frontiers in Genetics

METHODS
published: 10 September 2020
doi: 10.3389/fgene.2020.00979

Kernel Fusion Method for Detecting Cancer Subtypes via Selecting Relevant Expression Data

Shuhao Li[1], *Limin Jiang*[1], *Jijun Tang*[1,2], *Nan Gao*[3]* *and Fei Guo*[1]*

[1] School of Computer Science and Technology, College of Intelligence and Computing, Tianjin University, Tianjin, China, [2] Department of Computer Science and Engineering, University of South Carolina, Columbia, SC, United States, [3] School of Computer Science and Technology, Zhejiang University of Technology, Hangzhou, China

Recently, cancer has been characterized as a heterogeneous disease composed of many different subtypes. Early diagnosis of cancer subtypes is an important study of cancer research, which can be of tremendous help to patients after treatment. In this paper, we first extract a novel dataset, which contains gene expression, miRNA expression, and isoform expression of five cancers from The Cancer Genome Atlas (TCGA). Next, to avoid the effect of noise existing in 60, 483 genes, we select a small number of genes by using LASSO that employs gene expression and survival time of patients. Then, we construct one similarity kernel for each expression data by using Chebyshev distance. And also, We used SKF to fused the three similarity matrix composed of gene, Iso, and miRNA, and finally clustered the fused similarity matrix with spectral clustering. In the experimental results, our method has better *P*-value in the Cox model than other methods on 10 cancer data from Jiang Dataset and Novel Dataset. We have drawn different survival curves for different cancers and found that some genes play a key role in cancer. For breast cancer, we find out that HSPA2A, RNASE1, CLIC6, and IFITM1 are highly expressed in some specific groups. For lung cancer, we ensure that C4BPA, SESN3, and IRS1 are highly expressed in some specific groups. The code and all supporting data files are available from https://github.com/guofei-tju/Uncovering-Cancer-Subtypes-via-LASSO.

Keywords: cancer subtype, similarity Kernel fusion, LASSO, gene expression, miRNA expression, isform level

OPEN ACCESS

Edited by:
Wen Zhang,
Huazhong Agricultural University, China

Reviewed by:
Qi Zhao,
Liaoning University, China
Cangzhi Jia,
Dalian Maritime University, China

***Correspondence:**
Nan Gao
gaonan@zjut.edu.cn
Fei Guo
fguo@tju.edu.cn

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *08 July 2020*
Accepted: *03 August 2020*
Published: *10 September 2020*

Citation:
Li S, Jiang L, Tang J, Gao N and Guo F (2020) Kernel Fusion Method for Detecting Cancer Subtypes via Selecting Relevant Expression Data. Front. Genet. 11:979. doi: 10.3389/fgene.2020.00979

1. INTRODUCTION

Numerous studies have shown that cancer is a heterogeneous disease (Wang et al., 2005). Today, doctors can use the special information contained in different cancers for more targeted treatment (Fedele et al., 2014; Fu et al., 2014; Marino et al., 2017). Therefore, it is very meaningful to be able to accurately identify cancer subtypes, including molecular subtyping as well as clinical outcome-based clustering. For breast cancer, four major molecular subtypes include Luminal A, Luminal B, Triple negative/basal-like, and HER2-enriched. However, clustering samples based on therapy response and the aggressiveness level may not overlap with these subtypes. With the development of whole-genome sequencing techniques in recent years, the diagnosis and treatments have gained great development (Wang K. et al., 2014; Haase et al., 2015). We have obtained massive cancer expression from database as The Cancer Genome Atlas (TCGA) (Tomczak et al., 2015). Thus, these expression data have positive influence on the development of the cancer subtype identification tools (Sohn et al., 2017; Guo Y. et al., 2018).

Generally, the machine learning method is now widely used to solve clustering problem for cancer subtypes (Kourou et al., 2015; Li et al., 2016; Mirza et al., 2019). Wang et al. (2018) combined Monte Carlo feature selection (MCFS), random forest (RF), and rough set-based rule learning to identify breast cancer. Li and Ruan (2005) used support vector machine for cancer recognition. Monti et al. (2003) combined resampling consensus clustering. Also, there are many tools based on deep learning method (Wang et al., 2016; Esteva et al., 2017; Miotto et al., 2017). Chen et al. (2019) used RNN to identify some genes that have an impact on cancer. Neighbor Ensemble-based Detection (NED) proposed by Zhou et al. identified lung cancer cells (Zhou et al., 2002). Karabatak and Ince (2009) identified breast cancer through association rules (AR) and neural network (NN). Brunet et al. (2004) proposed non-negative matrix factorization to find cancer subtype.

Furthermore, many predictive models can identify cancer subtypes by using single expression data (Verhaak et al., 2010; Chen et al., 2013; Zhang et al., 2017). Verhaak et al. (2010) employed gene expression to identify four subtypes in glioblastoma multiforme (GBM). Brunet et al. (2004) used gene expression to uncover subtypes on three datasets, including Myelogenous leukemia, Medulloblastomas, and Central Nervous System Tumors. Wong et al. (2012) proposed the Feature Set Reduction method to select more important single nucleotide polymorphism and classify cancer subtypes on three diseases as sarcoma, lymphoma, and leukemia. Zhang et al. (2017) used DNA methylation to find cancer subtypes on breast cancer. Pan et al. (2018) used copy number variants to identify four cancer subtypes on breast cancer. Zhao et al. (2009) used single-stranded DNA (ssDNA) to find cancer subtypes on lung cancer.

However, since cancer is a heterogeneous disease, independent analysis of a single type of data often results in unsatisfactory consequence. Some studies take advantage of various popular multiple kernel learning methods (Ding et al., 2017; Jiang et al., 2018), mainly through the integration of similarity networks among patients from multiple expression data. Wang B. et al. (2014) integrated three expression data, including gene expression, DNA methylation data, and miRNA expression data, to calculate the patient similarity network by using the similarity network fusion (SNF). Ma and Zhang (2017) improved the SNF and proposed the affinity network fusion (ANF) to cluster multiple cancer patients. The unsupervised multiple kernel learning (UMKL) for multiple datasets was proposed by Mariette and Villa-Vialaneix (2017). Jiang et al. (2019) improved the SNF and proposed the similarity kernel fusion (SKF) to combine three expression data including gene expression, isoform data, and miRNA expression data, and first collected five cancer datasets to verify the performance of model. Jiang et al. used the Euclidean distance when constructing the similarity kernels. The dimensionality of DNA and other features is very large. The use of Euclidean distance may have a great impact on the clustering results.

