Embedding Multiomics Data for Correlation Network Construction and Pathway Analysis

Chung Suhwan, Wang Conghao, Ong Zhi Lin, Charlene

ABSTRACT

Background The learning of embeddings has been an effective method of dimensionality reduction and to illustrate the relationships between patients and multi-omics data in multiple fields of cancer research, However, there is limited research in breast cancer that utilises learnt embeddings from the gene expression for correlation network construction and pathway analysis. In this study, we demonstrated that multi-omics embeddings could be used for correlation network construction and pathway analysis in breast invasive carcinoma, the most common breast cancer diagnosis.

Methods This study implements embeddings of gene using samples of gene expression (GE), micro-RNA (miRNA) data. After learning gene expression, miRNA and sample embeddings of multi-omics dataset, the study performs correlation network construction and pathway analysis to identify enriched biological processes and pathways that comprised of associated genes related to breast cancer.

Results Co-expression networks were constructed based on gene and miRNA embeddings. Then we detected the hub genes and miRNAs in each module of the network, and discovered some entities such as mir-100 and mir-16 that are believed to play a significant role in breast cancer. Eventually, pathway analysis was performed on both modules and the whole gene set, and a wealth of cell regulation, protein process, and immune system related pathways were discovered to be enriched.

INTRODUCTION

The recent advances in high-throughput screening technologies provided an unprecendented amount of molecular data. In particular, one of the biggest databases is The Cancer Genome Atlas (TCGA) project [18.19], which is a valuable source of multi-omics data, and includes datasets from breast invasive carcinoma. Through the analysis of multi-omics datasets, novel biomarkers or interesting insights about the pathways could be gleaned. At the same time, there is also an advancement in machine learning, and in particular the learning of embeddings using machine learning methodologies. However, the use of these technologies such as artificial neural networks have rarely been applied to learn representations of multi-omics data. Given the success of artificial neural networks for dimensionality reduction and to learn relations in other domains, we hypothesize that the use of embeddings would assist in revealing biologically relevant insights pertaining to biological pathways. Through characterizing multi-omics data as embeddings, we explored the utility of embeddings in correlation network construction and pathway analysis.

DATASET

The dataset currently consists of multi-omics data from The Cancer Genome Atlas Breast Invasive Carcinoma (TCGA-BRCA), including gene expression (GE) and micro-RNA (miRNA) expression, as shown in Table 1.

	Sample amount	Feature dimension
Gene expression	1222	60483
miRNA expression	1207	1881

Table 1 Dataset of TCGA-BRCA

Despite that there are known biomarker discovery methods involving a number of these omics types, pathway analysis methods always integrate no more than two omics types (e.g., GE + ME, or GE + miRNA). Thus we plan to start with GE and miRNA at first so that we can take advantage of existing pathway analysis software. Our GE data are measured by RNA-Seq, including raw read counts and Fragments per Kilobase of transcript per Million mapped reads (FPKM). miRNA data also comprise raw read counts and normalized counts in reads-per-million miRNA mapped.

METHODS

DATA PRE-PROCESSING

miRNA expression counts were filtered as miRNA with 0 values in 90% of the samples were removed. After filtering, the number of miRNA was reduced from 1881 to 975. Then the filtered counts were normalized with Counts per Million (CPM) method, and log2 transformed.

RNA-Seq read counts are used as gene expression data in this project. As there are over 60k genes in the raw

counts data, we decided to perform gene selection with DESeq2 package. After selecting FDR-adjusted p-value < 0.05, 11,532 genes were preserved. Then counts of the selected genes were normalized and log2 transformed.

EMBEDDINGS

Collaborative filtering

Despite the demonstrated applications of collaborative filtering (CF) from lots of research, there is relatively lack of CF implementations that use neural networks approach. The most frequently implemented approach is memory-based technique based on cosine similarity or correlation coefficients that calculate proximity of metrics of data. Non-parametric machine learning techniques that use such algorithms as KNN come under the memory-based approach. In this study, we implemented a shallow artificial neural network for CF to learn parameters via gradient descent.

