WORKSHOP 16: LIBRARY PREP FOR NGS

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)
1. INTRO TO SEQUENCING TECHNOLOGIES
2. LIBRARY PREP 101
3. LIBRARY PREP 102
4. SEQUENCING 101
5. TARGETED SEQUENCING & PCR duplicates

MINI-QUIZ + ATTENDANCE
KAHOOOT!?!?
MINI-QUIZ + ATTENDANCE
FINAL

It depends on the

#1
#2
#3
#4
#5
#6
### QUESTIONS #1 - INTRO: SEQUENCING TECHNOLOGIES

<table>
<thead>
<tr>
<th>FIRST NAME</th>
<th>LAST NAME</th>
<th>email</th>
<th>QUESTION</th>
<th>ANSWER</th>
</tr>
</thead>
<tbody>
<tr>
<td>MARCO</td>
<td>MORSELLI</td>
<td><a href="mailto:mmorselli@ucla.edu">mmorselli@ucla.edu</a></td>
<td>How old are you?</td>
<td></td>
</tr>
</tbody>
</table>

**WRITE HERE**

### QUESTIONS #2 - LIBRARY PREP 101

<table>
<thead>
<tr>
<th>FIRST NAME</th>
<th>LAST NAME</th>
<th>email</th>
<th>QUESTION</th>
<th>ANSWER</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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
Third-Generation Sequencing:
- Single-molecule sequencing
- Real time
- Longer Reads
- Higher Error-rate (but..)
- Lower number of Reads
- Read native modifications!!!
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!!!
e.g. HiSeq3000/4000/X Flow cell = 8 lanes → each lane has ≈480M 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*
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

Qubit Fluorometer

Pro: more accurate and specific + low concentration
Cons: does not detect contaminants

<table>
<thead>
<tr>
<th>Substance</th>
<th>Absorbance (nm)</th>
<th>Purity indicator</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nucleic acids</td>
<td>260 nm</td>
<td>A260:280 &gt; 1.8 DNA</td>
</tr>
<tr>
<td>Proteins</td>
<td>280 nm</td>
<td>A260:280 &gt; 2.0 RNA</td>
</tr>
<tr>
<td>Organics</td>
<td>230 nm</td>
<td>A260:230 &gt; 2.0 DNA/RNA</td>
</tr>
</tbody>
</table>

Agarose Gel

Agilent TapeStation Genomic DNA - DIN
Purified gDNA

Fragmentation

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

1. Nebulization
2. Sonication
3. Enzymatic
4. Chemical

RNA: Magnesium or Zinc
2. Sonication

Purified gDNA

Fragmentation

3. Enzymatic
Fragmented DNA

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)

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

<table>
<thead>
<tr>
<th>Sample</th>
<th>Index 1 (i7)</th>
<th>Index 2 (i5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample 1</td>
<td>ATATGCGC</td>
<td>CTGATCGT</td>
</tr>
<tr>
<td>Sample 2</td>
<td>TGGTACAG</td>
<td>ACTCTCGA</td>
</tr>
<tr>
<td>Sample 3</td>
<td>AACCGTTG</td>
<td>TGAGCTAG</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
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% 100%
Illumina Adapters

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

2. Amplification Regions:
Flowcell binding + PCR Primer

3. Y-shaped

4. Ligation (T4 DNA Ligase)
Illumina Adapters

4. Ligation

(T4 DNA Ligase)

P5

Index 2

Index 1

P7

overhang 3’-T

5’-P

5’-P

3’-A overhang

INSERT...
Fragmented DNA

1. DNA
2. Short (average 200-800 bp)
3. Modified with adapters

End Repair and A-tailing!!!

What is needed for Adapter Ligation?
1. DNA
2. short (average 200-800 bp)
3. modified with adapters

Fragmented DNA → Adapter Ligation

EndRepair & A-Tailing

ATCGCCGATATCG
GCGGCTATAGC

5'
3'

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

ATCGCCGATATCG
TAGCGGCTATAGC

T4 PNK

ATCGCCGATATCG
TAGCGGCTATAGC

P

3'

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

CCGATATCG
GCGGCTATAGC

5'
3'

Klenow exo- or Taq pol

T4 PNK

→ 5'-P + 3'-OH

CCGATATCG
GGCTATAGC

P

3'

P

A

T4 PNK
1. DNA
2. short (average 200-800 bp)
3. modified with adapters
KEEP CALM IT'S BREAK TIME

QUESTIONS #2
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

PCR - why?
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

PCR Primer Dimers

A

B

C

Adapter Dimers

Final Libraries

POLL 5
Final Libraries:

- 200-700 bp
- 120-140 bp
- < 100 bp

GOAL: Less than 1%
- 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

Load → Bind → Wash → Elute → Purified DNA
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

PURIFICATIONs

Workflow Overview

DNA Frag and Polishing

Adapter Ligation

Cleanup

Library Amplification

Cleanup

Enzymatic Frag

Tunable – 60 mins

15 min

25 min

25 min

25 min

Total Time 2.5 hr
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
- Sample-multiplexing
- Single-indexed
- Dual-indexed
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
How to choose the sequencer?!?

