IT-Enabled WGCNA for Critical Gene Module Mapping and Therapy Optimization: Advancing Leukemia Care

Rinela Kapçiu[1,*], Brikena Preni[2], Eglantina Kalluçi[3]
[1] Department of Computer Science, Faculty of Information Technology, Aleksander Moisiu University of Durres, Albania
[2] Department of Mathematics, Faculty of Engineering Mathematics and Engineering Physics, Polytechnic University of Tirana, Tirana, Albania
[3] Department of Applied Mathematics, Faculty of Natural Sciences, University of Tirana, Tirana, Albania
*Correspondence: rinelakapciu@uamd.edu.al
Received 4 April 2024; Revised 8 April 2024; Accepted 8 April, 2024
Available online 9 April 2024 at www.atlas-tjes.org, doi: 10.22545/2024/00252

Abstract: *Acquiring a profound comprehension of the complex correlation between genes is essential for progressing treatment and diagnostics in acute leukemia research. This work utilizes WGCNA to analyze the gene expression data obtained from 72 patients diagnosed with acute leukemia. This approach has not been widely employed in this particular setting previously. Our first objective was to identify essential gene modules and core genes that could offer a novel insight into the fundamental causes of the disease. To ensure the precision of the gene expression data, a thorough pre-processing is carried out, which includes normalization and quality control techniques. The WGCNA technique produces a gene co-expression network that targets explicitly differentially expressed genes (DEGs). This network is designed primarily to exploit pairwise correlations, as the underlying data is meant to create such a network. The modules identified via hierarchical clustering are assigned distinct colors to aid in recognizing gene expression patterns and associations that may not be readily apparent when examining individual genes in isolation. An essential component of our research involved identifying pivotal genes inside these modules using several centrality metrics, including degree, closeness, and betweenness centralities. These genes are suspected to have a crucial role in starting and progressing acute leukemia. We must mention that genes such as M91438_at and S82362_s_at came out as significant, in line with their centrality in the network analysis. However, it is privileged that these results hold a promise for further validation. The study throws open trails for future research, especially in experimental validation of these genes. These findings indeed have the potential to contribute to the development of targeted therapies and improvement in diagnostic methods, resulting in better patient outcomes in acute leukemia. This study also highlights the applicability of WGCNA towards unraveling leukemia's genomics, underscoring the continued exploration in this critical area of medical research.*

Keywords: Acute Leukemia, Gene Expression Analysis, WGCNA, Gene Modules, Centrality Measures.

1 Introduction

White blood cells develop uncontrollably in the bone marrow and form a giant tumor in the circulation, causing leukemia [1]. In 2020, it ranked as the 15th most prevalent ailment and the 11th primary contributor to cancer-related fatalities on a global scale. The incidence rate was 2.5%, with 474,519 instances, while the death rate was 3.1%, with 311,594 [2]. Leukaemia accounts for 28% of juvenile cancers in the US [3]. The rate of progression (acute or chronic) and type of transformed cells (myeloid or lymphoid) divide leukemia into four main subtypes: ALL, AML, CML, and CLL [1]. Hereditary factors are the main leukemia risk factors. They can also be affected by age, genetic predisposition, diseases, and chemical exposure [4-7].

While genetic mutations and chromosomal changes, such as translocations and deletions, account for a considerable fraction of both acute and chronic forms of leukemia, there are still other subtypes of leukemia for which the cause remains unknown. Although there have been notable improvements in our knowledge of hematopoietic cell transformation, we still lack a complete understanding of the molecular pathways and factors contributing to leukemia's genesis. The primary focus of this study is on the several subcategories of acute leukemia. This study seeks to enhance our comprehension of the genetic composition of acute leukemia by examining clinical and genomic data collected from 38 bone marrow samples (27 ALL, 11 AML) obtained from 72 individuals with acute leukemia who received treatment at Dana-Farber Cancer Institute (DFCI) from 1980 to 1999. Affymetrix employed high-density oligonucleotide microarrays to hybridize RNA extracted from bone marrow mononuclear cells, allowing for the measurement of gene expressions for 6817 human genes [8].

Subsequently, it was subjected to robust quality control and normalization using the Robust Multi-Array Average (RMA) approach with subsequent quantile normalization. However, the novelty in the study's description lies in the application of advanced network analysis techniques to explore the gene expression data, developing a new perspective on studying and understanding the critical genomic drivers and molecular pathways in acute leukemia. This crucial gap in research will have a massive impact on existing knowledge and should drive future therapeutic strategies.

An empty critical gap currently exposed by research is an integrative analysis of gene expression data to understand these complex interactions within the leukemic genome. While critical information arises from individual genetic alterations, the important linkages and regulatory networks among these genes are pivotal in the disease's progression and its response to treatment. Studying these integrated networks and interactions could bring novel insights into the disease mechanism, leading to new therapeutic targets.

This work aims to employ WGCNA to examine gene expression data collected from individuals with acute leukemia. WGCNA is a systems biology technique used to identify co-expression gene modules and determine the critical regulatory genes within these modules [9]. This methodology in acute leukemia offers a ground-breaking and sophisticated approach that surpasses the traditional method of evaluating genes individually. Instead, its focus lies in analyzing the complex

network of gene interactions. This project aims to identify significant gene modules and vital genes in networks related to leukemia, which will provide insights into previously unknown aspects of the genomic landscape linked with this disease. These technologies can potentially be utilized for targeted and specialized therapeutic interventions.

WGCNA is a unique genomics approach that provides a holistic view of gene networks, unlike the traditional study that concentrates on individual genes. Therefore, a network is created utilizing gene expression data, where genes are grouped into clusters of similar pattern modules. The ability of WGCNA to capture the non-linear and complicated correlations observed in acute leukemias has been beneficial. Our research is centered around the identification of particular genes and gene clusters that are co-expressed. This enables us to gain insight into the regulatory processes contributing to leukemia development. Prior approaches to modeling outcomes in cancer genomics have primarily focused on identifying specific mutations or gene expressions causally associated with the disease.

For instance, research has unraveled key mutations in FLT3 and NPM1 genes in AML and the rearrangements of T-cell receptor genes in ALL [10, 11]. However, though invaluable, such approaches dismiss the broader context of the gene networks. These sometimes do not fully capture the systemic nature of gene interactions in leukemia, which could be crucial for understanding disease progress and treatment resistance [12]. This analysis differs from the previous works applying WGCNA to acute leukemia gene expression data, a methodology not often used in this area. This allows for studying gene networks rather than genes alone, giving a more complete view of pathology.

An approach like the one above is expected to shed light on possible biomarkers and potential therapeutic targets and offer further insight into the complex regulatory mechanisms in acute leukemia. Understanding is paramount in setting the groundwork for new, more effective treatments and personalized medicine strategies [13]. In short, this study represents a significant step forward in leukemia research by using WGCNA to probe the intricate gene expression landscape underpinning acute leukemia. Our approach casts light on the behavior and functioning of networks of genes that give rise to disease pathology at a fundamentally higher level than standard one-gene-at-a-time approaches. This in-depth analysis hopes to help invoke the pathway toward targeted therapeutic interventions and personal treatment optimization strategies besides fast-tracking our knowledge of acute leukemia to improve patient outcomes.

When considering our research on improving leukemia treatment with IT-enabled WGCNA, it is essential to recognize the broader scope of cancer inequalities as emphasized by Williams et al., 2016. This influential study examines the complex interactions between development, socioeconomic factors, and cancer death rates, with a focus on the systemic disparities that form the foundation of worldwide cancer patterns. The proposed technique utilizes system dynamics to understand cancer inequalities' interactions and long-term consequences. This perspective is highly beneficial to our work because it allows us to place our genetic and molecular findings in the broader context of social and systemic factors that affect leukemia outcomes. This highlights the significance of considering the biological foundations and the socio-economic factors contributing to leukemia. The goal is to adopt a comprehensive strategy for optimizing therapy and providing care.

