BMC Bioinformatics

METHODOLOGY

Open Access

A statistical network pre-processing method to improve relevance and significance of gene lists in microarray gene expression studies

Giuseppe Agapito[1,2*], Marianna Milano[2,3] and Mario Cannataro[2,3]

From The 17th International Conference on Computational Intelligence Methods for Bioinformatics and Biostatistics (CIBB 2021)
Virtual. 15-17 November 2021

*Correspondence:
agapito@unicz.it

[1] Department of Law, Economics and Sociology Sciences, University Magna Græcia, 88100 Catanzaro, Italy
[2] Data Analytics Research Center, University Magna Græcia, 88100 Catanzaro, Italy
[3] Department of Medical and Surgical Sciences, University Magna Græcia, 88100 Catanzaro, Italy

Abstract

Background: Microarrays can perform large scale studies of differential expressed gene (DEGs) and even single nucleotide polymorphisms (SNPs), thereby screening thousands of genes for single experiment simultaneously. However, DEGs and SNPs are still just as enigmatic as the first sequence of the genome. Because they are independent from the affected biological context. Pathway enrichment analysis (PEA) can overcome this obstacle by linking both DEGs and SNPs to the affected biological pathways and consequently to the underlying biological functions and processes.

Results: To improve the enrichment analysis results, we present a new statistical network pre-processing method by mapping DEGs and SNPs on a biological network that can improve the relevance and significance of the DEGs or SNPs of interest to incorporate pathway topology information into the PEA. The proposed methodology improves the statistical significance of the PEA analysis in terms of computed *p* value for each enriched pathways and limit the number of enriched pathways. This helps reduce the number of relevant biological pathways with respect to a non-specific list of genes.

Conclusion: The proposed method provides two-fold enhancements. Network analysis reveals fewer DEGs, by selecting only relevant DEGs and the detected DEGs improve the enriched pathways' statistical significance, rather than simply using a general list of genes.

Keywords: Biological pathways, Differential expressed genes, Pathway enrichment analysis, Statistical analysis, Data mining network, Network analysis, SNPs

Introduction

The advent of microarrays [1] allowed for efficient investigation of genetic matter, making it possible to improve both real-time polymerase chain reaction (RT-PCR) [2] and the Sanger methods [3], allowing large scale studies of differential expressed genes (DEGs) and even single nucleotide polymorphisms (SNPs) [4].

In this manner, microarrays allow for screening of thousands of genes for a single experiment. Sanger-method, RT-PCR and microarrays rely on the extension of small segments of DNA through the *polymerase biological process*. All the cited methods will extend the genetic sequence of interest by adding on the complementary nucleotide from the template DNA strand. These methods allow a relative and accurate quantification of DNA and mRNA molecules with a sufficiently high reproducibility and low variability, and they are all well suited to study gene expression.

However, after the initial fervor, it became apparent that even the lists of DEGs or SNPs were mainly as enigmatic as the first nucleotide sequence of the genome. The main reason being that these lists of DEGs and SNPs are independent from the affected biological context. To overcome this limitation, several statistical software tools [5–9] have been developed to help researchers analyze this enormous amount of microarray data to elucidate more valuable and suitable outcome for clinical activities. In addition, several data mining software tools [10–12] are available that allow computation of multiple associations among SNPs. The produced results, from both categories, provide lists of DEGs or SNPs that are unlikely to be directly used in clinical activities, because results are still disconnected from the affected biological processes.

Pathway enrichment analysis (PEA) can facilitate the interpretation of such a list of DEGs or SNPs, linking both to the affected biological pathways and consequently to the underlying biological functions and processes. Although, PEA can help figure out the affected biological pathways starting from the DEGs or SNPs of interest, poor quality and relevance of the employed input can produce pathways that are not directly related to the condition under investigation. This is due to the fact that, a poor quality list of DEGs or SNPs can enrich a general pathway such as *disease,* rather than a more specific one like *cellular responses to external stimuli*, a well-known pathway involved in the progression of colorectal cancer, for example. These biases prevent researchers from figuring out the proper affected biological pathways and the related functional interactions.

To improve the enrichment analysis results, it is necessary to determine the relevant DEGs that can both improve the *p* value (i.e. relevance) of the enriched pathways and reduce the number of enriched pathways, consequently improving their relevance with respect to the condition under investigation.

For these reasons, we developed a new DEG preprocessing method based on statistical and networks analysis. The proposed method identifies, from the whole DEGs list of interest, the most relevant genes with which to perform PEA. In short, the proposed method follows these steps: *(i)* DEG filtering relies on the *Kruskal–Wallis test* [13] to select only DEGs with similar behaviours from the provided input list, splitting DEGs in up- and down-regulated gene groups. In addition, *Kruskal–Wallis test* returns results in the form of matrices. The provided matrices contain the *p* values for each group, that will be used to build gene interaction networks. In this model, the computed *Kruskal–Wallis p* values are considered as a similarity measure among gene pairs [14, 15]. *(ii)* Next, the computed similarity matrices are converted into networks from which the essential DEGs are extracted. *(iii)* Finally, both essential DEGs groups are mapped separately on the human protein-protein interaction (PPI) network obtained from the Integrated

Interactions Database (*IID*) database [16] to discover additional relevant genes to perform PEA analysis.

The rest of the paper is organized as follows. Section 2 describes the provenance of the downloaded gene expression data sets, the methods employed to obtain the list of DEGs, and the threshold used to select DEGs. Section 2.6 highlights and details the major phases of the DEG preprocessing methodology. Section 3 describes and discusses the preliminary results as a validation of our approach, highlighting the principal benefits. Section 4 validates the enrichment results by manually exploring the literature and finally, Sect. 5 concludes the paper.

Methods

Data set

Microarray assays are extensively used in many omics data analyses for several reasons. First microarrays analyse are cheaper than Next-Generation Sequencing (NGS), RNAseq. Second, extensive microarray studies are available in the literature and cover a variety of different phenotypes. Microarray data are curated, providing well-documented criteria, making it easy to verify the accuracy and reproducibility of the research. In addition, microarray data sets can be used as benchmarks to validate data analysis workflows. Hence, we have chosen to use GEO microarray data sets to perform the preliminary tests of our methodology.

We downloaded from the Gene Expression Omnibus (GEO) database [17] the following data sets:

- **GSE1297** [18] provides microarray correlation analysis of hippocampal gene expression deemed to be responsible for incipient Alzheimer's disease (AD). The data set contains data from approximately 31 subjects: 9 controls, and 22 cases affected by AD. Expression profiles were collected using Affymetrix Human Genome U133A Array. For further details see https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE1297.
- **GSE5281** [19–22] contains gene expression profiling data collected from brain samples. The Affymetrix Human Genome U133 Plus 2.0 Array was used to yield the expression profiles. The data set is comprised of data from about 161 subjects: 100 Alzheimer subjects, and 61 controls. Both samples groups are related to six brain regions that are histopathologically or metabolically relevant to AD and aging. For further details see https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE5281.
- **GSE16759** [23] contains a combination of profiled messenger RNA (*mRNA*) and *microRNA* (*miRNA*) expressions to define the role of miRNAs in AD. Expression profiles were obtained using Affymetrix Human Genome U133 Plus Array and the USC/XJZ Human 0.9 K miRNA-940-v1.0. The overall design of the *GSE16759* data set is parietal lobe tissue from 4 Alzheimer's subjects and 4 age-matched controls. For further details see https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE16759.
- **GSE9476** [24] describes the use of microarrays to identify previously unrecognized expression changes that occur only in acute myeloid leukemia (AML) blasts. Expression profiles were obtained using Affymetrix Human Genome U133A Array. The

overall design includes gene expression profiles between normal hematopoietic cells from 38 healthy controls, and leukemic blasts from 26 AML patients. Eighteen normal hematopoietic samples included CD34+ selected cells, 10 unselected bone marrows cells, and 10 unselected peripheral blood cells. For further details see https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE9476.

- **GSE14924** [25] attempts to prove that *T* cells from patients with chronic lymphocytic leukemia (CLL) show differentially regulated genes compared with healthy *T* cells. Expression profiles were obtained using Affymetrix Human Genome U133 Plus 2.0 Array. The overall design includes gene expression profiles of four groups of samples: 10 AML CD4, 10 AML CD8, 10 Healthy CD4, and 11 Healthy CD8. AML samples were chosen to represent the range of prognostic groups and patient outcomes. For further details see https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE14924.
- **GSE24739** [26, 27] encompasses gene expressions of normal and chronic myelogenous leukemia. The differentially expressed genes were grouped according to their reported functions, and correlations were sought with biological differences previously observed between the same groups. Expression profiles were obtained using Affymetrix Human Genome U133 Plus 2.0 Array. The overall design includes gene expression profiles of 8 AML samples and 4 normal samples. For further details see https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE24739.

The main features of the six downloaded data sets are listed in Table 1.

