TOPIC 1: Introduction to 2nd and 3rd-Gen Sequencing

Biol 525D - Bioinformatics for Evolutionary Biology
2020

Instructors

Dr Tom Booker

Julia Kreiner

Professor Kathryn Hodgins

booker@zoology.ubc.ca

julia.kreiner@mail.utoronto.ca

kathryn.hodgins@monash.edu

WEBSITE: https://ubc-biol525d.github.io/

Overview of the week

1. Introduction: Scope of course and overview of technology [Tom]
2. Introduction to command line programming [Tom]
3. Fastq files and quality checking/trimming [Kay]
4. Alignment: algorithms and tools [Tom]
5. Assembly: transcriptome and genome assembly [Kay]
6. RNAseq + differential expression analysis [Kay]
7. SNP and variant calling [Julia]
8. Population genomics and plotting in R (Part 1) [Julia]
9. Population genomics and plotting in R (Part 2) [Julia]
10. Case studies [Tom/Julia]

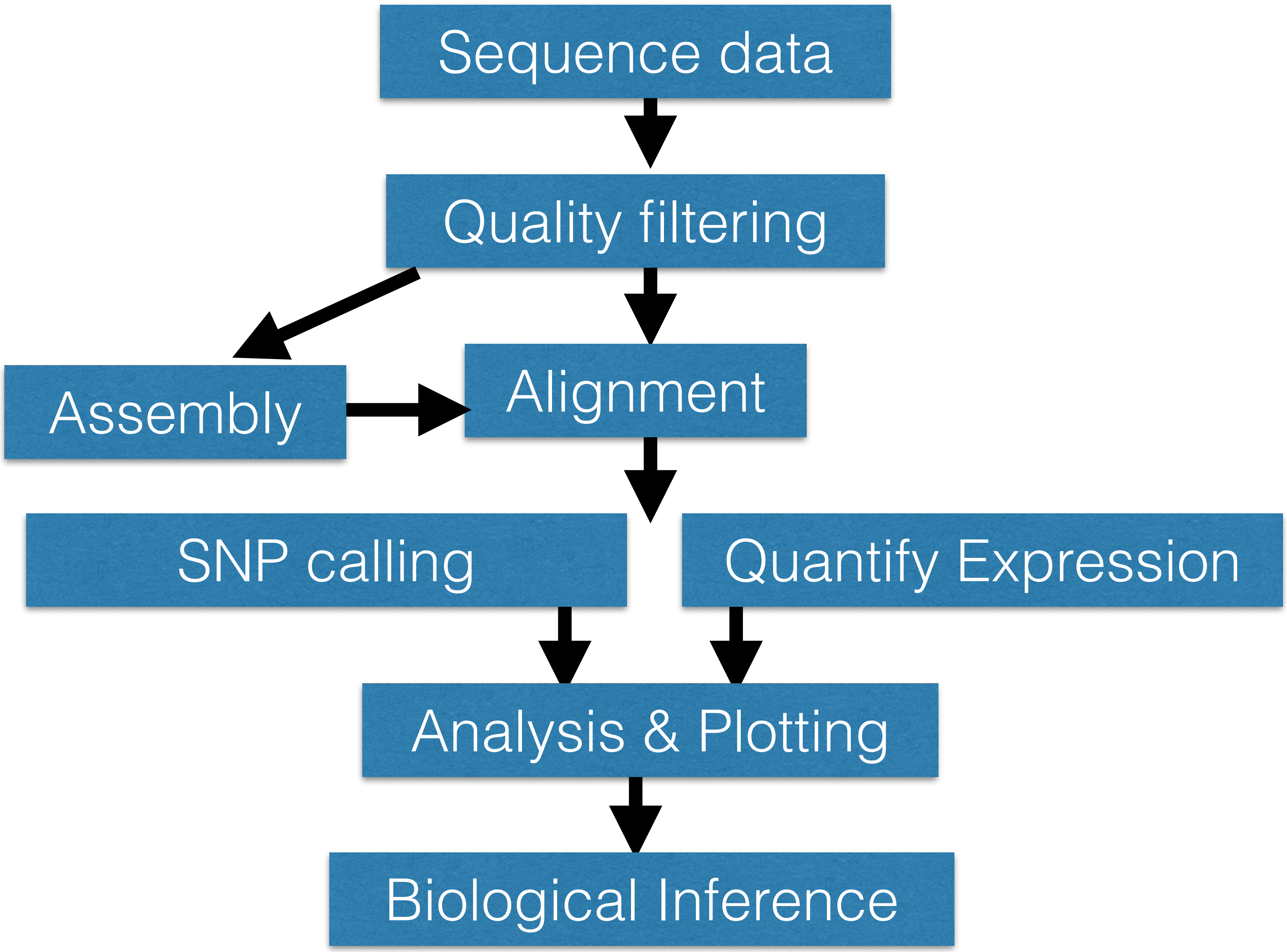
Goal of any sequencing project

Raw sequence data

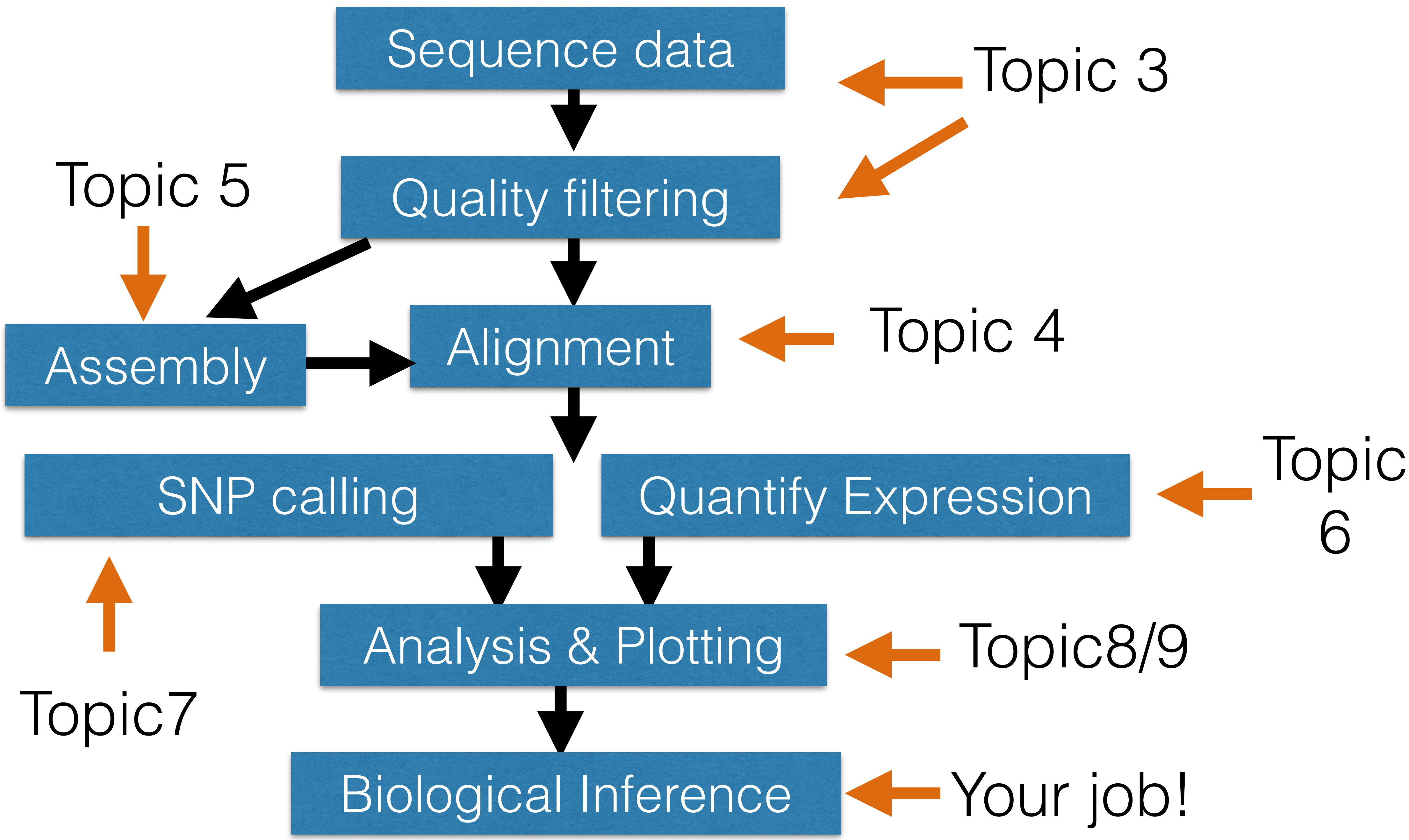
????

Biological inference

Rough outline



Rough outline



A brief history of DNA sequencing

Genome milestones

1977: *Bacteriophage* ΦX174
1982: *Bacteriophage lambda*
1995: *Haemophilus influenzae*
1996: *Saccharomyces cerevisiae*
1998: *Caenorhabditis elegans*
2000: *Drosophila melanogaster*
2000: *Arabidopsis thaliana*
2001: *Homo sapiens*
2002: *Mus musculus*
2004: *Rattus norvegicus*
2005: *Pan troglodytes*
2005: *Oryza sativa*
2007: *Cyanidioschyzon merolae*
2009: *Zea mays*
2010: Neanderthal
2012: Denisovan
2013: The HeLa cell line
2013: *Danio rerio*
2017: *Xenopus laevis*

Excerpted and edited from Box 1 and 2 - Shendure et al 2017 Nature

A brief history of DNA sequencing

Technological milestones

1953: Sequencing of insulin protein
1965: Sequencing of alanine tRNA
1968: Sequencing of cohesive ends of phage lambda DNA
1977: Maxam–Gilbert sequencing
1977: Sanger sequencing
1990: Paired-end sequencing
2000: Massively parallel signature sequencing by ligation
2003: Single-molecule massively parallel sequencing-by-synthesis
2003: Zero-mode waveguides for single-molecule analysis
2003: Sequencing by synthesis of in vitro DNA colonies in gels
2005: Four-colour reversible terminators
2005: Sequencing by ligation of in vitro DNA colonies on beads
2007: Large-scale targeted sequence capture
2010: Direct detection of DNA methylation during single-molecule sequencing
2010: Single-base resolution electron tunnelling through a solid state detector
2011: Semiconductor sequencing by proton detection
2012: Reduction to practice of nanopore sequencing
2012: Single-stranded library preparation method for ancient DNA

