(Shotgun) sequencing

Titus Brown
Three basic problems

Resequencing, counting, and assembly.
1. Resequencing analysis

We know a reference genome, and want to find variants (blue) in a background of errors (red)
2. Counting

We have a reference genome (or gene set) and want to know how *much* we have. Think gene expression/microarrays.
3. Assembly

We don’t have a genome or any reference, and we want to construct one.
(This is how all new genomes are sequenced.)
Outline

• Shotgun sequencing
• The magic of polonies, and how Illumina sequencing works
• Sequencing depth, read length, and coverage
• Paired-end sequencing and insert sizes
• Coverage bias
• Long reads: PacBio and Nanopore sequencing
Shotgun sequencing

It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness

It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness
Ion Torrent

1e8 wells
Each one is a mini pH meter

Add A
Did H get released for this well?

L Yes? Then next base was A.

~ 6 hrs for sample prep plus run => data
$500 or so.
Two specific concepts:

• First, sequencing everything at random is very much easier than sequencing a specific gene region. (For example, it will soon be easier and cheaper to shotgun-sequence all of *E. coli* then it is to get a single good plasmid sequence.)

• Second, if you are sequencing on a 2-D substrate (wells, or surfaces, or whatnot) then any increase in density (smaller wells, or better imaging) leads to a squared increase in the number of sequences yielded.
Random sampling => deep sampling needed

Typically 10-100x needed for robust recovery (300 Gbp for human)
“Coverage” is simply the average number of reads that overlap each true base in genome.

Here, the coverage is ~10 – just draw a line straight down from the top through all of the reads.
Illumina yields the *deepest* sequencing available

- **MiSeq**
  - 30 million reads per run
  - 300 base paired-end reads

- **HiSeq 2500 RR/X 10**
  - 6 billion reads per run
  - 150 base paired-end reads

- **PacBio**
  - 44,000 reads per run
  - 8500 bp in length

Illumina basics


Bridge amplification and Sequencing-by-synthesis

http://ted.bti.cornell.edu/cgi-bin/epigenome/method-1.cgi
A movie of Illumina sequencing:

https://www.youtube.com/watch?v=tuD-ST5B3QA#t=61
What goes wrong with basic assumptions?

- Not all sequence is as easily sequenced as other, depending on your sequencing technology (e.g. GC/AT bias);

- Some RNA not be as accessible as others (secondary structure);
FASTQ

@895:1:1:1246:14654/1
CAGGCGCCACCACCGTGCCCTCCAACCTGATGGT
+
][aaX___aa[`ZUZ[NONNFNNNNNO____^RQ_
@895:1:1:1246:14654/2
ACTGGGGCTAGACGGTGTCCTCATCGGCACCAGC
+
\UJUWSSV[JQQWNP]]SZZWU^]ZX][^TXR`
@895:1:1:1252:19493/1
CCGGCGTGGTTGGTGAGGTCACTGAGCTTCCATGTC
+
OOOKONNNNN__`R]O[TGTRSY[IUZ]]]___X___
Read length and reconstructability

**Figure 3.** Percentage of the *E.coli* genome covered by contigs greater than a threshold length as a function of read length.

Whiteford et al., Nuc. Acid Res, 2005
“Reconstructability”

• Assembling new genomes or transcriptomes...

• *Haplotyping* - think human genetics & viruses, both.
Repeats! (and shared exons)
Longer reads ... OR ...
Paired-end/mate pair sequencing

A  R  B

longer reads

C  R  D

paired ends
Paired-end sequencing

http://vallandingham.me/RNA_seq_differential_expression.html
Genomic DNA

1. Fragment (200–500bp)
2. Ligate Adaptors
3. Generate Clusters
4. Sequence First End
5. Regenerate Clusters and Sequence Paired End
Mate Pair library preparation is designed to generate short fragments that consist of two segments that originally had a separation of several kilobases in the genome. Fragments of sample genomic DNA are end-biotinylated to tag the eventual mate pair segments. Self-circularization and refragmentation of these large fragments generates a population of small fragments, some of which contain both mate pair segments with no intervening sequence. These Mate Pair fragments are enriched using their biotin tag. Mate Pairs are sequenced using a similar two-adapter strategy as described for paired-end sequencing.

Mate-pair sequencing (long insert)
Longer reads

- PacBio
- Moleculo
- Nanopore
Next-gen sequencing: Pacific Biosciences

1. DNA polymerase wrapped around DNA chain

2. Phospholinked nucleotides

3a. Milliseconds

3b. Microseconds

Phospholinked nucleotide binds, fluoresces and detaches as nucleotide base is read

http://www.melanieswan.com/FOLS.html
Moleculo (Illumina)

A SOLID STATE NANOPORE ARTICULATED WITH PROBES

http://labs.mcb.harvard.edu/branton/projects-NanoporeSequencing.htm
Actual yields

- **MiSeq**
  - 30 million reads per run
  - 300 base paired-end reads

- **HiSeq 2500 RR/X 10**
  - 6 billion reads per run
  - 150 base paired-end reads

- **PacBio**
  - 44,000 reads per run
  - 8500 bp in length

Your basic data (FASTQ)

