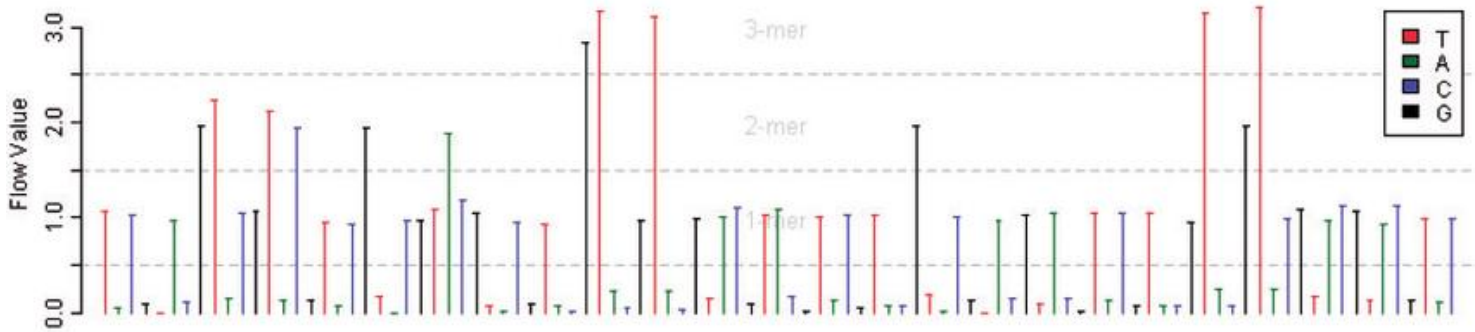
	SOLiD	454	Ion Torrent
Template preparation	EM PCR	EM PCR	EM PCR
Technology	SBL	SBS	SBS
Detection	Fluorescent double base marking	Chemiluminescence - pyrosequencing	Detection of released protons - pH change

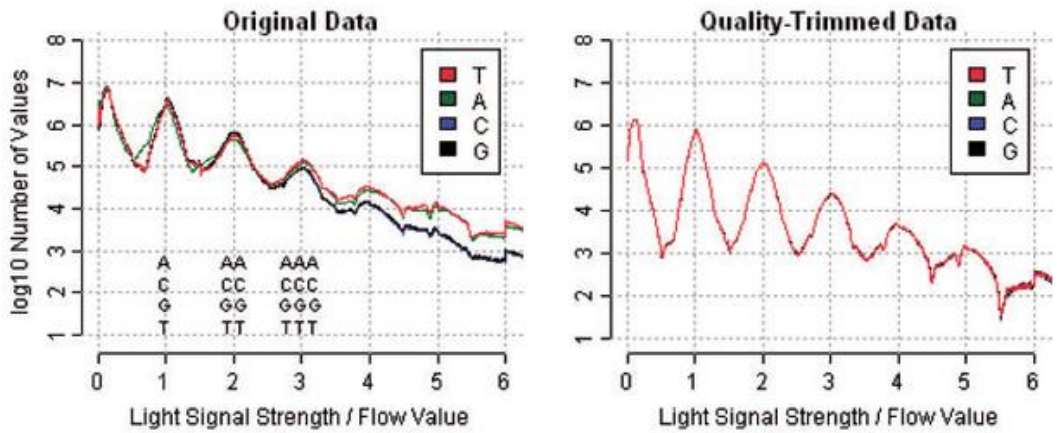
454 - pyrosequencing





454 - pyrosequencing
homopolymer-related sequence errors



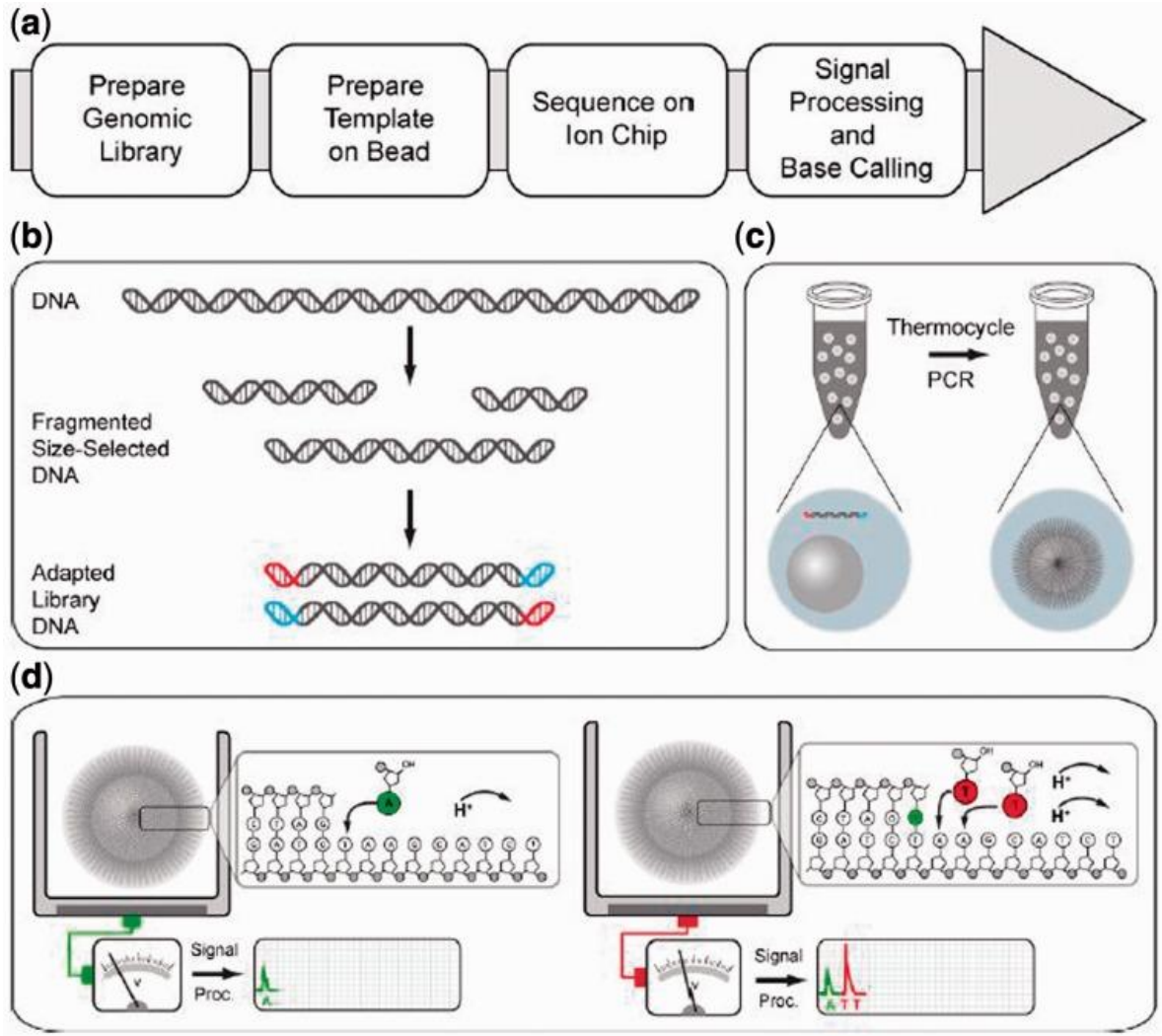


Direct Comparisons of Illumina vs. Roche 454 Sequencing Technologies on the Same Microbial Community DNA Sample

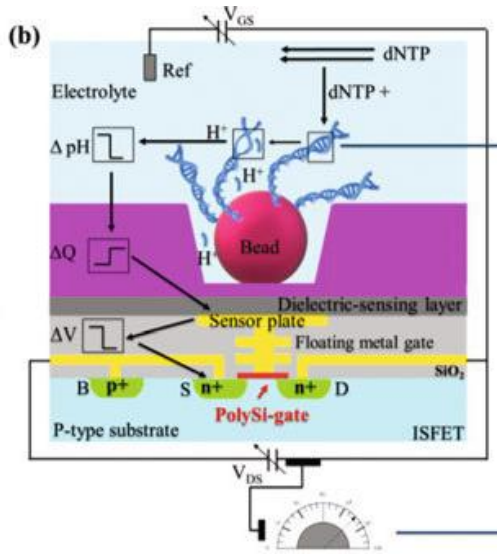
https://doi.org/10.1371/journal.pone.0030087

Ion Torrent

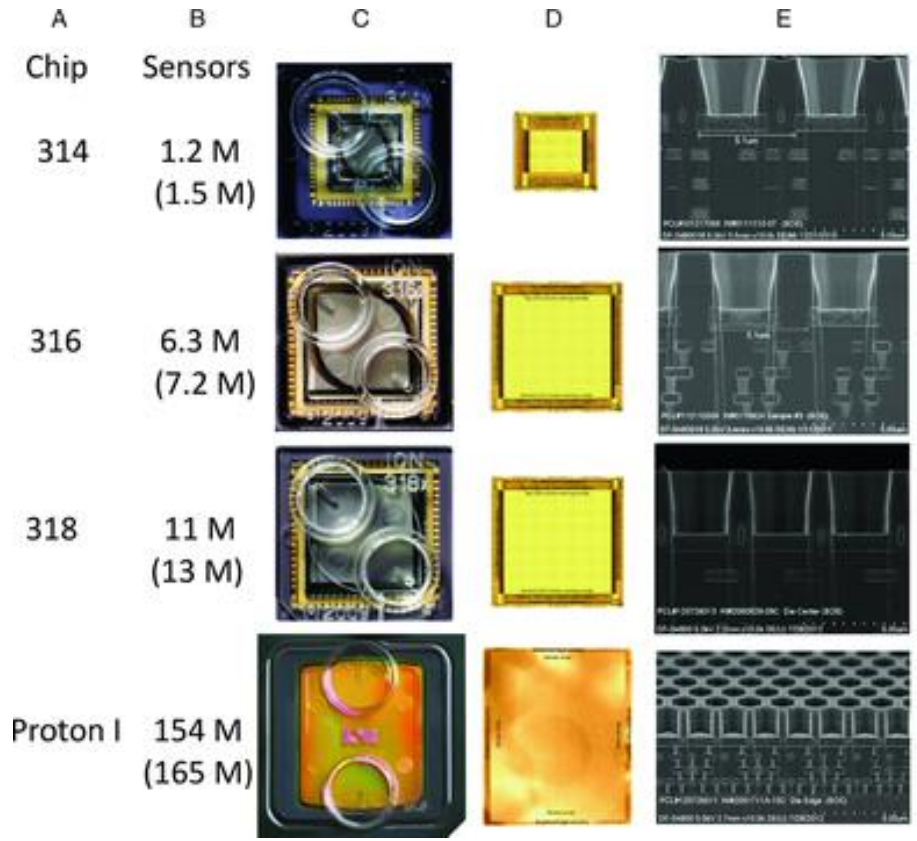
Semiconductor ion sequencing:
Semiconductor sequencing is a SBS technology based on **detecting the hydrogen ion** that is released during the DNA synthesis reaction by a very sensitive pH meter—a microchip sensor.



