

博士論文

Ensemble feature selection approaches for discovery of candidate cancer biomarkers

(癌バイオマーカー候補発見のためのアンサンブル特徴選択技法)

文 銘 填

Table of Contents

Abstract	3
Introduction	5
Section 1: Stable feature selection based on the ensemble l_1-norm support vector machine for biomarker discovery	8
Background.....	9
Methods	12
Materials and pre-processing	12
Ensemble l_1 -norm SVM-RFE.....	14
Backward feature elimination for optimal feature subset selection.....	22
Performance test.....	23
Results	25
Feature selection	25
Performance evaluation.....	30
Section 2: Integrative analysis of gene expression and DNA methylation using unsupervised feature extraction for detecting candidate cancer biomarkers	38
Background.....	39
Methods	41
Materials and pre-processing	42
Normalization and rescaling.....	44
Feature extraction and selection	45
Performance test.....	48
Results	49
Normalization and feature extraction	49
Feature selection	51
Performance evaluation.....	56
Conclusion	60
References	63
Acknowledgements	70

Abstract

Lately, biomarker discovery has gained prominence as a significant research issue in the biomedical field. Owing to the presence of high-throughput technologies, genomic data, such as microarray data and RNA-seq, have become widely available. As biomarker discovery is typically modeled to determine the most discriminating features from the datasets, it can be described as a feature selection problem regarding class. Many kinds of feature selection techniques have been applied to retrieve significant biomarkers from these kinds of data. However, they tend to contain high-dimensional features with only a small number of samples; thus, conventional feature selection approaches may be problematic in terms of reproducibility. In addition, conventional studies mostly focus on genes that are differentially expressed in different states of cancer only; however, noise in gene expression datasets and insufficient information in limited datasets impede precise analysis of novel candidate biomarkers.

In this thesis, I propose ensemble feature selection approaches for discovery of candidate cancer biomarkers. Ensemble feature selection involves integration of different kinds of feature selectors to obtain more robust results through instance perturbation and merging of multiple datasets. First, I propose an ensemble l_1 -norm support vector machine to efficiently reduce irrelevant features by considering the feature stability. I define the stability score for each feature by aggregating the ensemble results, and utilize backward feature elimination on a purified feature set based on this score; therefore, it is possible to acquire an optimal set of features for improved performance without the need to set a specific threshold. The proposed

methodology is evaluated by classifying the binary stage of renal clear cell carcinoma with RNA-seq data. A comparison with established algorithms enables me to prove the superior performance of my method in terms of classification as well as stability in general. It is also shown that the proposed approach performs moderately on high-dimensional datasets consisting of a very large number of features and a smaller number of samples. In addition, I propose an integrative analysis of gene expression and DNA methylation using normalization and unsupervised feature extractions to identify candidate cancer biomarkers. Gene expression and DNA methylation datasets are normalized through Box-Cox transformation and integrated into a one-dimensional dataset that retains the major characteristics of the original datasets through unsupervised feature extraction methods. Differentially expressed genes are then selected from the integrated dataset. Use of the integrated dataset yields improved performance as compared with conventional approaches that utilize gene expression or DNA methylation datasets alone. Validation based on literature shows that a considerable number of top-ranked genes from the integrated dataset have known relationships with cancer, implying that novel candidate biomarkers can also be acquired from the proposed analysis method. The proposed approaches are expected to be applicable to various research studies that aim at candidate cancer biomarker discovery.

Introduction

At present, biomarker discovery is one of the most important research topics worldwide. In particular, detection of significant biomarker genes related to cancer is important for early diagnosis and treatment of cancer. With the prevalence of high-throughput technologies, genomic data such as RNA-seq have become widely available for studies targeting biomarker discovery. Generally, biomarker discovery is modeled as a feature selection process that determines the most discriminating features from datasets. However, because of the availability of a limited number of samples as compared to the number of features and the existence of a large amount of noise in genomic datasets, conventional feature selection techniques cannot be applied directly.

In this study, I propose ensemble feature selection approaches for discovery of candidate cancer biomarkers. Ensemble feature selection can be defined as the use of different selectors through integration of feature selection methods, feature perturbation, instance perturbation, or dataset perturbation, as shown in Figure 1. Ensemble feature selection involves integration of different kinds of feature selectors to obtain more robust results through instance perturbation or merging of multiple datasets. Section 1 describes a novel ensemble feature selection method based on the l_1 -norm support vector machine (SVM) [1]. To be specific, l_1 -norm SVM, which efficiently reduces the number of irrelevant or redundant features and produces sparse feature sets, is applied over bootstrap samples produced by random sampling of the original dataset. Through this process, high stability as well as high classification performance can be achieved. Most of the research described in

section 1 has been referenced from [1]. Section 2 focuses on integrative analysis of gene expression and DNA methylation. It is known that DNA methylation is a key regulator of gene expression; thus, it is expected that better understanding of gene-regulatory mechanisms can be acquired by integrating DNA methylation and gene expression datasets. I apply Box-Cox normalization and unsupervised feature extraction methods to merge DNA methylation and gene expression datasets into a one-dimensional dataset containing their main characteristics. Integrative analysis shows generally superior performance as compared to use of the gene expression or DNA methylation dataset only. I believe that the proposed method can be applied efficiently to several kinds of biomarker discovery task.

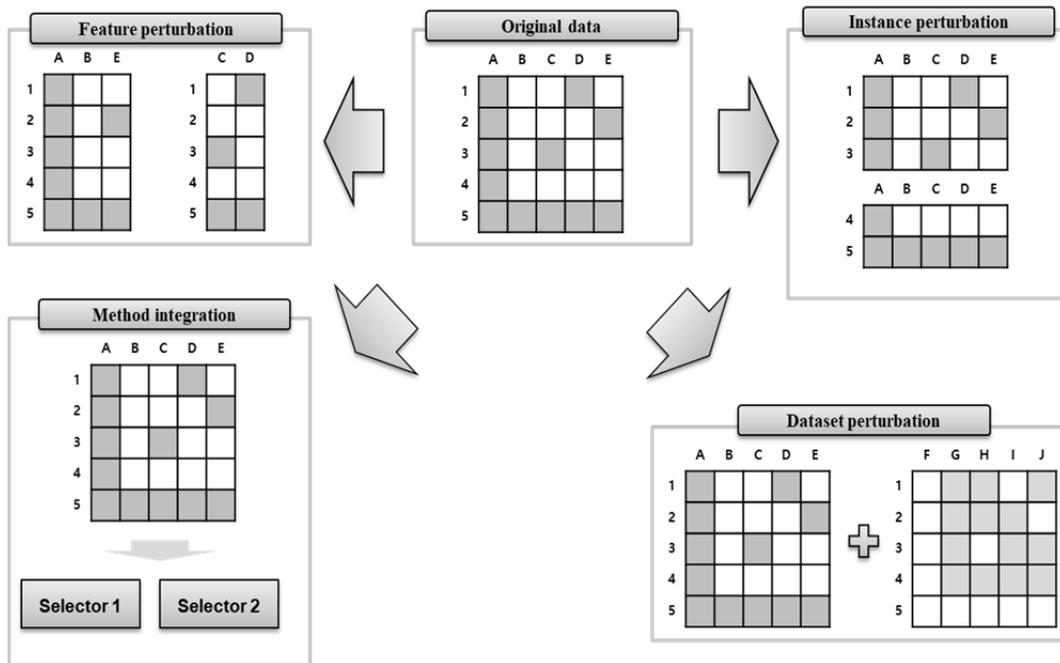


Figure 1 – Examples of ensemble feature selection methods

The ensemble feature selection can be implemented through the integration of feature selection methods, feature perturbation, instance perturbation, and dataset perturbation. In this research, instance perturbation (Section 1) and dataset perturbation (Section 2) are utilized as ensemble selection methods.

