LEARNING-BASED ROBUST SAMPLE SELECTION TO REDUCE NOISE IN HIGH DIMENSIONAL TRANSCRIPTOME DATA

Nehir Kızılilsoley, Ezgi Tanıl, and Emrah Nikerel

Yeditepe University, Department of Genetics and Bioengineering, Istanbul, Turkey
email: emrah.nikerel@yeditepe.edu.tr

ABSTRACT

To reduce inherent noise in high dimensional transcriptome data from a lung cancer cohort, a learning based sub-sample selection approach is adopted. Focusing on consensus clustering analysis, TCGA network data on lung cancer reached its maximum cluster stability when divided into three, which matches with the number of actual groups (adenocarcinoma, squamous cell carcinoma and normal). Using silhouette width as well as naive inspection of clustering performance to filter out samples, 840 out of 1145 samples were selected as core samples. The contribution of using consensus clustering analysis as a sample selection method was assessed by comparing the subtype classification accuracies of informative genes discovered from the "initial" set (1145 samples), "reduced" set (901 samples) and core set (840 samples). The list of candidate markers obtained from initial samples and core samples were similar, with a great increase in the prediction accuracy. Taken together, the results suggest that learning based sample selection can aid in sample filtering while retaining most of the information and reducing the noise.

1. INTRODUCTION

Transcriptome data, the set of all RNA transcripts from an individual or a population, is always noisy due to inherent noise in the gene expression process or the source being a heterogeneous group of cells. The noise amplifies more when cohort data is used, i.e. from a group of disease/normal samples, in the context of e.g. cancer, which in turn makes the biomarker discovery studies challenging. A key challenge is then to obtain mechanistic information from the available noisy data.

One way to reduce the noise from population data is the sample selection from the initial set for improved data quality. Briefly, the approach consists of iteratively selecting most informative samples among an initial set, that would allow best separation among groups using feature selection algorithms and building classifiers. Among available options, consensus clustering analysis (CCA) is a technique used to run a (collection of) selected clustering algorithms (k-means, hierarchical, biclustering etc.) recursively on subsamples to obtain a consensus from all the results of each iteration; to determine the optimum number of clusters, to evaluate the stability of the found clusters, to reduce data dimension while keeping the information content [1]–[3].

Lung cancer is the most lethal cancer type in both men and women, as it comes first in cancer-related deaths worldwide; with a survival rate of 15% in the first 5 years and 7% in 10 years after diagnosis [4]. The high rates of mortality arise from not only the lack of early diagnosis strategies but also the lack of efficient treatments specialized for the stage and subtype of lung cancer the patients are suffering from [5], [6]. These limitations reflect an urgent need for biomarkers that allow for early-stage diagnosis and prognosis of lung cancer, which may improve the treatment patients receive [7]. Along with the advance of omic technologies, transcriptomics has assisted greatly in the identification of biological markers for lung cancer. Identification of sensitive and reproducible gene markers for accurate diagnosis is of great interest in precision medicine.

This study focuses on sample (re)selection problem and uses the consensus clustering analysis to select core samples representing the 2 main subtypes of lung cancer and normal samples: adenocarcinoma (LUAD) and squamous cell carcinoma (LUSC) and normal cells (N). For the case study, lung cancer from The Cancer Genome Atlas network consisting 65048 genes and 1145 lung cancer samples (LUAD: 535, LUSC: 502, N: 108) were used.

2. METHODS

2.1 Data retrieval and preprocessing

The transcriptome data as HTSeq counts for two main subtypes of lung cancer were retrieved from TCGA Research Network Data Portal from two projects TCGA-LUAD (with 535/59 tumor/normal samples), and TCGA-LUSC (with 502/49 tumor/normal samples) for adenocarcinoma and squamous cell carcinoma samples respectively.

