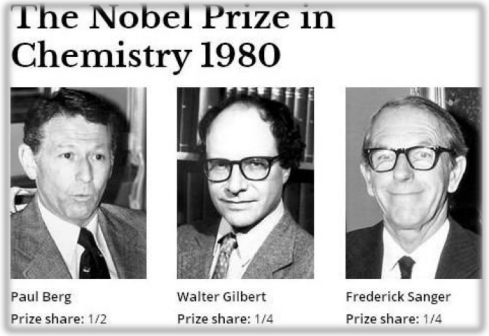
Sequencing technologies for metagenomics

Espen Åberg

Marine Metagenomics Workshop
26-30 November 2018- Tromsø

www.elixir-europe.org

Early metagenomic sequencing

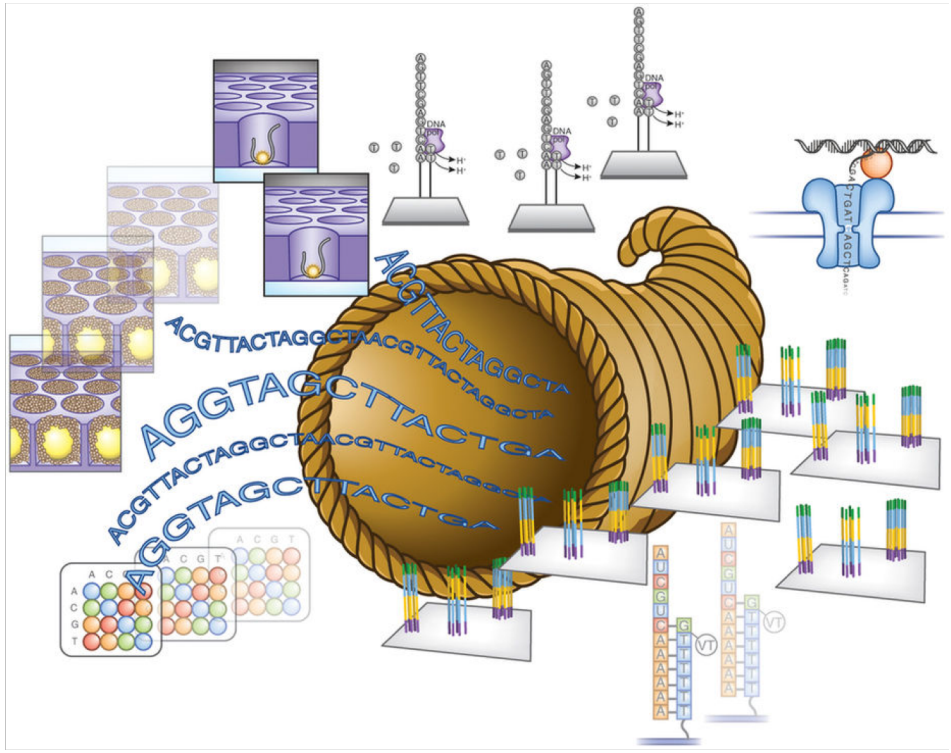


- Pioneering metagenomic studies used the Sanger platform
  - i.e Venter, J.C. et al. Environmental genome shotgun sequencing of the Sargasso Sea. Science 304, 66–74 (2004).
    - 1800 genomic species , 148 novel bacterial phylotypes

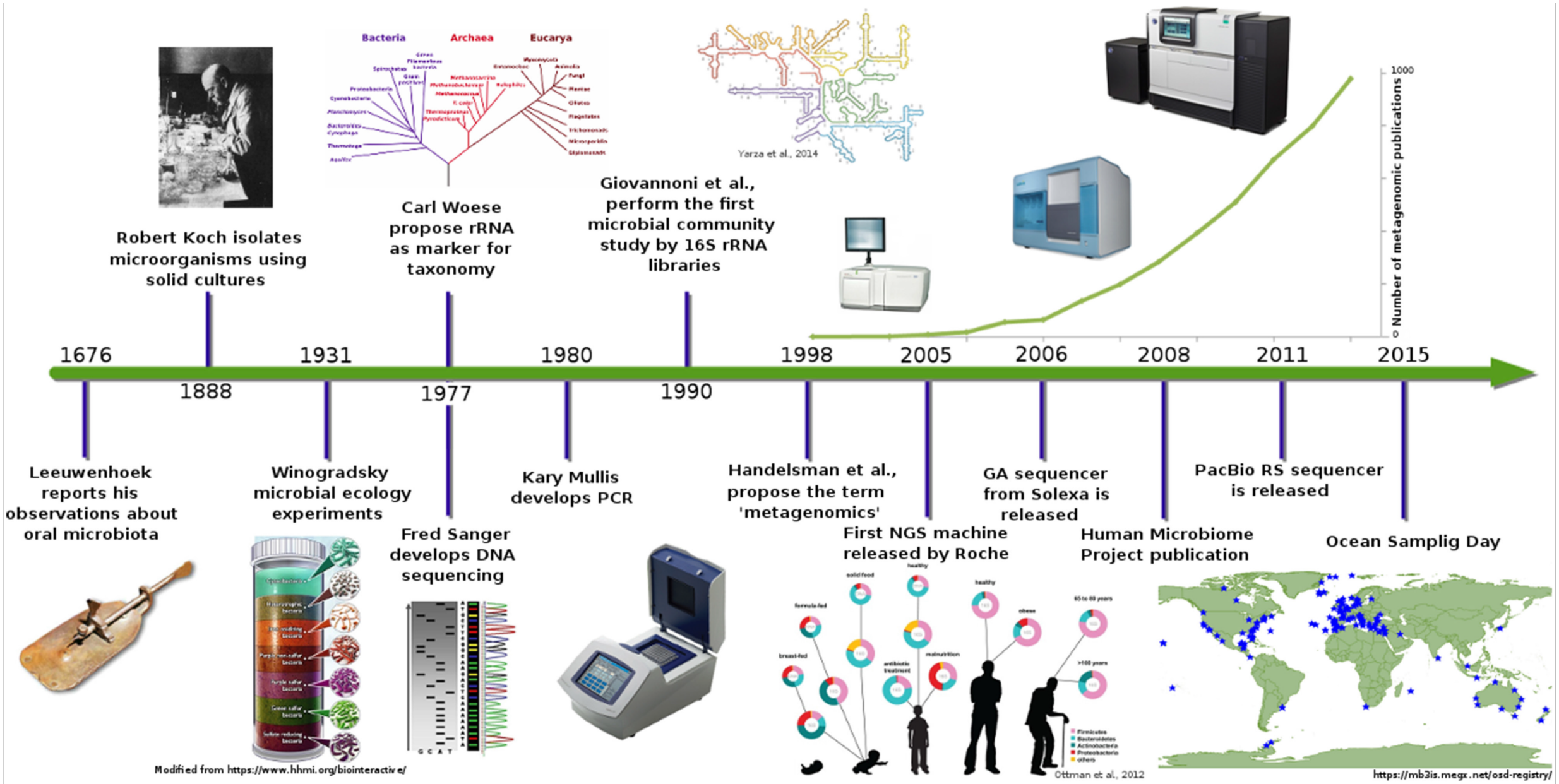
- This technology can not provide sufficient read depth to saturate moderately diverse communities
  - Sanger-based metagenomic projects are often limited to:
    - Fosmid or bacterial artificial chromosome libraries
    - low-diversity microbial communities.

Next-generation sequencing (NGS)

- Overcome several of the disadvantages of Sanger sequencing
  1. Substantially higher throughput
  2. Cheaper cost per base sequencing
  3. Simpler library preparation
  4. No cloning step
  5. Real time

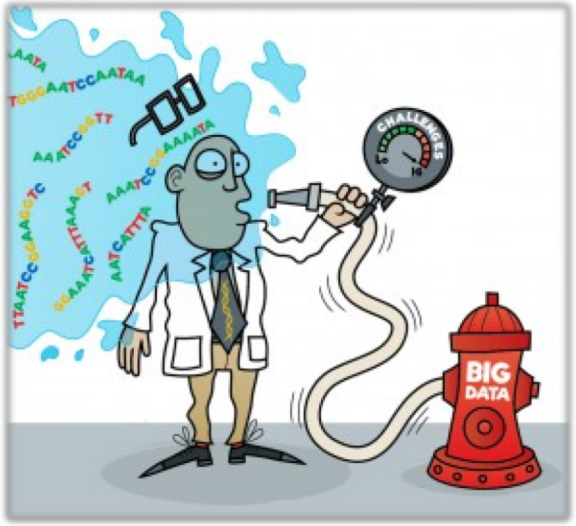


Metagenomics: a dominant contributors to sequence databases

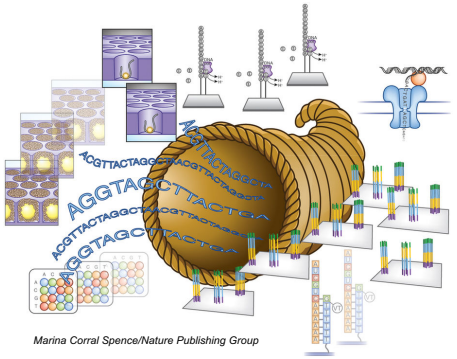


Next-generation sequencing (NGS)

- Not without new challenges…
  - Each new technology has a different error model and biases that need to be considered during experimental design and sequence analysis
  - Errors that occur in the output sequence on NGS
    - Indels (insertion/deletion)
    - Base substitutions
  - Increased coverage can overcome errors but absolute number of sequencing errors will increase with coverage

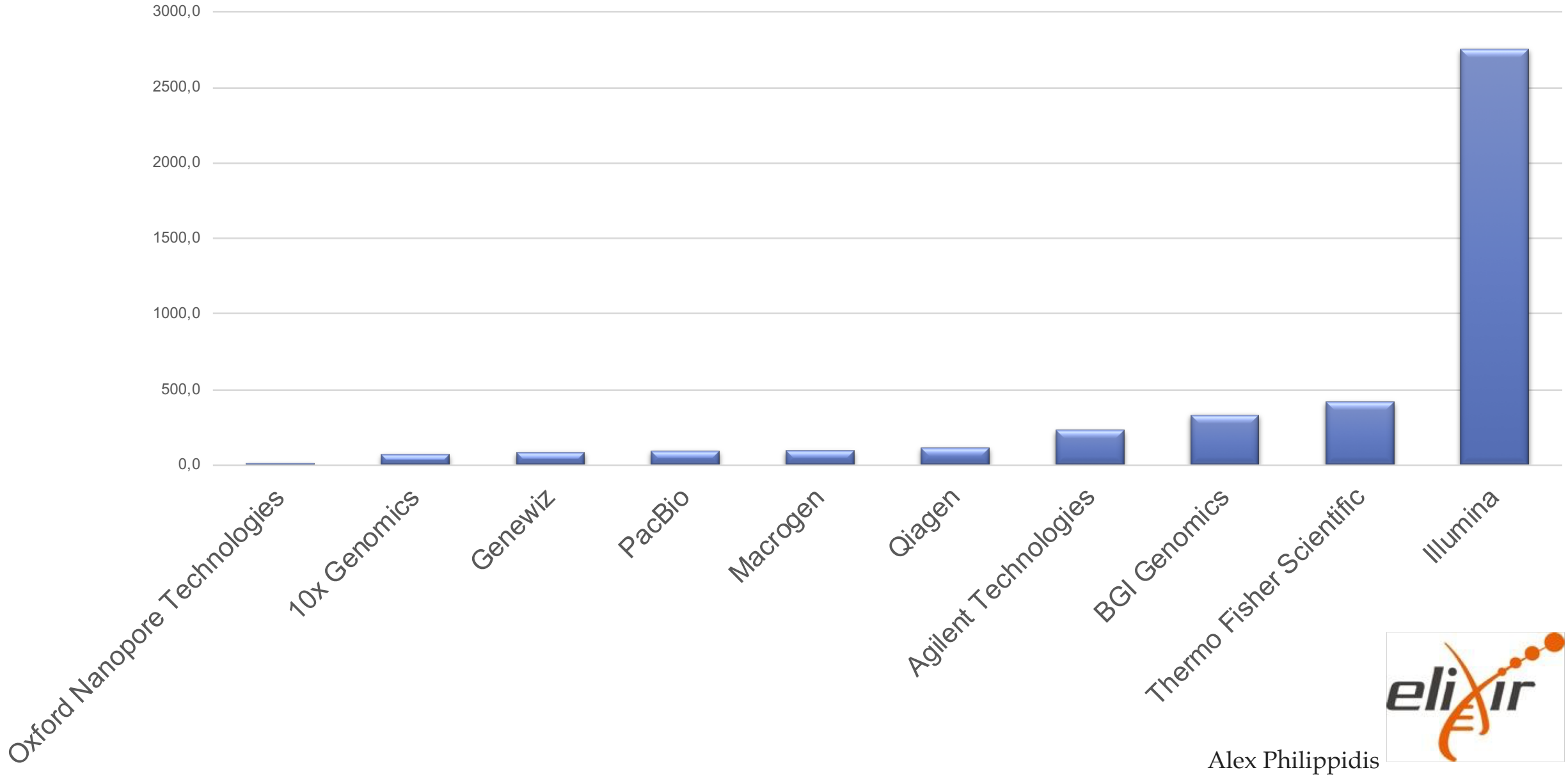


NGS?



- Next Generation Sequencing
  - **Second** Generation Sequencing
    - Sequencing By Synthesis
      - Pyrosequencing
        - 454
      - Sequencing by reversible terminator
        - Illumina
      - Sequencing by Detection of Hydrogen Ion
        - Ion Torrent
    - Sequencing By Ligation
      - Sequencing by hybridisation and ligation
        - SOLiD
  - **Third** Generation Sequencing
    - Single Molecule real-time long reads
      - Pacific Biosciences Single molecule real time (SMRT) sequencing
      - Oxford Nanopore
    - Synthetic long reads
      - Illumina Synthetic Long-Read
      - 10X Genomics
    - Sequencing By Ligation
      - Sequencing by hybridisation and ligation
        - Complete Genomics (BGI)

Top 10 Sequencing Companies - 2017 Revenues (mill $)



Alex Philippidis

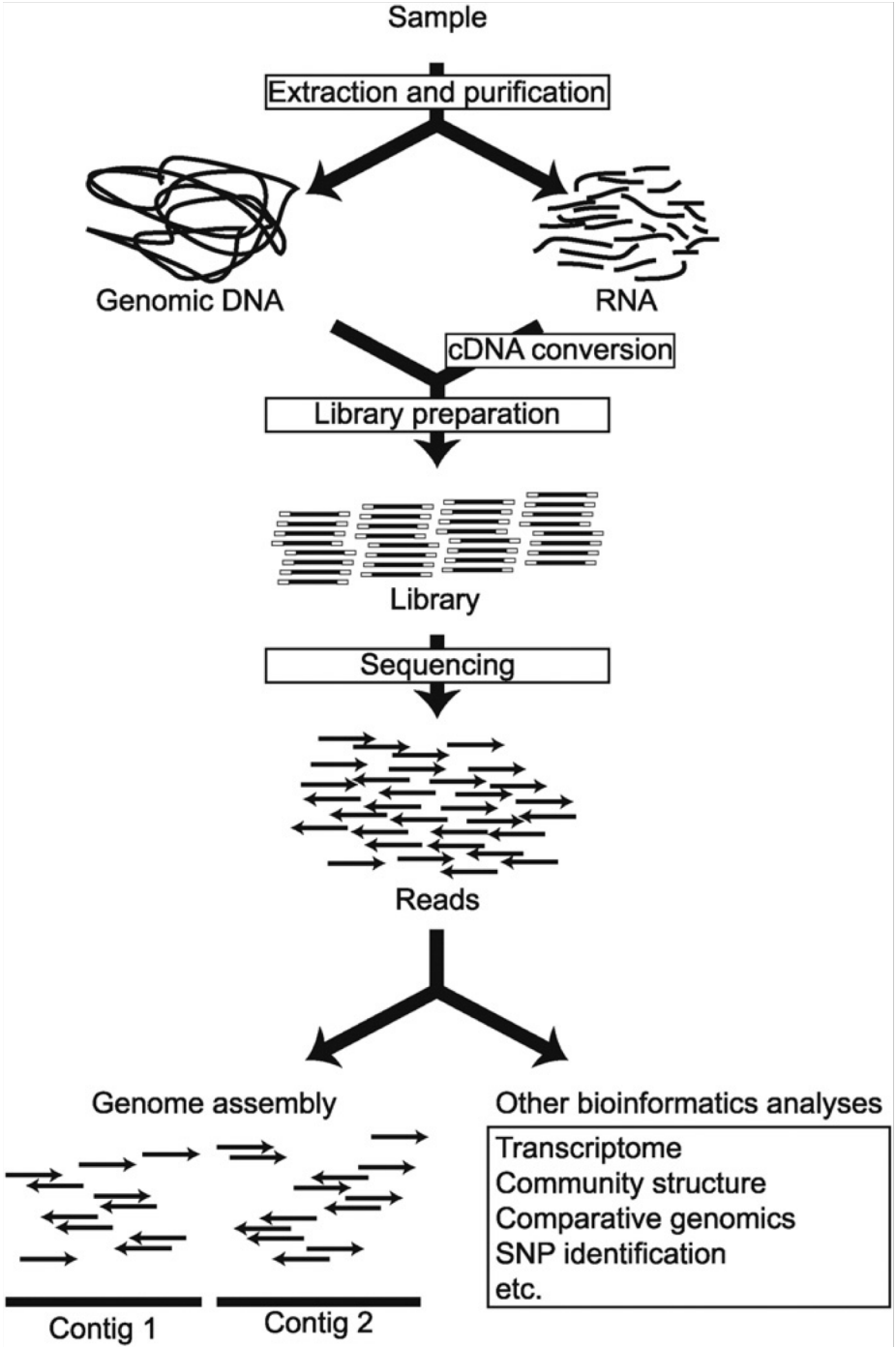
Illumina

- Market leader
  - Latest addition Novaseq 6000
  - S4 chemistry, 500+ Gb of data per lane
    - 100$ genomes?
  - iSeq 100 (benchtop sequencer)

- Long-read sequencing market?

