TOPIC 1: Introduction to Next-Gen Sequencing

Bill 525D - Bioinformatics for Evolutionary Biology

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]

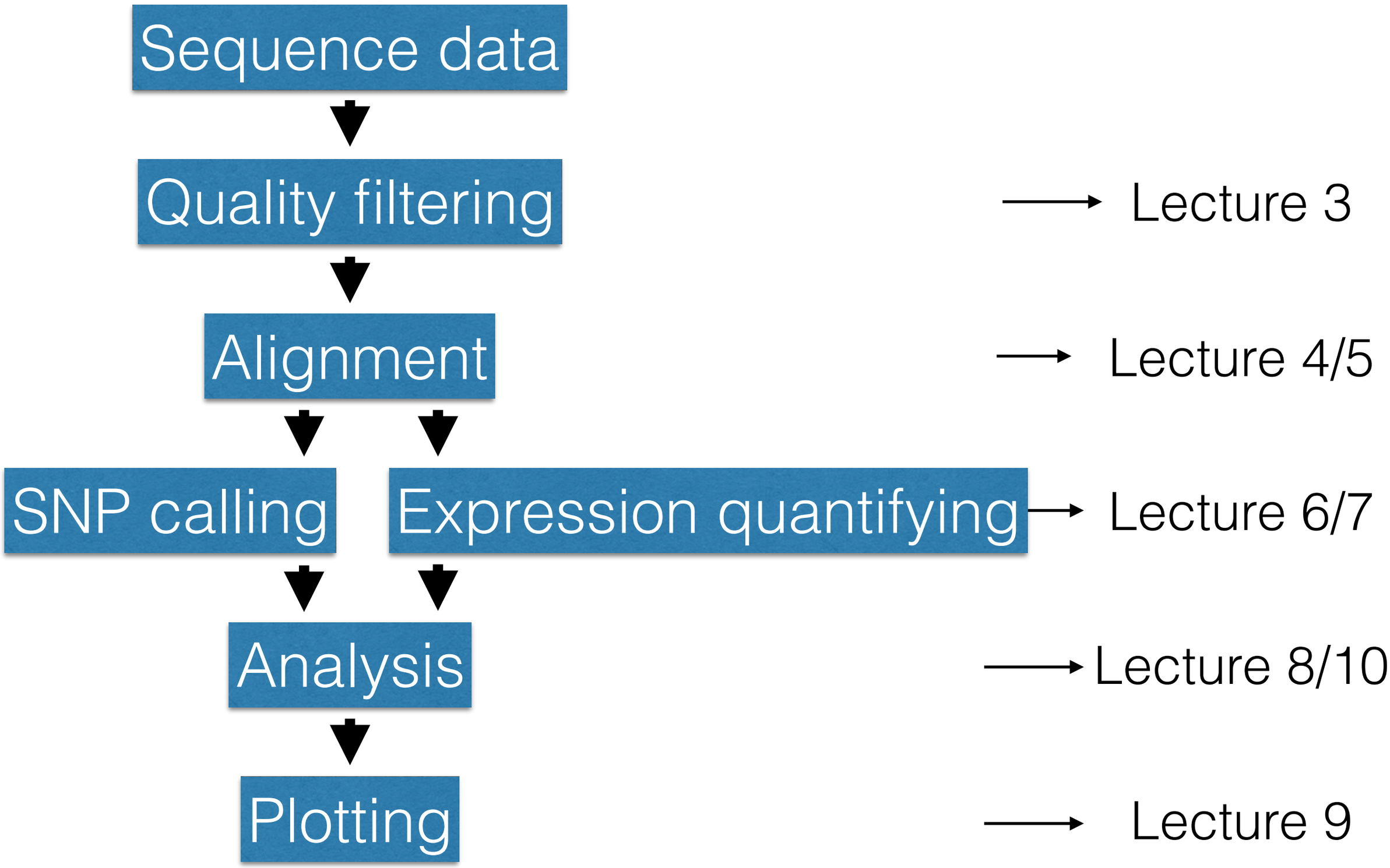
Goals

Raw sequence data

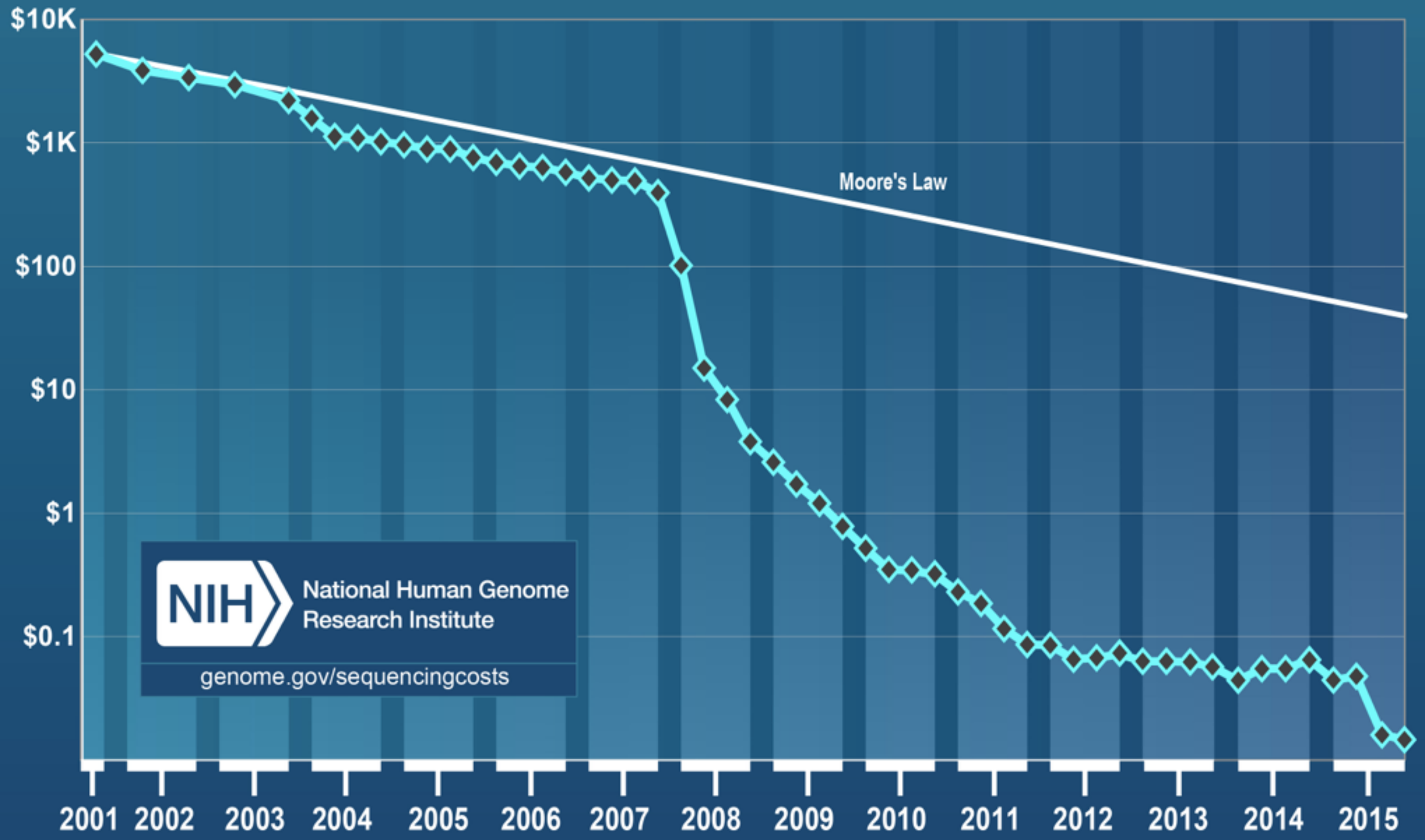
