



RRBS-SEQ
ICE
OS-SEQ
MAINE-SEQ
RIBO-SEQ
CLASH-SEQ
TIF-SEQ/PEAT
IN-SEQ
TAB-SEQ
5-C
SMM
OXBS-
RIP-SEQ

CHIP-SEQ

WORKSHOP 16: LIBRARY PREP FOR NGS

pt1

Marco Morselli

Pellegrini Lab

QCBio Collaboratory Fellow

April 14th, 2020

TODAY (day 1)

- **INTRO to SEQUENCING**
- **LIBRARY PREP 101**

Specialty DNA Library Preps

TOMORROW (day 2)

- **RNA Library Preps**
 - RNAseq
 - small RNA-seq
 - Low-input RNAseq

THURSDAY: ONLINE LAB Experiments (day 3)

(Use the TapeStation for QC; Library Preparation; submit libraries for sequencing)

POLL 1



**1. INTRO TO SEQUENCING
TECHNOLOGIES**



#1

2. LIBRARY PREP 101



#2

3. LIBRARY PREP 102



#3

4. SEQUENCING 101



#4

**5. TARGETED SEQUENCING &
PCR duplicates**



#5



It depends
on the 



FINAL



#6

CHECK FOR QUESTIONS



**MY
ANSWER**

- 1. NO QUESTIONS IN THE CHAT, PLEASE!!!!**
- 2. USE THE CHAT ONLY TO REPORT TECH PROBLEMS...like audio/video**
- 3. MAKE SURE YOUR MIC IS MUTED, PLEASE!!!!**

1. POST QUESTIONS in GOOGLE DRIVE. I will periodically answer to them



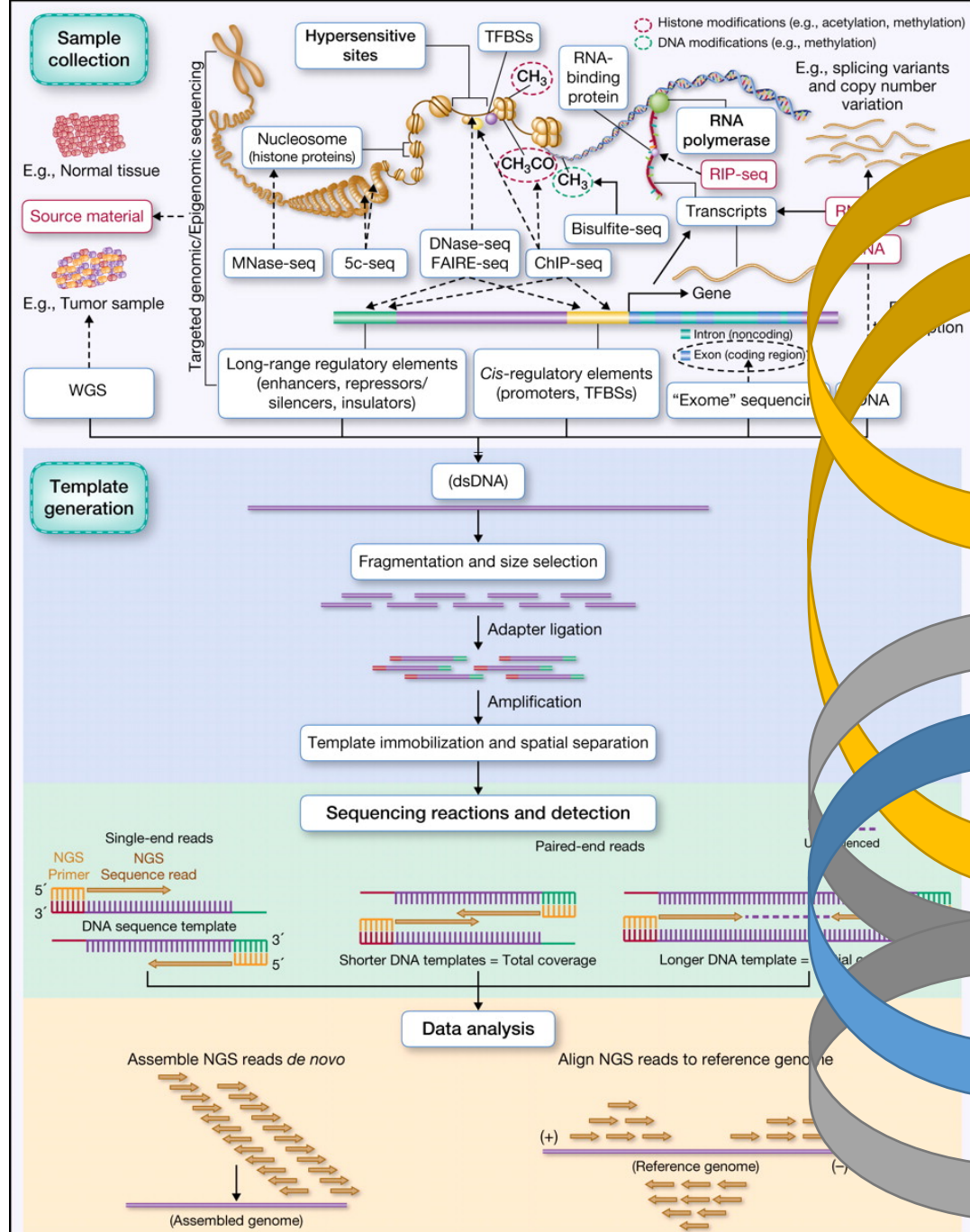
2. POLLS during the class

3. This class will be recorded. There are also videos available online for the past session.
(W16 – Collaboratory webpage)



SUPER-BRIEF

INTRO TO SEQUENCING TECHNOLOGIES



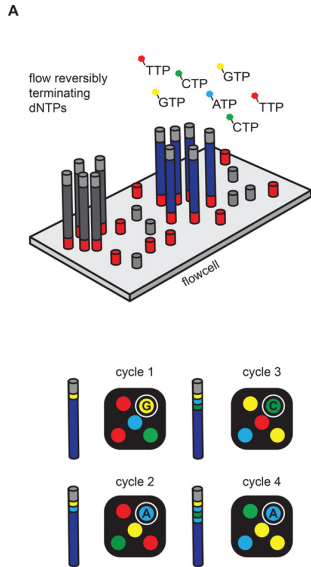
Starting Material

Library Preparation

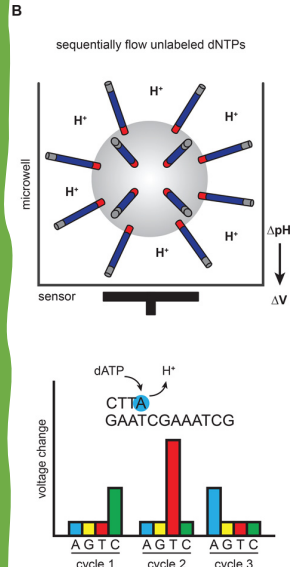
Sequencing

Analysis

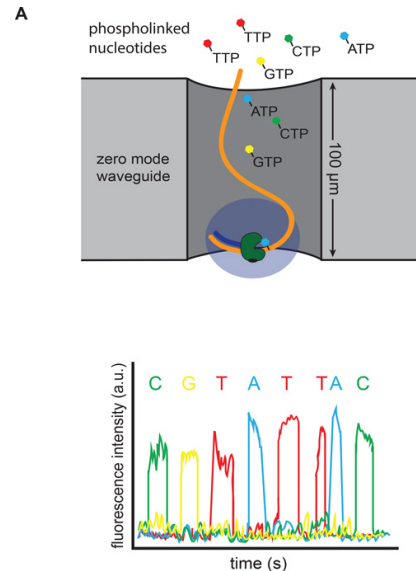
Illumina



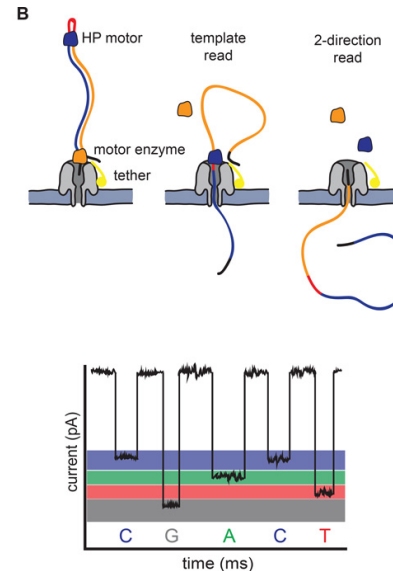
IonTorrent



Pacific Biosciences



Oxford NanoPore



Third-Generation Sequencing:

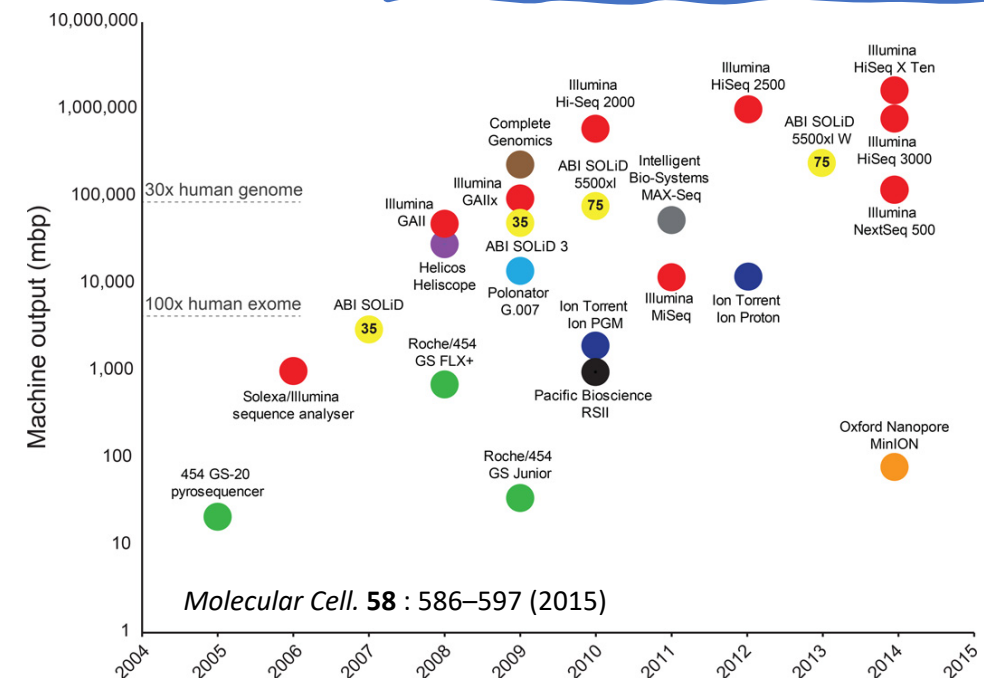
- Single-molecule sequencing
- Real time
- Longer Reads
- Higher Error-rate (but..)
- Lower number of Reads
- Read native modifications!!!

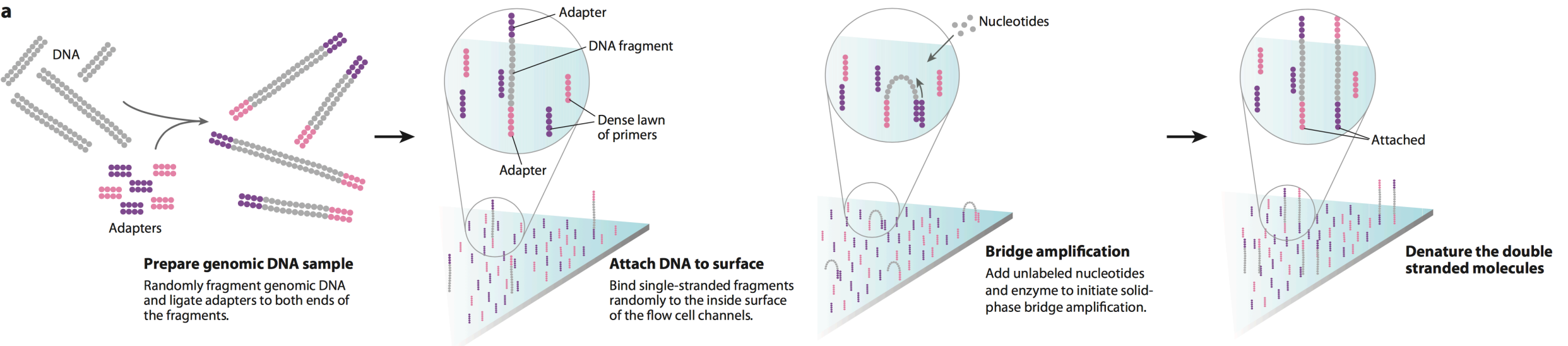
First-Generation Sequencing:

Maxam-Gilbert & Sanger

Next-Generation Sequencing or Second-Generation Sequencing:

- Amplification is required before Sequencing
- Sequencing by Synthesis
- Short Reads
- Millions of Reads
- Low Error-rate
- CANNOT read native modifications!!!

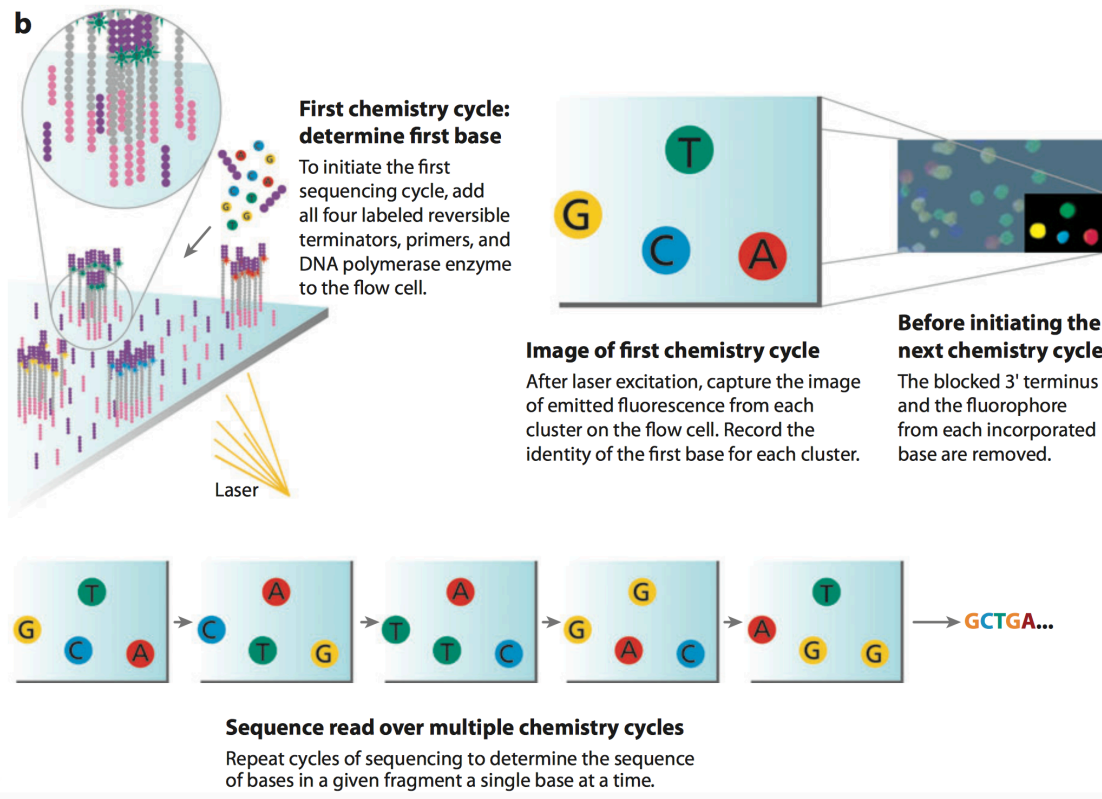




Annual Review of Genomics and Human Genetics, 2008
Molecular Cell. **58** : 586–597 (2015)

Next-Generation Sequencing or Second-Generation Sequencing:

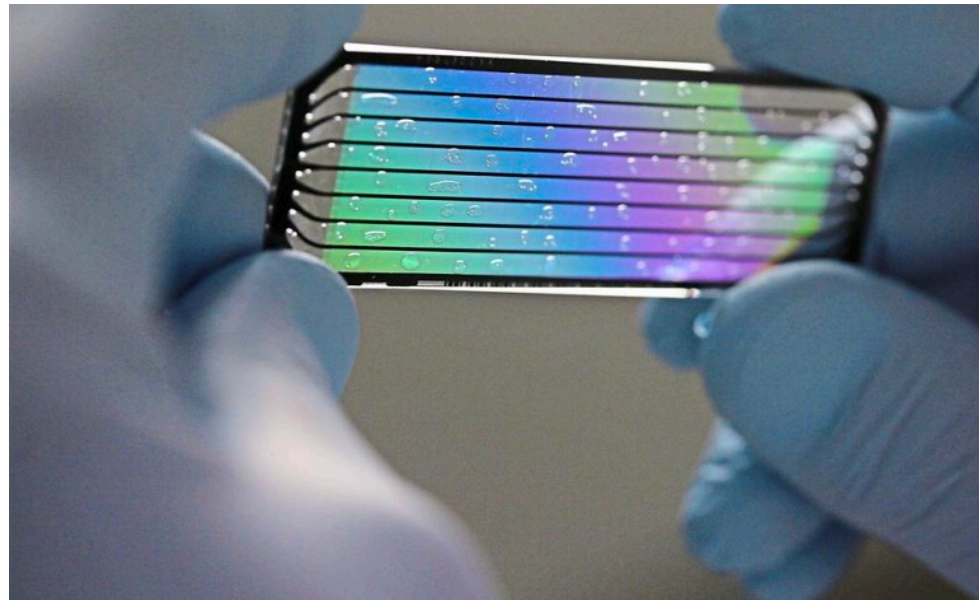
- Amplification is required before Sequencing
- Sequencing by Synthesis
- Short Reads
- Millions of Reads
- Low Error-rate
- CANNOT read native modifications!!!





