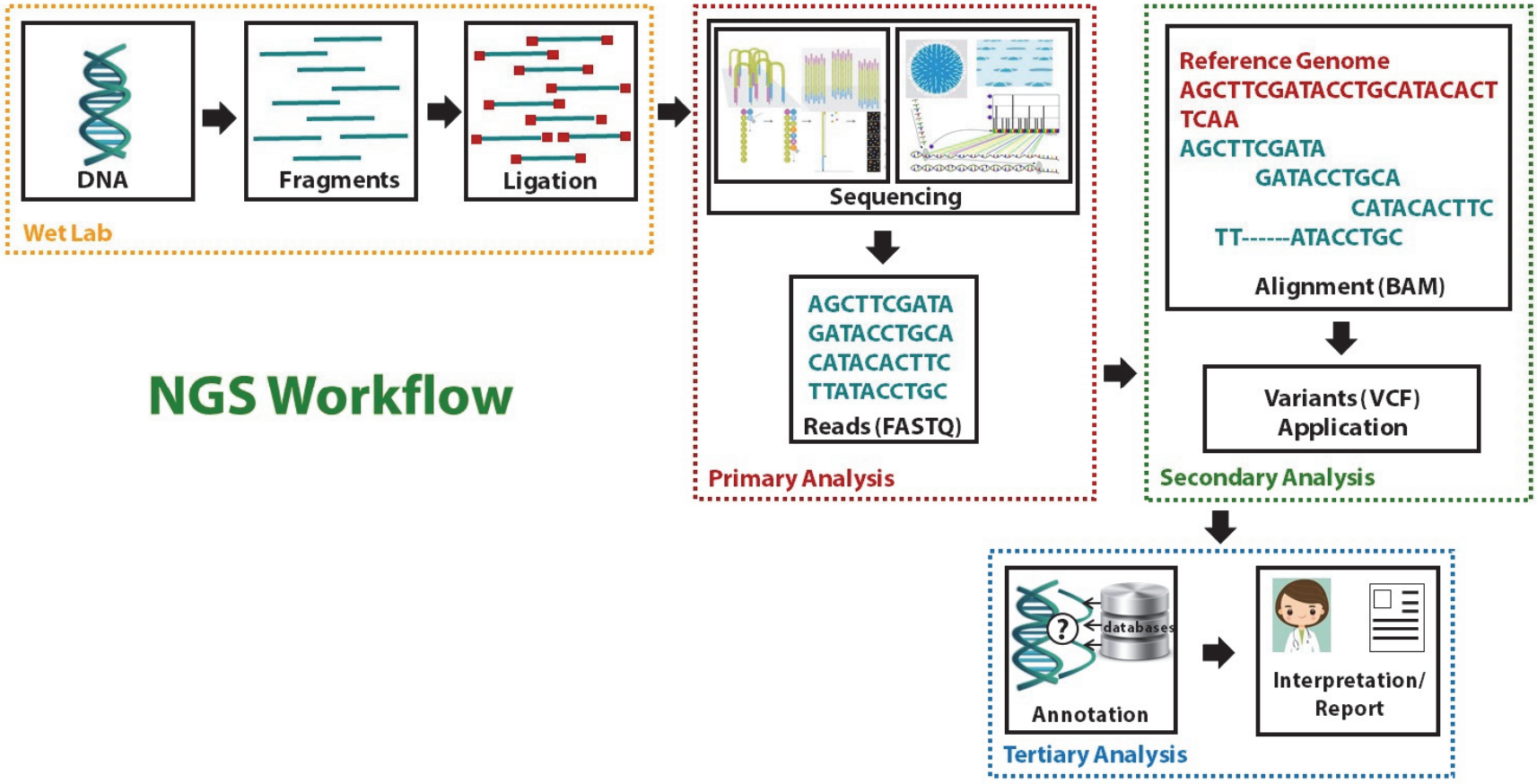
Prepared by the Association for Molecular Pathology Training and Education Committee
For More Educational Resources: *www.amp.org/AMPEducation*

Molecular In My Pocket™…

Bioinformatics: Basic File Formats Part I

Analysis Schema and FASTQ (Primary Analysis)



NGS Workflow

Revised 7/2023

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

Line1: A sequence identifier with information about the sequencing run and cluster.
Line2: The sequence of the read (the base calls; A, C, T, G and N).
Line3: A "+" and is optionally followed by sequence related information.
Line4: 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.

ASCII Character	Phred Quality Score	Probability of Incorrect Base Call	Base Call Accuracy
+	10	1 in 10	90%
5	20	1 in 100	99%
?	30	1 in 1,000	99.9%
I	40	1 in 10,000	99.99%

Example 2:

@NB551591:7:HJT5CBGX9:1:11101:9719:1040 1:N:0:CATCAAGT
TACTGNTAATGGTGGCAGGGTCTGTTCTTTCCAATCCTGAGGACTCCATGGGCACAAAATTCTCCTG
+
AAAAA#EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE

@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

ASCII Character	Phred Quality Score	Probability of Incorrect Base Call	Base call Accuracy	Error Probability from ASCII_BASE 33
#	2	1 in 2	37%	0.63096
A	32	1 in 1,587	99.94%	0.00063
E	36	1 in 4,000	99.98%	0.00025

References

1. https://support.illumina.com/bulletins/2016/04/fastq-files-explained.html
2. http://drive5.com/usearch/manual/quality_score.html

Revised 7/2023