Sequence Analysis of Next Generation Sequences

Fawwaz Chirag Sofyan

November 5, 2023

Abstract

Next Generation Sequencing (NGS) technology has revolutionized genomics and molecular biology, enabling researchers to generate massive amounts of genetic data with unprecedented speed and accuracy. However, the reliability and validity of the insights derived from NGS data are contingent upon rigorous quality control (QC) measures. This abstract provides an overview of the essential elements of NGS data QC. It highlights the critical steps in assessing data quality, including the evaluation of raw sequence data, library preparation, alignment, coverage depth, detection of PCR duplicates, and the identification of contaminants. Furthermore, it emphasizes the significance of QC in both DNA-Seq and RNA-Seq experiments. Effective QC measures are essential to ensure that NGS data accurately reflects the underlying biology, free from technical artifacts and errors. This abstract underscores the importance of a well-executed QC pipeline in NGS research, promoting data integrity, reproducibility, and the generation of meaningful biological insights.

1 Introduction

Next Generation Sequencing (NGS) has become a cornerstone of modern genomics, enabling researchers to unravel the intricacies of genetic information on an unprecedented scale. This revolutionary technology has the potential to provide a wealth of data for various applications, from deciphering the genetic basis of diseases to understanding complex biological processes. However, with the power of NGS also comes the responsibility of rigorous quality control, as the accuracy and reliability of the data generated are paramount for meaningful scientific discoveries.

Quality control (QC) in NGS is a multifaceted process that involves a series of assessments and measures aimed at identifying and rectifying issues, anomalies, and errors in the generated sequence data. The ultimate goal of QC is to ensure that the data accurately represents the biological reality being studied, free from technical artifacts or biases.

The importance of quality control in NGS cannot be overstated. NGS data can be affected by various sources of error, such as base-calling inaccuracies, contamination, library preparation issues, and more. These errors, if left unaddressed, can lead to incorrect conclusions and misinterpretations of genomic or transcriptomic information. Thus, a robust QC pipeline is crucial for safeguarding the integrity of the data and for generating reliable results.

Quality control in NGS encompasses a wide range of activities, including the assessment of raw sequence data, alignment to a reference genome or transcriptome, evaluation of coverage depth, identification of PCR duplicates, detection of potential contaminants, and the examination of variant calling (for DNA-Seq) or gene expression quantification (for RNA-Seq). Each of these steps requires careful scrutiny and appropriate corrective measures to ensure data quality.

You have got to check out Fig. 3: it is amazing.

Here is an example citation when you want an author name like **?** to appear in the text. And here's how to do a parenthetic citation, when you want to mention a reference at the end of a sentence or part of a sentence (**?**).

It is possible to cite multiple references at the same time (**??????**).

2 Methods

Quality control of Next Generation Sequencing (NGS) data is essential to ensure the reliability and accuracy of the results obtained from genomic or transcriptomic analyses. Here are the key steps and methods for controlling the quality of NGS data:

Assess Raw Sequence Data Quality: Evaluate the FastQ files containing the raw sequence reads using tools like FastQC or MultiQC. Check for per-base sequence quality scores, sequence duplication levels, adapter contamination, and overrepresented sequences. Identify and remove low-quality reads and adapters using trimming tools like Trimmomatic or Cutadapt.

Validate Library Preparation: Ensure the integrity of the library preparation process, as issues here can lead to biases in downstream analysis. Verify that the library size distribution is as expected and consistent with the experimental design.

Examine Sequence Alignment: Map the reads to a reference genome or transcriptome using alignment tools like BWA, Bowtie, or STAR. Check alignment statistics to assess mapping rates and the distribution of reads across different genomic regions.

Evaluate Coverage Depth: Analyze the depth and uniformity of sequencing coverage. Ensure that regions of interest have sufficient coverage for the intended analysis. Tools like BEDTools can help calculate coverage statistics.

Detect PCR Duplicates: PCR duplicates can inflate read counts and skew downstream analyses. Use tools like Picard MarkDuplicates to identify and remove them.

Assess Variant Calling (for DNA-Seq): If you are performing DNA-Seq, evaluate the quality of variant calling using tools like GATK (Genome Analysis Toolkit) or samtools. Pay attention to variant quality scores, allele balance, and allele frequency.