For our implementation of neural network based CF, we used 1207 samples of 11,532 gene counts and 975 miRNA matrices from pre-processed data. This study implements models using the Fastai package in Python in Linux server environment which is hosted on Nanyang Technological University server.

In implementing model training, we create a learner for collaborative filtering which binds the 75% of the training data with EmbeddingDotBias model which is identical to the SVD model. The model is trained in a way to minimise mean squared error. The sizes of embeddings vectors are set to 20 and 50 for miRNA and GE respectively. Learning rate of the training model is defined by a bit of trial and error and all models are trained with 30 epochs.

Visualizing embedding dimensions

To visualise high-dimensional embedding matrices, we use Python's Scikit-Learn package to implement principal component analysis (PCA) and t-distributed stochastic neighbor embedding (t-SNE) that are well-suited to model high-dimensional genes. We set perplexity of five and 1,000 iterations for t-SNE and x components (n_components=x) for PCA.

CO-EXPRESSION NETWORK CONSTRUCTION

Weighted gene co-expression network analysis (WGCNA) technique is known to have capabilities to define clusters, intramodular hubs, and network nodes to illustrate relationships between co-expression. WGCNA uses network methodologies, hence the technique is well-suited for integration of complementary genomic datasets. Our study is based on the integration of multi-omics datasets including miRNA and RNA sequence from TCGA breast cancer database, therefore WGCNA approach is an appropriate approach in performing pathway analysis.

There are relatively few previous research efforts that implemented weighted gene co-expression network analysis (WGCNA) techniques to perform pathway analysis (PA). According to Kao, et al (2017)[6], most of the pathway analysis approaches are based on Genome-Wide association Study (GWAS) which uses single nucleotide polymorphism (SNP). Unfortunately, SNP data is unavailable in TCGA breast cancer dataset. Other than GWAS, WGCNA is also known to be used for single cell RNA-seq and miRNA datasets and capable of constructing gene co-expression networks with gene expression dataset. In such a network, genes being observed with co-expression pattern are connected with edges (distances between genes are calculated by Pearson Correlation Coefficient) and similar genes are clustered into modules.

In our study, we propose a novel approach to build gene co-expression networks with WGCNA based on gene embeddings and miRNA embeddings. Conventional pathway analysis methods via WGCNA depend on the expression matrix over all samples for network construction. However, we believe our embedding model developed by CF is capable of distilling the significant statistical information from the expression matrix and will enable us to identify biologically significant modules. Consequently, our co-expression networks were built based on gene and miRNA embeddings, respectively.

Gene co-expression network

Firstly, we selected an appropriate soft-thresholding power to endow our network with scale-freeness. Correlations between genes were computed accordingly and transformed into a topological overlap matrix (TOM). The corresponding dissimilarity was calculated as 1-TOM. Then we produced the gene clustering trees with the hierarchical clustering algorithm. At last, we employed the dynamic tree cut algorithm to identify the

modules with similar expression profiles and thus merged these modules.

miRNA co-expression network

miRNA co-expression network was constructed in a very similar way with gene co-expression network. The major difference was that we did not apply the tree cut algorithm to merge any modules, since the number of miRNA is extremely less than the number of genes and relatively less modules were detected in the first place.

PATHWAY ANALYSIS

Eventually, we performed pathway enrichment analysis on the clustered hub genes and miRNAs in modules to detect the biological functions. After obtaining pathway terms and definitions from gene ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) database, this study applies hypergeometric test to detect enriched entries. After identification of miRNA-enriched modules, we detect miRNA's target genes in each module and testify with the known database recording miRNAs.

OUTCOMES

EMBEDDING VISUALIZATION

In this study, we apply principal component analysis (PCA) on raw log 2 expression across miRNA and RNA-seq datasets to compare with embeddings. In our implementation of PCA, we obtain component matrices from 20-component for miRNA and 50-component for RNA-seq, which are in turn projected into three dimensional space.

	miRNA	RNA-seq
PCA	20-component	50-component
t-SNE embedding	perplexity of 5 and 1000 iterations	

Table 2 | PCA and t-SNE embedding methods

As illustrated in Table 2, we obtain the same component matrix from 50-component PCA on RNA-seq and 20-component PCA on miRNA embedding matrix to understand detailed information from gene expression data and from different methods. As a result of the component matrices, we project sample entity matrices and vitality status into 3-dimensional space and color coded according to the vitality status (alive: green, dead: blue).

