ISSN: 0975-8585

Research Journal of Pharmaceutical, Biological and Chemical Sciences

Integrative RNA-Seq Pipeline for Transcriptome Profiling of Colorectal Cancer.

Preeti Mate, Sakshi Indulkar, Pratiksha Thakre, Pratiksha Bhoi, and Sharddha Ranpise*.

Dr. D. Y. Patti Arts, Commerce and Science College, Sant Tukaram Nagar, Pimpri, Pune, Maharashtra, India, 411018.

ABSTRACT

Colorectal cancer (CRC) is a major global health concern, ranking among the leading causes of cancer-related deaths. Despite advancements in treatment, challenges in early detection and understanding of its complex molecular mechanisms persist. This project focuses on transcriptome profiling of CRC using RNA sequencing (RNA-Seq), a high-throughput technology that enables precise analysis of gene expression. The workflow includes quality assessment of raw FastQ files using FastQC, trimming with Trimmomatic, alignment with HISAT2, quantification using FeatureCounts, and differential gene expression analysis via DESeq2. MultiQC is used for integrated quality reporting and visualization. Through this pipeline, we aim to identify differentially expressed genes and pathways that contribute to CRC progression, offering insights into potential diagnostic biomarkers and therapeutic targets, and enhancing the overall understanding of colorectal cancer biology.

Keywords: Colorectal cancer, RNA-Seq, Transcriptome profiling, differential gene expression, FastQC, Trimmomatic, HISAT2, FeatureCounts, DESeq2, MultiQC

**Corresponding author*

ISSN: 0975-8585

INTRODUCTION

Colorectal cancer (CRC) remains a critical global health concern, representing the third most diagnosed malignancy and the second leading cause of cancer-related mortality worldwide (WHO, 2022). In 2020 alone, CRC accounted for approximately 1.9 million new cases and 935,000 deaths globally (Sung et al., 2021). A multitude of factors—including dietary patterns, sedentary lifestyle, and an aging population—have contributed to the increasing incidence of CRC across both developed and developing nations (Arnold et al., 2017). Despite advancements in screening and therapeutic strategies, the prognosis for advanced CRC remains poor, often due to late-stage diagnosis, tumor heterogeneity, and resistance to conventional therapies (Dekker et al., 2019).

At the molecular level, CRC is characterized by a complex interplay of genetic mutations, epigenetic alterations, and dysregulated signaling pathways that promote uncontrolled cell growth, invasion, and metastasis (Fearon & Vogelstein, 1990; Guinney et al., 2015). Traditional investigative methods—such as histopathological analysis and targeted molecular assays—offer limited resolution in capturing the transcriptomic complexity of CRC.

High-throughput RNA sequencing (RNA-Seq) has emerged as a transformative technology in transcriptomics, offering a comprehensive, unbiased assessment of gene expression, alternative splicing events, and novel transcript discovery (Wang et al., 2009; Stark et al., 2019). Unlike microarray technologies, RNA-Seq is probe-independent, thereby allowing the detection of previously unannotated transcripts and isoforms (Marguerat & Bähler, 2010). In cancer research, RNA-Seq facilitates the discovery of diagnostic biomarkers, therapeutic targets, and mechanisms of drug resistance (Hrdlickova et al., 2017).

This study presents a robust RNA-Seq analysis pipeline designed for transcriptomic profiling of colorectal cancer. Utilizing public datasets from the European Nucleotide Archive (ENA), the pipeline incorporates multiple quality control, alignment, quantification, and statistical analysis tools. The primary objective is to identify differentially expressed genes (DEGs) and pathways involved in CRC progression, thereby contributing to biomarker discovery and personalized medicine strategies.

MATERIALS AND METHODS

Data Collection

RNA-Seq datasets specific to colorectal cancer were retrieved from the ENA using keywords such as "colorectal cancer," "RNA-Seq," and "human." Accession numbers were used to download raw FASTQ files representing tumor and control samples.

Quality Control and Preprocessing

Initial quality assessment of raw reads was conducted using **FastQC** (Andrews, 2010) to identify low-quality bases, adapter contamination, and other artifacts. Subsequently, **Trimmomatic** and **Fastp** were used for adapter trimming and removal of low-quality sequences (Bolger et al., 2014). Post-trimming quality was re-evaluated with FastQC.

