Cleaning Illumina reads

Torsten Seemann

ARC CoE in Coral Reef Studies - Magnetic Island - 7 Mar 2011
Outline

1. About the VBC
2. Motivation for cleaning
3. Cleaning steps
4. Let's clean a read together!
5. Results
6. Conclusions
Victorian Bioinformatics Consortium

- Monash University
  - Faculty of Medicine, Nursing & Health Sciences
    - School of Biomedical Sciences
      - Victorian Bioinformatics Consortium (VBC)

- Bacterial pathogenomics
  - antibiotic resistance (hospital and community)
  - large comparative genomics (100s of strains)
  - software tools for high throughput sequencing

- Worked on assembly and annotation of first bacteria fully sequenced in Australia - *Leptospira borgpetersenii*
Where is the VBC?

VBC in Melbourne

You are here
VBC Collaborators

- Monash University
  - ARC CoE - Structural & Functional Microbial Genomics
  - Depts: Microbiology, Physiology, Computer Science, ...

- National
  - Universities: UniMelb, UWA, U.Syd, UQ, IMB
  - CSIRO: AAHL, FNS, Livestock Industries, Dairy CRC
  - ARC CoE - Coral Reef Studies / JCU

- International
  - USDA, TIGR (now JCVI), Pasteur Institut, EBI/EMBL
  - Universities: Copenhagen, UC Davis, UCSD

- Consulting
  - Biota, Merck, DPI Victoria
Illumina short reads

- **Length**
  - 35 to 150bp, typically 100bp today
- **Attributes**
  - High quality at 5' start, lowers toward 3' end
  - Indels & homopolymer run errors are rare
- **"Single end"**
  - Just a shotgun read sequenced from one end
- **"Paired end"**
  - Typically 250-500bp fragments sequenced at both ends
  - Very reliable
- **"Mate pair"**
  - Circularized 2-10 kbp fragments, paired sequencing
  - Variable reliability
Why clean reads?

● Erroneous data may cause software to:
  ○ run more slowly
  ○ use more RAM
  ○ produce poor / biased / incorrect results

● Cleaning can:
  ○ improve overall average quality of the reads
    ■ hopefully giving a better result
  ○ reduce the volume of reads
    ■ some algorithms are $O(N \cdot \log N)$ or $O(N^2)$
    ■ enable processing when otherwise couldn't

● (some software does handle them appropriately)
The FASTQ format

Combines the sequence and quality into a 4 line record:

@HWUSI-EAS-100R:6:1:9646:1115#GATCAG/1
GGACCTGAGAGTGTGCATGAAGAGGGCAGCCCTCGCGCACCGCTG
+
HWUSI-EAS-100R:6:1:9646:1115#GATCAG/1
ccf^_cdf_d^dddddafaaf\^a_a_fff]dd[dya^] ]daBBBB

1. @ Machine : Lane : Tile : X : Y # Mux / Direction
2. DNA sequence
3. + [ copy of 1. ]
4. Quality string (encoded, see next slide)
FASTQ quality string

- Encodes Phred qualities (Q) between 0 and 40
  - Q = 10 \log_{10} p \quad (higher \ is \ better)
  - p = estimated probability that the base call is incorrect

- Uses 41 "readable" characters
  - ASCII 64 '@' to 104 'h'
  - "B"=BAD "s"=satisfactory "g"=good "h"=high
  - Beware there are 3 other alternate encodings :-(
Ambiguous bases

● If there is ambiguity in the base call, an "N" is used

@ILLUMINA:6:1:964:115#GATCAG/1
GGACCTGAGAGTGTGCATGAAGGGCAGCGCGCACNGCA
+
ccf^_cdf_d^ddddddfaaf\^a_a_fff]dd[dyPFBBB

● Possible software responses:
  ○ Crash!
  ○ Ignore it
  ○ Silently convert to fixed or random base (Velvet)
  ○ Handle it appropriately

● Small proportion overall, safer to discard
Homopolymers

- A read consisting of all the same base

@ILLUMINA:6:1:964:115#GATCAG/1
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
+
ccf^_cdf_d^ddddfafaaf\^a_a_ff]d[dYPFDEDCBBBBB

- Often occur from clusters at edge of flowcell lane
- Early Illumina software called 'blank' as 'A'
- Unlikely to be present in real DNA
- Best to discard
Quality trimming

● Remove low quality sequence
  ○ Q=13 corresponds to 5% error (p=0.05)
  ○ Q=0..13 encoded by @ABCDEFGHIJKLMNOPQRSTUVWXYZ

@ILLUMINA:6:1:9646:1115#GATCAG/1
GACCTGAGAGTGTGCAAGAGAGCCAGCCCAGCCACTGCATG
+
cdf^_cdf_d^dddddfaf\^a_a_fff]ddPFDEDCBBBBBB

● Can trim per
  ○ each base
  ○ window moving average eg. 3 base mean
  ○ minimum % good per window eg. need 4 of 5
Illumina Adaptors

- Used in the sequencing chemistry
- Can appear at ends of read sequences
- Worse for mate-pair than for paired-end reads