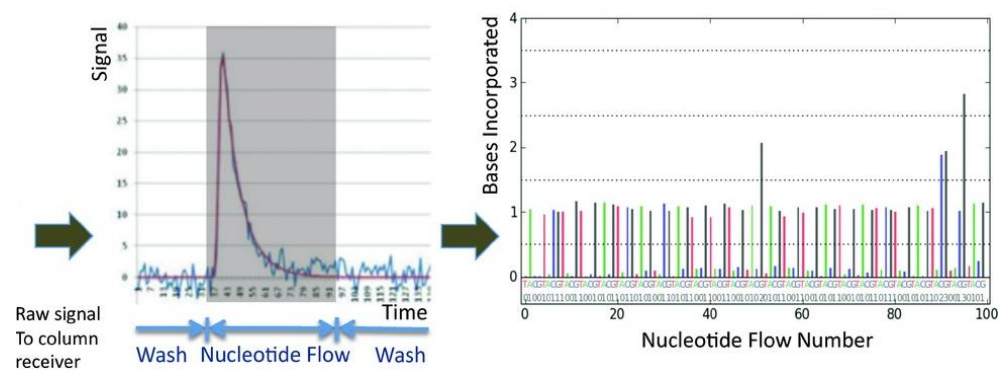
Ion Torrent



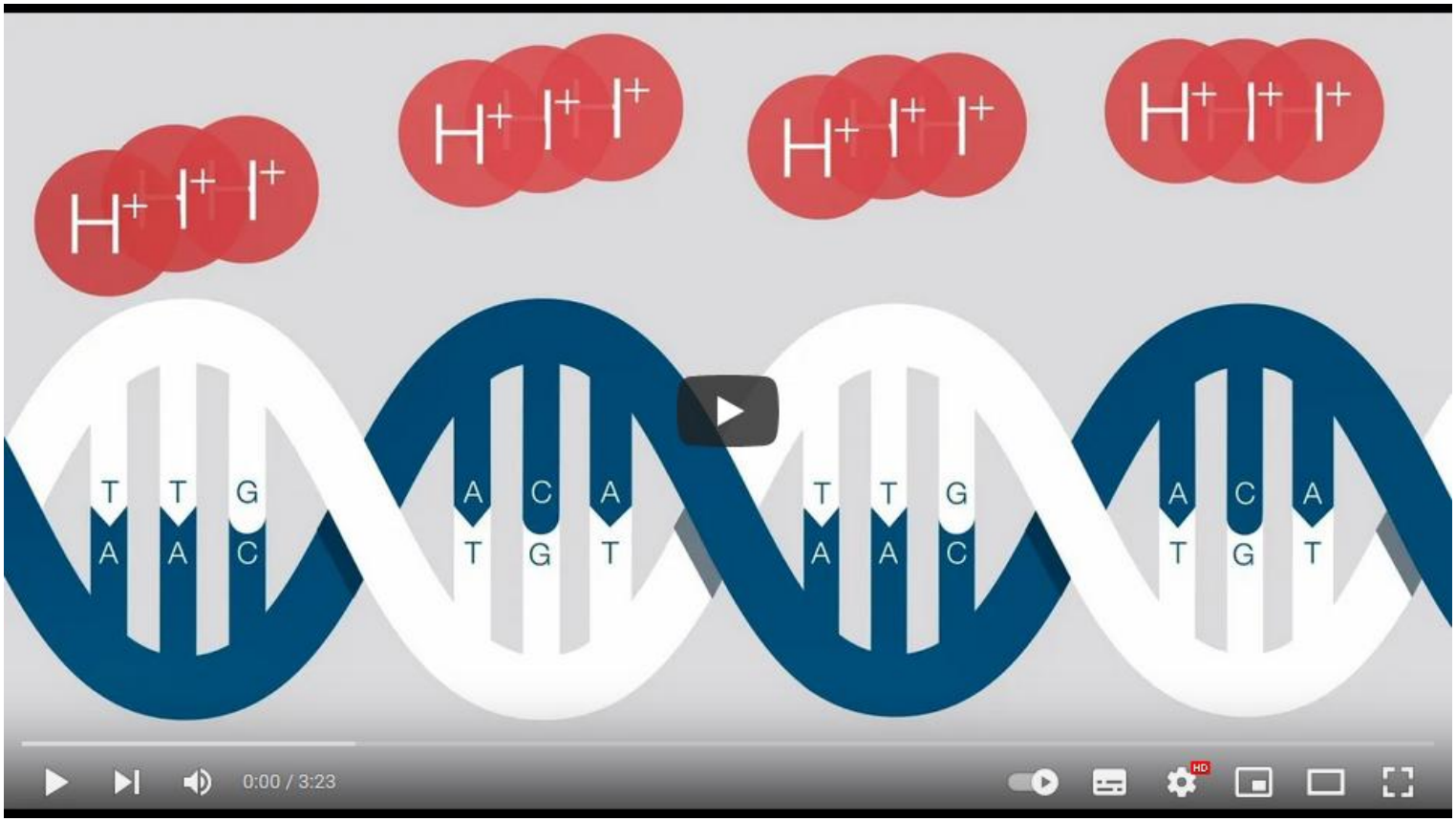
Single pH Sensor
Chemical to Digital Sequence



- template synthesis
- dNTP incorporation releases H^+ changing pH followed by change in voltage
- semiconductor with millions of pH sensors

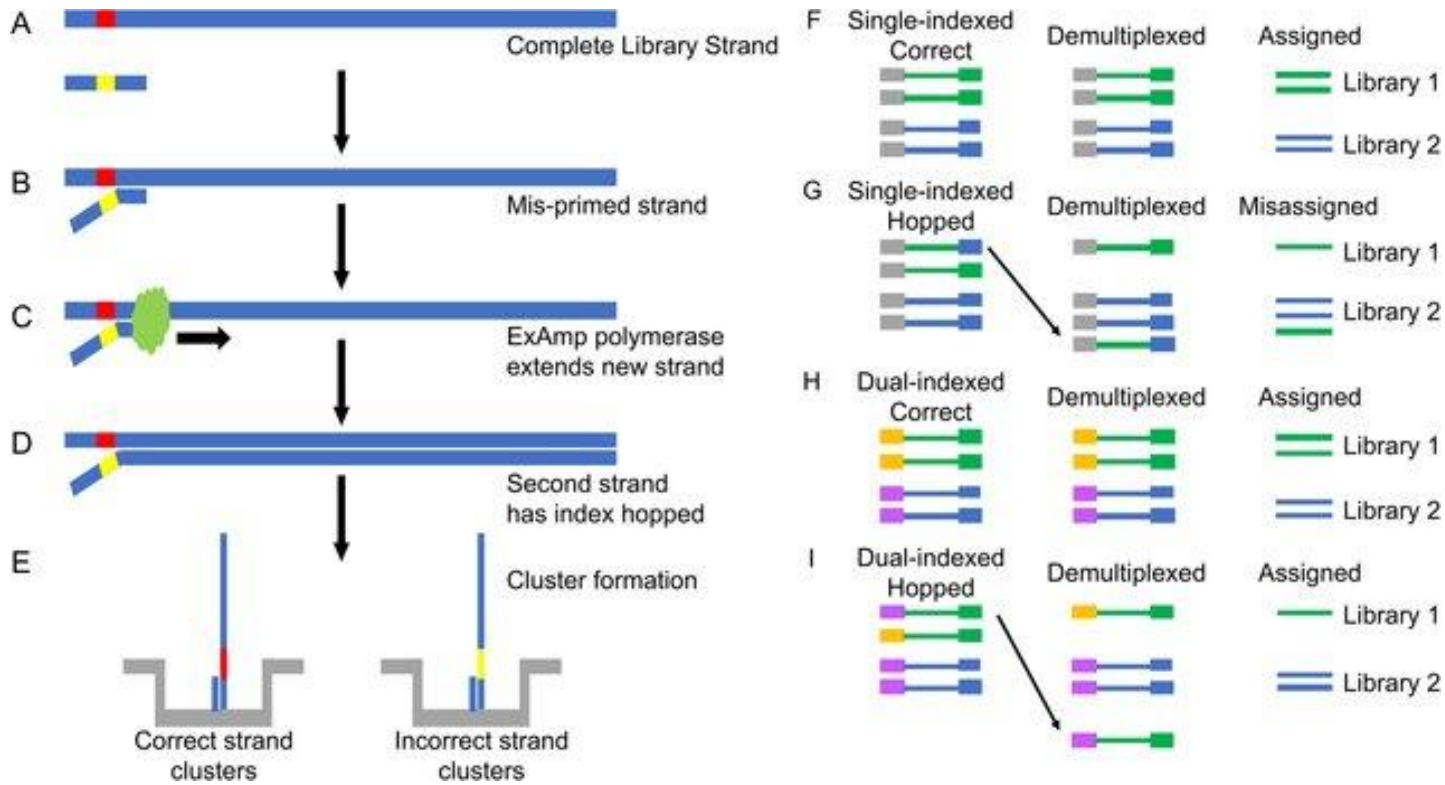


Ion Torrent



https://www.youtube.com/watch?v=zBPKj0mMcDg

Sequencing Glossary

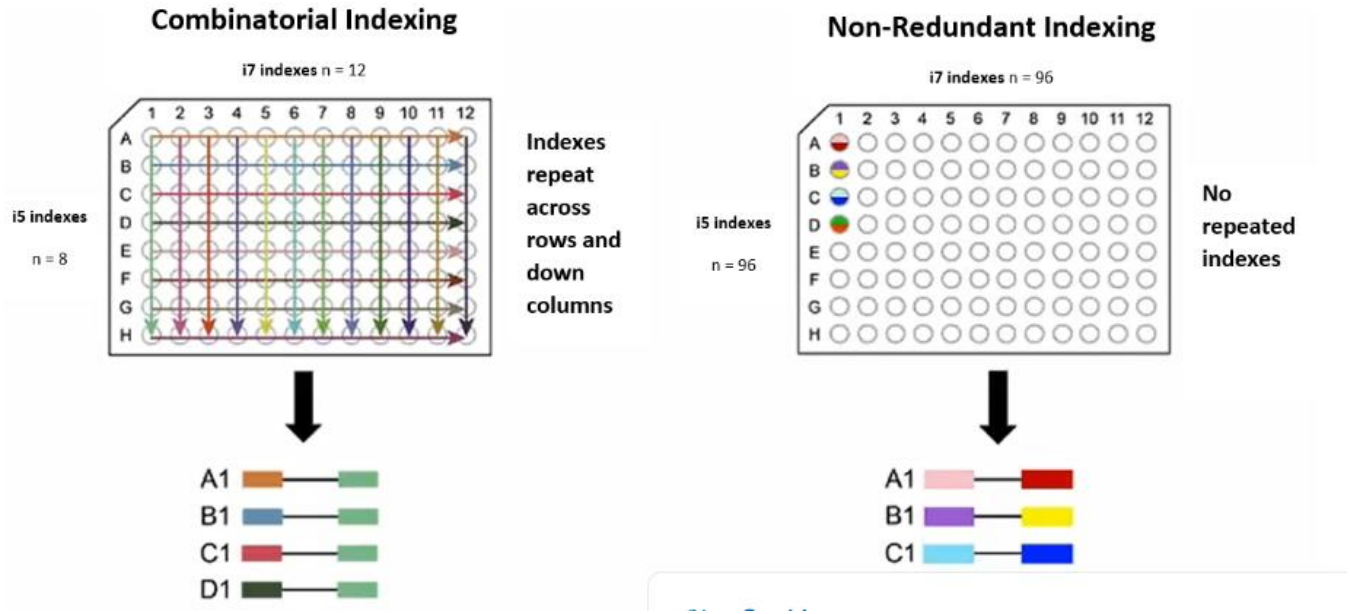


Index hoping

Index hopping occurs due to the **physical incorporation of a sample index from one library into a sample molecule from another library**. The end result is an incorrect read assignment between samples.

