

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

Bill 525D - Bioinformatics for Evolutionary Biology  
2018

# Instructors

Dr. Gregory Owens



gregory.owens@alumni.ubc.ca

Dr. Kathryn Hodgins



kathryn.hodgins@monash.edu

WEBSITE: <https://github.com/owensgl/biol525D>

# Course Objective

1. Introduction: Scope of course, goals and overview of technology [GREG]
2. Programming for biologists [GREG]
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]

# Goals

Raw sequence data

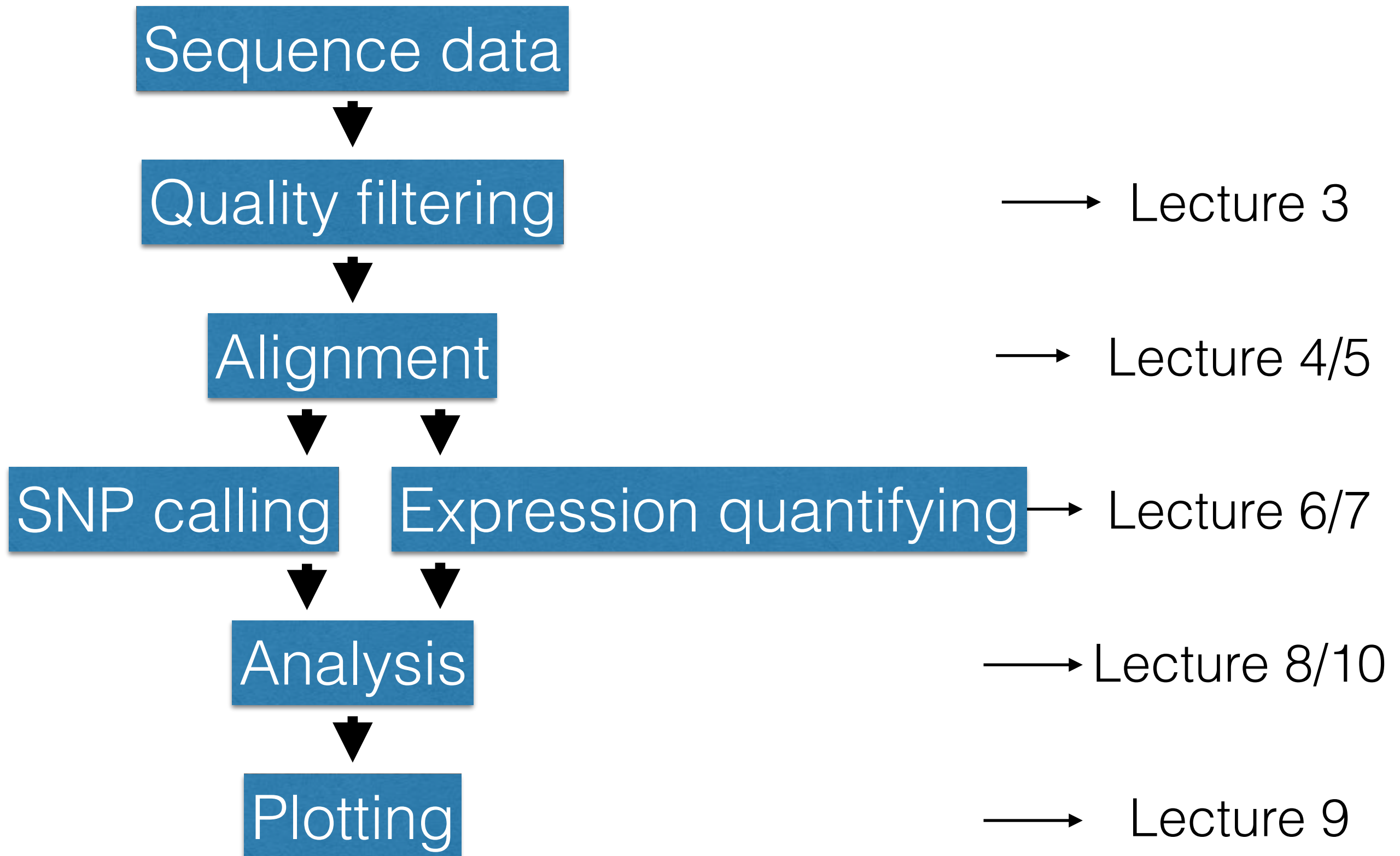


????

