Overview of Genomics Data Generation

Fred W. Kolling IV, PhD
Co-Director, Genomics Shared Resource

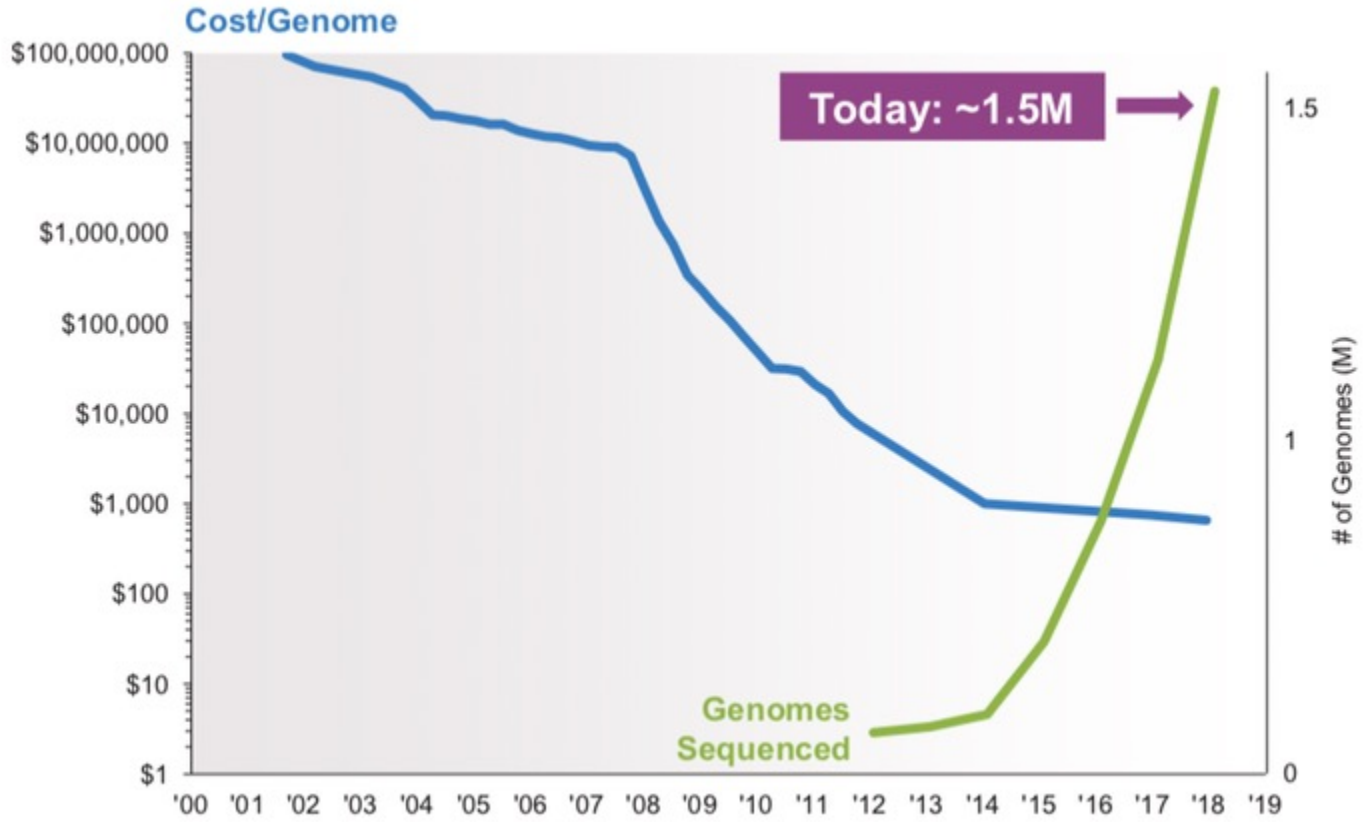
Dartmouth
GEISEL SCHOOL OF
MEDICINE

Genetics vs Genomics: What's the difference?

- Genetics: the study of one or a few genes and their patterns of inheritance
  - Gregor Mendel's pea experiments
  - Pedigree analysis of familial disease

- Genomics: The study of all genetic material within an organism
  - Only made possible in last 20 years
    - High throughput sequencing
    - Computational biology

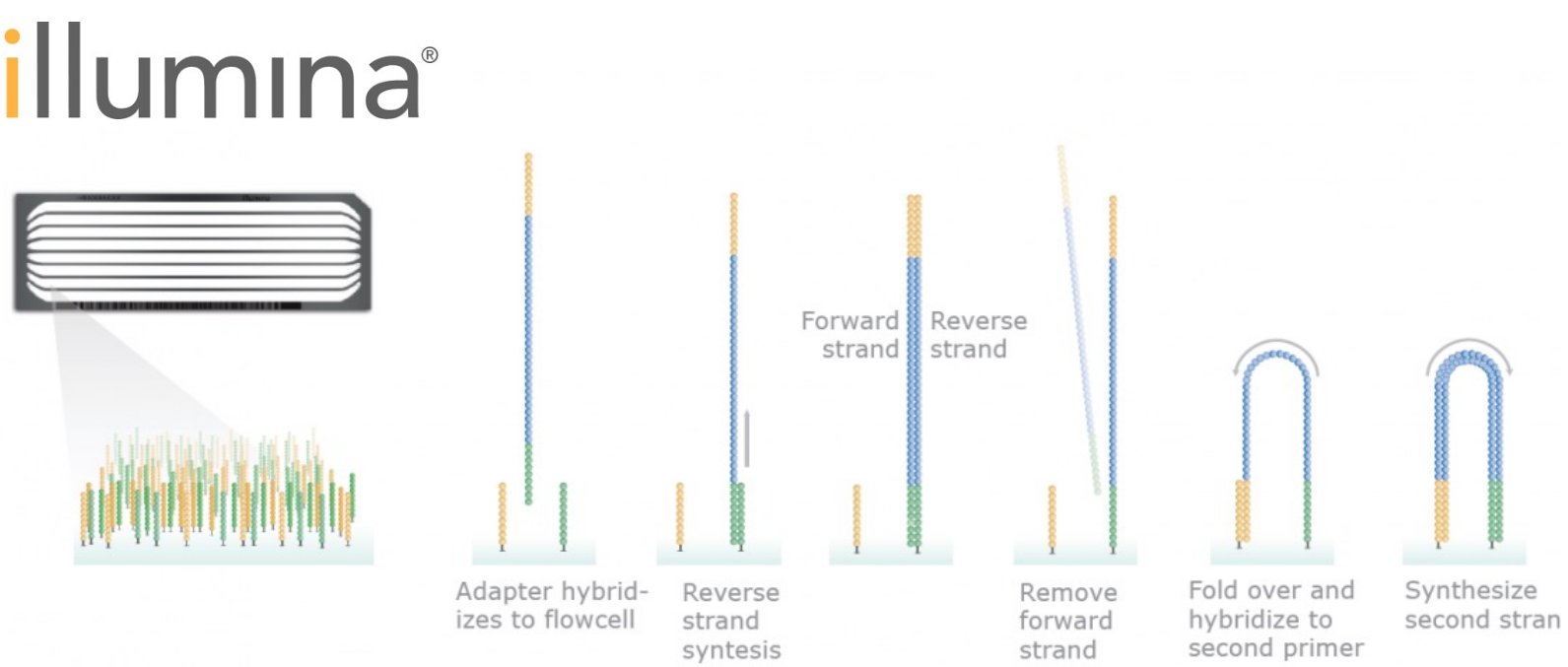
Next Generation Sequencing reduces costs, expands access to genomics technologies

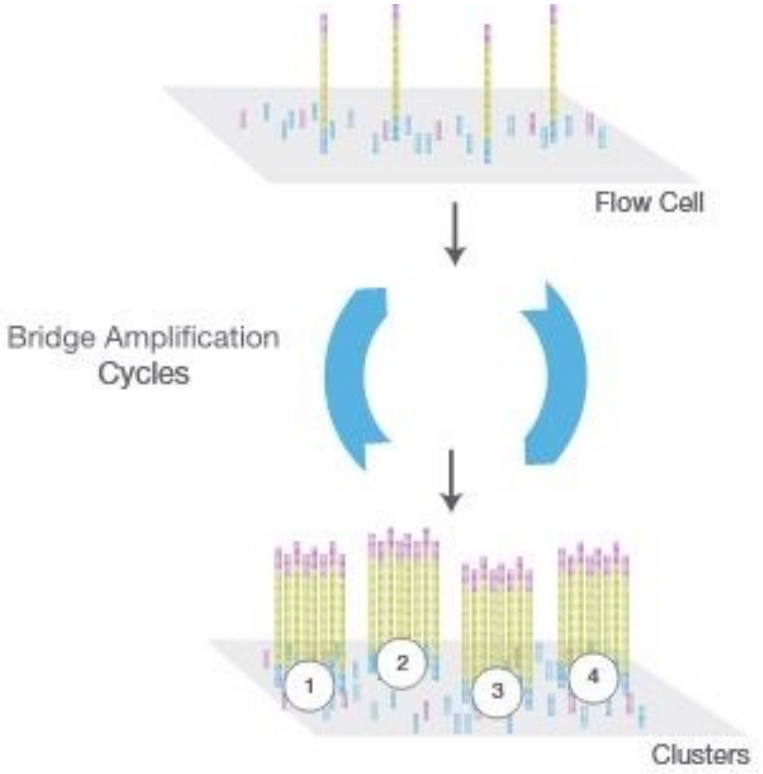


What is Next Generation Sequencing (NGS)?