Sequencing Glossary

UDI (Unique Dual Indexes)



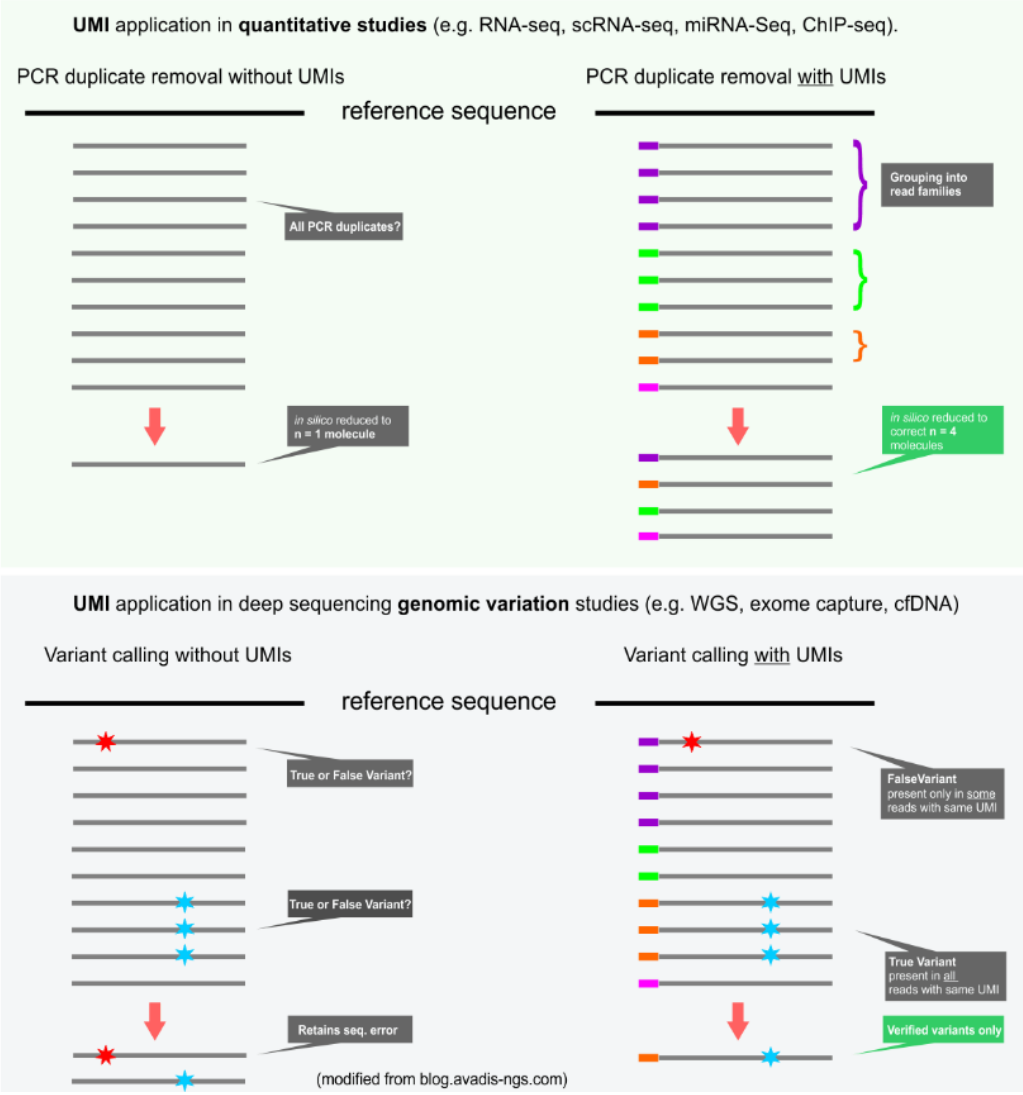
https://knowledge.illumina.com/library-preparation/general/library-preparation-general-reference_material-list/000002344

there is a **different index on each side of the sequenced strand**, allowing you to filter sequences where index hopping has occurred

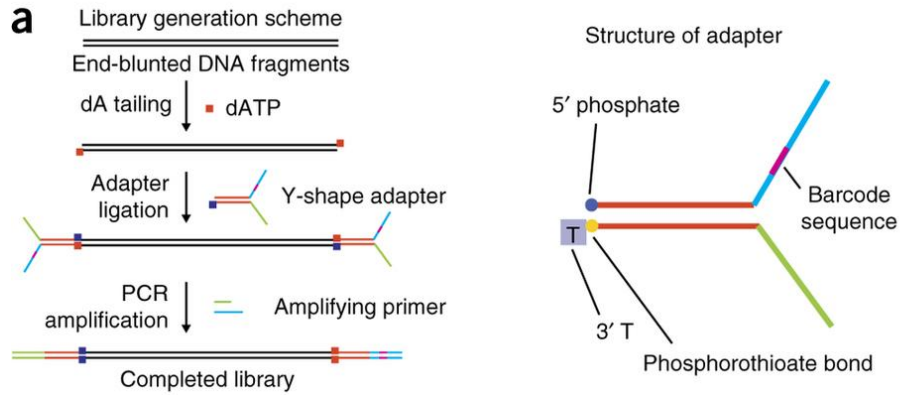
Sequencing Glossary

UMI (Unique Molecular Identifier)

- unique sequences used to label individual DNA fragments prior to amplification
- allows fragment tracking during library preparation, targeted enrichment and data analysis
- allows differentiation of PCR duplicates and rare variants

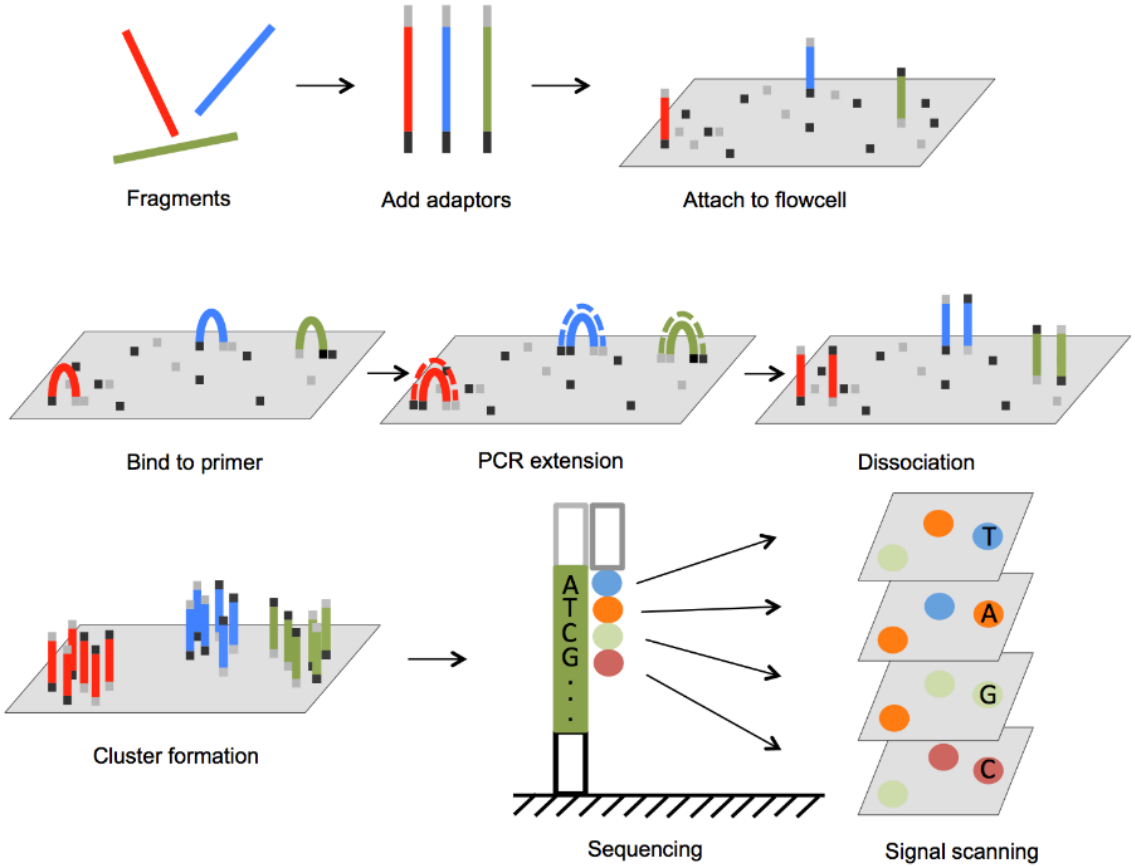


Next (Second) Generation Sequencing - short fragments, high throughput, low price…

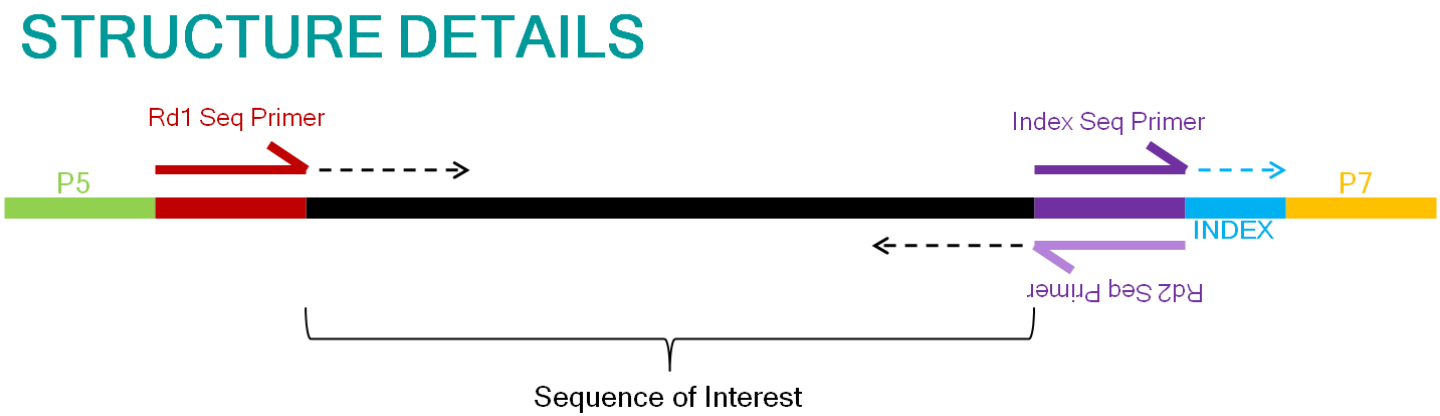


Illumina sequencing

- reversible dye-terminators
- identification of single bases introduced into DNA strands
- DNA attaches to the flow cell via complementary sequences.
- strand bends over and attaches to a second oligo forming a bridge
- polymerase synthesizes the reverse strand
- two strands release and straighten. Each forms a new bridge (bridge amplification)
- result is a cluster of DNA forward and reverse strands clones.



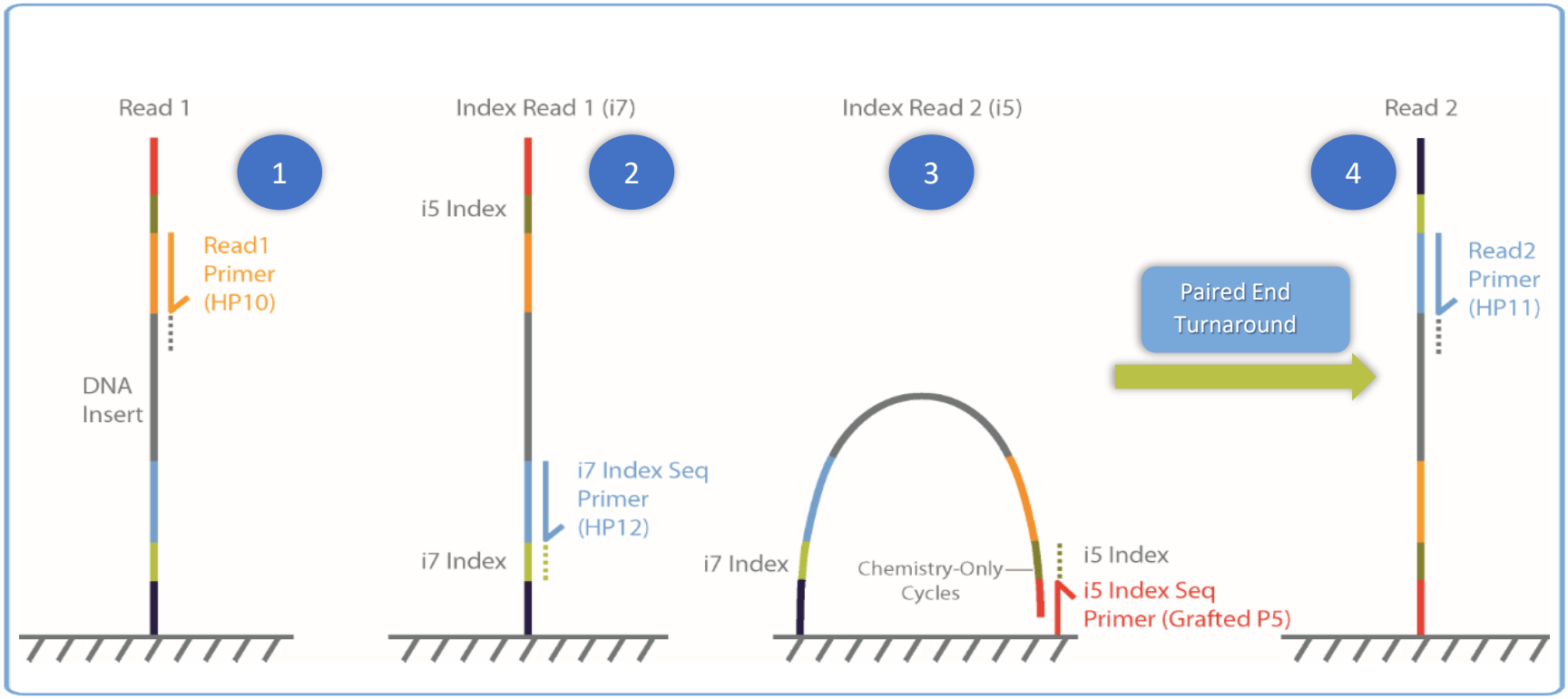
Illumina Sequencing



Adapters must include:

- sequence **complementary to oligonucleotides (grafts) occurring on the flowcell**, thus immobilizing the DNA. There are **two types of grafts on the flowcell, P5 and P7**, so our fragments must have a different adapter on each side.
- **tag (index, tag, mid) to help distinguish the samples from each other** after sequencing. While index 1 always has its own primer and thus there must be a region on the adapter for it, for index 2 the graft from flowcella serves as a "primer" in some instruments, in others it has its own special primer and thus the situation is similar to index 1
- place where the **sequencing primer** sits both **for read 1** and after flipping the sequence **for read 2**.