Section 1:

**Stable feature selection based on the ensemble l_1 -norm
support vector machine for biomarker discovery**

Background

Biomarker discovery has gained prominence as a significant research objective in recent years. Because biomarker discovery is typically modeled to determine the most discriminating features for classification, it can be described as a feature selection problem regarding class from the viewpoint of machine learning [2-4]. Feature selection is a step that involves identification of the most salient features for learning [5] and enables the performance of the classifier to be enhanced by eliminating irrelevant features that cause inaccurate prediction or over-fitting problems. In addition, the time required for learning is reduced as feature selection serves to lower the dimensionality. Although classification without a feature selection process may improve the classification performance, many features complicate result interpretation. In short, feature selection not only enhances the classification performance, but also improves understanding and analysis of the data. The emergence of new high-throughput technologies has made genomic data, such as microarray data and RNA-seq, widely available for biomarker discovery. However, the distinct characteristics of biomedical data, which often contain far more features than the number of samples, ensure that applying conventional feature selection approaches may be problematic, especially regarding reproducibility. That is, small changes in the dataset could induce large changes in the feature selection result, emphasizing some features that should not be considered as candidates for important biomarkers. Feature selection in the biomedical field should consider feature stability as well as influence on the classification performance. Some researchers suggested use of ensemble feature selection based on instance perturbation to address this problem [6-8]. Their

investigations proved that the stability of the selected features was significantly improved by performing feature selection on slightly different datasets and aggregating their results. Furthermore, research combining lasso regression with resampling was performed [9,10]. This showed that the l_1 -norm of lasso regression tends to force the solution to sparsity, and exhibits high efficiency for feature selection in regression problems.

In this study, I propose a stable feature selection method based on the l_1 -norm support vector machine (SVM). The basic concept of the l_1 -norm SVM is similar to that of lasso regression, but the former is tailored for classification tasks, which is the model for many biomarker discoveries. In addition, as SVM tries to maximize the margin between the closest support vectors, the classification model becomes more robust. The l_1 -norm SVM efficiently reduces the number of irrelevant or redundant features to fewer than the number of samples; thus, it is appropriate for biomedical high-dimensional data. As the proposed method is applied over instance perturbation steps, the stability issue, which is one of the most critical problems of the l_1 -norm, can also be managed. Furthermore, the optimal subset of selected features is detected by applying backward feature elimination to the proposed own ranking criteria. By eliminating features one by one based on the ranking criteria as generated by the l_1 -norm SVM, a cross-validated classification score is calculated and a subset of features that maximizes classifier performance is acquired.

Here, the proposed method is tested for renal clear cell carcinoma stage classification. I use an RNA-seq gene expression dataset of renal clear cell carcinoma samples from the Cancer Genome Atlas (TCGA) for my study. I compare my approach with three established feature selection methods, namely, a

fast correlation-based filter (FCBF) [11], random forest [12], and an ensemble version of SVM-based recursive feature elimination (SVM-RFE) [13], from the viewpoint of classification performance and stability of the selected features. This experiment shows that my method is capable of good classification performance and stability.

Methods

Materials and pre-processing

The RNA-seq gene expression dataset of renal clear cell carcinoma was obtained from Broad GDAC Firehose, one of the genome data analysis centers of the TCGA project [14]. Level-3 RNAseqV2 datasets of kidney renal clear cell carcinoma (KIRC) were used for the experiments. Several studies have used reads per kilobase million (RPKM) or fragments per kilobase million (FPKM) to measure gene expression levels from RNA-seq data. However, RPKM and FPKM are unsuitable for comparisons among different samples [15]; therefore, I selected transcripts per million (TPM) to measure gene expression. TPM allows comparisons among different samples, as the sums of all TPMs in each sample are the same. TPM can be calculated using FPKM by the following expression:

$$TPM_i = \frac{FPKM_i}{\sum_{j=1}^n FPKM_j} \cdot 10^6 \quad (1)$$

TPM vectors obtained through expectation maximization (RSEM) [16] and normalized by z -score were used as estimates for the gene expression level. I utilized only tumor samples and discarded genes and samples that contained invalid or null values. The pathogenic stage information of the renal clear cell carcinoma samples was retrieved from TCGA clinical dataset biotab files and set as the class labels of the gene expression data. Basically, the stage was divided into four stages, namely, stages I, II, III, and IV, based on the tumor-node-metastasis (TNM) stage groupings, which are decided by the size of the tumor, the lymph nodes involved, and distant metastasis [17]. I considered only two stages, i.e.,

stages I and IV, as stage-I renal clear cell carcinoma involves local tumors that only exist in the kidney, whereas tumors at stage-IV have grown into other tissues outside the kidney or have spread widely to other lymph nodes. Thus, the use of these stages could provide significant clues regarding tumor advancement and tumor metastasis. Samples for which the stage information was unclear were excluded from the test. After the filtering steps, the dataset consisted of 352 samples, of which 268 and 84 were stage-I and stage-IV samples, respectively. Each sample consisted of the TPM vector for 20199 genes.

Ensemble l_1 -norm SVM-RFE

SVM is an effective and popular method in machine learning, and its usage includes applications of this method to biomedical problems [18]. Conventionally, SVM has been used for classification tasks, but it can be also applied to feature selection by considering the weights of the classifier [4]. In addition, instead of using the general l_2 -norm for SVM, application of the l_1 -norm, which tends to produce sparse solutions, makes it possible to considerably reduce the number of features of a large feature set [19,20]. The maximum number of features selected by the l_1 -norm SVM is bounded by the number of samples [21]. Thus, this method is particularly suitable for biomarker selection from the RNA-seq data considered in this study, which usually contain far more features than the number of samples. However, application of the l_1 -norm SVM to a single dataset may produce a result that is excessively dependent on the sample set, even causing reproducibility uncertainty for datasets that are only slightly different. Moreover, difficulties are known to arise when applying the l_1 -norm to select closely correlated factors, as it tends to select only a single feature from among them and ignore the rest [22]. These problems can be addressed by applying the l_1 -norm SVM to a perturbed dataset. Here, I propose ensemble l_1 -norm SVM feature selection combined with data perturbation to consider stability. The flow of the algorithm is illustrated in Figure 2. As the l_1 -norm SVM is recursively applied until the optimal subset of the features is found, the proposed methodology can be described as l_1 -norm SVM-RFE.