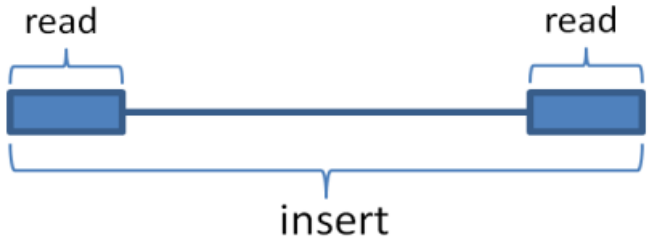
General NGS Principle

- Sequence a large number of DNA fragments (thousands to millions) in parallel in a single machine run
- Possible downstream analyses depends on:
  - The way libraries are prepared
  - Choice of the sequencing instrument and associated technology



Vincent AT, et al; J Microbiol Methods; vol138:p60-71 (2017)

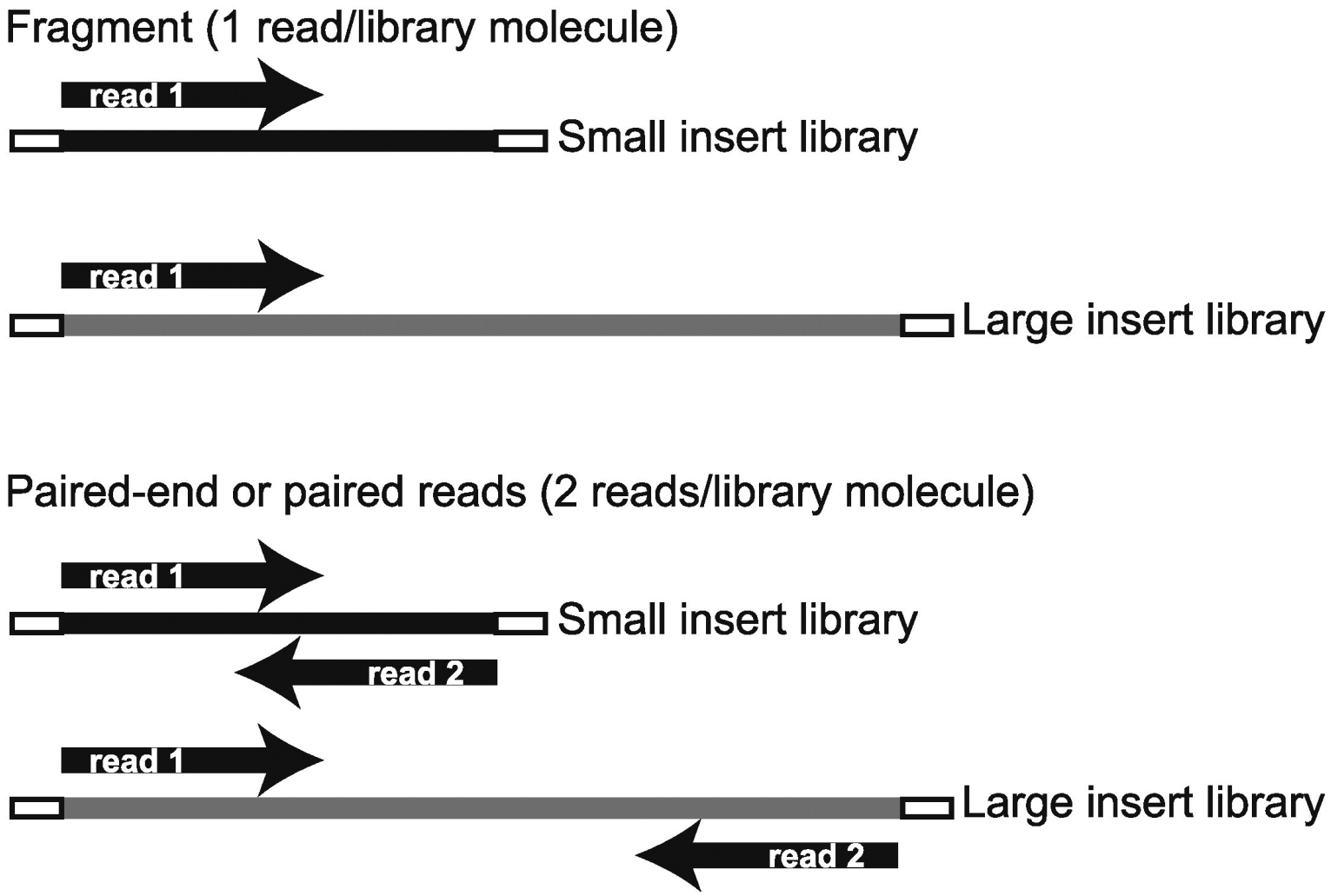
Basic concepts



Insert: The DNA fragment that is used for sequencing.

Read: The part of the insert that is sequenced.

Single-end or Paired-end reads…



…Determines:

- Sequencer choice
- How the libraries are produced

Vincent AT, et al; J Microbiol Methods; vol138:p60-71 (2017)

Sequence read assembly

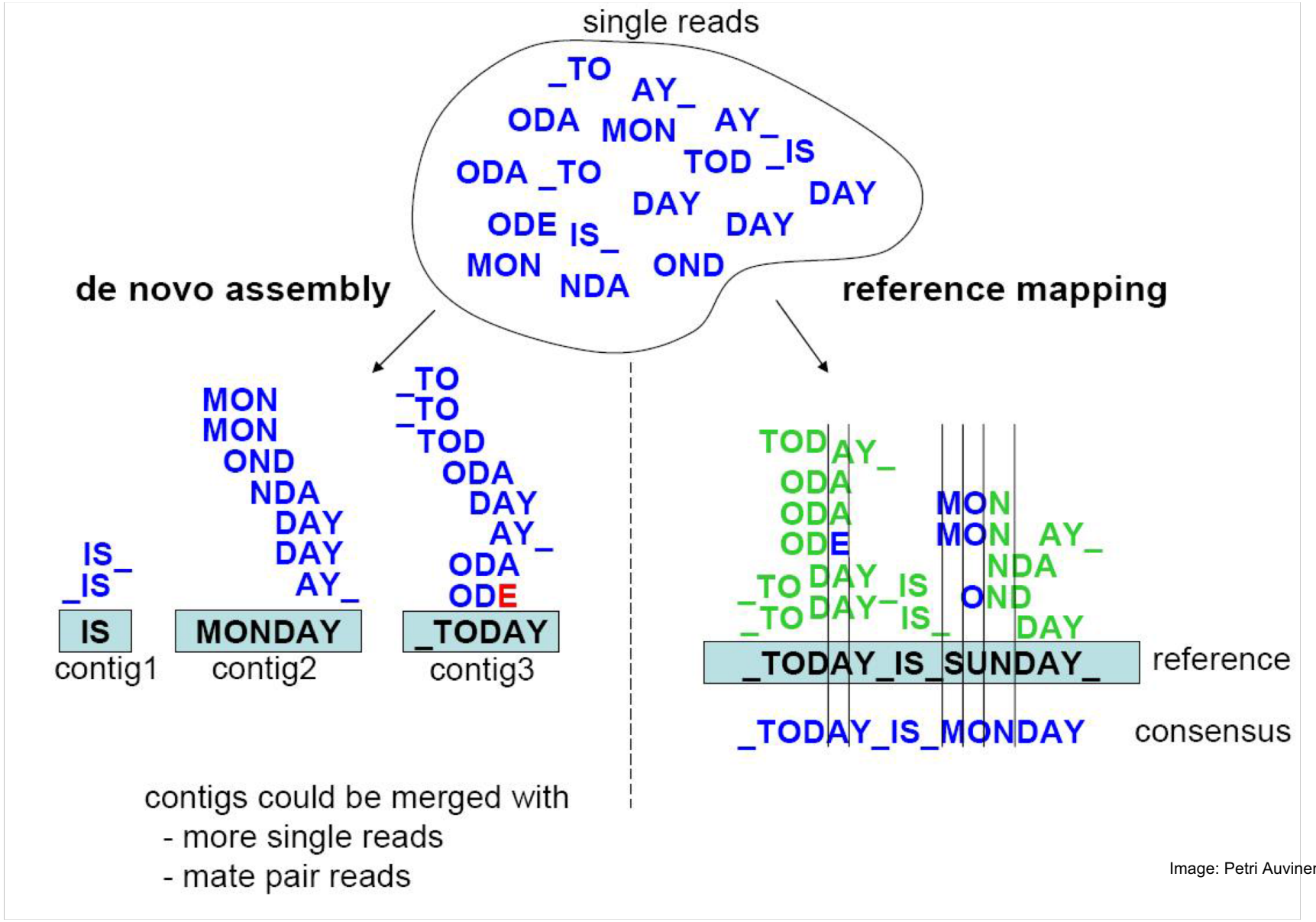
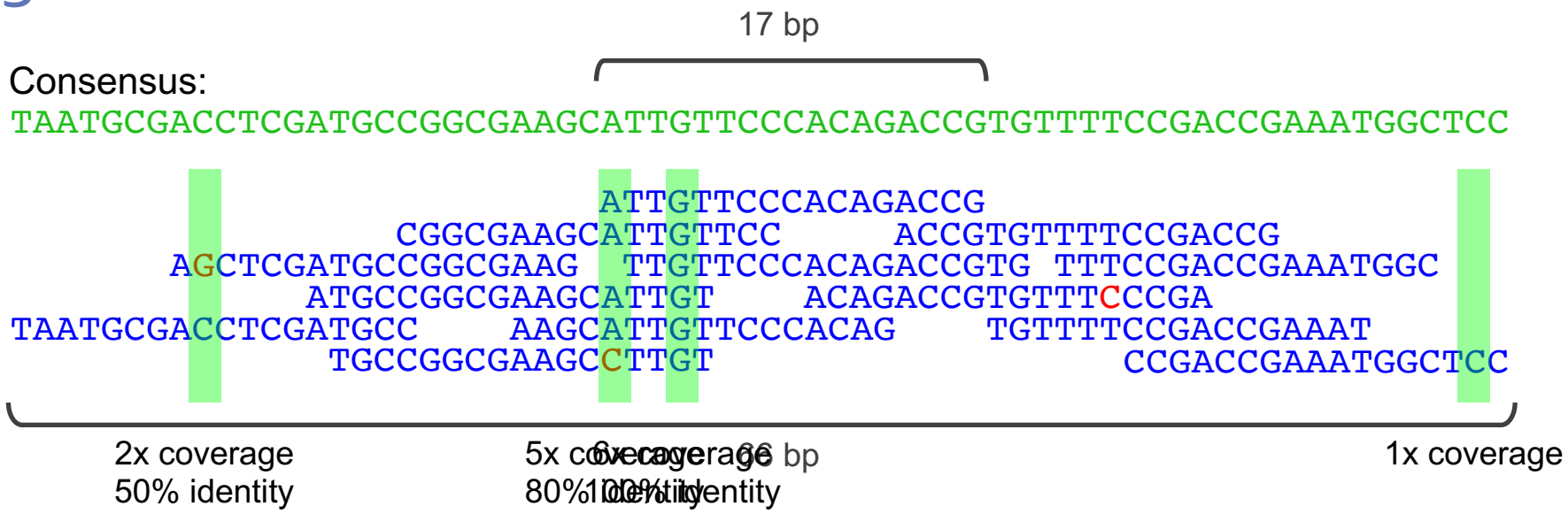


Image: Petri Auvinen

Coverage



Coverage: # of reads underlying the consensus

General overview of NGS library construction

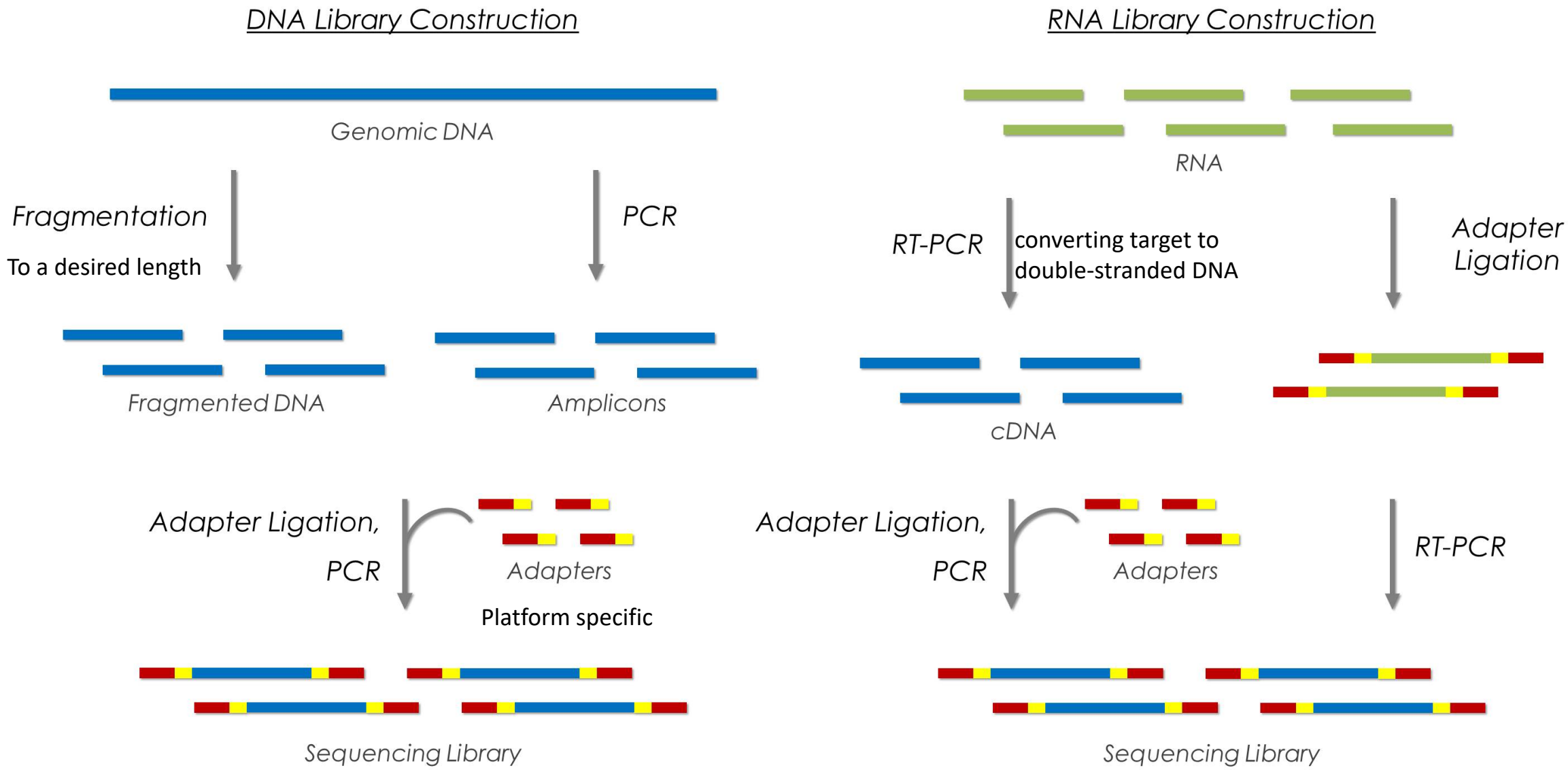
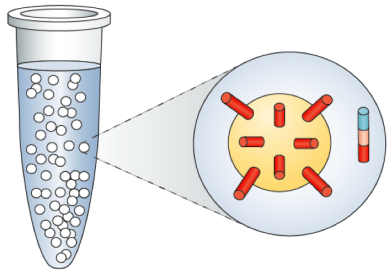