Illumina Sequencing

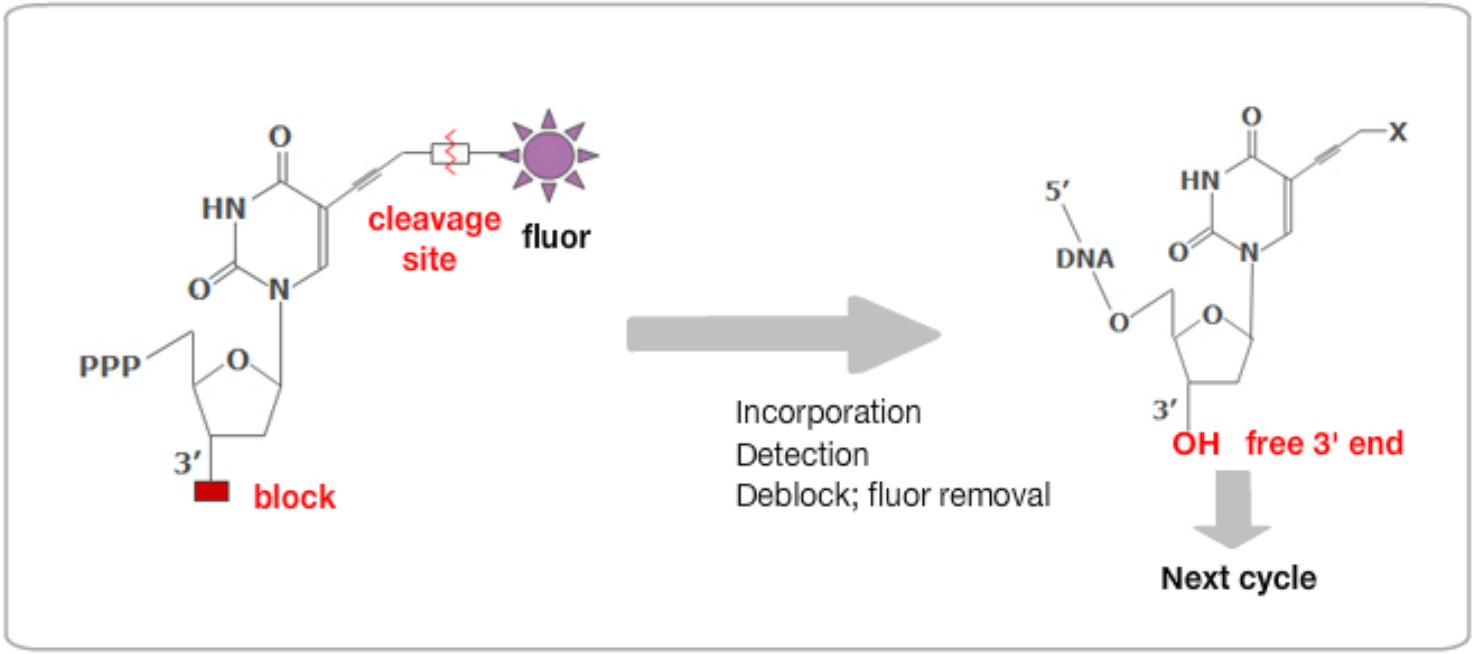


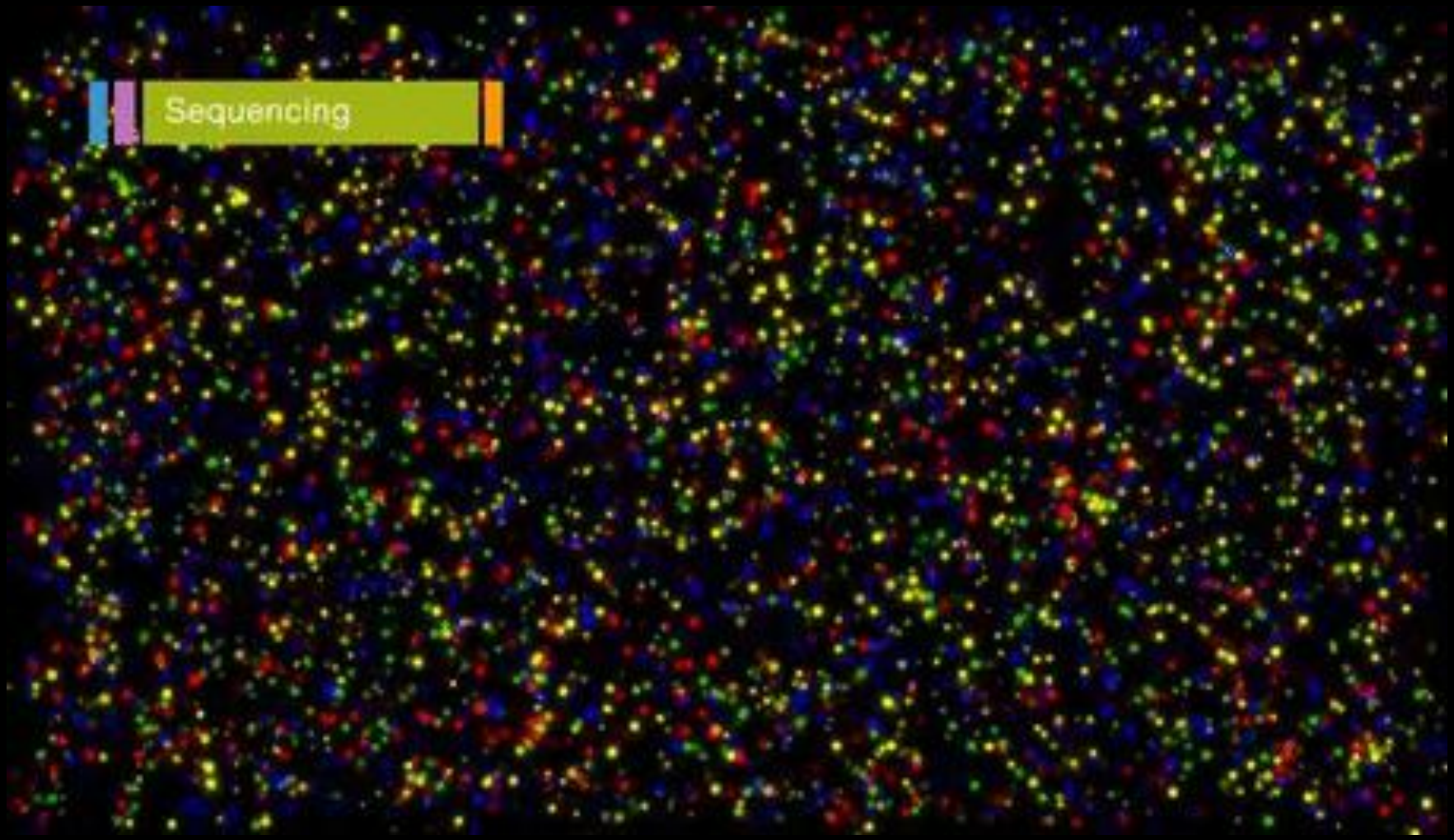
Read one is always read first, which is done by the aforementioned sequencing primer for read 1. Then **index 1 is read using its own primer**. After synthesizing the second string, we wash the original string (thus flipping the sequence to the opposite orientation), and we can sequence read 2 using the primer for read 2. Therefore, in the results, we see all the quality ratings for the 4 reads, whose order is read 1, index 1, index 2 and read 2. The advantage of reading the indices separately is that they are always read with sufficient quality, even if for some reason the reading quality of the actual sequence decreases.

Detection: fluorescently labelled bases with reversible terminator

The MiSeq sequences the DNA clusters using Illumina's Sequencing By Sythesis (SBS) Chemistry which relies on Reversible Terminator Chemistry (RTC).