????

Results and Figures

Goals

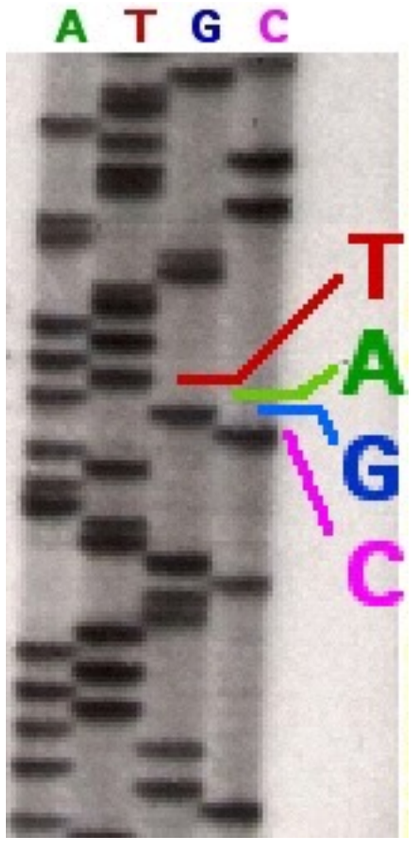

Sequence data
Quality filtering
Alignment
SNP calling
Expression quantifying
Analysis
Plotting
Lecture 3
Lecture 4/5
Lecture 6/7
Lecture 8/10
Lecture 9