Alignment to the Human Reference Genome

Trimmed reads were aligned to the GRCh38 human reference genome using **HISAT2**, a spliced aligner optimized for RNA-Seq data (Kim et al., 2019). The resulting SAM/BAM files were indexed for downstream quantification.

Quantification of Gene Expression

Feature Counts was employed to generate a gene-level count matrix from the aligned BAM files using GTF annotation (Liao et al., 2014). This matrix formed the basis for differential expression analysis.

ISSN: 0975-8585

Differential Expression Analysis

Gene expression differentials between tumor and normal tissues were determined using **DESeq2**, an R package that models RNA-Seq count data using the negative binomial distribution (Love et al., 2014). Significantly altered genes were identified based on fold-change thresholds and adjusted p-values (FDR < 0.05).

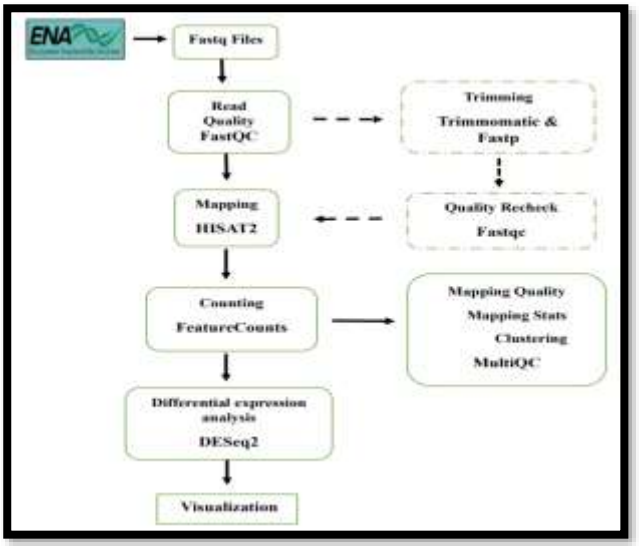


Figure: RNA-Seq Data Analysis Workflow

Data Integration and Visualization

MultiQC was used to compile and visualize QC reports across multiple tools and samples (Ewels et al., 2016). Volcano plots, read coverage histograms, and paired-end alignment summaries were generated for comprehensive data visualization.

RESULTS AND DISCUSSION

Read Quality Assessment

FastQC analysis indicated acceptable read quality across samples, with mean Phred scores exceeding 30 across most read positions. GC content and sequence duplication levels were within acceptable ranges.

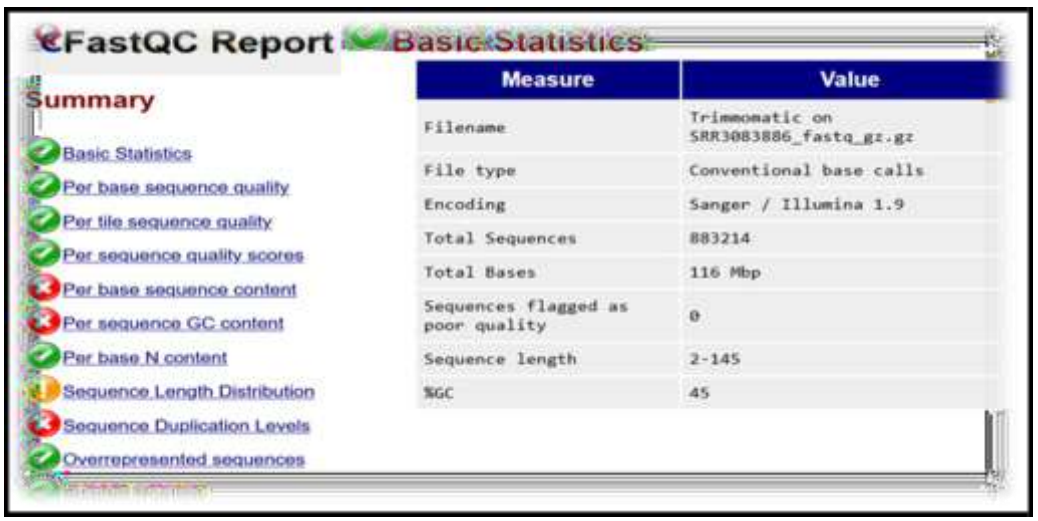


Figure 2: Basic Statistic of FastQC Report

ISSN: 0975-8585

Trimming and Filtering

Trimmomatic efficiently removed adapter sequences and filtered low-quality bases. The average sequence length post-trimming remained consistent, and short reads (<20 bp) were eliminated to minimize alignment errors.