Examine Gene Expression (for RNA-Seq): If you are conducting RNA-Seq, assess gene expression using software such as STAR, HISAT2, or Salmon. Check the distribution of read counts, library size, and transcript coverage.

Detect and Address Contaminants: Perform a BLAST search or use dedicated contaminant detection tools to identify any non-sample sequences or contaminants. Exclude or filter out contaminants to prevent them from interfering with downstream analysis.

Monitor Batch Effects: For multi-sample experiments, watch out for batch effects that can confound your analysis. Tools like ComBat or SVA can help mitigate batch effects.

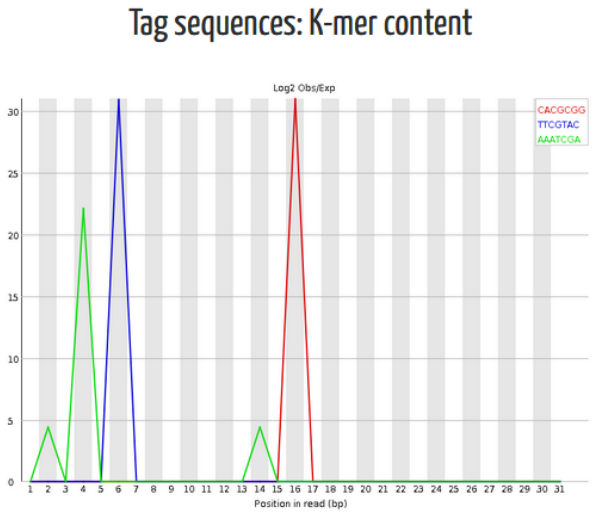

Figure 1: Enter Caption

Quality Control Metrics: Compute relevant quality control metrics, including mean quality scores, base composition, GC content, and more, depending on the specific analysis requirements.

Visualization: Visualize your QC results using plots or charts to quickly identify any anomalies or trends in the data.

Documentation: Keep detailed records of all QC steps, parameters, and any data manipulations performed. This documentation is crucial for transparency and reproducibility.

Iterative QC: Perform QC at multiple stages of the analysis pipeline to catch and address issues early.

By systematically applying these quality control measures, you can ensure the accuracy and reliability of your NGS data, leading to more robust and meaningful biological insights. Additionally, the choice of tools and software may vary depending on the specific NGS platform, the nature of your experiment, and the biological questions you seek to answer.

3 Results

In this tutorial we checked the quality of FASTQ files to ensure that their data looks good before inferring any further information. This step is the usual first step for analyses such as RNA-Seq, ChIP-Seq, or any other OMIC analysis relying on NGS data. Quality control steps are similar for any type of sequencing data: Quality assessment with tools like: Short Reads: FASTQE (Galaxy version 0.2.6+galaxy2) Short+Long: FASTQC (Galaxy version 0.73+galaxy0) Long Reads: Nanoplot (Galaxy version 1.41.0+galaxy0) Nanopore only: PycoQC (Galaxy version 2.5.2+galaxy0) Trimming and filtering for short reads with a tool like Cutadapt Heres the result about quality control of NGS:

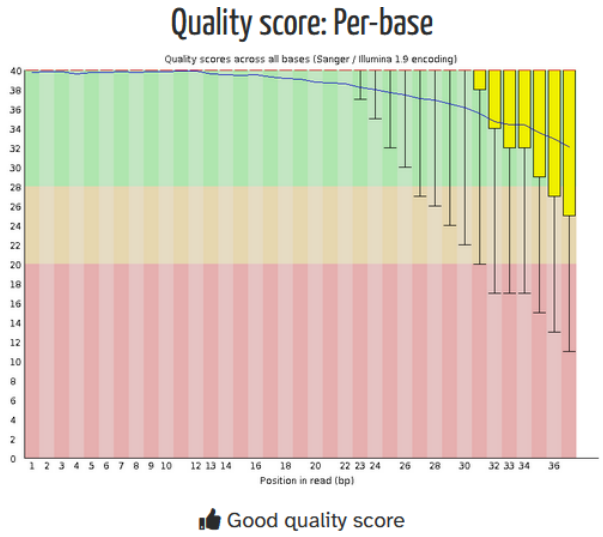

Figure 2: Quality Score-Per Base.