Excerpted and edited from Box 1 and 2 - Shendure et al 2017 Nature

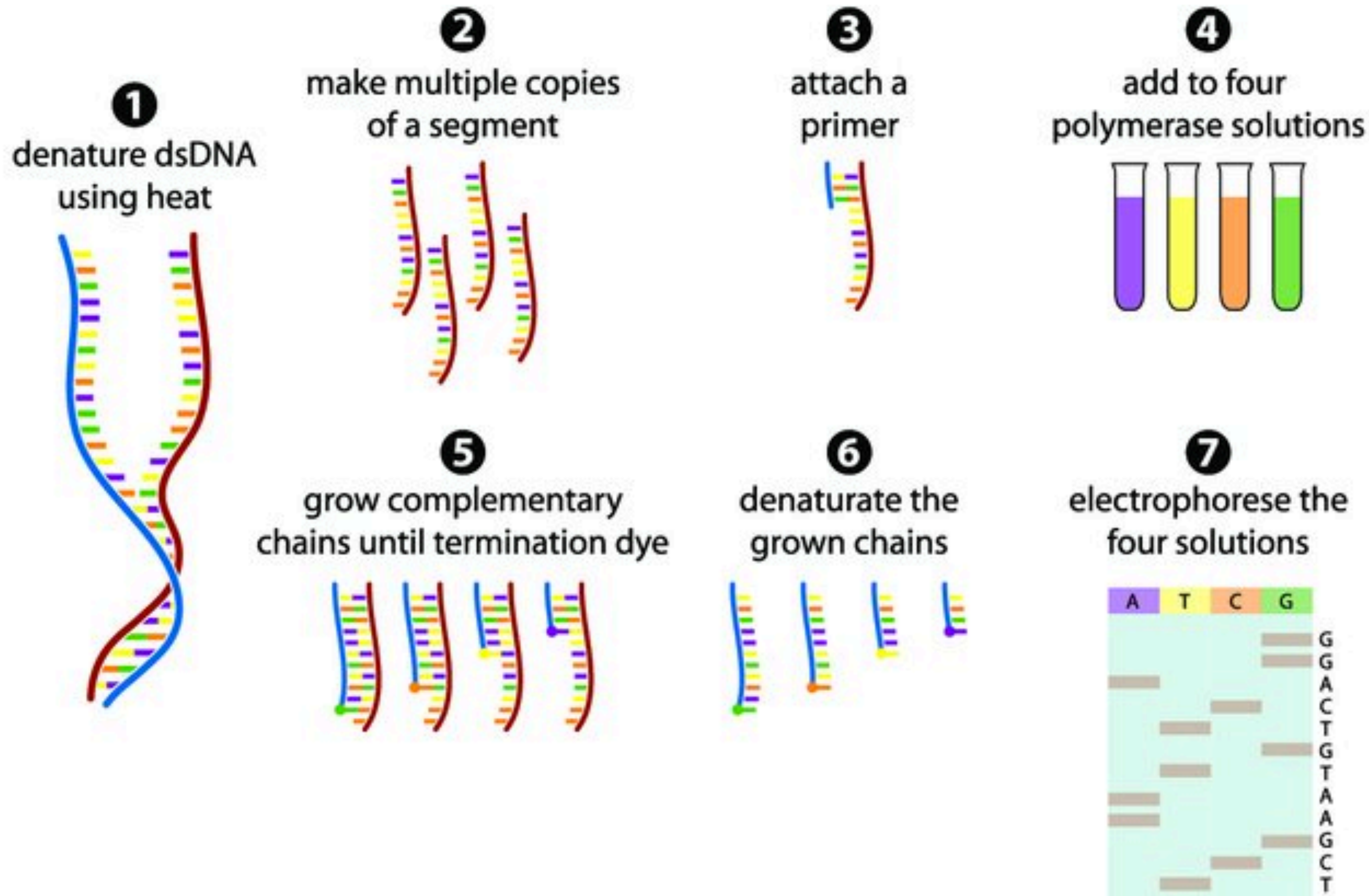
First Generation Sequencing

Maxam-Gilbert: Chemical modification and cleavage followed by gel electrophoresis

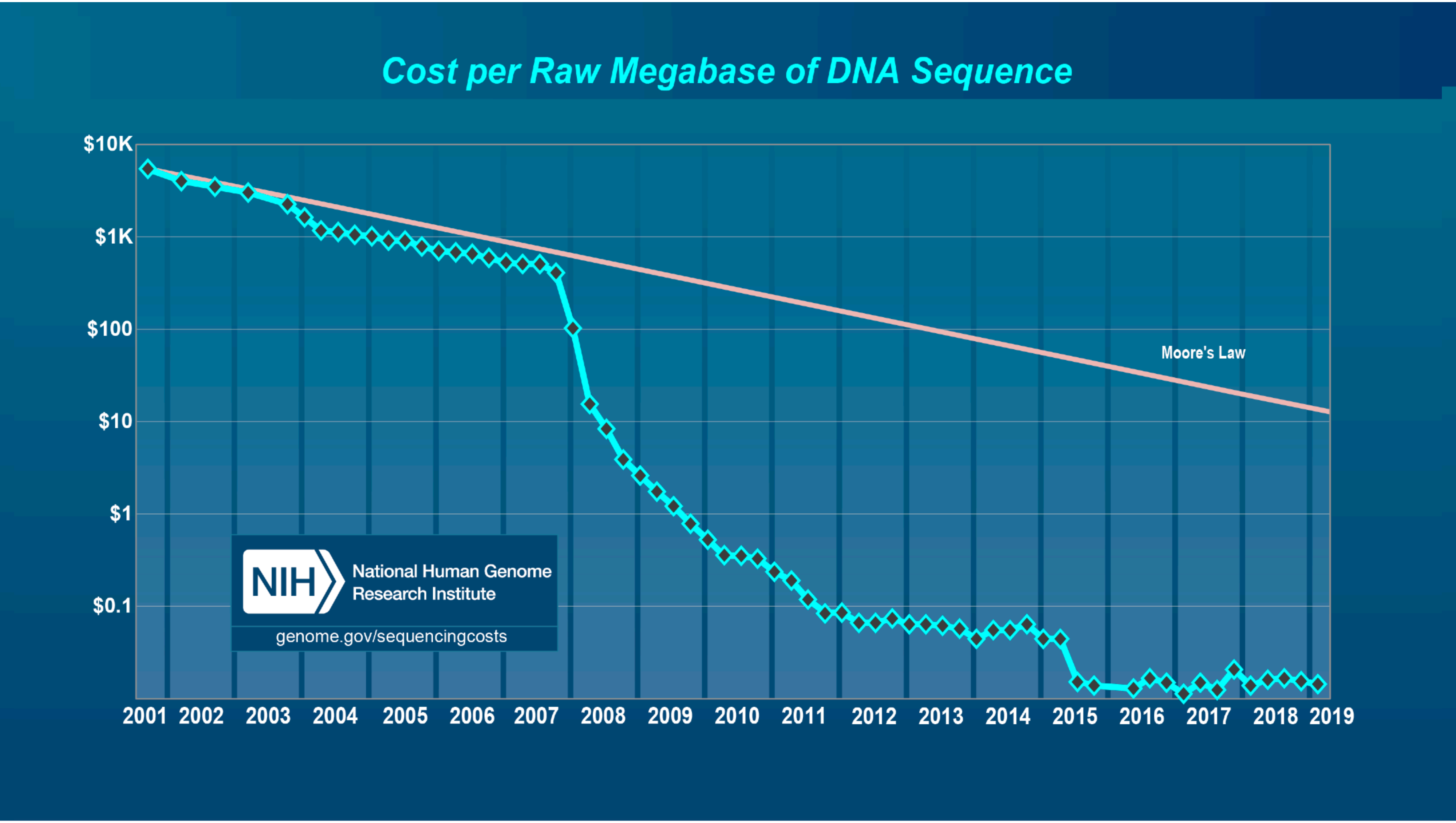
Sanger: Selective incorporation of chain-terminating dideoxynucleotides followed by gel electrophoresis

- Became full automated using flourescently labeled dideoxy bases
- Dominant sequencer up until 2007
- Only one fragment sequenced per reaction
- Still used for sequencing individual PCR products

Sanger sequencing



From PhD thesis of Michel G. Gauthier



*Moore’s law stated that the number of transistors on a microchip doubled every two years, while costs halved

Second (Next-gen) and third generation sequencing

Sequences many molecules in parallel

Don’t need to know anything about the sequence to start

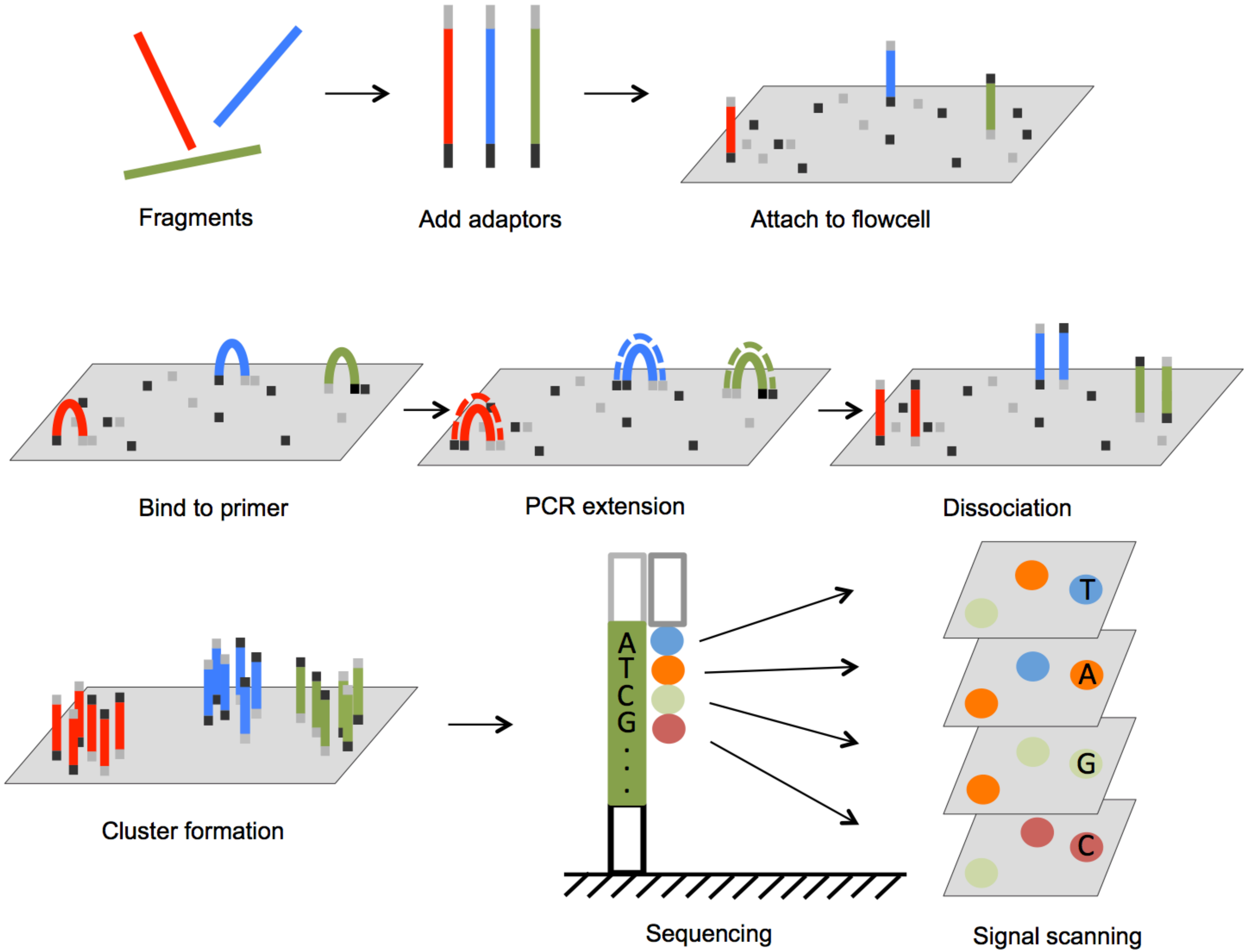
Main technologies:

- Illumina
- Ion torrent
- 454 (Pyrosequencing)
- PacBio

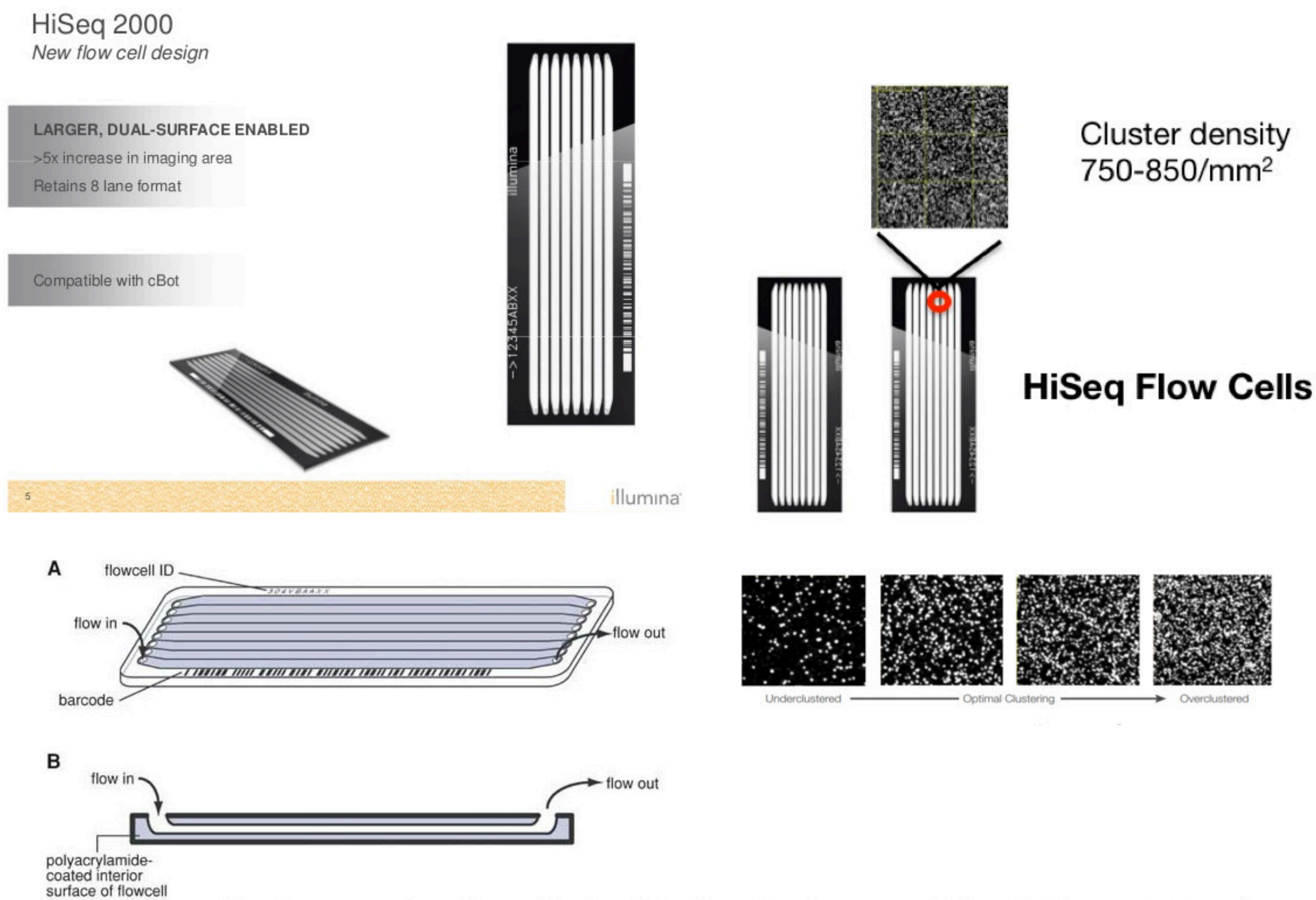
Second generation sequencing

Technology	Read Length	Accuracy	Bases/run	Uses
Illumina	50-600bp	99.9%	500-600 GBase	Resequencing General depth
Oxford Nanopore	5kb-100kb	85-95%	10-30GBase	Microbial genomes Genome assembly
PacBio	10kb-40kb	85-90%	5-10Gbase	Genome assembly Structural variants

Illumina sequencing



Illumina sequencing



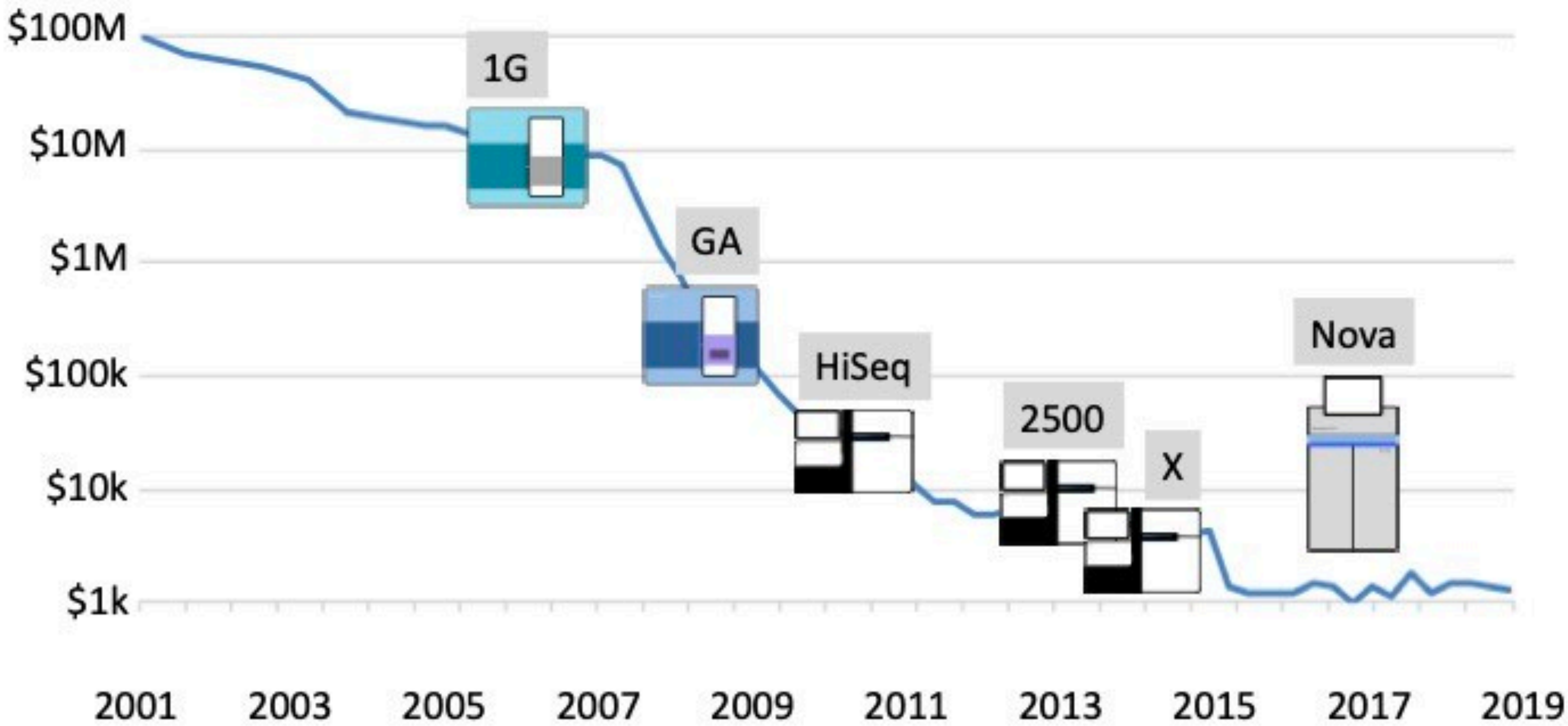
Illumina uses a glass ‘flowcell’, about the size of a microscope slide, with 8 separate ‘lanes’.

The HiSeq instrument scans both upper and lower surfaces of each flowcell lane.

From [hackteria.org](https://www.hackteria.org/wiki/File:FlowCell.jpg)

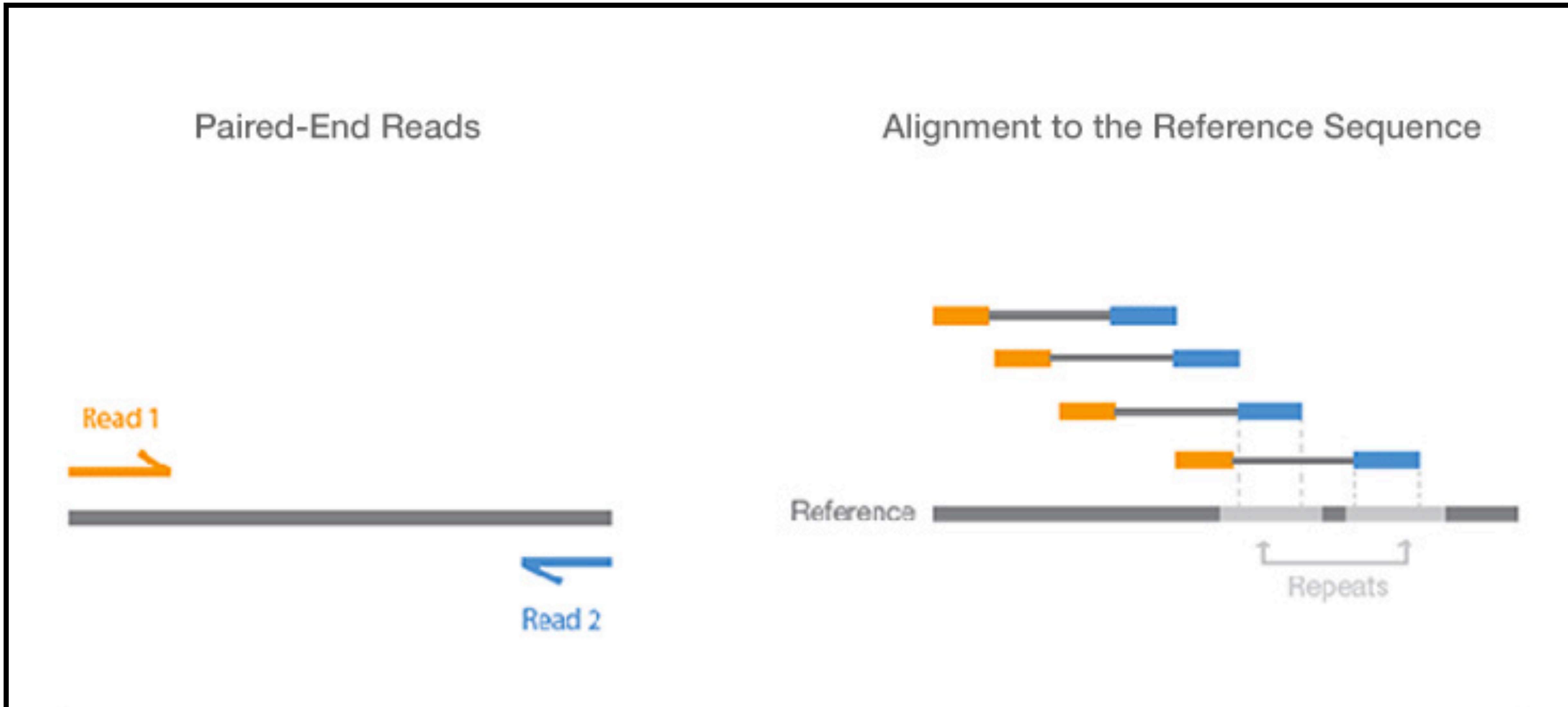
https://www.hackteria.org/wiki/File:FlowCell.jpg