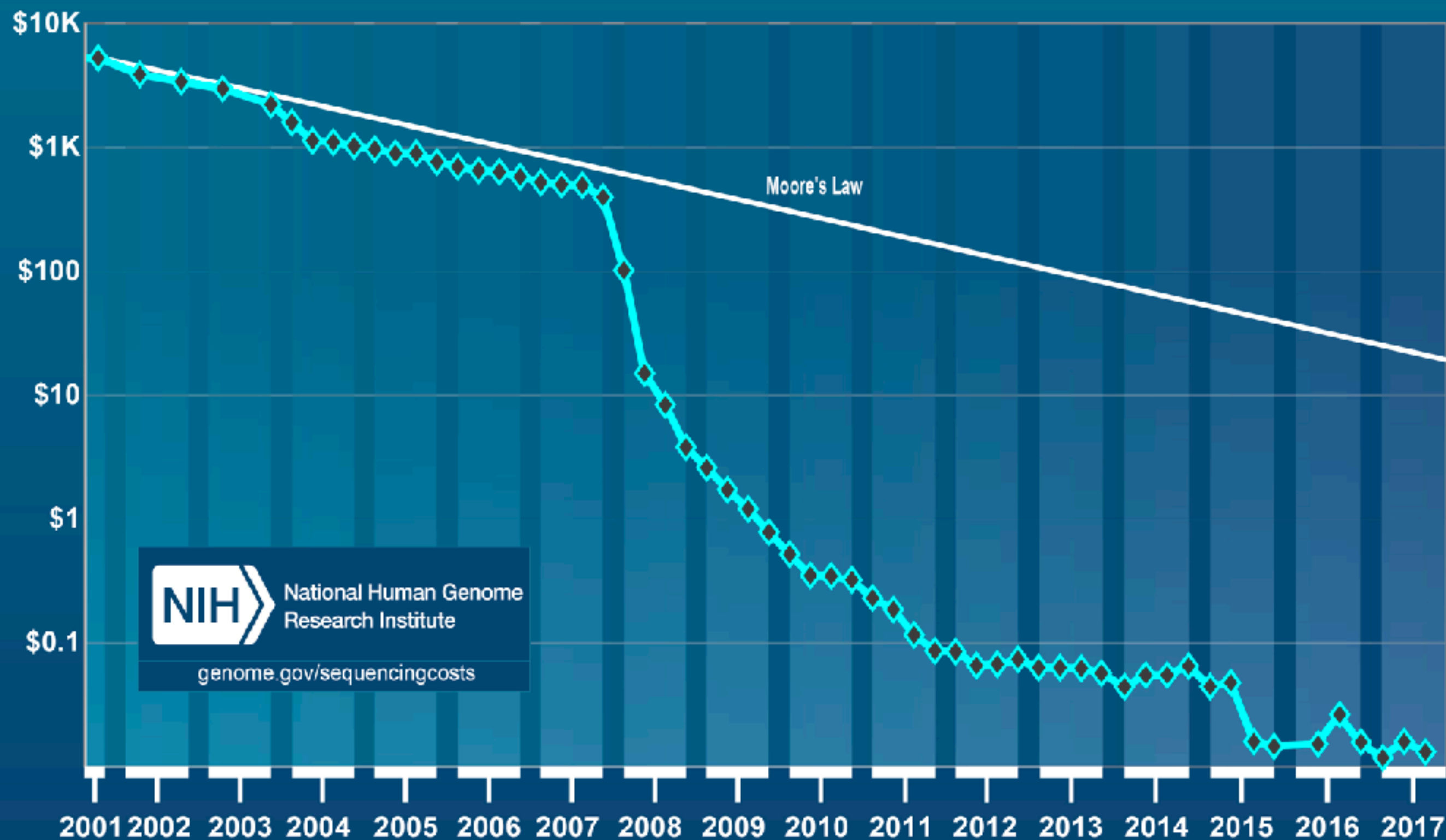


Results and Figures

# Goals

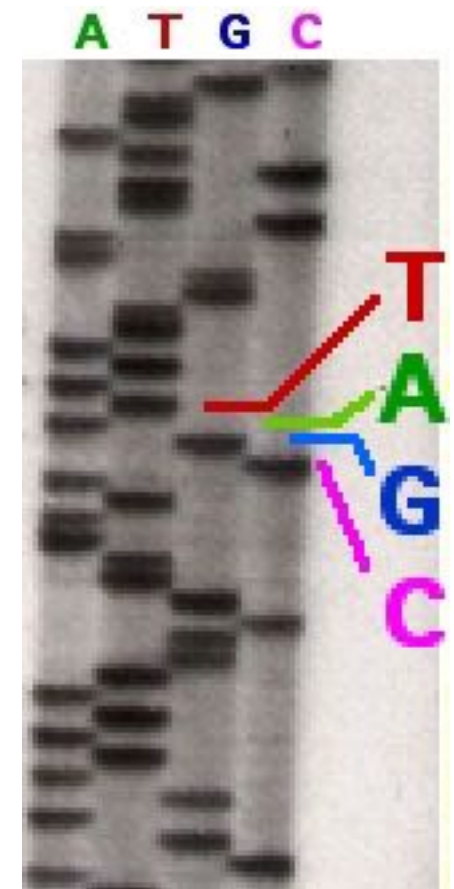


# Cost per Raw Megabase of DNA Sequence



# 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 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

