BI I: Shotgun sequencing/assembly, then & now

November 26, 2013

Dr. Eric Stone
Announcements

• HW #3 due today

• Will be graded by next week
  • Obviously!

• I am (am I?) all that stands between you and Thanksgiving
  • And likewise…

• In advance of that, have a great break!
Before whole-genome alignments…

• How do we get a whole genome?

• Whole-genome alignment =
  • Order one genome with respect to another

• How do we get the order of the first genome?

• This is the problem of sequence assembly
Sequencing and assembly

• DNA sequencing
  • Determining the number and order of nucleotides that make up a given molecule of DNA

• Practical limitations
  • Not possible (yet) to continuously sequence long stretches of DNA
  • Length depends on technology

• DNA assembly
  • Merging these shorter sequences into longer ones
Milestones

• First isolation of DNA (1867; Freidrich Meisher)

• Composition of nucleic acids (1909-1940; Phoebus Levine)

• G=C and A=T however, the G/C and A/T content of different organisms vary (1950; Edwin Chargaff)

• G/C content by annealing (1968; Mandel and Marmur)

• Maxam-Gilbert and Sanger Sequencing (1977)

• Next-Generation Sequencing (2005)
Sanger Sequencing

- "Sanger Sequencing" developed by Fred Sanger et al in the mid 1970’s

- Uses dideoxynucleotides for "chain termination", generating fragments of different lengths ending in ddATP, ddGTP, ddCTP or ddTTP

http://openwetware.org/wiki/BE.109:Bio-material_engineering/Sequence_analysis
Sanger Sequencing Reactions

• For given template DNA, like PCR except
  • Uses only a single primer and polymerase to make new ssDNA pieces
  • Includes regular nucleotides (A, C, G, T) for extension, but also includes dideoxy nucleotides.

Regular Nucleotides

Dideoxy Nucleotides

1. Labeled
2. Terminators
Sanger Sequencing

Primer

5' T G C G C G G C C C A
    A C G C G C C G G G T
Sanger Sequencing

5' Primer
T G C G C G C G G C C C A G T C T T T G G G G C T

5'
Sanger Sequencing

Primer


21 bp
Sanger Sequencing

5' T G C G G C C G G G T C A G A A C C C G A T C G C G 3'
5' C T G C G C G G C C C A

21 bp

5' Primer

3' T G C T T G G G C T A G C G

26 bp

5' T G C G C G C G G C C C A G T C T T G G G C T 3'

5' T G C G C G G C C C A G T C T T G G G C T 3'

5' T G C G C G G C C C A G T C T T G G G C T A G C G 3'

21 bp

26 bp
Sanger Sequencing

5' Primer

T G C G C G G C C C A G

3'

G T C T T G G G C T A G C G
C

5'

T G C G C G G C C C A
G T C T T G G G C T

21 bp

5'

T G C G C G G C C C A G T C T T T G G G C T

26 bp

5'

T G C G C G G C C C A G T C T T T G G G C T A G C G C

22 bp

5'

T G C G C G G C C C A G T C T T T G G G C T A
Sanger Sequencing

Primer

5' T G C G C G G C C C C A T C T T T G G G G G C
3' G T C T T G G G C T A G C G

21 bp

5' T G C G C G G C C C C A G T C T T T G G G G C T
3' G T C T T G G C T A

26 bp

5' T G C G C G G C C C C A G T C T T T G G G G C T A G C G C
3' T C G C G

22 bp

5' T G C G C G G C C C C A G T C T T T G G G C T A G
3' T C G C G

12 bp
Sanger Sequencing


5' T G C G C G C G G C G G C C C A G T C T T T G G G G C T 21 bp

5' T G C G C G C G G C C C A G T C T T T G G G G C T A G C G C 26 bp

5' T G C G C G C G G C C C A G T C T T T G G G G C T A G C G C T A 22 bp

5' T G C G C G C G G C C C A G 12 bp

5' T G C G C G C G G C C C A G T C T T T G G G G C 20 bp
Sanger Sequencing


5' 3'

G T C T T G G G C T A G C G

5'

C

T G C G C G G C C C A

G T C T T G G G C T

21 bp

T G C G C G G C C C A

G T C T T G G G C T A G C G

26 bp

T G C G C G G C C C A

G T C T T G G G C T A G C G

22 bp

T G C G C G G C C C A

G T C T

12 bp

T G C G C G G C C C A

G T C T T G G G C T A G C G

20 bp

T G C G C G G C C C A

G T C T

16 bp
Sanger Sequencing

```


5' T G C G C G G C C C A G 12 bp


```
Sanger Sequencing
Sanger Sequencing Output

- Each sequencing reaction yields a chromatogram
- Practical upper bound of around 1000 bp
What is a base quality?