Flow-cell

Patterned
Flow-cell



8 Lanes

e.g. **HiSeq3000/4000/X** Flow cell = 8 lanes → each lane has $\approx 480\text{M}$ cells



**KEEP
CALM
IT'S
BREAK
TIME**



QUESTIONS
#1

LIBRARY PREP 101

1. DNA

2. short (average 200-800 bp)

3. modified with adapters*



1. DNA

2. short (average 200-800 bp)

3. modified with adapters*

CONCENTRATION
& QC

POLL 3



CONCENTRATION
& QC



NanoDrop Spectrophotometer

Pro: 260/280 and 260/230 ratio

Cons: Inaccurate, especially if DNA is not clean and low concentration.

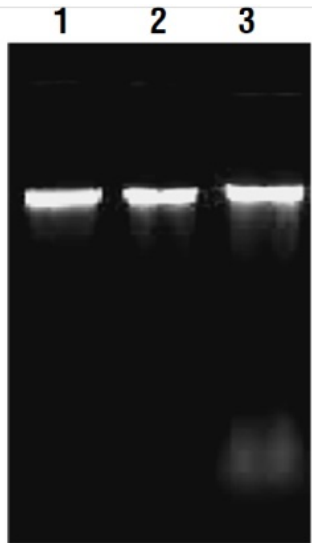
IT'S OK FOR RNA

Substance	Absorbance (nm)	Purity indicator
Nucleic acids	260 nm	A260:280 > 1.8 DNA
Proteins	280 nm	A260:280 > 2.0 RNA
Organics	230 nm	A260:230 > 2.0 DNA/RNA

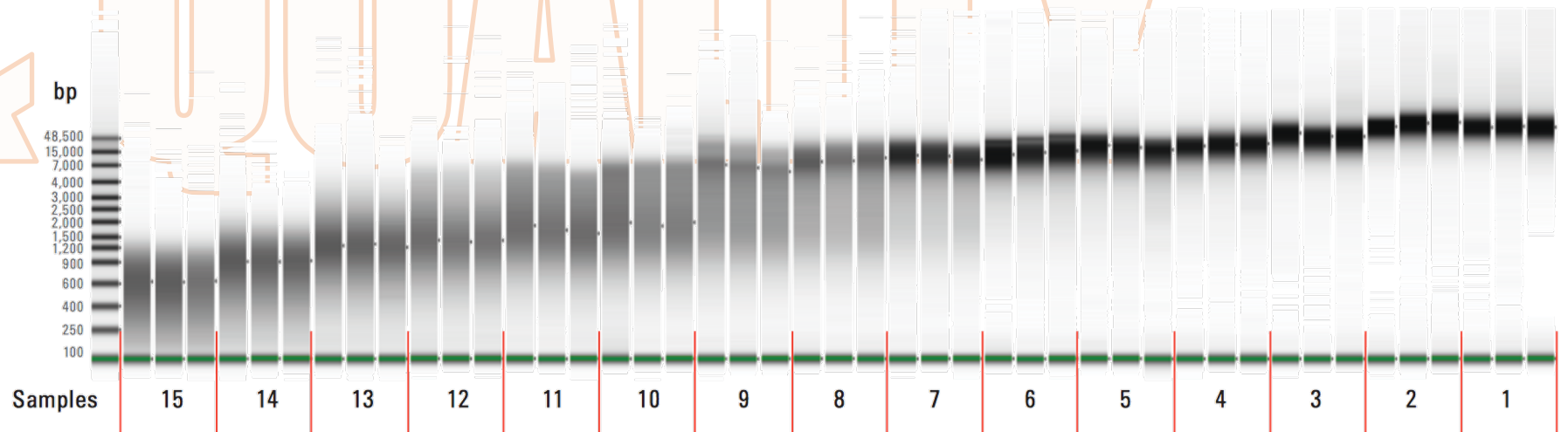
Qubit Fluorometer

Pro: more accurate and specific + low concentration

Cons: does not detect contaminants

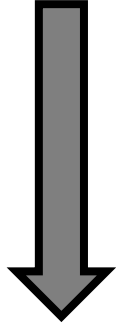


Agarose Gel

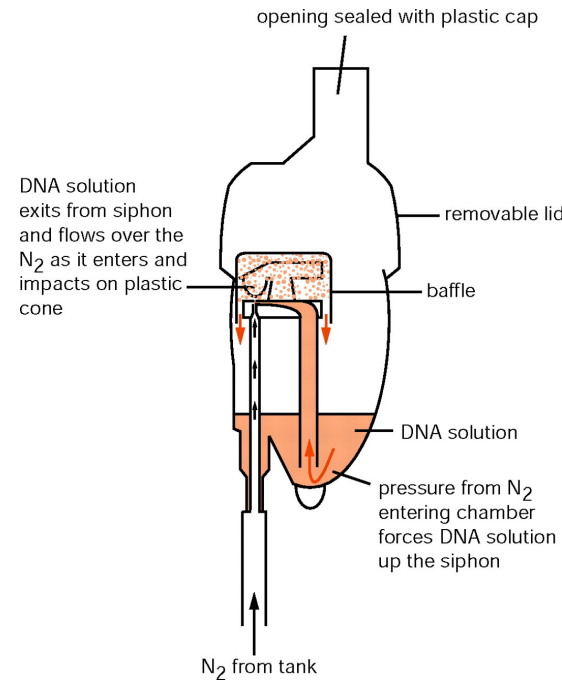


Agilent TapeStation Genomic DNA - DIN

Purified gDNA



Fragmentation



1.

1. DNA

2. short (average 200-800 bp)

3. modified with adapters

1. Nebulization

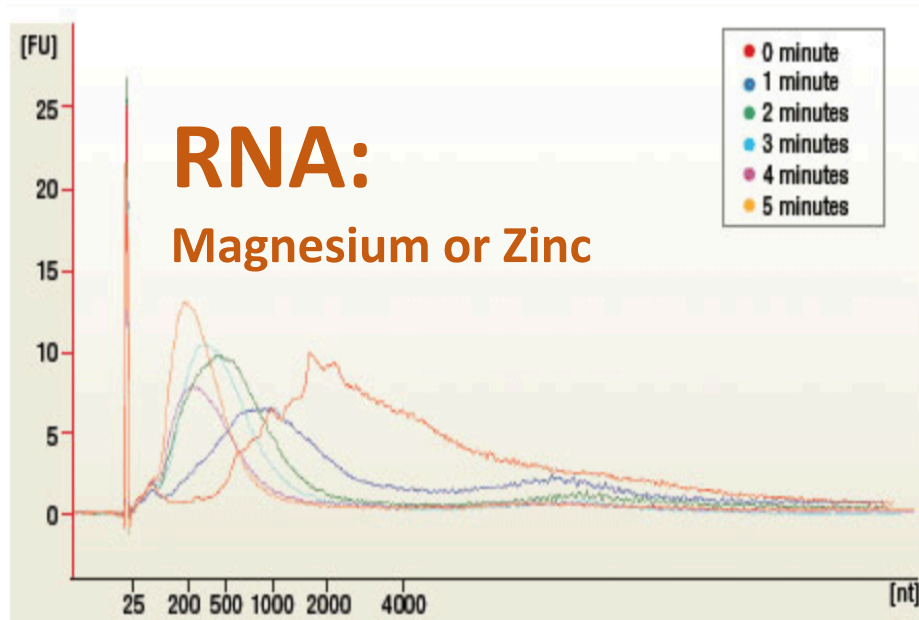
2. Sonication

3. Enzymatic

4. Chemical

LAB*

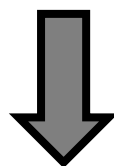
4.





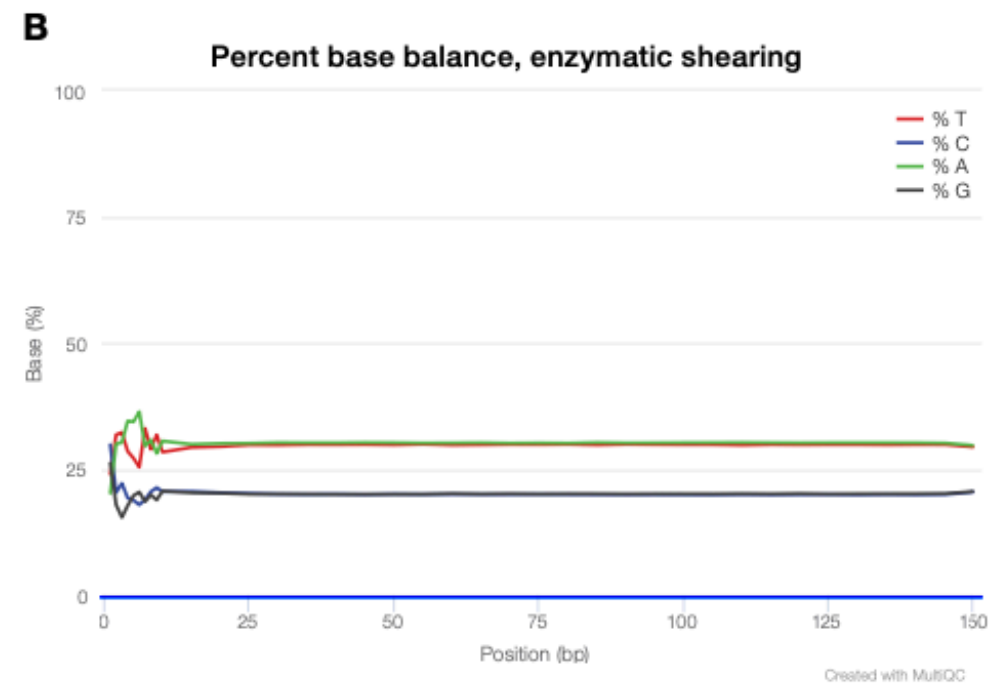
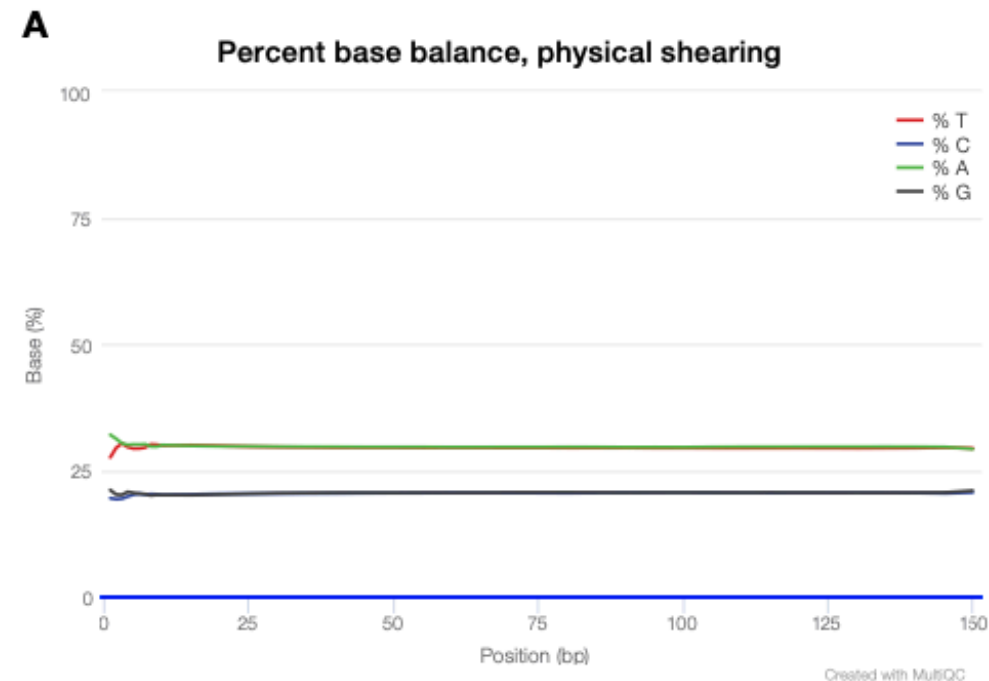
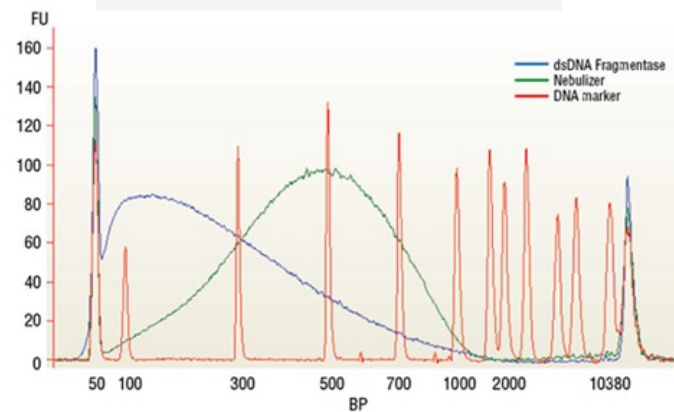
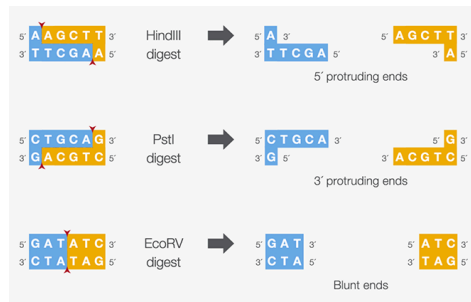
2. Sonication