Figures 1 and 2 show the Uniform Manifold Approximation and Projection (UMAP) and Volcano plot related to the downloaded data sets.

Detection of DEGs with GEO2R

To identify the differential expressed genes between cases and controls, we used *GEO2R* (https://www.ncbi.nlm.nih.gov/geo/geo2r/). *GEO2R* is an interactive online analysis tool used to detect DEGs enclosed in expression profile array data sets. *GEO2R* allows classification of subjects in several groups, using the *define groups* command. The panel *options* available in *GEO2R*, allow straightforward analysis customization. The option panel enables users to select the statistical corrector, the data normalization method, and the cut-off value to filter out the genes not holding the defined cut-off. In addition, *GEO2R* exploits the `limma` package to perform inter- and intra-sample normalization. To perform DEG analysis, we selected the false discovery rate (FDR) *p* value adjustment

Table 1 A summarization of the main features of the downloaded data sets

	Disease	Cases	Controls
GSE1297	Alzheimer's disease	22	9
GSE5281	Alzheimer's disease	100	61
GSE16759	Alzheimer's disease	4	4
GSE9476	Acute myeloid leukemia	26	38
GSE14924	Acute myeloid leukemia	20	21
GSE24739	Acute myeloid leukemia	8	4

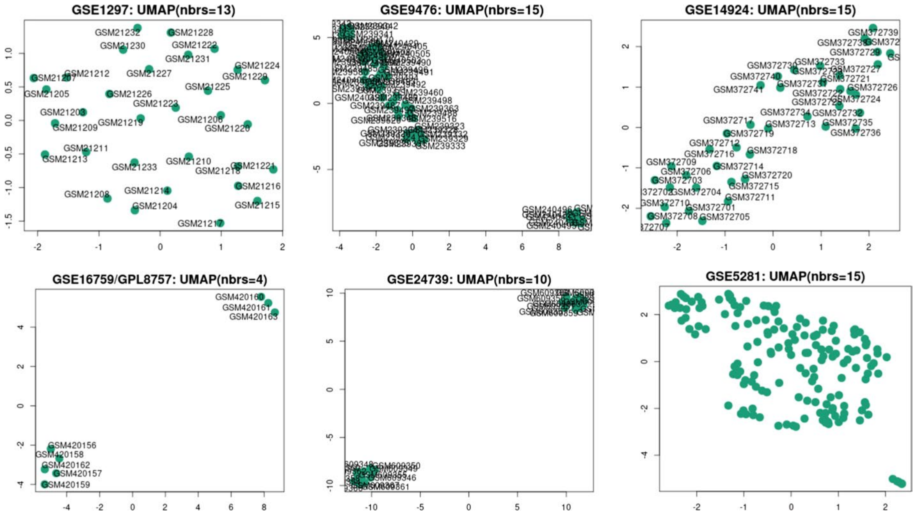

Fig. 1 The UMAP of the six downloaded data sets

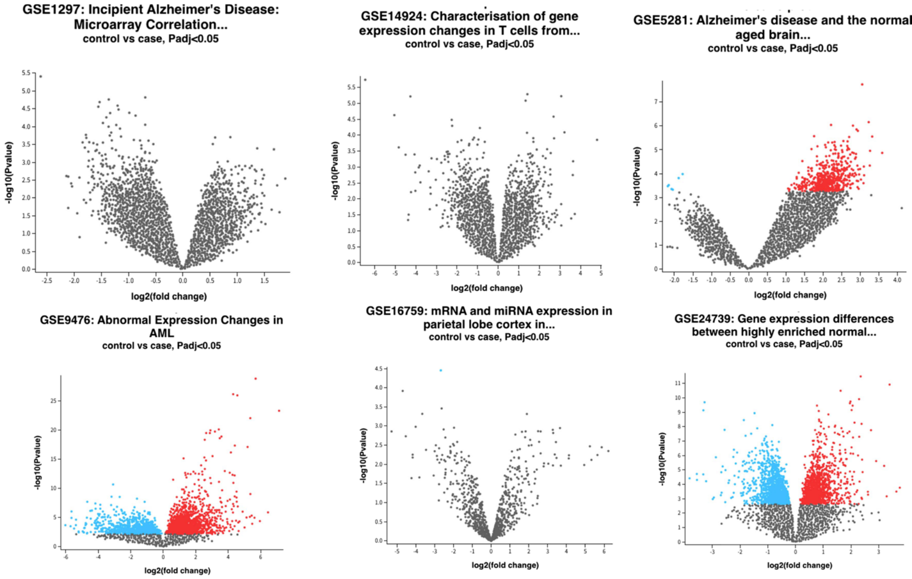

Fig. 2 The Volcano Plot of the six downloaded data sets

for multiple testing, and the *log data transformation* method to normalize the results. Finally, we selected and downloaded the following results: adjusted p value, p value, *logFC*, gene symbol, and title.

Selection of DEGs

A cohort of DEGs was obtained by filtering out all the genes that do not meet the adjusted threshold criteria (p value ≤ 0.005 and $|(logFC)| \geq 1.5$). The fold change logarithm (*logFC*) is a metric to assess the change in the ratio between the expression levels of two genes. Hence, the genes meeting both criteria were designated as DEGs. DEGs associated to negative *logFC* values are classified as down-regulated DEGs, otherwise they were classified as up-regulated DEGs. Next, the DEGs are investigate using the *Kruskal–Wallis* test [13], to differentiate genes with similar behaviour. The Kruskal–Wallis test evaluates the similarity between pairs of genes, assessing if two genes are correlated [14, 15]. In general, Kruskal–Wallis test is applied to test the null hypothesis which states that *k* number of samples have been starved from the same population or an identical population with the same or identical median. In this manner, accepting the null hypothesis, e.g., a *p* value greater than 0.005, allow coupling of genes with the same median among them enabling identification of genes with the same statistical behaviours.

Pathway enrichment analysis

To identify the connections among DEGs with the affected biological functions, we can use PEA, making it possible to take advantage of the pathway database's information to discover connections with biological mechanisms. This approach helps researchers interpret gene lists, or other biological entity lists of interest, disconnected from the biological context, facilitating and validating their findings [28, 29]. To perform PEA, we used the BioPAX-Parser (BiP) software tool [30], an automatic and graphics-based tool to achieve PEA by using pathways data encoded in BioPAX format. BioPAX-Parser is fully developed using Java 8, and helps perform PEA by merely loading a list of proteins/genes of interest. Enrichment in BiP implements the Hypergeometric test, False Discovery Rate (FDR), and Bonferroni multiple-test statistical correctors.

Pathway data

Pathway data were collected from the *Reactome* database [31] (version 79) with BiP. Reactome is an open source, open access, manually curated, and peer-reviewed database of human pathways, biological processes and biochemical reactions. Reactome is the result of the joint efforts of several international research institutes. In the current version, Reactome contains the whole known pathways coming from 22 different organisms including the *Homo sapiens*. Reactome includes over 2, 000 pathways and about 10, 000 annotated proteins for the *Homo sapiens*. Reactome allows to browse pathways through the graphical web interface, as well as download the data in different formats comprising Systems biology markup language (SBML) Level 2, BioPAX Level 2 and Level 3 and other graphical formats for local analysis.

The DEGs preprocessing method

The proposed statistical network pre-processing methodology automatically determines significant *DEGs* to use in PEA analysis in order to obtain more relevant biological pathways with respect to the condition under investigation. The proposed method consists of the following steps:

1. *Similarity matrix computation. Similarity matrices* As a preliminary step, the input DEG list is filtered by using the criteria introduced in Sect. 2.3. The remaining, DEGs are automatically grouped into up- and down- regulated genes, to yield the related up- and down- regulated similarity matrices' *UpSM* and *DownSM*. The Kruskal–Wallis test [13] is used to compute both the *UpSM* and *DownSM* matrices by using

the grouped DEGs. Kruskal–Wallis test is a non-parametric version of a parametric one-way ANOVA with the data substituted by their scores [32]. It works on two or more independent populations whose dimensions, e.g., the number of elements in each population, can be different. Equation 1 shows the formal definition of the Kruskal–Wallis test.

$$K = (N-1)\frac{\sum_{i=1}^{g} n_i(\bar{s}_{i\cdot} - \bar{s})^2}{\sum_{i=1}^{g}\sum_{j=1}^{n_i}(s_{ij} - \bar{s})^2}, \quad \bar{s}_{i\cdot} = \frac{\sum_{j=1}^{n_i} s_{ij}}{n_i}; \quad \bar{s} = \frac{1}{2}(N+1) \tag{1}$$

In Eq. 1, N is the total number of elements, g is the groups number, n_i is the number of elements in group i, s_{ij} is the observation of element j from group i, $\bar{s}_{i\cdot}$ is the average similarity of all elements in group i, and $\bar{s}$ is the average of all the s_{ij} similarities. A generic *SM*'s cell (i, j) contains the value of the similarity obtained comparing two genes by means of the Kruskal–Wallis test. The Kruskal–Wallis test assesses if two genes are correlated. In this manner, the Kruskal–Wallis test compares the genes with the aim to elucidate statistically similar behaviours among them. Other models [33, 34] used the Wilcoxon test [35, 36] to compute the *SM*. To compute the *SM*, the Wilcoxon test requires that the number of elements in each populations, e.g., the assessed expression levels n for each gene is $n \geq 20$. Conversely, the Kruskal–Wallis test works on 2 or more independent populations which may have different number of elements. A lower score (i.e, p value) implies that two genes are different according to the *logFC*. Otherwise, a higher score implies that genes show a similarity. The threshold was set to 0.005. Hence, the *SM*'s will contain only p values ≥ 0.005, 0 otherwise.