<table>
<thead>
<tr>
<th>Base Quality</th>
<th>$P_{\text{error}}(\text{obs. base})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>50.12%</td>
</tr>
<tr>
<td>5</td>
<td>31.62%</td>
</tr>
<tr>
<td>10</td>
<td>10.00%</td>
</tr>
<tr>
<td>15</td>
<td>3.16%</td>
</tr>
<tr>
<td>20</td>
<td>1.00%</td>
</tr>
<tr>
<td>25</td>
<td>0.32%</td>
</tr>
<tr>
<td>30</td>
<td>0.10%</td>
</tr>
<tr>
<td>35</td>
<td>0.03%</td>
</tr>
<tr>
<td>40</td>
<td>0.01%</td>
</tr>
</tbody>
</table>
Human Genome Project

- How many bp are in the human genome?
  - ~3 billion (haploid)

- How much $$$ to sequence first human genome?
  - ~$2.7 billion

- How long did it take?
  - ~13 years

- When was it completed?
  - 2000-2003
## Declines in sequencing cost

![Cost per Raw Megabase of DNA Sequence](image)

<table>
<thead>
<tr>
<th>Date</th>
<th>Cost per Mb</th>
<th>Cost per Genome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sep-01</td>
<td>$5,292.39</td>
<td>$95,263,072</td>
</tr>
<tr>
<td>Sep-02</td>
<td>$3,413.80</td>
<td>$61,448,422</td>
</tr>
<tr>
<td>Oct-03</td>
<td>$2,230.98</td>
<td>$40,157,554</td>
</tr>
<tr>
<td>Oct-04</td>
<td>$1,028.85</td>
<td>$18,519,312</td>
</tr>
<tr>
<td>Oct-05</td>
<td>$766.73</td>
<td>$13,801,124</td>
</tr>
<tr>
<td>Oct-06</td>
<td>$581.92</td>
<td>$10,474,556</td>
</tr>
<tr>
<td>Oct-07</td>
<td>$397.09</td>
<td>$7,147,571</td>
</tr>
<tr>
<td>Oct-08</td>
<td>$3.81</td>
<td>$342,502</td>
</tr>
<tr>
<td>Oct-09</td>
<td>$0.78</td>
<td>$70,333</td>
</tr>
<tr>
<td>Oct-10</td>
<td>$0.32</td>
<td>$29,092</td>
</tr>
<tr>
<td>Oct-11</td>
<td>$0.09</td>
<td>$7,743</td>
</tr>
<tr>
<td>Oct-12</td>
<td>$0.07</td>
<td>$6,618</td>
</tr>
<tr>
<td>Jan-13</td>
<td>$0.06</td>
<td>$5,671</td>
</tr>
</tbody>
</table>

### Rise of next-generation sequencing

#### Table 1: (a) Advantage and mechanism of sequencers. (b) Components and cost of sequencers. (c) Application of sequencers.

##### (a)

<table>
<thead>
<tr>
<th>Sequencer</th>
<th>454 GS FLX</th>
<th>HiSeq 2000</th>
<th>SOLiDv4</th>
<th>Sanger 3730xl</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequencing mechanism</td>
<td>Pyrosequencing</td>
<td>Sequencing by synthesis</td>
<td>Ligation and two-base coding</td>
<td>Dideoxy chain termination</td>
</tr>
<tr>
<td>Read length</td>
<td>700 bp</td>
<td>50SE, 50PE, 101PE</td>
<td>50 + 35 bp or 50 + 50 bp</td>
<td>400~900 bp</td>
</tr>
<tr>
<td>Accuracy</td>
<td>99.99%*</td>
<td>98%, (100PE)</td>
<td>99.94% *raw data</td>
<td>99.999%</td>
</tr>
<tr>
<td>Reads</td>
<td>1 M</td>
<td>3 G</td>
<td>1200~1400 M</td>
<td>—</td>
</tr>
<tr>
<td>Output data/run</td>
<td>0.7 Gb</td>
<td>600 Gb</td>
<td>120 Gb</td>
<td>1.9~84 Kb</td>
</tr>
<tr>
<td>Time/run</td>
<td>24 Hours</td>
<td>3~10 Days</td>
<td>7 Days for SE</td>
<td>20 Mins~3 Hours</td>
</tr>
<tr>
<td>Advantage</td>
<td>Read length, fast</td>
<td>High throughput</td>
<td>Accuracy</td>
<td>High quality, long read length</td>
</tr>
<tr>
<td>Disadvantage</td>
<td>Error rate with polybase more than 6, high cost, low throughput</td>
<td>Short read assembly</td>
<td>Short read assembly</td>
<td>High cost low throughput</td>
</tr>
</tbody>
</table>