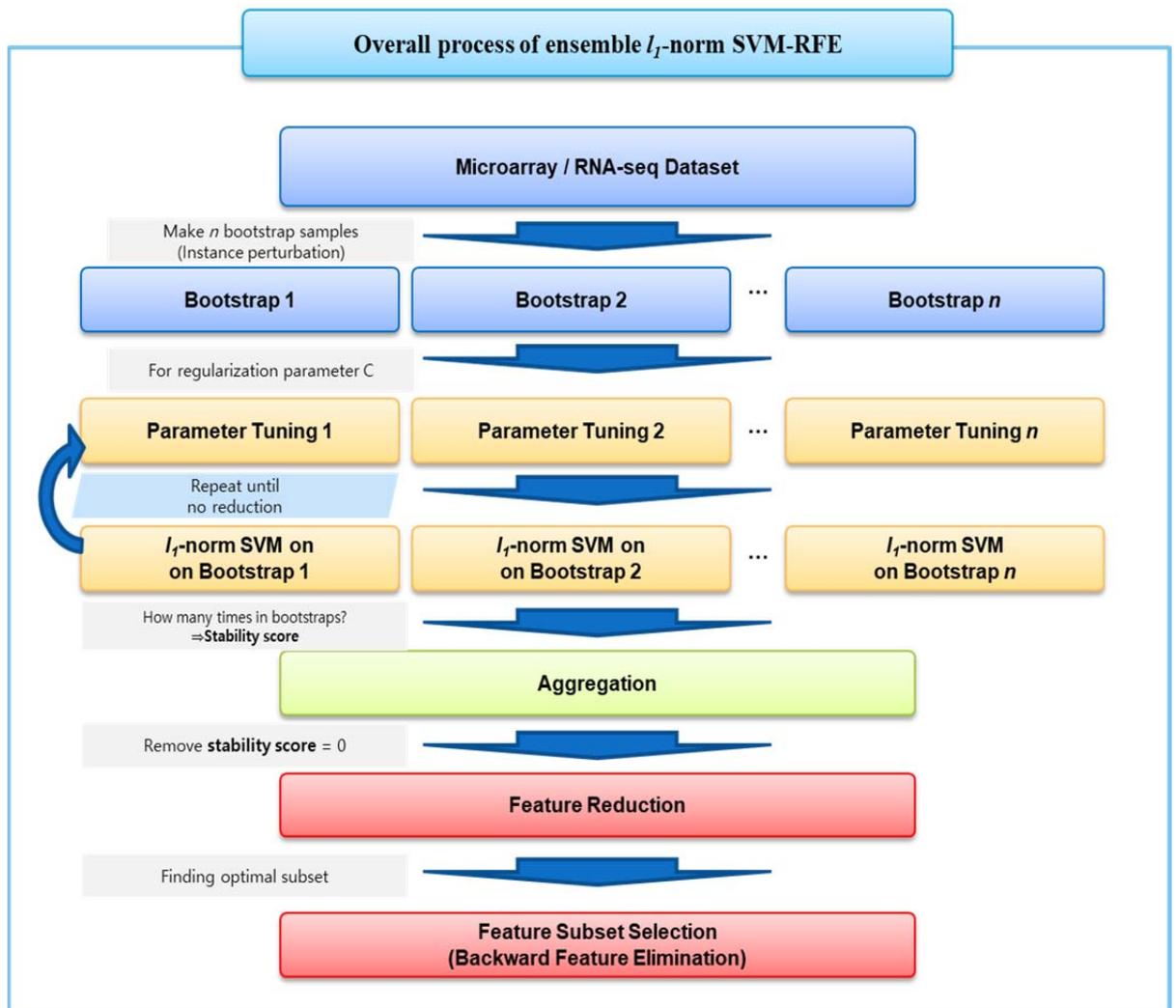


Figure 2 - The flow of the proposed method

- (1) Generate n random bootstrap samples, X_1, X_2, \dots, X_n , containing $i\%$ of the data of the whole training dataset X .
- (2) Perform a cross-validation test on each bootstrap sample to set regularization parameter C .
- (3) Apply the l_1 -norm SVM to each bootstrap sample. Then, the weight vector w is calculated for each feature.
- (4) Eliminate features for which the coefficient $w = 0$ in each bootstrap sample.
- (5) Record the cross-validation score for each bootstrap sample for step (7).
- (6) Repeat steps (2)–(5) until no more features with $w = 0$ are available for any bootstrap.

- (7) Select optimal feature subsets for each bootstrap sample, which maximize the cross-validation score recorded in step (5).
- (8) Produce the integrated feature set of size k by aggregating all the remaining features in the bootstrap samples.
- (9) Convert X to the reduced dataset X' that consists of features in k . Here, the number of bootstrap samples that contain a given feature is considered the “stability score” S for that feature ($1 \leq S \leq n$).

For stable feature selection, I use bagging to generate n bootstrap samples from the training dataset [23,24]. A simple example of the proposed bootstrap is shown in Figure 3. There is no solid rule for setting the optimal value of n , and a related study showed that adjusting n only marginally affects the classification performance or stability [7]. Hence, a moderate value of n can be set, although it is expected that a slightly more converged result can be acquired by using a larger value. After generating n bootstrap samples, the regularization parameter of the l_1 -norm SVM is optimized for each of the samples. Unlike the l_2 -norm SVM, the number of features to be reduced is automatically selected by the regularization parameter. The differences between l_1 -norm and l_2 -norm are listed in Table 1.

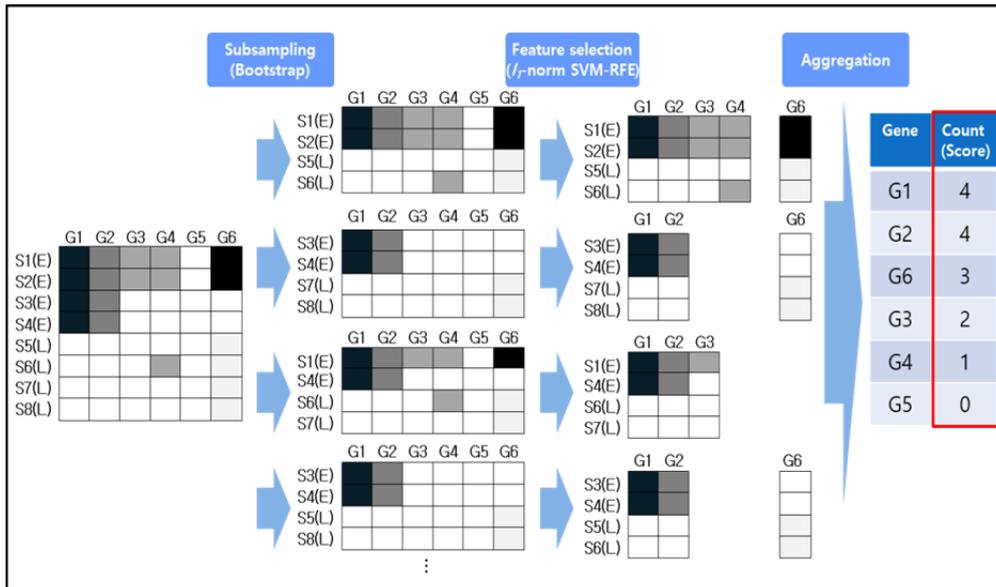


Figure 3 – An example of bootstrap aggregation

Table 1 – The difference between l_1 -norm and l_2 -norm SVM-RFE

l_1 -norm SVM-RFE	l_2 -norm SVM-RFE
Uses l_1 -norm penalty	Uses l_2 -norm penalty
$w_i = 0$ for less-important features ↳ remove features with $w_i = 0$	$w_i \geq 0$ for less-important features ↳ k features to remove per iteration
Automatic stopping criteria ↳ no more features of $w_i = 0$ (the only parameter - C)	Criteria for stopping iteration ↳ n features to select finally (multiple parameters – C, k, n)
Leads to sparsity ($p \leq n$)	Maintains density

The optimization problem of the l_1 -norm SVM can be described as follows.

$$\min_{w,b} \|w\|_1 + C \sum_{i=1}^n \max(0, 1 - y_i(w^T x_i + b))^2 \quad (2)$$

Here, parameter C reflects a regularization parameter that solves the trade-off problem between the training error and the complexity of the model, and $\|w\|_1$ denotes the l_1 -norm of the weight vector w . I apply a linear kernel as a kernel function, which is defined as follows.

$$k(x, y) = x^T y \quad (3)$$

If the number of features is large, the linear kernel is efficient, because mapping data to high-dimensional space usually does not improve the performance [25]. Hence, it is possible to obtain a comparable result at much lower cost. In addition, the linear kernel is less prone to over-fitting than non-linear kernels, and the only parameter that requires optimization is C . I find the optimal value of C by applying a grid search using 10-fold cross-validation on the training dataset, as demonstrated in Figure 4. The number of selected features are automatically adjusted by C as described in Figure 5.

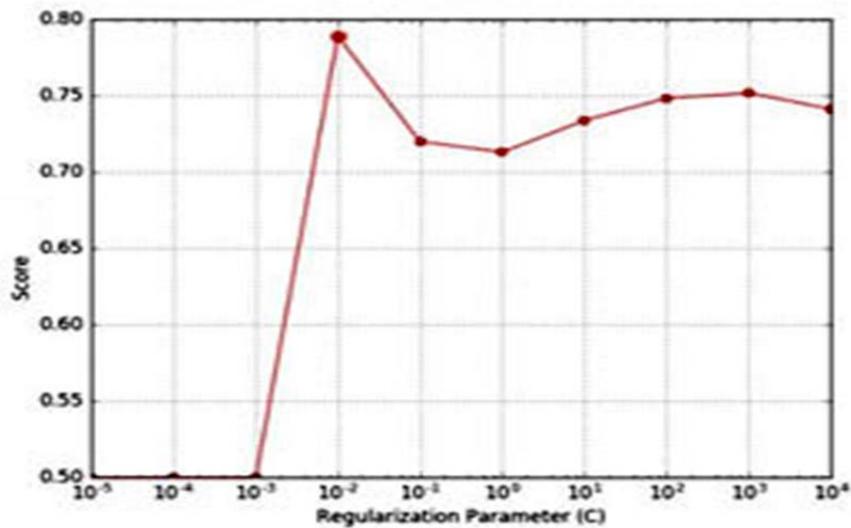


Figure 4 – An example of C optimization through grid search

The X- and Y-axes denote the value of C and the 10-fold cross-validation score, respectively. Here, 10^{-2} is selected as the optimal C value as its classification performance is the best.

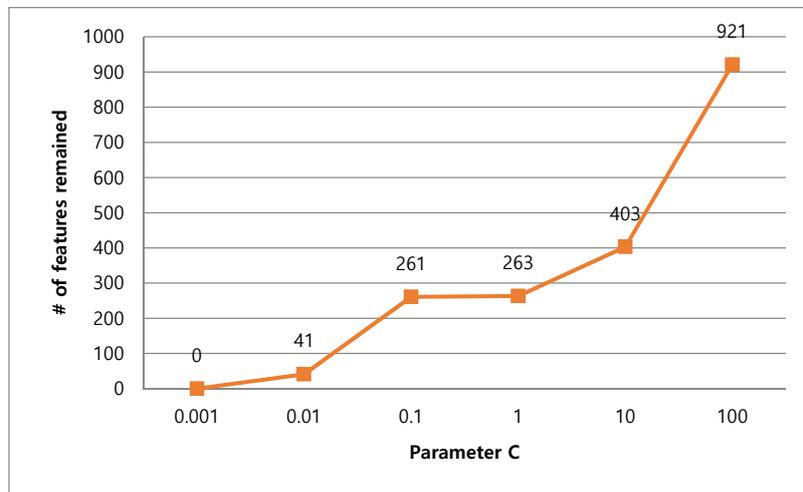


Figure 5 – An example of the correlation between the number of features selected and the regularization parameter C

The X- and Y-axes denote the value of C and the number of features remaining after the application of l_1 -norm SVM, respectively. If the larger C is selected, the number of selected features increases, and vice versa.

