Online ISSN: 2682-2628
Print ISSN: 2682-261X

IJCBR

International Journal of Cancer and Biomedical Research

https://jcbr.journals.ekb.eg

Using network analysis and machine learning to identify genes implicated in spinal muscular atrophy

Islam M. Nofal, Elsayed E. Hafez, Amira Y. Haikal and Mostafa A. Elhosseini

PUBLISHED BY
EACR EGYPTIAN ASSOCIAN FOR CANCER RESEARCH
Since 2014

INTERNATIONAL JOURNAL OF CANCER AND BIOMEDICAL RESEARCH (IJCBR)

RESEARCH ARTICLE

Using network analysis and machine learning to identify genes implicated in spinal muscular atrophy

Islam M. Nofal[1], Elsayed E. Hafez[2], Amira Y. Haikal[3] and Mostafa A. Elhosseini[3,4]

[1] Egyptian Armed Forces
[2] City of Scientific Research and Technological Applications New Borg Al Arab City, 21934, Alexandria, Egypt
[3] Computers and Control Systems Engineering Department, Faculty of Engineering, Mansoura University, Egypt
[4]College of Computer Science and Engineering, Taibah University, Yanbu 46421, Saudi Arabia

ABSTRACT

Background: Spinal Muscular Atrophy (SMA) is a genetic disease that causes the loss of a survival motor neuron (SMN), leading to vital muscle atrophy. **Aim:** Despite numerous studies to find a cure for this disease, the best of these treatments is still suffering from some limitations and difficulties. It was found that treatments that focus on just one gene are not usually effective. Consequently, the current study investigates gene impacts and interactions by gathering an appropriate microarray dataset for various human SMA instances. In addition, embryonic stem cell samples, which are anticipated to play a significant role in the future treatment of the majority of incurable diseases. **Materials and Methods:** By using linear models for microarray data analysis (LIMMA), highly differentially expressed genes (DEG) were identified. Then, cluster these genes into modules using machine learning and weighted gene co-expression network analysis (WGCNA) algorithms. **Results:** By using the preservation methods, the foundation of interesting modules was evaluated between the collected cases. Moreover, the results of previous studies on SMN1, SMN2, NAIP, DYNC1H1, and PLS3 genes have proved that they are direct causes or modifiers of SMA disease severity. However, the change in the expression of these genes did not come at the forefront of the changed genes, which is the exact opposite of what is expected. Accordingly, other interesting modules were determined here as highly correlated modules with these genes. These modules' genes were imported into Cytoscape for generating SMA networks, and finding their hub genes. **Conclusion:** These genes can be used as key genes for better analysis, diagnosis, and therapy development, such as BCL2, Cntn1, TYRP1, N4Bp2, and PFDN2.

Keywords: Survival Motor Neuron; co-expression network; Key genes; LIMMA; Microarray

Editor-in-Chief: Prof. M.L. Salem, PhD - Article DOI: 10.21608/JCBR.2022.173808.1283

ARTICLE INFO

Article history
Received: November 09, 2022
Revised: December 12, 2022
Accepted: December 13, 2022

Correspondence to
Islam M. Nofal
Egyptian Armed Forces
Tel.: 01066857535
Email: zeiad85eslam@gmail.com

INTRODUCTION

A neurological disease, SMA runs in families as an autosomal recessive trait. The disease is characterized by the loss of muscle mass and the weakness of muscles that control movements such as crawling, walking, sitting, and moving the head. Furthermore, in severe cases of the disease, the breathing and swallowing muscles may be affected, resulting in the patient's death within a short period after suffering from the disease symptoms. These muscle problems cause progressive degeneration of specialized nerve cells called Survival motor neurons (SMN). This neuron can be found in the spinal cord and brain stem (Chaytow et al., 2018). It is estimated that 1/10,000 live births are affected by SMA globally, which makes it one of the most common hereditary diseases causing childhood deaths (Anderson et al., 2003). At the same time, SMA carrier frequency varies per racial group from 1 in 38 to 1 in 72. Across all levels of severity, the disease is on the rise in countries where consanguineous marriage is on the rise (Chaytow et al., 2018) (Anderson et al., 2003).

SMA results from a deficiency of an essential protein for keeping the survival motor neurons called SMN protein, which is generated by two very homologous genes. Telomeric SMN1 and its centromeric homology SMN2. These genes

are part of a 500 kbp inverted duplication on chromosome 5q13 with at least four genes. Therefore, they are susceptible to deletions and rearrangements. The SMN1 gene is the dominant gene for the production of the SMN protein. Hence, there must be at least one functional copy to avoid the disease. Otherwise, SMN protein derived from the gene SMN2 is the only source of protein for the affected individual in this case (Genetics Home Reference, 2017) (SMA Care series, 2009). Notably, these two genes differ only by five nucleotide changes, and the difference at codon 280 is the most important. In addition, the c-to-T substitution at position 6 of exon 7 in SMN2 affects the splicing of SMN2. This splicing change yields only around 10% of the full-length protein per each SMN2 copy, and the remaining is an unstable and truncated protein (Genetics Home Reference, 2017). Accordingly, the severity degree of SMA disease inversely correlates with the SMN2 copy number. Therefore, SMN2 may be an important target for SMA therapy due to its role as a significant SMA disease modifier (Ludolph et al., 2018).

The severity of SMA disease is influenced by other genes besides SMN2, including zinc finger protein 1 (Zpr1), plastin 3 (Pls3), Dynein Cytoplasmic 1 Heavy Chain 1 (DYNC1H1), and Ubiquitin Like Modifier Activating Enzyme 1 (UBA1) (Shawky et al., 2001) (Anderson et al., 2003). In addition, some reports have demonstrated a correlation between SMA severity and neuronal apoptosis inhibitory protein gene (NAIP gene) deletion. However, the functional role of NAIP in the pathogenesis of SMA has not been fully elucidated (Anderson et al., 2003) (Hassan et al., 2020) (Shawky et al., 2001) (Shawky & El-Sayed, 2011). Not the previous genes only. But also, some other modifier genes may result in less common types of SMA, such as X-linked SMA. This appears when mutations occur in the UBA1 gene, which participates in protein degradation within cell. Consequently, reducing this amount leads to protein build-up inside the cell and damages motor neurons (U.S. National Library of Medicine, 2017). SMA-LED is another type; Mutations in the DYNC1H1 gene caused the cell to lose the Dynein protein that transports cellular components from the junctions between neurons (synapses) to the center. The signals are transmitted from one neuron to another through this mechanism. As a result, DYNC1H1 mutations decrease and prevent neuron signal transfer with time, resulting in missed control over muscle movement (Wirth et al., 2020).

SMA patients are divided into five phenotype sub-types of SMA (0-IV). In addition, the character traits of the patient may identify the degree of muscular weakness and the age at which the muscle difficulties first appear (U.S. National Library of Medicine, 2017). An overview of possible diagnostic pathways for a person with symptoms similar to SMA disease is shown in Figure 1. Moreover, it illustrates the characteristics of the five types of illness, in which the modifier genes are the leading player in determining what degree of disease we are (Prior et al., 2019).

Figure 1 illustrates the role of the SMN1 gene in differentiating SMA patients and the effect of modifiers genes on determining the severity of the disease and the characteristics of each grade of SMA. Stem cells have the capacity for self-renewal and cell-type differentiation. Stem cells support a variety of tissues, including the nervous system, heart, skeletal muscles, meniscus, tendons, ligaments, and labrum. To target certain tissues or organs, stem cell administration in combination with other therapeutic agents has the ability to improve, alter, or start local or systemic healing processes. Therefore, stem cells can be employed to convey chemicals created artificially for therapeutic purposes (Xiaowen, 2020).

Research Gap: Numerous studies have been conducted to find a cure for this disease, but even the best treatments still have limitations and difficulties. Genes typically cooperate rather than work alone; therefore, therapies that target one gene exclusively are ineffective in some cases (Chen & Tai-Heng, 2020). In addition, all genes associated with SMA disease and its severity have been linked directly or indirectly by previous studies. Therefore, there seems to be a significant correlation between these genes and certain SMA modules.

