Paper PP19

Data Cleaning and Gene Expression Visualization in Hematological Conditions

Ashwika Papasani[1], Nasreen Bano[2] Micheal Atchison[2]

[1]Downingtown STEM Academy, Downingtown, PA, USA

[2]University of Pennsylvania, Philadelphia, PA, USA

*The work is done at the University of Pennsylvania

ABSTRACT

A new technology that offers a view into gene expression at the individual cell level or across the entire cell population is RNA sequencing. In clinical research, it is becoming increasingly crucial for diagnosing diseases and understanding how patients respond to treatment. In preclinical studies, this approach has already shown promise in assisting researchers in understanding the expression of genes and the functioning of cells. Large volumes of data are produced by this technology, necessitating the use of specialized bioinformatics tools to identify, clean, and eliminate low-quality cells, group cells according to gene expression, and produce visualizations that show variations. In order to investigate blood cell formation, we examine in this presentation how this data is collected, processed, and visualized.

INTRODUCTION

RNA sequencing has emerged as a powerful technology for examining gene expression, offering researchers a detailed view of transcriptional activity across diverse biological systems. Whether applied at the single-cell or bulk level, RNA-seq enables the detection of coordinated changes in gene expression and has become increasingly important in both clinical and preclinical research. Although it is widespread, there are some challenges in the analysis of bulk RNA-seq. There is a large amount of raw sequencing files, complexity of steps, and a need for reproducible computational workflows. In this study, we showcase a pipeline using bioinformatic software or tools for bulk RNA-seq that combines quality control, alignment, read count generation and differential expression analysis between two conditions. In the present study, RNA-seq analysis was performed on HCT116 cells that were divided into two groups: a control group transfected with the empty vector (pcCon) and an experimental/treated group transfected with the YY2 expression plasmid (pcYY2). Each group included three biological replicates. Comparative transcriptomic analysis between these groups was conducted to identify YY2-regulated genes associated with ferroptosis. Ferroptosis is a form of programmed cell death driven by the disruption of redox homeostasis and is known to exert tumor-suppressive effects, as well as play critical roles in cardiovascular and neurological disorders.

METHODOLOGY

We analyzed the RNA-seq data using a combination of command-line tools and RStudio. The pipeline accepts raw. data in either Sequence Read Archive (SRA) or FASTQ format. If the files are in SRA format, we first convert them to FASTQ using the SRA Toolkit. We then check the read quality with FastQC (1), which provides easy-to-view HTML reports. Next, we align the reads to the human genome (GRCh38) using Hierarchical Indexing for Spliced Alignment of Transcripts (HISAT2), a fast and sensitive spliced-read aligner. After alignment, we use featureCounts to count how many reads map to each gene. These counts are then loaded into RStudio, where we use the edgeR package to calculate counts per million (CPM) and filter out genes with little or no expression. We then assess the correlation between samples to evaluate how well the replicates are reproducible with each other.

using the ggcorrplot and heatmap packages. After reproducibility checks, we identify significantly up- and down-regulated differentially expressed genes (DEGs) between the control and treatment groups. Finally, we use the list of significantly up- and downregulated genes for functional enrichment and pathway analysis to understand the biological meaning of the results (Fig.1).