- @895:1:1:1246:14654/1
- CAGGCGCCCCACCCACCGTGCCCTCCAACCTGATGGT
- +
- ][aaX__aa[\`ZUZ\[NONNFNNNNNO_____\^RQ_]
- @895:1:1:1246:14654/2
- ACTGGGCCTAGACGGTGTCCTCATCGGCACCAGC
- +
- \UJUWSSV[JQQWN][SZZWU\]ZX][\^TXR`
- @895:1:1:1252:19493/1
- CCGGCGTGTTGTTGGTGAGGTCACTGAGCCTTCATGTC
- +
- OOOKONNNNNN__`R][O[TGTRSY[\[IUZ]][\]X__
Mapping

- Many fast & efficient computational solutions exist.
- You have to figure out how to choose parameters to maximize sensitivity/specificity, and when to validate.

U. Colorado
http://genomics-course.jasondk.org/?p=395
Assembly

Reassemble random fragments computationally.

UMD assembly primer (cbcb.umd.edu)
Shotgun sequencing

It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness.
Where does # of reads count?

Resequencing, counting, and assembly.
Where does reconstructability matter?

Resequencing, counting, and assembly.
Summary

- Coverage matters for SNP calls and assembly;
- # of reads matters for counting;
- Length of reads matters for reconstructability (assembly & haplotyping);
- Illumina is still “best” for high coverage;
- PacBio and Molecule => genome assembly;
- Nanopore: still tricky but lots of progress being made.
Bad data

I asked:
https://twitter.com/ctitusbrown/status/624721875252420608

I received:

• http://www.bioinfo-core.org/index.php/Interesting_NGS_failures
• http://bioinfo-core.org/index.php/9th_Discussion-28_October_2010
• https://biomickwatson.wordpress.com/2013/01/21/ten-things-to-consider-when-choosing-an-ngs-supplier/
Sequencing Bloopers

Simon Andrews
Tim Stevens
Technical sequencer problems
Manifold burst in cycle 26
Specific cycles lost
No priming /signal
(Wrong adapters used)

Read 1

Read 2
(barcode)
Tile Problems - Overclustering
Tile Problems – Consistent tile fail
Tile problems – transient tile fail
Incorrect Phred Scores

“the NCBI SRA makes all its data available as standard Sanger FASTQ files (even if originally from a Solexa/Illumina machine)”

Found LOTS of examples of this in the SRA
Data Extraction
Wrong barcode annotation

![Bar chart showing expected and not expected barcode annotations](chart.png)
Contaminated Barcode Stocks

Barcode Frequency

<table>
<thead>
<tr>
<th>Barcode</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>GCCAAT</td>
<td>20</td>
</tr>
<tr>
<td>ACTTGA</td>
<td>18</td>
</tr>
<tr>
<td>CAGATC</td>
<td>16</td>
</tr>
<tr>
<td>TTAGGC</td>
<td>15</td>
</tr>
<tr>
<td>CTTGTA</td>
<td>14</td>
</tr>
<tr>
<td>ACAGTG</td>
<td>12</td>
</tr>
<tr>
<td>GATCAG</td>
<td>4</td>
</tr>
<tr>
<td>TTAAGC</td>
<td>1</td>
</tr>
<tr>
<td>CGGATC</td>
<td>0</td>
</tr>
</tbody>
</table>

- **Expected barcode**
- **Not expected barcode**
Odd sequence composition
Read through adapter
Adapter dimer overload

<table>
<thead>
<tr>
<th>Sequence</th>
<th>%</th>
<th>Possible Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCTAAGGAGATCGGAAGAGCGTCGTGTAAGGGAAAGGTAGTAGTCGATCTCGGTGTCGCCGTATCTATTAAAAA</td>
<td>9.42</td>
<td>Illumina Single End PCR Primer 1</td>
</tr>
<tr>
<td>TCAATGAAAGATCGGAAGAGCGTCGTGTAAGGGAAAGGTAGTAGTCGATCTCGGTGTCGCCGTATCTATTAAAAA</td>
<td>7.30</td>
<td>Illumina Single End PCR Primer 1</td>
</tr>
<tr>
<td>GAGACTCAGATCGGAAGAGCGTCGTGTAAGGGAAAGGTAGTAGTCGATCTCGGTGTCGCCGTATCTATTAAAAA</td>
<td>5.65</td>
<td>Illumina Single End PCR Primer 1</td>
</tr>
</tbody>
</table>

(Single-cell Hi-C)
Positional Sequence Bias Application Specific – BS-Seq
Positional Sequence Biases Expected - RRBS

Also reports of a ‘Chinese CRO’ whose RRBS libraries have the MspI sites missing due to their proprietary and unexplained pre-processing
Positional Sequence Biases
Unavoidable – RNA-Seq

Sequence content across all bases

Position in read (bp)
Positional Sequence Biases
Unexpected – Doubled Adapters
Overrepresented Individual Sequences

• Adapter dimers
• rRNA
• Satellite sequences
My data doesn’t map well...
Contaminated with guessable sequence

www.bioinformatics.babraham.ac.uk/projects/fastq_screen
Contaminated with guessable sequence