- All 4 labeled nucleotides in 1 reaction
- Higher accuracy



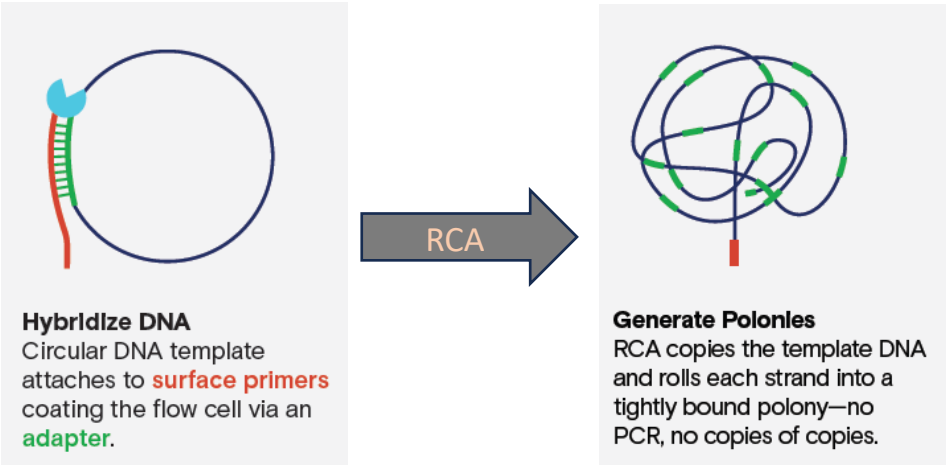

Sequencing

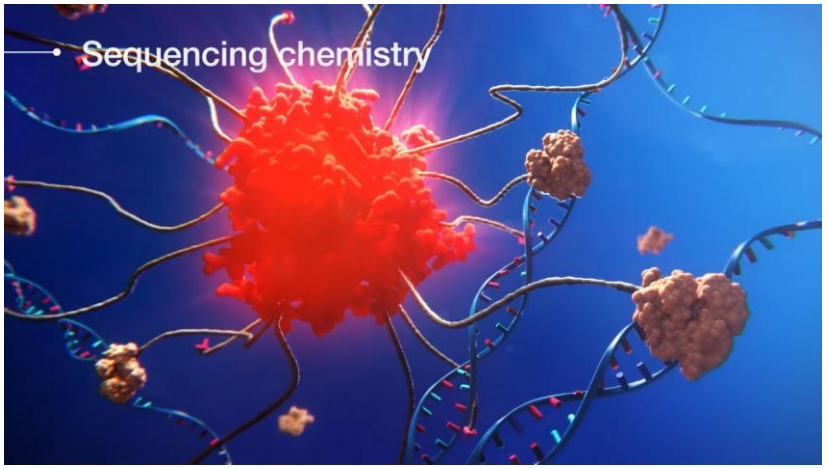
Illumina

https://www.youtube.com/watch?v=fCd6B5HRaZ8

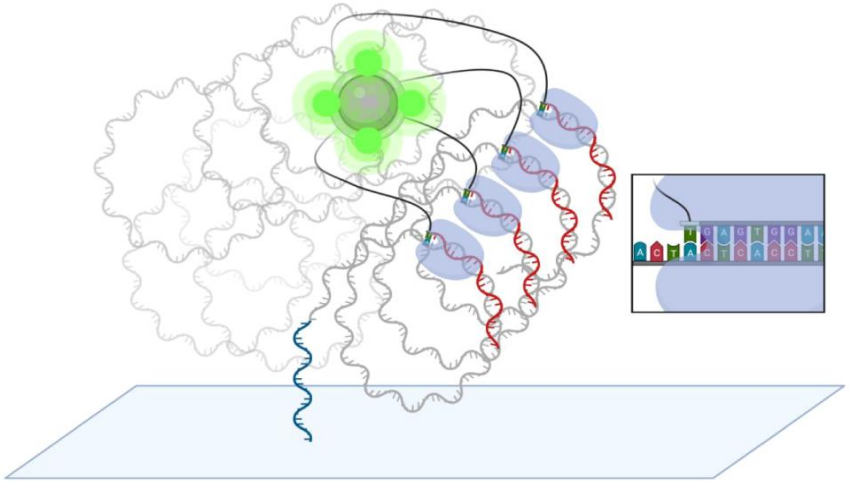
New 2nd gen sequencer - AVITI

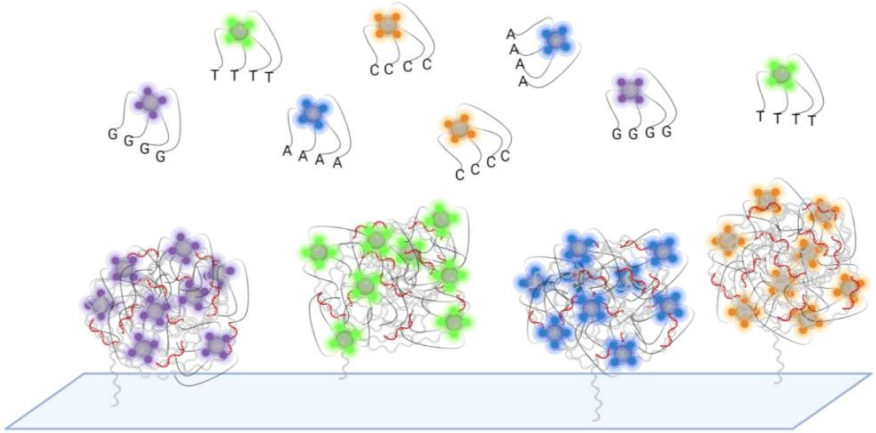
- also by template synthesis and optical detection
- other 'dyeing' methods uses monovalent labels, AVITI uses <u>polyvalent</u> labels – multiple dNTPs bound to a single fluorescent 'core'
- Rolling Circle Amplification (RCA)
  - isothermal and no PCR error propagation
  - polymerase continuously adds dNTPs to a primer annealed to a circular template

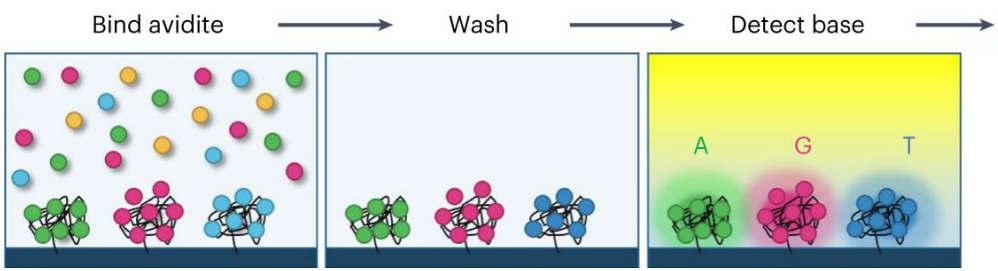




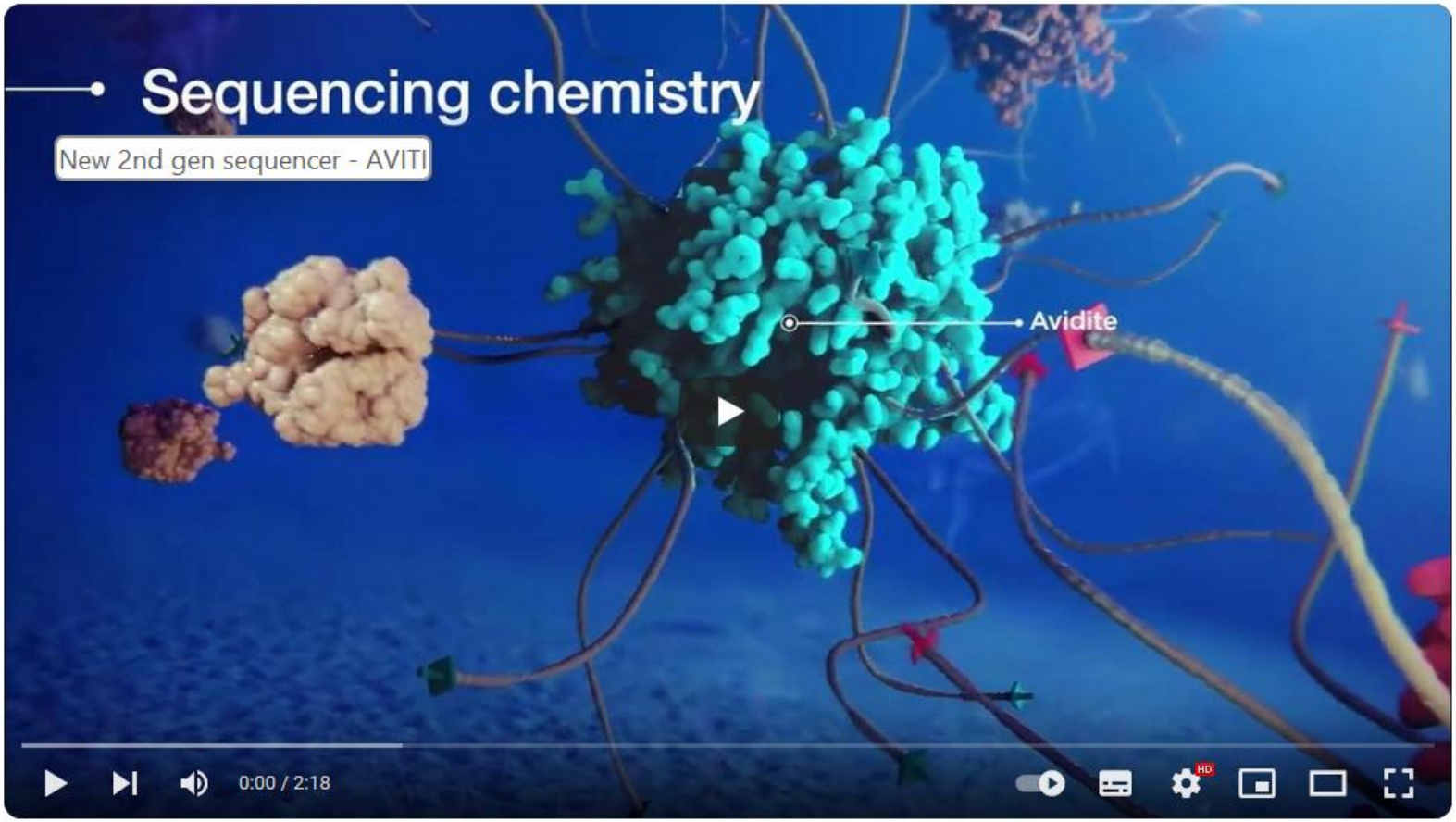
New 2nd gen sequencer - AVITI







AVITI



https://www.youtube.com/watch?v=b_cC5wi2OYg

3rd gen sequencing – Single molecule sequencing

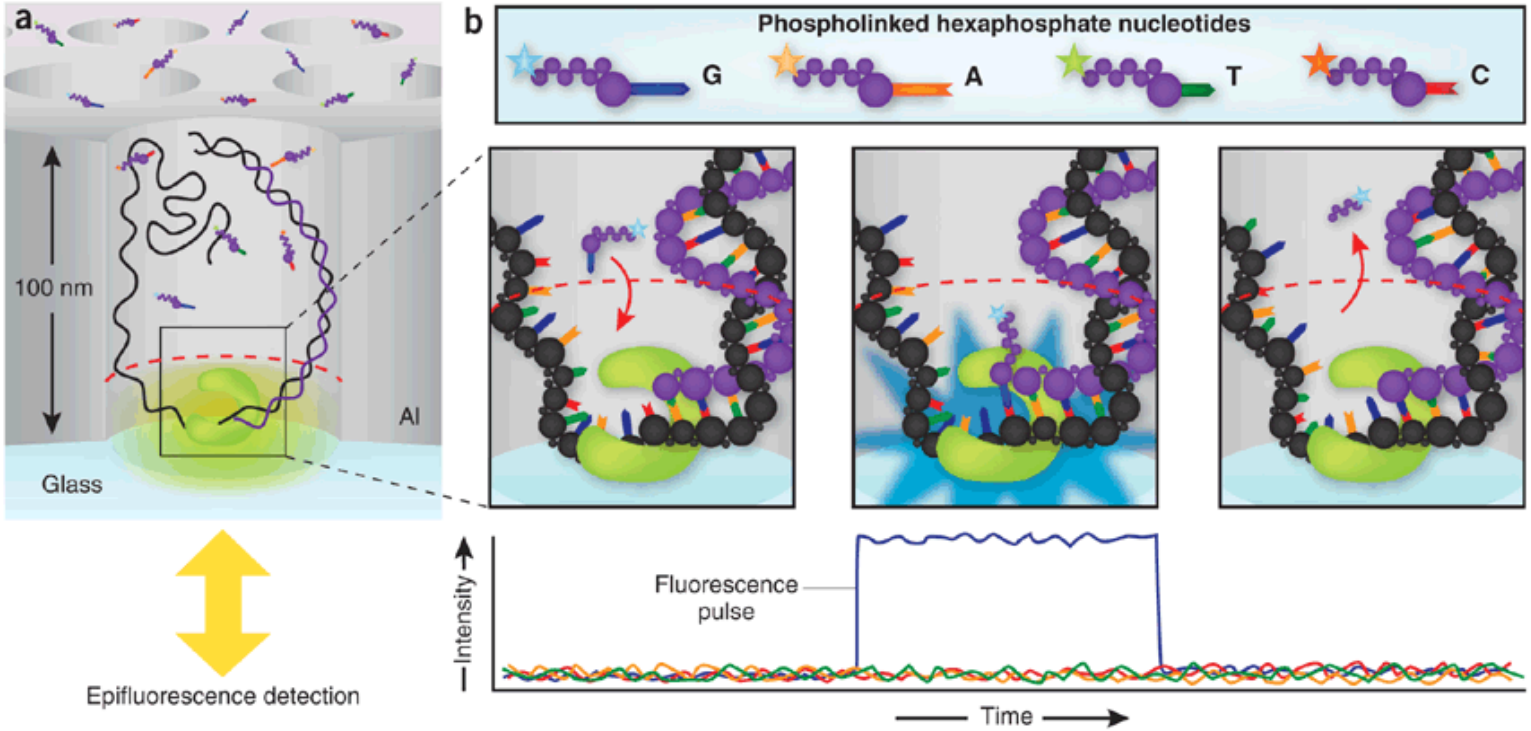
- initially very low accuracy: <u>single molecule</u> sequencing leads to ↓ signal to noise ratio

- still lower throughput than 2nd generation

- main technologies are:
  - PacBio CLR/HiFi – sequencing by template synthesis
  - Oxford Nanopore Sequencing (ONT) – sequencing by 'flow'

