NGS Quality Control

Sandra Smit
RNAseq course, June 5-7
NBIC/EPS/WUR
NGS data

DATA: sequences (reads) and quality values
FASTQ

Many formats:
- FASTQ
- FASTA & QUAL
- SCARF
- SFF

Sequence ID

Sequence

Quality values
FASTQ quality values

<table>
<thead>
<tr>
<th>0</th>
<th>33</th>
<th>59</th>
<th>64</th>
<th>73</th>
<th>104</th>
<th>126</th>
</tr>
</thead>
<tbody>
<tr>
<td>26</td>
<td>31</td>
<td>40</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-5</td>
<td>0</td>
<td>9</td>
<td>40</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>9</td>
<td></td>
<td>40</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>9</td>
<td>40</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

S - Sanger  Phred+33, raw reads typically (0, 40)
X - Solexa  Solexa+64, raw reads typically (-5, 40)
I - Illumina 1.3+ Phred+64, raw reads typically (0, 40)
J - Illumina 1.5+ Phred+64, raw reads typically (3, 40)
      with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
      (Note: See discussion above).
L - Illumina 1.8+ Phred+33, raw reads typically (0, 41)
Quality Control (QC) of NGS data

• Quality assessment of data
• Improvement of quality (trimming and filtering)
  – selection: remove problematic reads
  – trimming: discard regions of low fidelity
  – correction: replace improbable basecalls
• Quality assessment again, see improvement?
What could be wrong?

- Base calling errors
- Uncalled bases
- GC bias
- Homopolymers
- Low quality bases (3’ end)
- Clonal duplicates
- Contamination (pathogens)
- Nuclear vs. organellar reads
- Sequencing artifacts (adaptors, vectors, clones, chimeric reads)
QC Tools

• FastQC
  – http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

• PINSEQ
  – http://prinseq.sourceforge.net/

• FASTX Toolkit
  – http://hannonlab.cshl.edu/fastx_toolkit/

• NGS QC Toolkit
  – http://www.nipgr.res.in/ngsqctoolkit.html

• NGS QC in Galaxy
QC and manipulation in Galaxy

Manipulation of FASTQ data with Galaxy
Blankenberg et al.
Bioinformatics 2010
FastQC Quality Checking Tool

- Contamination screen
- Per base sequence quality
- Sequence length distribution
- Per base GC content
- Per sequence quality
- Per base sequence content
- Per base N-content
- Sequence duplication

FastQC is a quality checking tool that helps to ensure the quality of sequencing data. It provides various metrics such as base content, sequence quality, and GC content distribution, which are crucial for downstream analyses.
Good

Quality scores across all bases (Illumina 1.5 encoding)
Bad
Good

Quality score distribution over all sequences

Average Quality per read

Mean Sequence Quality (Phred Score)
Bad

Quality score distribution over all sequences

Average Quality per read
Tools in galaxy

• FASTQ groomer
  – Verify and convert between the known FASTQ variants
• Quality statistics
• Read Trimmer
• Quality filter
• Dealing with paired-end data
Tutorial

- Assess the quality in a good and bad data set
- Clean up the bad data set
- Examine the improvements
### Strategy

**Tool: Filter by quality**

<table>
<thead>
<tr>
<th>Name</th>
<th>Filter by quality on data 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Created</td>
<td>Jun 04, 2013</td>
</tr>
<tr>
<td>Filesize</td>
<td>19.2 MB</td>
</tr>
<tr>
<td>Dbkey</td>
<td>?</td>
</tr>
<tr>
<td>Format</td>
<td>fastqsanger</td>
</tr>
<tr>
<td>Galaxy Tool Version</td>
<td>1.0.0</td>
</tr>
<tr>
<td>Tool Version</td>
<td></td>
</tr>
<tr>
<td>Tool Standard Output</td>
<td>stdout</td>
</tr>
<tr>
<td>Tool Standard Error</td>
<td>stderr</td>
</tr>
<tr>
<td>Tool Exit Code</td>
<td>0</td>
</tr>
<tr>
<td>API ID</td>
<td>5abf71dc00cfff9f5</td>
</tr>
</tbody>
</table>

**Input Parameter**

- **Library to filter**: 5: FASTQ, Groomer on data 2
- **Quality cut-off value**: 20
- **Percent of bases in sequence that must have quality equal to / higher than cut-off value**: 80

**Tool: FASTQ Quality Trimmer**

<table>
<thead>
<tr>
<th>Name</th>
<th>FASTQ Quality Trimmer on data 7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Created</td>
<td>Jun 04, 2013</td>
</tr>
<tr>
<td>Filesize</td>
<td>18.9 MB</td>
</tr>
<tr>
<td>Dbkey</td>
<td>?</td>
</tr>
<tr>
<td>Format</td>
<td>fastqsanger</td>
</tr>
<tr>
<td>Galaxy Tool Version</td>
<td>1.0.0</td>
</tr>
<tr>
<td>Tool Version</td>
<td></td>
</tr>
<tr>
<td>Tool Standard Output</td>
<td>stdout</td>
</tr>
<tr>
<td>Tool Standard Error</td>
<td>stderr</td>
</tr>
<tr>
<td>Tool Exit Code</td>
<td>0</td>
</tr>
<tr>
<td>API ID</td>
<td>af7fa13876122ee5</td>
</tr>
</tbody>
</table>

**Input Parameter**

- **FASTQ File**: 7: Filter by quality on data 5
- **Keep reads with zero length**: False
- **Trim ends**: 3' only
- **Window size**: 1
- **Step Size**: 1
- **Maximum number of bases to exclude from the window during aggregation**: 0
- **Aggregate action for window**: min score
- **Trim until aggregate score is**: >=
- **Quality Score**: 20.0
After cleaning (50989 reads)