2. *Converting similarity matrix to network* The *UpSM* and *DownSM* matrices are converted to networks N_{up} and N_{down}, where nodes are the genes and the edges connect them when the similarity value among two genes in the (*i-th*, *j-th*) cell exceeds the similarity threshold (e.g., p values ≥ 0.005). The *Closeness Centrality (CC)* measure determines from the N_{up} and N_{down} networks the genes to include in the respective *Essential Gene sets* EG_{up} and EG_{down}. The CC is tightly related to the notion of distance between nodes and indicates how close a node is to all other nodes in the network. It is calculated as the average of the shortest path length from a node to every other node in the network. Only the nodes with CC values less than or equal to the computed average CC (e.g., $CC(n_i) \leq CC_{avg}(N)$) were included in the respective EG_{up} or EG_{down} gene sets.
3. *Improving genes relevance* The essential genes in both *EG* sets identified in the previous step are mapped onto the protein-protein interaction (PPI) network obtained from the *IID* [16] database. All DEGs that do not exist in the *iidNetwork* i.e., N_{iid} are filtered out. For each mapped gene from the respective gene sets, EG_{up} and EG_{down}, we computed from the N_{iid}, the neighborhood with a radius equal to 1, yielding respectively the *up-regulated gene community* GC_{up}, and the *down-regulated gene community* GC_{down}. In this way, it is possible to identify similar genes, and similar genes tend to interact among them to complete biological tasks. Finally, from both neighborhoods, all the nodes with a *Bottleneck* value greater than the *average Bottleneck* value, were selected to compute PEA.

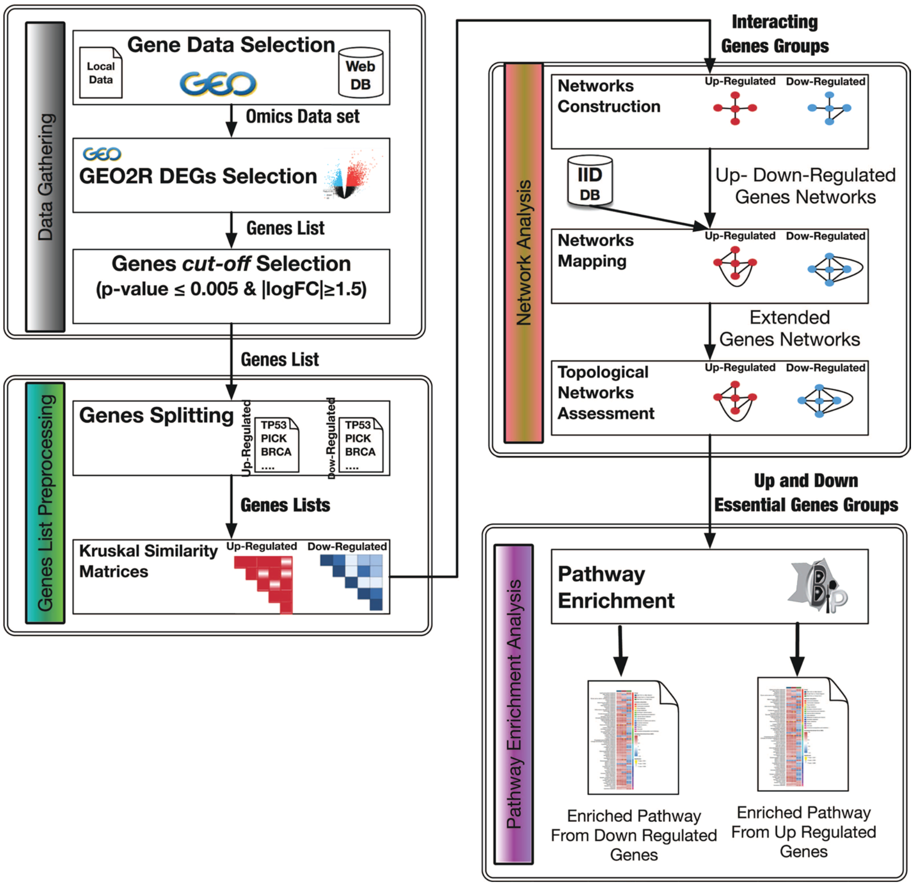

Fig. 3 The main steps of the proposed method

Table 2 *Tot#Genes* refers to the total number of filtered differential expressed gene from GEO2R tool, after removing duplicate genes

Data sets name	*Tot#Genes*	$\#Do_{Reg}G$	$\#Up_{Reg}G$	$\#G$
GSE1297	22,283	6	24	30
GSE5281	54,675	185	122	307
GSE16759	54,675	2	5	7
GSE9476	22,283	350	379	729
GSE14924	22,283	124	440	564
GSE24739	54,613	105	69	174

$\#Do_{Reg}G$ is the number of down regulated genes obtained employing the proposed methodology. $\#Up_{Reg}G$ indicates the number of up regulated genes obtained employing the proposed methodology. Finally, $\#G$ is the total number of extracted genes for each data set, holding all criteria

Figure 3 shows the main steps of the proposed method.

Results

The six data sets obtained from GEO and analyzed through the GEO2R framework were used as benchmark data. Table 2 contains the information the preprocessing of the original data sets of Table 1 referring to acute myeloid leukemia and Alzheimer's disease

Analysis of all data sets began by filtering out genes using the threshold values defined Sect. 2.3. Next, the genes holding the threshold criteria were split into up and down-regulated gene sets, EG_{up} and EG_{down}. It is worth noting that up- and down-regulated gene sets do not overlap, e.g., $\{EG_{up} \cap EG_{down} = \emptyset\}$. Third, genes in both gene sets were analyzed using the Kruskal–Wallis test to identify genes with the same behavior to compute similarity matrices, *UpSM* and *DownSM*. The *UpSM* and *DownSM* matrices were converted into networks N_{up} and N_{down}, and CC is used to determine the genes to include in the respective *Essential Gene sets*, EG_{up} and EG_{down}. Only the nodes with CC values less or equal to the computed average CC (e.g., $CC(n_i) \leq CC_{avg}(N)$) were included in the respective EG_{up} or EG_{down} gene set.

All six data sets contain duplicate genes that must be removedso they do not compromise the analysis. In many investigations, researchers manually remove the duplicate genes through some customized scripts. This long, tedious, and error-prone process introduce biases, and potentially entangle the PEA results. To overcome this limitation, the proposed preprocessing methodology automatically removes the duplicate genes, and retaining unique genes for further analysis (see Table 2*Tot#Genes* columns).

To improve both relevance and specificity of the selected genes within both the EG_{up} and EG_{down} regulated gene sets, each gene was matched to the *IID* network obtained from, filtering out all the unmatched genes. After the mapping on the *IID*, we computed the neighborhoods with distance 1. In this way, the neighborhoods allow identification of new relevant genes exploiting topological information, as reported in Table 3. In order to limit the number of potential genes to use in PEA, all the DEGs holding the following threshold: $Bottleneck(g_i) \geq AvgBottleneck$ were selected. The number of relevant selected DEGs is summarized in Table 4.

The descriptive statistics of the six data sets are summarized in Figs. 4 and 5 the number of relevant selected DEGs are listed in Table 4.

It is important to mention that the proposed methods limit the number of DEGs in the PEA to elucidate more relevant biological pathways to the condition under investigation. Tables 6, 7, 10 and 11 report the respective enriched pathways for each gene group.

Tables 6, 7, 10 and 11 report the enriched pathways for each genes' group.