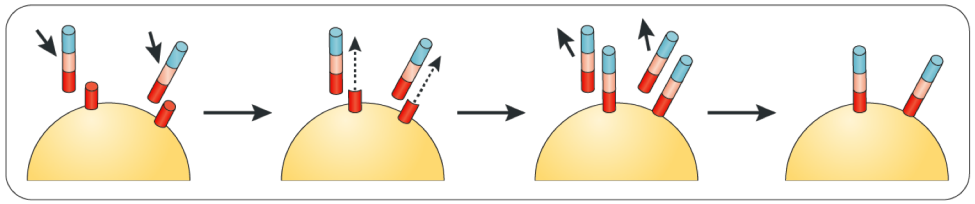
Image: Biocompare

Template amplification strategies

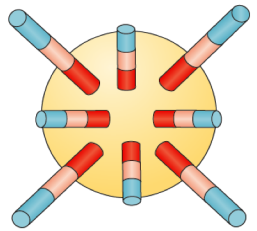
Emulsion PCR
(454 (Roche), SOLiD (Thermo Fisher), GeneReader (Qiagen), Ion Torrent (Thermo Fisher))



Emulsion
Micelle droplets are loaded with primer, template, dNTPs and polymerase

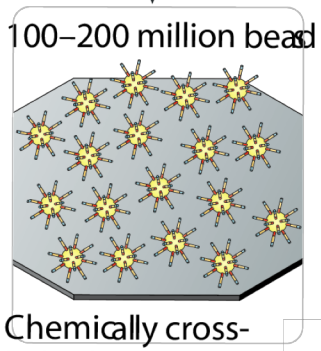


On-bead amplification
Templates hybridize to bead-bound primers and are amplified; after amplification, the complement strand disassociates, leaving bead-bound ssDNA templates



Final product
100–200 million beads with thousands of bound template

PCR is carried out within the micelle

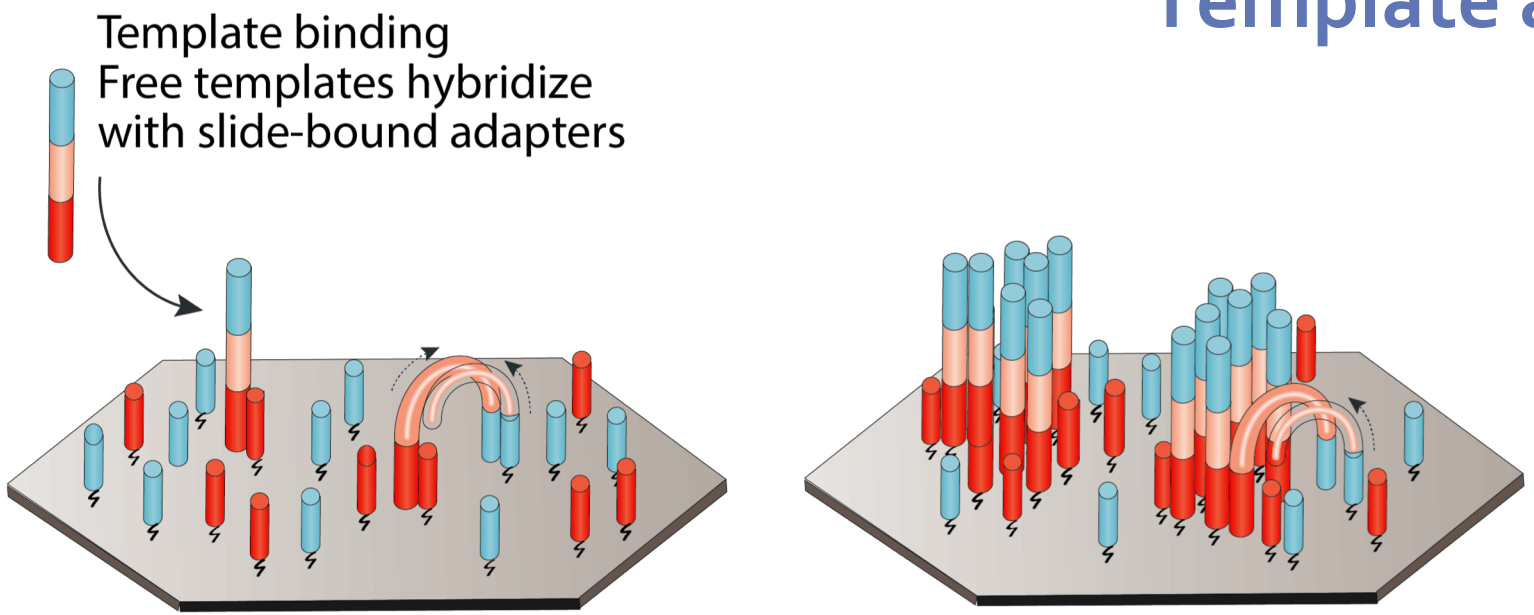


Chemically cross-linked to a glass slide

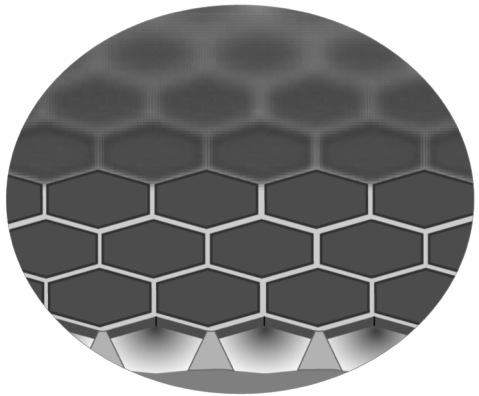
Goodwin S, et al; Nature Reviews Genetics vol17,p333–351(2016)

Template amplification strategies

b Solid-phase bridge amplification (Illumina)



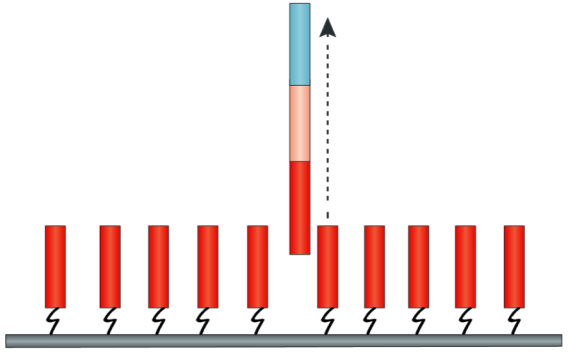
Patterned flow cell
Microwells on flow cell direct cluster generation, increasing cluster density



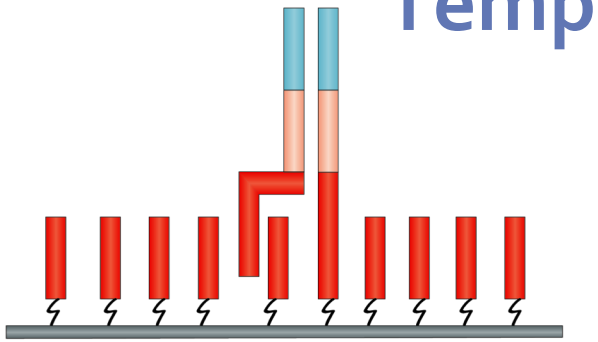
Goodwin S, et al; Nature Reviews Genetics vol17,p333–351(2016)

Template amplification strategies

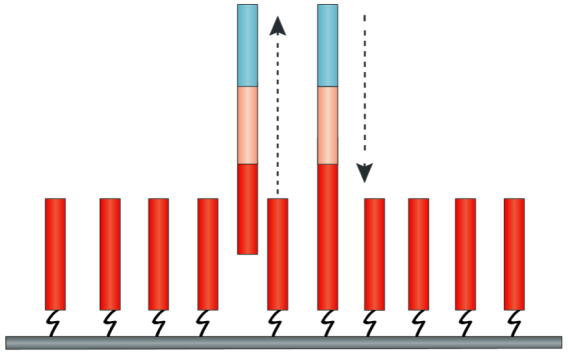
c Solid-phase template walking
(SOLiD Wildfire (Thermo Fisher))



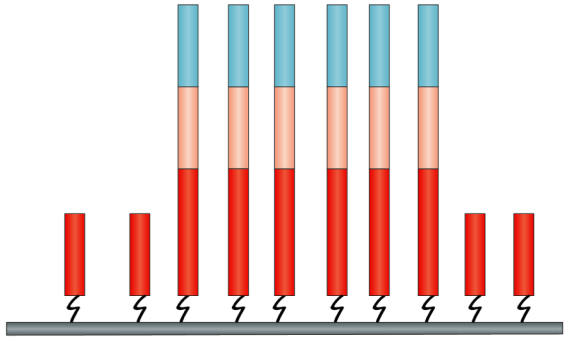
Template binding
Free DNA templates hybridize to bound primers and the second strand is amplified



Primer walking
dsDNA is partially denatured, allowing the free end to hybridize to a nearby primer



Template regeneration
Bound template is amplified to regenerate free DNA templates

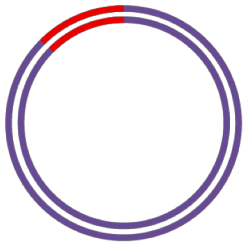


Cluster generation
After several cycles of amplification, clusters on a patterned flow cell are generated

Goodwin S, et al; Nature Reviews Genetics vol17,p333–351(2016)

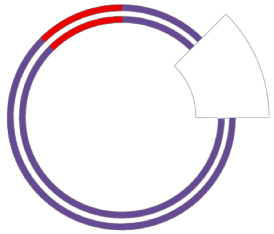
Template amplification strategies

d In-solution DNA nanoball generation (Complete Genomics (BGI))



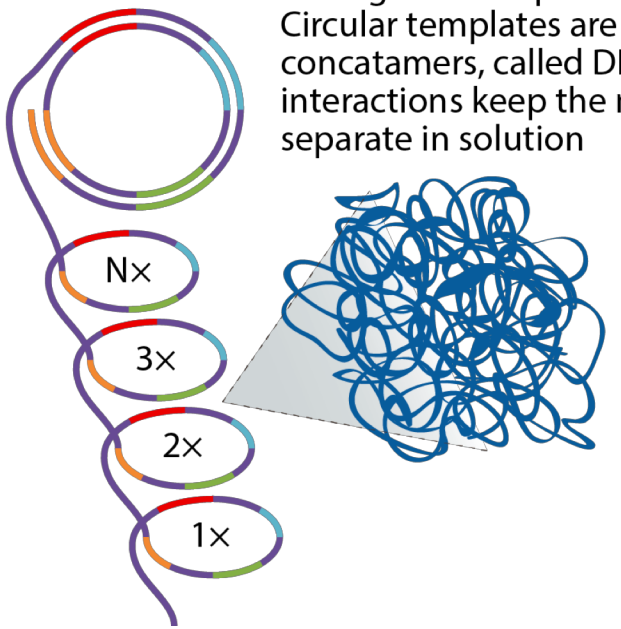
Adapter ligation
One set of adapters is ligated to either end of a DNA template, followed by template circularization

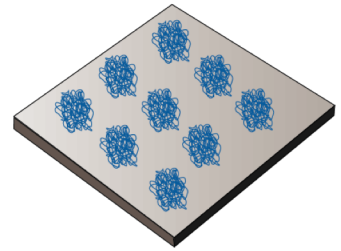
Cleavage
Circular DNA templates are cleaved downstream of the adapter sequence



Iterative ligation
Three additional rounds of ligation, circularization and cleavage generate a circular template with four different adapters

Rolling circle amplification
Circular templates are amplified to generated long concatamers, called DNA nanoballs; intermolecular interactions keep the nanoballs cohesive and separate in solution