Sequencing by Synthesis (SBS) Technology

illumina®

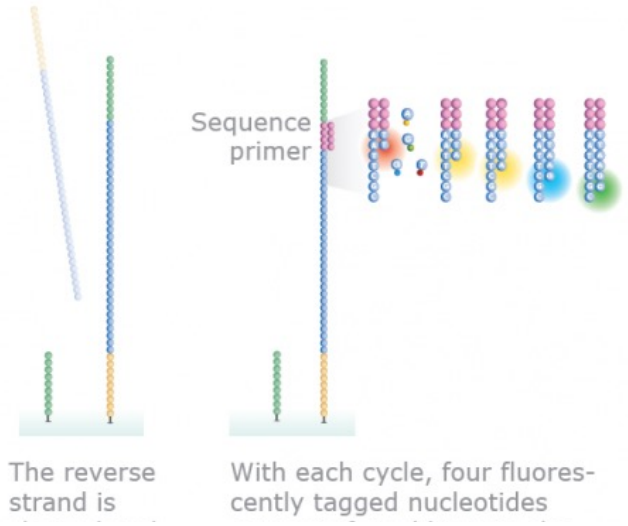


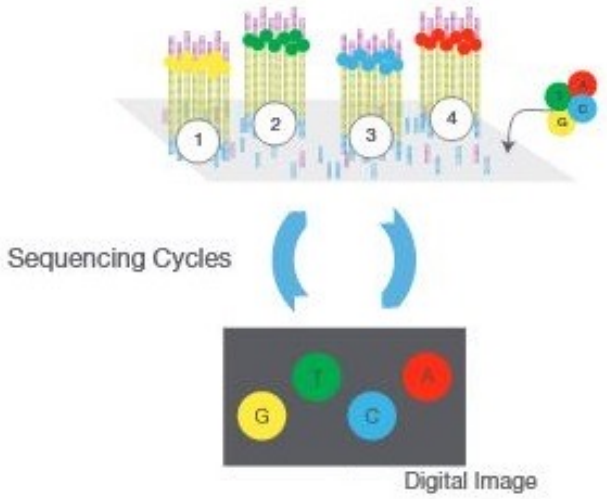


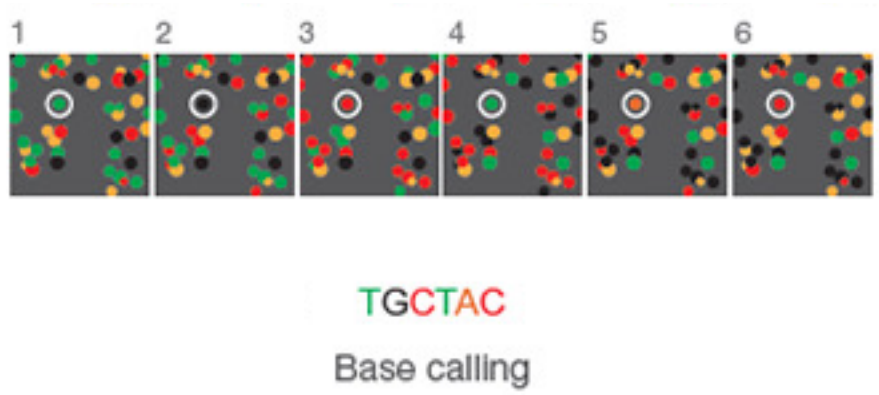
What is Next Generation Sequencing (NGS)?

illumina®

Sequencing by Synthesis (SBS) Technology

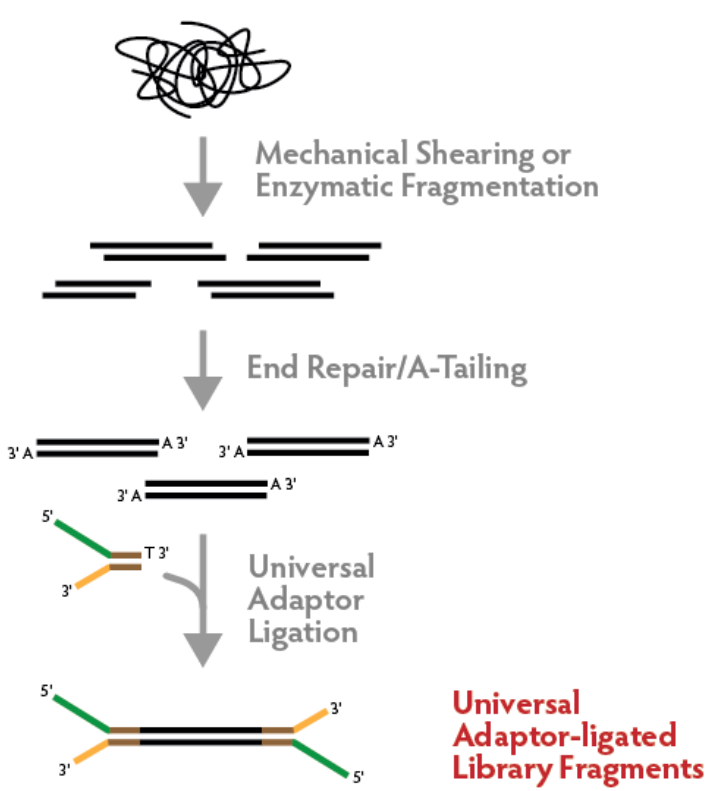


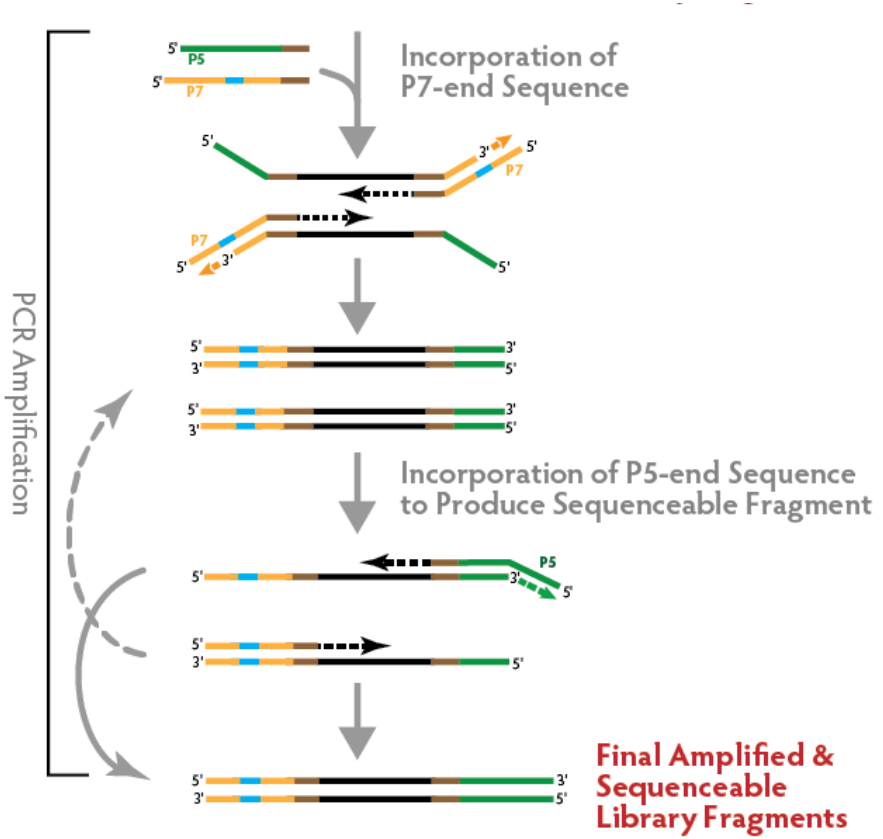




Preparing DNA for NGS

- DNA must be made into a "library" prior to sequencing
  - Add sequences required for binding to flow cell, priming sites for sequencing

