



Article

Relevant and Non-redundant Feature Selection for Cancer Classification and Subtype Detection

Pratip Rana ^{1,†} , Phuc Thai ^{1,}, Thang Dinh ^{1,}, and Preetam Ghosh ^{1,*} 

¹ Department of Computer Science, Virginia Commonwealth University, Richmond, VA 23284, USA; ranap@vcu.edu (P.R.); thaipd@vcu.edu (P. T.); tndinh@vcu.edu (T. D.); pghosh@vcu.edu (P.G.)

* pghosh@vcu.edu

Abstract: Biologists seek to identify a small number of significant features that are important, non-redundant, and relevant from diverse omics data. For example, statistical methods like LIMMA and DEseq distinguish differentially expressed genes between a case and control group from the transcript profile. Researchers also apply various column subset selection algorithms on genomics datasets for a similar purpose. Unfortunately, genes selected by such statistical or machine learning methods are often highly co-regulated, making their performance inconsistent. Here, we introduce a novel feature selection algorithm that selects highly disease-related and non-redundant features from a diverse set of omics datasets. We successfully applied this algorithm to three different biological problems: a) disease to normal sample classification, b) multiclass classification of different disease samples, and c) disease subtypes detection. Considering classification ROC-AUC, False-positive, and False-negative rates, our algorithm outperformed other gene selection and differential expression (DE) methods for all six types of cancer datasets from TCGA considered here for binary and multiclass classification problems. Moreover, genes picked by our algorithm improved the disease subtyping accuracy for four different cancer types over the state-of-the-art methods. Hence, we posit that our proposed feature reduction method can support the community to solve various problems, including the selection of disease-specific biomarkers, precision medicine design, and disease sub-type detection.

Keywords: feature subset selection; disease classification; subtype detection

1. Introduction

Omics data usually comprises thousands of features; however, most of these features are redundant, irrelevant, or noisy. Experimental noise, multiple intrinsic interconnections between the biological units, and co-regulation between the features are possible reasons for redundancy. For example, typical RNA-seq measurements catalog the expression of thousands of transcripts; but most of them are redundant (i.e., highly correlated) or noisy. Moreover, due to the experimental costs, the number of samples available is lower than the number of features, making the traditional machine learning and statistical algorithms easily overfit the biological data. Another problem is the lack of control/normal samples; this is mainly because there are fewer chances to collect data from healthy patients. Therefore, selecting a small number of relevant and non-redundant features among the complete set of features is a significant research problem.

Yu et al. classified the genes in a disease into the following four categories: a) irrelevant or noisy genes, b) weakly relevant and redundant genes, c) weakly relevant and non-redundant genes, and d) strongly relevant genes [1]. Generally, researchers wish to select a small number of genes that are either strongly relevant or weakly relevant and non-redundant. The selection of this subset of relevant genes is essential for several biological problems, such as identifying causal disease-related genes, early detection of diseases, designing precision medicine, and disease sub-type detection [2]. In machine learning, a similar problem is termed as the feature selection problem, where the goal is to select the most informative and small subset of features from a larger number of features. Feature selection becomes critical, where only a few samples are available compared to the number of features in the dataset; it reduces noise, improves the training time of the machine learning models, and avoids over-fitting. Ang, et al. [2] classified the feature selection

techniques into the following five categories: a) Filter b) Wrapper c) Embedded d) Hybrid, and e) Ensemble. Several of these feature selection techniques have been used in the past for different purposes. For example, *twoPhase* [3], *iterFS* [4], *WeiBi* [5] are based on information gain, LASSO, or Fisher score. Feature selection algorithms in the biological domain is discussed briefly in [2] and [6].

In this paper, we propose a new feature selection method and demonstrate its performance in (i) classifying disease samples from normal samples, (ii) classifying the different types of disease samples, and also (iii) disease subtype detection. First, we evaluated the performance of our method in disease classification considering six different large cancer gene expression datasets from The Cancer Genome Atlas (TCGA) [7]. Then, we benchmarked our results using three different feature/geneset selection methods and a few differential expression (DE) methods on the same dataset. From the results, we found that our proposed method outperformed all the popular geneset selection methods and can identify essential genes and functional pathways in diseases in general and cancer in particular.

2. Materials and Methods

2.1. Method overview

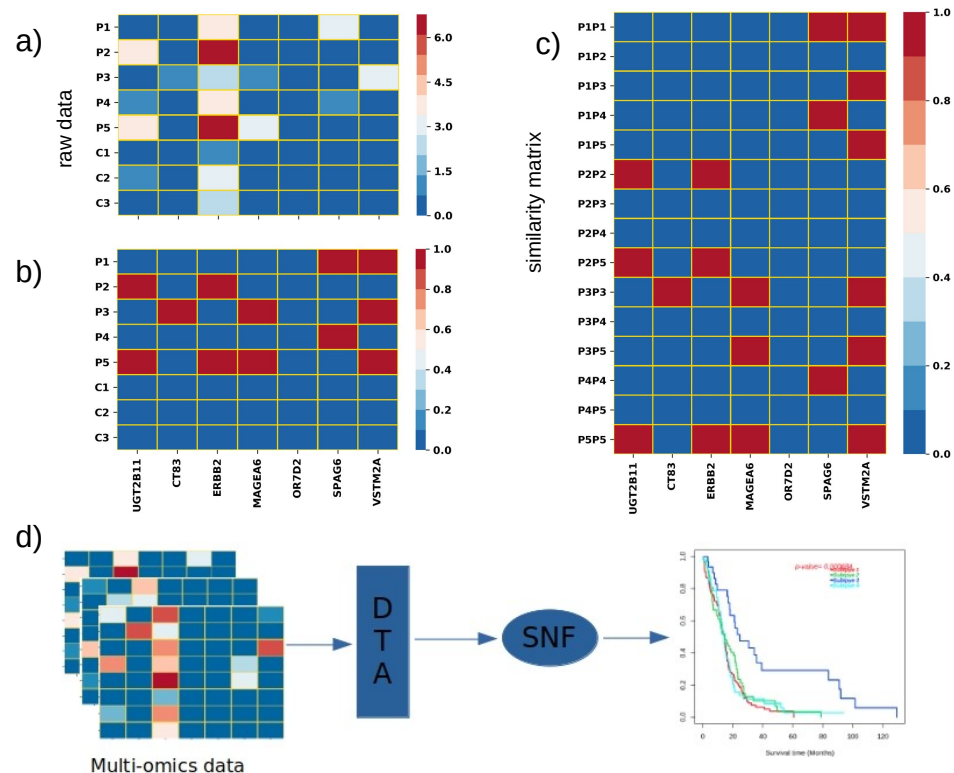


Figure 1. Overview of the method. a) Gene expression profiles of the patients; here, P_i denotes disease sample, and C_i denotes control sample. b) Construction of PEEP matrix from (a). c) Creating patient-to-patient (denoted by P_iP_j) similarity matrix and selecting the column using an approximate k-cover algorithm. d) An overview of the subtype detection pipeline.