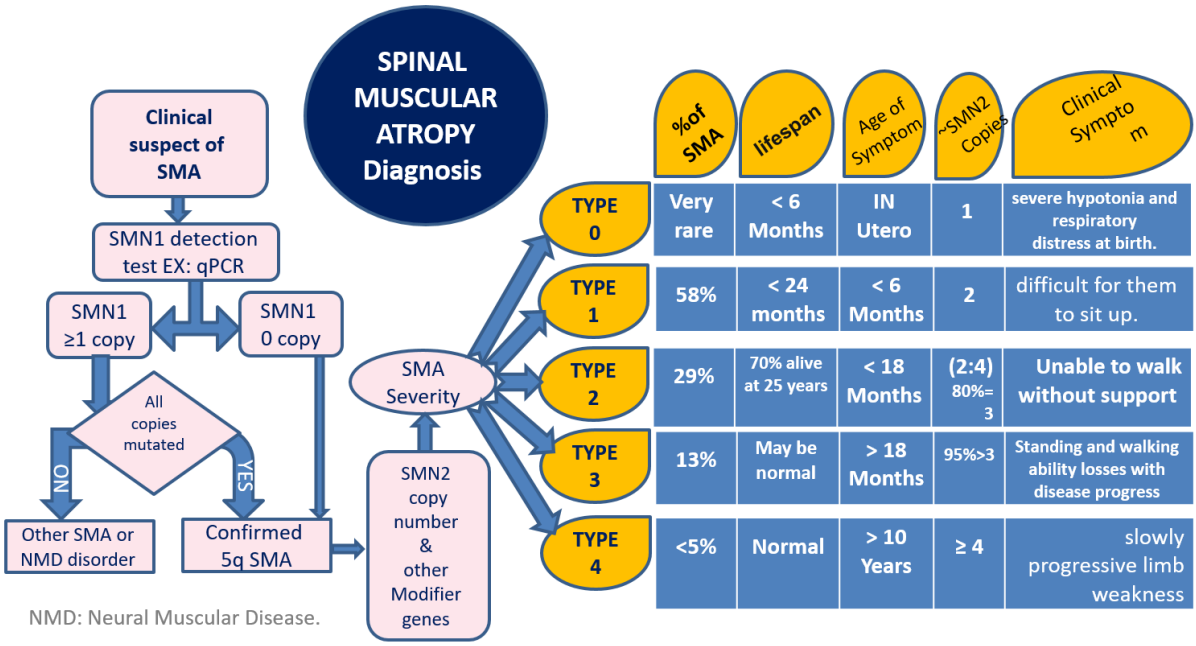


Figure 1. SMA diagnosis methods and their types

Undoubtedly, the hub genes of these modules will contribute to solving the mystery of this disease, diagnosing it, or developing treatments for it. Moreover, further investigation should be made into the role of stem cells in this disease and its treatments.

SMN-dependent and SMN-independent therapies are the two treatment approaches available for SMA (Mendell et al., 2017) (Corti et al., 2018). SMN-independent strategies aim to strengthen and maintain motor neurons, neural muscular junctions, and affected muscles to reduce disease symptoms. While, in the SMN-dependent approach, the focus is on increasing the amount of SMN protein as the cause of the problem, either by editing the deleted SMN1 gene or replacing the mutated one (Zolgensma). Furthermore, by treating SMN2 gene splicing issues (Nusinersen). On the whole, the purpose of this is to provide the body with enough SMN protein, a multifunctional and universally expressed protein (Ludolph et al., 2018) (Chen & Tai-Heng, 2020) (Mendell et al., 2017) (Corti et al., 2018). Although all previous treatments already have Food and Drug Administration (FDA) approval, and other continuous attempts have significantly impacted the severity of the disease, it still suffers from some limitations. Such as the very high costs, the danger and pain of repeated intrathecal injections in the back chain, and treatment under a certain age and weight, not above it. Not only the previous therapeutic strategies but also stem cells have an important share of studies on genetic diseases. However, numerous studies show that primary neural stem cells injected into the spinal canal engraft the spinal cord, enhance motor function, and prolong survival in patients with SMA. Unfortunately, until now, all the studies about stem cells being used to treat SMA and other genetic diseases are just experimental because this type of therapy has still not been thoroughly tested in clinical trials (YANG et al., 2019).

Identifying and characterizing genes that result in SMA disease is considered a very important requirement for the successful development of new SMA therapies. Moreover, understanding the disease mechanism. Bioinformatics methods are considered the best and fastest to get this essential required information (Vamshi K. Rao et al., 2018) (Hensel et al., 2020). The data used in the current study was gutted from microarray experiments. Microarray technology significantly monitors the gene expression levels for each gene on the genome-scale under different conditions. All of these measured values are recorded in a matrix known as the Gene Expression matrix. Experience. Theoretically, the weighted network is the most suitable way to describe the relation and interaction between these genes

together. But computationally, the analysis of a network is often restricted to a limited number of nodes (e.g., 2000 nodes) with top differential expressed genes (Zhang & Horvath, 2004). Various DEG tools are available, including the Empirical Analysis of Digital Gene Expression Data (Edge) R package, DESeq, DESeq2, baySeq, SAMSeq, and limma (Linear models for microarray data analysis).

Due to its stability, LIMMA (https://limma.html) was preferred in the current study, even for experiments with a few arrays, complex experiments, or varying conditions. Furthermore, it can be applied to any quantitative gene expression technology such as microarrays, RNA-sequencing, or quantitative PCR (Ritchie et al., 2015).

Despite the weighted network construction eliminating the information loss found in an unweighted network (Goh et al., 2007), biological significance is essential in biological networks (Goh et al., 2007). The network must be compatible with the scale-free topology feature to accomplish this. The number of hub genes within the network is far greater than the number of non-hub genes, according to its content. Furthermore, model fitting index R2 of the linear model, which correlates the frequency distribution of the connectivity p(k) to connectivity (k) itself, can be used to visually inspect whether approximate scale-free topology is satisfied or not (Zhang & Horvath, 2004). Our current research used the WGCNA R package (https://cran.r/WGCNA/html) to achieve what we wanted in our designed network and divide the network into highly correlated modules using machine learning. Moreover, it interfaced with the ***Cytoscape*** program to visualize and identify the modules' hub genes.

The clustering method is one of the unsupervised machine learning methods used for categorizing data into sets (Eisen et al., 1998). Consequently, the current data is organized into sets of samples and sets of genes with shared patterns that are representative of the group (Modules). Objects can be grouped into hierarchical clusters with relationships between them specified, much like a phylogenetic tree. Furthermore, they can be grouped into non-hierarchical clusters (arranging items into clusters without defining the links between them). For instance, in the current work, each object (sample/ gene) is regarded as a cluster by a hierarchical agglomerative clustering algorithm (Kapp & Tibshirani, 2007). Significantly, calculating pairwise distance estimates for the items that will be grouped is the first stage. Then, clusters are created from related items based on the pairwise distances between them. Following this, the pairwise distances between the clusters are once again determined. Further, related clusters are iteratively joined until all the items are included in a single group. Indeed, a dendrogram may be used to visualize this information, as shown in Figure 3 and Figure 7. Accordingly, the distance from the branch point is a reference for separating two groups or items (Zhang & Horvath, 2004) (Eisen et al., 1998).

This research paper contributes to the enrichment of stem cell studies for SMA treatment by using embryonic stem cell data and correlating effective SMA genes with the modules of different SMA cases. The method of module preservation evolution was used here to determine the hub genes of various cases of SMA. Using these genes as key genes may contribute to the development of better diagnostic procedures and therapies.