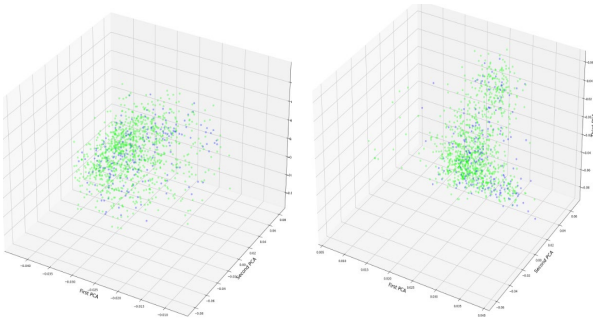

Figure 1. 3D PCA plot of sample embeddings (Left: miRNA, Right: RNA-seq)

As shown in Figure 1, the 3-dimensional plot of sample embeddings of miRNA and RNA-seq matrices do not reveal distinct clusters for samples of vital status. On the other hand, samples with the same vitality type have shown clusters from 3-dimensional t-SNE embeddings plot in Figure 2.

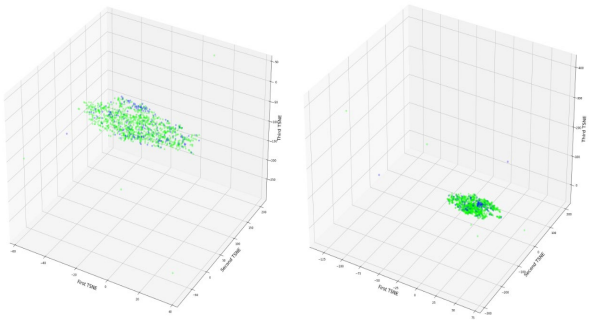

Figure 2. 3D t-SNE plot of sample embeddings (Left: miRNA, Right: RNA-seq)

Using learnt embeddings, we investigated the semantic of biological relationship between GE and RNA-seq data. [1]. Figure 3 illustrates 3D PCA projection of learnt embeddings from miRNA and RNA-seq. The relationship between genes from miRNa and RNA-seq is illustrated by distance in entity space of the 3-dimensional plot, which shows distinct clustering of a subset of genes.

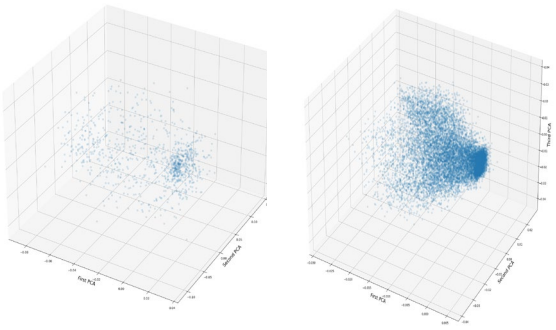

Figure 3. 3D PCA gene embeddings (Left: miRNA, Right:RNA-seq)

In order to deep dive into the embedding dimensions further, this study generates cluster maps to identify how different vital status can be differentiated in embedding dimension. Each row of the cluster map denotes a patient's embedding. Patients which have a vital status=0, i.e. alive, are denoted as dark blue while patients with a vital status=1, i.e. dead, are denoted as light blue.

