Cancer Biomarkers

Biomarker identification in hematopoietic stem cells micro-array gene expression data in humans

--Manuscript Draft--

Manuscript Number:	
Full Title:	Biomarker identification in hematopoietic stem cells micro-array gene expression data in humans
Short Title:	
Article Type:	Research Article
Section/Category:	
Keywords:	biomarker; differentially expressed genes; cancer; gene ontology pathway enrichment; hippo signaling pathway
Corresponding Author:	Emine Güven, Ph.D. at West Virginia University Duzce University - Konuralp Campus: Duzce Universitesi Düzce, TURKEY
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	Duzce University - Konuralp Campus: Duzce Universitesi
Corresponding Author's Secondary Institution:	
First Author:	Emine Güven, Ph.D. at West Virginia University
First Author Secondary Information:	
Order of Authors:	Emine Güven, Ph.D. at West Virginia University
	Sevinç Akçay, Ph.D.
Order of Authors Secondary Information:	
Abstract:	BACKGROUND: An identification of the cellular and molecular properties of hematopoietic stem cells (HSCs) has made studies in clinical use, stem cell identification and use highly effective. This study uncovers biomarkers using GSE32719 data set publicly reachable at NIH/NCBI Gene Expression Omnibus database. The data set contains a total of expression of 54,676 genes of healthy human bone marrow hematopoietic stem cells in groups of 14 young (20–31 years), 5 middle age (42–61), 8 old (65–85) groups. OBJECTIVE: The differentially expressed genes (DEGs) were subjected to biological process which deciphers the increase in hematopoietic stem cell population and functional decline. METHODS: The data set is analyzed by using the GEOquery package in Bioconductor following standard procedures in R studio. Using Biobase, GEOquery, gplots packages out of which 453 genes chosen. The gene ontology of pathway enrichments and KEGG enrichment analyses of DEGs were studied. RESULTS: The results reveal UBC, PTK2, and TCF7L2 genes arei dentified as hub genes. UBC plays a critical role in maintaining ubiquitin (Ub) homeostasis which includes a delay in cell-cycle progression and increased susceptibility to cellular stress. Omics alterations of PTK2 in human cancers such Glioblastoma Multiforme and Glioma. Diseases associated with TCF7L2 include type 2 Diabetes Mellitus and Colorectal Cancer. CONCLUSIONS: Hippo signaling pathway was observed to be significant in acute myeloid leukemia (AML). UBC, PTK2, and TCF7L2 genes may be used as potential biomarkers for patients with AML and related diseases diagnosis and treatment.
Suggested Reviewers:	Dilek Pirim, Ph.D. dilekpirim@uludag.edu.tr
	Ye Chen, Ph.D.

	Ye.Chen@nau.edu
	Mustafa Unal, Ph.D. mustafaunal@kmu.edu.tr
Order of Authors (with Contributor Roles):	Emine Güven, Ph.D. at West Virginia University (Conceptualization; Formal analysis; Investigation; Methodology; Project administration; Resources; Software; Supervision; Validation; Visualization; Writing – original draft; Writing – review & editing)
	Sevinç Akçay, Ph.D. (Conceptualization; Formal analysis; Methodology; Supervision; Visualization; Writing – original draft; Writing – review & editing)
Additional Information:	
Question	**Response**

Cancer Biomarkers – submitted article

Biomarker identification in hematopoietic stem cells micro-array gene expression data in humans

Emine Güven[1*] and Sevinç Akçay[2]

1 Department of Biomedical Engineering, Düzce University, Düzce, Turkey

2 Department of Molecular Biology and Genetics, Ahi Evran University, Kırşehir, Turkey

*Corresponding author: Emine Güven, Assistant Professor of Bioinformatics, Engineering Building B, Konuralp Campus, Konuralp, Duzce, Turkey.

Tel.: + 90 (380) 5421036

E-mail: emine.guven@duzce.edu.tr

ORCID: 0000-0001-9324-0879

Abstract

BACKGROUND: An identification of the cellular and molecular properties of hematopoietic stem cells (HSCs) has made studies in clinical use, stem cell identification and use highly effective. This study uncovers biomarkers using GSE32719 data set publicly reachable at NIH/NCBI Gene Expression Omnibus database. The data set contains a total of expression of 54,676 genes of healthy human bone marrow hematopoietic stem cells in groups of 14 young (20–31 years), 5 middle age (42–61), 8 old (65–85) groups.

OBJECTIVE: The differentially expressed genes (DEGs) were subjected to biological process which deciphers the increase in hematopoietic stem cell population and functional decline.

METHODS: The data set is analyzed by using the GEOquery package in Bioconductor following standard procedures in R studio. Using Biobase, GEOquery, gplots packages out of which 453 genes chosen. The gene ontology of pathway enrichments and KEGG enrichment analyses of DEGs were studied.

RESULTS: The results reveal UBC, PTK2, and TCF7L2 genes are identified as hub genes. UBC plays a critical role in maintaining ubiquitin (Ub) homeostasis which includes a delay in cell-cycle progression and increased susceptibility to cellular stress. Omics alterations of PTK2 in human cancers such Glioblastoma Multiforme and Glioma. Diseases associated with TCF7L2 include type 2 Diabetes Mellitus and Colorectal Cancer.

CONCLUSIONS: Hippo signaling pathway was observed to be significant in acute myeloid leukemia (AML). UBC, PTK2, and TCF7L2 genes may be used as potential biomarkers for patients with AML and related diseases diagnosis and treatment.

Keywords: biomarker; differentially expressed genes; cancer; gene ontology pathway enrichment; hippo signaling pathway

1 Introduction

Bone marrow is a special site where blood cells are covered with structural stromal cells. It is a spongy adipose tissue found in the bones such as femurs, rib cage, ribs, pelvis and human skull. Bone marrow is fed with specialized blood vessels and contributes to circulation. The specialized phenestra capillary, called sinusoid, penetrates the extracellular / extracellular matrix (ECM), and ECMs are sponge-like matrices produced by reticular fibroblasts [1].

These proteins and cells place the hematopoietic cells in separate compartments. Similarly, hematopoietic cells, bidothelial cells, stromal cells, ECM, cytokines, growth factors, and chemokines contain special microenvironment in the bone marrow [2,3]. Recent improvements cover new markers for hematopoietic stem cells (HSCs) and niche stem cells, systematic analysis of expression datasets of niche factors, implementation of genetics for functional in vivo identification of niche cells, and improved imaging techniques. Stem cells have the ability to divide for a long time in the living body, to be able to regenerate and transform into other tissue cells by differentiating according to the needs of the body [4].

A detailed examination of the cellular and molecular properties of HSCs has made studies in clinical use, stem cell identification and use highly effective. Bone marrow microenvironment is an ideal place to support healthy and yet it might be also support malignant hematopoiesis that performs a fundamental position in the growth and development of leukemia and different cancer types [1,4–6].