Purified gDNA

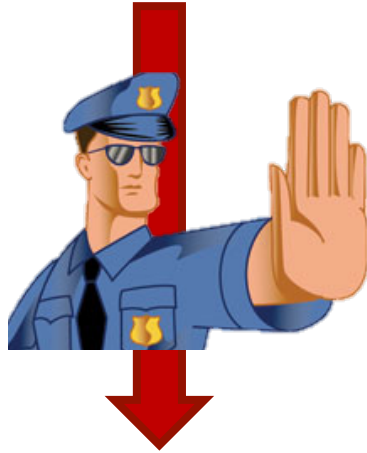


Fragmentation

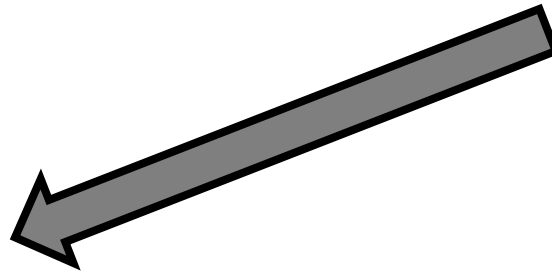
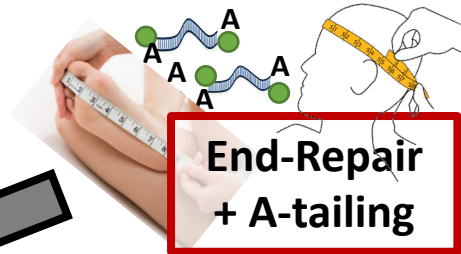
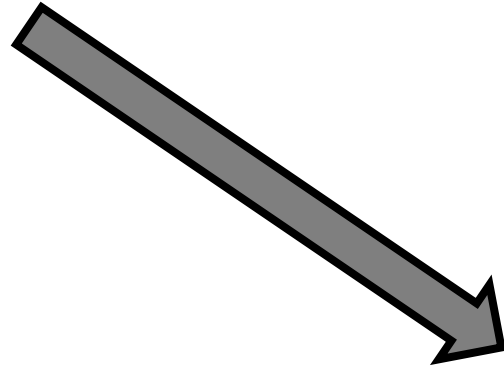
3. Enzymatic



Fragmented DNA



Adapter
Ligation



1. DNA

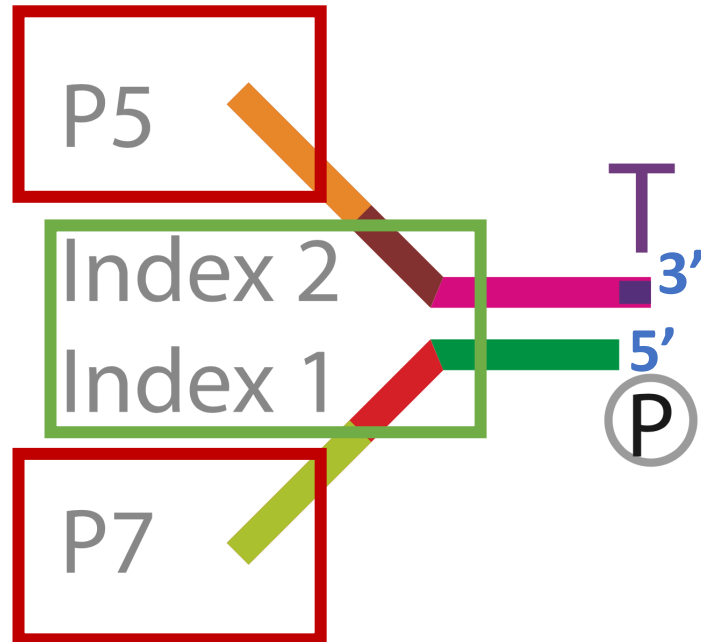
2. short (average 200-800 bp)

3. modified with adapters

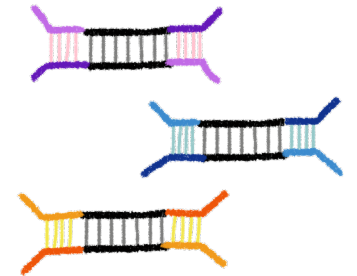
How do we perform
Adapter Ligation?
...let's understand how
Adapters are made...

Illumina Adapters

1. Sample multiplexing (it's not the only way)

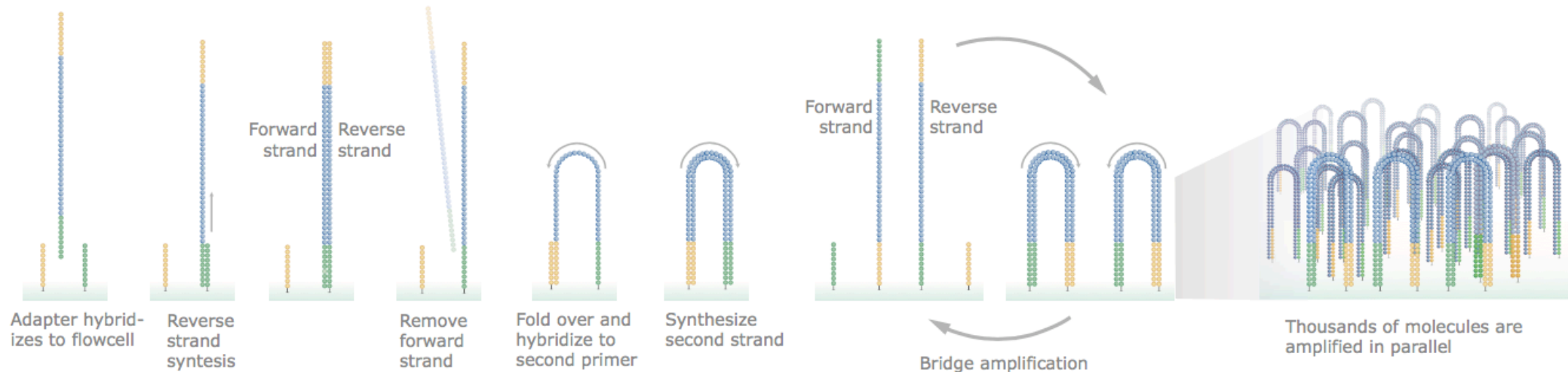
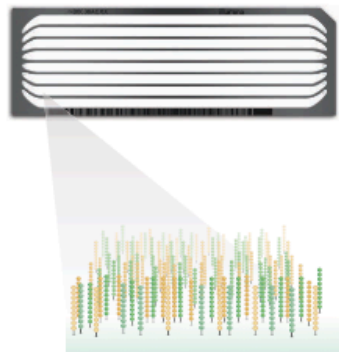
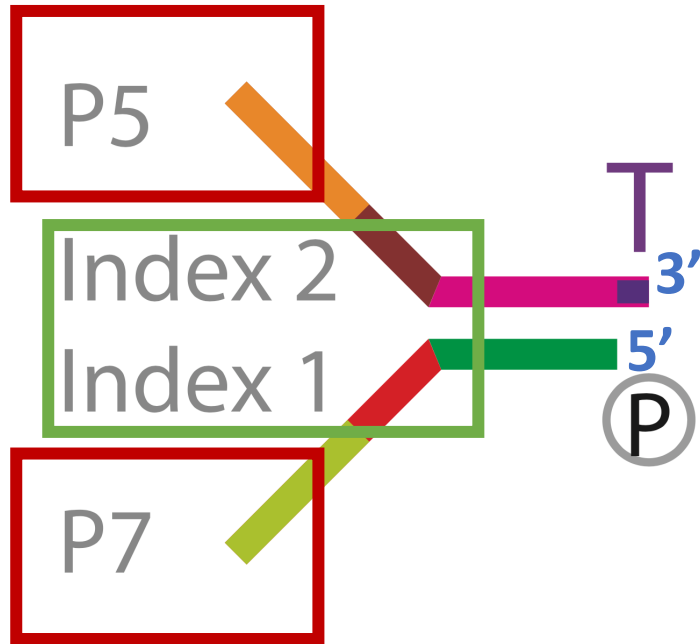


Sample	Index 1 (i7)	Index 2 (i5)
Sample 1	ATATGCGC	CTGATCGT
Sample 2	TGGTACAG	ACTCTCGA
Sample 3	AACCGTTC	TGAGCTAG
...	...	...



The combination
of Index 1 and
Index 2 is different
for each sample

Illumina Adapters



1. **Sample multiplexing**
(it's not the only way)
2. **Amplification Regions:**
Flowcell binding+ PCR Primer

Illumina Adapters



1. **Sample multiplexing**
(it's not the only way)

2. **Amplification Regions:**
Flowcell binding+ PCR Primer

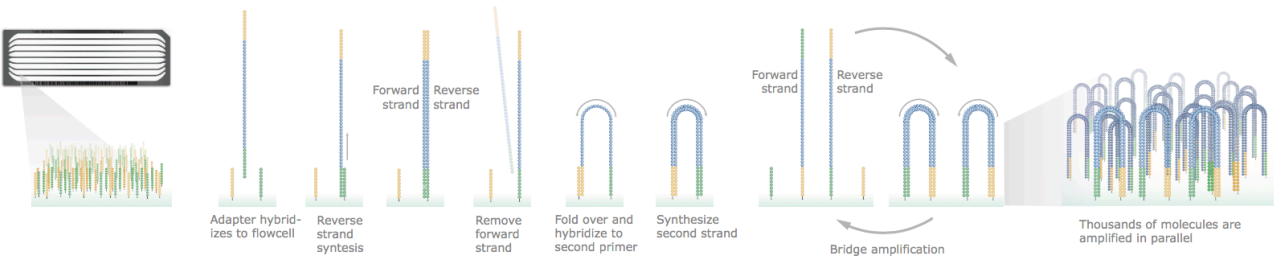
3. **Y-shaped**

Why?

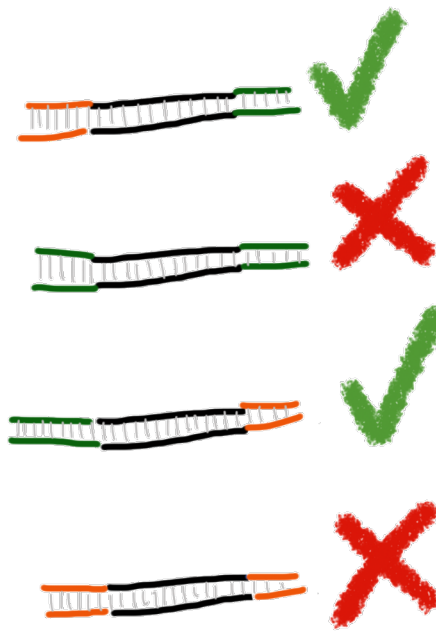
Illumina Adapters



3. Y-shaped

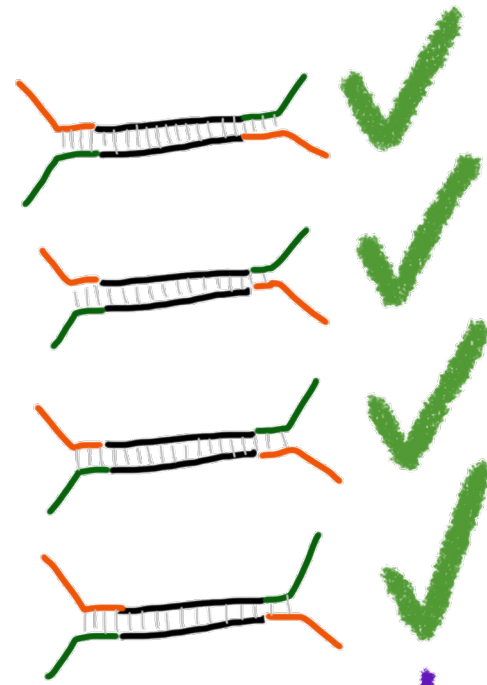


P5 P7



linear dsDNA

50%

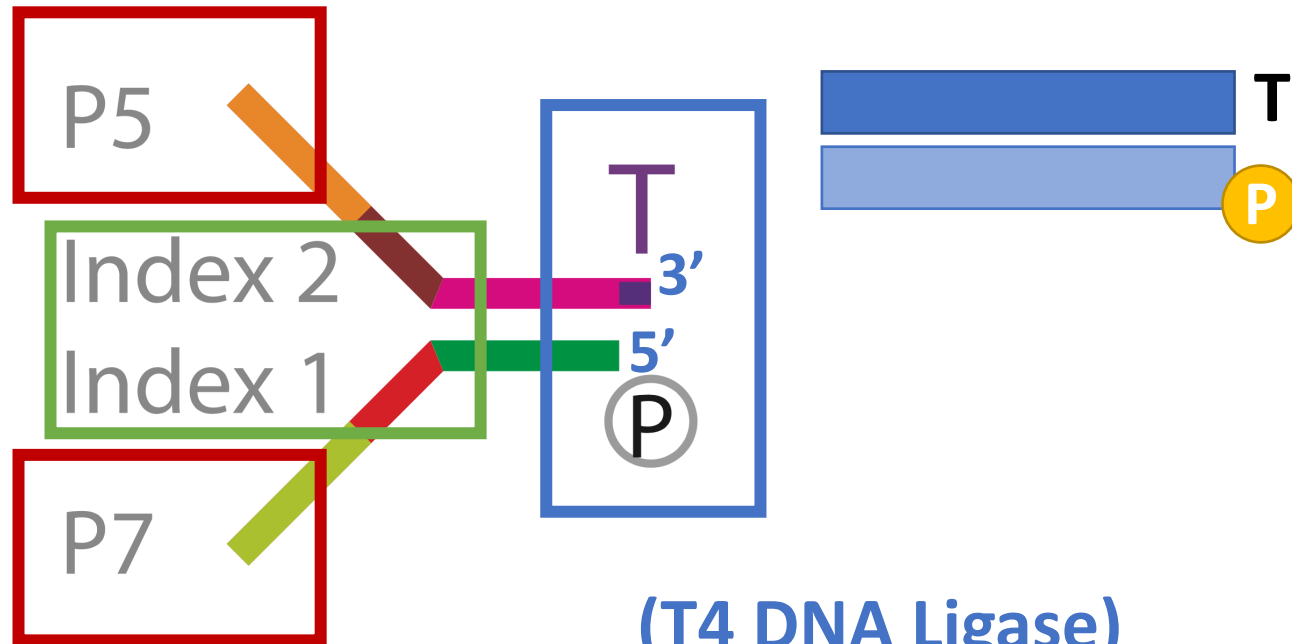


Y-shaped

100%

Illumina Adapters

POLL 4



(T4 DNA Ligase)

4. Ligation

**1. Sample
multiplexing**
(it's not the only way)

2. Amplification

Regions:

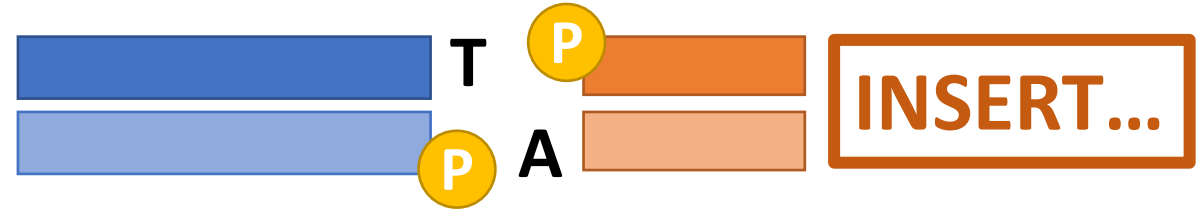
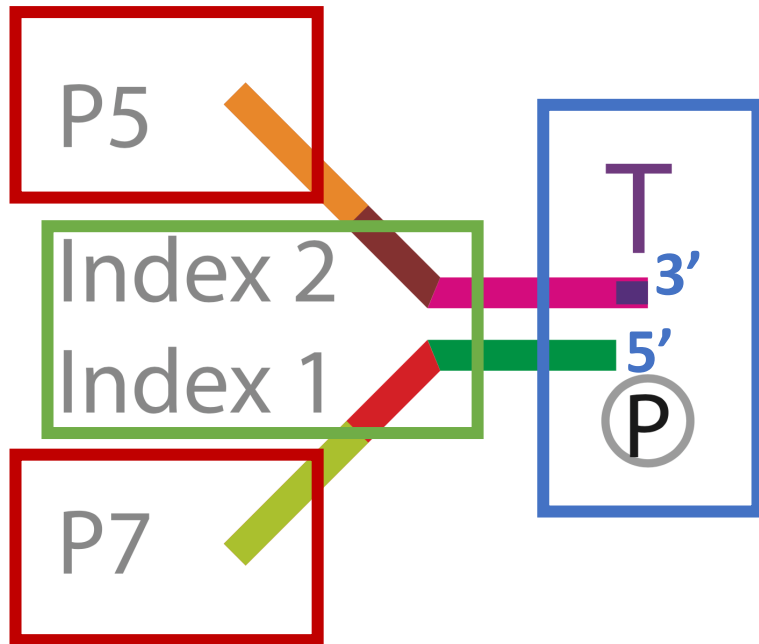
Flowcell binding+ PCR Primer

