Bioinformatics, 2019, 1–7
doi: 10.1093/bioinformatics/btz772
Advance Access Publication Date: 11 October 2019
Original Paper

Genome analysis

Cancer classification of single-cell gene expression data by neural network

Bong-Hyun Kim[1,2,†], Kijin Yu[1] and Peter C. W. Lee [1,*]

[1]Department of Biomedical Sciences, University of Ulsan College of Medicine, ASAN Medical Center, Seoul 05505, Korea and [2]Advanced Bio Computing Center, Frederick National Laboratory for Cancer Research, Frederick, MD 21702, USA

*To whom correspondence should be addressed.

†Present address: BlueSphere Bio, Pittsburgh, PA 15219, USA.

Associate Editor: Janet Kelso

Received on April 28, 2019; revised on August 13, 2019; editorial decision on October 4, 2019; accepted on October 8, 2019

Abstract

Motivation: Cancer classification based on gene expression profiles has provided insight on the causes of cancer and cancer treatment. Recently, machine learning-based approaches have been attempted in downstream cancer analysis to address the large differences in gene expression values, as determined by single-cell RNA sequencing (scRNA-seq).

Results: We designed cancer classifiers that can identify 21 types of cancers and normal tissues based on bulk RNA-seq as well as scRNA-seq data. Training was performed with 7398 cancer samples and 640 normal samples from 21 tumors and normal tissues in TCGA based on the 300 most significant genes expressed in each cancer. Then, we compared neural network (NN), support vector machine (SVM), k-nearest neighbors (kNN) and random forest (RF) methods. The NN performed consistently better than other methods. We further applied our approach to scRNA-seq transformed by kNN smoothing and found that our model successfully classified cancer types and normal samples.

Availability and implementation: Cancer classification by neural network.

Contact: pclee@amc.seoul.kr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Due to the recent interest in next-generation sequencing techniques, the genomes and transcriptomes of cancer patients have been investigated with whole genome sequencing, whole exome sequencing and RNA sequencing (Agarwal and Owzar, 2014; Cancer Genome Atlas Network, 2012a, b; Cancer Genome Atlas Research Network, 2008, 2011; Fritz *et al.*, 2011; Meyerson *et al.*, 2010; Wang *et al.*, 2013, 2015). One of the most significant efforts is The Cancer Genome Atlas (TCGA) program, which has generated and deposited multi-platform genomic analyses, such as gene expression, protein expression, mutation, DNA methylation and copy number variation, of over 11 000 patients representing 33 cancer types (Tomczak *et al.*, 2015). The large volume of data typically have led to discoveries of cancer-specific tumorigenic features through comparisons between tumor and non-tumor data (Cancer Genome Atlas Network, 2012a, b; Cancer Genome Atlas Research Network, 2008, 2011). The accumulation of these tumorigenic events allows us to further analyze their biological relevance across multiple cancer types as well as further investigate the downstream events. Cancer-specific features help us to understand other types of less-studied cancers, and we can predict that the cancers with a similar set of tumorigenic features might share similar characteristics and prognoses (Aran *et al.*, 2015; Cancer Genome Atlas Research Network, 2013; Chen *et al.*, 2018; Hoadley *et al.*, 2014, 2018; Martinez *et al.*, 2015; Peng *et al.*, 2015; Wan *et al.*, 2015; Zhang and Zhang, 2017).

Single-cell RNA sequencing (scRNA-seq) methods have provided a powerful analytical basis for understanding tumor heterogeneities or microenvironments. However, scRNA-seq gene expression measurements exhibit natural heterogeneity and detection noise, so-called 'dropouts'. This occurs due to stochastic fluctuations and poor conversion rates in capturing RNA in the single cell and converting them to cDNA as well as downstream sequencing reads. To overcome statistical fluctuations or dropout problems, scRNA-seq denoising methods have been developed in smoothing and machine learning (ML) (Chen and Zheng, 2018; Eraslan *et al.*, 2019; Ronen and Akalin, 2018; Wagner *et al.*, 2018). ML methods have been utilized to preprocess scRNA-seq data and determine cell types as well as to predict diagnoses, symptoms and prognoses from various types of data. ML methods such as k-nearest neighbors (kNN), naïve Bayesian (NB), support vector machine (SVM), neural network (NN) and other methods have been implemented based on multi-type data like sequencing, patient clinic and tissue images (Angermueller *et al.*, 2016; Kim and Kim, 2018; Kourou *et al.*, 2015; Li *et al.*, 2017; Zararsiz *et al.*, 2017).

For binary classifications of a breast cancer (BRCA) and healthy samples, a study was conducted in which the deep learning method was applied to identify BRCA-specific gene set and separate cancer samples and normal samples (Danaee *et al.*, 2017). Gene relationships were extracted from high dimensional gene expression data using stacked denoising autoencoders and deeply connected genes (DCGs) were selected from 500 genes for interpretation of gene set. NN resulted in 91.74% accuracy on DCGs enriched in pathways whose functional roles are related with cell proliferation and tumor suppression.

For the multi-class classification of 20 cancer types, Kim and Kim (2018) used kNN on 868 023 SNP loci whose significant values were selected from genome-wide association studies. The level of accuracy spans from 33% to 88% corresponding to Thyroid Carcinoma and Pheochromocytoma and Paraganglioma (PCPG), and median accuracy was 44% for predicting 21 phenotypes. Li *et al.* (2017) classified 32 cancer types in RNA-seq data of TCGA and applied kNN as the pan-cancer classifier with a subset of 20 genes iteratively selected by genetic algorithm (GA). Unfortunately, they ran the GA/kNN one thousand times to search for the best gene set, and the predictions were not consistent among cancer types. Also, this method required high computation time to retrieve the best classification.

