**ABSTRACT**

In **“Colorectal cancer Gene Expression Analysis”** is the analysis of colorectal cancer microarray data using E-MEXP-3756 dataset from the ArrayExpress. The main objective is to identify significant differentially expressed genes (DEGs) which may serve as biomarkers for early detection. The study uses Rstudio and Bioconductor packages for statistical and visualization analysis, normalization, annotation, PCA and enrichment gene analysis and machine learning model for feature selection using SHAP.

1. **INTRODUCTION**
2. **Colorectal cancer**
   1. **History**

Colorectal cancer is one of the leading causes of cancer-related deaths globally, with increasing prevalence over the years. Advances in genomics have facilitated the use of microarray and microarray and RNA-seq technologies to study gene expression patterns, offering insights into underlying molecular mechanisms. The dataset E-MEXP-3756, sourced from ArrayExpress, provides transcription profiling data from Homo sapiens, enabling bioinformatics studies to uncover biomarkers linked to colorectal cancer progression.

* 1. **Requirement Analysis**

To analyze the E-MEXP-3756 dataset, preprocessing microarray data is essential for extracting biologically meaningful insights. This project was conducted in Rstudio using a comprehensive suite of packages, ensuring a robust and reproduceable workflow. **Requirements include:**

High-quality gene expression data with normalization using packages. Raw microarray files (CEL files) are imported using imported using specialized tools such as the Oligo package. The data then undergoes a log2 transformation to stabilize the variance followed by quantile normalization commonly performed via the Robust Multi-array Average method. These steps are crucial to correct for systematic technical biases and to ensure that expression levels across samples are comparable.

comprehensive computational tools for annotation, visualization and enrichment analysis: The gene expression must be accurately annotated to map probe identifiers to gene symbols and biological functions. Using annotation packages such as hugu133plus2.db and org.Hs.eg.db to ensure that each probe is correctly linked to its corresponding gene.

For visualization packages like ggplot, pheatmap and enhancedVolcano are utilized to generate PCA plots, heatmaos and volcano plots.

Enrichment analysis is performed using clusterProfiler which examines Gene Ontology (GO) and KEGG pathway enrichments.

Machine learning techniques for feature selection and interpretability using random forest and frameworks like fastshap.

* 1. **Objective**

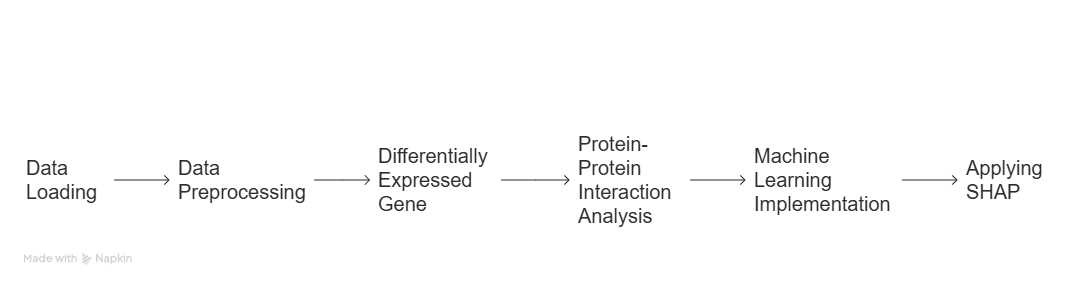
The objective of this study is to identify significant DEGs in colorectal cancer and understand their biological roles and reveal potential biomarkers for early detection and therapeutic targets. Performed explainable AI technique for extracting the feature importance of input features in the machine learning model.

1. **SYSTEM ANALYSIS**
   1. **Motivations**

The motivation is to identify early detection biomarkers in colorectal cancer and to explore the explainable AI techniques like SHAP that allows understanding in biomarkers.

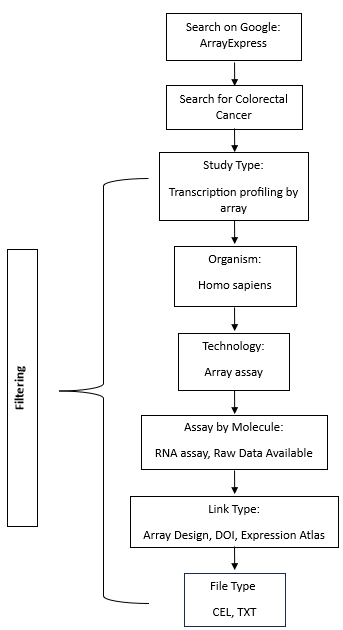
* 1. **Modules**

The analysis of workflow consists of the following modules:



* + 1. **Data loading**

Here is the pipeline for data collection below.



RStudio utilizes functions from packages oligo and Biobase to read the raw CEL files and their associate SDRF metadata file "E-MEXP-3756.sdrf.txt." The SDRF file becomes an AnnotatedDataFrame through read.delim() function then conversion. By using oligo::read.celfiles() the program imports CEL files while linking each file to its respective metadata. This step creates a foundation since it verifies that all further procedures will be conducted on data that has been properly annotated.

* + 1. **Data preprocessing**
    2. The initial step includes data loading which is then processed by performing log2 transformation before implementing quantile normalization through RMA (Robust Multi-array Average) protocol. This correction method enables technicians to minimize technical sample biases. Boxplots along with PCA plots provide visualization methods to monitor quality before and after normalization by displaying log2-intensity distributions and confirming cluster formation of similar samples. The module implements procedures that select and retain high-quality interchangeable gene expression values essential for subsequent analysis.
    3. **Differentially Expressed Gene**

The statistical analysis tools lead to gene expression comparison through the limma package. The normalized data receives linear model analysis to identify differentially expressed genes (DEGs) through pre-defined criteria (adjusted p-value < 0.05 and log fold change > specified threshold). The analysis filter process checks for both genes showing increased expression and those showing decreased expression. The EnhancedVolcano package generates volcano plots for displaying DEGs that show biologically important patterns for colorectal cancer.

* + 1. **Protein-Protein Interaction Analysis**

The DEGs are converted to proteins while the STRINGdb package seek known and predicted protein–protein interactions. The ipgraph library generates visual network maps to display protein interactions from the retrieved data using STRINGdb. The computational method computes centrality measures including degree and betweenness and closeness to find proteins that serve as important disease mechanism hubs.

* + 1. **Machine Learning implementation Applying SHAP**

The randomForest package constructs a random forest classifier for normal and colorectal cancer sample discrimination by using chosen gene expression features. The dataset undergoes separation into training data and testing data for measuring model performance through computational metrics that include the confusion matrix. Model interpretability receives additional enhancement through the computation of SHAP (SHapley Additive exPlanations) values using the fastshap package. SHAP analysis calculates the exact influence that each feature (such as logFC or average expression or t-statistic) holds toward achieving prediction results. After the visualization methods with beeswarm plot and box plot which provide the knowledge of feature importance to get explainability.

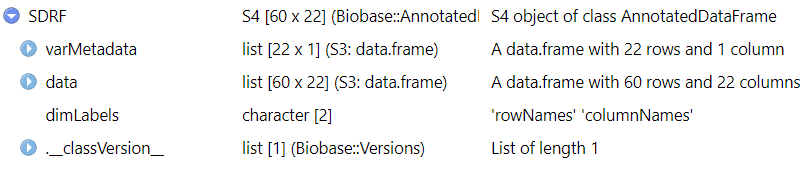
1. **IMPLEMENTATION**
   1. **Data Preprocessing**

**3.1.1. Data Loading**

Set the working directory in RStudio and load library “Biobase” which provides foundational classes and functions to handle biological data.

