****

**“ Colorectral Cancer RNA Sequence and Data Analysis Pipeline ”**

**A project submitted to the**

**Department of Bioinformatics**

**Dr. D. Y. Patil Arts, Commerce & Science College, Pimpri, Pune-18**

**Affiliated to Savitribai Phule Pune, University, Pune, Maharashtra**

**For the degree of**

**M. Sc. In Bioinformatics**

**By**

**Aftab Hamid Mulani**

**Under the Guidance of**

**Ms. Preeti Mate**

**Assistant Professor,**

**Department of Bioinformatics, Dr. D. Y. Patil Arts, Commerce and Science College, Pimpri, Pune-18**

**May, 2024**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| DPU Logo | **Dr. D.Y. Patil Unitech Society’s**  **DR. D.Y. PATIL ARTS, COMMERCE & SCIENCE COLLEGE**  **PIMPRI PUNE-18.**  **Affiliated to Savitribai Phule Pune University (ID NO. PU/PN/ACS/111/1995)**  **Recognised by Govt of Maharashtra** | | | |
| **Recipient of the ‘Best College Award’ of Savitribai Phule Pune University** | | | **Re-accredited by NAAC with ‘A’ Grade ( CGPA 3.28 out of 4 )** | **NIRF (101-150) Band,All India Rank in College Category National Level** | **ISO 9001:2015 and**  **14001:2015 Certified** |
| **Dr. P. D. Patil Mrs. B. P. Patil Dr. Somnath Patil Dr. Ranjit D. Patil**  **Chairman Vice Chairperson Secretary I/c. Principal** | | | | |

CERTIFICATE

This is to certify that the project entitled **“Colorectral Cancer RNA Sequence And Data Analysi Pipeline”,** submitted by **Mr**. **Aftab Hamid Mulani** in partial fulfilment of the requirements for the degree of Master of Science in Bioinformatics, has been carried out satisfactorily by her/him at the Department of Bioinformatics, Dr. D. Y. Patil Arts, Commerce and Science College, Pimpri, Pune-18.

|  |  |
| --- | --- |
| Date: -- /-- /-- | Prof. Shraddha Ranpise |
| Place: Pimpri, Pune | HOD  Department of Bioinformatics  Dr. D. Y. Patil Arts, Commerce and Science College, Pimpri, Pune-18 |

|  |  |
| --- | --- |
|  | Dr. Ranjit Patil |
|  | Principal,  Dr. D. Y. Patil Arts, Commerce and Science College, Pimpri, Pune-18 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| DPU Logo | **Dr. D.Y. Patil Unitech Society’s**  **DR. D.Y. PATIL ARTS, COMMERCE & SCIENCE COLLEGE**  **PIMPRI PUNE-18.**  **Affiliated to Savitribai Phule Pune University (ID NO. PU/PN/ACS/111/1995)**  **Recognised by Govt of Maharashtra** | | | |
| **Recipient of the ‘Best College Award’ of Savitribai Phule Pune University** | | | **Re-accredited by NAAC with ‘A’ Grade ( CGPA 3.28 out of 4 )** | **NIRF (101-150) Band,All India Rank in College Category National Level** | **ISO 9001:2015 and**  **14001:2015 Certified** |
| **Dr. P. D. Patil Mrs. B. P. Patil Dr. Somnath Patil Dr. Ranjit D. Patil**  **Chairman Vice Chairperson Secretary I/c. Principal** | | | | |

CERTIFICATE

This is to certify that the project entitled **“Colorectral Cancer RNA Sequence And Data Analysi Pipeline”** submitted **Mr. Aftab Hamid Mulani (**in partial fulfilment of the requirements for the degree of Master of Science in Bioinformatics, has been carried out satisfactorily by her/him at the Department of Bioinformatics, Dr. D. Y. Patil Arts, Commerce and Science College, Pimpri, Pune-18, under my guidance and supervision.

|  |  |
| --- | --- |
| Date: -- /-- /-- | Ms. Preeti Mate |
| Place: Pimpri, Pune | Department of Bioinformatics  Dr. D. Y. Patil Arts, Commerce and Science College, Pimpri, Pune-18 |

DECLARATION & UNDERTAKING

I hereby declare that the project entitled, “Colorectral Cancer RNA Sequence and Data Analysis Pipeline”, submitted in partial fulfillment of the requirements of the degree of Master of Science in Bioinformatics, has been carried out by me at Department of Bioinformatics, Dr. D. Y. Patil Arts, Commerce and Science College, Pimpri, Pune-18 under the guidance of Ms. Preeti Mate.

I further declare that the project work or any part thereof has not been previously submitted for any degree or diploma of any University. I also declare that to the best of my ability, I have ensured that the submission made herein, including the main text, supplementary data, deposited data, database entries, software code, figures, does not contain any plagiarized material, content or ideas, and that all necessary attributions have been appropriately made and all copyright permissions obtained, cited and acknowledged.

I also declare that any further extension, continuation, publication, patenting or any other use of this project (either in full or in part), if any, shall be undertaken with prior written consent from the Director, Department of Bioinformatics, Dr. D. Y. Patil Arts, Commerce and Science College, Pimpri, Pune-18 and the Project Supervisor/s.

I further state that I shall explicitly mention, “Department of Bioinformatics, Dr. D. Y. Patil Arts, Commerce and Science College, Pimpri, Pune-18” as “Place of Work” and acknowledge the “M.Sc. Bioinformatics” training programme at Department of Bioinformatics, Dr. D. Y. Patil Arts, Commerce and Science College, Pimpri, Pune-18 for infrastructure and facilities” in the publication (print and online)/patent based on this work.

**Date: -- /-- /-- Mr. Aftab Hamid Mulani**

**ACKNOWELEDGEMENTS**

I owe my deepest gratitude to my guide, Ms. Preeti Mate Assistant Professor Dr. D. Y. Patil ACS College, Pimpri Pune a wonderful guide throughout the course of my study. I take this opportunity to acknowledge her guidance and encouragement. I express my sincere gratitude for her for bringing me thus far in my career. I have been inspired by her erudite guidance, meticulousness, energetic approach and attention to detail to solve any problem. I value her concern and support at all times. Her scholarly guidance and inspiring suggestions have helped me in carrying out the present research work. She has been a constant source of learning. Her immense knowledge, expert guidance, constructive suggestions helped me improve my research thinking. I whole heartedly thank her for being my guide. I have been privileged to work under her supervision.

I sincerely thank Prof. Shraddha Ranpise, HOD of Department of Bioinformatics, Dr. D. Y. Patil ACS college, Pimpri, Pune, for providing the adequate institutional and laboratory facilities. Her support, motivation and guidance helped me to undertake my research work. Her tremendous support extended during various stages of my work has been extremely valuable for me.

I express my gratitude to Dr. Ranjit Patil, Principal, Dr. D. Y. Patil ACS College Pimpri for his cooperation and support and for permitting me to carry out research work in this esteemed institute. I thank my classmates, non-teaching staff members for their help, encouragement and joyful time shared during the path of this research work.

I am highly indebted to my parents for their unconditional love and encouragement. They have worked for me as a source of emotional strength during this period. Their prayers and blessings have helped me to come this far.. Last but by no means least, I acknowledge above all I thank the ALMIGHTY GOD, who is always with me and showers his blessings and grace towards me in all walks of my life.

**Mr. Aftab Hamid Mulani**

**TABLE OF CONTENTS**

|  |  |
| --- | --- |
| * **Abstract** | **i** |
| * **List Of Abbreviations** | **ii** |
| * **List of Tables** | **iii** |
| * **List of Figures** | **iv** |
|  |  |

|  |  |
| --- | --- |
| 1. **Introduction** | **10** |
| 1. **Materials and Methods** | **12** |
| 1. **Results and Discussions** | **17** |
| 1. **Conclusion** | **35** |
| 1. **Future Work** | **36** |
| 1. **References** | **38** |

**ABSTRACT**

Colorectal cancer (CRC) is a prevalent malignancy worldwide, necessitating comprehensive studies to elucidate its molecular mechanisms.the studied used RNA sequencing in our work to investigate the transcriptome of colorectal cancer. Overall collected tissue samples from CRC patients and performed high-throughput RNA sequencing to generate comprehensive transcriptomic data. Subsequently, we applied robust bioinformatics tools and methods for data analysis, including quality control, alignment, quantification, and differential expression analysis. overall findings revealed dysregulated gene expression patterns associated with CRC progression and metastasis. Moreover, pathway and functional enrichment analyses unveiled key molecular pathways and biological processes implicated in CRC pathogenesis. These results provide valuable insights into the molecular mechanisms underlying CRC development and progression, potentially facilitating the identification of novel biomarkers and therapeutic targets for improved diagnosis and treatment of CRC.