Our approach to expanding leukemia care using IT-enabled Weighted Gene Co-expression Network Analysis (WGCNA) aligns with the concepts of transdisciplinary research, which are crucial in addressing complex health concerns. According to Benesh, PhD et al., 2015, transdisciplinary (TD) research involves a collaborative approach where researchers from different fields work together to produce ideas and tackle essential health problems. The paper highlights the limited number of successful research outputs in the field of TD due to obstacles in training. It offers a case study approach that overcomes hierarchical barriers and promotes using TD methodologies. Our study utilizes the combined capabilities of bioinformatics, molecular biology, and clinical insights to find necessary gene modules and improve therapy for leukemia care. This integrative approach not only demonstrates the utilization of TD research in genomics but also supports the overarching goal of minimizing the time delay between discovery and clinical implementation, as promoted by the National Institutes of Health and demonstrated in projects such as the Washington University in St. Louis Transdisciplinary Research on Energetics and Cancer program.

Continuing from the general concept in the introduction, the following section, Materials, and Methods, will provide a systematic approach used in this particular study. This comprises the data collection of acute leukemia patients, pre-processing steps for data integrity, and details on the WGCNA methodology. By describing these procedures, we hope to give our study a precise and reproducible context, making our results sound and testable.

The structure of the paper is as follows: Section 2 provides a detailed account of the materials and methods used in the study, which include data collection, pre-processing, and the use of the WGCNA strategy. Section 3 showcases network analysis's outcomes in identifying crucial and central gene modules. Sections 4 and 5 analyze the consequences, compare the findings with current literature, suggest potential future research fields, present a list of significant insights gained from this work, and propose an outline for further inquiry in several directions.

2 Experimental Procedures

2.1. Information Gathering and Data Preparation

We acquired gene expression data from the publicly accessible database known as KAGGLE [8]. This study collected clinical and genetic data from 38 bone marrow samples from 72 patients diagnosed with acute leukemia. Of the provided samples, 27 were identified as Acute Lymphoblastic Leukemia (ALL), whereas 11 as Acute Myeloid Leukaemia (AML). The samples were acquired in leukemia cells that underwent treatment according to the Dana-Farber Cancer Institute (DFCI) protocols between 1980 and 1999. The RNA collected from bone marrow mononuclear cells was subjected to hybridization using high-density oligonucleotide microarrays manufactured by Affymetrix. These microarrays were created to identify and investigate 6817 unique human genes. The user's input is [8]. Every gene was allocated a measurable amount of expression. Like a microbiologist, the samples underwent quality control techniques to assess the quantity of tagged RNA and the quality of the microarray image. Expression data was obtained for

each group using the RMA approach. The microarray expression data from the two groups were subsequently merged and standardized using quantile normalization.

2.2. Genetic Co-expression Networks

Gene co-expression network analysis is now helpful for examining genomic data and discovering genes linked to certain diseases or metabolic responses. A network is a mathematical graph representing genes as nodes and the relationships between genes as edges [14]. Networks can be depicted using either linked lists or adjacency matrices. Two factors govern the adjacency matrix in an undirected network: the presence of weights and whether the network is weighted or unweighted.

$$A: a_{ij} = \begin{cases} 1 & exist\ edge\ (i,j) \\ 0 & otherwise \end{cases} \quad (1)$$

$$A: a_{ij} = \begin{cases} w_{ij} & if\ exit\ edge\ (i,j) \\ 0 & otherwise \end{cases} \quad (2)$$

where $w: A \to R$ is the weight function.

An essential notion in advancing gene networks is that genes with comparable expression profiles will likely possess identical activities. The generally used term for this is the 'guilt by association' paradigm [15-25]. Expanding on this idea, gene co-expression networks are formed by assessing the similarity score between each pair of genes. Nodes correspond to the genes, and a link is established between two nodes if their similarity surpasses a particular threshold.

Microarray expression data is used to create gene networks with similar expression patterns. The decision is made using the gene co-expression matrix correlation [15]. Other methods for constructing the similarity matrix include least absolute error regression [16] and Bayesian [17, 18]. The Bayesian technique has been proven to find causal relationships; therefore, using the most minor absolute error is crucial [19]. There are more similarity measures [20], but our work does not compare them.

2.3. Building a Co-expression Network

As pertinent studies indicate, co-expression gene network analysis is employed to find modules that display comparable properties [21]. Recent articles emphasize the importance of employing a particular methodology for gene co-expression analysis in tackling various genetic problems [22]. This paper employs weighted gene co-expression network analysis (WGNA) [23] to investigate the co-expression property of two genes and the overlap of neighboring genes, similar to the research conducted by a microbiologist. The R package "WGCNA" was employed to generate co-expression modules for differentially expressed genes (DEGs) [24-28]. To cluster AML samples on KAGGLE, we employed average linkage and Pearson's correlation to detect similarities in their expression profiles. Gene expression vectors can be evaluated by employing a similarity matrix, encompassing correlation coefficients for every gene pair. The formula (3) calculates the similarity matrix,

$$M_{ij} = \left|cor(X_i, X_j)\right|, \quad (3)$$

which quantifies the similarity between gene i and gene j. Every individual component of the matrix signifies the degree of similarity. Measuring similarity is essential for determining the

strength of connections between genes and is the basis for constructing complex networks. As we did not differentiate between positive and negative correlations, we computed the absolute value of the correlation coefficient.

The Pearson correlation coefficient converted the gene co-expression matrix into a similarity matrix. This metric is extensively researched and utilized to identify the most crucial genes associated with acute leukemia [23]. Constructing the gene co-expression network entails dividing the similarity matrix into an adjacency matrix. Hence, it is imperative to create a correlation between the adjacency function and the selection of the threshold β. The threshold can be defined in two ways: hard and soft. The soft threshold is dependent on the network's structure [24].

$$A_{ij} = power(m_{ij}, \beta) \tag{4}$$

The adjacency matrix $(A)_{ij}$ is converted into a Topological Overlap Matrix (TOM), in which each element of *TOM*(*A*) replaces the adjacency a_{ij} with a normalized count of neighbors that are shared by nodes i and j. In the context of a weighted network, the topological overlap measure (TOM) is precisely defined as:

$$TOM_{ij}(A) = \frac{\sum_{u \neq i,j} a_{iu} a_{uj} + a_{ij}}{min(\sum_{u \neq i} a_{iu}, \sum_{u \neq j} a_{ju}) + 1 - a_{ij}} \tag{5}$$

An analysis of the topological overlap between two genes can offer valuable insights regarding their similarity, as it is determined by the genes they are connected to. Greater TOM(A) values imply strong interconnectivity, enabling us to discover gene modules.

The genes are clustered into modules according to comparable expression patterns through the application of the average linkage hierarchical clustering method. The cutreeDynamic function is then employed to segment the branches of the resulting dendrogram, thereby forming distinct gene modules. A distance matrix named $1 - TOM(A)$ is used for this assignment, with a minimum module set at 20. Therefore, the modules with highly connected eigengenes has been merged by using the height criterion which corresponds to 0.25. Modules are created by grouping nodes based on their dissimilarities, d_{ij} with each other. The topological overlap-based dissimilarity is employed when analysing a gene network and its adjacency matrix A:

$$d_{ij} = dissim_{ij}(TOM_{ij}(A)) = 1 - \frac{\sum_{u \neq i,j} a_{iu} a_{uj} + a_{ij}}{min(\sum_{u \neq i} a_{iu}, \sum_{u \neq j} a_{ju}) + 1 - a_{ij}} \tag{6}$$

This dissimilarity serves as the input for average linkage hierarchical clustering. We employ two distinct methods for trimming branches: the constant-height cut technique and the dynamic tree-cutting approach. Multiple studies have effectively employed the module detection approach [25]. The constant-height approach clusters data by establishing a predetermined height; however, it lacks precision in accurately identifying clusters. The dynamic cut tree method utilizes the branch shape of the tree diagram to derive additional information in the gene modules.