Read the SDRF file and convert it into an annotated data frame.

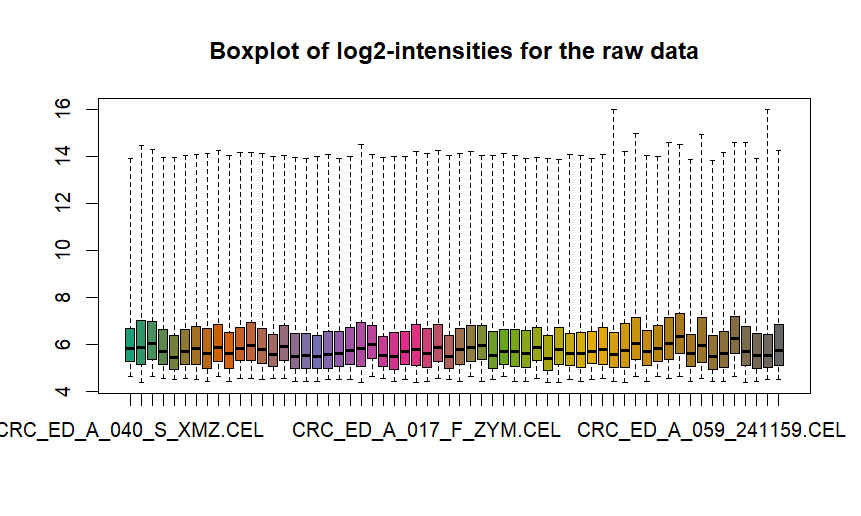
Process CEL files using the oligo::read.celfiles() function.



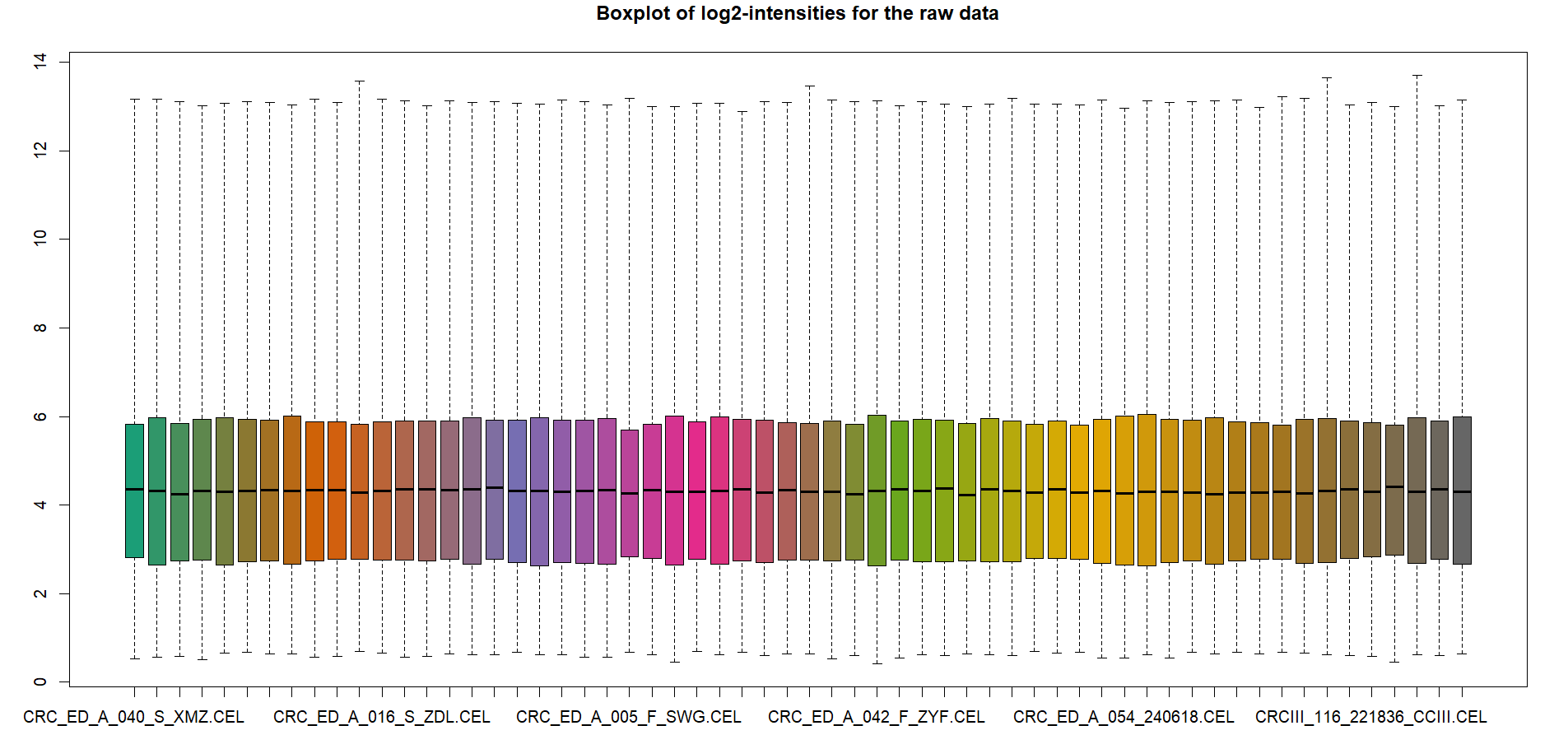
**3.1.2. Normalization and Transformation**

Performed Log2 transformation was applied to stabilize variance across expression values and Quantile normalization (RMA – Robust Multi-array Average) to adjust for technical variations.