Here, we define the smallest number of genes necessary to reliably classify 21 cancer types from normal samples using transcriptome in TCGA. This set of pan-cancer gene features is generated by selecting differentially expressed genes in each cancer by comparing it against the adjacent normal tissues. We compared the performances of classifications with linear SVM (L-SVM), radial basis function-kernel SVM (RBF-SVM), kNN, random forest (RF) and NN. Then, the best classifier in classifying bulk RNA-seq data was applied on the scRNA-seq data based on whether it can predict cancer types of single-cell samples accurately with the proposed of gene set (Fig. 1). We also demonstrated that these selected sets of genes are annotated with functional roles that are related to tumorigenic processes. By finding the correlation of tumorigenic features between different types of cancers, efficient therapeutics for cancer treatment can be established.

2 Materials and methods

2.1 Data preparation

The RNA-seq data of 21 cancer types and tumor-adjacent normal tissues from TCGA were used to build and validate the ML models. Additional single-cell RNA-seq data of 2 cancer types were also applied to test our cancer classifiers.

The pan-organ cancer types, such as colorectal adenocarcinoma, pan-kidney cohort (KICH + KIRC + KIRP), glioma and stomach and esophageal carcinoma, were excluded to prevent sample duplication on training data since they are integrated with two or more cancer types. Finally, all available samples via TCGA2STAT (ver 1.3.1), total of 7398 tumor samples of 21 cancer types that have adjacent normal tissues and 640 samples of normal tissue adjacent to 21 tumor tissues were downloaded for our classification (Table 1).

As the test set of the single-cell samples, breast cancer (GSE75688) (Chung *et al.*, 2017) and skin melanoma (GSE72056) (Tirosh *et al.*, 2016) data, normalized by the RSEM (Li and Dewey, 2011) were downloaded from Gene Expression Omnibus (GEO) (https://www.ncbi.nlm.nih.gov/geo/). The cancer and normal samples were annotated by the authors; 317 breast cancer (BRCA) cells and 198 normal cells were isolated from 11 patients and 1257 cancer cells and 3256 normal cells from were obtained from 19 melanoma (SKCM) patients.

Expression values calculated by the RSEM from the RNA-seq data covering 20 501 genes were obtained through TCGA2STAT (Wan *et al.*, 2016), an open source R package. Total 20 501 gene expressions in each sample were normalized by Z score (Zill *et al.*, 2011) to manipulate different expression levels depending on the individual. For instance, the expression of the TRIM28 gene in 517 samples of LUAD ranged from 1076 to 20 289 transcripts per million. These variances may affect the performance of each classifier and increase the computational complexity of classification.

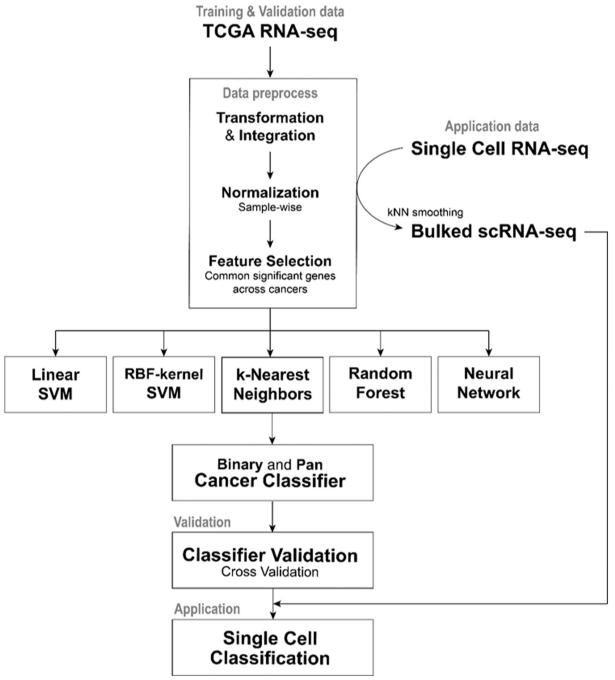

Fig. 1. Pan-cancer classification workflow

2.2 Smoothing scRNA-seq data by k-nearest neighbors

To remove stochastic noise and heterogeneity, we smoothed scRNA-seq data using the k-nearest neighbors (kNN) method. First, the scRNA-seq datasets were divided into three classes: BRCA, SKCM and NORMAL (selected from breast cancer and melanoma patients scRNA-seq datasets). For each class, we selected 100 initial cells randomly, then Manhattan distance (Riesz, 1910) was used to select k-nearest cells for the average expression values from the initial cell. The number of single cells, including the initial and closest cells for smoothing, was tested with 12 values as follows: 1, 3, 5, 10, 20, 30, 50, 100, 150, 200, 250 and 300. Since cancer classification results may vary depending on the selected initial cells, 10 datasets were generated in the same way to confirm the robustness of the classification criteria. To determine the amount of data required for the classification test, a total of 360 smoothed test datasets were generated by selecting the initial cell count of each cancer type as 100, 200 and 300.

2.3 ML models for cancer classifier

Five ML algorithms, L-SVM, RBF-SVM (Cortes and Vapnik, 1995), kNN (Altman, 1992), RF (Barandiaran, 1998) and NN (Hopfield, 1982), were built to determine which best classifies cancer types. All models were learned from the same training data generated by selecting 75% of the TCGA data, and the remaining 25% were used as validation data to measure and compare the performance of the model. Each algorithm was also tested with combinations of parameters; finally, we found that $c = 10$ for the L-SVM, $c = 100$ for the RBF-SVM, $k = 7$ for the kNN, 100 trees for RF and two hidden layers for the NN produced the best results. To evaluate overall performance of each model, 10-fold cross validation was performed. Of 10 divided sets from TCGA data, the process by which the learned model predicts remaining one set was repeated 10 times, and eventually, all data were used for validation. All ML algorithms were implemented in the scikit-learn package (Pedregosa *et al.*, 2011).