MATERIALS AND METHODS

It has been reported that Polymer Chain Reactions (PCRs) have been used in several SMA studies to correlate phenotype with genotype (Shawky et al., 2001) (Hassan et al., 2020) (Shawky & El-Sayed, 2011). Still, the PCR method does not consider the interactions between all genes. In contrast, other studies have used gene expression technologies such as RNA-sequencing or microarrays to study gene interactions. This is more accurate and precise than studying the effects of a single gene alone on disease diagnosis (Hensel et al., 2020).

The current study uses a publicly available microarray dataset for SMA disease (Carriers and several severity degrees) (https://www.ncbi.nlm.nih.gov/geo/) to get the variation in gene expression levels, which might reflect the pathogenic process of illness.

Moreover, to support stem cell research in SMA disease analysis, we gathered several stem cells samples that were deposited on the same platform as our SMA data (GPL6947) for this publication. These samples, however, were taken from tests that were conducted in various settings. In accordance with reasonable exceptions that permit the completion of the study, abnormal samples were removed and samples with circumstances that were comparable to those of the samples taken for the aforementioned disease were preserved. accordingly, as shown in Figures 3 and 4, by using Hierarchical sample clustering, principal component analyses (PCA), and Box plots to detect outlier samples. This may ensure a proper proportion of the compatibility of the experimental conditions of the collected samples. An overview of the current work steps is shown in Figure 2, starting with the preprocessed data and moving on to its aim.

In order to construct a gene-weighted network, which is widely regarded as the most effective method for describing the interaction between genes, there are a number of steps that must first be completed.

These steps involve meeting certain topological network criteria in order to account for biological criteria. Following these steps, shared and unshared modules between different cases are identified for the purpose of further study using machine learning and preservation techniques.

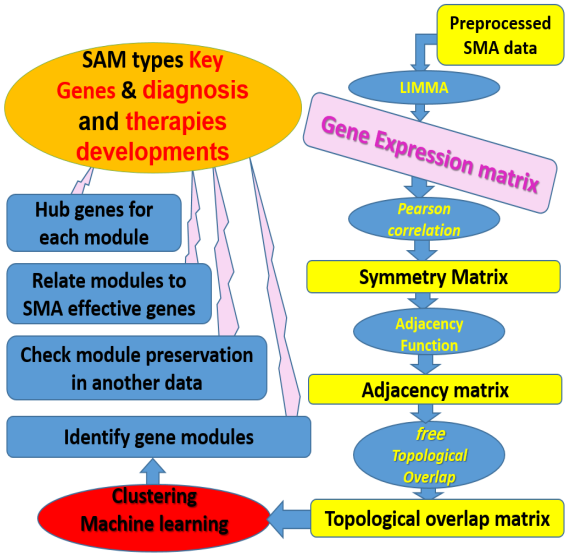


Figure 2. Graphical abstract

Using the package 'limma' that was built under free R version 4.0.3, a significance test is used to identify highly differentially expressed genes (DEGs) between test and control samples (YANG et al., 2019). The LIMMA package provides stable analysis even for experiments with a few arrays and complex experiments with various conditions. Furthermore, it applies to data obtained via different quantitative gene expression technologies, including microarray, RNA-sequencing, and quantitative PCR (Zhang & Horvath, 2004) (Eisen et al., 1998). The output genes of the LIMMA package would be collected in a gene expression matrix for further analysis.

Understanding how genes interact is the first step to understanding a complex gene interaction system. Based on intuitive network concepts (e.g., modules and connectivity), networks are considered the most effective method for analyzing complex interactions between nodes. There are many domains of networks covered by these nodes. Some of them can be applied to the biological field. Such as protein-protein interaction networks (Ritchie et al., 2015), cell-cell interaction networks (Goh et al., 2007), and gene co-expression networks (Eisen et al., 1998) (Yang et al., 2016). Moreover, they can be applied to personal accounts on social networks, Internet sites, and other fields.

For constructing the current study network between the gutted DEGs, a similarity matrix (Sij) between gene expression profiles was calculated using a correlation method to measure their degree of concordance. Indeed, ***Pearson*** correlation coefficients are generally utilized as co-expression measures in linear relationships to establish the relevance of gene co-expression networks (Albert & Réka, 2005). Then, the similarity matrix is transformed into the adjacency matrix (aij). It does describe the strength of the link between each pair of genes. In reality, this transformation is accomplished by threshold the coefficients of the similarity matrix with a suitable adjacency function. Significantly, this function depends on the type of the constructed network, unweighted or weighted. For the first network, the used adjacency function, such as the signum function, provides a hard threshold, leading to the connection strength between any two

points having a discrete value (0 or 1). Not only this type of network makes the value of the threshold very sensitive. But this also leads to a significant amount of information loss in the upcoming analyses. The disadvantage of the previous type was eliminated in the weighted network by using a soft threshold provided by a different kind of adjacency function. For example, the sigmoid function and the power adjacency function, both of which were used in the current study with only a single parameter (β), as shown in Eq. (1) (Zhang & Horvath, 2004).

a_{ij} = power ($S_{ij,}$ β) ≡ | S_{ij} |$^{\beta}$. Eq. (1)

The adjacency function parameters a (i,j) are chosen to use mean connectivity criteria (mean(k)= $\sum_1^n$ ki/n). Biological networks were created using topological criteria that combined biological significance with statistical significance. It takes into consideration the influence of neighboring genes on each gene (topological overlap matrix (TOM)) Eq. (2) (Zhang & Horvath, 2004) (Albert & Réka, 2005).

$$TOM(i,j) = \frac{|N1(i) \cap N1(j)| + aij}{\min(|N1(i)| \& |N1(j)|) + 1 - aij} \quad \text{Eq. (2)}$$

Scale-free topology refers to the frequency distribution of the connection $P(k)$. Significantly, it implies hub nodes are connected to many other nodes. The goodness of fit of linear model fitting R^2 index, which measures the goodness of fit for linear model fitting $p(k)$, is tested using scatter plots, log transformation, and k Scale Free Topology criteria for network development (Yang et al., 2016) (Albert & Réka, 2005). In addition, several biological issues reported that intramodular connectivity correlates with gene importance more strongly than the whole network (Horvath S. , 2011), thus adding validity and confidence to the network-based approach to identifying diseases' molecular signatures (Chen & Tai-Heng, 2020). Notably, it was found that the node dissimilarity method (1-TOM) was an effective distance measurement for biologically meaningful modules. Moreover, it was combined with a hierarchical clustering method and a dynamic tree-cutting algorithm to detect gene subsets that had strong relationships (modules) (Stuart, 2003) (Carter, 2004). Significantly, it is noteworthy that the corresponding modules' eigengene (E) can be correlated to each other. As the first component of a specific module, it can indicate the gene expression patterns of that module (Jeong, 2001) (Hartwell, 1999). Therefore, dynamic tree-cutting algorithms may identify modules whose expression profiles are semi-identical (Stuart, 2003). Thus, the final modules were obtained by merging the modules whose gene expressions are strongly correlated (Merging Modules). The current study used the package 'WGCNA' that was built under R version 4.0.4 for network construction, topological properties calculations, and module detection.

Furthermore, it can be interfaced with external software such as Cytoscape for network visualization and identifying the hub genes of all gutted modules. Often, it is found that the relationship between R2 and the threshold parameter (β) of the adjacency function follows approximately a saturation curve. Therefore, based on the scale-free topology criterion, the initial threshold parameter value at which saturation is achieved should be used as long as it is higher than (0.8). Alternatively, the default threshold parameters for unsigned and signed correlation networks are ($\boldsymbol{\beta} = \mathbf{6}$) and ($\boldsymbol{\beta}$= 12), respectively (Jeong, 2001).

