Raw Illumina Next Generation Sequencing data files and Quality Control

Gilgi Friedlander
Bioinformatics Unit, Biological Services
Illumina Workflow

Sample preparation
(Application-specific)

Load on Flow-Cell

Generate clusters

Sequence base-by-base

Run pipeline + QC
Next Generation Sequencing (NGS) experiments

- Plan the experiment!
  - Design
  - How to perform
  - Single end/paired end
  - read length
  - Replicates
  - Data analysis

We encourage to have a “kick-off” meeting

Lab- High Throughput Sequencing Unit:
Dr. Shirley Horn-Saban - Head of Genomic Technologies
Dr. Daniela Amann-Zalcenstein
Muriel Chemla

Bioinformatics (NGS analysis):
Dr. Dena Leshkowitz
Dr. Ester Feldmesser
Dr. Gilgi Friedlander
Today:

- Illumina’s pipeline
- How is the data organized?
- The format of the output files
- Quality control
- Viewing the data in a genomic browser

Exercise
Illumina’s pipeline

**CASAVA (v1.8.2)**

**Consensus Assessment of Sequence And Variation**

Irit Orr
Today:

- Illumina’s pipeline
- How is the data organized?
- The format of the output files
- Quality control
- Viewing the data in a genomic browser

Exercise
How is the data organized?

- **htsdata**
  - **Run_ID**
    - 23456_SN808_0051_BA0BC9ABXX
    - **Unaligned**
    - **Aligned**
  - **Aligned**
    - **Sample_name1**
      - SampleSheet.csv
      - fastq.gz files divided to multiple files (4M reads in each file)
    - **Export**
  - Summary_Stats
  - Barcode_Lane_Summary.htm
How is the data organized?

- **htsdata**
  - **Run_ID**
    - 23456_SN808_0051_BA0BC9ABXX
  - **Unaligned**
  - **Aligned**
- **Unaligned**
  - SampleSheet.csv
  - fastq.gz files divided to multiple files (4M reads in each file)
- **Aligned**
  - Summary_Stats
  - Barcode_Lane_Summary.htm
  - Sample_name1
  - Export

**What is the fastq format?**
Today:

- Illumina’s pipeline
- How is the data organized?
- The format of the output files
- Quality control
- Viewing the data in a genomic browser

Exercise
Output files

The FastQ format
(standard text representation of short reads)

- A FASTQ text file normally uses four lines per sequence.

- Example
  
  ```
  @SEQ_ID
  GATTTGGGGTTCAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
  +
  !'**((****++)%%%%++)%%%%).1****-++''))**55CCF>>>>>>CCCCCCC65
  ```

- Line 1 begins with a '@' character and is followed by a sequence identifier and an optional description (like a FASTA title line).
- Line 2 is the raw sequence letters.
- Line 3 begins with a '+' character (optionally followed by SEQ_ID).
- Line 4 encodes the quality values for the sequence in Line 2, and must contain the same number of symbols as letters in the sequence. The letters encode Phred Quality Scores.
Output files

fastq files—cont’

Quality scores

Each base has a quality score that measures the probability that a base is called incorrectly.

Illumina's base scoring is similar to Phred scores—a way of expressing estimates of sequencing error probabilities.

The quality score is in ASCII format: ASCII character code - 33

\[ Q_{\text{phred}} = \text{ASCII code} - 33 = -10 \log_{10}( Pe ) \]

\( Pe \) = error probability of a particular base call

\( Q_{20} = 1 \text{ error in 100 bases} \)

