2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) | 978-1-6654-6819-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/BIBM55620.2022.9995086

Identification of Cancer Biomarkers for Multi-class Diagnostics through Network Analysis of RNAseq Data of Tumor-Educated Platelets

1st Ali Toccacieli
MSc Program in Engineering in Computer Science
Sapienza University of Rome
Rome, Italy
alitoccacieli@gmail.com

2nd Manuela Petti
dept. of Computer, Control, and Management Engineering
Sapienza University of Rome
Rome, Italy
manuela.petti@uniroma1.it

***Abstract*—Tumor-educated platelets (TEPs) are circulating blood cells with a distinct tumor-driven phenotype and act as carriers and protectors of metastases. To date, some studies have shown that the TEPs transcriptome can be used for cancer diagnostics. The objective of this study is to propose a procedure based on differential gene expression and differential gene co-expression analyses to identify a set of key genes for multi-class cancer diagnostics. To reach this aim, we analyzed RNA-seq data (57736 genes) of 130 subjects, of whom 40 patients with glioblastoma multiforme (GBM), 35 patients with pancreatic adenocarcinoma (PAAD), and 55 healthy donors (HC). We focused our analysis on the subset of differentially expressed genes (DEGs), and we used these genes to build and analyze the differential co-expression networks, identifying the hub genes. With this procedure, we obtained a restricted set of DEGs that maximize the accuracy in classifying patients according to their conditions (GBM, PAAD, or HC). Indeed, we validated our results by comparing the achieved classification accuracy with that resulting from random selections of DEGs and we obtained that genes selected by differential co-expression (DCE) network analysis have greater predictive power than any other set of differentially expressed genes, including using all of them.**

***Index Terms*—network medicine, tumor-educated platelets, differential gene expression, differential co-expression networks, feature selection**

I. Introduction

Platelets, originating as anucleate cells from megakaryocytes, have emerged as central players in the systemic and local responses to tumor growth [1], [2]. Liquid biopsies offer a minimally invasive and sensitive alternative to tissue biopsies for cancer management. Specifically, tumor-educated platelets (TEPs) provide a comprehensive characterization of the heterogeneous tumor profile offering an alternative approach for detecting and monitoring cancer. In fact, liquid biopsies could enable early detection (screening), as well as “real-time” assessment of treatment effectiveness and early detection of disease recurrence [1]. Recent studies have demonstrated the success of TEPs RNA-seq data analysis in classifying samples affected by different tumors, obtaining relevant results in terms of high accuracy. The seminal study was conducted by Best et al. in 2015 [3] and highlighted distinct onco-signatures for six primary cancer types identifying differentially expressed genes from TEPs transcriptome. Later, other studies proposed tumor-educated platelet biomarkers of several tumor types including glioblastoma [4], pancreatic cancer [5], non-small-cell lung cancer [6] and renal cell carcinoma [7]. All these studies are based on differential gene expression (DGE) analysis, and thus on the study and interpretation of differences in abundance of gene transcripts.

Recently, network analysis has been established as one of the most successful and promising computational approaches in studying human diseases [8], [9]. In particular referring to gene expression data, gene co-expression networks highlight genes that have the tendency to exhibit similar expression profile in a group of samples and differential co-expression (DCE) network identify alterations of gene co-expression patterns observed in two different conditions [10], [11].

In this work, we aim to propose a procedure for the identification of a set of key genes for multi-class cancer diagnostics that goes beyond the study of differential expression. Indeed, we propose to exploit also a network-based approach to the analysis of TEPs transcriptome. For the aim of the study, we consider three different classes (glioblastoma multiforme, GBM; pancreatic adenocarcinoma, PAAD; and healthy donors, HC), and we validate the proposed procedure by comparing the obtained multi-class accuracy with that resulting from the use of differential expression analysis alone.