Not only are preservation methods used to evaluate the clustering efficiency, but also to find the previous gutted modules in other SMA types, carriers, or stem cells. These methods can discover interesting modules. As shown in Eq. (3,4), and Table 1, Composite preservation statistics ($Z_{summary}$), which can be used to quickly evaluate several modules across many networks with different preservation scales based on specific thresholds, can be applied to this task (Eisen et al., 1998) (Kapp & Tibshirani, 2007) (Horvath & Steve., 2011).

$$Z = \frac{Z\ connectivity\ +\ Z\ density}{2} \quad \text{Eq. (3)}$$

$$Z_{summary} = \frac{observed - mean\ permuted}{sd\ permuted} \quad \text{Eq. (4)}$$

The $\boldsymbol{Z_{summary}}$, however, may not be acceptable when comparing modules with extreme differences in size since it largely depends on module size (Horvath & Steve., 2011). Accordingly, the Median Rank preservation statistics (Eq. 5) will be more appropriate.

Table 1. Composite preservation (Zsummary) statistics threshold guidelines

$Z_{summary}$>10	Strong evidence supports the module's preservation.
2<$Z_{summary}$<10	Weak to moderate preservation evidence is present.
$Z_{summary}$ < 2	There is no proof the module was preserved.

$$Median\ \ Rank = \frac{Median\ Rank\ .\ density + median\ Rank\ .connectivity}{2} \quad \text{Eq. (5)}$$

Furthermore, since it uses observed preservation statistics instead of Z statistics or p −values, it is less dependent on module size. Compared to a module with a higher median rank, one with a lower median rank usually shows superior observed preservation statistics. These statistics make it possible to detect which modules are highly preserved in most cases. It also helps identify the similarities and differences between these cases to understand the disease mechanism better, saving time, money, and effort (Horvath S. , 2011) (Shengni et al., 2020) (Horvath et al., 2007). SMA disease severity and its likelihood of occurrence can also be determined by other critical modules that are highly correlated with highly effective genes such as SMN1, SMN2, NAIP, Pls3, and DYNC1H1 (Anderson et al., 2003) (Shawky et al., 2001). According to previous studies, these genes are the causative and modifier genes. However, they did not appear in front of differentially expressed genes between different SMA cases. Therefore, the current study's primary objective was to examine these genes' mutational effects on the network modules' genes with the highest correlations to these genes.

Here, Cytoscape program v 3.8.0 was used for recreating molecular networks and detecting hub genes. Gene lists produced by highly connected hub gene selection methods are known to be more meaningful than those produced by standard statistical analysis (Jeong, 2001) (Zhang& Horvath., 2005) (Horvath et al., 2007). Therefore, another objective of this study is to use the hub genes in the selected modules as novel genes for studying SMA pathogenesis or as key genes for improving patients, carriers, prenatal diagnosis, and therapy development.

Data collection and preprocessing

In the present study, SMA datasets were collected from the (GEO) databases (https://www.ncbi.nlm.nih.gov/geo/), which are deposited on the (GPL6947) Illumina HumanHT-12 V 3.0 expression bead chip. A total of 12 samples of GEO (GSE58316) were recruited from 12 people without any treatment. SMA severity degrees are categorized as follows: (5: SMA Carrier), (2: Severe SMA), and (5: Mildly SMA). In addition, for purposes of subsequent study, all samples were considered processed and hybridized under identical circumstances.

SMA treatment with stem cells and SMN restoration may be complementary. Consequently, understanding how stem cells work in SMA treatment is a great goal. In spite of the fact that there are a significant number of preclinical research involving stem cells for the treatment of SMA, we were almost unsuccessful in our search for samples from these studies. The author searches for stem cell samples that have been put on an Illumina HumanHT-12 V3.0 expression beads chip using the identical processing procedures as before (GPL6947). In light of the aforementioned, our interest in stem cell research as a potential treatment for SMA led to our willingness to contribute to the expansion of stem cell research for SMA treatment. For example, six samples of embryonic stem cells were obtained from separate investigations (GSE29784 (4 samples), GSE35029 (1 sample), and GSE31845 (1 sample), and they were subsequently placed on the Illumina HumanHT-12 V3, version (GPL6947).

Moreover, anomalies were detected by hierarchical clustering of machine learning science for all 18 samples with vision cutting at different heights. To improve the hypothesis, the same hybridization and processing conditions are applied to all samples. Accordingly, 16 samples were identified, as shown in Figure 3. In addition to hierarchical clustering, principal component analysis and box plots were used here to verify the validity of the assumption. Also, both methods remove only the same previous two samples, as shown in Figure 4.

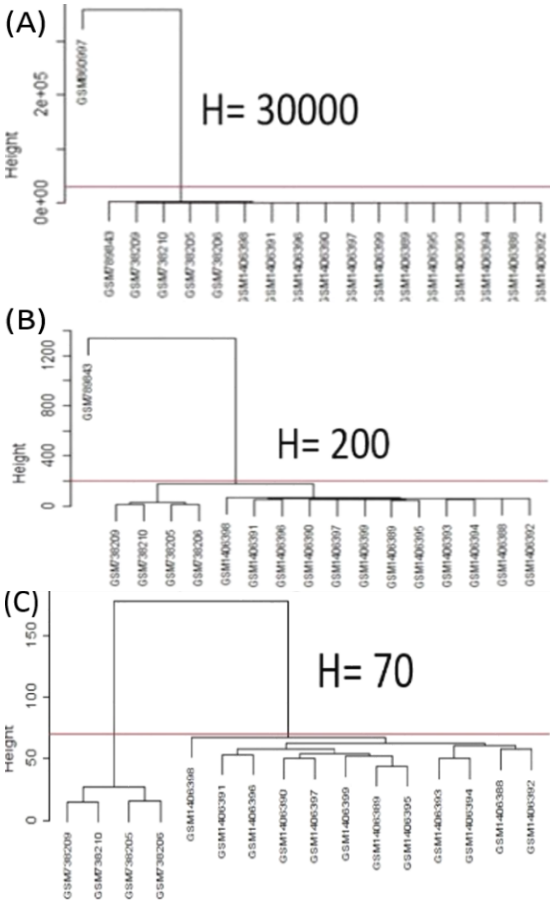


Figure 3. Hierarchical samples clustering to detect outlier samples. A) 1 sample is outer at cutting high=30000 and 17 samples are remaining, B) 1 sample is outer at cutting high = 200 and 16 samples are remaining, C) up to cutting high = 70 no outer samples and 16 samples are remaining.

Furthermore, ensure that the 16 samples have the same hybridization and processing conditions for future research.

COMPUTER RESULTS AND SIMULATION

According to Figures 3 and 4, two anomaly samples with a height of 70 were removed. Outlier samples were determined using principal component analyses and box plots for all samples. Outliers were then removed from the same two samples. As a result, the remaining 16 samples consist of five carriers, two severe carriers, four mild carriers, and four stem cell carriers.

Therefore, it can be assumed that the hybridization and processing conditions for all 16 samples are adequate for further analysis within acceptable limits. Furthermore, these 16 samples were preprocessed as follows.

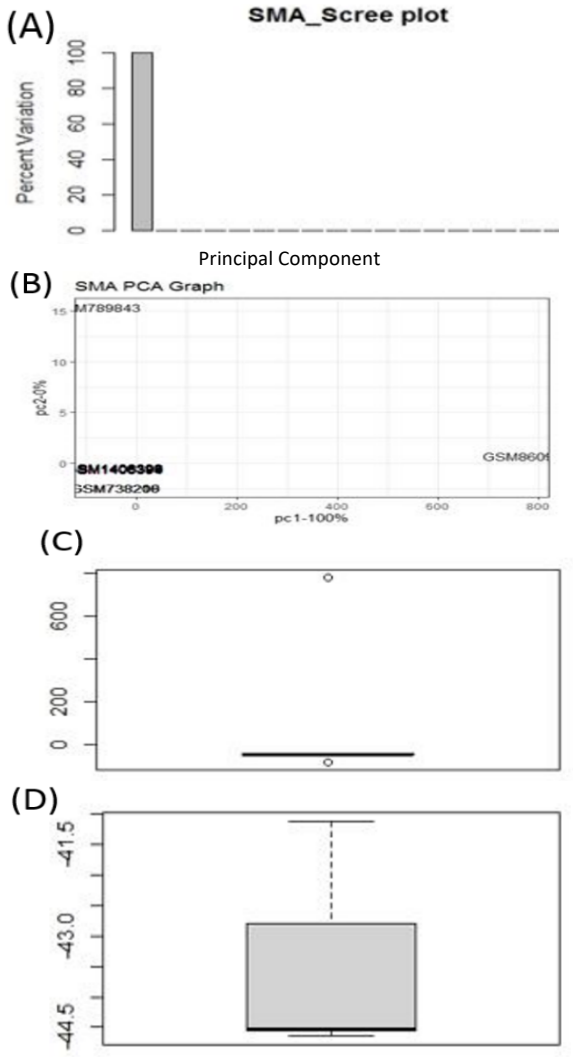


Figure 4. principal component analyses and Box plot to detect outlier samples. A) first principal component is suitable for analysis of all 18 samples, B) all samples have an equivalent first and second principal component analysis except 2 samples, C) box plot for all 18 samples showing 2 outlier samples, D) box plot for the same remaining 16 samples with assumption that the hybridization and processing conditions were confident for further analysis.