3. Y-shaped

Illumina Adapters

(T4 DNA Ligase)

4. Ligation



overhang 3'-T

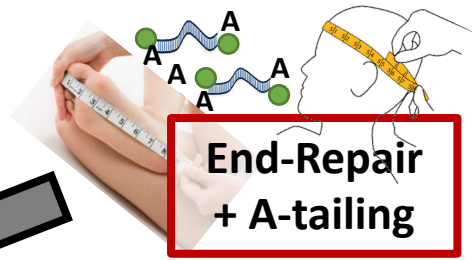
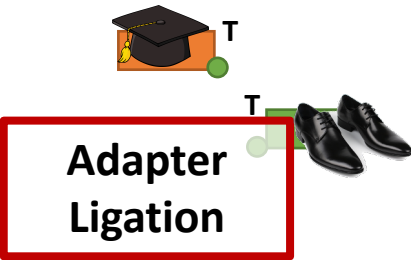
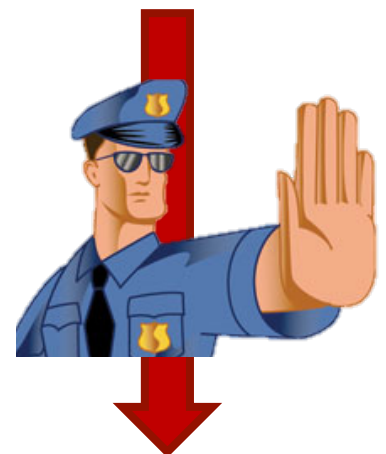
5'-P

5'-P

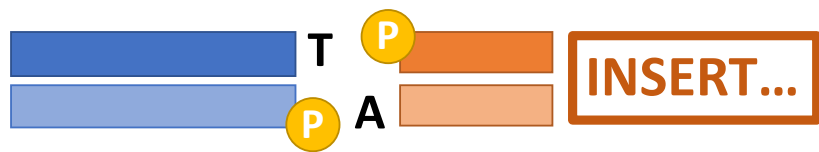
3'-A overhang

Fragmented DNA

- 1. DNA
- 2. short (average 200-800 bp)
- 3. modified with adapters

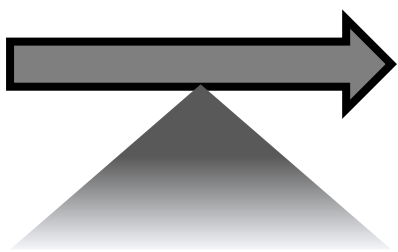


What is needed for Adapter Ligation?



End Repair and A-tailing!!!

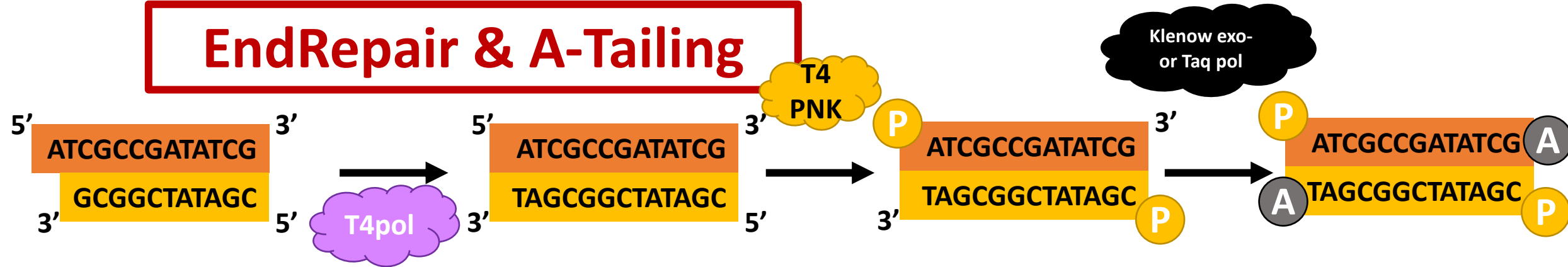
Fragmented
DNA



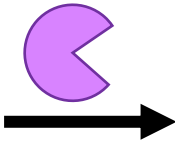
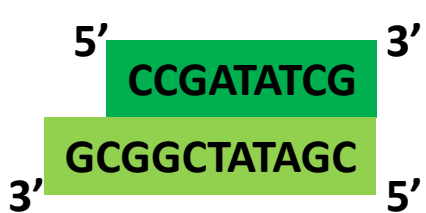
Adapter
Ligation

- 1. DNA
- 2. short (average 200-800 bp)
- 3. modified with adapters

EndRepair & A-Tailing



T4 DNA Polymerase
(fill-in and chew back)
→ blunting



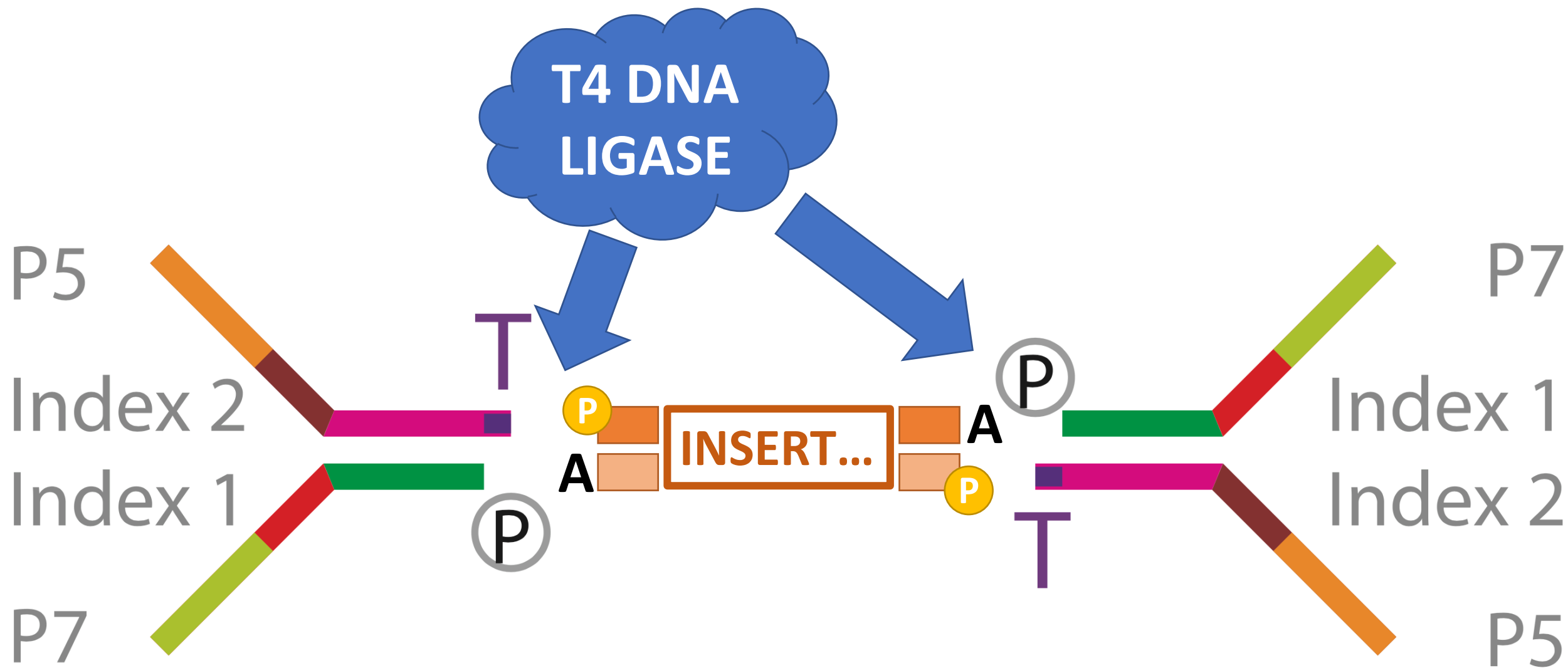
T4 PNK
→ 5'-P + 3'-OH



A-tailing

(Klenow exo- or Taq polymerase)
any polymerase lacking 3'->5' exonuclease activity





FINAL LIBRARY



1. DNA

2. short (average 200-800 bp)

3. modified with adapters



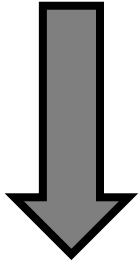
**KEEP
CALM
IT'S
BREAK
TIME**



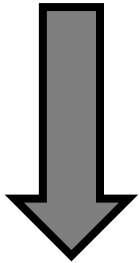
QUESTIONS
#2

LIBRARY PREP 10²~~1~~

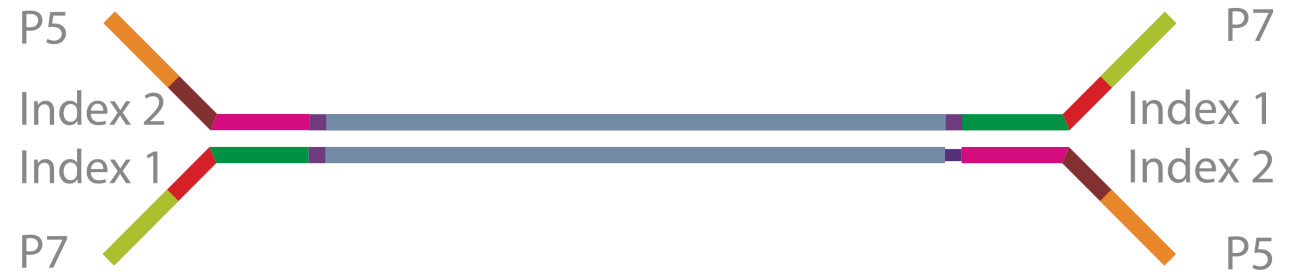
End-Repaired &
dA-tailed DNA



Adapter Ligation



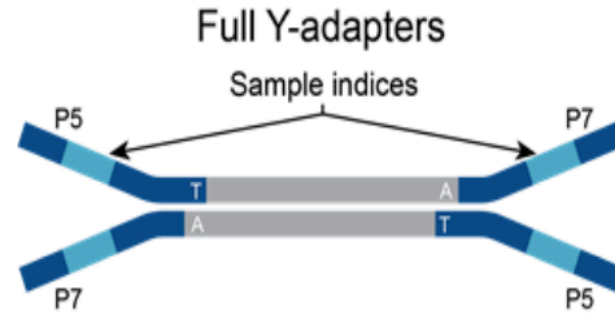
PCR - why?



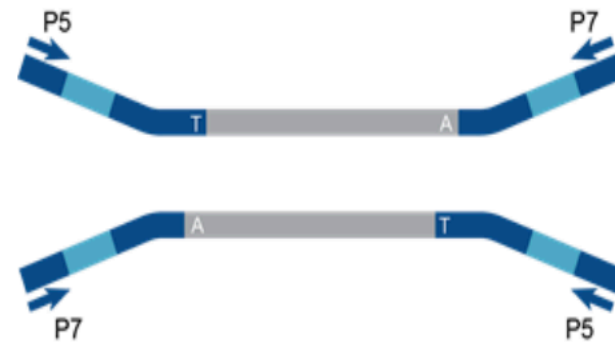
PCR introduces biases and **USELESS duplicated reads**
The fewer, the better. Monitor with qPCR

Indexing strategies using ligation

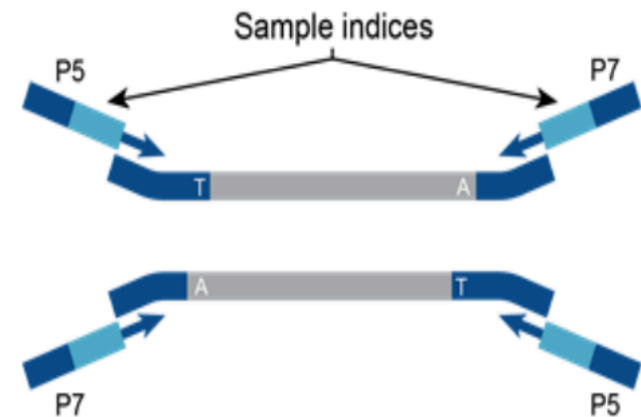
Ligation



Library amplification



Stubby Y-adapters



PCR - why?

