A Novel Method for the Analysis of Gene Expression Microarray Data with K-Means Clustering: Sorted K-Means

George I. Lambrou[1], Maria Braoudaki[2]

[1,2]National and Kapodistrian University of Athens, First Department of Pediatrics, Choremeio Research Laboratory, Thivon&Levadeias, 11527 Goudi, Athens, Greece
Email: glamprou@med.uoa.gr

[2]University Research Institute for the Study and Treatment of Childhood Genetic and Malignant Diseases, National and Kapodistrian University of Athens, «Aghia Sophia» Children's Hospital, Athens, Greece
Email: mbraoudak@med.uoa.gr

Abstract—

***Background**: Microarray technology has revolutionized the way genomic analysis has been performed. High-throughput data acquisition, brought up a challenge in data comprehension i.e. in gene expression.*

***Methods**: k-means cluster obtained after analysis of miRNA expression data have been sorted by an algorithmic procedure. **Results**: The proposed method managed to sort k-means centroids and manifest a more simple way of drawing conclusions on studied tumor samples. miRNAs were unraveled that changed in expression levels with respect to tumor aggressiveness. **Conclusions**: In the present work we presented a new and simple approach in data analysis using a new analysis approach, which we termed sorted-k-means analysis.*

Keywords—k-means, microarrays, sorted k-means, gene expression, miRNA, tumors.

I. INTRODUCTION

Microarray technology is based on the basic property of nucleic acids, which is the selective binding of two complementary chains/sequences. The basic technological idea on the exploitation of this property already existed [1, 2], but what gave a huge rash into this technology was the discovery of microarrays. However, this could not have been feasible without the unraveling of the sequence of the human genome. At the same time, advances in technological aspects such as the miniaturization of arrays and high-density printing on a solid substrate have allowed the appearance of microarray chips [3].The advent of this technology has given the initial idea that questions arising for cellular and molecular events could easily be answered through a comparison between the "control" and the "investigated" samples. Microarray technology was initially applied in molecular investigations on the genome level at the end of 90's[4, 5]. DNA microarrays detect patterns of gene expression, therefore they can be used for acquiring such "images" and the induction of conclusions on cell state [6]. cDNA microarrays have been used for a plethora of experiments; virtually any property of a DNA sequence which can be experimentally modified may be determined as far as its differential expression is concerned, and this can be performed on thousands of sequences simultaneously. Research questions that can be answered with DNA microarrays are related mainly to the investigation of gene expression. They can compare the relative abundance of mRNA (or microRNA; miRNA) of a gene under investigation, between different cells or tissue samples. For example, an experiment could compare cells before and following an experimental intervention, or at successive moments of a specific process, or between stages of differentiation or mRNA expressed in a mutant cell compared to that of wild type. This would be the simplest type of experiment. In particular, microarrays have been applied for the diagnosis of cancer [7, 8]. They have been used to investigate the hypothesis that the classification of cancers can be based on their gene expression profiles, subsequently eliminating the need for histopathological diagnosis [9]. Microarray analysis has been used to predict "tumor grades" or subtypes of cancers, regardless of prior knowledge of their biology [10]. They can also provide an opportunity to study the possibility of tumor gene expression in correlation with the prediction of the disease outcome. In general, microarray platforms afford attractive methodologies for discovery-based investigations [11].

In the current study, we attempted to present a novel methodology for the analysis of gene expression data using a unique k-means approach which we defined as *sorted k-means*. We manifested the suggested algorithm by using a previously studied dataset, from childhood Central Nervous System (CNS) tumors.

II. Materials and Methods

2.1 Samples

The previously studied childhood CNS malignancies were investigated for their miRNA expression profiles [12]. The aforementioned study included in total,, 26 resected brain tumors from children diagnosed with pilocyticastrocytomas (PA) (n=19) and ependymomas (EP) (n=7). Additionally, we included glioblastomas (GB) (n=12) (online data: E-MEXP-1029[1][13]), germinomas (GE) (n=12) (GSE19347[2]) and dysembryoblasticneuroepithelial tumors (DNET) (n=4). All were diagnosed according to the 2007 WHO criteria [14]. As controls, 17 samples were used; The First-Choice Human Brain Reference RNA was used (Ambion, Austin, TX, USA) and 16 samples were obtained from deceased children who underwent autopsy and were not present with any brain distortion, including the following anatomic locations: cerebellum (n=4), medulla oblongata (n=4), parietal lobe (n=4) and temporal lobe (n=4).

