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

Biol 525D - Bioinformatics for Evolutionary Biology
2019

Instructors

Dr. Gregory Owens

gregory.owens@alumni.ubc.ca

Dr. JS Légaré

jslegare@cs.ubc.ca

Dr. Kathryn Hodgins

kathryn.hodgins@monash.edu

WEBSITE: https://owensgl.github.io/biol525D/

Course Objective

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

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?

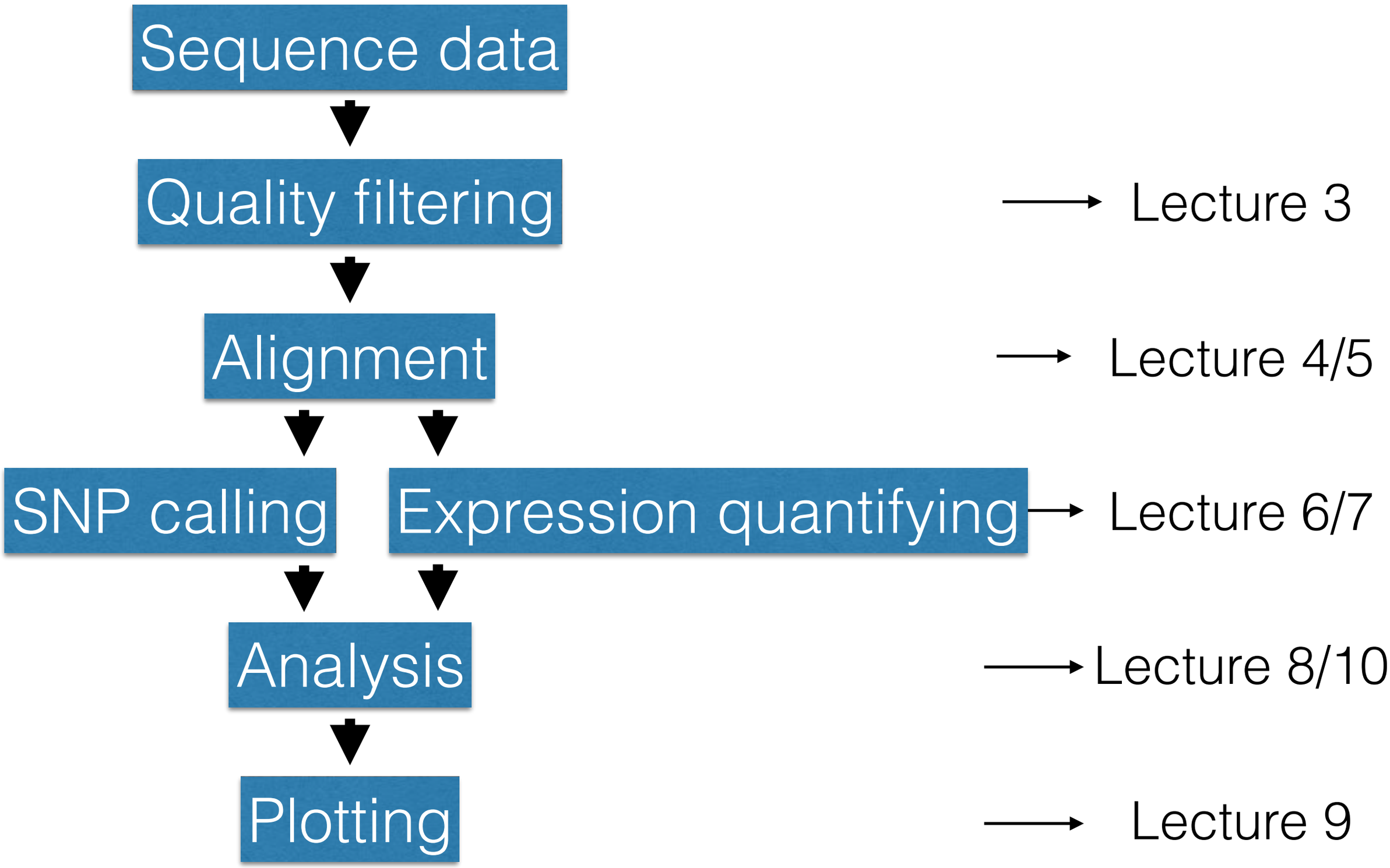
Goals

Raw sequence data

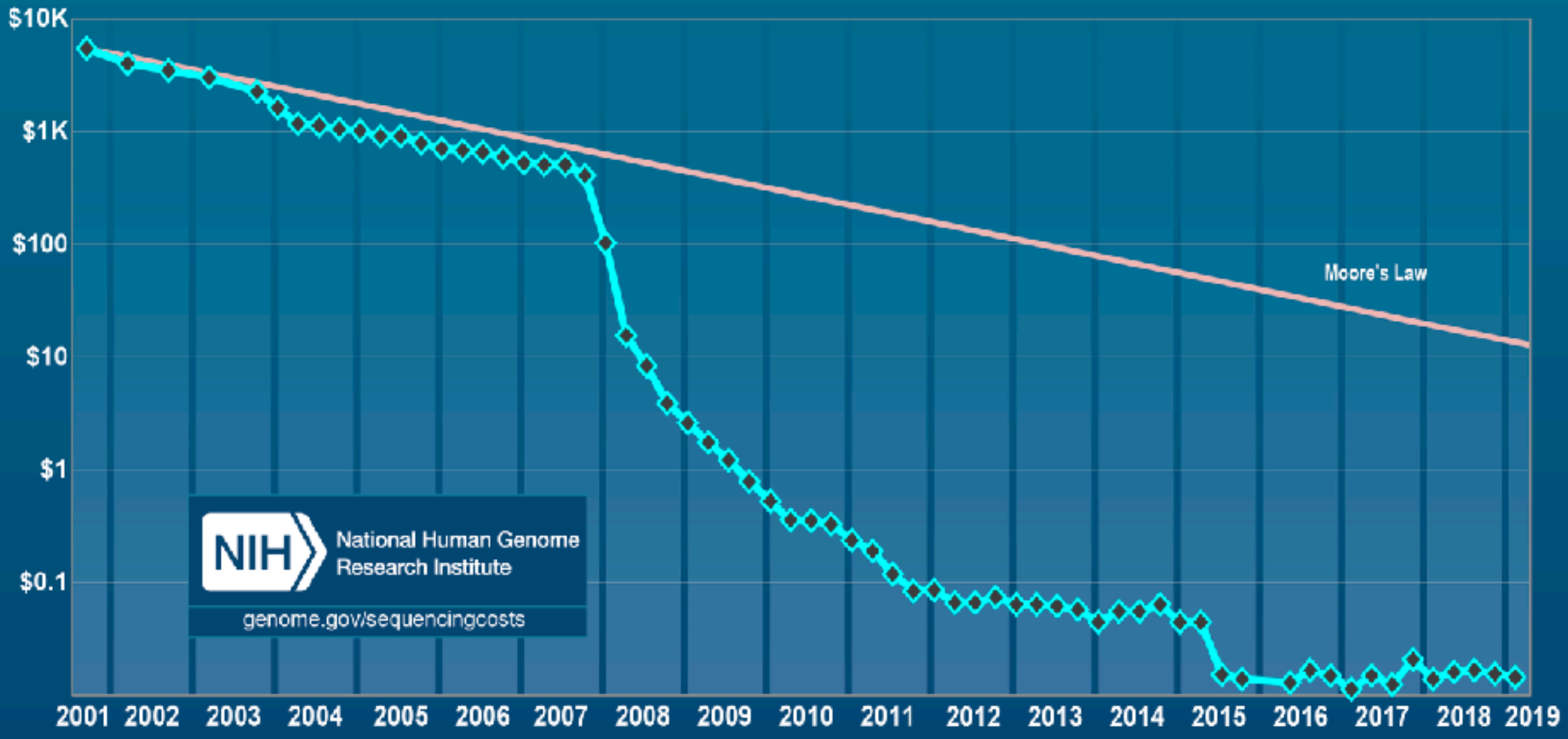
????

Results and Figures

Goals

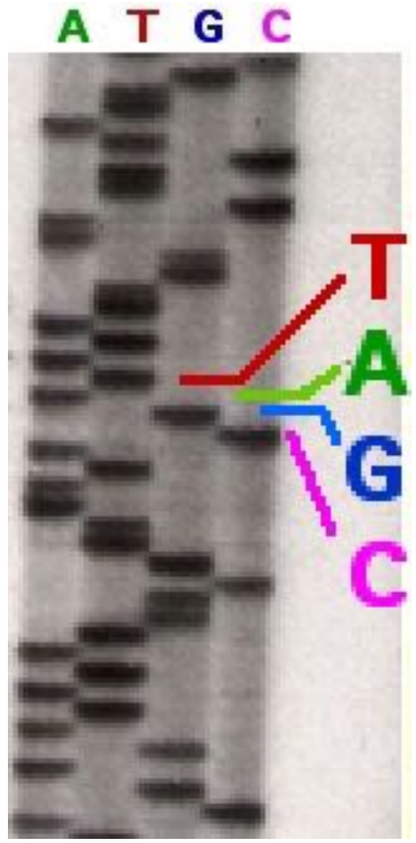

Sequence data
Quality filtering
Lecture 3
Alignment
Lecture 4/5
SNP calling
Expression quantifying
Lecture 6/7
Analysis
Lecture 8/10
Plotting
Lecture 9