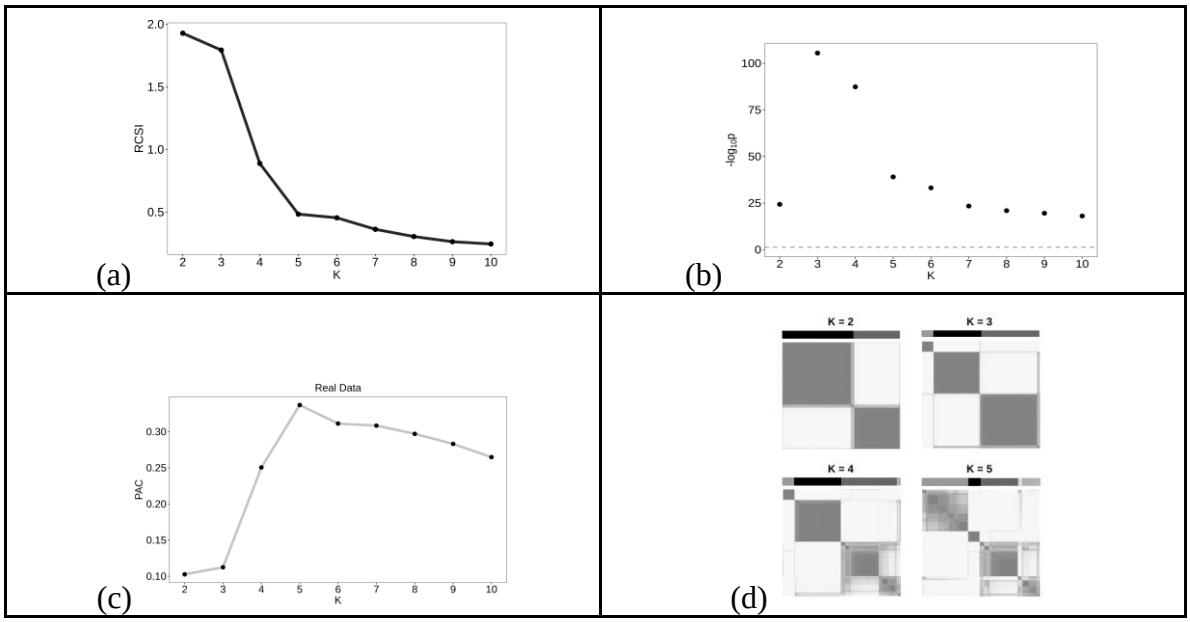

Figure 1. (a) the relative cluster stability index (RCSI) plot (b) P-value distribution for each cluıster number (c) the portion of ambiguously clustering (PAC) plot (d) Consensus matrices for clusters K=2,3,4 and 5.

Raw matrix of gene counts was annotated using biomaRt R package [8] and gene summarization was performed by keeping the transcript with highest overall count out of transcripts of the same genes. Then, lowly expressed genes were filtered out. The data was normalized to counts per million (cpm) using the edgeR package [9], and log-transformed. Genes were then ranked according to their sample-wise variance and the genes with a variance greater than a selected threshold are kept for further analysis.

2.2 Consensus clustering analysis

Consensus clustering analysis was performed using the R package M3C [3]. Robust clusters were detected using agglomerative hierarchical clustering as the basis. Euclidean distance was chosen as the distance metric, and the procedure was repeated 100 times with a subsampling rate of 80%. Optimum number of clusters are decided based on the cumulative distribution function (CDF) graph, the relative cluster stability index (RCSI), the portion of ambiguously clustering (PAC), and the p-value for the null hypothesis that the groups are the same (Fig.1).

2.3 Selecting core samples and evaluating classification performances

To select the samples that best represent the clusters, two different strategies were followed: (i) removing samples with negative silhouette widths [10] and (ii) removing samples clustered into "false clusters". Removal of samples with negative silhouette widths left 901 samples and this data set will be referred as the "reduced" set. The "core" set was the one where samples were further filtered by removing "misclustered" ones, down to 840 samples.

In order to measure the benefit of this filtering pipeline and see it improve subtying classification, a ML classification pipeline was built and the three data sets ("initial", "reduced" and "core" set) with 1277 genes were registered to this pipeline. This pipeline firstly split the given data set into training and test set by partitioning it by 80:20. Then, the training data was scaled and the test set was transformed with mean and standard deviation calculated from the training data.

