

# Genomics & Bioinformatics

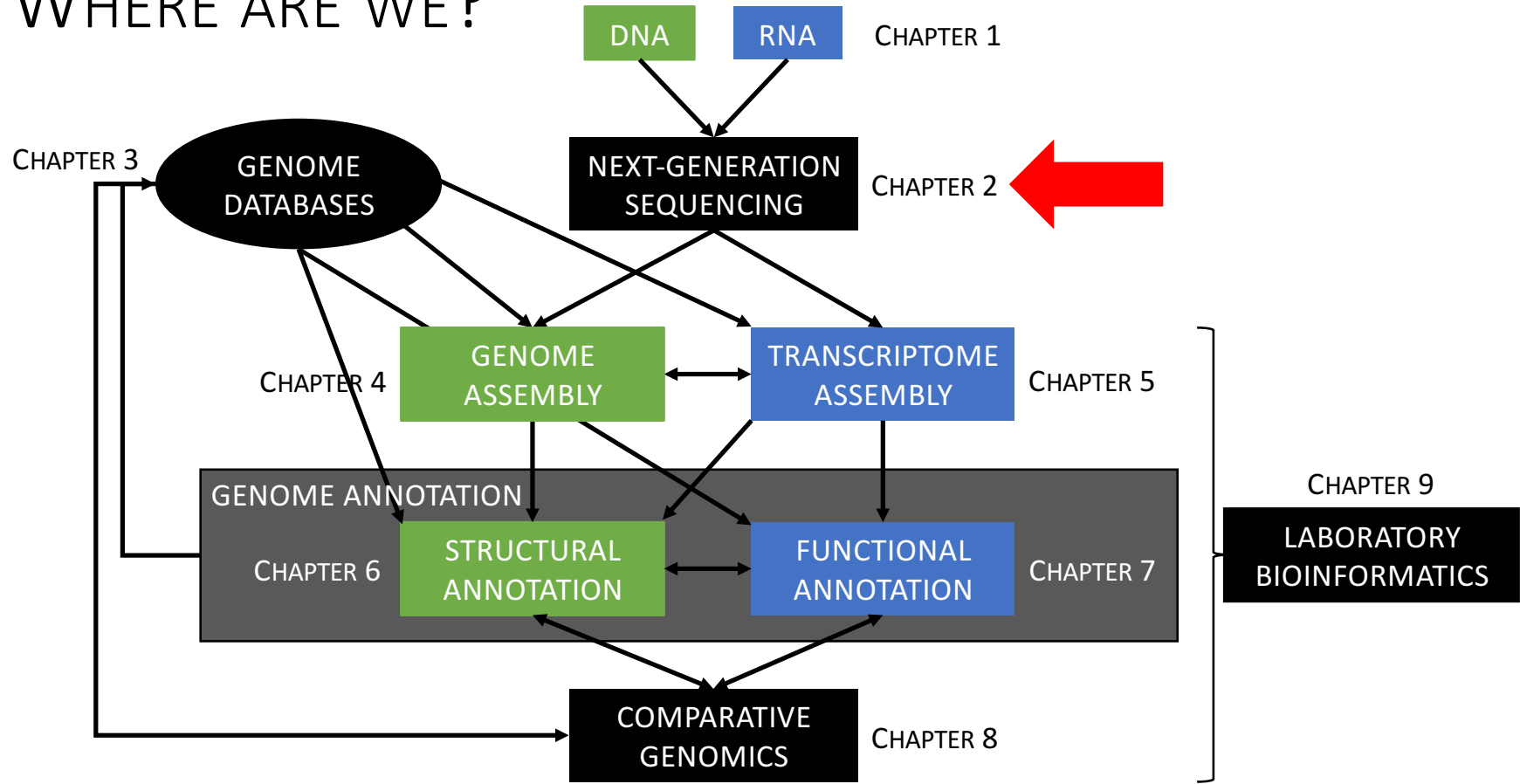


BIOL 497, 597

Boise State University

Spring

# WHERE ARE WE?



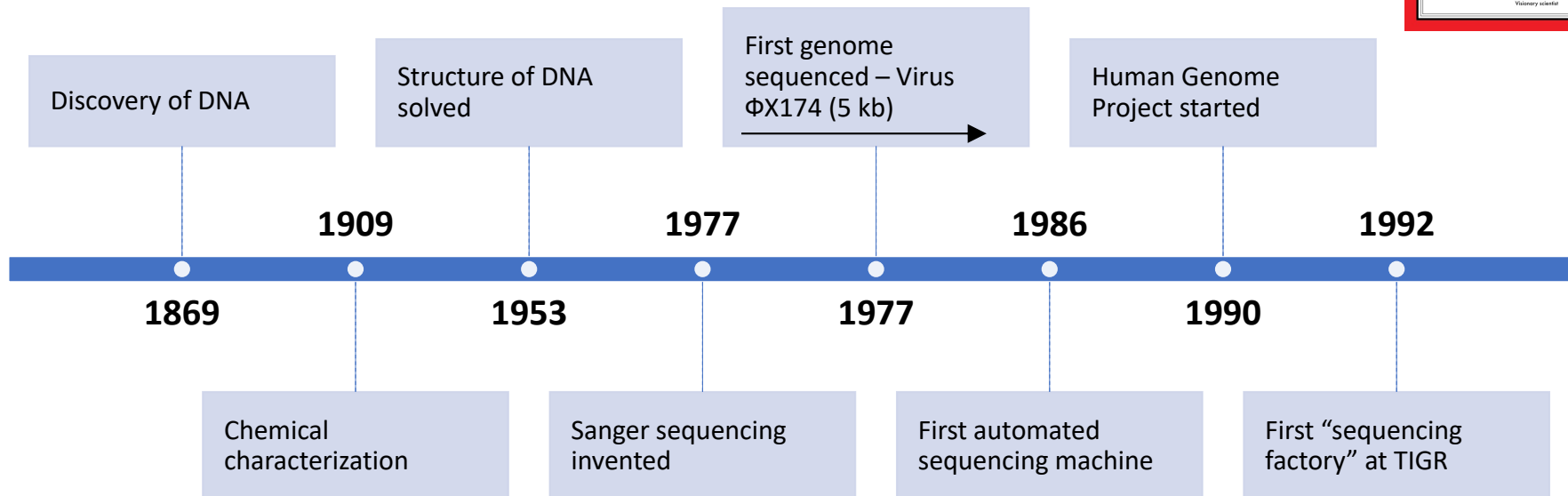
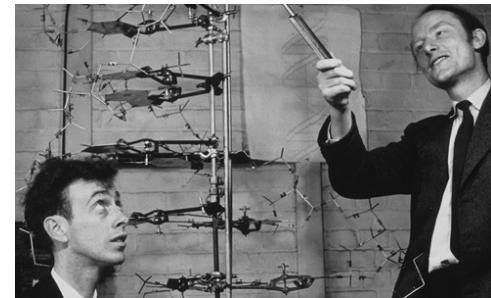
## Question (Review Chapter 1)

- What are the 2 main pieces of evidence required to develop a sequencing strategy to assemble (*de novo*) an eukaryote reference genome?