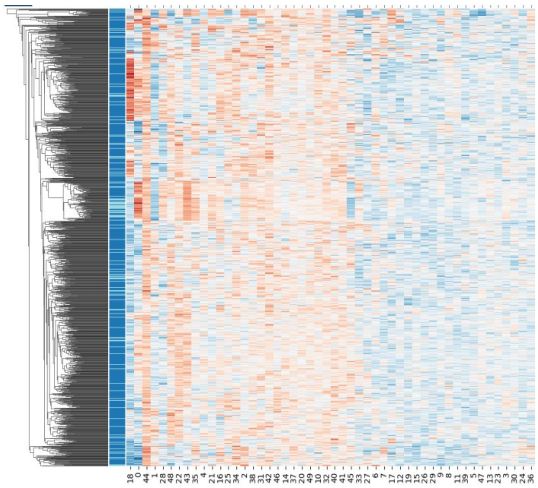

Figure 4. Sample cluster map (miRNA)

CO-EXPRESSION NETWORK CONSTRUCTION

Taking advantage of WGCNA, co-expression networks were constructed on gene expression and miRNA expression data, respectively.

To build a scale-free network, we applied WGCNA's network topology analysis function and selected the best soft-thresholding power for network construction.

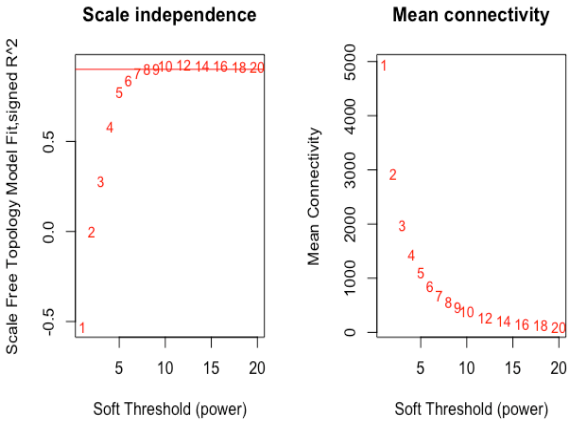

Figure 5. Relationship between soft-thresholding power and network properties

Figure 5 shows how different soft-thresholding powers affect the network built on gene expression data. When the power is 7, the network almost meets scale-freeness and mean connectivity becomes stable. Therefore, we chose 7 as the soft-thresholding power to calculate the adjacency in our gene co-expression network. 2 modules were merged when we set the height cut to be 0.25. As shown in Figure 6, we obtained 9 modules after merging, i.e. black, blue, brown, green, grey, magenta, pink, turquoise, yellow. There are over 9,000 genes in the largest turquoise module, and 92 genes in the smallest magenta module.

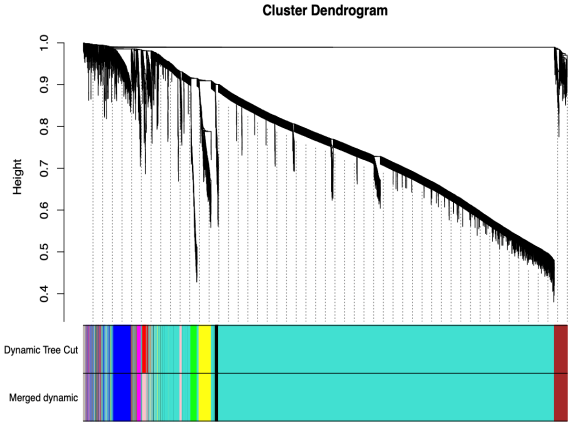

Figure 6. Cluster Dendrogram of gene network

As for miRNA co-expression network, dynamic clustering tree cut and module merging were dismissed. Selecting the soft-thresholding power to be 5, a total of 8 modules were detected.

Then we detected the salient genes in each module with module membership over 0.8 as hub genes. Top 30 hub

genes with TOM correlation over 0.3 in gene co-expression modules are visualized using VisANT software [12]. Only turquoise, green, black and yellow modules possess such correlated hub genes, as shown in Figure 7.