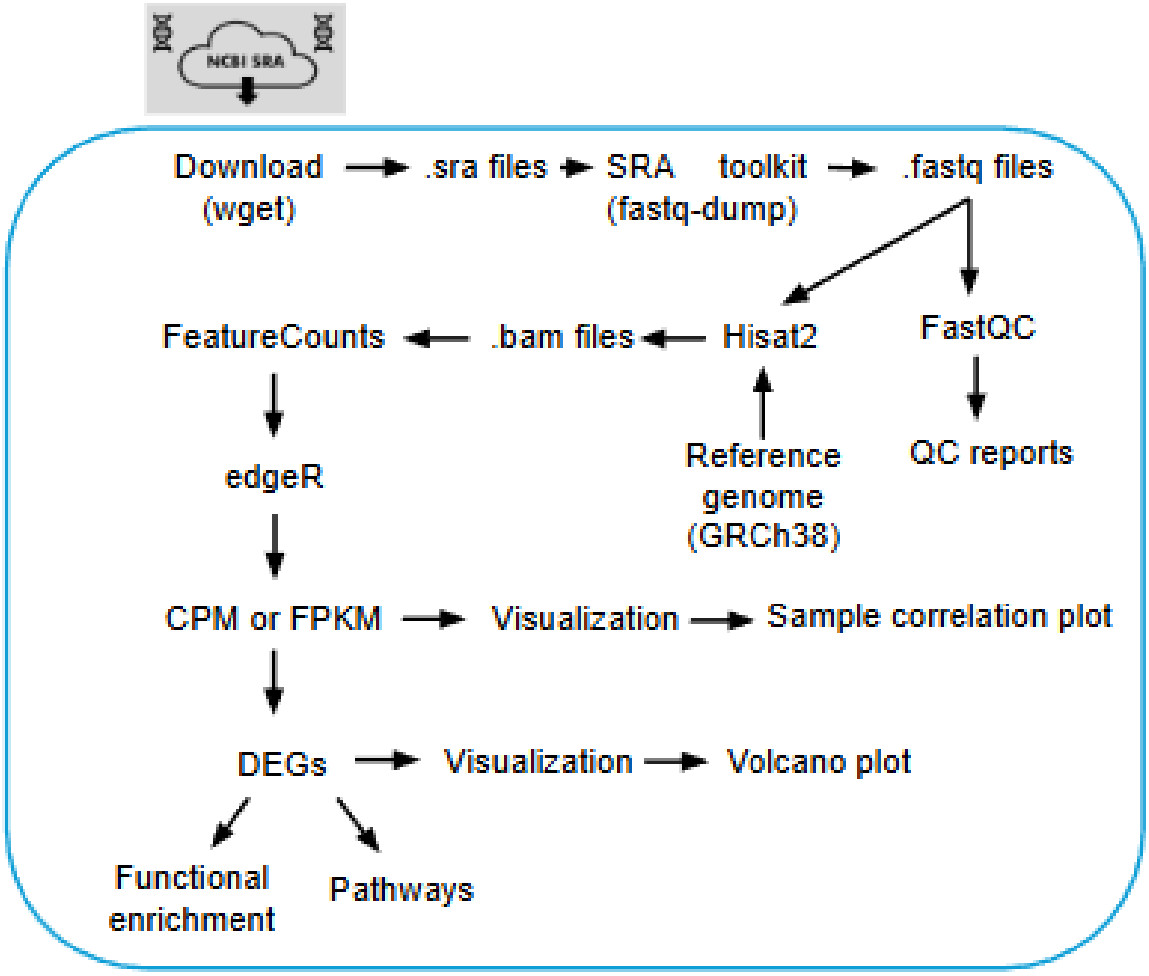

Figure 1. Overview of the bulk RNA-seq analysis pipeline, including data preprocessing, quality control, read alignment, gene-level quantification, differential expression analysis, and downstream functional enrichment to interpret biological pathways altered between control and treatment samples.

RESULTS

DIFFERENTIALLY EXPRESSED GENE ANALYSIS

Pairwise Pearson correlation coefficients were calculated using normalized gene expression values across all detected genes. Biological replicates from control (pcCon) and YY2-overexpression (pcYY2) groups cluster tightly, demonstrating high reproducibility and clear separation between experimental conditions (Fig. 2.-left panel). Differential gene expression analysis comparing pcYY2 versus pcCon identified 217 up-regulated genes, 317 down-regulated genes, and 21,633 non-significant genes. Each point represents a gene plotted by log2 fold change and −log10 adjusted *p*-value, highlighting genes significantly altered by YY2 overexpression (Fig. 2.-right panel). Further heatmap of top 20 up and downregulated genes are presented using heat map (Fig.3)

