GSE24514

Abstract

In this study, we delved into the GSE24514 dataset, which encompasses the expression profiles of 34 microsatellite instability (MSI) colorectal cancers and 15 normal colonic mucosas. Generated in 2002, the dataset focused on comparing malignant and healthy tissues. Microsatellite instability, arising from defective mismatch repair, is observed in a specific subset of colorectal cancers (CRCs). We assessed somatic mutations in microsatellite repeats of genes chosen due to reduced expression in MSI CRC and the presence of a coding mononucleotide repeat.

A variety of analyses were carried out utilizing this gene expression profile. Genes that underwent significant alterations were pinpointed, and enrichment analyses were executed on these genes. Additionally, the clustering status of the data was scrutinized. Interactions among genes were thoroughly examined to uncover insights into their relationships. The results of these analyses provided a comprehensive understanding of the compatibility status within each studied cell population, as well as the distinctions and similarities between them.

The primary aim of this study was to meticulously examine the selected dataset using diverse analytical approaches. Through this investigation, we sought to illuminate the extent of gene alterations in specific conditions within the dataset, determine the significance of gene expressions, and evaluate the impact of condition changes on the expression of these genes.

1. Literature Information

1.1. Article Information

Colorectal cancers (CRCs) with microsatellite instability (MSI) arise from deficiencies in DNA mismatch repair (MMR) and occur in 15% of non-metastatic cases and 5% in metastatic conditions. Approximately 30% of MSI CRCs are associated with constitutional mutations in the MMR system, such as Lynch syndrome. Pathogenic alterations in MMR genes result in the accumulation of somatic mutational events, endowing these tumors with a high antigen burden and intense infiltration of cytotoxic T-cell lymphocytes. The MSI/dMMR status holds prognostic and predictive significance in both non-metastatic and metastatic CRCs. While the prognostic value of MSI status in non-metastatic CRCs has been extensively studied, data regarding its predictive value for adjuvant chemotherapy efficacy remain more limited. In both metastatic and non-metastatic settings, treatment with immune checkpoint inhibitors (ICIs) has demonstrated remarkable effectiveness in the context of MSI/dMMR status.

Recent data from prospective cohorts and randomized trials have indicated a substantial improvement in survival with immunotherapy (programmed death-ligand 1 [PD-(L)1] cytotoxic T-lymphocyte-associated antigen 4 [CTLA-4] blockade) for metastatic or non-metastatic MSI/dMMR CRC. This review provides insights into testing methods for the MSI/dMMR phenotype, the prognostic value of this phenotype, and new treatment recommendations for this unique CRC population. Despite their efficacy, primary and secondary resistance to immune checkpoint inhibitors (ICIs) are observed in over 50% of MSI-H/dMMR CRC patients, posing an important challenge for identifying these patients and overcoming resistance in the future.

In this study, expression data from human MSI colorectal cancer and normal colonic mucosa were investigated in a dataset belonging to the Homo sapiens species. The dataset comprises a total of 49 samples, with 34 belonging to the tumor group and 15 representing the normal (control group).

1.2. Analysis Information

The data analysis employed various analytical tools and methodologies, including Principal Component Analysis (PCA). PCA, a method for dimensionality reduction, seeks to identify principal components (PCs), which are linear combinations of the original measurements. These PCs, being orthogonal to one another, can effectively represent the impacts of the original measurements with substantially reduced dimensions. The widespread utilization of PCA in diverse statistical domains is attributed to its computational simplicity and favorable statistical properties. Notably, in recent bioinformatics investigations, particularly gene expression studies, PCA has been instrumental in reducing the dimensionality of high-throughput measurements.

To identify significantly altered genes in the data, a t-test was employed. The Student's t-test assesses the mean and standard deviation of two samples to determine if a statistically significant difference exists between them. In experimental settings, a t-test aids in discerning whether differences between the control and experimental groups result from the manipulated variable or are merely random fluctuations.

The elucidation of associations among substantially altered genes was conducted through Gene Ontology (GO) analysis. GO enrichment analysis is widely employed to evaluate high-throughput molecular data, offering hypotheses about the underlying biological processes in experiments. GO, a prominent ontology, categorizes genes based on cellular location, molecular function, and biological process. In gene expression microarray results, GO analysis begins by compiling genes exhibiting differential expression. Subsequently, GO enrichment analysis determines whether specific biological processes, molecular functions, or cellular components are over- or under-represented in the gene set, providing valuable insights into the biological significance of changes in gene expression levels.