Cost per Raw Megabase of DNA Sequence



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

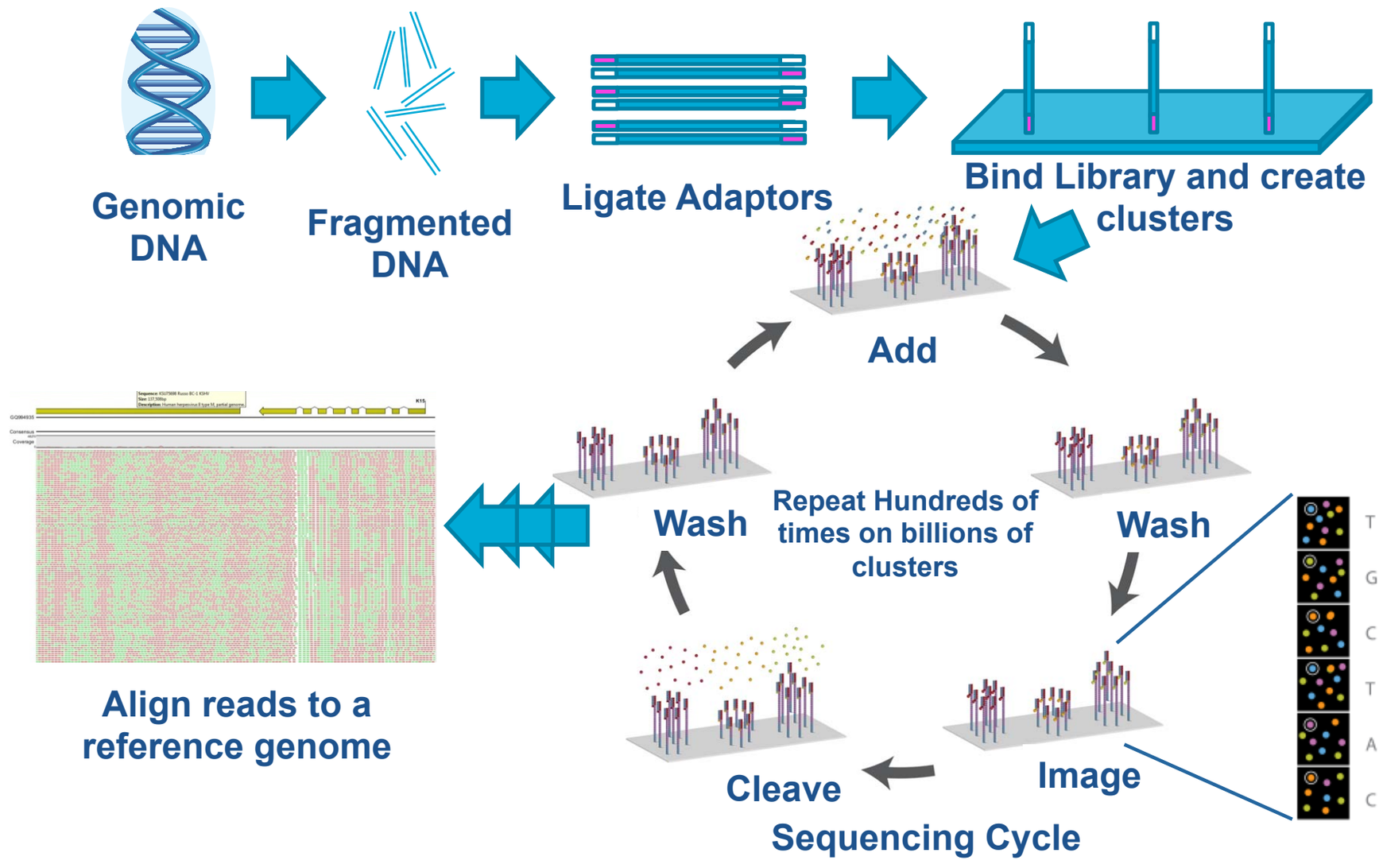
Second/third generation sequencing

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

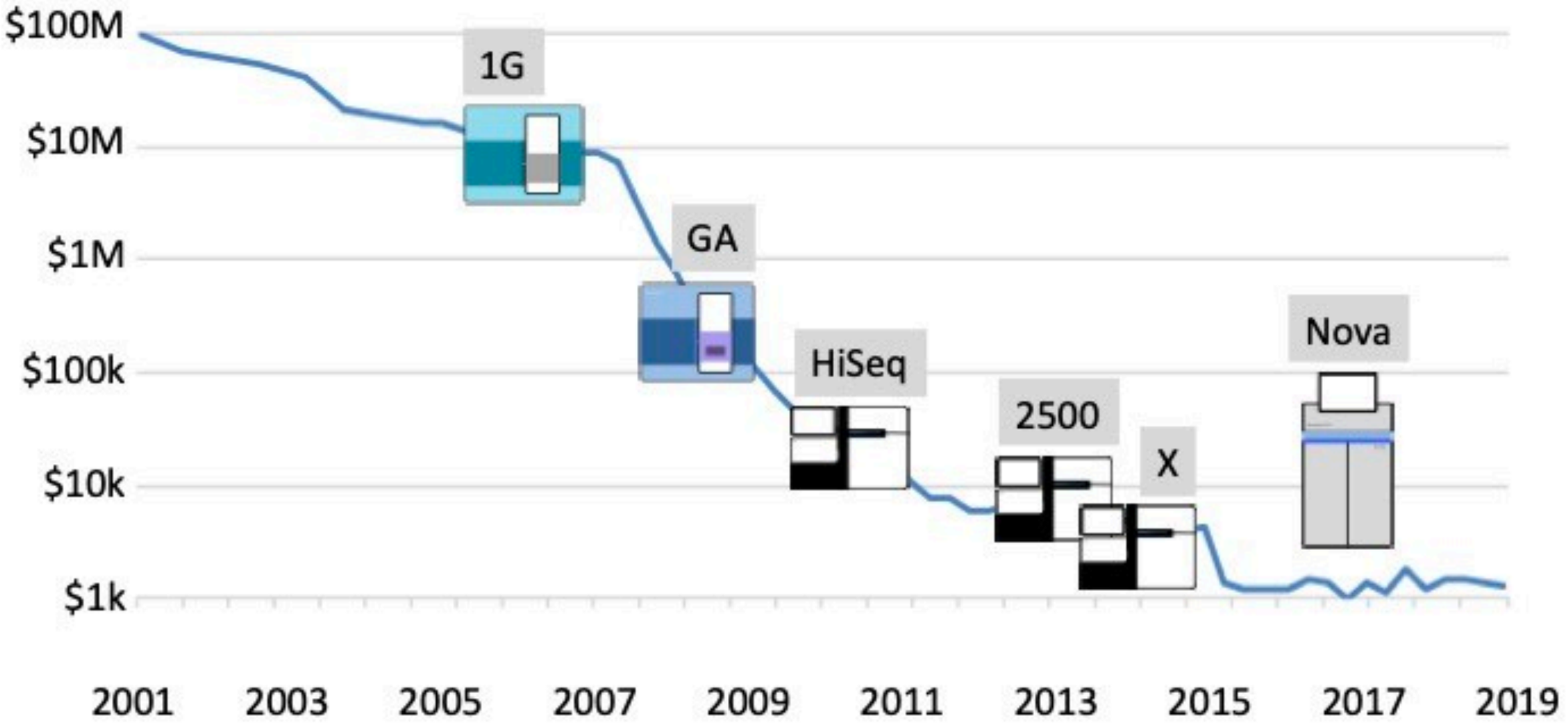
Second generation sequencing

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