II. Material and Methods

A. Data

mRNA sequencing data of tumor-educated blood platelets were obtained from the Gene Expression Omnibus (GEO) database [12] under accession code GSE68086, using the GEOquery R package function getGEO. We downloaded 40 GBM samples, 35 PAAD samples, and 55 healthy control (HC) samples, each of which had an expression profile of 57736 genes. Genes that did not satisfy the condition that any sample

This research was funded by Sapienza University of Rome, “Progetto di Ateneo”, grant number: RM12117A8B3C3033.

related to that specific gene had a read count greater than 5 were removed, thus reducing the set of genes to 14411.

B. Differential Gene Expression Analysis

To identify the differentially expressed genes (DEGs), we used the R package edgeR (version 4.1.2). In detail, we initiated a DGEList with a count matrix containing the read counts for 130 samples and 14411 genes. We built the group factor requested by the DGEList with our 3 class denominations (GBM, PAAD, and HC). The DGEList object holds the dataset to be analyzed by edgeR and the subsequent calculations performed on the dataset (e.g., normalization). Once the DGEList object was initiated, we applied the Trimmed Mean of M-Values (TMM) normalization. TMM normalization is a simple and effective method for estimating relative RNA production levels from RNA-seq data [12]. After fitting the negative binomial model and estimating the common, tagwise, and trended dispersion, differentially expressed transcripts were determined using a generalized linear model (GLM) likelihood ratio test.

At this point, for each gene, we obtained the following measures:

- logarithmic fold change (logFC) defined as $log_2(A/B)$, where A and B are the two conditions under investigation.
- false discovery rate (FDR).
- logarithmic counts per million (logCPM).

We performed a selection of genes by filtering out those genes that did not satisfy a $logCPM > 3$ and $FDR < 0.01$. The number of genes that satisfied these constraints is 2045.

C. Differential Co-Expression Network Analysis

Once the differentially expressed genes were identified, we studied the differentially co-expressed networks. The first step was to compute the two co-expression networks, one related to tumor samples (GBM or PAAD) and one related to healthy samples. The genes used are those selected with the DGE analysis (n = 2045). We obtained the co-expression networks (tumor and healthy) using the R function cor (method: Pearson). At this point, we applied the Fisher z-Transformation:

$$z_{AorB} = \frac{1}{2} ln \left(\frac{1 + p_{AorB}}{1 - p_{AorB}} \right) \tag{1}$$

Where z_{AorB} is the Fisher's transformation for condition A or condition B.

At this point, we computed the z-scores by applying the following formula:

$$Z = \frac{z_A - z_B}{\sqrt{\frac{1}{n_A - 3} + \frac{1}{n_B - 3}}} \tag{2}$$

Here nA refers to the sample size of the respective tumor (GBM or PAAD) and nB refers to the sample size of HC.

To highlight significant and stronger differences in the comparison between co-expression networks, we set the threshold t equal to 3. Thus, filtering the z scores, we obtained the adjacency matrix A as follows:

$$A(v) = \begin{cases} 1 & \text{if } v > t \\ -1 & \text{if } v < -t \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

To identify the key genes in the DCE networks, we computed the degree index, which, in this application, informs about the number of significant chances in which each node is involved. Then we built the degree distribution by removing isolated genes, computed the 90^{th} quantile, q, and selected those genes whose degree was greater than q.

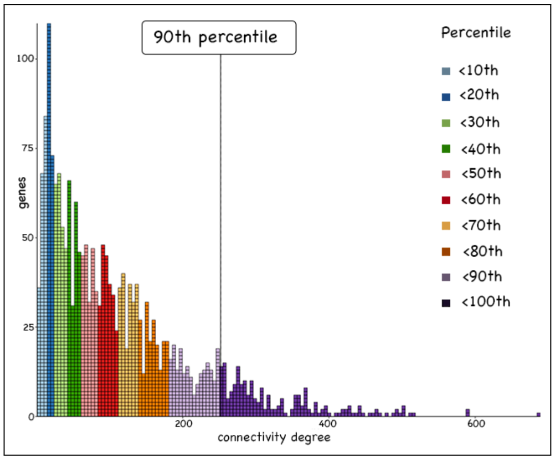

Fig. 1. Degree distribution of GBM differential co-expression network: each color represents a quantile cutoff in the range of 0% to 90%.