In this paper, we employ LASSO for gene selection and use Chebyshev distance for constructing similarity kernels. The main process of this article is roughly introduced as follows. First, we extract five novel datasets (bladder cancer, blood cancer, brain cancer, ovary cancer, and pancreas cancer) from The Cancer Genome Atlas (TCGA). It's worth noting that each cancer has three expression data, including gene expression, isoform expression, and miRNA expression. Second, we employed LASSO (Tibshiranit, 1996) to identify the high-efficiency gene expression data and fit survival time, in order to achieve the purpose of feature selection. Since the original gene expression data has high dimensions, the high dimensionality of the data has a very negative effect on the clustering results of small sample size. Third, the Chebyshev distance replaces the Euclidean distance to construct the kernel of the patient's similarity, which can further mitigate the impact of the high-dimensional data. Forth, we used similarity kernel fusion (SKF) to fuse three similarity kernels into one synthetical kernel. Finally, we used spectral clustering on the fused kernel to predict the patient's cancer subtype. In the experimental results, we found that our method achieves outstanding *P*-value in the Cox model on five existing datasets and five novel datasets. We also find the survival curve and the heat map preform outstandingly well on each cancer subtype according to our model.

2. MATERIALS AND METHODS

We select a group of significant gene expression to construct three similarity kernels. Also, we fuse three similarity kernels into one kernel for cancer subtype clustering. The whole process of our method is shown in **Figure 1**.

2.1. Novel Dataset

Wang B. et al. (2014) have already extracted five datasets from TCGA, but the number of patients is too small for each dataset. The datasets of Jiang et al. (2019) have alleviated the problem of fewer samples. To better verify the performance of model, we extract five novel data sets, in addition to Jiang's dataset. For each dataset, we select three types of expression data, including gene expression, miRNA expression, and isoform level. The number of expression data is shown in **Table 1**. We can see that the Jiang's dataset includes stomach cancer, lung cancer, kidney cancer, breast cancer, and colon cancer, Our Novel Dataset add five novel cancer data to Jiang Dataset, which are bladder cancer, blood cancer, brain cancer, ovary cancer, and pancreas cancer.

2.2. Gene Selection

The gene expression data have high dimensions in our novel extracted datasets. Due to the curse of dimensionality, high-dimensional data have a great influence on the experimental results. Therefore, We use LASSO to select a part of important genes. We give a formalized description of LASSO, as Equation (1).

$$\min \frac{1}{2}\sum_{n=1}^{N}(y(X_{n\cdot},\omega)-T_n)^2+\frac{\lambda}{2}\|\omega\|_1 \tag{1}$$

We represent patient data as $X \in R^{n\times m}$, where n is the number of patients and m is the number of expression factors. Patient survival time is defined as $T \in R^{n\times 1}$. We choose the gene

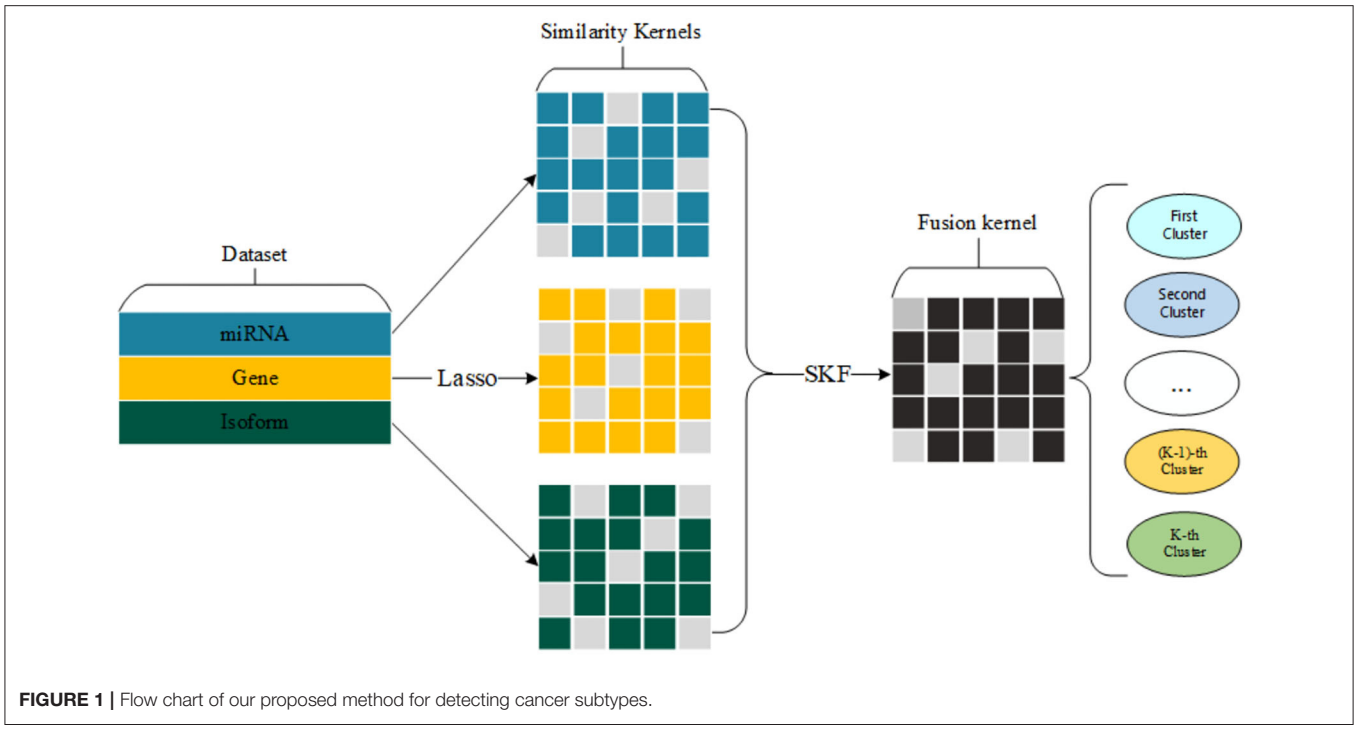

FIGURE 1 | Flow chart of our proposed method for detecting cancer subtypes.

TABLE 1 | Description of Jiang dataset and Novel dataset from TCGA.

	Disease	Patients	Gene	Isoform	miRNA
Jiang dataset	Stomach	1,071	60,483	183	1,881
	Lung	981	60,483	174	1,881
	Kidney	868	60,483	176	1,881
	Breast	1,071	60,483	183	1,881
	Colon	426	60,483	186	1,881
	Bladder	427	60,483	211	1,881
Novel dataset	Blood	165	60,483	166	1,881
	Brain	532	60,483	239	1,881
	Ovary	374	60,483	175	1,881
	Pancreas	177	60,483	262	1,881

expression with the coefficient more than zero as the selected gene features.

2.3. Similarity Kernel Construction

We make use of Chebyshev distance (Krivulin, 2011) instead of traditional Euclidean distance to construct the similarity between two patients. The Chebyshev distance is a metric defined on a vector space where the distance between two vectors is the greatest of their differences along any coordinate dimension. The Chebyshev distance between two vectors p and q, with standard coordinates p_i and q_i, is defined as Equation (2):

$$D_{Chebyshev}(p,q) = \max_i(|p_i - q_i|) \tag{2}$$

The expression data are denoted as $E \in R^{n\times m}$, where n is the number of patients and m is the number of expression factors. The expression data have been centered and scaled to unit variance, as Equation (3):

$$x' = \frac{x - \overline{X}}{S} \tag{3}$$

where x is an element of E, x' is corresponding elements of E after standardization, $\overline{X}$ is the mean of E and S is standard deviation of E. Here, we denote normalized expression data as E'.