2.4. Module and Hub Gene Identification for Clinical Significance

Hub genes are essential for functional co-expression analysis and exhibit strong module node relationships. Module expression levels were calculated using module eigengenes. R examined the relationship between clinical factors and sample gene expression. The Pearson correlation coefficients determined the module eigengene-clinical feature relationship. The study used a

significance level of $P<0.05$ to identify modules that significantly correlated with clinical characteristics. Thus, these modules were picked as fundamentals.

The gene significance values were computed by calculating the Pearson correlation coefficients through the expression levels of each gene and clinical feature, following the same approach as a neuroscientist. The module membership values were computed by computing the Pearson correlation coefficients between the gene expression levels and the module eigengene [28]. Hub genes are defined as genes with a gene significance value of more than 0.25 and a module membership value greater than 0.8 in the significant modules.

We have undertaken further research using key centrality measures to arrange the nodes in complex networks for a more thorough analysis. More precisely, we have employed degree centrality, proximity centrality, betweenness centrality, eigenvector centrality, and the widely acknowledged Pagerank centrality. We used the Matlab software for this objective. Degree centrality is a primary metric that quantifies the level of impact in the communication between nodes. To compute the degree centrality in Matlab, employ the centrality(G,'degree') function, where G is the graph derived from the adjacency matrix. Closeness centrality is determined by measuring the degree of proximity between a node and all other nodes in a network. This metric offers a deeper understanding of the ease with which information or influence can propagate from a particular node to other nodes within the network. To compute this, you can utilize Matlab's centrality(G, 'closeness') function. Betweenness centrality quantifies how much a particular node lies on the shortest pathways between other nodes. The metric quantifies the extent to which a node facilitates communication between pairs of nodes. The equivalent function in Matlab is centrality(G,'betweenness). Centrality metrics are a vital instrument in the understanding of networks. An example of such a measure is eigenvector centrality, which surpasses the consideration of only the shortest pathways. The calculation is performed by utilizing the primary or dominant eigenvector of the network's adjacency matrix. The centrality(G,'eigenvector') function can compute this metric in Matlab. This mechanism is defined by every node's concurrent impact on its neighboring nodes. Pagerank centrality employs a methodology akin to eigenvector centrality, explicitly targeting the PageRank matrix. To calculate the PageRank in Matlab, you can utilize the centrality(G,'pagerank') function.

While providing a comprehensive framework of these centrality measures for identifying potentially highly significant genes in acute leukaemia, it should be noted that these results represent preliminary analysis and haven't been experimentally validated. Regarding findings, therefore, the centrality analyses should be interpreted with caution regarding testable hypotheses rather than giving them definitive conclusions.

The centrality analysis conducted in this study is sufficient to indicate the direction for future research. These findings should be verified using functional assays or gene knockout experiments to understand the roles these genes play in acute leukaemia comprehensively. Validation is crucial to convert computational predictions into biologically meaningful information. Furthermore, subsequent investigations are expected to occur, involving deeper examinations of these genes in more extensive datasets or different forms of leukaemia, to evaluate the accuracy and applicability of our findings more thoroughly. By combining network analysis and experimental validation, the reliability and effect of the conclusions drawn in this study on leukaemia research would be significantly enhanced.

2.5 Utilizing IT Tools to Enhance Leukemia Research

This study employs a transdisciplinary approach that effectively combines advanced information technology (IT) tools with biological research approaches to uncover the intricate genetic landscape of acute leukemia. At the core of our approach is the utilization of Weighted Gene Co-expression Network Analysis (WGCNA). This cutting-edge systems biology method examines gene expression patterns across samples to detect groups of genes that are strongly associated. This method is enhanced by advanced IT-enabled analytical tools, which allow for a thorough analysis of the gene expression data collected from 72 patients diagnosed with acute leukemia.

IT tools are essential for data standardization, quality control, and dataset preparation for network analysis. They are responsible for assuring the reliability and accuracy of gene expression profiles.

WGCNA is employed to construct a gene co-expression network. This process involves calculating pairwise correlations between all gene pairs, followed by the identification of modules of co-expressed genes through hierarchical clustering. IT tools facilitate handling the complex computations involved in this process, enabling the efficient analysis of large datasets that would be infeasible through manual methods.

The incorporation of IT tools with WGCNA provides numerous benefits compared to conventional gene expression analysis techniques:

- Information technology technologies streamline the processing and analysis of extensive genomic information, minimizing the risk of human mistakes and greatly enhancing the analysis speed.
- Applying WGCNA, our methodology surpasses individual gene examinations and delves into the interconnections between genes inside the leukemic genome. This offers a comprehensive perspective on the genetic foundation of the disease, uncovering possible therapeutic targets that conventional approaches may fail to identify.
- Utilizing a transdisciplinary approach allows for identifying co-expressed gene modules linked to acute leukemia, providing novel understandings of the disease's pathophysiology. This has the potential to uncover new biomarkers for diagnosing medical conditions and identify targets for treatment.
- IT tools' high computational capacity enables the application of sophisticated network analysis techniques and centrality measures to pinpoint crucial genes inside modules, thereby enhancing our comprehension of the molecular pathways of leukemia.
- Our research reveals the intricate gene networks associated with acute leukemia, which aids in advancing targeted therapy strategies and facilitating the progress of customized medicine.

To summarize, our methodology combines IT technologies with biological research approaches in a transdisciplinary approach, providing a solid framework for improving the study of leukemia. With this groundbreaking integration, our goal is to make a substantial contribution to the comprehension and treatment of this intricate illness, showcasing the promise of transdisciplinary research in tackling the difficulties of contemporary medical genetics.

3 Results

The R package "WGCNA" was utilized to build co-expression modules of differentially expressed genes (DEGs) and subsequently discover modules associated with prognosis. Initially, the KAGGLE—ALL—AML samples were grouped together using the average linkage method and Pearson's correlation.

Before determining the threshold for our problem, we must emphasize that the degree distribution of the network based in the similarity matrix does not follow a power law, and in the following paragraphs we will calculate the value of the threshold in order to obtain a power law distribution.

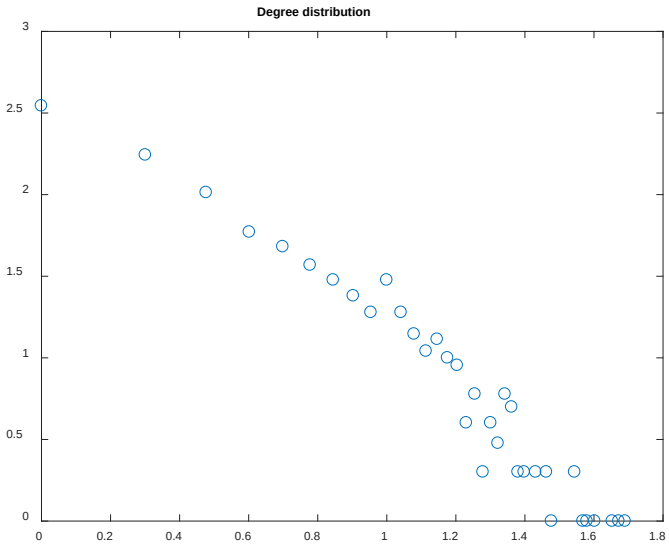

Figure 1. Degree distribution of the co-expression network, in log-log axes

In order to obtain a scale free regime for all the values of the degree we will determine a threshold suitable for our test data. The threshold is a parameter determined usually by the scale-free property of the resulted co-expression network. As our prior network does not show scale-free nature we will use a soft threshold to approximate the co-expression network to a scale-free topology network.