2.4 Performance evaluation

Due to binary classification on imbalanced data, which consists of 13 times more cancer samples than normal samples, the performance of data validation was calculated according to the accuracy

Table 1. TCGA cancer types and the number of RNA-seq samples used in this paper

Cancer name	Code	Cases	Cancer name	Code	Cases
Bladder urothelial carcinoma	BLCA	408	Liver hepatocellular carcinoma	LIHC	373
Breast invasive carcinoma	BRCA	1100	Lung adenocarcinoma	LUAD	517
Cervical and endocervical cancers	CESC	306	Lung squamous cell carcinoma	LUSC	501
Cholangiocarcinoma	CHOL	36	Pancreatic adenocarcinoma	PAAD	179
Colon adenocarcinoma	COAD	287	Pheochromocytoma and paraganglioma	PCPG	184
Esophageal carcinoma	ESCA	185	Prostate adenocarcinoma	PRAD	498
Glioblastoma multiforme	GBM	166	Rectum adenocarcinoma	READ	95
Head and neck squamous cell carcinoma	HNSC	522	Sarcoma	SARC	263
Kidney chromophobe	KICH	66	Skin cutaneous melanoma	SKCM	472
Kidney renal clear cell carcinoma	KIRC	534	Stomach adenocarcinoma	STAD	415
Kidney renal papillary cell carcinoma	KIRP	291	Adjacent to 21 tumors	NORMAL	640

(ACC) from 0 to 1 and the Matthews Correlation Coefficient (MCC) (Matthews, 1975) from −1 to 1.

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN} + \mathrm{TN}} \tag{1}$$

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}} \tag{2}$$

In Equations (1) and (2), TP is the number of samples correctly predicted as CANCER in cancer samples, FN is the number of samples incorrectly predicted as NORMAL in cancer samples, FP is the number of samples incorrectly predicted as CANCER in normal samples and TN is the number of samples correctly predicted as NORMAL in normal samples. ACC ranges from 0 to 1, with a perfect classification at 1 and a completely incorrect classification at 0. MCC ranges from −1 to 1, with a completely wrong classification at −1 and perfect classification at 1.

As with the ACC of binary classification, the ACC of pan-cancer classification is a precisely categorized rate of the whole sample and has a value from 0 to 1.

$$\mathrm{ACC} = \frac{1}{n_{\mathrm{samples}}} \sum_{i=0}^{n_{\mathrm{samples}}-1} 1(\hat{y}_i = y_i) \tag{3}$$

In Equation (3), n_{samples} is the total number of samples, and ACC is 1 if y_i, the number of i-th class samples, is equal to $\hat{y}_i$, the number of samples predicted by the i-th class.

MCC of pan-cancer classification is defined as:

$$\mathrm{MCC} = \frac{c \times s - \sum_k^K p_k \times t_k}{\sqrt{\left(s^2 - \sum_k^K p_k^2\right) \times \left(s^2 - \sum_k^K t_k^2\right)}} \tag{4}$$

In Equation (4), s is the total number of samples, c is the total number of correctly predicted samples, $t_k = \sum_i^K C_{ik}$ is the number of all samples in class k and $p_k = \sum_i^K C_{ki}$ is the number of correctly predicted samples in class k. Unlike ACC, MCC of pan-cancer classification for perfect prediction is 1, but the minimum is somewhere between −1 and 0, depending on the number and distribution of the actual labels.

2.5 Pan-cancer gene set

In order to determine the optimal number of genes for training ML models, we run ANOVA test comparing cancer and normal samples for each cancer types. Then we selected top n genes of highest F-values. As we selected highest n genes, we did not perform any multiple test correction. Then calculated frequencies of selection per genes for all 21 cancer types and determine final top n genes. Here, we used 12 different n values 5, 10, 15, 20, 25, 30, 50, 100, 150, 200, 250 and 300 (Fig. 2).

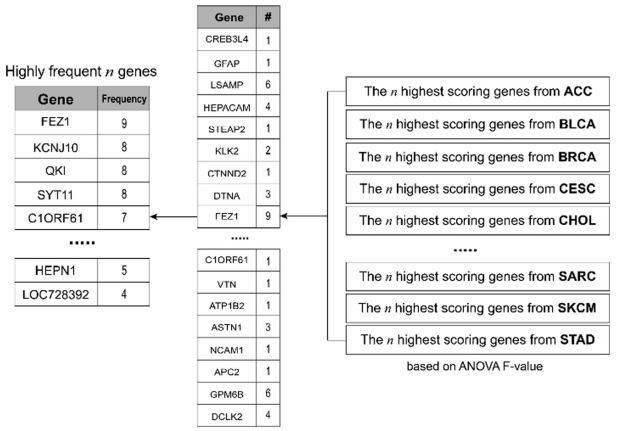

Fig. 2. Composition procedure of pan-cancer gene set for classification. Genes with significant variance in expression were selected from each cancer type by ANOVA and frequently expressed genes comprised the significant gene sets

2.6 Functional analysis of gene set

The functional roles of the proposed gene set were annotated by gene set enrichment analysis (GSEA) (Subramanian *et al.*, 2005) using GSEAPY 0.9.7 library (https://pypi.org/project/gseapy/), canonical pathways obtained from Gene Ontology (GO) (Ashburner *et al.*, 2000), KEGG (Kanehisa and Goto, 2000) and BioCarta (Nishimura, 2001) organized in the Molecular Signatures Database (MSigDB version). The heatmaps across cancer types in TCGA and single-cell data were generated to compare gene expressions of significantly enriched pathways ordered by adjusted P-values.

3 Results

We generated 12 sets of different sizes—from 5 to 300—of significant genes by selecting cancer-specific genes that exhibit different gene expression values as compared to the adjacent normal samples. For example, the tumor protein p53 gene (TP53) are expressed in similar levels among many cancer types (Hoadley *et al.*, 2014) and the serine proteinase inhibitor, member 3 gene (SERPINB3) is highly expressed in squamous cell carcinomas, CESC, HNSC and LUSC (Sheshadri *et al.*, 2014). We built the binary classifier for cancer-normal classification on these common genes since they can be used to classify samples as cancerous or non-cancerous efficiently, but these genes might not be effective in distinguishing a specific type of cancer from others by their expression values. For the multi-class classification of cancer types, we derived highly frequent genes across the significant gene lists of the 21 cancer samples and built pan-cancer classifier based on the gene set. Moreover, the precision performances of our classifiers were tested with transformed single-cell RNA-seq data by kNN smoothing.