FINAL LIBRARIES

1. QC → size distribution, presence of adapter dimers and primer dimers



- Agilent TapeStation/Bioanalyzer
- Agarose gel

2. Concentration



- Qubit BR
- ddPCR
- qPCR

3. Dilution



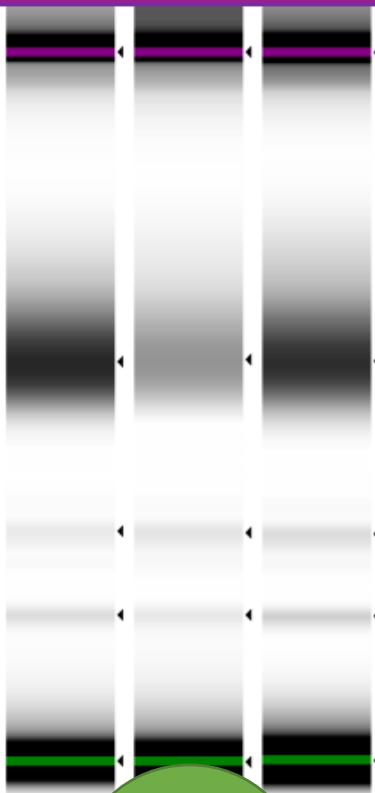
- Calculate molarity

4. Submission for Sequencing

FINAL LIBRARIES

LAB

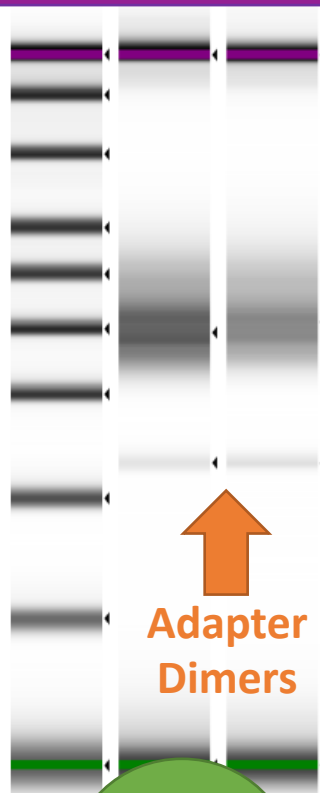
PCR
Primer
Dimers



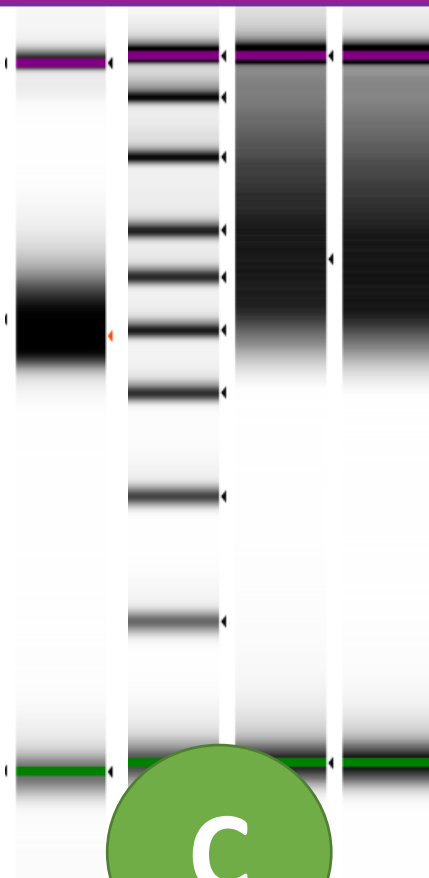
A

1,500
1,000
700
500
400
300
200
100
50
25

Adapter
Dimers



B



C

Final
Libraries

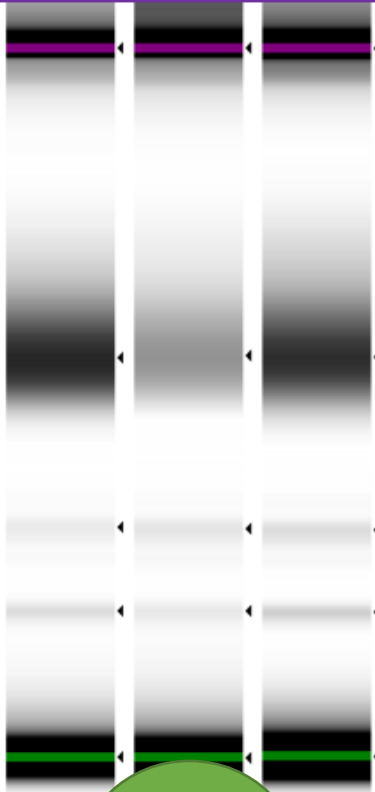
POLL 5



FINAL LIBRARIES

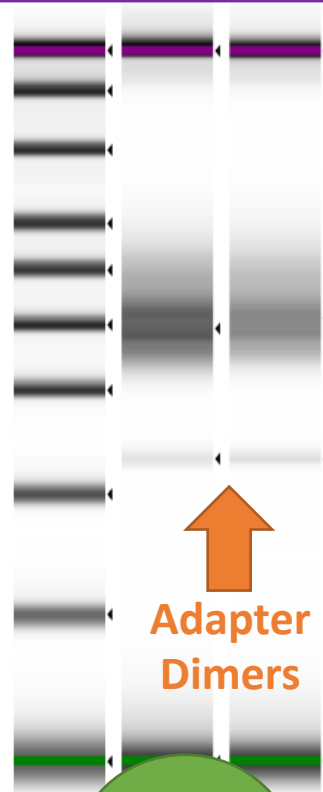
LAB

PCR
Primer
Dimers

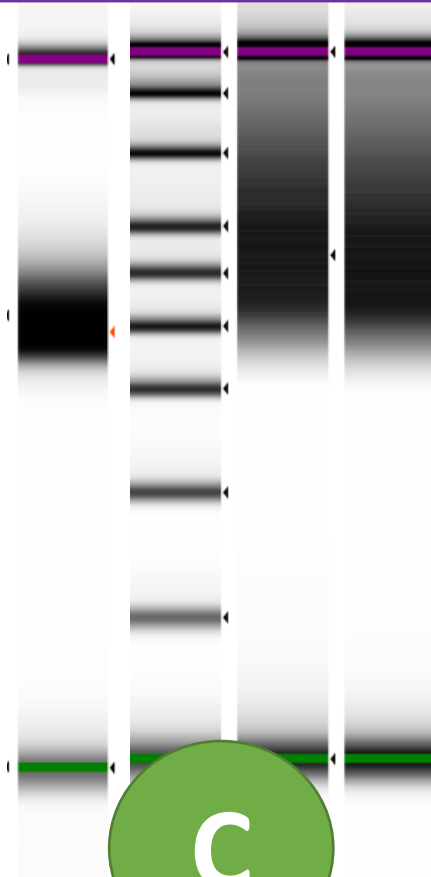


A

1,500
1,000
700
500
400
300
200
100
50
25



B



C

Final
Libraries

GOAL: Less than 1%

Final Libraries: 200-700 bp

Adapter Dimers: 120-140 bp

Primer Dimers < 100 bp

- WASTE OF READS (= WASTE OF \$\$\$\$)

- INDEX MISASSIGNMENT/INDEX HOPPING



GOAL: Less than 1%

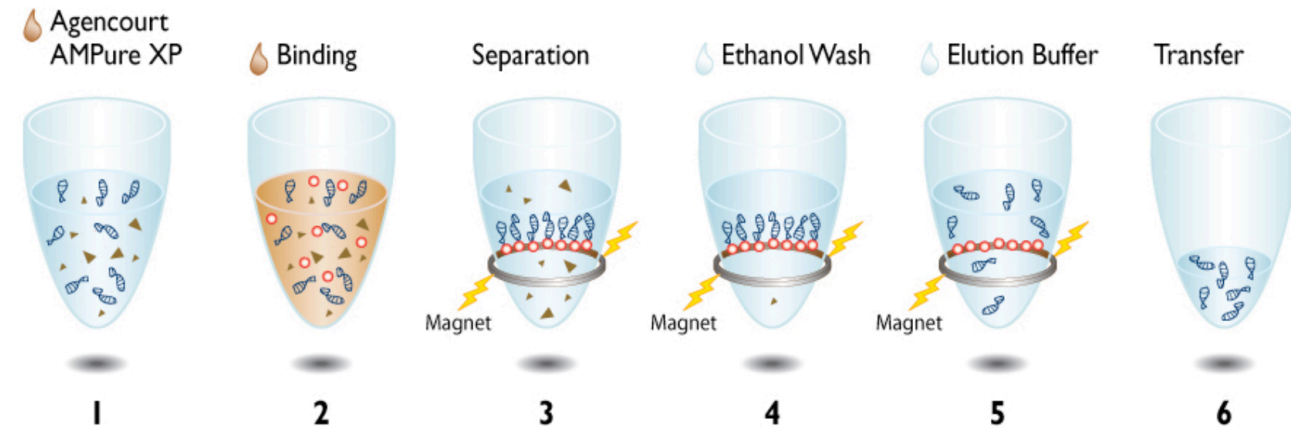
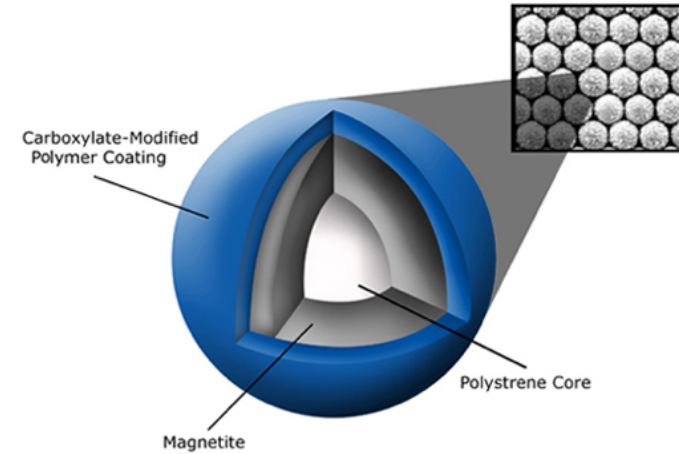
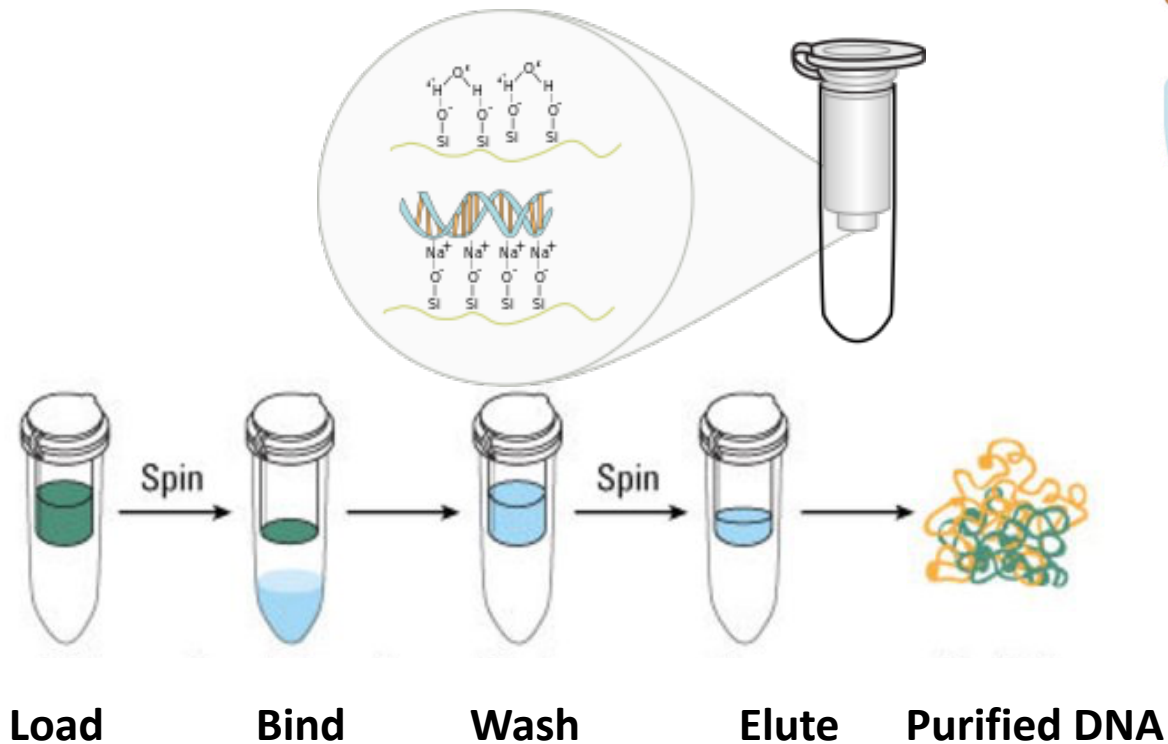
Final Libraries: 200-700 bp

Adapter Dimers: 120-140 bp

Primer Dimers < 100 bp

Purifications:

Spin column

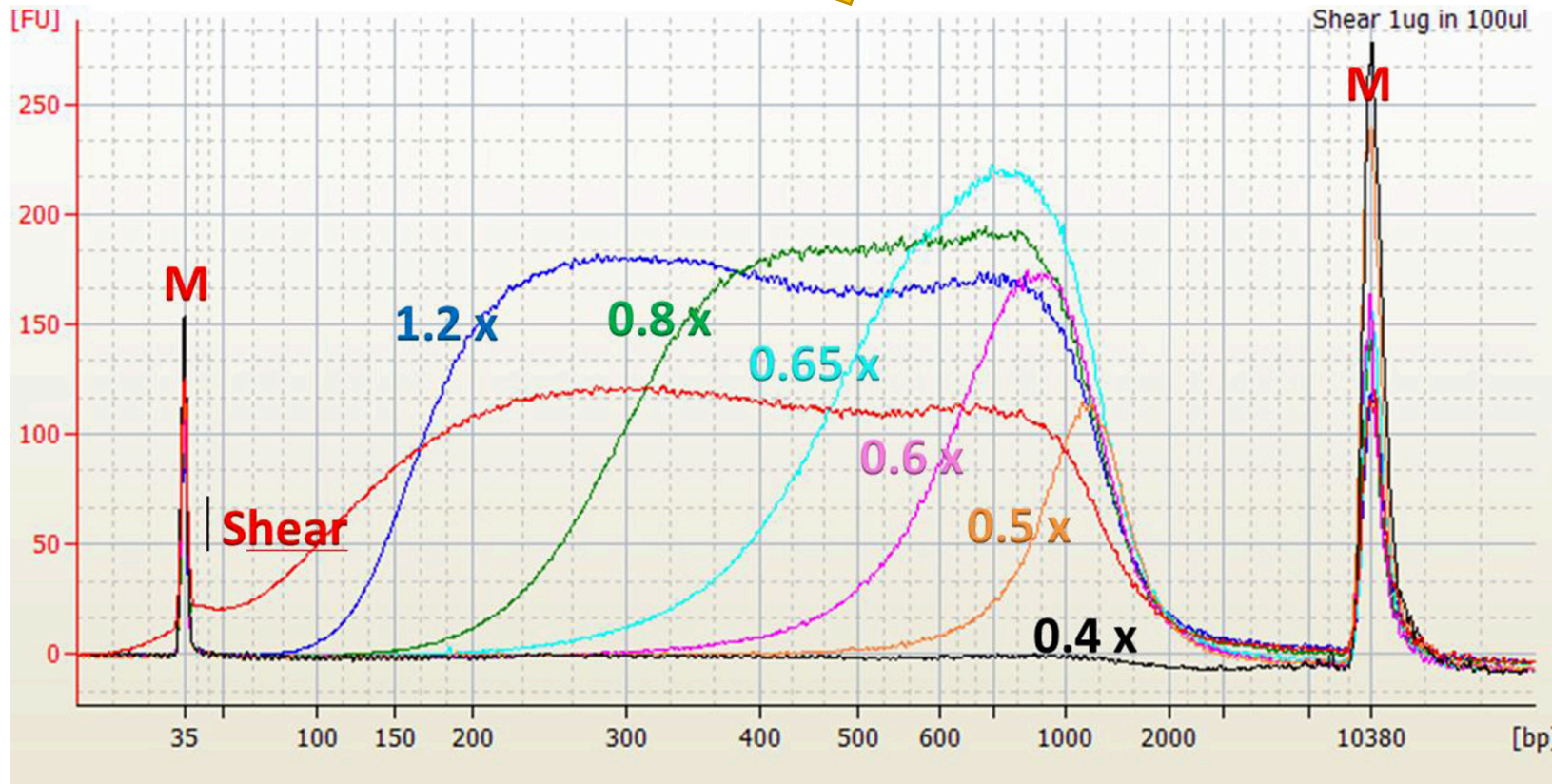


AMPure beads
Mag beads + PEG 8000 + NaCl

AMPure beads



→ Mag beads + PEG 8000 + NaCl



1.2x or 1:1.2 = 1 volume of DNA solution and 1.2 volumes of AMPure beads (50 µl:60 µl)

