wellcome connecting science

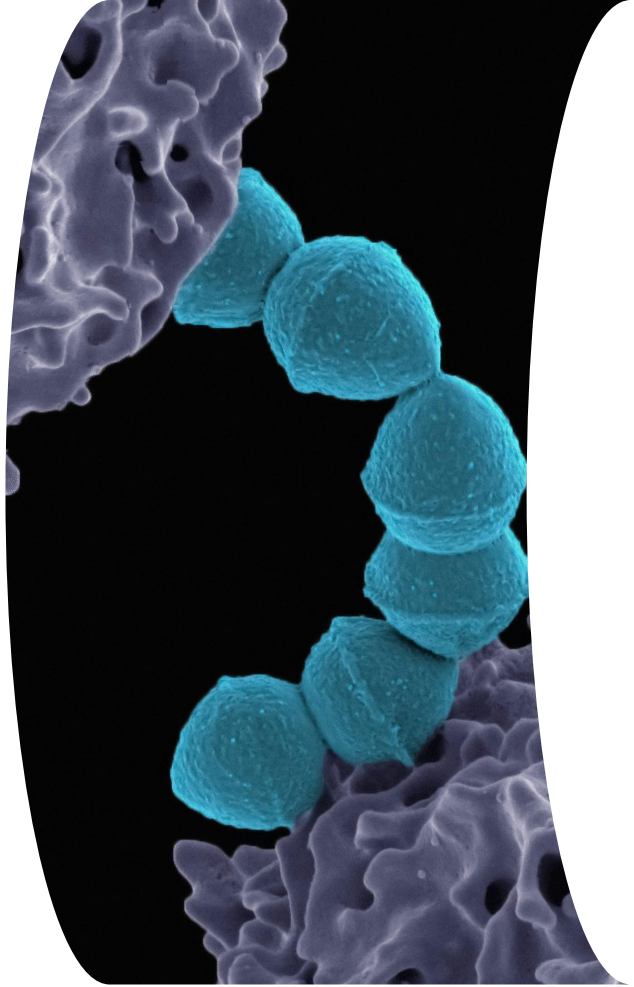
ACORN

Quality Control of Sequence Reads

21 May 2024

[Virtual, Across Africa and Asia]

Collins Kigen
Research Scientist



Streptococcus Pyogenes
Photo by National Institute of Allergy and Infectious Diseases on Unsplash

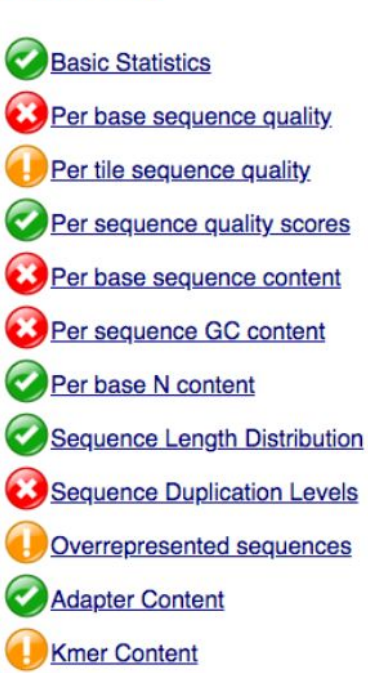
INTRODUCTION

- Errors at any of these steps can negatively impact the quality of the sequence
- If these errors are not removed from the raw reads, they might be incorporated into your analysis output and would be harder to resolve later on.
- Therefore, it is important to perform quality checks on the raw sequence reads before starting your analysis.

FastQC

- FastQC: A tool for quality control of high-throughput sequence data.
- The tool can be run by both command-line and also has a graphical user interface
- The tool provides you with a report on the quality of sequence reads using a traffic light system, red, amber and green.
- There are a number of parameters which we will learn in this module that help us in assessing the sequence data quality.

FastQC



- Green: Indicates no significant issues detected.
- Amber: Suggests potential problems that may require attention.
- Red: Indicates critical issues that need immediate investigation and possibly corrective action.