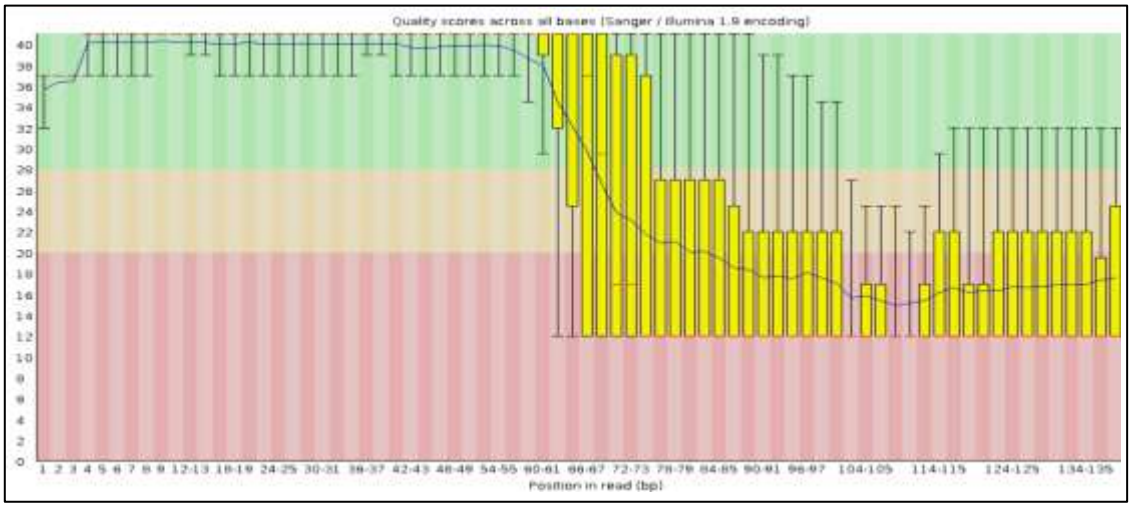


Figure 3: The Box Whisker Plot

Alignment Statistics

Using **HISAT2**, over 90% of reads aligned to the reference genome. BAM files indicated consistent mapping quality scores and uniform coverage across exonic regions, confirming successful transcriptome capture.