Next, the l_1 -norm SVM is implemented on each bootstrap sample to derive the sparse feature sets. In this step, the ranking criterion of the SVM for each feature, i.e., the size of w , is not considered. Therefore, all the features for which the coefficients have $w > 0$ are considered, which means that features that are selected for at least one bootstrap sample are considered. After eliminating features having coefficients with $w = 0$, I calculate and record the cross-validation score for each bootstrap sample. Similar to SVM-RFE, the above-mentioned steps are repeated until no more features can be reduced by the l_1 -norm SVM. However, reducing features does not always yield improved classification performance; sometimes, the classification performance may decrease through excessive filtering of features. Thus, I select the feature subsets that maximize the cross-validation scores as optimal features for each bootstrap sample.

Then, the features that remain in all the bootstrap samples are aggregated. In the aggregation step, I only consider the frequency of each feature among all the bootstrap samples. It is plausible that the features remaining in a greater number of bootstrap samples are more stable. Therefore, for each feature, I regard the number of bootstrap samples that contain that feature as its “stability score.” This score is valuable for regularization of the number of features, for e.g., to ignore the features selected from less than 10% of the bootstrap samples when there is a need to reduce the features to less than the number of samples; this scenario commonly arises in the biomedical field.

In the present study, instead of setting a specific cutoff point, I applied backward feature elimination based on the stability score to select the optimal subset, which is described in the next section. Finally, the dimension of the training dataset was

reduced to the number of features that remained after the previous steps.

Backward feature elimination for optimal feature subset selection

A stable l_1 -norm SVM provides an efficient solution for eliminating and selecting features. However, there might be a subset for which the performance exceeds the one achieved by using all the selected features. Therefore, an additional process capable of extracting the particular feature subset that can improve the classification performance is required. For this purpose, I utilize the stability score described in the previous section, which reflects the number of bootstrap samples that contain a given feature. The stability score is aggregated from multiple bootstrap samples, thereby managing the stability issue that can arise when only a single dataset is used. I apply backward feature elimination to find one optimal subset of features for which the classification performance is best.

In the experiment conducted in this study, I used SVM as the classifier; however, other classifiers can also be applied. Here, as the number of features decreased considerably because of the previous feature selection step, I applied the radial basis function (RBF) kernel, which generally performs better than the linear kernel.

$$k(x, y) = \exp(-\gamma |x - y|^2) \quad (4)$$

Again, a grid search approach using 10-fold cross-validation was applied to optimize the parameters of the RBF kernel, C , and γ . The classification score was calculated on each fold of the cross-validation test and then aggregated. By eliminating the features with the lowest stability scores one by one, the classification score was calculated on all progressively smaller subsets until the subset size reached 1. Finally, the one feature subset that succeeded in maximizing the classification performance over the cross-validation test was acquired.

Performance test

Conventionally, the feature selection performance is evaluated by measuring the classification performance. However, recent studies have placed great importance on the stability of the selected features, which indicates the reliability and reproducibility of the features. Instability of feature selection is mainly caused by the association of many features with a small number of data samples, which complicates the proper reduction of features. These characteristics are common for biomedical data such as RNA-seq; thus, a biomarker discovery study should consider stability as well as classification performance.

Therefore, I evaluated the performance of the proposed method by measuring the classification performance and feature stability. Because the proposed feature selection procedure contains a resampling step, I employed an independent training and test set in addition to the cross-validation test. I applied a well-known classification algorithm after feature selection to evaluate the classification performance. Statistical measures in the form of the accuracy, F1 score, Matthews correlation coefficient (MCC), and area under the curve (AUC) were measured together to evaluate the classification performance.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (5)$$

$$F1 = \frac{2TP}{2TP + FP + FN} \quad (6)$$

$$MCC = \frac{(TP \cdot TN) - (FP \cdot FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (7)$$

$$AUC = \int_0^1 ROC(x) dx \quad (8)$$

Here, TP , TN , FP , FN , and ROC represent true positive, true negative, false positive, false negative, and the receiver operating characteristic curve, respectively.

Then, I tested the stability of the selected features by calculating the Tanimoto distance T , which has been applied in several previous studies [26,27]. T is a statistical measure for calculating overlaps between two sets of elements of arbitrary cardinality and is calculated as follows.

$$T(S_i, S_j) = 1 - \frac{|S_i| + |S_j| - 2|S_i \cap S_j|}{|S_i| + |S_j| - |S_i \cap S_j|} \quad (9)$$

Here $|S_i|$ and $|S_j|$ denote the number of elements in sets S_i and S_j , respectively. The Tanimoto distance T_n was obtained over multiple n sets of samples by calculating the arithmetic mean of T for each set of pairs, as described below.

$$T_n(S_i, S_j) = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^n (T(S_i, S_j)) \quad (10)$$

T_n takes values between 0 and 1, where 0 means there is no overlap between the two sets, and 1 indicates that the two sets have identical elements.

Results

Feature selection

The RNA-seq dataset was obtained from the Broad GDAC Firehose and filtered as described in the Methods section. Although cross-validation is one of the most popular methods for classification tests on a biological dataset, it tends to produce a dataset-dependent result, especially with a small sample size [28]. Hence, the classification performance should be evaluated through not only a cross-validation test, but also an independent data test. In this study, I ensured that the test remained independent of the dataset by randomly dividing the original dataset into a training set (80%) and a test set (20%). As a result, the training set consisted of 214 and 67 stage-I and stage-IV samples, respectively, whereas the test set contained 54 and 17 samples, respectively. Only the training dataset was utilized for the cross-validation test. The stability test was performed by randomly generating 20 subsets from the original dataset, each of which contained 80% of the entire number of samples. The FCBF feature selection test was implemented using Weka 3.7.13 [29], and the other experiments were all implemented in Python, using the scikit-learn 0.17 library [30].

Data perturbation was achieved by producing 1000 bootstrap samples containing 80% of the data from the training set. Then, feature selection was performed by first calculating the C of the l_1 -norm SVM through a grid search, using 10-fold cross-validation on each bootstrap sample of the training data. I determined the best value in the range of $C \in \{10^{-5}, 10^{-4}, \dots, 10^3, 10^4\}$. Because the dataset contained bias in class proportions, simply considering accuracy as a performance

estimator was inappropriate. Instead, I regarded the AUC as the main criterion for evaluating the experiments. Thus, the value of C resulting in the best AUC score was selected for each bootstrap sample.

Then, the l_1 -norm SVM was applied to 1000 bootstrap samples to filter the genes for which the coefficients were 0. I calculated and recorded the 10-fold cross-validation score of the reduced feature sets at this point, to choose optimal feature sets for each bootstrap in the later step. These steps were repeated several times until no further feature reduction was possible. The remaining genes were aggregated to one gene set and ranked by the number of bootstrap samples that finally remained.

Subsequently, as a revising step, backward feature elimination was performed to find the optimal feature subset for which the classification performance was best. Again, I used the AUC score as the main criterion for the classification performance. The SVM was selected as a classifier and 10-fold cross-validation was applied for the test. The grid-search method was also applied to set C and γ in the range of $C \in \{10^{-5}, 10^{-4}, \dots, 10^9, 10^{10}\}$ and $\gamma \in \{10^{-9}, 10^{-8}, \dots, 10^2, 10^3\}$, respectively.

Then, the mean AUC score obtained from the cross-validation test was calculated. The score was recursively calculated by reducing the genes one by one, starting from the full gene sets, until a subset consisting of only one gene was tested. Figure 6 demonstrates the alteration of the mean AUC score by use of backward feature elimination.