Production cost per 30x Human genome over 18 years



Illumina sequencing

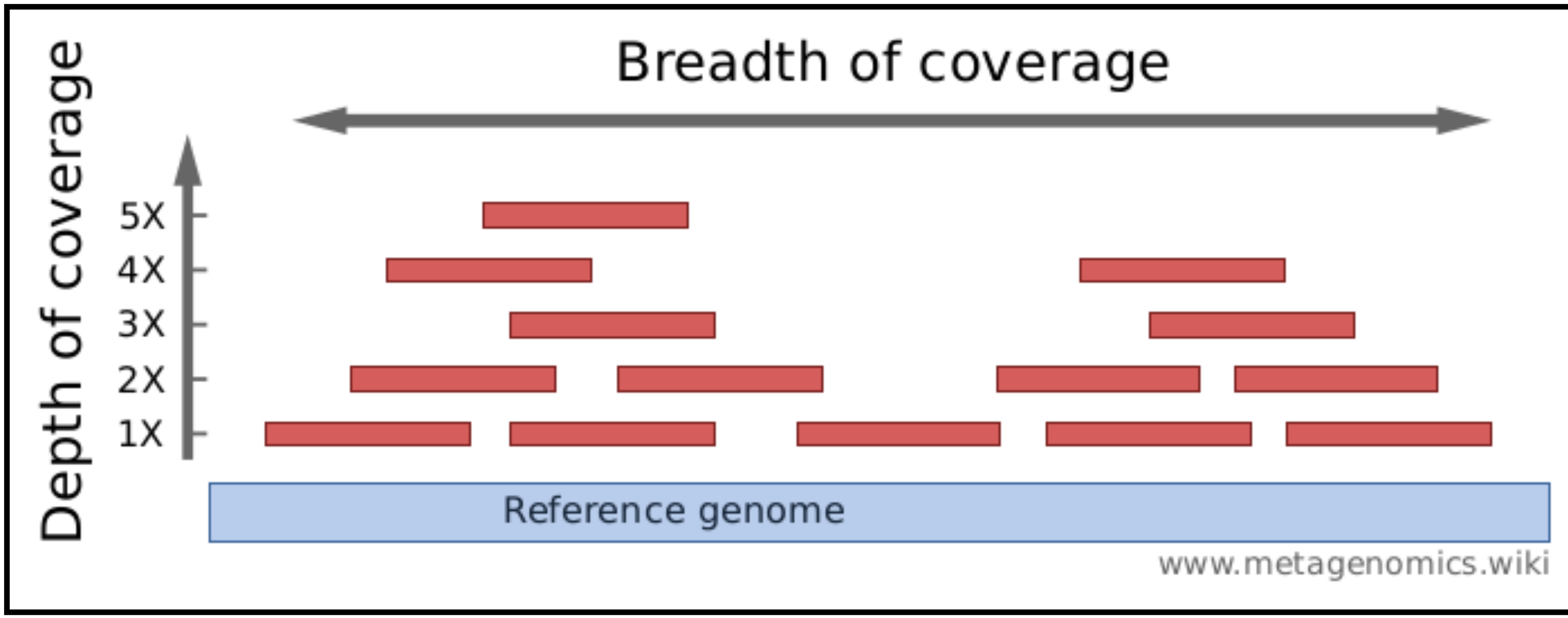
Why paired ends?



From Illumina website

Illumina sequencing

Important concepts



Illumina Machines

Name	MiSeq	HiSeq 4000	NovaSeq 6000
Sequencing Capacity	8Gbp	50Gbp	500-600Gbp
Cost (/lane)	~$1,500	~$3,000	~$8,000

Long read sequencing

Two dominant companies are PacBio and Oxford Nanopore

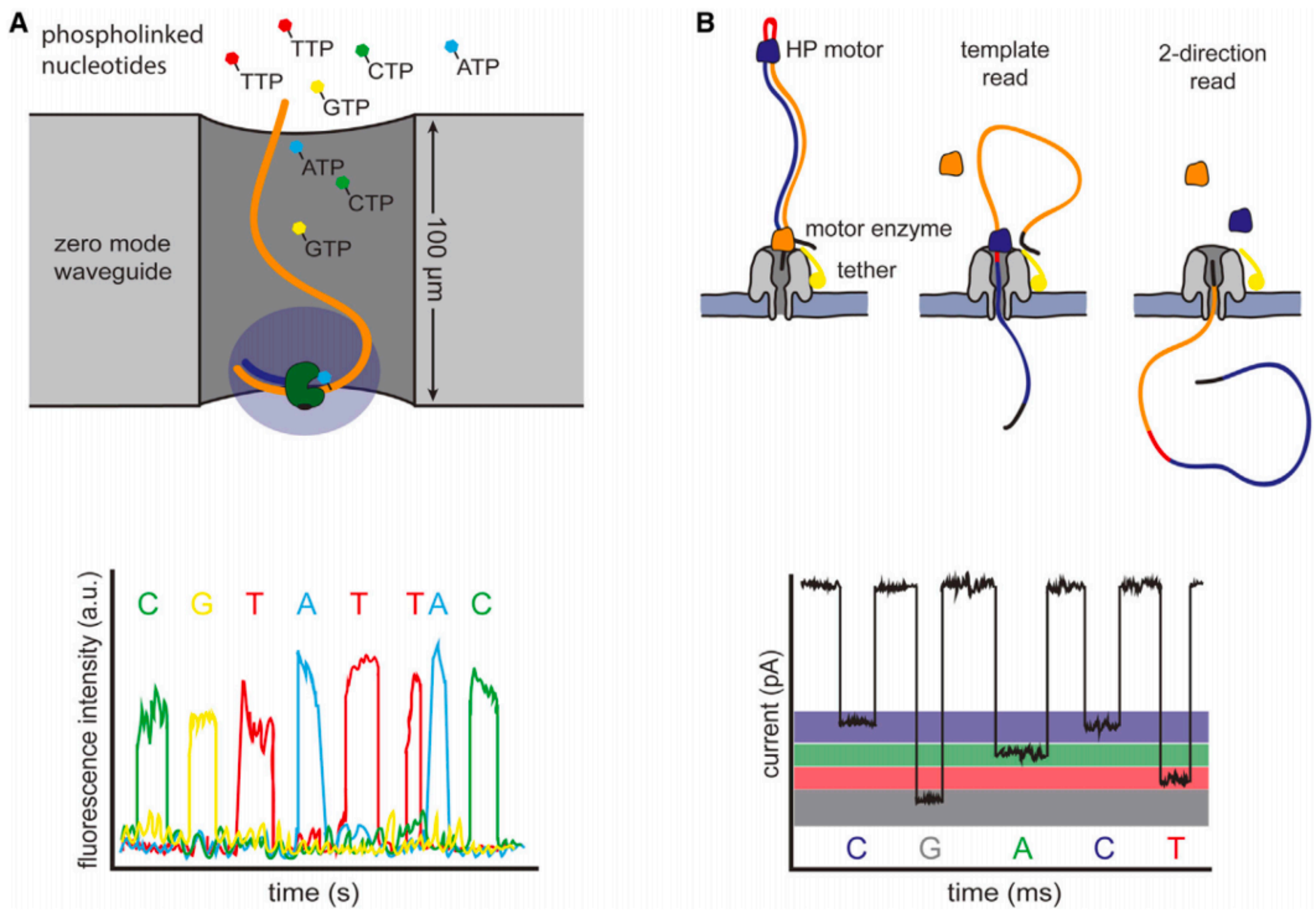
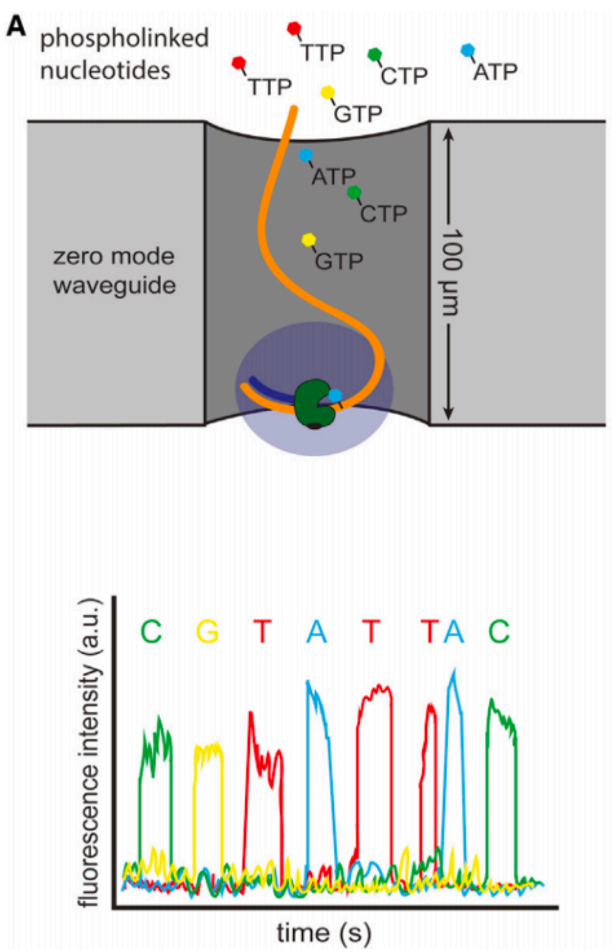


Figure 3. Single Molecule Sequencing Platforms
(A) Pacific Bioscience's SMRT sequencing. A single polymerase is positioned at the bottom of a ZMW. Phosphate-labeled versions of all four nucleotides a
present, allowing continuous polymerization of a DNA template. Base incorporation increases the residence time of the nucleotide in the ZMW, resulting i
detectable fluorescent signal that is captured in a video.
(B) Oxford Nanopore's sequencing strategy. DNA templates are ligated with two adapters. The first adaptor is bound with a motor enzyme as well as a teth
whereas the second adaptor is a hairpin oligo that is bound by the HP motor protein. Changes in current that are induced as the nucleotides pass through the p
are used to discriminate bases. The library design allows sequencing of both strands of DNA from a single molecule (two-direction reads).