Blast cells are declined during normal maturation, begin to accumulate in the blood together with the bone marrow in acute myeloid leukemia (AML). The body remains vulnerable because white blood cells cannot form. Red blood cell and thrombocyte production is impaired in the bone

marrow due to abnormal proliferation of myeloblasts. Thus, anemia infection and platelet count reduces [7].

AML is not limited to a particular part of the body since its onset, but can spread to the blood, lymphatic tissue and all other organ systems from the bone marrow. As with many other leukemia diseases, it is defined as a malignant systemic disease. HSC and other hematopoietic progenitor cells are evaluated and it was discovered that the elderly HSC increased frequently [8]. For several years, the discovery of AML-particular antigen has been utilized classified among the indicative regulators for checking AML.

Microarray has become a significant tool in investigating cancer genes and target therapeutic drugs. Recent studies suggest an extensive gene expression analysis of HSCs by reviewing publicly available gene expression data to detect age-related diseases gene expression altering related with AML [9]. Moreover, relative study of the differentially expressed genes (DEGs) stays moderately constrained, and a dependable biomarker profile would be a need to develop different drugs for different age groups [6,10]. The protein expression alterations in the development and growth of AML and related diseases require comprehensive analysis. Furthermore, the relations among the detected DEGs, specifically protein-protein interaction (PPI) networks and underlying signaling pathways should be clarified.

Raw data (GSE32719) is retrieved through GEO that is a public data repository for the storing of microarray and sequence-based data and their retrieval [8]. The DEGs of HSCs data between different age groups were identified by comparing gene expression by fold change and t-test. Afterwards, the DEGs were screened using Gorilla and DAVID Gene Ontology (GO) and the analysis of pathway enrichments [11,12]. By studying their hub nodes globally and between different age groups constructing PPI networks, the objective of this project is to investigate the

molecular mechanisms AML and related diseases growth and to come up with candidate biomarkers for diagnosis, therapeutic targets and predictions.

Earlier studies tackling age-related changes in human HSC, due to the implicit evaluation of source and progenitor populations, need to support experimental studies with numerical analysis and statistical methods in addition to previous HSC studies in mice. Various agents are assessed and employed to prevent AML, including monoclonal antibodies, new formulation of old drugs, FLT3 and IDH1/2 inhibitors [13,14]. Much work has been done with microarray gene expression technology to reveal the central mechanism of AML formation and progression and focus these methods for therapeutic approach. It still persists a request for more efficient treatments or methods that can improve curative responses to AML medication. In this project, we used microarray data sets of public transcriptome datasets of human bone marrow tissue and conduct connective analysis on the DEGs with bioinformatics tools. Although experimental studies are needed to confirm our findings, our results will reveal potential biomarkers and bright therapeutic objectives for AML. The present study also focused on hippo signaling pathway transcription factors that takes a key task in regulating organ growth, regeneration, homeostasis, and gene expression control [15].

2 Materials and methods

2.1. Preprocessing of the data set

The publicly available gene expression data set from human bone HSCs were pull out from the GEO database with GSE32719 [8]. Genomic information ranging from gene sequences to protein structure predictions were obtained. As described by Pang et al, the data set contains a total of expression of 54,676 genes of healthy human bone marrow hematopoietic stem cells in

groups of 14 young (20–31 years), 5 middle age (42–61), 8 old (65–85) groups. The GSE32719 data set is analyzed by using the GEOquery package in Bioconductor following standard procedures in R studio [16–19]. The other packages we used in R studio are as the following; Biobase, biomaRT and gplots packages [17,18,20]. To estimate the adjusted *p value* and avoid Type I errors, we used Bejamini- Hochberg Procedure to correct multiple testing. In order to adapt the statistical tests locally, hypergeometric model was performed for both of the down-regulated and up-regulated DEGs in the functional GO and pathway enrichment analysis, and false discovery rate (FDR) were computed [21–23].

2.2. Experimental data and analysis codes

Analysis were conducted in the R statistical environment. Sample codes and analysis of GSE32719 data can be found https://github.com/mathbioGVN/GSE32719.HSC.microarray.project repository. We separate samples into three groups provided that young-old aged, young-middle aged, and middle-old aged. The data set was normalized by computing the means of the samples of each group in R programming language. The process on separated samples which is grouped by categories was performed as computing fold-change (biological significance) difference between the means of the categories. A broadly performed statistical model is the t-distribution and its versions. A t-test compares the discrepancy of the average gene expression levels between the two samples or subgroups, given the noisiness of the data i.e. the difference in means between samples divided by the standard deviation. The genes are filtered in accordance with both fold change and *p value* criterion. Despite the fact that, methods to correct for multiple comparisons have been applicable for a long time such as Bonferroni correction, most of these methods are not appropriate to

analyze gene expression data sets [19]. We highlight statistical significance performing *t-test* by taking *p value* cutoff 0.01 and |Log2 (fold cut-off)| > 1.2 to identify down and up-regulated DEGs between each category understudy.

2.3. Differentially expressed genes and clustering analysis

Using GEOquery package in Bioconductor, gene expression values were pull out for each sample and converted to base-2 logarithmic scale using R language. We used gplots package of R to create heatmaps of DEGs with heatmap.2 function. Clustering analysis of DEGs was performed to compare the expression pattern of DEGs in each bi-group i.e. young-old aged, young-middle aged, and middle-old aged groups.

2.4. GO terms and analysis of the pathway enrichments

Expression measurements annotations for up-regulated and down regulated DEGs for each group probes mapped to gene names using Ensemble Biomart package in R. All of the DEGs were characterized by their biological processes, molecular functions, and cellular components of gene ontology (GO) enrichment of the database for Annotation and DAVID which stands for Visualization and Integrated Discovery [12,24]. All classified genes were cautiously examined and further parts like the Universal Protein resource, and physical properties Gene Ontology (GO) and annotation types were taken using Gorilla Gene Ontology Enrichment analysis and Visualization Tool, DAVID, and KEGG Kyoto Encyclopedia of Genes and Genomes [11,25]. We then compared the results of DAVID with NetworkAnalyst enrichments performed with KEGG [26,27].

2.5. Protein-Protein Interaction Network

NetworkAnalyst, publicly accessible on the web, provides analysis of PPI networks for single gene lists using STRING Interactome [28] . To comprehensively decipher the regulatory mechanisms in AML and related diseases, DEGs from young-old aged, young-middle aged, and middle-aged groups were analyzed to form a PPI network with previously reported GO classification and enrichment (Table 1).

3 Results

3.1. Experimental Data Analysis

With gene expression result of the GSE32719 data set, we detect differentially expressed genes (**DEGs**) in total 453 genes from **young-old, young-middle**, and **middle-old** aged groups which was demonstrated in volcano plot (Fig. 1). We find the down-regulated and up-regulated DEGs in each group. The expression values were pull out, and a heatmap was created to show the young-old**,** young-middle, and middle-old differentially expressed genes by groups. (Fig. 2). DEGs were selected with common *t* test, and labelled with |Log2 (FC)| >1.2 and $p < 0.01$ which is a the screening cut-off for each group. Numbers of up and down regulated gene expressions between young-old aged, young-middle aged, and middle-old aged groups shown in Table 1. Here, we detected 326 differentially expressed genes of up regulation, whereas we find 127 down-regulated gene. Although, we could not assert if the samples belonged to any specific gender, i.e., male or females.