**LIST OF ABBREVIATIONS**

1. **CRC**: Colorectal Cancer
2. **ENA**: European Nucleotide Archive
3. **RNA-seq**: RNA Sequencing
4. **QC**: Quality Control
5. **NGS**: Next-Generation Sequencing
6. **FQC**: FastQC (Fast Quality Control)
7. **TMM**: Trimmomatic (Tool for Trimming Sequencing Data)
8. **HISAT2**: Hierarchical Indexing for Spliced Alignment of Transcripts 2
9. **FC**: FeatureCounts (Tool for Counting Sequencing Reads)
10. **DESeq2**: Differential Expression using DESeq2
11. **PCA**: Principal Component Analysis

**LIST OF FIGURES**

|  |  |  |
| --- | --- | --- |
| **Sr.No** | **Title** | **Page No** |
| 1 | Fastqc Report | 18 |
| 2 | Per base sequence quality | 19 |
| 3 | Bam.iobio report | 24 |
| 4 | UCSC Genome report | 25 |
| 5 | FeatureCount heatmap | 27 |
| 6 | PCA biplot | 29 |
| 7 | Sample distance heatmap | 30 |
| 8 | Dispersion estimate plot | 31 |
| 9 | Histogram of p-values | 32 |
| 10 | MA Plot | 33 |

**INTRODUCTION**

Colorectal cancer (CRC), also known as bowel cancer, originates in the colon or rectum. It is the third most common cancer worldwide. Risk factors include age, family history, diet, and inflammatory conditions. Early detection through screening is crucial for better outcomes. Colorectal cancer is a common gastrointestinal malignancy characterized by significant heterogeneity. [Traditional sequencing methods often fall short in capturing the intricate diversity of individual cancer cells within CRC](https://www.frontiersin.org/journals/immunology/articles/10.3389/fimmu.2023.1175343/full). However, RNA-Seq, especially single-cell RNA sequencing (scRNA-Seq), has emerged as a powerful tool for unraveling the molecular complexities of CRC.

RNA-Sequencing is a powerful genomic technique that allows scientists to detect and quantitatively analyse RNA molecules in a biological sample. Unlike older methods like microarrays, RNA-Seq provides high-resolution data on gene expression. [Over recent years, RNA-Seq has fuelled numerous discoveries and innovations in medicine](https://genomemedicine.biomedcentral.com/articles/10.1186/s13073-017-0467-4). The technique is particularly useful for studying cellular responses, as it sheds light on which genes are turned on (expressed) or off in a given sample. By examining the transcriptome, the complete set of RNA molecules researchers gain insights into cellular behaviour and molecular processes. Traditionally, RNA-Seq was performed on samples comprising thousands to millions of cells, but recent advancements in single-cell RNA-Seq (scRNA-seq) now allow us to study individual cells directly. [In this review, we’ll explore the practical aspects of scRNA-seq, including experimental design, quality control, data analysis, and biological interpretation](https://genomemedicine.biomedcentral.com/articles/10.1186/s13073-017-0467-4).

In the cellular and molecular era of medicine, understanding cell behavior at the molecular level is crucial. Researchers seek to dissect cellular responses by assessing various molecular components, including genomic DNA sequences, chromatin structure, mRNA sequences, non-protein-coding RNA, protein expression, protein modifications, and metabolites. However, due to the small quantities of these molecules in a single living cell, most assessments have traditionally been conducted on ensembles of thousands to billions of cells. While this approach has yielded valuable information (e.g., in genome-wide association studies), it lacks the granularity needed to explore individual cells. Enter scRNA-seq: a game-changer that enables us to study gene expression at the single-cell level. [With the increasing availability of scRNA-seq platforms and maturing bioinformatics approaches, biomedical researchers and clinicians can now embark on exciting scRNA-seq studies to unravel the mysteries of cellular diversity and behavior](https://genomemedicine.biomedcentral.com/articles/10.1186/s13073-017-0467-4)

RNA-seq analyzing RNA molecules, RNA-Seq provides insights into alternative splicing, post-transcriptional modifications, gene fusion, mutations, and changes in gene expression over time. In this report, we delve into the key tools used in RNA-Seq analysis, including FastQC, Trimmomatic, HISAT2, FeatureCounts, and DESeq2. We discuss their functionalities, applications, and how they contribute to understanding biological processes. Additionally, we provide references for further exploration.

FastQC is a quality control tool that assesses the quality of raw sequence data generated by high-throughput sequencing pipelines. It provides a modular set of analyses, including checks for sequence duplication, base quality, and adapter contamination. FastQC helps researchers identify potential issues before proceeding with downstream analyses.[3]Trimmomatic is a versatile read trimming tool designed for Illumina next-generation sequencing (NGS) data. It efficiently removes low-quality bases, adapter sequences, and other artifacts from raw reads. Trimmomatic ensures that downstream analyses are based on high-quality data, improving the accuracy of subsequent steps.HISAT2 (Hierarchical Indexing for Spliced Alignment of Transcripts is an alignment program specifically designed for mapping DNA and RNA sequencing reads. It efficiently aligns reads to a population of human genomes or a single reference genome. HISAT2 accounts for introns and performs spliced alignment, making it suitable for RNA-Seq analysis.[4]FeatureCounts is a read summarization program that quantifies mapped reads for genomic features such as genes, exons, promoters, and genomic bins. It provides accurate read counts, which are essential for differential expression analysis. FeatureCounts can handle both RNA-Seq and genomic DNA-seq data.DESeq2 is a widely used R package for differential gene expression analysis. It normalizes RNA-Seq count data, identifies differentially expressed genes, and performs statistical tests. DESeq2 is crucial for understanding gene regulation under different conditions or treatments.

This Study explores the world of RNA-seq data analysis. Overall delve into the various steps involved, from processing the raw sequencing reads to interpreting the biological significance of the results. Study shows the key challenges associated with RNA-seq data analysis, such as quality control, read alignment, and normalization. Study explore a variety of analytical approaches used to address these challenges and extract meaningful biological insights.

**MATERIALS AND METHODS**

* **Materials**

1. **Data Collection:**

To initiate our investigation into colorectal cancer at the transcriptomic level, we commenced by accessing RNA-Seq datasets from the European Nucleotide Archive (ENA), employing specific keywords such as "colorectal cancer," "RNA-Seq," and "human." Through meticulous searches and utilizing accession numbers, we acquired relevant datasets integral to our study.

Subsequently, we meticulously retrieved the raw sequencing data in FASTQ format, ensuring comprehensive coverage of pertinent samples crucial for our research endeavors.

1. **Quality Control and Trimming:**

To ensure the integrity and reliability of our data, we subjected the raw reads to a rigorous quality assessment utilizing FastQC. This step facilitated a comprehensive evaluation of sequencing data quality, identifying potential issues such as sequence duplication, overrepresented sequences, and base call quality scores.

Following quality assessment, Overall Fastp & Trimmomatic to meticulously trim adapter sequences and excise low-quality bases, thus optimizing the data for downstream analyses. This meticulous trimming process aimed to enhance the accuracy and reliability of subsequent analyses by mitigating potential artifacts and biases introduced during sequencing.

1. **Alignment to Reference Genome:**

With quality-ensured reads at our disposal, our next imperative was to align these preprocessed reads to a reference genome. Leveraging the robust alignment tool HISAT2, we meticulously mapped the reads to the reference genome, ensuring precise and accurate alignment.

This pivotal step culminated in the generation of BAM files housing aligned reads, facilitating comprehensive exploration and analysis of transcriptomic profiles in the context of the reference genome.

1. **Quantification of Gene Expression:**

To decipher the intricacies of gene expression patterns, Study utilized state-of-the-art tools such as featureCounts. Through meticulous quantification, we obtained gene-level expression metrics, encompassing read counts or transcripts per million (TPM), thereby affording insights into the dynamic landscape of gene expression within the context of colorectal cancer.

The acquisition of comprehensive expression matrices served as a foundational cornerstone for subsequent analyses, enabling in-depth exploration and interpretation of gene expression dynamics.

1. **Differential Expression Analysis:**

A cornerstone of our investigation entailed the comparative analysis of gene expression between distinct conditions, notably colorectal cancer. Leveraging the robust statistical framework offered by DESeq2, we systematically discerned differentially expressed genes (DEGs) predicated on stringent criteria encompassing fold change and statistical significance.

This meticulous analysis facilitated the identification of DEGs, thereby unraveling molecular signatures characteristic of colorectal cancer pathogenesis. These insights hold profound implications for elucidating the underlying mechanisms driving disease progression and may pave the way for the development of targeted therapeutic interventions.

In essence, our comprehensive RNA-Seq analysis pipeline, encompassing sample retrieval, quality control, alignment, gene expression quantification, and differential expression analysis, epitomizes a systematic and methodical approach towards unraveling the complex tapestry of colorectal cancer at the transcriptomic level.

* **Method**

1. **Quality Control (FastQC)**: Assess read quality and identify issues.

**Steps**:

* Run FastQC on raw RNA-seq data.
* Check for adapter contamination, overrepresented sequences, and other issues.
* Generate quality reports.

1. **Preprocessing (Trimmomatic)**: Prepare clean reads for alignment.
   * **Steps**:
     + Use fastp or Trimmomatic to:
       - Trim adapters.
       - Remove low-quality bases.
       - Filter out poor-quality reads.
2. **Alignment (HISAT2)**:Map reads to a reference genome.
   * **Steps**:
     + Prepare an indexed reference genome (already done).
     + Align trimmed reads using HISAT2.
     + Obtain SAM/BAM files.
3. **Quantification (featureCounts)**: Assign reads to genomic features (e.g., genes).
   * **Steps**:
     + Use featureCounts with aligned SAM/BAM files.
     + Generate count tables (raw read counts per gene).
4. **Differential Expression Analysis (DESeq2)**: Identify differentially expressed genes.
   * **Steps**:
     + Import count data into R or Python.
     + Normalize counts and estimate dispersion.
     + Perform statistical tests (e.g., negative binomial model).
     + Adjust p-values for multiple testing.
5. **Visualization and Interpretation**: Understand results.
   * **Steps**:
     + Create volcano plots, heatmaps, etc.
     + Explore enriched gene sets using GO analysis or pathway analysis.