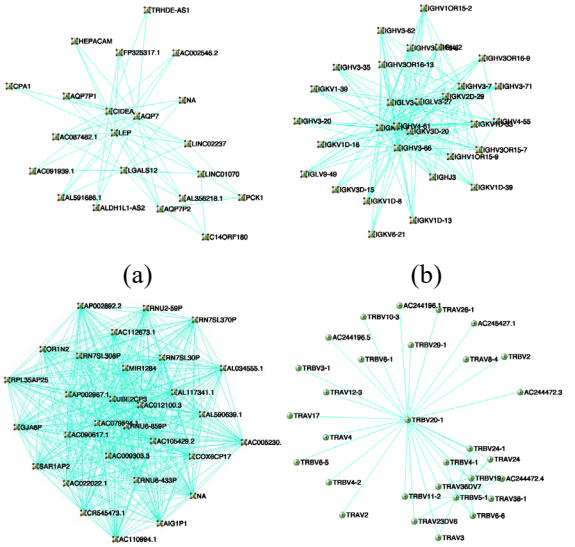

Figure 7. Topological interaction network of hub genes. (a) Black module; (b) Green module; (c) Turquoise module; (d) Yellow module

GO ENRICHMENT ANALYSIS ON MODULES

We performed GO and KEGG enrichment analysis with StringDB [13] and clusterProfiler [14] packages on all the modules identified in the gene co-expression network, respectively. We found that the amount of enriched pathways was not related to the number of genes in certain modules. For instance, despite that pink module gathered only 333 genes, it owned the most enriched GO terms up to 611, while the turquoise module who possessed over 9,000 genes only had 200 enriched GO terms. No pathways were discovered to be enriched in the grey module.

After visualizing with the aid of Enrichplot [15], we found that different modules possessed enriched pathways of distinct functions. For example, the yellow module was abundant in pathways regarding immunoregulation, and the magenta module was enriched with immunocyte regulation pathways. But some modules were not found to possess enriched pathways related to breast cancer. As we performed the GO enrichment analysis from three aspects, i.e., cellular component (CC), biological process (BP), and molecular function (MF), the potential pathways that most likely give rise to breast cancer were found to be enriched in the brown module from BP's perspective. The specific pathways are shown in Figure 8.

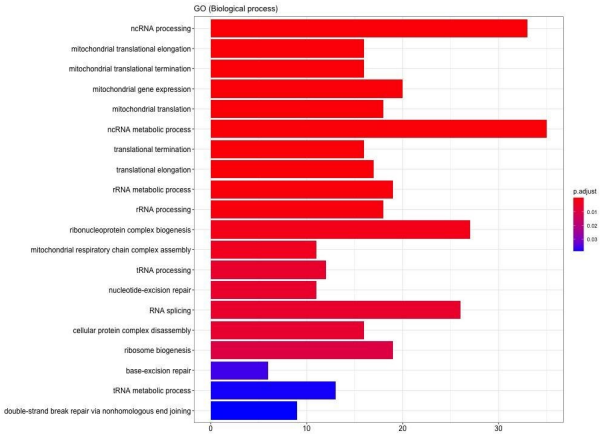

Figure 8. Enriched BP GO pathways in brown module

We also plotted the protein-protein interaction (PPI) figure of the brown module, which is presented in Figure 9. In the PPI network, the green halo represents downregulated genes.

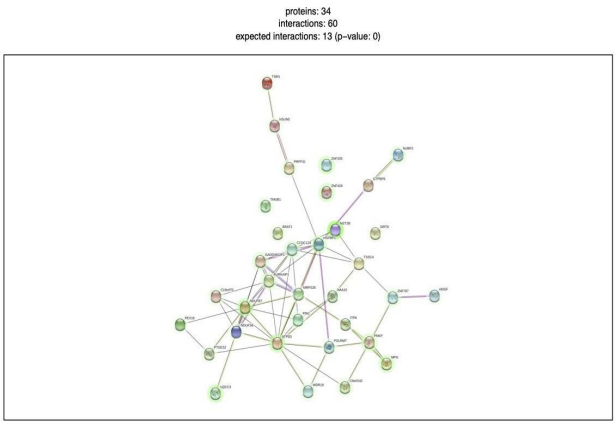

Figure 9. PPI network of brown module

GSEA ON WHOLE GENE SET

From the GO and KEGG enrichment analysis implemented on modules clustered by WGCNA, potential breast cancer-related pathways were merely found in the brown module. In order to testify the pathway enrichment condition, we then performed GSEA on the whole gene set with the log2 fold change values computed by DESeq2.