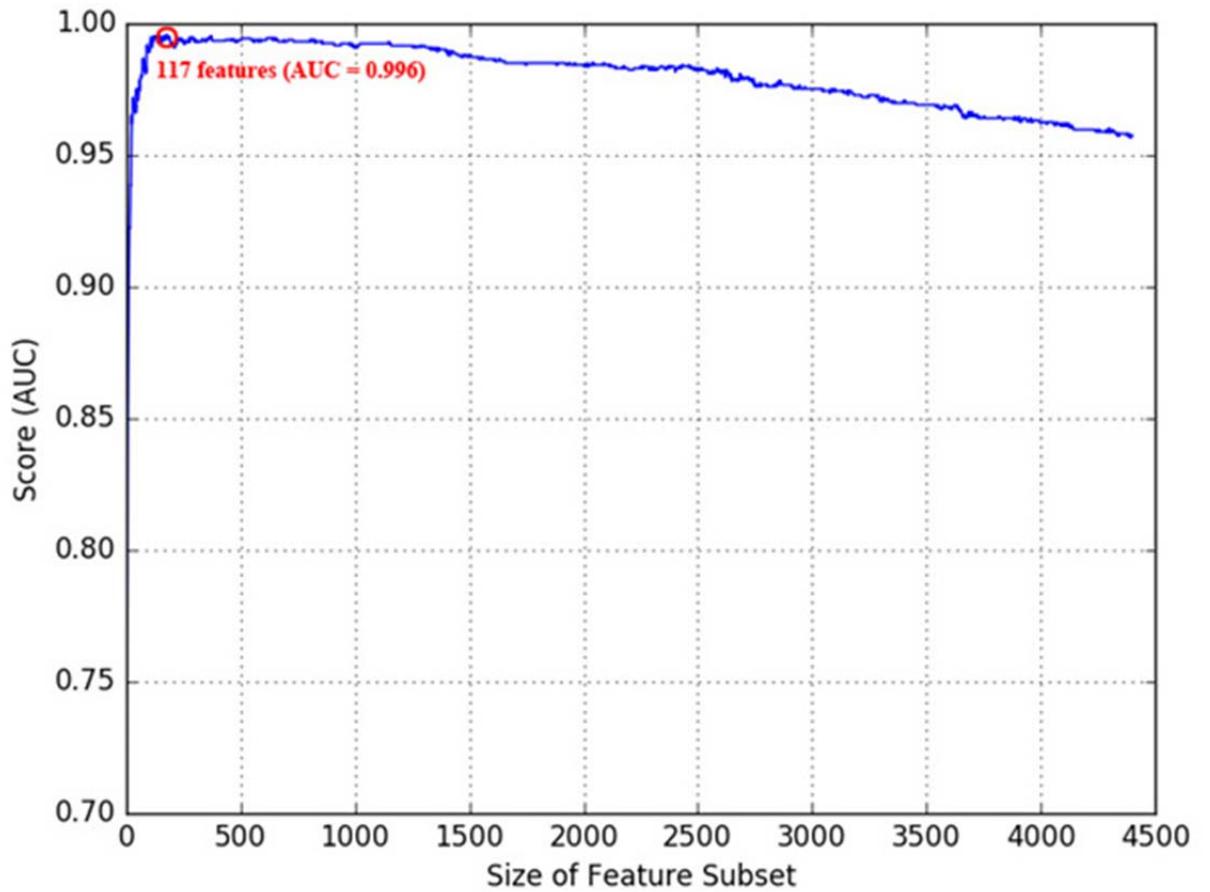


Figure 6 - Backward feature elimination for optimal subset

Backward feature elimination was performed based on the ranking criteria. SVM with an RBF kernel was used as a classifier for calculating the cross-validation score. The X- and Y-axes denote the size of the feature subset and 10-fold cross-validation AUC score, respectively. The red circle indicates the number of features with the highest AUC score in the experiment, i.e., 177 features with AUC = 0.996.

The performance of the proposed model was compared with three well-known feature selection methods, namely, FCBF, random forest, and an ensemble version of SVM-RFE. FCBF is a feature selection method that is used to remove irrelevant features based on symmetric uncertainty. Although FCBF does not consider the correlations among features, I selected it as a comparison target because this feature selection algorithm was adapted from a previous study that classified stage progression of renal clear cell carcinoma based on the RNA-seq dataset [31]. Random forest is a method based on decision trees that has been frequently used for both feature selection and classification. Random forest handles the stability issue by bootstrap aggregation (bagging), which is included in the algorithm. SVM-RFE is a feature selection method that recursively eliminates features for which the weight magnitudes of the l_2 -norm SVM are smallest [13]; this method has been proven to deliver superior performance, however, SVM-RFE is problematic in terms of stability when applied to a single dataset. Some previous researchers used the ensemble version of SVM-RFE based on instance perturbation to address this problem [32]. For the ensemble SVM-RFE employed in the comparison experiment, the fraction ratio for elimination in each step was set to 20%, and a linear aggregation method that sums the rank over all bootstraps was used, as detailed in Ref. [7]. The number of bootstraps was set to 1000, and C was optimized in the same way as for the proposed method. As random forest and SVM-RFE provide only the ranking list of all features, I applied the backward feature elimination method to find the optimal subset of features, in a manner similar to the proposed method. The 10-fold cross-validation score was calculated by the classifier while features were added one by one based on the feature

rankings. Then, the optimal feature set that maximized the cross-validation score was acquired. The random forest classifier and SVM classifier with the RBF kernel were used as classifiers for the random forest and ensemble SVM-RFE, respectively.

Performance evaluation

For evaluating the performance, I first performed a stability test by creating 20 random subsamples from the original dataset, which was constructed with 80% of the data. The stability test was performed using the Tanimoto distance T . I also tested the proposed method without bagging for comparison. Figures 7 and 8 show the mean and standard deviations of T for each method, respectively. As is apparent from the figures, the proposed method generally demonstrated higher stability when compared to the examined methods. The performance of the ensemble SVM-RFE was similar to that of the proposed method. However, the l_1 -norm SVM without instance perturbation exhibited remarkably lower scores than the other methods, which proves the significance of ensemble selection for l_1 -norm-based methods. FCBF also resulted in a lower stability score than the other methods as it basically does not consider the stability issue. However, it exhibited the least variance, as it selects almost the same number of features within subsamples.

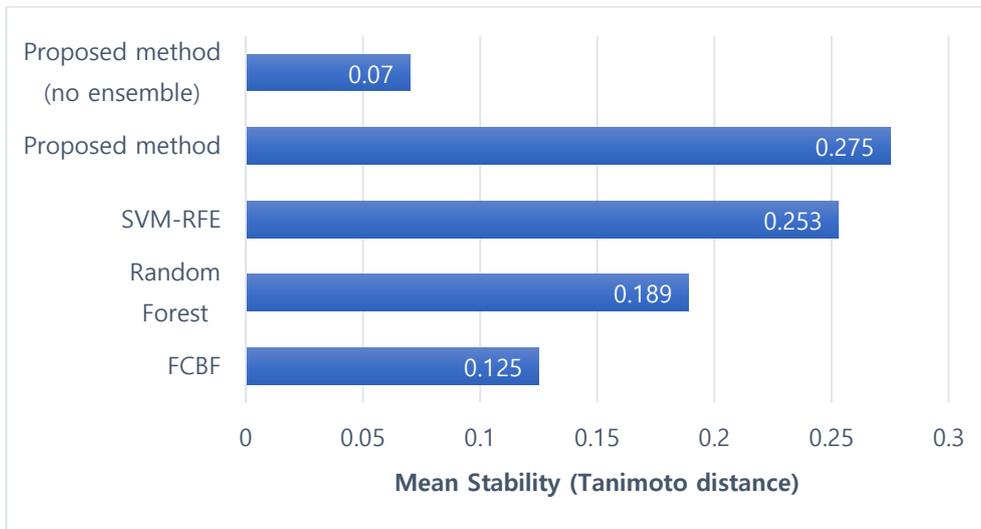


Figure 7 - Mean Tanimoto distance

The arithmetic mean of the Tanimoto distance on 20 random subsamples was calculated as the stability score. The X-axis denotes the arithmetic mean of the Tanimoto distance of each method.