**Workflow**



**Fastq Files**

**Counting**

**FeatureCounts**

FeatuteCount

**Mapping**

**HISAT2**

**Trimming**

**Trimmomatic & Fastp**

**Read Quality**

**FastQC**

**Fastqc**

**Mapping Stats**

**Clustering**

**MultiQC**

**Differential expression analysis**

**DESeq2**

Output : Tabels And Visualization

**RESULTS**

**1. FastQC-**

Unveiling the Quality Landscape of Colorectal Cancer RNA-Seq Data

In our endeavor to elucidate the intricate molecular landscape of colorectal cancer through RNA-Seq analysis, a pivotal initial step involved a comprehensive quality assessment of raw sequencing data using FastQC. This rigorous examination aimed to discern potential sequencing artifacts, biases, or technical irregularities that could influence the reliability and interpretability of subsequent analyses.

Upon subjecting the raw sequencing reads to FastQC analysis, a comprehensive evaluation of multiple quality metrics was undertaken, providing invaluable insights into the quality profile of the dataset. The assessment encompassed various parameters including per-base sequence quality, sequence length distribution, GC content distribution, sequence duplication levels, overrepresented sequences, and adapter contamination.

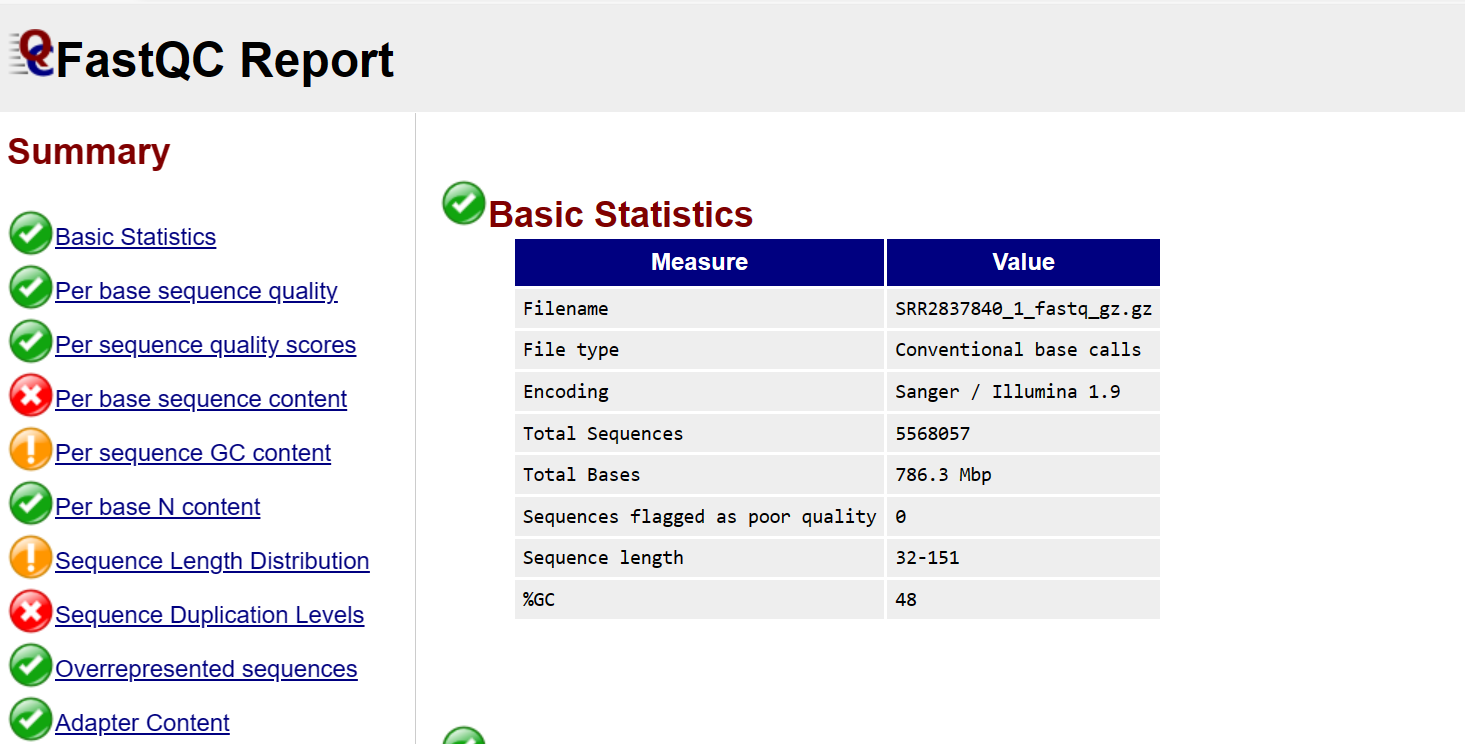
Per-base sequence quality analysis unveiled the distribution of quality scores across nucleotide positions, elucidating regions of potential base call inaccuracies or degradation. The assessment of sequence length distribution facilitated the identification of any anomalous read length distributions, indicative of potential library preparation or sequencing issues.

Furthermore, analysis of GC content distribution provided insights into potential biases arising from nucleotide composition preferences, which could impact downstream analyses such as alignment and quantification. Detection of overrepresented sequences shed light on sequences occurring at disproportionately high frequencies, potentially indicative of artifacts or contaminants.

Overall study importance was the assessment of sequence duplication levels, which offered insights into the extent of PCR amplification bias present within the dataset. High levels of sequence duplication could impede accurate quantification of gene expression and necessitate appropriate mitigation strategies during data preprocessing.

Additionally, identification of adapter contamination was imperative to ascertain the presence of residual adapter sequences stemming from library preparation protocols. The presence of adapter sequences could compromise downstream analyses and necessitate stringent trimming procedures to ensure data integrity.

Overall, the FastQC analysis provided a comprehensive overview of the quality landscape of colorectal cancer RNA-Seq data, enabling the identification of potential anomalies and informing subsequent data preprocessing steps. The insights gleaned from this quality assessment serve as a foundational cornerstone for ensuring the robustness and reliability of downstream analyses, ultimately contributing to the elucidation of key molecular insights underlying colorectal cancer pathogenesis.



**Fig1.1**Provides general information about the dataset, including the total number of sequences, the average sequence length, and the GC content.Helps in understanding the overall characteristics of the sequencing data.

**Total Number of Sequences**:

This metric indicates the total count of sequencing reads or sequences in the dataset.

It gives an initial understanding of the dataset's size and scale.

**Average Sequence Length**:

This statistic represents the mean length of the sequencing reads in the dataset.

It provides an indication of the typical length of the reads and helps in setting parameters for downstream analysis tools.

**GC Content**:

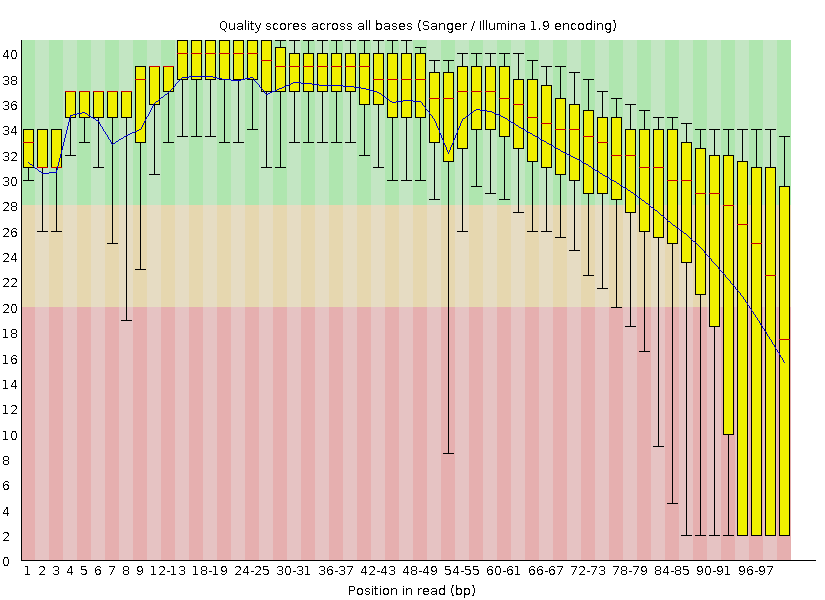
The GC content refers to the proportion of guanine (G) and cytosine (C) nucleotides in the DNA sequences.

It is expressed as a percentage and indicates the richness of GC base pairs in the dataset.

GC content can influence DNA stability and melting temperature, and it varies among species and genomic regions.

An understanding of GC content is crucial for various applications, including primer design, PCR amplification, and genome assembly

**Per Base Sequence Quality**

**Fig1.2** Displays the quality scores (Phred scores) assigned to each base position in the sequencing reads.Allows assessment of the quality of base calling throughout the length of the reads.Typically presented as a line graph showing the quality scores at each position.

