CS342: Bioinformatics
Assembling a Genome
DNA Sequencing Technologies

- **Sanger Sequencing** (1977): "First Generation Sequencing"
- **1990’s**: Massively Parallel Signature Sequencing at Lynx Therapeutics (only performed in-house)
- **~2007**: "Next Generation Sequencing" – Massively Parallel
- **Today**: “Third Generation Sequencing” – Single Cell Sequencing
Next-Gen Sequencing!

illumina®

NextSeq Series

HiSeq Series

454 Sequencing

Applied Biosystems™ by Life Technologies™
Next-Gen Sequencing

DNA is fragmented and adapters are ligated to both ends.

Fluorescently labeled nucleotides are added.

Fragments hybridize to flow cell and are amplified.

Output is a “Bag of reads.”

Color indicates nucleotide.
Two Different Protocols

**Single Read Sequencing**

Length \( n \) reads taken from one end of a DNA fragment.

- \( n = 7 \)

All output reads have length \( n \)

- \( \text{ACTTCTATCTGATAGTCAATGTAG} \)
- \( \text{TGAAGATAGACTATCAGTTACATC} \)

**Paired-End Sequencing**

Length \( n \) reads taken from both ends of a DNA fragment.

- \( n = 7 \)

All output reads have length \( n \) and are part of a pair of reads

- \( \text{ACAGATC} \)
- \( \text{TATGATC} \)
- \( \text{ATTGATC} \)
- \( \text{CGTGATC} \)
- \( \text{TCCGATC} \)
- \( \text{ATTGATC} \)
- \( \text{TATGATC} \)
- \( \text{ATTGATC} \)

Unknown length

- \( \text{ACTTCTATCTGATAGTCAATGTAG} \)
- \( \text{TGAAGATAGACTATCAGTTACATC} \)
- \( \text{ACAGATC} \)
- \( \text{ATTGATC} \)
- \( \text{CGTGATC} \)
- \( \text{TCCGATC} \)
- \( \text{ATTGATC} \)
- \( \text{TATGATC} \)
- \( \text{TCCGATC} \)
## A Few Examples

<table>
<thead>
<tr>
<th></th>
<th>MiSeq</th>
<th>HiSeq 3000</th>
<th>HigSeq 4000</th>
<th>HiSeq X</th>
<th>NovaSeq</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Max Read Length</strong></td>
<td>2 x 300 bp</td>
<td>2 x 150 bp</td>
<td>2 x 150bp</td>
<td>2 x150bp</td>
<td>2 x 250bp</td>
</tr>
<tr>
<td><strong># Reads per run</strong></td>
<td>1-25 Million</td>
<td>2.1 Million – 5 Billion</td>
<td>Up to 10 Billion</td>
<td>5.3-6 Billion</td>
<td>32-40 Billion</td>
</tr>
<tr>
<td><strong>Run Time:</strong></td>
<td>4-56 hrs</td>
<td>&lt; 1-3.5 days</td>
<td>&lt; 1-3.5 days</td>
<td>&lt; 3 days</td>
<td>13-44 hrs</td>
</tr>
<tr>
<td><strong>Output:</strong></td>
<td>540 Mb - 15Gb</td>
<td>650 – 750 Gb</td>
<td>1300 – 1500 Gb</td>
<td>1.6 – 1.8 Tb</td>
<td>4800 – 6000 Gb</td>
</tr>
</tbody>
</table>

~ 1% error rate

** Numbers updated from: [https://www.illumina.com/systems.html](https://www.illumina.com/systems.html) on 10/6/2019 **
DNA Sequencing Technologies

- Sanger Sequencing “First Generation Sequencing” (1977)
- Massively Parallel Signature Sequencing at Lynx Therapeutics (only performed in-house) (1990’s)
- “Next Generation Sequencing” – Massively Parallel (~2007)
- “Third Generation Sequencing” – Single Cell Sequencing (Today)
Single Molecule Real Time (SMRT)

~ 15% error rate

http://www.nature.com/nbt/journal/v28/n5/fig_tab/nbt0510-426_F1.html
DNA data from Africans reveals sequences that we’d missed

One reference genome doesn't capture the huge variation in human DNA.

CATHLEEN O'GRADY - 11/24/2018, 4:00 PM

Let’s try this out