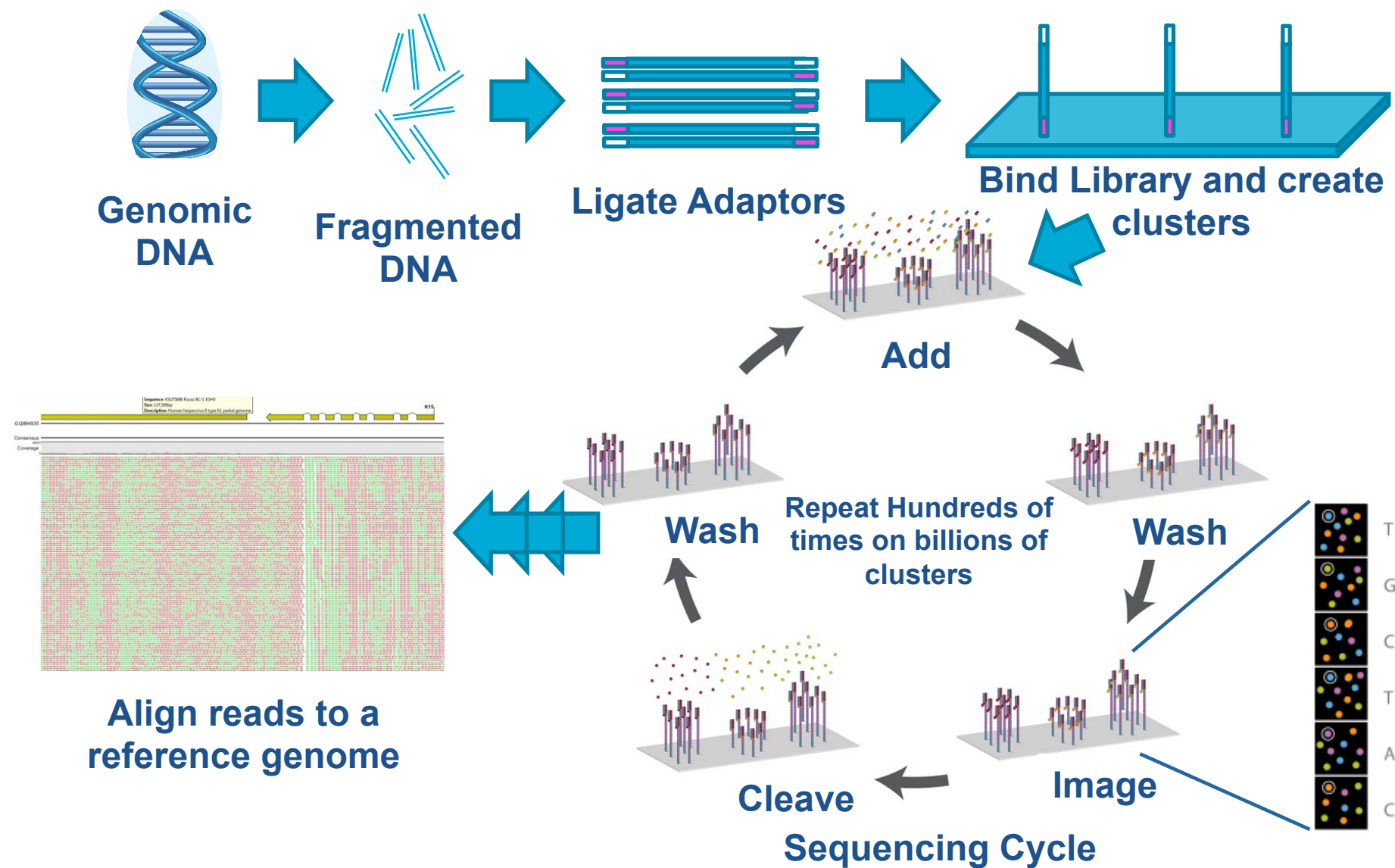
- Sequences many molecules in parallel
- Don't need to know anything about the sequence to start.
- Main technologies:
  - Illumina
  - Ion torrent
  - 454 (Pyrosequencing)
  - PacBio