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

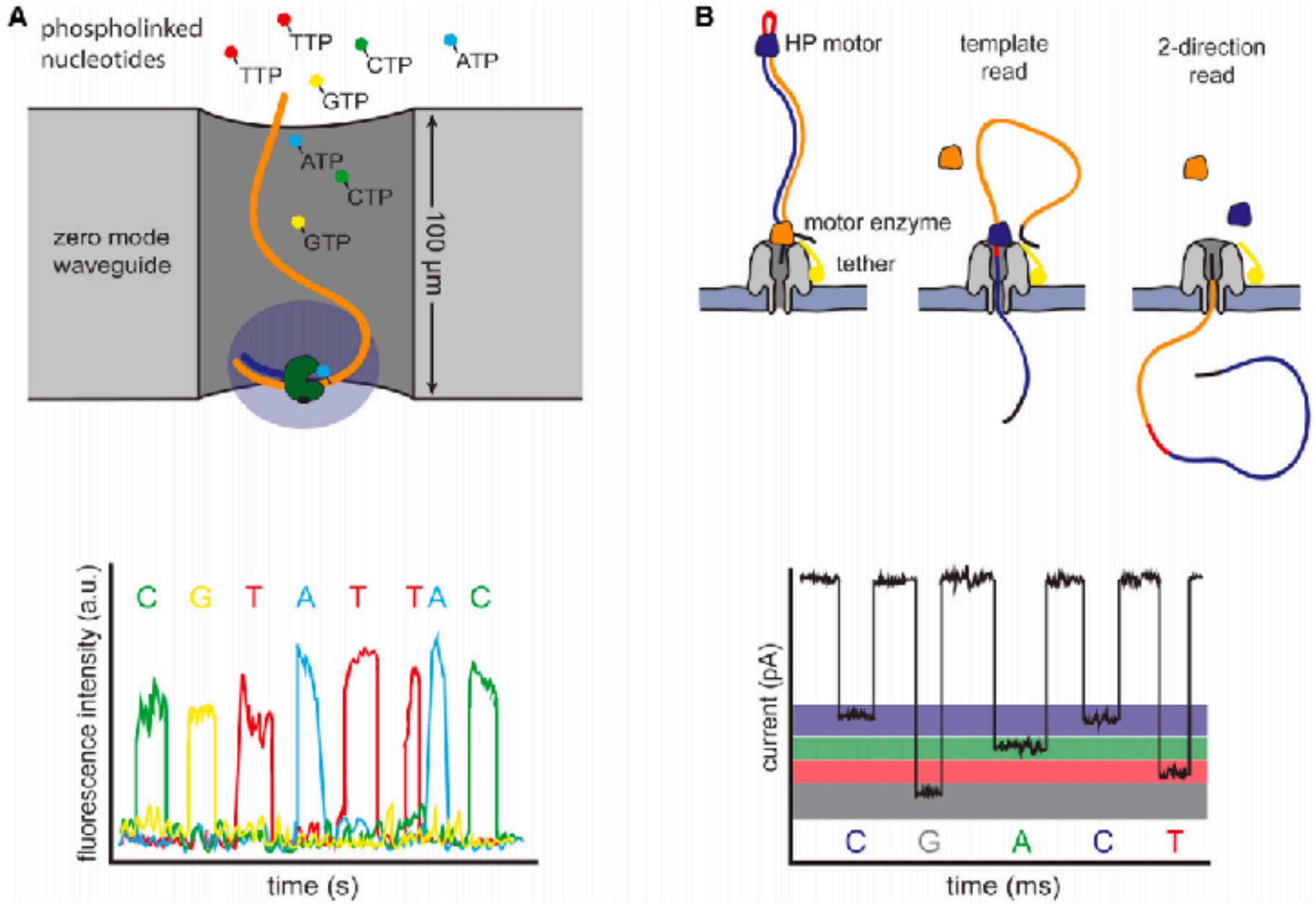


Figure 3. Single Molecule Sequencing Platforms

(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 adaptor is bound with a motor enzyme as well as a tether, 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 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