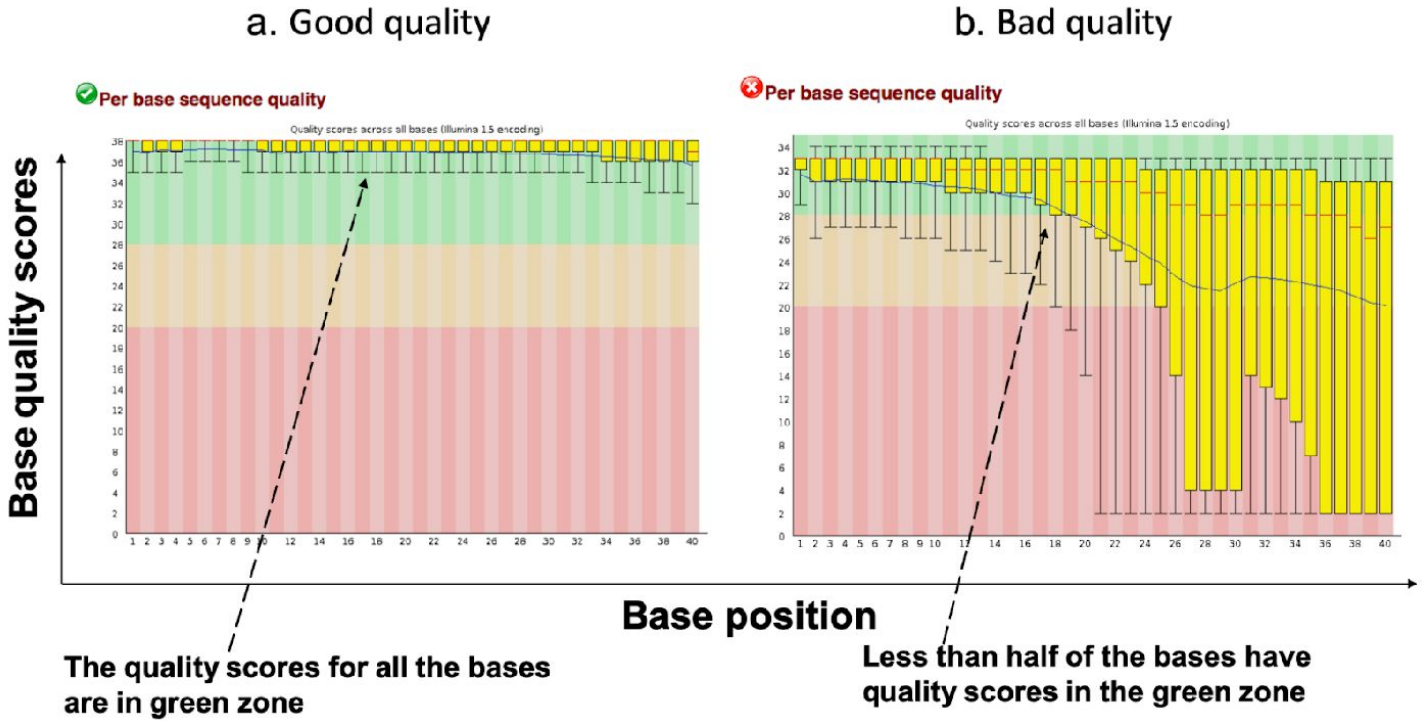
Basic Statistics

Basic Statistics

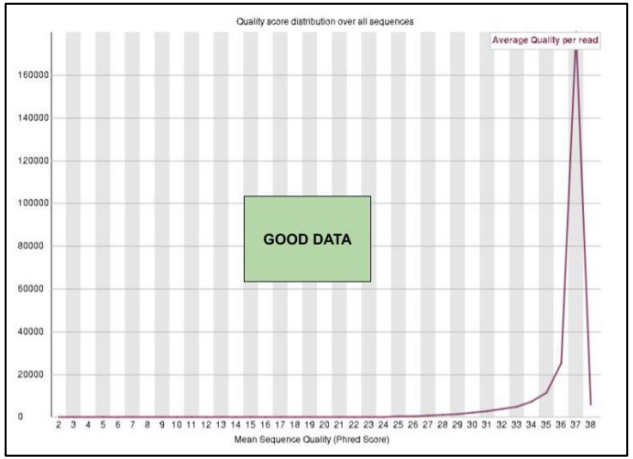
Measure	Value
Filename	ERR2093245_1.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	1339517
Sequences flagged as poor quality	0
Sequence length	250
%GC	51