3.1 Binary classification of cancer and normal

Five types of ML models were trained with 12 different sizes of gene expression datasets—5, 10, 15, 20, 25, 30, 50, 100, 150, 200, 250 and 300 genes—to determine the optimal number of genes for binary classification performance. First, 7398 samples from 21 cancer types were merged into the CANCER class, and 640 samples of adjacent normal tissues were labeled as the NORMAL class. All samples were split in a ratio of 0.75 and 0.25 as training and validation data, and five ML models were trained with identical 12 gene sets.

The results indicate that as the number of genes increases, the accuracy of the models increases in kind. NN achieved the best performance of MCC 0.92 and ACC 0.99 when learning with 300 genes (Fig. 3A and Supplementary Fig. S1A). kNN and RF achieved MCC 0.8 and 0.83, respectively, when learning with 200 genes, and L-SVM and RBF-SVM achieved MCC 0.69 and 0.83, respectively, using 300 genes. Figure 3A shows that NNs results with 10 genes performed better than the best results of kNN.

Since NN exhibited the best performance, we used cross validation to better measure our model. The performance of the NN model scored MCC 0.68 and ACC 0.94 (Fig. 3C) by 10-fold cross validation. Because 7025 cancer samples out of 7398 were correctly classified, high ACC was achieved, but only 541 of 640 normal samples were correctly classified and exhibited relatively low MCC.

Among total number of misclassified as NORMAL (or false normal) of 373 samples, 85% were from LUAD, LUSC, SKCM and STAD where false normal samples are 25, 107, 62 and 124, respectively. To analyze the relationship between these samples, we performed t-distributed Stochastic Neighbor Embedding (tSNE), which conducts two-dimensional data reduction and mapped it to the binary classification result (Fig. 4A). Most of the cancer samples had tSNE1 values of −50 to 45, whereas 318 samples of the four cancer types that were misclassified as NORMAL were distributed over tSNE1 values of 30–40, shown in Fig. 4C (Supplementary Fig. S3M, N, T and U). However, the samples classified as CANCER (false cancer) have no distribution difference from most normal samples, and all have tSNE1 values of 10–45. Seventy-one percentage of 99 normal samples categorized as CANCER are samples of BRCA (14), HNSC (24), PRAD (21) and STAD (11).

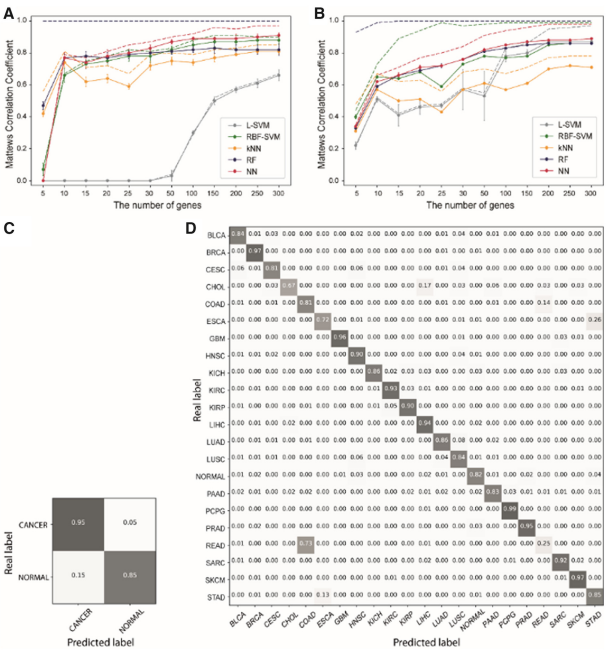

Fig. 3. The classification results of TCGA RNA-seq data. Graphs show the performance of the five ML models according to the number of genes for binary classification (**A**) and multiple classification (**B**). NN achieved better MCC than other four ML models in both classifications. (**C**) The confusion matrix shows the accuracy of NN model with 300 genes for binary classification. It is identified 95% of cancer samples and 85% of non-cancer samples correctly. (**D**) The confusion matrix shows the accuracy of NN model using 300 genes for multi-classification. It is identified 10 classes with over ACC 0.9 and achieved MCC 0.88 and ACC 0.88 on average

3.2 Pan-cancer classification

For pan-cancer classifiers, five ML models were trained with 12 different gene set sizes from 22 phenotypes—21 cancer samples and 1 normal type. Similar to the result of the binary classifier, increasing the size of gene sets generally led to higher classification accuracy. NN, at 300 significant genes, resulted in the best performance with MCC 0.89 and ACC 0.9; kNN was the worst classifier with MCC 0.71 (Fig. 3B and Supplementary Fig. S1B). In pan-cancer classification, L-SVM (MCC 0.87) and RF (MCC 0.86) perform more accurately than binary classification, and NN (MCC 0.89), RBF-SVM (MCC 0.87) and kNN (MCC 0.71) performs more poorly than binary classification.

The classifier was evaluated by 10-fold cross validation for the dataset-independent validation, and the results are represented in Fig. 3D. The pan-cancer classifier scores were MCC 0.88 and ACC 0.88. Over 90% samples from 10 classes, including NORMAL, were correctly predicted by the pan-cancer classifier. Although largely different sample sizes per single cancer types in range from 36 to 1100, the pan-cancer classifier positively predicted over 80% samples of 19 classes.

