

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

Website

<https://khodgins.github.io/BioinformaticsWorkshopMonash2021/>

Bioinformatics for HDR students in the School of Biological Sciences, Monash University

INTRODUCTION TO BIOINFORMATICS – 2021



General info

Questions

Topics:

- 1. Servers
- 2. Cmdline
- 3. Preprocessing
- 4. Sequence Alignment
- 5. Assembly
- 6. SNP Calling
- 7. Population genomics
- 8. Plotting in R
- 8.1. PCA

Description

The purpose of this course is to provide biology graduate students with the theoretical knowledge and practical skills for the analysis of next generation sequence data. The course will provide an overview of sequencing approaches, genome assembly, sequence alignment techniques and variant calling. The manipulation of genomic datasets using the commandline will be taught through interactive tutorials. The unit will be presented as a series of short lectures and lab exercises over a three day period.

Instructors

Dr Kathryn Hodgins and Dr Paul Rathnay

Unit Objectives

1. Introduction: Scope of course, goals and overview of technology
2. Introduction to command line programming
3. Sequence file formats and quality checking/trimming
4. Alignment: algorithms and tools
5. Assembly: transcriptome and genome assembly
6. SNP and variant calling

Meet your neighbor!

Introduce yourself to the people beside you

What is your experience with genomics?

What do you want to get out of the course?

Learning outcomes

- Understand the differences between major sequencing technologies
- Identify pros and cons of different sequencing platforms
- Understand different sequencing approaches
- Identify the pros and cons of using different sequencing approaches

Goals

Raw sequence data

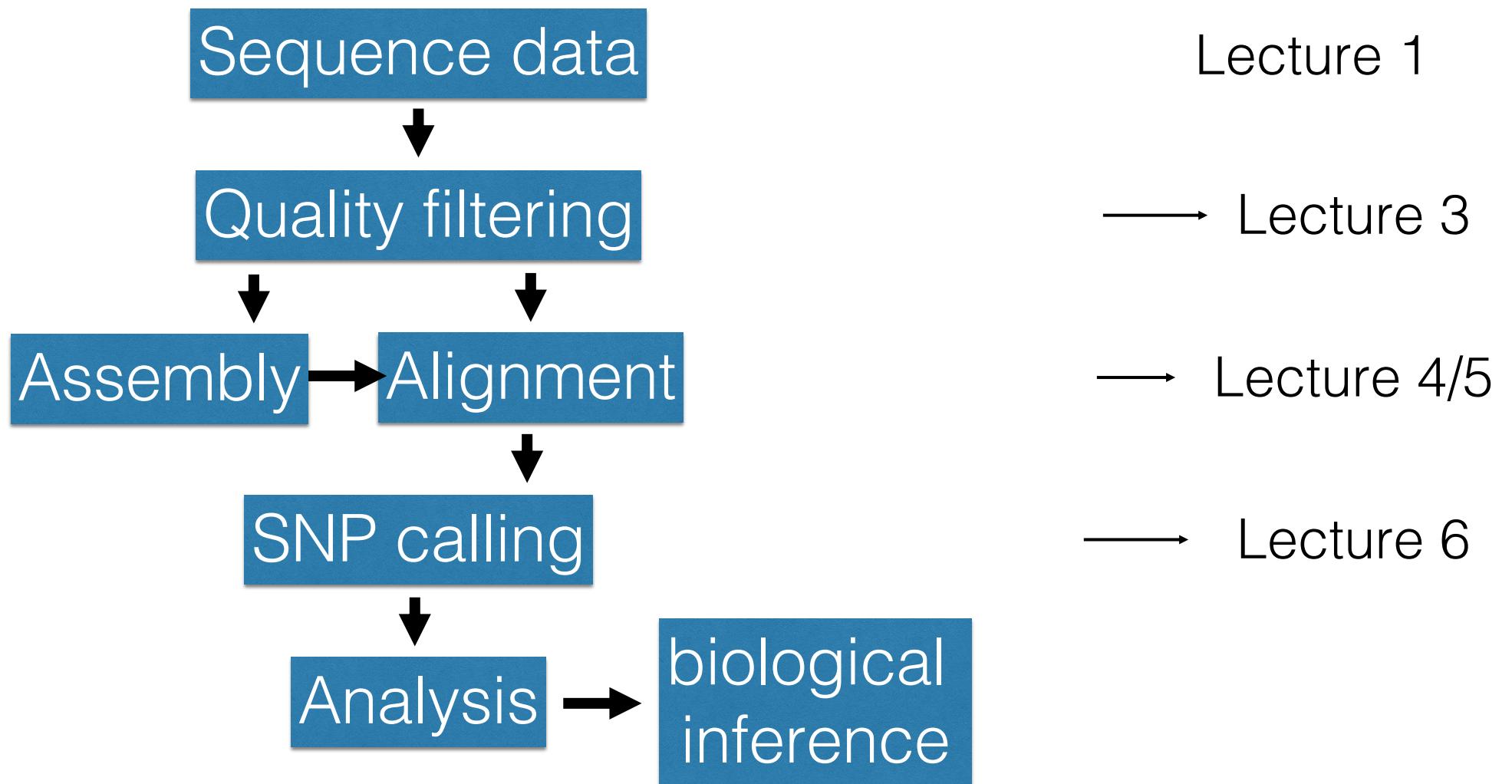


????



Results and Figures

Goals



A brief history of DNA sequencing

Technological milestones (Box 1 and 2 - Shendure et al 2017 Nature)

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*

A brief history of DNA sequencing

Technological milestones (Box 1 and 2 - Shendure et al 2017 Nature)

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

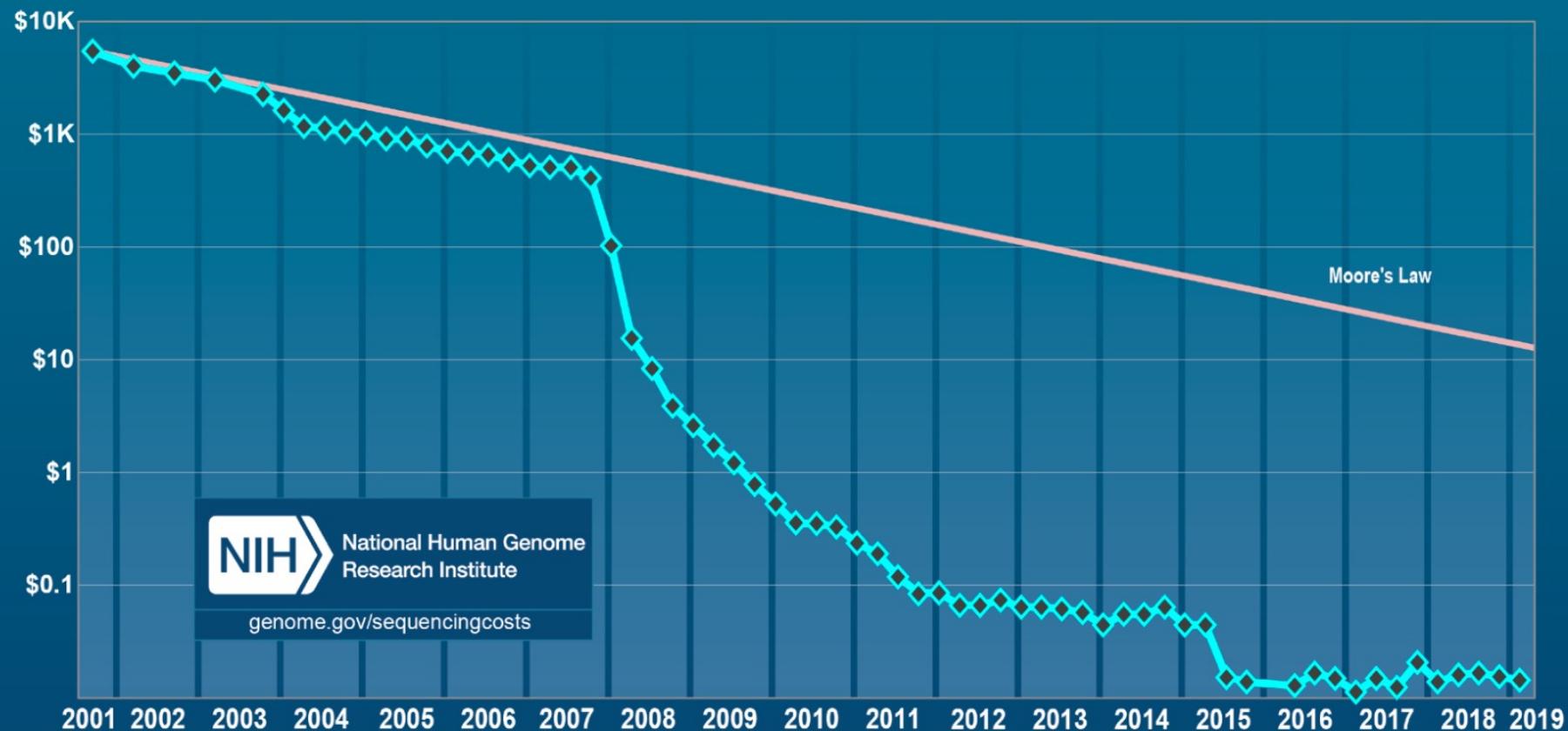
2010: Single-base resolution electron tunnelling through a solid state detector