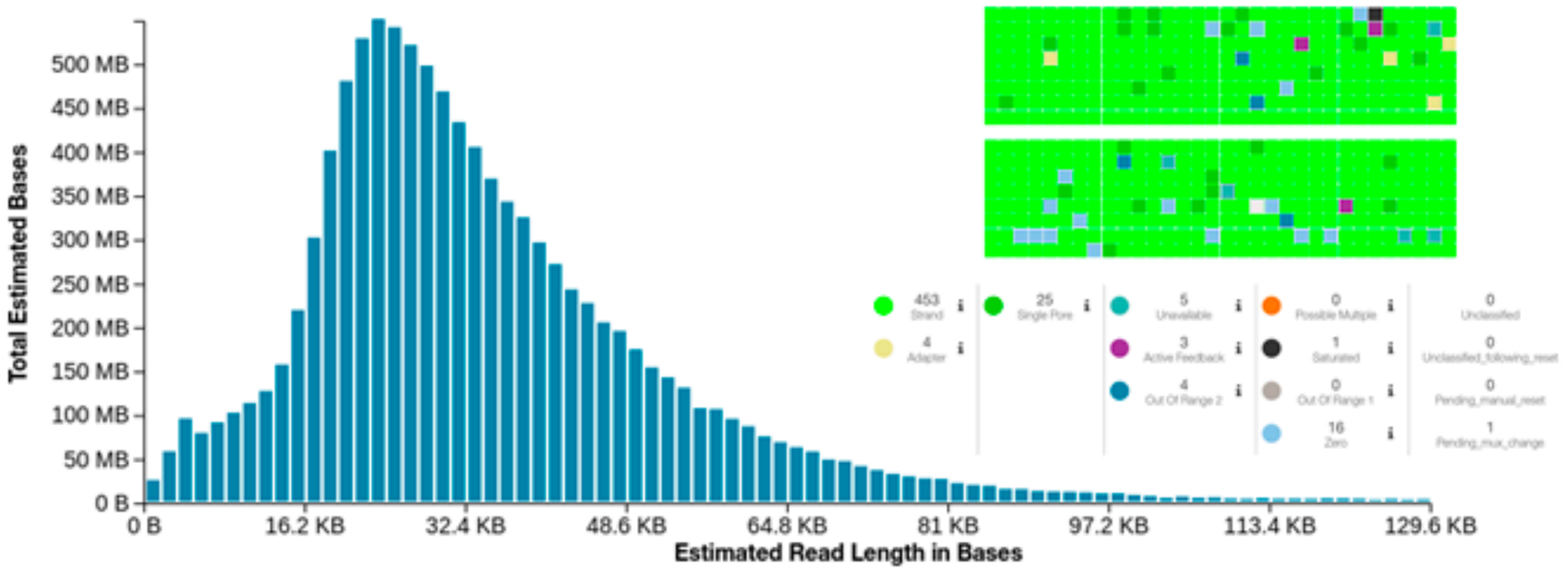
100-180Gb/flowcell

~$2000/flowcell

2-13% error rate

Oxford Nanopore

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



Pacific Biosciences

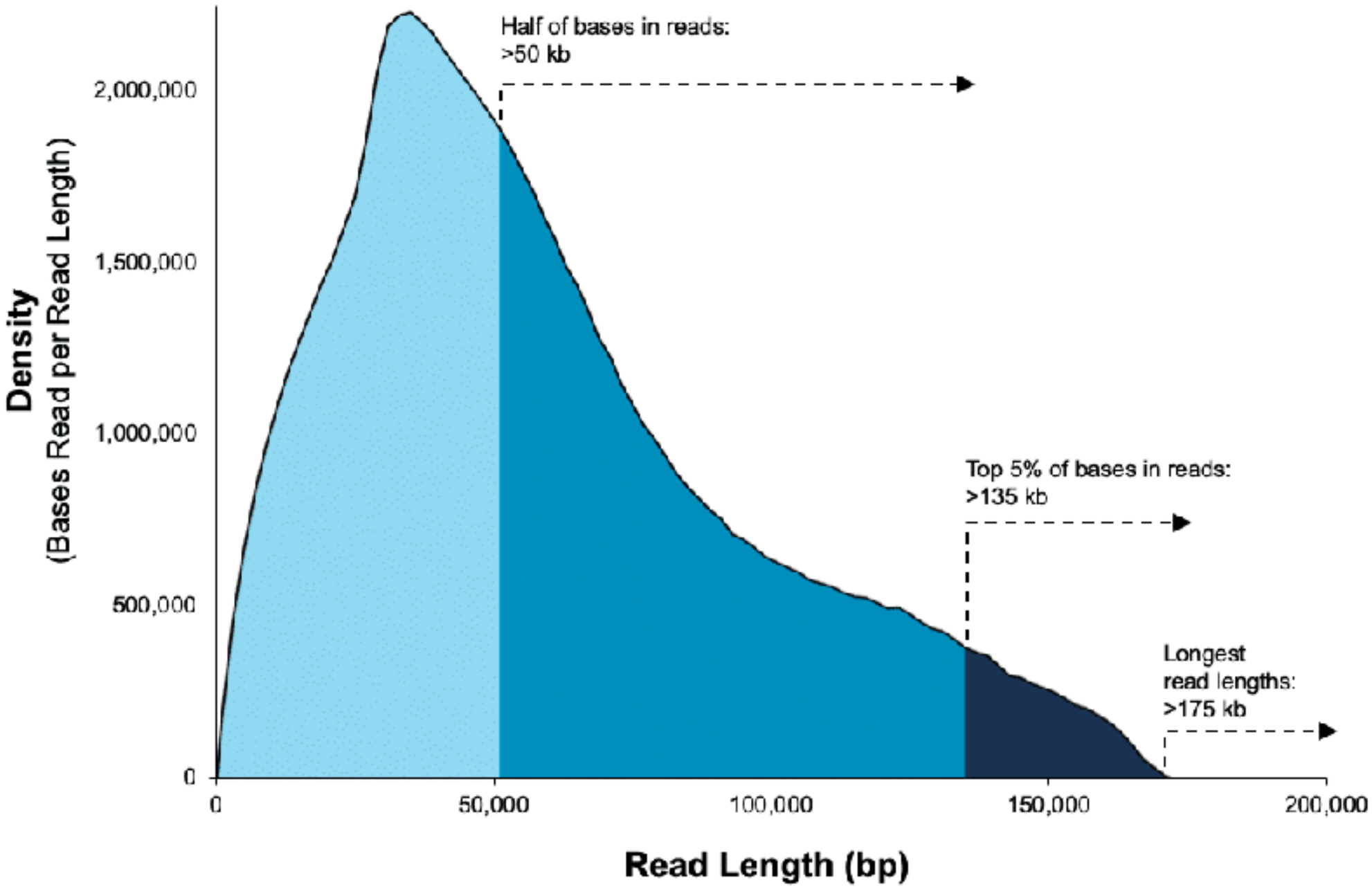
Sequel II

1-10Gb/flowcell

~$500/flowcell

13% error rate

Pacific Biosciences



Challenges of long read technology

- Too expensive to be used for population level sequencing.

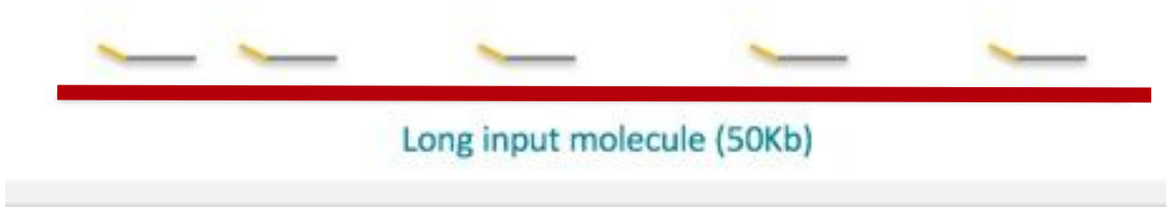
- High error rate.