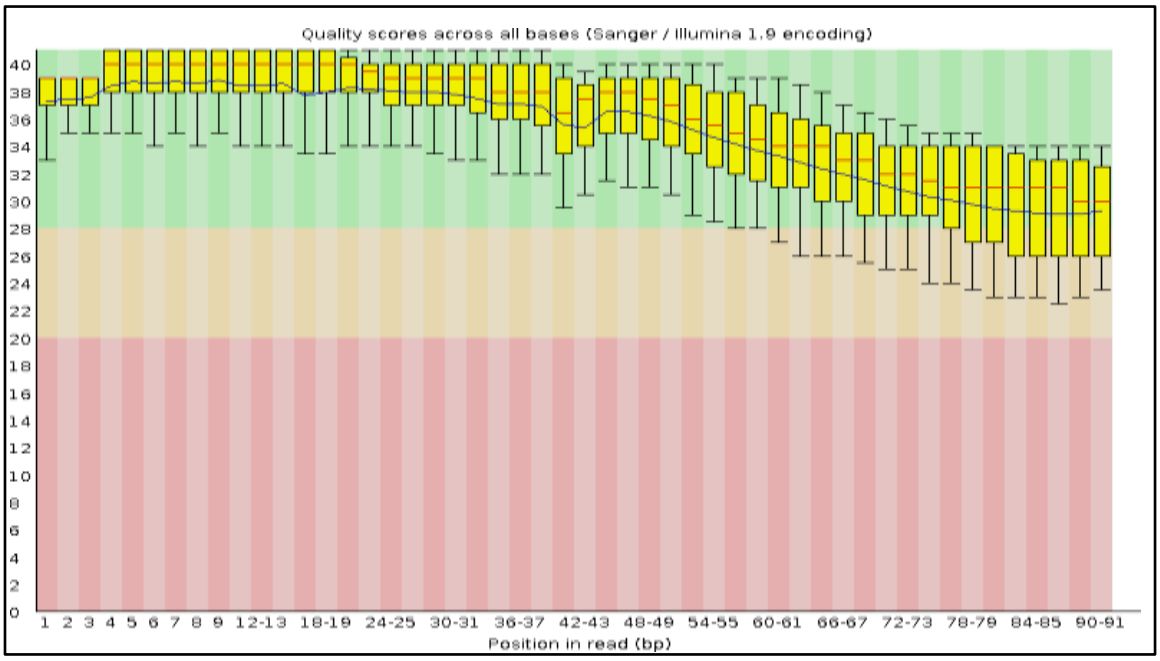


Figure 4: Quality Recheck across reads.

Gene Count Generation

FeatureCounts yielded high-quality count matrices with minimal ambiguous or unassigned reads. Strand-specific and paired-end sequencing options were accounted for during count assignment.

ISSN: 0975-8585

Quality Aggregation via MultiQC

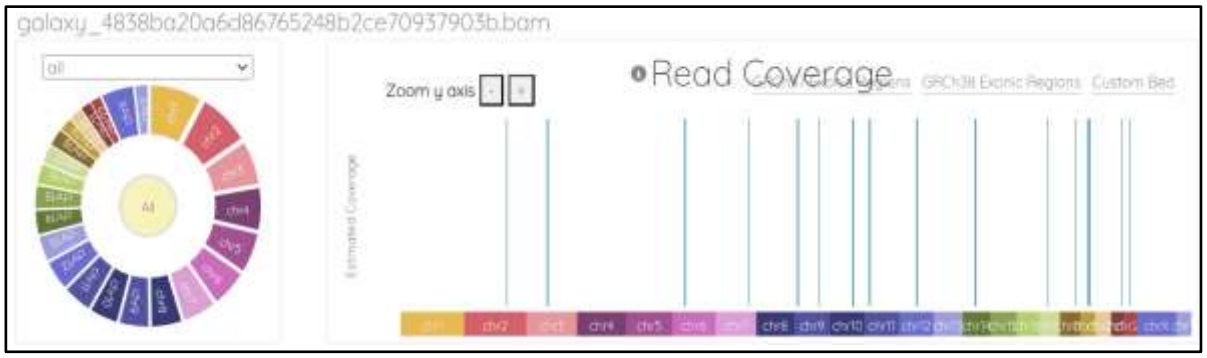
MultiQC provided a consolidated view of QC metrics from FastQC, HISAT2, and FeatureCounts, enabling streamlined interpretation of dataset integrity and processing steps.

Differential Expression Insights

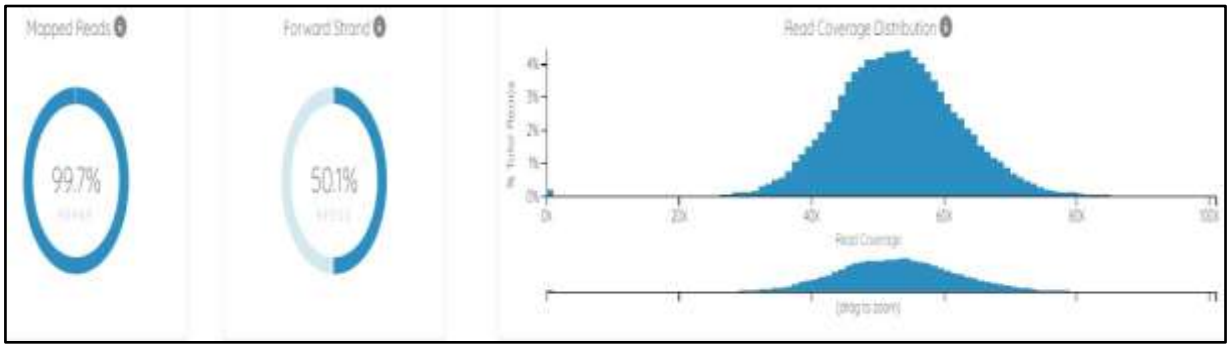
DESeq2 identified multiple DEGs with significant expression shifts in CRC samples. Volcano plots illustrated gene expression changes, highlighting both upregulated oncogenes and downregulated tumor suppressors.