**SEQUENCING #4**

---

**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**
- **Channel:** 1-channel
- **SBS:**
- **Max Output:** 1.2 Gb
- **Max Output:** 4 M
- **Max Read Length:** 2x150 bp

**MiniSeq**
- **Channel:** 2-channels
- **SBS:**
- **Max Output:** 8 Gb
- **Max Read Length:** 2x150 bp

**MiSeq**
- **Channel:** 4-channels
- **SBS:**
- **Max Output:** 15 Gb
- **Max Read Length:** 2x300 bp

**NextSeq**
- **Channel:** 2-channels
- **SBS:**
- **Max Output:** 120 Gb
- **Max Read Length:** 2x150 bp

**HiSeq 4000**
- **Channel:** 4-channels
- **SBS:**
- **Max Output:** 1500 Gb
- **Max Read Length:** 2x150 bp

**HiSeq X Ten**
- **Channel:** 4-channels
- **SBS:**
- **Max Output:** 1600 Gb
- **Max Read Length:** 2x150 bp

**NovaSeq**
- **Max Output:** 2400-3000 Gb
- **Max Read Length:** 2x150 bp
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

<table>
<thead>
<tr>
<th>NovaSeq 6000 System</th>
<th>SP</th>
<th>S1</th>
<th>S2</th>
<th>S4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flow Cell Type</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Human Genomes per Run</td>
<td>~4</td>
<td>~8</td>
<td>~20</td>
<td>~48</td>
</tr>
<tr>
<td>Exomes per Run</td>
<td>~40</td>
<td>~80</td>
<td>~200</td>
<td>~500</td>
</tr>
<tr>
<td>Transcriptomes per Run</td>
<td>~32</td>
<td>~64</td>
<td>~164</td>
<td>~400</td>
</tr>
</tbody>
</table>
SEQUENCING #5

Sample Multiplexing - simplified

- Single-End (SE) 50, 100...
- Paired-End (PE) 50, 100...
- Single-indexed
- Dual-indexed

<table>
<thead>
<tr>
<th>Size</th>
<th>Cost</th>
<th>Area</th>
<th>Cost/in²</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small 8”</td>
<td>6.95$</td>
<td>50.2 in²</td>
<td>13.8 ¢</td>
</tr>
<tr>
<td>Medium 14”</td>
<td>15.50$</td>
<td>153.9 in²</td>
<td>10.0 ¢</td>
</tr>
<tr>
<td>Large 18”</td>
<td>18.95$</td>
<td>254.5 in²</td>
<td>7.5 ¢</td>
</tr>
<tr>
<td>X-Large 23”</td>
<td>23.95$</td>
<td>415.5 in²</td>
<td>5.8 ¢</td>
</tr>
</tbody>
</table>

Cost 3.4 x total

Area 8.3 x bigger

Cheaper/in²

POLL 6
SEQUENCING #5

Cost/Human Genome
- 1’150 $
- 890 $
- 845 $
- 520 $

Flow-cell cost @UCLA (not a lane)
- 4’600 $
- 7’100 $
- 16’900 $
- 25’000 $

NovaSeq 6000 System
<table>
<thead>
<tr>
<th>Flow Cell Type</th>
<th>SP</th>
<th>S1</th>
<th>S2</th>
<th>S4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human Genomes per Run</td>
<td>~4</td>
<td>~8</td>
<td>~20</td>
<td>~48</td>
</tr>
<tr>
<td>Exomes per Run</td>
<td>~40</td>
<td>~80</td>
<td>~200</td>
<td>~500</td>
</tr>
<tr>
<td>Transcriptomes per Run</td>
<td>~32</td>
<td>~64</td>
<td>~164</td>
<td>~400</td>
</tr>
</tbody>
</table>

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**
4. **Ligation**

**Cost/Human Genome**
- 1’150 $
- 890 $
- 845 $
- 520 $

**Flow-cell cost @UCLA**
- 4’600 $
- 7’100 $
- 16’900 $
- 25’000 $

**NovaSeq 6000 System**

<table>
<thead>
<tr>
<th>Flow Cell Type</th>
<th>SP</th>
<th>S1</th>
<th>S2</th>
<th>S4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human Genomes per Run</td>
<td>~4</td>
<td>~8</td>
<td>~20</td>
<td>~48</td>
</tr>
<tr>
<td>Exomes per Run</td>
<td>~40</td>
<td>~80</td>
<td>~200</td>
<td>~500</td>
</tr>
<tr>
<td>Transcriptomes per Run</td>
<td>~32</td>
<td>~64</td>
<td>~164</td>
<td>~400</td>
</tr>
</tbody>
</table>