Our proposed feature selection method consists of two main steps. First, it creates a binary patient-specific perturbation profile (PEEP) from the genomics dataset using data normalization and imposes a cut-off. Second, it selects non-redundant features that maximize the similarity between each patient pair by an approximate k-cover algorithm. The k-cover problem in a graph $G = (V, E)$ is an NP-complete problem, which seeks a set of size k nodes that cover the maximum number of edges. A standard greedy algorithm can approximate this problem with $(1-1/e)$ approximation. However, due to the large space

consumption and time complexity which can be a bottleneck for a large graph, several algorithms have been proposed for approximate k -cover. We use the Dynamic threshold algorithm (DTA) [8] to find the subset C of k genes that guarantee a $1 - \frac{1}{e} - \epsilon$ approximation solution to the k -cover problem. Fig 1 shows a brief overview of the DTA algorithm.

2.2. Creating personalized perturbation profile (PEEP)

First, we convert the raw gene expression dataset to a log scale. Consider a set of n patients $\mathcal{X} = \{X_1, X_2, \dots, X_n\}$, where a patient, X_i , is represented by m features. First, we compute the mean and standard deviation of each feature from the control sample. Next, we perform a column-wise normalization using the following formula.

$$z_{ij} = \frac{x_{ij} - \mu_j^N}{\sigma_j^N} \quad (1)$$

Here, x_{ij} refers to the log-transformed raw expression of patient i and feature j , z_{ij} is the Z-score transformed value of expression of patient i and feature j . μ_j^N and σ_j^N are the mean and standard deviation of feature j in the control samples.

In the next step, we convert the Z matrix to the binary '0' / '1' PEEP matrix $P = p_{ij}$ by imposing a threshold of z_t . For example, if a feature in a subject has an expression that is less or greater than z_t , we consider that as '1', otherwise '0'.

$$p_{ij} = \begin{cases} 0 & \text{if } -z_t < z_{ij} < z_t \\ 1 & \text{if } -z_t > z_{ij} \text{ or } z_t < z_{ij} \end{cases} \quad (2)$$

Thus, the features having '1' can be viewed as over-expressed/under-expressed in one subject. Similar normalization is performed as before, and we use the same z_t cutoff of 2.5 for our analysis [9]. We also performed the same analysis using two other cutoffs (2, 3) and observe similar results (not shown here).

2.3. Feature-selection problem formulation

In our method, we only select a subset of features to run the classification. Here, we select k features so that we can preserve the maximum information of the patients. By performing gene-selection, we reduce the overfitting of the classification, thereby improving accuracy.

Similar to the feature-selection problem [10], we consider the dataset as the set of pairs of patients. We define that two patients are similar if there exists a feature that is up-regulated/ down-regulated in both patients. We build a similarity matrix $S = s_{ij}$ as

$$s_{ij} = \begin{cases} 1 & \text{if } \exists h : p_{ih} = p_{jh} = 1 \\ 0 & \text{if otherwise} \end{cases}$$

The goal is to select a subset C of k features so that we can preserve the similarity matrix. More concretely, we want to maximize the similarity between patients on the subset C as follows.

$$C = \arg \max_{C \subset G, |C|=k} \sum s_{ij}^C$$

where

$$s_{ij}^C = \begin{cases} 1 & \text{if } \exists h \in C : p_{ih} = p_{jh} = 1 \\ 0 & \text{if otherwise} \end{cases}$$

2.4. DTA algorithm

The feature-selection problem can be considered as a k -cover problem. Let's say a feature h covers a pair of patients (X_i, X_j) if $p_{ih} = p_{jh} = 1$; we use the DTA algorithm [8] to find the subset C of k genes that guarantee a $1 - \frac{1}{e} - \epsilon$ approximation solution. However,

note that our proposed framework can work seamlessly with other existing algorithms for the k -cover problem.

In the DTA algorithm, we sample random hyperedges (each hyperedge consists of a feature and a pair of patients covered by that feature) to select a subset of k features.

In particular, we iterate k times to select the subset of k features. At a specific iteration, we select a feature as follows:

- Add new hyperedges by repeating the following steps.
 - randomly choose two patients X_i, X_j ;
 - for each feature $h \in G$, if $p_{ih} = p_{jh} = 1$, add $(h, (i, j))$ to the hyperedge.
- Select a feature h^* that covers the most pairs of patients. For each pair of patients (i, j) that is covered by h^* , remove all hyperedges that consist of (i, j) .

The DTA algorithm has a run-time of $O(\epsilon^{-2}km \log m)$ (where m is the total number of genes, and k is the number of selected genes) and can quickly scale to thousands of patients' data.

2.5. Classification workflow

We perform five-fold cross-validation (5-CV) with an 80-20 split on the raw dataset. For each split of the training dataset, we first estimate the mean (μ) and standard deviation (σ) from the healthy samples. Next, we select a set of genes from the training set using our method. Finally, we train the model with a linear-SVM on the selected genes and classify the samples as healthy or disease samples. Unfortunately, the TCGA dataset is imbalanced as they have fewer healthy samples than disease samples. Hence, we use the ROC-AUC score, false-positive, and false-negative rates to evaluate the classification performance.

We also benchmark our results with two other feature selection methods, *twoPhase* [3], *iterFS* [4], and one geneset selection method *Barabasi* [9], and one DE analysis method, *LIMMA*. For *twoPhase*, *iterFS*, and *Barabasi* we follow the same preprocessing as our method to select the subset of features. For *LIMMA* we use raw data directly as the input as *LIMMA* takes continuous inputs.