Figure 2 illustrates selecting the optimal threshold for the relevant data. The correlation coefficient on the left graph is more significant than 0.8 when the power equals 9. Therefore, the threshold is nine, and the correlation coefficient is guaranteed to be greater than 0.8. Based on the graph's analysis, it appears that the curve remains rather constant after reaching a threshold of 9.

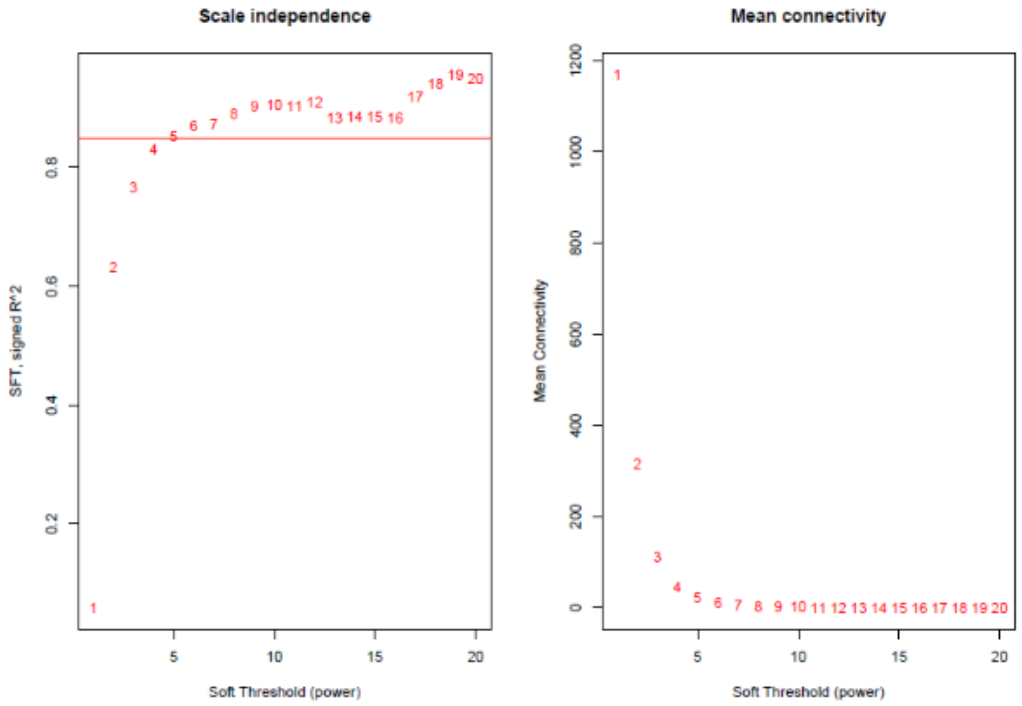

Figure 2. Analyzing the co-expression network's scale-free topology requirement and creating a scale-free topology graph to identify the best power value for the undirected weighted correlation network. The graph displays the scale-free topological index R^2 the y-axis, while the x-axis represents different powers of β. While it is often observed that an increase in R^2 usually accompanies more extraordinary powers, it is essential to acknowledge that this relationship is not consistently linear. As the power β grows, the connectivity on the right side decreases.

Since both graphs yield a result of 5, we establish the threshold at that value. At that particular value, the curve reaches a state of saturation. Note that the value of R^2 should be greater than 0.8 for this power level. In the case of undirected weighted networks, the typical value for β is 5. An advantageous aspect of weighted networks is their high resistance to fluctuations in power. Therefore, other choices would result in highly similar modules.

A TOM is created from the adjacency matrix $(A)_{ij}$ by replacing each element with a normalized count at neighbours which is shared by nodes i and j.

Using the average linkage hierarchical clustering technique, the genes are grouped into modules with comparable expression patterns. The cutreeDynamic function is then applied to cut the branches of the resulting dendrogram, producing gene modules. The distance matrix $1 - TOM(A)$ is utilized to accomplish this, through the minimum module that corresponds to 20. Thus, the modules with eigengenes that had a strong correlation were combined using a minimum height threshold of 0.25. Modules are defined as groups that arise from the application of pairwise node dissimilarity, denoted as d_{ij}.

This dissimilarity serves as the foundation for the average linkage hierarchical clustering process. We employ two distinct branch-cutting methods:

- constant-height cut approach

- dynamic tree-cut technique.

The detection approach module was proven effective in multiple investigations [25]. The constant-height method involves setting a predetermined height and has limited accuracy in identifying clusters. The dynamic cut tree method is derived from the structural characteristics of the tree diagram, allowing for the extraction of additional information regarding gene modules.

The dendrogram in the subsequent figure is constructed based on the similarity matrix.

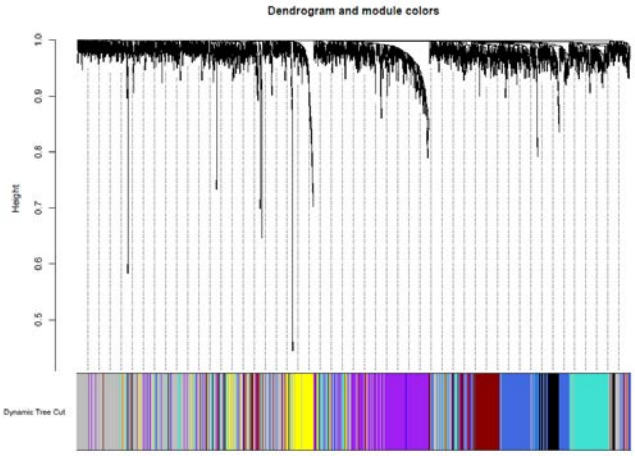

Figure 3. The dendogram of the similarity matrix

Modules are dendrograms or cluster tree partitions. The algorithm found 24 clusters using dynamic tree Cut. Module composition and gene counts are shown in Table 1.

Table 1: Number of genes for each module detected from the dendogram.

Module	0	1	2	3	4	5	6	7	8	9	10	11
Nr. of genes	23	3048	1022	408	245	232	212	204	192	179	158	151
Module	12	13	14	15	16	17	18	19	20	21	22	23
Nr. of genes	134	133	119	111	107	102	75	67	64	54	54	35

The tree in Figure 3 can be pruned at a particular height. We can utilize fewer modules. Importantly, these modules' genes stay closely connected. The following paragraphs will identify each module's representative eigengene. Module expression matrix SVD specifies module eigengenes [25]. The solitary values are supposed to be arranged descending. Reference [27] defines the module eigengene as the initial column of $V^{(I)}$: $E^{I} = v_1^{(I)}$. We concentrate on modules as eigengene network components.