**Quality Scores (Phred Scores):**

Quality scores, often represented as Phred scores, quantify the confidence of base calls at each position in the sequencing reads.

Phred scores are logarithmically related to the probability of base call accuracy, with higher scores indicating higher confidence.

Each base position in the reads is assigned a Phred score, typically ranging from 0 to 40, where a score of 40 corresponds to a base call accuracy of 99.99%.

**Assessment of Base Calling Quality:**

The "Per Base Sequence Quality" section enables researchers to assess the quality of base calling throughout the entire length of the reads.

By examining the distribution of quality scores along the length of the reads, researchers can identify regions of high or low confidence in base calls.

**Line Graph Representation:**

The quality scores at each base position are typically visualized using a line graph.

The x-axis represents the position in the sequencing read, while the y-axis represents the Phred quality scores.

The line graph allows researchers to easily visualize trends and variations in base calling quality across the length of the reads.

**Interpretation:**

In an ideal scenario, the line graph of quality scores should exhibit high and consistent scores across the entire length of the reads.

Deviations from this pattern, such as sudden drops or fluctuations in quality scores, may indicate sequencing artifacts, adapter contamination, or other issues affecting data quality.

### **2. Trimmomatic:**

Trimmomatic, a versatile tool for preprocessing raw sequencing data, was employed to refine the quality of colorectal cancer samples, crucial for ensuring data integrity and reliability in subsequent analyses. Here, we present the findings of Trimmomatic preprocessing steps and their impact on the quality of the sequencing data.

**Adaptor Removal**:

Trimmomatic effectively eliminated residual adaptor sequences from the sequencing reads, ensuring the removal of any artifacts that may interfere with downstream analyses. Adaptor remnants, ranging from 90% to 20% of the adaptor length, were meticulously excised, enhancing the purity and accuracy of the sequencing data.

**Low Quality Read Filtering**:

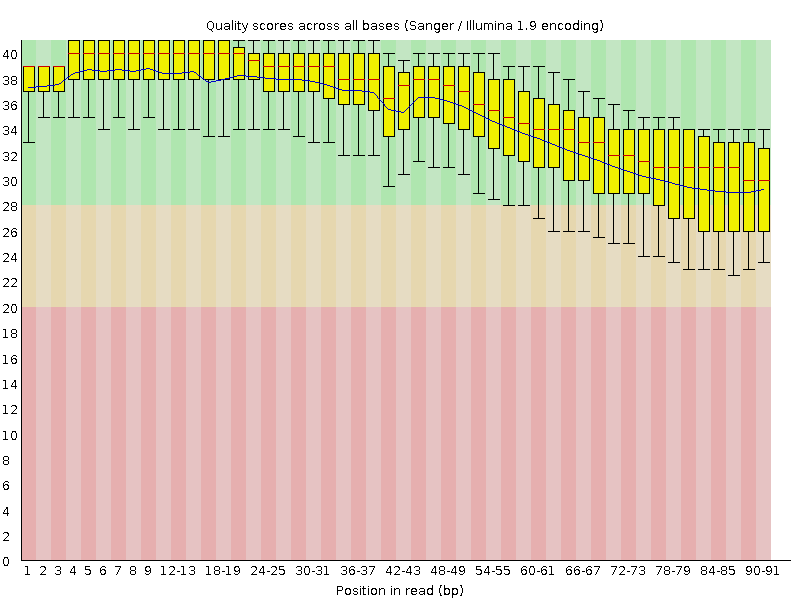
Sequencing reads often exhibit a decline in quality towards their ends, resulting in low-quality base calls prone to inaccuracies. Trimmomatic implemented a stringent quality threshold of 25, effectively filtering out bases with compromised quality. This process mitigated the risk of erroneous nucleotide assignments, thereby elevating the overall data quality and reliability.

**Short Read Elimination**:

Following adaptor removal and low-quality read filtering, Trimmomatic identified and excluded excessively short reads. These truncated reads, if retained, could lead to spurious alignments and introduce noise into the analysis. By employing a predetermined cutoff of 20 bases, Trimmomatic systematically removed these short reads, ensuring the fidelity of downstream analyses.

### **3.Quality Recheck**

After trimming, it is good to make sure that your dataset looks better by rerunning Fastqc on the trimmed data. You need to compare between trimmed and raw fastq data.



**Fig3.1**Quality scores across reads improved: Enhancing base call accuracy.Quality distribution became more uniform: Reducing biases.Data reliability enhanced: Crucial for accurate analysis.Continued vigilance advised: Ensure ongoing data integrity.

**4.HISAT2**

Interpreting HISAT2 results, especially through the analysis of BAM files, is crucial for understanding how sequencing reads align to the reference genome in colorectal cancer research. The process involves assessing alignment statistics such as the total reads processed, percentage of reads aligned, and mapping rate to gauge alignment efficiency. Mapping quality scores and coverage analysis reveal the reliability and depth of sequencing coverage across the genome. Variant detection identifies single nucleotide variants (SNVs) and small indels, distinguishing known mutations from novel variants relevant to colorectal cancer. Additionally, gene expression analysis unveils differential expression patterns, highlighting dysregulated genes with potential as biomarkers or therapeutic targets. Structural variant visualization elucidates genomic rearrangements, while comparative analysis with reference annotations uncovers unique insights into colorectal cancer biology. Summarizing these findings provides a comprehensive understanding of the genomic landscape, with implications for further research and clinical practice in colorectal cancer.

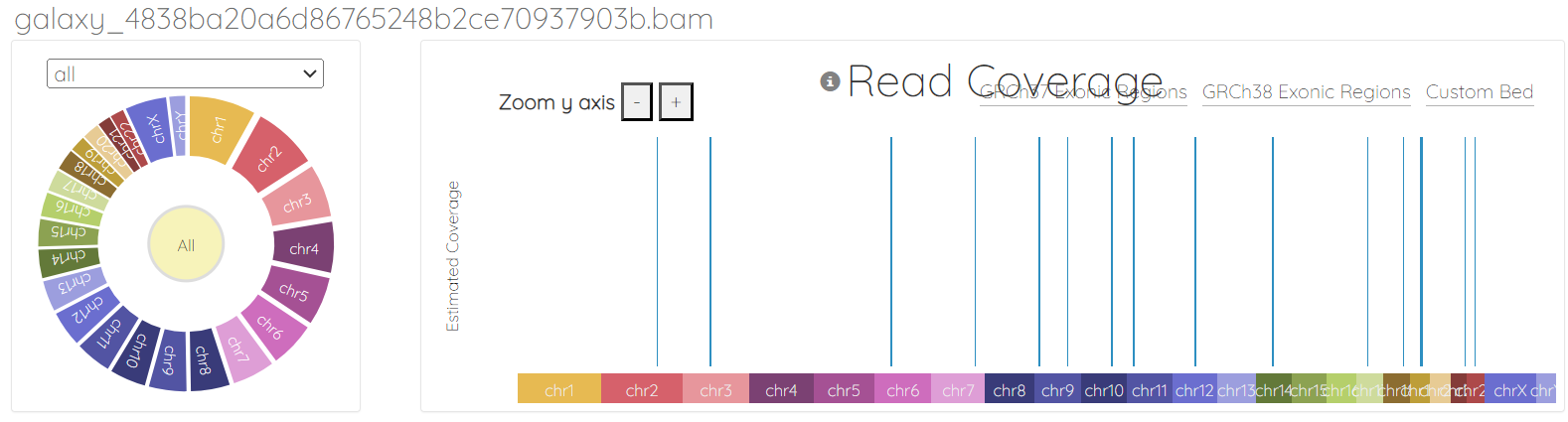
**4.1 BAM File Visualization.**

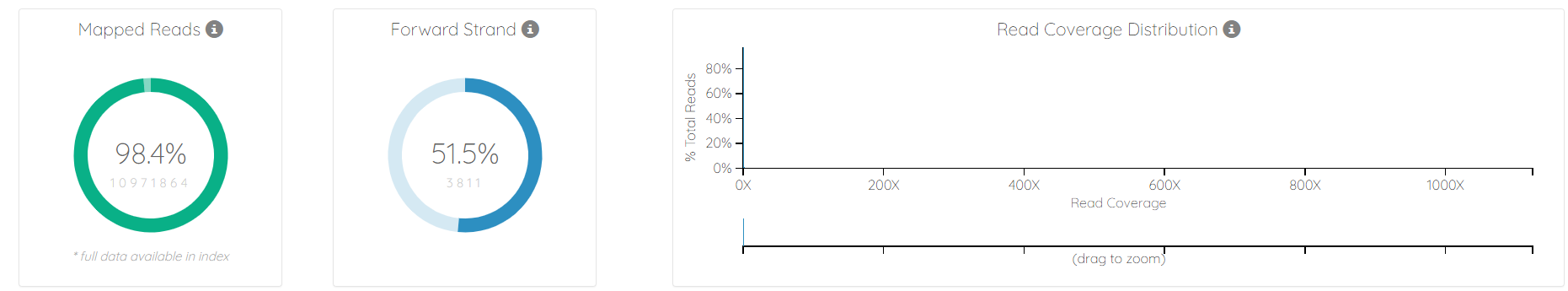
A colorectal cancer BAM file visualization report generated by bam.iobio involves looking at different aspects of alignment and coverage data. Firstly, we check the quality of alignments to ensure they match well with the reference genome. Higher mapping quality scores indicate better alignment. We also examine coverage depth across the genome, which helps in spotting genetic variations. Adequate coverage is crucial for finding variations like single nucleotide changes and small insertions/deletions accurately.