It is worth noting that the gene groups classification e.g., down- and up- regulated gene groups, provides a two advantages. First, it limits the number of possible enriched pathways by employing fewer more specific genes. Second, it highlights which genes

Table 3 $\#U_{DEGs}$ refers to the total number of filtered out up-regulated DEGs, $\#D_{Reg}G$ is the number of down-regulated genes in each data set

Dataset	$\#U_{DEGs}$	$\#D_{DEGs}$	$\#R_{UDEGs}Ext$	$\#R_{DDEGs}Ext$	$\%Sel_{UDEGs}$	$\%Sel_{DDEGs}$
GSE1297	24	6	2119	1047	0.1077	0.0469
GSE9476	379	350	10,132	11,908	1.7008	0.5343
GSE24739	69	105	4785	5939	0.1263	0.1087
GSE5281	122	185	9113	7013	0.2231	0.1282
GSE16759	5	2	1091	64	0.0091	0.0011
GSE14924	440	124	11,429	4473	0.80473	0.0818

$\#R_{UDEGs}Ext$ indicates the number of up-regulated genes for each gene that successfully mapped onto the IID network. $\#R_{DDEGs}Ext$ represents the down regulated genes for each gene that successfully mapped onto the IID network. Finally, $\%Sel_{DDEGs}$ indicates the percentage of detected relevant genes with respect to the total number of available genes in each data set

Table 4 In the table, *#R_{UDEGs}Ext* indicates the number of up regulated genes for each gene for it which was possible to obtain a mapping and subsequent extension in *IID* network

Dataset name	#RUDEGsExt	#RDDEGsExt	BS_{UDEGs}	BS_{DDEGs}	$\%BS_{UDEGs}$	$\%BS_{DDEGs}$
GSE1297	2119	1047	37	11	0.0175	0.0105
GSE9476	10,132	11,908	331	360	0.0327	0.0302
GSE24739	4785	5939	77	115	0.0161	0.0194
GSE5281	9113	7013	118	197	0.0129	0.0281
GSE16759	1091	64	9	2	0.0082	0.0313
GSE14924	11,429	4473	368	133	0.0322	0.0297

#R_{DDEGs}Ext represents the down regulated genes for each gene for which it was possible to obtain a mapping and subsequent extension in *IID* network. BS_{DDEGs} indicates the percentage of detected relevant gene with respect to the total number of available genes in each data set

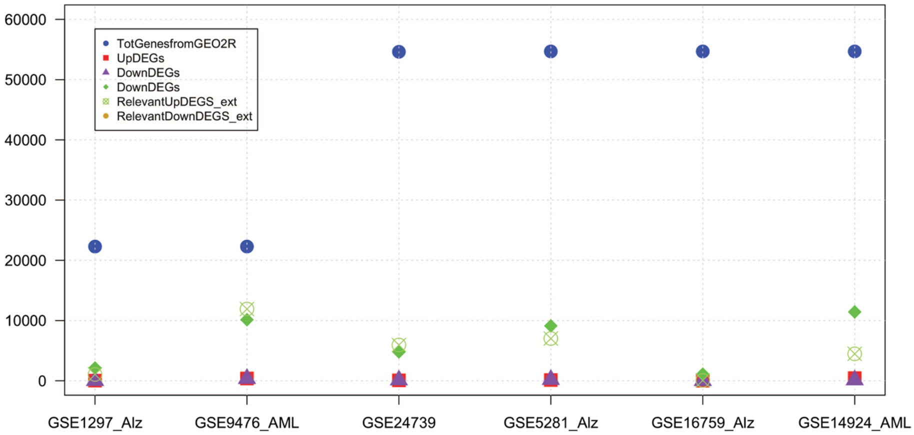

Fig. 4 The figure shows the descriptive statistics of six data sets about the Top Gene from GEO2R, Up DEGs, Down DEGs, Relevant Up DEGS, Relevant Down DEGS

affect the underlining biological functions. This aspect is more evident when analysing the enrichment results in Tables 5 (AML) and 9 (AD) from the ungrouped DEGs. In fact, Table 5 contains the same enriched pathways (in a different order) with respect to Tables 6 and 7, and Table 9 contains the same enriched pathways (in a different order) with respect to Tables 10 and 11, further complicating determination of which DEGs underlie biological mechanisms and functions.

Analysing the data contained in Tables 6, 7, 10, and 11, it should be noted that using of down- and up-regulated gene groups provides a better indication of which group of genes are affecting the pathways responsible for the current phenotype with respect to all unclassified genes.

Comparing the enrichment results obtained using the DEGs without preprocessing and the enrichment results obtained by employing the computed essential DEGs, it supports the proposed approach's effectiveness in identifying crucial DEGs related to the phenotype under investigation. The effectiveness of the proposed method in selecting proper DEGs is indicated by the obtained from the enrichment's statistical function, e.g., *Hyper-geometric* function *p* values for each enriched pathway. In fact, higher *p* values refer to more specific biological pathways for the condition under investigation. For example, in Table 6, the first enriched pathway is *Signaling by Interleukins*, a

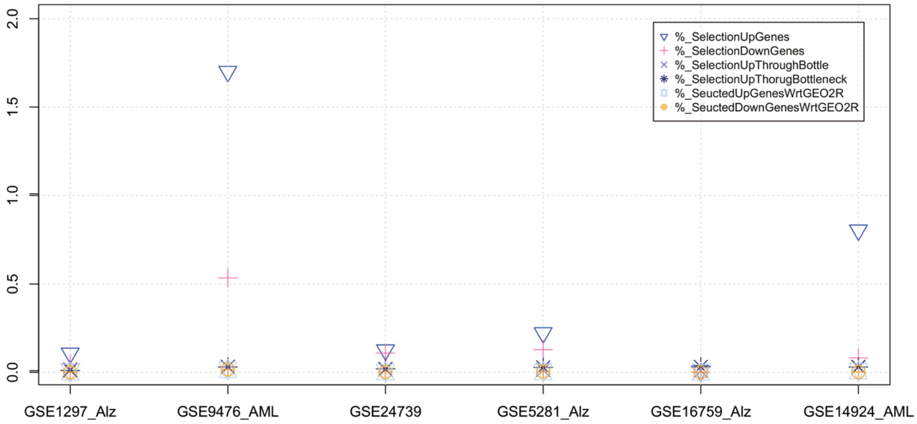

Fig. 5 The figure shows the descriptive statistics of six data sets about Selection Up Genes, Selection Down Genes, Selection Up Through Bottle, Selection Up Through Bottleneck Seucted Up Genes Wrt GEO2R, Seucted Down Genes Wrt GEO2R

Table 5 The first 10 enriched pathways using the whole list of relevant DEGs obtained from the AML GSE24739 data set without using DEG group classification

PathwayName	Pvalue	FDRC	BonfC	$\lvert lg_2(P_{value})\rvert$	$\lvert lg_2(FDR(P_{value})\rvert$	$\lvert lg_2 Bonf(P_{value})\rvert$
(1) Signaling pathways	5.29E−13	4.87E−10	4.87E−10	30.94	30.94	30.94
(2) Hemostasis	5.36E−13	2.47E−10	4.93E−10	31.92	30.92	30.92
(3) Developmental biology	9.86E−10	3.02E−07	9.07E−07	21.66	20.07	20.07
(4) Signaling by interleukins	1.37E−09	3.15E−07	1.26E−06	21.60	19.60	19.60
(5) Cell surface interactions at the vascular wall	1.94E−09	3.56E−07	1.78E−06	21.42	19.10	19.10
(6) Cytokine signaling in immune system	5.40E−09	8.28E−07	4.97E−06	20.20	17.62	17.62
(7) Muscle contraction	9.65E−09	1.27E−06	8.88E−06	19.59	16.78	16.78
(8) Signaling by GPCR	1.83E−08	2.10E−06	1.68E−05	18.86	15.86	15.86
(9) GPCR downstream signalling	2.09E−08	2.14E−06	1.93E−05	18.83	15.66	15.66
(10) Cardiac conduction	2.12E−08	1.95E−06	1.95E−05	18.97	15.65	15.65

The *FDRC* indicates the corrected *p* value using *FDR* statistical corrector. The *BonfC* represents the corrected *p* value using *Bonferroni* statistical corrector. Finally, the last three columns contain the $\lvert lg_2(\cdot)\rvert$ of each *p* value, for ease comparison

very well-known pathway involved in the generation of AML [37]. In Table 7, the first enriched pathway is *Hemostasis*, a well-known pathway involved in the development and progression of AML [38]. Whereas in Table 8, all ten enriched pathways are generic pathways that do not provide any additional information to the researchers about the DEGs and their involvement in AML. This shows the DEGs role in the PEA. In this enrichment, the *Signaling by Interleukins* pathway is shifted to the 50*th* position in the enrichment ranking, while the *Hemostasis* pathway has been moved to 30*th* position. These results highlight the importance of the chosen DEGs for the PEA. Many DEGs provides many general enriched pathways, challenging researchers to obtain new clues about the relationship between DEGs and biological functions. Finally, it is worthy noting that the use of gene groups along with the identification of essential genes, makes it straightforward to understand which genes are responsible for affecting the underlying biological functions, by linking DEGs with more specific pathways. More information about the other investigated GSE data sets are reported in Additional file 1.

Discussion

In order to place long lists of differential genes into the context of biological processes and pathways, enrichment pathway analysis is widely used. In this work we proposed a new statistical network pre-processing method to improve the relevance and significance of the DEGs or SNPs of interest when performing PEA attempting to incorporate pathway topology information into the analysis.