For multi-class classification, we merge the PEEP transformed patient subjects data of six cancer types. Also, we perform a 5-CV with an 80-20 split. For the performance evaluation, we first measure the ROC-AUC score of one-to-all classification of one cancer type to all other cancer types. Next, we average the ROC-AUC score of one-to-all classifications of the cancer types. Finally, for the multi-class classification, we benchmark only *twoPhase* and *iterFS*, as *Barabasi* and *LIMMA* are not designed for the multi-class problem.

2.6. Disease subtyping workflow

A heterogeneous disease such as cancer is activated through several pathways and shows a high level of molecular heterogeneity. Therefore, causal oncogenes express only in a subset of patients for a cancer type. For example, in the TCGA-LUSC (Lung squamous cell carcinoma) dataset, the popular linear DE method *LIMMA* identified around 13852 genes as differentially expressed ($p \leq 0.05$) [11]. Interestingly, most of these differentially expressed genes are perturbed in a few patients only. For a heterogeneous disease like cancer, different perturbations in multiple oncogenes lead to a common phenotypic outcome. The outcome of patients with the same cancer type also differs significantly based on the phenotypic outcome. Thus, disease subtyping is a crucial method to predict disease variability, identify associated molecular pathways, and design a personalized treatment plan for a heterogeneous disease.

The disease subtype prediction from omics data mainly consists of the following two steps. First, it computes patient-to-patient similarity (e.g., Euclidean distance, Pearson correlation) from omics data. Next, it performs an unsupervised clustering (e.g., k-means, consensus) on that similarity matrix to group similar patients for identifying the sub-types. Subtype detection is complex and often requires multi-omics data integration of the same patients to achieve better clustering/subtyping. Some of the popular multi-omics data

integration methods are iCluster [12], Similarity network fusion (SNF) [13], PINS [14], CIMLR [15] and autoencoders [16]. However, these methods often depend on selecting important features from diverse high-dimensional omics datasets (e.g., gene expression, methylation, copy number, miRNA expression). Our proposed DTA method can identify such important features from these multiple datasets and improve the existing sub-type detection methods.

We analyzed the performance of DTA as a feature selection method in the standard subtype detection pipeline using the data integration method SNF. SNF is a network fusion method that generates a patient-to-patient similarity network first from all datatypes individually using a non-linear kernel function. Then, it fuses all individual networks into a single comprehensive network using an iterative cross-network diffusion algorithm. We use gene expression, miRNAs expression, and DNA methylation profiles from TCGA [7] of the same patients to perform subtype detection. Next, we use our feature selection method on the multiview data to individually select essential features from each datatype. Based on this, we calculate the Euclidean distance (i.e., similarity) for every patient pair. Then, we integrate these three distance matrices into a single comprehensive dataset using SNF. Later, we perform a spectral clustering on this euclidean distance to group similar patients. Finally, we validate disease subtypes using the Kaplan–Meier curve of the survival rate of the patients.

We use the R tool CancerSubtype for subtyping a cancer type [17]. For feature selection, we use DTA and the other two feature selection methods based on principal component analysis (PCA) and variance (VAR) for comparison. For further analysis, we use linear model LIMMA to identify DE genes [11], and the ClusterProfiler package for the functional analysis of the DE geneset [18].

3. Results

3.1. TCGA dataset

The Cancer Genome Atlas (TCGA) program integrates various molecular profile information of more than 33000 samples across 68 different cancer types [7]. We performed our classification analysis for RNA-seq data for six different types of cancer from TCGA: BRCA (Breast Invasive Carcinoma), LUAD (Lung adenocarcinoma), LUSC (Lung squamous cell carcinoma), PRAD (Prostate adenocarcinoma), COAD (Colon Adenocarcinoma), and KICH (Kidney Chromophobe). We discarded other cancer types in TCGA from our analysis, as they had fewer control samples. We used the TCGA dataset of four cancer types BRCA, COAD, LUSC, and GBM (glioblastoma multiforme), for subtype detection. Three different data types, gene expression, DNA methylation, and miRNA expression of the same patient set, were used here. To perform survival analysis, we also downloaded clinical data of the patients from the TCGA. We used the TCGAbiolink package to load data from TCGA [19].

3.2. Classifying Disease Samples with Normal samples

First, we evaluated the performance of DTA for cancer with normal sample classification using gene expression data. We have selected a comparatively small k ($1 \leq k \leq 20$) number of genes using DTA for each cancer type. DTA performed remarkably well in classifying the diseases of the normal samples compared to the baseline (classification without any feature reduction). Additionally, its performance was quite consistent over different disease types in achieving a high ROC-AUC score with a low FN rate, as shown in Fig 2. We also benchmarked four other methods *twoPhase* [3], *iterFS* [4], *Barabasi* [9], and *LIMMA*. We found the performance of these algorithms was inconsistent, and the performance degrades, especially when a small number of features are selected (Fig. 2). This shows that these standard feature selection algorithms cannot find the most non-redundant and relevant features that reduce the classification accuracy. The average of the FP, FN, and ROC-AUC scores for the classification of six different cancer types are shown in Fig 2a. DTA achieved almost perfect ROC-AUC using only three genes, while the other methods struggle to achieve an 0.7 ROC-AUC for the same number of selected genes. The FP, FN,