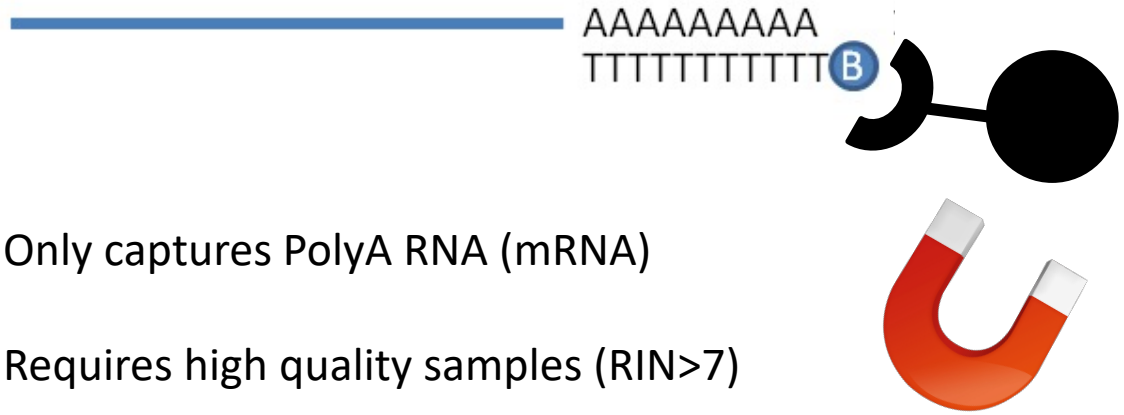




Preparing RNA for NGS

- rRNA accounts for ~95% of RNAs in the cell
  - Not interesting and must remove before sequencing

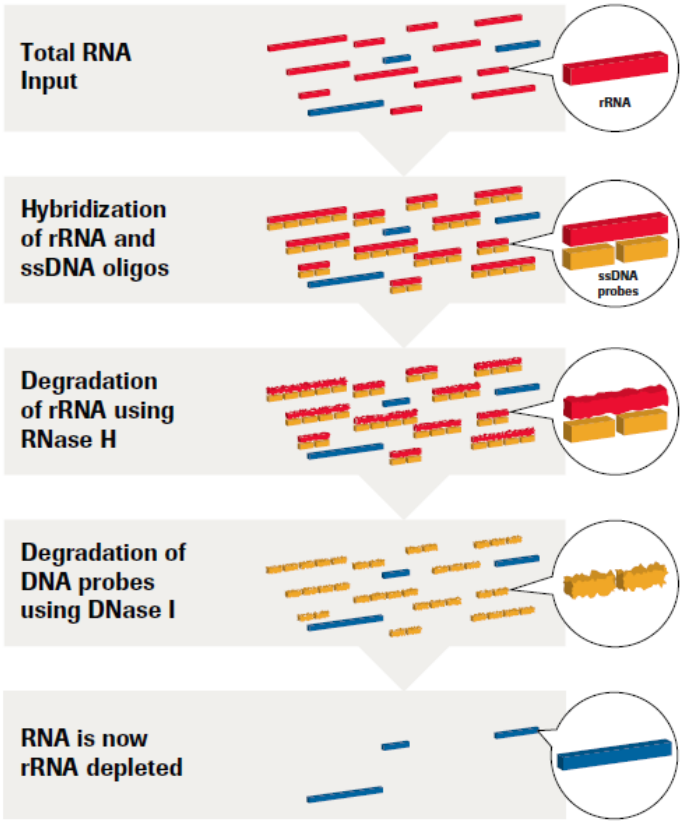
PolyA Capture



Only captures PolyA RNA (mRNA)

Requires high quality samples (RIN>7)

Ribodepletion

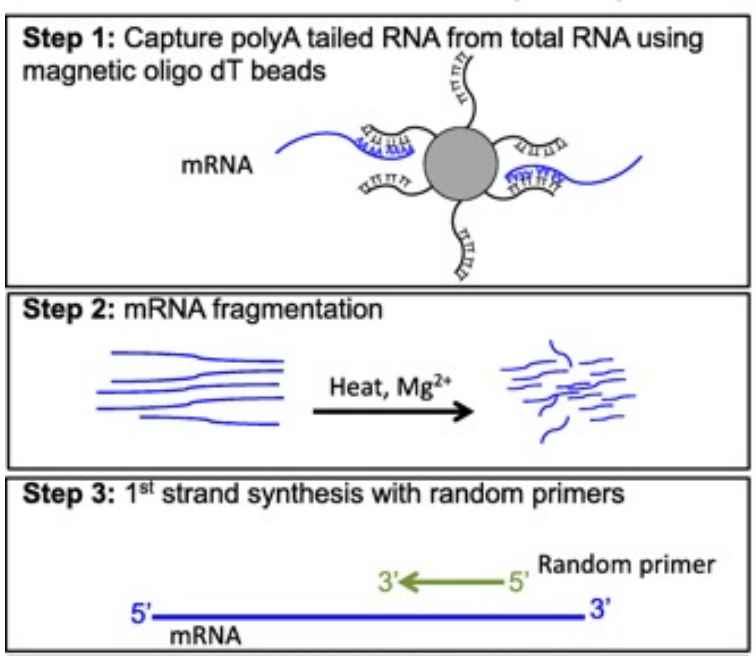


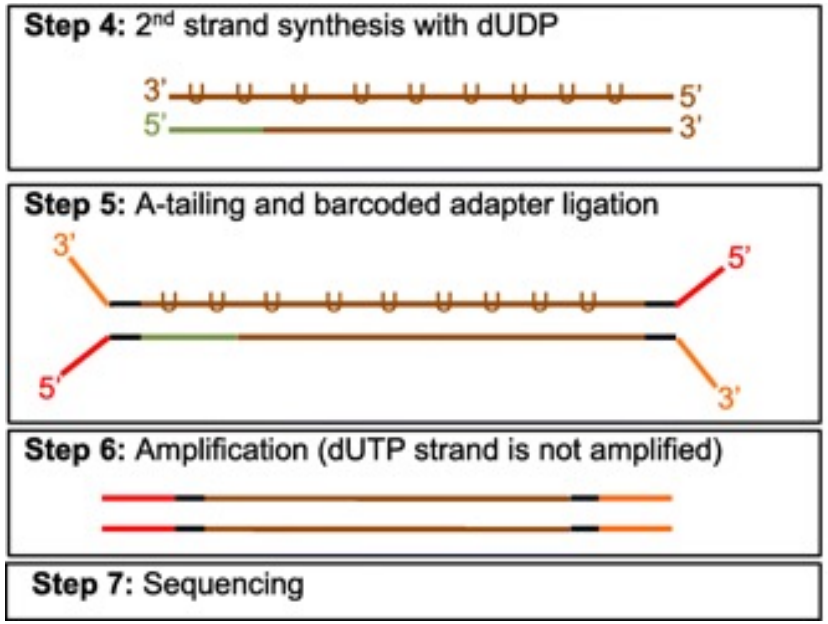
Captures all RNAs >200nt (ncRNAs + mRNAs)

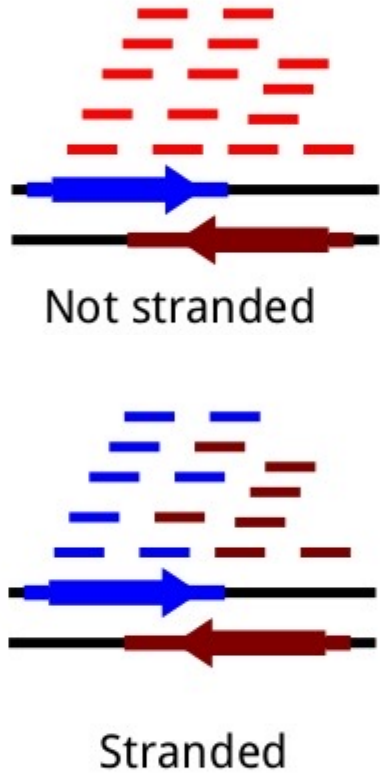
Good for low quality samples

Preparing RNA for NGS

- RNA must first be converted into cDNA, then prepared the same as DNA libraries



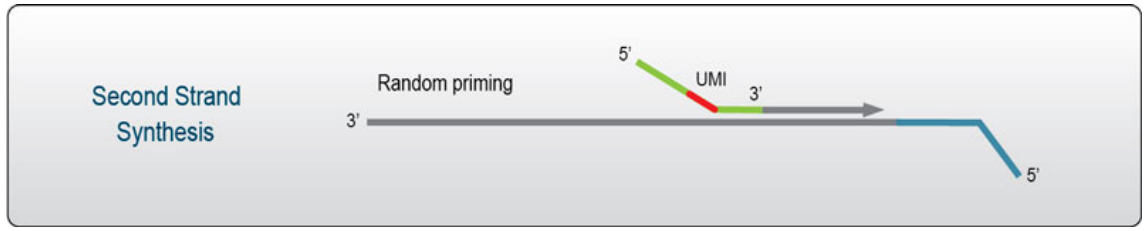




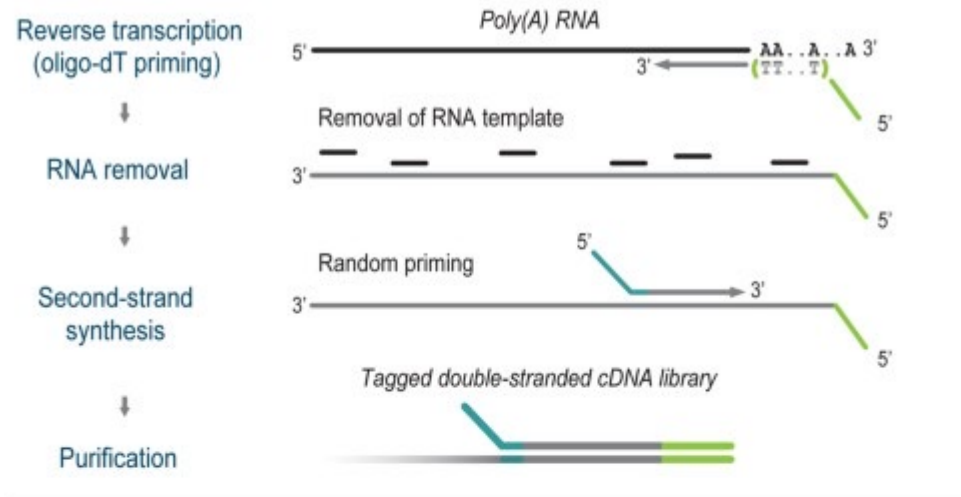
Stranded libraries are essential for identifying overlapping genes, especially antisense transcripts