**Before normalization**

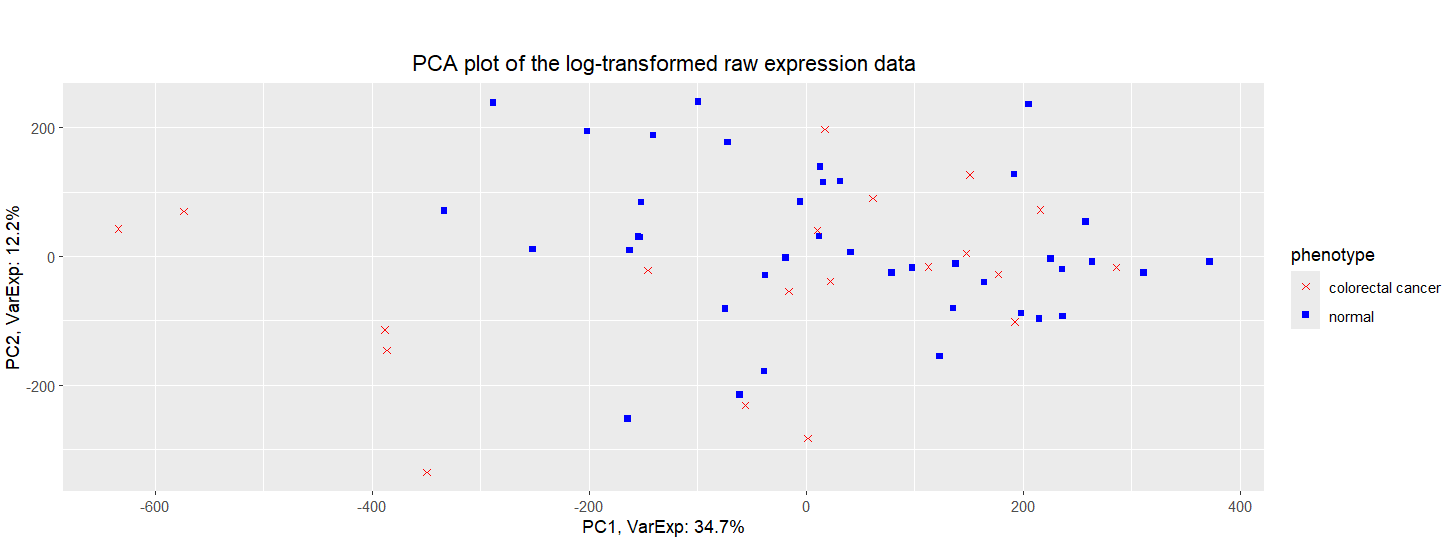


**After normalization**

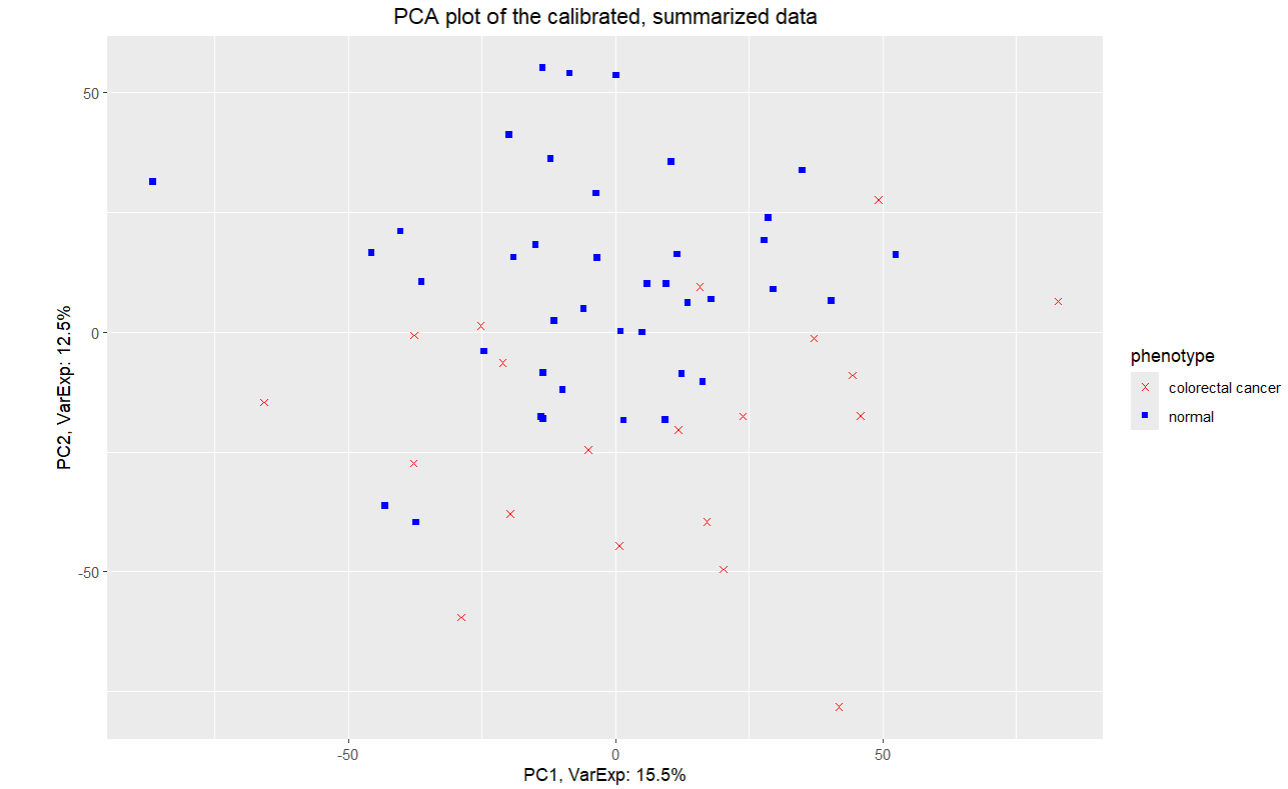


**PCA plots to visualize clustering of samples.**

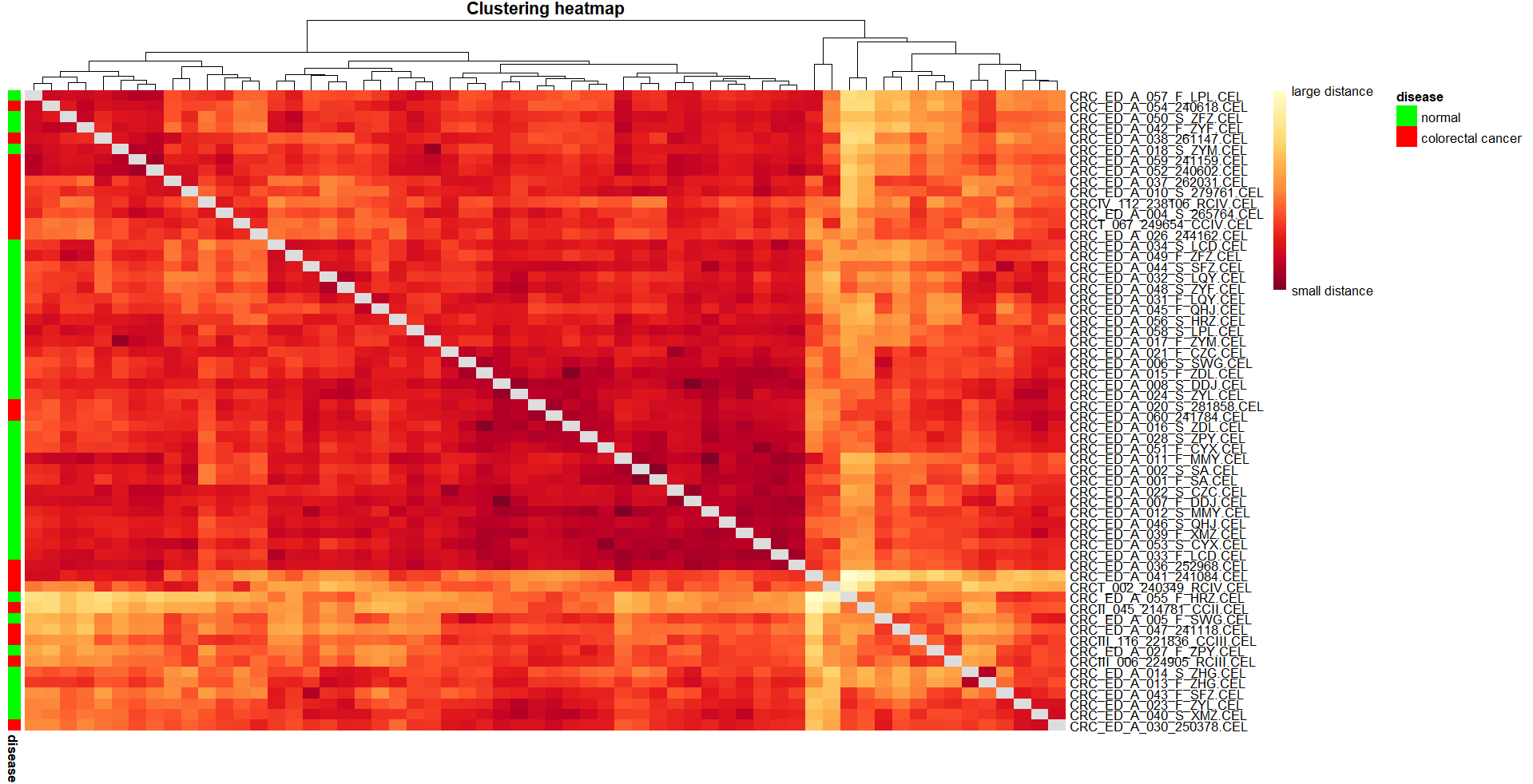
**Before normalization**



**After normalization**



**Heatmap**



The top dendrogram shows hierarchical shows clustering of samples based on their distance. Samples that are close together in the dendrogram have similar expression profiles. In center color, the matrix shows pairwise distances between smaples.