In the next step, a feature selection applied to reduce 1277 genes down to set k value. To do this, mutual information, which is a measure of dependence between two random variables, were used to select top k most informative genes for the classification. The selected genes were used as features in the next step's support vector machine classification (SVM). Five-fold cross-validated linear SVM machines were used as classifiers. Predictions of the classifiers on unseen data were performed on held-out test data.

3. RESULTS

3.1 Preprocessing and consensus clustering analysis

Raw count data was pre-processed and filtered. The number of genes dropped from 65048 to 27016 by removing lowly expressed and non-annotated genes. The second preprocessing was performed based on sample-wise variation. The threshold value was selected by considering the variance of some known reference genes. The median expression variance of the keratin, GAPDH, beta-actin and YWHAZ reference genes [13], [14] and all isoforms of these genes were found to be 1.11, 1.37, 2.99 and 0.84, respectively. Thus, a safe threshold value of minimum 5.00 was chosen. 1277 genes remained from this step. Finally, highly correlated healthy tissue samples from both projects (LUAD and LUSC), a total of 108, were combined. The final input matrix consisted of 1277 genes and 1145 samples.

In this work, CCA was used to re-find the subtypes of lung cancer from a given data, and find samples that create noise in the definition process of the group they belong to. Even though all clusters from 2 to 10 were found to be significant ($p < 0.05$), the number of clusters found most significant when tested against the null hypothesis K=1 was 3 (Fig. 1a, b and c). The consensus matrices (Fig. 1d) also did not show a stable distribution after K=3, supporting the finding that the data is clustered best in 3, matching the readily known and anticipitated number of classes. The entire analysis was also performed for 100, 500 and 1000 iterations, but no change was observed in the results.

3.2 Selecting core samples

Silhouette method is a quality measure for a clustering task, which visualizes silhouette coefficients of each sample. Silhouette coefficient, in turn, quantifies how much a given sample belong to that cluster as a function of distance. The silhouette plot of the transcriptome data divided into the three clusters shows the samples in each sample and their "fidelity" to that cluster. In a perfect case where all samples would have been clustered to their actual group, the first group would have 535 samples (LUAD), the second group would have 502 samples (LUSC), and the last group would have 108 (Normal) samples. However, the group numbers predicted by cluster analysis were found to be 585, 449 and 111, respectively. When examined carefully, it was seen that all normal samples were clustered correctly along with 3 LUAD and 1 LUSC samples, but 59 LUSC samples were clustered together with LUAD samples and 7 LUAD samples assigned as LUSC.

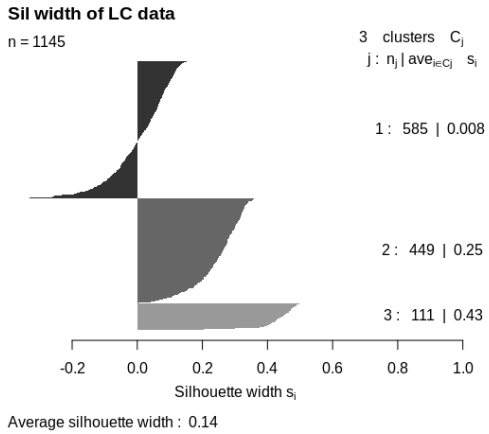

Figure 2. Silhouette graph of optimum clustering at K=3.

Verhaak et al. (2010) demonstrated the application of CCA in finding yet-to-known molecular subtypes of glioblastoma multiforme transcriptome data, wherein they removed samples with negative silhouette widths [10]. To achieve a refined sample set also in this work, 244 samples with negative silhouette width were extracted from the dataset (Fig. 2). The remaining 901 samples (345 LUAD, 448 LUSC and 108 Normal) were further examined for "mis-clusterings": the samples that do not agree with the label majority of the cluster they are assigned. To identify mis-clustered samples, confusion matrices were created with the cluster assignments of the CCA algorithm (Table 1).

Using the confusion matrix of "reduced" set (Table 1b), falsely clustered samples were removed manually. By subtracting 8 samples from LUAD data and 53 samples from LUSC data, 840 samples were selected as representative samples and named as "core" set. This resulted in a great reduction in the overlap between classes in the principal component space (Fig. 3).

