DNASeq: Analysis pipeline and file formats

Sumir Panji, Gerrit Boha and Amel Ghouila

Bioinformatics analysis and annotation of variants in NGS data workshop

Cape Town, 4th to 6th April 2016
Pipeline and formats

1. FastQ reads
   - Cleaning/trimming
   - Assembly
   - Mapping

2. SAM files
   - Mapping

3. BAM files
   - CNV, read depth
   - Structural variation
   - Variant detection
   - SNP call

4. VCF files
   - SNV call (FastQC)
   - CNV, read depth (BWA, Bowtie2, etc.)
   - Structural variation (Bedtools, Picard)

5. Visualization
6. Annotation
7. VCF files (GATK)
Quality control of output file (FastQC)

- Before the alignment or mapping of sequence reads to a reference sequence, 3 characteristics of NGS data complicate this task:

  1- If Read lengths are relatively short (36–250 bp) compared to traditional capillary-based sequencing \(\rightarrow\) less information to use for mapping + decreases likelihood that a read can be mapped to a single, unique location.

  2- If Reads are of imperfect quality = reads contain higher rates of sequencing error.

  3- Sheer volume of data. A single run produces millions of sequencing reads, whose alignment to a large reference sequence requires significant computing power.
Quality Control

FastQC: Quality Control for FastQ files

Command line
Available on Galaxy
Graphic reports

http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
What is FastQC?

• Before analyzing generated sequence to draw biological conclusions, a quality control check should be performed to make sure there is no biases in the data.

• Most sequencers will generate a QC report as part of their analysis pipeline focusing on identifying problems generated by the sequencer.

• FastQC aims to provide a QC report which can spot problems which originate either in the sequencer or in the library material.
## Basic statistics

<table>
<thead>
<tr>
<th>Measure</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Filename</td>
<td>good_sequence_short.txt</td>
</tr>
<tr>
<td>File type</td>
<td>Conventional base calls</td>
</tr>
<tr>
<td>Encoding</td>
<td>Illumina 1.5</td>
</tr>
<tr>
<td>Total Sequences</td>
<td>250000</td>
</tr>
<tr>
<td>Sequences flagged as poor quality</td>
<td>0</td>
</tr>
<tr>
<td>Sequence length</td>
<td>40</td>
</tr>
<tr>
<td>%GC</td>
<td>45</td>
</tr>
</tbody>
</table>
QC reports

- Basic Statistics
- Per base sequence quality
- Per sequence quality scores
- Per base sequence content
- Per base GC content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences

Quality scores across all bases (Sanger encoding)
FastQC reports

A summary of the modules which were run, and a quick evaluation of whether the results of the module seem entirely normal (green tick), slightly abnormal (orange triangle) or very unusual (red cross).
Per Base Sequence Quality
Per Base Sequence Quality

- The background of the graph divides the y axis:
  - ✓ very good quality calls (green)
  - ✓ calls of reasonable quality (orange)
  - ✓ calls of poor quality (red)

The quality of calls on most platforms will degrade as the run progresses, so it is common to see base calls falling into the orange area towards the end of a read.
FastQC reporting

• In addition to providing an interactive report, FastQC also has the option to create an HTML version of this report for a more permanent record.
Examples of Quality reports

• Good quality FastQC report:
  http://www.bioinformatics.babraham.ac.uk/projects/fastqc/good_sequence_short_fastqc.html

• Bad quality FastQC report
  http://www.bioinformatics.babraham.ac.uk/projects/fastqc/bad_sequence_fastqc.html
Trimming

- Poor quality in the beginnings and the ends can be cut using trimming tools (remove tags/adapter sequences)
Quality filtering

• Flexible tools for Illumina NGS data:
  ✓ Cutadapt
  ✓ Trimmomatic
  ✓ Seq TK, etc.

**Interest:**
✓ Can align more reads: may increase mapping rates
✓ Improve probably de novo assembly

**Warning:**
✓ Loss of information if stringent trimming