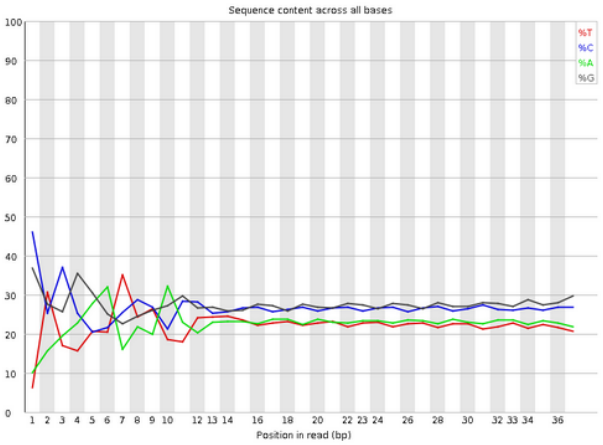

Figure 3: Perbase Quality.

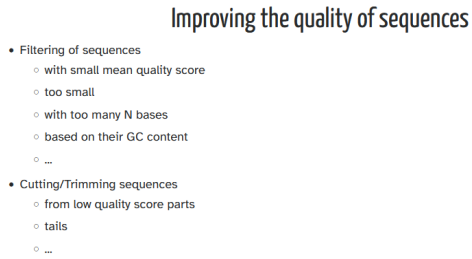

Figure 4: Improving the quality sequences

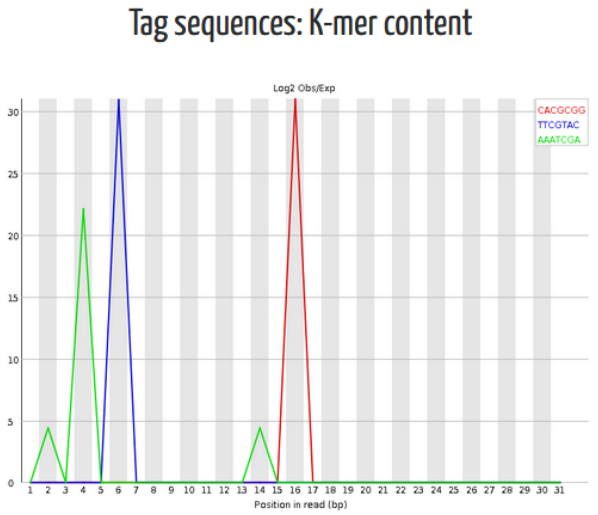

Figure 6: K-mer content

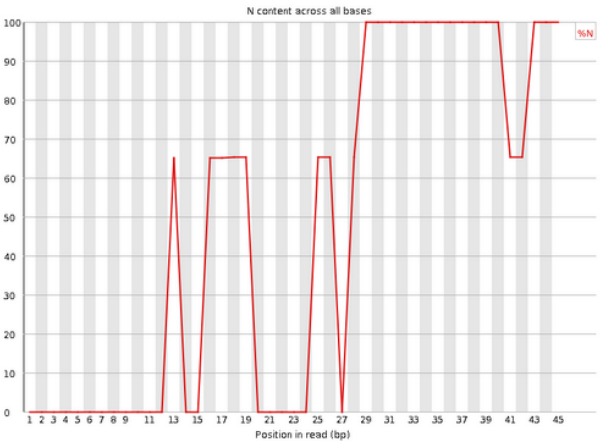

Figure 7: Per base n-content

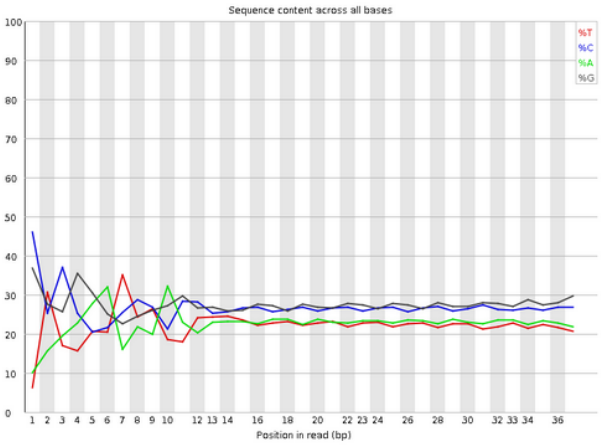

Figure 8: Per base sequence content