# Second generation sequencing

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

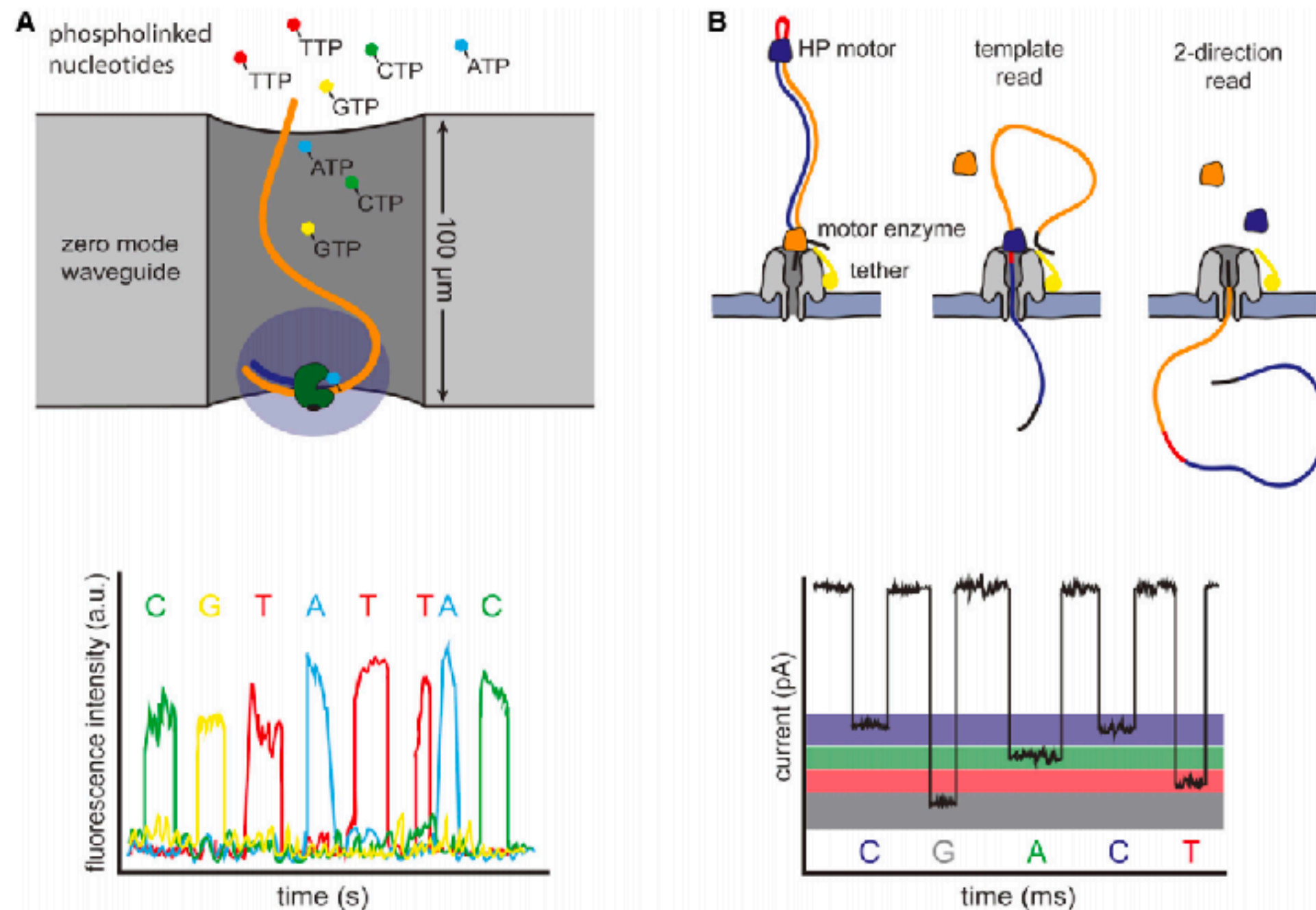
# Illumina sequencing



# 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 adaptors. 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).