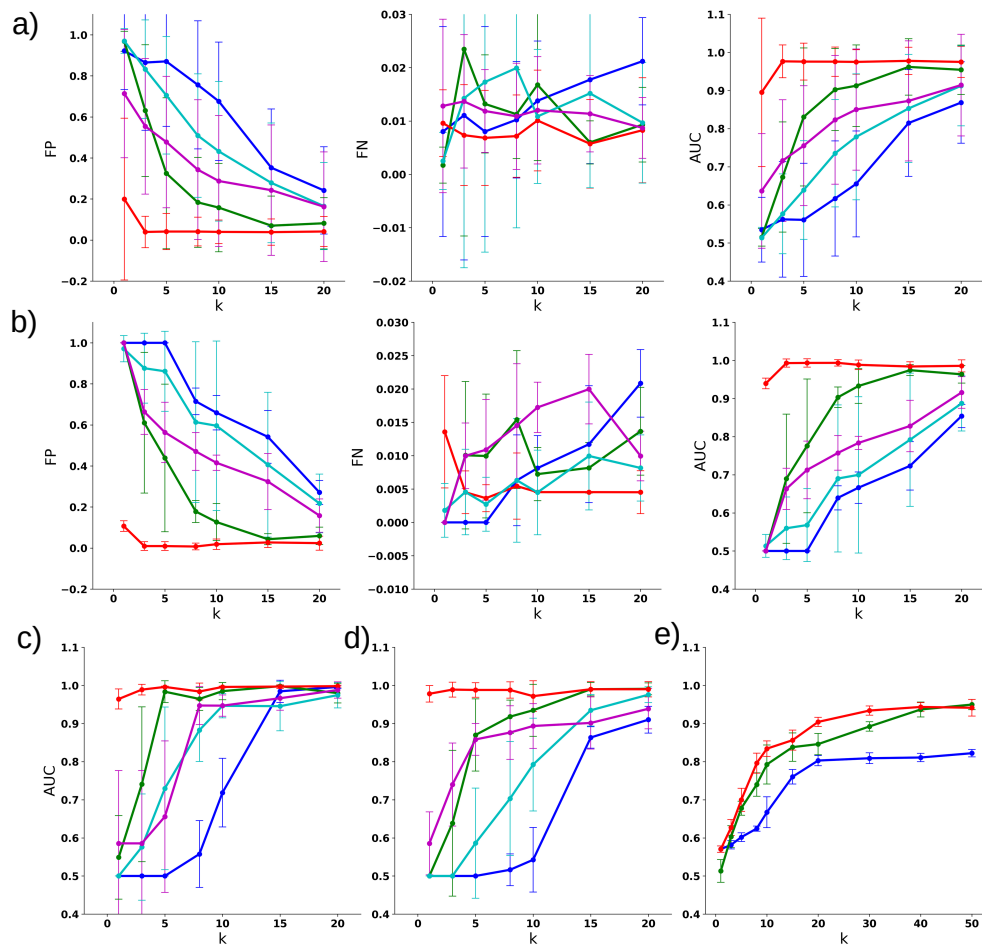


Figure 2. Performance benchmarks of DTA compared to other feature selection methods for disease classification. The blue color denotes PCA, green denotes two-phase, cyan denote Barabasi, magenta denote LIMMA, while red denotes DTA geneset. a) The average False-positive, False-Negative, and ROC-AUC score of the SVM classifier over six different disease datasets from TCGA. b) False-positive (FP), False-Negative (FN), and ROC-AUC score for the TCGA-BRCA dataset. c) ROC-AUC score of TCGA-LUSC dataset. d) ROC-AUC score of TCGA-LUAD dataset. e) ROC-AUC score of multiclass classification for six cancer types from TCGA.

and ROC-AUC score of BRCA classification are shown in Fig 2b, and the ROC-AUC of LUSC, LUAD is shown in Fig 2 c,d.

3.3. Multiclass Disease Classification

Next, we extended our analysis to evaluate the performance of DTA in a multi-class disease classification problem. We trained a multi-class classification model using SVM with a linear kernel and also performed 5-CV. DTA showed a remarkable improvement in ROC-AUC classification compared to the other gene selection methods and the original data. In particular, DTA achieved about 0.9 ROC-AUC for the selected 20 genes. Especially for a small number of selected genes, the performance of the other methods was quite poor. The comparison of ROC-AUC of DTA to the other two feature selection methods *twoPhase* and *iterFS* is shown in Fig 2e. These feature selection algorithms performed poorly for a small number of genes compared to DTA.

3.4. Disease Subtype Detection

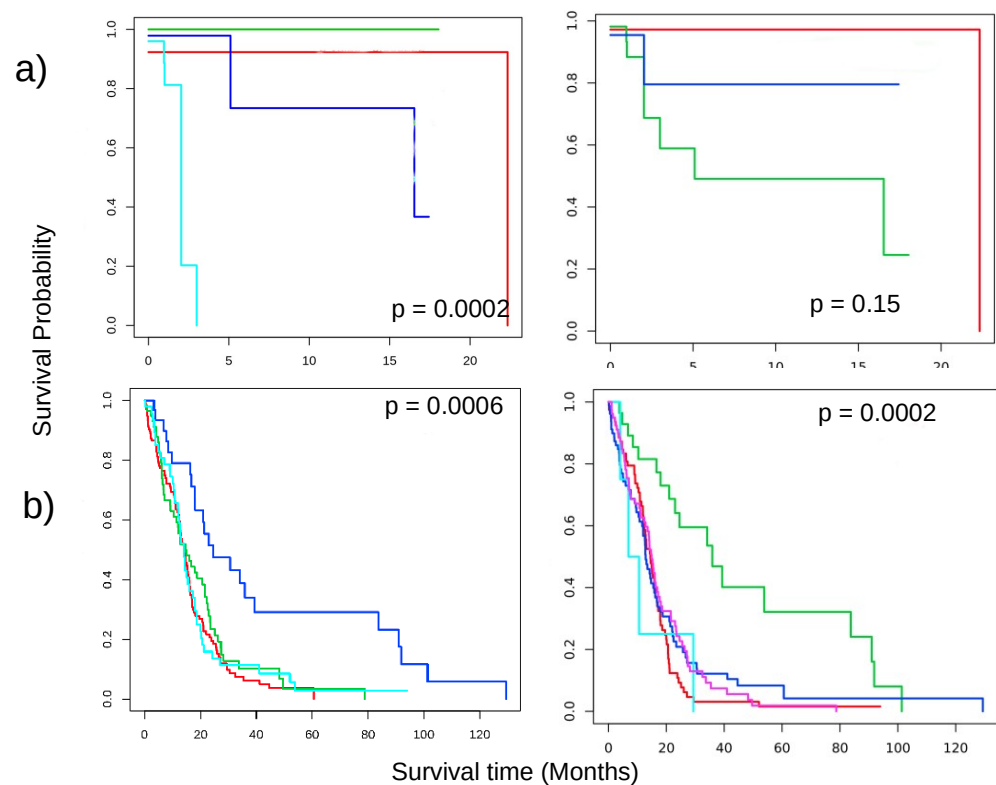


Figure 3. Disease subtype detection using DTA as a feature reduction step. Kaplan–Meier curve of detected subtypes using DTA (left) and maximum variance (right) selected genes of a) COAD cancer, b) GBM cancer. p-value signifies that the patients in different subtypes have different survival profiles.