Table 1. Confusion matrix of actual labels and the abundant labels the samples clustered into (top) before (bottom) after the removal of samples with negative silhouette widths ("reduced" set).

		True Classes			
		LUAD	**LUSC**	**Normal**	
Assigned Cluster	**LUAD**	526	59	0	
	LUSC	7	442	0	
	Normal	2	1	108	
		535	502	108	1145

		True Classes			
		LUAD	**LUSC**	**Normal**	
Assigned Cluster	**LUAD**	337	52	0	
	LUSC	6	395	0	
	Normal	2	1	108	
		345	502	108	901

3.3 Signature gene identification and classification performance evaluation

The effect of sample selection via CCA for subtyping biomarker discovery was elaborated by using the three data sets, the "initial" set (1145 samples), "reduced" (901 samples) and "core set" (840 samples) in ML classification pipeline and comparing each classifiers performance. Training accuracies are averaged over cross validation folds (Fig. 4). The removal of samples with negative silhouette widths ("reduced" set) clearly reduced the variance within model performances, although no appearent improvement in mean accuracies compared to the SVMs built with "initial" set. The "core" set accuracies are generally higher than the other sets' results.

The testing accuracies for each data set with increasing feature numbers are given in Figure 5. There is evident increase in accuracies for all feature numbers (k) when went from "initial" set to "reduced" and "core" set, respectively, apart from the expected increase within each set's accuracies with

increasing k. Also, the accuracies of "reduced" and "core" set are more stable than the "initial" set.

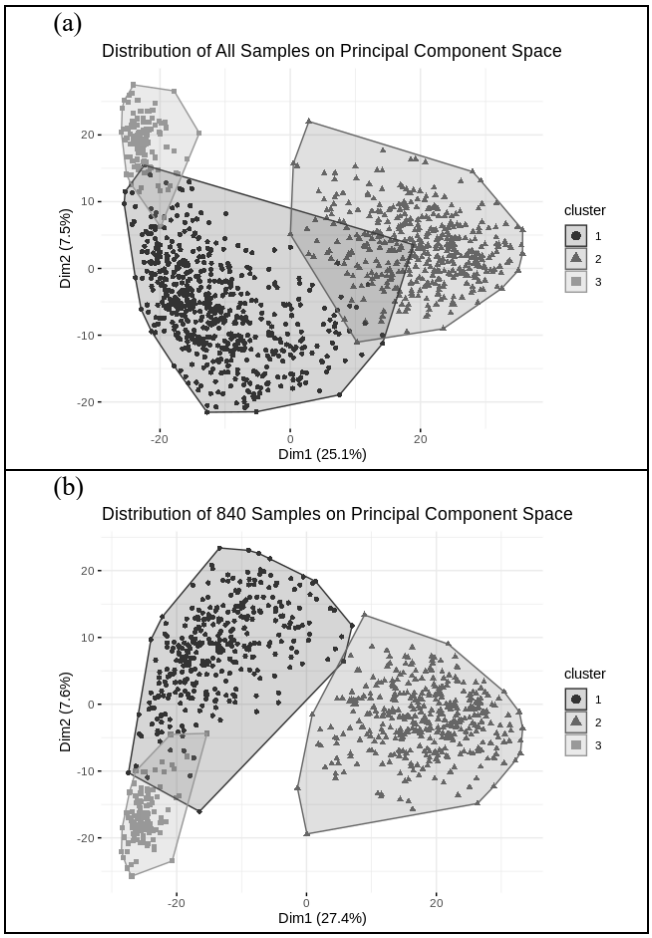

Figure 3. Distribution of samples on principal component space. (a) 1145 samples, (b) 804 "core" samples.