- Contain basic information from sequence reads like number of reads, length, GC%

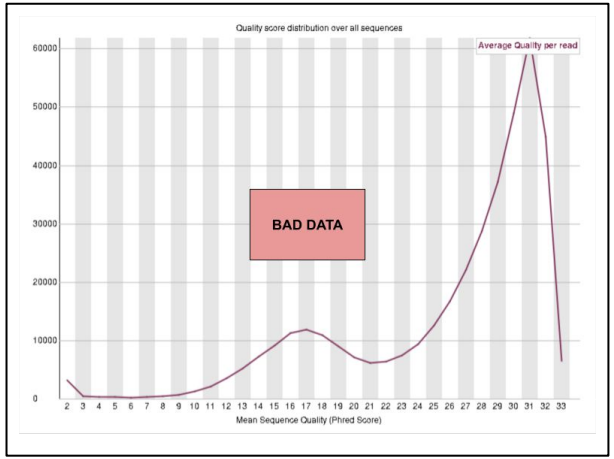
Quality


a. Good quality
b. Bad quality
Per base sequence quality
Per base sequence quality
Quality scores across all bases (Illumina 1.5 encoding)
Quality scores across all bases (Illumina 1.5 encoding)
Base quality scores
Base position
The quality scores for all the bases are in green zone
Less than half of the bases have quality scores in the green zone