For classes with <100 samples, the accuracy is relatively variable: KICH exhibited ACC 0.9 while CHOL exhibited ACC 0.67 and READ exhibited ACC 0.25. Several cancer types were incorrectly classified as one another as they were grouped together. Of all the BLCA samples, 7% were falsely predicted to be CESC and LUSC; 16% of the CESC samples were categorized as BLCA, HNSC and LUSC and 8% of LUSC samples were categorized as BLCA, CESC and HNSC. Although BLCA, CESC, HNSC and LUSC occur in different organs, these four types of cancers originate in squamous cells that form the surface of skin and organs (Cancer Genome Atlas Research Network, 2013; Zack *et al.*, 2013). Thus, LUSC have relatively lower true values for prediction than others overall due to 4% of the samples LUSC being predicted as LUAD. Subtypes of lung cancer, KICH, KIRC and KIRP, occurred in kidney tissues and were also misclassified as one another (Hoadley *et al.*, 2018).

Interestingly, we found the frequently misclassified cancers have similar tissues of origins. For example, cancers related in upper digestive tracks, ESCA and STAD showed high false classification

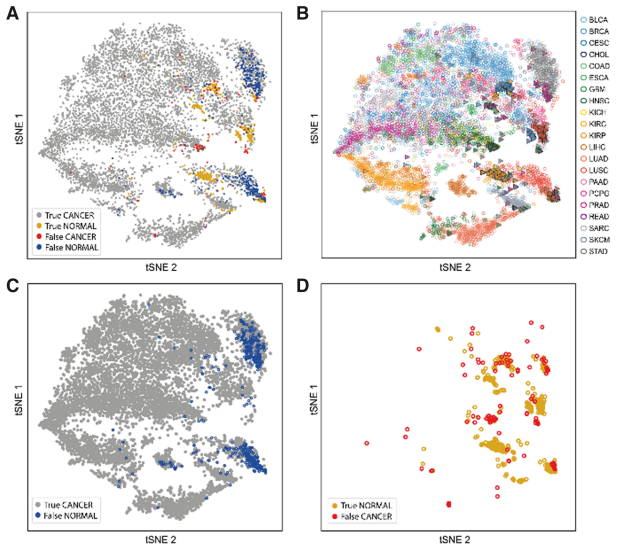

Fig. 4. tSNE plots of TCGA data. (**A**) The result of the binary classification has been mapped in tSNE plot. (**B**) For the 21 cancer types, samples are colored according to class and notably normal samples are in triangles. From the binary classification, (**C**) cancer samples are plotted in gray and blue for correctly and incorrectly classified samples, respectively and (**D**) normal samples are plotted in yellow and red for correctly and incorrectly classified samples, respectively

rates between them. Of the ESCA samples, 26% were misclassified to STAD, 13% of STAD samples were misclassified to ESCA. Similarly, colorectal cancers, 14% of COAD samples were misclassified to READ, and 73% of READ samples were misclassified to COAD. Clearly, the tissue of origin affects the accuracy of prediction. It would be interesting if we can develop a methodology to tease out the source of the mis-classification is the normal tissue 'contamination' or the remaining 'normal' signals in the cancers. Moreover, tSNE was performed to compare sample distributions between misclassified cancer types (Fig. 4B). Squamous cell cancers, BLCA, CESC, HNSC and LUSC (Supplementary Fig. S3A, C, H and M); kidney cancers, KICH, KIRC and KIRP (Supplementary Fig. S3I, J and K) and colorectal cancers, COAD and READ (Supplementary Fig. S3E and R), are tumor-origin cancer types with similar distribution patterns. However, LUAD and LUSC, lung cancer subtypes, exhibited different distribution patterns across most samples (Supplementary Fig. S3M and N).

3.3 Application to single-cell data

Due to intratumor heterogeneity (Poirion *et al.*, 2016; Shalek *et al.*, 2014), our classifiers trained with bulk RNA-seq data barely classified cancer types of single-cell expression data (Fig. 5A and B). As shown in Fig. 6A, tSNE analysis revealed that 198 normal cells extracted from breast tissue are located near 317 BRCA cells and are distant from 3256 normal cells of melanoma (skin) tissue. In addition, 1257 SKCM cells are not clustered with normal skin cells, but rather sporadically distributed near breast cells, making it difficult to predict the cancer type of original scRNA-seq. To deal with this, single-cell data were transformed to bulked data that use averages of expression values from selected cells by kNN for smoothing data. We generated several datasets to determine how many single cells must be aggregated to achieve optimized classifier performance.

Figure 5A is the binary classification result of 360 datasets generated from single-cell samples. The graph indicates that as the number single cells used for the average increases, the accuracy of the classifications increase as well. For instance, when the original expression data of single cells (one single-cell data) were classified, the binary classifier merely attained MCC 0.5 (Fig. 5A). The score does not increase from the time the dataset is averaged over more than 50 single-cell data. There is no significant change in accuracy between the numbers of initial cells. Based on this result, we selected the dataset consisting of 300 cells in which the data of each cell was smoothed using the 150 nearest cells (bulked SC data), resulting in stability and increased performance for the single-cell classification test of our binary classifier.

The binary classifier predicted the cancer type of test data with MCC 0.85 and ACC 0.93. All smoothed cancer cells were perfectly classified as CANCER, and only 79% smoothed normal cells were classified as NORMAL (Fig. 5C). The results of the tSNE analysis (Fig. 6B) showed that 21% smoothed normal cells classified as CANCER were distributed between cancer cells (gray spots) and normal cells (yellow spots). Separately clustered original BRCA and SKCM cells (Fig. 6A) were closely transferred by kNN smoothing (Fig. 6C) and classified as cancer with 100% accuracy.

Multiple classifications of 360 datasets smoothed in the same manner as binary classification were executed to retrieve test data suitable for the pan-cancer classifier. The pan-cancer classifier categorized original single cells as performing more poorly (MCC 0.4) than binary classification, but the smoothed dataset of 300 single cells was classified the best (Fig. 5B). Therefore, we used smoothed data with 300 nearest cells based on 100 initial cells of each cancer type as a test dataset of pan-cancer classifier. The pan-cancer classifier distinguished 95 smoothed BRCA cells, 87 smoothed normal cells and 89 smoothed SKCM cells into each class (Fig. 5D), which resulted in MCC 0.86 and ACC 0.9. As shown in Fig. 6D, smoothed single cells were redistributed to each cancer type by kNN smoothing.