Figure 10 depicts the top GO pathways and their enrichment condition in the whole gene set. Top enriched

terms include regulation of cell communication and signaling, protein associated biological process, immune system regulation, etc. Figure 11 visualizes GSEA results of the top GO pathways enriched in our pre-ranked gene set.

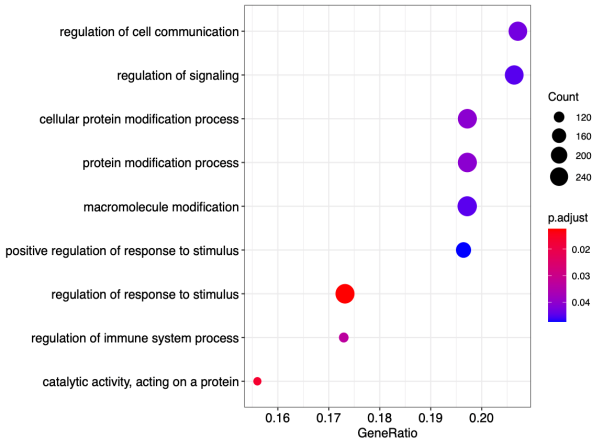

Figure 10. Dot plot of enriched GO pathways in the whole gene set.

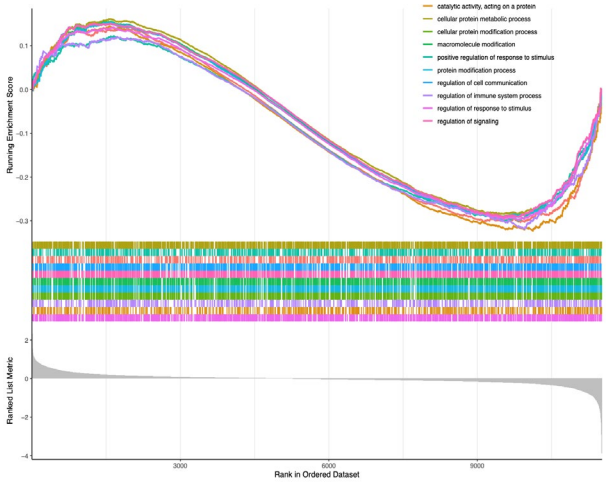

Figure 11. GSEA running scores and ranked list metrics of top 10 GO pathways.

We also employed GSEA for KEGG pathway analysis. However, unlike GO terms, there is only one KEGG pathway being identified as an enriched term on our gene set, which is Endocytosis (map04144). The GSEA result is shown in Figure 12.

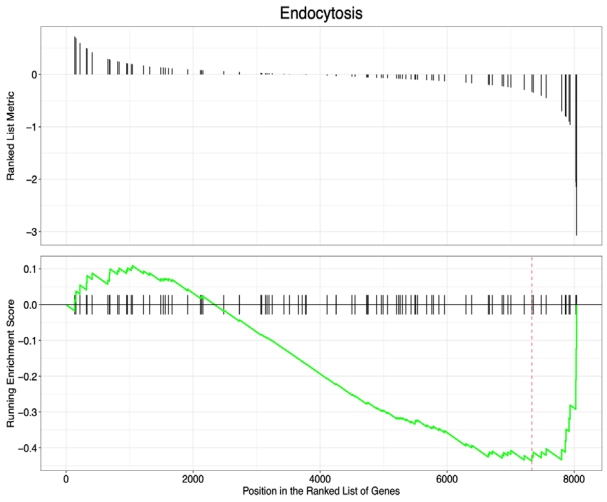

Figure 12. GSEA running scores and ranked list metrics of Endocytosis

miRNA CO-EXPRESSION NETWORK ANALYSIS

With the aid of WGCNA, a miRNA co-expression network was constructed with 8 modules. Specifically, the largest turquoise module gathered 249 miRNAs, grey, brown and blue modules gathered around 140 miRNAs respectively, etc. Then we detected the hub miRNAs in each module to further explore the role of miRNA in breast cancer.