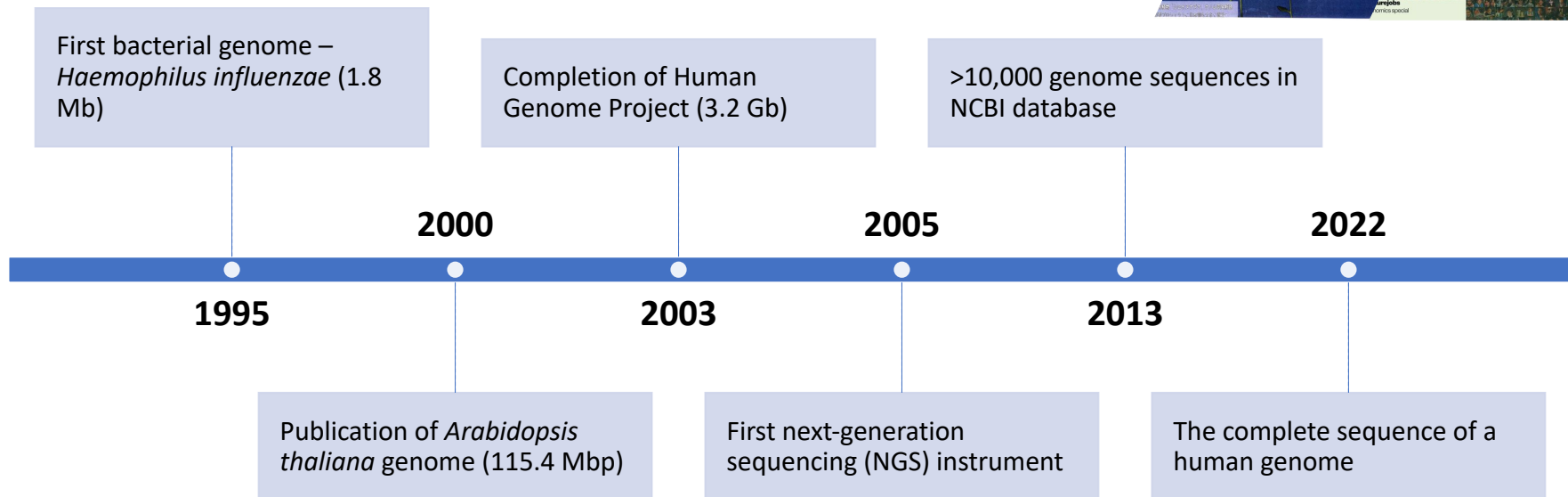
# A QUICK HISTORY OF SEQUENCING

Franklin

Watson & Crick

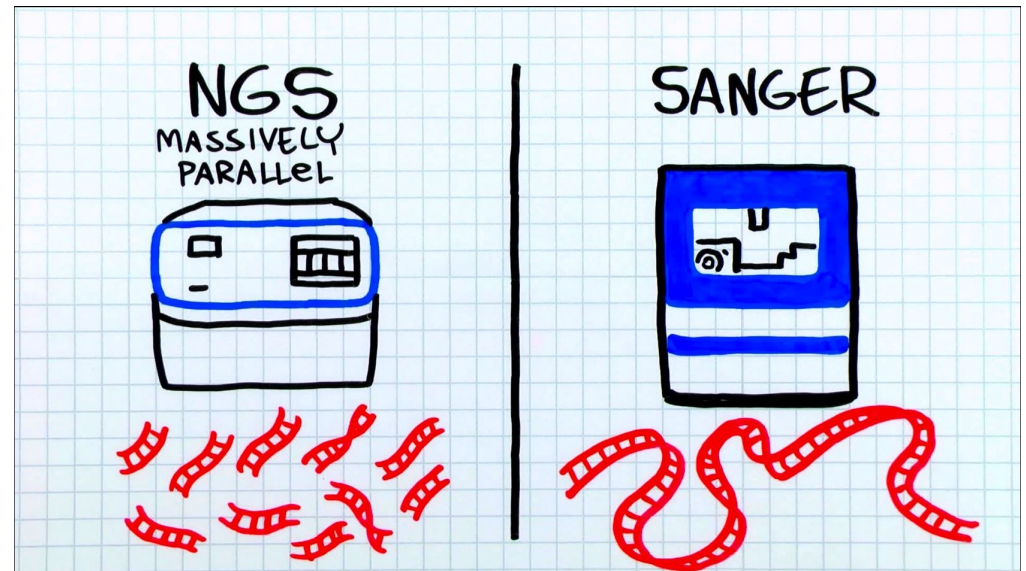


# A QUICK HISTORY OF SEQUENCING



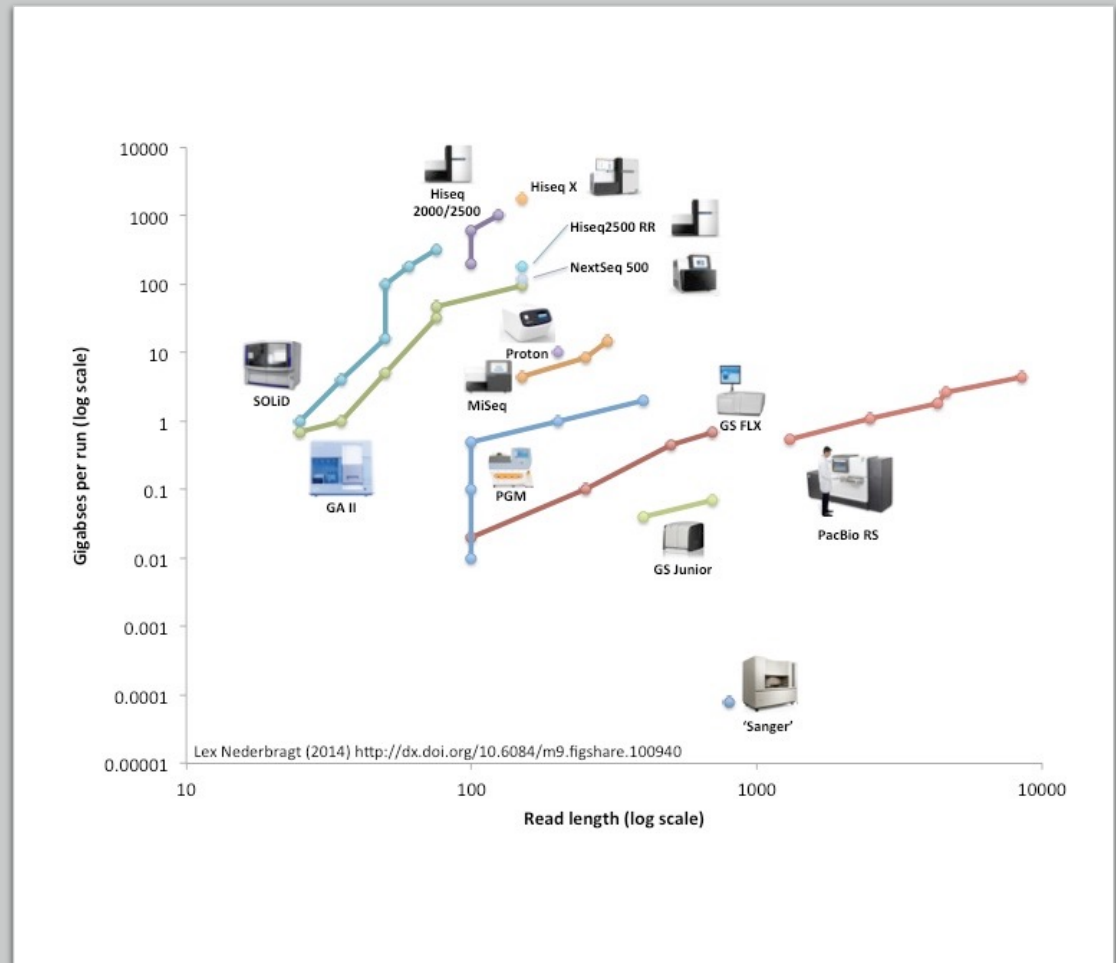
# WHAT IS NGS?

- **NGS** is the term applied to methods enabling **thousands of millions of DNA fragments to be sequenced in parallel in a single experiment.**
- NGS outperformed Sanger sequencing, which was able to sequence only individual DNA fragments, each fragment obtained by a different PCR.



# WHAT IS NGS?

- **NGS platforms can produce gigabases (billion bases) of data per run.**



Type	Instrument	Primary Errors	Single-pass Error Rate (%)	Final Error Rate (%)
Short reads	3730xl (capillary)	substitution	0.1-1	0.1-1
	454 All models	indel	1	1
	<b>Illumina All Models</b>	substitution	~0.1	~0.1
	Ion Torrent – all chips	indel	~1	~1
	SOLiD – 5500xl	A-T bias	~5	≤0.1
Long reads	<b>Oxford Nanopore</b>	deletions	≥4*	4*
	<b>PacBio RS</b>	Indel	~13	≤1

## WHAT IS NGS?

- NGS methods have higher base calling error rates than Sanger (here 3730xl), especially methods producing long reads (but they are improving very fast!).





RNA Sequencing



Whole Genome Sequencing



Exome Sequencing



Targeted Sequencing



Amplicon Sequencing



Sequencing Only



Single-Cell Services



PacBio Services



Metagenomics



Immunogenomics



Epigenomics



Spatial Genomics



AAV Sequencing



Custom



Plasmid-EZ



Precision Antibody Discovery

SEQUENCING COMPANIES PROVIDE A WIDE ARRAY OF SERVICES — WHICH ONE SUITS MY NEEDS?

# WHAT ARE THE DIFFERENT TYPES OF SEQUENCING?

**Whole-genome sequencing (WGS):** The process of determining the complete DNA sequence of an organism's genome at a single time.

**Targeted re-sequencing:** Sequencing of selected regions of a genome; for instance, a subset of genes (using either baits or amplicons). This approach is now commonly applied for phylogenetic inferences.

## WHAT ARE THE DIFFERENT TYPES OF SEQUENCING?

**Exome sequencing:** Exome sequencing is the targeted enrichment and sequencing of the whole exome. Exome sequencing is the most powerful tool available for the identification of genetic variations associated with a phenotype, such as a disease.

**RNA-Seq:** Sequencing RNAs in a sample. RNA sequencing is carried out by reverse transcribing the RNAs, and sequencing the complementary DNA (cDNA).

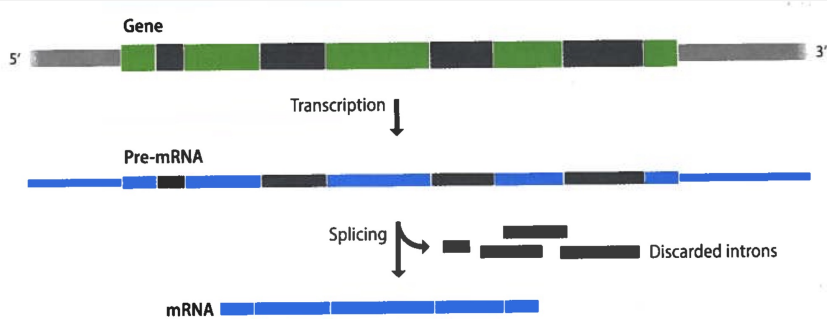
# COMPARING EXOME SEQUENCING AND RNA-SEQ

Exome sequencing looks at the DNA contained in exonic regions of the genome, whereas RNA-Seq looks at RNA transcribed from DNA, much of which, but not all, derive from exonic regions (mRNA).

The exome is the protein-coding portion of the genome (~2-4% of mammalian DNA), comprised of exons. Sequencing the exome involves a targeted approach.

Exome sequencing is often used to identify SNPs (correlated to e.g. diseases). Exome sequencing is preferred for SNP and mutation analyses.

RNA-seq provides information regarding alternative spliced forms, relative gene expression level.



# OVERVIEW OF GENOME SEQUENCING & ASSEMBLY

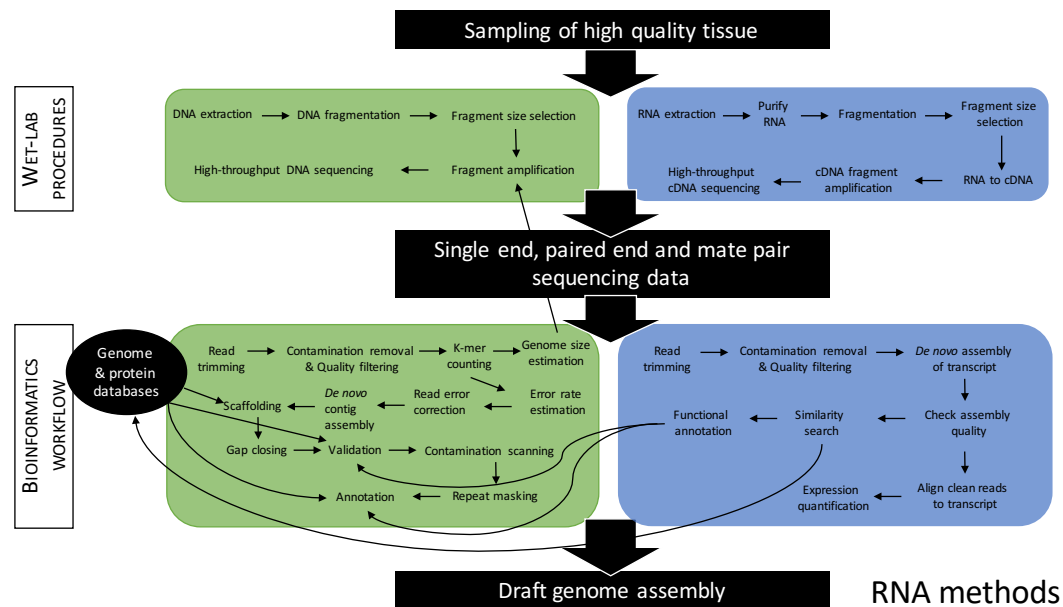
Independently on the approach used to obtain a whole-genome sequence, all sequencing projects share the same major steps:

## 1. Wet-lab procedures:

- DNA extraction,
- DNA library preparation,
- Sequencing.

## 2. Bioinformatics workflow:

- Genome assembly,
- Genome annotation.