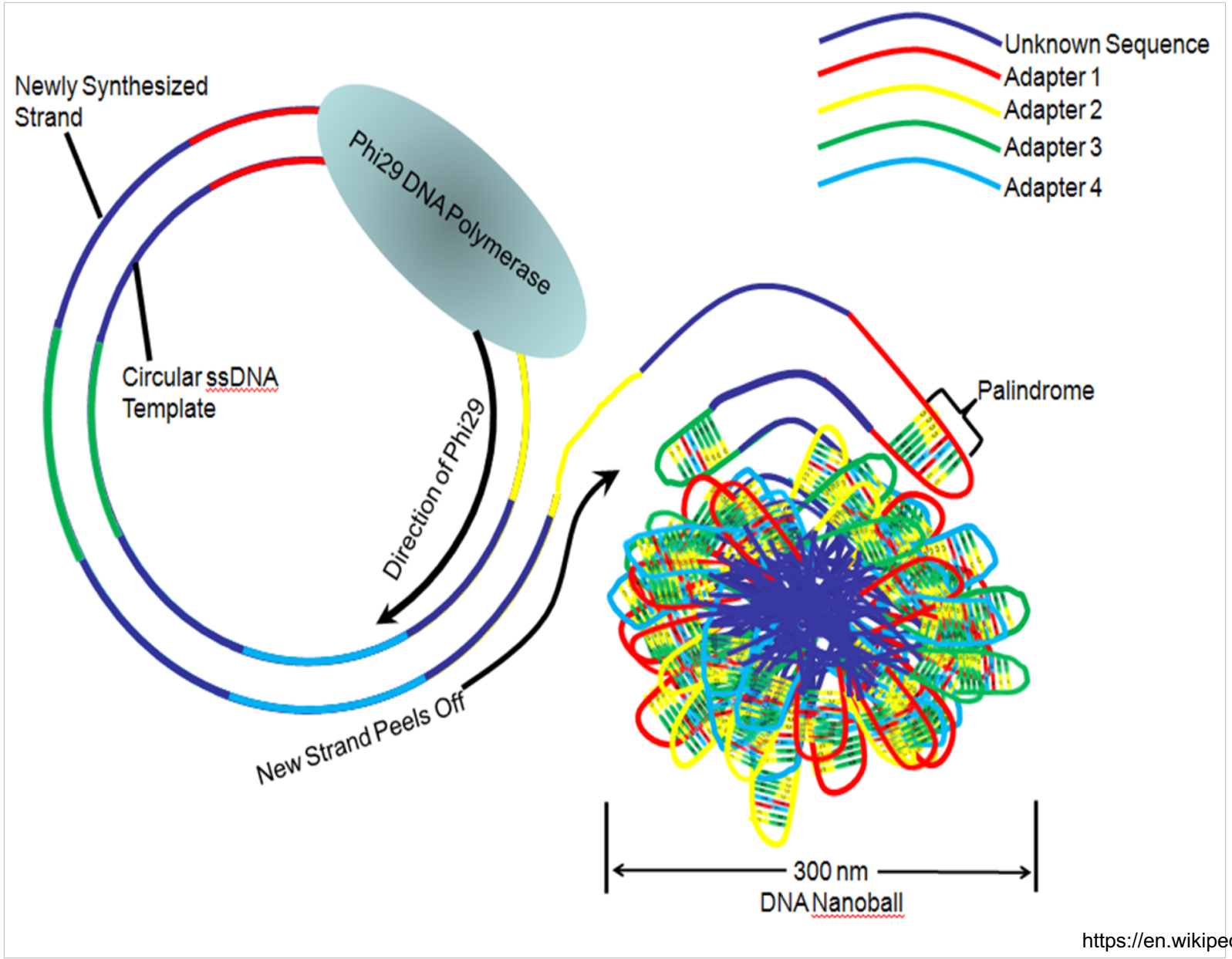




Hybridization
DNA nanoballs are immobilized on a patterned flow cell

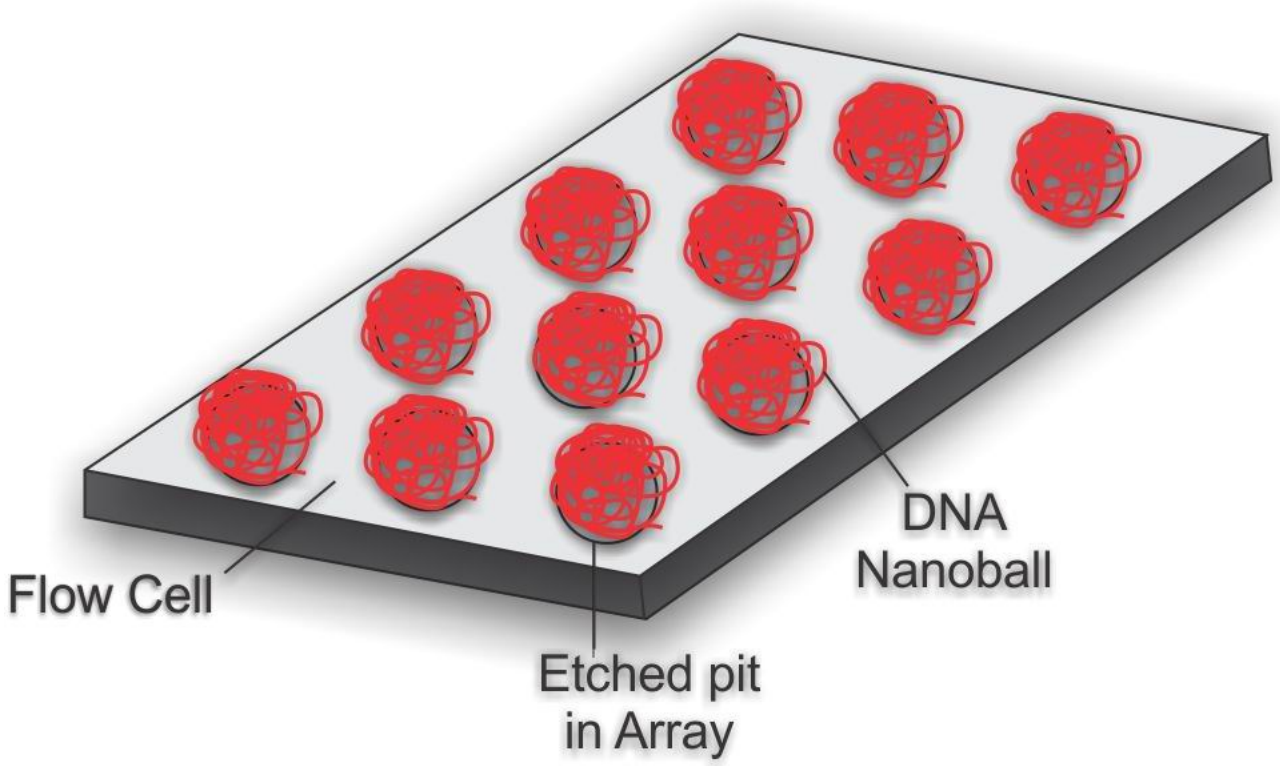
Goodwin S, et al; Nature Reviews Genetics vol17,p333–351(2016)

Template amplification strategies

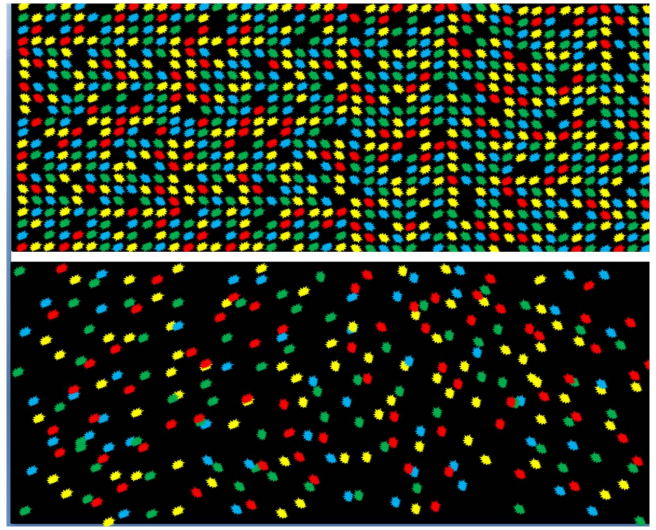


https://en.wikipedia.org/wiki/DNA_nanoball_sequencing

Template amplification strategies



high density of sequencing reads maximizing the number of reads per flow cell



https://en.wikipedia.org/wiki/DNA_nanoball_sequencing

Bejing Genomic Institute (BGI)

- Biggest sequencing centre on earth.
- Short-read sequencing platform, the **BGISEQ-500,MGI-200, MGI-2000**
- An initial study suggests it may produce data of a comparable quality to Illumina (Mak *et al.* 2017).

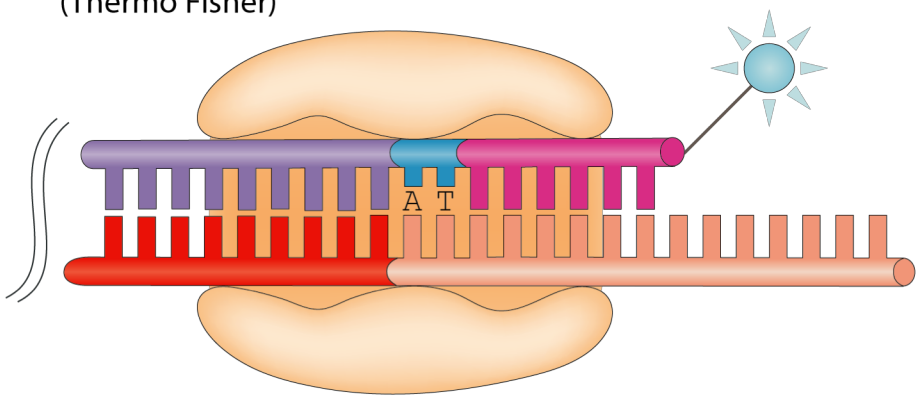
NGS Sequencing strategies

- Sequencing platforms can collect information from many millions of reaction centres at the same time

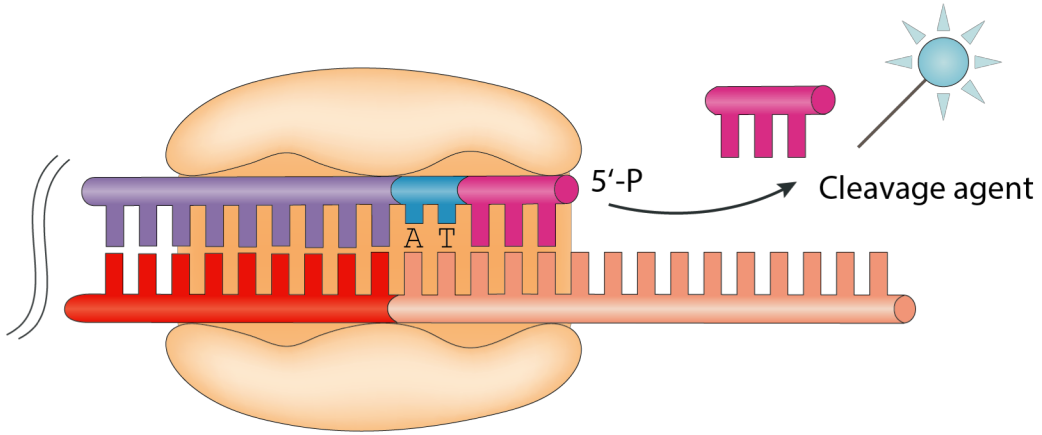
- Two broad categories
  - Sequencing by ligation (SBL)
  - Sequencing by synthesis (SBS)

Sequencing by ligation methods.

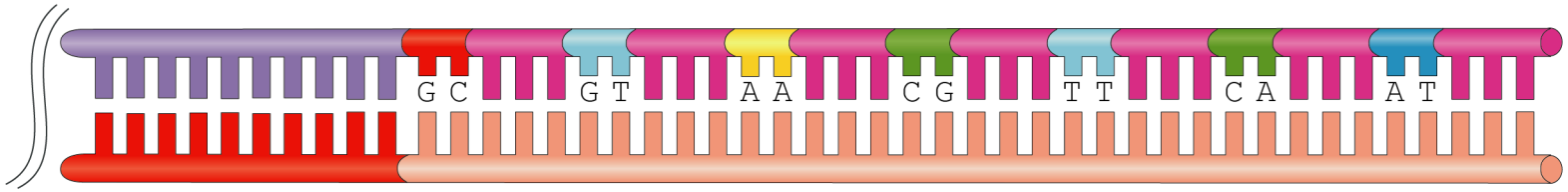
a SOLiD
(Thermo Fisher)



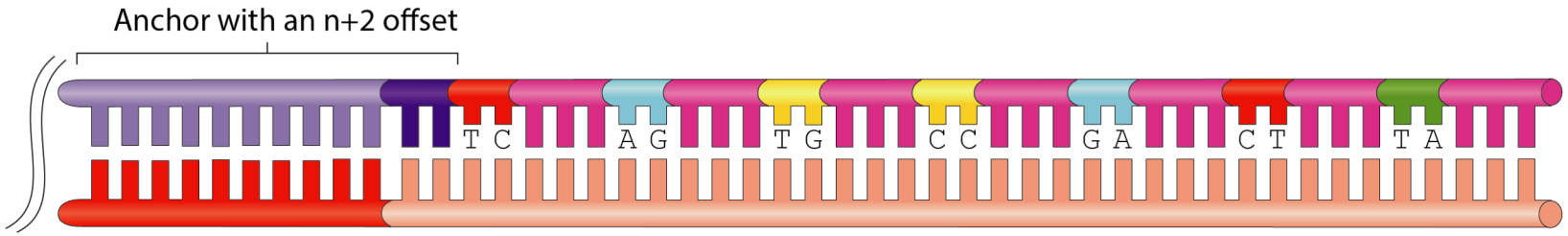
Two-base-encoded probes
Probes with two known bases followed by degenerate or universal bases hybridize to a template; ligase immobilizes the complex and the slide is imaged



Cleavage
The fluorophore is cleaved from the probe along with several bases, revealing a 5' phosphate

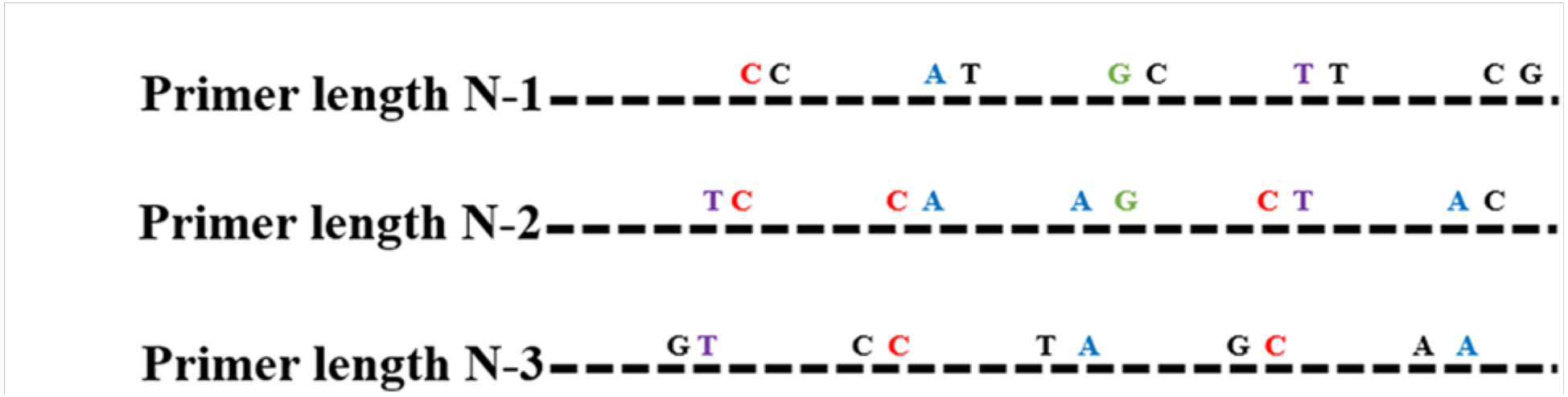


Probe extension
10 rounds of hybridization, ligation, imaging and cleavage identify 2 out of every 5 bases

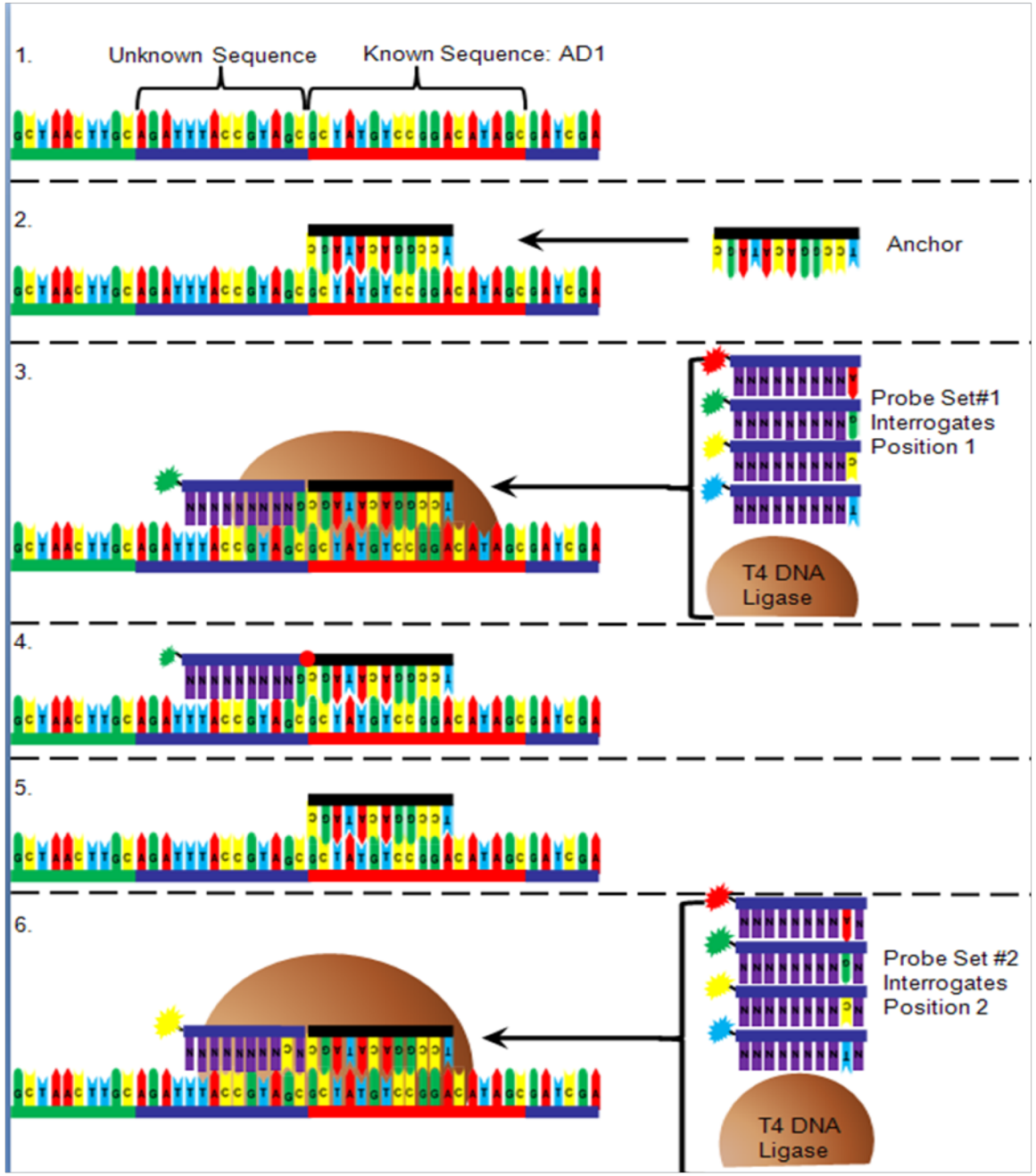


Reset
After a round of probe extension, all probes and anchors are removed and the cycle begins again with an offset anchor