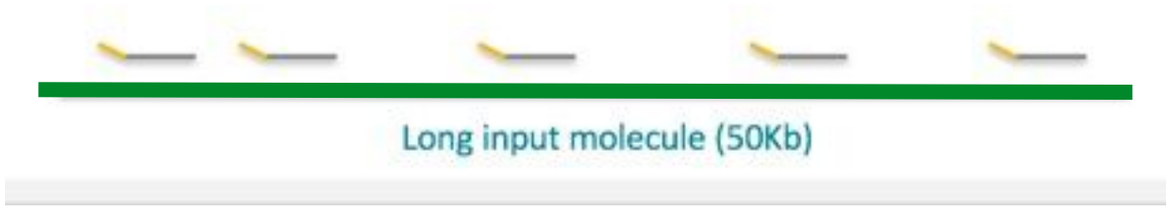
Uses of long reads

- 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

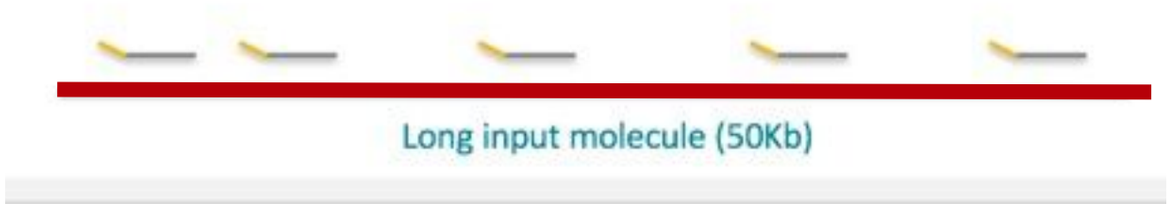
- Barcodes read originating from individual DNA molecules
- Uses Illumina sequencing
- Tells you which reads are physically nearby and on the same strand

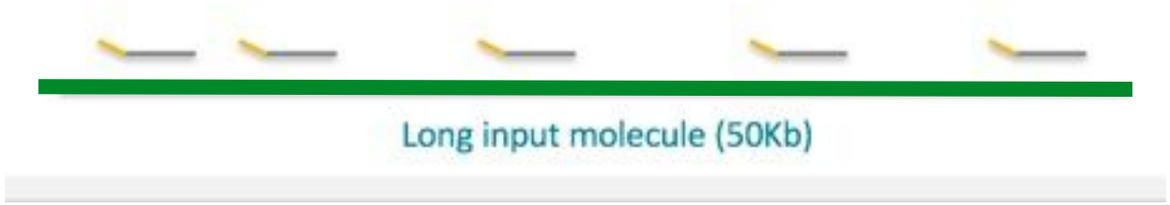




Synthetic long reads

- 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
-Bioinformatic challenges at high sample number

Number of SNPs: 10+ million

Pool Seq

- Whole genome sequencing with pooled DNA of multiple individuals

- Produces a measure of allele frequency but not individual genotypes

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

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

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

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

Cons:
-Limited to one or few loci.
-Mutations in primer site don't sequence

Number of SNPs: <100

Sequence Capture

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

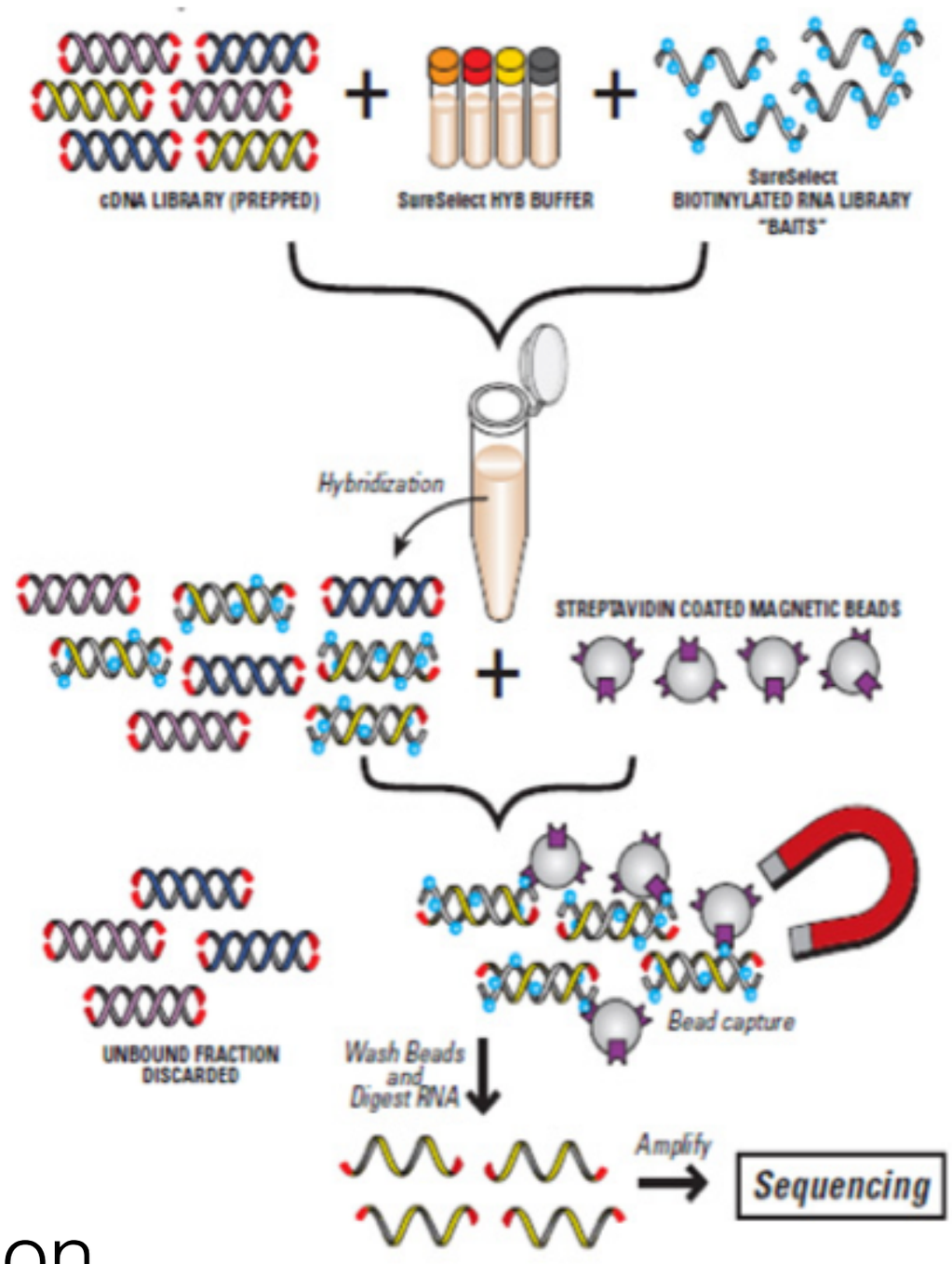
- Make WGS library, hybridize with probe set. Matching sequence will be captured, all others washed away.