# Challenges of long read technology

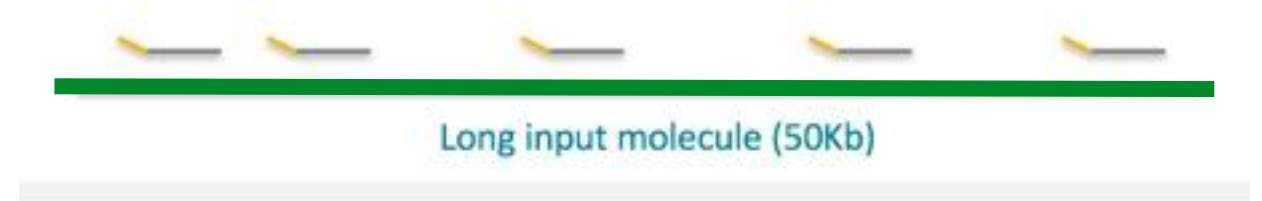
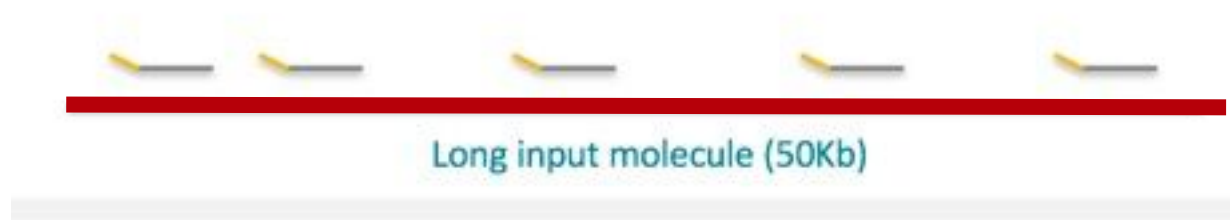
- Too expensive to be used for population level sequencing.
- High error rate.

# 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.

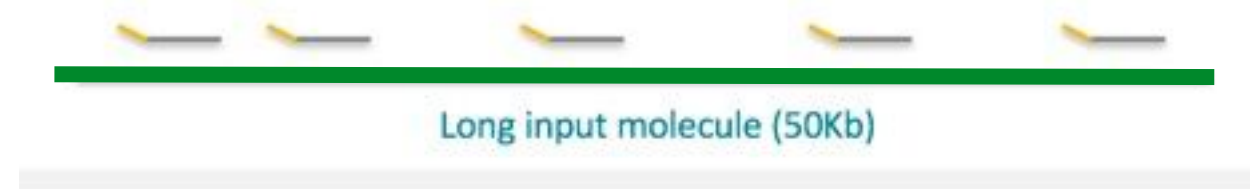
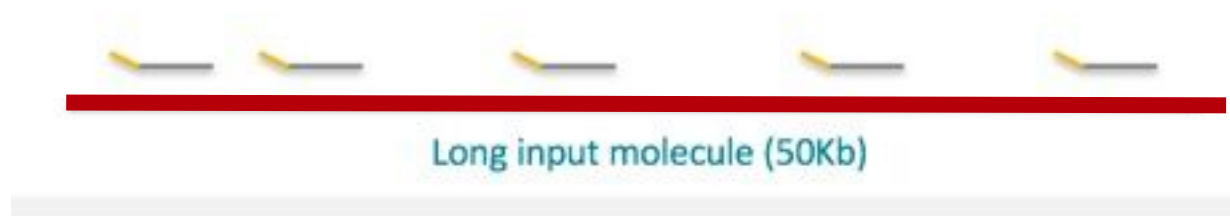
# Synthetic long reads

- 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





# Illumina = more and more data

2019?