3.4 Gene set enrichment analysis

To investigate the correlated functional roles of the selected gene set, we analyzed enriched pathways and their differential expression

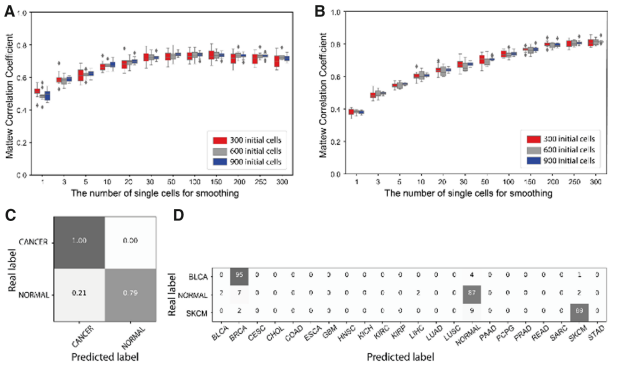

Fig. 5. The classification results of bulked scRNA-seq data. The graphs show the classification results using the binary classifier (**A**) and pan-cancer classifier, (**B**) which depend on the different sizes of the nearest single cells for smoothing. (**C**) The binary classifier distinguished cancer cells from normal cells with MCC 0.85 and ACC 0.93 in 300 initial cells smoothed using the nearest 150 cells. (**D**) The pan-cancer classifier predicted cancer types with MCC 0.86 and ACC 0.9 in 300 smoothed cells using the nearest 300 cells

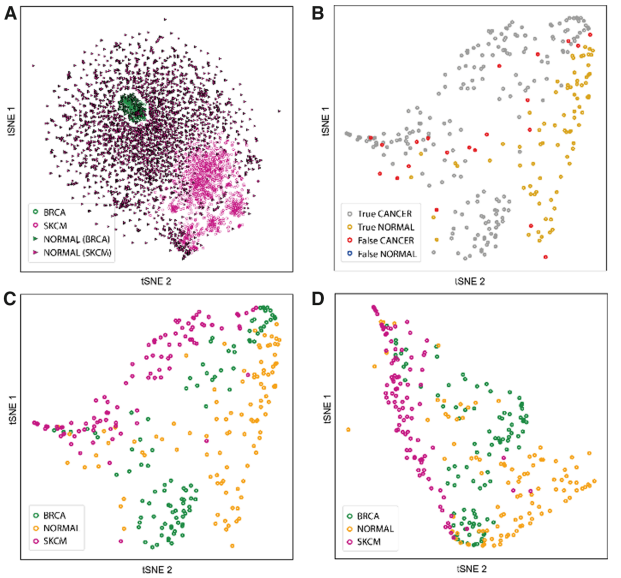

Fig. 6. tSNE plots of scRNA-seq data. (**A**) In the original scRNA-seq data, BRCA and SKCM samples are shown in green and magenta, respectively, and normal samples are shown as triangles. (**B**) Result of the binary classification using 300 initial single cells smoothed with the closest 150 cells (**C**), in which gray dots are cancer samples correctly classified as CANCER, yellow dots are normal samples classified as NORMAL, red dots are normal samples incorrectly classified as CANCER, and blue dots are cancer samples misclassified as NORMAL. (**D**) Result of the pan-cancer classification using 300 initial cells smoothed by the 300 nearest cells, distributed in the tSNE plot. The samples from BRCA, NORMAL and SKCM are colored in green, yellow and magenta, respectively (C, D)

patterns across cancer types. As a result, we found pan-cancer 300 genes are enriched in 10 pathways: RNA binding (GO: 0003723), DNA replication (hsa03030), cell cycle (has04110), role of Ran in mitotic spindle regulation (h_ranMSpathway), mRNA splicing via spliceosome (GO: 0000398), nuclear chromosome, telomeric region (GO: 0000784), RNA splicing (GO: 0008380), positive regulation of ubiquitin-protein ligase activity involved in regulation of mitotic cell cycle transition (GO: 0051437), DNA replication (GO: 0006260) and G1/S transition of mitotic cell cycle (GO: 0000082) (Fig. 7A).

Particularly, five pathways are involved in the cell cycle. It is known that the cell cycle is very closely related with tumorigenesis; thus, cancer is fundamentally caused by uncontrolled cell growth

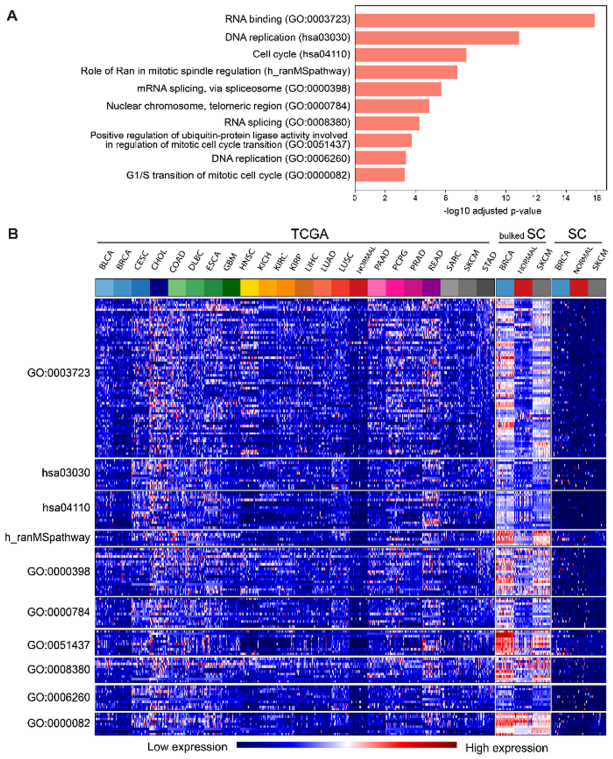

Fig. 7. GSEA results. (**A**) Ten cell cycle-related pathways that 300 genes enriched were searched. (**B**) Heatmaps of TCGA and single-cell data are available to compare expression levels of genes in enriched pathways across cancer types

that deregulates cell proliferation and division (Hartwell and Kastan, 1994; Horning *et al.*, 2018; Martinez *et al.*, 2015). In Fig. 7B, in accelerating cancer cell division, genes in these pathways involved in the cell cycle are highly expressed in the samples of the 21 cancer types more frequently than normal samples. Genes of enriched pathways have different expression levels across cancer types not only in the TCGA data but also in the bulked SC data, as indicated by the summarized expression values of 300 single cells. However, different gene expression patterns are not detected among cancer types in original scRNA-seq. Figure 7B represents that increased gene expression in pathways related to the cell cycle result in increased proliferation activity of tumor cells.