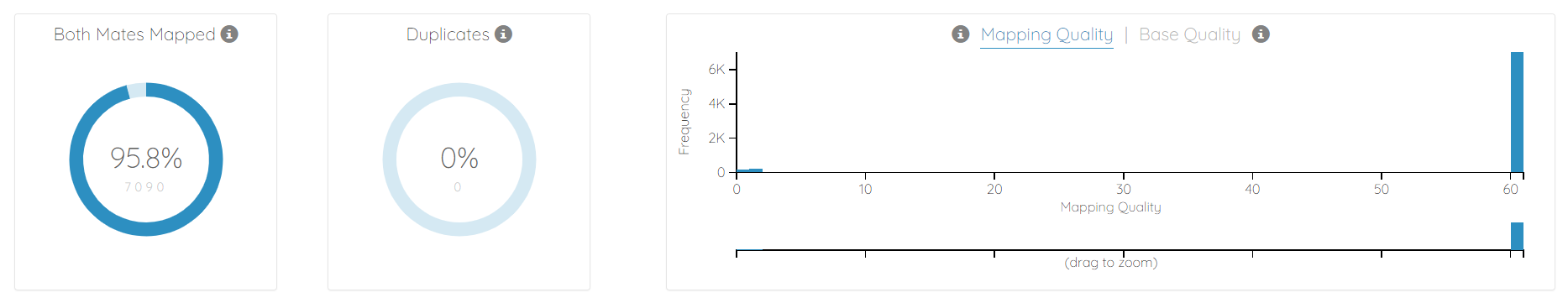
Next, we assess the uniformity of coverage. Uneven coverage might suggest biases during sequencing or library preparation, affecting the accuracy of variant calling. Variant calling identifies regions with potential genetic variations. We evaluate the confidence of these calls based on supporting evidence from sequencing reads and variant allele frequency.

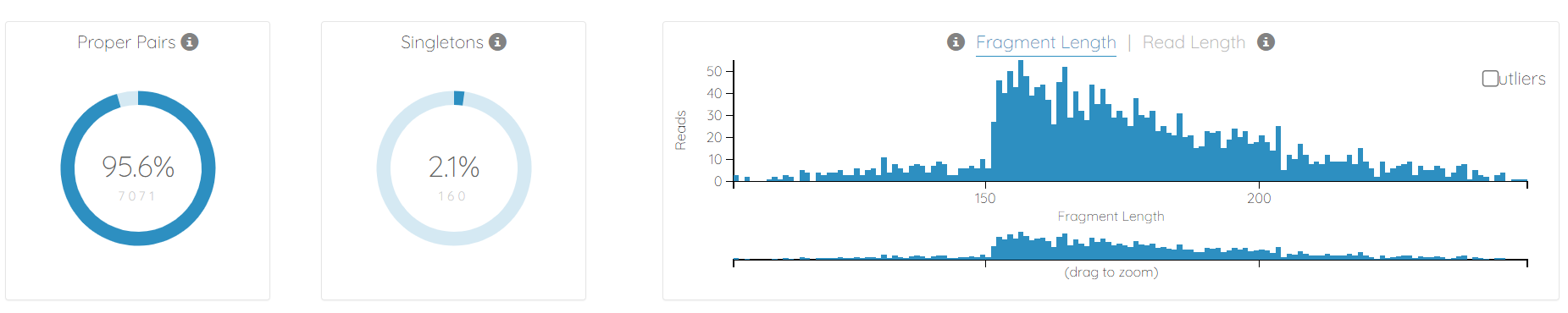
It's also essential to identify genomic features like exons and introns and ensure they're accurately represented in the data. Tumor heterogeneity, or differences within the tumor sample, can impact variant calling. Comparing alignment and coverage patterns to a reference dataset helps spot differences, like specific mutations in the tumor sample.

Lastly, we pay attention to any quality control metrics or warning flags provided by the visualization tool. These could indicate issues with the data or analysis pipeline that need further investigation. Overall, interpreting this report helps uncover insights into the genetic landscape of colorectal cancer, guiding potential biomarker discovery and treatment strategies.









**Fig 4.1** The bam.iobio.io web application. We selects a Colorectal Cancer alignment file, and the application rapidly samples the entire file to estimate genome-wide averages for a set of informative alignment metrics. Additionally, the user is able to select specific regions of interest and redo the analysis in those regions in seconds.

**4.2 UCSC Genome Browser Visualization**

The UCSC Genome Browser visualization of a colorectal cancer BAM file provides a comprehensive overview of the tumor's genetic profile. It highlights regions of high or low coverage, detects genetic variations like SNPs and indels, and identifies structural alterations such as insertions, deletions, and rearrangements. By pinpointing genomic features like exons and regulatory regions, it enhances our understanding of the tumor's biology and potential treatment targets. Comparisons to reference datasets reveal somatic mutations unique to the tumor, while quality control metrics ensure data reliability. Overall, this visualization aids in studying the molecular characteristics of colorectal cancer, guiding further research and personalized treatment approaches.Top of Form

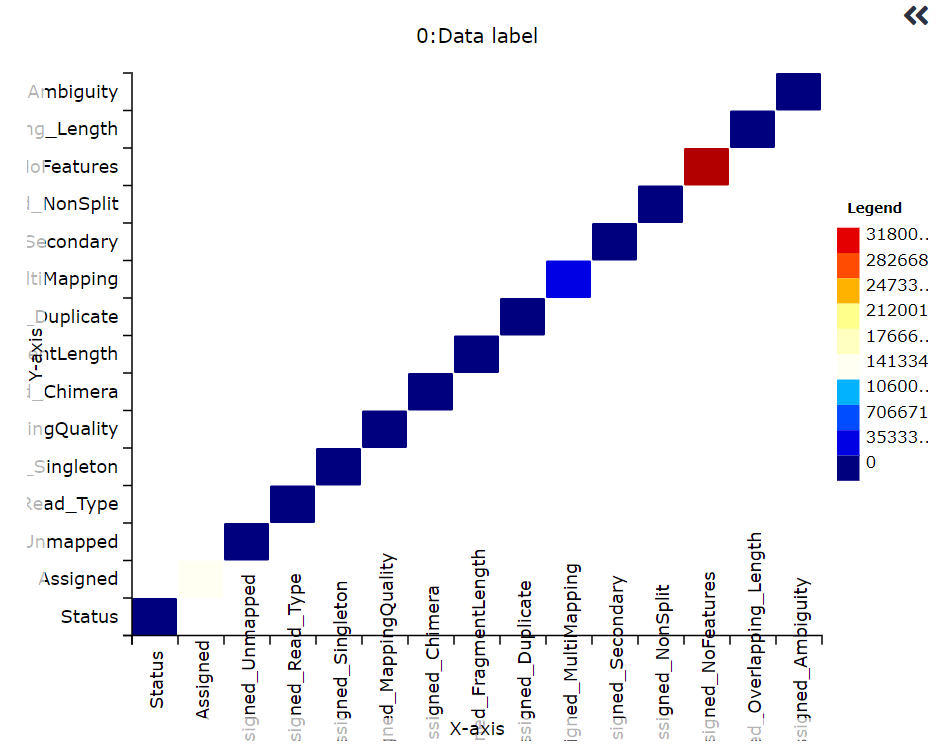




**Fig 4.2** Visualizing a BAM file in the UCSC Genome Browser provides a graphical representation of aligned sequencing reads on the reference genome. It displays alignment coverage, identified variants, and other annotations, aiding in the interpretation of genomic data.