Nature Reviews | Genetics
www.nature.com/nrg



Ambardar S. Et al, Indian J Microbiol (Oct–Dec 2016) 56(4):394–404



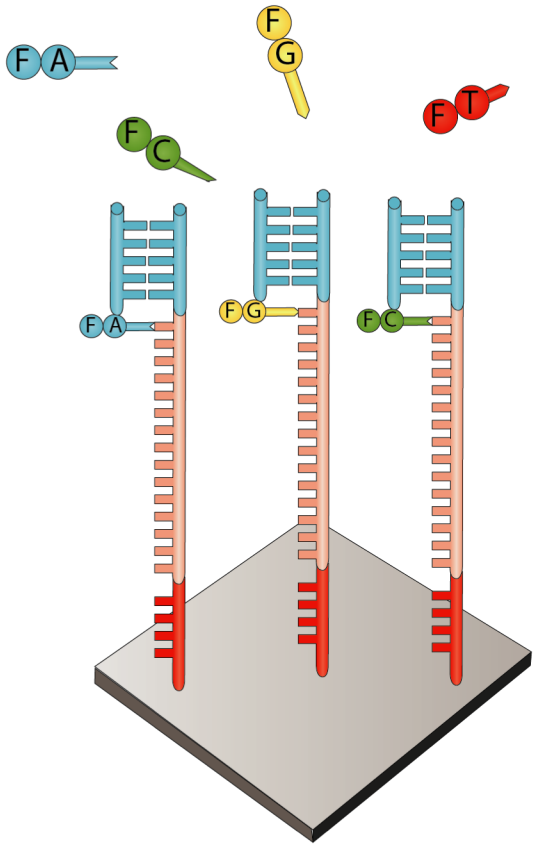
Sequencing by ligation methods

Complete Genomics - combinatorial probe–anchor ligation (cPAL) approach

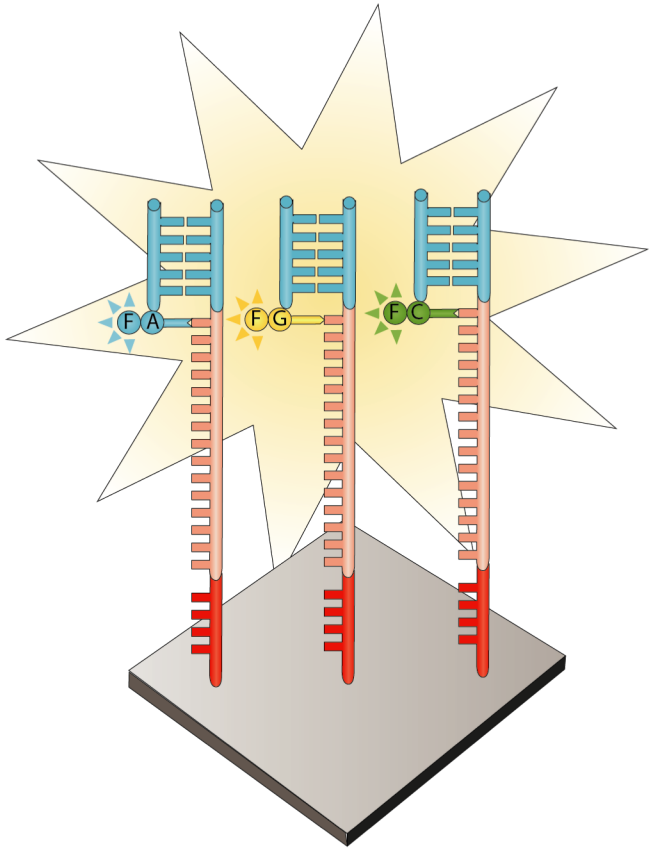
Short reads - bad for repeated elements
PCR bias due to many rounds of PCR

Sequencing by synthesis: cyclic reversible termination

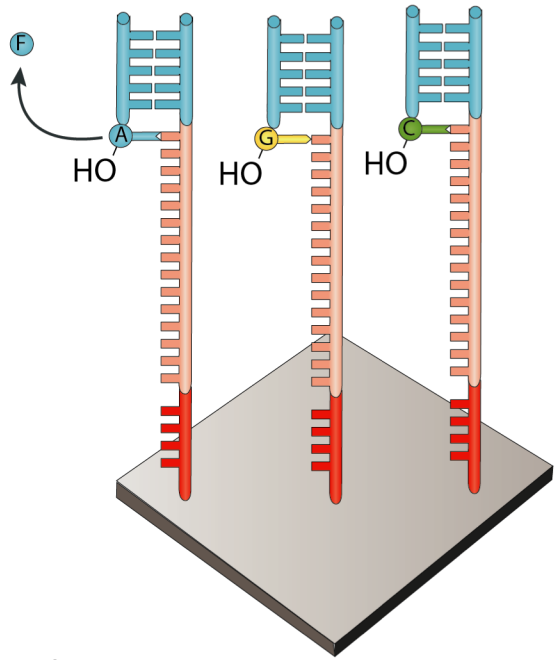
Illumina



Nucleotide addition
Fluorophore-labelled, terminally blocked nucleotides hybridize to complementary base. Each cluster on a slide can incorporate a different base.



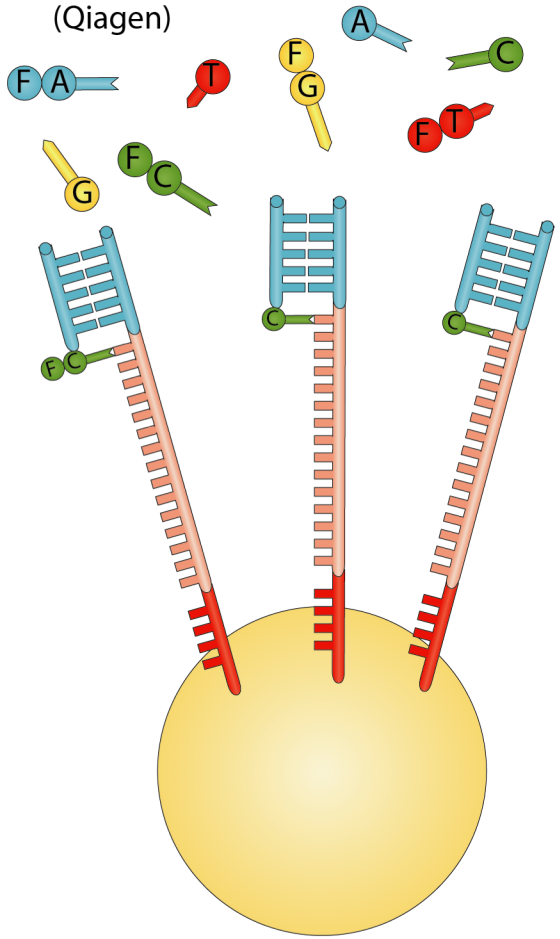
Imaging
Slides are imaged with either two or four laser channels. Each cluster emits a colour corresponding to the base incorporated during this cycle.



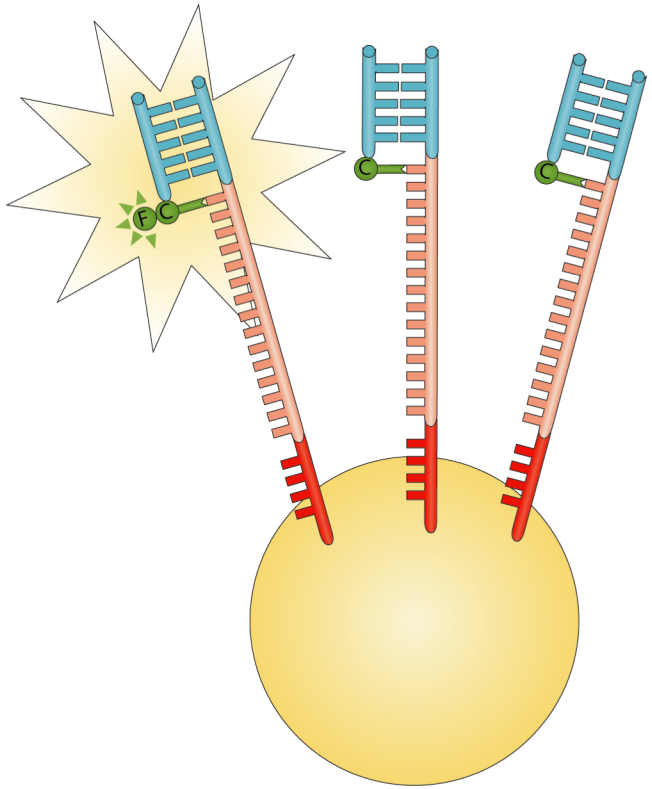
Cleavage
Fluorophores are cleaved and washed from flow cells and the 3′OH group is regenerated. A new cycle begins with the addition of new nucleotides.

Sequencing by synthesis: cyclic reversible termination

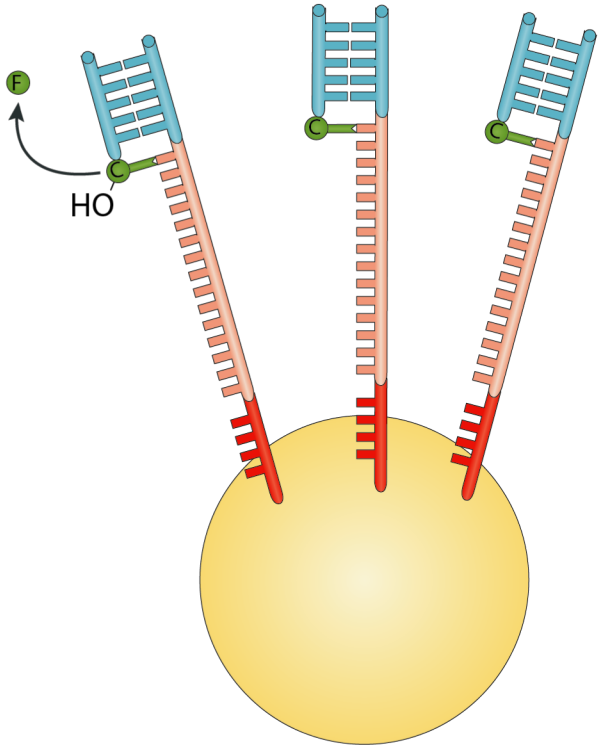
b GeneReader (Qiagen)



Nucleotide addition
A mixture of fluorophore-labelled, terminally blocked nucleotides and unlabelled, blocked nucleotides hybridize to complementary bases. Each bead on a slide can incorporate a different base.



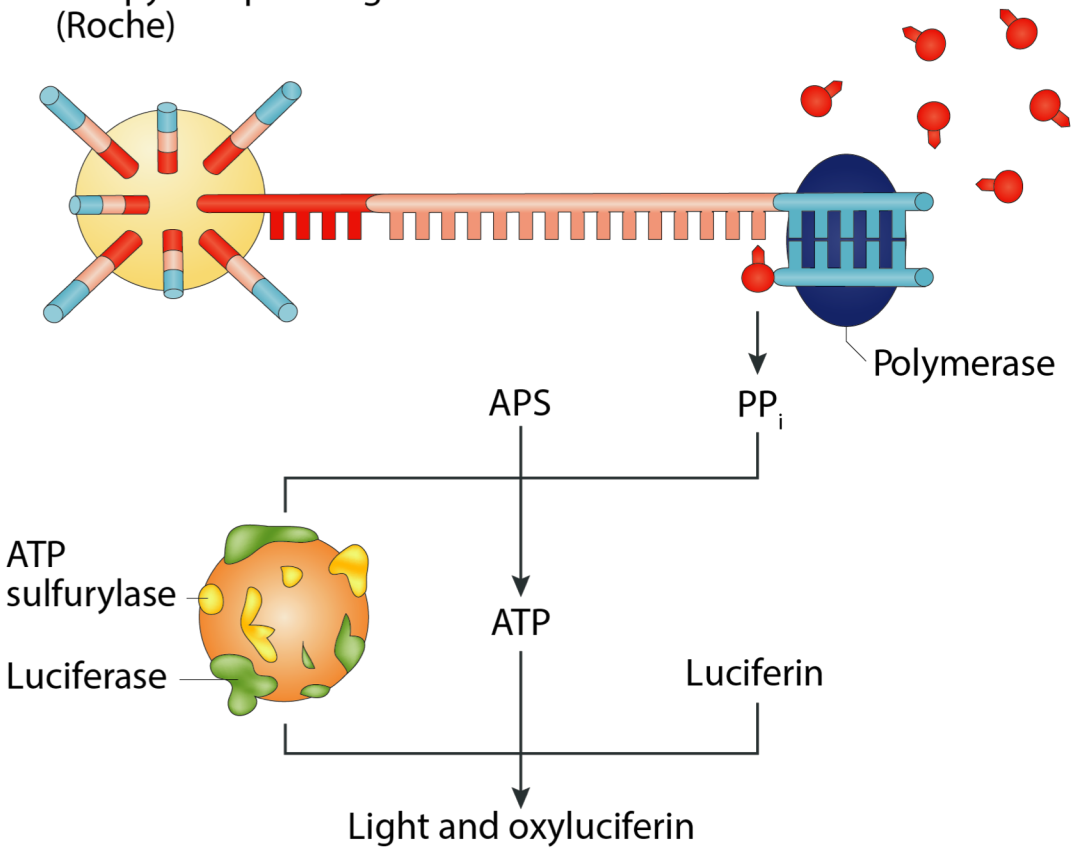
Imaging
Slides are imaged with four laser channels. Each bead emits a colour corresponding to the base incorporated during this cycle, but only labelled bases emit a signal.



Cleavage
Fluorophores are cleaved and washed from flow cells and the 3′-OH group is regenerated. A new cycle begins with the addition of new nucleotides.

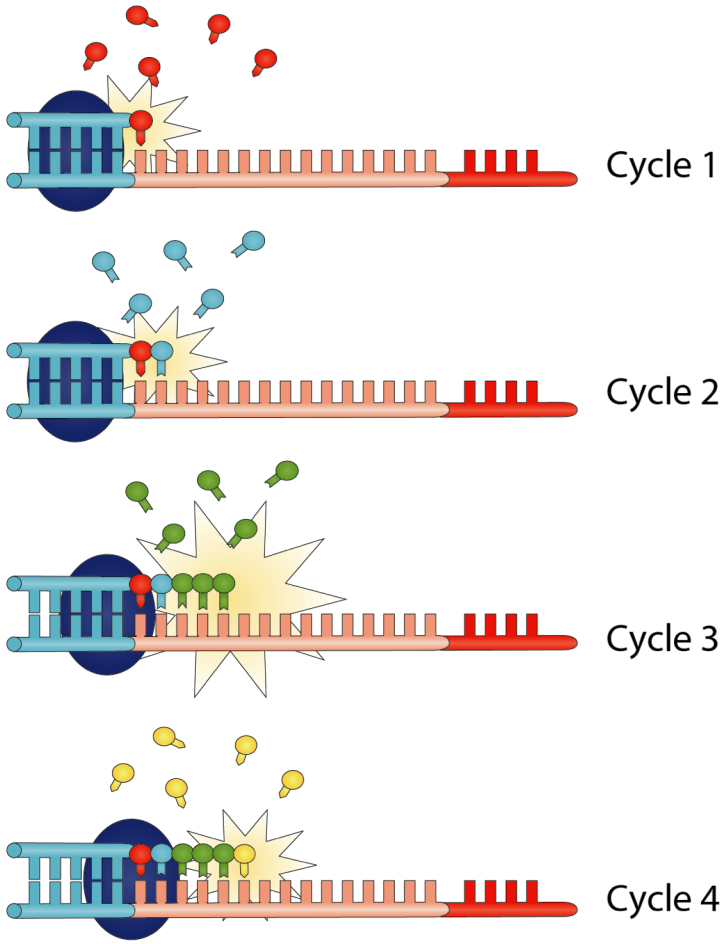
Sequencing by synthesis: single-nucleotide addition

454 pyrosequencing
(Roche)