- 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.

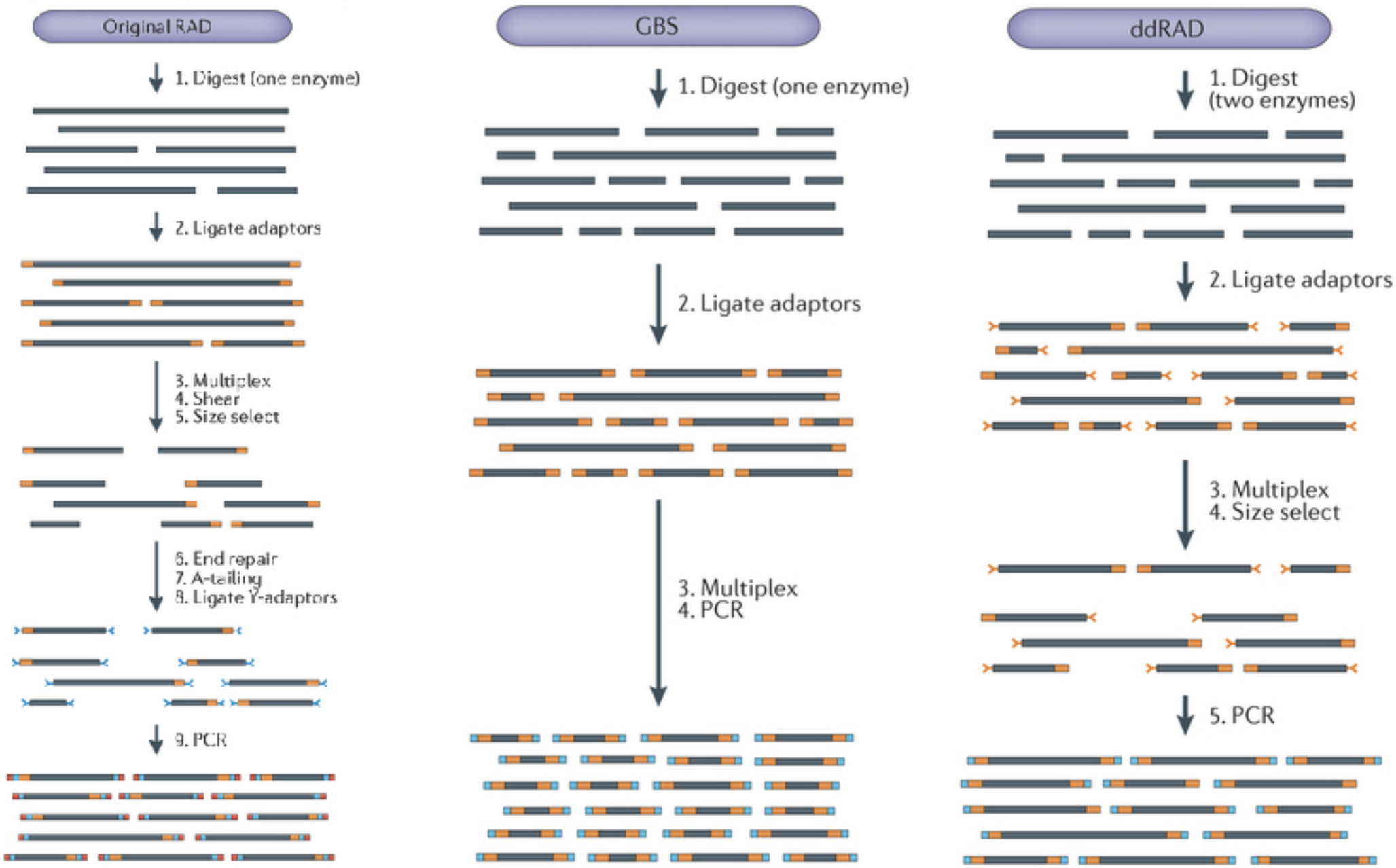
- Don’t need to know genome sequence to sequence exome, or design targets.