2011: Semiconductor sequencing by proton detection

2012: Reduction to practice of nanopore sequencing

2012: Single-stranded library preparation method for ancient DNA

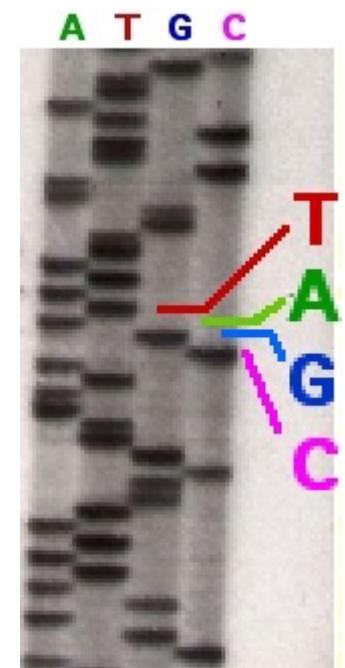
Cost per Raw Megabase of DNA Sequence



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

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 fluorescently labeled dideoxy bases
 - Dominant sequencer up until 2007
 - Only one fragment sequenced per reaction
 - Still used for sequencing individual PCR products



Sanger

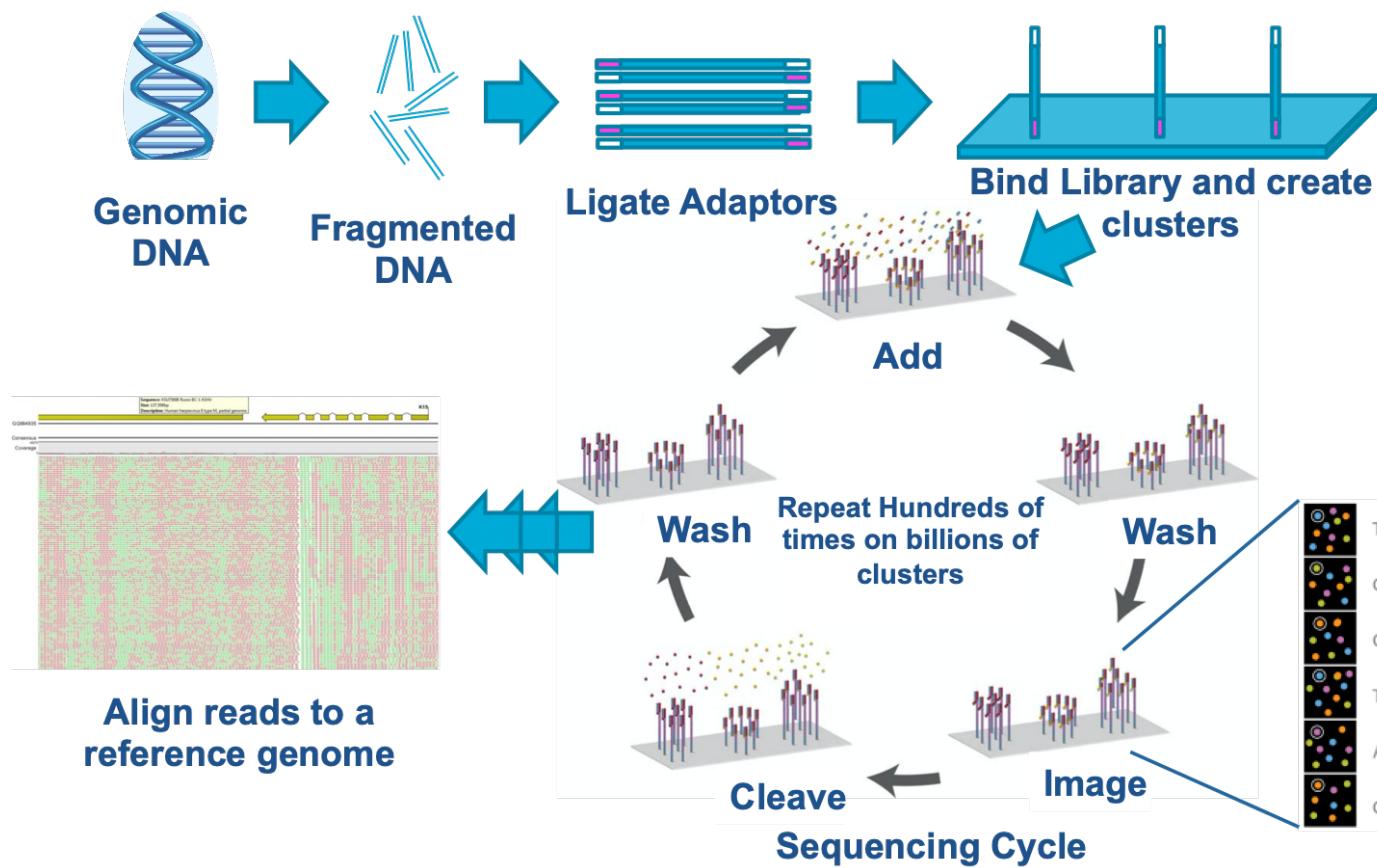
Second/third generation sequencing

- Sequences many molecules in parallel
- Don't need to know anything about the sequence to start.
- Main technologies:
 - Illumina
 - Ion torrent
 - 454 (Pyrosequencing)
 - PacBio
 - Nanopore