Yield (Gbases): 8.11
Sequences: 162,116,141
Sampled: 100,000

<table>
<thead>
<tr>
<th>Reference ID</th>
<th>Species/Reference Genome</th>
<th>Aligned</th>
<th>Aligned %</th>
<th>Error rate</th>
<th>Assigned</th>
<th>Assigned %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hsa.GRCh37</td>
<td>Homo sapiens (human)</td>
<td>91337</td>
<td>91.3%</td>
<td>0.28%</td>
<td>91337</td>
<td>91.3%</td>
</tr>
<tr>
<td>phix174</td>
<td>Phi X 174</td>
<td>4483</td>
<td>4.5%</td>
<td>0.15%</td>
<td>4483</td>
<td>4.5%</td>
</tr>
<tr>
<td>Ptc.CHIMP2</td>
<td>Pan troglodytes (chimpanzee)</td>
<td>86152</td>
<td>86.2%</td>
<td>1.21%</td>
<td>641</td>
<td>0.6%</td>
</tr>
<tr>
<td>Ggo.gorGor3</td>
<td>Gorilla gorilla</td>
<td>83354</td>
<td>83.4%</td>
<td>1.47%</td>
<td>104</td>
<td>0.1%</td>
</tr>
<tr>
<td>fungi.RefSeq</td>
<td>Fungi</td>
<td>3092</td>
<td>3.1%</td>
<td>4.17%</td>
<td>33</td>
<td>0.0%</td>
</tr>
<tr>
<td>Nle.Nleu1</td>
<td>Nomascus leucogenys (northern white cheeked gibbon)</td>
<td>67793</td>
<td>67.8%</td>
<td>2.40%</td>
<td>27</td>
<td>0.0%</td>
</tr>
<tr>
<td>Cja.calJac1</td>
<td>Callithrix jacchus (marmoset)</td>
<td>29717</td>
<td>29.7%</td>
<td>3.08%</td>
<td>7</td>
<td>0.0%</td>
</tr>
<tr>
<td>Mml.MMUL1</td>
<td>Macaca mulatta (mecaque)</td>
<td>52073</td>
<td>52.1%</td>
<td>2.90%</td>
<td>6</td>
<td>0.0%</td>
</tr>
<tr>
<td>Hsa.NCBI36</td>
<td>Hsa.NCBI36</td>
<td>91106</td>
<td>91.1%</td>
<td>0.28%</td>
<td>5</td>
<td>0.0%</td>
</tr>
<tr>
<td>Cfa.BROAD2</td>
<td>Canis familiaris (dog)</td>
<td>3093</td>
<td>3.1%</td>
<td>3.95%</td>
<td>1</td>
<td>0.0%</td>
</tr>
<tr>
<td>Xtr.GGI4_1</td>
<td>Xenopus tropicalis (Western clawed frog)</td>
<td>4796</td>
<td>4.8%</td>
<td>3.81%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hsa.GRCh37.assembled</td>
<td>Hsa.GRCh37.assembled</td>
<td>91065</td>
<td>91.1%</td>
<td>0.29%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hsa.NCBI36.assembled</td>
<td>Hsa.NCBI36.assembled</td>
<td>90978</td>
<td>91.0%</td>
<td>0.29%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>20 others</td>
<td>126</td>
<td>0.1%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unmapped</td>
<td></td>
<td>3230</td>
<td>3.2%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adapter</td>
<td></td>
<td>0</td>
<td>0.0%</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

CRUK Multi-genome alignment system (MGA)
Contamination with unguessable sequence

![](chart.png)

>AF431889 AF431889.1 **Acinetobacter lwoffii** type II modification

Query: 1  cggtgacaggcattgaatgatatttttaaaggttgtttaagaaacactggccgctt 60

Sbjct: 4661 cggtgacaggcattgaatgatatttttaaaggttgtttaagaaacactggccgctt 4720

>GQ352402 GQ352402.1 **Acinetobacter baumannii** strain AbSK-17 plasmid

Query: 1  ggtgacagtggttcatgttttaaaggttgtttaagaaacactggccgctt 55

Sbjct: 8213 ggtgacagtggttcatgttttaaaggttgtttaagaaacactggccgctt 8159

>AF431889 AF431889.1 **Acinetobacter lwoffii** type II modification

Query: 1  acttgctgcgattaaagcagaaaaaacacttgctgaattgagtgct 46

Sbjct: 4484 acttgctgcgattaaagcagaaaaaacacttgctgaattgagtgct 4529
TAGC Plots

Assemble

Filter contigs

Plot %GC vs Coverage

Sample and blast

https://github.com/blaxterlab/blobology
Reagent contamination

Molbio grade water is not the same as DNA free water – heat treated but DNA survives
Later this week --

Many different approaches to evaluating quality/mismatches:

1. Quality-score based (FastQC etc)
2. Composition based (FastQC etc)
3. Reference based (“I know what the answer should look like”)
4. Assembly-graph / k-mer based
Reference & quality-score independent approaches (k-mers)

Zhang et al., https://peerj.com/preprints/890/
...from a well known data set...

Zhang et al., https://peerj.com/preprints/890/