2.2 MicroRNA Profiling

The miRNA profiling was performed as described by Braoudaki *et al.*, 2014 [11, 15]. In brief, total RNA and miRNAs were extracted using the Trizol standard protocol (Invitrogen, Carlsbad, CA) and the mirVANA miRNA isolation kit (Ambion, Austin, TX). Labelling and hybridization were carried out using the LabelIT miRNA labelling kit (Mirus Bio LLC, USA) following the manufacturer's instructions. All specimens were hybridized to n Applied MicroArrays (miRlinkBioarray 300054-3PK) platform and all images were scanned using the Agilent Microarray Scanner (G2565CA) controlled by Agilent Scan Control 7.0 software. The total gene signals were extracted using the Imagene 6.0 software (Biodiscovery Inc., USA). MicroRNAs were significantly differentially expressed (DEx) when they obtained *a p-value*<0.05 and a false discovery rare; FDR≤0.05. Overall, our analysis revealed 70 DE miRNAs.

2.3 Data Analysis

The multiparameter analyses were performed with MATLAB® simulation environment (The Mathworks, Inc., Natick, MA). Microarray data were processed as previously reported [15]. In brief, filtering was performed based on the signal intensity. Background correction was carried out by subtracting the median local background from the signal intensity as previously reported [16]. Normalization was performed using the quantile normalization algorithm. The two tailed student t-test was used to test the mean differences between two groups. MicroRNAs were considered to be significantly differentially expressed (DEx) if they obtained a p-value<0.05 and an FDR≤0.05. MiRNA expression levels were further analyzed with the k-means methodology. *K*-means is a method of cluster analysis**,** which partitions***n*** observations into *k* clusters**,** in which each observation belongs to the cluster with the nearest mean. Given a set of observations ($x_1, x_2, \ldots, x_n$), where each observation is a *d*-dimensional real vector, then *k*-means clustering aims to partition the *n* observations into *k* sets ($k<n$) S={S_1, S_2, …, S_k} so as to minimize the within-cluster sum of squares. Further on, centroids were calculated and were sorted in an ascending order. Sorting was performed with the MATLAB computing environment. Each cluster was transformed in a DataMatrix structure, where rows indicated miRNAs and columns designated the tumor samples. The DataMatrix was then sorted with respect to the column and plotted, respectively. This was repeated for each k-means cluster separately. The k-means implementation in MATLAB has a randomized component, which is the selection of initial centers. This implies that every time the methodology will yield different results. Yet, our methodology sorts the produced centroids every single time accounting for the random effect of the MATLAB k-means algorithm. MiRNA annotation was performed with the *Webgestalt*[3] on-line tool [17, 18].

III. Results

MiRNA expression profiles were clustered with k-means with respect to all CNS tumor samples, in random order (**Fig. 1A**). At the same time, centroids were also randomly calculated (**Fig. 1B**). The calculated centroids were sorted and samples appeared in ascending order with respect to ascending miRNA expression levels (**Fig. 1C**). Samples were sorted and manifested a pattern with respect to all samples. This type of sorting gave the opportunity to examine patterns of expression with respect to the complete sampling. The next step in the evaluation of k-means clustering was to calculate the mean of each tumor type with respect to each miRNA. Those miRNAs and tumor types were clustered in random order, where it

[1] http://www.ebi.ac.uk/arrayexpress/experiments/E-MEXP-1029/

[2] http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE19347

[3] http://bioinfo.vanderbilt.edu/webgestalt/

appeared that it was not easy to extract rapid conclusions (**Fig. 2A**, **2B**). Sorting the k-means clusters' centroids made easier to identify patterns of expression. In particular, it appeared that sorted c*entroids in an ascending order revealed specific patterns. In particular*, *red*-shaded clusters indicated clusters where sample size coincided with tumor aggressiveness (**Fig. 2C**). In sorted clusters 13, 20, 22, 29, 31 and 34, tumors were presented from the most aggressive tumor (GE) to the most benign (DE) (**Fig. 2C**). In addition, in cluster 33, miRNAs were classified with respect to aggressiveness from the benign (DE) to the most aggressive tumor type (GE) (**Fig. 2C**).This type of analysis also revealed significant differences between tumor types. For example, in cluster 13, the miRNA expression levels were significant between germinoma (GE) and DNET (DE) samples. The suggested algorithm could successfully sort tumor types with respect to aggressiveness, both in descending as well as ascending order. In particular, genes appeared to increase from aggressive to benign neoplasms. Additionally, annotation analysis showed that miRNAs that are expressed in such a pattern participated in both hematological malignancies and in neuroectodermal tumors (**Supplementary Table 1**). The individual k-means clusters are provided as supplementary data (file: kmeans_Group_Quantile.xlsx). In addition, the code that generated the suggested algorithm is provided as supplementary data (file: MATLAB Code for Sorted k-means.docx).