3.2. Gene ontology (GO) and enrichment analysis

Table 2 demonstrates the significant enrichments of DEGs using biological processes (BP) involving vasculogenesis, leukocyte migration, oligosaccharide metabolic process, and defense response to virus. The significant enrichment of DEGs in cellular component (CC) contains DNA replication factor A complex, nucleoplasm, apical part of the cell for cell component. Finally, the significant enrichments GO terms in molecular function (MF) is revealed protein binding, GTPase activity and translation factor activity, RNA binding. KEGG signaling pathway study outcomes demonstrated that the DEGs were considerably enriched in hepatitis C pathway, arrhythmogenic right ventricular cardiomyopathy (ARVC), and hippo signaling pathway (Table 4). Further, hippo signaling pathway is vital in stem cell and tissue particular progenitor cell and genes self-regeneration and growth (Table 2).

3.3. GO pathways results of biological processes of DEGs from the groups between different aged groups

The most and most significant GO pathway results are conducted from the down-regulated DEGs between young and middle aged samples in Fig. 3(left). It demonstrates the most important pathways that play roles in HSCs between young-middle and young-old aged samples. The long lists of GO terms were outlined and pictured using REVIGO tool embedded R-script. Among 51 ontology terms, by visualizing in the scatter plot, 23 of the biological process *(BP)* terms were found, the most significant ones are c*ellular metabolic process, metabolic process,* and *negative regulation of cellular macromolecule biosynthetic process*. Similar to down-regulated DEGs, *BP terms* of the DEGs between young-old aged samples in Fig. 3(right) can be summarized as the following; *cellular metabolic process*, *cellular component organization* or *biogenesis*, and *negative regulation of multicellular organismal process.*

3.4. GO terms and analysis of the pathway enrichments

Expression measurements annotations for up-regulated and down regulated DEGs for each group probes mapped to gene names using Ensemble Biomart package in R. All of the DEGs were characterized by their biological processes, molecular functions, and cellular components of gene ontology (GO) enrichment of the database for Annotation and DAVID which stands for Visualization and Integrated Discovery. All classified genes were cautiously examined and further parts like the Universal Protein resource, and physical properties Gene Ontology (GO) and annotation types were taken using Gorilla Gene Ontology Enrichment analysis and Visualization Tool, DAVID, and KEGG Kyoto Encyclopedia of Genes and Genomes [11,25]. We then compared the results of DAVID with NetworkAnalyst enrichments performed with KEGG [26,27].

Fig.5 demonstrates PPI network of DEGs in different bi-groups. To conclude, PIP-network of young-middle group **(**Fig. 5B) is a sample of all the DEGs (Fig. 4 and Table 3) in the data set whereas STAT1 and E2F1 is common proteins for Fig. 5B and C. Main genes of young-old aged group in Fig. 5A can be listed as EZR, GJA1, IFIT1 , ZNRF3, and KDM6A. EZR gene is active in biological processes such as cell surface structure adhesion, migration and organization, and it has been involved in several human cancers. IFIT1 can be associated with hepatitis C and hepatitis. It involves in biological process such as *RNA binding*.

3.5. Analysis of the Hippo signaling pathway

The core of this project due to it is close association with cancer within all of the significantly (*p value* < 0.01) enriched pathways of DEGs is the hippo signaling pathway. There were 6 DEGs particularly engaged in this pathway, containing TCF7L2, PPP2R1B, PPP2R2B, PPP2R2D, BMPR1B, TEAD2 (Fig. 6 and Table 2 and 4). We have performed the primary DEGs associated with hippo signaling pathway in Fig. 6. We implemented hippo signaling pathway genes to construct Fig. 6A showing subnetwork 1 and Fig. 6B showing subnetwork 2 protein-protein interaction networks.

Associated genes with the DEGs of the data set enriched with the hippo signaling pathway deciphered as new hub genes. In subnetwork 1 of hippo signaling pathway (Fig. 6A), PPP2R1B is in the center and the most significant gene in terms of BC. Moreover, protein phosphatase 2 regulatory subunits gene family also performs a key position which is demonstrated in the subnetwork of the hippo signaling pathway. As shown in subnetwork 2, TEAD2 is a gene that is associated with hippo signaling pathway as previously reported by Zheng & Pan, 2019. In Fig. 6B associated proteins with TEAD2 in hippo signaling pathway of HSCs gene expression data set are revealed using NetworkAnalyst. Drug therapies such as verteporfin has side effects on cells in terms of toxicity, and stopping YAP/TAZ consistently utilizing little molecules could have fall outs in other cells and tissues.

We suggest an alternative approach on the basis of the structure of the NCOA1-TEAD compound. which has been previously formulated based on the organizations of YAP-TEAD and VGLL4-TEAD compounds. These outcomes confirm the key task of the hippo signaling pathway engaged in AML and related diseases treatment, offering new molecular therapeutic targets to improve fundamental drug agents.

4 Discussions

Even though many research have been conducted over the past 10 years to reveal the causes and potential mechanisms of acute myeloid lymphoma (AML) and related diseases, the result still remains bleak. While most cases arise among adults, AML is a widespread kind of leukemia detected in children and old aged people. AML accounts for 32% of all adult and 20% of all children leukemia cases [29].

Discovering specific biomarkers to advance early diagnosis of AML which is a main task in assisting patients the best possible result is vital. Equally essential, it is a requirement to determine and confirm novel molecular therapeutic targets to improve fundamental drug agents that may be prosperous in curing AML and AML related diseases. Healthy, hematologically normal young aged, middle aged and old human bone marrow specimens of hematopoietic progenitor populations are evaluated to uncover potential biomarkers that may affect age-related hematopoietic dysfunction in the elderly human hematopoietic system.

Astonishingly, most studies focus on an multiple genetic results that are derived from multiple sample studies through microarray analysis that are not uniform with each other [30–33]. Our study performed bi-group comparison profile from three category in which young aged, middle aged, and old aged samples within one data set and employed bioinformatics techniques to strongly analyze the data set and detected 453 DEGs in total. The number of down-regulated DEGs was importantly less than the up-regulated DEGs (127 versus 326).

Gene ontology analysis reveal information about these genes is used to identify function and disease associated with proteins. The highest interactive 13 proteins UBC, EP300, CREBBP, TP53, PTK2, E2F1, STAT1, NFYB, RAD18, TCEB1, and ANAPC5, VEGFA are predicted that are involved in several types of cancers like lung cancer, leukemia, breast cancer, glioma,

ovarian cancer, and colorectal cancer [34–36]. Some genes like GTF3A, STAT1, TFC7L2, and XIAP proteins are related with mental retardation, type 2 diabetes, and mycobacterial and viral infections [37–40]. These genes can be employed as biomarkers for diagnosis of early stage of diabetes mellitus and can also operate as promising drug targets for the drug to strengthen immunity to viral diseases [40–43].