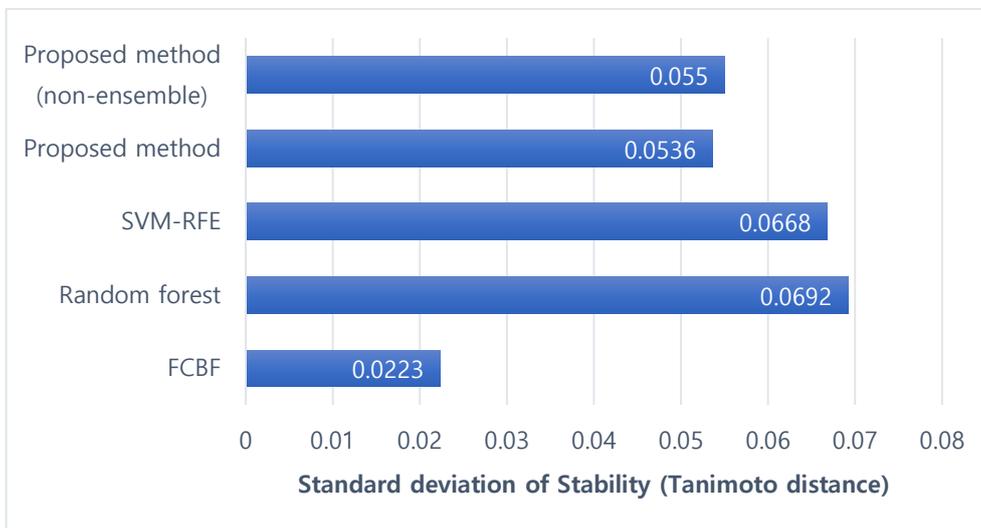


Figure 8 - Standard deviation of Tanimoto distance

The standard deviation of the Tanimoto distance on 20 random subsamples was calculated. The X-axis denotes the standard deviation of the Tanimoto distance of each method.

Then, a cross-validation test as well as an independent dataset test were conducted for evaluating the classification performance. As described in the subsection on datasets, the data included in the independent test set were not used for the feature selection process. Only the training set was used for the cross-validation test. Four popular classification methods, adaptive boosting (AdaBoost), logistic regression, random forest, and SVM with an RBF kernel, were used for performance evaluation. To deal with the overfitting problem, l_2 -norm was applied to logistic regression and SVM. Gene selection was performed by FCBF, random forest, ensemble SVM-RFE, and the proposed method before executing the classification tests. Figure 9 demonstrates the number of genes selected by each method. As random forest selected far more genes than the other methods and the number of genes was larger than that of the samples, there was a possibility of overfitting. Thus, an additional test with 180 genes was also conducted. The value of 180 was similar to the number of genes I used for the ensemble SVM-RFE method and the proposed method, and it resulted in only a minute difference in classification performance when compared to that obtained using 766 genes. I assessed four performance measures, namely, the accuracy, f1 score, MCC, and AUC for each classifier. Tables 2 and 3 compare the classification performance of the examined approaches for the independent data test and cross-validation test, respectively. As is apparent from the tables, the proposed algorithm exhibited the overall best performance as compared to most classifiers in terms of the performance indices in both the cross-validation test and independent dataset test. In particular, the proposed method with the SVM classifier exhibited the best performance among all techniques.

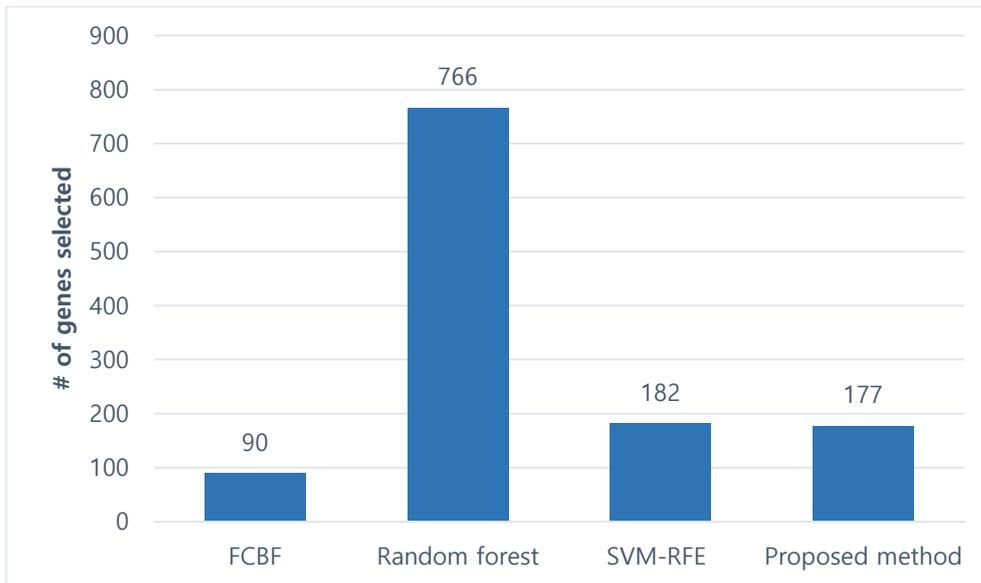


Figure 9 - The number of genes selected by each method

The chart shows the number of genes selected by each feature selection method, namely, FCBF, random forest, ensemble SVM-RFE, and the proposed method.

Table 2 - Classification performance test with the independent dataset

The performance scores of different methods were calculated on the independent dataset. The items that yielded the best scores are highlighted in bold. (The numbers below the random forest classifier denote the number of genes selected for the performance test.)

Classifier	Performance Measure	FCBF	Random forest (766)	Random forest (180)	Ensemble SVM-RFE	Proposed method
AdaBoost	Accuracy	0.789	0.845	0.859	0.887	0.901
	F1-score	0.545	0.593	0.667	0.75	0.774
	MCC	0.408	0.532	0.588	0.68	0.717
	AUC	0.7	0.717	0.766	0.825	0.834
Logistic regression	Accuracy	0.789	0.789	0.817	0.845	0.845
	F1-score	0.651	0.595	0.667	0.718	0.732
	MCC	0.533	0.456	0.552	0.623	0.646
	AUC	0.801	0.74	0.799	0.838	0.858
Random forest	Accuracy	0.817	0.817	0.803	0.831	0.845
	F1-score	0.48	0.48	0.462	0.5	0.56
	MCC	0.426	0.426	0.381	0.479	0.531
	AUC	0.658	0.658	0.649	0.667	0.697
SVM	Accuracy	0.803	0.831	0.859	0.831	0.901
	F1-score	0.632	0.571	0.583	0.684	0.811
	MCC	0.504	0.489	0.589	0.577	0.749
	AUC	0.77	0.708	0.706	0.808	0.895

Table 3 - Classification performance test with cross-validation

The mean performance scores of different methods were calculated for the 10-fold cross-validation test. The items that yielded the best scores are highlighted in bold. (The numbers below the random forest classifier denote the number of genes selected for the performance test.)

Classifier	Performance measure	FCBF	Random forest (766)	Random forest (180)	Ensemble SVM-RFE	Proposed method
AdaBoost	Accuracy	0.872	0.882	0.85	0.886	0.889
	F1-score	0.737	0.749	0.647	0.738	0.735
	MCC	0.668	0.677	0.568	0.676	0.686
	AUC	0.902	0.923	0.868	0.936	0.944
Logistic regression	Accuracy	0.833	0.853	0.822	0.957	0.978
	F1-score	0.704	0.722	0.664	0.915	0.958
	MCC	0.609	0.636	0.566	0.894	0.947
	AUC	0.904	0.893	0.853	0.994	0.997
Random forest	Accuracy	0.871	0.84	0.844	0.83	0.833
	F1-score	0.614	0.553	0.625	0.459	0.45
	MCC	0.579	0.504	0.557	0.473	0.457
	AUC	0.918	0.869	0.851	0.924	0.928
SVM	Accuracy	0.879	0.854	0.84	0.95	0.968
	F1-score	0.762	0.659	0.589	0.895	0.933
	MCC	0.692	0.58	0.514	0.865	0.914
	AUC	0.915	0.885	0.871	0.992	0.996

As the numbers of samples included in the datasets were quite small, even the reduced number of genes obtained after feature selection could be excessive. Thus, I conducted an additional classification performance test using only the top 20 genes selected by each feature selection method. Figure 10 demonstrates the AUC scores of each feature selection method for these top 20 genes. The results show that the proposed method demonstrated the best performance with limited number of features. Then, I subjected the top 20 genes selected by the proposed method, specifically, *C2orf55*, *CTAGE4*, *CPZ*, *ASAPIIT1*, *OASL*, *FABP7*, *NEGRI*, *SH2D1B*, *GARNL3*, *PLIN1*, *ZNF382*, *THEM220*, *SPATA7*, *LOC285733*, *CXCL11*, *KIR2DL4*, *LOC100134868*, *CENPBD1*, *KLRF1*, and *GLBIL3*, to further analysis. By searching the literature, databases, and gene annotations, I found that 14 of the 20 genes, namely, *C2orf55*, *CTAGE4*, *CPZ*, *OASL*, *FABP7*, *NEGRI*, *SH2D1B*, *GARNL3*, *PLIN1*, *ZNF382*, *THEM220*, *CXCL11*, *KIR2DL4*, and *KLRF1*, have known relationships with tumors. Further, 8 genes among those 14, *OASL*, *FABP7*, *NEGRI*, *GARNL3*, *PLIN1*, *ZNF382*, *THEM220*, and *CXCL11*, have known relationships with tumor progression and metastasis, which are directly related to the cancer stage. Thus, the other 6 genes that have no known relationships with tumors may also be considered as candidate genes for tumor or tumor progression and metastasis.