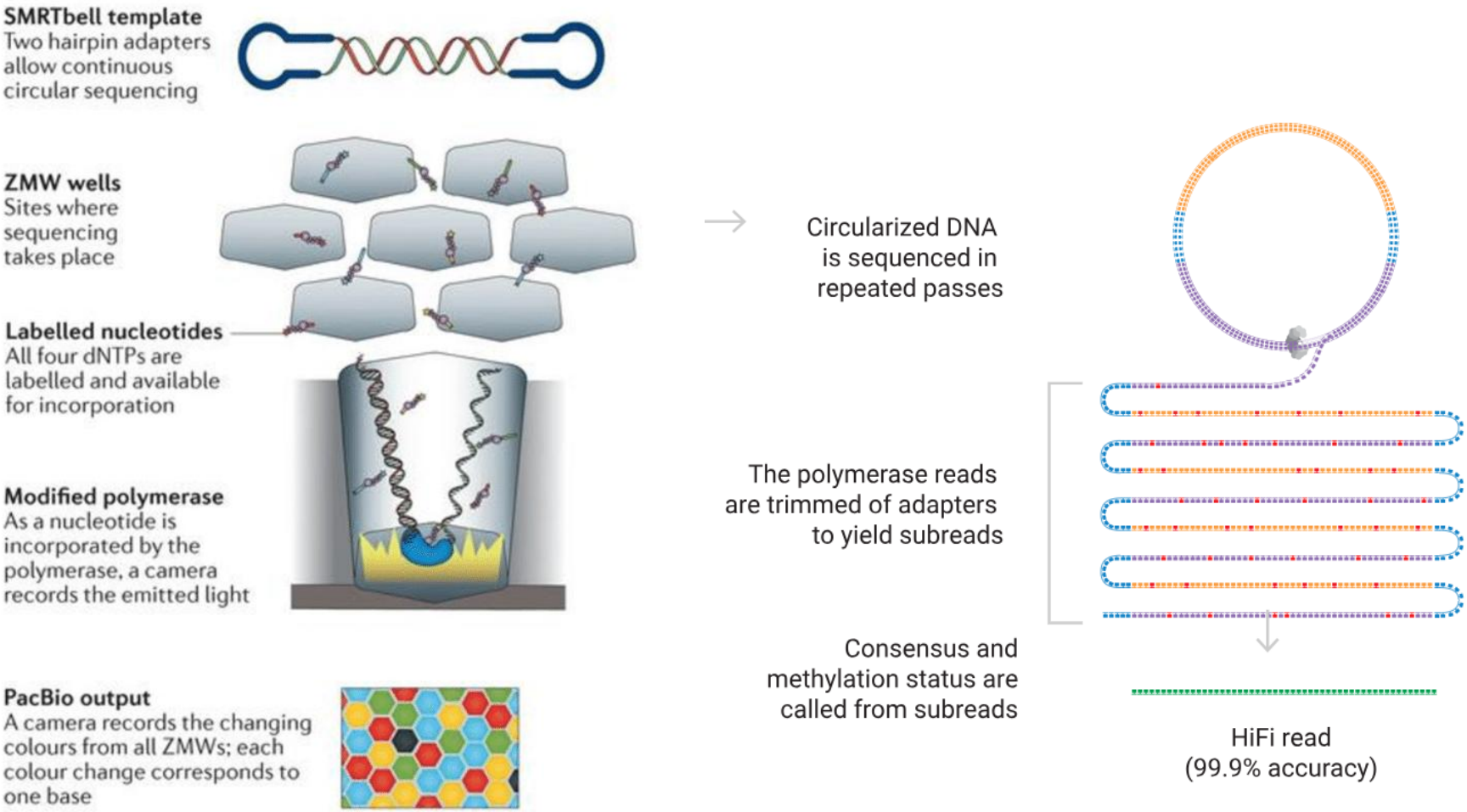
3rd gen sequencing – Single molecule sequencing

Single molecule real time sequencing (**SMRT**) is a parallelized single molecule real time DNA sequencing method. **DNA polymerase enzyme is affixed at the bottom of a zero-mode waveguide** (ZMW). The ZMW is a structure that creates an illuminated observation volume that is small enough to observe only a single nucleotide of DNA being incorporated by DNA polymerase. **Each of the four DNA bases is attached to one of four different fluorescent dyes**. When a nucleotide is incorporated by the DNA polymerase, the fluorescent tag is cleaved off and diffuses out of the observation area of the ZMW where its fluorescence is no longer observable.



PacBio – HiFi long reads

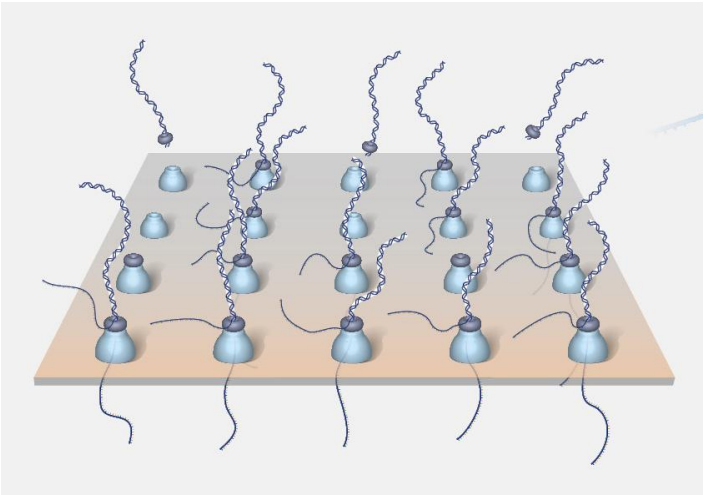
- initially accuracy was ~80%
- sequences up to 25kb
- same template is sequenced several times
- longer reads, less passes, lower accuracy