**5. FeatureCount**

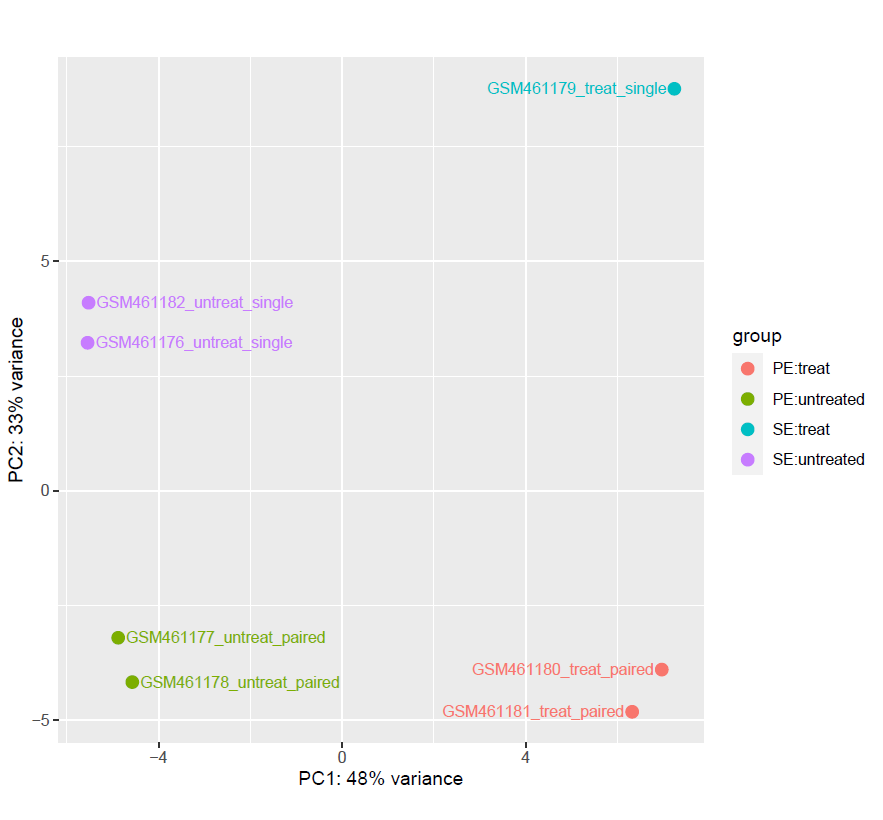
In a FeatureCounts result for colorectal cancer samples, we typically encounter a comprehensive table or matrix comprising gene counts. Each row of this matrix corresponds to a distinct gene, identified by its official gene symbol or ID. The columns, on the other hand, represent different samples derived from colorectal cancer patients, including tumor samples and matched normal tissue samples. Within this matrix, the values in each cell denote the number of RNA-seq reads that have aligned to a specific gene in a particular sample. These count values serve as a quantitative measure of gene expression levels across the samples, providing valuable insights into the transcriptional activity of individual genes in the context of colorectal cancer. By comparing these count values between tumor and normal samples, researchers can identify genes exhibiting differential expression patterns. Genes with notably higher or lower counts in tumor samples compared to their normal counterparts may signify their involvement in key biological processes underlying colorectal cancer development or progression. To further elucidate gene expression patterns and potential regulatory mechanisms, researchers often employ visualization techniques such as heatmaps or scatterplots. These visualizations help identify distinct expression profiles, clusters of co-regulated genes, and potential biomarkers or therapeutic targets associated with colorectal cancer. Additionally, statistical analyses can be applied to the FeatureCounts data to identify differentially expressed genes and pathways, facilitating a deeper understanding of the molecular mechanisms driving colorectal cancer pathogenesis and aiding in the discovery of novel diagnostic or therapeutic strategies. Ultimately, FeatureCounts results play a pivotal role in unraveling the intricate gene expression landscape of colorectal cancer, offering crucial insights that guide further research endeavors and clinical applications.



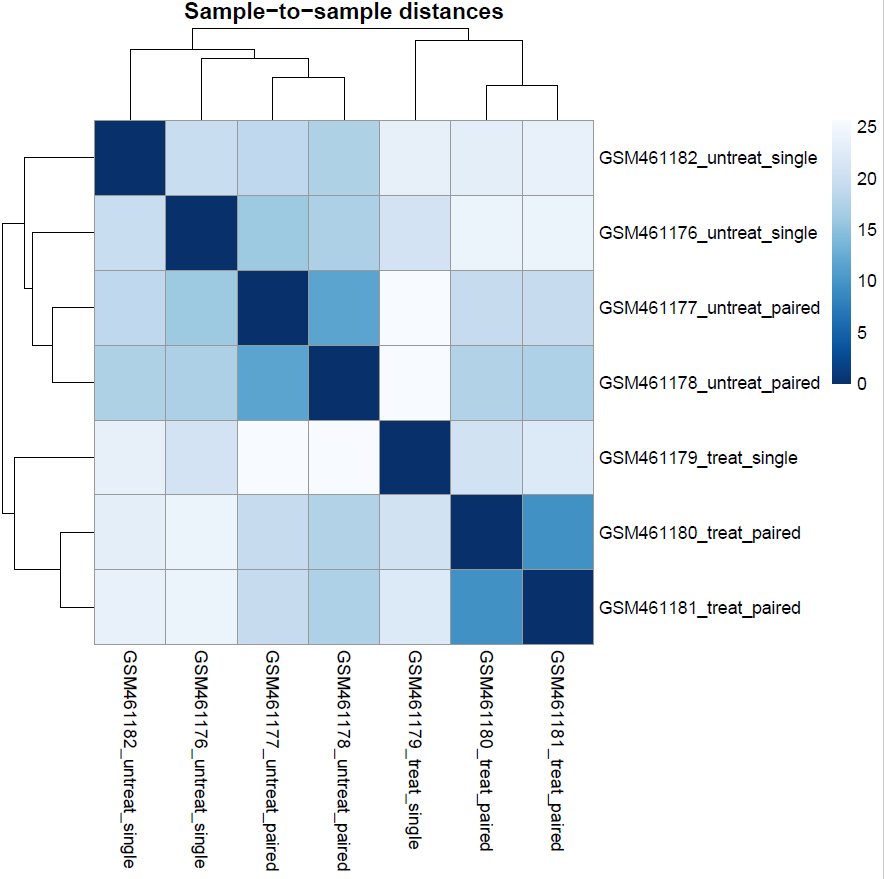
**Fig 5** The FeatureCounts heatmap provides a visual depiction of gene expression dynamics across various samples. Rows depict individual genes, while columns represent distinct samples. Color intensity within each cell reflects the relative expression level of a gene in a specific sample, facilitating the identification of expression patterns and co-regulated gene clusters. This visualization technique unveils insights into gene co-regulation, sample similarities, and differential expression patterns, thereby contributing to biomarker discovery efforts, particularly in the context of colorectal cancer research.

**6.DESeq2**

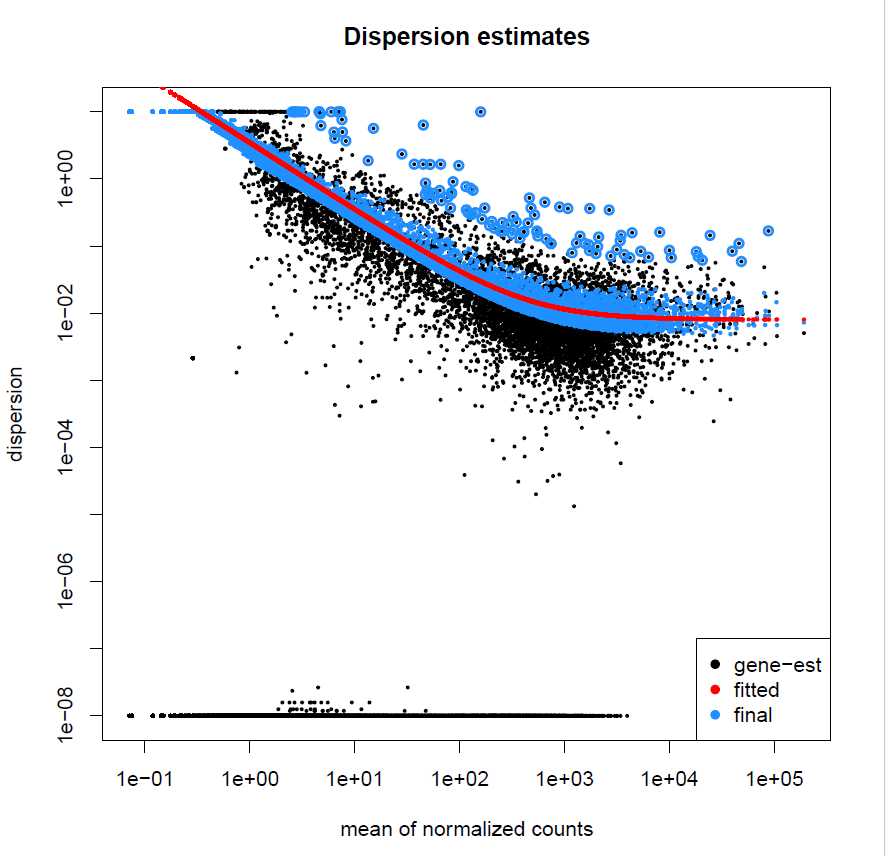
In a Result focusing on analyzing RNA sequences in colorectal cancer using DESeq2, the interpretation of results encompasses several crucial aspects. Firstly, it involves identifying genes that show significant differences in expression between colorectal cancer samples and normal tissue controls, noting whether these genes are upregulated or downregulated. The volcano plot generated from DESeq2 results visually represents these significant expression changes, highlighting genes with notable fold changes and statistical significance. Furthermore, Gene Ontology (GO) enrichment analysis helps explain the biological processes, molecular functions, and cellular components enriched among these differentially expressed genes. By understanding these enriched terms, we gain insights into how gene expression changes may affect colorectal cancer biology. Pathway analysis supplements this by identifying dysregulated signaling pathways and molecular networks associated with colorectal cancer pathogenesis and progression, like Wnt signaling and cell cycle regulation. Additionally, exploring the association between differentially expressed genes and clinicopathological features of colorectal cancer, such as tumor stage and patient prognosis, aids in identifying potential biomarkers for diagnosis, prognosis, and treatment response prediction. However, it's crucial to validate these findings through further experimental studies to confirm their biological relevance and elucidate their roles in colorectal cancer biology. Such validation studies are necessary to understand the underlying molecular mechanisms driving colorectal cancer development and progression fully. Finally, discussing the clinical implications of the DESeq2 results for colorectal cancer diagnosis, prognosis, and treatment, and proposing future research directions based on identified differentially expressed genes and dysregulated pathways, provides a roadmap for translating these findings into clinical practice and advancing research in the field. By addressing these components comprehensively, the thesis aims to offer a nuanced understanding of the gene expression landscape in colorectal cancer and its implications for clinical care and further scientific inquiry.



