LECTURE 1

Closer Look: Quality Checks in Bacterial Genomics

Daniel Gyamfi Amoako, **PhD**
University of Guelph, Canada (damoako@uoguelph.ca)
University of KwaZulu-Natal, South Africa (amoakod@ukzn.ac.za)
Amoako Analytics Hub *

* "Empowering One-Health in Africa through Bioinformatics: Free Machine Learning and AI Analytics Data Solutions”

Outline

- Introduction
- Importance of quality checks
- Quality checks overview
- Read filtering
- Trimming
- Contamination assessment
- Quality control tools and metrics
- Quality check best practices
- Summary
- References
- Q&A

Introduction

This presentation will;

- focus on the importance of quality checks in bacteria genomics, and provide an overview of the key concepts and best practices related to quality checks.

- provide a better understanding of how quality checks can help ensure the accuracy and reliability of genomics data.

- equip you with the knowledge and tools needed to conduct effective quality checks in your own research.

The importance of quality checks in genomics data

- Genomics surveillance is crucial for tracking bacterial antimicrobial resistance, detecting outbreaks, and guiding treatment decisions.

- However, accurate and reliable genomics data is essential for effective interventions. Even minor errors in sequencing, assembly, or annotation can result in incorrect conclusions.

- Quality checks are necessary to ensure the accuracy and reliability of genomics data in bacterial antimicrobial resistance surveillance.

- Quality checks help researchers and public health officials identify and address data errors/biases, leading to more informed decisions and effective actions.

Quality checks overview

- **Definition:** A set of procedures used to assess the quality of genomics data and ensure its accuracy and reliability.
- **Types:** pre-processing, processing, and post-processing.
- Quality checks can be performed at various stages of the genomics workflow, including **DNA extraction**, **library preparation**, **sequencing**, and **data analysis**.
- Each step is essential for ensuring the accuracy and reliability of the genomics data.

Read Filtering

- **Definition:** Is a pre-processing quality check that involves removing low-quality or irrelevant reads from the sequencing data.

- **Primary reasons:** To remove sequencing errors, adapter contamination, and low-quality reads, which can affect downstream analyses and interpretation of the data.

- **Types:** quality-based filtering (removes reads with low-quality scores), adapter trimming (removes reads with adapter contamination) and length filtering (removes reads that are too short or too long).

- **Tools:** Trimmomatic, Fastp, BBDuk, and Cutadapt.

Trimming

- **Definition:** Is a pre-processing quality check that involves removing low-quality or irrelevant bases from the ends of sequencing reads.

- **Primary reasons:** To remove sequencing errors, adapter contamination, and low-quality bases, which can affect downstream analyses and interpretation of the data.

- **Types:** quality-based trimming (removes bases with low-quality scores), adapter trimming (adapter sequences), and sliding window trimming (low-quality bases from the ends of reads).

- **Tools:** Trimmomatic, Fastp, and Sickle.

Contamination assessment

- **Definition:** Is a post-processing quality check that involves identifying and quantifying the presence of contamination in the sequencing data.

- **Sources:**

a. **Host contamination**: occurs when the genomic DNA of the host organism is present in the sequencing data.
b. **Cross-sample contamination**: occurs when DNA from one sample is mixed with DNA from another sample during library preparation or sequencing.
c. **Reference contamination**: occurs when the reference genome used for analysis is contaminated with foreign DNA.

- **Tools:** Kraken, Kraken2, and BlobTools.

These tools use a variety of methods, such as taxonomic classification and mapping to a reference genome, to identify and quantify contamination in the sequencing data.

Quality control tools

- **Definition:** Tools are used to evaluate the quality of sequencing data and ensure that it is accurate and reliable based on their metrics.

- **Tools:** FastQC, MultiQC, Trimmomatic, Trim_galore, BBDuk,Sickle Kraken, ConFindr, CLC workbench, Geneious, checkM, cutadapt

- **Quality control metrics:** They are quantifiable measurements used to assess the quality of genomic data or processes.

Examples: read length, read quality, GC content, base composition, duplication rate, mapping rate etc.

Quality control metrics-1

- **Base composition:** The frequency of each base type (adenine, guanine, cytosine, and thymine) in the DNA sequence. Deviations from expected base composition can indicate contamination or other issues.
- **Duplication rate:** The percentage of reads that are identical to another read. High duplication rates can indicate biases or issues with library preparation.
- **Mapping rate:** The percentage of reads that can be aligned to a reference genome or transcriptome. Low mapping rates can indicate poor quality or contamination.
- **Alignment rate**: The percentage of reads that can be aligned to the reference genome. A high alignment rate is generally desirable, as it indicates that the reads are of high quality and are representative of the target genome.