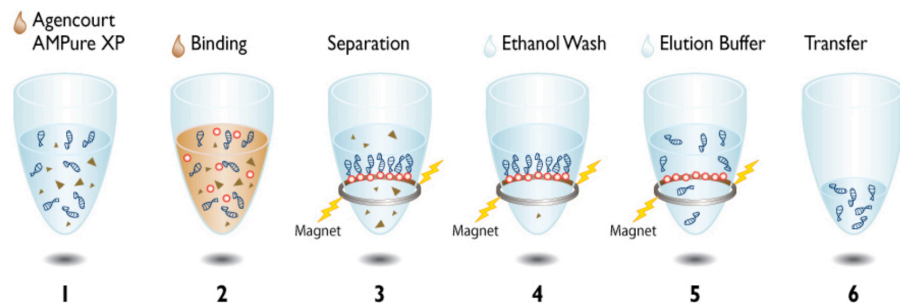
0.8x

0.65x

0.6x

0.5x

0.4x

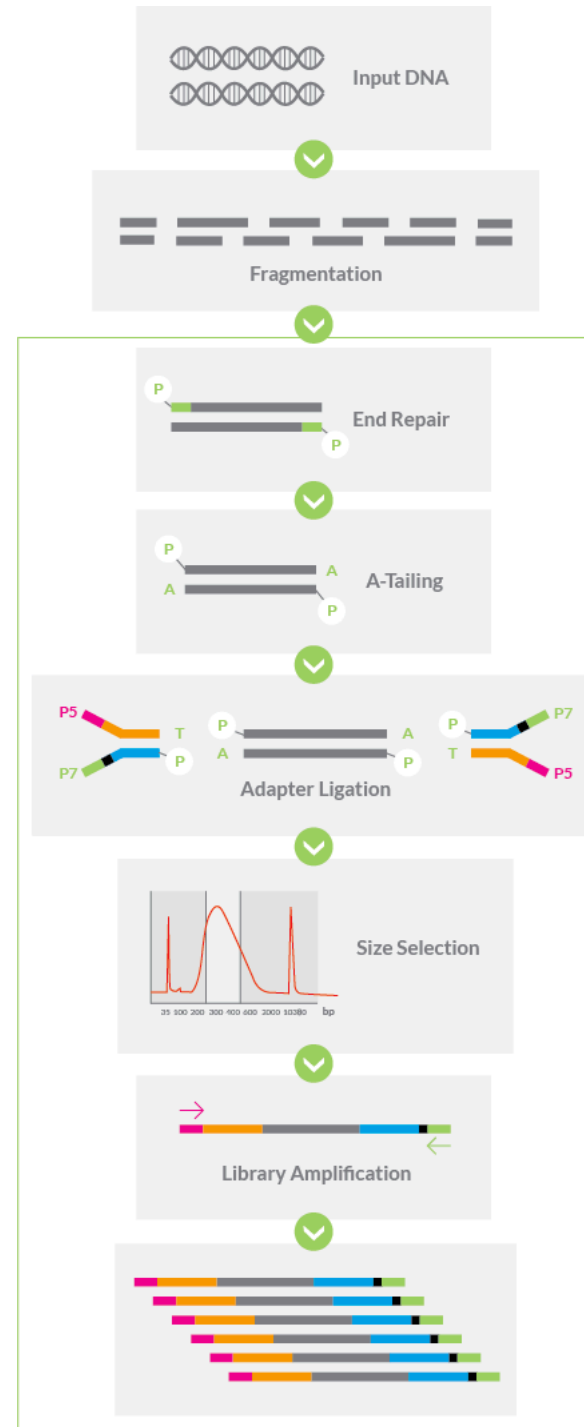


→ Beads preferentially bind long DNA

→ The higher the **buffer**:DNA ratio, the shorter the fragments bound

STANDARD (OLD) WORKFLOW

PURIFICATIONS



Total: 4.5 hours

Fragmentation: 10-30 min

End Repair: 30 min

Purification: 30-45 min

dA-tailing: 30 min

Adapter Ligation: 15 min

Purification: 30-45 min

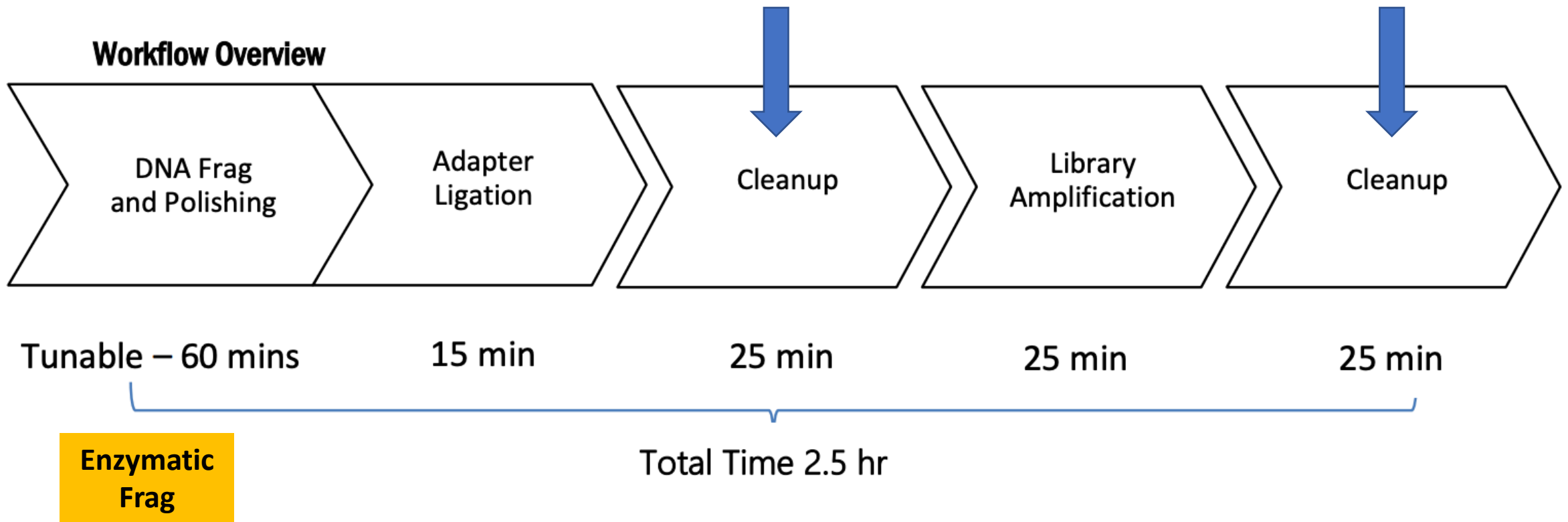
PCR: 30 min

Purification: 30-45 min

STREAMLINED WORKFLOW

LAB

PURIFICATIONS





**KEEP
CALM
IT'S
BREAK
TIME**



QUESTIONS
#3

**GOOGLE
FORMs for
ATTENDANCE
and mini-QUIZ
(DAY 1, 1st)**

SEQUENCING 101

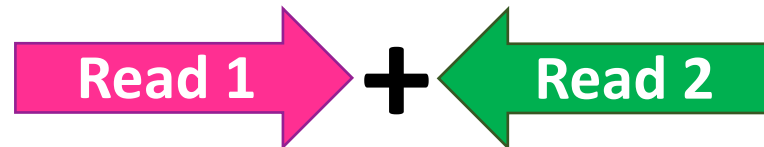
SEQUENCING



Single End - SE 50, 100...



Paired End - PE 50, 100...



Insert
Sequencing

Single-indexed



Dual-indexed



Sample-
multiplexing

SEQUENCING #2

Read Length #1



Some Examples:

SE50: ChIPseq, RNAseq, smallRNAseq

SE100: BS-seq

PE50 (or longer): ATACseq, Hi-C

PE100 (or longer): RNAseq (splicing or *de novo* transcriptome), WGS, WES

SEQUENCING #3

Read Length #2



WGS → any length (the longer the better: cheaper/base, but be aware of the fragment size) – **start at 100PE**

ChIPseq → SE50 – 20M uniquely mapped: 25-30M reads [or PE50]

RNAseq → SE50 – regular DiffExp polyA: 20-25M mapped reads

RNAseq → PE100/or longer – splicing/*de novo* transcriptome: 50/100M reads or more

ATACseq → PE50 or PE100 – 50M+ mapped reads (high content in chrM – improved protocols)

Hi-C → PE50 or PE100 – 300M/1B reads

BSseq/WGBS → SE100/PE100* – 400/800M reads (10-30x coverage)

RRBS → SE100 – 15-20M mapped reads

SEQUENCING #4

How to choose the sequencer?!?

4-Channel Chemistry				
				
	A	G	T	C
Image 1				
Image 2				
Image 3				
Image 4				
Result	A	G	T	C

2-Channel Chemistry				
				
	A	G	T	C
Image 1				
Image 2				
Result	A	G	T	C

1-Channel Chemistry				
				
	A	G	T	C
Image 1				
Image 2				
Result	A	G	T	C

----- Intermediate chemistry step

Figure 2: Four-, Two-, and One-Channel Chemistry—Four-channel chemistry uses a mixture of nucleotides labeled with four different fluorescent dyes. Two-channel chemistry uses two different fluorescent dyes, and one-channel chemistry uses only one dye. The images are processed by image analysis software to determine nucleotide identity.



iSeq100	MiniSeq	MiSeq	NextSeq	HiSeq 4000	HiSeq X Ten
					
1-channel SBS	2-channels SBS	4-channels SBS	2-channels SBS	4-channels SBS	4-channels SBS
MAX OUTPUT 1.2 Gb	MAX OUTPUT 8 Gb	MAX OUTPUT 15 Gb	MAX OUTPUT 120 Gb	MAX OUTPUT 1500 Gb	MAX OUTPUT 1800 Gb
MAX READ NUMBER 4 M	MAX READ NUMBER 25 million	MAX READ NUMBER 25 million	MAX READ NUMBER 400 million	MAX READ NUMBER 5 billion	MAX READ NUMBER 6 billion
MAX READ LENGTH 2x150 bp	MAX READ LENGTH 2x150 bp	MAX READ LENGTH 2x300 bp	MAX READ LENGTH 2x150 bp	MAX READ LENGTH 2x150 bp	MAX READ LENGTH 2x150 bp

MAX OUTPUT
2400-3000 Gb

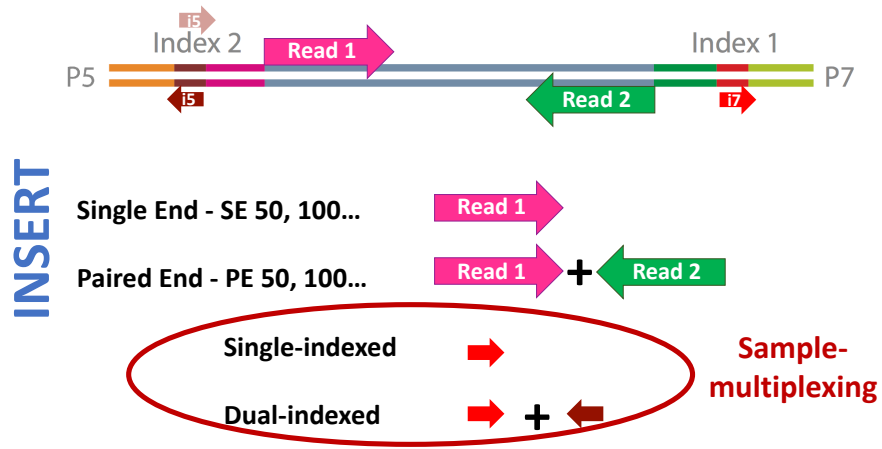
MAX OUTPUT
16-20 billion

MAX READ LENGTH
2x150 bp

NovaSeq

SEQUENCING #5

Sample Multiplexing - simplified



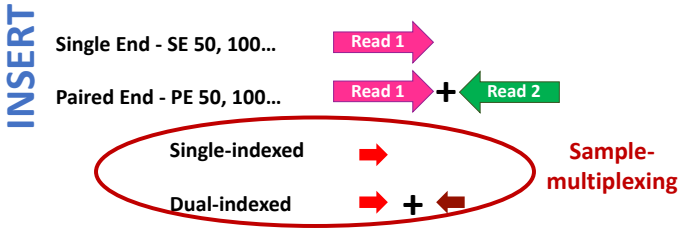
Strategy to combine multiple libraries
in a single run → **WHY?!?!?!?!?!?**

Illumina machines: the higher the
throughput, the lower cost/base
sequenced

	NovaSeq 6000 System			
Flow Cell Type	SP	S1	S2	S4
Human Genomes per Run	~4	~8	~20	~48
Exomes per Run	~40	~80	~200	~500
Transcriptomes per Run	~32	~64	~164	~400

SEQUENCING #5

Sample Multiplexing - simplified



Small 8"	6.95\$	50.2 in ²	13.8 ¢
Medium 14"	15.50\$	153.9 in ²	10.0 ¢
Large 18"	18.95\$	254.5 in ²	7.5 ¢
X-Large 23"	23.95\$	415.5 in ²	5.8 ¢
	Cost	Area	\$/in ²

Cheaper/in²



3.4 x total
cost

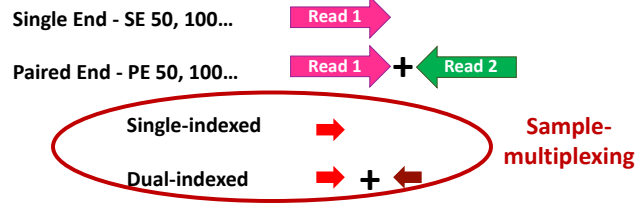
8.3 x bigger

POLL 6



SEQUENCING #5

INSERT



Cost/Human Genome	1'150 \$	890 \$	845 \$	520 \$
Flow-cell cost @UCLA <i>(not a lane)</i>	4'600 \$	7'100 \$	16'900 \$	25'000 \$

Small 8"	6.95\$	50.2 in ²	13.8 ¢
Medium 14"	15.50\$	153.9 in ²	10.0 ¢
Large 18"	18.95\$	254.5 in ²	7.5 ¢
X-Large 23"	23.95\$	415.5 in ²	5.8 ¢

Cheaper/in²



3.4 x total cost
8.3 x bigger

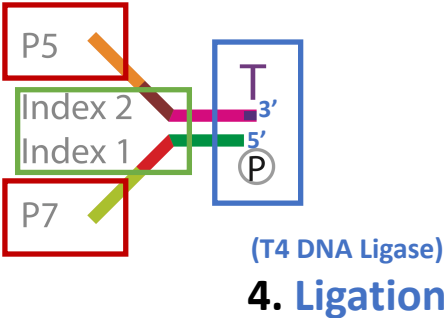
	NovaSeq 6000 System			
Flow Cell Type	SP	S1	S2	S4
Human Genomes per Run	~4	~8	~20	~48
Exomes per Run	~40	~80	~200	~500
Transcriptomes per Run	~32	~64	~164	~400

Output/flow-cell 2x150 (=150PE)	250 Gb	500 Gb	1250 Gb	3000 Gb
------------------------------------	--------	--------	---------	---------



SEQUENCING #5

Illumina Adapters



1. Sample
multiplexing
(it's not the only way)

2. Amplification
Regions:
Flowcell binding+ PCR Primer

3. Y-shaped

Sample
multiplexing

Cost/Human Genome

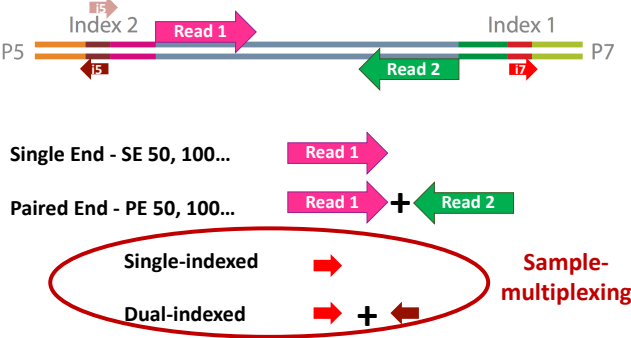


1'150 \$ 890 \$ 845 \$ 520 \$

Flow-cell cost @UCLA
(not a lane)

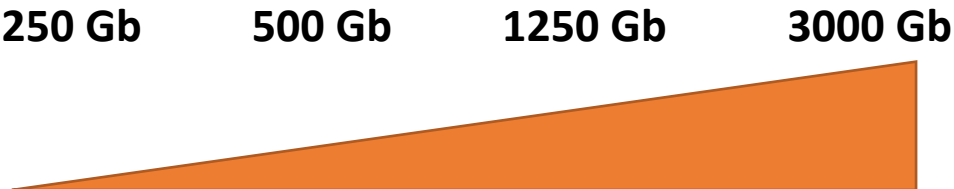
4'600 \$ 7'100 \$ 16'900 \$ 25'000 \$

INSERT



	NovaSeq 6000 System			
Flow Cell Type	SP	S1	S2	S4
Human Genomes per Run	~4	~8	~20	~48
Exomes per Run	~40	~80	~200	~500
Transcriptomes per Run	~32	~64	~164	~400

Output/flow-cell
2x150 (=150PE)



250 Gb 500 Gb 1250 Gb 3000 Gb



**KEEP
CALM
IT'S
BREAK
TIME**



QUESTIONS
#4

TruSeq Nano

WGS #2

Nextera Library Preparation

Double-stranded DNA



Fractionate
Size select



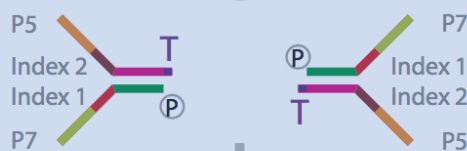
End repair
Phosphorylate



A-overhang



Add Adaptors



Adaptor ligation

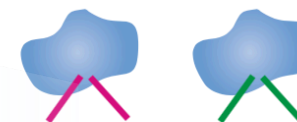


Denature and amplify



Product ready for
cluster generation

VS.