Based on the processed expression data E', we construct similarity kernel $K \in R^{n\times n}$ for patients. Here, the similarity between two patients is defined as Equation (4):

$$K_{i,j} = D_{Chebyshev}(e_i, e_j) \tag{4}$$

where $K_{i,j}$ is the similarity between i-th patient and j-th patient, e_i and e_j are two vectors of i-th row and i-th row of E'.

Finally, we get three similarity kernels for a special cancer, including similarity kernel $K_1 \in R^{n\times n}$ by using gene expression, similarity kernel $K_2 \in R^{n\times n}$ by using miRNA expression, and similarity kernel $K_3 \in R^{n\times n}$ by using isoform level.

2.4. Similarity Kernel Fusion

We construct three similarity kernels for patients in the above section. Then, we use similarity kernel fusion (SKF) to combine these kernels into one kernel $K^* \in \mathbb{R}^{n\times n}$.

First, we construct two kernels $P \in \mathbb{R}^{n\times n}$ and $S \in \mathbb{R}^{n\times n}$ for each similarity kernel, where P is a normalized kernel and S is

a sparse kernel that eliminates weak similarity, as Equations (5) and (6):

$$P(i,j) = \frac{K_{i,j}}{\sum_{k=1}^{n} K_{k,j}} \tag{5}$$

where P satisfies $\sum_{k=1}^{n} P(k,j) = 1$.

$$S(i,j) = \begin{cases} 0 & if\ j \notin N_i \\ \frac{K_{i,j}}{\sum_{k \in N_i} K_{i,k}} & if\ j \in N_i \end{cases} \tag{6}$$

where S satisfies $\sum_{k=1}^{n} S(i,j) = 1$, and N_i is a set of top k nearest neighbors of i-th patient including itself.

Second, we uncover more information by using multiple iterations (Wang B. et al., 2014), as Equation (7):

$$P_l^{t+1} = \alpha(S_l \times \frac{\sum_{r \neq l} P_r^t}{2} \times S_l^T) + (1-\alpha)(\frac{\sum_{r \neq l} P_r^0}{2}) \tag{7}$$

where P_l^t $(l = 1, 2, 3)$ is the status of l-th kernel after t iterations, α is a coefficient and satisfies $\alpha \in [0, 1]$, and P_r^0 $(r = 1, 2, 3)$ represents the initial status of P_r.

After $t + 1$ iterations, the overall kernel can be computed as Equation (8):

$$K^* = \frac{1}{3} \sum_{l=1}^{3} P_l^{t+1} \tag{8}$$

2.5. Mining Subtypes Using Spectral Clustering

Through SKF, we have obtained the fusion kernel containing multi-angle information, and the invention of spectral clustering is to cluster through the kernel. So, We employ spectral clustering on integrated similarity kernel to divide all patients into multiple clusters. In order to ensure that the difference between each pair of classes should be as large as possible, also the similarity within one class should be as large as possible, this problem is a relaxation of the NCut problem (Von Luxburg, 2007). The detailed processes of spectral clustering model is introduced as follows.

First, we calculate the Laplacian matrix L based on K^*. Then, we compute the first k generalized eigenvectors $\{u_1, \ldots, u_k\}$ from the generalized eigenproblem $Lu = \lambda Du$, D is a diagonal matrix whose diagonal element is the sum of the row elements of K^*. We define $U \in \mathbb{R}^{n \times k}$ as the matrix containing k vectors $\{u_1, \ldots, u_k\}$ as columns, and $y_i \in \mathbb{R}^k$ as the vector corresponding to the i-th row of U. Finally, we cluster the points $\{y_i\}_{i=1,\ldots,n}$ in $\mathbb{R}^k$ with the k-means clustering algorithm into clusters $\{C_1, \ldots C_k\}$.

We define a matrix $Y \in \mathbb{R}^{n \times k}$, $Y_j = (y_{1,j}, ..., y_{n,j})$ to represent the cluster result (Von Luxburg, 2007), where $y_{i,j} = \frac{1}{vol(\sqrt{Cluster_j})}$ if patient p_i belongs to j-th cluster, otherwise $y_{i,j} = 0$. The whole issue can be transformed into solving the optimization problem, as Equation (9):

$$\min_{T \in \mathbb{R}^{n \times k}} Tr(T' D^{-\frac{1}{2}} L D^{-\frac{1}{2}} T) \quad s.t. T'T = I \tag{9}$$

where D is the degree matrix of K^*, L is the Laplacian matrix of K^*, $T = D^{-\frac{1}{2}} Y$, $vol(A) = \sum_{i \in A} \sum_{j=1}^{n} K_{i,j}^*$.

Here, our proposed method can be shown in Algorithm 1.

Algorithm 1: Algorithm of our proposed method

Require: A patient data matrix $X \in R^{n \times m}$, Patient survival time vector $T \in R^{n \times 1}$.

Ensure: $Y \in \{0, 1\}^{n \times k}$ to represent cluster result, where $Y(i, j) = 1$ if patient p_i belong to j-th cluster.

1: Feature selection through LASSO, as Equation (1);
2: Normalize X and denote expression data as E';
3: Get the similarity kernels $K1, K2, K3 \in R^{n \times n}$, as Equation (4);
4: Use SKF algorithm for kernel fusion, as $K^* \in R^{n \times n}$;
5: Minimize Equation (9) to obtain $Y \in \{0, 1\}^{n \times k}$.

3. RESULTS

In this section, we analyze the performance of our method on the dataset in several ways. First, we introduce an evaluation criteria and a verification method that are used to evaluate the significant performance of cancer subtypes prediction. Second, we analyze the performance of SKF on the Jiang's dataset. Third, we analyze the performance of LASSO on the Jiang's dataset. Fourth, we compare our method with other methods. Fifth, we apply five novel data sets to evaluate our new method. Finally, we plot survival curves and heat maps for some cancers.

TABLE 2 | Comparison between SKF and the single kernel on Jiang Dataset.

Cancer	Gene	miRNA	Isoform	SKF
Stomach (C = 10)	0.196	0.076	0.327	0.002
Lung (C = 7)	0.005	0.173	0.241	1.586×10^{-6}
Kidney (C = 10)	0.082	0.642	0.585	0.018
Breast (C = 5)	0.009	0.680	0.322	1.116×10^{-5}
Colon (C = 11)	0.046	0.050	0.099	1.117×10^{-6}

TABLE 3 | Comparing SKF with different kernels.