##### (b)

<table>
<thead>
<tr>
<th>Sequencers</th>
<th>454 GS FLX</th>
<th>HiSeq 2000</th>
<th>SOLiDv4</th>
<th>3730xl</th>
</tr>
</thead>
<tbody>
<tr>
<td>Instrument price</td>
<td>Instrument $500,000, $7000 per run</td>
<td>Instrument $690,000, $6000/(30x) human genome</td>
<td>Instrument $495,000, $15,000/100 Gb</td>
<td>Instrument $95,000, about $4 per 800 bp reaction</td>
</tr>
<tr>
<td>CPU</td>
<td>2* Intel Xeon X5675</td>
<td>2* Intel Xeon X5560</td>
<td>8* processor 2.0 GHz</td>
<td>Pentium IV 3.0 GHz</td>
</tr>
<tr>
<td>Memory</td>
<td>48 GB</td>
<td>48 GB</td>
<td>16 GB</td>
<td>1 GB</td>
</tr>
<tr>
<td>Hard disk</td>
<td>1.1 TB</td>
<td>3 TB</td>
<td>10 TB</td>
<td>280 GB</td>
</tr>
<tr>
<td>Automation in library preparation</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Other required device</td>
<td>REM e system</td>
<td>cBot system</td>
<td>EZ beads system</td>
<td>NO</td>
</tr>
<tr>
<td>Cost/million bases</td>
<td>$10</td>
<td>$0.07</td>
<td>$0.13</td>
<td><strong>$2400</strong></td>
</tr>
</tbody>
</table>
**Ex: next-gen vs. Sanger**

- Consider the domesticated silkworm genome
  - The reference genome is about 432Mb large
  - Assembled from approximately 8.5 fold coverage

<table>
<thead>
<tr>
<th>Platform</th>
<th>ABI3730xl Genome Analyzer</th>
<th>Roche (454) FLX</th>
<th>Illumina Genome Analyzer</th>
<th>ABI SOLiD</th>
<th>Helicos Heliscope</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sequencing Speed</strong></td>
<td>0.03-0.07 Mb/h</td>
<td>13 Mb/h</td>
<td>25 Mb/h</td>
<td>21–28 Mb/h</td>
<td>83 Mb/h</td>
</tr>
<tr>
<td><strong>Time to sequence (days)</strong></td>
<td>2185.7</td>
<td>11.8</td>
<td>6.1</td>
<td>5.5</td>
<td>1.8</td>
</tr>
<tr>
<td>Method</td>
<td>Read Length</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>--------------</td>
<td>-------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sanger</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>454</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Illumina</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ion Torrent</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Method</td>
<td>Read Length</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>--------------</td>
<td>----------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sanger</td>
<td>600-1000 bp</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>454</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Illumina</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ion Torrent</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Method</td>
<td>Read Length</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>--------------</td>
<td>----------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sanger</td>
<td>600-1000 bp</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>454</td>
<td>300-500 bp</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Illumina</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ion Torrent</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Method</td>
<td>Read Length</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---------------</td>
<td>-------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sanger</td>
<td>600-1000 bp</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>454</td>
<td>300-500 bp</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Illumina</td>
<td>~100 bp</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ion Torrent</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Method</td>
<td>Read Length</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---------------</td>
<td>--------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sanger</td>
<td>600-1000 bp</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>454</td>
<td>300-500 bp</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Illumina</td>
<td>~100 bp</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ion Torrent</td>
<td>~200 bp</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**But...**

- Phage Genome: 30,000 to 500,000 bp
- Bacteria: Several million bp
- Human: 3 billion bp
Shotgun genome sequencing