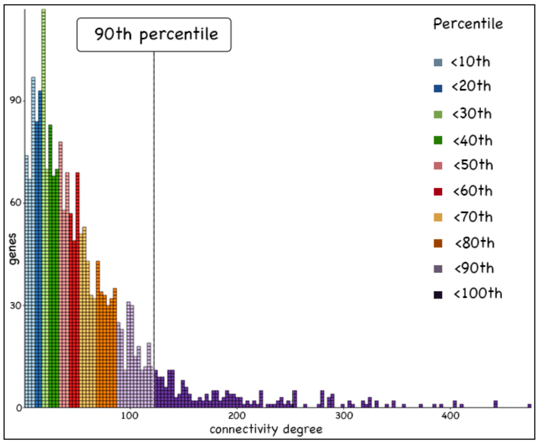

Fig. 2. Degree distribution of PAAD differential co-expression network: each color represents a quantile cutoff in the range of 0% to 90%.

As we can see in Fig.1 and Fig.2, a line shows the 90^{th} percentile q computed over the degree values of each comparison (tumor vs healthy). This value was then used to select only those genes with a degree above it. We computed this step for both differential networks, GBM vs HC and PAAD vs HC. Then these genes were merged together ($S1 \cup S2$), and genes that belonged to both groups (intersection, $S1 \cap S2$) were removed, resulting in a final set of 332 genes. This set of genes was then used for classification.

D. Classification with DE and DCE genes

For classification tasks, we used Python libraries such as Numpy, Pandas, and Scikit-learn. An SVM machine was used, with a One-vs-One (OVO) approach. OVO is a heuristic method for using binary classification algorithms for multi-class classification. Like one-vs-rest, one-vs-one splits a multi-class classification dataset into binary classification problems. The ratio between the training set and the validation set was of 70:30, respectively. This means that, of the 130 samples, 70% were used for training and 30% for validation. We used the function ”trainTestSplit” from the library scikit-learn in order to have the same percentage of sample selection for each class. Since the results could depend on the sample selection, we made 100 iterations of unique and different training and validation sets, and at each iteration, we stored the results in a vector. Once the 100^{th} iteration was reached, we returned the average results.

We applied this procedure to three different sets of genes. The first one involved all DEGs (n = 2045). The second one, instead, involved only DCE genes, that is, genes with a connectivity degree above the 90^{th} percentile in the differential co-expression network. The third genes selection was performed for validation purposes. Indeed, we tested the classification performances associated to a random selection of a set of genes with the same cardinality of the DCE genes (i.e. n=332). To perform this phase, we picked 100 different sets of 332 genes from the 1713 DEGs (difference set between DEGs and DCE genes).
Figure 3 shows an overview of the proposed procedure.

III. Results

In this section, we show the predictive power of differentially expressed and differentially co-expressed genes, with a connectivity degree in the differential co-expression network greater than the 90^{th} percentile of the degree distribution (see Fig.1 and Fig.2). As mentioned before, 332 genes (features) were selected for multi-class classification. The values of the performance metrics (accuracy, sensitivity and specificity) were averaged across the 100 different shuffles of samples (70:30, training:validation). The results show that the here proposed procedure for genes selection is able to distinguish the 3 classes (GBM, PAAD, and HC) with an overall accuracy of 91% (SD = 0.044), whereas using all of the differentially expressed genes (n = 2045) returns an overall accuracy of 90% (SD = 0.040). We additionally computed the performance metrics for each class. The scores (see Table I) indicate a sensitivity of 78%, a specificity of 99% and an accuracy of 92% for GBM. For PAAD, instead of a sensitivity of 99%, a specificity of 94% and an accuracy of 96% were obtained.