**Output/flow-cell**
- 250 Gb
- 500 Gb
- 1250 Gb
- 3000 Gb
KEEP CALM
IT'S BREAK TIME

QUESTIONs #4
TruSeq Nano vs. WGS #2

**TruSeq Nano**
- Double-stranded DNA
  - Fractionate
  - Size select
  - End repair
  - Phosphorylate
  - A-overhang
  - Add Adaptors
  - Adaptor ligation
  - Denature and amplify
  - Product ready for cluster generation

**WGS #2**
- Nextera Library Preparation
  - Transposase
  - DNA
  - Tagmentation
  - ~300bp
  - Amplification
  - Product ready for cluster generation
Tagmentation = Fragmentation and Adapter Ligation* Simultaneously

VS.

TruSeq Nano

Double-stranded DNA

Fractionate
Size select

End repair
Phosphorylate

Adaptors

Adaptor ligation

Denature and amplify

Product ready for cluster generation

WGS #2

Nextera Library Preparation

Transposase

DNA

~300bp

Adaptor ligation

Adaptors

Adaptor ligation

Amplification

Product ready for cluster generation
Go to Kahoot.it then follow on-screen instruction.
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
# Whole Exome Sequencing (WES)

1-2% of the **Entire Genome**

## Exome-Seq Kits

<table>
<thead>
<tr>
<th>Kit</th>
<th>Targeted Region</th>
<th>Number of Probes</th>
<th>Probe Type</th>
<th>Genomic DNA input required</th>
<th>Adapter addition</th>
<th>Probe Length (mer)</th>
<th>Probe Design</th>
<th>Price per capture (negotiable)</th>
<th>Designed on build</th>
<th>Hybridization time (hours)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agilent SureSelect XT2 V6 Exome</td>
<td>60 Mb</td>
<td>~758,086</td>
<td>biotinylated cRNA baits</td>
<td>100 ng</td>
<td>Ligation</td>
<td>120</td>
<td>Non-overlapping, paired-end reads used to fill gaps</td>
<td>$270</td>
<td>GRCh37 (hg19)</td>
<td>16</td>
</tr>
<tr>
<td>Agilent SureSelect XT2 V5 Exome</td>
<td>51 Mb</td>
<td>~655,872</td>
<td>biotinylated cRNA baits</td>
<td>100 ng</td>
<td>Ligation</td>
<td>120</td>
<td>Non-overlapping, paired-end reads used to fill gaps</td>
<td>$200</td>
<td>GRCh37 (hg19)</td>
<td>16</td>
</tr>
<tr>
<td>IDT xGEN Exome Panel</td>
<td>39 Mb</td>
<td>429,826</td>
<td>biotinylated DNA baits</td>
<td>500 ng</td>
<td>Ligation</td>
<td>not described</td>
<td>Non-overlapping</td>
<td>$250</td>
<td>GRCh37 (hg19)</td>
<td>4</td>
</tr>
<tr>
<td>Illumina Nextera Rapid Capture Expanded Exome</td>
<td>62 Mb</td>
<td>&gt;340,000</td>
<td>biotinylated DNA bait</td>
<td>50 ng</td>
<td>Transposase</td>
<td>95</td>
<td>Non-overlapping (adjacent to each other)</td>
<td>$250</td>
<td>GRCh37 (hg19)</td>
<td>24-48</td>
</tr>
<tr>
<td>Roche Nimblegen SeqCap EZ Exome v3.0</td>
<td>64 Mb</td>
<td>&gt;2,100,000</td>
<td>biotinylated DNA bait</td>
<td>1 ug</td>
<td>Ligation</td>
<td>60 - 90</td>
<td>Overlapping baits</td>
<td>$600</td>
<td>GRCh37 (hg19)</td>
<td>72</td>
</tr>
<tr>
<td>Probe Type</td>
<td>Target Region</td>
<td>Number of Probes</td>
<td>Genomic DNA Input Required</td>
<td>Adapter Addition</td>
<td>Ligation Length (bp)</td>
<td>Probes per Ligation</td>
<td>Price per Capture (negotiable)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------------------</td>
<td>---------------</td>
<td>------------------</td>
<td>-----------------------------</td>
<td>------------------</td>
<td>----------------------</td>
<td>---------------------</td>
<td>-------------------------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Agilent SureSelect XT2 V6 Exome</td>
<td>60 Mb</td>
<td>~758,086</td>
<td>100 ng</td>
<td>Ligation</td>
<td>120</td>
<td>Non-overlapping, paired-end reads used to fill gaps</td>
<td>$270</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Agilent SureSelect XT2 V5 Exome</td>
<td>5 Mb</td>
<td>~397,872</td>
<td>100 ng</td>
<td>Ligation</td>
<td>20</td>
<td>Non-overlapping, paired-end reads used to fill gaps</td>
<td>$300</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IDT xGEN Exome Panel</td>
<td>39 Mb</td>
<td>429,826</td>
<td>500 ng</td>
<td>Ligation</td>
<td>60</td>
<td>Non-overlapping (adjacent to each other)</td>
<td>$250</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Illumina Nextera Rapid Capture Expanded Exome v3.0</td>
<td>6 Mb</td>
<td>~2,106,000</td>
<td>50 ng</td>
<td>Transposase</td>
<td>90</td>
<td>Overlapping baits</td>
<td>$1000</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Agilent’s SureSelect XT2 V6 Exome
- High-density, high-quality probes
- Nonoverlapping, paired-end reads used to fill gaps
- Designed on build
- Hyb dif time (hours) 16