RNA methods are presented in chapter 5

# OVERVIEW OF WET-LAB PROCEDURES

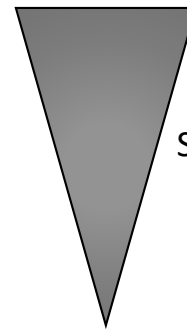
## 1. DNA extraction

## 2. DNA library preparation

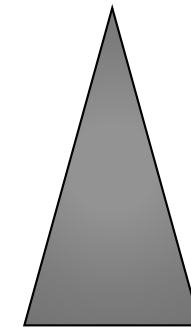
- a. Whole genome sequencing (WGS)
- b. Targeted sequencing
  - amplicon sequencing
  - hybrid capture
- c. RAD sequencing
- d. Multiplexing

## 3. Sequencing (technologies)

- a. Illumina
- b. PacBio
- c. Oxford Nanopore/Minion



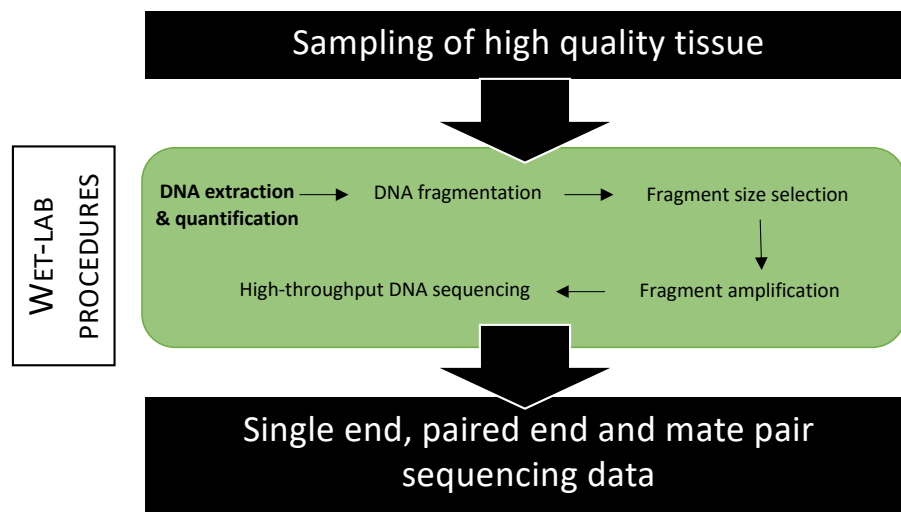
Sequencing  
depth



Number of  
sample(s) / run

# WET-LAB PROCEDURES – DNA EXTRACTION

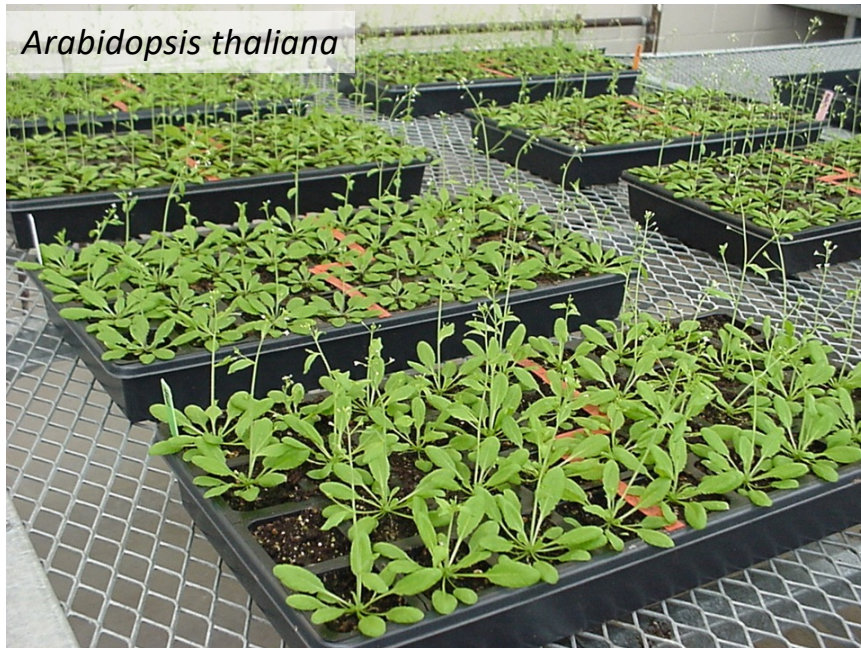
- The first crucial step for genome sequencing is the isolation and quality control (quantification) of DNA.
- This key step greatly influences the building of the DNA library (and which technology can be applied).



# WET-LAB PROCEDURES – DNA EXTRACTION

- Obtaining enough biomass for DNA extraction may sometimes be compromised by the very nature of the sample, especially with non-model and/or extinct organisms.

*Arabidopsis thaliana*



**VS.**

*Trochetiopsis* spp.

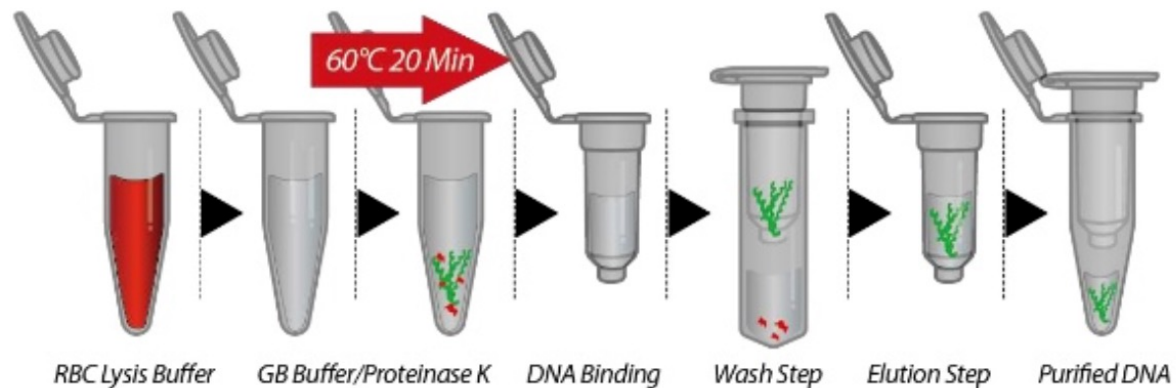




# WET-LAB PROCEDURES – DNA EXTRACTION

The DNA extractions should provide sufficient quantities of gDNA meeting the following criteria to ensure successful library prep.:

- ✓ Pure,
- ✓ Double-stranded,
- ✓ Highly-concentrated,
- ✓ Uncontaminated,
- ✓ Intact (especially important for long-read sequencing).



# WET-LAB PROCEDURES – DNA EXTRACTION

Current methods for quantifying DNA (before library preparation) are using a variety of techniques:

- ✓ UV absorption (e.g. Nanodrop)
- ✓ Intercalating dyes (e.g. QuBit) → widely used
- ✓ 5' hydrolysis probes (e.g. Taqman) coupled with qPCR
- ✓ Droplet digital emulsion PCRs (ddPCR; Bio-Rad)



## WET-LAB PROCEDURES – DNA EXTRACTION

Current methods for quantifying DNA (before library preparation) are using a variety of techniques:

- ✓ UV absorption (e.g. Nanodrop)
- ✓ Intercalating dyes (e.g. QuBit) → widely used
- ✓ 5' hydrolysis probes (e.g. Taqman) coupled with qPCR
- ✓ Droplet digital emulsion PCRs (ddPCR; Bio-Rad)



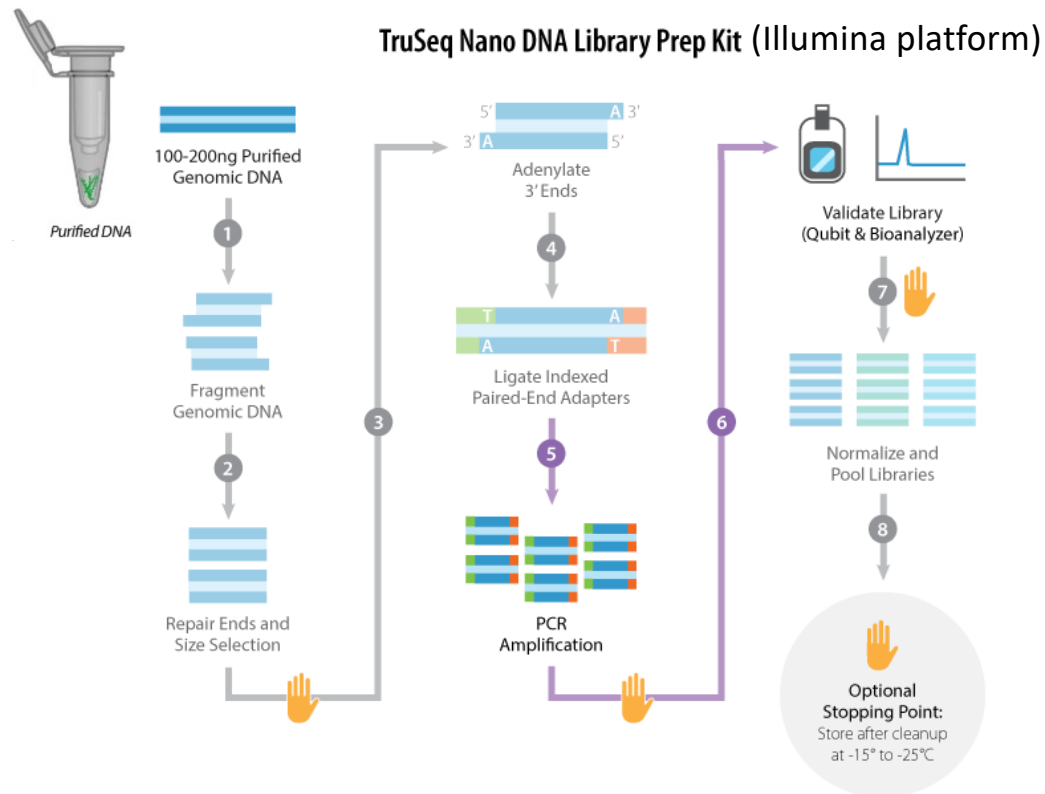
Purified DNA

For most of these methods, large amounts of DNA are required to provide accurate titrations!