\( Q_{30} = 1 \text{ error in 1000 bases} \)
<table>
<thead>
<tr>
<th>Char</th>
<th>ASCII</th>
<th>Qphred ASCII-33</th>
<th>P(error)</th>
</tr>
</thead>
<tbody>
<tr>
<td>!</td>
<td>33</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>&quot;</td>
<td>34</td>
<td>0.7943</td>
<td>282</td>
</tr>
<tr>
<td>#</td>
<td>35</td>
<td>0.6309</td>
<td>573</td>
</tr>
<tr>
<td>$</td>
<td>36</td>
<td>0.5011</td>
<td>872</td>
</tr>
<tr>
<td>%</td>
<td>37</td>
<td>0.3981</td>
<td>072</td>
</tr>
<tr>
<td>&amp;</td>
<td>38</td>
<td>0.3162</td>
<td>278</td>
</tr>
<tr>
<td>(</td>
<td>39</td>
<td>0.2511</td>
<td>886</td>
</tr>
<tr>
<td>)</td>
<td>40</td>
<td>0.1995</td>
<td>262</td>
</tr>
<tr>
<td>*</td>
<td>41</td>
<td>0.1584</td>
<td>893</td>
</tr>
<tr>
<td>+</td>
<td>42</td>
<td>0.1258</td>
<td>925</td>
</tr>
<tr>
<td>.</td>
<td>43</td>
<td>0.1</td>
<td>0.1</td>
</tr>
<tr>
<td>H</td>
<td>72</td>
<td>0.00001</td>
<td>259</td>
</tr>
<tr>
<td>l</td>
<td>73</td>
<td>0.00001</td>
<td>259</td>
</tr>
</tbody>
</table>

SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAAT
+
!'"*(((((**+))%%%++)()%%()%%%.1**+-++'))

Q phred = -10 log10( Pe )
Output files

**fastq files— cont’**

Divided: each contains 4M read
Zipped
Contains only passed filtered reads

* Read number: 1 can be single read or read 2 of paired-end
How is the data organized?

- htsdata
  - Run_ID
    - 23456_SN808_0051_BA0BC9ABXX
  - Unaligned
  - Aligned

- Unaligned
  - Sample_name1
    - SampleSheet.csv
    - fastq.gz files divided to multiple files (4M reads in each file)
  - Aligned
    - Summary_Stats
      - Barcode_Lane_Summary.htm
    - Sample_name1
      - Export
## Output files

### SampleSheet.csv

<table>
<thead>
<tr>
<th>Header</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>FCID</td>
<td>Flowcell ID</td>
</tr>
<tr>
<td>Lane</td>
<td>Positive integer indicating lane number (1-8)</td>
</tr>
<tr>
<td>SampleID</td>
<td>ID of sample</td>
</tr>
<tr>
<td>SampleRef</td>
<td>The reference sequence to be used for Sample</td>
</tr>
<tr>
<td>Index</td>
<td>Index sequence</td>
</tr>
<tr>
<td>Description</td>
<td>Description of the sample</td>
</tr>
<tr>
<td>Control</td>
<td>Y indicates lane is control lane N means sample</td>
</tr>
<tr>
<td>Recipe</td>
<td>Recipe used for sequencing</td>
</tr>
<tr>
<td>Operator</td>
<td>Name or ID of operator</td>
</tr>
<tr>
<td>SampleProject</td>
<td>The project the sample belongs to</td>
</tr>
</tbody>
</table>
How is the data organized?

- htsdata
  - Run_ID
    - 23456_SN808_0051_BA0BC9ABXX
  - Unaligned
  - Aligned
- Unaligned
- Sample_name1
  - SampleSheet.csv
  - fastq.gz files divided to multiple files (4M reads in each file)
- Aligned
  - Summary_Stats
    - Barcode_Lane_Summary.htm
  - Sample_name1
    - Export
Today:

- Illumina’s pipeline
- How is the data organized?
- The format of the output files
- Quality control

Exercise
Quality Control

What is the quality of my data?
Quality Control

Two levels:

1. The qualities of the bases of the reads

2. If there is an available genome:
   What is the fraction of the reads that align to the genome? What is the error rate?
Quality Control

When you get your data, you get a mail with the location of the files and with a link to some tables and plots, that looks something like:

http://dapsas.weizmann.ac.il/ngsreports/110922_SN808_0058_BB07HNABXX/
Open the link and go to the Summary tab

The PhiX reads are mapped on the PhiX reference genome, the error rate is then estimated by the number mismatches, over the total number of bases of mapped PhiX reads should be below 1.5.