Quality control metrics-2

- **Mean depth of coverage:** Average number of times a given base is sequenced. A high mean depth of coverage is generally desirable, as it indicates that the assembly is of high quality and has low error rates.
- **GC content:** Percentage of bases that are either guanine (G) or cytosine (C). In general, bacterial genomes have a relatively stable GC content, and deviations from the expected GC content can be a sign of contamination or other issues with the sample.
- **Genome completeness:** Measure of how much of the genome has been sequenced. In general, a genome assembly is considered to be complete when it covers at least 95% of the expected genome size.
- **N50:** Measure of the length of the largest contiguous sequence (contig) in an assembly. A high N50 value is generally desirable, as it indicates that the assembly is composed of longer contigs, which may be more accurate and easier to interpret.

Quality control metrics-3

- **Phred score:** Score of ≥20 corresponds to a base error rate of less than 1 in 10,000, which is generally considered to be of high quality.

- **Coverage:** Coverage is a measure of the depth of sequencing, or the number of times a given base is sequenced.

- **Error rate:** The error rate is the percentage of base calls that are incorrect. A low error rate is generally desirable, as it indicates that the reads are of high quality and can be trusted for downstream analysis.

General acceptable QC metric scores *

- Phred score: ≥20 (at least 99.99% accuracy)
- Coverage: >70
- Contamination: <5%
- Mapping rate: ≥90%
- Error rate: <1%
- GC content: within a narrow range around the expected value for the organism
- Genome completeness: ≥95%
- N50: ≥10,000 bp
- Mean depth of coverage: ≥30x
- Alignment rate: ≥90%
- Duplication rate: <5%

*** They are acceptable but not standards**

Quality check best practices -1

- **Tips for successful quality checks include:**

➢ Ensuring high-quality sequencing data

➢ Selecting appropriate quality check parameters

➢ Optimizing the quality check workflow to reduce errors and increase efficiency.

- **Quality check parameters should be tailored to the type of sequencing data and research question being addressed:**

➢ Different read lengths and quality scores may be required for different sequencing applications.

Quality check best practices -2

- **Quality check optimization involves:**

➢Selecting appropriate tools

➢Setting appropriate thresholds for quality check parameters

➢Ensuring the quality check workflow is efficient and effective.

- **Workflow management is also essential for successful quality checks:**

➢Creating a clear and reproducible workflow

➢Tracking sample status throughout the workflow

➢Documenting all quality check results and decisions

Summary

- Quality checks are essential for ensuring the accuracy and reliability of genomics data in surveillance of bacterial antimicrobial resistance.
- By identifying and addressing errors and biases in the data, quality checks can help researchers and public health officials make more informed decisions and take more effective actions.
- Successful quality checks require careful selection of quality check parameters, optimization of the quality check workflow, and effective workflow management.
- Quality checks play a critical role in the future of genomics surveillance of bacterial antimicrobial resistance, and will continue to be important as new sequencing technologies and analytical methods are developed.

References

- Andrews, S. (2010). FastQC: A quality control tool for high throughput sequence data. Bioinformatics, 32(19), 2503-2505.
- Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics, 30(15), 2114-2120.
- Bushnell, B. (2014). BBMap: A fast, accurate, splice-aware aligner. Lawrence Berkeley National Lab (LBNL), Berkeley, CA (United States).
- Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: Quality assessment tool for genome assemblies. Bioinformatics, 29(8), 1072-1075.
- Wood, D. E., & Salzberg, S. L. (2014). Kraken: Ultrafast metagenomic sequence classification using exact alignments. Genome Biology, 15(3), R46.
- Jia, B., Raphenya, A. R., Alcock, B., Waglechner, N., Guo, P., Tsang, K. K., ... & McArthur, A. G. (2017). CARD 2017: Expansion and model-centric curation of the comprehensive antibiotic resistance database. Nucleic Acids Research, 45(D1), D566-D573.
- Kuczynski, J., Stombaugh, J., Walters, W. A., González, A., Caporaso, J. G., & Knight, R. (2012). Using QIIME to analyze 16S rRNA gene sequences from microbial communities. Current protocols in microbiology, 27(1), 1E.5.1-1E.5.20.
- Loman, N. J., & Pallen, M. J. (2015). Twenty years of bacterial genome sequencing. Nature Reviews Microbiology, 13(12), 787-794.
- Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. Bioinformatics, 30(14), 2068-2069.
- Weissman, J. S., Snyder, M., & Gerstein, M. B. (2009). Personalized genomic responses to cancer. New England Journal of Medicine, 360(24), 2492-2505.

Q & A

Open forum for discussion and questions from the audience.

THANK YOU