As there are not as many useful tools for miRNA analysis as RNA-seq data analysis. We manually compared the identified hub miRNA in all the modules with the miRNA listed in several credible literature [16] [17] that are believed to play a paramount role in breast cancer. Table 3 illustrates the information of 11 miRNAs that not only serve as hub entities in the co-expression network, but also have a significant impact on breast cancer.

CONCLUSION

Through this study, we demonstrated that the use of embeddings could assist in detecting significant entities such as mir-100 and mir-16, which is believed to play a significant role in breast cancer. The multi-momics embeddings can also be used to discover modules which also correspond to enriched pathways of different functions. The brown module, in particular, is believed to be likely to give rise to breast cancer. Pathway analysis on the whole gene set also revealed a wealth of enriched pathways relating to cell regulation, protein process, and immune system.

Name	Module	Accession Num	Major role in BRCA	Post-transcriptional regulatory interactions	Target Gene/Pr	Events	miRbase Ref
hsa-mir-100	Yellow	MI0000102	Tumor suppressor	Inducing angiogenesis	VEGF, mTOR/HIF-1α	Shuttling of miRNA enriched in MSC-derived exosomes, anti-angiogenesis and anti-tumorigenesis	http://www.mirbase.org/cgi-bin/mirna_entry.pl?acc=MI0000102
hsa-mir-101	Yellow	MI0000103 MI0000739	Tumor suppressor	Resisting apoptotic response and cell death	EYA1, jagged1, Hes1, Hey1, SOX2	Promotion of apoptotic response by negatively regulating Notch pathway	http://www.mirbase.org/cgi-bin/mirna_entry.pl?acc=MI0000103 http://www.mirbase.org/cgi-bin/mirna_entry.pl?acc=MI0000739
hsa-mir-143	Yellow	MI0000459	Tumor suppressor	Sustaining growth and proliferative signals	ERK5, MAP3K7, Cyclin D1	Anti-proliferative	http://www.mirbase.org/cgi-bin/mirna_entry.pl?acc=MI0000459
hsa-mir-335	Yellow	MI0000816	Tumor suppressor	Activating metastasis and invasion	EphA4	Anti-metastasis and anti-invasion	http://www.mirbase.org/cgi-bin/mirna_entry.pl?acc=MI0000816
hsa-mir-512	Black	MI0003140 MI0003141	Tumor suppressor	Replicative immortality	hTERT	Reduction of telomerase activity, impairment of telomere maintenance and activation of replicative senescence and apoptosis programs	http://www.mirbase.org/cgi-bin/mirna_entry.pl?acc=MI0003140 http://www.mirbase.org/cgi-bin/mirna_entry.pl?acc=MI0003141
hsa-mir-519a-2	Black	MI0003182	Oncogenic	Resisting apoptotic response and cell death	TRAIL-R2 (TNFRSF10B), caspase-8, caspase-7, MICA, ULBP2	Promotion of apoptosis resistance and escape from natural killer cell recognition	http://www.mirbase.org/cgi-bin/mirna_entry.pl?acc=MI0003182
hsa-mir-16	Green	MI0000070 MI0000115	Tumor suppressor	Sustaining growth and proliferative signals	Cyclin E1, E2F7	Anti-proliferative and G1–S cell cycle arrest, restores tamoxifen sensitivity	http://www.mirbase.org/cgi-bin/mirna_entry.pl?acc=MI0000070 http://www.mirbase.org/cgi-bin/mirna_entry.pl?acc=MI0000115
hsa-mir-191	Green	MI0000465	Oncogenic	Inducing angiogenesis	HuR, TGFβ2, SMAD3, BMP4, JUN, FOS, PTGS2, CTGF, VEGFA	Hypoxia-inducible miRNA and stimulator of TGFβ-signaling pathways	http://www.mirbase.org/cgi-bin/mirna_entry.pl?acc=MI0000465
hsa-mir-122	Turquoise	MI0000442	Oncogenic	Reprogramming energy metabolism	pyruvate kinase (PK) and citrate synthase (CS)	Promotion of metastasis by reprogrammed glucose metabolism	http://www.mirbase.org/cgi-bin/mirna_entry.pl?acc=MI0000442
hsa-mir-331	Red	MI0000812	Oncogenic	Activating metastasis and invasion	HER2, HOTAIR, E2F1, DOHH, PHLPP	Promotion of metastasis and invasion by elevation in plasma of metastatic breast cancer patients	http://www.mirbase.org/cgi-bin/mirna_entry.pl?acc=MI0000812
hsa-mir-135b	Grey	MI0000810	Oncogenic	Sustaining growth and proliferative signals	LATS2, CDK2, p-YAP	Promotion of cell proliferation and S–G2/M cell cycle progression	http://www.mirbase.org/cgi-bin/mirna_entry.pl?acc=MI0000810