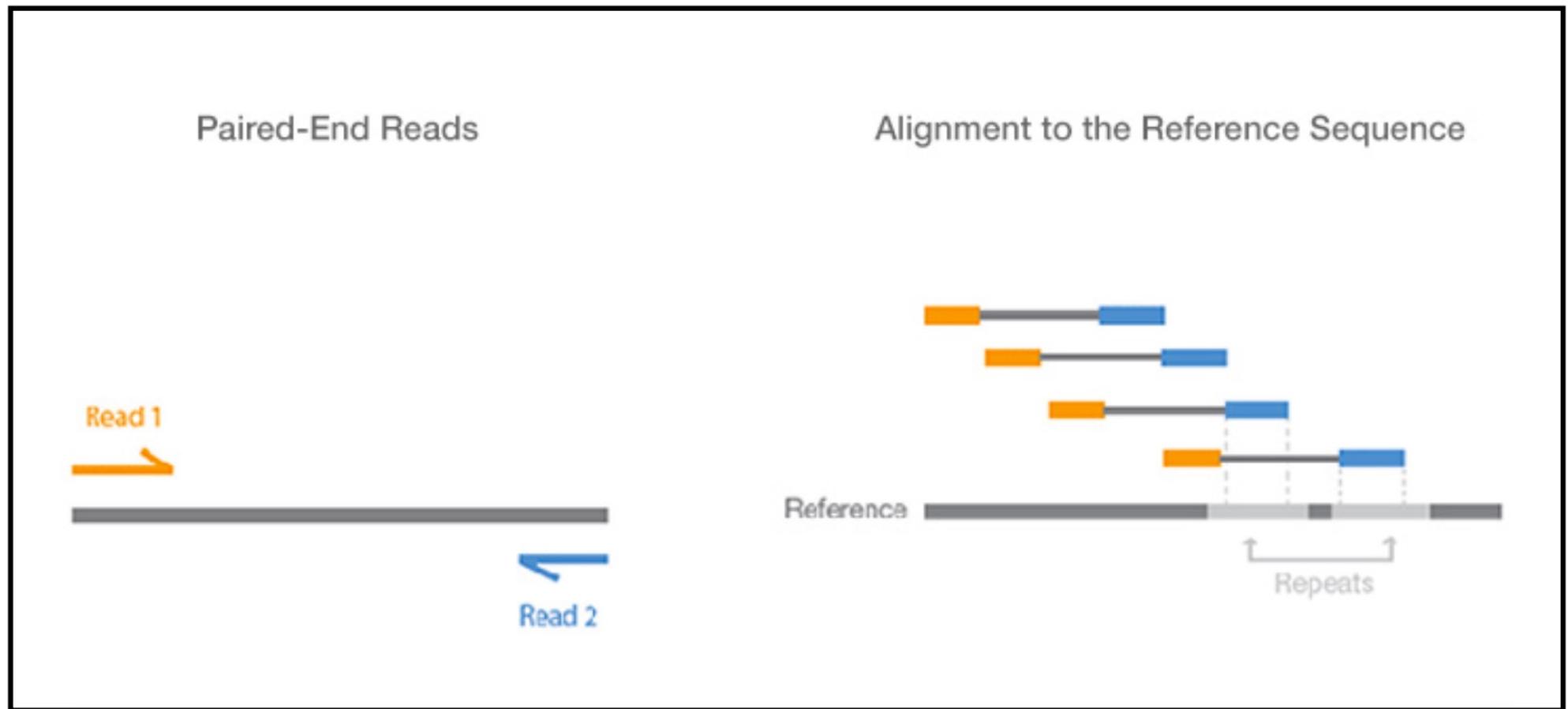
Second/third generation sequencing

Technology	Read Length	Accuracy	Bases/run	Uses
Illumina	50-600bp	99.9%	500-600GBase	Resequencing General depth
Oxford Nanopore	5kb-100kb	85-95%	10-30GBase	Microbial genomes Genome assembly
PacBio	10kb-40kb	85-90% 99.5% (HiFi)	5-10Gbase	Genome assembly Structural variants