Although PEA is an essential part of DEG data analysis, the absence of suitable standards force validation of enrichment results. The following is a literature review for the results from our new statistical network pre-processing methodology.

Analysis of the enriched pathways using the up-regulated relevant DEGs obtained from the AML data set with identifier GSE24739, indicates that the *Signaling by Interleukins* pathway can promote the generation of AML as reported in [37]. *Hemostasis* is a well-known pathway involved in the development and progression of AML [38]. Since both up- and down-regulated gene sets are involved in the *Hemostasis* pathway, all the enriched pathways capture this information. On the other hand, analysis that does not incorporate gene groups, makes it difficult to understand which group of genes are affecting the *Hemostasis* pathway. The third enriched pathway in Table 5 which does not use the gene, groups is the *Developmental Biology* pathway and its relationship with AML as described in [39]. However, the third enriched pathway in Table 7 is the *Muscle contraction* pathway, whose role in AML is explained in [40].

The role of the *signaling pathways* family in the progression and developing of AML is well-known in the literature. In fact, the enriched pathways in Table 6 highlight this peculiarity of enriching the following signalling pathways: *Cytokine Signaling in Immune system*[41], *POU5F1 (OCT4), SOX2, NANOG activate genes related to proliferation* [42], *Signaling by EGFR* [43], and *Signaling Pathways* [44]. The last enriched pathway in Table 6 is *Transcriptional regulation of pluripotent stem cells* whose involvement in AML is described in [45].

Analysis of the enriched pathways in Table 7 reveals the link between the *Metabolism of proteins* and AML as described in [46]. The role of *Amino acid and derivative metabolism* in developing of AML is discussed in [47]. The relation between AML and *Cell Cycle* and *Cell Cycle Mitotic* are described in [48]. In [49] the role *Mitotic G1 phase and G1/S transition* pathway in AML is introduced. In [50] it is clarified that AML can cause *Cardiac conduction* abnormalities in the elderly. Finally, the connection between the *Nervous system development* pathway and AML is documented in [51].

As proof of concept, Table 8 shows the first 10 enriched obtained using all the genes identifiers within the AML GSE24739 data set. Only one of the enriched pathways seems to have a connection with AML, the *Cellular Senescence* pathway [52]. To the best of our knowledge, we were unable to find any connection between the remaining nine enriched pathways in Table 8 and AML. This shows that using many less specific DEGs provides more enriched pathways but they are disconnected from the biological context of reference.

Table 6 The first 10 enriched pathways using the list of relevant down regulated DEGs obtained from the AML GSE24739 data set

PathwayName	Pvalue	FDRC	BonfC	$\|lg_2(P_{value})\|$	$\|lg_2(FDR(P_{value})\|$	$\|lg_2Bonf(P_{value})\|$
(1) Signaling by interleukins	8.61E−15	8.21E−12	8.21E−12	46.72	36.83	36.83
(2) Adaptive immune system	1.66E−14	7.92E−12	1.58E−11	45.78	36.88	35.88
(3) Signaling pathways	3.88E−14	1.23E−11	3.70E−11	44.55	36.24	34.65
(4) Cell surface interactions at the vascular wall	1.44E−13	3.44E−11	1.38E−10	42.65	34.76	32.76
(5) Cytokine signaling in immune system	1.48E−13	2.81E−11	1.41E−10	42.62	35.05	32.73
(6) POU5F1 (OCT4), SOX2, NANOG activate genes related to proliferation	4.47E−13	7.11E−11	4.27E−10	41.02	33.71	31.13
(7) Signaling by EGFR	4.47E−13	6.10E−11	4.27E−10	41.02	33.93	31.13
(8) Developmental biology	5.86E−13	6.99E−11	5.59E−10	40.63	33.74	30.74
(9) Hemostasis	5.88E−13	6.24E−11	5.61E−10	40.63	33.90	30.73
(10) Transcriptional regulation of pluripotent stem cells	6.67E−13	6.36E−11	6.36E−10	40.45	33.87	30.55

The *FDRC* indicates the corrected *p* value using *FDR* statistical corrector. The *BonfC* represents the corrected *p* value using *Bonferroni* statistical corrector. Finally, the last three columns contain the $|lg_2(\cdot)|$ of each *p* value, for ease comparison

Table 7 The first 10 enriched pathways using the list of relevant up regulated DEGs obtained from the AML GSE24739 data set

PathwayName	Pvalue	FDRC	BonfC	$\|lg_2(P_{value})\|$	$\|lg_2(FDR(P_{value})\|$	$\|lg_2Bonf(P_{value})\|$
(1) Hemostasis	1.82E−11	1.60E−08	1.60E−08	35.68	25.89	25.89
(2) Signaling pathways	4.32E−10	1.91E−07	3.81E−07	31.11	22.32	21.32
(3) Metabolism of proteins	4.09E−08	1.20E−05	3.61E−05	24.54	16.34	14.76
(4) Muscle contraction	4.13E−08	9.11E−06	3.64E−05	24.53	16.74	14.74
(5) Cell cycle, mitotic	1.18E−07	2.09E−05	1.04E−04	23.01	15.55	13.23
(6) Mitotic G1 phase and G1/S transition	1.18E−07	1.74E−05	1.04E−04	23.01	15.81	13.23
(7) Cardiac conduction	1.55E−07	1.95E−05	1.36E−04	22.62	15.65	12.84
(8) Cell Cycle	1.78E−07	1.96E−05	1.57E−04	22.42	15.64	12.64
(9) Nervous system development	2.29E−07	2.24E−05	2.02E−04	22.06	15.45	12.28
(10) Amino acid and derivative metabolism	2.88E−07	2.54E−05	2.54E−04	21.73	15.26	11.94

The *FDRC* indicates the corrected *p* value using *FDR* statistical corrector. The *BonfC* represents the corrected *p* value using *Bonferroni* statistical corrector. Finally, the last three columns contain the $|lg_2(\cdot)|$ of each *p* value, for ease comparison

Thus, it is worth noting that using the improved list of genes provides more relevant enriched pathways as demonstrated in Tables 5, 6, and 7, where all the enriched pathways have a connection with AML. Indeed, using a generic list of genes, the ratio between enriched pathways with the biological context of reference, e.g., AML, drops to 10%.

The connection among the first three enriched pathways in Table 9 and Alzheimer's are the following. The association between the *Post-translational protein modification* pathway and AD is reported in [53]. The relationship between the *Metabolism of proteins* pathway and AD is provided in [54]. While in [55], the authors describe the role of the *Neurophilin interactions with VEGF and VEGFR* pathway and AD.

Table 8 The first 10 enriched pathways using all genes obtained from GEO2R enclosed in the AML GSE24739 data set, without using the proposed pre-processing method

PathwayName	Pvalue	FDRC	BonfC	$\lvert lg_2(P_{value})\rvert$	$\lvert lg_2(FDR(P_{value})\rvert$	$\lvert lg_2Bonf(P_{value})\rvert$
(1) Signaling pathways	7.67E−303	3.01E−301	1.92E−299	1003.606	998.313	992.313
(2) Generic transcription pathway	7.67E−303	3.01E−301	1.92E−299	1003.606	998.313	992.313
(3) Gene expression (transcription)	7.67E−303	3.01E−301	1.92E−299	1003.606	998.313	992.313
(4) SLC-mediated transmembrane transport	7.67E−303	3.01E−301	1.92E−299	1003.606	998.313	992.313
(5) Cellular responses to stimuli	7.67E−303	3.01E−301	1.92E−299	1003.606	998.313	992.313
(6) Cellular responses to stress	7.67E−303	3.01E−301	1.92E−299	1003.606	998.313	992.313
(7) Cellular senescence	7.67E−303	3.01E−301	1.92E−299	1003.606	998.313	992.313
(8) DNA damage-telomere stress induced senescence	7.67E−303	3.01E−301	1.92E−299	1003.606	998.313	992.313
(9) Carbohydrate metabolism	7.67E−303	3.01E−301	1.92E−299	1003.606	998.313	992.313
(10) Metabolism	7.67E−303	3.01E−301	1.92E−299	1003.606	998.313	992.313

.The *FDRC* indicates the corrected *p* value using *FDR* statistical corrector. The *BonfC* represents the corrected *p* value using *Bonferroni* statistical corrector. Finally, the last three columns contain the $|lg_2(\cdot)|$ of each *p* value, for ease comparison

Table 9 The first 10 enriched pathways using the whole list of relevant DEGs obtained from the Alzheimer GSE16759 data set without using genes' group classification