Most sequencing platforms require micrograms of DNA for library prep., but most of it goes for quantification and **12 pM is used per Illumina lane** (1000-fold less than what is used for quantification)

# WET-LAB PROCEDURES — LIBRARY PREPARATION

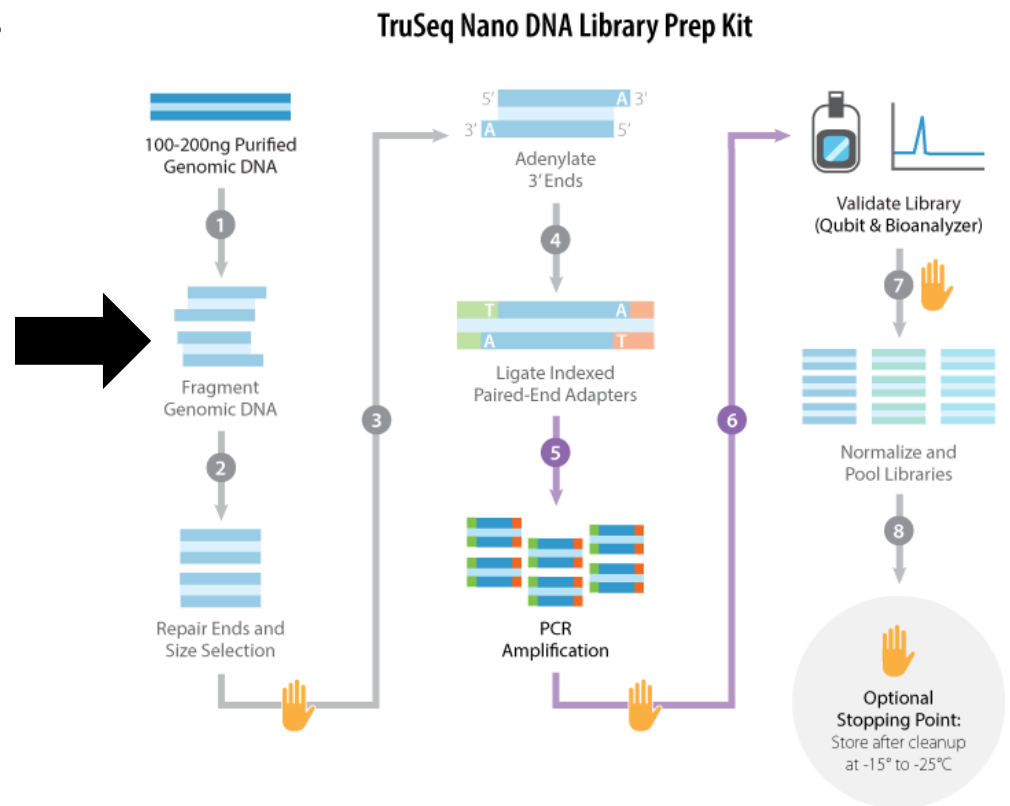
Overview of steps required to prepare a library for WGS (short-reads):



# WET-LAB PROCEDURES — LIBRARY PREPARATION

- The first step of a library prep. is DNA fragmentation, which can be done following three methods:

1. Nebulization,
2. Sonication,
3. Enzymatic DNA digestion.



## WET-LAB PROCEDURES — LIBRARY PREPARATION

- **Nebulization:** Compressed nitrogen forces gDNA repeatedly through a small hole producing random mechanically sheared fragments leading to a heterogeneous mix of double-stranded DNA molecules containing 3'- or 5' overhangs as well as blunt ends.



## WET-LAB PROCEDURES — LIBRARY PREPARATION

- **Sonication:** Samples are subjected to ultrasonic waves, whose vibrations produce gaseous cavitations in the liquid that shear or break high molecular weight DNA molecules through resonance vibration.



## WET-LAB PROCEDURES — LIBRARY PREPARATION

- **Sonication:** Samples are subjected to ultrasonic waves, whose vibrations produce gaseous cavitations in the liquid that shear or break high molecular weight DNA molecules through resonance vibration.

High molecular weight DNA corresponds to DNA fragments between 100-300kb





## WET-LAB PROCEDURES — LIBRARY PREPARATION

- **Enzymatic DNA digestion (dsDNA Fragmentase):** It might be an effective alternative to the random shearing methods. A commercial enzymatic fragmentation kit (NEBNext) has become available, which generates random DNA fragments between 100 and 800 bp length.

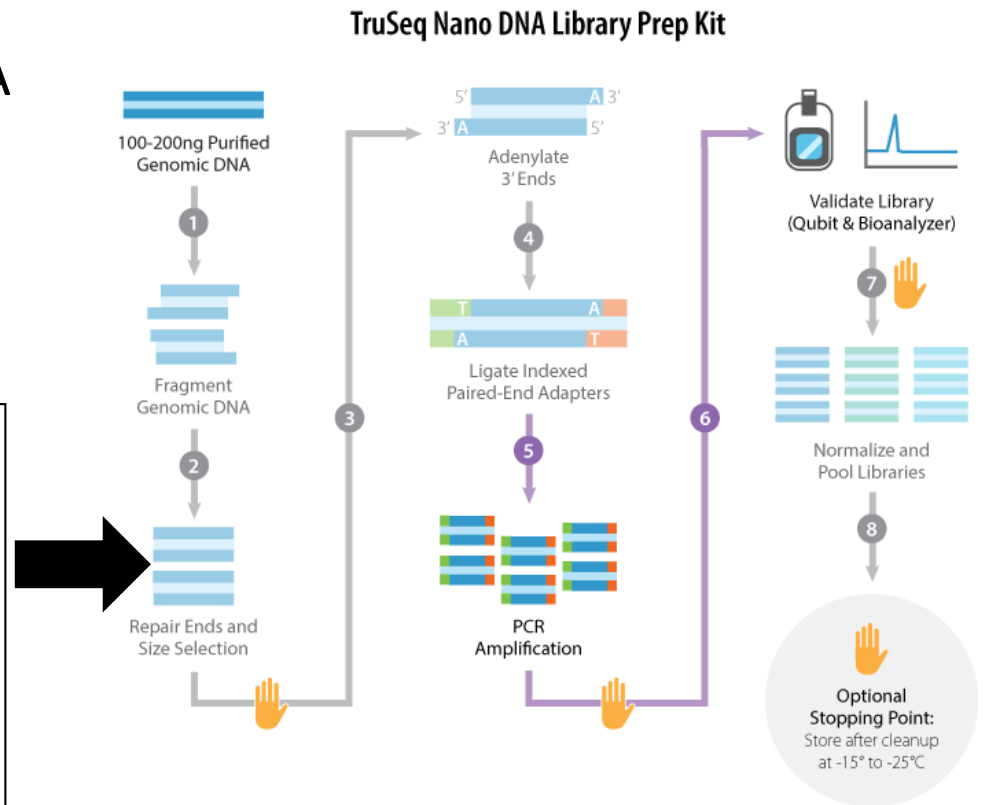


Ignatov *et al.* (2019)

# WET-LAB PROCEDURES — LIBRARY PREPARATION

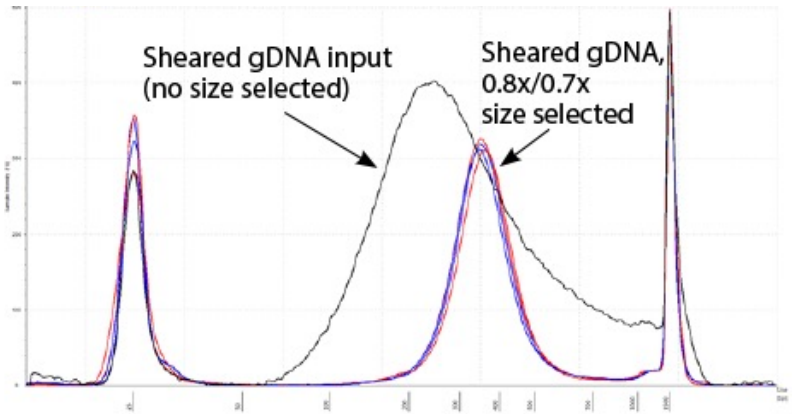
- **End-repair:** fragmentation produces double-stranded DNA with a mixture of blunt ends, recessed 3' ends and recessed 5' ends, with and without a 5' phosphate moiety.

→ Fragments must be made uniform before adapters can be ligated using a mixture of enzymes to generate blunt-ended fragments with phosphorylated 5' termini.

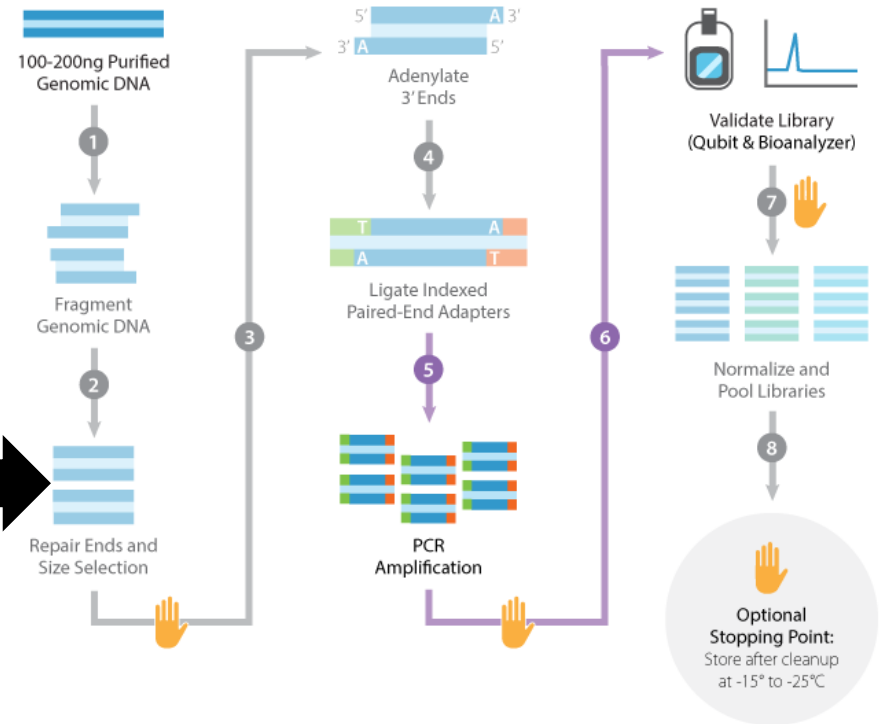


# WET-LAB PROCEDURES — LIBRARY PREPARATION

- **Size selection (optional):** e.g. AMPure XP has developed kits using beads to remove small sized fragments.