Illumina sequencing



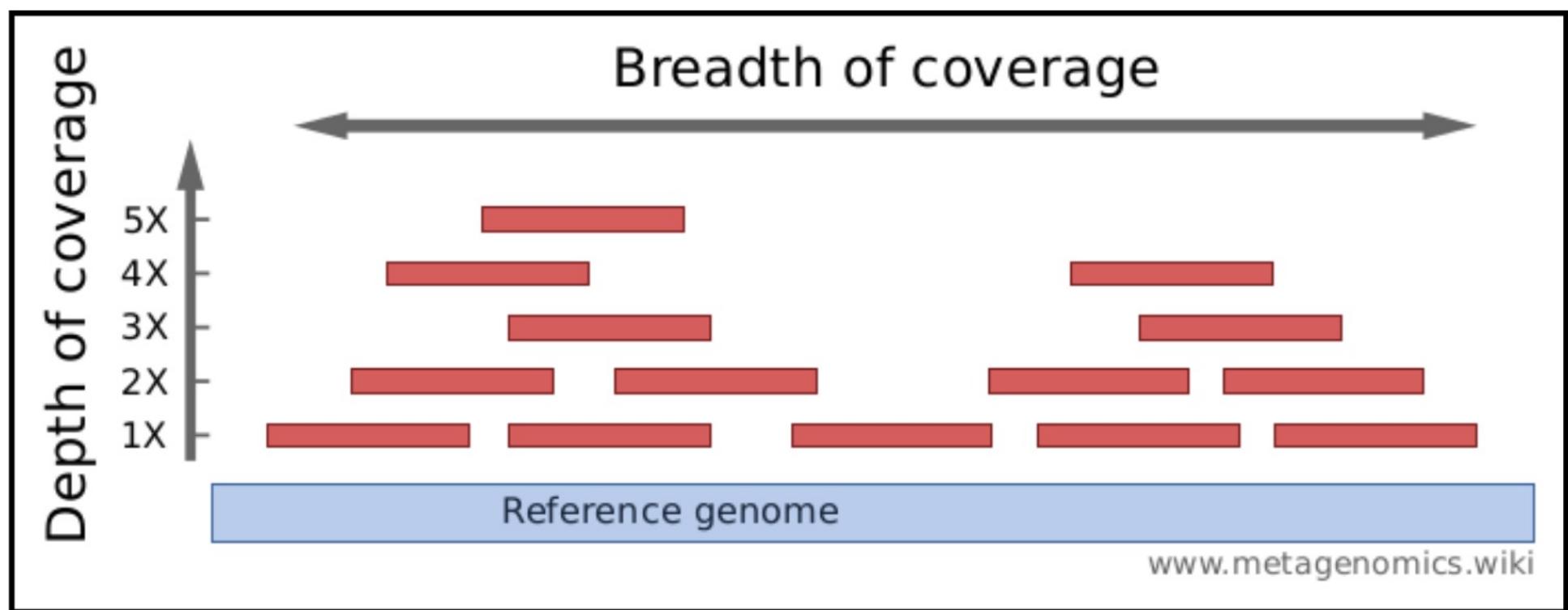
Paired-end reads improve mapping



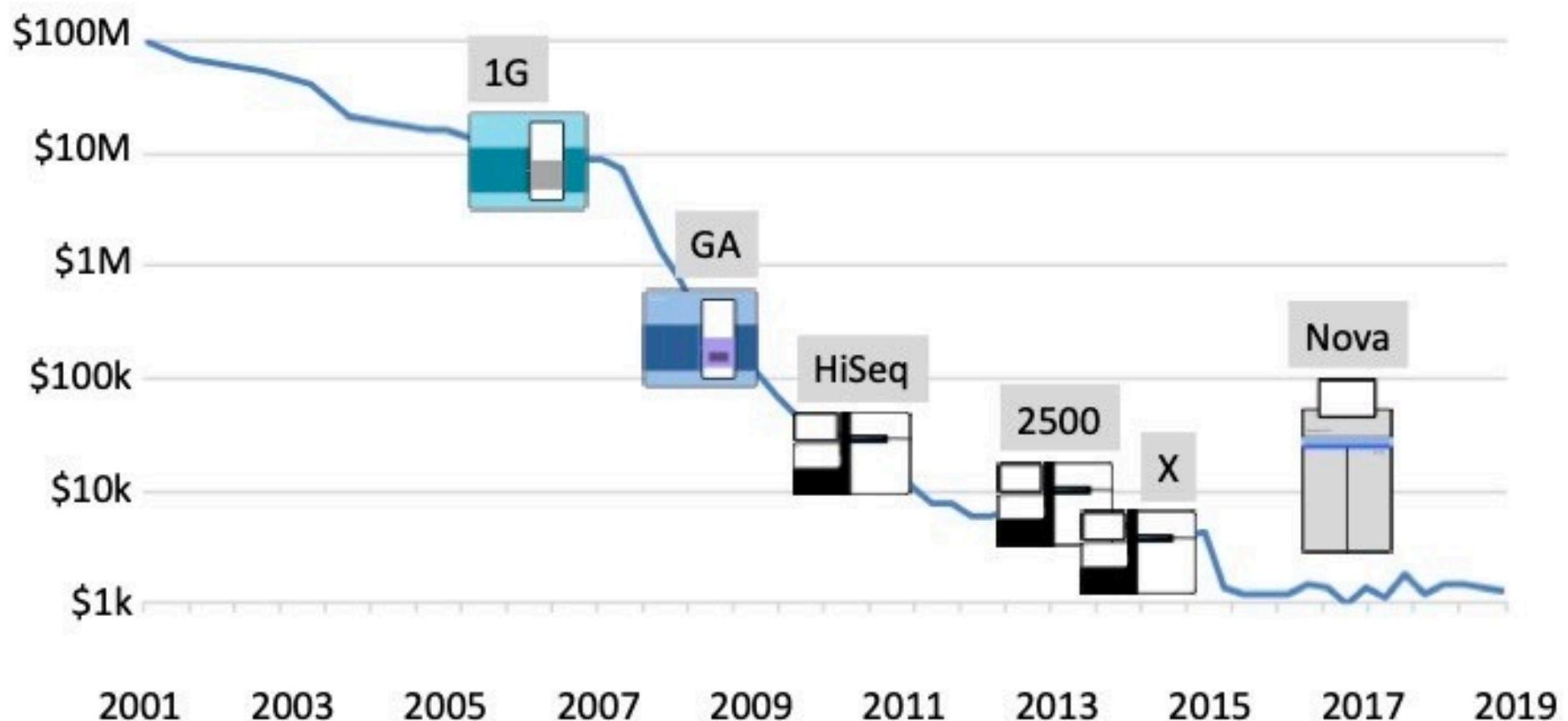
From Illumina website

Short reads provide high depth and breadth at lower cost

Important concepts



Production cost per 30x Human genome over 18 years



Illumina Machines



MiSeq

8Gb/run

~\$1500/lane



HiSeq 4000

50Gb/lane

~\$3000/lane



Novaseq 6000

500-600Gb/lane

~\$8000/lane

Challenges of short read technology

- Rely on amplification, which can introduce errors (10^{-6} - 10^{-7}).
- Assembling and aligning reads challenging in repetitive regions
- Difficulty with both large and small structural variants.

Long read sequencing

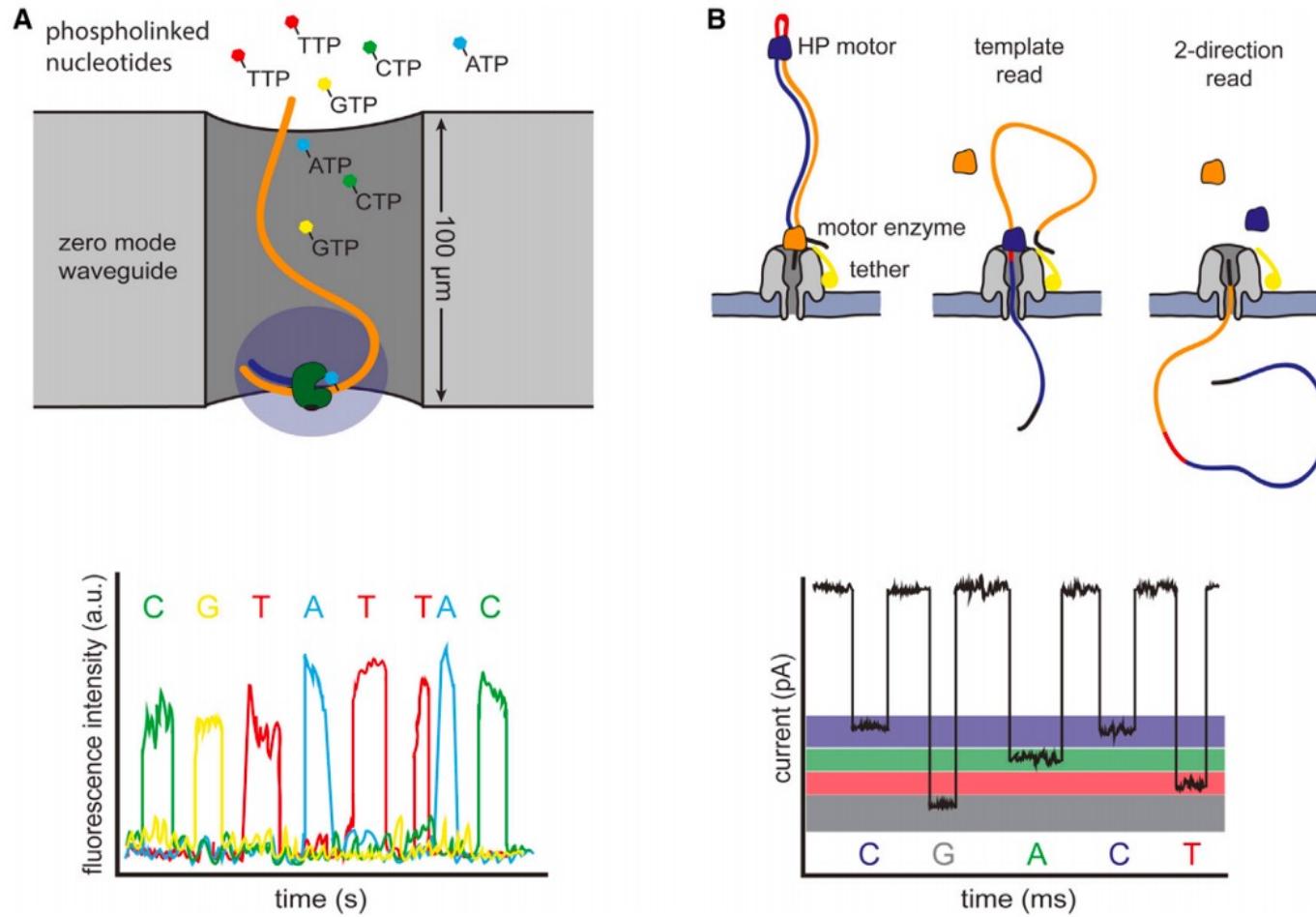


Figure 3. Single Molecule Sequencing Platforms

(A) Pacific Bioscience's SMRT sequencing. A single polymerase is positioned at the bottom of a ZMW. Phosphate-labeled versions of all four nucleotides are present, allowing continuous polymerization of a DNA template. Base incorporation increases the residence time of the nucleotide in the ZMW, resulting in a detectable fluorescent signal that is captured in a video.

(B) Oxford Nanopore's sequencing strategy. DNA templates are ligated with two adapters. The first adapter is bound with a motor enzyme as well as a tether, whereas the second adapter is a hairpin oligo that is bound by the HP motor protein. Changes in current that are induced as the nucleotides pass through the pore are used to discriminate bases. The library design allows sequencing of both strands of DNA from a single molecule (two-direction reads).