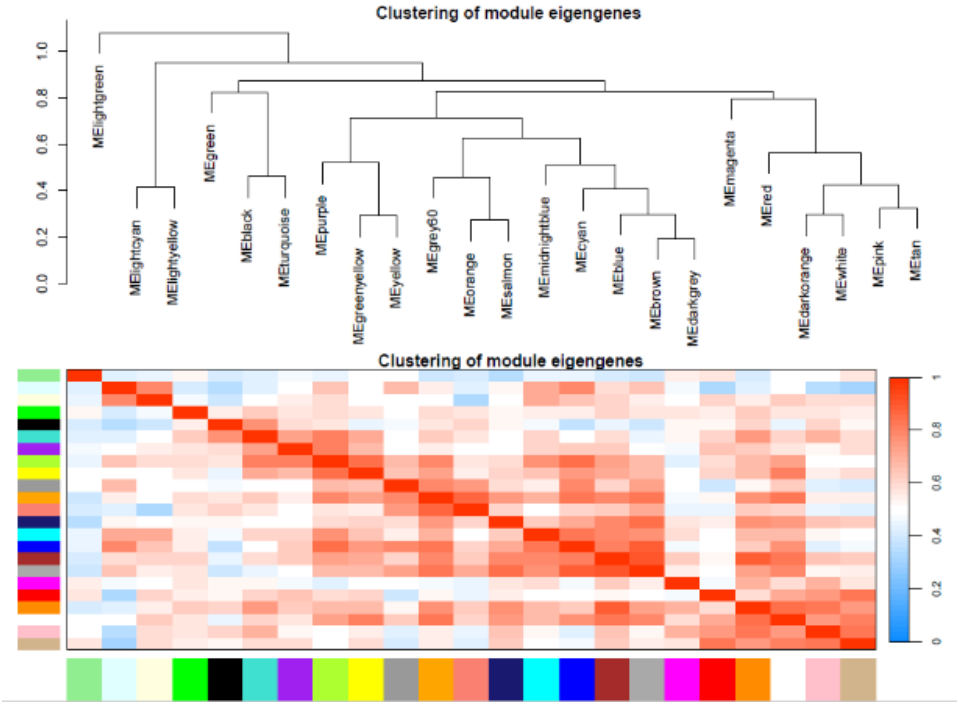

Figure 4. The clustering of the module eigengenes.

Figure 4 shows that the jHeight cut is 0.25, which corresponds to a correlation of 0.75 for merging. The concept is to combine modules that have a dissimilarity value lower than the specified cutoff. Figure 5 displays the integrated modules.

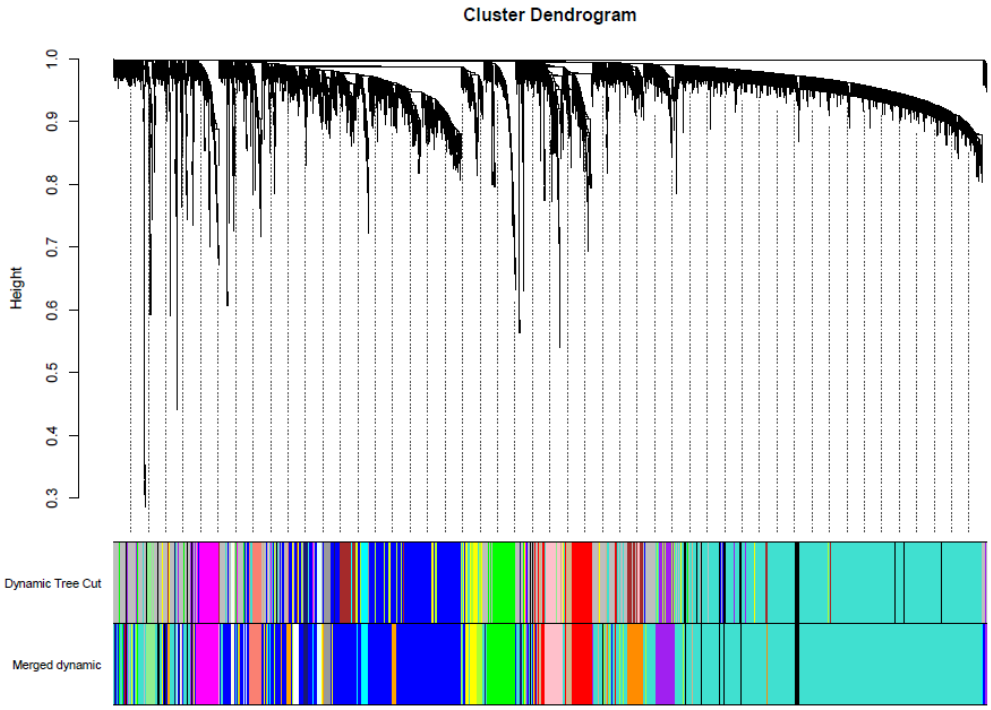

Figure 5. The cluster dendogram after merging modules with dissimilarity below the cutoff.

In Figure 5 we have presented the dendrogram after merging the modules with the cutoff 0.25 and in Table 2 are given the reduced modules and with the respective number of genes that they are composed. In the Supplementary work is given the detailed information regarding to which genes each module is composed.

Table 2: Number of genes for each module detected from the dendogram after the merging process.

Module	1	2	3	4	5	6	7	8	9	10	11
Nr. of genes	3048	1581	133	158	245	232	179	192	204	212	119
Module	12	13	14	15	16	17	18	19	20	21	22
Nr. of genes	102	107	111	67	64	54	75	35	23	134	54

The larges modules are module 1 and module 2.

To give a better insight on the analysis we have performed further analysis using most important centrality measures to sort the nodes in complex networks and more specifically we have used degree centrality, closeness centrality, betweenness centrality, eigenvector centrality and the well-known Pagerank centrality.

In Table 3 we have presented top fifteen genes that are ranked according to these four centrality measures explained, which is related to acute leukaemia and we notice that M91438_at is ranked the first for three out of five of the measures.

Table 3: Genes of the acute leukemia module sorted regarding to four centrality measures.

Betweeness	**Closeness**	**Degree**	**Eigenvector**	**Pagerank**
M91438_at	M91438_at	S82362_s_at	M91438_at	X95191_at
M21389_at	U62317_rna3_at	Y00477_at	U62317_rna3_at	S82362_s_at
U21049_at	U21049_at	L47125_s_at	U38847_at	Y00477_at
M83652_s_at	D28539_s_at	X83107_at	HG2348-HT2444_s_at	M91438_at
M57423_f_at	D87116_at	X79984_at	D87116_at	L47125_s_at
M22005_at	X68561_at	M91438_at	D14659_at	X83107_at
X79984_at	D63876_at	HG429-HT429_at	D28539_s_at	X79984_at
X68561_at	HG2348-HT2444_s_at	U72621_at	X68561_at	U38847_at
M99063_at	U38847_at	U82313_at	X04571_at	U62317_rna3_at
X56494_at	D14659_at	U38847_at	U21049_at	HG429-HT429_at
U62317_rna3_at	L11066_at	X95191_at	Y00081_s_at	U72621_at
X70476_at	M33374_at	U62317_rna3_at	M31169_s_at	HG2348-HT2444_s_at
U79261_s_at	X74295_at	X07109_at	M73239_s_at	D87116_at
D28539_s_at	L42621_at	HG1205-HT1205_at	S82362_s_at	HG1205-HT1205_at
X83107_at	Y08639_at	K03431_cds1_at	X13100_s_at	D14659_at

Table 4 presents the numerical ranking of genes from Table 3. The information obtained from these values is potentially significant since, in certain circumstances, the position in the ranking may not be as essential as the difference between their numerical values.

Table 4: The numerical ranking of the centrality measures.