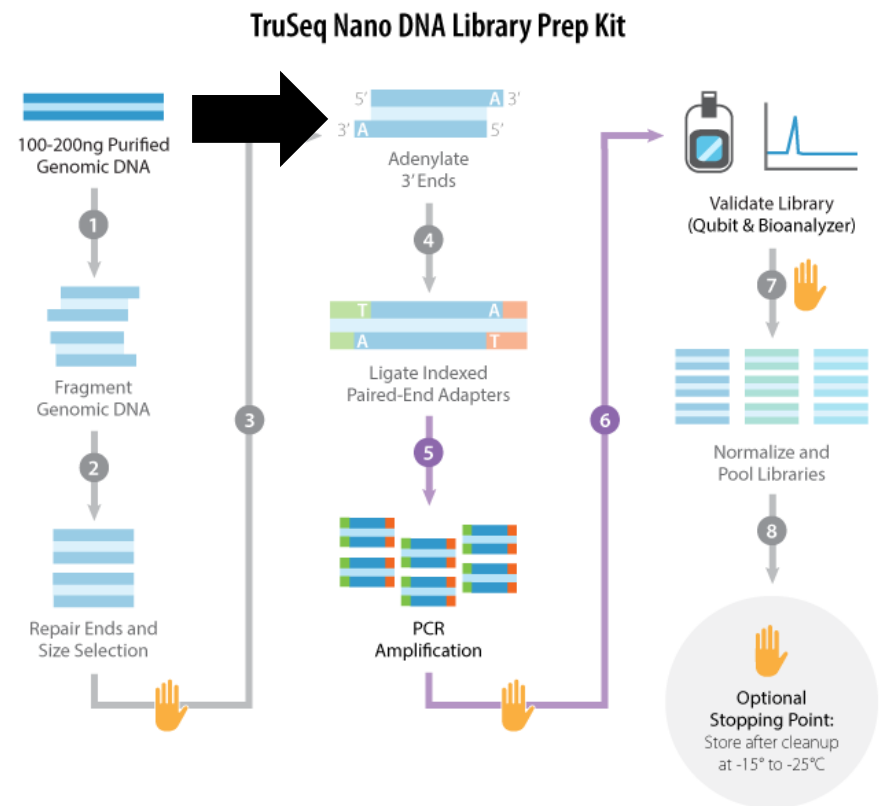


## TruSeq Nano DNA Library Prep Kit



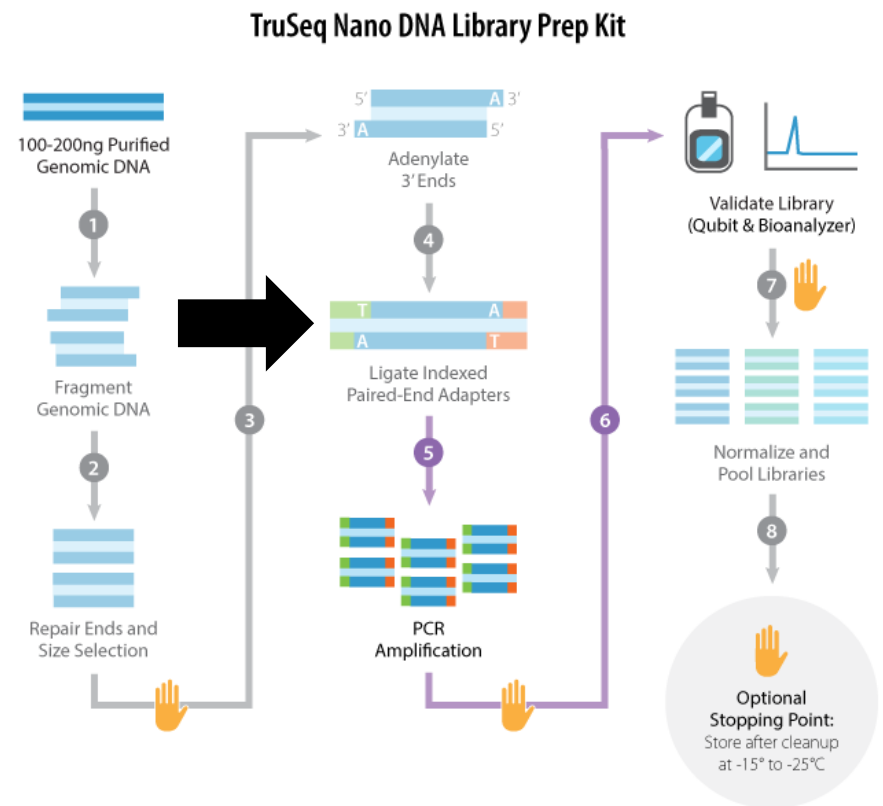
# WET-LAB PROCEDURES — LIBRARY PREPARATION

- **A-tailing:** Addition of a single A nucleotide to the 3' ends of fragments before adapter ligation deters concatemerization of templates and because the adapters are t-tailed, it increases the efficiency of adapter ligation.



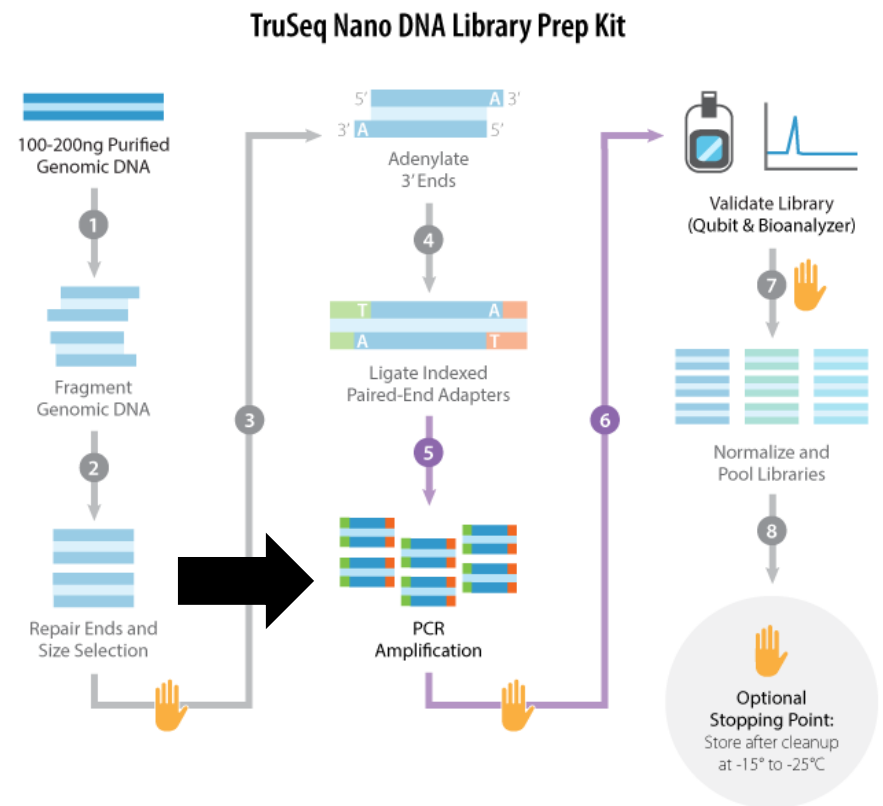
# WET-LAB PROCEDURES — LIBRARY PREPARATION

- **Adapter ligation:** Template strands must receive a different adapter sequence at either end to participate successfully in the cluster amplification and sequencing reactions.
- Will will shortly discuss three following sequencing strategies:
  1. Single-end,
  2. Paired-end,
  3. Mate Pair.



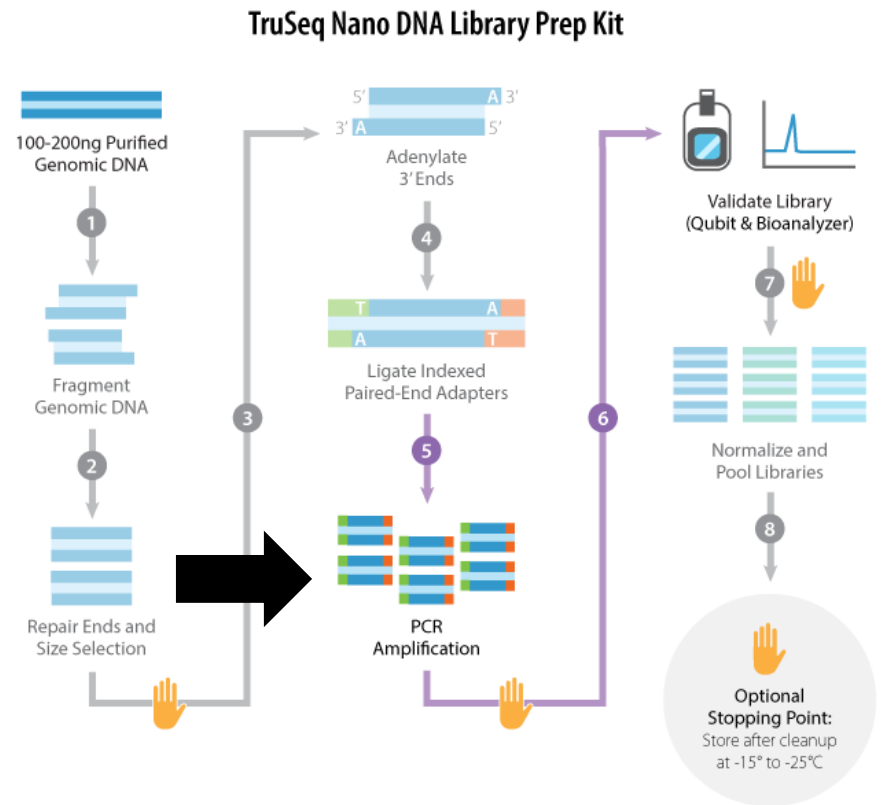
# WET-LAB PROCEDURES — LIBRARY PREPARATION

- **PCR (optional):** Libraries are amplified by PCR to
  - ✓ enrich for properly ligated template strands—those that have an adapter at both ends,
  - ✓ increase the amount of library available for sequencing
  - ✓ generate enough DNA for accurate quantification.



# WET-LAB PROCEDURES — LIBRARY PREPARATION

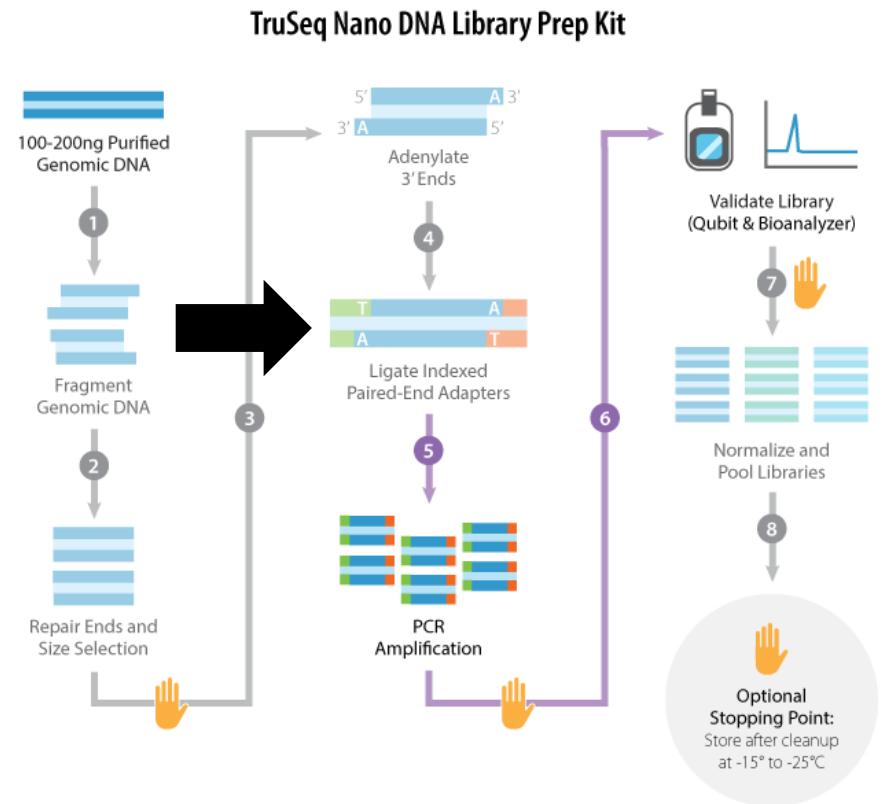
**To your knowledge, why adding a PCR step in the library preparation could negatively impact (bias) your genome assembly?**



# WET-LAB PROCEDURES — LIBRARY PREPARATION

## Sequencing strategies (decided at the adapter ligation step):

- **Single-end:** Technique in which sequence is reported from only one end of a fragment.
- **Paired-end:** Technique in which sequence is reported from both ends of a fragment (with a number of undetermined bases between the reads that is known only approximately).
- **Mate pair:** Technique in which sequence is reported from two ends of a DNA fragment, usually several thousand base-pairs long.