First, using the library of (AnnotationDbi) and (org.Hs.eg.db) packages in package 'BiocGenerics' that was built under R version 4.0.3 (https://annotation/html/), all duplicate probe sets in each microarray dataset were identified as important features (GO, Entrez ID). In the next step, delete probe sets that do not have "Entrez id" or "Gene name". Then, the collected samples are further grouped, and probes of each group are checked for excessive missing values. Finally, the remaining probe set of genes (SMA_ FII) represents the current gene set that will be analyzed further.

Currently, six DEG cases are available, which can be divided into two groups. Each of them contains three DEG cases. For the first group, embryonic stem cells were compared to the SMA types tested (Carrier SMA (DEG), Sever_ SMA (DEG2), and Mildly SMA (DEG3)) using a threshold (P= 0.99 and FC= 1.5).

The resulting gene number is sufficient for each type, as shown in Figure 5A and C. In the second group, the test and control data were mutually compared with three types of SMA data to determine the DEGs for the three cases (DEG4: (Mildly_SMA vs. Carrier SMA), DEG5: (Carrier SMA vs. Sever_ SMA), and DEG6: (Mildly_ SMA vs. Sever_ SMA)) to determine why the severity discrepancy exists.

However, when the same threshold was used in the first group, the results showed that the gene number was insufficient, as shown in Figure 5B. Consequently, this group would be further analyzed using the top 2000 differentially expressed genes with a different threshold for each case as shown in Figure 5D. As illustrated by two volcano plots (Figure 5C and D). Determining the common genes and the different genes between the different SMA types and stem cells will help analyze the pathological mechanism of SMA in other conditions at the molecular level.

Furthermore, explain more why there is a severity discrepancy in SMA types. We would take the expressed values of SMA carriers' genes derived from DEG4 (Mildly SMA vs. Carrier SMA) for clustering genes to modules and identifying the hub genes for each module. Moreover, identifying the interesting modules concerned with different cases for further analysis. These interesting modules were determined here as highly correlated modules with effective genes in SMA, which are already mentioned in previous studies.

Network construction and modules detection

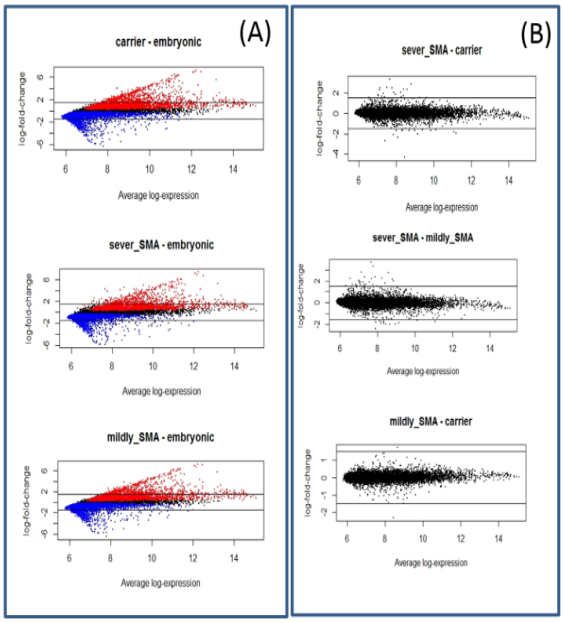
Using the WGCNA R package, the similarity matrix was calculated by applying the Pearson correlation method to the gutted carrier gene expression matrix. Then, by applying the protocol of (WGCNA) to determine the parameter of the adjacency function ($\boldsymbol{\beta}$) as in equation (1).

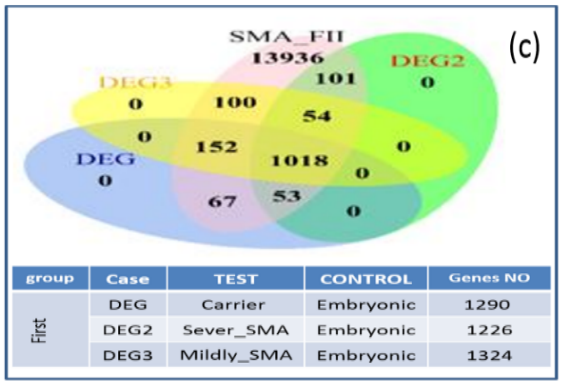
Accordingly, Figure 6 illustrates the value of the soft threshold at the first saturation curve ($\boldsymbol{\beta} = \mathbf{7}$) at which the constructed weighted network has good free topological criteria. After that, the similarity matrix was transformed into an adjacency matrix. Using equation (2), the adjacency matrix was transformed into a topological overlap matrix, which considers biological significance in addition to statistical significance. According to the protocol of the WGCNA method, the soft threshold of the adjacency function that ensures the free topological criteria of the designed weighted gene co-expression network ($\boldsymbol{\beta} = \mathbf{7}$) at the first saturation curve and R2 >0.8, and its equivalent mean connectivity of the network.

Unsupervised machine learning was used to generate the initial modules named dynamic modules in Figure 7B. As it happens, the node dissimilarity method (1- TOM) in conjunction with a dynamic tree-cutting algorithm for the hierarchical clustering with a minimum module size of 30 genes was used. However, the dynamic tree-cutting algorithm combines the advantage of both hierarchical and k-mean non-hierarchical clustering. But some modules are highly correlated among all the resulting modules (Zhang & Horvath, 2004).

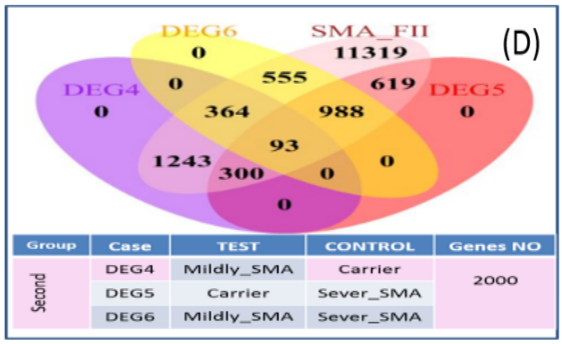
Dynamic Module Eigengenes was calculated as the first principal component. Moreover, these modules are related by correlating the corresponding modules' Eigengenes (E) to each other in a hierarchical clustering shape, as shown in Figure 7A. Furthermore, the red line in Figure 7A represents a cutting height=0.25. Finally, the modules at this line were merged since their genes are highly co-expressed (more than 0.75 correlation), obtaining the final modules (Merging modules) as shown in Figures. 7B and C.

Figure 7 illustrates a hierarchical clustering of the initial clustered modules using the corresponding modules' Eigengenes (E) to represent the correlation between each other and merging the highly correlated modules (>0.75) at the red line (A). The number of genes per each merging module is shown in (C).





group	Case	TEST	CONTROL	Genes NO
First	DEG	Carrier	Embryonic	1290
	DEG2	Sever_SMA	Embryonic	1226
	DEG3	Mildly_SMA	Embryonic	1324



Group	Case	TEST	CONTROL	Genes NO
Second	DEG4	Mildly_SMA	Carrier	2000
	DEG5	Carrier	Sever_SMA	
	DEG6	Mildly_SMA	Sever_SMA	

Figure 5. Differentially Expressed Genes (DEGs) with certain thresholds (A, B) and Volcano Plots for Differentially Expressed Genes (DEGs) (C, D).