Fragmulated genome opinions

Fragmulated genome opinions
Shotgun genome sequencing

Fragmented genome chunks

NOT REALLY DONE BY DUCK HUNTERS
Hydroshearing, sonication, enzymatic shearing
The basic idea

• Assembling fragments into contigs requires the detection of significant overlapping regions at the ends of each fragment:

  2 contigs _________  

• How much overlap is required?
  • *Hope/assume* fragment is so long that suffix/prefix sequences appear exactly once in the genome
  • Then if we see such a sequence in two fragments, they must overlap
Putting the pieces back together

17 bp

ATTGTTCCCACAGACCG
CGGCGAAGCATTGTTCC
AGCTCGATGCCGGCGAAG
ATGCCGGCGAAGCATTGCT
TAATGCGACCTCGATGCC

66 bp

AGCTCTGTTCCACAGACCG
ACCGTGTTTTCCGACCG
TTTCCGACCGAAATGGC
ACAGACCGTGGTTCCCAGA
TGTTTTCCGACCAGAAAT
CCGACCAGAAATGGCTCC
Putting the pieces back together

Consensus:
TAATGCGACCTCGATGCCC GCCGAAGCATTGTTCCCACAGACCGTG TTTTCCGACCGAAATGGCTCC

ATTGTTCCACAGACCG
CGGCGAAGCATTGTTCCC
ACCGTGGTCTCCGACCG
AGCTCGATGCCC GCCGAAGA
TGTTTCCACAGACCGTG TTTCCCACAGACC AAAATGGCC
ATGCCGGCGAAGCATTGTT
ACAGACC CGTGGTCTCC GCA
TAATGCGACCTCGATGCC AACCGATTGTTCCACAGA
TGTTTCCGACCGAAAT
TGCCGGCGAAGCCTTGT
CGGACCGAATGGCTCC
Putting the pieces back together

Consensus:

```
TAATGCAGACCTCGATGCCGGCGAAGCATTTGTTCCACAGACCGTGTTTTTCCGACCGAAATGGCTCC
```

```
ATTGTTCACCCACAGACCG
CGCGGAAGCATTGTTCC
ACCCTGTGGTTTCCGACCG
AGCTCGATGCCGGCGAAG
TTGTTCACAGACCGTG
TTACGACCGAAATGGCTCC
```

Coverage: # of reads underlying the consensus
Putting the pieces back together

Consensus:
TAATGCAGACCTTGATGCGGGCGCAAGCATTGTTCCCACAGACCGTTGTTTTCCGACCGAAATGGCTCC

ATGTTCCCACAGACCG
CGGGAAGCATTGTTCCACGTGTTTTCCGACCGAAATGGCTCC
AGCTCGATGCGGGCGCAAGTTTCCACAGACCGTGTTTCCGACCGAAATGGCTCC
ATGCCGGCGGCGCAAGCATTGTTTCCGACCGAAATGGCTCC

6x coverage
100% identity

Coverage: # of reads underlying the consensus
Putting the pieces back together

Consensus:

```
TAATGCAGACCTCGATGCGCGGCAACG ACTTTGTTCCACAGACCG TGTTTTCCGACCGAAATGGGCTCC
```

5x coverage

80% identity

**Coverage:** # of reads underlying the consensus
Putting the pieces back together

Consensus:
TAATGCACCTCGATGCCGCGCGAACAGCATTTGTTCACAGACCCTTCCGACGAAATGGCTCC

Consensus:
ATTGTCCCACAGACCG
CGCGAAGCCATTGCTCC
ACCGGATTTTCCGACCG
AGCTCGATGCCGCGCAG
TTGTTCCACAGACCCTG
TTTCCGACGAAATGGCTCC

2x coverage
50% identity

Coverage: # of reads underlying the consensus
Putting the pieces back together

Consensus:
TAATGCGACCTCGATGCCGTCGAAGCATTTGTTCCACAGACCGGTGTTTTCCGACCAGAAATGGCTCC

ATTGTTCCACACAGACCG
CGCGCAAGCATTGTTCC
ACCGTGTTTTCCGACCG
AGCTCGATGCGGCGGAAG
TTGTTCCACAGACCGTG
TTTCCGACCAGAAAATGGA
ATGCCGGCGAAGCATTGTT
ACAGACCAGTGTTCGCCGA
TAATGCGACCTCGATGCC
AAGCATTTGTTCCACAG
TTGTTTCCGACCAGAAAAT
TGCCGGCGAAGCCTTGT
CCGACCAGAAATGGCTCC