Figure 10 - Classification performance test using top 20 genes

The AUC scores for the top 20 genes selected by each feature selection method were calculated for the independent dataset and 10-fold cross-validation test.

Section 2:

Integrative analysis of gene expression and DNA methylation using unsupervised feature extraction for detecting candidate cancer biomarkers

Background

In recent decades, detection of candidate biomarkers has been the focus of cancer research to support timely and accurate diagnosis and prognosis of the disease [33,34]. In particular, detection of diagnostic biomarkers is essential to help prevent cancer progression [35]. Conventionally, gene expression analysis using microarray or RNA-seq data has been an effective method for cancer biomarker discovery, and several studies have focused on identifying genes that are differentially expressed in different cancer states [36-39]. However, gene expression profiles do not contain a sufficient number of samples, and large quantities of noise exist in the datasets. Therefore, common strategies based solely on gene expression datasets hold limited potential for identifying novel candidate biomarkers. Recently, genome-wide variations in DNA methylation levels were shown to contribute significantly to cancer development [40,41]. Because DNA methylation is a key regulator of gene expression, it is expected that better understanding of gene-regulatory mechanisms can be acquired by integrating gene expression and DNA methylation datasets. Therefore, while several recent studies have focused on integrative analysis of gene expression and DNA methylation, most of them have separately analyzed gene expression and DNA methylation, or focused solely on genes that exhibited high levels of both expression and methylation [42-44].

In this part of the study, I propose an integrative analysis of gene expression and DNA methylation using normalization and unsupervised feature extraction. Feature extraction is a method that derives a set of features that efficiently represent significant portions of the original multi-dimensional data. Hence, a reduced one-

dimensional dataset is obtained following application of feature extraction to gene expression and DNA methylation data, while retaining the main characteristics of the original datasets. This approach enables discovery of relationships between gene expression and DNA methylation for each gene that cannot be observed using single datasets. Therefore, analysis of differentially expressed genes in an integrated dataset enables discovery of novel, previously undetectable, candidate biomarkers.

Methods

The overall methodology followed in this section is illustrated in Figure 11.

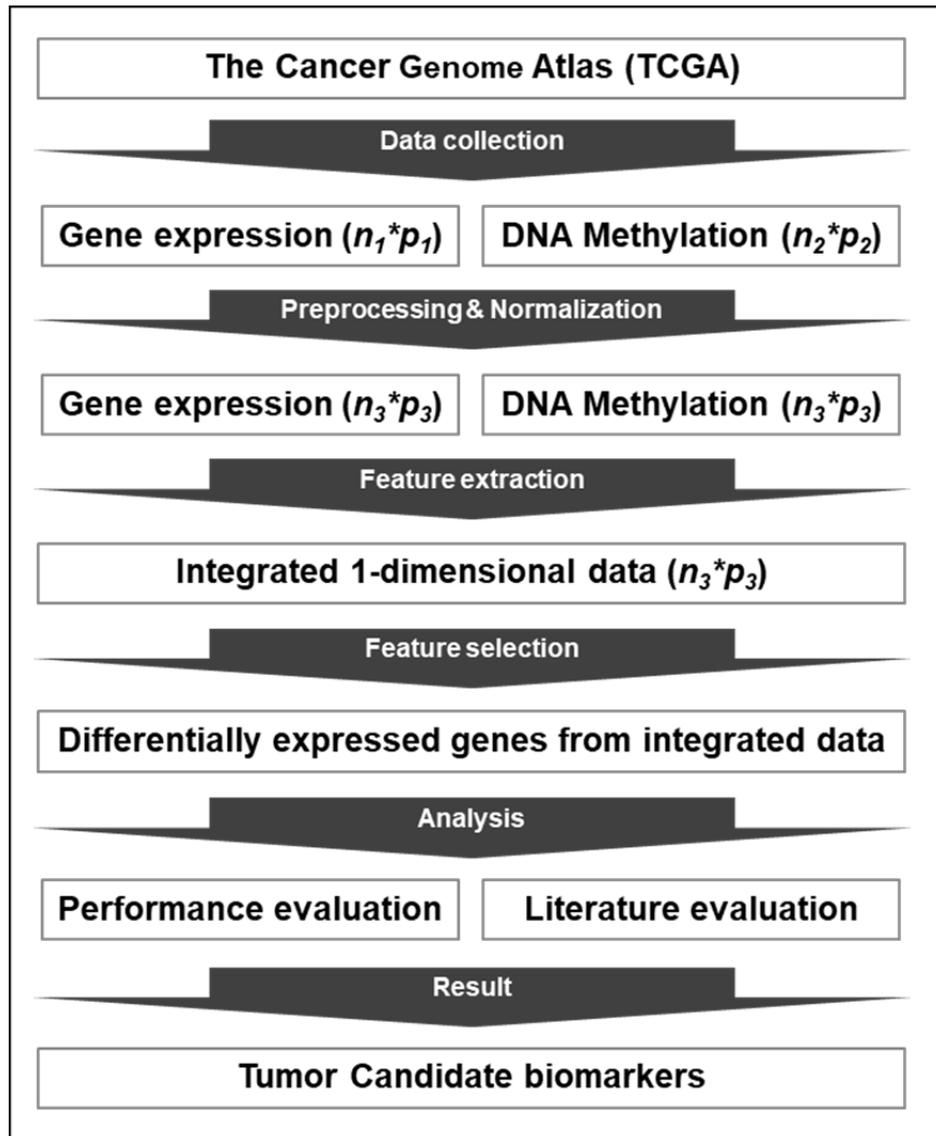


Figure 11 - Overall process

The process flowchart for the proposed method. Here, n and p denote the number of samples and genes in the datasets, respectively.

Materials and pre-processing

The dataset used for the experiments conducted in this part of the study included gene expression and DNA methylation datasets, as well as clinical datasets, acquired from the pan-kidney cohort (KIPAN) of the Broad GDAC Firehose, a genome data analysis center associated with The Cancer Genome Atlas (TCGA) project [14]. The KIPAN dataset comprises three major types of renal cell carcinomas (RCCs): clear cell, papillary, and chromophobe RCCs, each of which can be distinguished microscopically by their cancer-cell morphology [45]. The pathogenic stages of the RCCs are divided into stages I, II, III, and IV based on tumor-node-metastasis (TNM) grouping according to tumor size, involvement of lymph nodes, and distant metastasis [16]. RCCs in stages I and II are local tumors that exist in the kidneys and differ in size only, whereas stages III and IV metastasize to tissues and lymph nodes outside the kidney. Therefore, stage information can be converted into binary stages of early (stages I and II) and late (stages III and IV), which are set as the class labels. The goal of this part of the study was to identify genes highly correlated with a specific cancer stage, to present them as significant candidate biomarkers of cancer and its progression.

For pre-processing, normal samples were removed from each dataset, and genes and samples existing in only one dataset were discarded. Additionally, genes lacking variance between samples and samples without clinical information were removed from the datasets. After the pre-processing step, each dataset contained 18,981 genes and 633 samples, which consisted of 417 early stage samples and 216 late-stage samples, respectively. Descriptions of the datasets after pre-processing are provided in Table 4.

Table 4 - Details of the datasets after pre-processing

Category	Description
Data source	The Broad GDAC Firehose (TCGA)
Data type	Gene expression - Level 3 RNAseqV2 (TPM) DNA methylation - HumanMethylation450K (β value)
Number of genes	18981

Transcripts per million (TPM) was used to measure gene expression level. To represent the DNA methylation level of the genes, the β value from 1500 bp upstream of the transcription start site to the end of the gene was used. The β value describes the ratio of the methylated array intensity to the total array intensity:

$$\beta = M / (M + U + \alpha) \quad (11)$$

where M and U denote the methylated and unmethylated intensities, respectively, and α denotes a constant offset, usually set to 100.