For visual representation, scatter charts were employed due to their efficacy in quickly comparing two data sets. However, identifying gene groups with similar expression profiles across studies is often crucial. Cluster analysis, a method frequently utilized to unveil functionally connected genes, is employed to uncover such relationships, with resulting clusters often indicative of biological processes. Hierarchical clustering, the most popular mathematical method, arranges genes into small clusters and clusters into higher-level systems, resembling a dendrogram in its hierarchical tree structure. The current project's analysis utilized hierarchical clustering to unravel meaningful patterns in the data.

2. Methods

Initially, a research was conducted on the Gene Expression Omnibus (GEO) database site to identify appropriate data for analysis. The selected dataset was sourced from the article titled Candidate driver genes in microsatellite-unstable colorectal cancer' (GSE24514) and the study titled Expression data from human MSI colorectal cancer and normal colonic mucosa. Subsequently, the data with the GSE24514 identifier was retrieved using the R programming language. Following this, the expression of this data was transformed into a matrix, and its logarithm (base 2) was taken for subsequent analysis, serving as the foundational data for the analysis steps.

The dataset encompasses a total of 49 samples, with 34 representing tumor cell and 15 normal cell. To observe the grouping behaviors of these samples, subsets of samples were subjected to Principal Component Analysis (PCA). Subsequently, genes that exhibited significant changes between conditions, specifically between lung tissue and spleen tissue, were identified using the Benjamini-Hochberg corrected t-test. A list of these genes was compiled.

Following this, g:Profiler was employed to conduct Gene Ontology (GO) Analysis on the genes in this list. The analysis focused on biological processes and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways, providing results that facilitated the comparison of the most affected metabolic pathways and biological processes.

Hierarchical clustering of samples, initially performed using Pearson correlation with the Centroid method, aimed to assess whether samples from the same condition clustered together. Additionally, to better visualize the distinction, a heatmap was generated..

Subsequently, a Support Vector Machine (SVM) model was trained, and two unknown samples were subjected to the model for class prediction, utilizing known samples. The accuracy of predictions was verified. Network inference was then conducted for the top five genes ranked by p-values from the list of significant variable genes. This analysis offered additional insights into molecular interactions and regulatory relationships among the top-ranked genes.

In the final stage, the GGm-based network interaction method was employed to examine the interactions of the top 5 genes based on their relevance, and reference consistency was verified.

3. Results

3.1. Principal Component Analysis

PCA is a useful approach for reducing the dimensionality of huge datasets such as gene expression profiles, allowing us to depict complex data in a more understandable way. To visualize the data, a scatter plot was created using the first two principal components. Result of the principal component analysis dataset GSE24154 in Figure 1.

```
Importance of components:
                          PC1     PC2      PC3      PC4      PC5      PC6      PC7      PC8      PC9     PC10     PC11     PC12     PC13
Standard deviation     74.2970 53.0366 37.48700 35.04652 30.28784 25.52497 23.98781 23.51524 21.64336 20.99879 19.75006 19.58949 18.93163
Proportion of Variance  0.2477  0.1262  0.06306  0.05512  0.04117  0.02924  0.02582  0.02482  0.02102  0.01979  0.01751  0.01722  0.01608
Cumulative Proportion   0.2477  0.3740  0.43702  0.49214  0.53331  0.56255  0.58837  0.61319  0.63421  0.65400  0.67151  0.68873  0.70481
                          PC14     PC15     PC16     PC17     PC18     PC19     PC20     PC21     PC22     PC23     PC24     PC25     PC26
Standard deviation     17.73874 17.52710 17.22280 16.45089 16.21015 16.06952 15.71309 15.41212 15.06222 14.88346 14.66078 14.33524 14.30258
Proportion of Variance  0.01412  0.01379  0.01331  0.01215  0.01179  0.01159  0.01108  0.01066  0.01018  0.00994  0.00965  0.00922  0.00918
Cumulative Proportion   0.71893  0.73272  0.74603  0.75818  0.76997  0.78156  0.79264  0.80330  0.81348  0.82342  0.83307  0.84229  0.85147
                          PC27     PC28     PC29     PC30     PC31     PC32     PC33     PC34     PC35     PC36     PC37     PC38     PC39
Standard deviation     14.04861 13.69641 13.66201 13.44898 13.34485 13.17267 13.11823 12.86731 12.6668 12.57056 12.47067 12.3996 12.29522
Proportion of Variance  0.00886  0.00842  0.00838  0.00812  0.00799  0.00779  0.00772  0.00743  0.0072  0.00709  0.00698  0.0069  0.00678
Cumulative Proportion   0.86033  0.86874  0.87712  0.88524  0.89323  0.90102  0.90874  0.91617  0.9234  0.93046  0.93744  0.9443  0.95112
                          PC40     PC41     PC42     PC43     PC44     PC45     PC46     PC47     PC48     PC49
Standard deviation     11.99292 11.88904 11.68377 11.60086 11.41541 10.58620 10.44212 9.73177 9.31074 2.579e-13
Proportion of Variance  0.00645  0.00634  0.00613  0.00604  0.00585  0.00503  0.00489 0.00425 0.00389 0.000e+00
Cumulative Proportion   0.95758  0.96392  0.97005  0.97609  0.98194  0.98697  0.99186 0.99611 1.00000 1.000e+00
```