Modifications to traditional RNA-seq methods

Unique Molecular Identifiers (UMIs) to improve quantification

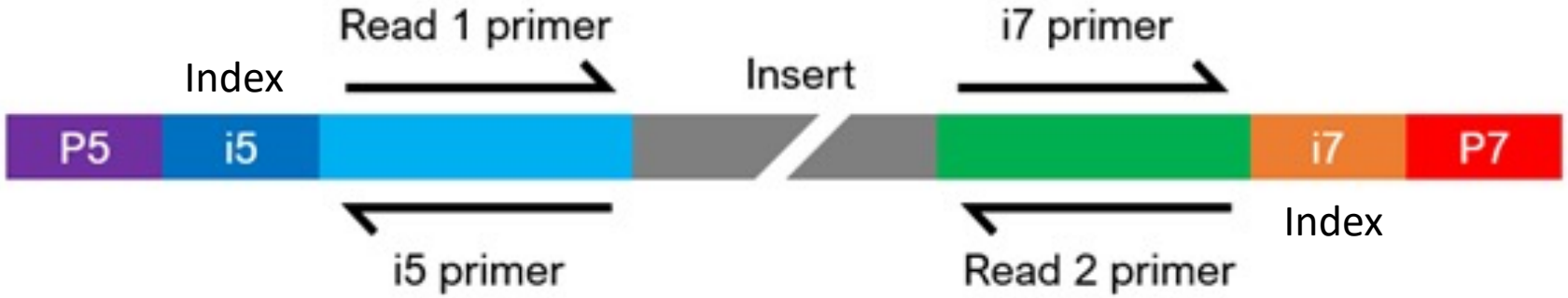


3'-end sequencing to reduce costs for differential expression



Reading libraries on the sequencer

- Inserts can be read in one (single-end) or both (paired-end) directions
- Libraries may contain "index" sequences that are unique to each sample, allowing multiplexing on the same run



Data output from the sequencer: bcl and fastq files

- Basecall files (bcl)
  - Contains raw data regarding fluorescence color and intensity at each cluster for each sequencing cycle
- Convert to fastq →bcl2fastq
- Fastq
  - Nucleotide sequences
  - Basecall quality scores in PHRED format
  - Sequencer information

Example fastq entry

```
AAAAA#EEEEEEEEEEAEEEEEEEEE#E#EEEEEEEEEEEEEEEEEAEEEEEAEEEEE<EEEE/E</<EE<A/A
@NB501031:515:H77NWBGXH:1:11101:9502:1057  1:N:0:AAGAGGCA+TCTTACGC
GAATANGATCAAGTTCAACGGTTTTTCCACAGTGCATTCTGCATCATGCTTCCATGGAGAATAATAGAAATAAGT
```

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%

- Sample1_S1_L001_R1.fastq.gz
- Sample1_S1_L001_R2.fastq.gz
- Sample1_S1_L001_I1.fastq.gz
- Sample1_S1_L001_I2.fastq.gz

Data access and storage

- Data files are large!

- Files are stored on DartFS
  - High speed, redundant disk space
  - Accessible via Discovery/Polaris
  - Can be mounted to your PC/Mac

- Each lab has a folder that is protected with your NetID and password
  - We can add/remove people from your lab group

- Data can also be shared via a web link for external users

- Fastqs are stored for 5 years in read-only format

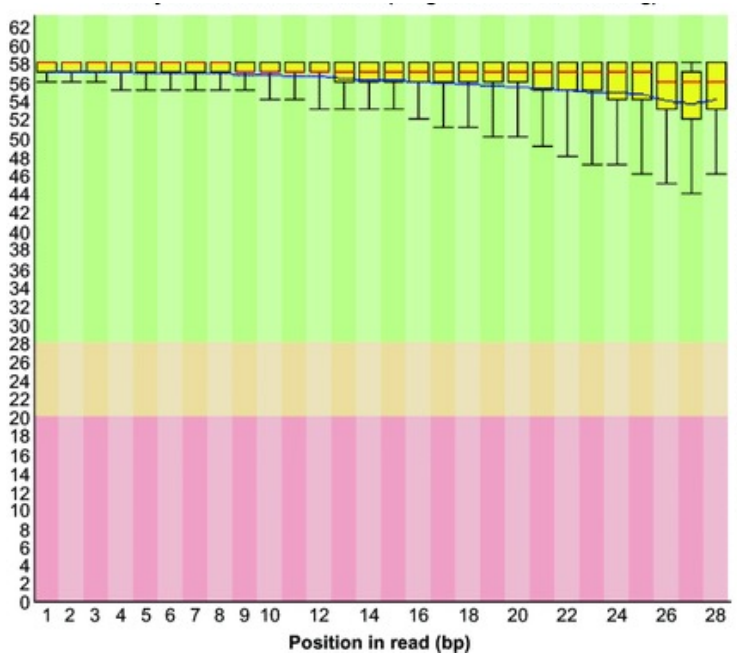
Run Type	Size (Gb)
MiniSeq Mid Output 300 cycle	5Gb
MiniSeq High Output 300 cycle	14Gb
NextSeq Mid Output 300 cycle	30Gb
NextSeq High Output 300 cycle	100Gb

Storage Speed	Time
High Speed (T3)	1yr
Archive (T4)	4yrs

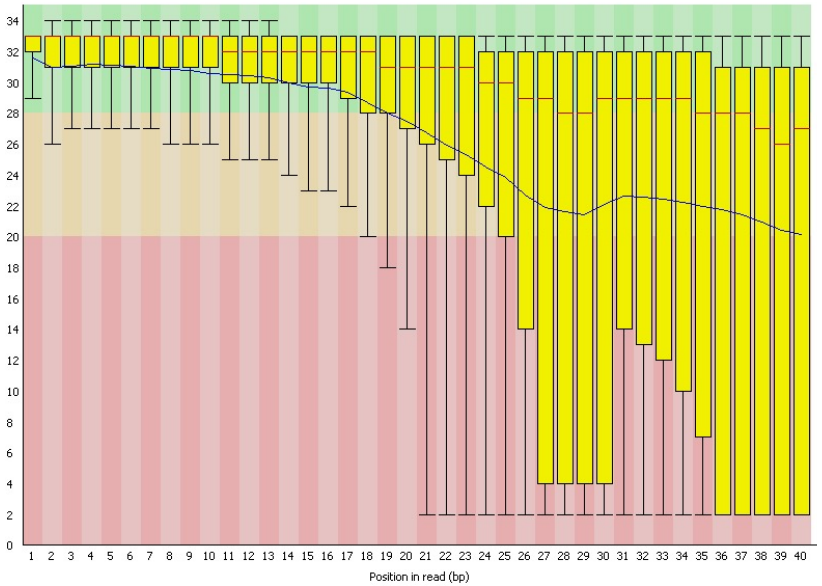
Data output from the sequencer: bcl and fastq files

- Basecall files (bcl)
  - Contains raw data regarding fluorescence color and intensity at each sequencing cycle
- Convert to fastq →bcl2fastq
- Fastq
  - Nucleotide sequences
  - Basecall quality scores in PHRED format
  - Sequencer information