	HiSeq 2500	HiSeq 4000	HiSeq X	NovoSeq S1	NovoSeq S2	NovoSeq S4
Reads per lane (million)	250	350	430	800	1650	2500
Throughput per lane (Gb)	62.5	105	129	240	495	750
500 Mb genome coverage	120X	210X	260X	480X	990X	1500X
500 Mb 10X genomes	12	21	26	48	99	150

# 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

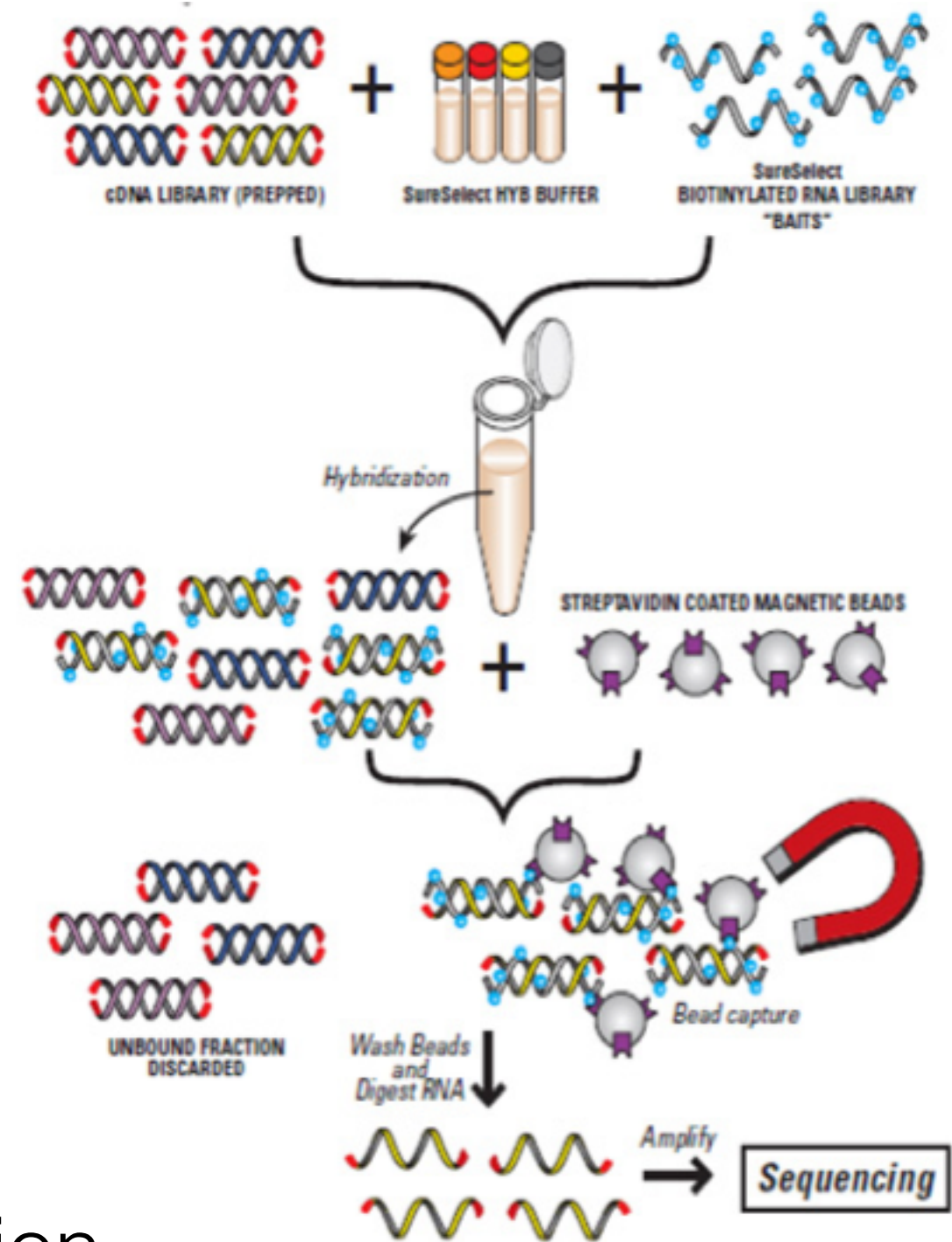
Pros:

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

Cons:

- Requires designing probes.
- Long library prep.