Excerpted from Reuter et al 2015 - Molecular Cell

Long read sequencing

PacBio - Pacific Biosciences



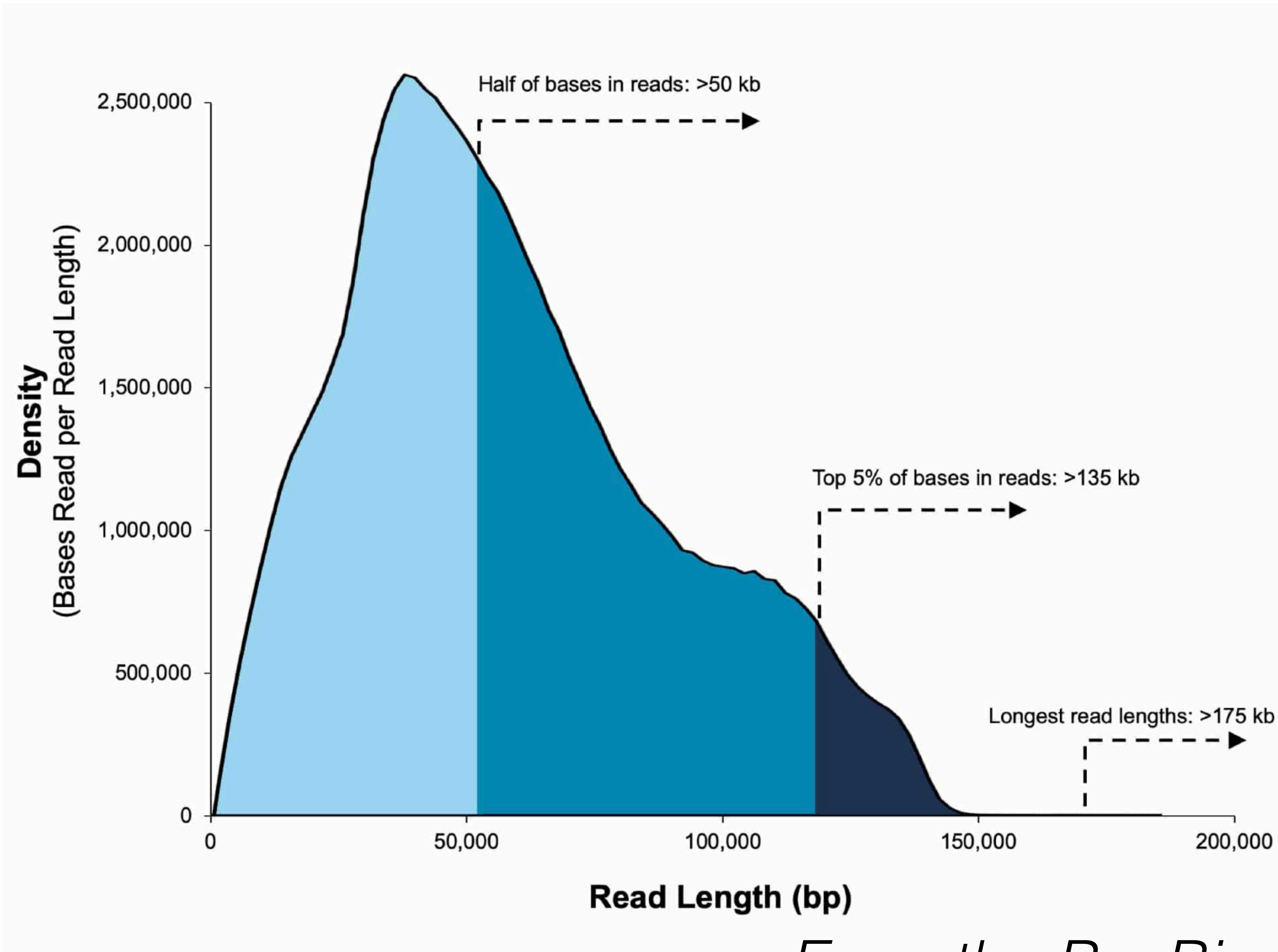
Sequel II

1-10Gb/flowcell

~$500/flowcell

13% error rate

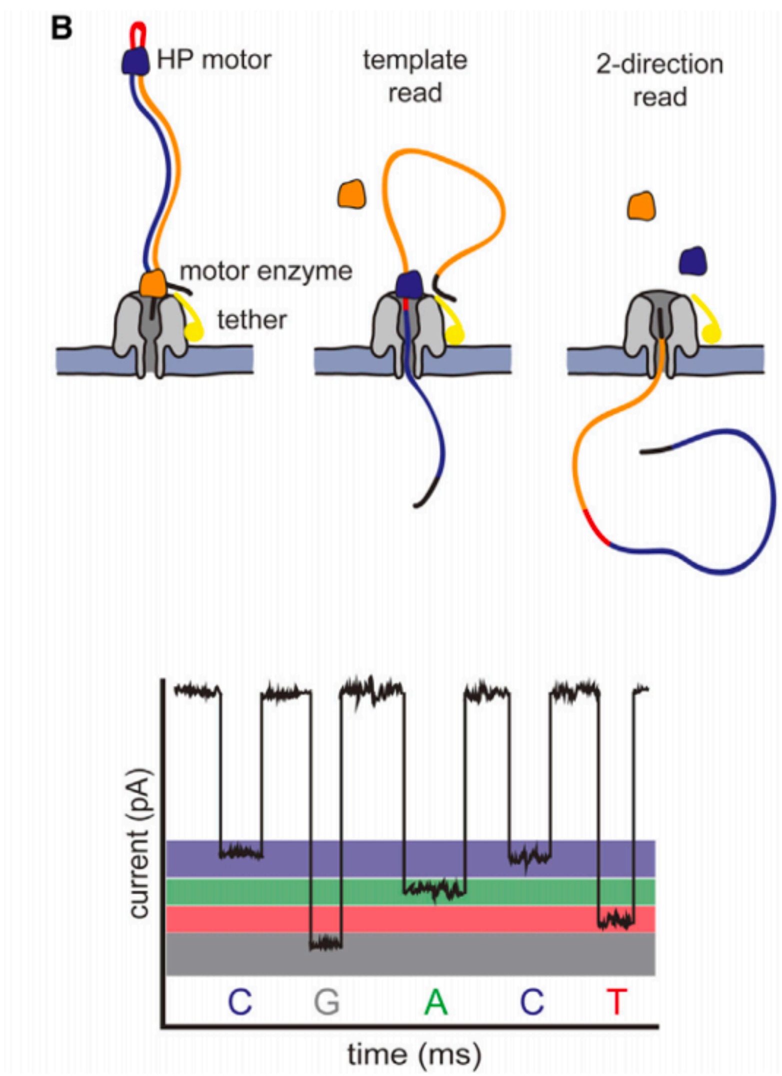
Pacific Biosciences



From the PacBio website

Long read sequencing

Oxford Nanopore

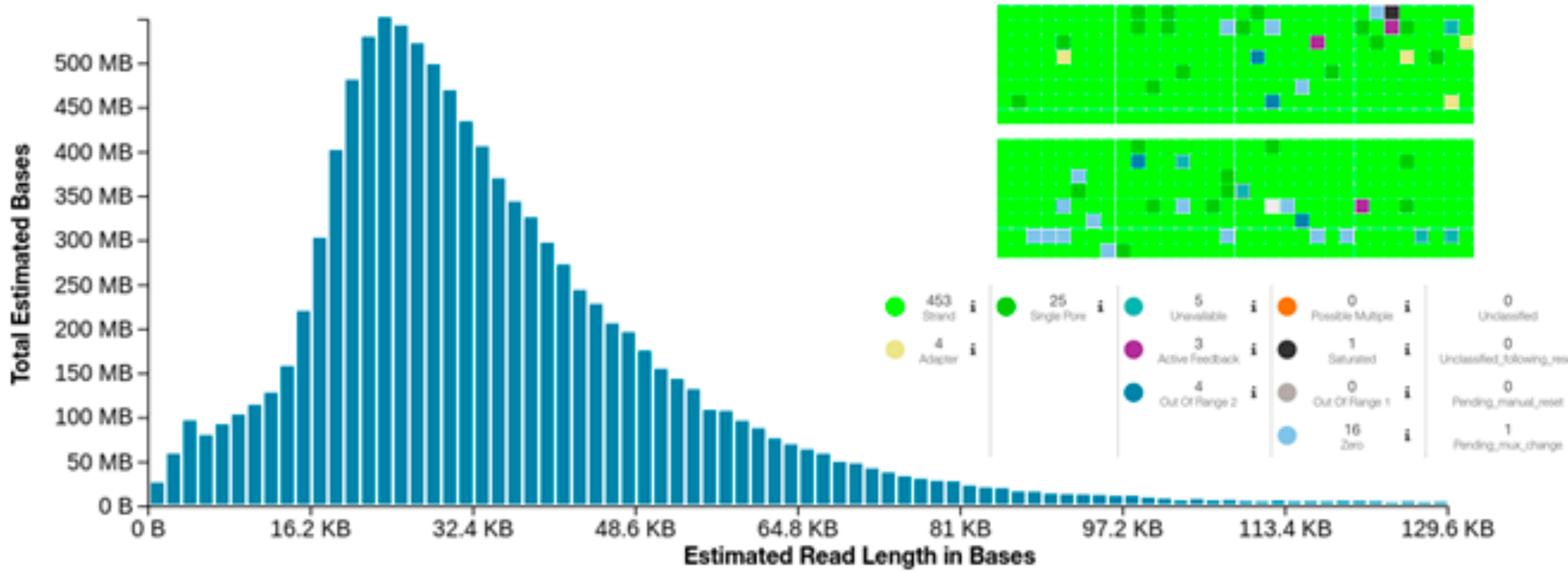


MinION	**PromethION 24**
15-30Gb/flowcell	100-180Gb/flowcell
~$1000/flowcell	~$2000/flowcell

2-13% error rate

Oxford Nanopore

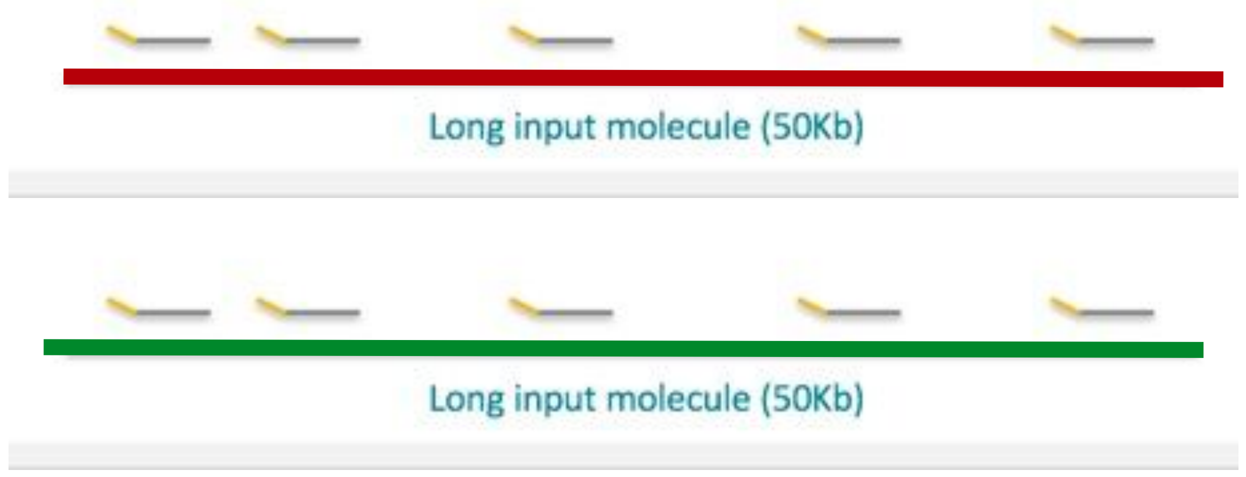
(C) *Eucalyptus albens;* end ligation library prep (SQK-LSK109). Output: 12.50 Gb.



Comparing short and long read technologies

Short Reads		Long Reads	
Pros	**Cons**	**Pros**	**Cons**
Extremely accurate for complex regions	Rely on amplification, which can introduce errors (at a rate of around 10^{-6} - 10^{-7}/bp).	Great for genome assembly • 30-60X coverage from ion torrent or PacBio will produce a nice draft genome.	More difficult library prep
Allele frequencies can be scored at many sites across the genome	Assembling and aligning short reads in repetitive regions is very challenging -> impossible	Can characterise alternate splicing of genes.	Too expensive to be used for population level sequencing.
Very cost-effective	Both large and small structural variants pose difficulties	Structural rearrangement discovery and genotyping.	High error rate.