DTA-selected genes performed quite well to identify different subtypes in a cancer type. For most of the cancer types, the identified clusters have a different survival profile. The DTA method improved the p-value of the survival profile/Kaplan–Meier curve of the clusters over the baseline (without feature selection). Here, a low p-value confirmed that the survival profile of patients in different clusters is significantly different. Furthermore, we benchmarked the result with the other two feature selection techniques PCA and maximum variance (VAR). DTA showed a better and consistent performance in most cancer types in terms of the p-value of the Kaplan–Meier curve and average silhouette score of the clusters. We have also performed functional analysis on the DE gene set of the identified clusters.

Table 1: List of previously validated genes selected by our algorithm

Disease	Genes	Comments
BRCA	EN1	Involved with Basal-like breast tumors [20].
	ERBB2	Involvement is reported in several studies for the past 30 years [21].
	MAGEA6	Associated with poor survival of breast cancer patients [22].
COAD	CLDN18	Encodes gastric type adhesion molecule and known biomarker for gastric cancer [23].
	MUC5AC	Associated with tumorigenesis in colorectal cancer via a serrated neoplasia pathway [24].
	NPC1L1	Key regulator of lipid homeostasis [25].
	SLC14A1	Associated with poor survival of COAD patients [26].
GBM	NDRG1	Associated with the hypoxia-associated molecule and is expressed in GBM cell [27].
LUAD	ABCC2	Important gene candidate for LUAD [28].
	CSAG1	Encodes cancer-germline antigens (CGAs) [29].
	INSL4	Related to LKB1-Inactivated Lung Cancer [30]
	CHRNA9	Nicotinic Receptor and is related to smoking induced tumor formation [31].

This further confirmed that the identified clusters are related to very different biological pathways.

3.5. Key findings on BRCA:

Breast cancer is the most commonly diagnosed cancer among US women after skin cancer. In the TCGA-BRCA dataset, the DE method, LIMMA, identified 13493 genes to be differentially expressed out of 18934 genes (adj p-value < 0.05), most of which are however redundant and co-regulated. To identify the non-redundant genes, we applied the DTA method on this TCGA dataset. DTA achieved almost near perfect ROC-AUC only using 3 genes. While the other methods struggled to achieve 0.7 ROC-AUC using the same number of genes. The genes selected by DTA for $k = 10$ are *ANO3*, *CT83*, *ERBB2*, *MAGEA6*, *OR7D2*, *SPAG6*, *TDRD12*, *TDRD9*, *UGT2B11*, *VSTM2A*. However, surprisingly out of these top 10 genes, *OR7D2*, *TDRD1*, and *TDRD12* genes were not found to be differentially expressed although the involvement of some of these genes was already mentioned in the literature. For example, *EN1* has been reported to be involved with Basal-like breast tumors [20]. *ERBB2* is another oncogene whose involvement is reported in several studies for the past 30 years [21]. Expression of *MAGEA6* was found to be associated with poor survival of breast cancer patients [22]. Thus DTA could find important features (i.e., genes) from the dataset even though their expression profiles did not exhibit significant differential expression.

Subtype detection using DTA identified four subtypes of BRCA on 302 patients. DTA identified subtypes did not show different survival profiles p-value = 0.07. The other feature selection methods also failed to achieve the subtype with significant difference in survival profiles. The subtype 1 has 115 patients, whereas, a total of 116 and 71 patients were identified as subtype 2 and subtype 3 in BRCA. *COL10A1*, *MMP11*, *DMD*, *C10orf90*, and *CNTNAP3* are the top five differentially expressed genes (based on lowest p-value) of BRCA subtype 1. Similarly, the genes *COL10A1*, *MMP11*, *CXCL2*, *CA4*, and *LRRC3B* are the top five differentially expressed genes in subtype 2. Top 5 differentially expressed genes of subtype 3 are *TPX2*, *KIF4A*, *NEK2*, *CDCA5*, and *COL10A1*.

3.6. Key findings on COAD:

Colorectal cancer is the fourth-ranked in terms of cancer-related deaths globally. DTA selected the following genes as important indicators of colorectal cancer: *ABCA12*, *CLDN18*, *MAGEA11*, *MAGEA6*, *MTRNR2L1*, *MTTP*, *NPC1L1*, *PRSS21*, *SLC14A1*, and

SPESP1. *CLDN18* encodes gastric type adhesion molecule and is a known biomarker for gastric cancer [23]. *MUC5AC* expression is associated with tumorigenesis in colorectal cancer via a serrated neoplasia pathway [24]. Another study found *MUC5AC* was expressed in pancreatic ductal and various gastrointestinal tract tumors [32]. Colorectal cancer is strongly associated with lipid metabolism; *NPC1L1* is a sterol transporter that is a key regulator of lipid homeostasis. *NPC1L1* knockout mice was found to have reduced number of tumor than wild type mice [25]. *SLC14A1* was identified in intestinal stem cell signature, which is associated with poor survival of COAD patients [26]. Thus the genes detected by DTA are significant predictors of colorectal cancer.

DTA also identified four different clusters in COAD by using only 10 features from each of the three data types. These clusters showed an entirely different survival profile (p -value < 0.00022). The clusters consisted of 26, 37, 48, and 25 patients, respectively. The top five DE genes of the first cluster are *CDH3*, *CA7*, *PHLPPL*, *GCNT2*, and *ENPP6*. The most related functional pathways considering these genes are cell division, G2/M transition of the mitotic cell cycle, mitotic nuclear division, regulation of protein serine/threonine kinase activity, leukocyte migration. Top differentially expressed genes in the second cluster are *MYOC*, *ABHD7*, *ABCA8*, *SLC30A10*, and *CDH3*, and the corresponding enriched biological processes are cell division, G2/M transition of the mitotic cell cycle, cell cycle G2/M phase transition, ncRNA processing, ncRNA metabolic process. *CA7*, *CDH3*, *CLEC3B*, *CLDN8*, *SLC30A10* are the most differentially expressed genes in cluster 3, corresponding to the enriched biological processes: mitotic cell cycle phase transition, cell cycle G2/M phase transition, cell cycle phase transition, G2/M transition of the mitotic cell cycle, and mitotic nuclear division. In the last cluster, the most expressed genes are *ABCA8*, *SFRP1*, *CDH3*, *GCNT2*, *KIAA1199*, and the corresponding enriched GO terms are regulation of leukocyte migration, leukocyte migration, cell division, mitotic nuclear division, and cell cycle G2/M phase transition.