Transposase



DNA



Tagmentation

~300bp



Amplification



Product ready for
cluster generation

TruSeq Nano

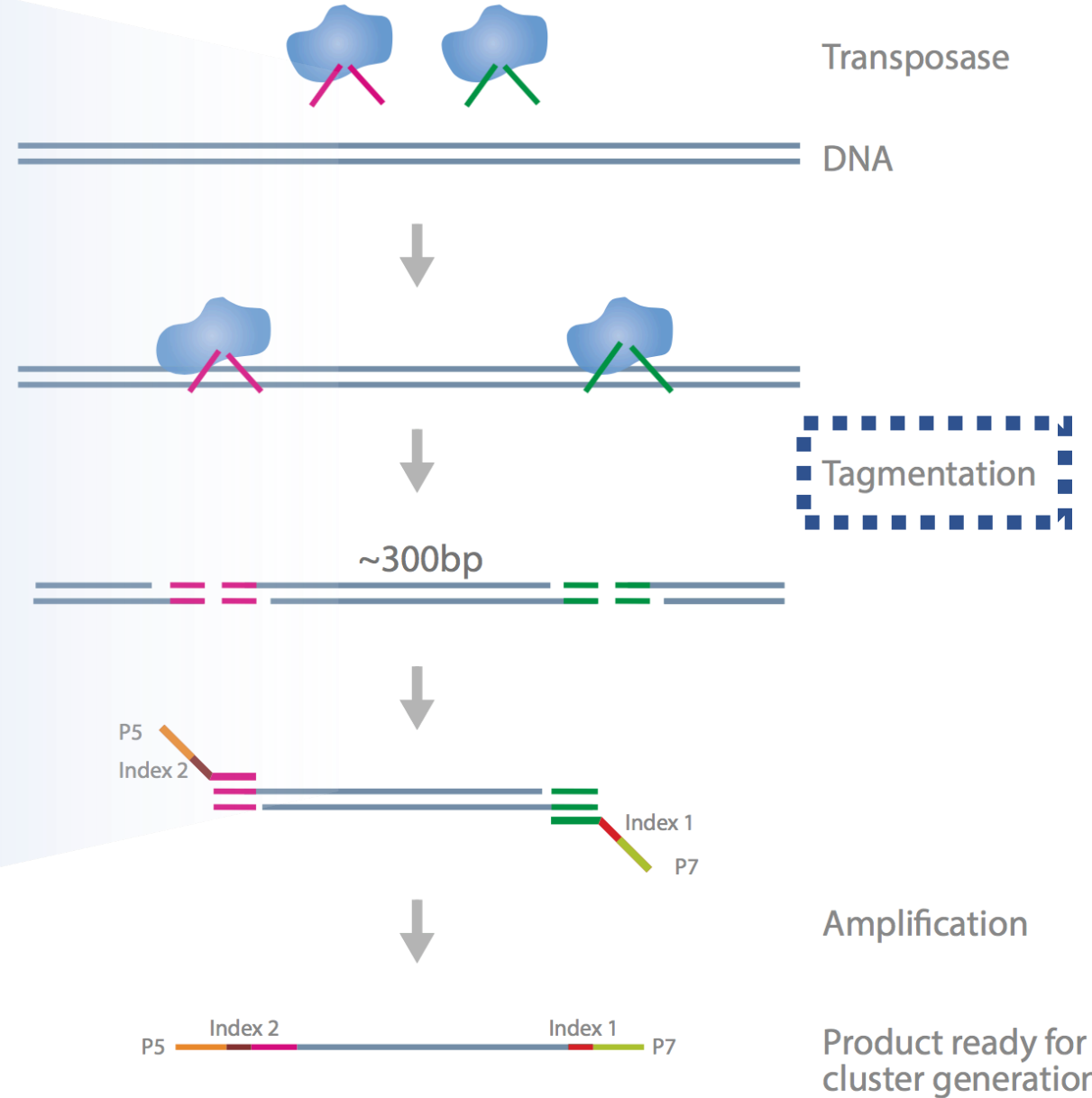
WGS #2

Nextera Library Preparation

Tagmentation =

Fragmentation and
Adapter Ligation*
Simultaneously

VS.

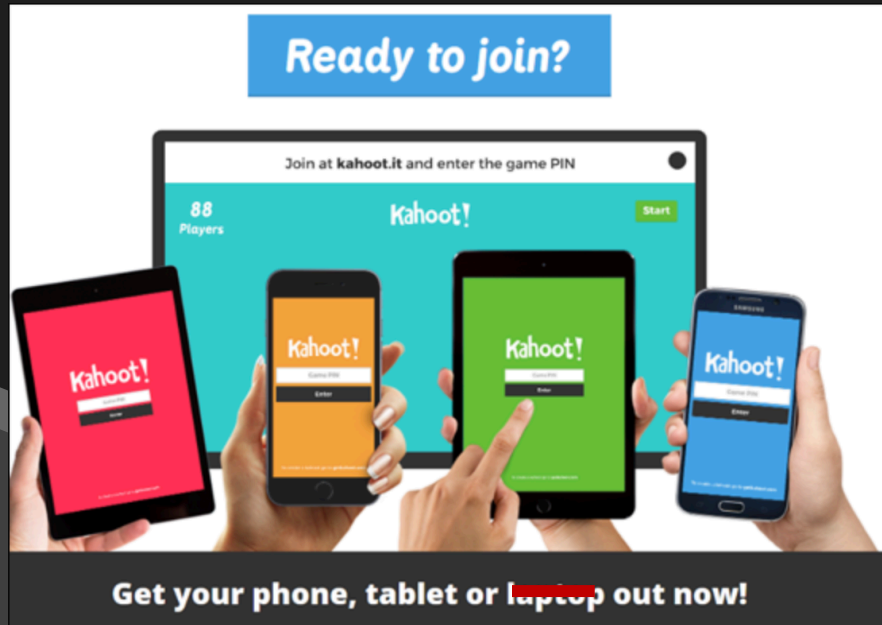


Denature and amplify

Product ready for
cluster generation

→ Go to Kahoot.it

then follow on-
screen instruction



QUIZ DNA

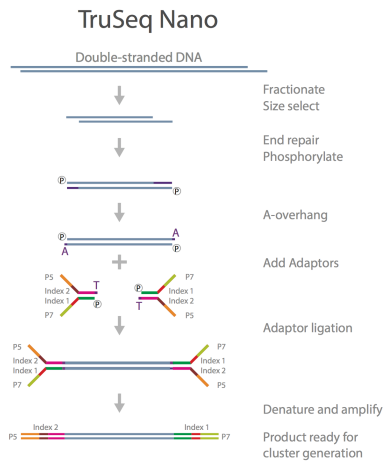
Kahoot!

Game PIN

Enter

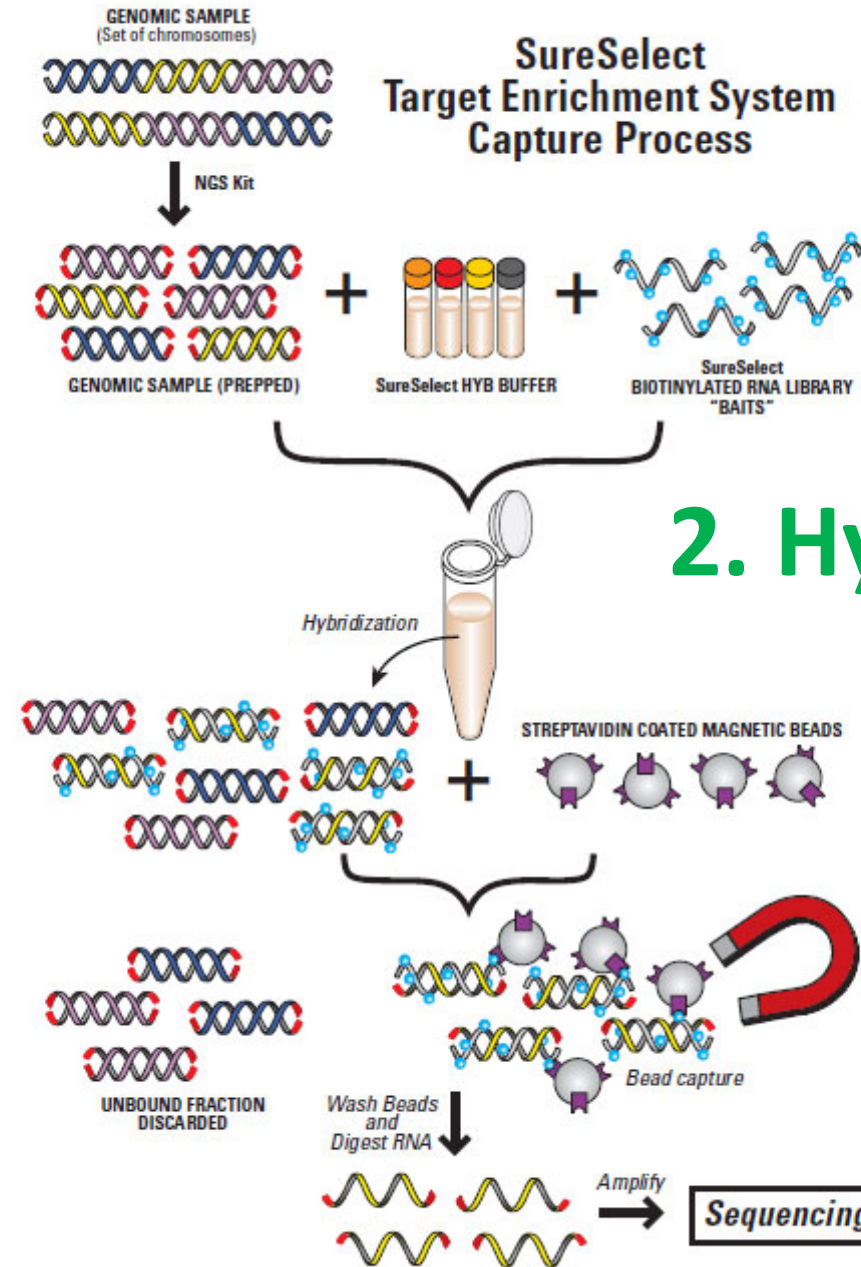
Common Variations on the DNA Sequencing (DNA-seq) Theme

- **Target Capture Sequencing (e.g. Exome-Seq):** mutations
- **Whole Genome Bisulfite-Seq (WGBS) & Co** DNA Methylation
- **Chromatin Immunoprecipitation (ChIP-seq):** sequencing your favorite protein-associated DNA
- **Assay for Transposase-Accessible Chromatin (ATACseq):** open chromatin



1. DNA Library Prep ➔

**Target Capture
Sequencing
e.g. Exome-Seq**



2. Hybridization Capture

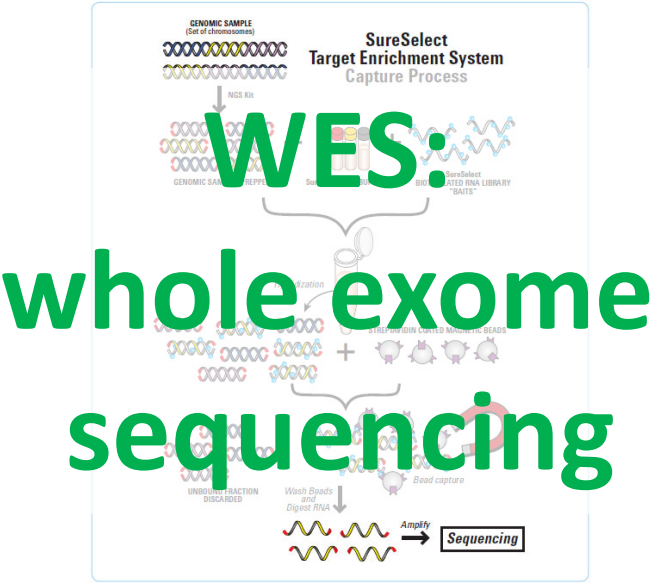
	NimbleGen's SeqCap EZ Exome Library	Agilent's Sure Select Human All Exon Kit	Illumina's TruSeq Exome Enrichment Kit	Illumina's Nextera Rapid Capture Exome Kit
Probe size, bp	55–105	114–126	95	95
Probe type	DNA	RNA	DNA	DNA
Coverage strategy	High-density, overlapping probes	Adjacent probes	Gaps between probes	Gaps between probes
Fragmentation method	Ultrasonication	Ultrasonication	Ultrasonication	Transposomes
Target region size (human), Mb	64	50	62	62
Reads remaining after filtering	66%	71.7%	54.8%	40.1%
Major strengths	(i) High sensitivity and specificity	(i) Better coverage of indels	(i) Good coverage of UTRs and miRNAs	(i) Good coverage of UTRs and miRNAs
	(ii) Most uniform coverage in difficult regions	(ii) High alignment rate		
		(iii) Fewer duplicate reads than other platforms		
Major weaknesses	(i) More duplicate reads than Agilent	(i) Fewer high-quality reads than NimbleGen	(i) High off-target enrichment	(i) High off-target enrichment
	(ii) Lower alignment rate than Agilent			(ii) Coverage bias for high GC content areas reducing uniformity
Nonhuman supported	Yes	Yes	No	No

Exome-Seq Kits	Targeted Region	Number of Probes	Probe Type	Genomic DNA input required	Adapter addition	Probe Length (mer)	Probe Design	Price per capture (negotiable)	Designed on build	Hybridization time (hours)
Agilent SureSelect XT2 V6 Exome	60 Mb	~758,086	biotinylated cRNA baits	100 ng	Ligation	120	Non-overlapping, paired-end reads used to fill gaps	\$270	GRCh37 (hg19)	16
Agilent SureSelect XT2 V5 Exome	51 Mb	~655,872	biotinylated cRNA baits	100 ng	Ligation	120	Non-overlapping, paired-end reads used to fill gaps	\$200	GRCh37 (hg19)	16
IDT xGEN Exome Panel	39 Mb	429,826	biotinylated DNA baits	500 ng	Ligation	not described	Non-overlapping	\$250	GRCh37 (hg19)	4
Illumina Nextera Rapid Capture Expanded Exome	62 Mb	>340,000	biotinylated DNA bait	50 ng	Transposase	95	Non-overlapping (adjacent to each other)	\$250	GRCh37 (hg19)	24-48
Roche Nimblegen SeqCap EZ Exome v3.0	64 Mb	>2,100,000	biotinylated DNA bait	1 ug	Ligation	60 - 90	Overlapping baits	\$600	GRCh37 (hg19)	72