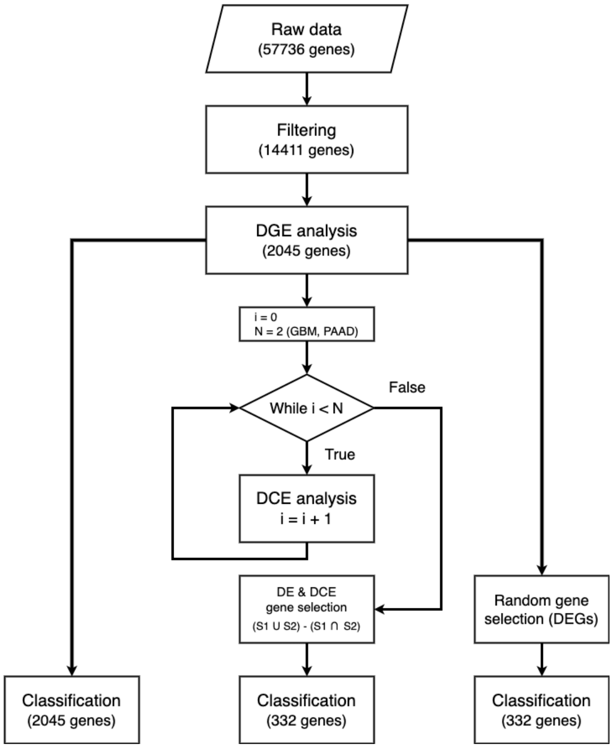

Fig. 3. Overview of the proposed procedure

TABLE I
DEGs with a degree above the 90^{th} percentile

Genes 332	Individual class scores		
	Sensitivity	***Specificity***	***Accuracy***
GBM	78%	99%	92%
PAAD	99%	94%	96%

Overall accuracy 91%

TABLE II
DEGs with a degree below the 90^{th} percentile

Genes 332 (100 iterations)	Individual class scores		
	Sensitivity	***Specificity***	***Accuracy***
GBM	72%	95%	88%
PAAD	96%	95%	95%

Overall accuracy 86%

To validate these results, we used random sets of 332 genes between the DEGs, excluding DCE genes. This was done, as already explained, 100 times. The results presented in the following paragraph refer to the average value of performance metrics (overall accuracy and individual class scores) across the random iterations.

The overall accuracy returned by the SVM (with the OVO approach) was of 86% (SD = 0.019). We additionally computed scores for each class (see Table II). The classification returned a sensitivity of 72%, a specificity of 95% and an accuracy of 88% for GBM. For PAAD, instead, a sensitivity of 96%, a specificity of 95% and an accuracy of 95% were returned.

Figure 4 shows the differential co-expression network obtained for GBM: the hubs are hilighted in red.

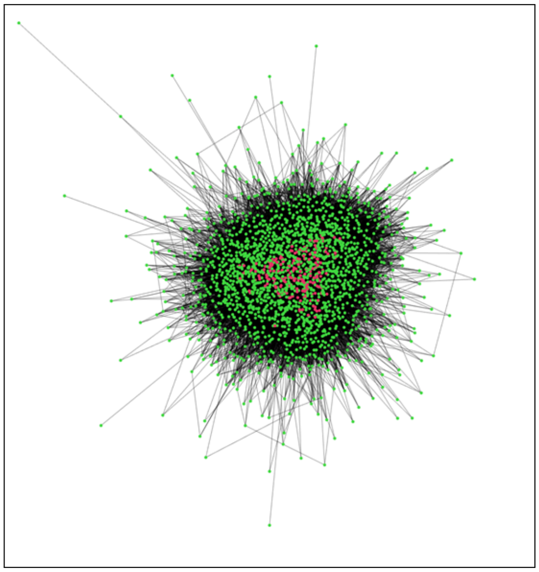

Fig. 4. Differential network for GBM.