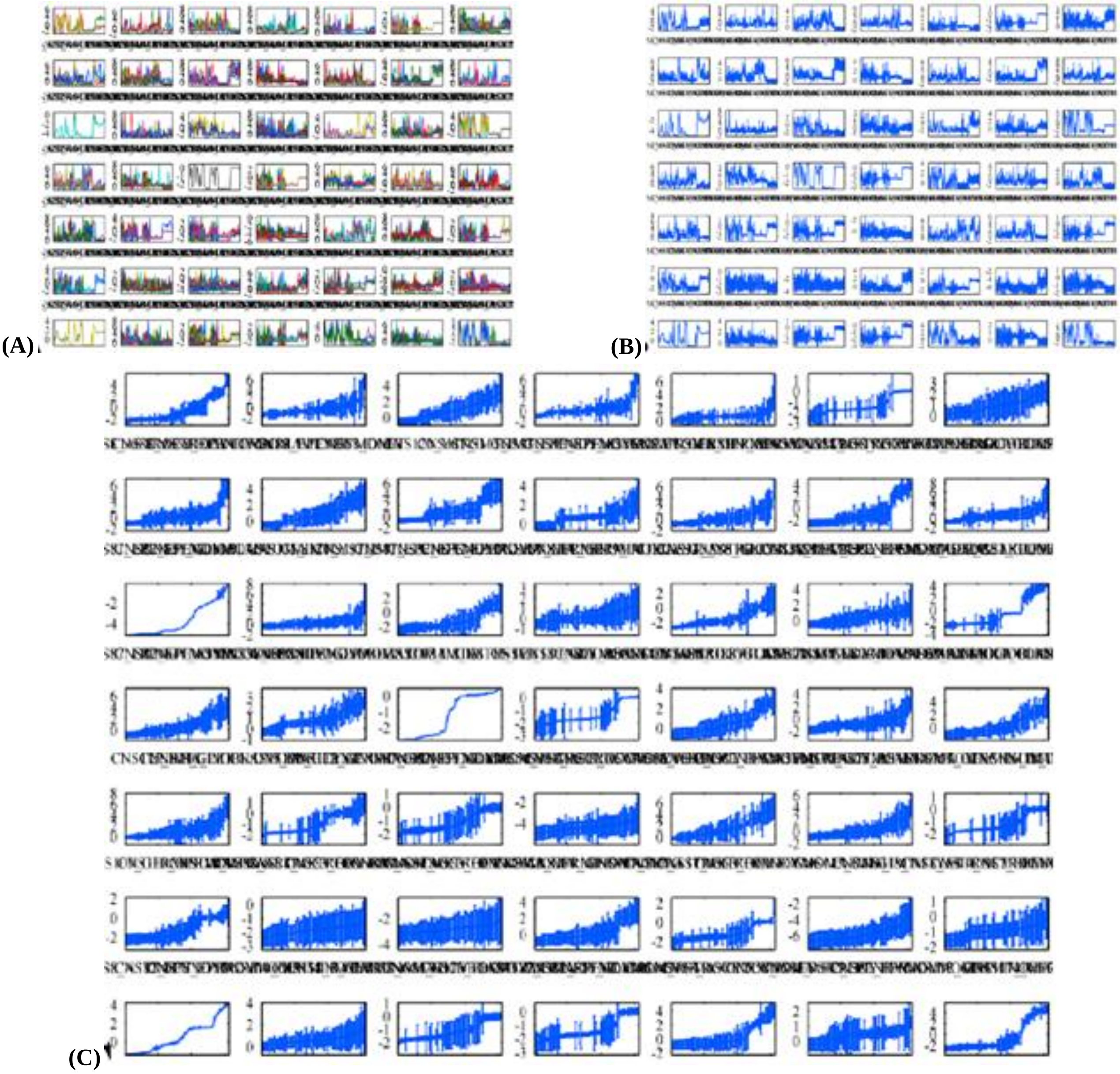

FIGURE 1. K-MEANS CLUSTERING OF THE COMPLETE SAMPLE SIZE. ALL INDIVIDUAL SAMPLES WERE CLUSTERED (A). CENTROIDS WERE CALCULATED IN THE SAME ORDER AS SAMPLES WERE CALCULATED (B) AND FURTHER ON, CENTROIDS WERE SORTED IN AN ASCENDING ORDER (C)