PathwayName	Pvalue	FDRC	BonfC	$\lvert lg_2(P_{value})\rvert$	$\lvert lg_2(FDR(P_{value})\rvert$	$\lvert lg_2Bonf(P_{value})\rvert$
(1) Post-translational protein modification	8.20E-05	0.0390	0.0390	4.68	4.68	4.68
(2) Metabolism of proteins	1.67E−04	0.0396	0.0793	4.66	3.66	3.66
(3) Neurophilin interactions with VEGF and VEGFR	7.62E−04	0.1208	0.3625	3.05	1.46	1.46
(4) Disease	0.0013	0.1574	0.6299	2.67	0.67	0.67
(5) VEGF binds to VEGFR leading to receptor dimerization	0.0028	0.2679	1	1.90	0.00	0.00
(6) VEGF ligand–receptor interactions	0.0028	0.2233	1	2.16	0.00	0.00
(7) Signaling by VEGF	0.0029	0.1976	1	2.34	0.00	0.00
(8) Synthesis of 5-eicosatetraenoic acids	0.0047	0.2827	1	1.82	0.00	0.00
(9) Signaling by receptor tyrosine Kinases	0.0047	0.2534	1	1.98	0.00	0.00
(10) Synthesis of leukotrienes (LT) and Eoxins (EX)	0.0049	0.2352	1	2.09	0.00	0.00

The *FDRC* indicates the corrected *p* value using *FDR* statistical corrector. The *BonfC* represents the corrected *p* value using *Bonferroni* statistical corrector. Finally, the last three columns contain the $|lg_2(\cdot)|$ of each *p* value, for ease comparison

Searching the scientific literature, we find the following connection between the first three enriched pathways in Table 10 and AD. The role of the *Synthesis of 5-eicosatetraenoic acids* and *Synthesis of Leukotrienes (LT) and Eoxins (EX)* pathways with the AD is reported in [56] and [57], respectively. The implication of *HIV Transcription Initiation* pathway in AD is explained in [58].

Table 10 The 5 enriched pathways using the list of relevant down regulated DEGs obtained from the Alzheimer GSE16759 data set

PathwayName	Pvalue	FDRC	BonfC	$\lvert lg_2(P_{value})\rvert$	$\lvert lg_2(FDR(P_{value})\rvert$	$\lvert lg_2 Bonf(P_{value})\rvert$
1) Synthesis of 5-eicosatetraenoic acids	0.002	0.013	0.735	6.23	0.45	0.45
2) Synthesis of leukotrienes (LT) and eoxins (EX)	0.002	0.014	0.764	6.20	0.39	0.39
3) HIV transcription initiation	0.003	0.026	1.000	5.28	0.00	0.00
4) RNA Polymerase II HIV promoter escape	0.003	0.025	1.000	5.30	0.00	0.00
5) Transcription of the HIV genome	0.005	0.037	1.000	4.76	0.00	0.00

The *FDRC* indicates the corrected *p* value using *FDR* statistical corrector. The *BonfC* represents the corrected *p* value using *Bonferroni* statistical corrector. Finally, the last three columns contain the $\lvert lg_2(\cdot)\rvert$ of each *p* value, for ease comparison

Table 11 The 6 enriched pathways using the list of relevant up regulated DEGs obtained from the Alzheimer GSE16759 data set

PathwayName	Pvalue	FDRC	BonfC	$\lvert lg_2(P_{value})\rvert$	$\lvert lg_2(FDR(P_{value})\rvert$	$\lvert lg_2 Bonf(P_{value})\rvert$
1) PERK regulates gene expression	4.81E−05	0.02	0.02	14.34	5.42	5.42
2) Signaling pathways	1.10E−04	0.03	0.05	13.15	5.23	4.23
3) Unfolded protein response (UPR)	3.33E−04	0.05	0.16	11.55	4.21	2.63
4) Neurophilin interactions with VEGF and VEGFR	0.0011	0.14	0.56	9.77	2.85	0.85
5) ATF6 (ATF6-alpha) activates chaperone genes	0.0028	0.28	1.00	8.45	1.85	0.00
6) ATF6 (ATF6-alpha) activates chaperones	0.0034	0.28	1.00	8.19	1.85	0.00

The *FDRC* indicates the corrected *p* value using *FDR* statistical corrector. The *BonfC* represents the corrected *p* value using *Bonferroni* statistical corrector. Finally, the last three columns contain the $\lvert lg_2(\cdot)\rvert$ of each *p* value, for ease comparison

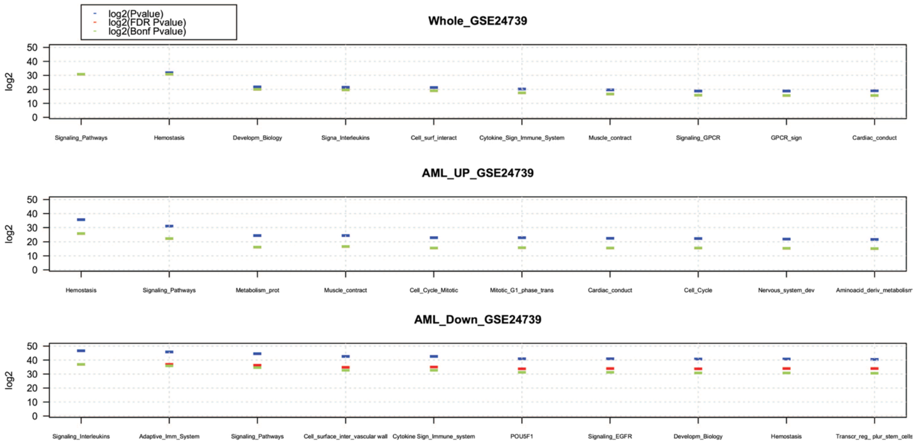

Fig. 6 The first 10 enriched pathways using the whole list of relevant DEGs, up-regulated DEGs, and down regulated DEGs obtained from the AML GSE24739. The blue mark represents the $log_2(pvalues)$, the red mark represents $log_2(FDRpvalues)$, and the green mark represents $log_2(Bonferronipvalues)$

Table 11 lists the enriched pathways using the up-regulated relevant DEGs obtained from the *GSE16759* data set. In [59], the authors describe the implication of the *PERK regulates gene expression* pathway with AD. In [60] the authors clarify

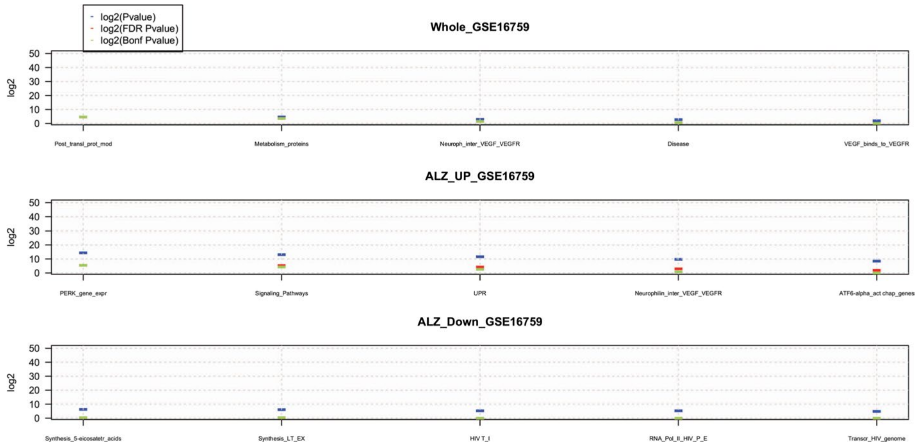

Fig. 7 The first 10 enriched pathways using the whole list of relevant DEGs, up-regulated DEGs, down-regulated DEGs obtained from the AD GSE16759. The blue mark represents the $log_2(pvalues)$, the red mark represents $log_2(FDRpvalues)$, and the green mark represents $log_2(Bonferronipvalues)$

the implication of *Signaling Pathways* in the development of many human diseases including AD disease. The authors in [61] characterize the association between *unfolded protein response (UPR)* with onset of familial Alzheimer's disease (Fig. 5).

Comparing the enriched pathways in Figs. 6 and 7, further highlights the benefits of the proposed approach, revealing more specific pathways affecting the biological functions and mechanisms.

Finally, we performed PEA using the grouped and ungrouped DEG sets to assess the effectiveness of the proposed DGE preprocessing and selection method. Analyzing the obtained pathway enrichment results using both data sets highlight that DEGs critically impact PEA, since employing an ungrouped DEG set can lead to poor enrichment results. Also, the first ranked enriched pathway using grouped DEGs is related to the condition under investigation, which may induce new biological discoveries and simplify research.

Conclusions

In this work, we proposed a new statistical network pre-processing approach to identify relevant DEGs that can improve PEA results, helping researchers identify the affected underlying biological functions and processes. The proposed method provides a twofold improvement. First, network analysis yields fewer DEGs, choosing only relevant DEGs that directly involved with the condition under investigation. Second, the detected DEGs improve the enriched pathways' statistical significance over a more general list of genes. As a drawback, the number of enriched pathways is still too large; thus, future research should be aimed at developing a method to further reduce the number of enriched pathways.