This investigation further highlighted the hippo signaling pathway involving differentially expressed genes in a broad various kinds of cancer cells and immune cells [15]. The significant role of the hippo signaling network to cell life and tissue growth has been increasingly investigated in the last 20 years. Simultaneously, our knowledge of the entanglement of this network and its partners in other pathways has enhanced enormously, and hippo signaling pathway is currently thought an important integrator of tag from the biophysical and biochemical surroundings of cells [44]. Hippo signaling pathway is regulated by negative regulation of YAP transcription factors which is a protein partner of TEAD2. Drug therapies can disturb the connection between YAP proteins and TEAD, and inhibit expression of YAP genes target [45]. However, certain drug therapies such as verteporfin has high cytotoxicity, and stopping YAP/TAZ consistently utilizing little molecules could have fall outs in other cells and tissues. A different technique has been formulated based on the organizations of YAP-TEAD and VGLL4-TEAD complexes [46] .

Additional studies is required for clinical lab confirmation of predicted genes that are expressed in HSC data set and express at the developmental stage AML and related diseases. More research is needed in the field of cancer biology to detect AML and subset diseases at its early stage. This paper also emphasizes the importance of microarray experiment in comprehending AML and related diseases and approach to study several results of gene expression data, like differentially

expressed genes analysis, pathway and process identification, and protein-protein interaction network study.

Author Contributions

CONCEPTION: Emine Güven

INTERPRETATION OR ANALYSIS OF DATA: Emine Güven

PREPARATION OF THE MANUSCRIPT: Emine Güven and Sevinç Akçay

REVISION FOR IMPORTANT INTELLECTUAL CONTENT: Sevinç Akçay and Emine Güven

SUPERVISION: Emine Güven and Sevinç Akçay

Conflicts of Interest

The Authors declare no conflicts of interest.

Data and Methods Procedure Availability

The data sets used and analyzed in this present study are available in the NIH GEO (http://www.ncbi.nlm.nih.gov/geo) public repository. The following supplementary information available at https://github.com/mathbioGVN/GSE32719.HSC.microarray.project for data analysis codes, supplementary figures, and drafts of manuscript.

References

[1] Pinho S, Frenette PS. Haematopoietic stem cell activity and interactions with the niche. Nat Rev Mol Cell Biol. 2019;20(5):303–20.

[2] Morrison SJ, Scadden DT. The bone marrow niche for haematopoietic stem cells. Nature. 2014;505(7483):327–34.

[3] Dar A, Kollet O, Lapidot T. Mutual, reciprocal SDF-1/CXCR4 interactions between hematopoietic and bone marrow stromal cells regulate human stem cell migration and development in NOD/SCID chimeric mice. Exp Hematol. 2006;34(8):967–75.

[4] Fraga M, Esteller M. Fraga, M. F. & Esteller, M. Epigenetics and aging: the targets and the marks. Trends Genet. 23, 413-418. Trends Genet TIG. 2007 Sep 1;23:413–8.

[5] Perry JM, Li L. Disrupting the stem cell niche: good seeds in bad soil. Cell. 2007;129(6):1045–7.

[6] Kuranda K, Vargaftig J, de la Rochere P, Dosquet C, Charron D, Bardin F, et al. Age-related changes in human hematopoietic stem/progenitor cells. Aging Cell. 2011;10(3):542–6.

[7] Boyd AL, Reid JC, Salci KR, Aslostovar L, Benoit YD, Shapovalova Z, et al. Acute myeloid leukaemia disrupts endogenous myelo-erythropoiesis by compromising the adipocyte bone marrow niche. Nat Cell Biol. 2017;19(11):1336–47.

[8] Pang WW, Price EA, Sahoo D, Beerman I, Maloney WJ, Rossi DJ, et al. Human bone marrow hematopoietic stem cells are increased in frequency and myeloid-biased with age. Proc Natl Acad Sci U S A. 2011;108(50):20012–7.

[9] Roushangar R, Mias GI. Multi-study reanalysis of 2,213 acute myeloid leukemia patients reveals age-and sex-dependent gene expression signatures. Sci Rep. 2019;9(1):1–17.

[10] Appelbaum FR, Gundacker H, Head DR, Slovak ML, Willman CL, Godwin JE, et al. Age and acute myeloid leukemia. Blood. 2006;107(9):3481–5.

[11] Eden E, Navon R, Steinfeld I, Lipson D, Yakhini Z. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. BMC Bioinformatics. 2009;10(1):1–7.

[12] Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2009;4(1):44.

[13] Luppi M, Fabbiano F, Visani G, Martinelli G, Venditti A. Novel Agents for Acute Myeloid Leukemia. Cancers. 2018 Nov 9;10(11):429.

[14] Bohl SR, Bullinger L, Rücker FG. New Targeted Agents in Acute Myeloid Leukemia: New Hope on the Rise. Int J Mol Sci. 2019 Apr 23;20(8):1983.

[15] Zheng Y, Pan D. The Hippo Signaling Pathway in Development and Disease. Dev Cell [Internet]. 2019 Aug 5 [cited 2020 Oct 5];50(3):264–82. Available from: http://www.sciencedirect.com/science/article/pii/S1534580719305210

[16] Durinck S, Spellman PT, Birney E, Huber W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Nat Protoc. 2009;4(8):1184.

[17] Durinck S, Moreau Y, Kasprzyk A, Davis S, De Moor B, Brazma A, et al. BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. Bioinformatics. 2005;21(16):3439–40.

[18] Warnes GR, Bolker B, Bonebakker L, Gentleman R, Huber W, Liaw A, et al. gplots: Various R programming tools for plotting data. R Package Version. 2009;2(4):1.

[19] Tarca AL, Romero R, Draghici S. Analysis of microarray experiments of gene expression profiling. Am J Obstet Gynecol. 2006 Aug;195(2):373–88.

[20] Davis S, Meltzer P. GEOquery: A bridge between the Gene Expression Omnibus (GEO) and BioConductor. Bioinforma Oxf Engl. 2007 Aug 1;23:1846–7.

[21] Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B Methodol. 1995;57(1):289–300.

[22] Hochberg Y, Tamhane AC. Multiple comparison procedures. John Wiley & Sons, Inc.; 1987.

[23] Dudoit S, Shaffer JP, Boldrick JC. Multiple hypothesis testing in microarray experiments. Stat Sci. 2003;71–103.

[24] Huang DW, Sherman BT, Tan Q, Kir J, Liu D, Bryant D, et al. DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. Nucleic Acids Res. 2007;35(suppl_2):W169–75.