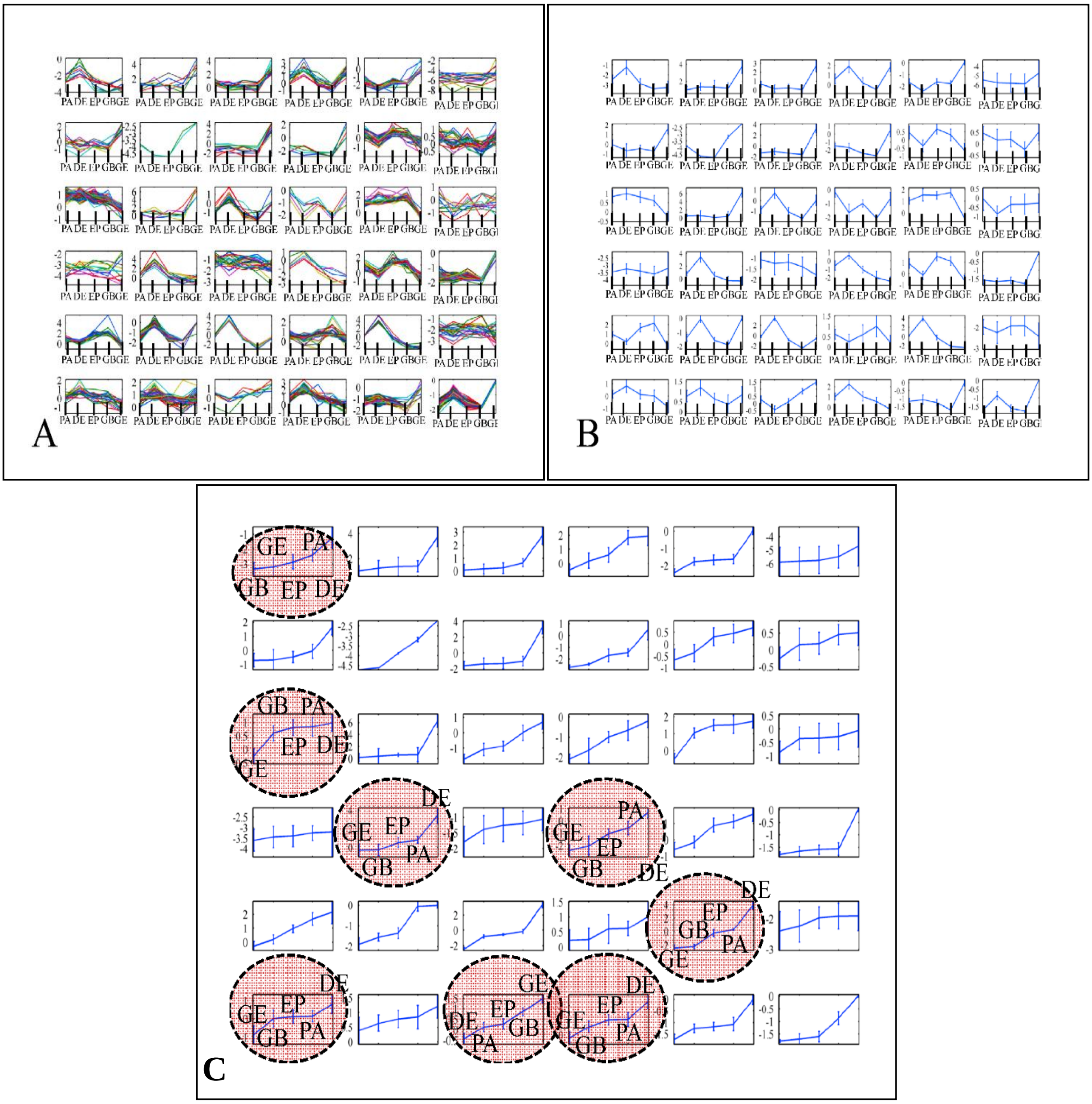

Fig. 2. The mean expression values of each miRNA were calculated with respect to the tumor type i.e. Dysembryoblastic Neuroepithilial Tumors (DNET designated as DE), Glioblastoma (Designated as GB), Pilocytic Astrocytoma (designated as PA), Ependymomas (designated as EP) and Germinomas (Designated as GE). Mean miRNA expression levels were clustered in random order (A), whereas the centroids were also clustered in random order (B). On the other hand, centroids were sorted in an ascending order where specific patterns were revealed. In particular, red-shaded clusters indicate clusters where sample size coincides with tumor aggressiveness (C). In sorted clusters 13, 20, 22, 29, 31 and 34 tumors were presented from the most aggressive tumor (GE) to the most benign (DE) (C). In addition, in cluster 33, miRNAs were classified with respect to aggressiveness from the benign (DE) to the most aggressive tumor type (GE) (C).