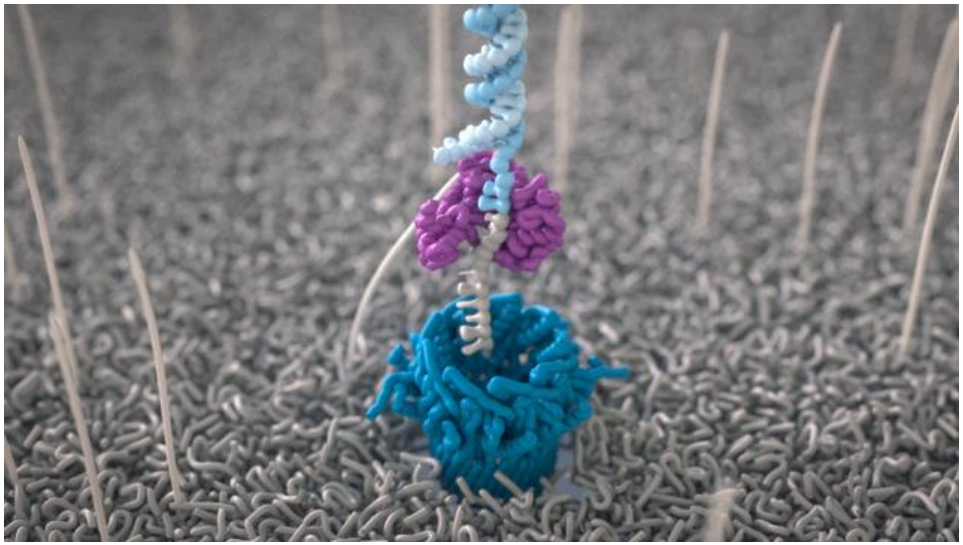


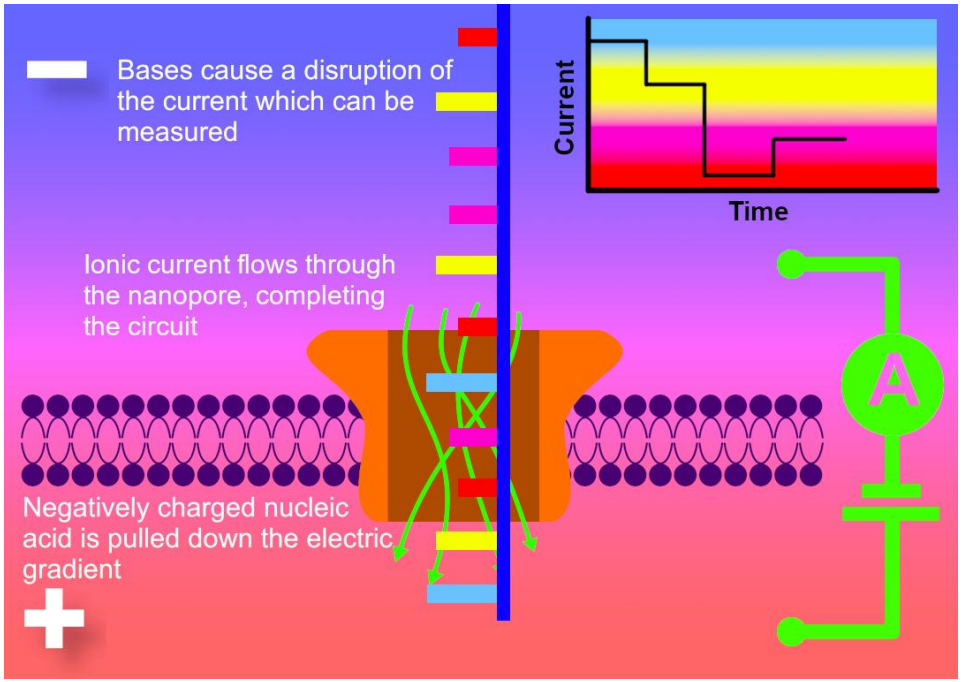
PacBio

https://www.youtube.com/watch?v=_ID8JyAbwEo

Oxford Nanopore sequencing

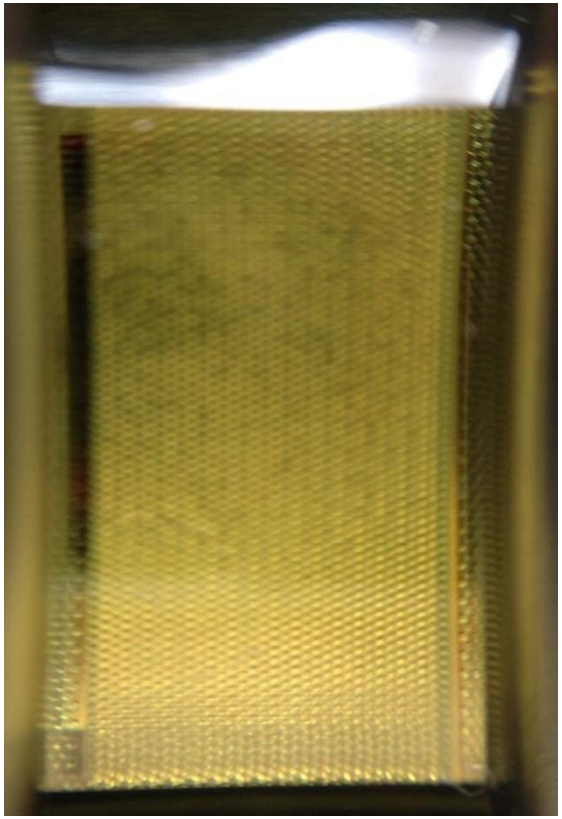


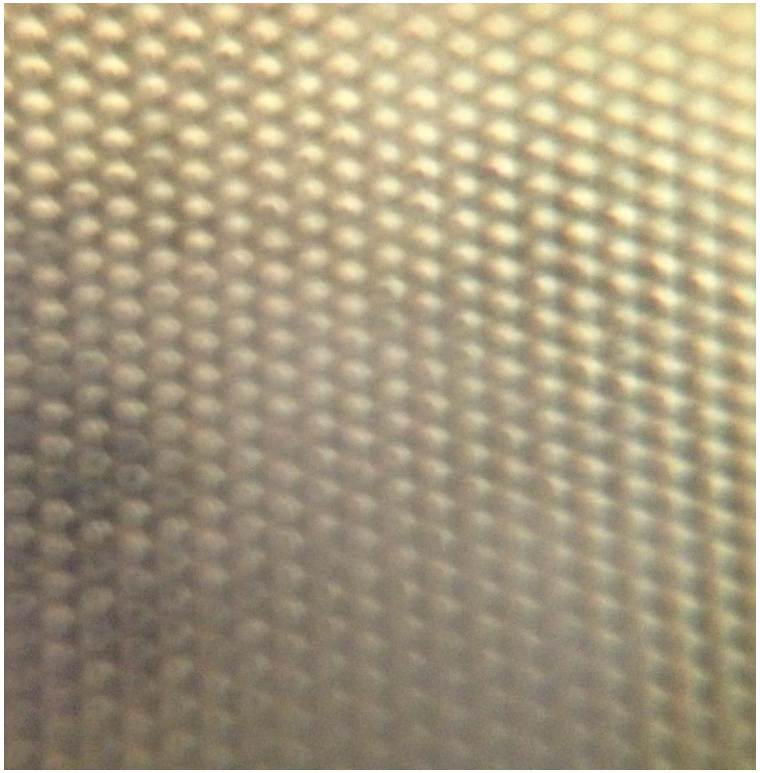




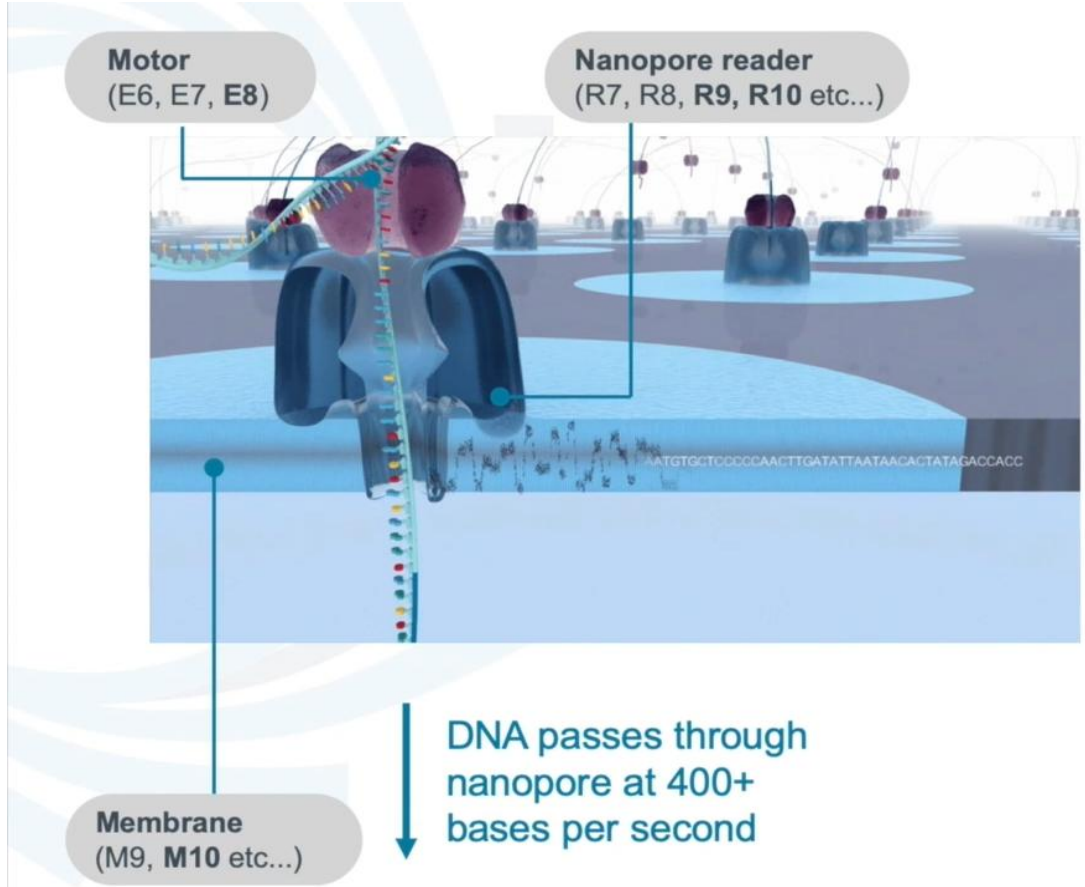
- Sequences native DNA and RNA molecules
- No template amplification; no immobilization
- Single strand moves through pore by potential difference: same principle of electrophoresis

Nanopore membrane

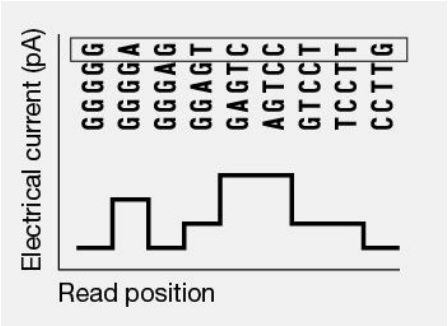




P2Solo – high throughput low scale



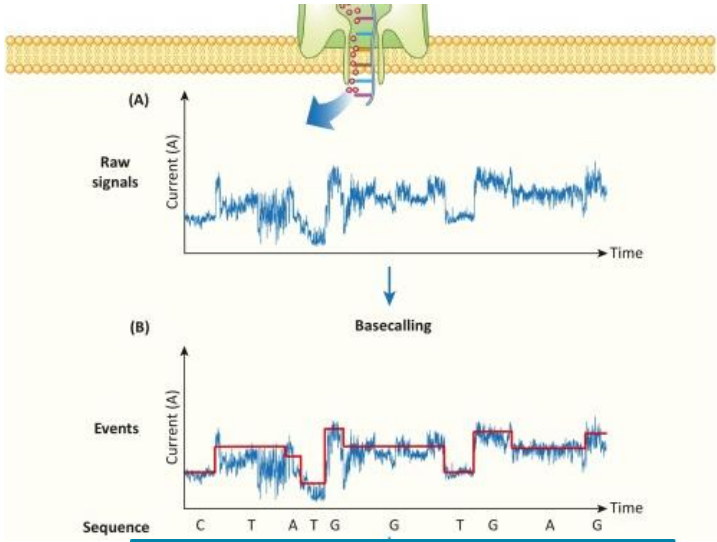
- Promethion flowcell
- Array of nanopores
- ~ 3000 channels, each with 4 pores
- Up to 12000 pores

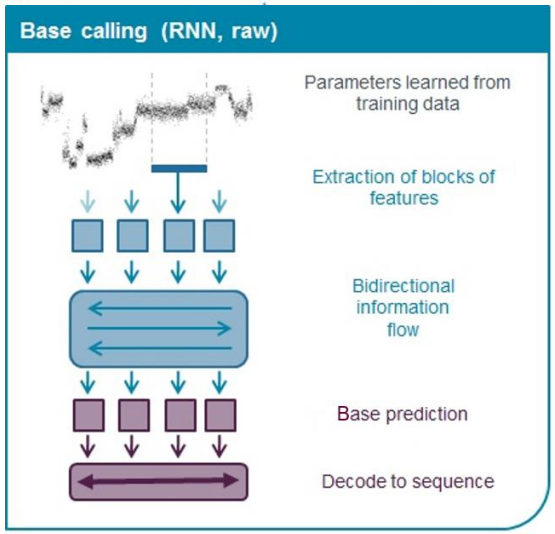


Signal decoding?

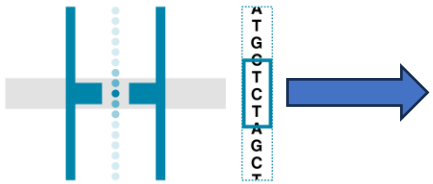
- Basecalling is the decoding of current changes to nucleotide sequences
- Basecaller is based on neural networks –

  bi-directional RNN – Recurrent Neural Network
- RNN keeps internal memory of previously seen data, and bi-directional can set a data in the context of what comes before and after the signal
- This aspect is continuously improved, increasing accuracy without needing new chemistries updates

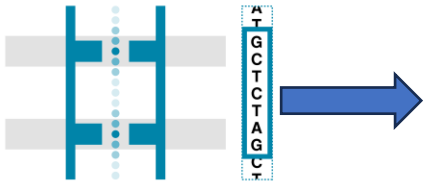




Improving on accuracy



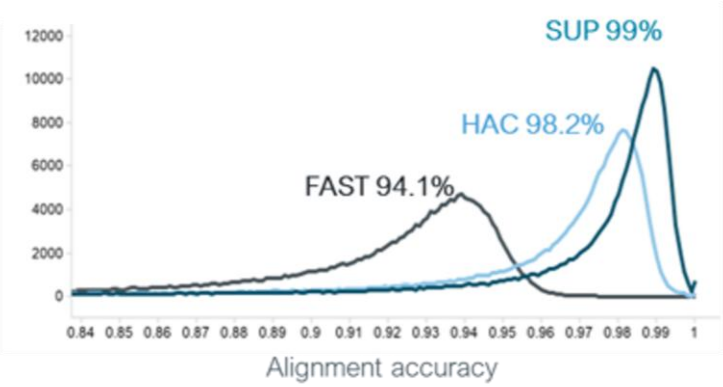
- R9 pores – single detection point



- R10 (current) – two readers, improved accuracy around homopolymer regions

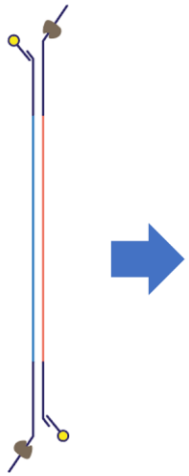
Basecalling models drastically change accuracy:

- **fast**: less computationally demanding, lowest accuracy
- **hac**: high accuracy, high computation demand
- **sup**: highest possible accuracy, highest computations demand

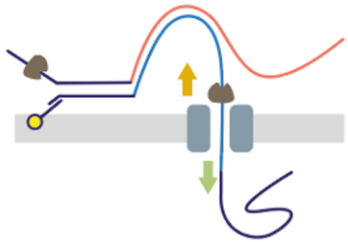


Duplex vs. Simplex reads

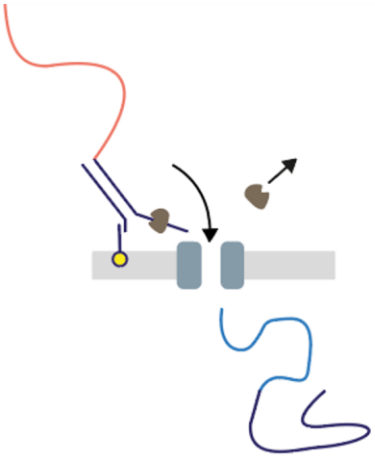
- Duplex method can increase mean read accuracy to >99%
- Low duplex rate in current chemistry version
- Optimizing duplex rate decreases throughput



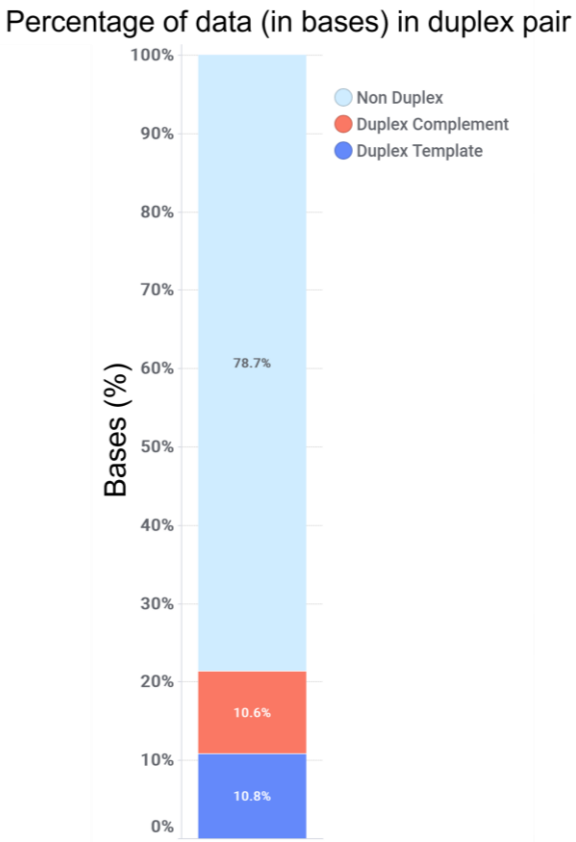
Linear dsDNA molecule adapted on both ends