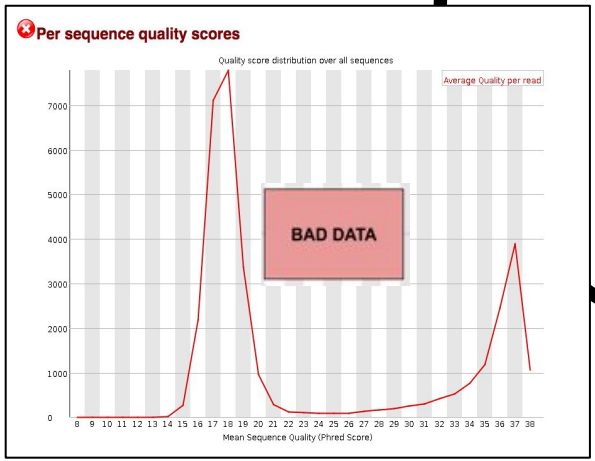
Per sequence quality scores



Single peak at av. quality score > 27

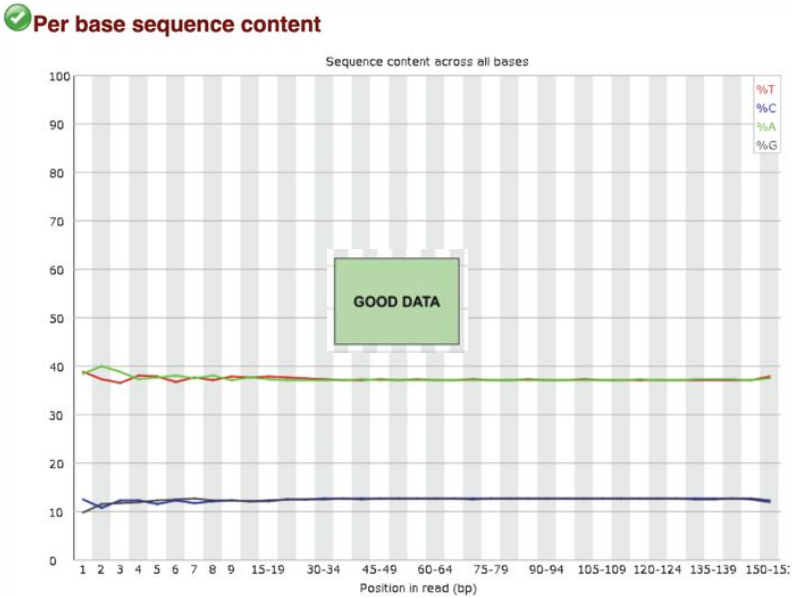


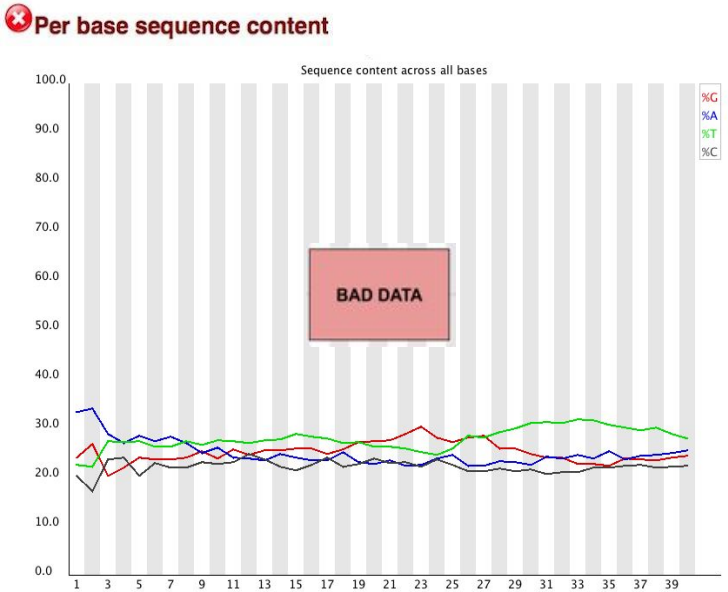
Bimodal distribution – Contaminating artifacts?