Synthetic long reads



Barcodes read originating from individual DNA molecules

Sequence with Illumina reads

Original molecule can be reconstructed using the barcodes

Potentially very useful for genome assembly and phasing

Synthetic long reads

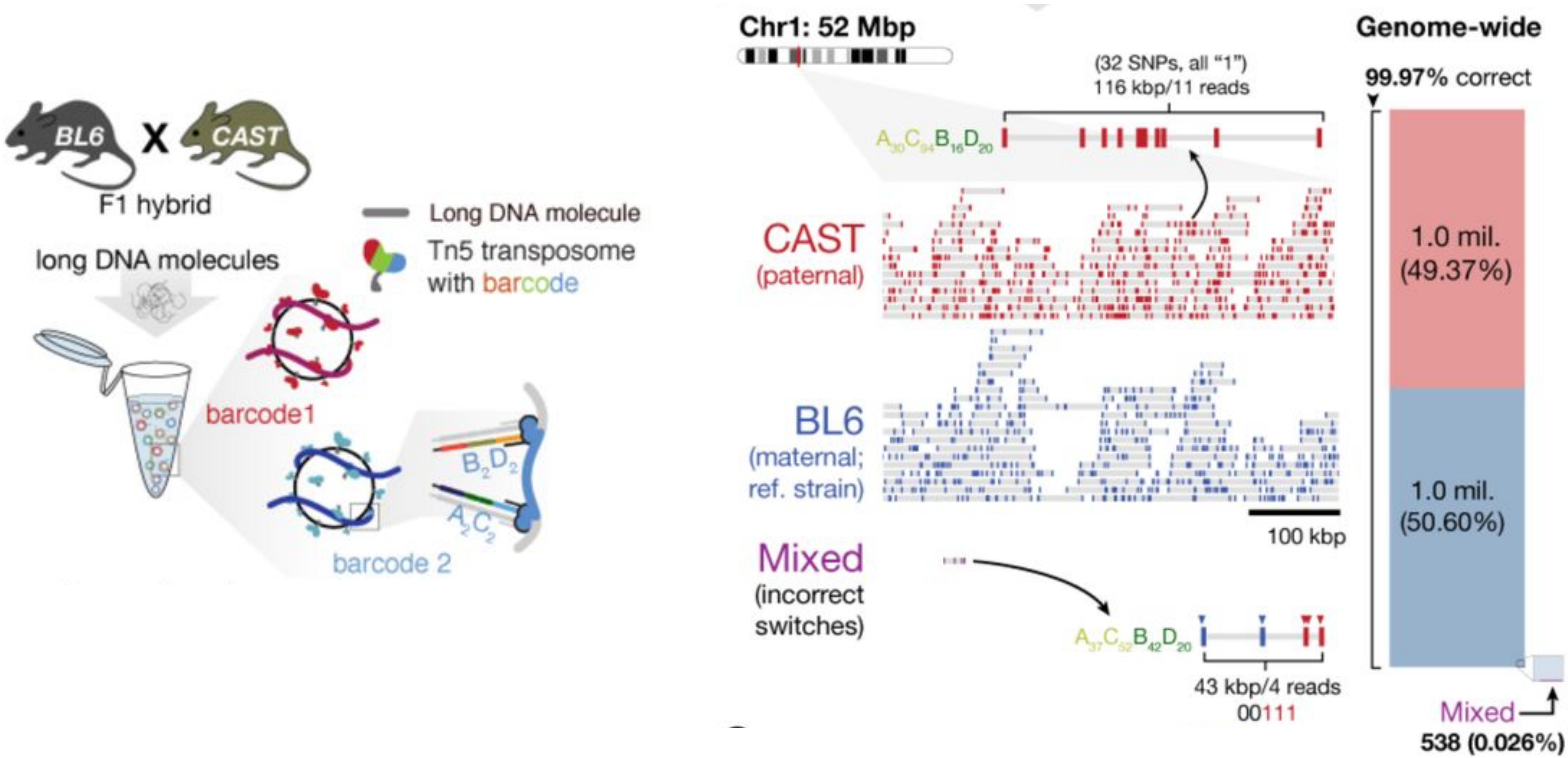


Figure adapted from Meier et al *Preprint* https://www.biorxiv.org/content/10.1101/2020.05.25.113688v2.full

Flavours of DNA/RNA sequencing

Whole Genome Sequencing

Pool Seq

RNAseq

Amplicon Sequencing (GT-seq)

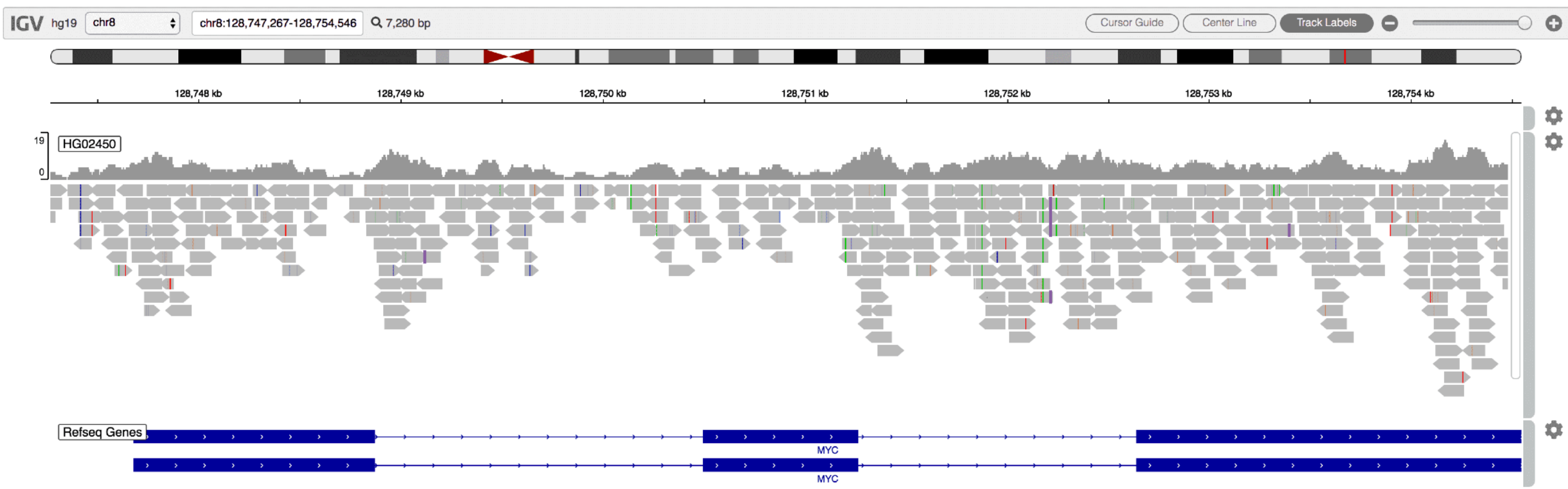
Sequence Capture

Reduced-Representation Sequencing (RADseq/GBS/ RADcapture)

Whole Genome Sequencing

Randomly sheer DNA and sequence all fragments

May use double-stranded nuclease treatment to reduce repetitive elements

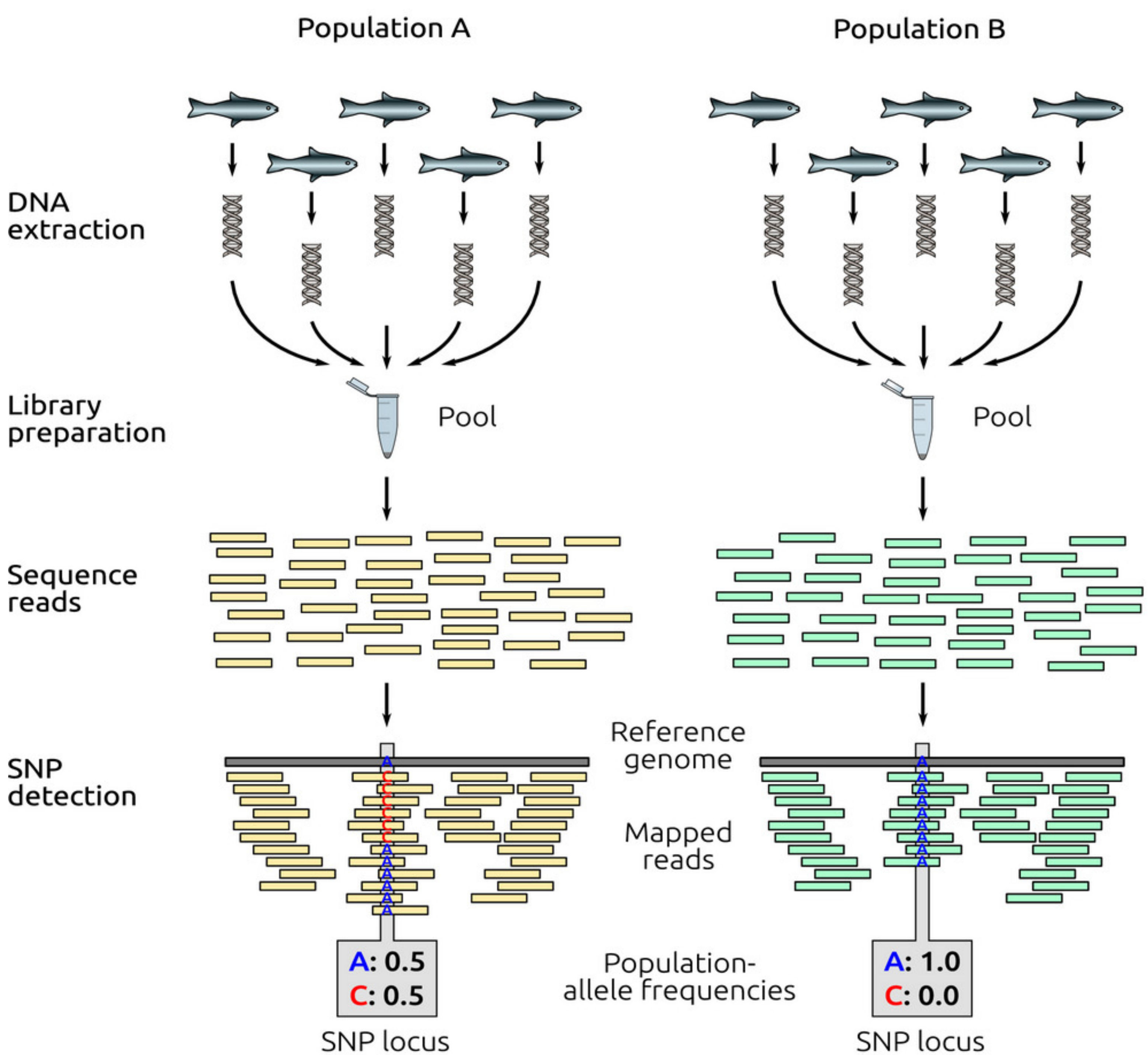


Screen shot from the Integrated Genomics Viewer

Whole Genome Sequencing

Pros	Cons
All sites possible	Comparatively expensive per sample
Simple library prep	Storage and bioinformatics challenging with lots of samples