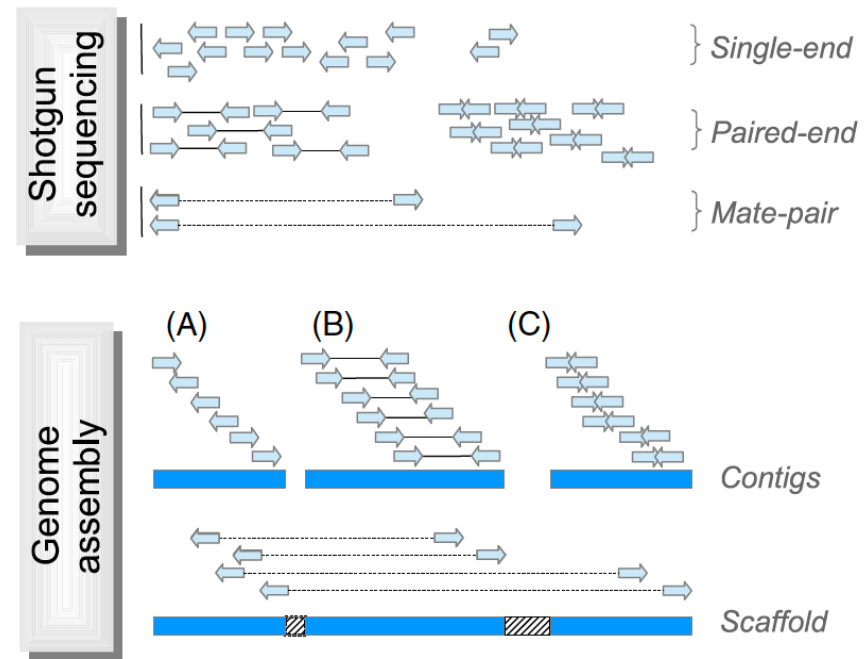




# WET-LAB PROCEDURES — LIBRARY PREPARATION

## Sequencing strategies (decided at the adapter ligation step):

- **Single-end:** Technique in which sequence is reported from only one end of a fragment.
- **Paired-end:** Technique in which sequence is reported from both ends of a fragment (with a number of undetermined bases between the reads that is known only approximately).
- **Mate pair:** Technique in which sequence is reported from two ends of a DNA fragment, usually several thousand base-pairs long.



The choice of library preparation will impact *de novo* reconstruction

## WET-LAB PROCEDURES – LIBRARY PREPARATION

- In single-end sequencing the fragment is read only from one end.



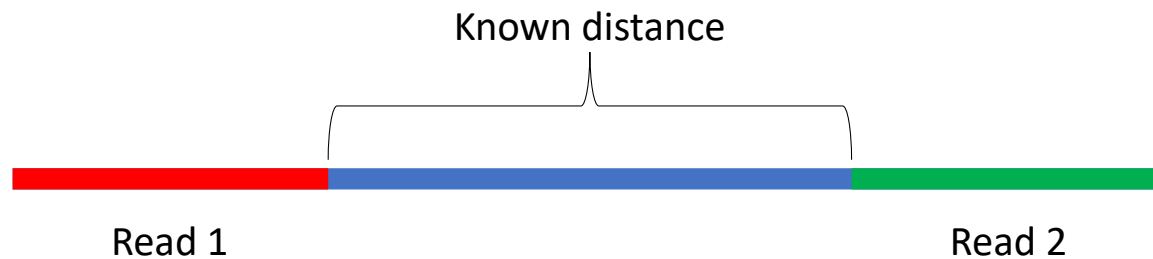
## WET-LAB PROCEDURES – LIBRARY PREPARATION

- In paired-end sequencing when the sequencer finishes one direction at the specified read length (i.e. 150 bp), it starts another round from the opposite end.



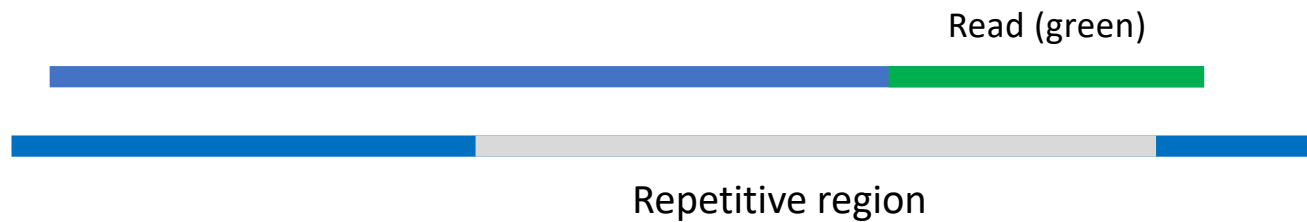
# WET-LAB PROCEDURES — LIBRARY PREPARATION

- Knowing the distance between paired reads helps improving their specificity and *de novo* genome assembly



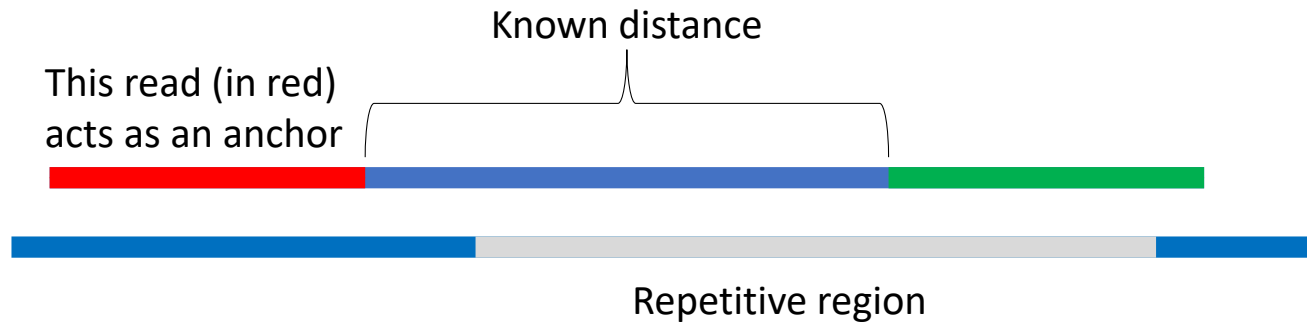
# WET-LAB PROCEDURES — LIBRARY PREPARATION

- Knowing this distance is also key to map reads over repetitive regions more precisely.
  - ✓ A single-end read (in green) might be unmappable because it falls in a very repetitive region.



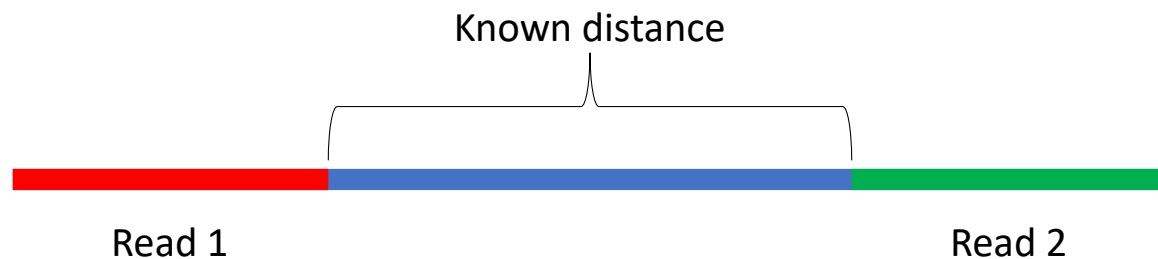
# WET-LAB PROCEDURES – LIBRARY PREPARATION

- Knowing this distance is also key to map reads over repetitive regions more precisely.
  - ✓ But, with paired-end reads (green and red), if one of the ends is unique (= it matches without ambiguities), you can use the distance information to map both ends and reconstruct the DNA sequence!



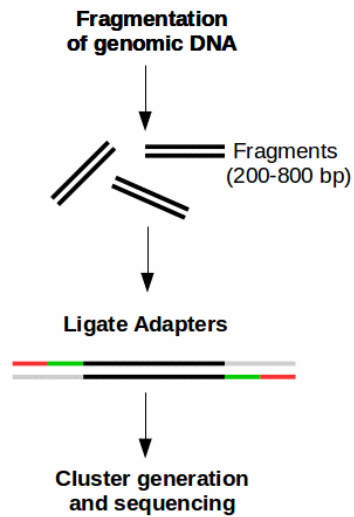
# WET-LAB PROCEDURES — LIBRARY PREPARATION

- To improve genome assembly, we want to minimize the “distance” between reads.
- Two approaches:
  - Paired-end sequencing: missing “distance” is minimum between reads (ca. 200-800 bp).
  - Mate pair sequencing: when the distance is larger than the fragment size.