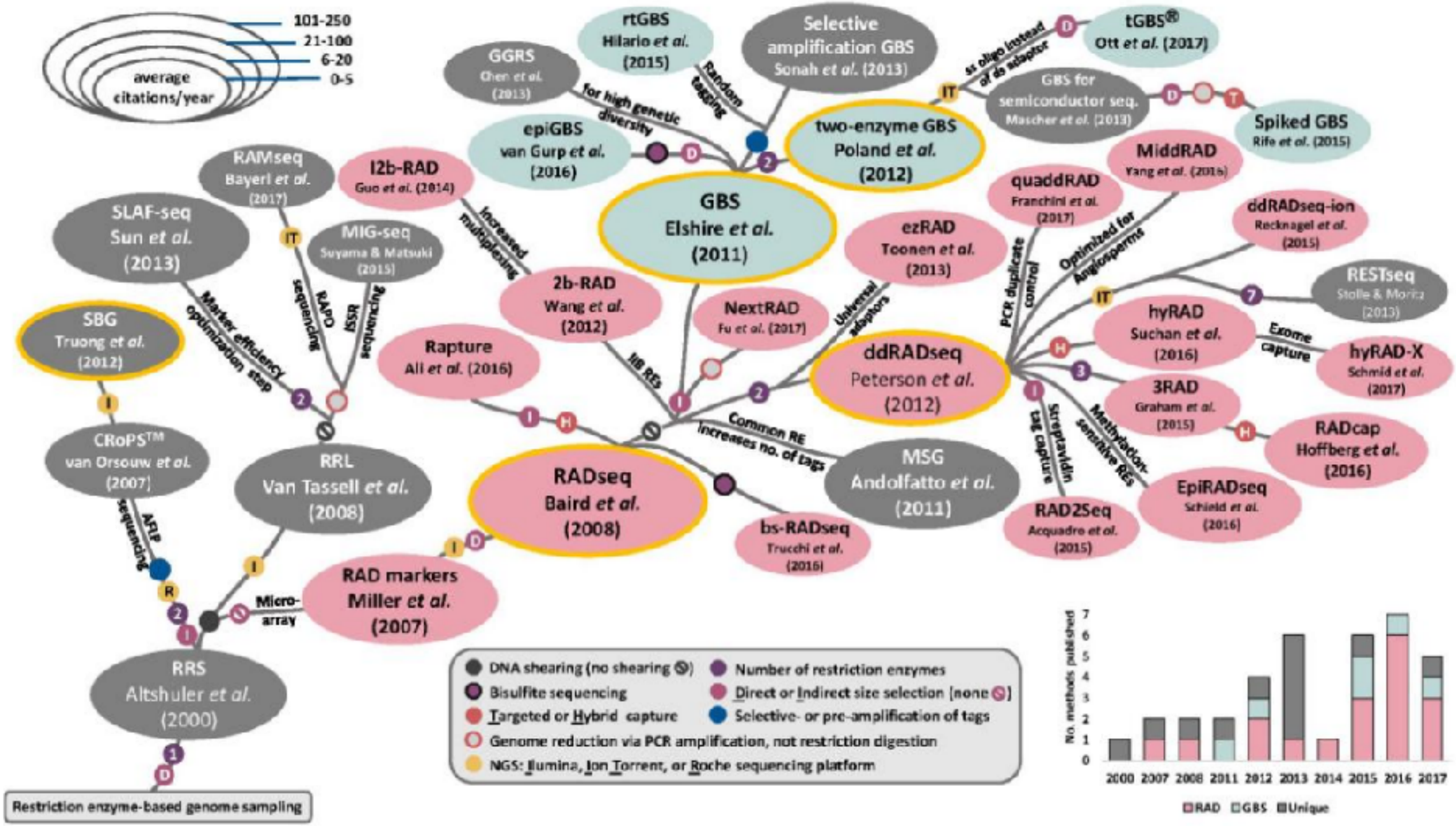
- ~$50 per sample.

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




average citations/year
101-250
21-100
6-20
0-5
Restriction enzyme-based genome sampling
RRS
Altshuler et al. (2000)
CRoPS™
van Orsouw et al. (2007)
SBG
Truong et al. (2012)
SLAF-seq
Sun et al. (2013)
RAMseq
Bayerl et al. (2017)
MIG-seq
Suyama & Matsuki (2015)
RRL
Van Tassell et al. (2008)
RAD markers
Miller et al. (2007)
I2b-RAD
Guo et al. (2014)
2b-RAD
Wang et al. (2012)
Rapture
Ali et al. (2016)
RADseq
Baird et al. (2008)
GBS
Elshire et al. (2011)
epiGBS
van Gurp et al. (2016)
GGRS
Chen et al. (2013)
rtGBS
Hilario et al. (2015)
Selective amplification GBS
Sonah et al. (2013)
two-enzyme GBS
Poland et al. (2012)
NextRAD
Fu et al. (2017)
ezRAD
Toonen et al. (2013)
ddRADseq
Peterson et al. (2012)
MSG
Andolfatto et al. (2011)
bs-RADseq
Trucchi et al. (2016)
tGBS®
Ott et al. (2017)
GBS for semiconductor seq.
Mascher et al. (2013)
Spiked GBS
Rife et al. (2015)
MiddRAD
Yang et al. (2016)
quaddRAD
Franchini et al. (2017)
ddRADseq-ion
Recknagel et al. (2015)
RESTseq
Stolle & Moritz (2013)
hyRAD
Suchan et al. (2016)
hyRAD-X
Schmid et al. (2017)
3RAD
Graham et al. (2015)
RADcap
Hoffberg et al. (2016)
EpiRADseq
Schield et al. (2016)
RAD2Seq
Acquadre et al. (2015)
AFLP sequencing
RAPD sequencing
ISSR sequencing
Micro-array
Marker efficiency optimization step
Increased multiplexing
IIB RES
Common RE increases no. of tags
Universal adaptors
Random tagging
for high genetic diversity
ss oligo instead of ds adaptor
PCR duplicate control
Optimized for Angiosperms
Exome capture
Streptavidin tag capture
Methylation-sensitive REs
DNA shearing (no shearing)
Bisulfite sequencing
Targeted or Hybrid capture
Genome reduction via PCR amplification, not restriction digestion
NGS: Illumina, Ion Torrent, or Roche sequencing platform
Number of restriction enzymes
Direct or Indirect size selection [none]
Selective- or pre-amplification of tags
No. methods published
0 1 2 3 4 5 6 7
2000 2007 2008 2011 2012 2013 2014 2015 2016 2017
RAD GBS Unique

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

GT-seq

- Genotyping by Thousands
- 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.

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
  - Small genome organisms

Recommendations

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