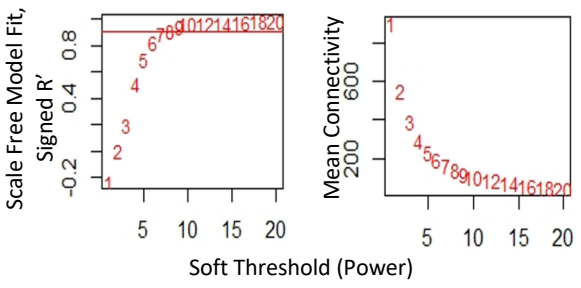


Figure 6. The soft threshold for adjacency function

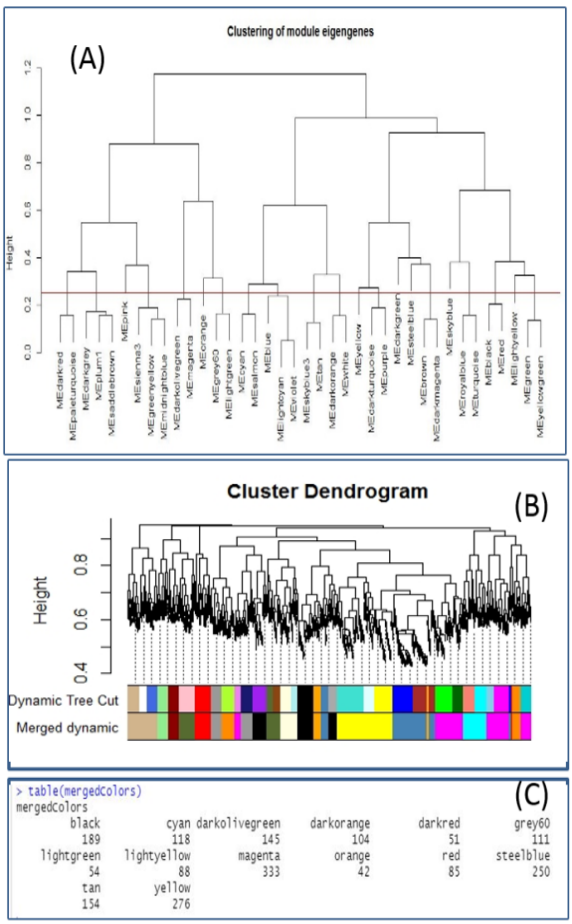


Figure 7. Modules detection

By considering the modules of the current network (Carrier SMA in DEG4) as a reference, module preservation methods were used here to evaluate the efficiency of the module's clustering process. Certainly, by setting the current network's data as test and control data. As shown in Figure 8, all modules are strongly preserved ($Z_{summary}$ > 10), indicating that the module clustering process is efficient.

Figure 8 illustrates that the $z_{summary}$ of all merging modules>10 which means high preservation that indicates good efficiency of the clustering process. The current network can also be compared with other test networks, such as the Mildly SMA patients' network and the embryonic stem cells network, to identify each preserved module (Table 2).

Also, other interesting modules were determined here as highly correlated modules (≥±0.7) with certain genes such as SMN1, SMN2, NAIP, DYNC1H1, and PLS3, which have a direct relationship with SMA disease and its severity.

Table 2. The interesting Modules in different study cases Vs. SMA carriers

Module	Mildly SMA	Embryonic stem cells	SMN1	SMN2	NAIP	DYNC1H1	PLS3
Tan							
Cyan					√	√	√
Light yellow	√						
Dark olive green				√			
Yellow	√	√					
Black			√	√		√	√
Grey60				√			√
Magenta	√	√					
Orange					√	√	√
Steel blue		√					
Dark orange				√	√		√
Dark red				√			
Light green			√	√			
Red					√		

Table 3. The top 10 hub genes in SMA Carriers Modules

Module	1	2	3	4	5	6	7	8	9	10
Tan	TAF8	SORCS2	MPEG1	ADH1B	NR2F6	SHROOM4	PROS1	ZSCAN4	TUBA3D	BMP10
Cyan	EEF2K	GRIA1	GFAP	ZNF691	SPRYD3	PDCD2	TAOK1	GGA2	SGK2	SMARCD1
Light yellow	ZMAT3	TMEM98	TCF4	SBF1	GRM6	NFAT5	PHF21B	SLC10A6	PAPOLG	NDUFS6
Dark olive green	TC2N	NFKBIL1	CHST3	RBMS3	CACNA1E	RPL8	TIE1	ZNF658B	ZNF304	TREH
Yellow	PARP4	PLEKHA3	Bcl2	NME6	RUVBL2	LPCAT1	SEMA4B	ZNF131	UBE2T	SERPINI2
Black	TRPM4	HYPK	SPRYD4	EPC2	FAM86B1	DLG5	NCF4	TDRD1	ATP6V0E2	ABCA5
Grey60	DSG3	OR1M1	ZBTB25	APOM	TMEM207	NMNAT3	CD247	RALGPS1	RGS8	DCTN6
Magenta	RANBP1	Cntn1	RASGRP3	CD300C	ENDOG	GPATCH4	ZNF589	COPS6	U2AF1	NCALD
Orange	THAP11	YIPF2	NUDCD2	FOXP1	OCEL1	NIN	MRPS18A	ABHD14A	GPC6	TP53
Steel blue	BAGE3	IREB2	FBXO4	CD63	TYRP1	UBE2J1	SNORD36A	RPS26P10	SERF1B	HDAC11
Dark orange	ARR3	TIMP3	STEAP1	IRF5	PPIB	ALG5	OR4C45	ADH5	MYOM2	PRRG2
Dark red	N4BP2	PFDN2	SPINK4	ARMC4	MMAB	SHISA3	OR8D4	GTF2H4	PI4KB	VIT
Light green	PAK2	DPRXP4	MTX2	GDF7	BAGE5	PARP14	SLC22A9	OR10A6	CKAP4	PTGIR
Red	CENPN	TAS2R38	RAD23B	OR4A15	SFMBT2	ZNF19	OR52E2	CHD7	TARS2	TULP1

These genes have been explained in several previous studies. The most highly connected intramodular genes (Module's hub genes) were determined using the Cytoscape program (https://cytoscape.org/) for all clustering modules (Table 3) (Shengni et al., 2020).

DISCUSSION

The recessive mutation in the SMN1 gene that gives rise to the hereditary illness known as spinal muscular atrophy (SMA) results in the loss of a survival motor neuron (SMN), which causes the muscles that are necessary for life to weaken and atrophy. It is the most prevalent cause of infant death associated with genetic defects. Despite extensive research to discover a solution for this illness, even the most effective medicines still have significant restrictions and challenges. Based on a human microarray dataset for various SMA instances from the Gene Expression Omnibus (GEO) databases. The current study investigates how genes impact and interact with one another. Stem cells have the ability to differentiate into numerous types of cells and to self-renew.

Stem cell injection in conjunction with other therapeutic treatments has the potential to enhance, modify, or initiate local or systemic

healing processes to target specific tissues or organs. As a result, stem cell can be used to deliver synthetic molecules for therapeutic purposes. Several stem cell samples were deposited on the same platform as our SMA data (GPL6947) as part of our efforts to support stem cell research in SMA disease analysis.

However, these samples were gathered under various experimental settings. Using hierarchical clustering, abnormal samples were eliminated, and those with conditions comparable to those of the samples taken for the aforementioned disease were retained, with acceptable allowances allowing the analysis to be completed. In addition, PCA and boxplot were used to identify outlier samples and approve the remaining samples based on the assumption that the hybridization and processing conditions for all 16 samples are sufficient for further analysis within acceptable limits.