[25] Supek F, Bošnjak M, Škunca N, Šmuc T. REVIGO summarizes and visualizes long lists of gene ontology terms. PloS One. 2011;6(7):e21800.

[26] Xia J, Gill EE, Hancock REW. NetworkAnalyst for statistical, visual and network-based meta-analysis of gene expression data. Nat Protoc. 2015;10(6):823–44.

[27] Zhou G, Soufan O, Ewald J, Hancock REW, Basu N, Xia J. NetworkAnalyst 3.0: a visual analytics platform for comprehensive gene expression profiling and meta-analysis. Nucleic Acids Res. 2019;47(W1):W234–41.

[28] Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, et al. STRING v10: protein–protein interaction networks, integrated over the tree of life. Nucleic Acids Res. 2015;43(D1):D447–52.

[29] Shallis RM, Wang R, Davidoff A, Ma X, Zeidan AM. Epidemiology of acute myeloid leukemia: Recent progress and enduring challenges. Blood Rev. 2019;36:70–87.

[30] Takahashi K, Wang F, Morita K, Yan Y, Hu P, Zhao P, et al. Integrative genomic analysis of adult mixed phenotype acute leukemia delineates lineage associated molecular subtypes. Nat Commun. 2018;9(1):1–12.

[31] Haferlach T, Kohlmann A, Bacher U, Schnittger S, Haferlach C, Kern W. Gene expression profiling for the diagnosis of acute leukaemia. Br J Cancer. 2007;96(4):535–40.

[32] Shivarov V, Bullinger L. Expression profiling of leukemia patients: key lessons and future directions. Exp Hematol. 2014;42(8):651–60.

[33] Figueroa ME, Wouters BJ, Skrabanek L, Glass J, Li Y, Erpelinck-Verschueren CA, et al. Genome-wide epigenetic analysis delineates a biologically distinct immature acute leukemia with myeloid/T-lymphoid features. Blood J Am Soc Hematol. 2009;113(12):2795–804.

[34] Tang Y, Geng Y, Luo J, Shen W, Zhu W, Meng C, et al. Downregulation of ubiquitin inhibits the proliferation and radioresistance of non-small cell lung cancer cells in vitro and in vivo. Sci Rep. 2015;5(1):9476.

[35] Attar N, Kurdistani SK. Exploitation of EP300 and CREBBP lysine acetyltransferases by cancer. Cold Spring Harb Perspect Med. 2017;7(3):a026534.

[36] Rasheed BKA, McLendon RE, Herndon JE, Friedman HS, Friedman AH, Bigner DD, et al. Alterations of the TP53 gene in human gliomas. Cancer Res. 1994;54(5):1324–30.

[37] Glorioso C, Oh S, Douillard GG, Sibille E. Brain molecular aging, promotion of neurological disease and modulation by Sirtuin5 longevity gene polymorphism. Neurobiol Dis. 2011;41(2):279–90.

[38] Grant SFA, Thorleifsson G, Reynisdottir I, Benediktsson R, Manolescu A, Sainz J, et al. Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of type 2 diabetes. Nat Genet. 2006;38(3):320–3.

[39] Wu H, Panakanti R, Li F, Mahato RI. XIAP gene expression protects β-cells and human islets from apoptotic cell death. Mol Pharm. 2010;7(5):1655–66.

[40] Durbin JE, Hackenmiller R, Simon MC, Levy DE. Targeted disruption of the mouse Stat1 gene results in compromised innate immunity to viral disease. Cell. 1996;84(3):443–50.

[41] Kamiyama N, Soma R, Hidano S, Watanabe K, Umekita H, Fukuda C, et al. Ribavirin inhibits Zika virus (ZIKV) replication in vitro and suppresses viremia in ZIKV-infected STAT1-deficient mice. Antiviral Res. 2017;146:1–11.

[42] Redondo MJ, Geyer S, Steck AK, Sosenko J, Anderson M, Antinozzi P, et al. TCF7L2 genetic variants contribute to phenotypic heterogeneity of type 1 diabetes. Diabetes Care. 2018;41(2):311–7.

[43] Groves CJ, Zeggini E, Minton J, Frayling TM, Weedon MN, Rayner NW, et al. Association analysis of 6,736 UK subjects provides replication and confirms TCF7L2 as a type 2 diabetes susceptibility gene with a substantial effect on individual risk. Diabetes. 2006;55(9):2640–4.

[44] Misra JR, Irvine KD. The Hippo signaling network and its biological functions. Annu Rev Genet [Internet]. 2018 Nov 23 [cited 2020 Oct 8];52:65–87. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6322405/

[45] Liu-Chittenden Y, Huang B, Shim JS, Chen Q, Lee S-J, Anders RA, et al. Genetic and pharmacological disruption of the TEAD–YAP complex suppresses the oncogenic activity of YAP. Genes Dev. 2012;26(12):1300–5.

[46] Jiao S, Wang H, Shi Z, Dong A, Zhang W, Song X, et al. A peptide mimicking VGLL4 function acts as a YAP antagonist therapy against gastric cancer. Cancer Cell. 2014;25(2):166–80.

Tables

Table 1. DEGs by groups and the number of genes.

groups compared	down-regulated DEGs	up-regulated DEGs
young-old	29	70
young-middle	59	232
middle-old	39	24

Rows represent comparison of groups within samples whereas columns portray down-regulated and up-regulated DEGs.

Table 2. HSCs gene expression data set retrieved with top significant pathways GO enrichment analysis of DEGs

Category	Term	Count	*p*-value	Genes
BP	GO:0001570~vasculogenesis	5	0.01	GJC1, ACKR3, PTK2, TEAD2, VEGFA
BP	GO:0050900~leukocyte migration	7	0.011	SELP, CXADR, ITGA4, PDE4B, SIRPA, SLC16A3, NKX2-3
BP	GO:0051607~defense response to virus	7	0.04	DNAJC3, RNASEL, CXADR, OAS1, STAT1, NCBP3, IFIT1
BP	GO:0007507~heart development	7	0.06	GJA1, CXADR, LOX, TRPS1, XIRP2, BCOR, FBN1
BP	GO:0009311~oligosaccharide metabolic process	3	0.056	ST6GAL2, ST8SIA4, ST3GAL6
BP	GO:0007159~leukocyte cell-cell adhesion	3	0.056	SELP, ITGA4, EZR
BP	GO:0009615~response to virus	4	0.09	GJA1, STMN1, TRIM13, GNG11
BP	GO:0030198~extracellular matrix organization	7	0.084	ITGA4, LOX, ABI3BP, ITGA2, SPP1, PTK2, FBN1
CC	GO:0005662~DNA replication factor A complex	4	1.4E-3	PURB, PURA, RPA4, ERCC5
CC	GO:0005654~nucleoplasm	60	0.7E-2	NUMA1, C2ORF88, ZBTB20, EFCAB13, AHR, CLINT1, PTPDC1, CDC14A, DCAF7, CDC14B, SPRED1, METTL14, TRPS1, NAMPT, UBXN7, TEAD2, C8ORF44, KDM6A, FNBP4, CXADR, RMI1, GTF3A, HCFC2, CDC25A, EMSY, SGO2, ZC3H11A, MORF4L2, BMP2K, AAGAB, ANAPC5, MCTP2, CASZ1, XIAP, NMD3, TGOLN2, CAND1, NAA25, E2F1, SRSF11, SCAI, TCF7L2, RBM39, NFYB, STAT1, PCIF1, ANKRD23, BTBD8, SMARCA2, SCAF4, FOSL2, SELP, RAD52, NFIA, RPA4, ZNF217, ERCC5, NANOG, RAD18, FERMT2
CC	GO:0005925~focal adhesion	14	0.8E-2	TPM4, ITGA4, ITGA2, LPP, TRIOBP, PTK2, GJA1, NFASC, NCSTN, CSRP2, ATP6V0A2, P4HB, EZR, FERMT2
CC	GO:0043197~dendritic spine	6	1.9E-2	FARP1, ARHGAP32, PDE4B, CRIPT, STRN, SHANK2
CC	GO:0045177~apical part of cell	5	2.9E-2	SRR, NUMA1, FAT4, EZR, ATP6V1C1
KEGG	hsa05160:Hepatitis C	7	1.4E-2	RNASEL, OAS1, PPP2R1B, PPP2R2B, STAT1, PPP2R2D, IFIT1