Cancer	Euclidean	Chebyshev
Stomach (C = 10)	0.043	0.002
Lung (C = 7)	0.089	1.586×10^{-6}
Kidney (C = 10)	0.175	0.018
Breast (C = 5)	0.042	1.116×10^{-5}
Colon (C = 11)	0.130	1.117×10^{-6}
Bladder (C = 5)	0.147	0.001
Blood (C = 7)	0.805	0.029
Brain (C = 9)	0.040	0.076
Ovary (C = 7)	0.681	0.001
Pancreas (C = 7)	0.243	0.008

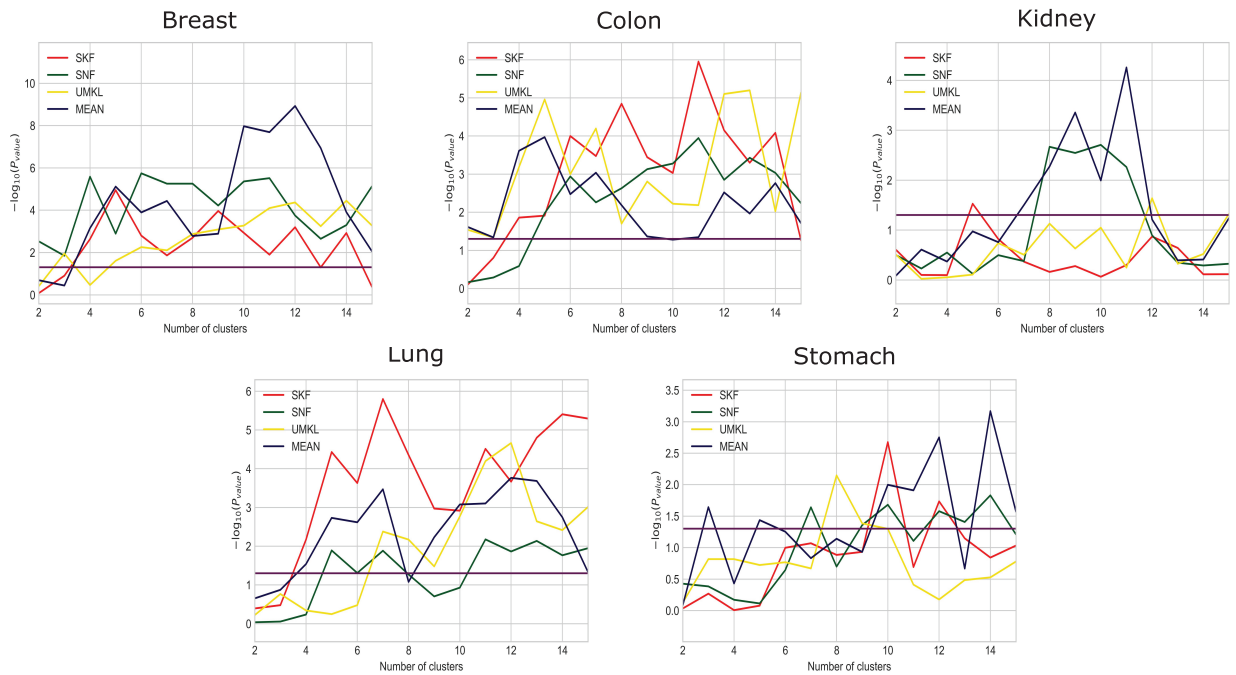

FIGURE 2 | Comparison between SKF and other fusion models.

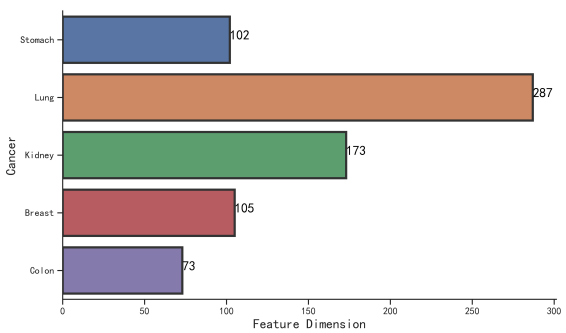

FIGURE 3 | The number of genes selected by LASSO.

3.1. Evaluation Criteria

In this paper, we use the *P*-value of Cox regression model and survival curve to evaluate the performance of our method, while the lower *P*-value indicates higher performance significance. Here, we use 0.05 as a standard for evaluating the performance of clustering results. The actual significance of *P*-value is the difference in survival rates among cancer subtypes. In addition, survival curve is the change of survival rate with survival time. We can find from the survival curve that different cancer subtypes have different survival odds. We can focus on cancer subtypes with high mortality.

3.2. Performance of SKF

In this section, we compare our approach on the use of SKF with the same model on the use of SNF, UMKL, the average kernel fusion or the direct use of single kernel on the Jiang's dataset. There are two important parameters α and K in SKF. We chose $K = 30$ and $\alpha = 0.9$ through experiments. Because the parameter space is very large, we mainly adjust K by fixing α first, and then fix K to adjust α to get an a local The optimal value.

3.2.1. Comparing SKF With Single Kernels

On the Jiang's dataset, we separately record the results of using SKF and using a single kernel, as shown in **Table 2**. We can see that the P value of some diseases is <0.05, despite using the single kernel. However, after using SKF for the kernel fusion, the effects of Lung, Breast, and Colon have been significantly improved. It can be seen that it is necessary to fuse the similarity kernels.

3.2.2. Comparing SKF With Different Kernels

In SKF, the choice of kernel is a very important factor. In most cases, we will choose Euclidean distance as the kernel generation formula, but considering that the dimensionality of biological data is generally large, using Euclidean distance will not have a good effect, we choose Chebyshev distance to construct the kernel. Specifically, it can be seen from the **Table 3** that choosing the Chebyshev distance has a significant improvement in the results.