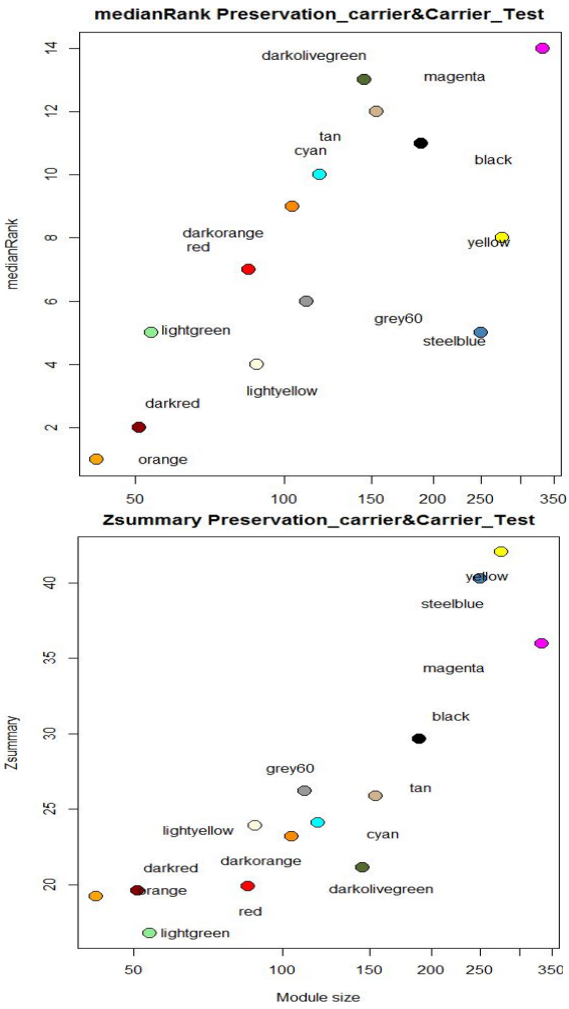


Figure 8. Modules preservation to check clustering efficiency

The preprocessing steps have been carried out on the genes of the SMA samples using Bioconductor software packages and libraries. The highest differentiated expression genes were determined for each of the four available SMA types (SMA carriers, mildly SMA, severe SMA, and embryonic stem cells) using the LIMMA package. Consequently, the Volcano plots will help professionals analyze the causes of SMA severity degree more thoroughly by identifying genes unique to every DEG case or common to two or more DEGs.

Networks are the best way to represent complex interactions between DEGs. The weighted network was preferred in the current work rather than the unweighted one. For the following reasons, the selection of the value of its threshold is less sensitive. It saves information from losses more than the unweighted one, and therefore, it provides more accurate results for further analysis. The WGCNA R software package calculated the similarity matrix by applying a Pearson correlation coefficient to the DEG matrix. Then, the WGCNA package's protocol was used to obtain the appropriate soft threshold of the adjacency function and accordingly transform the similarity matrix into an adjacency one that describes the strength of the connection between each pair of genes. However, the previous network construction steps were verified for statistical significance only.

It is better in the biological network to consider the biological relevance also by looking at shared neighbors. This significance can be taken into consideration as it occurs in the current study and in (Yang et al., 2016) study where they used the topological overlap criteria, which means that any two individuals may belong to the same clique (module) if they have the same friends. Moreover, the biological significance can be added as in (YANG et al., 2019) using the guilt-by-association method. It means that similar features such as genetic or physical interactions are commonly shared by genes with similar functions. However, the study presented on SMA disease (YANG et al., 2019) has some weaknesses as their data was based on Duchenne muscular dystrophy disease not

on SMA disease in accordance to their state. The analysis treated the DEGs network as a whole, not with clustered modules. However, intramodular connectivity correlates more strongly with gene significance than the whole network (Zhang & Horvath, 2004).

In the current work, the WGCNA package and the hierarchical clustering machine learning were used for gene module detection, then the module preservation methods were used to check the clustering efficiency. Further, the Database for Annotation, Visualization and Integrated Discovery (DAVID) v2022q1 https://david.ncifcrf.gov/ (Accessed on 15/ 5/ 2022) as a gene set enrichment analysis (GSEA) tool was used to remove genes with unknown functions from each module (Huang et al., 2007). In addition, the WGCNA package interfaced with the Cytoscape program to construct a visualized network and determine the hub genes for each module.

Finally, the interesting modules are defined per each case study as the common modules between SMA carriers and those SMA cases or Embryonic stem cells. Despite the proven significant effect of some genes such as SMN1, SMN2, NAIP, DYNC1H1, and PLS3 on the occurrence of the SMA disease and the change in its severity, these genes did not appear in the forefront of the differentially expressed genes among the different cases. Consequently, another consideration for the interesting modules is the highly correlated modules of the current SMA network with these specific SMA effective genes. Indeed, these modules and their hub genes are at the forefront of major contributions to the current work.

Early studies have linked several of these hub genes to SMA. Such as BCL2 (a pathway in skeletal development) (Horvath & Steve., 2011) (Tibshirani & Walther, 2005), Cntn1 (a path in nervous system development) (Horvath & Steve., 2011), and TYRP1. The effects of this may result in brain and nervous system injury (Chen & Tai-Heng, 2020) (Langfelder et al., 2011) (Kapp & Tibshirani, 2007). Several genes have not been found in previous SMA studies, yet they play a related role, such as the N4Bp2 gene, which regulates transcription-coupled DNA repair. In addition to multiple transcript variants encoding different isoforms of PFDN2, this gene is annotated as a chaperone for protein folding and one that binds unfolded proteins according to Gene Ontology (GO). Moreover, the *TAF8* gene is related to Neurodevelopmental disorder with severe motor impairment, absent language, cerebral hypomyelination, and brain atrophy. Hub genes of SMA modules were identified to study the causes of disease and its severity variation and to develop better diagnoses and therapies (Krebs et al., 2006).

CONCLUSIONS

SMA patients and their families bear a considerable burden in the absence of effective trials that treat the disease intensity. This is incompatible with our desire for everyone to live a healthy and safe life. Therefore, it seems likely, that this work will benefit these patients and their families. We join the convoys of researchers who have this noble goal of treating the diseases that afflict humanity in general and the SMA disease in particular. Furthermore, since the genes usually do not work alone but cooperate with each other, the differences in gene expression levels can be used to construct a weighted gene network that can reflect the interaction between genes. Further, combining machine learning, LIMMA, WGCNA, GSEA, and Cytoscape program can analyze SMA and stem cells data obtained from the GEO database and provide insight into the molecular mechanisms of disease. SMA key genes can be predicted in different cases to improve diagnosis and therapy development for patients, carriers, and prenatal diagnosis. This pilot study explores several genes to be used as key genes for several case studies. For each case study with respect to the carrier case, it's interesting modules are illustrated in Table 2. Then, the hub genes for these modules, which are considered the hub genes of this specific case, are found in table 3. This work offers critical information for future research on treatments and disease processes in SMA, even if the functions of these target genes in the pathogenesis of SMA are still not apparent.

LIMITATION OF THE STUDY

Embryonic stem cell data with identical experiment circumstances, SMA animal data for

animal model research, and extra-human SMA samples were not included in our research because they were not available. When all of these considerations are taken into account, the outcome will be more favorable. Under the initiative of the President of the Arab Republic of Egypt and NOVARTS to treat Egyptian SMA patients (less than two years): We wanted to conduct this pilot study on a sample of Egyptian SMA patients (less than 2 years old) before and after therapy with Zolgensma in order to gain a better understanding of the genes most affected by the drug, which would aid in SMA treatment and diagnosis. In addition, we want to conduct this pilot study on SMA patients of varying ages after therapy with Zolgensma in order to address irreversible SMN degenerative processes.