Figure 1. PCA Results of GSE24514 Dataset.

We used color coding to distinguish between tumor (red) and normal (green) samples in the plot. Each point represents a single sample. The PCA plot, titled "PCA of GSE24514", presents the distribution of the samples along the first two principal components (PC1 and PC2), which capture the most variance in the dataset (Figure 2). As the view of PCA plot,
The tumor and normal groups have been clearly distinguishable from each other.

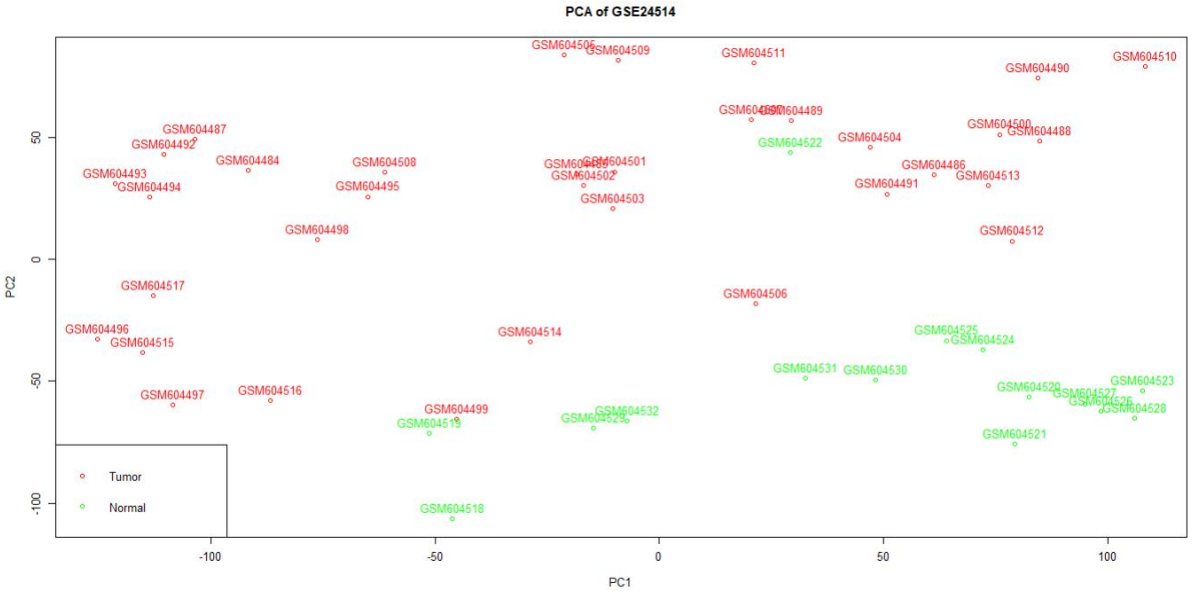

Figure 2. Represents Scatter Plot of PCA the GSE24514 Dataset.