Pool Seq

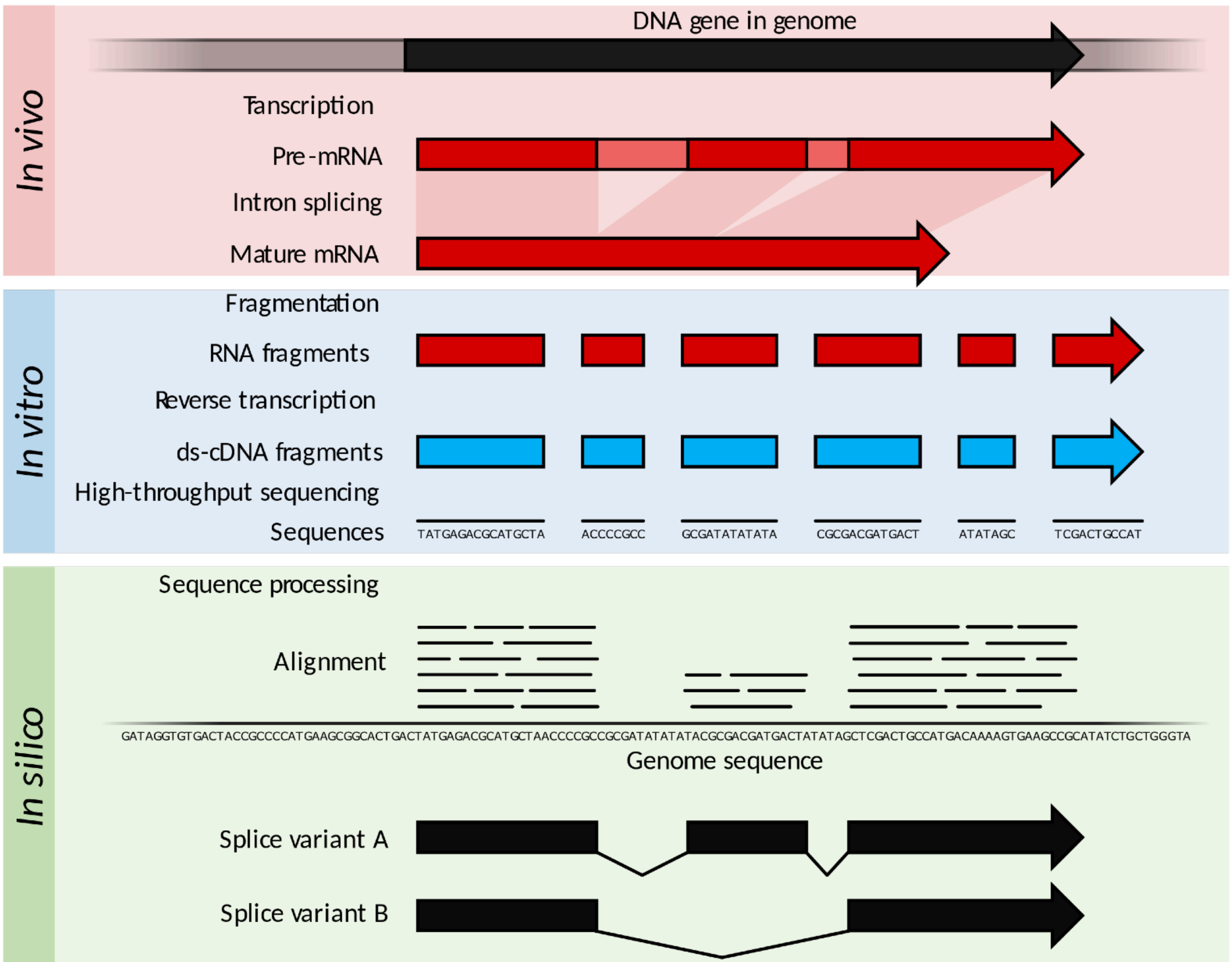


Adapted from Fuentes-Pardo & Ruzzante 2017 Mol. Ecol

Pool Seq

Pros	Cons
All sites possible	Limited analysis options
Simple library prep	No haplotype information
Cheaper than individual WGS	Best in cases where # samples > # reads

RNAseq



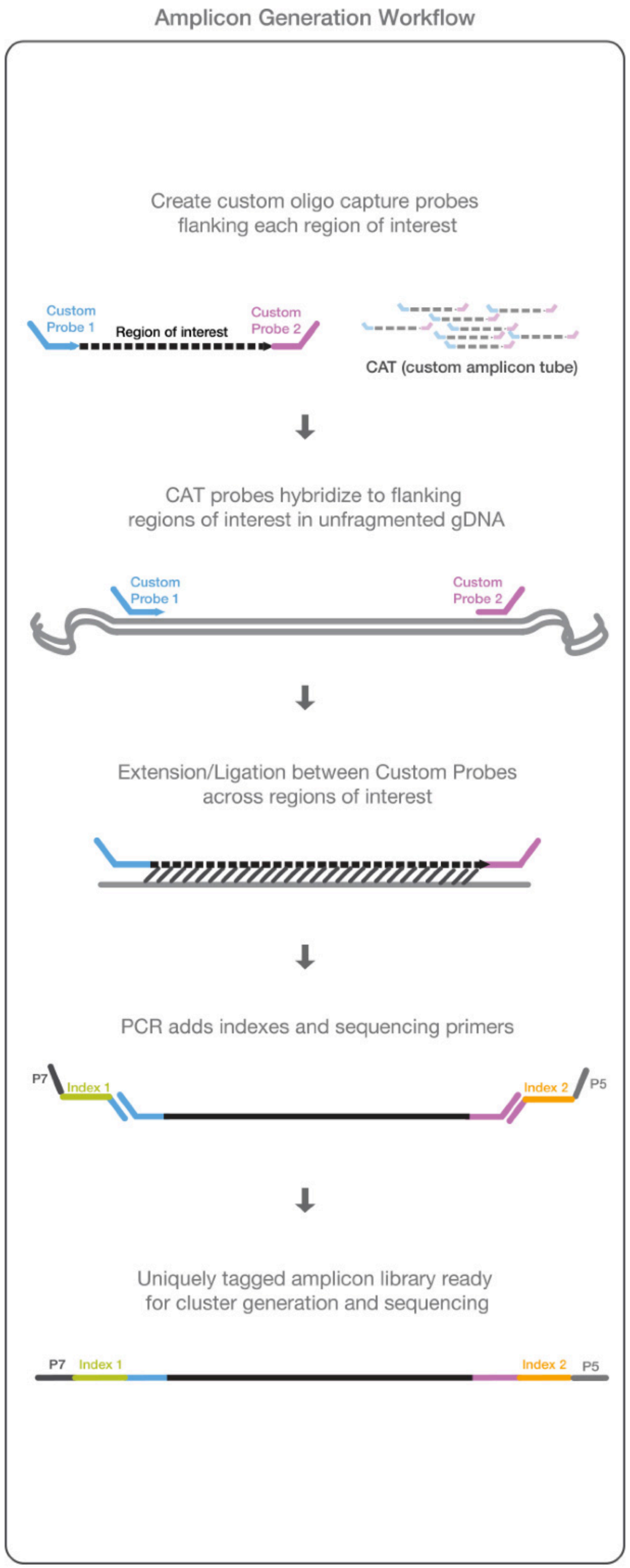
From Wikipedia

RNAseq

Pros	Cons
Many sites and only in genes	Expression differences complicate SNP calling
Also get expression information	Expensive for pop gen level sampling
Relatively easy to assemble	Difficult library prep (or so I'm told!)

Amplicon Sequencing

- Use PCR to amplify target DNA. Sequence many barcoded samples in one lane.
- Used to characterise microbiome by sequencing 16s rRNA



Amplicon Sequencing

Pros	**Cons**
Get incredible depth at single locus	Limited to one or few loci
Simple bioinformatics.	Mutations in primer site don't sequence

GT-seq

- Genotyping by Thousands
- Based on Amplicon sequencing
- Multiplex PCR amplify ~200 known SNPs and then sequence pooled PCR products.
- Very cheap ($1/sample), and bioinformatically simple.
- Useful for genotyping thousands or tens of thousands of samples.
- Complicated initial set-up.

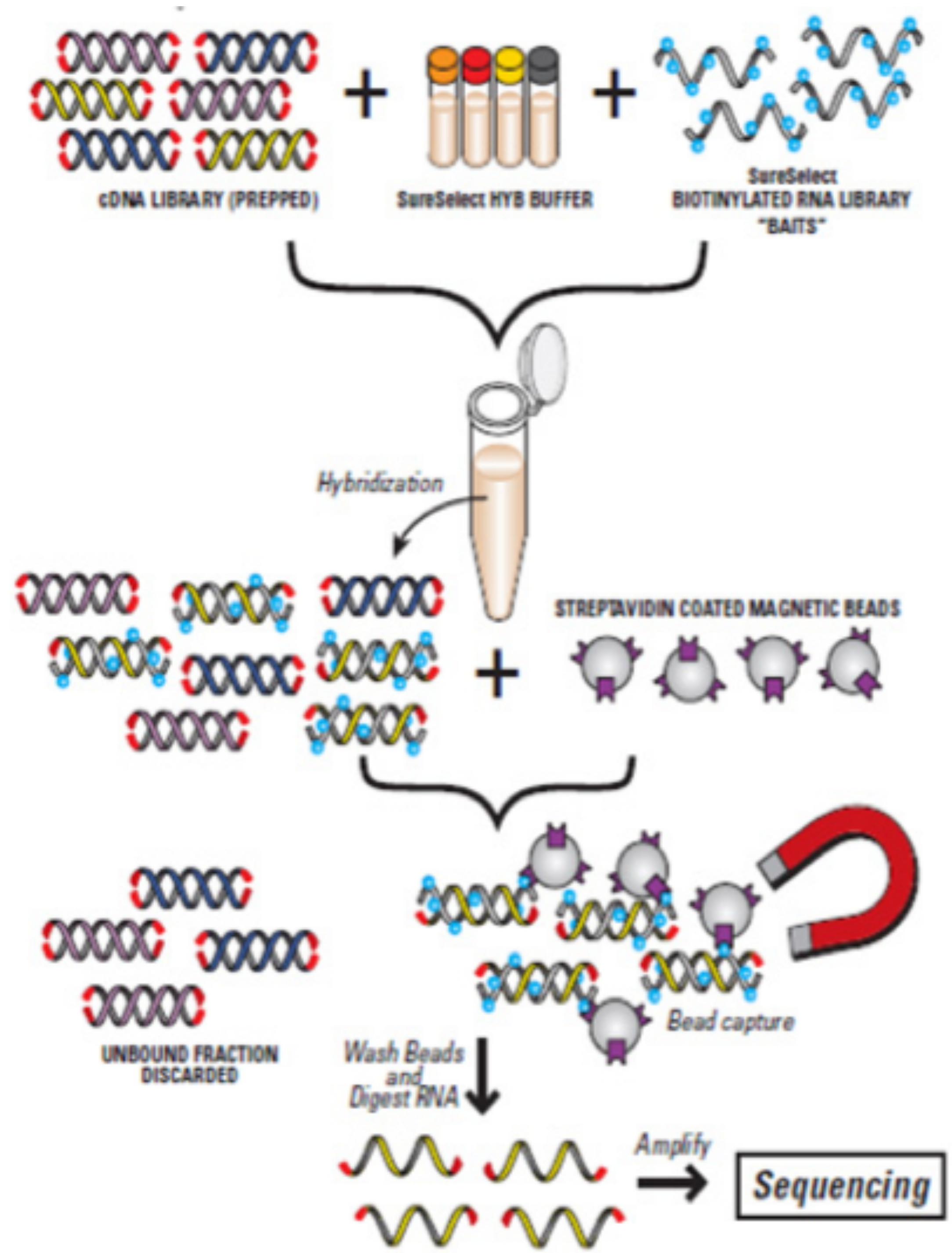
Campbell et al. 2015 *Mol Ecol Resour*.