IV. Discussion

K-means methodology is an extremely useful tool in the analysis of high-throughput gene expression data. Although useful, the output of k-means clustering makes it difficult to extract conclusions especially in the case of random pre-disposition of samples. The proposed method transforms k-means output data in such a way that cluster centroids were presented sorted with respect to gene expression levels. To the best of our knowledge, this is the first report, in which such an algorithm is proposed. In this sort of representation we were able to distinguish tumor types with respect to aggressiveness. Also, several miRNAs were found to ascend with respect to tumor aggressiveness, i.e. manifesting increasing expression levels as tumor grade decreased (from very aggressive to benign), which easily hinted towards groups of miRNAs that serve as possible tumor suppressor markers. The opposite pattern was also observed, i.e. miRNA levels increasing from benign neoplasms to more aggressive. This observation also led toward a group of putative tumor suppressor miRNAs. For example, the data used included tumors ranging from benign (e.g. DNET) to very aggressive (e.g. GB and GE), where both patterns were detected and miRNAs were identified as possible markers involved in tumor progression or tumor inhibition. For example, as previously reported, miR-184 and miR-766 potentially afford a oncogenic markers, which is consistent to our observations [19-21]. Similarly, we found that miR-1 was up-regulated as tumor grade increased, while other reports referred to this miRNA as a putative tumor suppressor molecule [22-24], and a possible novel therapeutic marker. Data sorting can be performed for several characteristics, besides tumor grade and it is possible to discover patterns based on other clinical phenotypes. The proposed algorithm is very simple and easy to use, yet it could be of important assistance in the comprehension of complicated datasets including microarray expression data.

V. CONCLUSION

In the present work we presented a new and simple approach in data analysis using a new analysis approach, which we termed *sorted-k-means* analysis. In several cases, k-means classification provides useful insight towards the understanding of biological data. Our analysis expands this potential and provides a further classification step, which might assist in the easier and more comprehensive understanding of complex microarray data.

ACKNOWLEDGEMENTS

GIL: This work was funded in part by the University of Athens, Medical School. **MB**: This work was funded in part by the University of Athens, Medical School. **Authors' Contributions**: **GIL**: conceived the idea, performed data analysis and drafted the manuscript, **MB**: performed experiments, proof-read the manuscript.

COMPETING INTEREST, DISCLOSURES AND CONFLICT OF INTEREST

The authors have nothing to disclose and no conflict of interest

SUPPLEMENTARY DATA

1. *Supplementary Table 1.xlsx*. Disease annotation table of miRNAs as derived from the *Webgestalt* website.
2. *MATLAB Code for Sorted k-means*.docx. Code used to perform the sorted k-means algorithm.
3. *kmeans_Group_Quantile.xlsx*. K-means cluster data as extracted from the algorithmic analysis. Each cluster corresponds to the same cluster in Figure 2.

REFERENCES

[1] D. Gillespie and S. Spiegelman, "A quantitative assay for DNA-RNA hybrids with DNA immobilized on a membrane," J Mol Biol, vol. 12, pp. 829-42, Jul 1965.

[2] E. M. Southern, "Detection of specific sequences among DNA fragments separated by gel electrophoresis," J Mol Biol, vol. 98, pp. 503-17, Nov 5 1975.

[3] H. C. Causton and L. Game, "MGED comes of age," Genome Biol, vol. 4, p. 351, 2003.

[4] J. DeRisi, L. Penland, P. O. Brown, M. L. Bittner, P. S. Meltzer, M. Ray, *et al.*, "Use of a cDNA microarray to analyse gene expression patterns in human cancer," Nat Genet, vol. 14, pp. 457-60, Dec 1996.

[5] J. L. DeRisi, V. R. Iyer, and P. O. Brown, "Exploring the metabolic and genetic control of gene expression on a genomic scale," Science, vol. 278, pp. 680-6, Oct 24 1997.

[6] M. Diehn, A. A. Alizadeh, and P. O. Brown, "Examining the living genome in health and disease with DNA microarrays," JAMA, vol. 283, pp. 2298-9, May 3 2000.