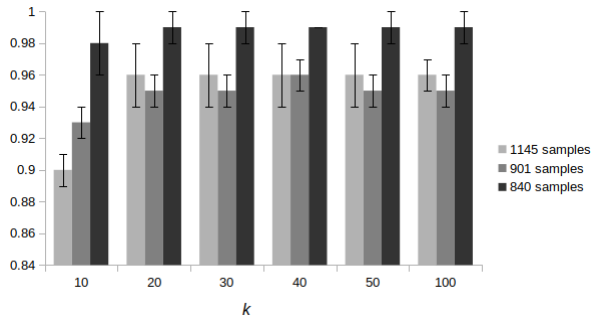

Figure 4. Mean training accuracies of SVM classifiers built with each data sets and different numbers of features (k).

To consider the direct effect of denoised data on the classification, the genes obtained from "core" sets were used to classify the "initial" set to see if the gene lists are more informative in the means of subtype classification when obtained from allegedly denoised "core" set (Fig. 5). This implementation resulted in considerable increase in the prediction performance for 10 genes (Table 2), and slight improvement for 40 and 50 genes. To avoid the information leakage between training samples of "core" set and testing of "initial" set, which would mislead the judgement on model performances, the training samples of "core" set used for feature selection were excluded from the testing of "initial" set. Here, as can be seen in the line named "1145 samples**" the accuracies were generally lower than it is for "initial" set model. Overall, performances of classifiers built with k genes selected from the information of 840 samples followed similar trend in classifying test samples of "initial" set.

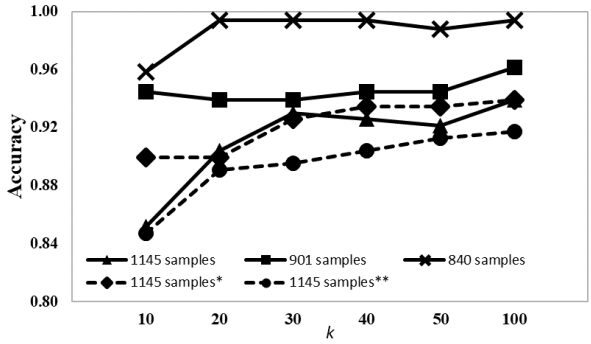

Figure 5. Overall testing accuracies of SVM classifiers built with one of three data sets and different numbers of features (k). The line named 1145 samples* represents the accuracies obtained from the classification of 1145 samples using features obtained from feature selection done on the "core" set. The 1145** line is again the classification of the "initial" set with features selected from the "core" set, where the training samples of "core" set are excluded from the testing of "initial" set to avoid information leakage.

4. DISCUSSION

This work focuses on the learning-based sample selection approach to reduce noise in high dimensional transcriptome data. Resulting dataset contains less noisy data suspected from classification performance, yet still yields a similar biomarker candidate list. The approach taken here yields improved biomarker discovery workflow, yielding robust biomarkers and the data from the remaining samples would yield crisper separation between groups.

A key challenge in clustering is the need for determining the number of clusters a priori. CCA scans a range of cluster numbers and selects the optimum K for a given data through evaluating various performance metrics. This method is prefered when the classes of data are unknown, as it was for Verhaak et. Al (2010) used it to find subtypes of glioblastoma multiforme, which were not readily known then [10]. To investigate this method in finding disease subtypes, we used TCGA gene expression data for lung cancer subtypes, whose classes were known. Indeed, CCA yielded 3 clusters as the optimum clustering and the clusters contained samples with majority of only one class. To come up with the refined samples that define its subtype robustly, first samples with negative silhouette widths were removed as [10]. Later, if a sample does not have the same actual label with the majority of its cluster, it was also removed from the dataset. However, as suggested in [10], samples with negative silhouette widths

did not really overlapped with mis-clustered ones, hence it may not be the correct way of finding subtype representative samples.

Once three data sets were produced, they were fed into ML classification pipeline, where most infomative k genes were selected by mutual information and SVM classification models were built using those genes. While 80% of samples from "core" set produced models with lowest variance and highest accuracy in training, removing of samples with negative silhouette widths ("reduced" set) had the effect in reduction of training variance without an improvement in accuracy of "initial" set. For testing, on the other hand, the accuracies for different gene numbers increased from "initial" set to "core" set, regardless of less samples being used for teaching the model.