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

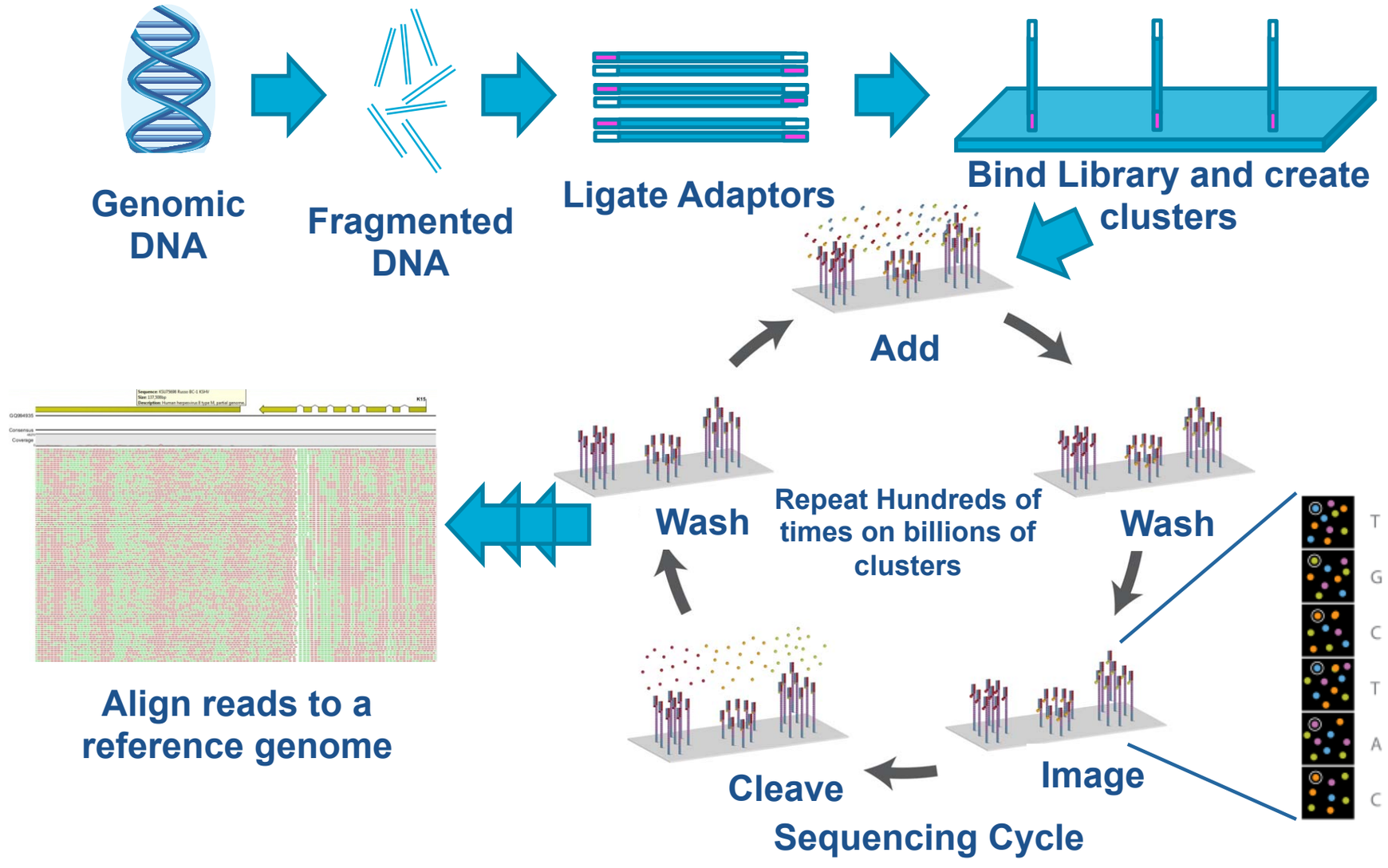
Second generation sequencing

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

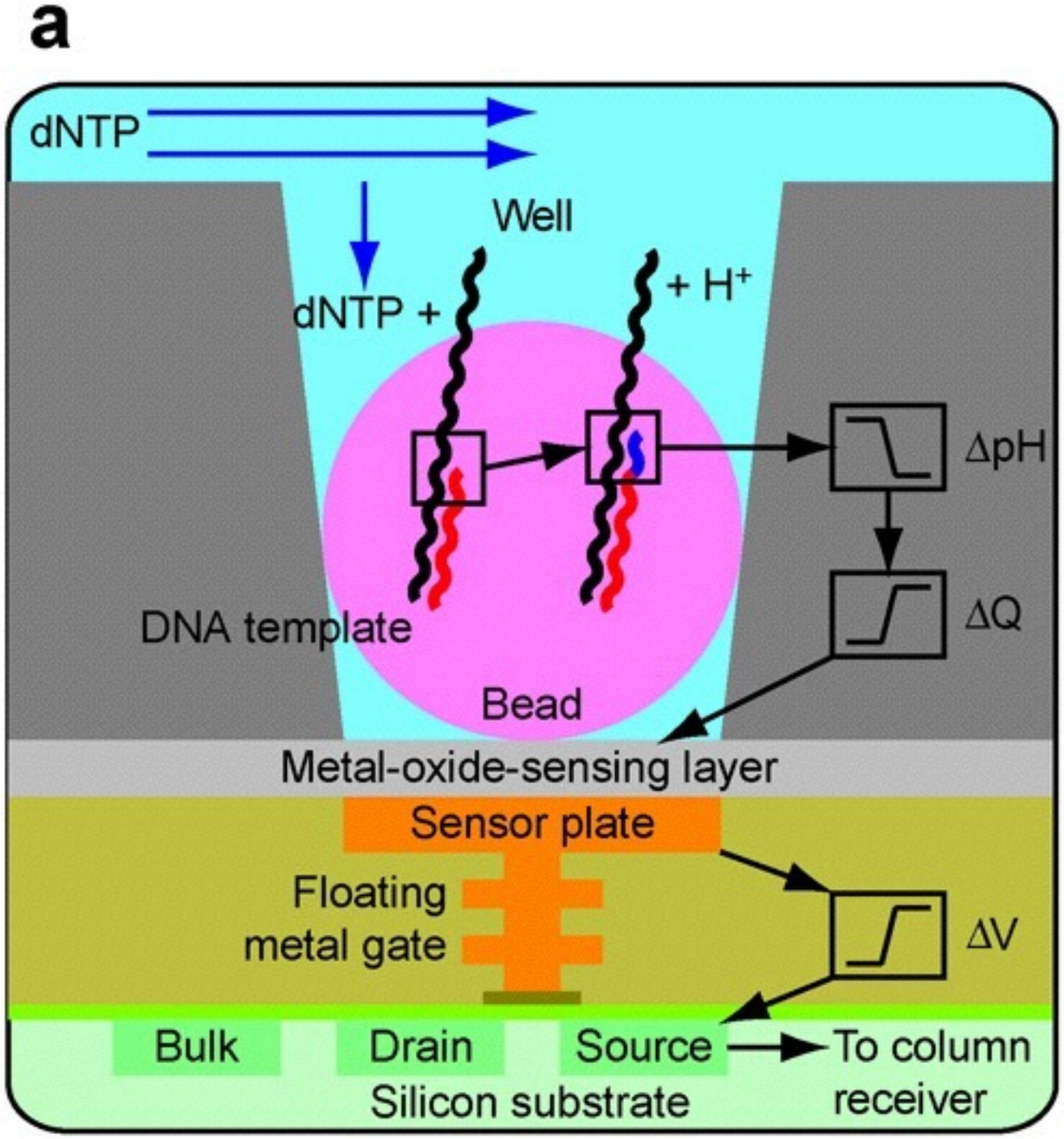
Second generation sequencing

Technology	Read Length	Accuracy	Reads/run	Uses
Illumina	50-300bp	99.9%	2-3 billion	Resequencing General depth
454	700bp	99.9%	1 million	Not currently used
Ion Torrent	400bp	98%	80 million	Cheaper equipment Fast
PacBio	10kb-40kb	87%	up to 1Gbase	Genome assembly Structural variants