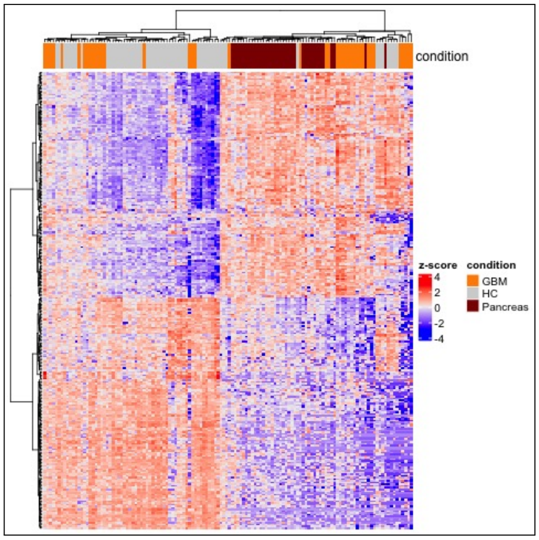

Fig. 5. Heatmap of unsupervised clustering of TEPs mRNA profiles of healthy donors (gray), GBM patients (orange) and PAAD patients (red).

IV. Discussion

We focused our analysis on two tumors GBM and PAAD, both with a five-year survival of less than 5% and for which an early diagnosis could have a great impact. From the heatmap we have generated in Fig.5, we can see how differentially expressed genes distinguish tumor patients (GBM and PAAD) from healthy donors; this result is consistent with the overall accuracy of 90% obtained by using a support vector machine (OVO approach) over the 2045 differentially expressed genes. We can observe from the heatmap that GBM samples tend to be distributed more separately than PAAD samples, which tend to be distributed close to each other. The same result is reflected in the individual classification scores. In fact, PAAD shows a higher sensitivity and specificity than what GBM offers. This is reflected in both sets of studies. In the first study, GBM scored a sensitivity of 78% while PAAD scored a sensitivity of 99% (Table I). The difference is slightly above 20%. In the second study, instead, GBM scores a sensitivity of 72% while PAAD scores 96% (Table II). As we can see, both GBM and PAAD sensitivity decreases, by losing 6% and 3% respectively.

In the previous section, we showed the results by using 2045 genes (DEGs), 332 genes (DCE genes), and a random set of 332 differentially expressed genes (average score over 100 different sets). The reduction of the features from 2045 genes (DEGs) to 332 genes (DCE) has determined a gain of 1% of accuracy. This means that with only 13% DEGs, selected by the differential co-expression analysis, we are able to obtain similar performance. In order to confirm these results, as explained in previous sections, we compared the DCE genes with random sets of 332 genes, selected from the 2045 differentially expressed ones. From the results, we can see an improvement in accuracy by using genes with a connectivity degree greater than the 90^{th} percentile (DCE genes). The difference in accuracy score between using these genes and using random sets of genes of the same size (n = 332) is 5%.

V. Conclusion

The obtained results confirm our initial assumption, which states that DGE-based DCE network analysis identifies a restricted set of differentially expressed genes (n=332) that are able to maximize accuracy in discriminating patients belonging to different conditions (in this study, GBM, PAAD, and HC). This was confirmed by the comparison with the performance metrics obtained by classifying all 2045 DEGs and validated by the comparison with the performance metrics obtained by classifying 100 randomly selected sets of DEGs (same size of DCE set,i.e. n=332). The accuracy obtained with our selection is grater than the one obtained by using all differentially expressed genes. Even if the increase was only of 1%, the number of genes involved in our selection was only the 13% of the total number of DEGs. This could raise some questions about whether DCE analysis could be a valid tool to increase the precision of identifying potential biomarkers for cancer diagnostics. This study wants to bring this new aspect to attention in order to provide a new approach that could highlight the importance of TEPs in cancer diagnostics.