To see if the gene lists are more informative for subtype classification when selected from allegedly denoised data, genes obtained from "core" set were used to classified all available data, the "initial" set. When the same samples used for testing the model of "initial" set were also used for the model that uses genes of "core" set, it resulted in information leakage from 140 intersecting samples of training data of "core" set into the testing prediction for this set up. Thus, genes from "core" samples seemed to performed better in "initial" set. However, when training samples of "core" set were excluded from the testing samples of this implementation, the accuracies were less than it was with genes selected from "initial" set. However, this difference in performance can be reduced by tuning classifiers for each input data to improve the learning, instead of using default settings for SVM. To further infer the effect of reducing samples with CCA-based procedure, other types of ML algorithms can also be used since there is no rule of thumb for selecting best classifier for different data.

Between the first 10 genes of "initial" set and the "core" set, RNF138 and DSC1 seem to be replaced with UBA5 and PTPDC1, which may be the features that contributed the difference in the accuracy (Table 2). RNF138 has shown to be a therapy and drug resistance indicator in various tumors as RNF family proteins are known to be involved in tumorigenesis [15-16]. Decreased desmocollin 1 (DSC1) expression was found to be associated with poor prognosis in human lung cancer [17]. There is no study indicating ubiquitin-like modifier-activating enzyme 5 (UBA5) is a marker of lung cancer, however it is reported that the inhibition of UBA5 can impede tumor development [18]. PTPDC1, being a regulator of signal transduction and cell cycle can be a tumor progressor and found to be linked to progression of gastric cancer [19]. The rest of the the gene lists obtained from 1145 samples and 840 samples were quite similar to each other. The most informative gene in all lists, AGBL4, the family of ATP/GTP Binding Protein Like proteins are also known to have a predictive value for lung cancer [20-21]. PLEKHM3 is another that appears only on "reduced" set and was found to be mutated in LUAD patients with overexpression of TMED2 and to be positively correlated in lymphoid neoplasm diffuse large B-cell lymphoma patients [22-23].

Table 2: Top 10 candidate gene marker lists obtained from all samples, reduced sample-set and "core" samples.

1145 samples	901 samples	840 samples
AGBL4	AGBL4	AGBL4
SLC37A4	APH1A	**UBA5**
ALPK1	SLC37A4	SLC37A4
RNF138	ALPK1	ALPK1
DSC1	RNF138	TRMT1L
TRMT1L	**PLEKHM3**	COQ8A
COQ8A	**PTPDC1**	**PTPDC1**
PRKCA	AC026954.2	PRKCA
GPR157	GPR157	GPR157
LONRF2	LONRF2	LONRF2

5. ACKNOWLEDGEMENTS

This work has been funded by Health Institutes of Turkey (TUSEB) project no: 2019-TA-01-4589.

6. REFERENCES

[1] S. Monti, P. Tamayo, J. Mesirov, and T. Golub, "Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data," Mach. Learn., vol. 52, no. 1, pp. 91–118, 2003.

[2] Y. Șenbabao\uglu, G. Michailidis, and J. Z. Li, "Critical limitations of consensus clustering in class discovery," Sci. Rep., vol. 4, no. 1, pp. 1–13, 2014.

[3] C. R. John et al., "M3C: Monte Carlo reference-based consensus clustering," Sci. Rep., vol. 10, no. 1, pp. 1–14, 2020, doi: 10.1038/s41598-020-58766-1.

[4] J. Ferlay et al., "Global cancer observatory: cancer today. International Agency for Research on Cancer," Lyon, Fr., 2020.

[5] P. Indovina, E. Marcelli, P. Maranta, and G. Tarro, "Lung cancer proteomics: recent advances in biomarker discovery," Int. J. Proteomics, vol. 2011, 2011.

[6] M. Saleem, S. K. Raza, and S. G Musharraf, "A comparative protein analysis of lung cancer, along with three controls using a multidimensional proteomic approach," Exp. Biol. Med., vol. 244, no. 1, pp. 36–41, 2019.

[7] C. H. Y. Cheung and H.-F. Juan, "Quantitative proteomics in lung cancer," J. Biomed. Sci., vol. 24, no. 1, pp. 1–11, 2017.