Good Quality

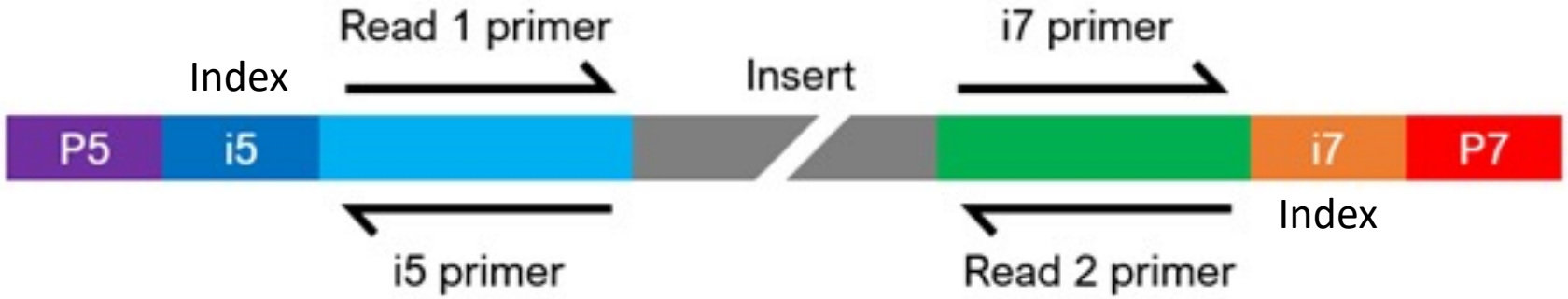


Poor Quality



Preprocessing reads for downstream analysis

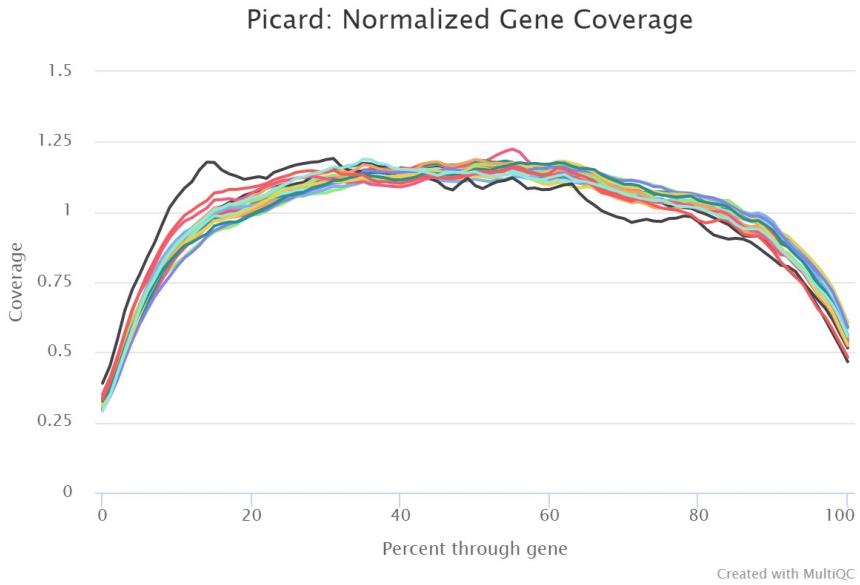
- Remove adapter sequences from resulting from read through
- Remove Poly-A/T sequences from RNA-seq data
- Quality trimming to remove low quality bases
- Parsing UMIs if present



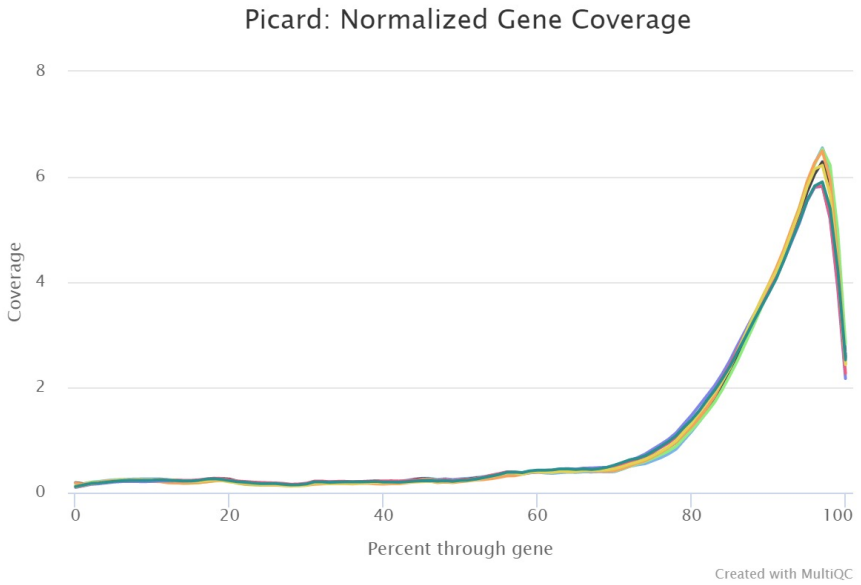
Evaluating coverage uniformity: RNA-seq

- Expect even gene body coverage for full-length methods
  - 3’ bias indicates low sample quality, especially in Poly-A protocol
- 3’-end sequencing methods will of course, have 3’ bias

Even Coverage

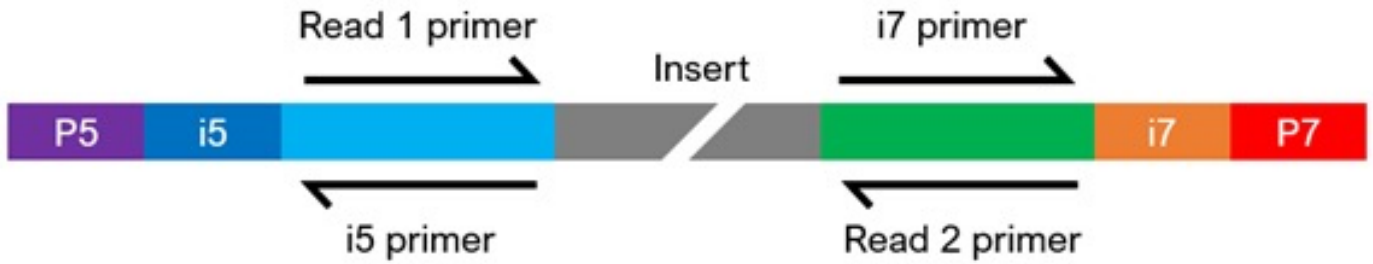


3’ Bias



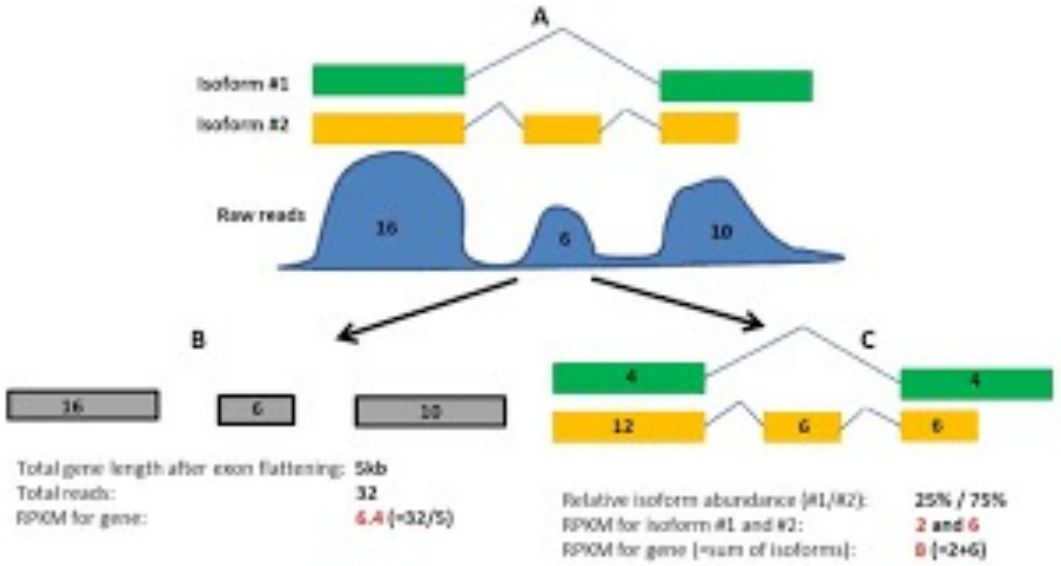
Single-End vs Paired-End Sequencing: DNA-seq