4 Discussion and conclusion

We identified pan-cancer gene sets that are efficient in both binary classification of cancer samples from normal tissue samples and multiple classification of 21 different cancer types. These 300 pan-cancer genes are distinguished by their differential expression in cancer samples in comparison to normal tissues, even though they have different levels of expression in different cancer types. For instance, *TP53* is a typical tumor-related gene that is highly expressed in cancerous tissues but is expressed at different levels across cancer types. The pathways enriched for these pan-cancer genes are functionally correlated with all cancers; these genes are expressed at higher levels in various types of cancers than in normal tissues.

For binary classification, we distinguished 95% cancer samples from normal samples obtained from TCGA and the prediction ACC was 94% by a 10-cross validation of NN. The binary classifier for BRCA has 91.74% ACC with NN based on 500 genes (Danaee *et al.*, 2017). Our binary classifier outperformed this BRCA classifier for all cancer types, even though our model used a smaller number of significant genes. Additionally, the performance of our pan-cancer classifier is almost constant for 22 phenotypes, in contrast with GA/kNN, which incorrectly predicts 18 cancer types several times in 1000 iterations (Li *et al.*, 2017). Furthermore, Lyu and Haque (2018) built CNN composed of three convolutional layers for filtering 10 381 genes and three fully connected layers for classification. Our pan-cancer classifier with NN, which only includes two hidden layers, resulted in an ACC of 94% using a relatively small gene set.

The typical ACC (0.88) of pan-cancer classification is less than the ACC (0.94) of binary classification derived, from correctly a predicted 'CANCER' class, with a sample size over 13 times that of 'NORMAL' class. Due to the imbalance in sample sizes across the cancer types, MCC needs to be applied to adjust accuracy and accurately identify the performance of our model. Therefore, the pan-cancer classifier shows a higher MCC of 0.88 than the MCC of 0.68 of the binary classifier.

We found that cancers of the same organ system have similar patterns, based on the tSNE plots (Supplementary Fig. S3) and are easily misclassified as one another, rather than as other types of cancers. ESCA, STAD, COAD and READ, which belong to gastrointestinal adenocarcinomas, were incorrectly predicted as each other. Other misclassifications occurred across BLCA, CESC, HNSC and LUSC due to similar expression patterns of squamous cell-specific genes. Additionally, we found some tissue-specific genes that might cause mispredictions between groups of lung cancer (LUSC and LUAD) and kidney cancer (KICH, KIRC and KIRP) (Hoadley *et al.*, 2014, 2018).

To examine whether our classifiers can perform well even with scRNA-seq data, we generated bulked SC data using kNN smoothing. With bulked SC data, the binary classifier exhibited a performance with MCC of 0.85 which is better than the performance with the MCC of 0.68 with the TCGA data, and the pan-cancer classifier showed an MCC of 0.86 which is similar to MCC of 0.88 with TCGA and more than twice as much as MCC 0.4 with original SC data. This result indicates that individual cells of the original SC data do not have the common characteristics of each type of cancer, but the smoothing method transforms the cells of the same cancer type into cells with similar characteristics. Thus, we removed the restriction on analysis of scRNA-seq data by transforming scRNA-seq data into a bulk form using the kNN smoothing method.

Our cancer classifiers, which are able to classify normal single cells and cancer cells, as well as predict different cancer types, can be helpful in cancer diagnosis through liquid biopsy. The development of a diagnostic kit using our cancer classifiers can simplify monitoring the prognosis and the prediction of primary and metastatic cancer. Accumulation of the discovered tumorigenic events may lead to further analysis of their biological relevance across the multiple cancer types and further study of downstream events. The cancer-specific features can help us to understand rare cancer types, since one can presume that cancers with a similar set of tumorigenic features might share similar prognosis.

Funding

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (NRF-2019R1A2C2084181) and a grant from the National R&D Program for Cancer Control, Ministry of Health & Welfare (HA17C0032).

Conflict of Interest: none declared.

References

Agarwal,P. and Owzar,K. (2014) Next generation distributed computing for cancer research. *Cancer Inform.*, **13(suppl)**, 97–109.

Altman,N.S. (1992) An introduction to kernel and nearest-neighbor nonparametric regression. *Am. Stat.*, **46**, 175–185.

Angermueller,C. *et al.* (2016) Deep learning for computational biology. *Mol. Syst. Biol.*, **12**, 878.

Aran,D. *et al.* (2015) Systematic pan-cancer analysis of tumour purity. *Nat. Commun.*, **6**, 8971.

Ashburner,M. *et al.* (2000) Gene ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25.