* White / light yellow = smaller distances (more similar)
* Red / Orange = larger distances (more dissimilar)

Color on left : Green -> Normal and Red -> colorectal cancer

**3.1.3. Probe annotation and feature selection**

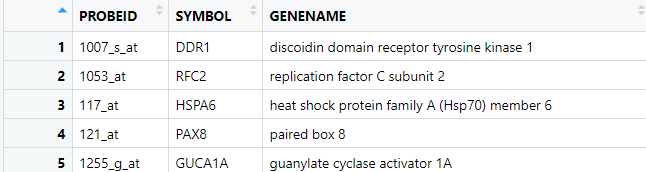
The analysis of microarray data heavily depends on the crucial annotation step known as probe annotation. The process of probe ID mapping to biological information allows researchers to interpret gene expression results. This research analyzed probe IDs by assigning them biological descriptors which included Gene Symbols and Gene IDs as well as Gene Names.

**Gene Symbols:** The genetic identification system utilizes Gene Symbols as standardized abbreviations to distinguish different genes.

**Gene IDs:** Unique numerical or alphanumerical identifiers specific to each gene.

**Gene Names**: The utilized dictionary uses Descriptive names to explain gene functions and their purposes.

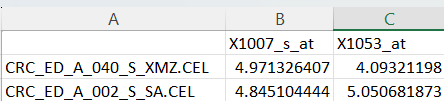
The accomplishment of the task used the hgu133plus2.db package. This Bioconductor annotation package delivers complete mappings of probes to genes through their symbols together with their IDs and names for our microarray platform.



The dataset underwent a filtering process following the annotation step with parameters that applied strict conditions.

Different gene symbols that remained unannotated were deleted from analysis to achieve full interpretation of downstream probes.

The analysis removed probes which linked to more than one gene entry. The objective behind this step was to achieve unique gene identification which improved the reliability of subsequent differential expression analysis results.



The dataset transformed into a high-quality research base through these data filtering procedures which retained uniquely mapped confident probe data points.

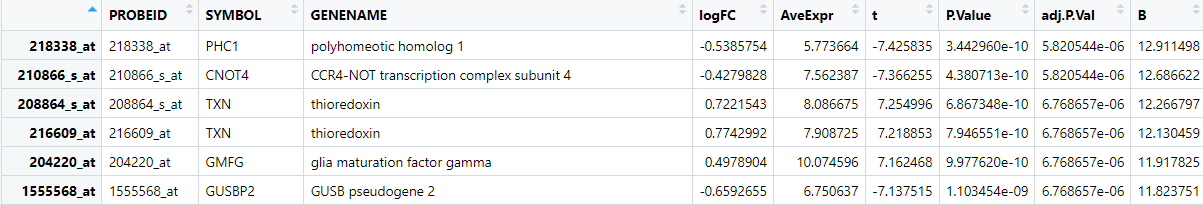
* 1. **Differentially Expressed Genes**

The objective of differential expression analysis consists in revealing genes that display substantial expression level differences between colorectal cancer samples and normal samples. The analysis reveals disease molecular processes while detecting possible biomarkers for detection.

* + 1. **Methodology**

The research design matrix enabled the evaluation of colorectal cancer expression profiles as a separate category from normal expression profiles through Condition = 0 and Condition = 1 classifications. The statistical model used for analysis is created through this matrix design.

The Limma package enabled differential expression analysis for normalized microarray data which researchers performed. The linear modeling capabilities of Limma gained additional power through eBayes moderation because this technique provides improved stability of gene expression variances. The researcher designed a contrast matrix that detailed the comparison of conditions in order to recognize genes with significant expression changes.



* + 1. **Key Statistics**

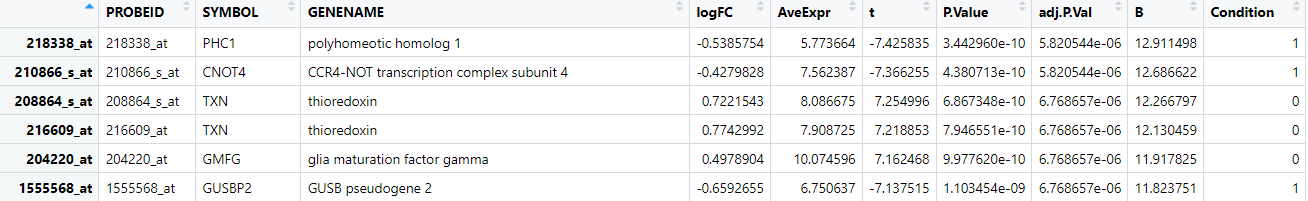
Total Probes Analyzed: 43,122.

DEG Selection Criteria:

Adjusted p-value (FDR) < 0.05.

Log fold change (logFC) threshold > 0.5.

The applied thresholds enabled the identification of 4,748 DEGs within the dataset which included both upregulated and downregulated genes.



* + 1. **Significant Gene Attributes**

The characterization of the selected DEGs utilized these main annotation elements:

The examination tool PROBEID establishes unique attributes to identify single genes or transcripts.

The official symbols like (PHC1, CNOT4 and TXN) function as data cross-referencing methods across biological databases.

Each gene received a functional descriptive name as its GENENAME designation.

The measurement of gene expression changes between conditions appears as Log2 fold change in the results.

The expression level of TXN gene increases by 0.77 within colorectal cancer specimens relative to healthy tissue.

The expression levels of PHC1 molecules decreased compared to normal samples based on its negative value of -0.53.

AveExpr: Represents the average expression value for each gene across all samples. Extremely stable genes can be distinguished from those with high expression variability through this metric.

A more powerful indicator showing differential expression evidence exists within the t-statistic and larger absolute values indicate robust statistical evidence.

A P.Value raw measure provides information about the chance that the noticed difference emerged by random occasion.

The Benjamini–Hochberg correction applied through adj.P.Val determined adjusted p-values for controlling the false discovery rate (FDR). Research identified statistical significance using genes that exhibited an adj.P.Val under 0.05.

Log-odds ratios of differential expression act as supplementary data for designating genes as significantly expressed among other candidates.