Sequence Capture

- Design probe sequences from genome resources, synthesis attached to beads

- Make WGS library, hybridize with probe set. Matching sequence will be captured, all others washed away

- Collect capture sequence, amplify and sequence



Sequence Capture

Pros	Cons
Relatively cheap per sample	Requires designing probes
Good depth at targeted sites	Long library prep

Reduced Representation Sequencing

Instead of sequencing the whole genome, it can be sufficient to sequence just a part of it

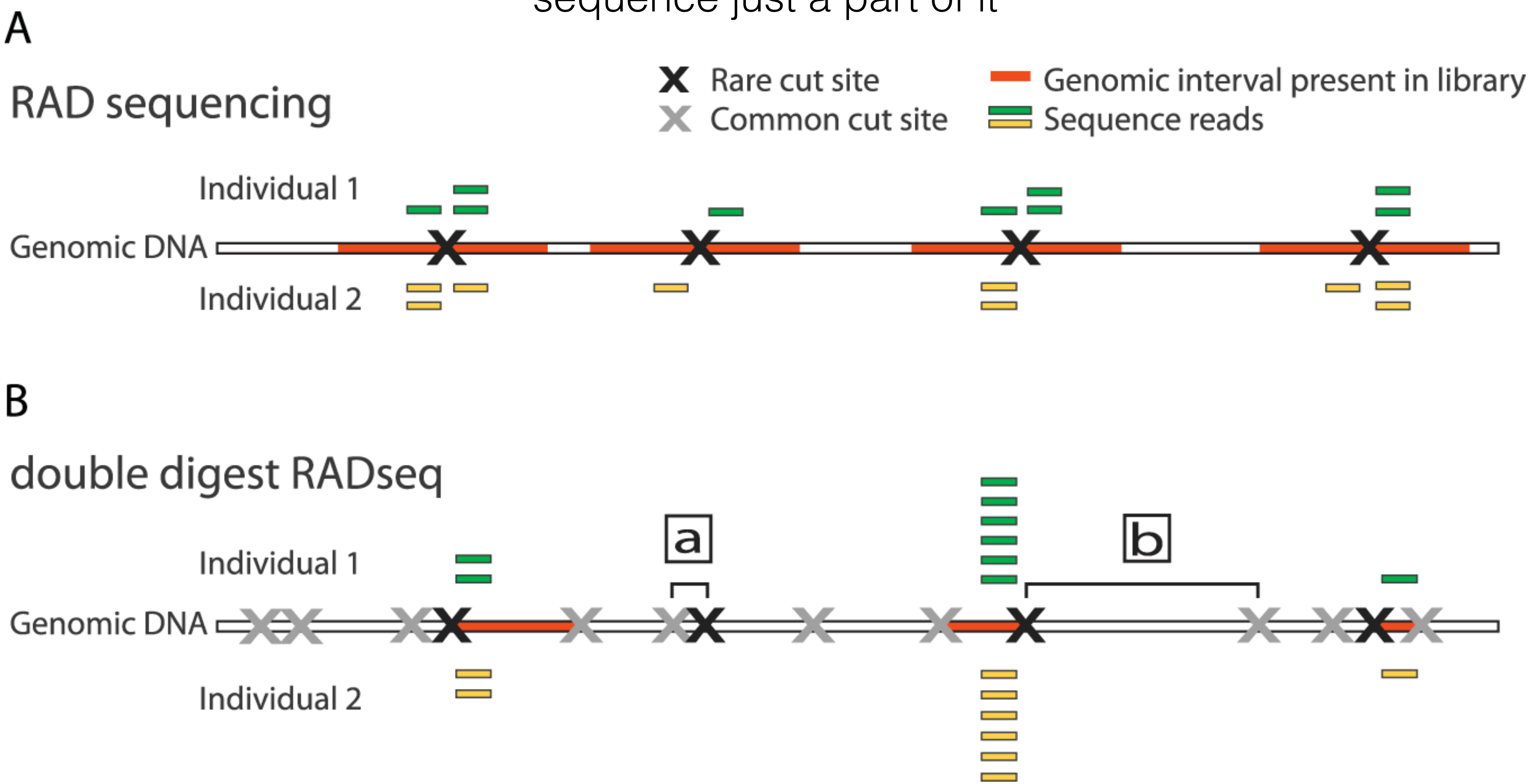
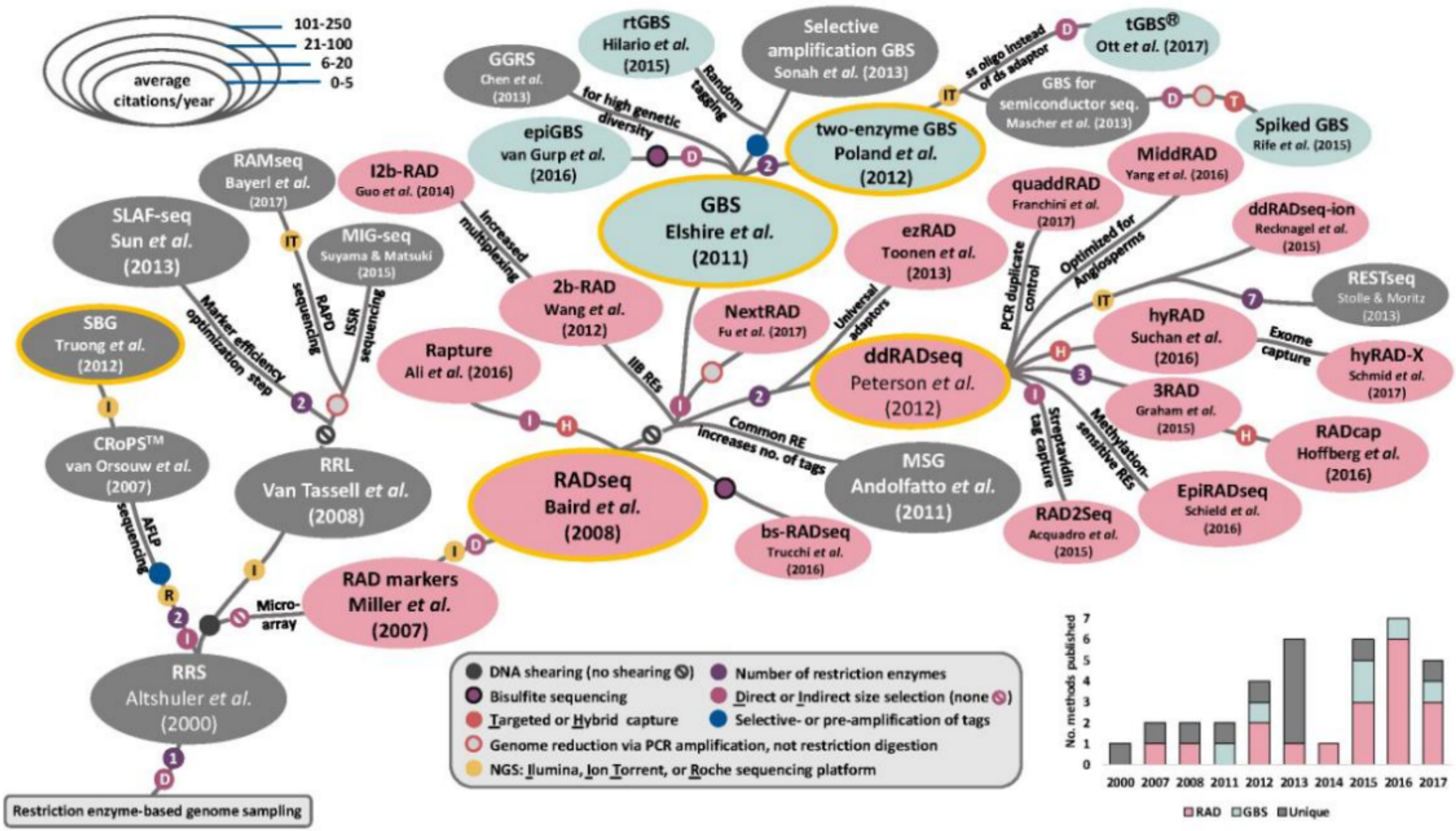


Figure from Peterson et al PLoS One 2012

Reduced Representation Sequencing

Pros	**Cons**
Quick library prep for hundreds of samples	Relatively sparse SNPs compared to other methods - limiting analysis options
Comparatively cheap per sample cost	Can have problems overlapping different library preps

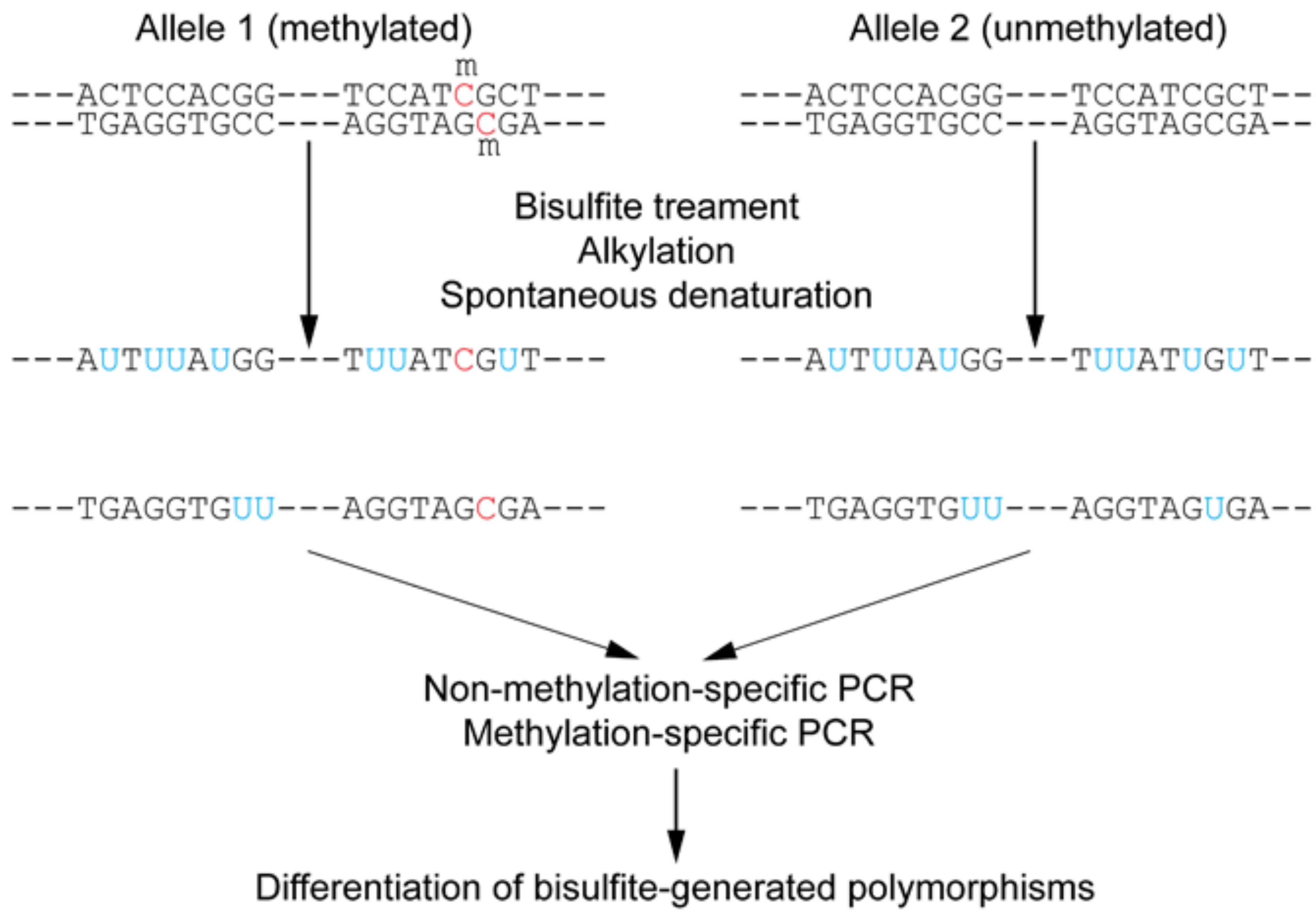
There is a huge diversity of reduced representation approaches



Bisulphite Sequencing

Unmethylated cytosines are converted to **U**racil

Methylated **C**pG sites are unchanged and are detected as polymorphisms



How to choose?

The different technologies and methodologies have different pros and cons

What you use will obviously be informed by budget, but the biological question should also drive your choice

How to choose?

For example,

If you wanted to estimate demographic history from the distribution of allele frequencies, a reduced representation method might suffice to obtain an estimate of the site frequency spectrum

Or, if you want to perform a genome scan, looking at how haplotype frequencies varied among populations, you'd probably need deeper, whole genome information - it all depends on the questions you are tackling

Further reading

PDFs are available on the GitHub page for this topic:

Andrews, K. R., Good, J. M., Miller, M. R., Luikart, G., & Hohenlohe, P. A. (2016). Harnessing the power of RADseq for ecological and evolutionary genomics. *Nature Reviews Genetics*, *17*(2), 81.

Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., & Hoekstra, H. E. (2012). Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PloS one*, *7*(5), e37135.

Shendure, J., Balasubramanian, S., Church, G. M., Gilbert, W., Rogers, J., Schloss, J. A., & Waterston, R. H. (2017). DNA sequencing at 40: past, present and future. *Nature*, *550*(7676), 345-353.