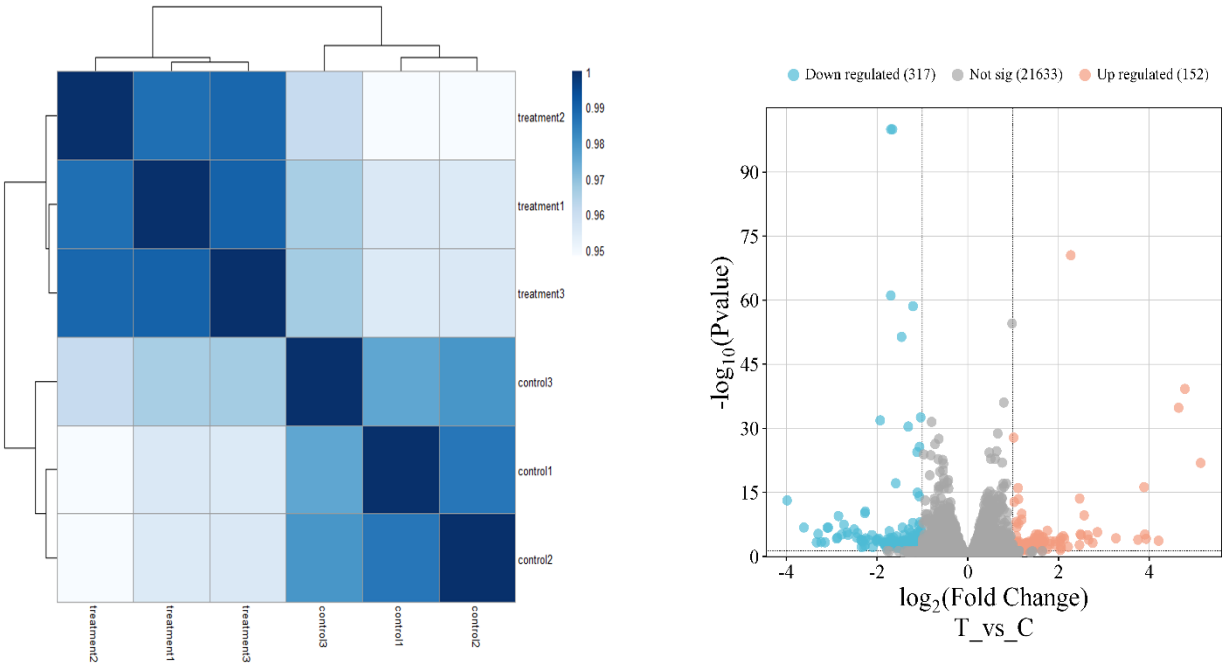

Fig.2. Correlation heatmap of RNA-seq samples and volcano plot of differentially expressed genes.

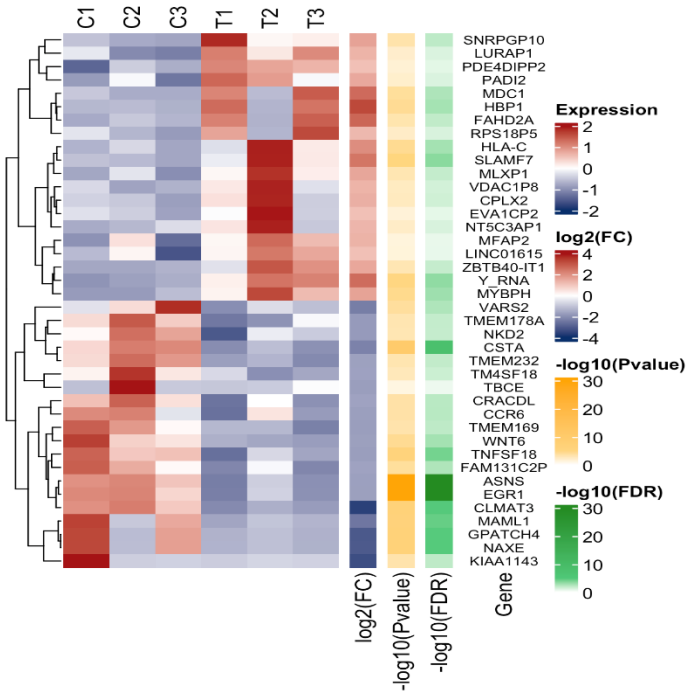

Fig.3. Heatmap representation of Top 20 up and down-regulated gene

GENE ONTOLOGY (GO) ENRICHMENT ANALYSIS

Significantly enriched GO terms are shown for Biological Process (BP), Cellular Component (CC), and Molecular Function (MF) categories. Bars represent −log10(*p*-value), indicating enrichment significance, while dot size reflects the number of genes associated with each term. Pink bars indicate up-regulated genes, and teal bars indicate down-

regulated genes. Enriched processes include epithelial morphogenesis, cell migration, cortical actin cytoskeleton organization, hematopoietic stem cell homeostasis, highlighting the key functional enrichment for up-regulated genes. Top enriched GO terms across BP, CC, and MF categories are displayed. Bar length represents −log10(*p*-value), dot size indicates gene count, and colors denote up- (pink) and down-regulated (teal) genes (Fig.4)