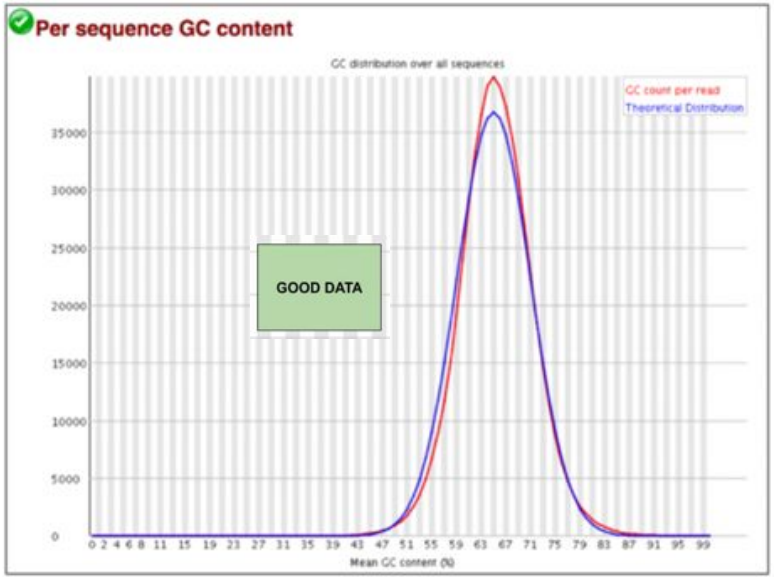
Majority of reads have low quality

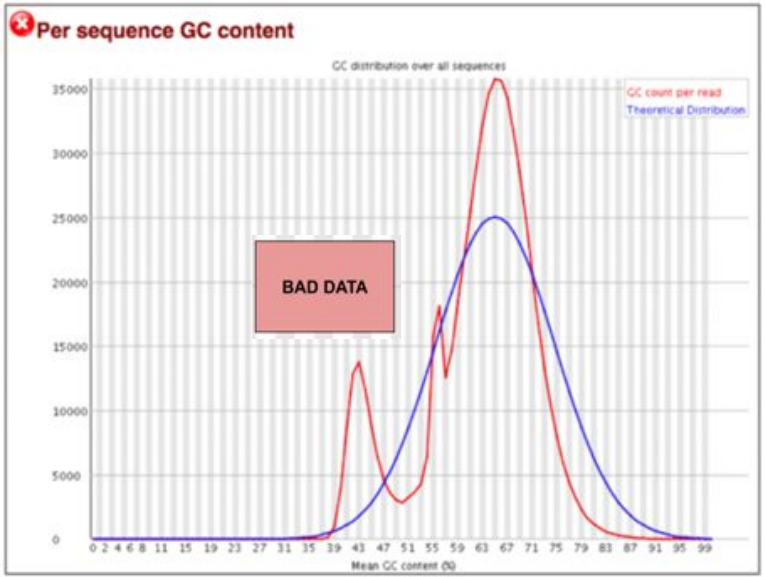
Per base sequence content



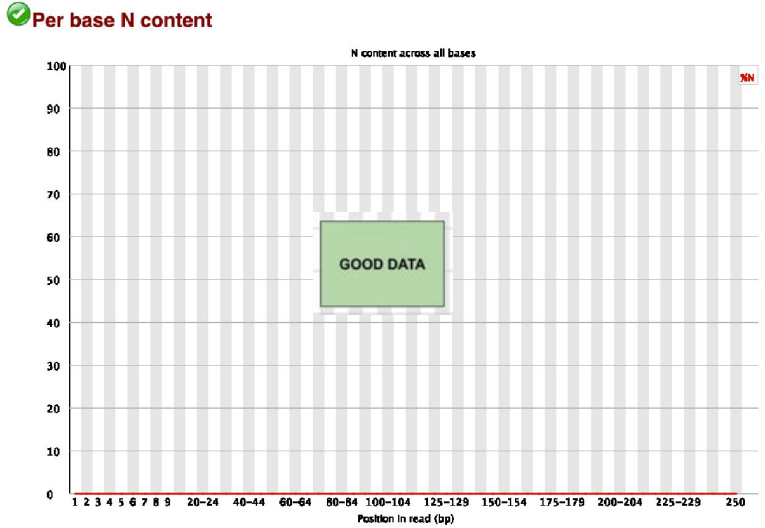


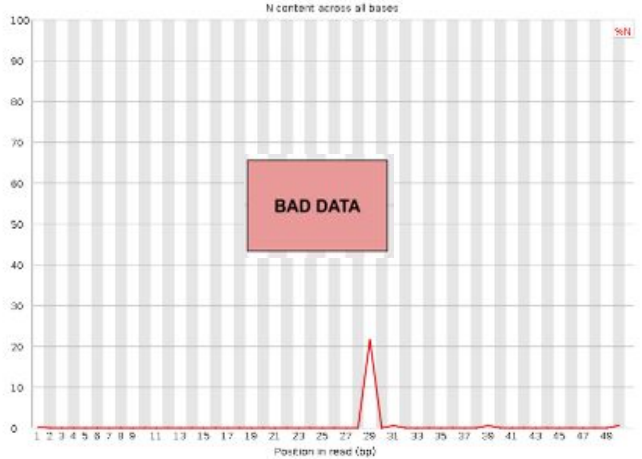
Per sequence GC content