Oxford Nanopore



MinION

15-30Gb/flowcell

~\$1000/flowcell



PromethION 24

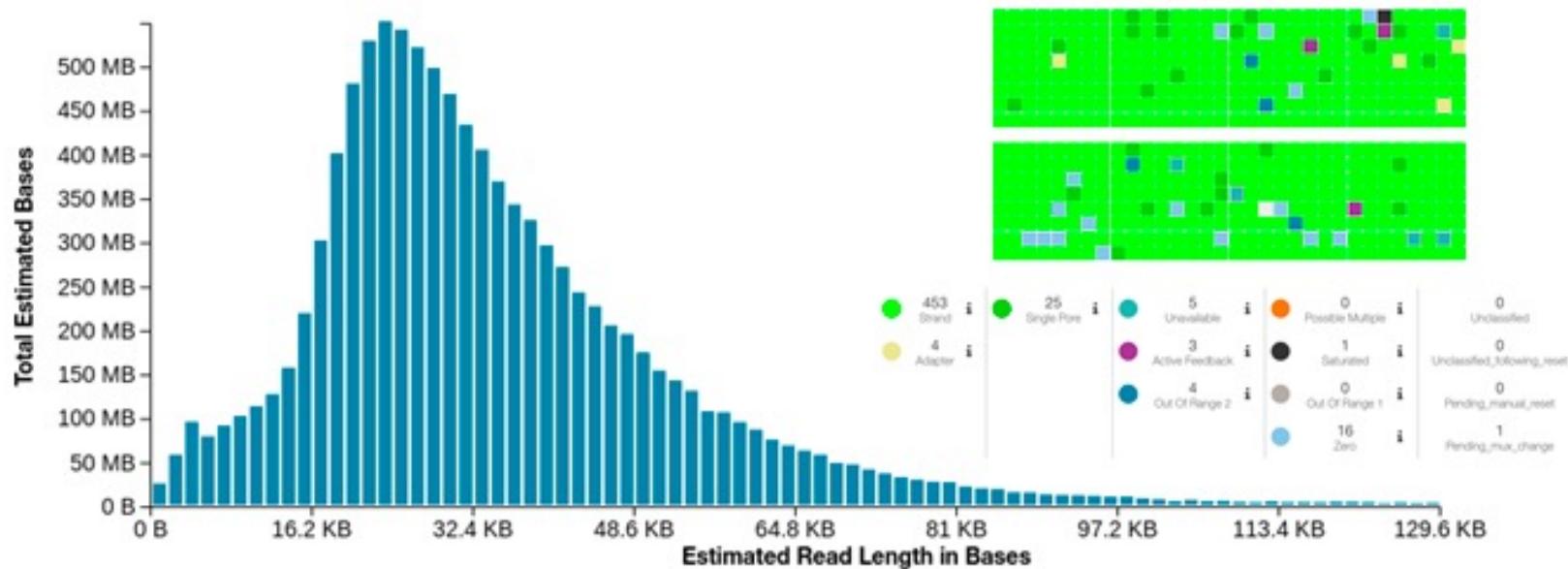
100-180Gb/flowcell

~\$2000/flowcell

2-13% error rate

Oxford Nanopore

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



Pacific Biosciences



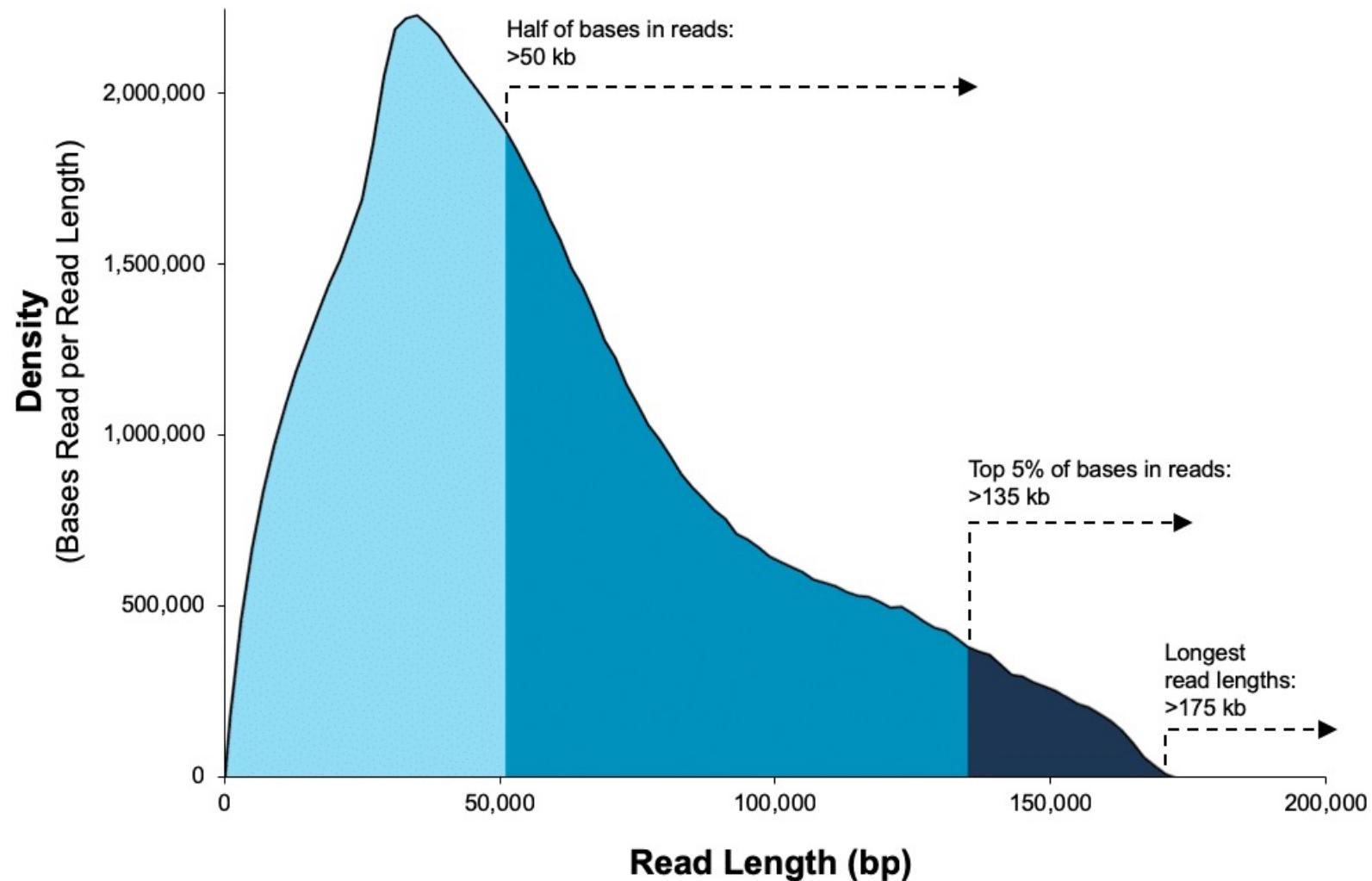
Sequel II

1-10Gb/flowcell

~\$500/flowcell

13% error rate

Pacific Biosciences



Challenges of long read technology

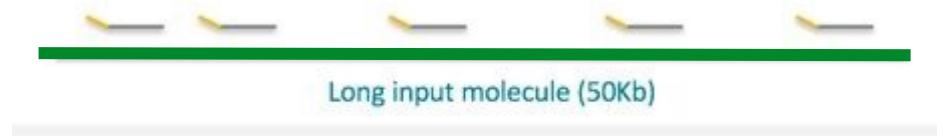
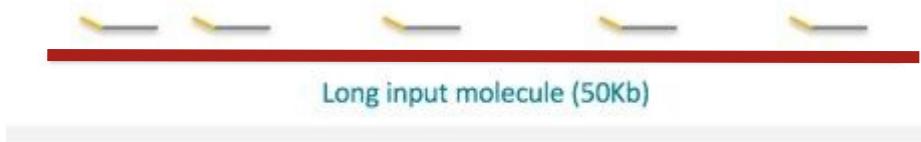
- Too expensive to be used for population level sequencing.
- High error rate.

Uses of long reads

- Genome assembly.
 - 30-60X coverage ion torrent or PacBio will produce a nice draft genome.
- Alternate splicing of genes.
- Structural rearrangement discovery and genotyping.