Abbreviations

AD	Alzheimer's disease
AML	Acute myeloid leukemia
API	Application program interface
BioPAX	Biological pathway exchange

BiP	BioPAX-parser
CC	Closeness centrality
CLL	Chronic lymphocytic leukemia
DEG	Differential expressed gene
DNA	Deoxyribonucleic acid
EG	Essential gene sets
FDR	False discovery rate
GEO	Gene Expression Omnibus
GWAS	Genome-Wide Association Studies
HT	High throughput
IID	Integrated interactions database
logFC	Logarithm of fold change
KEGG	Kyoto Encyclopaedia of Gene and Genome
miRNA	micro RNA
mRNA	messenger RNA
NGS	Next generation sequencing
ORA	Over represented analysis
PEA	Pathway enrichment analysis
RNA	Ribonucleic acid
SBML	Systems biology markup language
SM	Similarity matrix
SNPs	Single nucleotide polymorphism

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-022-04936-z.

Additional file 1: Provides detailed information about the other investigated GSE data sets.

Acknowledgements

This work has been partially founded by the Data Analytics Research Center, Department of Medical and Surgical Sciences, University of Catanzaro.

About thissupplement section

This article has been published as part of BMC Bioinformatics Volume 23 Supplement 6, 2022 Selected articles from the 17th International Conference on Computational Intelligence Methods for Bioinformatics and Biostatistics (CIBB 2021). The full contents of the supplement are available online at https://bmcbioinformatics.biomedcentral.com/articles/supplements/volume-23-supplement-6.

Author contributions

GA and MM contributed to the conceptual idea of the study. GA directed the writing of the manuscript. GA conceived and designed the experiments. GA and MM performed the experimental work and analyzed the results. GA, MM and MC wrote the paper. All authors read and approved the final manuscript.

Funding

Not applicable.

Availability of data and materials

The data sets used and analyzed in this study are freely available in GEO database. GEO data set links: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE1297; https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE5281; https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE16759; https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE9476; https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE14924; https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE24739; **Reactome** database link: https://reactome.org/download-data; **KEGG** database link:https://www.kegg.jp; **BiP** software tool link: https://gitlab.com/giuseppeagapito/bip; **GEO2R** software tool link: https://www.ncbi.nlm.nih.gov/geo/geo2r/. Also, all the links to the data sets and materials have been provided through the manuscript.

Declarations

Ethics approval and consent to participate

No ethics approval was required for the study.

Consent for publication

All the authors contributed to manuscript read, and approved the submitted version.

Competing interests

The authors declare that they have no competing interests.

Received: 15 September 2022 Accepted: 16 September 2022
Published online: 27 September 2022

References

1. Heller MJ. Dna microarray technology: devices, systems, and applications. Annu Rev Biomed Eng. 2002;4(1):129–53.
2. Saiki RK, Gelfand DH, Stoffel S, Scharf SJ, Higuchi R, Horn GT, Mullis KB, Erlich HA. Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. Science. 1988;239(4839):487–91.
3. Sanger F, Nicklen S, Coulson AR. Dna sequencing with chain-terminating inhibitors. Proc Natl Acad Sci. 1977;74(12):5463–7.
4. Bier FF, von Nickisch-Rosenegk M, Ehrentreich-Foerster E, Reiss E, Henkel J, Strehlow R, Andresen D. DNA microarrays. In: Renneberg R, Lisdat F, editors. Biosensing for the 21st century. Berlin: Springer; 2007. p. 433–53.
5. Guzzi PH, Agapito G, Di Martino MT, Arbitrio M, Tassone P, Tagliaferri P, Cannataro M. DMET-analyzer: automatic analysis of affymetrix DMET data. BMC Bioinform. 2012;13(1):1–10.
6. Agapito G, Cannataro M, Guzzi PH, Marozzo F, Talia D, Trunfio P. Cloud4snp: distributed analysis of SNP microarray data on the cloud. In: Proceedings of the international conference on bioinformatics, computational biology and biomedical informatics; 2013, p. 468–75.
7. Guzzi PH, Agapito G, Cannataro M. coreSNP: parallel processing of microarray data. IEEE Trans Comput. 2013;63(12):2961–74.
8. Agapito G, Milano M, Guzzi PH, Cannataro M. Extracting cross-ontology weighted association rules from gene ontology annotations. IEEE/ACM Trans Comput Biol Bioinf. 2015;13(2):197–208.
9. Agapito G, Guzzi PH, Cannataro M. DMET-miner: discovery of association rules from pharmacogenomic data. J Biomed Inform. 2015;56:273–83.
10. Agapito G, Guzzi PH, Cannataro M. Parallel and distributed association rule mining in life science: a novel parallel algorithm to mine genomics data. Inf Sci. 2021;575:747–61.
11. Agapito G, Guzzi PH, Cannataro M. Parallel extraction of association rules from genomics data. Appl Math Comput. 2019;350:434–46.
12. Milano M. Using gene ontology to annotate and prioritize microarray data. Berlin: Springer; 2022. p. 273–87.
13. Kruskal WH, Wallis WA. Use of ranks in one-criterion variance analysis. J Am Stat Assoc. 1952;47(260):583–621.
14. Bompais M, Ameur H, Pastor D, Dupraz E. The p-value as a new similarity function for spectral clustering in sensor networks. In: 2018 IEEE statistical signal processing workshop (SSP). IEEE; 2018. p. 95–9.
15. Bailey TL, Gribskov M. Combining evidence using p-values: application to sequence homology searches. Bioinformatics (Oxford, England). 1998;14(1):48–54.
16. Kotlyar M, Pastrello C, Malik Z, Jurisica I. Iid 2018 update: context-specific physical protein-protein interactions in human, model organisms and domesticated species. Nucleic Acids Res. 2019;47(D1):581–9.
17. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, et al. NCBI GEO: archive for functional genomics data sets-update. Nucleic Acids Res. 2012;41(D1):991–5.
18. Blalock EM, Geddes JW, Chen KC, Porter NM, Markesbery WR, Landfield PW. Incipient Alzheimer's disease: microarray correlation analyses reveal major transcriptional and tumor suppressor responses. Proc Natl Acad Sci. 2004;101(7):2173–8.
19. Liang WS, Dunckley T, Beach TG, Grover A, Mastroeni D, Walker DG, Caselli RJ, Kukull WA, McKeel D, Morris JC, et al. Gene expression profiles in anatomically and functionally distinct regions of the normal aged human brain. Physiol Genom. 2007;28(3):311–22.
20. Liang WS, Reiman EM, Valla J, Dunckley T, Beach TG, Grover A, Niedzielko TL, Schneider LE, Mastroeni D, Caselli R, et al. Alzheimer's disease is associated with reduced expression of energy metabolism genes in posterior cingulate neurons. Proc Natl Acad Sci. 2008;105(11):4441–6.
21. Readhead B, Haure-Mirande J-V, Funk CC, Richards MA, Shannon P, Haroutunian V, Sano M, Liang WS, Beckmann ND, Price ND, et al. Multiscale analysis of independent Alzheimer's cohorts finds disruption of molecular, genetic, and clinical networks by human herpesvirus. Neuron. 2018;99(1):64–82.
22. Liang WS, Dunckley T, Beach TG, Grover A, Mastroeni D, Ramsey K, Caselli RJ, Kukull WA, McKeel D, Morris JC, et al. Altered neuronal gene expression in brain regions differentially affected by Alzheimer's disease: a reference data set. Physiol Genom. 2008;33(2):240–56.
23. Nunez-Iglesias J, Liu C-C, Morgan TE, Finch CE, Zhou XJ. Joint genome-wide profiling of miRNA and mRNA expression in Alzheimer's disease cortex reveals altered miRNA regulation. PLoS ONE. 2010;5(2):1–9. https://doi.org/10.1371/journal.pone.0008898.
24. Stirewalt DL, Meshinchi S, Kopecky KJ, Fan W, Pogosova-Agadjanyan EL, Engel JH, Cronk MR, Dorcy KS, McQuary AR, Hockenbery D, et al. Identification of genes with abnormal expression changes in acute myeloid leukemia. Genes Chromosom Cancer. 2008;47(1):8–20.
25. Le Dieu R, Taussig DC, Ramsay AG, Mitter R, Miraki-Moud F, Fatah R, Lee AM, Lister TA, Gribben JG. Peripheral blood T cells in acute myeloid leukemia (AML) patients at diagnosis have abnormal phenotype and genotype and form defective immune synapses with AML blasts. Blood J Am Soc Hematol. 2009;114(18):3909–16.
26. Affer M, Dao S, Liu C, Olshen A, Mo Q, Viale A, Lambek C, Marr T, Clarkson B. Gene expression differences between enriched normal and chronic myelogenous leukemia quiescent stem/progenitor cells and correlations with biological abnormalities. J Oncol. 2011;2011:798592.
27. Abraham SA, Hopcroft LE, Carrick E, Drotar ME, Dunn K, Williamson AJ, Korfi K, Baquero P, Park LE, Scott MT, et al. Dual targeting of p53 and c-MYC selectively eliminates leukaemic stem cells. Nature. 2016;534(7607):341–6.
28. Agapito G, Cannataro M. Using biopax-parser (BIP) to enrich lists of genes or proteins with pathway data. BMC Bioinform. 2021;22(13):1–35.
29. Agapito G, Cannataro M. Using biopax-parser (BIP) to annotate lists of biological entities with pathway data. In: International conference on conceptual modeling. Springer; 2020. p. 92–101.
30. Agapito G, Pastrello C, Guzzi PH, Jurisica I, Cannataro M. Biopax-parser: parsing and enrichment analysis of biopax pathways. Bioinformatics. 2020;36(15):4377–8.
31. Joshi-Tope G, Gillespie M, Vastrik I, D'Eustachio P, Schmidt E, de Bono B, Jassal B, Gopinath G, Wu G, Matthews L, et al. Reactome: a knowledgebase of biological pathways. Nucleic Acids Res. 2005;33(suppl–1):428–32.
32. Girden ER. ANOVA: repeated measures, vol. 84. Thousand Oaks: Sage; 1992.