Future Work

Results shown in this paper confirm that DGE and DCE analyses could represent a valid tool for the identification of cancer biomarkers for multi-class diagnostics. In this study, we only studied GBM and PAAD, but in future works, more classes will be considered in order to investigate the potentiality of differential network analysis over many different tumors.

References

[1] In 't Veld SGJG, Wurdinger T. Tumor-educated platelets. Blood. 2019 May 30;133(22):2359-2364. doi: 10.1182/blood-2018-12-852830.

[2] Best MG, Wesseling P, Wurdinger T. Tumor-educated platelets as a noninvasive biomarker source for cancer detection and progression monitoring. Cancer Res. 2018;78(13):3407-3412.

[3] Best MG, Sol N, Kooi I, Tannous J, Westerman BA, Rustenburg F, Schellen P, Verschueren H, Post E, Koster J, Ylstra B, Ameziane N, Dorsman J, Smit EF, Verheul HM, Noske DP, Reijneveld JC, Nilsson RJA, Tannous BA, Wesseling P, Wurdinger T. RNA-Seq of Tumor-Educated Platelets Enables Blood-Based Pan-Cancer, Multiclass, and Molecular Pathway Cancer Diagnostics. Cancer Cell. 2015 Nov 9;28(5):666-676. doi: 10.1016/j.ccell.2015.09.018.

[4] Sol N, In 't Veld SGJG, Vancura A, Tjerkstra M, Leurs C, Rustenburg F, Schellen P, Verschueren H, Post E, Zwaan K, Ramaker J, Wedekind LE, Tannous J, Ylstra B, Killestein J, Mateen F, Idema S, de Witt Hamer PC, Navis AC, Leenders WPJ, Hoeben A, Moraal B, Noske DP, Vandertop WP, Nilsson RJA, Tannous BA, Wesseling P, Reijneveld JC, Best MG, Wurdinger T. Tumor-Educated Platelet RNA for the Detection and (Pseudo)progression Monitoring of Glioblastoma. Cell Rep Med. 2020 Oct 20;1(7):100101. doi: 10.1016/j.xcrm.2020.100101.

[5] Mantini G, Meijer LL, Glogovitis I, In 't Veld SGJG, Paleckyte R, Capula M, Le Large TYS, Morelli L, Pham TV, Piersma SR, Frampton AE, Jimenez CR, Kazemier G, Koppers-Lalic D, Wurdinger T, Giovannetti E. Omics Analysis of Educated Platelets in Cancer and Benign Disease of the Pancreas. Cancers (Basel). 2020 Dec 29;13(1):66. doi: 10.3390/cancers13010066.

[6] Sheng M, Dong Z, Xie Y. Identification of tumor-educated platelet biomarkers of non-small-cell lung cancer. Onco Targets Ther. 2018 Nov 14;11:8143-8151. doi: 10.2147/OTT.S177384.

[7] Xiao R, Liu C, Zhang B, Ma L. Tumor-Educated Platelets as a Promising Biomarker for Blood-Based Detection of Renal Cell Carcinoma. Front Oncol. 2022 Mar 7;12:844520. doi: 10.3389/fonc.2022.844520.

[8] Barabási AL, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. Nat Rev Genet. 2011 Jan;12(1):56-68. doi: 10.1038/nrg2918.

[9] Farina L. Network as a language for precision medicine. Ann Ist Super Sanita. 2021 Oct-Dec;57(4):330-342.

[10] de la Fuente A. From 'differential expression' to 'differential networking' - identification of dysfunctional regulatory networks in diseases. Trends Genet. 2010 Jul;26(7):326-33. doi: 10.1016/j.tig.2010.05.001.

[11] Ideker T, Krogan NJ. Differential network biology. Mol Syst Biol. 2012 Jan 17;8:565. doi: 10.1038/msb.2011.99.

[12] Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 2002 Jan 1;30(1):207-10

[13] Robinson, M.D., Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol 11, R25 (2010). https://doi.org/10.1186/gb-2010-11-3-r25

.