<table>
<thead>
<tr>
<th>Lane</th>
<th>Tiles</th>
<th>Clu. Dens. (#/mm²)</th>
<th>% PF Clusters</th>
<th>Clusters PF (#/mm²)</th>
<th>% Phas./Preph.</th>
<th>Cycles</th>
<th>% Aligned</th>
<th>% Error Rate 35 cycle</th>
<th>% Error Rate 75 cycle</th>
<th>% Error Rate 100 cycle</th>
<th>1st Cycle Int</th>
<th>% Intensity Cycle 20</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>32</td>
<td>401K +/-.40.5K</td>
<td>95.4 +/- 0.46</td>
<td>382.3K +/- 37.51K</td>
<td>0.519 / 0.208</td>
<td>49</td>
<td>99.05 +/- 0.016</td>
<td>0.17 +/- 0.024</td>
<td>0.12 +/- 0.003</td>
<td>0.00 +/- 0.000</td>
<td>4376 +/- 499.5</td>
<td>81.2 +/- 1.77</td>
</tr>
<tr>
<td>2</td>
<td>32</td>
<td>830K +/-.81.1K</td>
<td>85.4 +/- 2.89</td>
<td>683.2K +/- 49.39K</td>
<td>0.519 / 0.208</td>
<td>0</td>
<td>0.00 +/- 0.000</td>
<td>0.00 +/- 0.000</td>
<td>0.00 +/- 0.000</td>
<td>0.00 +/- 0.000</td>
<td>4617 +/- 513.5</td>
<td>81.1 +/- 1.01</td>
</tr>
<tr>
<td>3</td>
<td>32</td>
<td>636K +/-.94.6K</td>
<td>88.5 +/- 2.63</td>
<td>613.9K +/- 70.33K</td>
<td>0.519 / 0.208</td>
<td>0</td>
<td>0.00 +/- 0.000</td>
<td>0.00 +/- 0.000</td>
<td>0.00 +/- 0.000</td>
<td>0.00 +/- 0.000</td>
<td>4370 +/- 920.1</td>
<td>81.9 +/- 0.82</td>
</tr>
<tr>
<td>4</td>
<td>32</td>
<td>767K +/-.92.9K</td>
<td>86.5 +/- 2.55</td>
<td>661.8K +/- 63.45K</td>
<td>0.519 / 0.208</td>
<td>0</td>
<td>0.00 +/- 0.000</td>
<td>0.00 +/- 0.000</td>
<td>0.00 +/- 0.000</td>
<td>0.00 +/- 0.000</td>
<td>4533 +/- 462.2</td>
<td>81.2 +/- 1.24</td>
</tr>
<tr>
<td>5</td>
<td>32</td>
<td>682K +/-.95.7K</td>
<td>89.1 +/- 1.92</td>
<td>606.3K +/- 73.20K</td>
<td>0.519 / 0.208</td>
<td>0</td>
<td>0.00 +/- 0.000</td>
<td>0.00 +/- 0.000</td>
<td>0.00 +/- 0.000</td>
<td>0.00 +/- 0.000</td>
<td>4636 +/- 504.0</td>
<td>81.3 +/- 1.03</td>
</tr>
<tr>
<td>6</td>
<td>32</td>
<td>809K +/-.96.3K</td>
<td>84.2 +/- 3.65</td>
<td>677.8K +/- 56.22K</td>
<td>0.519 / 0.208</td>
<td>0</td>
<td>0.00 +/- 0.000</td>
<td>0.00 +/- 0.000</td>
<td>0.00 +/- 0.000</td>
<td>0.00 +/- 0.000</td>
<td>4119 +/- 479.0</td>
<td>86.6 +/- 1.53</td>
</tr>
<tr>
<td>7</td>
<td>32</td>
<td>701K +/-.93.2K</td>
<td>88.5 +/- 2.40</td>
<td>617.8K +/- 66.63K</td>
<td>0.519 / 0.208</td>
<td>0</td>
<td>0.00 +/- 0.000</td>
<td>0.00 +/- 0.000</td>
<td>0.00 +/- 0.000</td>
<td>0.00 +/- 0.000</td>
<td>4032 +/- 491.0</td>
<td>88.3 +/- 0.97</td>
</tr>
<tr>
<td>8</td>
<td>32</td>
<td>731K +/-.94.0K</td>
<td>86.0 +/- 2.71</td>
<td>627.8K +/- 70.65K</td>
<td>0.519 / 0.208</td>
<td>0</td>
<td>0.00 +/- 0.000</td>
<td>0.00 +/- 0.000</td>
<td>0.00 +/- 0.000</td>
<td>0.00 +/- 0.000</td>
<td>3870 +/- 529.4</td>
<td>87.5 +/- 1.33</td>
</tr>
</tbody>
</table>