[7] S. A. Armstrong, J. E. Staunton, L. B. Silverman, R. Pieters, M. L. den Boer, M. D. Minden, *et al.*, "MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia," Nat Genet, vol. 30, pp. 41-7, Jan 2002.

[8] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, *et al.*, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," Science, vol. 286, pp. 531-7, Oct 15 1999.

[9] S. Ramaswamy and T. R. Golub, "DNA microarrays in clinical oncology," J Clin Oncol, vol. 20, pp. 1932-41, Apr 1 2002.

[10] U. R. Kees, J. Ford, M. Watson, A. Murch, M. Ringner, R. L. Walker, *et al.*, "Gene expression profiles in a panel of childhood leukemia cell lines mirror critical features of the disease," Mol Cancer Ther, vol. 2, pp. 671-7, Jul 2003.

[11] M. Braoudaki and G. I. Lambrou, "MicroRNAs in pediatric central nervous system embryonal neoplasms: the known unknown," J Hematol Oncol, vol. 8, p. 6, 2015.

[12] M. Braoudaki, G. I. Lambrou, K. Giannikou, S. A. Papadodima, A. Lykoudi, K. Stefanaki, *et al.*, "Title," unpublished|.

[13] P. E. Blower, J. S. Verducci, S. Lin, J. Zhou, J. H. Chung, Z. Dai, *et al.*, "MicroRNA expression profiles for the NCI-60 cancer cell panel," Mol Cancer Ther, vol. 6, pp. 1483-91, May 2007.

[14] D. N. Louis, H. Ohgaki, O. D. Wiestler, W. K. Cavenee, P. C. Burger, A. Jouvet, *et al.*, "The 2007 WHO classification of tumours of the central nervous system," Acta Neuropathol, vol. 114, pp. 97-109, Aug 2007.

[15] M. Braoudaki, G. I. Lambrou, K. Giannikou, V. Milionis, K. Stefanaki, D. K. Birks, *et al.*, "Microrna expression signatures predict patient progression and disease outcome in pediatric embryonal central nervous system neoplasms," J Hematol Oncol, vol. 7, p. 96, 2014.

[16] A. Lycoudi, D. Mavreli, A. Mavrou, N. Papantoniou, and A. Kolialexi, "miRNAs in pregnancy-related complications," Expert Rev Mol Diagn, vol. 15, pp. 999-1010, Aug 2015.

[17] S. A. Kirov, B. Zhang, and J. R. Snoddy, "Association analysis for large-scale gene set data," Methods Mol Biol, vol. 408, pp. 19-33, 2007.

[18] B. Zhang, S. Kirov, and J. Snoddy, "WebGestalt: an integrated system for exploring gene sets in various biological contexts," Nucleic Acids Res, vol. 33, pp. W741-8, Jul 1 2005.

[19] W. S. Santhi, R. Prathibha, S. Charles, K. G. Anurup, G. Reshmi, S. Ramachandran, *et al.*, "Oncogenic microRNAs as biomarkers of oral tumorigenesis and minimal residual disease," Oral Oncol, vol. 49, pp. 567-75, Jun 2013.

[20] M. Sand, M. Skrygan, D. Georgas, D. Sand, S. A. Hahn, T. Gambichler, *et al.*, "Microarray analysis of microRNA expression in cutaneous squamous cell carcinoma," J Dermatol Sci, vol. 68, pp. 119-26, Dec 2012.

[21] X. F. Yu, J. Zou, Z. J. Bao, and J. Dong, "miR-93 suppresses proliferation and colony formation of human colon cancer stem cells," World J Gastroenterol, vol. 17, pp. 4711-7, Nov 14 2011.

[22] C. Han, Z. Yu, Z. Duan, and Q. Kan, "Role of microRNA-1 in human cancer and its therapeutic potentials," Biomed Res Int, vol. 2014, p. 428371, 2014.

[23] P. Letelier, P. Garcia, P. Leal, H. Alvarez, C. Ili, J. Lopez, *et al.*, "miR-1 and miR-145 act as tumor suppressor microRNAs in gallbladder cancer," Int J Clin Exp Pathol, vol. 7, pp. 1849-67, 2014.

[24] E. Osaka, X. Yang, J. K. Shen, P. Yang, Y. Feng, H. J. Mankin, *et al.*, "MicroRNA-1 (miR-1) inhibits chordoma cell migration and invasion by targeting slug," J Orthop Res, vol. 32, pp. 1075-82, Aug 2014.