3.2. Benjamini-Hochberg Correction(t-test)

We performed differential gene expression analysis to identify genes that show significant changes in expression between two distinct sample groups (samples 1-34 vs. samples 35-49). For each gene, we applied a two-sample t-test to compare its expression levels between the two groups. This yielded a p-value for each gene, representing the probability of observing the given difference in expression by chance. We considered genes with a Benjamini-Hochberg corrected p-value of less than 0.05 as significantly differentially expressed. Using this approach, we identified a total of 8,329 genes that exhibited significant changes in expression between the two groups. We described the most significant genes as **CXCL9, STIL,** and **MICB.**

3.3. GO Enrichment Analysis

We conducted a thorough enrichment analysis to determine the biological significance of the genes identified as significantly differentially expressed (**CXCL9, STIL,** and **MICB**). The analysis concentrated on two areas: GO Biological Process (BP) categories and KEGG pathway annotations. The GO enrichment analysis for Biological Processes identified key processes that are significantly overrepresented among differentially expressed genes. The terms identified include **'organelle organization', 'peptidyl-lysine modification','regulation of signaling', 'RNA 3'-end processing', 'positive regulation of biological process',** and **'regulation of intracellular signal transduction'** (Figure 3). Each of these processes is critical for cellular function and regulation.

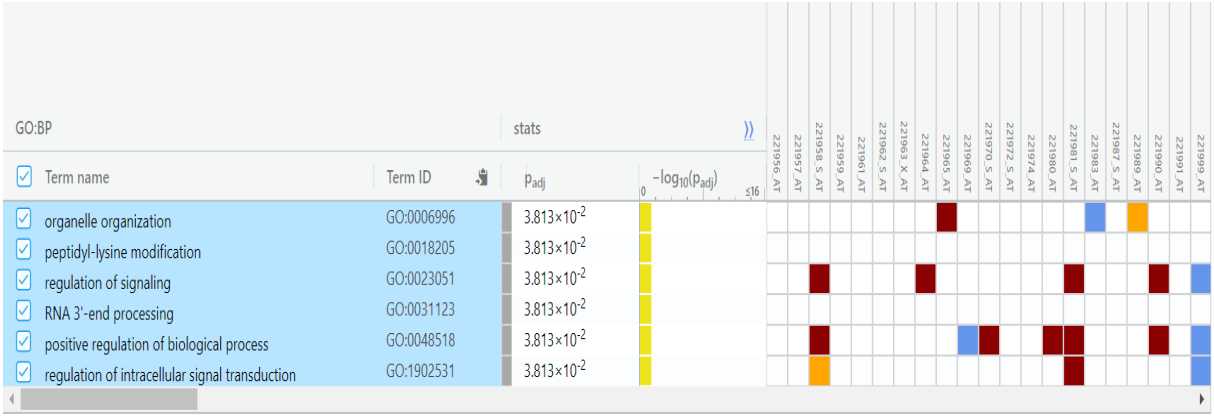

Figure 3. The figure illustrates the significant GO terms in the left panel, along with their associated adjusted p-values, and a heatmap in the right panel demonstrating the significance of gene-term associations. The heatmap enables quick identification of which genes are driving enrichment for each term.

KEGG pathway analysis was performed to identify significant pathways enriched among the differentially expressed genes. However, the analysis did not yield any significantly enriched pathways at the chosen level of statistical significance.

3.4. Hierarchical Clustering

To investigate the inherent groupings among our samples based on their expression profiles, we performed a hierarchical clustering analysis. The resulting dendrogram (Figure 4) illustrates the relationships among the samples. Each leaf on the dendrogram represents an individual sample, labeled according to our sample labels dataset. The height of the branches indicates the dissimilarity between clusters, with shorter branches representing more closely related samples. As depicted in the dendrogram, we can observe the formation of several distinct clusters, indicating the presence of different groups within our samples. The cluster formation suggests that there are significant differences in the expression profiles between these groups.