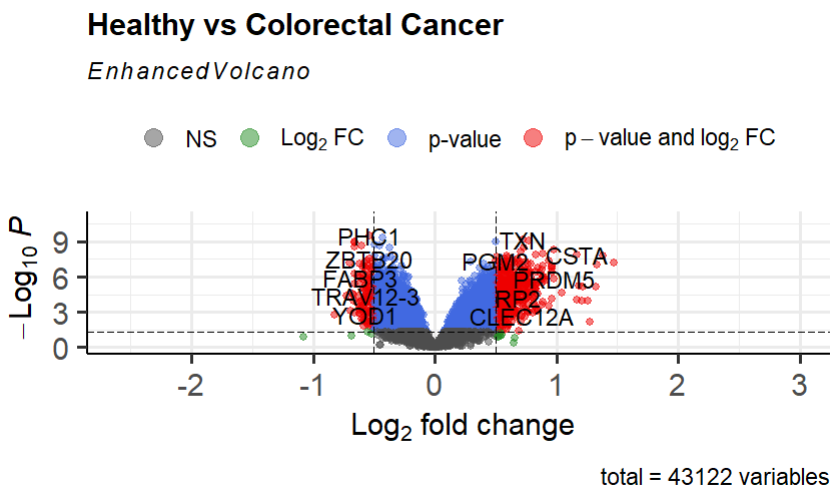
Genes receive annotations about the conditions where differential expression becomes evident (specified as "Normal" or "Colorectal Cancer").

* + 1. **Results and Visualization of DEGs**

Visualizing recently analyzed genes helps scientists easily understand both their significant statistical values and biological meanings. The study focuses on genes showing high fold changes combined with low p-values because these elements strongly indicate differential expression. The implementation of volcano plots creates a visual understanding between statistical significance and fold change evaluation for thousands of analyzed genes.

* + 1. **Volcano Plot**

The volcano plot divides genes into four distinct groups depending on their statistical significance and expression modifications.



The group of NS (Not Significant, Grey Dots) includes genes which do not pass the specified statistical threshold requirements for analysis.

Genes with meaningful expression changes can be found among the green dots although they did not pass statistical significance tests which would require reconsideration when using less strict cut-off points.

The blue dots represent the statistically significant genes which show minimal expression level changes despite their meaningful contribution to overall biological processes.

Red-dotted genes are significant targets due to their statistical significance (adjusted p-value < 0.05) and logFC > 0.5 threshold fulfillment for future investigation.

* + 1. **Observations**

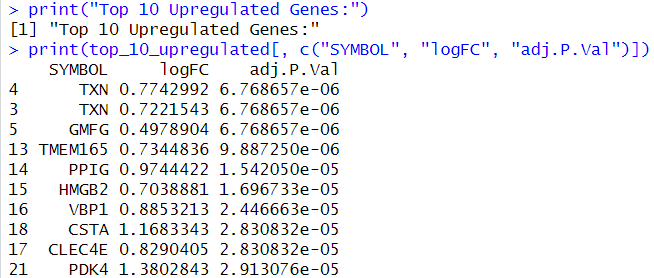
Observations:

• A total of 43,122 variables were analyzed.

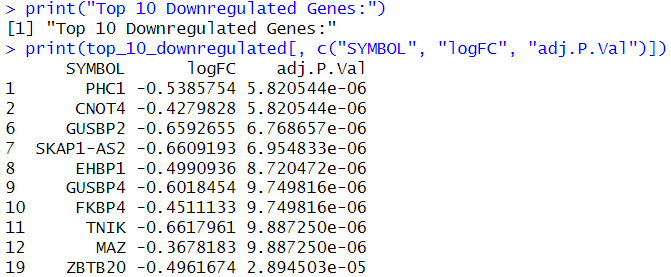
• Genes such as PHC1, CN0T4, GUSBP2, and SKAP1-AS2 are highlighted as significantly differentially expressed.

• Upregulated genes (right side) and downregulated genes (left side) indicate potential biomarkers or targets for further investigation in colorectal cancer research.

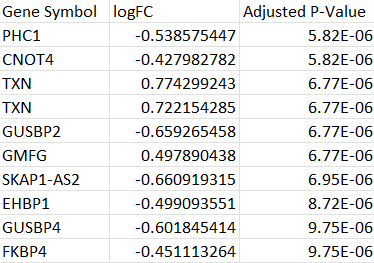
• Up regulated genes show higher expression in the condition of interest compared to the control. It means the gene is more active. They promote growth or prevent cell death.



• Down regulated genes show lower expression in the condition of interest. It means the gene is less active. These are the genes that includes tumor suppressors, whose inactivity allows cancer progression.



* + 1. **Top 10 DEGs**

****

These genes show the most statistically significant and biologically meaningful changes in expression between the normal and cancer samples.

**Significance (Adjusted P-Value):** These genes have the lowest adjusted p-values, meaning the likelihood of observing these expression differences by random chance is very low.

**Magnitude of change (Log Fold Change - logFC):** These genes exhibit the largest positive or negative changes in expression levels. A higher absolute value of logFC indicates a stronger difference in expression between these conditions.

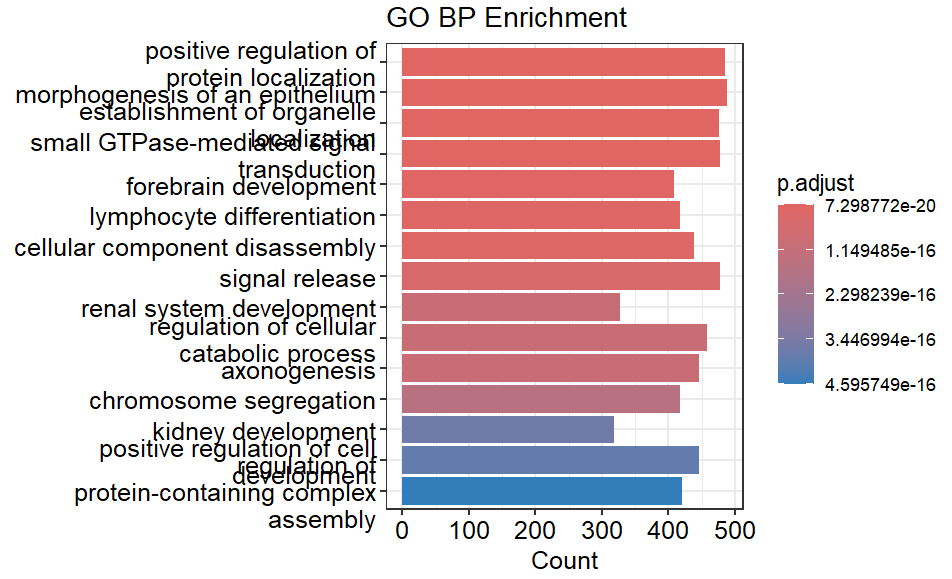
* + 1. **Biological Insights**

**•** Identified genes involved in colorectal cancer pathways.

• Potential biomarkers for early detection and targeted therapy.

* + 1. **Gene Enrichment Analysis**

It is a computational method used to determine whether specific genes, identified from an experiment are overrepresented in certain biological pathways, molecular functions or cellular components. This helps researchers interpret the biological meaning of their gene lists in the context of existing biological knowledge.

****

The above bar plot shows “Gene Ontology Biological Process” enrichment analysis, showing biological processes that are significantly enriched in the dataset.