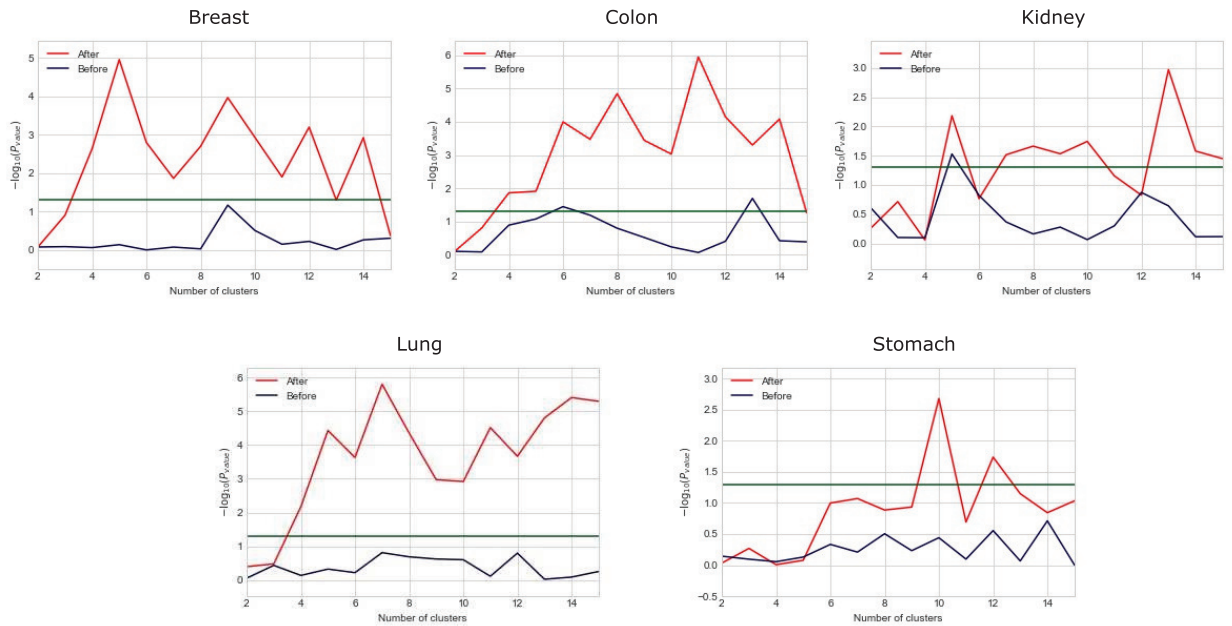

FIGURE 4 | Comparison between the performance of expression data before and after selection by LASSO.

TABLE 4 | Performance of different methods on Jiang's dataset.

Cancer	Our method	Jiang's method
Stomach (C = 10)	0.002	8.86×10^{-14}
Lung (C = 7)	1.586×10^{-6}	3.81×10^{-4}
Kidney (C = 10)	0.018	0.120
Breast (C = 5)	1.116×10^{-5}	6.1×10^{-6}
Colon (C = 11)	1.117×10^{-7}	0.025

TABLE 5 | Performance of our method on Novel dataset.

Cancer	Gene	miRNA	Isoform	SKF
Bladder (C = 5)	0.006	0.010	0.378	0.001
Blood (C = 7)	0.011	0.329	0.258	0.029
Brain (C = 9)	1.934×10^{-7}	0.392	0.585	0.076
Ovary (C = 7)	0.011	0.907	0.167	0.001
Pancreas (C = 7)	0.014	0.507	0.099	0.008

3.2.3. Comparing SKF With Other Fusion Models

We compare the results using SKF with the results of SNF, UMKL, and the average kernel fusion, as shown in **Figure 2** and X axis is the number of clusters and Y axis is the value of $-\log_{10}(P_{value})$. Red, green, yellow, and purple represent the results of using SKF, SNF, UMKL, and average kernel, respectively. And the horizontal line represents the p-value of 0.05. We can see that there is a better performance on Stomach, Lung, and Colon by using SKF. The use of SKF for the kernel fusion on Breast is very similar to that of SNF. It is not as good as SNF on Kidney, but similar to the results of other kernel fusions. Therefore, it can be found that the use of SKF for kernel fusion has an effect on most datasets.

3.3. Performance of LASSO

We observe that the original gene expression data has high dimension. Therefore, we use LASSO to identify the high-efficiency gene expression data. In **Figure 3**, we list the dimensions of gene reduction, and the size of gene is greatly reduced, which is very helpful for later experiments. In **Figure 4**, we compare the performance of expression data before and after selection by LASSO. The X-axis represents the number of clusters, the Y axis represents $-\log_{10}(P_{value})$, the red line represents the data obtained after selection, the blue line represents the data obtained before selection, and the horizontal line represents the P-value of 0.05. We can find that the selection of expression data has a certain influence on the P-value on Stomach, Kidney, while the P-value is greatly improved on Lung, Breast, Colon. Therefore, it can be found that the use of LASSO for selection of expression data has an effect on most datasets.

3.4. Comparing With Other Existing Methods

We compare our approach with the method of Jiang et al. (2019), as shown in **Table 4**. We find that the clustering results of Lung, Kidney, and Colon that using LASSO to select expression data before constructing the kernels and using Chebyshev distance instead of Euclidean distance to construct the kernels, have achieved outstanding performance.

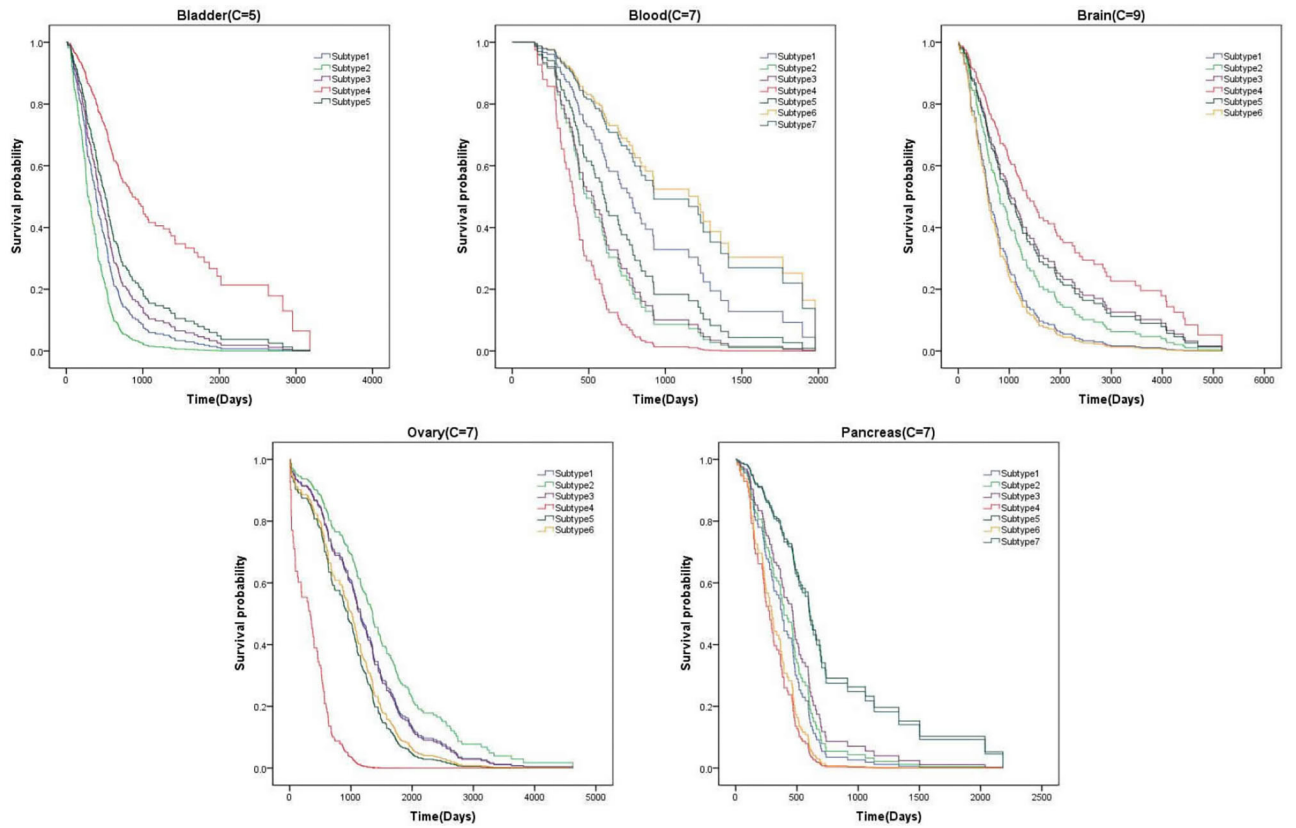

FIGURE 5 | Survival curve on the novel dataset.