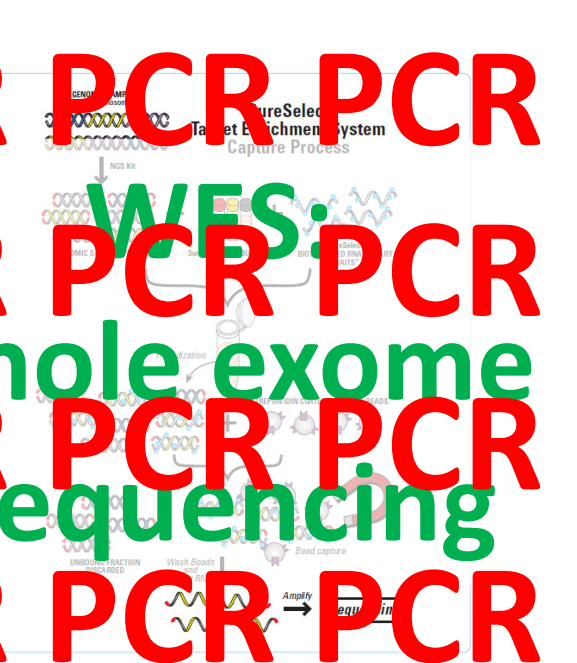
POLL 7



1-2% OF THE ENTIRE GENOME

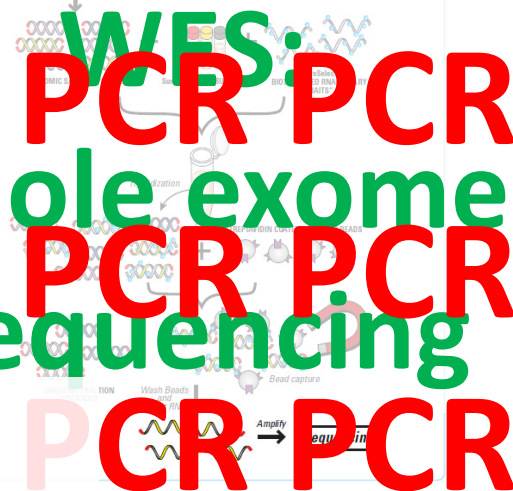


	NimbleGen's SeqCap EZ Exome Library	Agilent's SureSelect Human All Exon Kit	Illumina's TruSeq Exome Enrichment Kit	Illumina's Nextera Rapid Capture Exome Kit
Probe size, bp	110	110-126	95	95
Probe type	DNA	RNA	DNA	DNA
Coverage strategy	High-density, overlapping probes	Adjacent probes	Gaps between probes	Gaps between probes
Fragmentation method	Ultrasonication	Ultrasonication	Ultrasonication	Transposase
Target region size (human), Mb	64	50	62	62
Reads remaining after adapter removal	66%	77%	84.8%	40%
Major strengths	(i) High sensitivity and specificity (ii) Most uniform coverage in difficult regions	(i) Better coverage of indels (ii) High alignment rate (iii) Fewer duplicate reads than other platforms	(i) Good coverage of UTRs and miRNAs	(i) Good coverage of UTRs and miRNAs
Major weaknesses	(i) More duplicate reads than Agilent (ii) Lower alignment rate than Agilent	(i) Fewer high-quality reads than NimbleGen	(i) High off-target enrichment	(i) High off-target enrichment
Non-human supported	Yes	Yes	No	No



Exome-Seq kits	Targeted Region	Number of Probes	Probe type	Genomic DNA input required	Adapter addition	Probe length (bp)	Probe Design	Price per capture (negotiable)	Designed on build	Hybridization time (hours)
Agilent SureSelect XT2 V6 Exome	60 Mb	~758,086	biotinylated cRNA baits	100 ng	Ligation	120	Non-overlapping, paired-end reads used to fill gaps	\$270	GRCh37 (hg19)	16
Agilent SureSelect XT2 V5 Exome	51 Mb	~517,872	biotinylated cRNA baits	100 ng	Ligation	120	Non-overlapping, paired-end reads used to fill gaps	\$200	GRCh37 (hg19)	16
IDT xGEN Exome Panel	39 Mb	429,826	biotinylated DNA baits	500 ng	Ligation	not described	Non-overlapping	\$250	GRCh37 (hg19)	4
Illumina Nextera Rapid Capture Expanded Exome	52 Mb	~340,000	biotinylated DNA bait	10 ng	Transposase	75	Non-overlapping (adjacent to each other)	\$250	GRCh37 (hg19)	24-48
Roche NimbleGen SeqCap EZ Exome v3.0	64 Mb	>2,100,000	biotinylated DNA bait	1 µg	Ligation	60-90	Overlapping baits	\$600	GRCh37 (hg19)	72

	NimbleGen's SeqCap EZ Exome Library	Agilent's SureSelect Human All Exon Kit	Illumina's TruSeq Exome Enrichment Kit	Illumina's Nextera Rapid Capture Exome Kit
Probe size, bp	105	110-126	95	95
Probe type	DNA	RNA	DNA	DNA
Coverage strategy	High-density, overlapping probes	Adjacent probes	Gaps between probes	Gaps between probes
Fragmentation method	Ultrasonication	Ultrasonication	Ultrasonication	Transposase
Target region size (human), Mb	64	50	62	62
Reads remaining after adapter removal	66%	71-75%	71-8%	40%
Major strengths	(i) High sensitivity and specificity (ii) Most uniform coverage in difficult regions	(i) Better coverage of indels (ii) High alignment rate (iii) Lower duplicate read rate than other platforms	(i) Good coverage of UTRs and miRNAs (ii) High off-target rate	(i) Good coverage of UTRs and miRNAs (ii) High off-target rate
Major weaknesses	(i) More duplicate reads than Agilent (ii) Lower alignment rate than Agilent	(i) Fewer high-quality reads than NimbleGen (ii) Lower alignment rate than Agilent	(i) High off-target rate (ii) High duplicate read rate	(i) High off-target rate (ii) High duplicate read rate
Non-human supported	Yes	Yes	No	No



PCR DUPLICATES

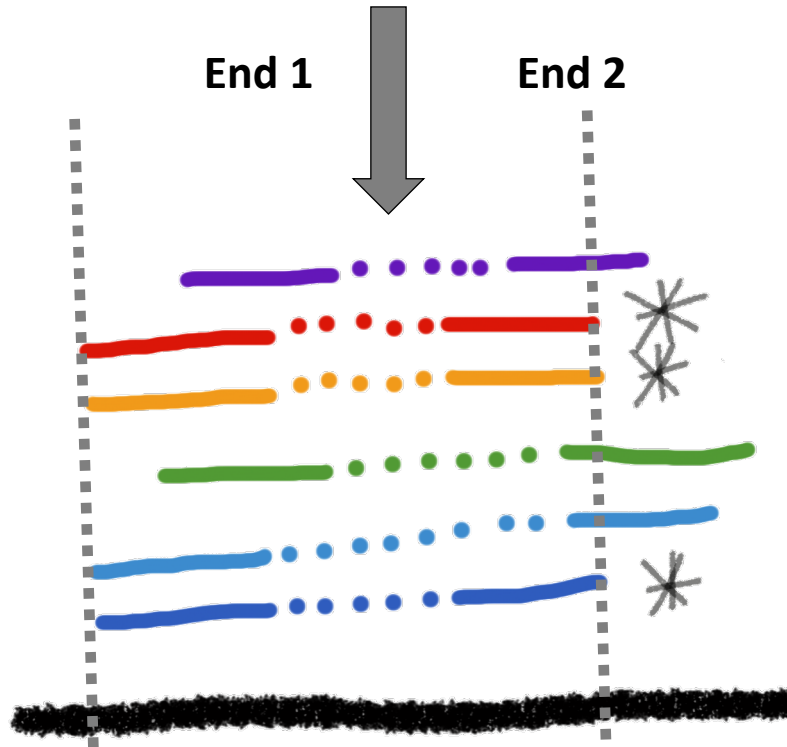
Exome-Seq kits	Targeted Region	Number of Probes	Probe type	Genomic DNA input required	Adapter addition	Probe length (bp)	Probe design	Price per capture (negotiable)	Designed on build	Hybridization time (hours)
Agilent SureSelect XT2 V6 Exome	60 Mb	~758,086	biotinylated cRNA baits	100 ng	Ligation	120	Non-overlapping, paired-end reads used to fill gaps	\$270	GRCh37 (hg19)	16
Agilent SureSelect XT2 V5 Exome	51 Mb	~517,872	biotinylated cRNA baits	100 ng	Ligation	120	Non-overlapping, paired-end reads used to fill gaps	\$200	GRCh37 (hg19)	16
IDT xGEN Exome Panel	39 Mb	429,826	biotinylated DNA baits	500 ng	Ligation	not described	Non-overlapping	\$250	GRCh37 (hg19)	4
Illumina Nextera Rapid Capture Expanded Exome	52 Mb	~340,000	biotinylated DNA bait	10 ng	Transposase	75	Non-overlapping (adjacent to each other)	\$250	GRCh37 (hg19)	24-48
Roche NimbleGen SeqCap EZ Exome v3.0	64 Mb	>2,100,000	biotinylated DNA bait	1 µg	Ligation	60-90	Overlapping baits	\$600	GRCh37 (hg19)	72

HOW DO YOU DETECT PCR DUPLICATES?

ALIGNMENT POSITION

in silico

(Paired-End is better than Single-End, especially for high coverage)



UMI

in silico + experimental

(Random or Pre-Designed)

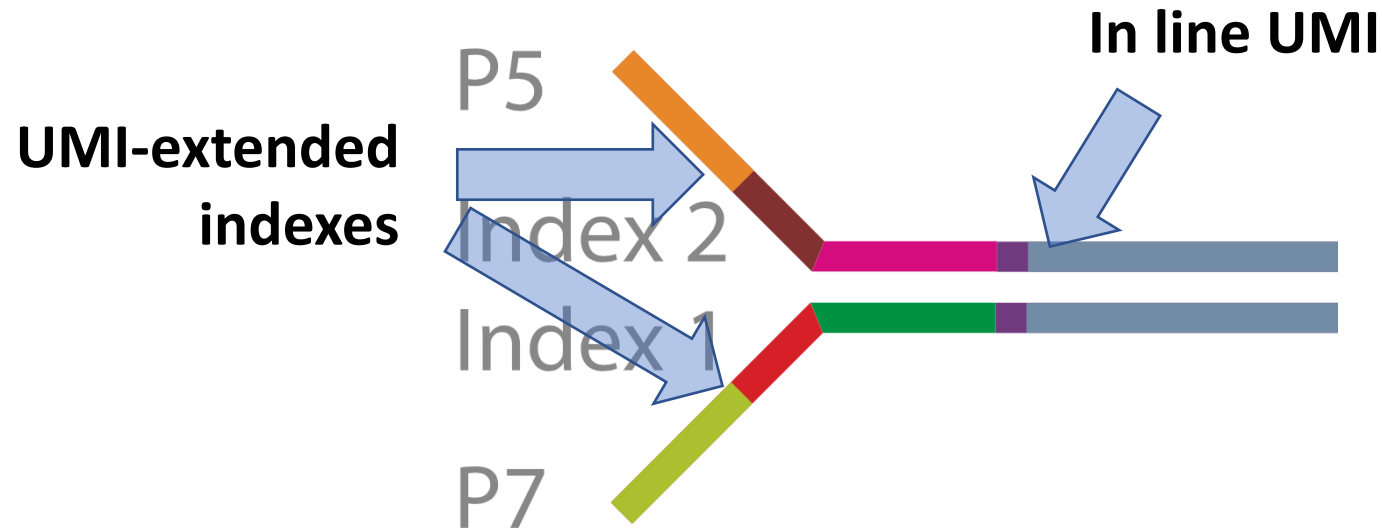
**Total Number of
Molecules = 4**

UMI: Unique Molecular Identifier

Short sequence (8-16 nt barcode) used to uniquely label a molecule of nucleic acid in order to reduce the bias introduced by PCR.

DIFFERENT FROM LIBRARY INDEXES/BARCODES!!!!!!!!!!!!!!!!!!!!!!

➔ low-input NGS applications and good practice

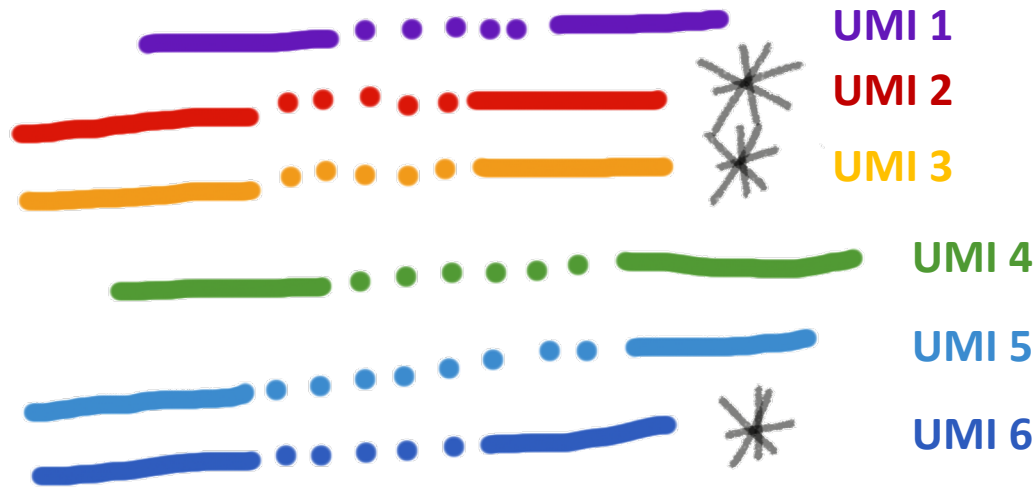


HOW DO YOU DETECT PCR DUPLICATES?

ALIGNMENT POSITION

in silico

(Paired-End is better than Single-End, especially for high coverage)



UMI

in silico + experimental

(Random or Pre-Designed)

Random
e.g. NNNNNNNNN

Total Number of
Molecules = 6

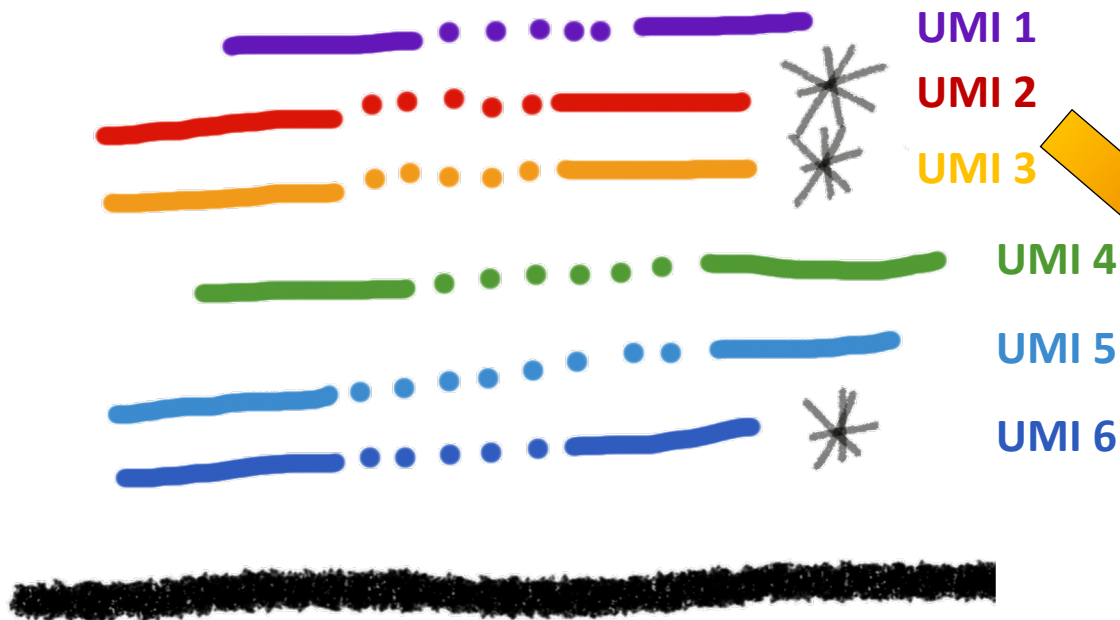
chr1:100'000-103'000

HOW DO YOU DETECT PCR DUPLICATES?

ALIGNMENT POSITION

in silico

(Paired-End is better than Single-End, especially for high coverage)



UMI

in silico + experimental

(Random or Pre-Designed)

Pre-Designed

ATCGATCGAT
TAGGCGTGTT
TGCCGTAATC
...

Total Number of
Molecules = 5

UMI 2 = ATCGATCGAT

UMI 3 = ATCGATCGAA

← Not a pre-designed sequence

GOOGLE FORMs
for ATTENDANCE
and mini-QUIZ
(DAY 1, 2nd)

TIME FOR ONLINE QUESTIONS

Mantua, Italy

QUESTIONS

#6



**...SEE YOU
TOMORROW**