In Y-axis, These are the genes that are overrepresented like

• Positive regulation of protein localization

• Morphogenesis of an epithelium

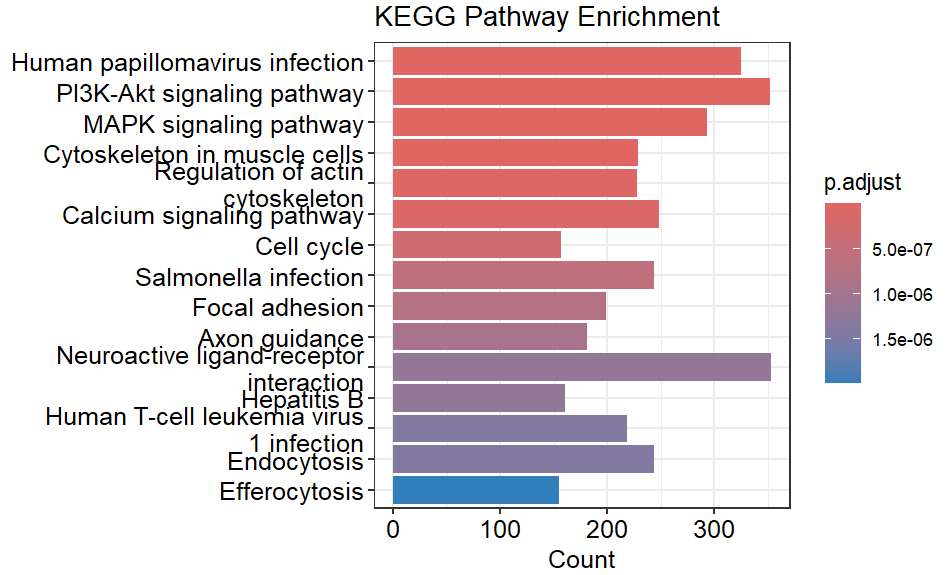
• Forebrain development

• Axonogenesis

• Chromosome segregation

In X-axis, The length of each bar represents the number of genes in dataset that are involved in the respective biological process.

Conclusion, “Positive regulation of protein localization” and “Chromosome segregation” indicates potential functional or regulatory mechanisms influenced by genes.

****

The above bar chart shows the KEGG Pathway Enrichment Analysis, highlighting that are significantly enriched in gene set.

In y-axis, These are the biological pathways where the genes in dataset show significantly involvement like

• Human papillomavirus infection

• PI3K-Akt signaling pathway

• MAPK signaling pathway

• Regulation of actin cytoskeleton

• Cell cycle

In x-axis, Each bar’s length reflects the number of genes from the dataset participating in that pathway. “Human papillomavirus infection” pathway has the highest count, suggesting strong gene representation.

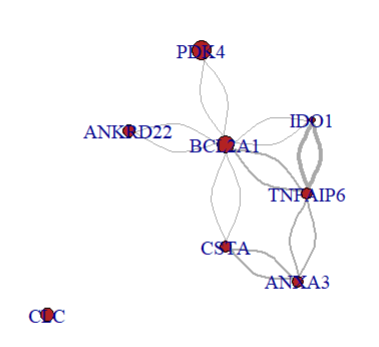
Conclusion, pathways such as “PI3K-Akt signaling pathway” and “MAPK signaling pathway” are often associated with key processes like cell growth, survival and differentiation.

Pathways like “Cell cycle” emphasize cellular regulation, which may be crucial in cancer-related studies.

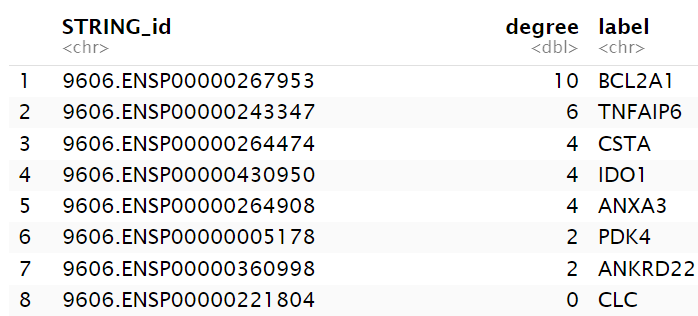
Pathways like “Human papillomavirus infection” suggest potential links to viral mechanisms influencing gene expression.

* 1. **Protein–Protein Interaction (PPI) Network**

It is used to visualize and analyze the relationships between proteins in the dataset.



This is the graph for these proteins in the dataset.



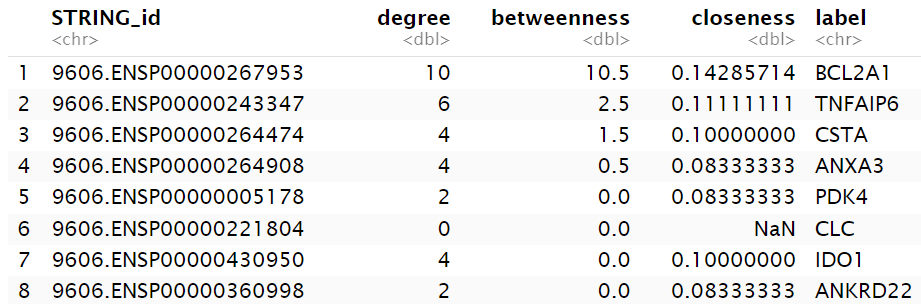
We can identify that BCL2A1 as the top hub gene, which has the highest degree 10, indicating that it is highly interconnected with other DEGs.

TNFAIP6 has degree of 6 and CSTA, IDO1, and ANXA3 each had a degree of 4.

High-degree genes are considered biologically important because they represent key regulatory nodes involved in disease mechanisms.

Hub genes such as BCL2A1 and TNFAIP6 can be biomarkers for colorectal cancer identification.

* + 1. **Centrality Analysis**

****

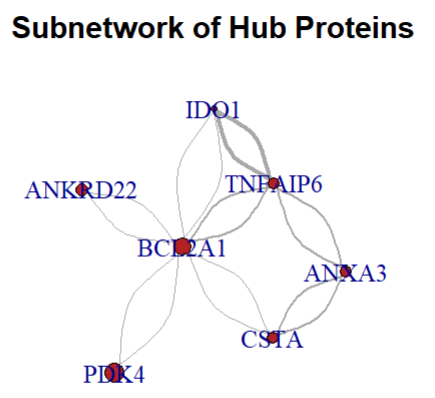
Centrality analysis revealed that BCL2A1 had the highest betweenness centrality (10.5) and degree (10), indicating its role as a key bottleneck and hub in the network.

High betweenness centrality implies that BCL2A1 lies on many shortest paths between other nodes, suggesting it may regulate communication and signal flow within the network. This highlights BCL2A1 as a potentially important regulator among DEGs.

TNFAIP6 also showed moderate betweenness (2.5) and a degree of 6, implying it may also play a significant but less dominant role in network communication.

Nodes like CSTA and ANXA3 showed low betweenness values (1.5 and 0.5) indicating more local connectivity but less global control over network flow.

Genes such as PDK4, IDO1 and ANKRD22 has low degrees and zero betweenness while CLC has a degree of 0. Thus, it is an isolated node with limited influence in the network.