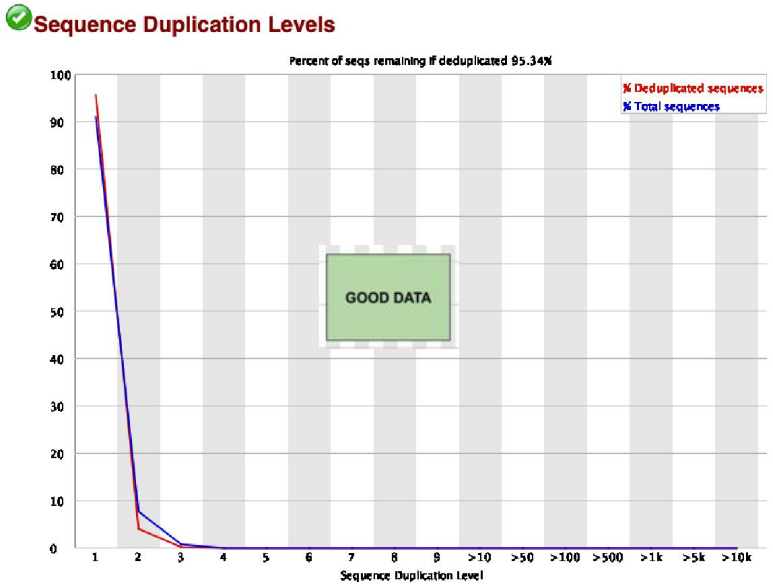


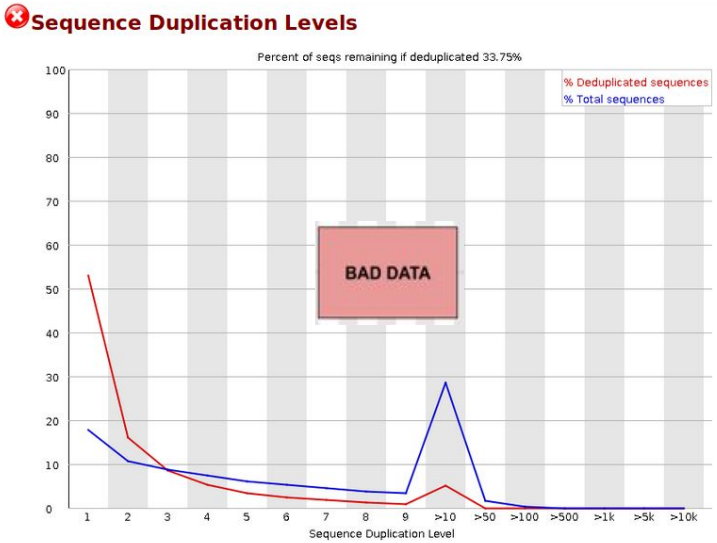
Per base N content



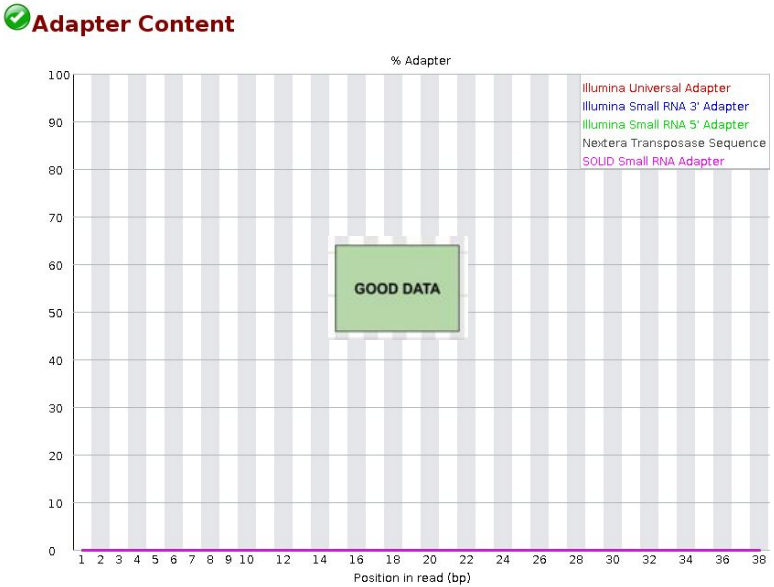


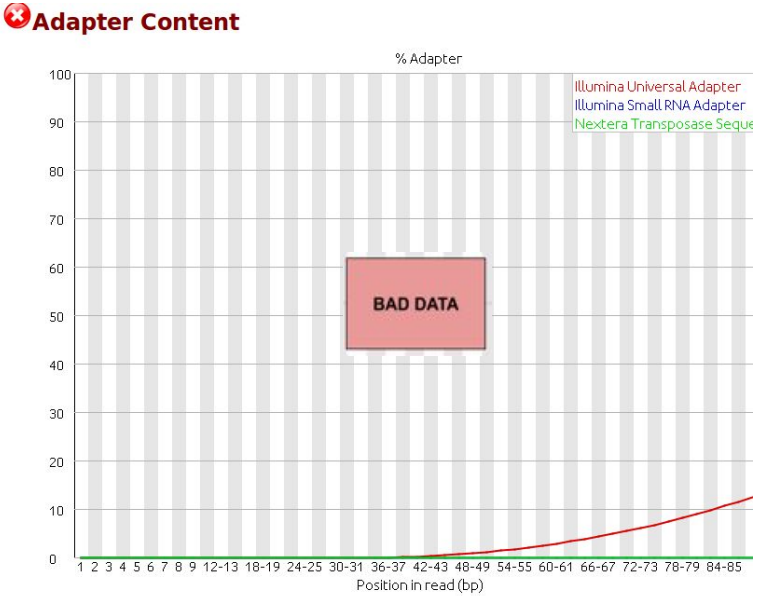
Sequence duplication levels





Adapter content





wellcome connecting science

ACORN

questions?