Barandiaran,I.J. (1998) The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **20**, pp. 832–844.
Cancer Genome Atlas Network (2012a) Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, **487**, 330.
Cancer Genome Atlas Network (2012b) Comprehensive molecular portraits of human breast tumours. *Nature*, **490**, 61.
Cancer Genome Atlas Research Network (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, **455**, 1061.
Cancer Genome Atlas Research Network (2011) Integrated genomic analyses of ovarian carcinoma. *Nature*, **474**, 609–615.
Cancer Genome Atlas Research Network (2013) The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.*, **45**, 1113–1120.
Chen,H. *et al.* (2018) A pan-cancer analysis of enhancer expression in nearly 9000 patient samples. *Cell*, **173**, 386–399.e2.
Chen,L. and Zheng,S. (2018) BCseq: accurate single cell RNA-seq quantification with bias correction. *Nucleic Acids Res.*, **46**, e82.
Chung,W. *et al.* (2017) Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer. *Nat. Commun.*, **8**, 15081.
Cortes,C. and Vapnik,V. (1995) Support-vector networks. *Mach. Learn.*, **20**, 273–297.
Danaee,P. *et al.* (2017) A deep learning approach for cancer detection and relevant gene identification. In: *Pacific Symposium on Biocomputing*. World Scientific, Singapore, pp. 219–229.
Eraslan,G. *et al.* (2019) Single-cell RNA-seq denoising using a deep count autoencoder. *Nat. Commun.*, **10**, 390.
Fritz,M.H.-Y. *et al.* (2011) Efficient storage of high throughput DNA sequencing data using reference-based compression. *Genome Res.*, **21**, 734–740.
Hartwell,L.H. and Kastan,M.B. (1994) Cell cycle control and cancer. *Science*, **266**, 1821–1828.
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.
Hoadley,K.A. *et al.* (2018) Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. *Cell*, **173**, 291–304e296.
Hopfield,J.J. (1982) Neural networks and physical systems with emergent collective computational abilities. *Proc. Natl. Acad. Sci. USA*, **79**, 2554–2558.
Horning,A.M. *et al.* (2018) Single-cell RNA-seq reveals a subpopulation of prostate cancer cells with enhanced cell-cycle-related transcription and attenuated androgen response. *Cancer Res.*, **78**, 853–864.
Kanehisa,M. and Goto,S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **28**, 27–30.
Kim,B.J. and Kim,S.H. (2018) Prediction of inherited genomic susceptibility to 20 common cancer types by a supervised machine-learning method. *Proc. Natl. Acad. Sci. USA*, **115**, 1322–1327.
Kourou,K. *et al.* (2015) Machine learning applications in cancer prognosis and prediction. *Comput. Struct. Biotechnol. J.*, **13**, 8–17.
Li,B. and Dewey,C.N. (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, **12**, 323.
Li,Y. *et al.* (2017) A comprehensive genomic pan-cancer classification using The Cancer Genome Atlas gene expression data. *BMC Genomics*, **18**, 508.
Lyu,B. and Haque,A. (2008) Deep learning based tumor type classification using gene expression data. In: *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*. ACM, Washington, DC, pp. 89–96.
Martinez,E. *et al.* (2015) Comparison of gene expression patterns across 12 tumor types identifies a cancer supercluster characterized by TP53 mutations and cell cycle defects. *Oncogene*, **34**, 2732–2740.
Matthews,B.W. (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta*, **405**, 442–451.
Meyerson,M. *et al.* (2010) Advances in understanding cancer genomes through second-generation sequencing. *Nat. Rev. Genet.* **11**, 685.
Nishimura,D. (2001) BioCarta. *Biotech Softw. Internet Rep.*, **2**, 117–120.
Pedregosa,F. *et al.* (2011) scikit-learn: machine learning in Python. *Mach. Learn.* **12**, 2825–2830.
Peng,L. *et al.* (2015) Large-scale RNA-Seq transcriptome analysis of 4043 cancers and 548 normal tissue controls across 12 TCGA cancer types. *Sci. Rep.*, **5**, 13413.
Poirion,O.B. *et al.* (2016) Single-cell transcriptomics bioinformatics and computational challenges. *Front. Genet.*, **7**, 163.
Riesz,F. (1910) Untersuchungen Über Systeme Integrierbarer Funktionen. *Math. Ann.*, **69**, 449–497.
Ronen,J. and Akalin,A. (2018) netSmooth: network-smoothing based imputation for single cell RNA-seq. *F1000Res.* **7**, 8.
Shalek,A.K. *et al.* (2014) Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. *Nature*, **510**, 363–369.
Sheshadri,N. *et al.* (2014) SCCA1/SERPINB3 promotes oncogenesis and epithelial-mesenchymal transition via the unfolded protein response and IL6 signaling. *Cancer Res.*, **74**, 6318–6329.
Subramanian,A. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA*, **102**, 15545–15550.
Tirosh,I. *et al.* (2016) Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science*, **352**, 189–196.
Tomczak,K. *et al.* (2015) The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemp. Oncol.*, **19**, A68.
Wagner,F. *et al.* (2018) K-nearest neighbor smoothing for high-throughput single-cell RNA-seq data. *BioRxiv*, **217737**.
Wan,Q. *et al.* (2015) BioXpress: an integrated RNA-seq-derived gene expression database for pan-cancer analysis. *Database*, **2015**, bav019.
Wan,Y.W. *et al.* (2016) TCGA2STAT: simple TCGA data access for integrated statistical analysis in R. *Bioinformatics*, **32**, 952–954.
Wang,E. *et al.* (2015) Predictive genomics: a cancer hallmark network framework for predicting tumor clinical phenotypes using genome sequencing data. *Semin. Cancer Biol.*, **30**, 4–12.
Wang,Q. *et al.* (2013) Detecting somatic point mutations in cancer genome sequencing data: a comparison of mutation callers. *Genome Med.*, **5**, 91.
Zack,T.I. *et al.* (2013) Pan-cancer patterns of somatic copy number alteration. *Nat. Genet.*, **45**, 1134–1140.
Zararsiz,G. *et al.* (2017) A comprehensive simulation study on classification of RNA-seq data. *PLoS One*, **12**, e0182507.
Zhang,J. and Zhang,S. (2017) Discovery of cancer common and specific driver gene sets. *Nucleic Acids Res.*, **45**, e86.
Zill,D. *et al.* (2011) *Advanced Engineering Mathematics*. Jones & Bartlett Learning, Burlington, MA, USA.