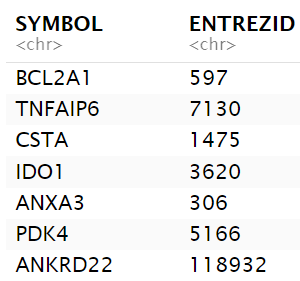


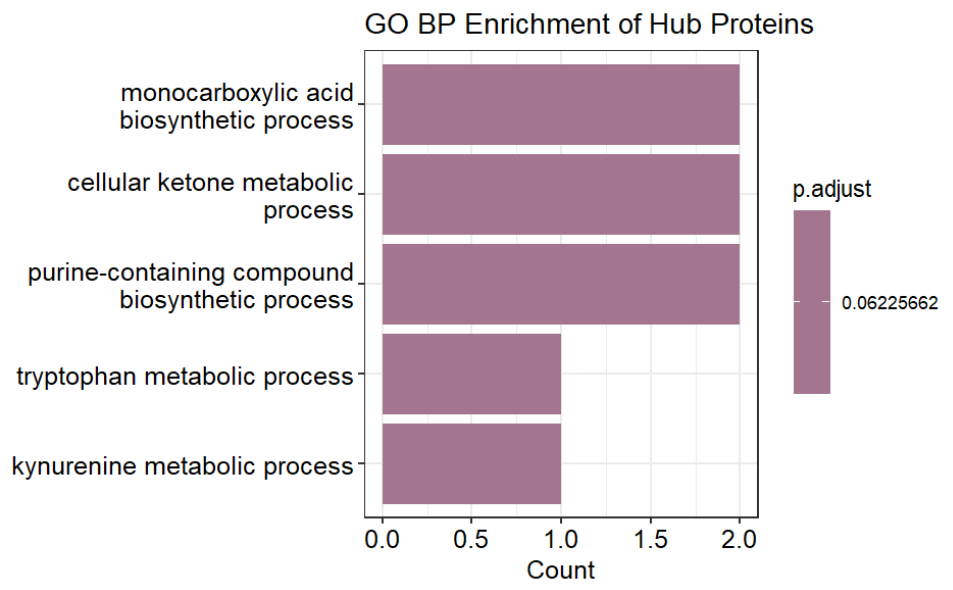
As shown in the figure, BCL2A1 as the core hub in the network, potentially playing a major regulatory role. TNFAI6 and ANXA3 form secondary connections suggesting they may cooperate with BCL2A1o involved in similar biological processes.

* + 1. **Gene Ontology Biological Process (BP) enrichment and KEGG pathway enrichment analysis**

To perform enrichment analysis, the hub gene symbols need to converted into standardized identifiers.

The bitr() function from the clusterProfiler package was used to map the hub genes from gene symbols to Entrez UDs, using the org.Hs.eg.db human gene annotation database.





The hub genes were first mapped to Entrez IDs and enrichment was performed focusing on the Biological Process (BP) category. A p-value cutoff of 0.1 and 1-value cutoff of 0.3 were applied to retain terms of potential biological relevance.

The bar plot displays the top five enriched biological processes:

• Monocarboxylic acid biosynthetic process

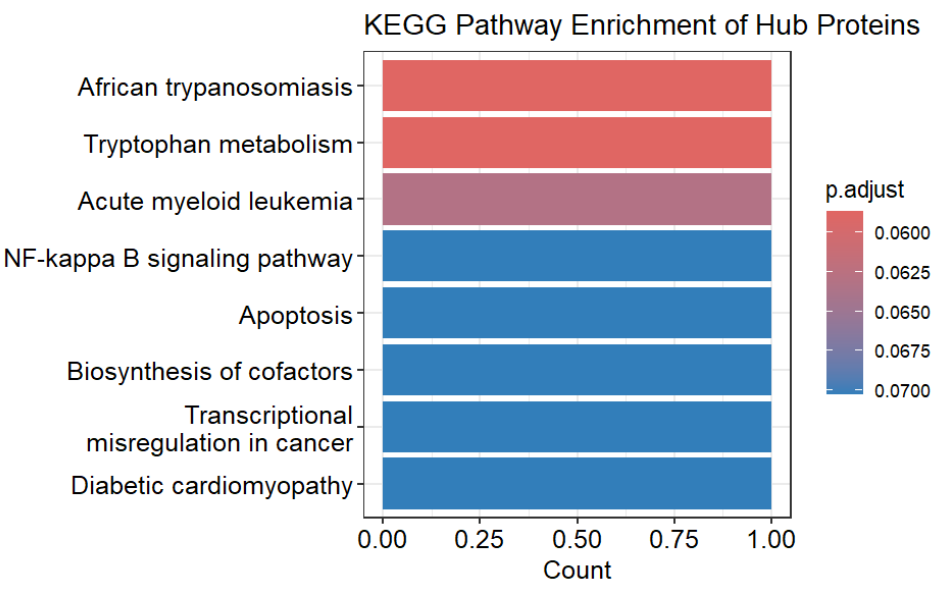
• Cellular ketone metabolic process

• Purine-containing compound biosynthetic process

• Tryptophan metabolic process

• Kynurenine metabolic process

These processes suggest that the hub proteins are primarily involved in metabolic and biosynthetic pathways, and are particularly related to small molecules and amino acid metabolism.



For further investigate the functional roles of the hub proteins, KEGG pathway enrichment analysis was performed using the enrichKEGG() function from the clusterProfiler package.

The bar plot shows top enriched KEGG pathways associated with the hub proteins:

• African trypanosomiasis

• Tryptophan metabolism

• Acute myeloid leukemia

• NF-kappa B signaling pathway

• Apoptosis

• Biosynthesis of cofactors

• Transcriptional misregulation in cancer

• Diabetic cardiomyopathy

NF-kappa B signaling pathway, Apoptosis, Biosynthesis of cofactors, Transcriptional misregulation in cancer are critically involved in immune response, programmed call death and cancer development.

Also, the identification of tryptophan metabolism is consistent with previous findings in the GO enrichment, highlighting its potential role in biological functions of the hub proteins.

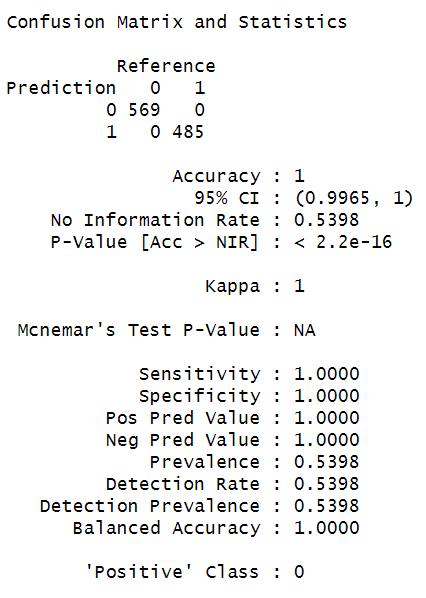
* 1. **Machine Learning model**

**Random Forest**

Implements random forest to check the importance of each feature in DEGs dataset.