% reads passed filter (chastity filter)

% aligned to reference using ELAND (the read-mapper supplied by Illumina)

Average of the four intensities at the first cycle

% intensity after 20 cycles should be 50% or more
Quality Control

Explore the quality of the data by looking at boxplots of various parameters.
Quality Control: viewing plots

Check one lane at a time

\[ Q_{\text{phred}} = -10 \log_{10}(P_e) \]

\[ Q_{34} \Rightarrow p(\text{error}) \sim 0.0004 \]
Quality Control: viewing plots

Qualities drop gradually  \( (Q_{30} \Rightarrow P(\text{error})=0.001) \)

For reads with 50 bases  \( \Rightarrow >90\% \)  For reads with 100 bases  \( \Rightarrow >75\% \)
Important:

These plots are created during the RUN in the HiSeq.

During CASAVA - the qualities are being calibrated.

⇒ It is a good idea to look at the quality scores also after the calibration (fastqc tool)
And decide whether we would like to filter our reads prior to our downstream analyses
In case we have alignment, it is important to check the % of reads that were aligned
How is the data organized?

htsdata

- Run_ID
  - 23456_SN808_0051_BA0BC9ABXX

- Unaligned
- Aligned

Unaligned

- Sample_name1
  - SampleSheet.csv
  - fastq.gz files
divided to multiple files
(4M reads in each file)

Aligned

- Summary_Stats
- Barcode_Lane_Summary.htm
- Export
Quality Control: viewing plots

Fastqc: a great tool for assessing the quality of the data

http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/

Simon Andrews, Cambridge - UK
Quality Control: viewing plots

Good dataset

Quality Control: viewing plots

Poor dataset

Quality scores across all bases (Illumina 1.5 encoding)

http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/bad_sequence_fastqc/fastqc_report.html
Genome Browser

After you make sure your data is of good quality:

Analysis step (beyond the scope of this workshop)

During the year the bioinformatics unit will give various workshops for specific NGS applications:
http://bip.weizmann.ac.il/ws/

During downstream steps of the analysis we can use a genome browser to view the data and assess specific, local quality, depending on the application.

Examples:
In RNA seq: investigating a newly identified transcripts
In genomic DNA seq: investigating a specific called SNP
Today:

- Illumina’s pipeline
- How is the data organized?
- The format of the output files
- Quality control

- Viewing the data in a genomic browser

Exercise
Genome Browser

There are many available genomic browsers. Among them:
UCSC browser
IGV (Integrative Genomics Viewer)

IGV:
A desktop application for integrated visualization of multiple data types and annotations in the context of the genome

http://www.broadinstitute.org/software/igv

Developed by Jim Robinson, Broad Institute
IGV provides a set of hosted genomes, but it is also possible to import other genomes