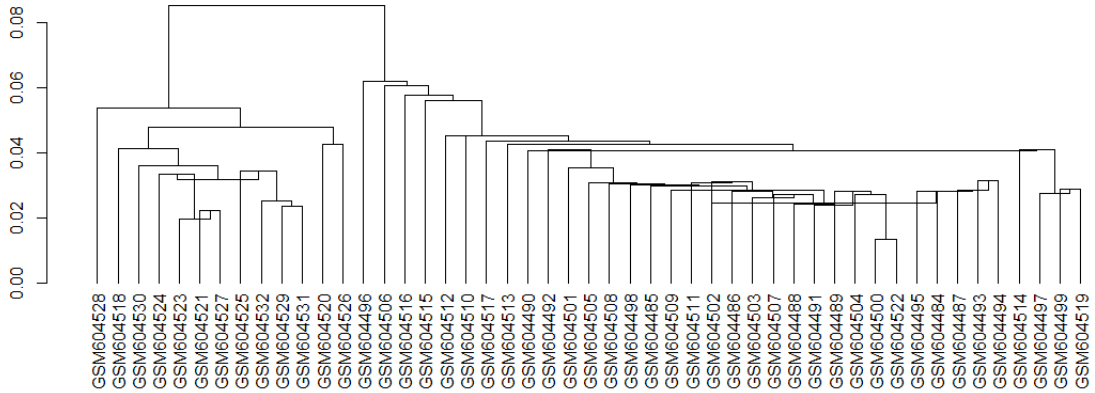

Figure 4. Dendrogram of GSE24514 dataset (Hierarchical Clustering). Normal samples are positioned on the left side, while tumor samples are situated on the right side.

We chose to use a heatmap to visualize the hierarchical clustering between genes and samples. In this context, rows represent genes, and columns represent sample groups. According to the generated heatmap graph, there is a clear distinction among different sample groups. Tumor samples are clustered on one side, while normal samples are displayed on the other side. Hierarchical cluster analysis effectively separates tumor and normal samples, highlighting significant differences in gene expression between these two groups (Figure 5).