1x coverage

Coverage: # of reads underlying the consensus
Parameters governing shotgun assembly

• $G = \text{genome length in nucleotides (e.g. } 3 \times 10^9) \]
• $L = \text{read length in nucleotides (e.g. 500)}$
• $N = \text{number of reads}$
• $NL = \text{total number of nucleotides sequenced}$
• $c = \frac{NL}{G} = \text{coverage (average, e.g. “10x”)}$

1 times coverage of the human genome requires

\[ N = \frac{cG}{L} = \frac{1(3 \times 10^9)}{500} = 6 \text{ million reads} \]

“10× coverage” requires $N = 60$ million reads
Key questions

• Questions:
  • How many contigs are there?
  • How big are the contigs?
  • How many reads are in each contig?
  • How big are the gaps between contigs?

• Mathematical model (Lander-Waterman)
  • Simplest case = reads uniformly distributed
  • Can use basic probability to gain intuition
Developing the intuition

• Suppose k chromosomes

• For each, read can start anywhere (left to right) except last L-1 positions

• So total of \( G - c(L-1) \) starting positions
  - Which is approximately equal to \( G \)

• With approximation, probability that one of the reads starts at any particular nucleotide is \( \frac{N}{G} \)
From nucleotide to interval of length L

• Let S be any specific interval of L consecutive nucleotides
  • What is the probability that at least one read starts in S?

• \( p = \Pr(\text{no reads start in } S) = (1 - \frac{N}{G})^L \) (coins!)

• Recall \( c = \frac{NL}{G} \). So \( p = (1 + \frac{-c}{L})^L \approx e^{-c} \)

• Poisson approximation to Binomial
  • Coverage \( \sim \) Poisson(c)
How much of the genome was sequenced?

• Probability of zero coverage at a site is \( p = e^{-c} \)

• Expected number of nucleotides in “gaps” is \( pG \)
  • Remaining \((1-p)G\) within contigs

• Suppose desire 99% in contigs, 1% in gaps
  • \( p = e^{-c} = .01 \Rightarrow c = -\ln(0.01) = 4.6 \)

• So 4.6x coverage (13.8 billion sequenced nt) yields an expectation of 1% gaps
How many contigs are there?

- Each contig has a unique rightmost read

- What is the probability that a read is rightmost?
  - Equal to prob that no other read starts within it! (so p)

- Label rightmost read “H” and remaining reads “T”
  - Then number of contigs is simply number of “H”

- Follows Binomial(N,p)
  - Expected number of contigs = Np = (cG/L)e^{-c}
  - So number of contigs (and gaps) decreases exp with c
How many reads per contig?

• The number of reads is the same as the position of the first “H”

• This is a geometric distribution!

• Expected number of reads per contig = $1/p = e^c$

• Equivalently just the number of reads divided by the number of contigs
How long are the contigs?

• The expected size of sequenced region is \((1-p)G\)

• The expected number of contigs is \(Np\)

• Mean contig size is the ratio \((1-p)G/(Np)\)

• In terms of coverage, \((e^c - 1)L/c\)

• So contig length grows exponentially in \(c\)
### Example for haploid human genome