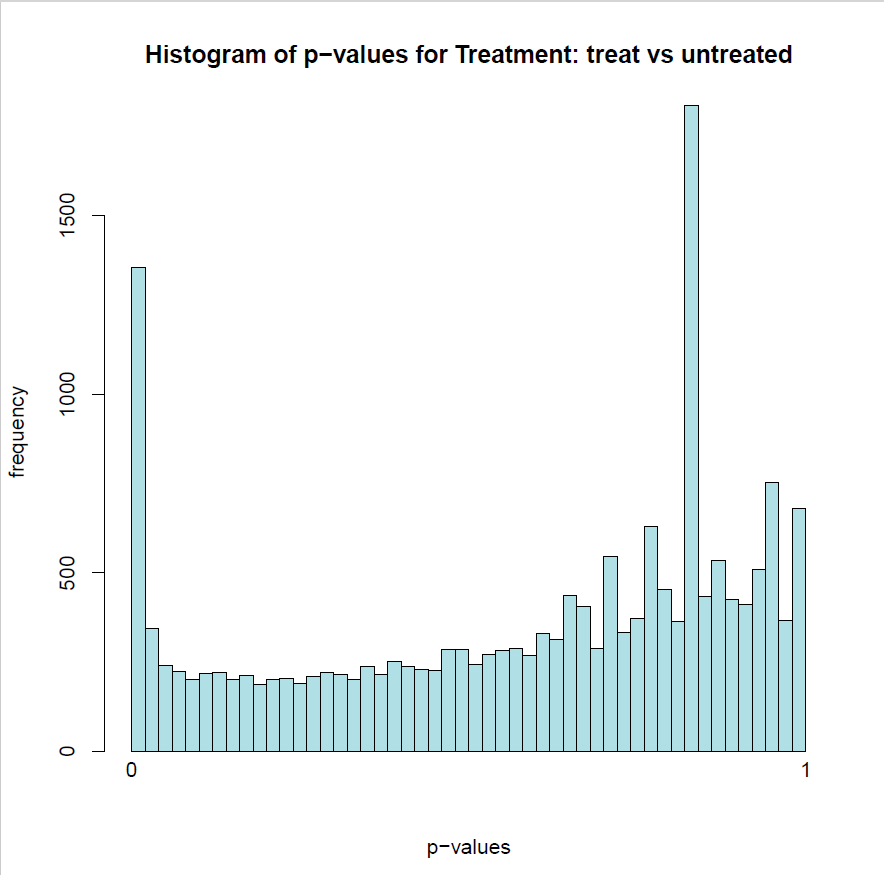
**Fig 6** .**1** In a thesis, the DESeq2 PCA biplot illustrates sample relationships using gene expression. Points represent samples, revealing similarities. Axes show principal components explaining variance. Colors or labels group samples by conditions. Outliers hint at issues, batch effects stress the need for correction. The PCA biplot guides quality check and sample grouping.



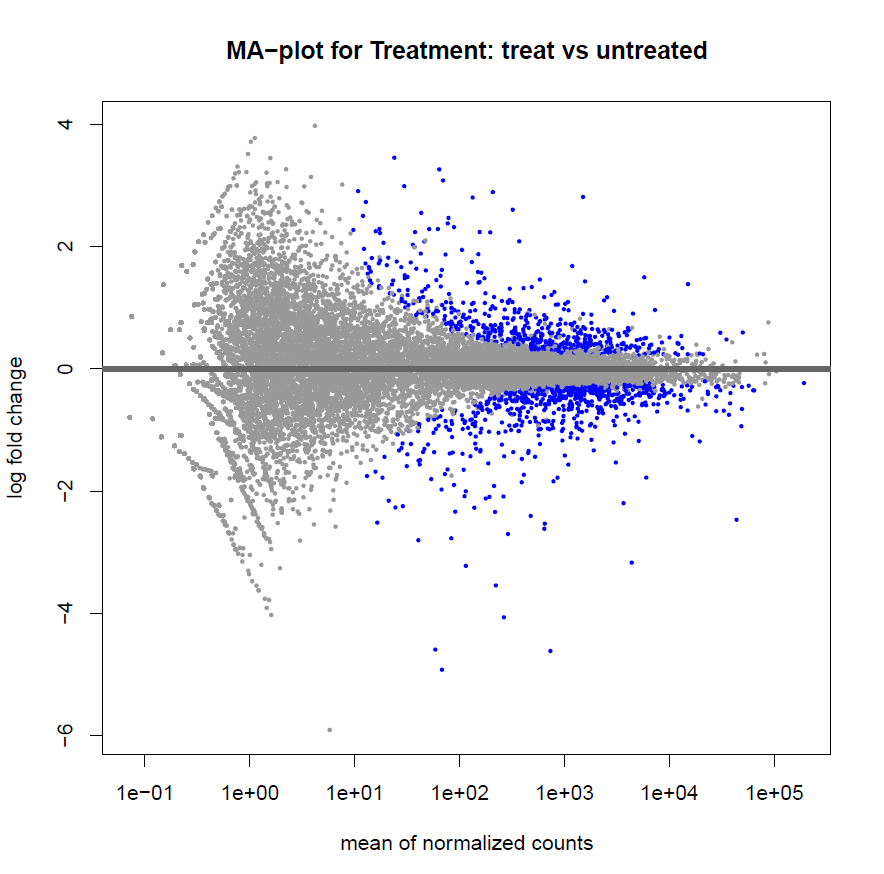
**Fig6.2** In DESeq2, the sample-to-sample distance heatmap shows how similar or dissimilar samples are based on gene expression. Each cell represents a pairwise comparison, with warmer colors indicating more similarity. This heatmap helps detect patterns, outliers, and batch effects, aiding quality assessment and guiding data interpretation.



**Fig6.3** DESeq2's dispersion estimate plot shows how dispersion varies with gene expression level. Dispersion, reflecting expression variability, typically decreases as expression increases. This plot aids in assessing dispersion reliability and identifying genes with unusual variability, crucial for accurate RNA-seq analysis.



**Fig6.4** DESeq2's p-value histogram for treatment comparison displays the distribution of p-values from differential expression testing. Each bin represents a range of p-values, showing gene counts. It helps assess significance, with low p-values indicating significant expression changes. Researchers use it to set significance thresholds and gauge analysis quality, identifying genes for further study.



**Fig6.5** DESeq2's MA plot compares log2 fold change (M) to the average normalized counts (A) for each gene. It highlights genes with significant expression changes, aiding in their prioritization for further study and providing insights into treatment effects.

**DISCUSSIONS**

The CRC RNA-seq and data analysis pipeline integrates a suite of powerful tools including FastQC, Trimmomatic, HISAT2, FeatureCounts, and DESeq2, providing researchers with a robust framework for exploring the intricate transcriptomic landscape of colorectal cancer (CRC). This pipeline begins with FastQC, which conducts initial quality assessments of raw sequencing data, ensuring the integrity and reliability of subsequent analyses. Trimmomatic follows, meticulously trimming low-quality bases and removing adapter sequences to enhance the quality of sequencing reads. HISAT2 is then employed for precise alignment of these processed reads to the reference genome, enabling accurate identification of gene expression patterns and alternative splicing events specific to CRC.Following alignment, FeatureCounts quantifies gene expression levels across samples, facilitating the identification of genes that are differentially expressed in CRC compared to normal tissue. This step is crucial for unraveling the molecular mechanisms driving CRC pathogenesis and progression. DESeq2 further refines the analysis by statistically evaluating differential gene expression, accounting for sample variability and experimental design factors. Through this comprehensive approach, researchers gain valuable insights into the dysregulated molecular pathways and biological processes underlying CRC development, potentially uncovering novel diagnostic biomarkers and therapeutic targets.Moreover, the integration of advanced bioinformatics tools ensures the rigor and reproducibility of research findings in CRC transcriptomics. By leveraging these tools within a cohesive pipeline, researchers can confidently explore the complexities of CRC at the transcriptomic level, paving the way for advancements in disease understanding and personalized treatment strategies. Ultimately, this integrated approach holds immense promise for accelerating discoveries in CRC research and improving patient outcomes.

**CONCLUSION**

The comprehensive RNA sequencing analysis for colorectal cancer, utilizing FastQC, Trimmomatic, HISAT2, FeatureCounts, and DESeq2, has yielded significant insights into the disease's transcriptomic landscape. FastQC and Trimmomatic ensured high data quality by identifying and correcting issues in raw reads. HISAT2 accurately aligned the cleaned reads to the reference genome, crucial for precise gene expression quantification. FeatureCounts provided a detailed gene count matrix, forming the basis for DESeq2's differential expression analysis. This robust pipeline identified numerous genes with altered expression in colorectal cancer, implicating key pathways in cancer development and progression, such as cell cycle regulation, apoptosis, DNA repair, and metabolism. These findings offer potential biomarkers for early diagnosis and therapeutic targets. Overall, this analytical approach proved effective and reliable, enhancing our understanding of colorectal cancer and paving the way for future research and improved patient outcomes.