KEGG	hsa04390:Hippo signaling pathway	6	7.5E-2	TCF7L2, PPP2R1B, PPP2R2B, PPP2R2D, BMPR1B, TEAD2
KEGG	hsa05412:Arrhythmogenic right ventricular cardiomyopathy (ARVC)	4	7.8E-2	TCF7L2, GJA1, ITGA4, ITGA2

Abbreviations: GO: gene ontology; BP: biological process; CC: cell component; KEGG: Kyoto Encyclopedia of Genes and Genomes (p-value<0.05)

Table 3. Top 15 genes of PIP network of DEGs in HSCs gene expression data.

Gene ID	Gene Name	ND	BC
6772	STAT1	84	175186.36
1869	E2F1	56	91229.24
57381	RHOJ	51	57433.11
5893	RAD52	50	40048.94
23097	CDK19	45	45505.58
26043	UBXN7	43	54800.8
54552	GNL3L	43	50432.01
993	CDC25A	41	59535.81
7430	EZR	38	97757.97
331	XIAP	38	63169.02
6595	SMARCA2	34	24992.22
7316	UBC	33	340166.09
6801	STRN	31	19748.33

Abbreviations: BC: Betweenness Centrality; ND: Node Degree

Table 4 Top 10 most excessive KEGG pathway enrichment analysis of global DEGs in HSCs micro-array gene expression data set.

Pathway	Gene Count	*p-value*
Hepatitis C	155	0.000445
ARVC	72	0.0214
Hippo signaling pathway	154	0.0259
Sphingolipid signaling pathway	119	0.0308
One carbon pool by folate	20	0.0344
Cell cycle	124	0.0359
N-Glycan biosynthesis	50	0.037
Small cell lung cancer	93	0.048

Collecting duct acid secretion	27	0.0595
Cell adhesion molecules (CAMs)	146	0.0641

Figure Captions

Fig. 1. Plots displaying the gene expression discrepancy in young-old, young-middle and middle-old comparison. Black illustrates no change (NO), blue illustrates down-regulated (Down), and red illustrates up-regulated (Up) DEGs. logYO, logYM, and logMO on the x-axis labels represents : Log2 fold changes respectively.

Fig. 2. Heat map demonstrating DEGs in (A) **young-old** and (B) **young-middle** (C) **middle-old** aged groups. Each columns present samples of aged groups, and rows present genes. Base-2 logarithmic values of the gene expression data are calculated. The progressive color changing from red to green represents the ranging from up to down-regulated DEGs.

Fig. 3. Scatter plots done with REVIGO that shows GO analyses in biological process from (A) down-regulated DEGs in the young-middle and (B) all the DEGs in the young-old aged groups. Relevant GO terms are aggregated and presented in bubbles of similar hues. Bubble colors show *p values*, and bubble sizes demonstrate relative frequency of GO terms.

Fig. 4. PPI network of DEGs was created and pictured using NetworkAnalyst. UBC with the largest Betweenness Centrality (BC) was suggested to be central to the PPI network associated with HSCs gene expression data. Moreover, STAT1, PTK2, E2F1, RHOJ, RAD52, and CDK19 with the secondary highest degree and VEGFA with the secondary BC might be involved in the development AML and other substance diseases.

Fig. 5. PPI networks of (A) young-old group DEGs (B) young-middle group DEGs and finally (C) middle-old aged DEGs was constructed and pictured using NetworkAnalyst. From light orange to red colors represent |Log2 (FC)| expression value for the DEGs in each bi-group. We can conclude from PIP-network of young-old comparison (B) reflects behavior of whole DEGs in the data set whereas STAT1 and E2F1 is common proteins for (B) and (C).

Fig. 6. Hub genes associated with hippo signaling pathway. (A) Subnetwork 1 and (B) Subnetwork 2 protein-protein interaction networks.

Fig. 1

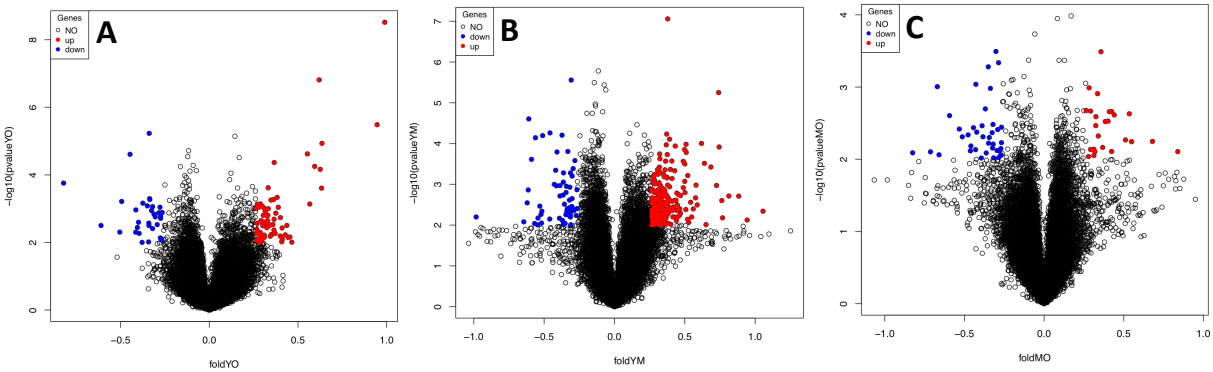

A
Genes
NO
up
down
−log10(pvalueYO)
foldYO
B
Genes
NO
down
up
−log10(pvalueYM)
foldYM
C
Genes
NO
down
up
−log10(pvalueMO)
foldMO