Illumina sequencing



Ion Torrent Sequencing



Challenges of short read technology

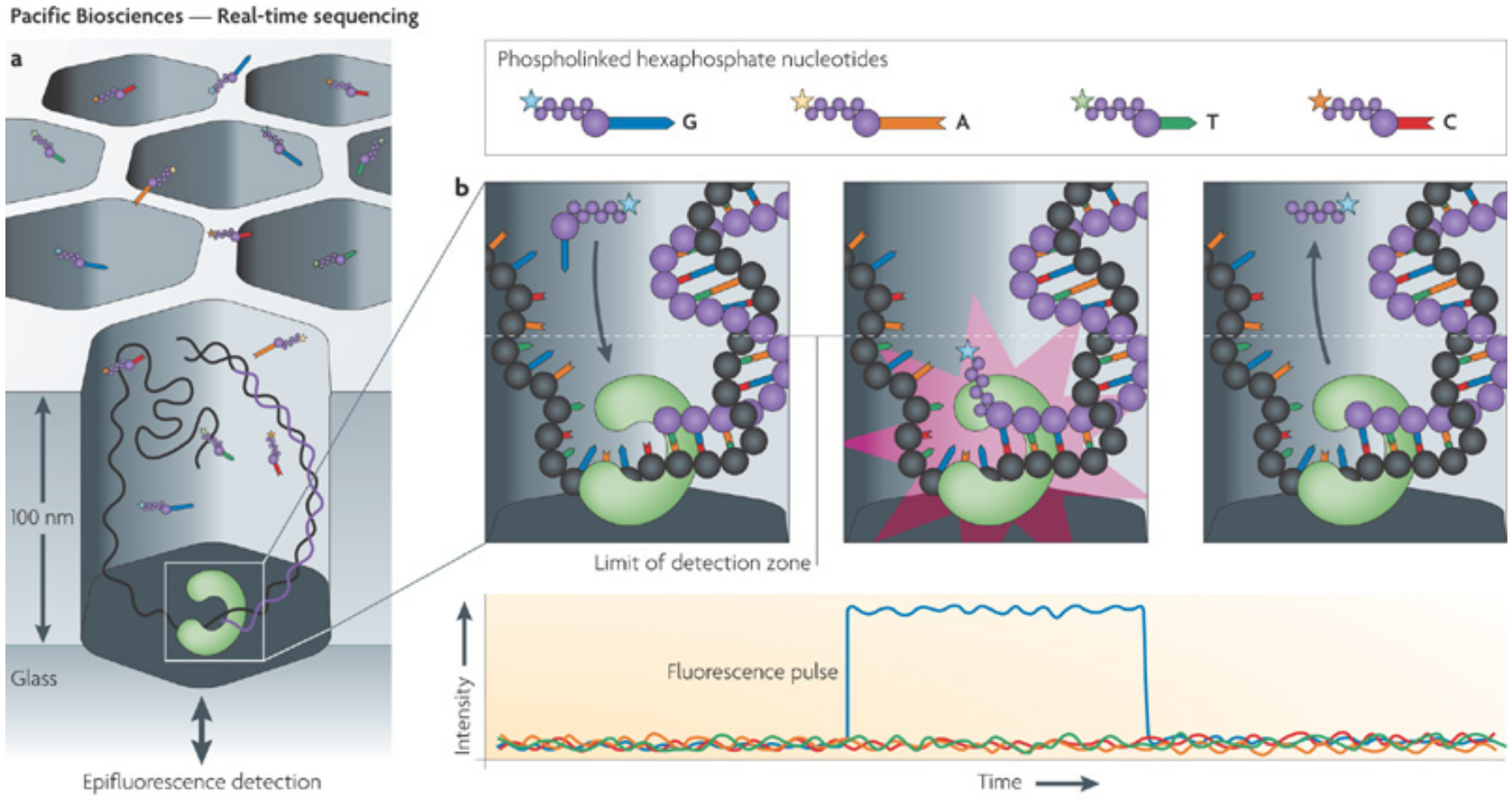
- Rely on amplification, which can introduce errors (10^{-6} -10^{-7}).

- Assembling and aligning reads challenging in repetitive regions

- Difficulty with both large and small structural variants.

Pacific Biosciences Sequencing

Used for making high quality genome assemblies



Challenges of long read technology

- Too expensive to be used for population level sequencing.