BAM File Visualization

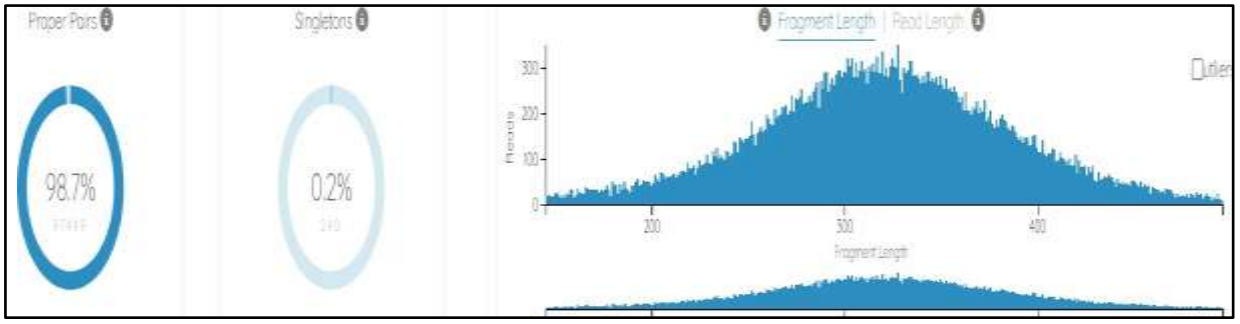
Using bam.iobio.io, read coverage was visualized across chromosomes, demonstrating consistent paired-end mapping and alignment depth. These visualizations affirmed the quality and reliability of the mapping process for downstream variant detection and transcriptomic analysis.



a. Chromosomal Read Coverage from RNA-Seq Alignment to Reference Genome (GRCh38)