- PCR Primer
  CAAGCAGAAGACGGCATACGAGCTCTTCCGATCT

- Genomic DNA Sequencing Primer
  CACTCTTTTCCCTACACGACGCTCTTCCGATCT

- TruSeq Universal Adaptor (newest chemistry)
  "CENSORED"
Adaptor clipping

● Method
  ○ Align 3' and 5' read end against all adaptor sequences
  ○ If there is an anchored "match", trim the read

● Minimum length of match?
  ○ want to remove adaptor, but not real sequence [10 bp]

● Allow substitutions in match?
  ○ as reads have errors, need some tolerance [1 sub]

● Allow gaps/indels in match?
  ○ indels are unlikely in Illumina reads [no]

● Slow to perform compared to other pre-processing steps
Decloning

- Illumina "mate pair" sequencing
  - Requires a lot of starting DNA
  - Challenging protocol to implement reliably
  - Not enough final DNA leads to PCR clones
  - Coverage is highly non-uniform and sporadic
  - Causes bias in analyses

- Decloning
  - Replace clones with a single representative
  - Choose representative with highest quality
  - Helps salvage usable information content
  - Implemented by Sylvain Foret
Read length

● Enforce a minimum read length $L$

● Choice is dependent on software
  ○ Short read assemblers eg, Velvet
    ■ Break reads into k-mers, so $L < k$ is pointless
  ○ Aligning reads to reference eg. BWA, Maq
    ■ Desire reasonable uniqueness of sequence
    ■ $L=24+$ is bare minimum
Walk-through

1. Original read + quality  = 43bp
   GTTAGCGCGCTGACCATGATTCAAGGAAC TTGGCC CCATTNATA
   hhhhhghfeefaa^a^[ [ [^X[ [XX^^^` SSTQPZZBBBBBBB
2. Homopolymer? No
   GTTAGCGCGCTGACCATGATTCAAGGAAC TTGGCC CCATTNATA
3. Ambiguous N bases? Yes, 1
   GTTAGCGCGCTGACCATGATTCAAGGAAC TTGGCC CCATTNATA
4. Quality < 20 ? Yes,  at 3’ end
   GTTAGCGCGCTGACCATGATTCAAGGAAC TTGGCC CCATTNATA
5. Adaptor sequences > 8bp ? Yes, 9 bp at 5’ end
   GTTAGCGCGCTGACCATGATTCAAGGAAC TTGGCC CCATTNATA
6. Combine all masks  Logical intersection
   GTTAGCGCGCTGACCATGATTCAAGGAAC TTGGCC CCATTNATA
7. Extract longest sub-sequence  = 19bp
   TGACCATGATTCAAGGAAC
Example

- **Raw data** (*A. millepora* Illumina)
  - 9 libraries - 3 x PE, 6 x MP - 200bp to 10kbp
  - 92.0 Gbp, 943M reads, average length 98bp

- **Method**
  - Decloned all MP libs, disallow Ns, reject homopolymers, trim Q < 20 + clip adaptors, minimum length 55bp

- **Cleaned data**
  - 42.5 Gbp, 478M reads, average length 88bp

- **Effect**
  - Good - *de novo* Velvet assembly improved overall
  - Bad - lower coverage
### Per library yields (Gbp)

<table>
<thead>
<tr>
<th>Library</th>
<th>Raw</th>
<th>Cleaned</th>
<th>%Kept</th>
</tr>
</thead>
<tbody>
<tr>
<td>pe_193</td>
<td>9.55</td>
<td>6.71</td>
<td>70</td>
</tr>
<tr>
<td>pe_463</td>
<td>19.19</td>
<td>13.89</td>
<td>72</td>
</tr>
<tr>
<td>pe_580</td>
<td>4.87</td>
<td>3.18</td>
<td>65</td>
</tr>
<tr>
<td>mp_2200</td>
<td>18.48</td>
<td>8.48</td>
<td>46</td>
</tr>
<tr>
<td>mp_2820</td>
<td>13.54</td>
<td>1.54</td>
<td>11</td>
</tr>
<tr>
<td>mp_4628</td>
<td>12.95</td>
<td>0.85</td>
<td>6</td>
</tr>
<tr>
<td>mp_5000</td>
<td>6.92</td>
<td>2.98</td>
<td>43</td>
</tr>
<tr>
<td>mp_8000</td>
<td>4.33</td>
<td>1.64</td>
<td>38</td>
</tr>
<tr>
<td>mp_10000</td>
<td>2.15</td>
<td>0.25</td>
<td>11</td>
</tr>
<tr>
<td>single</td>
<td>n/a</td>
<td>3.00</td>
<td>n/a</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td><strong>92.00</strong></td>
<td><strong>42.50</strong></td>
<td></td>
</tr>
</tbody>
</table>
Summary

GARBAGE IN, GARBAGE OUT!
Acknowledgements

● ARC CoE - Coral Reef Studies
  ○ Sylvain Foret
  ○ David Miller
  ○ Janet Swanson

● VBC
  ○ Paul Harrison

● Family
  ○ Naomi, Oskar, Zoe
Contact

● Email
  torsten.seemann@monash.edu

● Web
  http://bioinformatics.net.au/

● GoogleChat
  torsten.seemann