Fig. 2

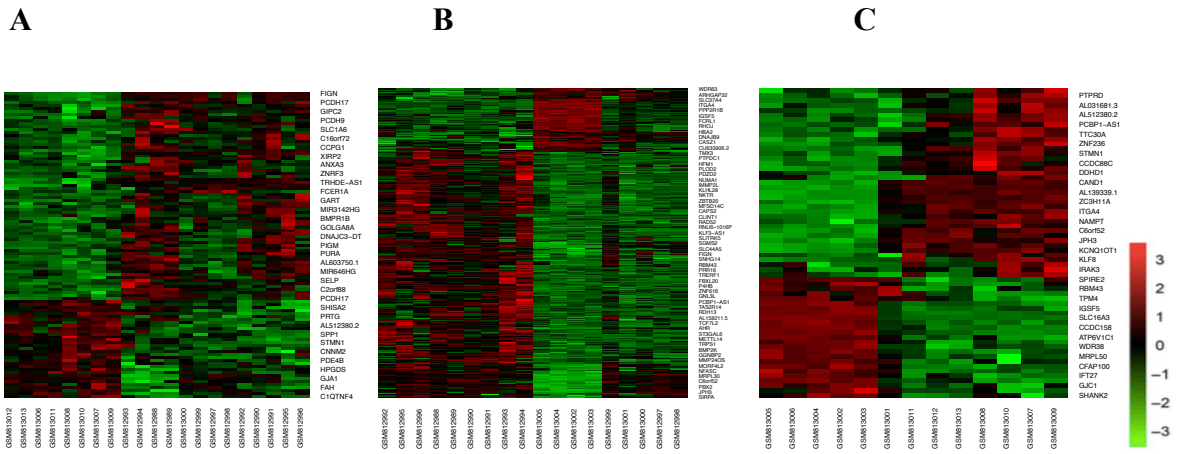

A
B
C
FIGN
PCDH17
GIPC2
PCDH9
SLC1A6
C16orf72
CCPG1
XIRP2
ANXA3
ZNRF3
TRHDE−AS1
FCER1A
GART
MIR3142HG
BMPR1B
GOLGA8A
DNAJC3−DT
PIGM
PURA
AL603750.1
MIR646HG
SELP
C2orf88
PCDH17
SH5A2
PRTG
AL512380.2
SPP1
STMN1
CNNM2
PDE4B
HPGDS
GJA1
FAH
C1QTNF4
PTPRD
AL031681.3
AL512380.2
PCBP1−AS1
TTC30A
ZNF236
STMN1
CCDC88C
DDHD1
CAND1
AL139339.1
ZC3H11A
ITGA4
NAMPT
C6orf52
JPH3
KCNQ1OT1
KLF8
IRAK3
SPIRE2
RBM43
TPM4
IGSF5
SLC16A3
CCDC158
ATP6V1C1
WDR38
MRPL50
CFAP100
IFT27
GJC1
SHANK2
3
2
1
0
−1
−2
−3

Fig. 3

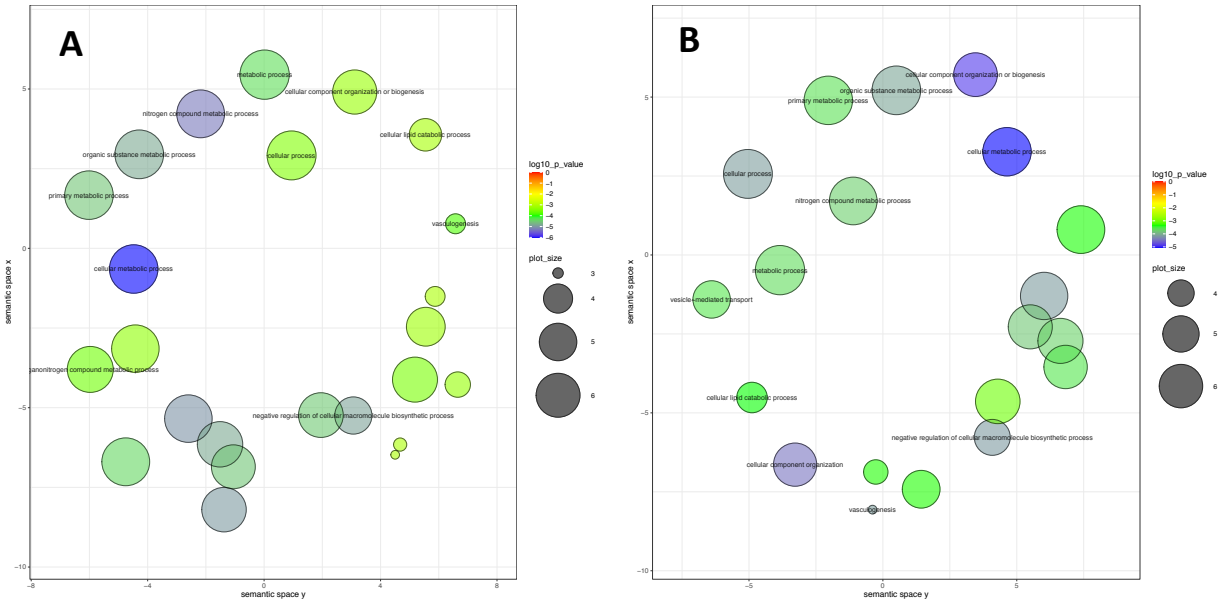

A
metabolic process
cellular component organization or biogenesis
nitrogen compound metabolic process
cellular lipid catabolic process
organic substance metabolic process
cellular process
primary metabolic process
vasculogenesis
cellular metabolic process
anonitrogen compound metabolic process
negative regulation of cellular macromolecule biosynthetic process
semantic space x
semantic space y
log10_p_value
plot_size
B
cellular component organization or biogenesis
organic substance metabolic process
primary metabolic process
cellular metabolic process
cellular process
nitrogen compound metabolic process
metabolic process
vesicle-mediated transport
cellular lipid catabolic process
negative regulation of cellular macromolecule biosynthetic process
cellular component organization
vasculogenesis
semantic space x
semantic space y
log10_p_value
plot_size