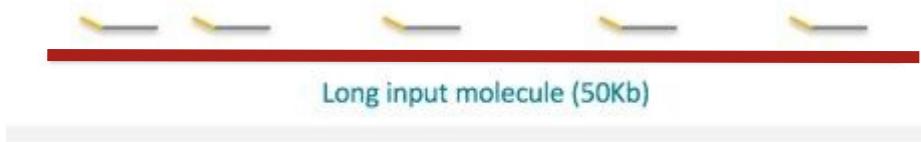
Synthetic long reads

- Barcodes read originating from individual DNA molecules
- Uses Illumina sequencing
- Tells you which reads are physically nearby and on the same strand



Synthetic long reads

- Used for genome assembly or phasing



Flavours of sequencing

- Whole Genome Sequencing
- Pool Seq
- RNAseq
- Amplicon Sequencing
- Sequence Capture
- Reduced-Representation Sequencing
(RADseq/GBS/RADcapture)
- GT-seq

Think - Pair - Share

What kind of sequencing are you planning on using
the future?

What are you using it for?

Why did you choose that method?

Whole Genome Sequencing

- Randomly sheer DNA and sequence all fragments
- May use double-stranded nuclease treatment to reduce repetitive elements

Pros:

- All sites possible
- Simple library prep

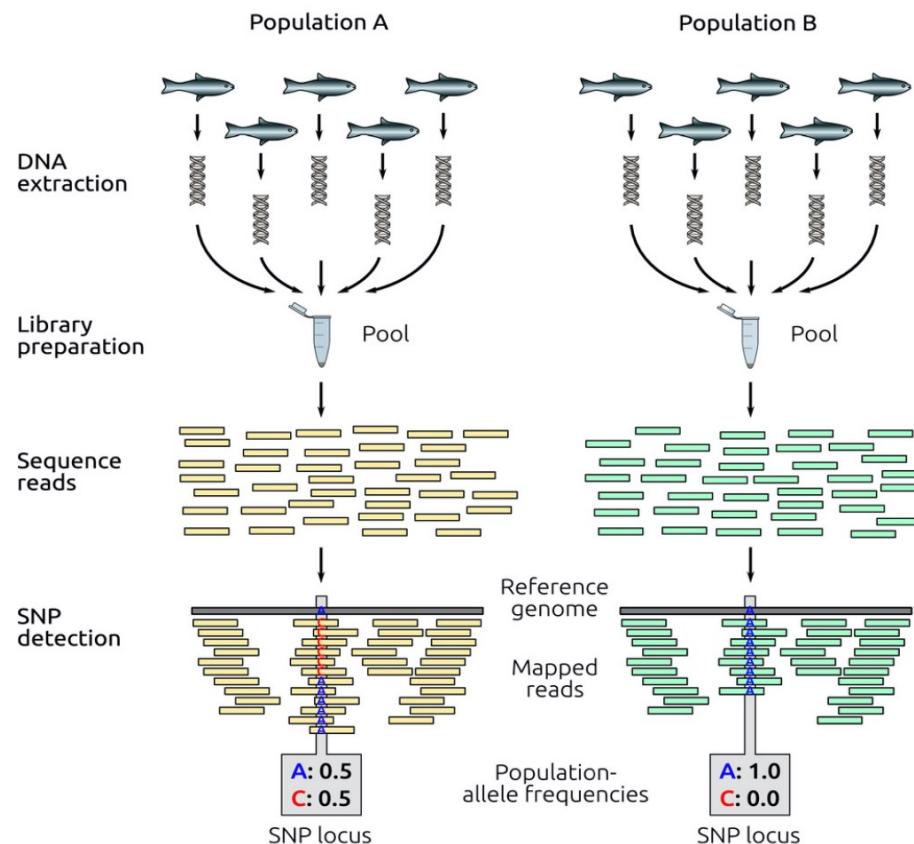
Cons:

- Expensive per sample
- Bioinformatics challenges at high sample number

Number of SNPs: 10+ million

Pool Seq

- Whole genome sequencing with pooled DNA of multiple individuals
- Produces allele frequency but not individual genotypes



Adapted from Fuentes-Pardo & Ruzzante 2017 Mol. Ecol

Pool Seq

Pros:

- All sites possible
- Simple library prep
- Cheaper than individual WGS

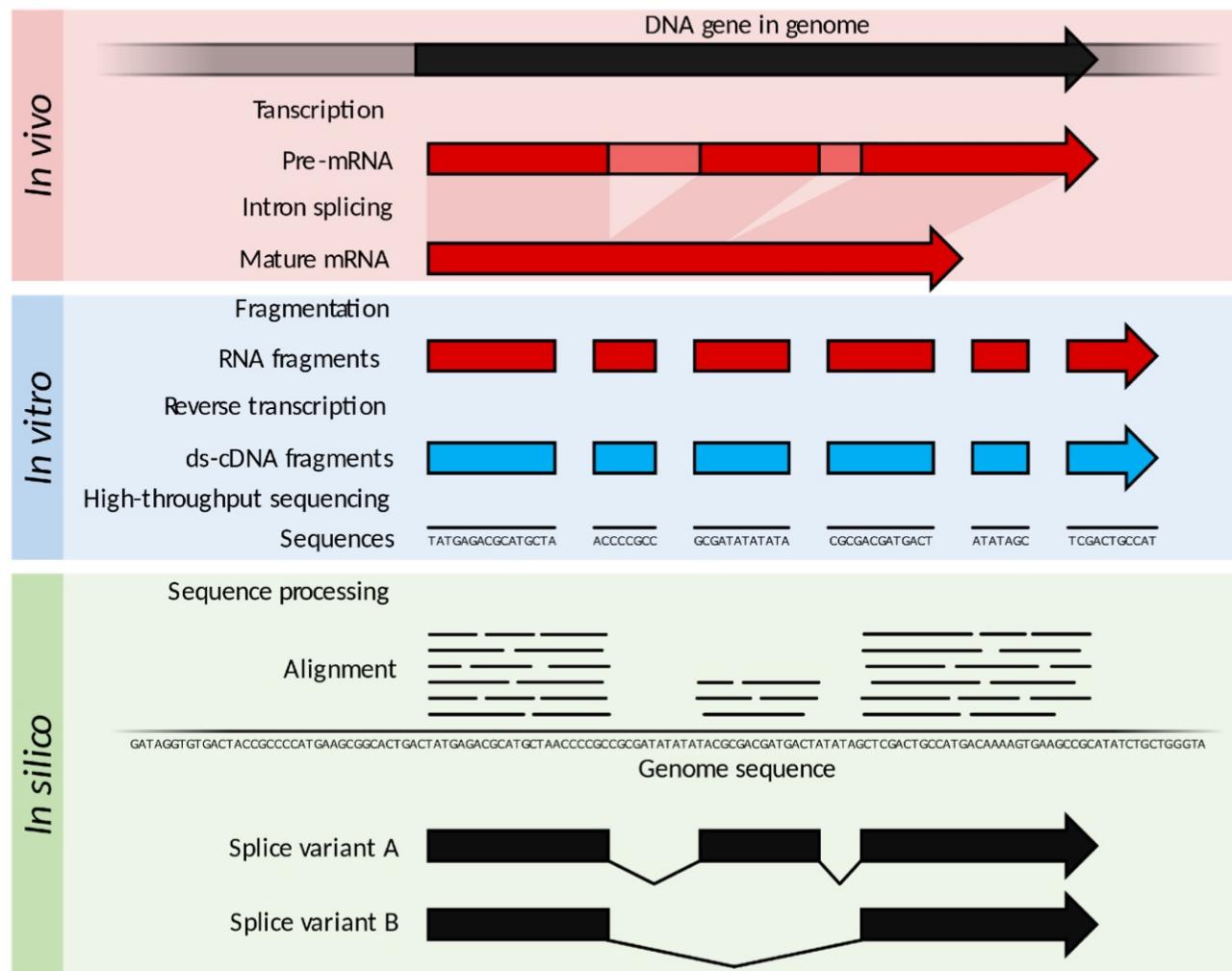
Cons:

- Limited analysis options
- No haplotype information

Number of SNPs: 10+ million

RNAseq

- Convert RNA to cDNA, randomly sheer and sequence.
- Only sequences expressed RNA



RNAseq

Pros:

- Many sites and only in genes.
- Also get expression information
- Relatively easy to assemble

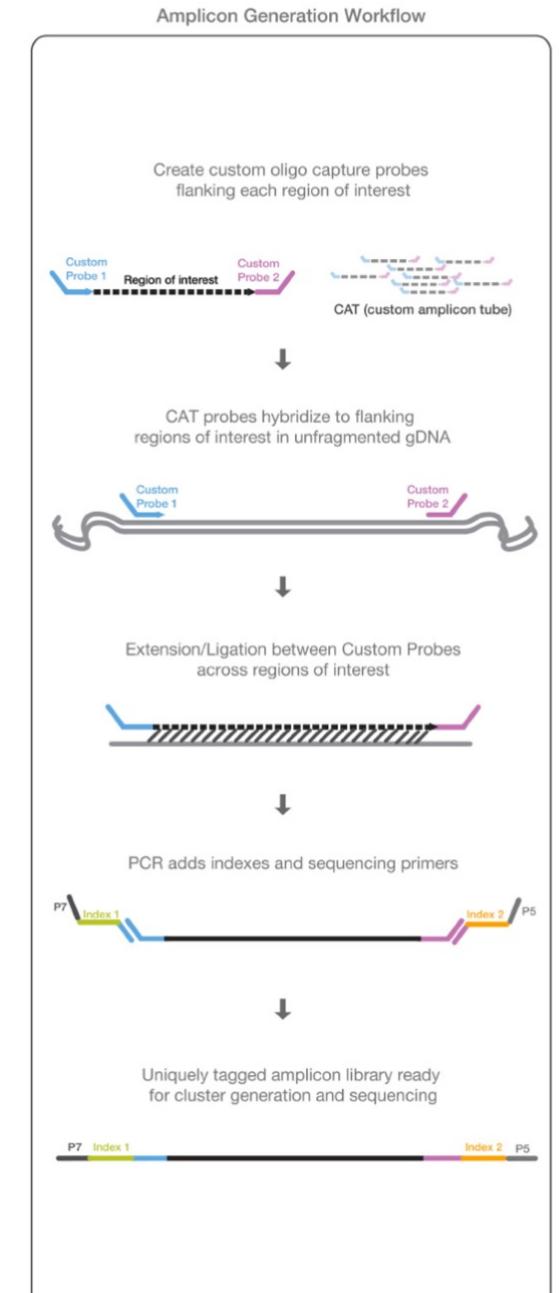
Cons:

- Expression differences complicate SNP calling
- Expensive for pop gen level sampling
- Difficult library prep

Number of SNPs: ~1 million

Amplicon Sequencing

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



Amplicon Sequencing

Pros:

- Get incredible depth at single locus.
- Simple bioinformatics.

Cons:

- Limited to a small number of loci
- Mutations in primer site don't sequence

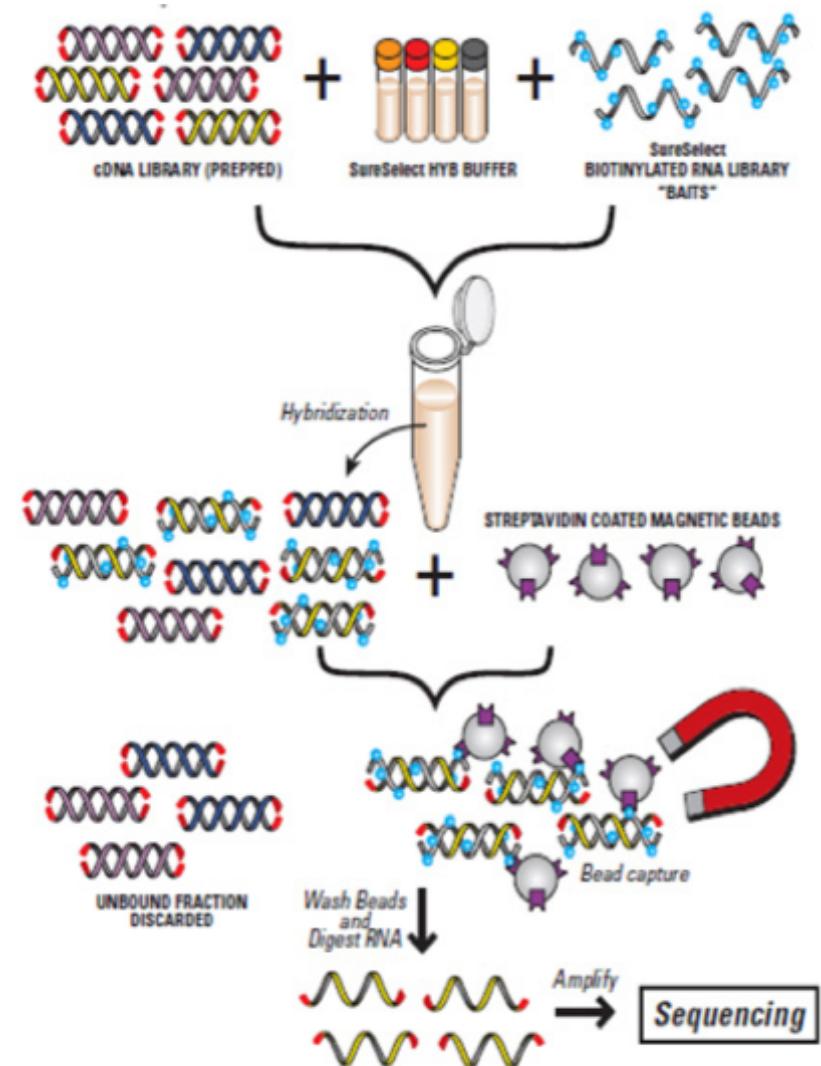
Number of SNPs: <100

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.

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:

- Relatively cheap per sample.
- Good depth at targeted sites

Cons:

- Requires designing probes.
- Long library prep.