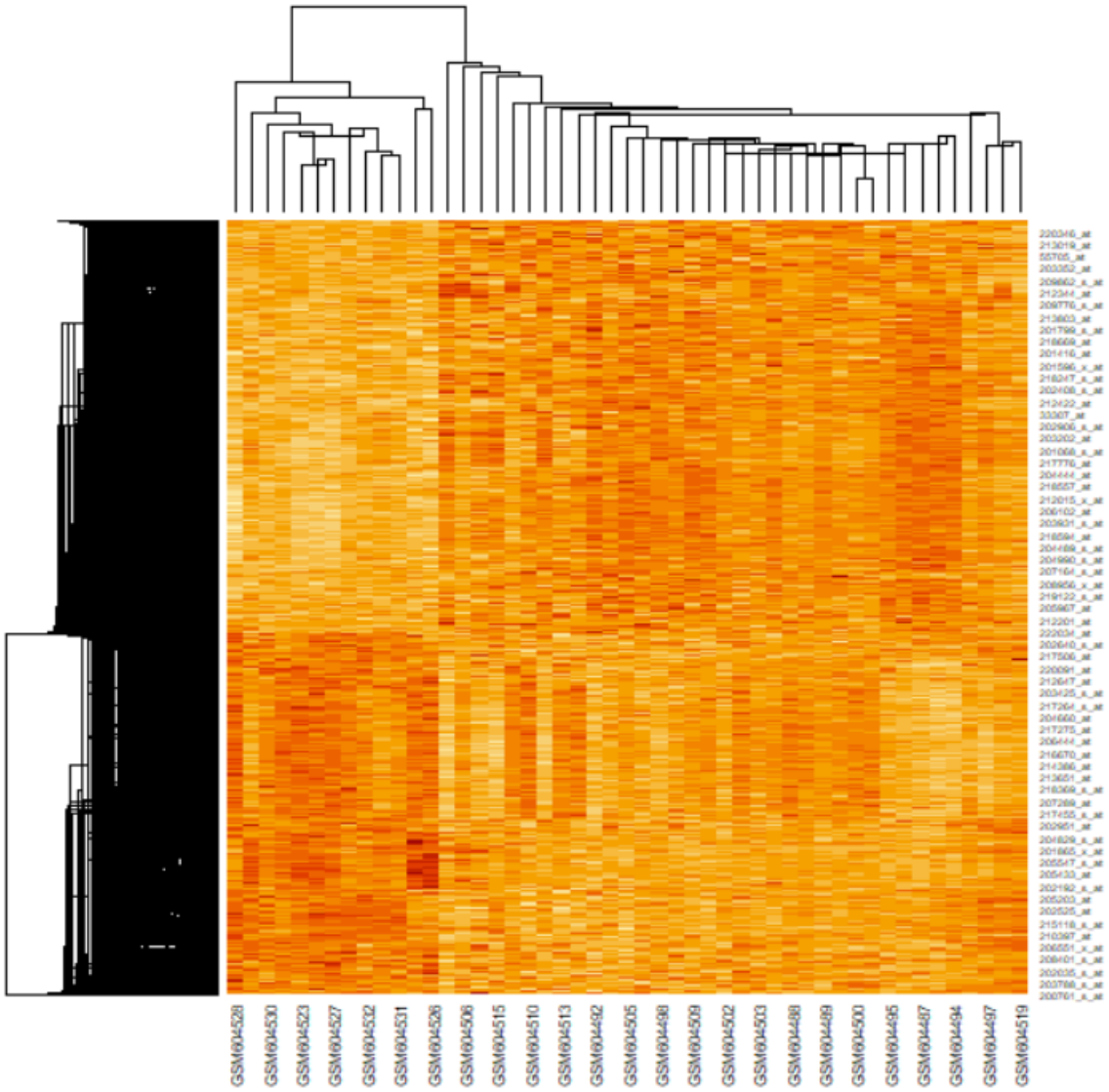

Figure 5. Heatmap of GSE24514 Dataset for Hierarchical Clustering.

3.5. Support Vector Machine(SVM)

A train_data and test_data group have been initially defined for the Support Vector Machine (SVM). The train_data includes samples from the 2nd to the 48th group; within this range, there are 33 samples in the tumor group and 14 in the normal group. For testing, the 1st group (tumor) and the 49th group (normal) have been specified. After training the SVM using train_data and the corresponding labels, it was applied to the test_data, and the results were examined. According to the outcome, GSM4484, representing the tumor sample, returned a result of 0 as expected, while GSM604532, representing the normal sample, returned a result of 1 as intended. Our SVM model is functioning correctly and possesses the capability to distinguish the relevant data (Figure 6).

```
Call:
svm(formula = label ~ ., data = train_data, type = "C-classification",
    kernel = "linear", cost = 10, scale = TRUE)


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  linear
       cost:  10

Number of Support Vectors:  19

 ( 13 6 )


Number of Classes:  2

Levels:
 0 1



>
> try_svm <- predict(svm_model, train_data) #control svm with train_data
> print(table(try_svm))
try_svm
 0  1
33 14
>
> predictions <- predict(svm_model, test_data) #test_data
> predictions
GSM604484 GSM604532
        0         1
Levels: 0 1
```

Figure 6. Support Vector Machine output of GSE24514 Dataset.

3.6. GGM-Based Network Inference

This analysis, conducted using the Real-Time Genetic Granger Causality Model (GGM), aims to understand the complexity of genetic regulation networks and evaluate correlations and interactions among genes. The correlation matrix highlights strong relationships between genes, while the interaction matrix illustrates whether these relationships translate into direct interactions.