Table 3. Salient miRNA information

REFERENCES

[1] 1 Choy C T, Wong C H, Chan S L. Embedding of genes using cancer gene expression data: biological relevance and potential application on biomarker discovery[J]. Frontiers in genetics, 2019, 9: 682.

[2] Gao Y L, Hou M X, Liu J X, et al. An integrated graph regularized non-negative matrix factorization model for gene co-expression network analysis[J]. IEEE Access, 2019, 7: 126594-126602.

[3] Liu F, Dong H, Mei Z, et al. Investigation of miRNA and mRNA Co-expression Network in Ependymoma[J]. Frontiers in bioengineering and biotechnology, 2020, 8: 177.

[4] Nicora G, Vitali F, Dagliati A, et al. Integrated multi-omics analyses in oncology: a review of machine learning methods and tools[J]. Frontiers in oncology, 2020, 10: 1030.

[5] Cirillo E, Parnell L D, Evelo C T. A review of pathway-based analysis tools that visualize genetic variants[J]. Frontiers in genetics, 2017, 8: 174.

[6] Kao P Y P, Leung K H, Chan L W C, et al. Pathway analysis of complex diseases for GWAS, extending to consider rare variants, multi-omics and interactions[J]. Biochimica et Biophysica Acta (BBA)-General Subjects, 2017, 1861(2): 335-353.

[7] Xiong Q, Ancona N, Hauser E R, et al. Integrating genetic and gene expression evidence into genome-wide association analysis of gene sets[J]. Genome research, 2012, 22(2): 386-397.

[8] Pers T H, Karjalainen J M, Chan Y, et al. Biological interpretation of genome-wide association studies using predicted gene functions[J]. Nature communications, 2015, 6(1): 1-9.

[9] Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis[J]. BMC bioinformatics, 2008, 9(1): 1-13.

[10] Ashburner M, Ball C A, Blake J A, et al. Gene ontology: tool for the unification of biology[J]. Nature genetics, 2000, 25(1): 25-29.

[11] Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes[J]. Nucleic acids research, 2000, 28(1): 27-30.

[12] Granger B R, Chang Y C, Wang Y, et al. Visualization of metabolic interaction networks in microbial communities using VisANT 5.0[J]. PLoS computational biology, 2016, 12(4): e1004875.

[13] Szklarczyk D, Gable A L, Lyon D, et al. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets[J]. Nucleic acids research, 2019, 47(D1): D607-D613.

[14] Yu G, Wang L G, Han Y, et al. clusterProfiler: an R package for comparing biological themes among gene clusters[J]. Omics: a journal of integrative biology, 2012, 16(5): 284-287.

[15] Yu G. Enrichplot: Visualization of functional enrichment result[J]. R package version, 2018, 1(2).

[16] Loh H Y, Norman B P, Lai K S, et al. The regulatory role of microRNAs in breast cancer[J]. International journal of molecular sciences, 2019, 20(19): 4940.

[17] Singh R, Mo Y Y. Role of microRNAs in breast cancer[J]. Cancer biology & therapy, 2013, 14(3): 201-212.

[18] Hoadley, K.A., et al., 2014. Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. Cell 158, 929–944.

[19] Yuan, Y., et al., 2014. Assessing the clinical utility of cancer genomic and proteomic data across tumor types. Nat. Biotechnol. 32, 644–652.