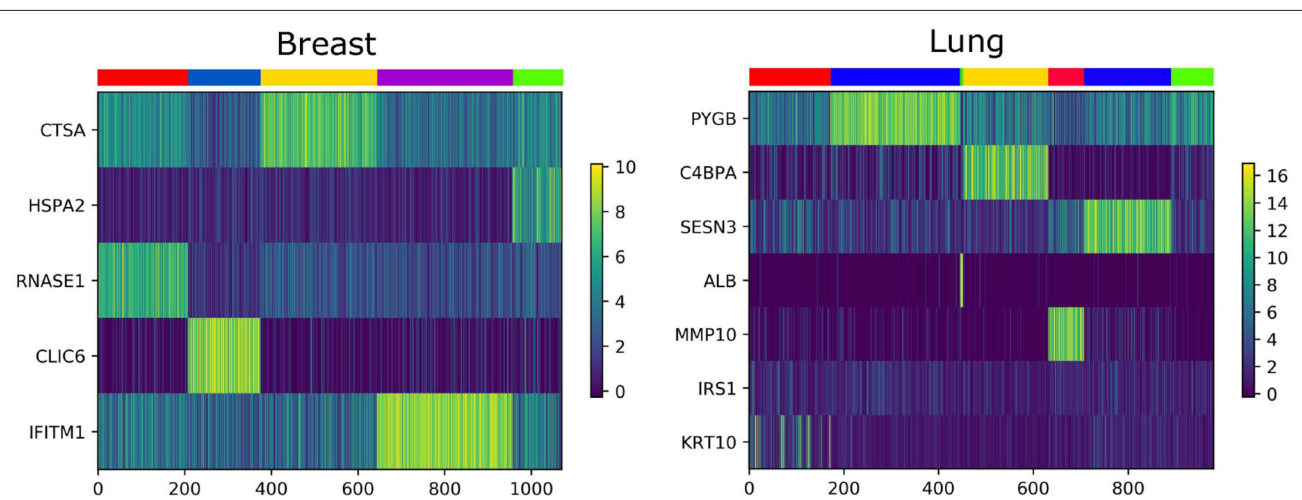

FIGURE 6 | The heat map of essential gene expression data.

3.5. Performance of Our Method on Novel Dataset

In the above section, our method has outstanding performance on Jiang Dataset. To further evaluate this model, we extract five novel datasets from the TCGA website and apply our method to these novel datasets. The detailed results are shown in **Table 5**. We can find that our method performs outstandingly well on Brain, and still has good performance on the remaining four diseases.

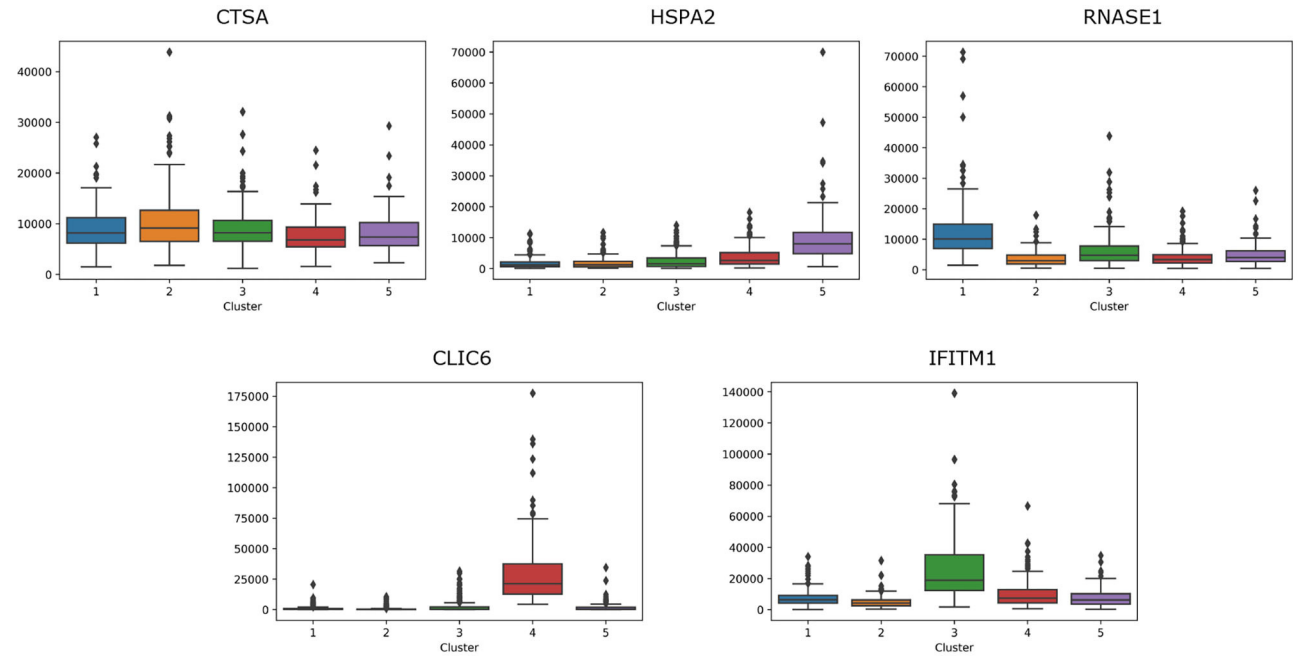

FIGURE 7 | The box plot of essential gene expression data for breast cancer.