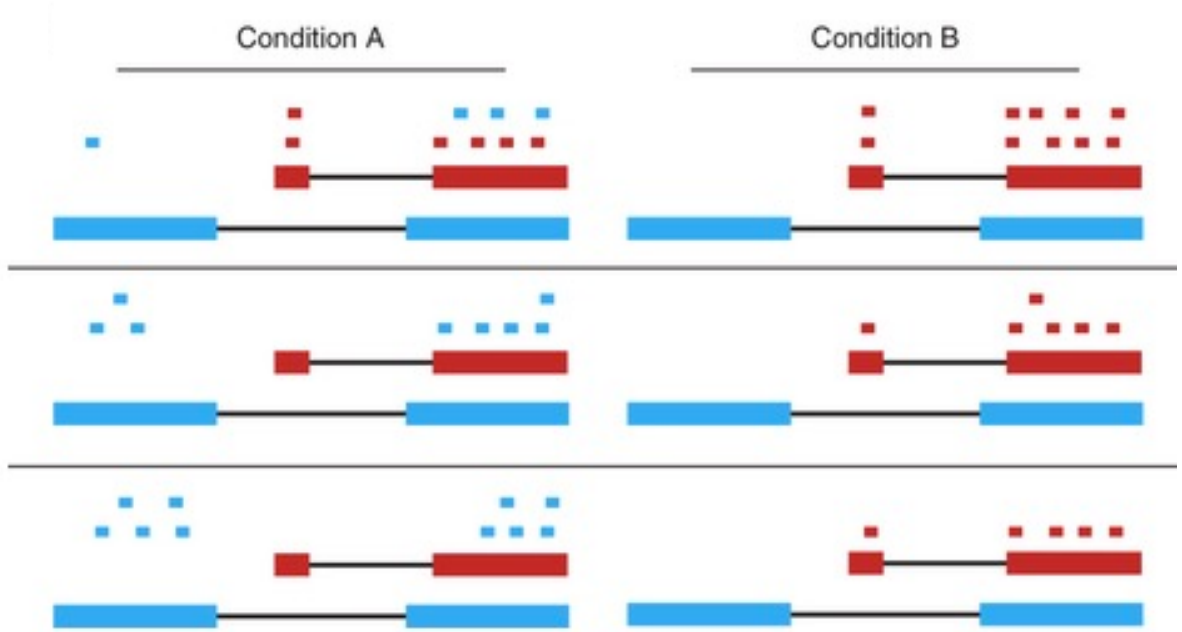
- Paired end sequencing is optimal for DNA-seq
  - Easier mapping to reference
  - Essential for de-novo assembly
  - Essential for identifying structural changes in the genome



Single-End vs Paired-End Sequencing: RNA-seq

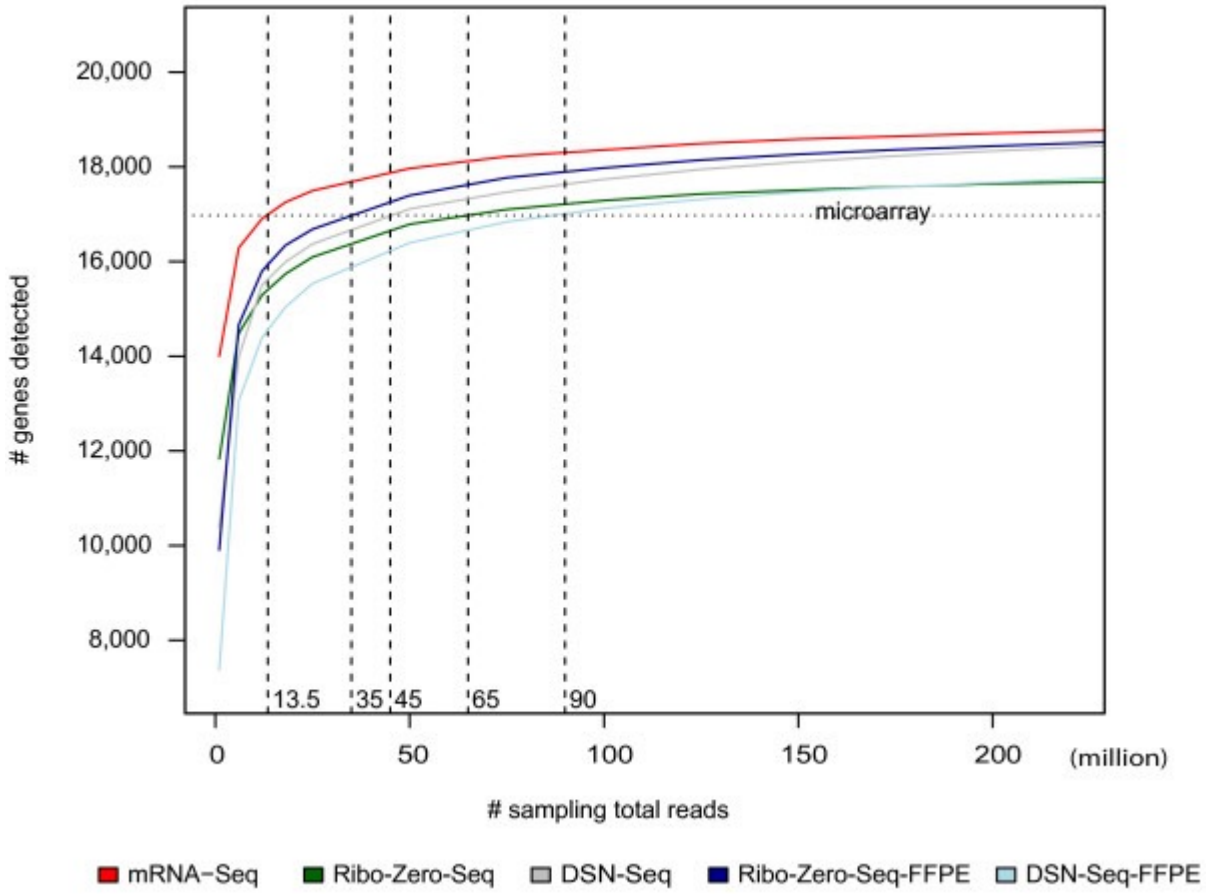
- Single-end vs paired-end
  - To measure overall levels of RNA, single end is usually sufficient
  - Paired-end is preferred if interested in alternative splicing, isoform-level information





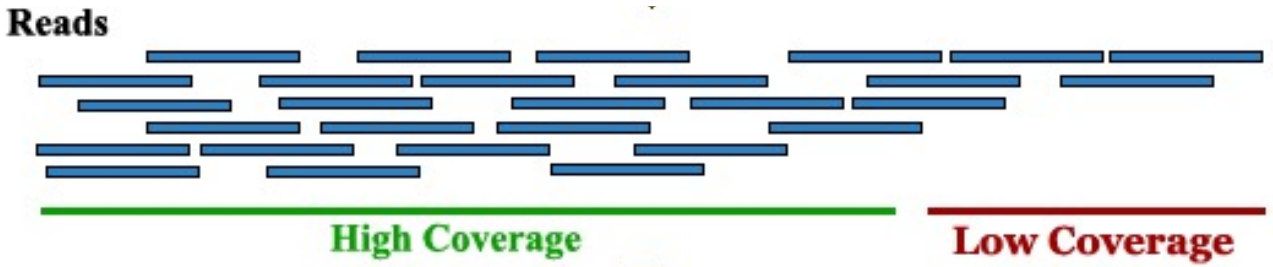
Sequencing coverage: RNA-seq

- Determined empirically
- Transcriptome size varies between cell types
- and across organisms
- Human/mouse/Rat:
  - mRNA → ~15M reads
  - mRNA +lncRNA → 30-40M reads
  - miRNA → 1-10M reads



Sequencing coverage: How much sequencing do I need?

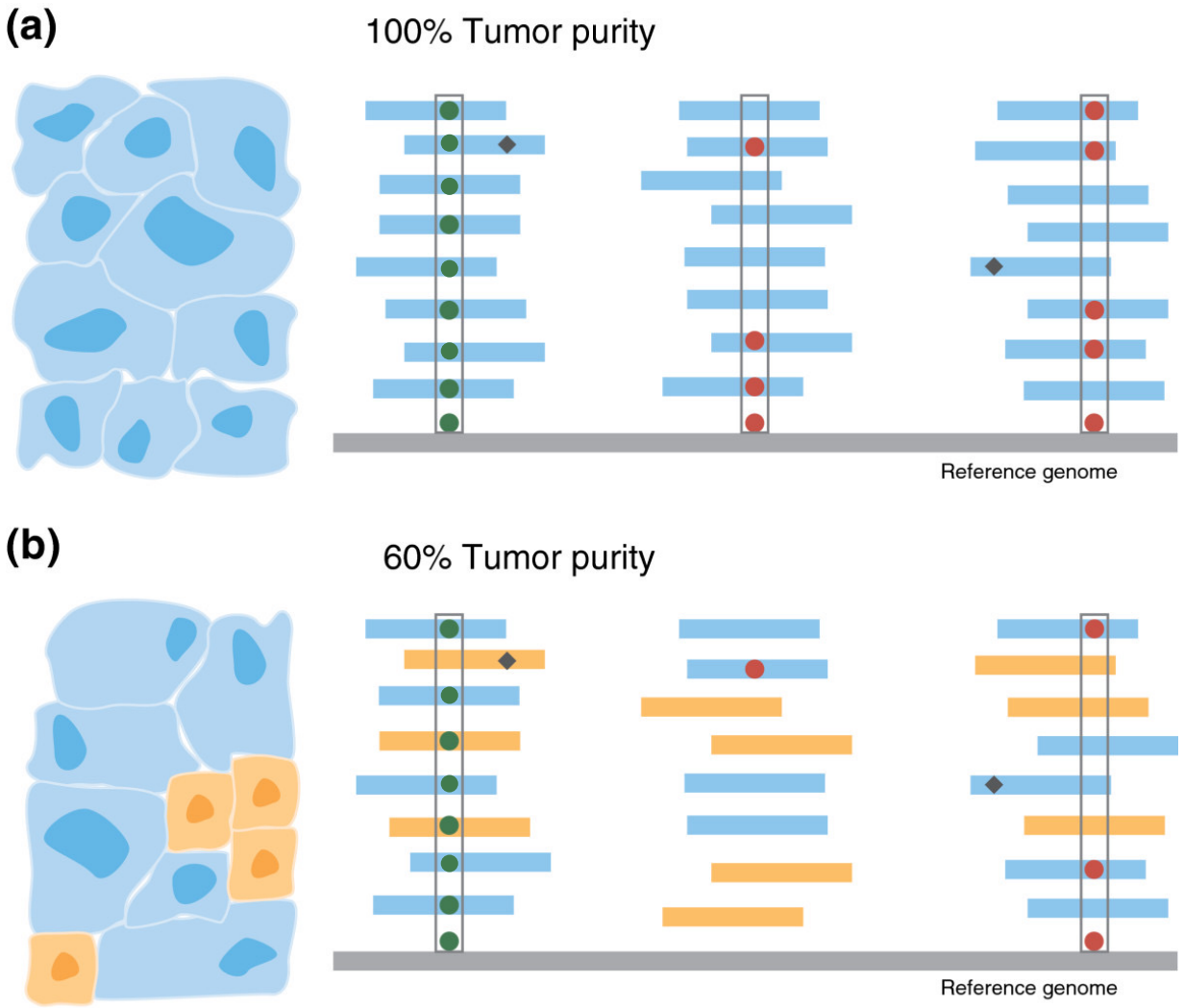
- Sequencing coverage: the average number of reads covering any given base in a dataset