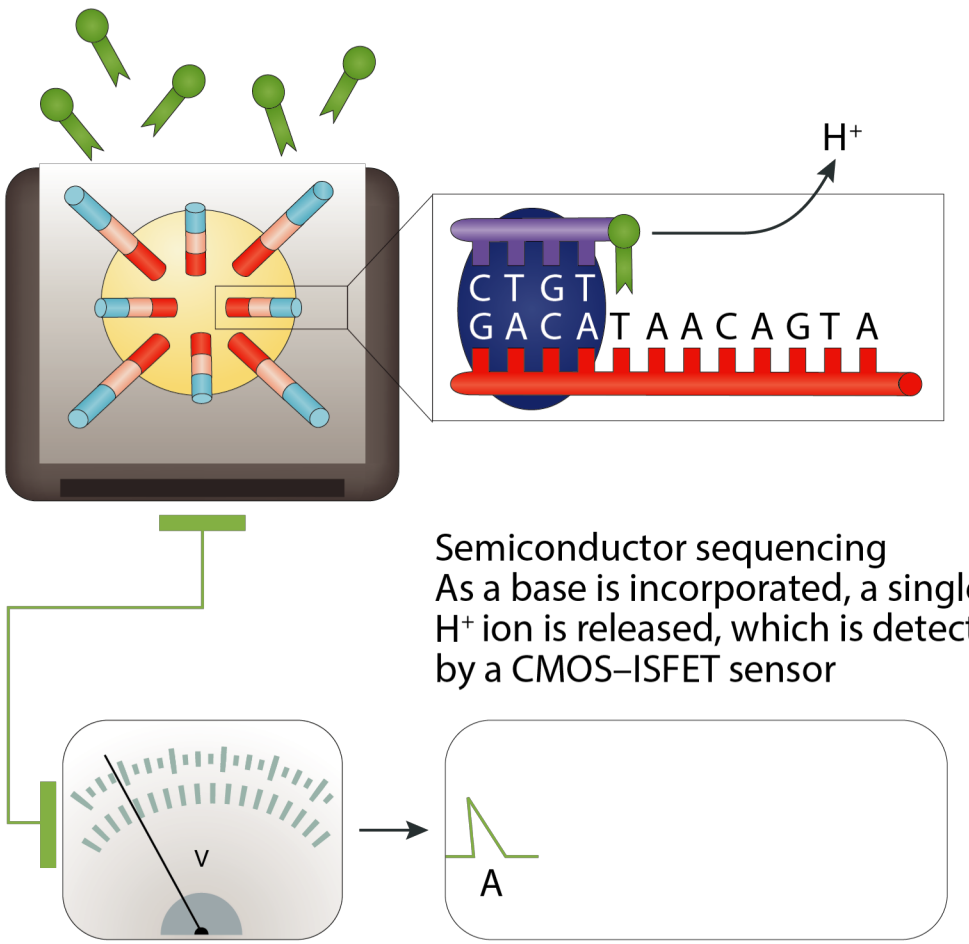
Pyrosequencing
As a base is incorporated, the release of an inorganic pyrophosphate triggers an enzyme cascade, resulting in light

Single nucleotide addition
Only one dNTP species is present during each cycle; multiple identical dNTPs can be incorporated during a cycle, increasing emitted light

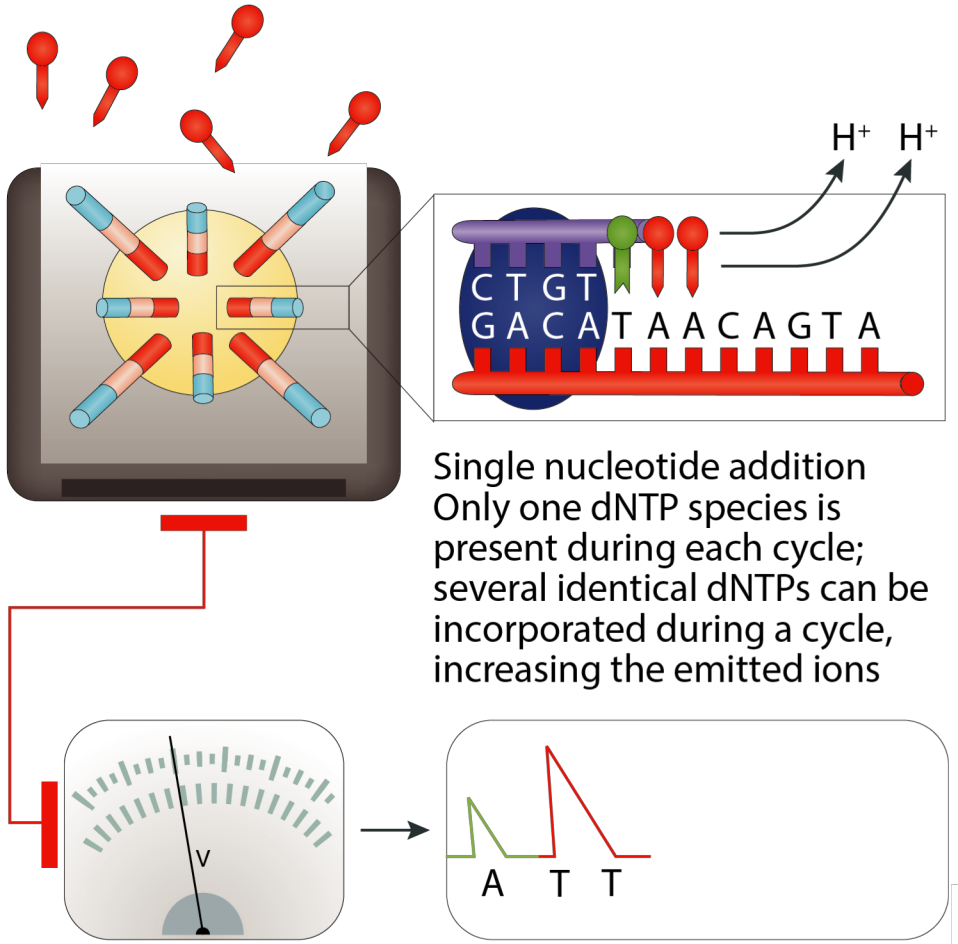


Sequencing by synthesis: single-nucleotide addition

Ion Torrent
(Thermo Fisher)

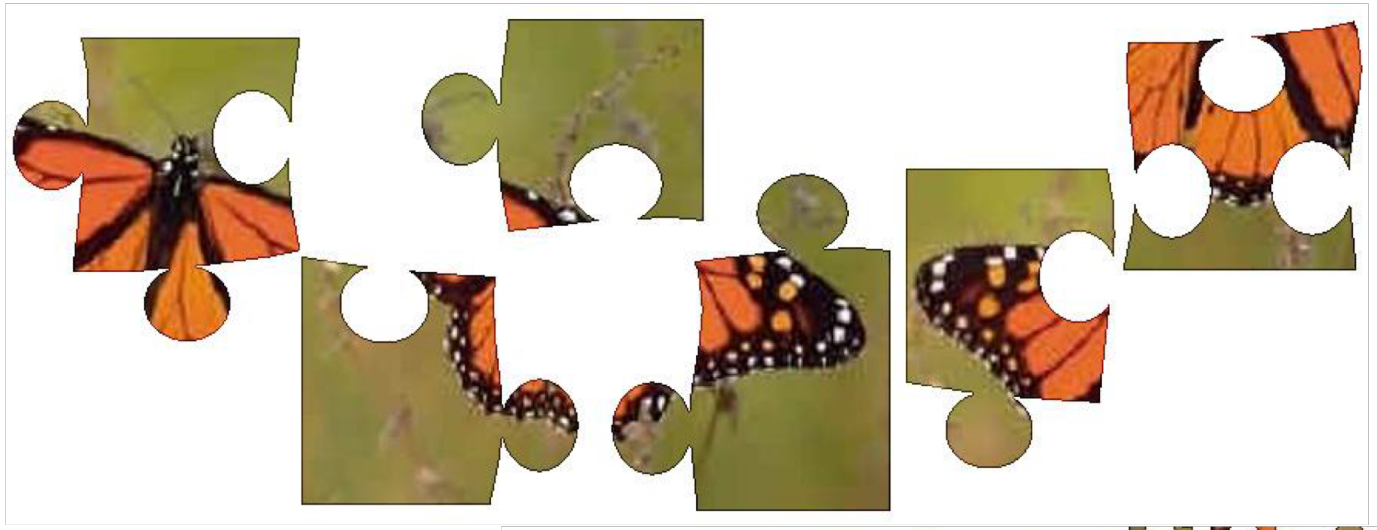


Semiconductor sequencing
As a base is incorporated, a single H^+ ion is released, which is detected by a CMOS–ISFET sensor



Single nucleotide addition
Only one dNTP species is present during each cycle; several identical dNTPs can be incorporated during a cycle, increasing the emitted ions

Long vs short reads



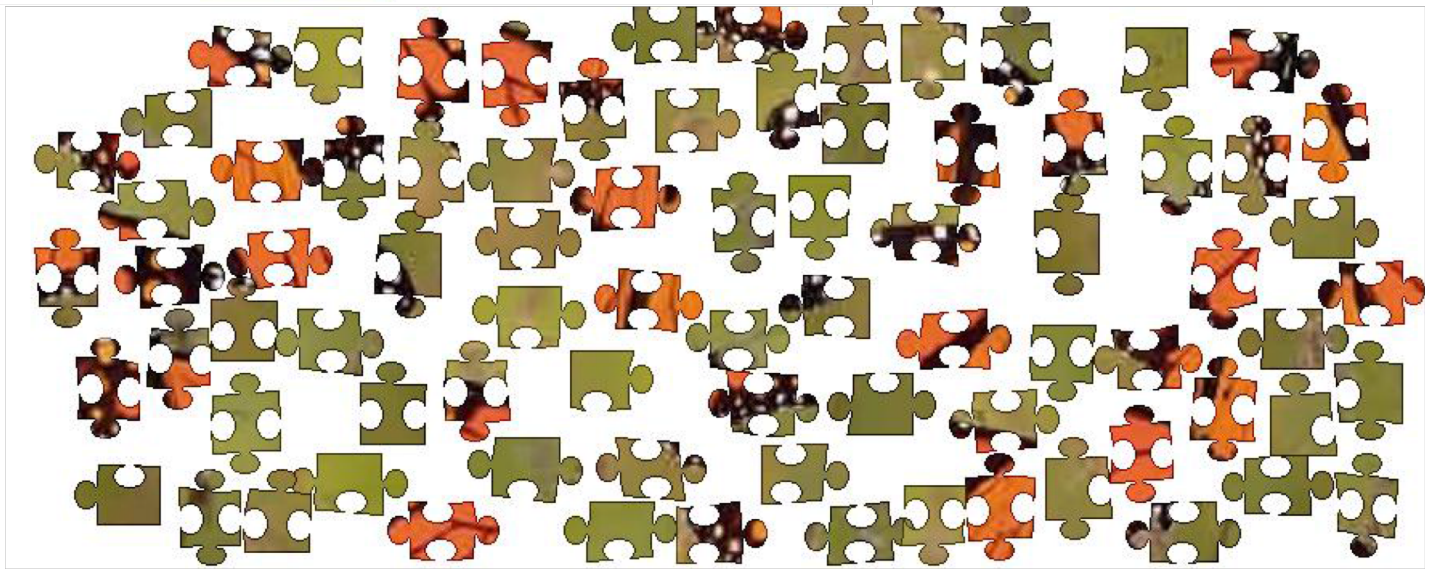


Image: Petri Auvinen

Challenges/limitations with short reads

- Cause **problems** for the assembler
- Produce **highly fragmented** genomes
- Fail to identify the full spectrum of **structural variation** seen in an individual genome

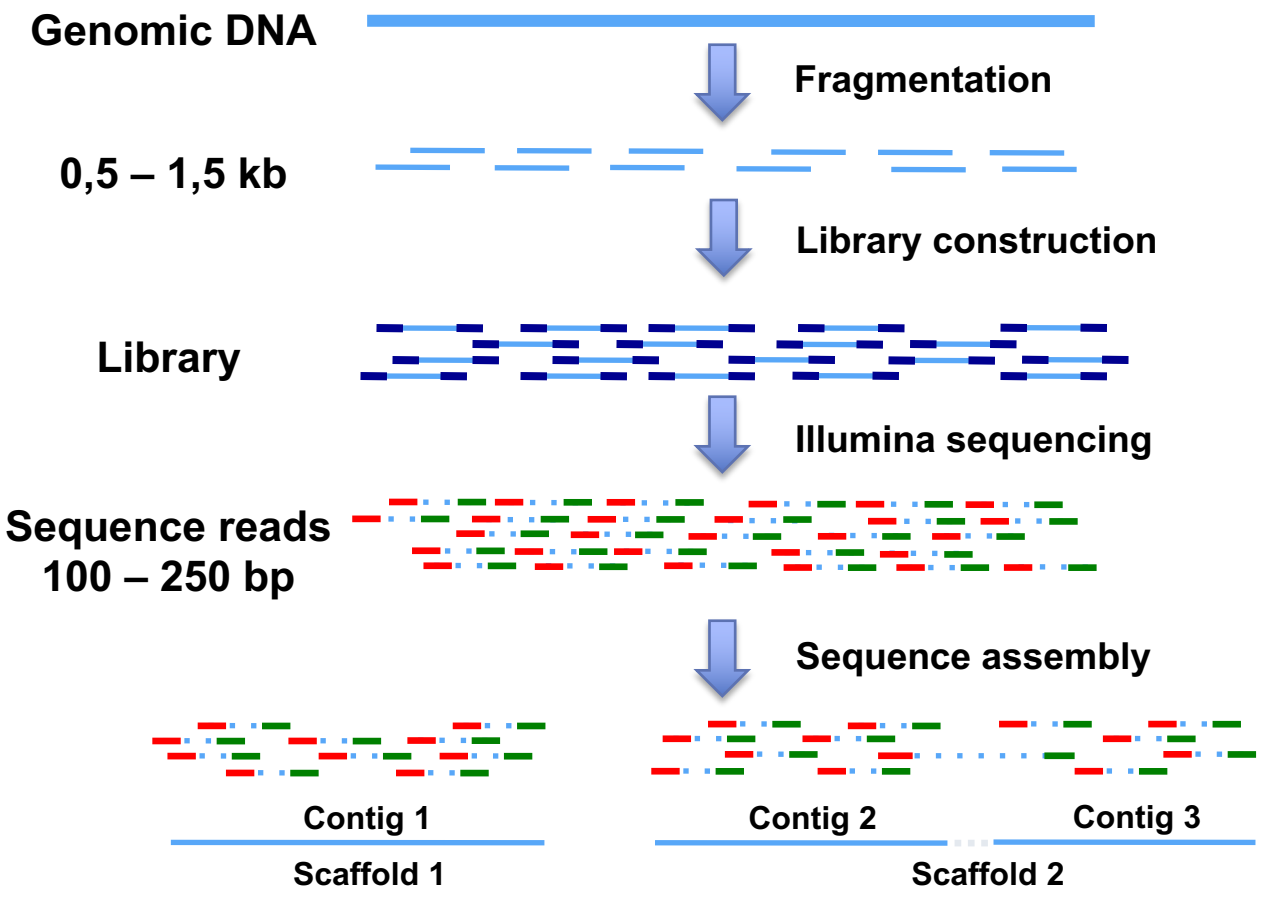


Image: Erik Hjerde

Why are repeats a problem?