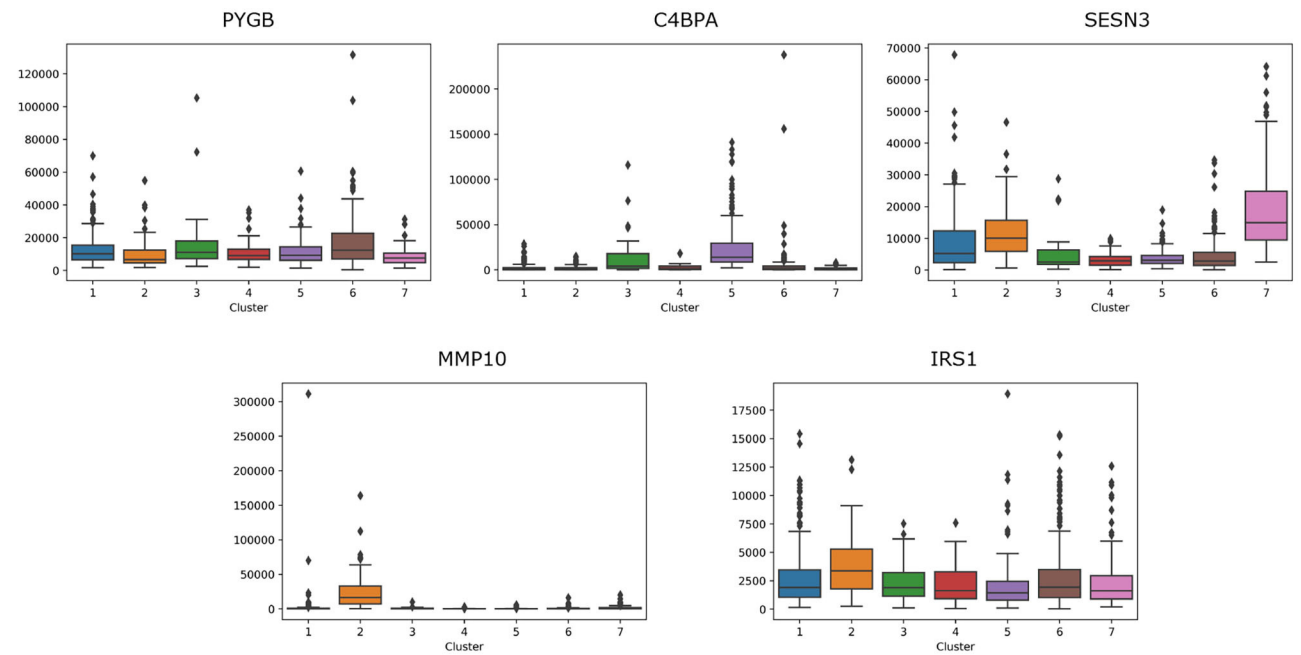

FIGURE 8 | The box plot of essential gene expression data for lung cancer.

3.6. Survival Analysis

From above, we have better measured the performance of clustering results on *P*-value. In this section, We list the survival curves of five cancers on the novel dataset, as shown in **Figure 5**. We can find that the difference of tendency between each subtype is very obvious on two cancers. It demonstrates that the clustering results have positive guidance for clinical treatment.

3.7. Analysis of Essential Genes

We analyze the importance of essential genes on Lung and Breast datasets. The association between clustering results and

expression data are shown in **Figure 6**. The X-axis is the patient, the Y-axis is the gene, and each color of the upper color block represents a category. We find that some essential genes have an effect on the identification of cancer subtypes, most of them can be confirmed by the GEO Profile Database.

For breast cancer, we select five essential genes, such as CTSA, HSPA2, RNASE1, CLIC6, IFITM1. We analyze the box plot of essential gene expression data in five categories, as shown in **Figure 7**. We find that, HSPA2A is highly expressed in 5-th group, RNASE1 is highly expressed in 1-th group, CLIC6 is highly expressed in 4-th group, and IFITM1 is highly expressed in 3-th group.

For lung cancer, we select five essential genes, such as PYGB, C4BPA, SESN3, MMP10, IRS1. We analyze the box plot of essential gene expression data in seven categories, as shown in **Figure 8**. We find that, C4BPA is highly expressed in 3-th and 5-th groups, SESN3 is highly expressed in 2-th and 7-th groups, and IRS1 is only highly expressed in 2-th group.

4. CONCLUSION

In this paper, we extract five novel datasets (bladder cancer, blood cancer, brain cancer, ovary cancer, and pancreas cancer) from the TCGA website. We find that our method not only works well on the Jiang's dataset, but also performs well on our newly extracted five datasets. In addition, we obtain some important genes that are related to a special cancer.

In the future, we will try to employ more kinds of expression data to further uncover cancer subtype because cancer is a multi-factors disease (Guo F. et al., 2018). We can also consider other machine learning methods or deep learning methods to uncover cancer subtype rather than spectral clustering (Ding et al., 2019; Shen et al., 2019).

DATA AVAILABILITY STATEMENT

All datasets presented in this study are included in the article/supplementary material.

AUTHOR CONTRIBUTIONS

FG, LJ, and SL conceived and designed the experiments. SL and LJ performed the experiments and analyzed the data. SL, NG, and FG wrote the paper. FG, NG, and JT supervised the experiments and reviewed the manuscript. All authors have participated in study discussion and manuscript preparation. All authors read and approved the final manuscript.

ACKNOWLEDGMENTS

This work was supported by a grant from the National Natural Science Foundation of China (NSFC 61772362, 61702456, 61972280), and National Key R&D Program of China (2018YFC0910405, 2017YFC0908400).

REFERENCES

Brunet, J., Tamayo, P., Golub, T., and Mesirov, J. (2004). Metagenes and molecular pattern discovery using matrix factorization. *Proc. Natl. Acad. Sci. U.S.A.* 101, 4164–4169. doi: 10.1073/pnas.0308531101

Chen, L., Pan, X., Zhang, Y.-H., Liu, M., Huang, T., and Cai, Y.-D. (2019). Classification of widely and rarely expressed genes with recurrent neural network. *Comput. Struct. Biotechnol. J.* 17, 49–60. doi: 10.1016/j.csbj.2018.12.002

Chen, P., Fan, Y., Man, T.-k., Hung, Y., Lau, C. C., and Wong, S. T. (2013). A gene signature based method for identifying subtypes and subtype-specific drivers in cancer with an application to medulloblastoma. *BMC Bioinformatics* 14:S1. doi: 10.1186/1471-2105-14-S18-S1

Ding, Y., Tang, J., and Guo, F. (2017). Identification of drug-target interactions via multiple information integration. *Inform. Sci.* 418–419, 546–560. doi: 10.1016/j.ins.2017.08.045

Ding, Y., Tang, J., and Guo, F. (2019). Identification of drug-side effect association via multiple information integration with centered kernel alignment. *Neurocomputing* 325, 211–224. doi: 10.1016/j.neucom.2018.10.028

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., et al. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 542:115. doi: 10.1038/nature21056

Fedele, C., Tothill, R. W., and Mcarthur, G. A. (2014). Navigating the challenge of tumor heterogeneity in cancer therapy. *Cancer Discov.* 4:146. doi: 10.1158/2159-8290.CD-13-1042

Fu, F., Nowak, M. A., and Bonhoeffer, S. (2014). Spatial heterogeneity in drug concentrations can facilitate the emergence of resistance to cancer therapy. *PLoS Comput. Biol.* 11:e1004142. doi: 10.1371/journal.pcbi.1004142

Guo, F., Wang, D., and Wang, L. (2018). Progressive approach for SNP calling and haplotype assembly using single molecular sequencing data. *Bioinformatics* 34, 2012–2018. doi: 10.1093/bioinformatics/bty059

Guo, Y., Qi, Y., Li, Z., and Shang, X. (2018). Improvement of cancer subtype prediction by incorporating transcriptome expression data and heterogeneous biological networks. *BMC Med. Genomics* 11:119. doi: 10.1186/s12920-018-0435-x