Fig. 4

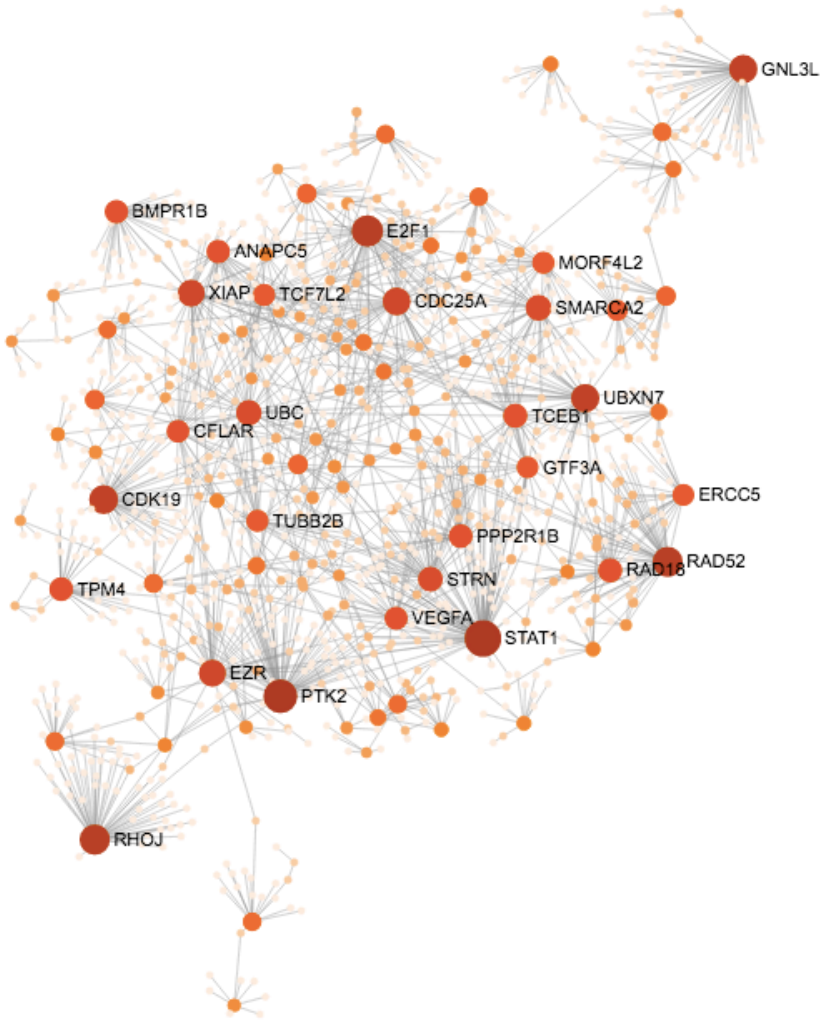

GNL3L
BMPR1B
E2F1
ANAPC5
MORF4L2
XIAP
TCF7L2
CDC25A
SMARCA2
UBXN7
UBC
TCEB1
CFLAR
GTF3A
ERCC5
CDK19
TUBB2B
PPP2R1B
RAD18
RAD52
TPM4
STRN
VEGFA
STAT1
EZR
PTK2
RHOJ

Fig. 5

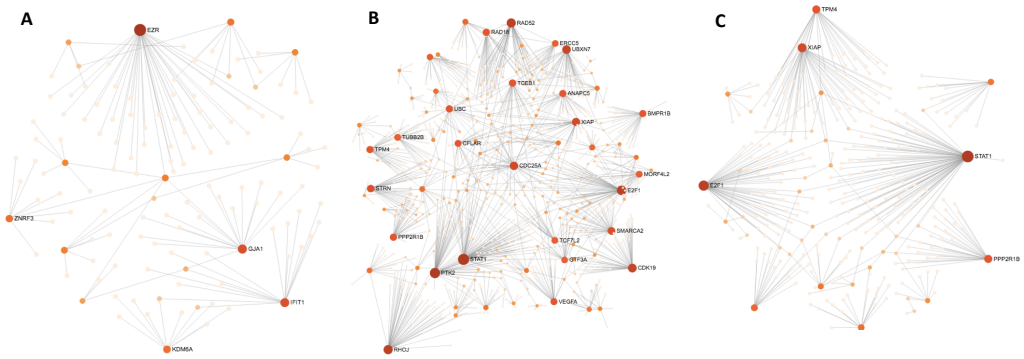
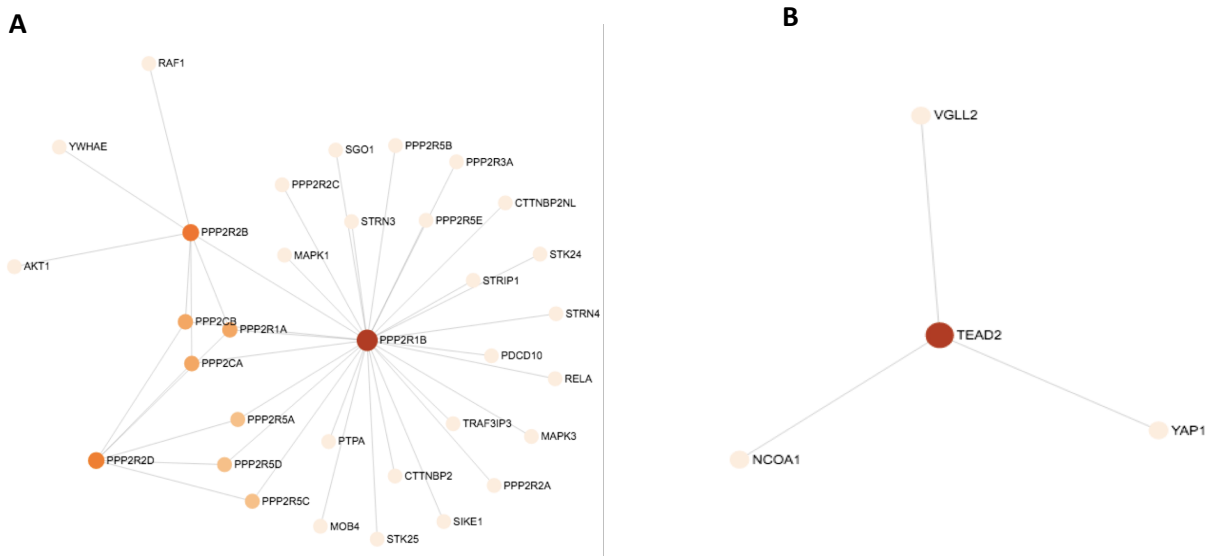

A
EZR
ZNRF3
GJA1
IFIT1
KDM6A
B
RAD52
RAD18
ERCC5
UBXN7
TCEB1
ANAPC5
UBC
BMPR1B
XIAP
TUBB2B
TPM4
CFLAR
CDC25A
MORF4L2
STRN
E2F1
PPP2R1B
SMARCA2
TCF7L2
STAT1
BRF1A
CDK19
PTK2
VEGFA
RHOJ
C
TPM4
XIAP
STAT1
E2F1
PPP2R1B

Fig. 6


A
RAF1
YWHAE
SGO1
PPP2R5B
PPP2R3A
PPP2R2C
CTTNBP2NL
STRN3
PPP2R5E
PPP2R2B
MAPK1
STK24
AKT1
STRIP1
STRN4
PPP2CB
PPP2R1A
PPP2R1B
PDCD10
PPP2CA
RELA
PPP2R5A
TRAF3IP3
MAPK3
PTPA
PPP2R2D
PPP2R5D
CTTNBP2
PPP2R2A
PPP2R5C
SIKE1
MOB4
STK25
B
VGLL2
TEAD2
YAP1
NCOA1