- The law of repeats
  - It is impossible to resolve repeats of length L unless you have reads longer than L
  - It is impossible to resolve repeats of length L unless you have reads longer than L

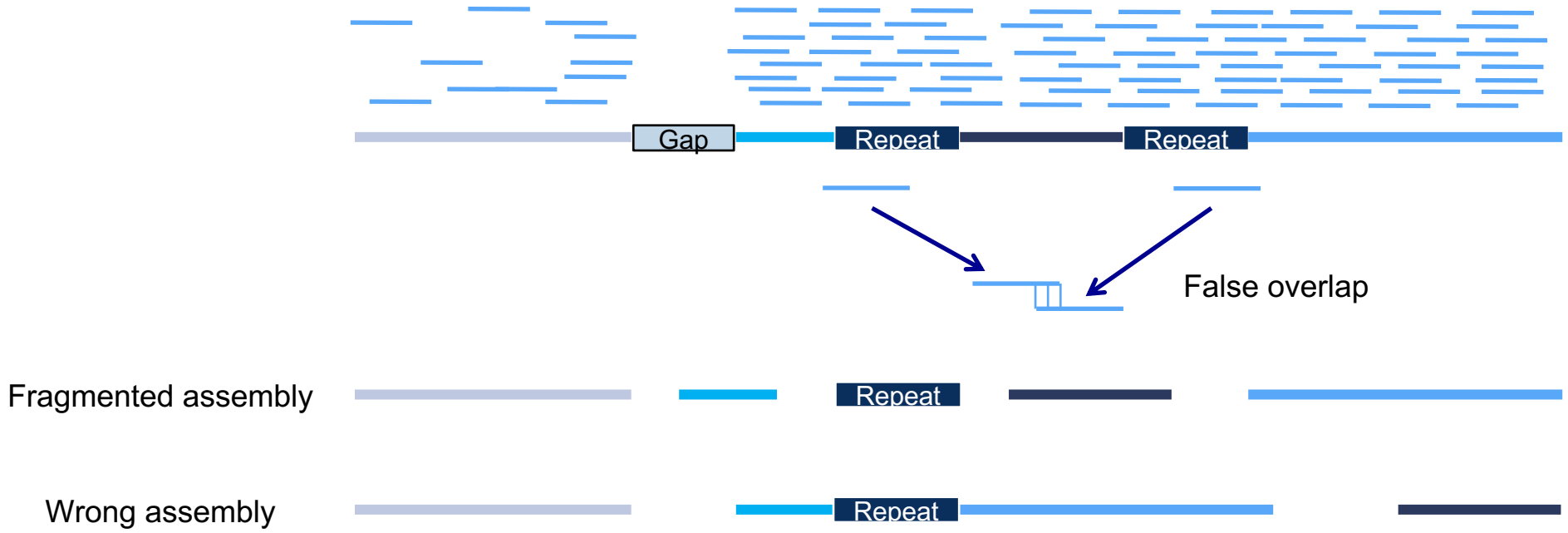
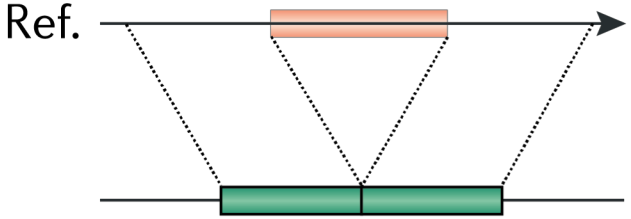


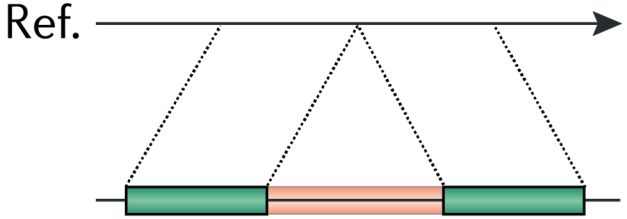
Image: Erik Hjerde

Challenges/Limitations with short reads: Structural variation

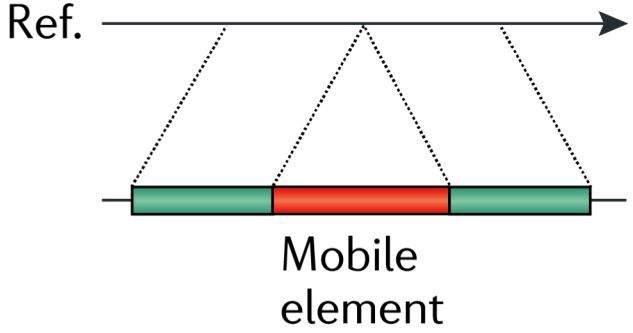
Deletion



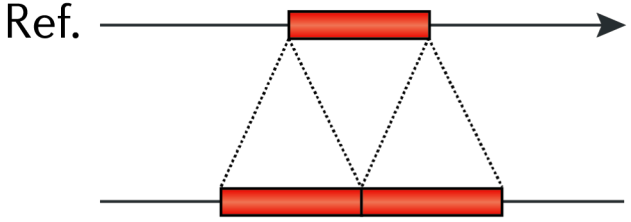
Novel sequence insertion



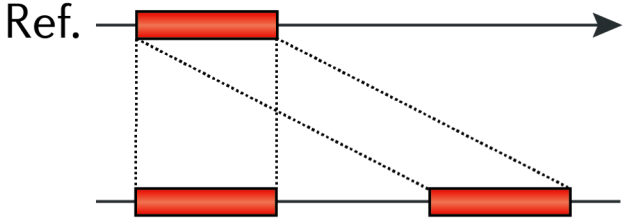
Mobile-element insertion



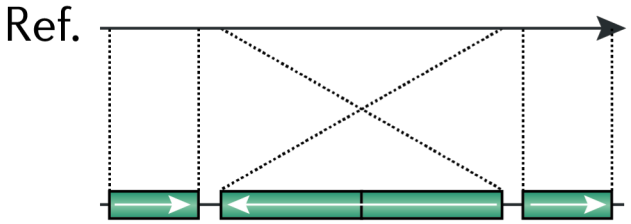
Tandem duplication



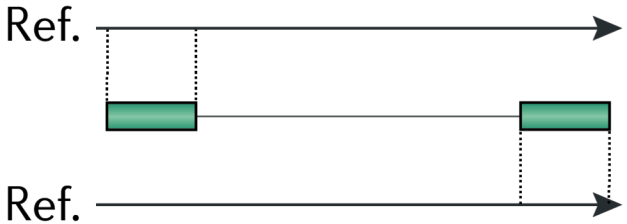
Interspersed duplication



Inversion



Translocation



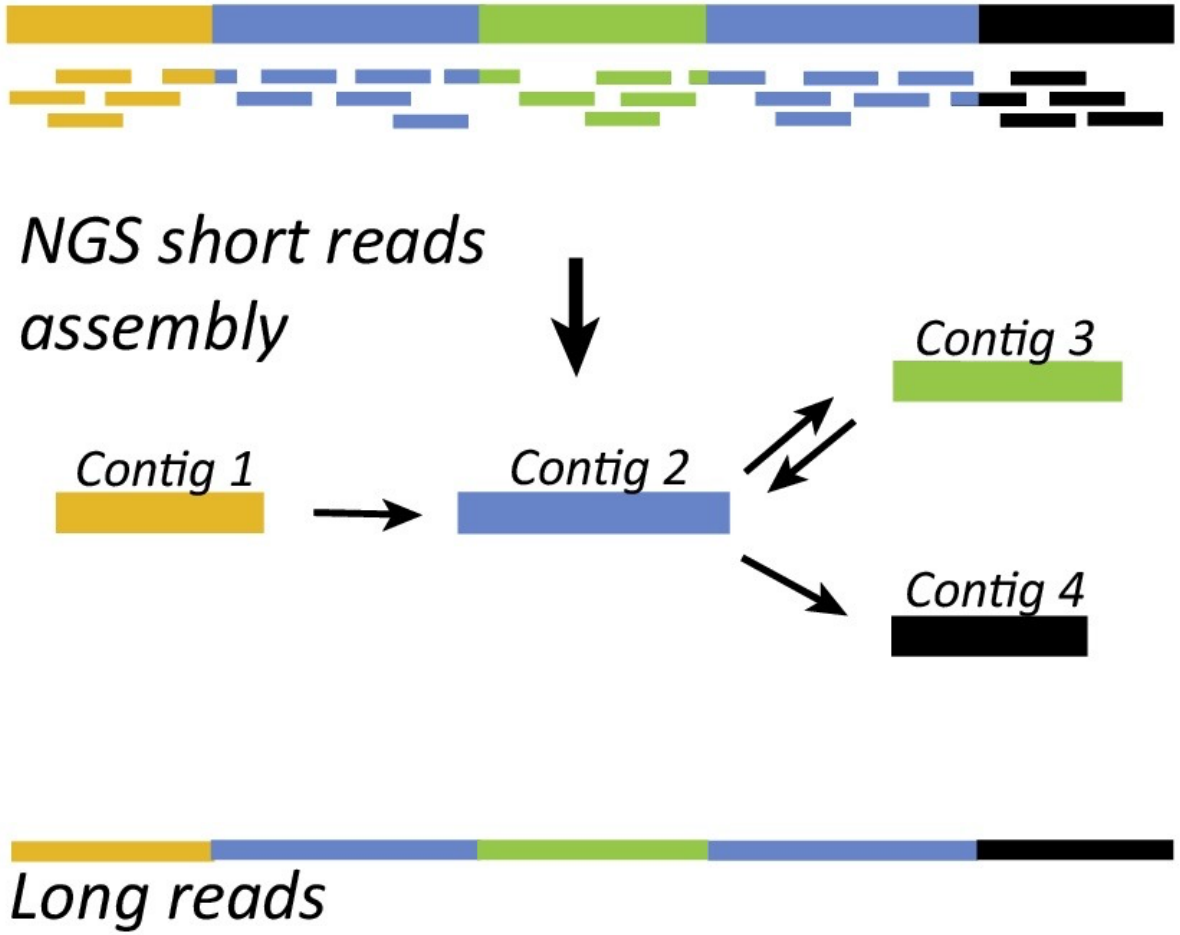
Solution = Long reads?

Sequencing lengths available

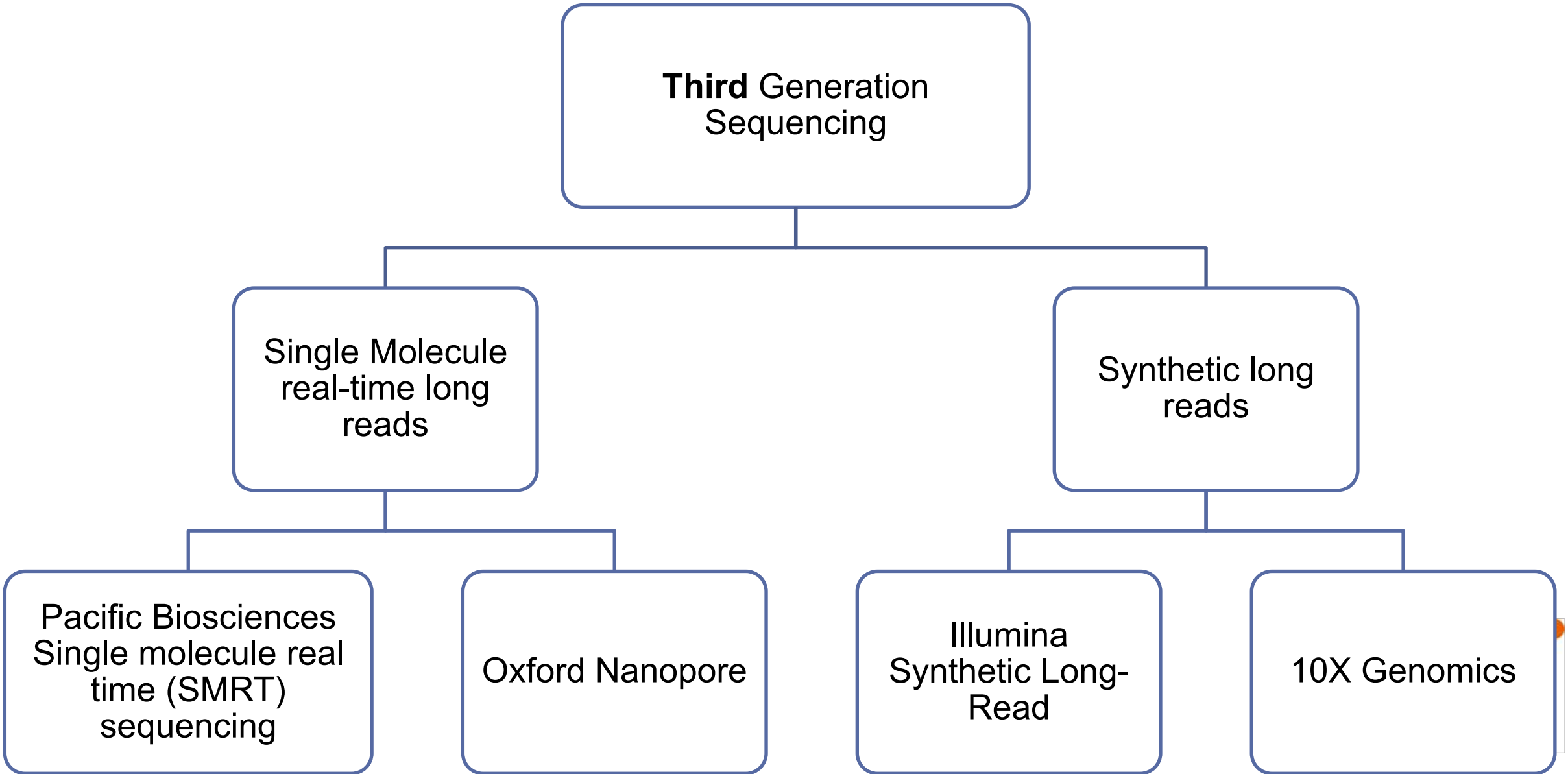
NGS Single End (50–300,Illumina)

NGS Paired End (2*75–300, total 150–600bp, Illumina)

Long Read (>10 000, no fixed upper limit)



The solution?

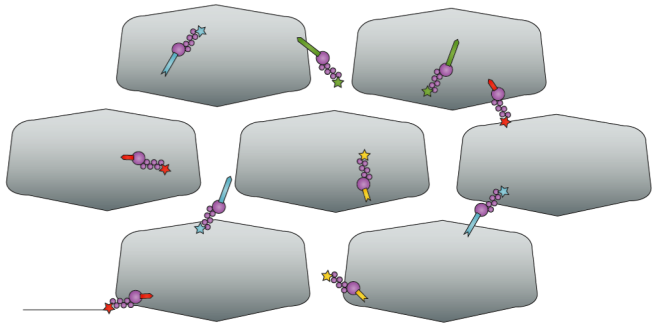


Pacific Biosciences

Real-time long-read sequencing approaches

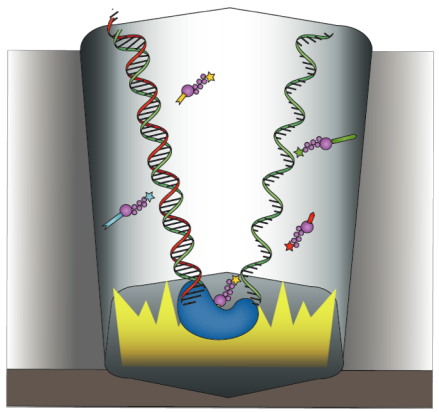
SMRTbell template
Two hairpin adapters allow continuous circular sequencing

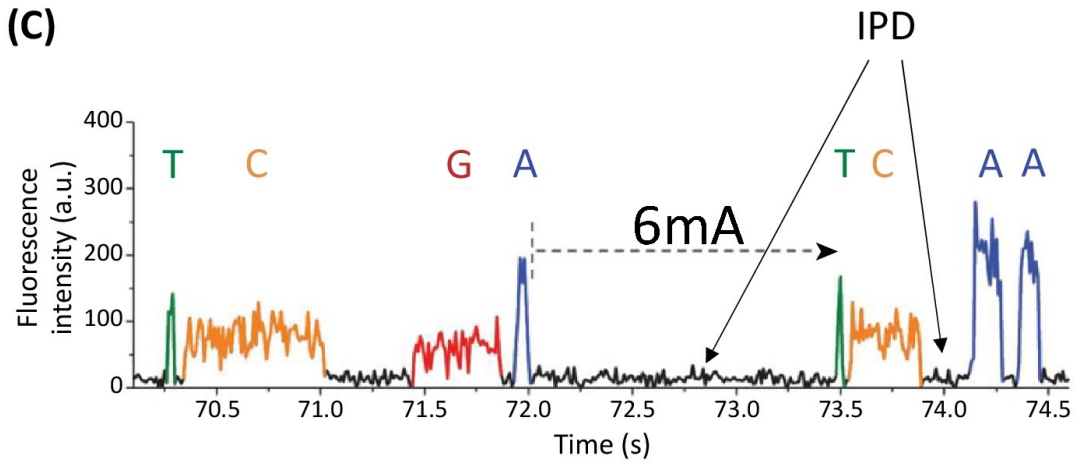
ZMW wells
Sites where sequencing takes place



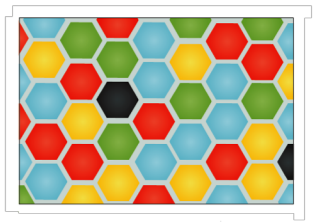
Labelled nucleotides
All four dNTPs are labelled and available for incorporation

Modified polymerase
As a nucleotide is incorporated by the polymerase, a camera records the emitted light



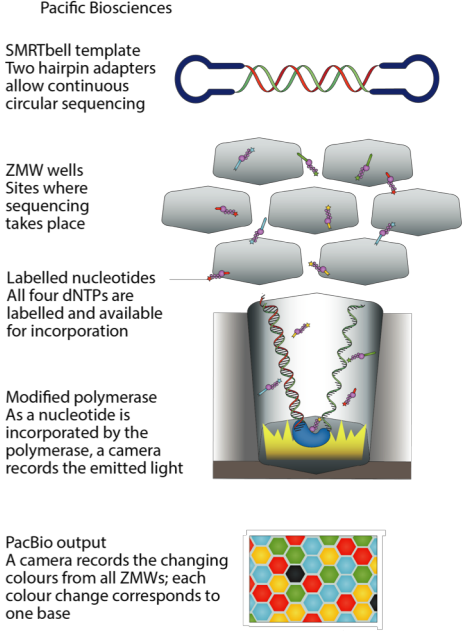


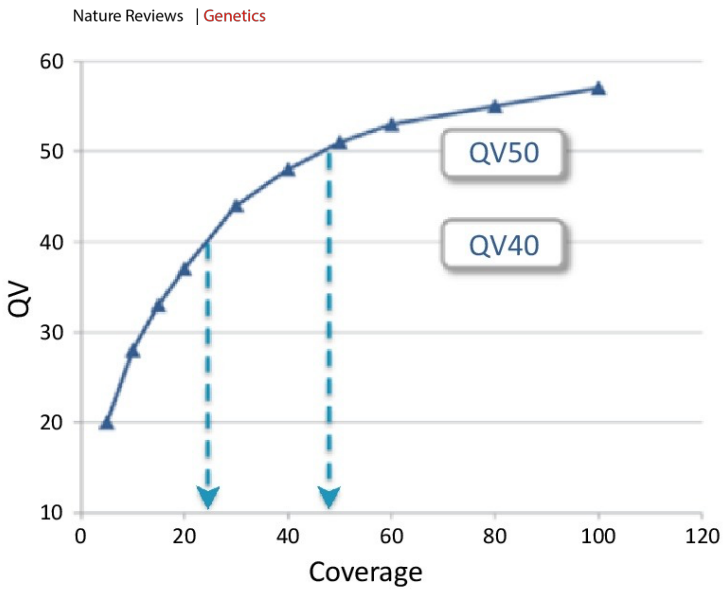
PacBio output
A camera records the changing colours from all ZMWs; each colour change corresponds to one base