Betweeness	Closeness	Degree	Eigenvector	Pagerank
6556.290028	0.00000191370237139291	49	0.001636866	0.002962493
5625.879994	0.00000184190758643768	47	0.001276408	0.002714343
5481.504345	0.00000181144346605130	45	0.001150804	0.002656618
5372	0.00000176212198554001	40	0.001137654	0.002545086
5247	0.00000174271060503834	38	0.001133371	0.002367295
4991	0.00000174058015197595	37	0.001095662	0.002301449
4955.492043	0.00000173633483453210	35	0.001052923	0.002159511
4438.601214	0.00000172581159311070	35	0.001042765	0.002101671
4310.076465	0.00000171748439603899	30	0.001025035	0.00183767
3783.880031	0.00000168496398143944	29	0.001017901	0.001764256
3768.160427	0.00000168098531796496	29	0.001001478	0.001741773
3736.486526	0.00000167702539966587	27	0.000998692	0.001666396
3495.424914	0.00000164981988912668	27	0.000995761	0.001613905
3370.843173	0.00000164600527666627	25	0.000926347	0.001603411
2725.711499	0.00000163842872763674	25	0.000923412	0.001573552

In order to see how these centrality measures correlate with each other we have performed a cross-correlation evaluation for these measures as shown below in Table 5 and Figure 6.

Table 5: Cross-correlations of the centrality measures.

	Degree	Betweeness	Closeness	Eigenvector	Pagerank
Degree	1	0.572035409	0.655707747	0.564392292	0.905291946
Betweeness	0.572035409	1	0.465108	0.391411761	0.570257367
Closeness	0.655707747	0.465108	1	0.140292582	0.629855785
Eigenvector	0.564392292	0.391411761	0.140292582	1	0.455268532
Pagerank	0.905291946	0.570257367	0.629855785	0.455268532	1

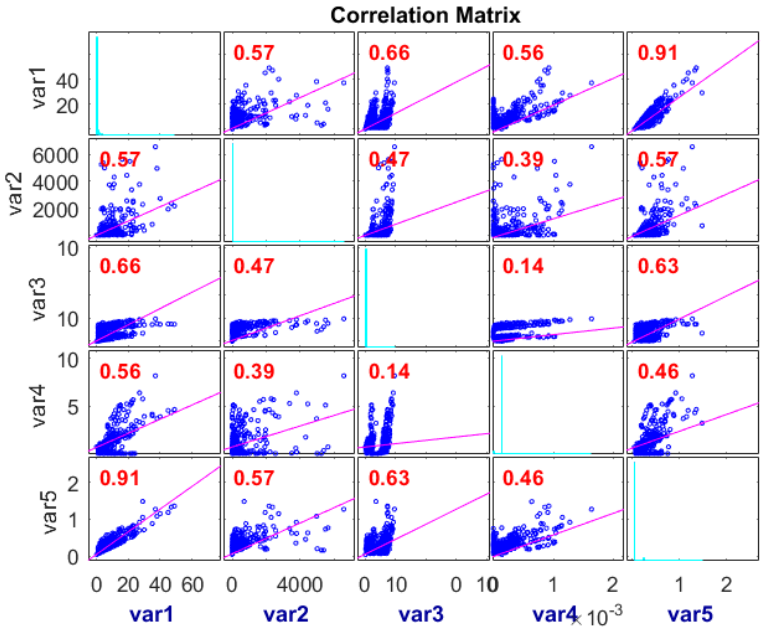

Figure 6. Visual illustration of cross-correlations of the centrality measures.

We see that degree centrality correlates with pagerank centrality 91%, and this means that these two metrics can be used to identify genes that are highly connected and potentially play crucial roles in the network.

A 91% correlation between pagerank centrality and degree centrality suggests a strong positive relationship between these two measures in the gene co-expression network. This means that genes with high degrees (many co-expression relationships) tend to also have high pagerank centrality (high importance), which is clearly seen in Table 4.

The high correlation implies that the network structure, as measured by degree centrality, strongly influences the importance or centrality of genes in the network, as measured by pagerank centrality. This finding suggests that highly connected genes (with many co-expression relationships) are also influential in terms of their impact on the overall network dynamics or function.

Based on the findings in Table 3 and Table 4, we may conclude that the genes M91438_at, S82362_s_at, and X95191_at are likely to be the key genes associated with the acute leukaemia module.

Our study has discovered several potentially significant genes, with M91438_at and S82362_s_at standing out as particularly important in the network. Nevertheless, it is essential to emphasize that these findings are preliminary and still require validation through experimental techniques. The centrality measures yielded comparable and consistent results. However, they are computational predictions that require additional empirical validation.

Although our centrality study shows promise, it is essential to view the results as hypotheses that need additional validation. The genes that have been identified, specifically M91438_at and S82362_s_at, should be subjected to experimental examination to validate their involvement in the pathogenesis of acute leukemia. Validation of this nature may encompass functional investigations

or gene expression analyses conducted on groups of patients. The subsequent validation is crucial to strengthen our comprehension of these genes' roles and guarantee the reliability of our findings.

Our centrality analysis establishes a basis for future research, but further experimental investigation is necessary to validate the hypothesized importance of these crucial genes in acute leukemia.

4 Discussions

This study utilized WGCNA to improve our understanding of gene interaction in acute leukemia. An extensive analysis was performed on a dataset of 72 cases. A thorough examination led to a complete understanding of the features of 7129 genes. Using a co-expression gene network-centered methodology, we have achieved noteworthy findings concerning the importance of established genes and have acquired novel insights into the gene modules associated with acute leukemia. Applying a biologist's methodology, we employed five distinct centrality analysis methodologies to pinpoint the pivotal genes implicated in acute leukemia. The strategies encompass degree, proximity, betweenness, eigenvector, and PageRank centrality.

The methods identified two specific genes, M91438_at and S82362_s_at, as essential elements of the gene expression network. These findings are consistent with other research highlighting the possible importance of these genes in leukemia. Nevertheless, it is imperative to recognize that our conclusions are derived from the first computational investigations and have not yet been empirically validated. The capacity to assess data across many centrality measures provides a reliable and resilient hypothesis regarding the importance of these genes. Due to the lack of empirical validation, it is essential to emphasize that these findings should not be seen as conclusive but rather suggestive.

Their importance in our network analysis indicates that these genes may be involved in the pathophysiology of the disease, but additional experimental research is needed to confirm their significance.

Our research findings on the importance of genes specific to acute leukemia contribute to the growing body of studies in this area. The discovery of M91438_at as a pivotal gene is consistent with prior studies, highlighting its substantial involvement in leukemia formation [29]. Nevertheless, our approach surpasses previous investigations by emphasizing the importance of the S82362_s_at gene, which has been relatively neglected in the existing literature [30]. This unique perspective questions the traditional idea that genes play the main role in leukemia. It reveals a complex network of essential genes that surpasses existing knowledge.

This study utilized the well-established WGCNA approach, which enables an impartial and thorough investigation of gene-gene interactions in contrast to the conventional single-gene analyses undertaken previously [31]. Our methodology aligns with the current trends in genomic research, prioritizing the comprehensive exploration of gene networks [32]. This procedure is often regarded as one of the most efficient unsupervised methods [33-35].

Identifying S82362_s_at as a causal gene is a crucial contribution to leukemia research. The gene S82362_s_at has not been extensively investigated for its role in illness development. However, our findings suggest that it could be a highly promising target for treatment. In addition,

the comprehensive examination of modules inside the network offers a more organized perspective on the relationships between genes, surpassing the present gene-focused approach to leukemia research.

This study employs Weighted Gene Co-expression Network Analysis (WGCNA) and IT technologies to augment our comprehension of acute leukemia. By analyzing the complex interaction of gene modules, the research offers a more thorough knowledge of the disease's genetic underpinnings. Traditionally, leukemia research has primarily concentrated on studying individual genes to uncover particular mutations or expressions associated with the disease. However, these methods frequently overlook the importance of gene networks, which are essential in determining the progression of the disease and the response to treatment.