Number of SNPs: 100k - 1 million



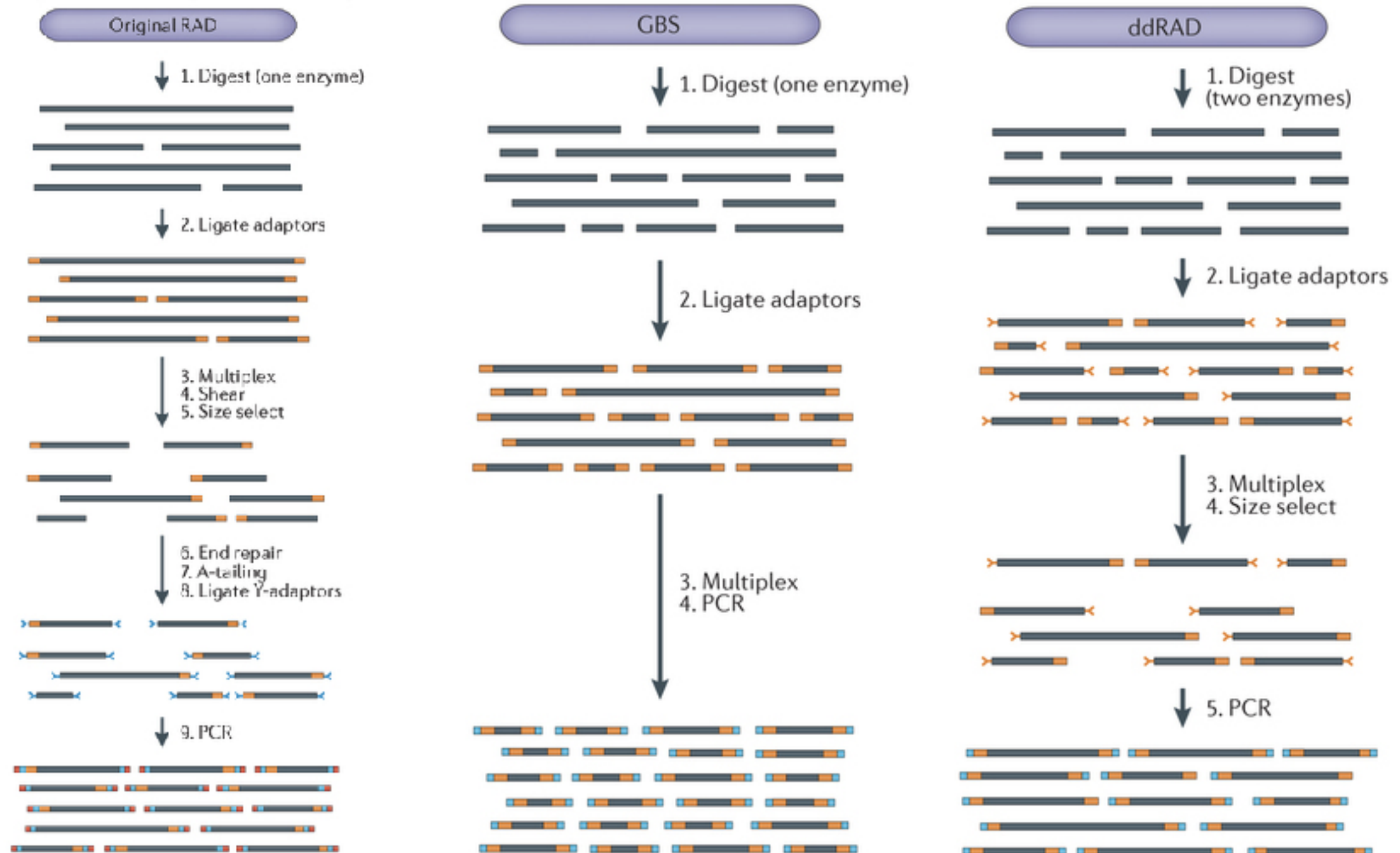
# EecSeq

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

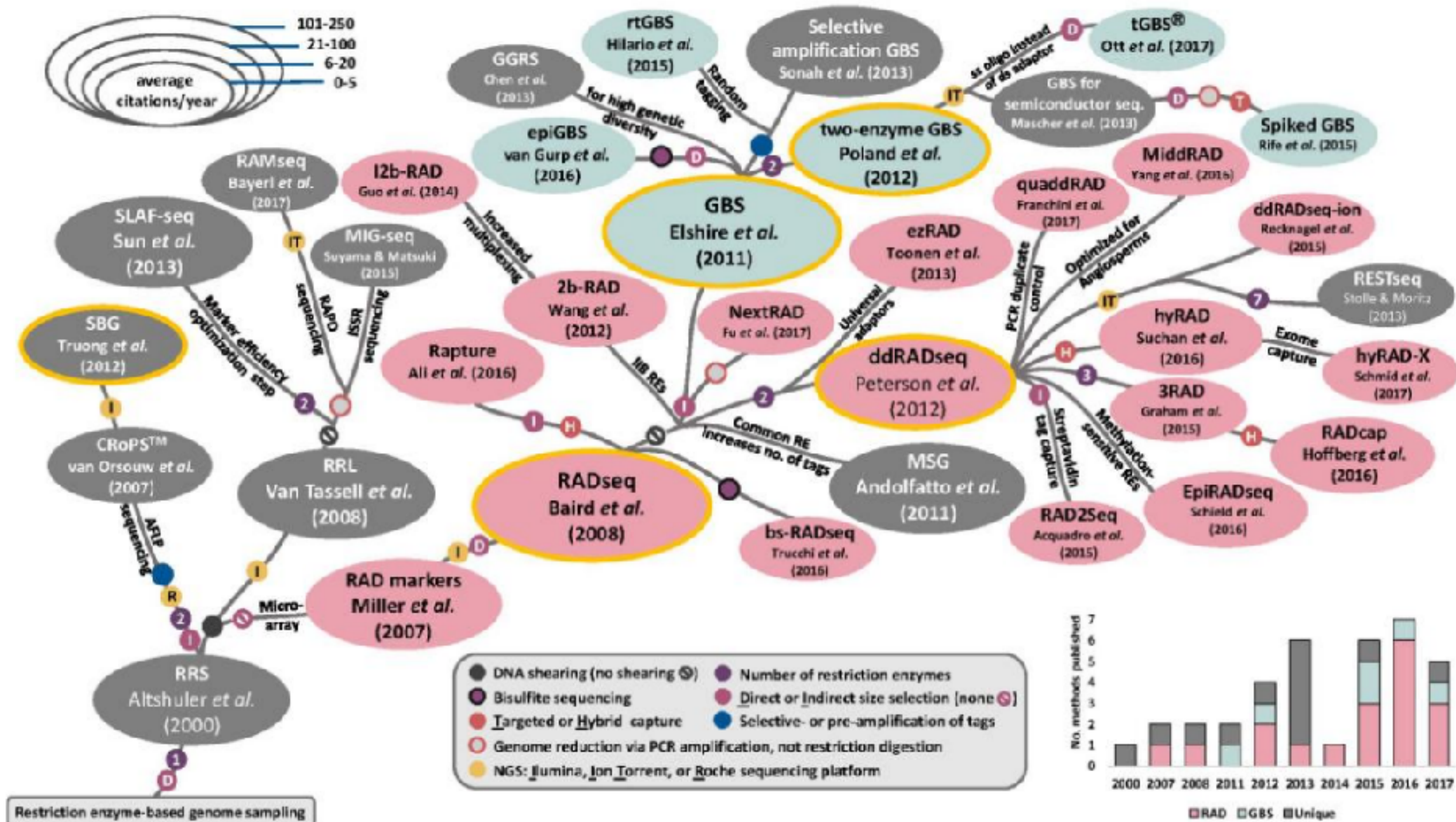
# Genotyping-by-Sequencing types

- 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







# 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

# Computing options

- Mid sized personal server (~30 cores, 200 GB ram)
  - Works for small/medium scale analyses
  - Hard to expand capacity
  - Upfront cost (\$5-10k)
  - Complete control

# Computing options

- Lab supercomputer (~100 cores, 500 GB ram)
  - Works for small to high scale analyses
  - Managing load between users can be troublesome
  - High upfront cost (\$20-50K)
  - Must be placed in Compute Canada facility.

# Computing options

- Compute Canada
  - Potentially thousands of cores
  - SLURM scheduler and 28 day limit on jobs
  - Free for base usage, applications for better access.
  - Virtual machines available, but limited.
  - HINT: Use ``salloc`` to get long interactive jobs for setting up your batch scripts.



# Computing options

- Zoology computing cluster
  - ~100 cores over several servers
  - Don't need to submit jobs, but limited installing privileges.
  - Storage space limitations
  - Often clogged by users
  - ~\$100 per year

# Computing options

- Cloud services (Google, Amazon)
  - Infinitely expandable
  - Can get expensive fast
  - Taking data off cloud servers is expensive