Input features: "LOGFC", "AVEEXPR", "T", "P.VALUE", "ADJ.P.VAL"

Output label: ‘’Condition’’

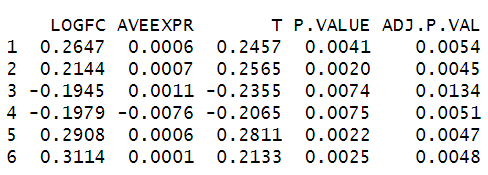


* 1. **Explainable Artificial Intelligence techniques**

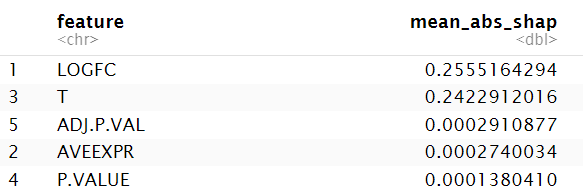
Applying SHAP,

Using fastShap library,

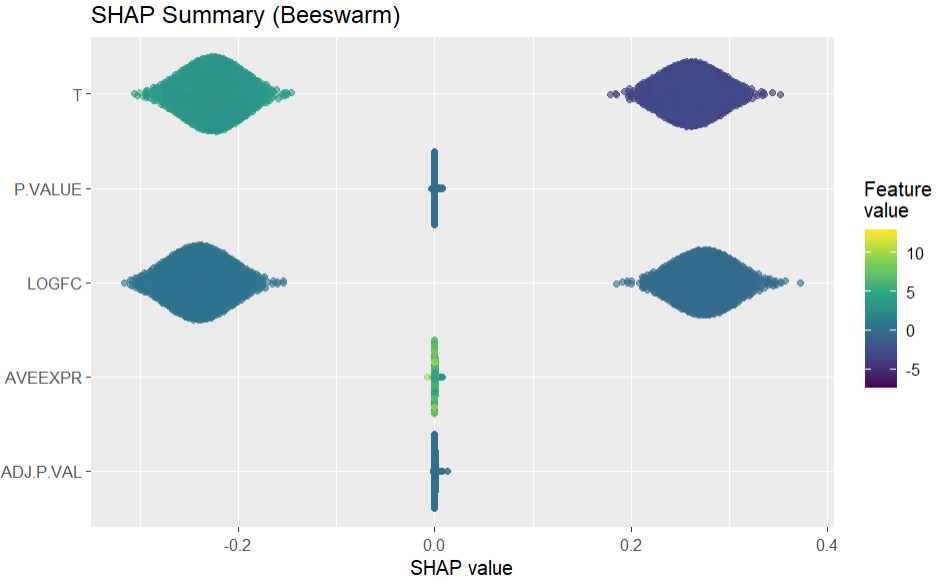
Calculating SHAP values for each features.



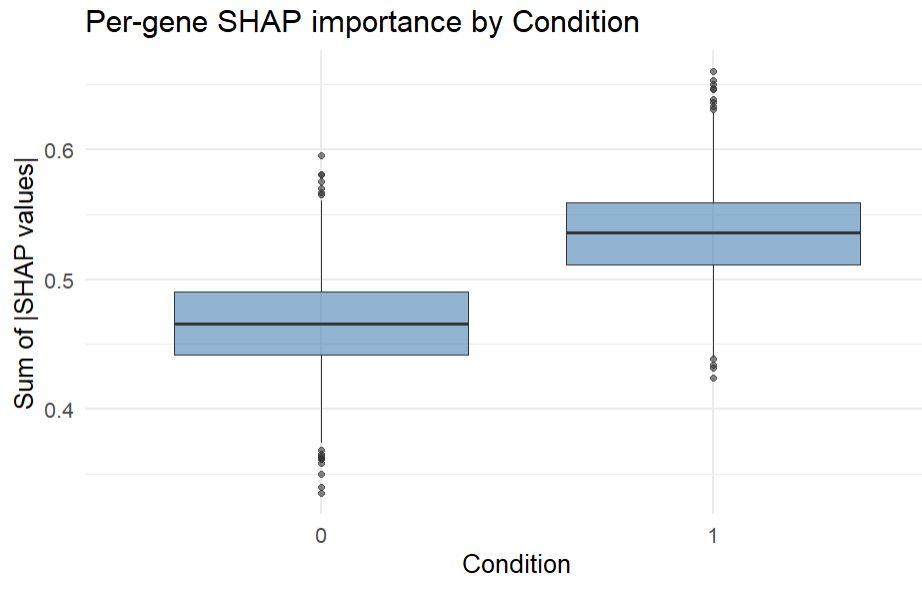
Calculating mean shap values



Plotting Beeswarm plot,



For t-static, a wide spread of SHAP values is on both sides and higher T values which are green in color push predictions higher and lower T values which are purple in color push predictions lower. So, T is a strong positive contributor to the model. High t-values increases model output. Similarly Log fold change is also strong predictive feature which upregulated genes (positive logFC) increase predicted probability. Whereas P.value, average expression (Aveexpr) and Adj.p.val have very low effect on model prediction.



This boxplot visualizes the pre-gene SHAP importance by condition 0 vs 1 based on the sum of absolute SHAP values per gene.

The median SHAP value sum is higher for condition indicating that the model relies more heavily on genes under this condition to make predictions. Condition 1 contributes more distinctive patterns to the model compared to Condition 0. It shows that gene expression patterns under this condition will provide more discriminative information.

1. **LIMITATIONS AND FUTURE ENHANCEMENTS**
   1. **Limitations**

Although the research was successful in detecting differentially expressed genes (DEGs) and studying their biological meaning using statistics, enrichment, and machine learning techniques, a number of limitations exist:

**Dataset Size and Diversity**: This study is based on one dataset (E-MEXP-3756) from ArrayExpress and its context may not be representative of diverse clinical conditions and genetic backgrounds. These results may be different with heterogeneous and larger datasets.

**Microarray technology** platforms have been known to lack dynamic range and sensitivity in comparison to other emerging technologies such as RNA-Seq, which may be capable of providing finer resolution insights.

**Annotation Bias**: Probe annotation depends on existing gene databases. Probes that map to out-of-date or poorly annotated genes can cause inaccuracies.

**Lack of Experimental Validation:** The identified biomarkers and pathway analyses are computationally derived. Without wet-lab validation (e.g., qPCR or Western blotting), biological relevance remains hypothetical.

**Limited clinical metadata**: This dataset does not contain elaborate clinical variables (e.g., tumor stage, age of patient), which limits one's capability to conduct stratified analyses or to correlate clinical outcomes with gene expression.

* 1. **Future enhancements**

Multiple frontiers of research are put forward to increase the translational utility and robustness of the results

**Cross-Platform Validation:** Involving more than one set of datasets from different platforms, e.g., RNA-Seq, may cross-validate and enhance the reproducibility of the DEGs identified to promote applicability in multiple technical and biological contexts.

**Clinical metadata integration:** By incorporating detailed clinical variables, we will be able to perform survival analysis and model prognosis and stratified biomarker discovery, and so enhance the clinical validity and utility of the biomarkers.

**Experimental Validation:** These hub genes including BCL2A1 and TNFAIP6 need to be validated by experimental methods (e.g., qPCR and Western blotting) to verify the biological importance and therapeutic significance of these genes.

**Advanced machine learning** may provide enhanced feature extraction and predictive performance with the use of methods from deep learning that can be used to build more reliable diagnostic and predictive models.

**Multi-Omics Integration:** Integrating other omics layers like proteomics and metabolomics will give a complete and multidimensional insight into the pathogenesis and development of colorectal cancer.

1. **CONCLUSION**

The E-MEXP-3756 microarray data was analyzed in this project to identify differentially expressed genes in colorectal cancer with the help of R and Bioconductor software. Statistical filtration was used to successfully identify 4,748 DEGs and visualized the major findings with PCA, heatmaps, and volcano plots. Gene enrichment and pathway studies identified biological processes and signaling pathways that were majorly implicated in colorectal cancer.

A protein–protein interaction network analysis revealed hub genes such as BCL2A1 and TNFAIP6 to be promising candidates playing a key regulatory role in cancer development. Machine learning deployment via Random Forest and SHAP yielded a readable model that represented the most significant influencing features with the use of T-statistic and log fold change.

Overall, the research illustrates the power of bioinformatics and explainable AI in unearthing biomarkers of early cancer detection and identifies various areas of further investigation and clinical translation.