33. Milano M, Zucco C, Cannataro M. Covid-19 community temporal visualizer: a new methodology for the network-based analysis and visualization of covid-19 data. Netw Model Anal Health Inform Bioinform. 2021;10(1):1–38.
34. Agapito G, Milano M, Cannataro M. A new parallel methodology for the network analysis of covid-19 data. In: Euro-Par 2020: parallel processing workshops. Nature Publishing Group; 2020. p. 333.
35. Wilcoxon F. Individual comparisons by ranking methods. Biom Bull. 1945;1(6):80–3.
36. Gehan EA. A generalized Wilcoxon test for comparing arbitrarily singly-censored samples. Biometrika. 1965;52(1–2):203–24.
37. Carey A, Eide CA, Newell L, Traer E, Medeiros BC, Pollyea DA, Deininger MW, Collins RH, Tyner JW, Druker BJ, et al. Identification of interleukin-1 by functional screening as a key mediator of cellular expansion and disease progression in acute myeloid leukemia. Cell Rep. 2017;18(13):3204–18.
38. Nadir Y, Katz T, Sarig G, Hoffman R, Oliven A, Rowe JM, Brenner B. Hemostatic balance on the surface of leukemic cells: the role of tissue factor and urokinase plasminogen activator receptor. Haematologica. 2005;90(11):1549–56.
39. Wang Y, Krivtsov AV, Sinha AU, North TE, Goessling W, Feng Z, Zon LI, Armstrong SA. The Wnt/β-catenin pathway is required for the development of leukemia stem cells in AML. Science. 2010;327(5973):1650–3.
40. Sun Y, Boyd K, Xu W, Ma J, Jackson CW, Fu A, Shillingford JM, Robinson GW, Hennighausen L, Hitzler JK, et al. Acute myeloid leukemia-associated mkl1 (mrtf-a) is a key regulator of mammary gland function. Mol Cell Biol. 2006;26(15):5809–26.
41. Camacho V, McClearn V, Patel S, Welner RS. Regulation of normal and leukemic stem cells through cytokine signaling and the microenvironment. Int J Hematol. 2017;105(5):566–77.
42. Picot T, Kesr S, Wu Y, Aanei CM, Flandrin-Gresta P, Tondeur S, Tavernier E, Wattel E, Guyotat D, Campos L. Potential role of oct4 in leukemogenesis. Stem Cells Dev. 2017;26(22):1637–47.
43. Almeida LYd, Rego EM. Is the EGFR pathway relevant for the pathogenesis but not for treatment of acute myeloid leukemia? J Cancer Metastasis Treat. 2021;7:57.
44. Rodrigues ACBdC, Costa RG, Silva SL, Dias IR, Dias RB, Bezerra DP. Cell signaling pathways as molecular targets to eliminate AML stem cells. Crit Rev Oncol/Hematol. 2021;160: 103277.
45. Yang J, Chai L, Fowles TC, Alipio Z, Xu D, Fink LM, Ward DC, Ma Y. Genome-wide analysis reveals sall4 to be a major regulator of pluripotency in murine-embryonic stem cells. Proc Natl Acad Sci. 2008;105(50):19756–61.
46. Grønningsæter IS, Reikvam H, Aasebø E, Bartaula-Brevik S, Tvedt TH, Bruserud Ø, Hatfield KJ. Targeting cellular metabolism in acute myeloid leukemia and the role of patient heterogeneity. Cells. 2020;9(5):1155.
47. Jones CL, Stevens BM, D'Alessandro A, Reisz JA, Culp-Hill R, Nemkov T, Pei S, Khan N, Adane B, Ye H, et al. Inhibition of amino acid metabolism selectively targets human leukemia stem cells. Cancer Cell. 2018;34(5):724–40.
48. Schnerch D, Yalcintepe J, Schmidts A, Becker H, Follo M, Engelhardt M, Wäsch R. Cell cycle control in acute myeloid leukemia. Am J Cancer Res. 2012;2(5):508.
49. Chae H-D, Sakamoto K. Replication factor c3 is a direct target of CREB, promotes g1/s transition of acute myeloid leukemia cells, and increases hematopoietic stem/progenitor cell self-renewal. Blood. 2013;122(21):3754.
50. Enjeti AK, D'Crus A, Melville K, Verrills NM, Rowlings P. A systematic evaluation of the safety and toxicity of fingolimod for its potential use in the treatment of acute myeloid leukaemia. Anticancer Drugs. 2016;27(6):560.
51. Wang Q, Stacy T, Binder M, Marin-Padilla M, Sharpe AH, Speck NA. Disruption of the Cbfa2 gene causes necrosis and hemorrhaging in the central nervous system and blocks definitive hematopoiesis. Proc Natl Acad Sci. 1996;93(8):3444–9.
52. Tang Y-L, Zhang C-G, Liu H, Zhou Y, Wang Y-P, Li Y, Han Y-J, Wang C-L. Ginsenoside rg1 inhibits cell proliferation and induces markers of cell senescence in CD34+ CD38- leukemia stem cells derived from kg1α acute myeloid leukemia cells by activating the sirtuin 1 (sirt1)/tuberous sclerosis complex 2 (tsc2) signaling pathway. Med Sci Monit Int Med J Exp Clin Res. 2020;26:918207–1.
53. Ramesh M, Gopinath P, Govindaraju T. Role of post-translational modifications in Alzheimer's disease. ChemBioChem. 2020;21(8):1052–79.
54. Kaddurah-Daouk R, Zhu H, Sharma S, Bogdanov M, Rozen S, Matson W, Oki N, Motsinger-Reif A, Churchill E, Lei Z, et al. Alterations in metabolic pathways and networks in Alzheimer's disease. Transl Psychiatry. 2013;3(4):244–244.
55. Zachary I. Neuroprotective role of vascular endothelial growth factor: signalling mechanisms, biological function, and therapeutic potential. Neurosignals. 2005;14(5):207–21.
56. Chen X-M, Feng M-J, Shen C-J, He B, Du X-F, Yu Y-B, Liu J, Chu H-M. A novel approach to select differential pathways associated with hypertrophic cardiomyopathy based on gene co-expression analysis. Mol Med Rep. 2017;16(1):773–7.
57. Magalhães KG, Luna-Gomes T, Mesquita-Santos F, Corrêa R, Assuncao LS, Atella GC, Weller PF, Bandeira-Melo C, Bozza PT. Schistosomal lipids activate human eosinophils via toll-like receptor 2 and pgd2 receptors: 15-lo role in cytokine secretion. Front Immunol. 2019;9:3161.
58. Minagar A, Shapshak P, Fujimura R, Ownby R, Heyes M, Eisdorfer C. The role of macrophage/microglia and astrocytes in the pathogenesis of three neurologic disorders: HIV-associated dementia, Alzheimer disease, and multiple sclerosis. J Neurol Sci. 2002;202(1–2):13–23.
59. Devi L, Ohno M. Perk mediates eif 2α phosphorylation responsible for bace1 elevation, CREB dysfunction and neurodegeneration in a mouse model of Alzheimer's disease. Neurobiol Aging. 2014;35(10):2272–81.
60. Mizuno S, Iijima R, Ogishima S, Kikuchi M, Matsuoka Y, Ghosh S, Miyamoto T, Miyashita A, Kuwano R, Tanaka H. Alzpathway: a comprehensive map of signaling pathways of Alzheimer's disease. BMC Syst Biol. 2012;6(1):1–10.
61. Hoozemans J, Veerhuis R, Van Haastert E, Rozemuller J, Baas F, Eikelenboom P, Scheper W. The unfolded protein response is activated in Alzheimer's disease. Acta Neuropathol. 2005;110(2):165–72.

Publisher's Note