**FUTURE WORK**

**Integration of Additional Omics Data:** Expand the analysis beyond RNA-seq by integrating other omics data such as whole-genome sequencing (WGS), DNA methylation profiling, or proteomics data. Integrating multi-omics datasets could provide a more comprehensive understanding of the molecular mechanisms underlying CRC and identify novel biomarkers or therapeutic targets.

**Single-Cell RNA-seq Analysis:** Incorporate single-cell RNA sequencing (scRNA-seq) data analysis to investigate the heterogeneity of tumor cell populations within CRC samples. By characterizing cellular diversity and identifying cell type-specific gene expression signatures, researchers can gain insights into tumor microenvironment dynamics, tumor progression, and potential therapeutic vulnerabilities.

**Longitudinal Studies and Clinical Correlations**: Conduct longitudinal studies to track changes in gene expression profiles over the course of CRC development, treatment, and recurrence. Additionally, correlate transcriptomic data with clinical outcomes, patient demographics, and treatment responses to identify predictive biomarkers of prognosis or therapeutic response.

**Functional Annotation and Pathway Analysis:** Perform functional annotation and pathway analysis to elucidate the biological processes and signaling pathways dysregulated in CRC. This could involve Gene Ontology (GO) enrichment analysis, pathway enrichment analysis, and network-based approaches to uncover key molecular pathways driving CRC pathogenesis and progression.

**Validation Studies:** Validate key findings from the RNA-seq analysis using independent CRC patient cohorts or experimental models such as cell lines or patient-derived xenografts (PDX). Experimental validation could include techniques such as qRT-PCR, immunohistochemistry, or functional assays to confirm the role of identified genes or pathways in CRC biology.

**Integration of Clinical Data and Machine Learning Approaches:** Integrate clinical data such as patient demographics, tumor stage, and treatment history with transcriptomic data using machine learning approaches. Develop predictive models to stratify patients based on their molecular profiles and predict clinical outcomes, treatment responses, or risk of disease recurrence.

**Exploration of Non-Coding RNAs:** Investigate the role of non-coding RNAs (e.g., microRNAs, long non-coding RNAs) in CRC pathogenesis using specialized bioinformatics tools and databases. Characterize non-coding RNA expression profiles, identify potential regulatory networks, and assess their functional significance in CRC progression.

**REFERENCES**

1.Tambonis, T., Boareto, M., Leite, V. B. P. [Differential Expression Analysis in RNA-seq Data Using a Geometric Approach.](http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed&cmd=Search&doptcmdl=Citation&defaultField=Title+Word&term=Differential+Expression+Analysis+in+RNA-seq+Data+Using+a+Geometric+Approach) Journal of Computational Biology. **25**, 1257-1265 (2018).

2.Rongbo Wen, Leqi Zhou, Zhiying Peng, Hao Fan, Tianshuai Zhang, Hang Jia, Xianhua Gao, Liqiang Hao, Zheng Lou, Fuao Cao, Guanyu Yu and Wei Zhang. Single-cell sequencing technology in colorectal cancer: a new technology to disclose the tumor heterogeneity and target precise treatment, (2023), DOI 10.3389/fimmu.2023.1175343

3.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. DOI: 10.1038/s41587-019-0201-4

4.Ashraful Haque1, Jessica Engel1, Sarah A. Teichmann2 and Tapio Lönnberg3 (2017). A practical guide to single-cell RNA sequencing for biomedical research and clinical application. Haque et al. Genome Medicine (2017) 9:75 DOI 10.1186/s13073-017-0467

5.Costa-Silva, J., Domingues, D., Lopes, F. M. [RNA-Seq differential expression analysis: An extended review and a software tool.](http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed&cmd=Search&doptcmdl=Citation&defaultField=Title+Word&term=RNA-Seq+differential+expression+analysis%3A+An+extended+review+and+a+software+tool) PloS One. **12**, 0190152 (2017).

6.Anders S., Reyes A., Huber W.. Detecting differential usage of exons from RNA-seq data. Genome Res. 2012; 22:2008–2017. - [PMC](http://www.ncbi.nlm.nih.gov/pmc/articles/pmc3460195/)- [PubMed](https://pubmed.ncbi.nlm.nih.gov/22722343/)

7.Varet, H., Brillet-Guéguen, L., Coppée, J. Y., Dillies, M. A. [SARTools: A DESeq2- and EdgeR-Based R Pipeline for Comprehensive Differential Analysis of RNA-Seq Data.](http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed&cmd=Search&doptcmdl=Citation&defaultField=Title+Word&term=SARTools%3A+A+DESeq2-+and+EdgeR-Based+R+Pipeline+for+Comprehensive+Differential+Analysis+of+RNA-Seq+Data) PloS One. **11**, 0157022 (2016).

8.Ritchie, M. E., et al. [limma powers differential expression analyses for RNA-sequencing and microarray studies.](http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed&cmd=Search&doptcmdl=Citation&defaultField=Title+Word&term=limma+powers+differential+expression+analyses+for+RNA-sequencing+and+microarray+studies) Nucleic Acids Research. **43**, 47 (2015).

9.Kim, D., Langmead, B., & Salzberg, S. L. (2015). HISAT: a fast spliced aligner with memory requirements. Nature Methods, 12(4), 357-360. DOI: 10.1038/nmeth.3317

10.Anders S., Pyl P.T., Huber W.. HTSeq—a Python framework to work with high-throughput sequencing data. Bioinformatics. 2015; 31:166–169. - [PMC](http://www.ncbi.nlm.nih.gov/pmc/articles/pmc4287950/)- [PubMed](https://pubmed.ncbi.nlm.nih.gov/25260700/)

11.Love, M. I., Huber, W., Anders, S. [Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.](http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed&cmd=Search&doptcmdl=Citation&defaultField=Title+Word&term=Moderated+estimation+of+fold+change+and+dispersion+for+RNA-seq+data+with+DESeq2) Genome Biology. **15**, 550 (2014).

12.Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina Sequence Data. Bioinformatics[, btu170](http://www.usadellab.org/cms/?page=trimmomatic)[1](http://www.usadellab.org/cms/?page=trimmomatic)

13.Andrews, S. (2010). [FastQC: A quality control tool for high throughput sequence data](http://www.usadellab.org/cms/?page=trimmomatic)[2](https://www.bioinformatics.babraham.ac.uk/projects/fastqc/)

14.Kim, D., Langmead, B., & Salzberg, S. L. (2015). HISAT: A fast spliced aligner with low memory requirements. Nature Methods[, 12(4), 357-360](http://www.usadellab.org/cms/?page=trimmomatic)[3](http://daehwankimlab.github.io/hisat2/)

15.Liao, Y., Smyth, G. K., & Shi, W. (2014). The Subread aligner: Fast, accurate and scalable read mapping by seed-and-vote. Nucleic Acids Research[, 41(10), e108](http://www.usadellab.org/cms/?page=trimmomatic)[4](https://rdrr.io/bioc/Rsubread/man/featureCounts.html)

16.Su Z., Labaj P.P., Li S., Thierry-Mieg J., Thierry-Mieg D., Shi W., Wang C., Schroth G.P., Setterquist R.A., Thompson J.F. et al. .. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. Nat. Biotechnol. 2014; 32:903–914. - [PMC](http://www.ncbi.nlm.nih.gov/pmc/articles/pmc4321899/)- [PubMed](https://pubmed.ncbi.nlm.nih.gov/25150838/)

17.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](http://www.usadellab.org/cms/?page=trimmomatic)[5](https://academic.oup.com/bioinformatics/article/30/7/923/232889)

18.Liao Y., Smyth G.K., Shi W.. featureCounts: an efficient general-purpose read summarization program. Bioinformatics. 2014; 30:923–930. - [PubMed](https://pubmed.ncbi.nlm.nih.gov/24227677/)

19.Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics, 30(15), 2114-2120. DOI: 10.1093/bioinformatics/btu170

20.Hoen PA et al. Reproducibility of high-throughput mRNA and small RNA sequencing across laboratories. Nat Biotechnol 31, 1015–1022 (2013).

21.Leinonen R, Sugawara H, Shumway M & International Nucleotide Sequence Database, C. The sequence read archive. Nucleic Acids Res 39, D19–21 (2011)

22.Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics, 10(1), 57-63.

23.Trapnell C., Pachter L., Salzberg S.L.. TopHat: discovering splice junctions with RNA-seq. Bioinformatics. 2009; 25:1105–1111. - [PMC](http://www.ncbi.nlm.nih.gov/pmc/articles/pmc2672628/)- [PubMed](https://pubmed.ncbi.nlm.nih.gov/19289445/)

24.Wang, Z., Gerstein, M., Snyder, M. [RNA-Seq: a revolutionary tool for transcriptomics.](http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed&cmd=Search&doptcmdl=Citation&defaultField=Title+Word&term=RNA-Seq%3A+a+revolutionary+tool+for+transcriptomics) Nature Reviews. Genetics. **10**, 57-63 (2009).