At this stage, we applied the pcor function to our data, obtaining statistical values such as estimates and p-values for the top 5 genes (Figure 7).

```
$estimate
             203915_at  205339_at   206247_at 209545_s_at   218404_at
203915_at    1.0000000 -0.3387418  0.46290321   0.0963501  0.53751235
205339_at   -0.3387418  1.0000000  0.57562316   0.1256896  0.33285317
206247_at    0.4629032  0.5756232  1.00000000   0.2558701 -0.06352695
209545_s_at  0.0963501  0.1256896  0.25587009   1.0000000  0.27194557
218404_at    0.5375124  0.3328532 -0.06352695   0.2719456  1.00000000

$p.value
               203915_at    205339_at    206247_at 209545_s_at    218404_at
203915_at   0.0000000000 2.129144e-02 1.198074e-03  0.52413869 0.0001171513
205339_at   0.0212914374 0.000000e+00 2.849173e-05  0.40523056 0.0238044721
206247_at   0.0011980741 2.849173e-05 0.000000e+00  0.08610026 0.6749038587
209545_s_at 0.5241386947 4.052306e-01 8.610026e-02  0.00000000 0.0675058206
218404_at   0.0001171513 2.380447e-02 6.749039e-01  0.06750582 0.0000000000

$statistic
             203915_at  205339_at 206247_at 209545_s_at 218404_at
203915_at    0.0000000 -2.3881477  3.464036   0.6421016  4.228199
205339_at   -2.3881477  0.0000000  4.669416   0.8403952  2.341408
206247_at    3.4640364  4.6694162  0.000000   1.7556952 -0.422243
209545_s_at  0.6421016  0.8403952  1.755695   0.0000000  1.874529
218404_at    4.2281989  2.3414084 -0.422243   1.8745289  0.000000

$n
[1] 49

$gp
[1] 3

$method
[1] "pearson"
```

Figure 7. The result of pcor Function of GSE24514 Dataset.

Subsequently, we attempted to answer the question of which genes have interactions by waiting for specific threshold values for estimate and p-value obtained. In this stage, an interaction is interpreted for pairs of genes that return a value of 1 in the matrix we obtained (Figure 8).

	203915_at	205339_at	206247_at	209545_s_at	218404_at
203915_at	1	0	1	0	1
205339_at	0	1	1	0	0
206247_at	1	1	1	0	0
209545_s_at	0	0	0	1	0
218404_at	1	0	0	0	1

Figure 8. Interaction Matrix for Most Significant 5 Genes.

The correlation matrix, designed to assess relationships between genes in the dataset, reveals robust connections. Specifically, when examining correlations with the MICB gene, it emphasizes its direct associations with other genes. For instance, the CXCL9 gene exhibits a positive correlation with MICB and SNX10 genes, suggesting potential involvement in similar biological processes or regulatory interactions. Similarly, the positive correlation between the STIL gene and CXCL9/MICB genes implies potential interactions along common biological pathways.

The interaction matrix delineates direct interactions among genes. For example, the SNX10 gene shows positive correlations with CXCL9 and MICB genes but displays a negative correlation with the RIPK2 gene. This suggests that SNX10 may collaborate with specific genes or be mutually regulated in distinct biological contexts. The absence of interactions involving the RIPK2 gene suggests it may have a particular isolation or play a specialized role in the genetic regulation network (Figure 9).

	CXCL9	STIL	MICB	RIPK2	SNX10
CXCL9	1	0	1	0	1
STIL	0	1	1	0	0
MICB	1	1	1	0	0
RIPK2	0	0	0	1	0
SNX10	1	0	0	0	1

Figure 9. Interaction Matrix Include Gene Names.

The visual representation of the relevant matrices and interactions between genes can be found in Figure 10.

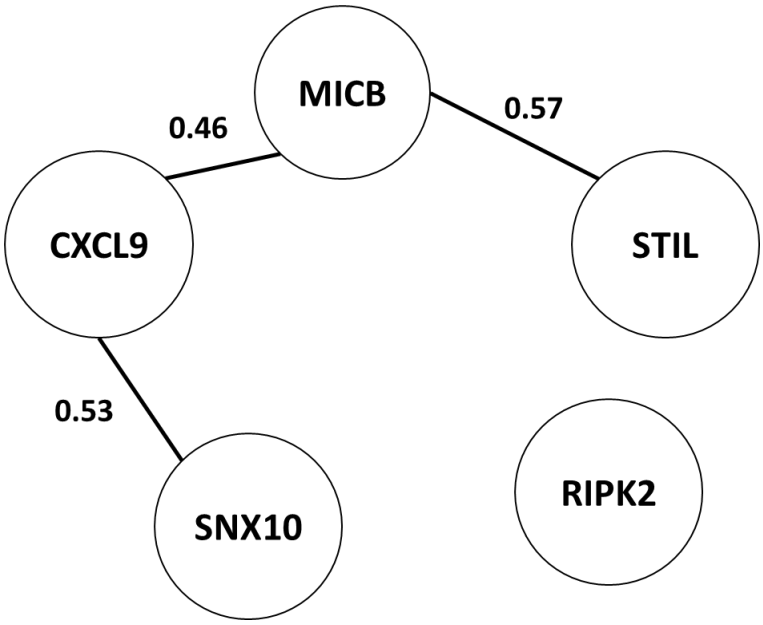

Figure 10. Visualization of Interaction between Significant 5 Genes.