3.7. Key findings on GBM:

Glioblastoma Multiforme is the most common malignant brain tumor in adults. DTA selected the following genes in this cancer: *ARHGEF2*, *FRS3*, *IRX2*, *VAT1L*, *NDRG1*, *PDPK1*, *PTK6*, *RAB3C*, *RPS4Y1*, and *WDR18*. *NDRG1* is a tumor suppressor gene with the ability of metastasis and migration of cancer cells. A study found *NDRG1* is associated with the hypoxia-associated molecule and is expressed in GBM cell [27]. GBM is also the most studied dataset for subtype detection. Here, we have identified four subtypes for 276 GBM patients. The first cluster consists of 137 patients, and the most expressed genes are *IL12RB2*, *CACNB1*, *ICAM5*, *BTN3A2*, and *INPP5F*. Regulation of vesicle-mediated transport, axodendritic transport, neuron death, mitotic cell cycle phase transition, and cell cycle phase transition are the most enriched processes in this cluster. The second cluster consists of 59 patients, and signature DE genes are *PI4KA*, *IL12RB2*, *CACNB1*, *MICAL2*, and *SLC17A7*. The third cluster has 33 patients with a high survival rate, and the top DE genes are *WSCD2*, *CACNB1*, *KDELR2*, *MICAL2*, and *RYR2*. Mitotic cell cycle phase transition, cell cycle phase transition, axodendritic transport, axonal transport, and regulation of mitotic cell cycle phase transition are the top enriched processes in this cluster. The fourth cluster consists of 47 patients with signature DE genes being *HRH3*, *TSPYL1*, *MAP2K1*, *GOT2*, and *FUT1*.

3.8. Key findings on LUAD:

TCGA-LUSC has a dataset of 59 control samples and 533 LUAD patient samples. Here, we have achieved almost near perfect AUC using only 3 genes using DTA. The top 10 genes selected using DTA are *ABCC2*, *CHRNA9*, *CSAG1*, *EPS8L3*, *INSL4*, *SLC13A2*, *SPESP1*, *TDRD1*, *ZFR2*, and *ZNF560*. One study identified *ABCC2* as an important gene candidate for LUAD using expression and network analysis [28]. *CSAG1* encodes cancer-germline antigens (CGAs) [29]. Aberrant *INSL4* signaling is related to LKB1-Inactivated Lung Cancer [30]. *CHRNA9* is a Nicotinic Receptor and is related to smoking induced tumor formation

[31]. We did not perform subtype detection for LUAD due to the lack of common patients in the three datatypes used in our analyses.

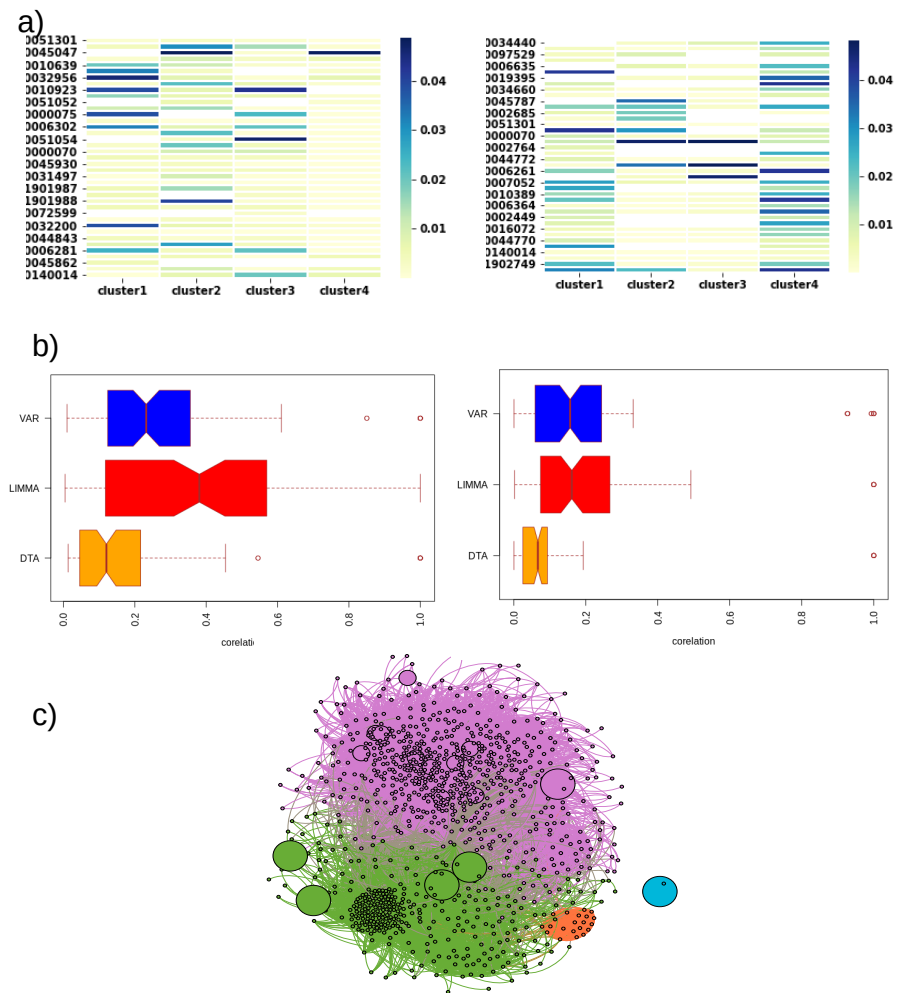


Figure 4. a) Enriched biological functions of each subtype of a cancer type where DTA was used as a feature reduction technique. xlabel denotes the enriched GO terms and color intensity represents the p-value of the association. b) Analysis of how DTA selected features are different from those identified by other feature selection methods. Here, we compared the correlation among the genes for different feature selection and DE methods. The left-side figure is for BRCA and the right side is for COAD. c) Predicted gene-gene interaction network of BRCA. The big circle shows the DTA selected genes, which are distributed over the network. The medium-sized circle denotes the genes selected by LIMMA, which mainly belong to one cluster.