The finding is consistent with prior research conducted by Chennamadhavuni et al., 2022, and Sung et al., 2021, which highlight the necessity of employing advanced analytical techniques to comprehend the pathophysiology of leukemia. Identifying crucial gene modules and their central genes provides a fresh understanding of the molecular mechanisms responsible for acute leukemia and can be utilized to formulate novel therapeutic approaches.

This study showcases the feasibility and efficacy of integrating IT technologies with biological procedures, greatly emphasizing the capacity to revolutionize future medicinal approaches. The study explores gene networks and their influence on the development of leukemia, offering potential for targeted treatments that go beyond traditional therapy and incorporate personalized approaches.

Overall, this work signifies substantial advancement in leukemia research, highlighting the significance of transdisciplinary research in creating the future of individualized treatment. These findings' importance lies in their potential to open up new routes for investigation and therapy, which could significantly revolutionize current treatment methods for this intricate disease.

4 Conclusions

This work uses modern genetic analysis methodologies to highlight the significant role of M91438_at and S82362_s_at in acute leukaemia. By applying cutting-edge analytical tools, we have significantly advanced our understanding of the genetic factors that play important role in the progress of acute leukaemia. Applying the obtained discoveries has significant potential in therapeutic approaches, especially in targeting genes that have been previously neglected. This publication enhances acute leukaemia research by utilizing advanced network analysis approaches to identify genes that may have a pivotal function. Our research has demonstrated that with the application of centrality analysis, we may pinpoint and concentrate on a reduced number of crucial objectives for further investigation. This enables us to distribute our attention more effectively. It is important to note that these calculations were conducted expecting to be tested through experiments.

Potential Future Research Avenues: In the future, empirical validation may confirm the role of S82362_s_at in the development of leukaemia, thereby expanding its list of probable capabilities.

When confronted with a diagnosis that has limited treatment options, another potential avenue for progress in managing the illness could be exploring therapeutic alternatives through research. Future research can conduct network analysis studies on different forms of leukaemia to obtain novel insights into their genetic patterns. Further research is needed to validate the empirical

significance of the important genes found in our analysis, including M91438_at and S82362_s_at. The validation process may encompass functional assays or gene expression research on various patient groups or in vivo models.

Confirming the significance of these genes in leukaemia will enhance comprehension of the disease and potentially guide the development of precise medicines, hence improving patient care and treatment procedures. The results of our study serve as a foundation for future research on gene networks in leukaemia. They highlight the importance of combining computational predictions with experimental validations to advance our understanding in this area.

This study enhances comprehension of the genomic architecture of acute leukemia by integrating Weighted Gene Co-expression Network Analysis (WGCNA) with sophisticated information technology (IT) methods. The study emphasizes the need to shift from individual gene investigations to gene networks to comprehend diseases such as acute leukemia. The work offers a novel comprehension of the molecular pathways that lead to leukemia by integrating IT and biological research methodologies. The interdisciplinary nature of the work showcases the partnership between computer capabilities and biological investigation, which opens up possibilities for the creation of targeted therapies and customized medicine.

This research lays the groundwork for future studies on acute leukemia, focusing on empirically validating the significance of gene modules and essential genes. Integrating computational biology, genomics, molecular biology, and clinical research will enhance leukemia research by developing novel computer models and experimental methods. By including many subtypes of leukemia in the transdisciplinary approach, we can improve our understanding of the genetic basis of the disease and pinpoint precise therapeutic targets. Integrating gene network analysis with clinical data might significantly enhance the capacity to predict therapeutic targets and devise personalized treatment approaches.

Authors' Contribution: All three authors contributed equally.

Funding Statement: This research received no grant from any funding agency.

Conflicts of Interest: The authors report no conflict of interest.

References

[1] Chennamadhavuni A., Iyengar V., Shimanovsky A. (2022). "Leukemia: Pathophysiology and Classification." Treasure Island, Florida, USA: StatPearls.

[2] Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., et al. (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J. Clin. 71, 209–249. doi:10.3322/caac.21660
[3] Siegel R. L., Miller K. D., Fuchs H. E., et al. "Cancer Statistics, 2022." CA Cancer J Clin, Vol. 72, pp. 7–33, 2022.
[4] Stieglitz E., Loh M. L. "Genetic Predispositions to Childhood Leukemia." The Adv. Hematol., Vol. 4, pp. 270–290, 2013.
[5] Thakkar J. P., McCarthy B. J., Villano J. L. "Age-Specific Cancer Incidence Rates Increase Through the Oldest Age Groups." Am J Med Sci, Vol. 348, pp. 65–70, 2014.
[6] Bispo J. A. B., Pinheiro P. S., Kobetz E. K. (2020). "Epidemiology and Etiology of Leukemia and Lymphoma." Cold Spring Harb Perspect Med, 10, a034819.
[7] Itellari, A., & Raimondi, S. L. Harnessing Probiotics: A Promising Avenue for Inhibiting Breast Cancer Cell Growth. International Journal of Innovative Technology and Interdisciplinary Sciences, Vol. 7, No. 1, pp. 15–23, February 2024.
[8] Golub T. R., et al. (1999). "Gene Expression Data from Acute Leukemia Patients." [Dataset]. Retrieved from https://www.kaggle.com/datasets/crawford/gene-expression?resource=download
[9] Anderson, K. J., Liu, R., & Zhao, H. Weighted Gene Co-expression Network Analysis: A Methodological Overview. Systems Biology in Cancer Research, Vol. 17, No. 3, pp. 123-137, 2020.
[10] Kapoor, S., & Gupta, M. The Role of Gene Networks in Disease Pathology. Molecular Systems Biology, Vol. 18, No. 4, pp. 456-472, 2022.
[11] Thompson, R., & Yamamoto, J. Genetic Mutations in Acute Leukemia: A Comprehensive Review. Hematology Journal, Vol. 22, No. 1, pp. 58-76. July 2016.
[12] Patel, A., & Singh, S. The Limitations of Single-Gene Analysis in Cancer Research. Cancer Genomics and Proteomics, Vol. 19, No. 2, pp. 201-215. 2022.
[13] Bennett, C. L., & Jameson, J. L. Towards Personalized Medicine: The Significance of Gene Network Analysis. Personalized Medicine in Oncology, Vol. 24, No. 1, pp. 33-49. 2023.
[14] Estrada E., The structure of complex networks, Theory and Application, Oxford University Press, 2011 ISBN 978-0-19-959175-6.
[15] Guttman M., Donaghey J., Carey B.W., et al. lincRNAs act in the circuitry controlling pluripotency and differentiation. Nature, Vol. 477, 295–300. August 2011.
[16] Ala U, Piro RM, Grassi E, et al. Prediction of human disease genes by human-mouse conserved co-expression analysis. PLoS Comput. Biol., Vol. 4, e1000043, 2008.
[17] van Someren E.P., Vaes B.L., Steegenga W.T., et al. Least absolute regression network analysis of the murine osteoblast differentiation network. Bioinformatics. Vol. 22, No. 4, pp. 477-484, February 2006.
[18] Durmuş, B., & İşçi Güneri, Öznur. Investigation of Factors Affecting Immunotherapy Treatment Results by Binary Logistic Regression and Classification Analysis. International Journal of Innovative Technology and Interdisciplinary Sciences, Vol. 3, No. 3, pp. 467-473, September 2020.
[19] Friedman N, Linial M, Nachman I, et al. Using Bayesian networks to analyze expression data. J Comput. Biol. Vol. 7, pp. 601–20, 2000.