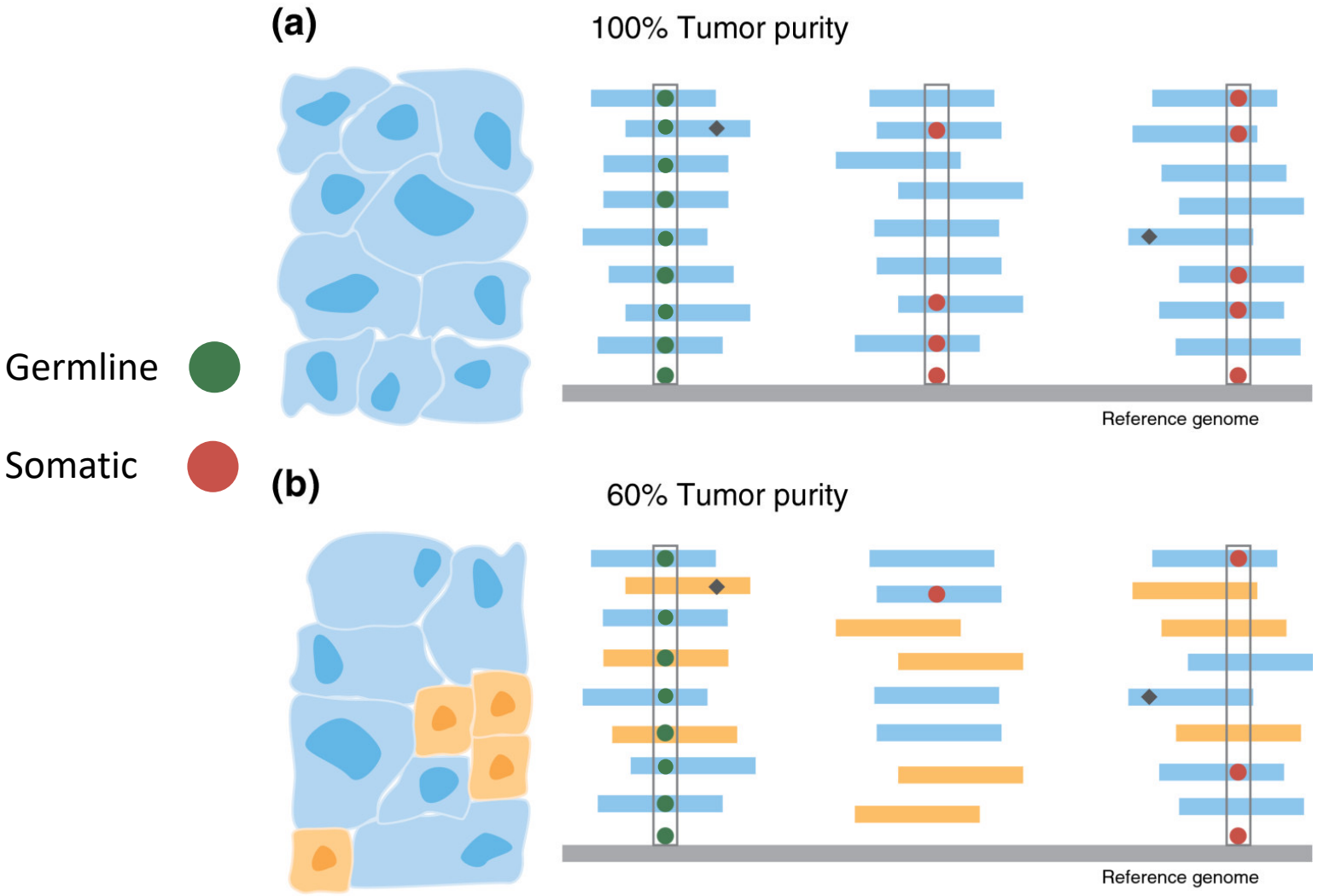
Sequencing coverage: DNA-seq

Germline

Somatic



Sequencing coverage: DNA-seq



What types of alternations am I looking for?

- Copy Number Variation (CNV)
  - Low coverage OK, lower coverage = lower resolution
- Single Nucleotide Variants (SNVs)
  - Germline
    - Less coverage required (20-50x)
  - Somatic
    - Depends on allele frequency (100-200x typical)

I can sequence the whole genome/transcriptome, should I?

- Whole Genome Sequencing (WGS)
  - Provides comprehensive view of the genome →Do I need all of this information?
    - Do I only care about changes in protein coding genes? Regulatory information important?
  - Expensive
  - Can I handle this much data?

- Targeted sequencing
  - Whole Exome Sequencing
    - Enrich and sequencing only protein-coding regions of the genome (i.e. exome)
  - Disease-specific or custom panels
    - Are there specific genes/mutations I am interested in?

- Do I need high coverage? (i.e. rare SNVs)
  - Focus sequencing efforts where needed

Cost considerations in WGS vs targeted sequencing

NextSeq 550 Series

Run Time	12–30 hours
Maximum Output	120 Gb
Maximum Reads Per Run	400 million
Maximum Read Length	2 × 150 bp

Cost = $5,500/run

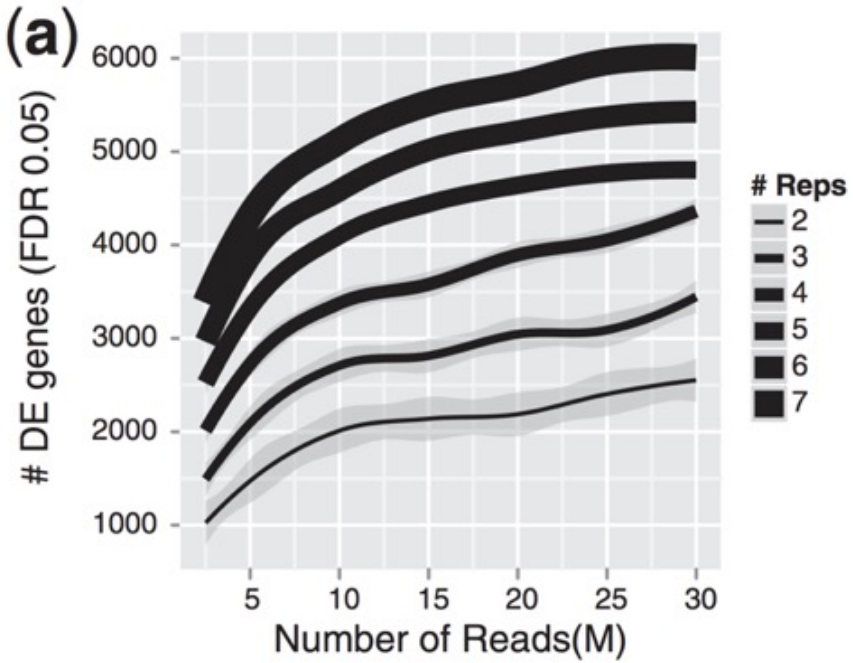
Sequencing read depth = ([Coverage] * [genome/panel size]*[1+duplication rate])/[read length(bp)]