- High error rate.

Flavours of sequencing

- Whole Genome Sequencing
- Pool Seq
- RNAseq
- Amplicon Sequencing
- 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

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

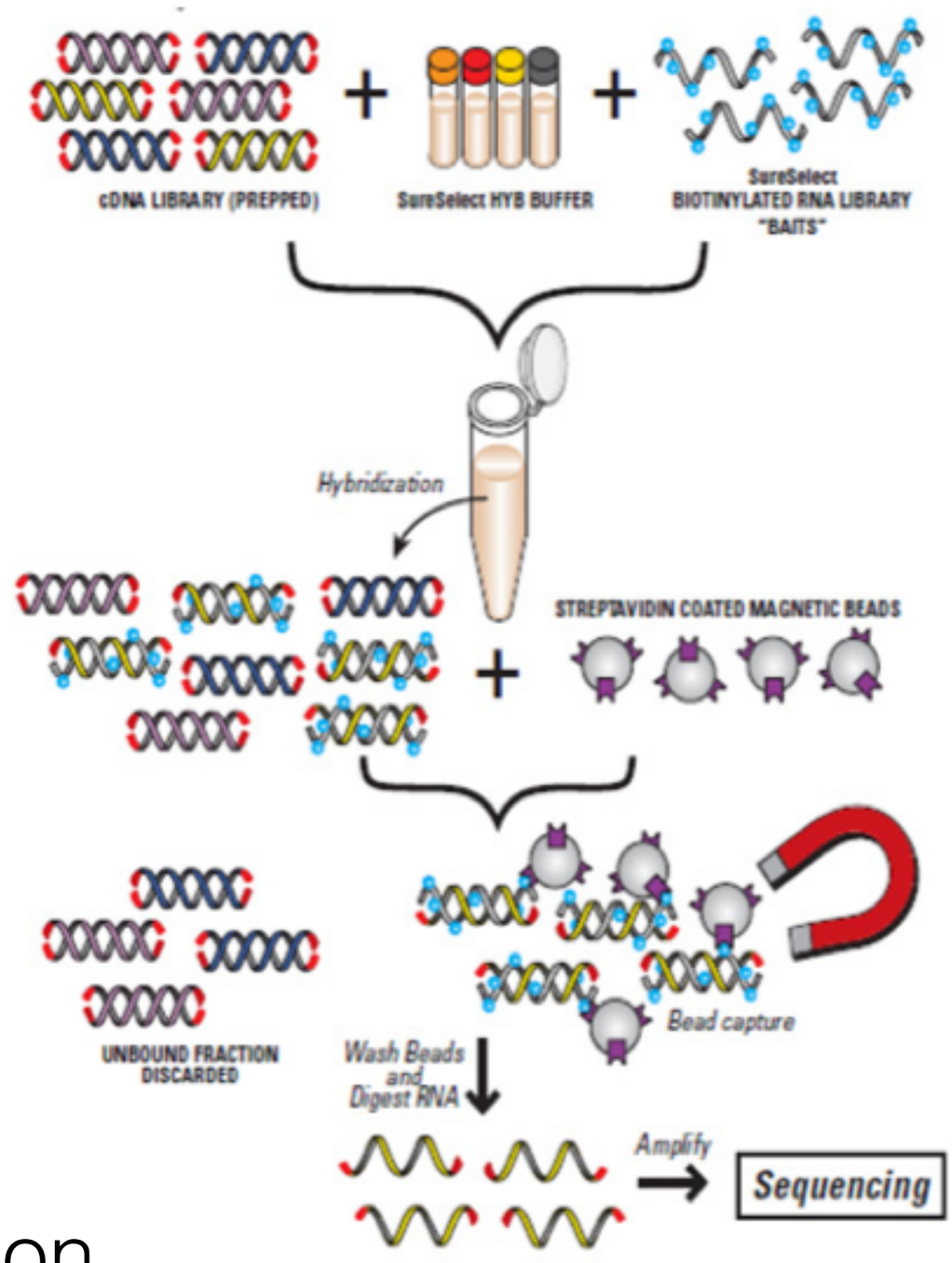
- Make WGS library, hybridize with probe set. Matching sequence will be captured, all others washed away.

- 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.