ACKNOWLEDGMENTS

We are indebted to Menna Allah M. Dosokey, Faculty of Science, Tanta University, Dr. Esraa M. Imam, Faculty of Medicine, Tanta University, and Moustafa A. Hassan; Medical Research Institute, Alexandria University for invaluable discussions and support.

FUNDING

This research did not receive any specific grant from funding agencies in public, commercial, or not-for-profit sectors.

HIGHLIGHTS

- The highly differentially expressed genes (DEG) were identified using LIMMA Package.
- stem cells have an important share of studies on genetic diseases.
- Network construction and module detection by using clustering and the WGCNA package.
- highly correlated modules with effective genes in SMA were identified.
- The common modules for different SMA types were identified by preservation methods.

REFERENCES

Albert, & Réka. (2005). Scale-free networks in cell biology. Journal of Cell Science, 118(doi:10.1242/jcs.02714), 4947-4957.

Anderson, Kirstie, Talbot, & Kevin. (2003). Spinal muscular atrophies reveal motor neuron vulnerability to defects in ribonucleoprotein handling. Current Opinion in Neurology.

Carter, S. L. (2004). Gene co-expression network topology provides a framework for molecular characterization of cellular state. Bioinformatics, 14(20), 2242–2250.

Chaytow, H., Huang, Y.-T., Gillingwater, T. H., & Faller., K. M. (2018). The role of survival motor neuron protein (SMN) in protein. Cellular and Molecular Life Sciences, 75, 3877–3894.

Chen, & Tai-Heng. (2020). New and Developing Therapies in Spinal Muscular Atrophy: From Genotype to Phenotype to Treatment and Where Do We Stand? Molecular Sciences, 3297(doi:10.3390/ijms21093297), 21.

Corti, Parente, V., & Stefania. (2018). Advances in spinal muscular atrophy therapeutics. Therapeutic Advances in Neurological Disorders, 11(4), 1-13.

Eisen, M. B., Spellman, P. T., Brown, P. O., & Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. USA, 95, 14863-14868.

Goh, Cusick, Valle, Childs, Vidal, & Barabási. (2007). The human disease network. PNAS, 1(104), 8685–8690.

Hartwell, L. H. (1999). From molecular to modular cell biology. Nature, 402(6761 Suppl), C47–52.

Hassan, H. A., Zaki, M. S., Issa, M. Y., & El-Bagoury., N. M. (2020). Genetic pattern of SMN1, SMN2, and NAIP. Egyptian Journal of Medical Human Genetics, 4(doi.org/10.1186/s43042-019-0044-z), 21.

Hensel, N., Kubinski, S., & Claus, P. (2020). The Need for SMN-Independent Treatments of Spinal Muscular Atrophy (SMA) to Complement SMN-Enhancing Drugs. Frontiers in Neurology, 11(10.3389), Article 45.

Horvath & Steve. (2011). EvaluatingWhether a Module is Preserved in Another Network. In H. &. Steve, Weighted Network Analysis Applications in Genomics and Systems Biology (pp. 207-220). New York Dordrecht Heidelberg London: Springer.

Horvath, Andy, & Steve. (2007). Gene network interconnectedness and the generalized topological overlap measure. BMC Bioinformatics, 2015(471), 8-22.

Horvath, S. (2011). Chapter 9: Evaluating Whether a Module is Preserved in Another Network. In Weighted Network Analysis Applications in Genomics and Systems Biology (pp. 207-247). New York: Springer.

Huang, D. W., Sherman, B. T., Tan, Q., Collins, J. R., Alvord, W. G., Roayaei, J., . . . Lempicki, R. A. (2007, 9 4). The DAVID Gene Functional Classification Tool: a novel biological module-

centric algorithm to functionally analyze large gene lists. Genome Biology Article R183, 8(9), 1-16.

Jeong, H. M. (2001). Lethality and centrality in protein networks. Nature, 411(may), 41–43.

Kapp, A., & Tibshirani, R. (2007). Are clusters found in one dataset present in another dataset? Biostat, 1(8), 9–31.

Krebs, S., Medugorac, I., Russ, I., Ossent, P., Bleul, U., & Schmahl, W. (2006). Fine-mapping and candidate gene. the International Mammalian Genome Society, 1(17), 67–76.

Langfelder, Luo, MC, O., & S, H. (2011). Is my network module preserved and reproducible? Plos Comput Biol 7(1):e1001057.

Ludolph, D, C., & C., W. a. (2018). Nusinersen for spinal muscular atrophy. Therapeutic Advances in Neurological Disorders, pp. Vol. 11: 1 –3.

Mendell, J., Shell, S. A., Arnold, W., Rodino-Klapac, L., Prior, T., Lowes, L., . . . Berry., K. (2017). Single-Dose Gene-Replacement Therapy for Spinal Muscular. The new england journal of medicine, 377(12), 1713- 1722.

Prior, T. W., PhD, F., Leach, M. E., & Erika Finanger, M. (2019). Spinal Muscular Atrophy. Gene Reviews.

Ritchie, M., Phipson, B., Yifang, & Law, C. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies and Gordon K. Smyth. Nucleic Acids Research, 43(7), e47.

Shawky, R. M., & El-Sayed, N. S. (2011). Clinico-epidemiologic characteristics of spinal muscular atrophy among Egyptians. The Egyptian Journal of Medical Human Genetics, 12(1), 25-30.

Shawky, R., Aleem, K. A., Rifaat, M., & Moustafa, A. (2001). Molecular diagnosis of Spinal Muscular Atrophy in Egyptians. Eastern Mediterranean Health Journal, 7(1/2), 229- 237.

Shengni, Zhonghua, Yingyao, Meixiao, Wang, & Ligong. (2020). Identification of hub genes in hepatocellular carcinoma using integrated bioinformatic analysis. AGING, 12(6), 5439- 5498.

SMA Care series. (2009). SMA Care. (SMA Care Series) Retrieved 2009, from www.cureSMA.org

Stuart, J. M. (2003). A gene-coexpression network for global discovery of conserved genetic modules. Science, 302(5643), 249–255.

Tibshirani, R., & Walther, G. (2005). Cluster validation by prediction strength. J Comput Graph Stat, 14, 511–528.

U.S. National Library of Medicine. (2017). Genetics Home Reference. Retrieved from U.S. National Library of Medicine, National Institutes of Health, Department of Health & Human Services: https://ghr.nlm.nih.gov/

Vamshi K. Rao, M. M., Daniel Kapp, P. B., & and Mary Schroth, M. (2018). Gene Therapy for Spinal Muscular Atrophy An Emerging Treatment Option for a Devastating Disease. Journal of Managed Care & specialty Pharmacy, 24(12-a), Supplement.

Wirth, B., Karakaya, M., & Kye., M. J. (2020, January 24). Twenty-Five Years of Spinal. Annual Review of Genomics and Human Genetics, pp. 12-47.

Xiaowen, B. (2020). Stem Cell-Based Disease Modeling and Cell Therapy. Cells, www.mdpi.com/journal/cells, 2193; doi:10.3390/cells9102193.

Yang, C.-W., Chien-Lin, Chou, W.-C., Ho-Chen, Jong, Y.-J., Tsai, & Chuang, C.-Y. (2016). An Integrative Transcriptomic Analysis for Identifying Novel Target Genes Corresponding to Severity Spectrum in Spinal Muscular Atrophy. journal.pone(DOI:10.1371), 1-18.

YANG, W., JING, JINFENG, FENG, Y., HOU, WANG, & TENGBO. (2019). Prediction of key gene function in spinal muscular atrophy using guilt by association method based on network and gene ontology. EXPERIMENTAL AND THERAPEUTIC MEDICINE, 17, 2561- 2566.

Zhang& Horvath. (2005). A network approach to detecting individual prognostic genes and therapeutic targets in brain cancer. De Gruyter. https://doi.org/doi.org/10.2202/1544-6115.1128

Zhang, B., & Horvath, S. (2004). A General Framework for Weighted Gene Co-Expression Network Analysis. Statistical Applications in Genetics and Molecular Biology, 4(1), Article 17.