235M reads = ([20x] * [3.2Gb] * [1.1]) / [300bp] → ~2 whole genomes/run

31M reads = ([200x] * [42Mb] * [1.1] / [300bp] → 13 whole exomes / run

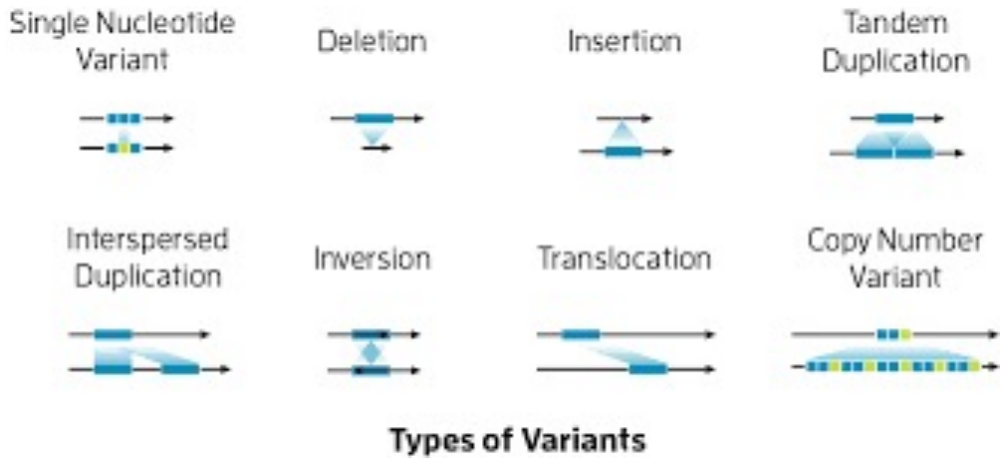
Targeted sequencing allows more replicates for better experimental design

- By focusing on targets of interest, sequencing costs are reduced
- Resources can be redirected towards additional biological replicates
- Replicates increase ability to detect differentially expressed genes

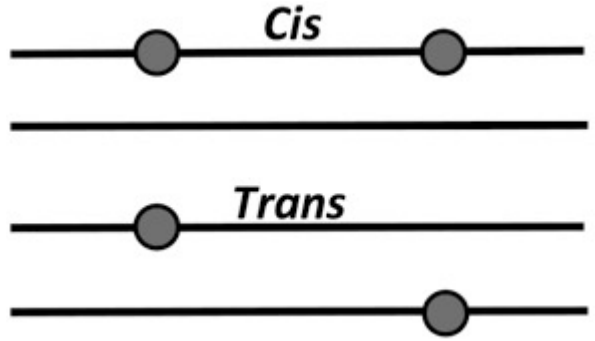


Disadvantages of short-read sequencing

- Challenging to detect large-scale structural variation
  - Inversion, non-reciprocal translocations, duplication
- Hard to map reads to repetitive regions
  - Telomeres, centromeres and others
- Hard to map reads to genes with closely related paralogs or pseudogenes
  - There are over 100s of clinically actionable variants that are missed by short read sequencing because reads cannot be confidently mapped
- Unable to determine for many genes, whether two mutations occur in cis or trans
  - Requires haplotype "phasing"

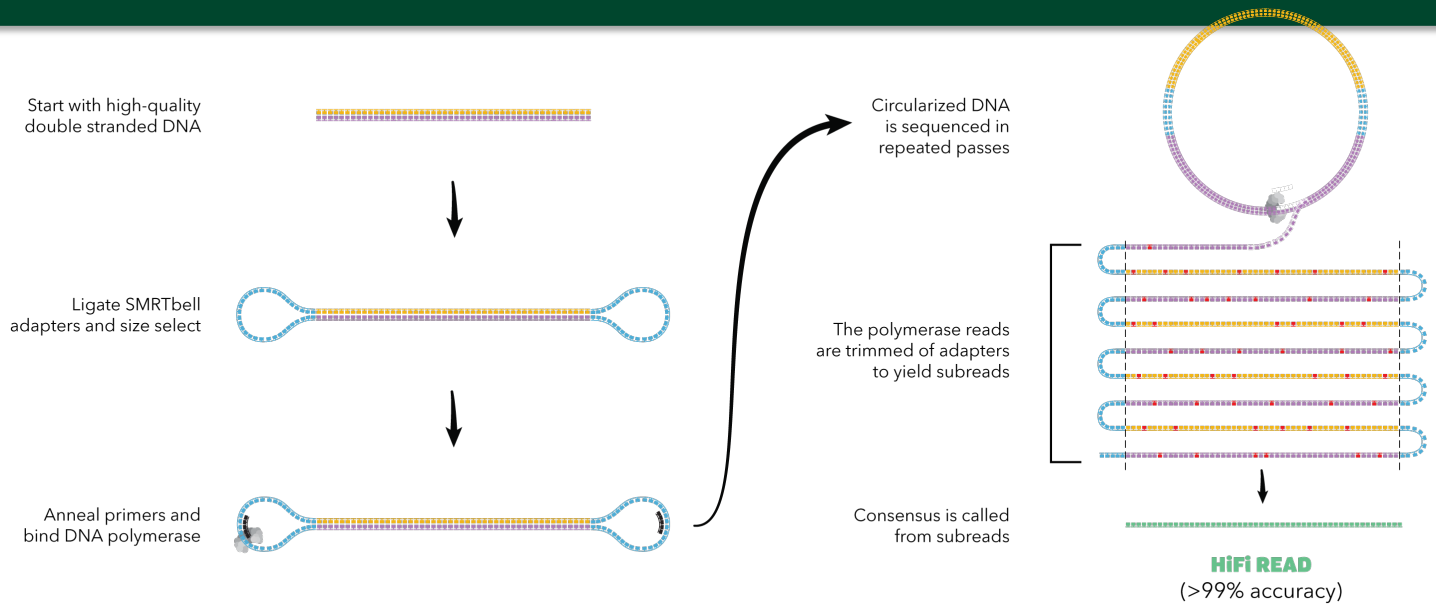


Types of Variants

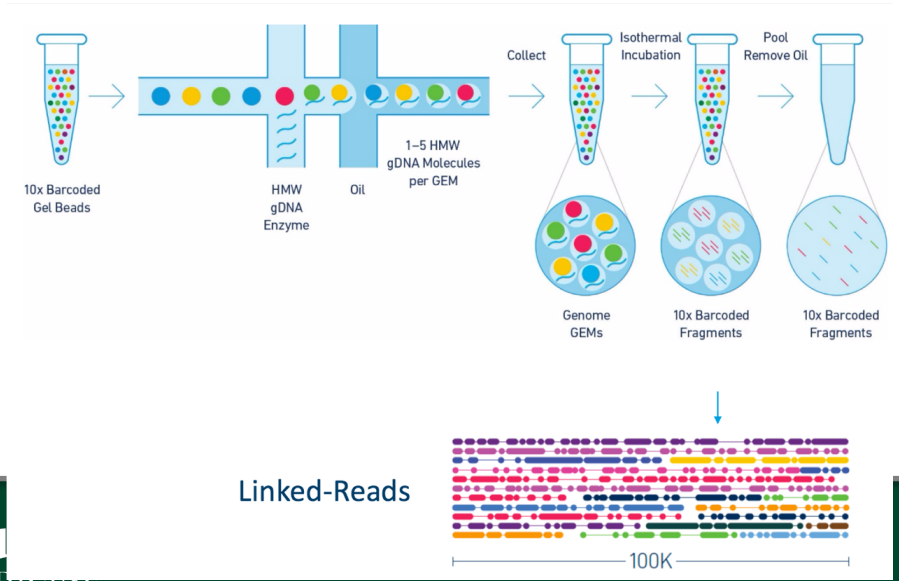


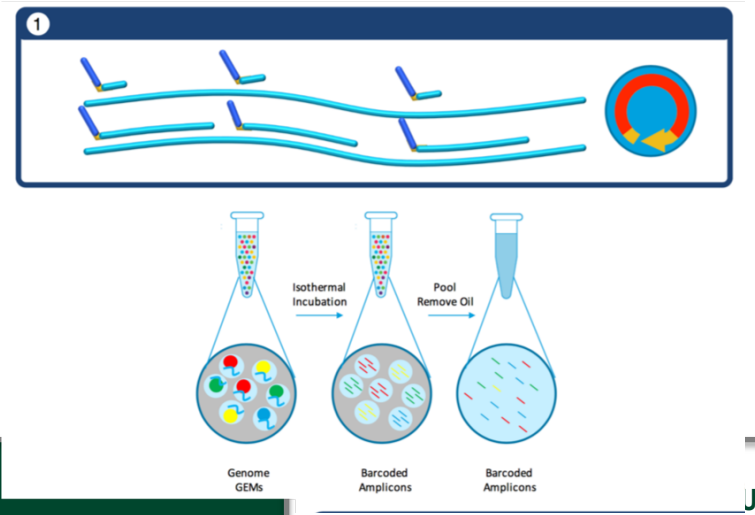
Long-read sequencing complements short-reads

Long Read Sequencing
(PacBio)



Synthetic Long-Reads
(Linked Reads)





3rd generation sequencing: direct sequencing using nanopores

- Advantages
  - VERY long reads → as long as the molecule you put in
  - Detection of base modifications
    - DNA methylation
    - RNA modifications
  - No need to convert RNA to DNA
  - Portable

- Disadvantages
  - High error rate (90bp/sec, too fast to read)
  - Need to prepare high molecular weight DNA