Number of SNPs: 100k - 1 million

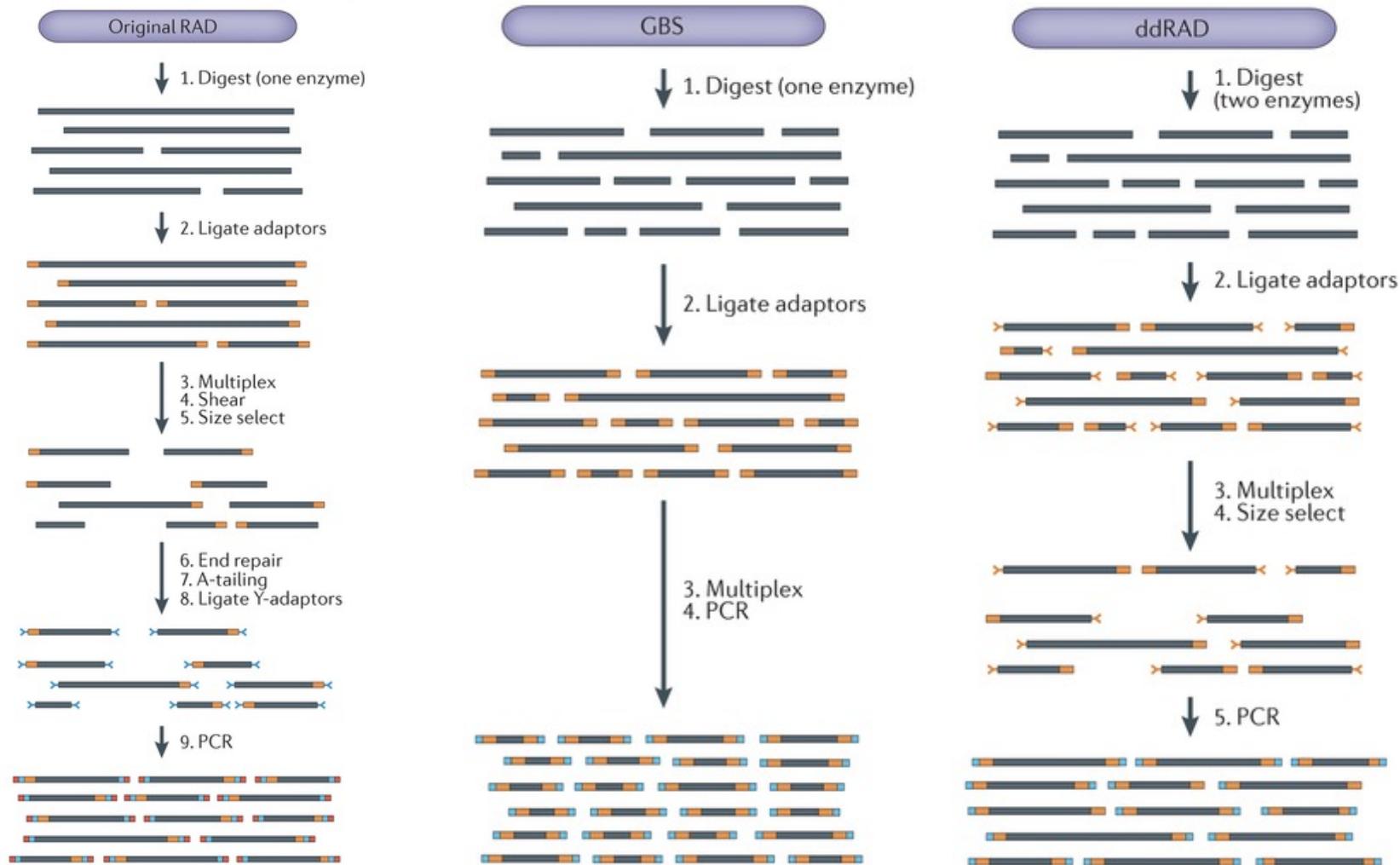
EecSeq

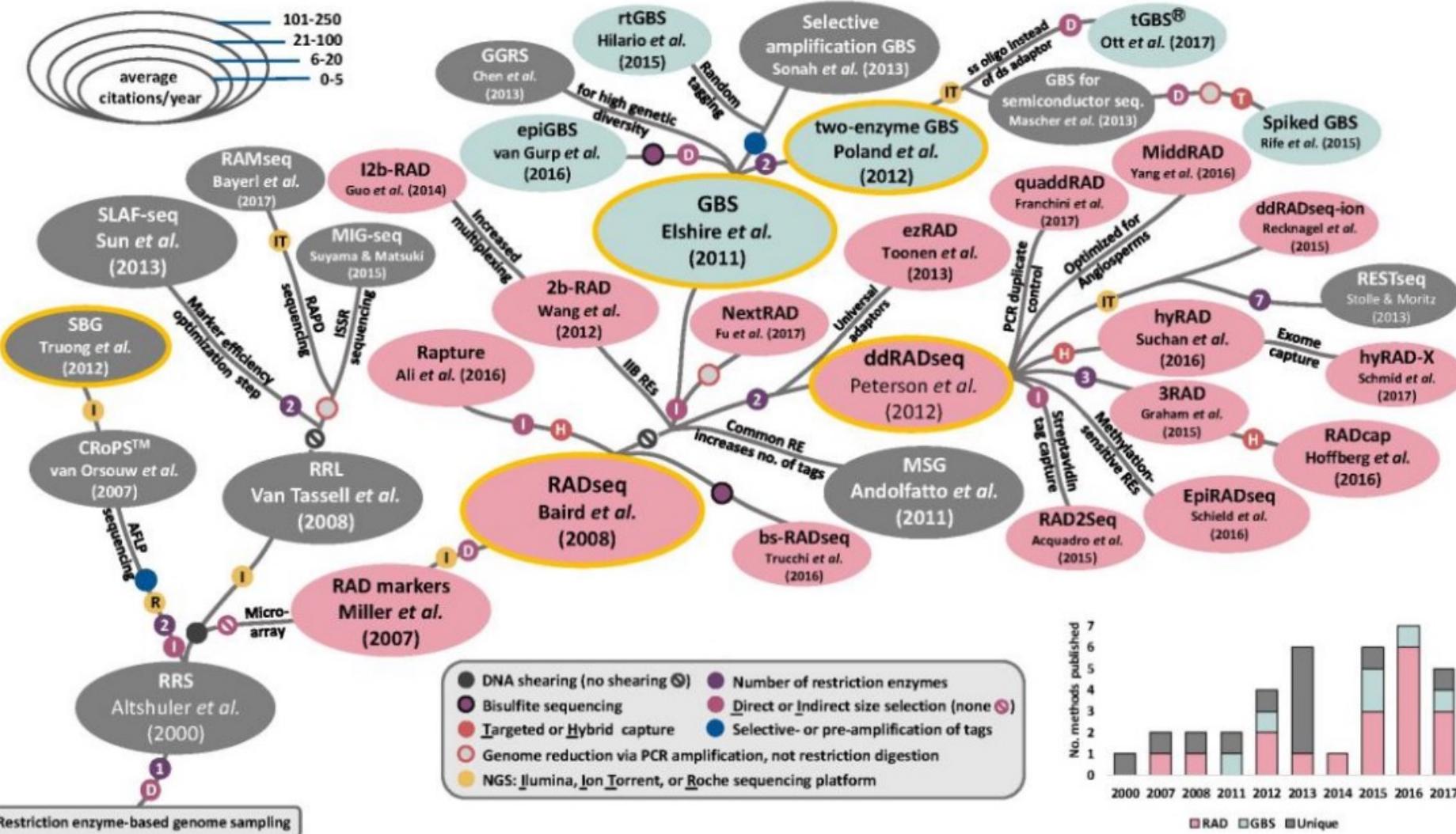
- Exome capture using cDNA targets.
- Don't need to know genome sequence to sequence exome, or design targets.
- ~\$50 per sample.

Genotyping-by-Sequencing types

- Digest DNA with restriction enzyme. Attach barcode and sequencing tags. Sequence many samples in one library.
- Many different flavours:
 - GBS, RAD, ddRAD

Genotyping-By-Sequencing





Genotyping-By-Sequencing

Pros:

- Quick library prep for hundreds of samples.
- Cheap per sample cost (<\$10/sample)

Cons:

- Relatively sparse SNPs compared to other methods
- Can have problems overlapping different library preps

Number of SNPs: 5k - 50k

RADcapture

- Digest DNA with restriction enzyme. Attach barcode and sequencing tags. **Sequence capture before sequencing.** Sequence many samples in one library.
- Different flavours
 - Rapture, RADcap

RADcapture

Pros:

- Quick library prep for hundreds of samples.
- Cheap per sample cost (<\$10/sample)
- More overlap of reads = more SNPs
- Can be good for poor quality samples (e.g. herbarium)

Cons:

- Relatively sparse SNPs compared to other methods
- Requires extra step to make capture probes
- Less well established

Recommendations

- GT-seq
 - **Large** scale genetic monitoring (e.g. fisheries)
 - Where you need many samples, but comparatively fewer markers.

Recommendations

- RAD/RADcapture
 - Short projects
 - Population structure
 - Phylogenetic
 - Genetic maps / QTL maps
 - Species ID
 - Genome scans?

Recommendations

- Whole genome sequencing
 - Fine scale genome analysis
 - Association mapping
 - Scans of selection
 - Small genome organisms

Recommendations

- Sequence capture
 - Large genomes
 - Bigger or longer projects
 - Fine scale genome analysis