This GGM-based analysis contributes to understanding the intricacies of genetic regulation networks. However, further statistical details and experimental studies are necessary to fully elucidate the meaning of genetic interactions. This assessment represents a significant step in comprehending causal relationships between genes and exploring the regulation of biological systems more comprehensively.

4. Discussions

We selected data for analysis from studies available in the omnibus database. The study was conducted on the Homo sapiens organism, specifically focusing on the gene expression profile titled "human MSI colorectal cancer and normal colonic mucosa," encompassing a total of 49 samples. Among these samples, 34 belong to the tumor cell group, while 15 are part of the normal or control cell group. The study's dataset was loaded into the R program using the access number GSE24514. We converted the loaded data into matrix format as the analysis will be performed in matrix format.

PCA analysis allowed us to observe the grouping behavior of the data in both different tissues. In the obtained output, it was consistently observed that samples from the same group tended to cluster together. At this stage, no group was excluded from the analysis as we did not observe any outliers.

To determine the biological significance of the genes identified as significantly differentially expressed, we conducted a comprehensive enrichment analysis. The analysis focused on two areas: GO Biological Process (BP) categories and KEGG pathway annotations. GO enrichment

analysis for Biological Processes revealed key processes significantly overrepresented among differentially expressed genes. Identified terms included 'organelle organization,' 'peptidyl-lysine modification,' 'regulation of signaling,' 'RNA 3'-end processing,' 'positive regulation of biological process,' and 'regulation of intracellular signal transduction' . Each of these processes plays a critical role in cellular function and regulation.

We performed hierarchical clustering on the data to group the samples, finding that samples under the same condition set tended to cluster together. In the dendrogram created as a result of clustering, tumor and normal cell groups are located on different branches, indicating that the data clustered as expected.

We created a Support Vector Machine (SVM) model based on known examples to predict unknown samples. By training the SVM model with 47 groups, we determined its ability to accurately predict which condition a given unknown sample belongs to. After building the SVM model, we selected one tumor sample and one test sample from unknown samples. We observed that the model provided accurate prediction results and correctly identified the chosen normal and tumor group samples at this stage.

In the GGM-based analysis, the interactions among five identified key genes were explored, revealing the presence of genes correlated with each other. While these genes did not directly interact, the functions of the genes displaying correlation in our study were found to be consistent with the functions and mechanisms influenced by the genes in the reference article.

5. References

1. Alhopuro P, Sammalkorpi H, et. al. Candidate driver genes in microsatellite-unstable colorectal cancer. Int J Cancer. 2012 Apr 1;130(7):1558-66. doi: https://doi.org/10.1002/ijc.26167 . Epub 2011 Aug 3. PMID: 21544814.
2. 2. https://doi.org/10.1016/j.ejca.2022.07.020
3. 3. https://doi.org/10.3748%2Fwjg.v20.i15.4230
4. Tomczak, A., Mortensen, J.M., Winnenburg, R. et al. Interpretation of biological experiments changes with evolution of the Gene Ontology and its annotations. Sci Rep 8, 5115 (2018).
5. https://doi.org/10.1038/s41598-018-23395-2
6. Loeb LA. A mutator phenotype in cancer. *Cancer Res* 2001; 61: 3230–3239.
7. Ferreira AM, Westers H, Wu Y, Niessen RC, Olderode-Berends M, van der Sluis T, van der Zee AG, Hollema H, Kleibeuker JH, Sijmons RH, Hofstra RM. Do microsatellite instability profiles really differ between colorectal and endometrial tumors? *Genes Chromosomes Cancer* 2009; 48: 552–557.
8. Wood LD, Parsons DW, Jones S, Lin J, Sjöblom T, Leary RJ, Shen D, Boca SM, Barber T, Ptak J, Silliman N, Szabo S, et al. The genomic landscapes of human breast and colorectal cancers. *Science* 2007; 318: 1108–1113.