Nature Reviews | Genetics

Real-time long-read sequencing approaches





Expensive method ($ 10.000 – 30 X)

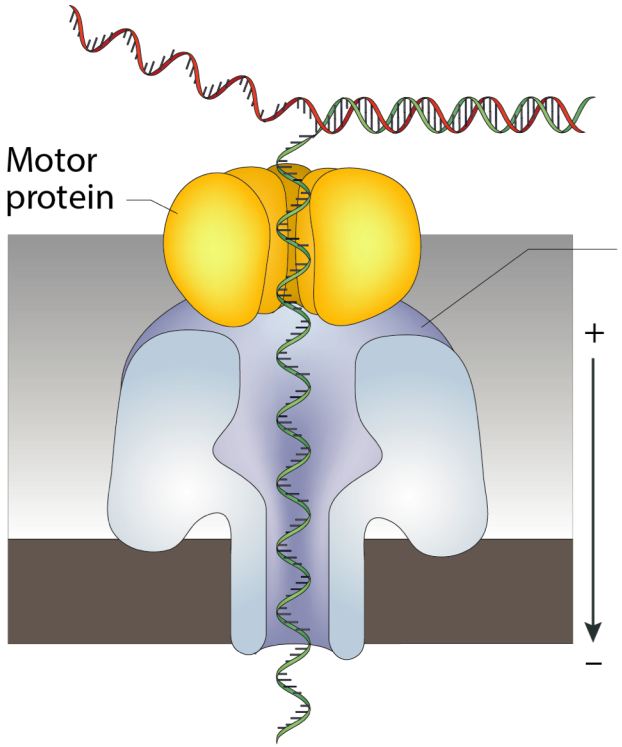
Problem: A single read is error prone (homopolymers)

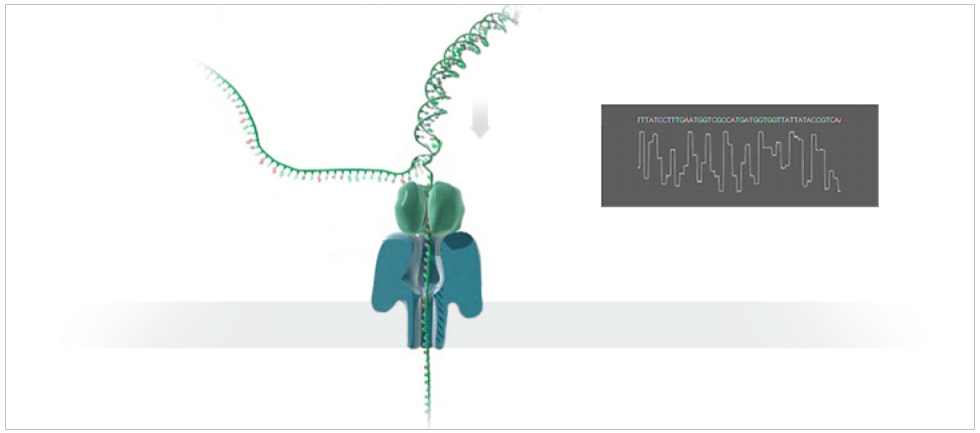
Solution: Fragments are sequenced continually to achieve basecall correction and very high accuracy (99.999%).

Oxford Nanopore Technologies

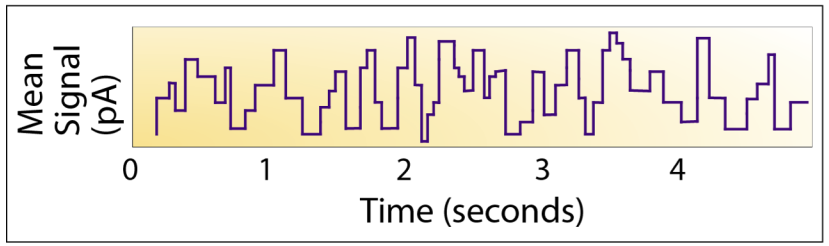
Real-time long-read sequencing approaches

Leader–Hairpin template
The leader sequence interacts with the pore and a motor protein to direct DNA, a hairpin allows for bidirectional sequencing





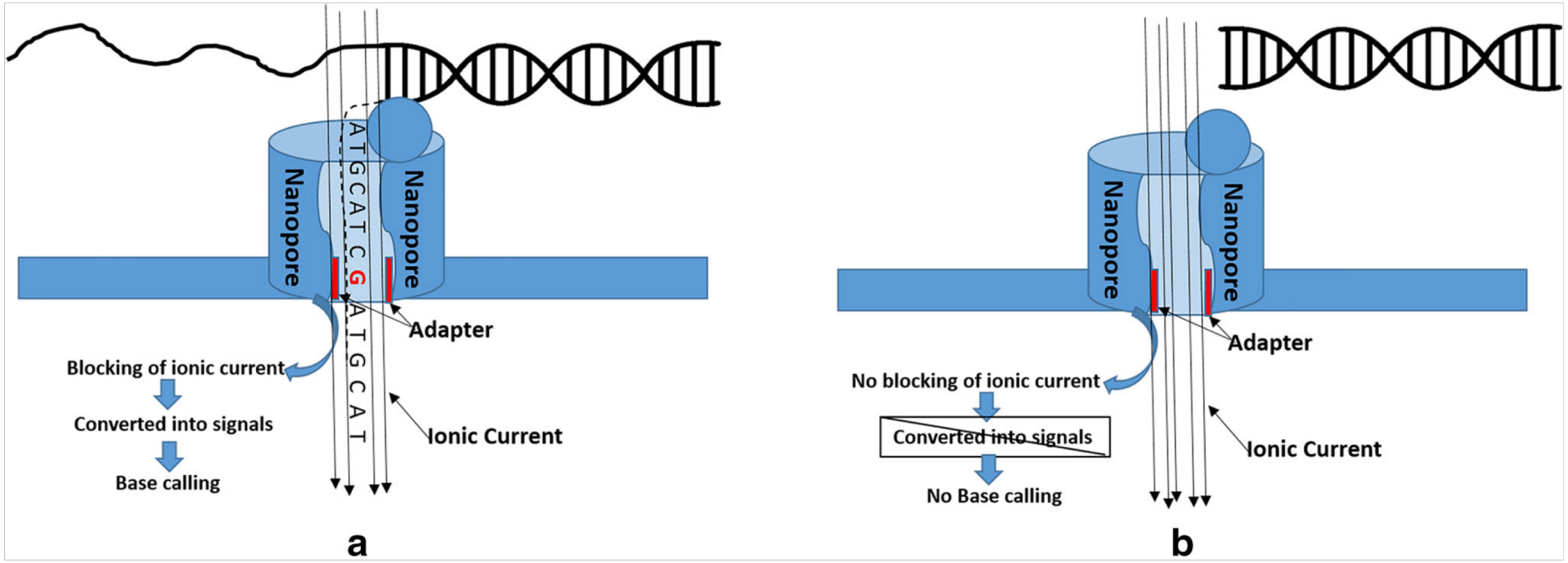
Problem: Has issues with homopolymers



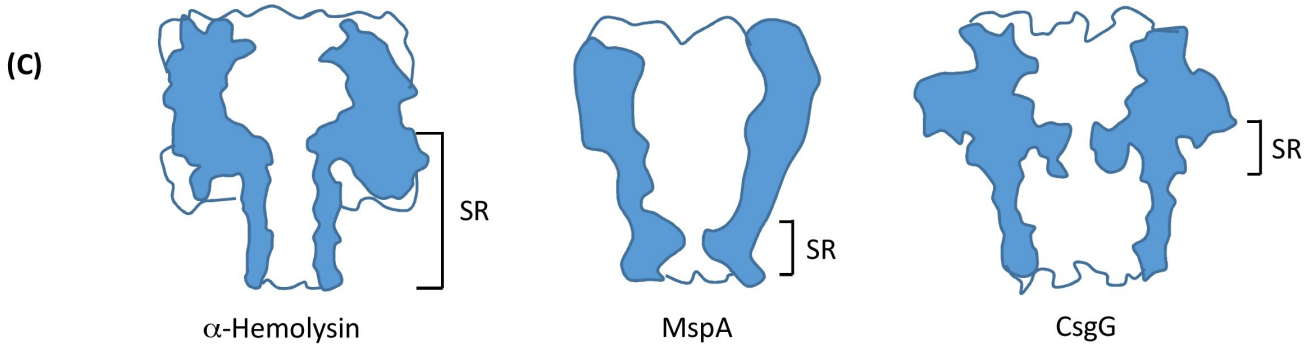
ONT output (squiggles)
Each current shift as DNA translocates through the pore corresponds to a particular k-mer

Solution: Optimal library preparation the DNA is double stranded, and both strands are read in succession, providing an internal control and an opportunity to create a consensus sequence of ~97% accuracy.

Oxford nanopore basecalling



Ambardar S. Et al, Indian J Microbiol (Oct–Dec 2016) 56(4):394–404



Trends in Genetics

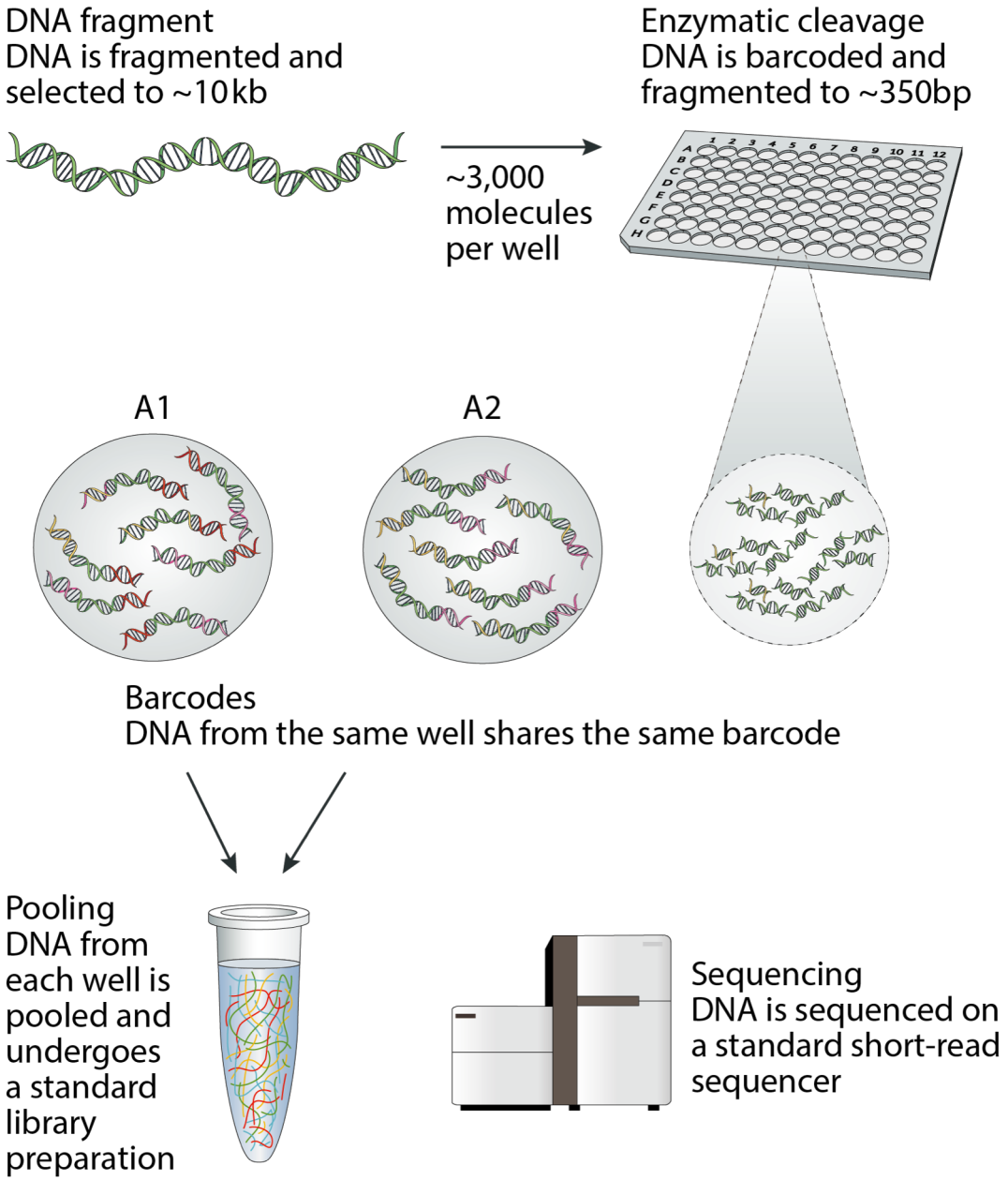
MinION (Oxford Nanopore)

- Very portable
- No special equipment to run
- Simple run
  - 10 minute prep
- Very cheap to run
  - $500-900 per (reusable flow-cell)
- Very long (100kb is not unusual)
- Reads RNA directly, (full-length transcripts)
- Data analysis is easier than for short-read sequencers,
- Reads appear in real-time (pull the USB plug when you have enough data)

Futuromics: SmidgION and the Flongle (Oxford Nanopore)

Synthetic long-read sequencing approaches

Illumina



Nature Reviews | Genetics

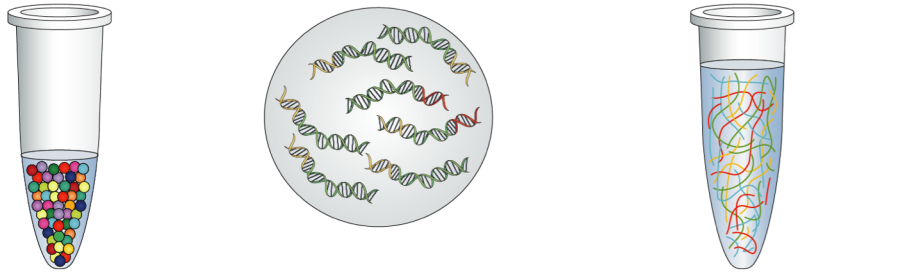
10X Genomics

Emulsion PCR
Arbitrarily long DNA is mixed with beads loaded with barcoded primers, enzyme and dNTPs

GEMs
Each micelle has 1 barcode out of 750,000

Amplification
Long fragments are amplified such that the product is a barcoded fragment ~350bp

Pooling
The emulsion is broken and DNA is pooled, then it undergoes a standard library preparation



Linked reads
- All reads from the same GEM derive from the long fragment, thus they are linked
- Reads are dispersed across the long fragment and no GEM achieves full coverage of a fragment
- Stacking of linked reads from the same loci achieves continuous coverage

Nature Reviews | Genetics

Synthetic long-read sequencing approaches

- Chromium prep ($ 600 – 30 X)
- 1 ng of starting material
- 1 mill GEMs
- 4 mill barcodes
- 50 kb «read clouds»

Metagenomic Sequencing

- Long read platforms:
  - Facilitates assembly and annotation
  - May fail to accurately quantify copy number and allelic variants
- Short read high coverage platforms
  - Accurate quantification of copy number and allelic variants of various genes
  - Assembly and annotation very challenging

Short vs long reads

- Illumina sequencing technology dominates
  - Short reads, 2x250 or 2x300 bp
  - Sequencing depth
  - Cheap
  - Lower error rates

- Longer reads (ex PACBIO, Nanopore)
  - High error rates
  - Lower sequencing depth
  - Higher costs
  - Epigenetics

Questions?