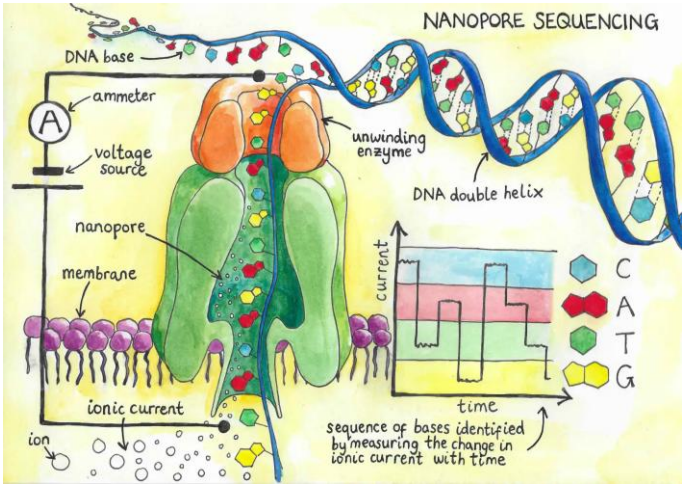
First strand sequenced



Second strand captured and sequenced immediately after



What is needed for sequencing



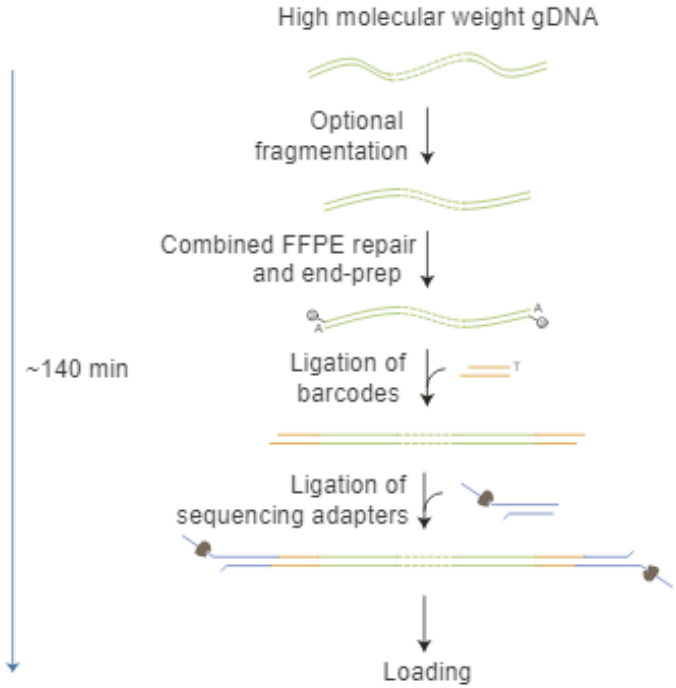
Minimal input required is 60ng/ul in volume 50ul

Ready for size selection? – SPRI AMPURE BEADS:

10-15kb should be the sweet spot

Library prep:

- native barcoding kit v14 (current version) 24 or 96 barcodes
- prep time 140min
- 400ng gDNA per sample for >4 barcodes
- free PCR method
- loading 10-20fmol library to the Flowcell



Considerations

Estimation of costs for metagenome sequencing:

- **PacBio HiFi: 2-3 samples at ~$2600 run -- ~$1000 per sample**
- **ONT Promethion: 10-20 samples at ~$1250 run -- up to $120 per sample**

	PacBio HiFi	ONT run1	ONT run2
Output QC reads (Gbp)	10.37	34.42	98.21
Largest assembled contig (bp)	5 835 850	5 368 224	7 010 576
Good quality MAGs	69	128	-

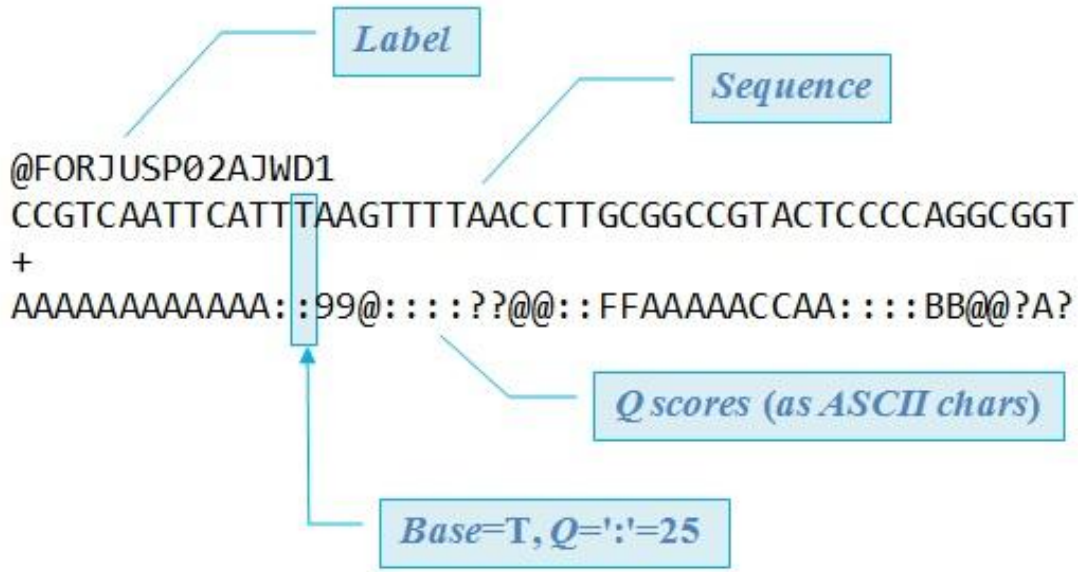
	Average length (bp)	Longest gene (bp)	Gene prediction rate
ONT MG assembly	602.4	15882	0.80
Illumina MG assembly	396	12621	0.85

Example of empirical results from ongoing analysis

Nanopore

https://www.youtube.com/watch?v=hs0FdiTHMbc

Output formates: Fastq files and phred score



ASCII_BASE=33 Illumina, Ion Torrent, PacBio and Sanger

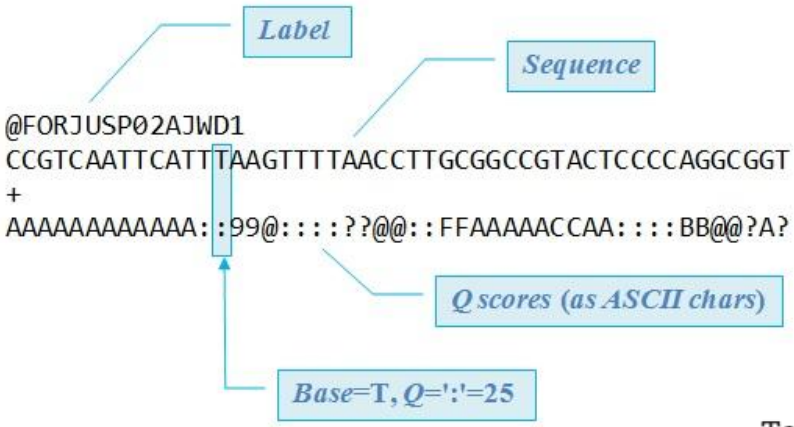
Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII
0	1.00000	33 !	11	0.07943	44 ,	22	0.00631	55 7	33	0.00050	66 B
1	0.79433	34 "	12	0.06310	45 -	23	0.00501	56 8	34	0.00040	67 C
2	0.63096	35 #	13	0.05012	46 .	24	0.00398	57 9	35	0.00032	68 D
3	0.50119	36 $	14	0.03981	47 /	25	0.00316	58 :	36	0.00025	69 E
4	0.39811	37 %	15	0.03162	48 0	26	0.00251	59 ;	37	0.00020	70 F
5	0.31623	38 &	16	0.02512	49 1	27	0.00200	60 <	38	0.00016	71 G
6	0.25119	39 '	17	0.01995	50 2	28	0.00158	61 =	39	0.00013	72 H
7	0.19953	40 (	18	0.01585	51 3	29	0.00126	62 >	40	0.00010	73 I
8	0.15849	41)	19	0.01259	52 4	30	0.00100	63 ?	41	0.00008	74 J
9	0.12589	42 *	20	0.01000	53 5	31	0.00079	64 @	42	0.00006	75 K
10	0.10000	43 +	21	0.00794	54 6	32	0.00063	65 A			

Output formates: Fastq files and phred score

$$Q_{illumina} = -10 \times \log_{10}\left(\frac{P_e}{1-P_e}\right),$$

where P_e is the probability of identifying a base incorrectly. For Sanger and other platforms, the formula is as follows [8]:

$$Q_{PHRED} = -10 \times \log_{10}(P_e).$$



base quality score for Illumina range from 0 to 40

$$Q_{illumina} = 10 \times \log_{10}\left(10^{\left\{\frac{Q_{PHRED}}{10}\right\}} + 1\right)$$

Table 2. Phred quality scores are logarithmically linked to error probabilities (http://en.wikipedia.org/wiki/Phred_quality_score)

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	99.90%
40	1 in 10 000	99.99%
50	1 in 100 000	99.999%
60	1 in 1 000 000	99.9999%

Fastq file header

@M02149:53:000000000-AANLH:1:1101:14924:1701 1:N:0:0
TACGGAGGGTGCAAGCGTTAATCGGAATCACTGGGCGTAAAGCGCACGTAGGCTGTCTGGTAA
GTCAGGGGTGAAATCCCGCGGCTCACCCGCGGAATTGCCCTTGATACTGCTGGACTTGAGTTC
GGGAGAGGGTGGCGGAATTCCAGGTGTAGGAGTGAAAGGCGTAGATAGCAGGAGGAACATC
AGGGGCGAAGGCGGCCACCTGGACCGATACTGACGCTGAGGTGCGAAAGCGTGGGGAGGAA
ACAGG
+
AAA??1>DDAAA11AFEGF00BGCEA0F1A1F10AAAFA//BAAA/AAB00ABGFF@F10BB@DG
G2B00/B//1@BF1F/>>>EEA<1B</<>///?F?DD<FGF>??<F1<F<??<FGHF?G<?CHHHHHFF
<::/0GHFB;:BFF0F;<1GG>BF2HHEB//?F@HGB@B110FFHFHGB1B0FB>/EE>HGFEEAA0/
1A011EEBA/2D2D/AEEABB1FHE00AAGFFEA1A1GGFFFB3@F>1AAA

Illumina header
@<instrument>:<run number>:<flowcell ID>:<lane>:<tile>:<x-pos>:<y-pos> <read>:<is filtered>:<control number>:<sample number>

http://support.illumina.com/content.htm

Fasta file

>M02149:53:000000000-AANLH:1:1101:14924:1701 1:N:0:0
TACGGAGGGTGCAAGCGTTAATCGGAATCACTGGGCGTAAAGCGCAC
GTAGGCTGTCTGGTAAGTCAGGGGTGAAATCCCGCGGCTCACCCGCG
GAATTGCCCTTGATACTGCTGGACTTGAGTTCGGGAGAGGGTGGCGG
AATTCCAGGTGTAGGAGTGAAAGGCGTAGATAGCAGGAGGAACATCA
GGGGCGAAGGCGGCCACCTGGACCGATACTGACGCTGAGGTGCGAA
AGCGTGGGGAGGAAACAGG

Quality check

FastQC Report

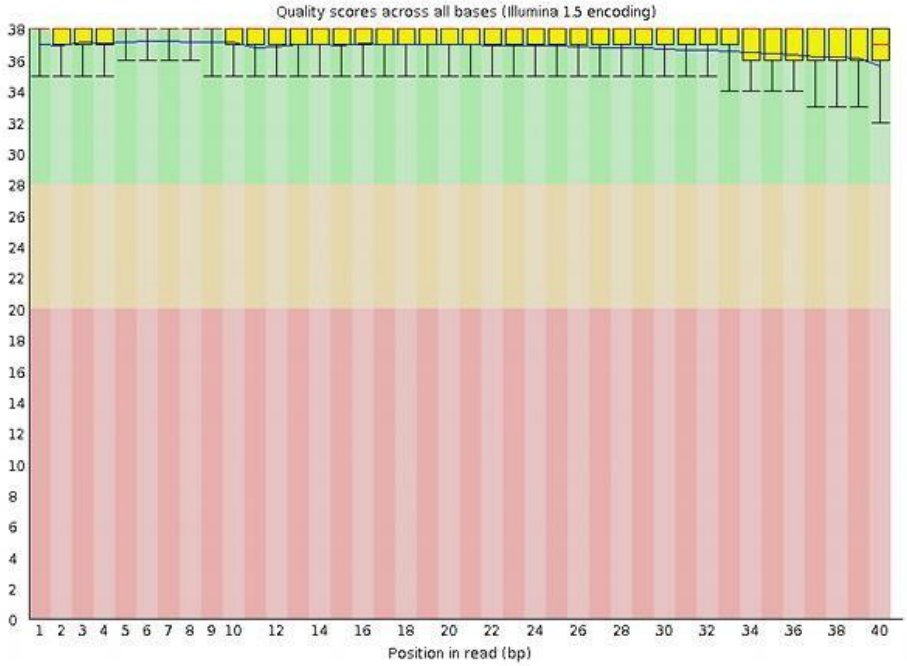
Summary

- Basic Statistics
- Per base sequence quality
- Per sequence quality scores
- Per base sequence content
- Per base GC content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Kmer Content

Basic Statistics

Measure	Value
Filename	good_sequence_short.fastq
File type	Conventional base calls
Encoding	Illumina 1.5
Total Sequences	250000
Filtered Sequences	0
Sequence length	40
%GC	45

Per base sequence quality



“Reality” check