Agilent’s SureSelect XT2 V5 Exome
- High-density, high-quality probes
- Nonoverlapping, paired-end reads used to fill gaps
- Designed on build
- Hyb dif time (hours) 16

IDT xGEN Exome Panel
- High-density, high-quality probes
- Nonoverlapping, paired-end reads used to fill gaps
- Designed on build
- Hyb dif time (hours) 4

Illumina Nextera Rapid Capture Expanded Exome v3.0
- High-density, high-quality probes
- Nonoverlapping, paired-end reads used to fill gaps
- Designed on build
- Hyb dif time (hours) 72
<table>
<thead>
<tr>
<th><strong>Vendor</strong></th>
<th><strong>Exome Panel</strong></th>
<th><strong>Target Region</strong></th>
<th><strong>Number of Probes</strong></th>
<th><strong>Probe Type</strong></th>
<th><strong>Probe Design</strong></th>
<th><strong>Genomic DNA Input</strong></th>
<th><strong>Adapter Addition</strong></th>
<th><strong>Probe Length</strong></th>
<th><strong>Hybridization Time</strong> (Hours)</th>
<th><strong>Price per capture (negotiable)</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Agilent</td>
<td>SureSelect XT2 V6 Exome</td>
<td>51 Mb</td>
<td>~655,872</td>
<td>biotinylated cRNA baits</td>
<td>Ligation</td>
<td>120</td>
<td>Non-overlapping, paired-end reads used to fill gaps</td>
<td>Non-overlapping, paired-end reads</td>
<td>16</td>
<td>$270</td>
</tr>
<tr>
<td>Agilent</td>
<td>SureSelect XT2 V5 Exome</td>
<td>51 Mb</td>
<td>~655,872</td>
<td>biotinylated cRNA baits</td>
<td>Ligation</td>
<td>120</td>
<td>Non-overlapping, paired-end reads used to fill gaps</td>
<td>Non-overlapping, paired-end reads</td>
<td>16</td>
<td>$200</td>
</tr>
<tr>
<td>IDT</td>
<td>xGEN Exome Panel</td>
<td>39 Mb</td>
<td>429,826</td>
<td>biotinylated DNA bait</td>
<td>Ligation</td>
<td>20</td>
<td>not described</td>
<td>Non-overlapping</td>
<td>4</td>
<td>$250</td>
</tr>
<tr>
<td>Illumina</td>
<td>Nextera Rapid Capture Expanded Exome v3.0</td>
<td>64 Mb</td>
<td>&gt;2,100,000</td>
<td>biotinylated DNA bait</td>
<td>Transposase</td>
<td>90</td>
<td>Non-overlapping (adjacent to each other)</td>
<td>Overlapping baits</td>
<td>4</td>
<td>$250</td>
</tr>
<tr>
<td>Roche</td>
<td>Nimblegen SeqCap EZ Exome v3.0</td>
<td>64 Mb</td>
<td>&gt;340,000</td>
<td>biotinylated DNA bait</td>
<td>Ligation</td>
<td>50</td>
<td>Non-overlapping, adjacent to each other</td>
<td>Overlapping baits</td>
<td>18</td>
<td>$600</td>
</tr>
</tbody>
</table>

**Note:** PCR duplicates are shown as a separate section.
HOW DO YOU DETECT PCR DUPLICATES?

ALIGNMENT POSITION

*in silico*

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

<table>
<thead>
<tr>
<th>End 1</th>
<th>End 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Total Number of Molecules = 4

UMI

*in silico* + experimental

(Random or Pre-Designed)

chr1:100'000-103'000
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)

chr1:100’000-103’000

Total Number of Molecules = 6

e.g. NNNNNNNNN

Random
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)

chr1:100’000-103’000

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, 2^{nd})