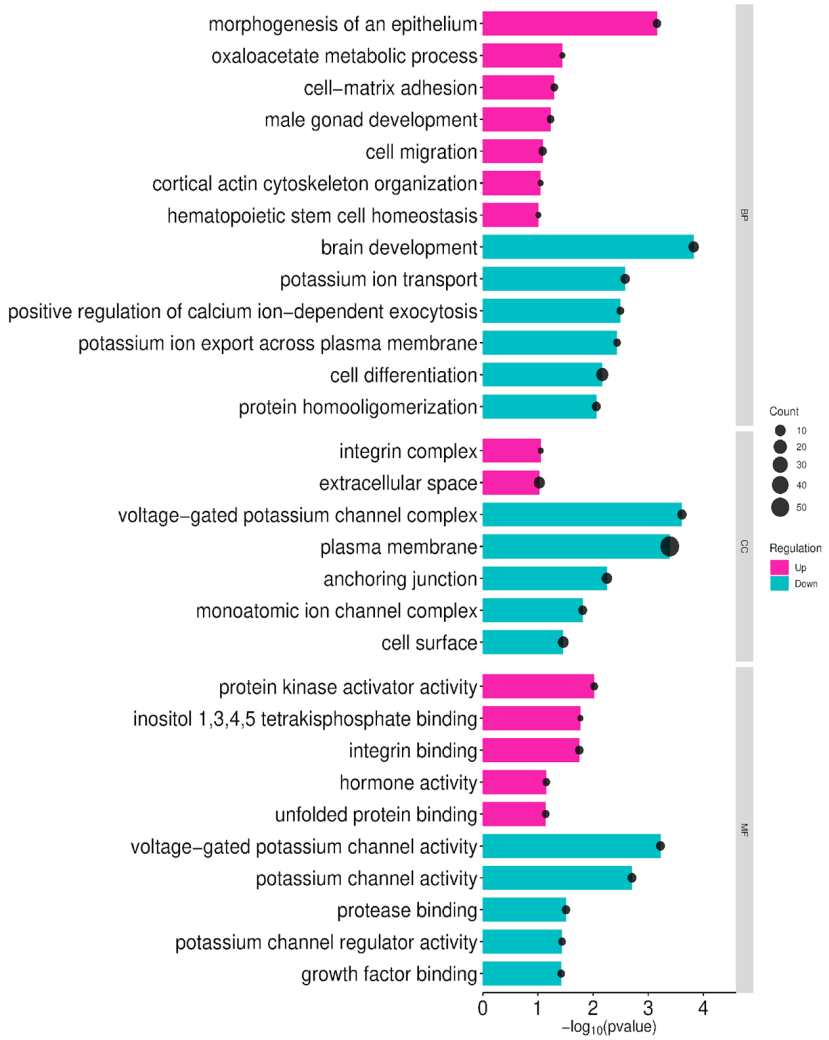

Fig.4. Gene Ontology enrichment of differentially expressed genes.

PATHWAY ENRICHMENT ANALYSIS OF DEGS

Significantly enriched pathways are shown for up-regulated (top panel) and down-regulated (bottom panel) gene sets. The x-axis represents fold enrichment, dot size indicates the number of genes mapped to each pathway, and color intensity corresponds to −log10(*p*-value) (Fig.5.). Enriched pathways include Cornified envelope formation, hormone signaling, extracellular matrix organization, and neutrophil degranulation, highlighting the key pathways for up-regulated DEGs.