<table>
<thead>
<tr>
<th>Coverage</th>
<th># reads N =</th>
<th># nuc. read</th>
<th>% genome sequenced</th>
<th>mean # contigs</th>
<th>mean contig length</th>
<th>mean # reads/contig</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>6M</td>
<td>3.0B</td>
<td>63.21%</td>
<td>2207276.6</td>
<td>859.1</td>
<td>2.7</td>
</tr>
<tr>
<td>1.5</td>
<td>9M</td>
<td>4.5B</td>
<td>77.69%</td>
<td>2008171.4</td>
<td>1160.6</td>
<td>4.5</td>
</tr>
<tr>
<td>2.0</td>
<td>12M</td>
<td>6.0B</td>
<td>86.47%</td>
<td>1624023.4</td>
<td>1597.3</td>
<td>7.4</td>
</tr>
<tr>
<td>3.0</td>
<td>18M</td>
<td>9.0B</td>
<td>95.02%</td>
<td>896167.2</td>
<td>3180.9</td>
<td>20.1</td>
</tr>
<tr>
<td>4.0</td>
<td>24M</td>
<td>12.0B</td>
<td>98.17%</td>
<td>439575.3</td>
<td>6699.8</td>
<td>54.6</td>
</tr>
<tr>
<td>5.0</td>
<td>30M</td>
<td>15.0B</td>
<td>99.33%</td>
<td>202138.4</td>
<td>14741.3</td>
<td>148.4</td>
</tr>
<tr>
<td>6.0</td>
<td>36M</td>
<td>18.0B</td>
<td>99.75%</td>
<td>89235.1</td>
<td>33535.7</td>
<td>403.4</td>
</tr>
<tr>
<td>7.0</td>
<td>42M</td>
<td>21.0B</td>
<td>99.91%</td>
<td>38299.0</td>
<td>78259.5</td>
<td>1096.6</td>
</tr>
<tr>
<td>8.0</td>
<td>48M</td>
<td>24.0B</td>
<td>99.97%</td>
<td>16102.2</td>
<td>186247.4</td>
<td>2981.0</td>
</tr>
<tr>
<td>9.0</td>
<td>54M</td>
<td>27.0B</td>
<td>99.99%</td>
<td>6664.1</td>
<td>450115.8</td>
<td>8103.1</td>
</tr>
<tr>
<td>10.0</td>
<td>60M</td>
<td>30.0B</td>
<td>100.00%</td>
<td>2724.0</td>
<td>1101273.3</td>
<td>22026.5</td>
</tr>
<tr>
<td>11.0</td>
<td>66M</td>
<td>33.0B</td>
<td>100.00%</td>
<td>1102.3</td>
<td>2721506.4</td>
<td>59874.1</td>
</tr>
<tr>
<td>12.0</td>
<td>72M</td>
<td>36.0B</td>
<td>100.00%</td>
<td>442.4</td>
<td>6781408.0</td>
<td>162754.8</td>
</tr>
<tr>
<td>13.0</td>
<td>78M</td>
<td>39.0B</td>
<td>100.00%</td>
<td>176.3</td>
<td>17015861.2</td>
<td>442413.4</td>
</tr>
<tr>
<td>14.0</td>
<td>84M</td>
<td>42.0B</td>
<td>100.00%</td>
<td>69.8</td>
<td>42950117.3</td>
<td>1202604.3</td>
</tr>
<tr>
<td>15.0</td>
<td>90M</td>
<td>45.0B</td>
<td>100.00%</td>
<td>27.5</td>
<td>108967212.4</td>
<td>3269017.4</td>
</tr>
<tr>
<td>16.0</td>
<td>96M</td>
<td>48.0B</td>
<td>100.00%</td>
<td>10.8</td>
<td>277690922.5</td>
<td>8886110.5</td>
</tr>
<tr>
<td>17.0</td>
<td>102M</td>
<td>51.0B</td>
<td>100.00%</td>
<td>4.2</td>
<td>710439757.5</td>
<td>24154952.8</td>
</tr>
<tr>
<td>18.0</td>
<td>108M</td>
<td>54.0B</td>
<td>100.00%</td>
<td>1.6</td>
<td>1823888003.8</td>
<td>65659969.1</td>
</tr>
<tr>
<td>18.5</td>
<td>111M</td>
<td>55.5B</td>
<td>100.00%</td>
<td>1.0</td>
<td>2925810452.7</td>
<td>108254987.8</td>
</tr>
</tbody>
</table>
Problems with the simple model

- **Sequencing errors:** Most sequencers have around 1% error in the best case.

- **Bias in the reads:** Not all genomic regions covered equally (due both to sampling and cloning bias).

- **Unknown orientation:** Could have sequenced either strand, so must consider read and complement.

- **Repeats:** There is much repetitive sequence, especially in human and higher plants. This is the most important problem in most cases.
Repeat problem

- How can we tell whether an overlap is due to a real match or a match to a repeating sequence somewhere?

- Repeats can be characterized by length, copy number & fidelity between copies
  - Human T-cell receptor: 5x of a 4kb gene w/ ~3% variation
  - ALUs. ~300bp w/5-15% variation, clustering to be 50-60% of many human sequence regions
  - Microsatellites, 3-6bp with thousands of repeats in centromeric and telomeric regions, 1-2% variation.
Repeat example 1

- Repeats at read ends can be assembled in multiple ways

```
TCTTG GTCAT GTCAT
GTCAT GTCAT ACGTC
ACGTC GTCAT GTCAT
GTCAT GTCAT TGGTCCC
```

or

```
TCTTG GTCAT GTCAT
GTCAT GTCAT TGGTCCC
```

```
ACGTC GTCAT GTCAT
GTCAT GTCAT ACGTC
```

correct

incorrect
Repeat example 2

- How long is a segment of short repeats?