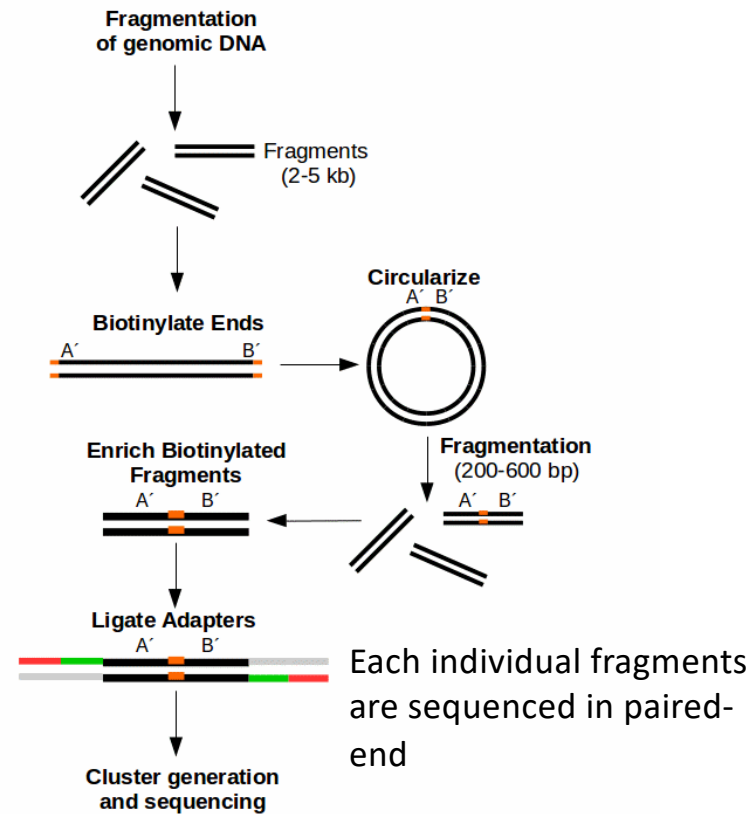


# PAIRED-END VS MATE PAIR SEQUENCING

## Paired-End Sequencing (Short-insert paired-end reads)



## Mate Pair Sequencing



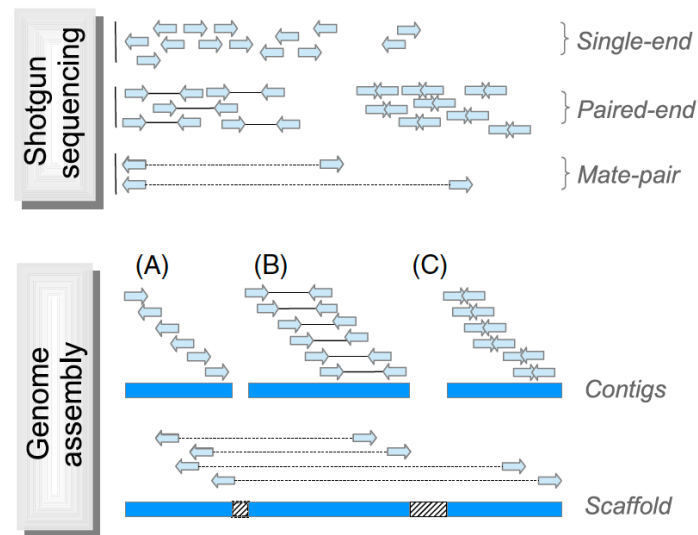


## WET-LAB PROCEDURES — LIBRARY PREPARATION

- Mate pair sequencing involves generating long-insert paired-end DNA libraries useful for a number of sequencing applications, including:
  - *De novo* sequencing,
  - Genome finishing,
  - Structural variant detection,
  - Identification of complex genomic rearrangements.

# WET-LAB PROCEDURES – LIBRARY PREPARATION

- A successful *de novo* sequencing project has to combine data from short-insert and long-insert sizes libraries therefore providing maximal sequencing coverage across the genome → Approach most applied for whole genome sequencing (e.g. Giant panda project).

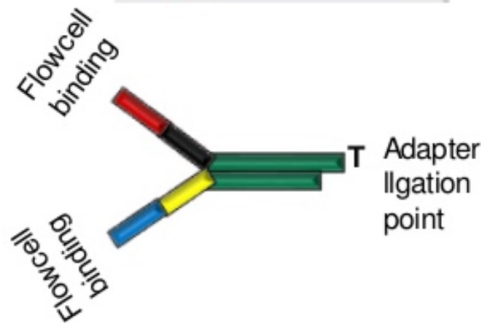
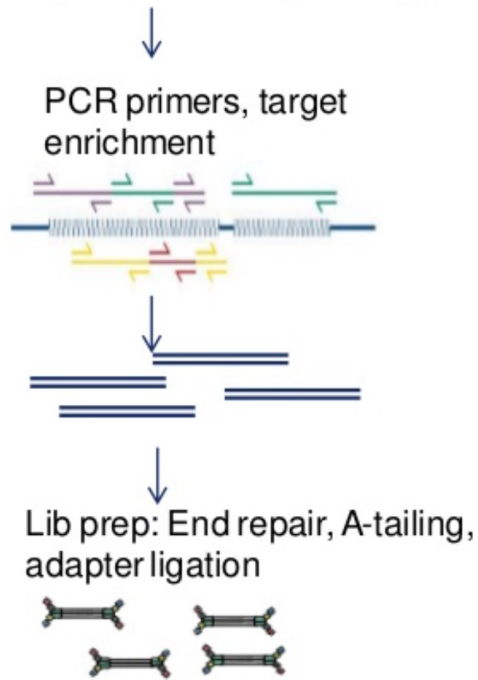


# Targeted Sequencing

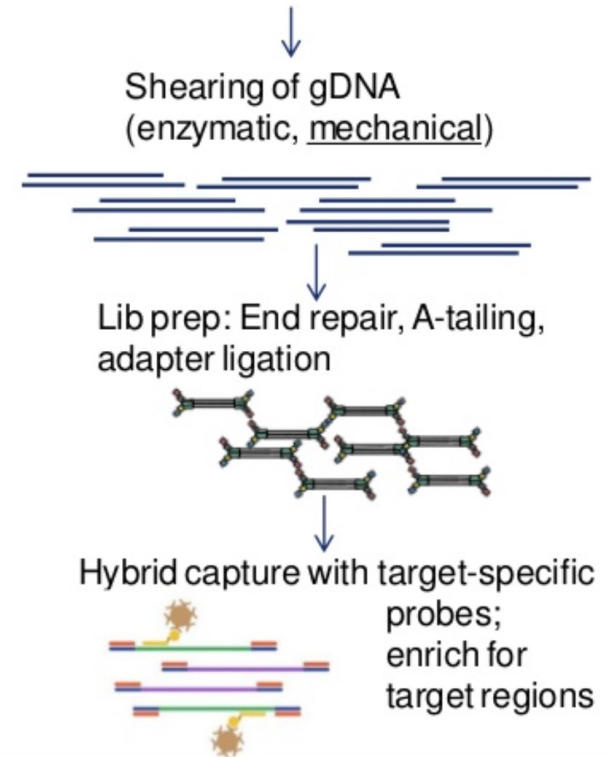


Extracted DNA

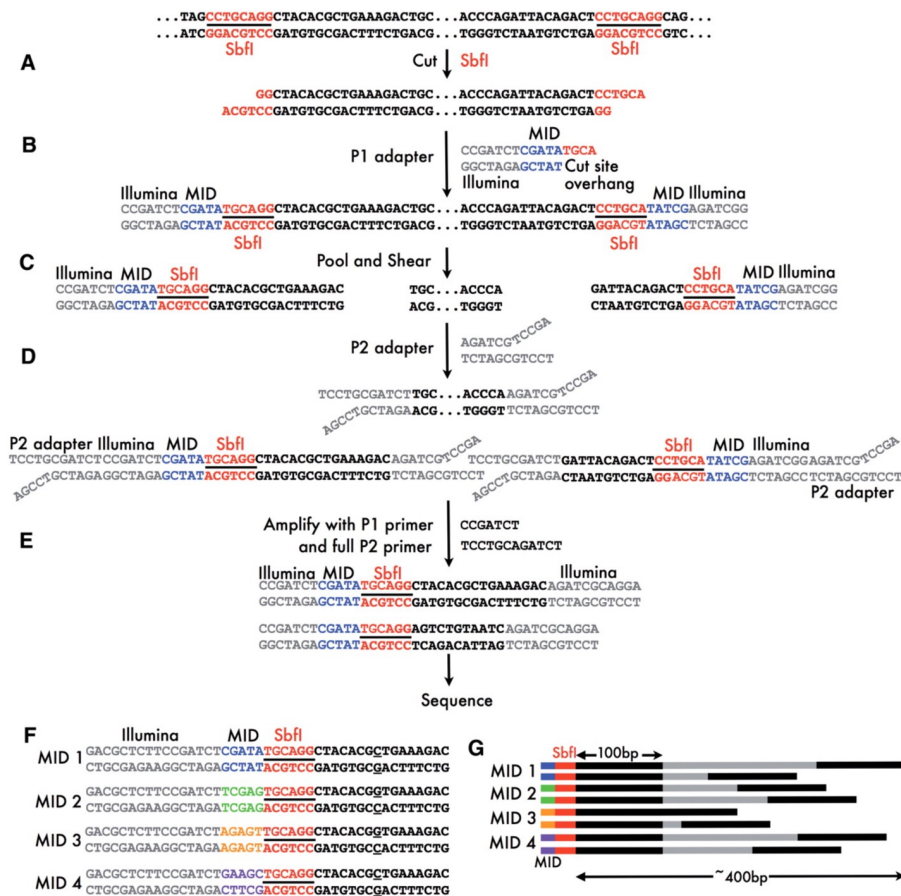
## PCR-based target enrichment (amplicon sequencing)



## Hybrid capture target enrichment (whole exome, targeted regions)

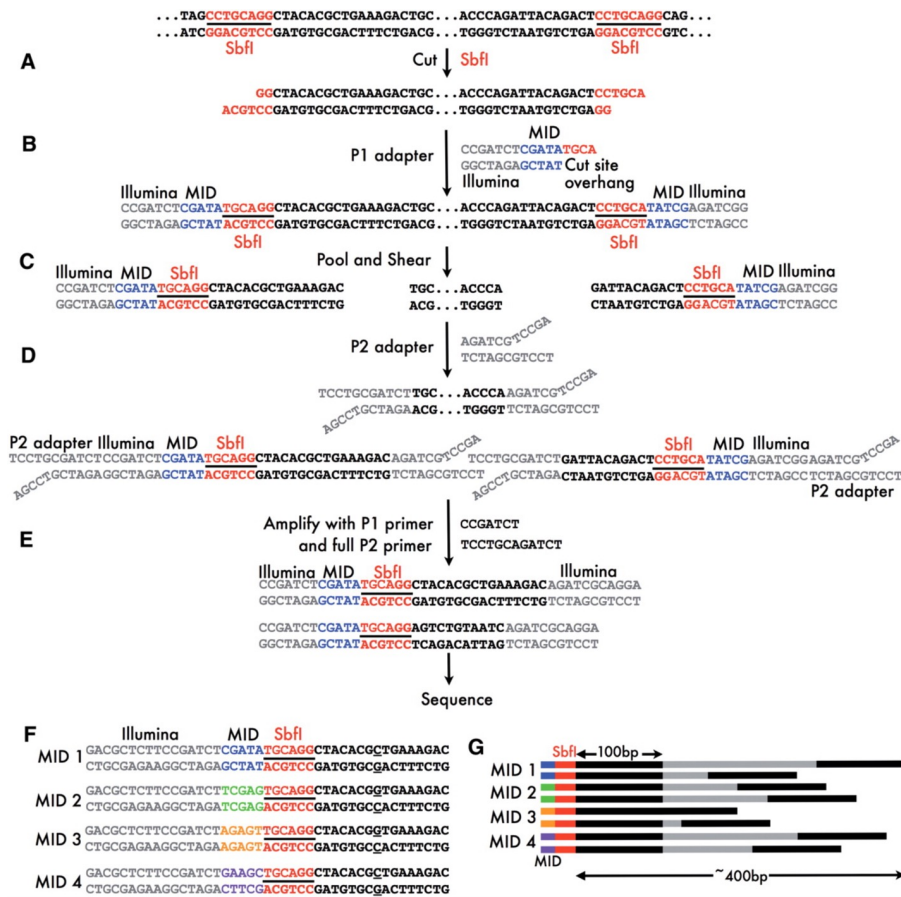


# WET-LAB PROCEDURES – RAD SEQUENCING



A method developed to identify single nucleotide polymorphism (SNP) sites using sequenced restriction-site associated DNA (RAD) markers