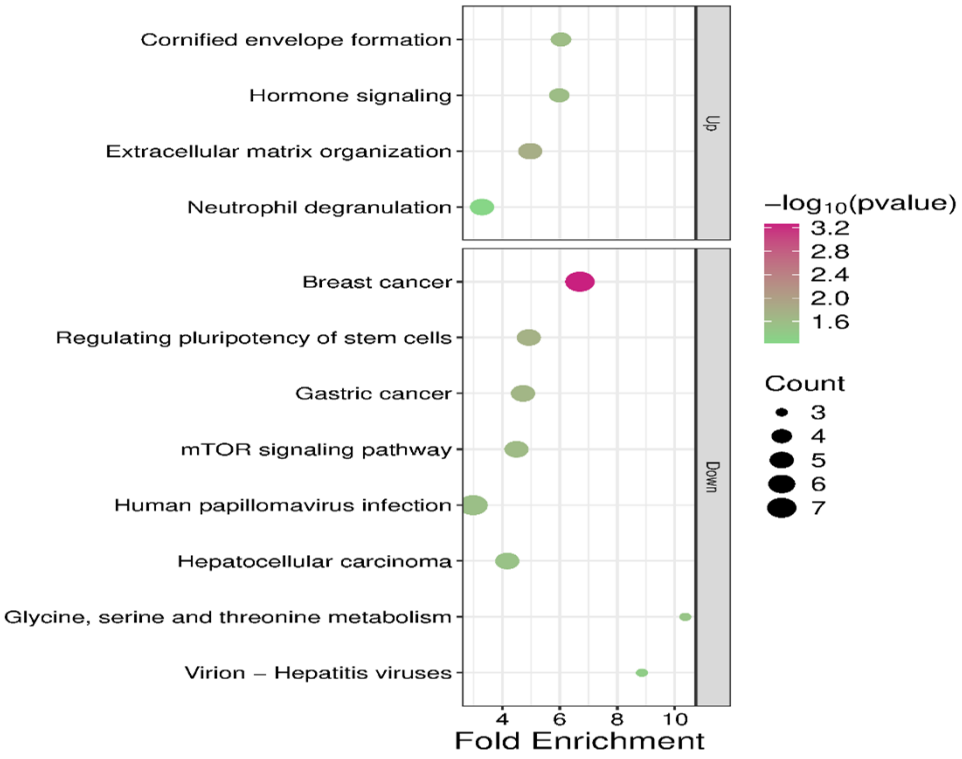

Fig.5. Pathway enrichment analysis of DEGs

CONCLUSION

RNA sequencing continues to transform modern biological research by enabling precise, high-resolution measurement of gene expression across diverse cellular contexts. In this study, we applied a reproducible and comprehensive bulk RNA-seq workflow—spanning quality control, genome alignment, read quantification, statistical modeling, and functional enrichment—to characterize transcriptional changes driven by YY2 overexpression in HCT116 cells. The strong correlation among biological replicates and the clear separation between experimental groups underscore the robustness of the analytical pipeline.

Differential expression analysis revealed distinct sets of up- and down-regulated genes, which, when examined through Gene Ontology and pathway enrichment, highlighted biological processes related to epithelial morphogenesis, cytoskeletal organization, extracellular matrix remodeling, and immune-associated pathways such as neutrophil degranulation. These findings not only demonstrate the power of RNA-seq and bioinformatic tools to uncover meaningful transcriptional signatures but also provide insight into potential mechanisms through which YY2 may influence ferroptosis-related pathways and broader cellular behavior. Together, this work illustrates how integrated RNA-seq analysis pipelines can illuminate gene regulatory networks and guide future investigations into disease mechanisms and therapeutic targets.

REFERENCES

1. Kim, D., Langmead, B. and Salzberg, S.L., 2015. HISAT: a fast spliced aligner with low memory requirements. *Nature methods*, *12*(4), pp.357-360.
2. Liao, Y., Smyth, G.K. and Shi, W., 2014. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics, 30(7), pp.923-930.
3. Robinson, M.D., McCarthy, D.J. and Smyth, G.K., 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. bioinformatics, 26(1), pp.139-140.

CONTACT INFORMATION

Contact the authors at:

Michael L. Atchison
Professor, Department of Biomedical Sciences
University of Pennsylvania
3800 Spruce Street
Philadelphia, PA 19104
Email: atchison@vet.upenn.edu

Nasreen Bano
Postdoctoral researcher
University of Pennsylvania
3800 Spruce Street
Philadelphia, PA 19104
Email: nbano@upenn.edu

Ashwika Papasani
Downingtown STEM Academy
335 Manor Avenue
Downingtown
PA 19335
Email: apapasani01@student.dasd.org