TCGATCGATCGATCGATCGATCGATCGATCGATCGATCGATCGATCGATCGATC
TCGATCGATCGATCGATCGATCGATCGATCGATCGATCGATCGATCGATCGATC
TCGATCGATCGATCGATCGATCGATCGATCGATCGATCGATCGATCGATCGATC
TCGATCGATCGATCGATCGATCGATCGATCGATCGATCGATCGATCGATCGATC
Repeat example 3

- Overcollapsing. Shortest common superstring is not always correct.
How do we avoid assembly errors?

• More reads?
  • Increase chance of substantial overlap

• Longer reads?
  • Increases chance of spanning a repeat unit
## Comparing different technologies

### Sanger Sequencing

<table>
<thead>
<tr>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lowest error rate</td>
<td>High cost per base</td>
</tr>
<tr>
<td>Long read length (~750 bp)</td>
<td>Long time to generate data</td>
</tr>
<tr>
<td>Can target a primer</td>
<td>Need for cloning</td>
</tr>
<tr>
<td></td>
<td>Amount of data per run</td>
</tr>
</tbody>
</table>
## Comparing different technologies

### 454 Sequencing

<table>
<thead>
<tr>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low error rate</td>
<td>Relatively high cost per base</td>
</tr>
<tr>
<td>Medium read length (~400-600 bp)</td>
<td>Must run at large scale</td>
</tr>
<tr>
<td></td>
<td>Medium/high startup costs</td>
</tr>
</tbody>
</table>
## Comparing different technologies

### Ion Torrent Sequencing

<table>
<thead>
<tr>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low startup costs</td>
<td>New, developing technology</td>
</tr>
<tr>
<td>Scalable (10 – 1000 Mb of data per run)</td>
<td>Cost not as low as Illumina</td>
</tr>
<tr>
<td>Medium/low cost per base</td>
<td>Read lengths only ~100-200 bp so far</td>
</tr>
<tr>
<td>Low error rate</td>
<td></td>
</tr>
<tr>
<td>Fast runs (&lt;3 hours)</td>
<td></td>
</tr>
</tbody>
</table>
## Comparing different technologies

**Illumina Sequencing**

<table>
<thead>
<tr>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low error rate</td>
<td>Must run at very large scale</td>
</tr>
<tr>
<td>Lowest cost per base</td>
<td>Short read length (50-75 bp)</td>
</tr>
<tr>
<td>Tons of data</td>
<td>Runs take multiple days</td>
</tr>
<tr>
<td></td>
<td>High startup costs</td>
</tr>
<tr>
<td></td>
<td>De Novo assembly difficult</td>
</tr>
</tbody>
</table>
## PacBio Sequencing

<table>
<thead>
<tr>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Can use single molecule as template</td>
<td>High error rate (~10-15%)</td>
</tr>
<tr>
<td>Potential for very long reads (several kb+)</td>
<td>Medium/high cost per base</td>
</tr>
<tr>
<td></td>
<td>High startup costs</td>
</tr>
</tbody>
</table>
How do we avoid assembly errors?

• More reads?
  • Increase chance of substantial overlap

• Longer reads?
  • Increases chance of spanning a repeat unit

• Break up the genome (BACs, Fosmids, etc)

• Leverage additional information (e.g. linkage maps)

• Other “tricks”
Human Genome Project revisited

- **Public:** Lander et al, *Nature*, 2/15/01
  - A consortium of government labs and universities
  - Used BAC-by-BAC sequencing

- **Celera:** Venter et al, *Science*, 2/16/01
  - Celera = private company
  - Used whole-genome random shotgun sequencing
  - Many skeptics due to large coverage required and the large number of repeats that make it impossible to detect overlaps correctly
Sequencing Types

Single Read

Paired-end read

Mate-pair read
Alternatives to assembly

• Thus far discussion has focused on one type of assembly (*de novo* assembly)

• Can alternatively use homology
  • Just like in the rest of bioinformatics

• Requires the use of a reference genome

• Align reads to reference to create “reference-guided assembly”