[8] S. Durinck, P. T. Spellman, E. Birney, and W. Huber, "Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt," Nat. Protoc., vol. 4, no. 8, pp. 1184–1191, 2009.

[9] M. D. Robinson, D. J. McCarthy, and G. K. Smyth, “edgeR: a Bioconductor package for differential expression analysis of digital gene expression data,” bioinformatics, vol. 26, no. 1, pp. 139–140, 2010.
[10] R. G. W. Verhaak et al., “Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1,” Cancer Cell, vol. 17, no. 1, pp. 98–110, 2010.
[11] T. Hastie, R. Tibshirani, B. Narasimhan, and G. Chu, “Pam: prediction analysis for microarrays,” R Packag. version, vol. 1, no. 1, 2019.
[12] V. G. Tusher, R. Tibshirani, and G. Chu, “Significance analysis of microarrays applied to the ionizing radiation response,” Proc. Natl. Acad. Sci., vol. 98, no. 9, pp. 5116–5121, 2001.
[13] O. Thellin et al., “Housekeeping genes as internal standards: use and limits,” J. Biotechnol., vol. 75, no. 2–3, pp. 291–295, 1999.
[14] P. Murthi, E. Fitzpatrick, A. J. Borg, S. Donath, S. P. Brennecke, and B. Kalionis, “GAPDH, 18S rRNA and YWHAZ are suitable endogenous reference genes for relative gene expression studies in placental tissues from human idiopathic fetal growth restriction,” Placenta, vol. 29, no. 9, pp. 798–801, 2008.
[15] C. Wu, L. Chen, H. Tao, L. Kong, and Y. Hu, “RING finger protein 38 induces the drug resistance of cisplatin in non-small-cell lung cancer,” Cell Biol. Int., vol. 45, no. 2, pp. 287–294, 2021.
[16] Y. Lu et al., “RNF138 confers cisplatin resistance in gastric cancer cells via activating Chk1 signaling pathway,” Cancer Biol. \& Ther., vol. 19, no. 12, pp. 1128–1138, 2018.
[17] Cui, T., Chen, Y., Yang, L., Mireskandari, M., Knösel, T., Zhang, Q., Petersen, I. (2012). “Diagnostic and prognostic impact of desmocollins in human lung cancer”. Journal of clinical pathology, 65(12), 1100-1106.
[18] Fang, B., Li, Z., Qiu, Y., Cho, N., & Yoo, H. M. (2021). “Inhibition of UBA5 Expression and Induction of Autophagy in Breast Cancer Cells by Usenamine A. Biomolecules”, 11(9), 1348.
[19] Li, Z., Cheng, Y., Fu, K., Lin, Q., Zhao, T., Tang, W., ... & Sun, Y. (2021). “Circ-PTPDC1 promotes the progression of gastric cancer through sponging Mir-139-3p by regulating ELK1 and functions as a prognostic biomarker.” International journal of biological sciences, 17(15), 4285.
[20] H. J. Kwak et al., “Expression of ATP/GTP Binding Protein 1 Has Prognostic Value for the Clinical Outcomes in Non-Small Cell Lung Carcinoma,” J. Pers. Med., vol. 10, no. 4, p. 263, 2020.
[21] L. Zhang, X. Li, X. Quan, W. Tian, X. Yang, and B. Zhou, “A case/control study: AGBL1 polymorphism related to lung cancer risk in Chinese nonsmoking females,” DNA Cell Biol., vol. 38, no. 12, pp. 1452–1459, 2019.
[22] L. Feng, P. Cheng, Z. Feng, and X. Zhang, “Transmembrane p24 trafficking protein 2 regulates inflammation through the TLR4/NF-$κ$B signaling pathway in lung adenocarcinoma,” World J. Surg. Oncol., vol. 20, no. 1, pp. 1–13, 2022.
[23] M. A. Qureshi et al., “Pan-cancer multiomics analysis of TC2N gene suggests its important role (s) in tumourigenesis of many cancers,” Asian Pacific J. Cancer Prev. APJCP, vol. 21, no. 11, p. 3199, 2020.