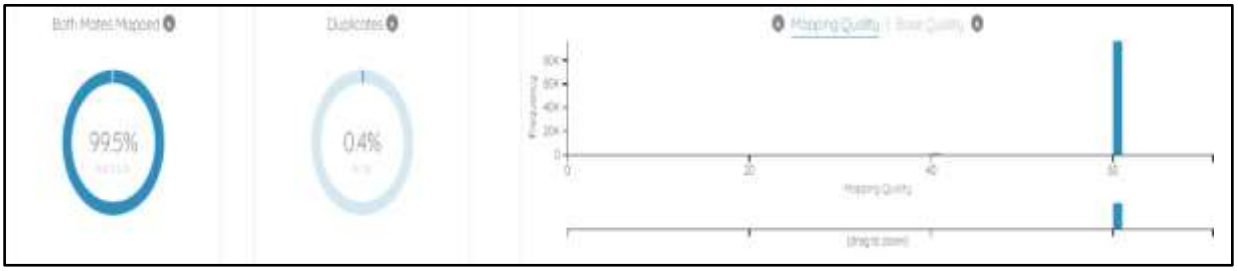


b. Doughnut chart of reads mapped to the reference genome



c. Paired-end sequencing and accurate alignment

ISSN: 0975-8585



d. Rate of paired-end sequencing Fig4: Bam File Visualization

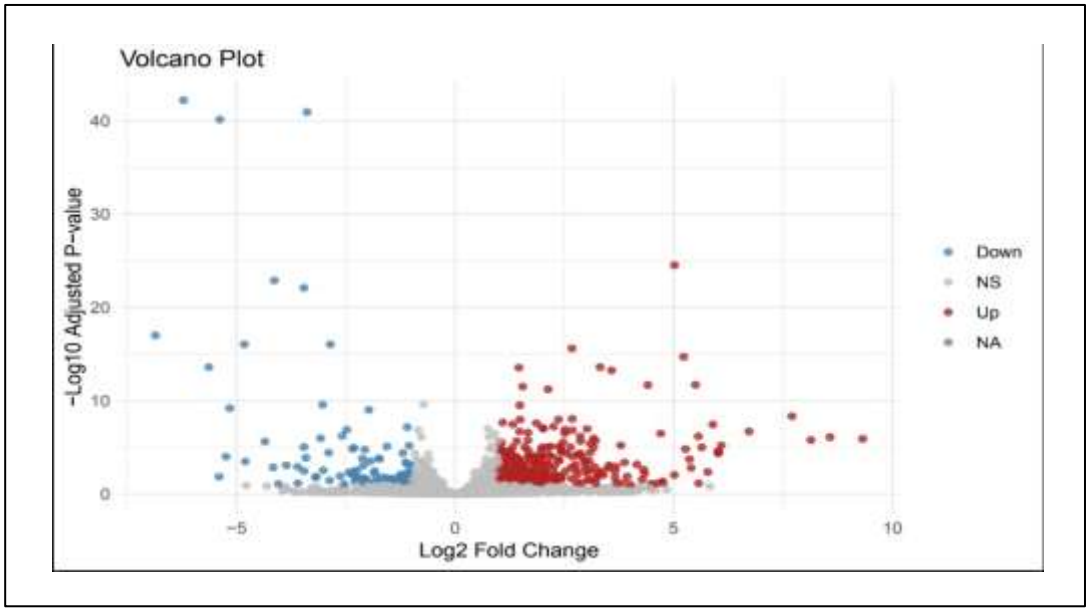


Figure 5: Volcano plot of log2 fold change against the –log10 adjusted p-value

CONCLUSION

This study demonstrates the application of a comprehensive RNA-Seq pipeline to profile gene expression in colorectal cancer. Utilizing a combination of open-source tools—FastQC, Trimmomatic, HISAT2, FeatureCounts, DESeq2, and MultiQC—we identified transcriptomic alterations and potential biomarkers relevant to CRC progression. The pipeline's modularity, scalability, and reproducibility render it adaptable for other transcriptomic studies, including those targeting different cancer types or integrated multi-omics analyses.

REFERENCES

[1] Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F. (2021). Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries.CA: A Cancer Journal for Clinicians, 71(3), 209–249.

[2] Kim, D., Paggi, J. M., Park, C., Bennett, C., & Salzberg, S. L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nature Biotechnology, 37(8), 907–915.

[3] Dekker, E., Tanis, P. J., Vleugels, J. L., Kasi, P. M., & Wallace, M. B. (2019). Colorectal cancer. The Lancet, 394(10207), 1467–1480.

[4] Pertea, M., Kim, D., Pertea, G. M., Leek, J. T., & Salzberg, S. L. (2016). Transcript- level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. Nature Protocols, 11(9), 1650–1667.

[5] Ewels, P., Magnusson, M., Lundin, S., & Käller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics, 32(19), 3047–3048.

[6] Anders, S., Pyl, P. T., & Huber, W. (2015). HTSeq—a Python framework to work with high-throughput sequencing data. Bioinformatics, 31(2), 166–169.

[7] Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics, 30(15), 2114–2120.
[8] Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology, 15(12), 550.
[9] Liao, Y., Smyth, G. K., & Shi, W. (2014). FeatureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics, 30(7), 923– 930.
[10] Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., ... & Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nature Biotechnology, 28(5), 511–515.
[11] Zhang, J., Baran, J., Cros, A., Guberman, J. M., Haider, S., Hsu, J., ... & Kasprzyk, A. (2011). International Cancer Genome Consortium Data Portal—a one-stop shop for cancer genomics data.Database, 2011, bar026.
[12] Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics.Nature Reviews Genetics, 10(1), 57–63
[13] Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., & Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: A Cancer Journal for Clinicians, 68(6), 394–424.
[14] National Center for Biotechnology Information (NCBI) & European Nucleotide Archive (ENA). Sequence Read Archive (SRA) and ENA Databases.
[15] World Health Organization (WHO). Colorectal cancer fact sheet.