Introduction

High-throughput sequencing technologies have greatly increased the volume of data generated in biological research[1]. This surge in data is primarily due to advancements in sequencing technologies and the decreasing cost of sequencing, which have led to an exponential growth in the amount of data produced. Many datasets remain underutilized due to the lack of accessible, automated tools that can handle the complexity and scale of the data involved. This work introduces a suite of UNIX-based scripts that simplify the quality control, alignment, and analysis of sequencing data. These tools are designed to be practical, scalable, and adaptable, serving as a foundational component for high-throughput data processing in bioinformatics.

Methods

Tables

The scripts that aggregate FastQC data output comma-separated values (CSV) files for each sample processed. These CSV files are then combined into a single table that summarizes the quality control metrics for all samples.

This table summarizes the quality control checks for each sample, indicating whether specific metrics have passed or failed the QC criteria. The data shown reflects the initial rows from the qcsummary.csv file generated by the FastQC scripts. The `multifastqc.sh` script processes multiple samples in parallel, then combines the individual QC summaries into a single table.

Table 1. Quality Control Summary

Sample	Basic Stats	Seq Quality	Tile Quality	Seq Scores	Seq Content	GC Content	N Content	Seq Length	Dup Levels	Adapter C
Sample1	PASS	PASS	PASS	PASS	FAIL	PASS	PASS	WARN	PASS	PASS
Sample2	PASS	PASS	PASS	PASS	FAIL	PASS	PASS	WARN	PASS	PASS
Sample3	PASS	PASS	PASS	PASS	FAIL	PASS	PASS	WARN	PASS	PASS
Sample4	PASS	PASS	PASS	PASS	FAIL	PASS	PASS	WARN	PASS	PASS
Sample5	PASS	PASS	PASS	PASS	FAIL	WARN	PASS	WARN	PASS	PASS
Sample6	PASS	PASS	PASS	PASS	FAIL	WARN	PASS	WARN	PASS	PASS
Sample7	PASS	PASS	PASS	PASS	FAIL	WARN	PASS	WARN	PASS	PASS
Sample8	PASS	PASS	PASS	PASS	FAIL	WARN	PASS	WARN	PASS	PASS

Note: The QC summary data displayed is extracted from the first few entries of the CSV outputs of the FastQC analysis scripts.

Detailed Basic Statistics

This table shows detailed basic statistics for each sample as produced by another script in the FastQC suite. Below are the initial rows from the ‘basic$_{s}tats.csv$‘.

Table 2. Detailed Basic Statistics

Sample ID	Filename	File type	Encoding	Total Sequences	Total Bases	Sequence length	%GC
Sample1	Sample1$_{R}$1.$fastq.gz$	Conventional	Sanger/Illumina 1.9	8,037,876	606 Mbp	35-76	51
Sample2	Sample2$_{R}$1.$fastq.gz$	Conventional	Sanger/Illumina 1.9	7,862,535	592.8 Mbp	35-76	51
Sample3	Sample3$_{R}$1.$fastq.gz$	Conventional	Sanger/Illumina 1.9	8,083,218	609.5 Mbp	35-76	51
Sample4	Sample4$_{R}$1.$fastq.gz$	Conventional	Sanger/Illumina 1.9	7,989,349	602.4 Mbp	35-76	51
Sample5	Sample5$_{R}$1.$fastq.gz$	Conventional	Sanger/Illumina 1.9	8,037,876	606 Mbp	35-76	51
Sample6	Sample6$_{R}$1.$fastq.gz$	Conventional	Sanger/Illumina 1.9	7,862,535	592.8 Mbp	35-76	51
Sample7	Sample7$_{R}$1.$fastq.gz$	Conventional	Sanger/Illumina 1.9	8,083,218	609.5 Mbp	35-76	51
Sample8	Sample8$_{R}$1.$fastq.gz$	Conventional	Sanger/Illumina 1.9	7,989,349	602.4 Mbp	35-76	51

Note: This table includes detailed statistics for the first eight samples processed. Each entry corresponds to an output from the FastQC report files.

Competing interests

No competing interest is declared.

[1] Deniz D, Ozgur A, Stallings CL. Applications and Challenges of High Performance Computing in Biology: Parallel Sequence Alignment. *Int J Comput Biol Drug Des.* 2010;3(2):124-134. `https://academic.oup.com/bib/article/15/3/390/186219`

Acknowledgments

Thank you to my labmates in the Palomero Lab for their guidance and advice.

References

B. Langmead and S. L. Salzberg, “Fast gapped-read alignment with Bowtie 2,” *Nature Methods*, vol. 9, no. 4, pp. 357–359, 2012. DOI: 10.1038/nmeth.1923 [URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3322381/].

R. Giancarlo, S. E. Rombo, and F. Utro, “Compressive biological sequence analysis and archival in the era of high-throughput sequencing technologies,” *Briefings in Bioinformatics*, vol. 15, no. 3, pp. 390–406, 2014. DOI: 10.1093/bib/bbt088 [URL: https://doi.org/10.1093/bib/bbt088].

J. K. Bonfield, “CRAM 3.1: advances in the CRAM file format,” *Bioinformatics*, vol. 38, no. 6, pp. 1497–1503, 2022. DOI: 10.1093/bioinformatics/btac010 [URL: https://doi.org/10.1093/bioinformatics/btac010].

H. Li et al., “The Sequence Alignment/Map format and SAMtools,” *Bioinformatics*, vol. 25, no. 16, pp. 2078–2079, 2009. DOI: 10.1093/bioinformatics/btp352 [URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2723002/].

A. Peste, A. Vladu, E. Kurtic, C. H. Lampert, and D. Alistarh, “CrAM: A Compression-Aware Minimizer,” *arXiv preprint arXiv:2207.14200*, 2022. [URL: https://arxiv.org/abs/2207.14200].

R. Nakato, T. Itoh, and K. Shirahige, “DROMPA: easy-to-handle peak calling and visualization software for the computational analysis and validation of ChIP-seq data,” *Genes & Cells*, vol. 18, no. 7, pp. 589–601, 2013. DOI: 10.1111/gtc.12058 [URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3738949/].