4. Discussion

Feature selection is a crucial step of biological data analysis as biological measurements contain a high number of features compared to samples. Here, we present an algorithm that showed a remarkable improvement over existing feature selection techniques for disease classification and subtype detection problems. Genes selected by our algorithm were previously validated as shown in Table 1. Furthermore, we performed a few tests to analyze some of the algorithm's properties, making it an excellent feature selection method. To ensure each identified cluster in a cancer type is functionally different, we enriched the functional GO terms from DE genes of each cluster. The enriched functions of each

cluster are quite different from each other (Fig. 4a). We also computed the correlation between the selected genes by our algorithm and compared them with the other feature selection methods. We observed a very low correlation between the selected genes, which ensures the capability to choose non-redundant features by our method. The mean gene to gene correlation is higher for LIMMA and VAR than DTA, as shown in Fig. 4b. Lastly, we predicted the gene regulatory network from gene expression data of a cancer type. We have used a consensus of six different gene-regulatory network prediction algorithms to get a high confidence regulatory network based on our previously developed pipeline [33]. Next, we performed a clustering to group similar genes in the network. So the genes belonging to a cluster regulate each other more than regulating genes from the other clusters. We found that the genes selected by DTA belonged to different clusters as shown in Fig. 4c. Whereas feature selection methods like LIMMA and VAR choose the features, i.e., genes, that belong to the same cluster.

5. Conclusions

In this work, we introduced a novel feature selection technique that selects important and non-redundant disease-related features. We applied DTA for three different biological problems. DTA outperformed other feature selection techniques for classifying healthy samples to cancer samples, multiclass classification of various cancers, and cancer subtype detection. Currently, DTA operates on a patient-specific binary perturbation matrix. Hence, some information can be lost due to discretization to create the patient-specific binary profile. Thus, one potential improvement over our current method is to design extensions of the DTA algorithm to work with continuous input. Theoretically, this can reduce the information loss due to discretization and further improve the results.

Author Contributions: Study conceptualization, P.R., P.T., T.D. and P.G.; methodology, P.R., P.T., T.D. and P.G.; software, P.R. and P.T.; validation, P.R., P.T., T.D. and P.G.; data curation and preprocessing P.R.; writing—original draft preparation, P.R., P.T., T.D. and P.G.; writing—review and editing, P.R., P.T., T.D. and P.G.; visualization, P.R.; supervision, T.D. and P.G.; funding acquisition, T.D. and P.G. All authors have read and agreed to the published version of the manuscript.

Funding: This study was financed in part by XXX

Institutional Review Board Statement: Not applicable

Informed Consent Statement: Not applicable

Data Availability Statement: The data and the codes are available online at https://github.com/bnetlab/DTA_feature_selection.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; nor in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

PCA	Principle component analysis
SVD	Singular value decomposition
LDA	Linear discriminant analysis
TCGA	The cancer genome atlas
DE	Differential expression
LASSO	Least absolute shrinkage and selection operator
DTA	Dynamic threshold algorithm
PEEP	Patient-specific perturbation profile
5-CV	five-fold cross-validation
SVM	Support vector machine
ROC	Receiver operating characteristic
AUC	Area under curve
LUSC	Lung squamous cell carcinoma
LIMMA	Linear models for microarray data
SNF	Similarity network fusion
BRCA	Breast Invasive Carcinoma
LUAD	Lung adeno-179carcinoma
PRAD	Prostate adenocarcinoma
COAD	Colon Adenocarcinoma
KICH	Kidney Chromophobe
GBM	Glioblastoma Multiforme
VAR	maximum variance

References

1. Yu, L.; Liu, H. Efficient feature selection via analysis of relevance and redundancy. *Journal of machine learning research* **2004**, 5, 1205–1224.

2. Ang, J.C.; Mirzal, A.; Haron, H.; Hamed, H.N.A. Supervised, unsupervised, and semi-supervised feature selection: a review on gene selection. *IEEE/ACM transactions on computational biology and bioinformatics* **2016**, 13, 971–989.

3. Boutsidis, C.; Mahoney, M.W.; Drineas, P. An improved approximation algorithm for the column subset selection problem. *Proceedings of the twentieth annual ACM-SIAM symposium on Discrete algorithms*. SIAM, 2009, pp. 968–977.

4. Ordozgoiti, B.; Canaval, S.G.; Mozo, A. A fast iterative algorithm for improved unsupervised feature selection. 2016 IEEE 16th International Conference on Data Mining (ICDM). IEEE, 2016, pp. 390–399.

5. Bi, W.; Kwok, J. Efficient multi-label classification with many labels. *International Conference on Machine Learning*, 2013, pp. 405–413.

6. Saeys, Y.; Inza, I.; Larrañaga, P. A review of feature selection techniques in bioinformatics. *bioinformatics* **2007**, 23, 2507–2517.

7. TCGA. <https://www.cancer.gov/tcga>. Accessed: 2010-09-30.

8. Nguyen, H.; Thai, P.; Thai, M.; Vu, T.; Dinh, T. Approximate k-Cover in Hypergraphs: Efficient Algorithms, and Applications. *arXiv preprint arXiv:1901.07928* **2019**.

9. Menche, J.; Guney, E.; Sharma, A.; Branigan, P.J.; Loza, M.J.; Baribaud, F.; Dobrin, R.; Barabási, A.L. Integrating personalized gene expression profiles into predictive disease-associated gene pools. *NPJ Systems Biology and Applications* **2017**, 3, 10.

10. Bateni, M.; Esfandiari, H.; Mirrokni, V. Optimal distributed submodular optimization via sketching. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018, pp. 1138–1147.

11. Smyth, G.K. Limma: linear models for microarray data. In *Bioinformatics and computational biology solutions using R and Bioconductor*; Springer, 2005; pp. 397–420.

12. Mo, Q.; Wang, S.; Seshan, V.E.; Olshen, A.B.; Schultz, N.; Sander, C.; Powers, R.S.; Ladanyi, M.; Shen, R. Pattern discovery and cancer gene identification in integrated cancer genomic data. *Proceedings of the National Academy of Sciences* **2013**, 110, 4245–4250.