# WET-LAB PROCEDURES – RAD SEQUENCING



**A.** Genomic DNA is digested with a restriction enzyme

**B.** P1 adapter are ligated to the fragments

**C.** Samples from multiple individuals are pooled and all fragments are randomly sheared to ca. 500 bp

**D.** P2 adapter are ligated to the fragments

**E.** PCR amplification with P1 and P2 primers

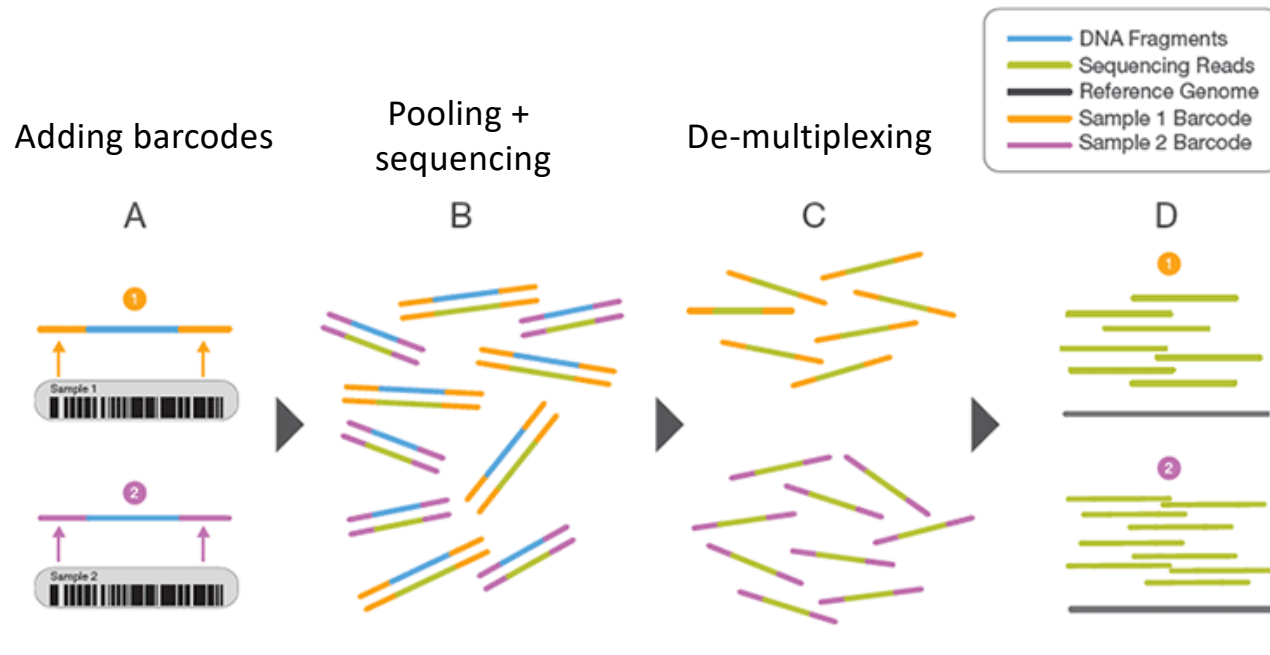
**F.** Pool samples to be sequenced

**G.** Paired-end sequencing covering fragments between 300-400 bp → identify SNPs

## WET-LAB PROCEDURES – MULTIPLEXING

- This process allows analyzing/pooling a large number of samples simultaneously on a high-throughput instrument.
- Sample multiplexing is a useful technique when targeting specific genomic regions or working with smaller genomes.

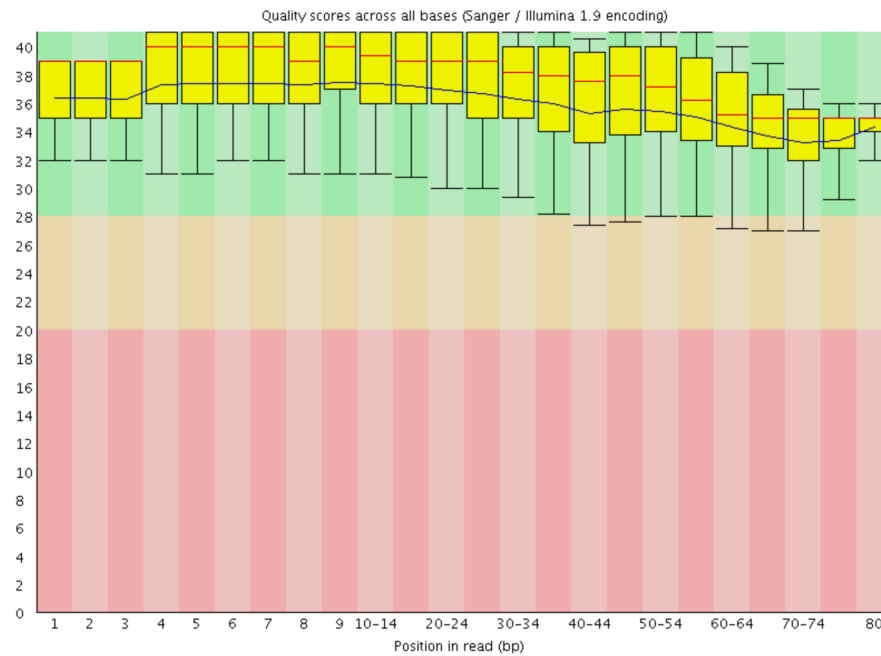
# WET-LAB PROCEDURES – MULTIPLEXING



- Two representative DNA fragments from two unique samples, each attached to a specific barcode sequence that identifies the sample from which it originated.
- Libraries for each sample are pooled and sequenced in parallel. Each new read contains both the fragment sequence and its sample-identifying barcode.
- Barcode sequences are used to de-multiplex, or differentiate reads from each sample.
- Each set of reads is aligned to the reference sequence.

# NGS OUTPUT DATA — FASTQ

- \*.fastq (sequence and corresponding quality score encoded with an ASCII character, phred- like quality score + 33)





## NGS OUTPUT DATA — FASTQ

- A fastq file normally uses four lines per sequence as follows:
  1. **Sequence identifier:** Begins with a '@' and is followed by an optional description.
  2. **Raw sequence letters.**
  3. **Optional line:** Begins with a '+' and is optionally followed by the same sequence identifier (and any description) again.
  4. **Quality scores (Phred scores):** Encodes the quality values for the sequence in line 2. The character '!' represents the lowest quality while '~' is the highest.

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!' '*((( (**+))%%%++)(%%%) .1***-+*'')**55CCF>>>>>>CCCCCCC65
```

## NGS OUTPUT DATA – PHRED QUALITY SCORE

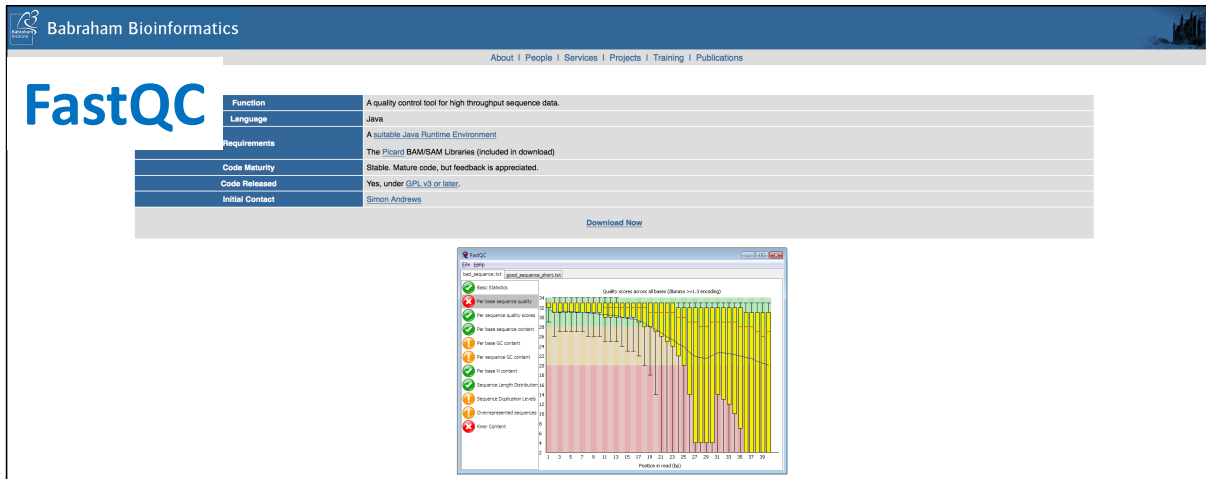
- **Phred quality score:** Measure of the quality of the identification of the nucleobases generated by automated DNA sequencing.
- Developed to support the Human Genome Project.
- Phred quality scores  $Q$  are defined as a property, which is logarithmically related to the base-calling error probabilities ( $P$ )
- $Q = -10 \log_{10} P$

Phred quality score	Probability that the base is called wrong	Accuracy of the base call
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1,000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%

Min. Phred score ( $Q$ ) of 30

→ Trim bases <30

# NGS OUTPUT DATA – SOFTWARE FOR QCs



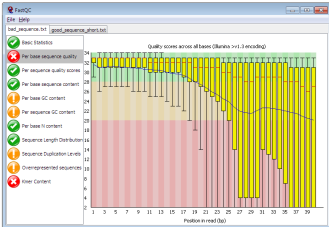
Babraham Bioinformatics

About | People | Services | Projects | Training | Publications

## FastQC

Function	A quality control tool for high throughput sequence data.
Language	Java
Requirements	A suitable Java Runtime Environment
Code Maturity	The <a href="#">Picked BAW/SAM Libraries</a> (included in download)
Code Released	Stable. Mature code, but feedback is appreciated.
Initial Contact	Yes, under <a href="#">GPL v3 or later</a> .
	<a href="#">Simon Andrews</a>

[Download Now](#)



## Trimmomatic: A flexible read trimming tool for Illumina NGS data

### Citations

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina Sequence Data. *Bioinformatics*, btu170.