Haase, R., Michie, M., and Skinner, D. (2015). Flexible positions, managed hopes: The promissory bioeconomy of a whole genome sequencing cancer study. *Soc. Sci. Med.* 130, 146–153. doi: 10.1016/j.socscimed.2015.02.016

Jiang, L., Xiao, Y., Ding, Y., Tang, J., and Guo, F. (2018). FKL-Spa-LapRLS: an accurate method for identifying human microRNA-disease association. *BMC Genomics* 19:911. doi: 10.1186/s12864-018-5273-x

Jiang, L., Xiao, Y., Ding, Y., Tang, J., and Guo, F. (2019). Discovering cancer subtypes via an accurate fusion strategy on multiple profile data. *Front. Genet.* 10:20. doi: 10.3389/fgene.2019.00020

Karabatak, M., and Ince, M. C. (2009). An expert system for detection of breast cancer based on association rules and neural network. *Expert Syst. Appl.* 36, 3465–3469. doi: 10.1016/j.eswa.2008.02.064

Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., and Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. *Comput. Struct. Biotechnol. J.* 13, 8–17. doi: 10.1016/j.csbj.2014.11.005

Krivulin, N. (2011). An algebraic approach to multidimensional minimax location problems with chebyshev distance. *WSEAS Trans. Math. Arch.* 10, 191–200.

Li, Y., Wu, F.-X., and Ngom, A. (2016). A review on machine learning principles for multi-view biological data integration. *Brief. Bioinformatics* 19, 325–340. doi: 10.1093/bib/bbw113

Li, Y.-X., and Ruan, X.-G. (2005). Cancer subtype recognition and feature selection with gene expression profiles. *Acta Electron. Sin.* 33, 651–655.

Ma, T., and Zhang, A. (2017). "Integrate multi-omic data using affinity network fusion (ANF) for cancer patient clustering," in *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (Kansas, MO), 398–403. doi: 10.1109/BIBM.2017.8217682

Mariette, J., and Villa-Vialaneix, N. (2017). Unsupervised multiple kernel learning for heterogeneous data integration. *Bioinformatics* 34, 1009–1015. doi: 10.1093/bioinformatics/btx682

Marino, F. Z., Accardo, M., and Franco, R. (2017). Crispr-barcoding in non small cell lung cancer: from intratumor genetic heterogeneity modeling to cancer therapy application. *J. Thorac. Dis*. 9:1759. doi: 10.21037/jtd.2017.06.27

Miotto, R., Wang, F., Wang, S., Jiang, X., and Dudley, J. T. (2017). Deep learning for healthcare: review, opportunities and challenges. *Brief. Bioinformatics* 19, 1236–1246. doi: 10.1093/bib/bbx044

Mirza, B., Wang, W., Wang, J., Choi, H., Chung, N. C., and Ping, P. (2019). Machine learning and integrative analysis of biomedical big data. *Genes* 10:87. doi: 10.3390/genes10020087

Monti, S., Tamayo, P., Mesirov, J., and Golub, T. (2003). Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Mach. Learn*. 52, 91–118.

Pan, X., Hu, X. H., Zhang, Y. H., Chen, L., Zhu, L. C., Wan, S. B., et al. (2018). Identification of the copy number variant biomarkers for breast cancer subtypes. *Mol. Genet. Genomics* 294:1–16. doi: 10.1007/s00438-018-1488-4

Shen, Y., Tang, J., and Guo, F. (2019). Identification of protein subcellular localization via integrating evolutionary and physicochemical information into chou's general PseAAC. *J. Theor. Biol*. 462, 230–239. doi: 10.1016/j.jtbi.2018.11.012

Sohn, B. H., Hwang, J. E., Jang, H. J., Lee, H. S., Oh, S. C., Shim, J. J., et al. (2017). Clinical significance of four molecular subtypes of gastric cancer identified by the cancer genome atlas project. *Clin. Cancer Res*. 23, 4441–4449. doi: 10.1158/1078-0432.CCR-16-2211

Tibshiranit, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B-Methodol.* 58, 267–288. doi: 10.1111/j.2517-6161.1996.tb02080.x

Tomczak, K., Czerwiska, P., and Wiznerowicz, M. (2015). The cancer genome atlas (TCGA): an immeasurable source of knowledge. *Contemp. Oncol*. 19, 68–77. doi: 10.5114/wo.2014.47136

Verhaak, R. G., Hoadley, K. A., Purdom, E., Wang, V., Qi, Y., Wilkerson, M. D., et al. (2010). Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell* 17, 98–110. doi: 10.1016/j.ccr.2009.12.020

Von Luxburg, U. (2007). A tutorial on spectral clustering. *Stat. Comput*. 17, 395–416. doi: 10.1007/s11222-007-9033-z

Wang, B., Mezlini, A. M., Demir, F., Fiume, M., Tu, Z., Brudno, M., et al. (2014). Similarity network fusion for aggregating data types on a genomic scale. *Nat. Methods* 11, 333–337. doi: 10.1038/nmeth.2810

Wang, D., Khosla, A., Gargeya, R., Irshad, H., and Beck, A. H. (2016). Deep learning for identifying metastatic breast cancer. *arXiv preprint arXiv:1606.05718*.

Wang, D., Li, J.-R., Zhang, Y.-H., Chen, L., Huang, T., and Cai, Y.-D. (2018). Identification of differentially expressed genes between original breast cancer and xenograft using machine learning algorithms. *Genes* 9:155. doi: 10.3390/genes9030155

Wang, K., Yuen, S. T., Xu, J., Lee, S. P., Yan, H. H., Shi, S. T., et al. (2014). Whole-genome sequencing and comprehensive molecular profiling identify new driver mutations in gastric cancer. *Nat. Genet*. 46, 573–82. doi: 10.1038/ng.2983

Wang, V., Li, C., Lin, M., Welch, W., Bell, D., Wong, Y. F., et al. (2005). Ovarian cancer is a heterogeneous disease. *Cancer Genet. Cytogenet*. 161, 170–173. doi: 10.1016/j.cancergencyto.2004.12.014

Wong, G., Leckie, C., and Kowalczyk, A. (2012). FSR: feature set reduction for scalable and accurate multi-class cancer subtype classification based on copy number. *Bioinformatics* 28:151. doi: 10.1093/bioinformatics/btr644

Zhang, W., Feng, H., Wu, H., and Zheng, X. (2017). Accounting for tumor purity improves cancer subtype classification from DNA methylation data. *Bioinformatics* 33:2651. doi: 10.1093/bioinformatics/btx303

Zhao, Z., Xu, L., Shi, X., Tan, W., Fang, X., and Shangguan, D. (2009). Recognition of subtype non-small cell lung cancer by Dna aptamers selected from living cells. *Analyst* 134, 1808–1814. doi: 10.1039/b904476k

Zhou, Z.-H., Jiang, Y., Yang, Y.-B., and Chen, S.-F. (2002). Lung cancer cell identification based on artificial neural network ensembles. *Artif. Intell. Med*. 24, 25–36. doi: 10.1016/S0933-3657(01)00094-X

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.