13. Wang, B.; Mezlini, A.M.; Demir, F.; Fiume, M.; Tu, Z.; Brudno, M.; Haibe-Kains, B.; Goldenberg, A. Similarity network fusion for aggregating data types on a genomic scale. *Nature methods* **2014**, 11, 333.

14. Nguyen, T.; Tagett, R.; Diaz, D.; Draghici, S. A novel approach for data integration and disease subtyping. *Genome research* **2017**, 27, 2025–2039.

15. Ramazzotti, D.; Lal, A.; Wang, B.; Batzoglou, S.; Sidow, A. Multi-omic tumor data reveal diversity of molecular mechanisms that correlate with survival. *Nature communications* **2018**, 9, 4453.

16. Franco, E.F.; Rana, P.; Cruz, A.; Calderón, V.V.; Azevedo, V.; Ramos, R.T.; Ghosh, P. Performance Comparison of Deep Learning Autoencoders for Cancer Subtype Detection Using Multi-Omics Data. *Cancers* **2021**, 13, 2013.

17. Xu, T.; Le, T.D.; Liu, L.; Su, N.; Wang, R.; Sun, B.; Colaprico, A.; Bontempi, G.; Li, J. CancerSubtypes: an R/Bioconductor package for molecular cancer subtype identification, validation and visualization. *Bioinformatics* **2017**, 33, 3131–3133.

18. Yu, G.; Wang, L.G.; Han, Y.; He, Q.Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *Omics: a journal of integrative biology* **2012**, 16, 284–287.

19. Colaprico, A.; Silva, T.C.; Olsen, C.; Garofano, L.; Cava, C.; Garolini, D.; Sabedot, T.S.; Malta, T.M.; Pagnotta, S.M.; Castiglioni, I.; others. TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic acids research* **2015**, *44*, e71–e71.
20. Beltran, A.; Graves, L.; Blancafort, P. Novel role of Engrailed 1 as a prosurvival transcription factor in basal-like breast cancer and engineering of interference peptides block its oncogenic function. *Oncogene* **2014**, *33*, 4767.
21. Kallioniemi, O.P.; Kallioniemi, A.; Kurisu, W.; Thor, A.; Chen, L.C.; Smith, H.S.; Waldman, F.M.; Pinkel, D.; Gray, J.W. ERBB2 amplification in breast cancer analyzed by fluorescence in situ hybridization. *Proceedings of the National Academy of Sciences* **1992**, *89*, 5321–5325.
22. Ayyoub, M.; Scarlata, C.M.; Hamaï, A.; Pignon, P.; Valmori, D. Expression of MAGE-A3/6 in primary breast cancer is associated with hormone receptor negative status, high histologic grade, and poor survival. *Journal of immunotherapy* **2014**, *37*, 73–76.
23. Matsusaka, K.; Ushiku, T.; Urabe, M.; Fukuyo, M.; Abe, H.; Ishikawa, S.; Seto, Y.; Aburatani, H.; Hamakubo, T.; Kaneda, A.; others. Coupling CDH17 and CLDN18 markers for comprehensive membrane-targeted detection of human gastric cancer. *Oncotarget* **2016**, *7*, 64168.
24. Walsh, M.D.; Clendenning, M.; Williamson, E.; Pearson, S.A.; Walters, R.J.; Nagler, B.; Packenas, D.; Win, A.K.; Hopper, J.L.; Jenkins, M.A.; others. Expression of MUC2, MUC5AC, MUC5B, and MUC6 mucins in colorectal cancers and their association with the CpG island methylator phenotype. *Modern Pathology* **2013**, *26*, 1642.
25. He, J.; Shin, H.; Wei, X.; Kadegowda, A.K.G.; Chen, R.; Xie, S.K. NPC1L1 knockout protects against colitis-associated tumorigenesis in mice. *BMC cancer* **2015**, *15*, 189.
26. Alajez, N.M. Large-scale analysis of gene expression data reveals a novel gene expression signature associated with colorectal cancer distant recurrence. *PloS one* **2016**, *11*, e0167455.
27. Said, H.M.; Safari, R.; Al-Kafaji, G.; Ernestus, R.I.; Löhr, M.; Katzer, A.; Flentje, M.; Hagemann, C. Time- and oxygen-dependent expression and regulation of NDRG1 in human brain cancer cells. *Oncology reports* **2017**, *37*, 3625–3634.
28. Murugesan, S.N.; Yadav, B.S.; Maurya, P.K.; Chaudhary, A.; Singh, S.; Mani, A. Expression and network analysis of YBX1 interactors for identification of new drug targets in lung adenocarcinoma. *Journal of genomics* **2018**, *6*, 103.
29. Shukla, S.A.; Bachiredy, P.; Schilling, B.; Galonska, C.; Zhan, Q.; Bango, C.; Langer, R.; Lee, P.C.; Gusenleitner, D.; Keskin, D.B.; others. Cancer-germline antigen expression discriminates clinical outcome to CTLA-4 blockade. *Cell* **2018**, *173*, 624–633.
30. Yang, R.; Li, S.W.; Chen, Z.; Zhou, X.; Ni, W.; Fu, D.A.; Lu, J.; Kaye, F.J.; Wu, L. Role of INSL4 Signaling in Sustaining the Growth and Viability of LKB1-Inactivated Lung Cancer. *JNCI: Journal of the National Cancer Institute* **2018**.
31. Lin, C.Y.; Lee, C.H.; Chuang, Y.H.; Lee, J.Y.; Chiu, Y.Y.; Lee, Y.H.W.; Jong, Y.J.; Hwang, J.K.; Huang, S.H.; Chen, L.C.; others. Membrane protein-regulated networks across human cancers. *Nature communications* **2019**, *10*, 3131.
32. Lau, S.K.; Weiss, L.M.; Chu, P.G. Differential expression of MUC1, MUC2, and MUC5AC in carcinomas of various sites: an immunohistochemical study. *American journal of clinical pathology* **2004**, *122*, 61–69.
33. Nalluri, J.J.; Barh, D.; Azevedo, V.; Ghosh, P. miRsig: a consensus-based network inference methodology to identify pan-cancer miRNA-miRNA interaction signatures. *Scientific reports* **2017**, *7*, 39684.