Molecular In My Pocket™ …

Bioinformatics: Basic File Formats Part I
Analysis Schema and FASTQ (Primary Analysis)

NGS Workflow

Reference Genome
AGCTTCGATA
GATACTGCA
CATACCTTC
TT-----ATACCTGC

Alignment (BAM)

Variants (VCF) Application

Secondary Analysis

Primary Analysis

AGCTTCGATA
GATACTGCA
CATACCTTC
TT-----ATACCTGC

Reads (FASTQ)

Sequencing

Ligation

Fragments

DNA

WetLab

*Tertiary Analysis

Annotation

Database

Interpretation/Report
**FASTQ format**: A text-based format for storing both biological sequences and their corresponding quality scores. FASTQ files contain 4 lines for each sequence.

Line 1: A sequence identifier with information about the sequencing run and cluster.
Line 2: The sequence of the read (the base calls; A, C, T, G and N).
Line 3: A “+” and is optionally followed by sequence related information.
Line 4: The ASCII encoded quality scores of the base calls in the sequence in Line 2.

**Example 1**: An ASCII characters and error probabilities table in ASCII_BASE 33 (the Phred quality score starts with ASCII code 33 to exclude the control characters). The Phred quality score calculation is described in [http://drive5.com/usearch/manual/quality_score.html](http://drive5.com/usearch/manual/quality_score.html).

<table>
<thead>
<tr>
<th>ASCII Character</th>
<th>Phred Quality Score</th>
<th>Probability of Incorrect Base Call</th>
<th>Base Call Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>+</td>
<td>10</td>
<td>1 in 10</td>
<td>90%</td>
</tr>
<tr>
<td>S</td>
<td>20</td>
<td>1 in 100</td>
<td>99%</td>
</tr>
<tr>
<td>?</td>
<td>30</td>
<td>1 in 1,000</td>
<td>99.9%</td>
</tr>
<tr>
<td>I</td>
<td>40</td>
<td>1 in 10,000</td>
<td>99.99%</td>
</tr>
</tbody>
</table>

**Example 2**: @instrument name:run ID:flowcell ID:lane number:tile number:x-coord within the tile:y-coord within the tile[space]member of a pair:if read filtered:control bits:index sequence

The Sequence of the Read

<table>
<thead>
<tr>
<th>ASCII Character</th>
<th>Phred Quality Score</th>
<th>Probability of Incorrect Base Call</th>
<th>Base Call Accuracy</th>
<th>Error Probability from ASCII_BASE 33</th>
</tr>
</thead>
<tbody>
<tr>
<td>#</td>
<td>2</td>
<td>1 in 2</td>
<td>37%</td>
<td>0.63096</td>
</tr>
<tr>
<td>A</td>
<td>32</td>
<td>1 in 1,587</td>
<td>99.94%</td>
<td>0.00063</td>
</tr>
<tr>
<td>E</td>
<td>36</td>
<td>1 in 4,000</td>
<td>99.98%</td>
<td>0.00025</td>
</tr>
</tbody>
</table>

**References**

1. [https://support.illumina.com/bulletins/2016/04/fastq-files-explained.html](https://support.illumina.com/bulletins/2016/04/fastq-files-explained.html)

"Molecular in My Pocket" reference cards are educational resources created by the Association of Molecular Pathology (AMP) for laboratory and other health care professionals. The content does not constitute medical or legal advice, and is not intended for use in the diagnosis or treatment of individual conditions. See [www.amp.org](http://www.amp.org) for the full "Limitations of Liability" statement.

Revised 7/2023