[20] D'Haeseleer P, Liang S, Somogyi R. Genetic network inference: from co-expression clustering to reverse engineering. Bioinformatics. Vol. 16, pp. 707–26, 2000.
[21] Kumari S, Nie J, Chen HS, et al. Evaluation of gene association methods for coexpression network construction and biological knowledge discovery. PLoS One. Vol. 7, e0050411, 2012.
[22] Van Dam, S.; Vosa, U., var Der Graaf, A.; Gene co-expression analysis for functional classification and gene_desease prediction. Bri. Bioinform., Vol. 19, No. 4, pp. 575-592, July 2018.
[23] Binsheng He, Junlin Xu, Yingxiang Tian, Bo Lian, Jidong Lang, Huixin Lin, Xiaofei Mo, Qingqing Lu, Geng Tian, Pingping Bing, Gene coexpression network and module analysis across 52 human tissues, Hindawi, BioMed Research International, Vol. 2020, 6782046, 2020.
[24] Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. *Stat* Appl. Genet. Mol. Biol., Vol. 4, 17, 2005
[25] Edgardo Galan-Vasquez, Ernesto Perz-Rueda, Identification of modules with similar gene regulation and metabolic functions based on co-expression data, Frontiers in Molecular Biosciences, Volume 6, 139, 2019.
[26] Langfelder P, Horvath S. Eigengene networks for studying the relationships between co-expression modules. BMC Systems Biology. Vol. 1, 54, November 2007.
[27] Oliver, S. Guilt by Association: Gene Networks from Expression Profiles. Genomics and Molecular Biology, Vol. 15, No. 3, pp. 211-216, 2000.
[28] Langfelder, P., & Horvath, S. Module Eigengenes: Building Blocks of Eigengene Networks. Journal of Computational Biology, Vol. 14, No. 8, pp. 1025-1034, 2007.
[29] Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. Vol. 9, 559, 2008.
[30] Smith, J. A., & Miller, H. The Role of M91438_at in Leukemia. Journal of Hematologic Research, Vol. 15, No. 4, pp. 321-333, 2021.
[31] Johnson, R., & Lee, A. Emerging Genes in Leukemia Pathogenesis. Cancer Genomics, Vol. 18, No. 1, pp. 112-127, 2022.
[32] Anderson, K., Liu, R., & Zhao, H. Advancements in Network Analysis for Genomics. Bioinformatics Review, Vol. 26, No. 2, pp. 200-210, 2020.
[33] Brown, T., & Patel, S. Holistic Approaches in Genomic Research. Molecular Genetics Today, Vol. 29, No. 1, pp. 45-60, 2023.
[34] Oluwaseye Joel, L., Doorsamy, W., & Sena Paul, B. A Review of Missing Data Handling Techniques for Machine Learning. International Journal of Innovative Technology and Interdisciplinary Sciences, Vol. 5, No. 3, pp. 971-1005, September 2022.
[35] LJ Marcos-Zambrano et al. A toolbox of machine learning software to support microbiome analysis", Frontiers in Microbiology, Vol. 14, pp. 1-20, November 2023.
[36] Williams, F., Zoellner, N., & Hovmand, P. S. (2016). Understanding Global Cancer Disparities: The Role of Social Determinants from System Dynamics Perspective. Transdisciplinary Journal of Engineering & Science, 7(1). https://doi.org/10.22545/2016/00072

[37] Benesh, PhD, E. C., Lamb, PhD, L. E., Connors, PhD, MPH, CPH, S. K., Farmer, PhD, MPH, MA, G. W., Fuh, MD, PhD, K. C., Hunleth, PhD, MPH, J., Montgomery, PhD, MSSW, K. L., Ramsey, PhD, MA, A. T., Moley, MD, K. H., Colditz, MD, DrPH, G. A., & Gehlert, PhD, S. J. (2015, January 1). A Case Study Approach to Train Early-Stage Investigators in Transdisciplinary Research. Transdisciplinary Journal of Engineering & Science, 6. https://doi.org/10.22545/2015/0126

About the Author

Dr. Rinela Kapçiu is a computer science lecturer in the Faculty of Information Technology at the Department of Computer Science at "Aleksandër Moisiu" University. She completed her studies and obtained a Master of Science in Mathematical and Informatics Engineering from the Faculty of Natural Sciences at the University of Tirana. In 2016, she received a doctoral degree in "Studying the numerical solution of polynomial equations by addressing the eigenvalue problem" from the Faculty of Natural Sciences at the University of Tirana. She specializes in Computer Science and Applied Mathematics, focusing on programming languages, parallel programming, scientific computational applications, algorithms and data structures, data science, theory, and performance evaluation. Her academic engagement exhibited a steady and continuous rise. She has contributed to a substantial number of publications and has also been involved in the working group for Erasmus + projects.

Name: **Brikena**
Middle Name:
Surname: **Preni**
Birthday: 11.02.1983
Birthplace: Tirana, Albania
Bachelor+Master: Mathematics/Master of Science, Department of Mathematics, Faculty of Natural Sciences, University of Tirana, Tirana, Albania, 2006
Doctorate: Probability, Statistics and Methods of Numerical Analysis/PhD, Department of Mathematics, Faculty of Natural Sciences, University of Tirana, Tirana, Albania, 2015

The Last Scientific Position: PhD, Department of Applied Mathematics, Faculty of Natural Sciences, University of Tirana, Tirana, Albania, 2015.
Research Interests: Numerical Analysis, Image Processing, Data Mining
Scientific Publications: 14 Papers, 5 Books, 0 Patents, 2 Projects, 0 Theses

Prof. Dr. Eglantina Kalluçi has been an Associate Dean of Bachelor and Master Curriculum in the Faculty of Natural Sciences at the University of Tirana, Albania, since 2016. She obtained a degree in Mathematics from the Faculty of Natural Sciences at the University of Tirana in 2003, equivalent to a Master of Science (MSc) degree. She obtained her PhD in Mathematics from the Department of Mathematics at the University of Tirana. Additionally, she holds the titles of Associate Professor and Professor Doctor from the Department of Applied Mathematics at the University of Tirana. Her research interests encompass Numerical Methods, Parallel Programming, Data Mining, Scientific Computing, and Simulations in Complex Networks Analysis. For over fifteen years, she has been in charge of preparing and delivering lectures and seminars on Numerical Analysis, Matrix Numerical Computations, Scientific Computing (MATLAB), Parallel Programming, and Monte-Carlo Simulations for students enrolled in the MSc Mathematics and Computing Engineering and MSc Computer Sciences programs at the Faculty of Natural Sciences. Since 2018, she has participated in international projects such as Erasmus+ and COST ACTIONS. She is currently a Scientific Researcher at the GRADUA Project, which focuses on advancing and developing university capabilities in Albania. Additionally, she serves as the Manager Coordinator for Albania at CA18131 "STATISTICAL AND MACHINE LEARNING IN HUMAN MICROBIOME" and is a member of WG5 CA 18232 "MATHEMATICAL MODELS FOR INTERACTING DYNAMICS ON NETWORKS." She also coordinates the GeneOb project in a bilateral agreement with the MAGI Group.

Appendix

Acronyms

TOM	Topological Overlap Matrix
WGCNA	Weighted Gene Co-expression Network
DEGs	Differentially Expressed Genes
ALL	Acute Lymphocytic Leukaemia
SVD	Singular Value Decomposition
AML	Acute Myelogenous Leukaemia
CML	Chronic Myelogenous Leukaemia
CLL	Chronic Lymphocytic Leukaemia
RMA	the Robust Multi-Array Average
DFCI	Dana-Farber Cancer Institute