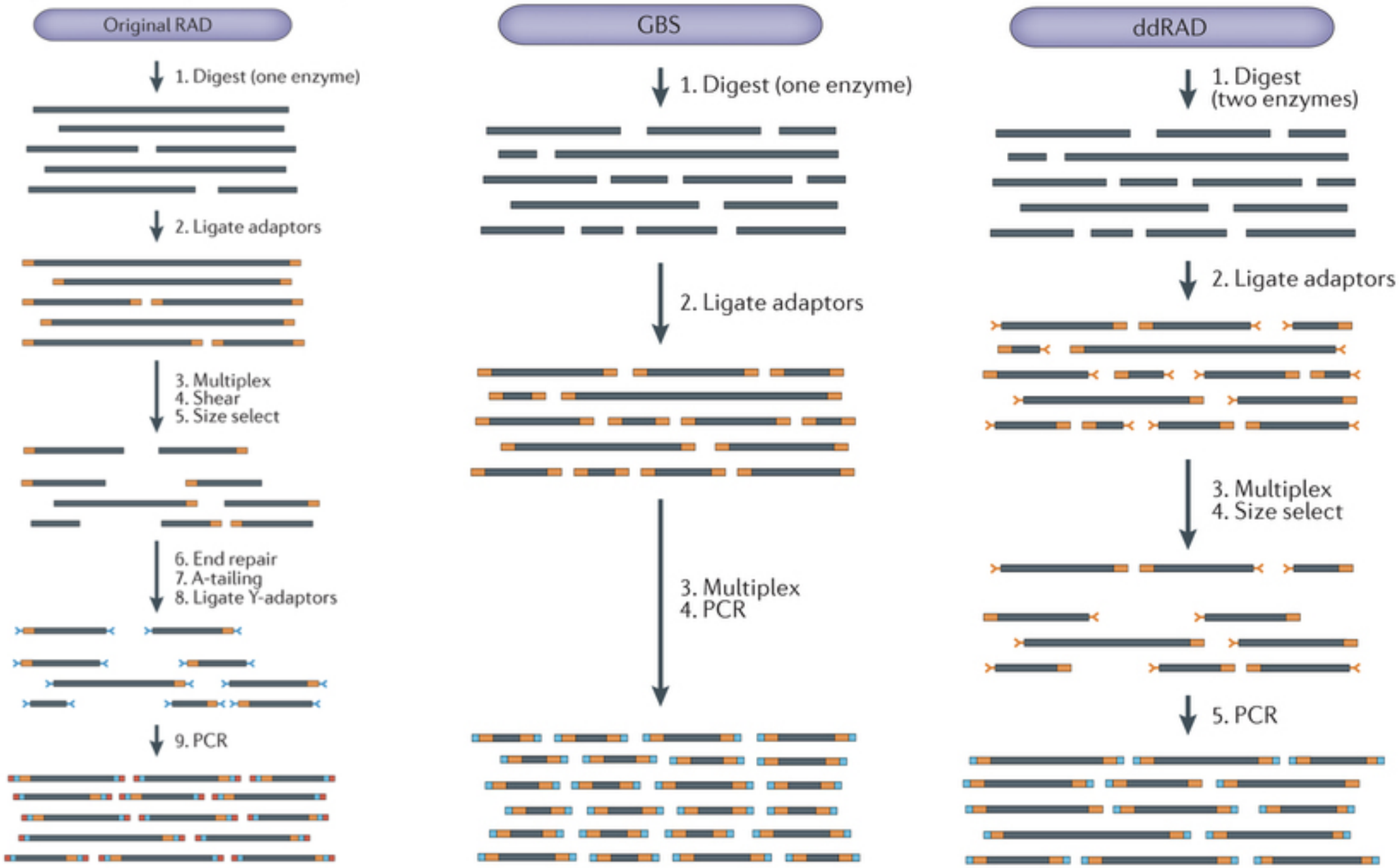


Number of SNPs: 100k - 1 million

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

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

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 (~12 cores, 30 GB ram)
  - Works for small/medium scale analyses
  - Too slow for genome assembly
  - Hard to expand capacity
  - Upfront cost ($5-10k)
  - Complete control

Computing options

- Lab supercomputer (~30 cores, 100 GB ram)
  - Works for small to high scale analyses
  - Managing load between users can be troublesome
  - High upfront cost ($50-100K)
  - Need server management

Computing options

- Westgrid
  - Potentially hundreds of cores
  - Less control and 3 day limit on jobs
  - Free

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
- Moving large amounts of data troublesome/ expensive