Normalization and rescaling

In the pre-processing step, only genes and samples present in both the gene expression and DNA methylation datasets were selected. Consequently, the sizes of the gene expression and DNA methylation datasets were identical. As the units denoting gene expression and DNA methylation levels differed (TPM and β values, respectively), standardization of the two datasets was essential. First, the Box-Cox transformation was applied to normalize the datasets as follows [46]:

$$y(\lambda) = \begin{cases} \frac{y^\lambda - 1}{\lambda}, & \text{if } \lambda \neq 0 \\ \log(y), & \text{if } \lambda = 0 \end{cases} \quad (12)$$

where λ denotes the regularization parameter, which maximizes the log-likelihood function. Additionally, as the scale, λ , was set differently for each gene, the datasets were rescaled into the range of $[0, 1]$, as follows:

$$f(x) = \frac{x - \min(x)}{\max(x) - \min(x)} \quad (13)$$

As a result, the gene expression and DNA methylation levels were converted into normalized forms of the TPM and β values in the range of $[0, 1]$.

Feature extraction and selection

After data normalization and rescaling, unsupervised feature extraction was applied to the gene expression and DNA methylation datasets to produce the integrated dataset. This was based on the hypothesis that the gene expression and DNA methylation status for each gene are correlated to a certain extent. Well-known algorithms, namely, principal component analysis (PCA) and an autoencoder, were applied as unsupervised feature extraction methods. PCA is a popular feature extraction method used to decompose a multivariate dataset into a set of components that express maximum variance [47]. Dimension reduction to one-dimensional data is achieved by deriving a 1st principal component, which can be described as follows:

$$v_1 = \arg \max \left\{ \frac{v^T (X^T X) v}{v^T v} \right\}, \quad (14)$$

where X and v denote the input data matrix and its eigenvectors, respectively. PCA is usually applied for linear feature extraction, although kernel functions can be applied to identify nonlinear relationships between features [48]. Therefore, a kernel PCA using a radial basis function was applied for the test in addition to linear PCA.

An autoencoder is an unsupervised learning algorithm based on a neural network that equalizes the output values to the inputs [49]. By setting the number of the intermediate nodes to 1, a one-dimensional integrated dataset can be acquired similar to that in PCA. Figure 12 demonstrates the schematic structure of the autoencoder. The goal of the proposed autoencoder was to acquire a one-dimensional representation of the intermediate hidden layer while preserving the

same amount of information as in the original two-dimensional data. To achieve this, the neural network was trained to minimize the mean square error between the input and output data. Here, a rectified linear unit was utilized as the activation function to determine nonlinear relationships.

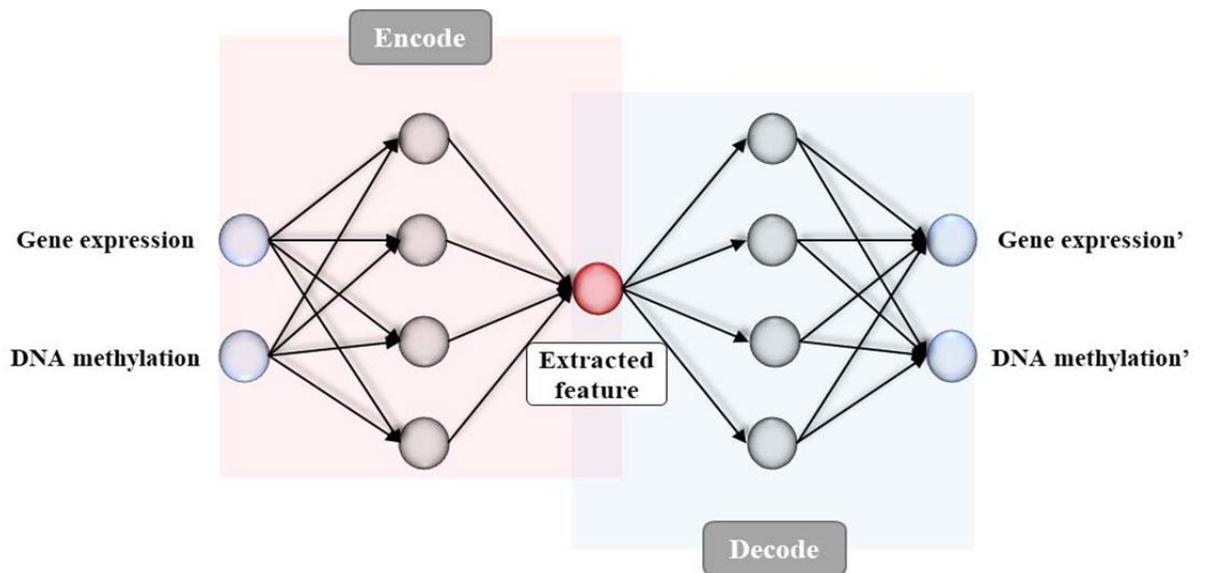


Figure 12 - Schematic structure of the proposed autoencoder

The schematic structure of the autoencoder applied for feature extraction is shown. By setting the number of intermediate nodes to 1 (the red node), a one-dimensional feature set containing the major characteristics of both gene expression and DNA methylation datasets can be acquired.

For feature selection from the integrated dataset, Welch's *t*-test was applied to derive top-ranked genes exhibiting high degrees of differential expression between two classes (the early and late stages) and to identify genes potentially highly correlated with cancer progression and metastasis. Welch's *t*-test is one of the most common methods for differentially expressed genes. Unlike Student's *t*-test,

Welch's t -test does not assume equal variance; therefore, it was suitable for application to the datasets, which had non-uniform sample sizes and variances.

$$t = \frac{\mu_1 - \mu_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}} \quad (15)$$

Here, N , s , and μ denote the size, standard deviation, and mean of the samples, respectively.

Performance test

The performance of the proposed method was measured based on the classification performance according to the RCC stage, using a limited number of top-ranked genes derived from the previous feature selection process. As determining the optimal number of genes required for obtaining the best prediction results is difficult, I applied forward feature selection in the proposed system. By adding genes one by one in rank order, I calculated an AUC score through 10-fold cross-validation testing for the performance evaluation. In addition, I considered the top 20 genes from each dataset and calculated four performance measures, the accuracy, f1 score, Matthews correlation coefficient (MCC), and area under the curve (AUC) using four classifiers, logistic regression, naïve Bayes, random forest, and support vector machine (SVM). This was followed by a literature evaluation in which the selected top-ranked genes were subject to analysis based on the expectation that genes with higher t -values are more highly correlated with cancer progression and metastasis.

Results

Normalization and feature extraction

The RNA-seq and DNA methylation datasets for the various RCCs types were obtained from the Broad GDAC Firehose, as described above. The datasets were normalized through a Box-Cox transformation and then rescaled to the range of [0, 1] for data integration. After normalization, both the gene expression and DNA methylation datasets showed identical value ranges with similar variance. Feature extraction was then applied to these two normalized datasets. Because each gene showed different correlations between the expression and methylation levels, feature extraction was applied to each gene; the number of features was two (gene expression and DNA methylation), with a sample size of 633 (the total number of samples) for each gene. Well-known feature extraction methods (linear PCA, kernel PCA, and an autoencoder) were used as the feature extraction methods. I derived only the 1st principal component after feature extraction for dimension reduction, resulting in the acquisition of an integrated one-dimensional dataset containing combinations of variables with relevant information from both gene expression and DNA methylation results. Here, the average percentage of variance that could be explained by the 1st principal component was 71.3%, which was calculated using the following expression.

$$\frac{\lambda_i}{\sum_{i=1}^n \lambda_i} \quad (16)$$

Here, λ_i denotes the eigenvalue of i th principal component.

For the autoencoder, dimensional reduction was achieved by setting the number of intermediate nodes to 1, as described previously. Tensorflow 1.2 was utilized for implementing the autoencoder [50], and all other feature extraction algorithms were implemented using Python and scikit-learn 0.18.1 [30].

Feature selection

As the structure of the integrated dataset was identical to those of the gene expression and DNA methylation datasets, conventional feature selection methods for differentially expressed gene analysis could be applied directly. I utilized Welch's t -test for each integrated dataset to calculate the feature importance and obtained a list of genes ordered by t -values. Here, I considered only the top 200 genes, a number smaller than the number of late-stage samples, to avoid overfitting. Figure 13 presents Venn diagrams showing the overlap in the genes selected from each dataset. Only three genes exhibited both differential expression and methylation, whereas the integrated datasets shared at least 126 and 25 genes with the gene expression and DNA methylation datasets, respectively. This implied that the integrated dataset successfully reflected the characteristics of the two original datasets. All three integrated datasets shared similar gene sets with 191 genes among all three datasets, because unsupervised feature extractions were applied for only two types of datasets, thereby enabling relatively simple determination of the principal variables. A list of the top 200 genes from each dataset, with their p -values, is given in Table 5.

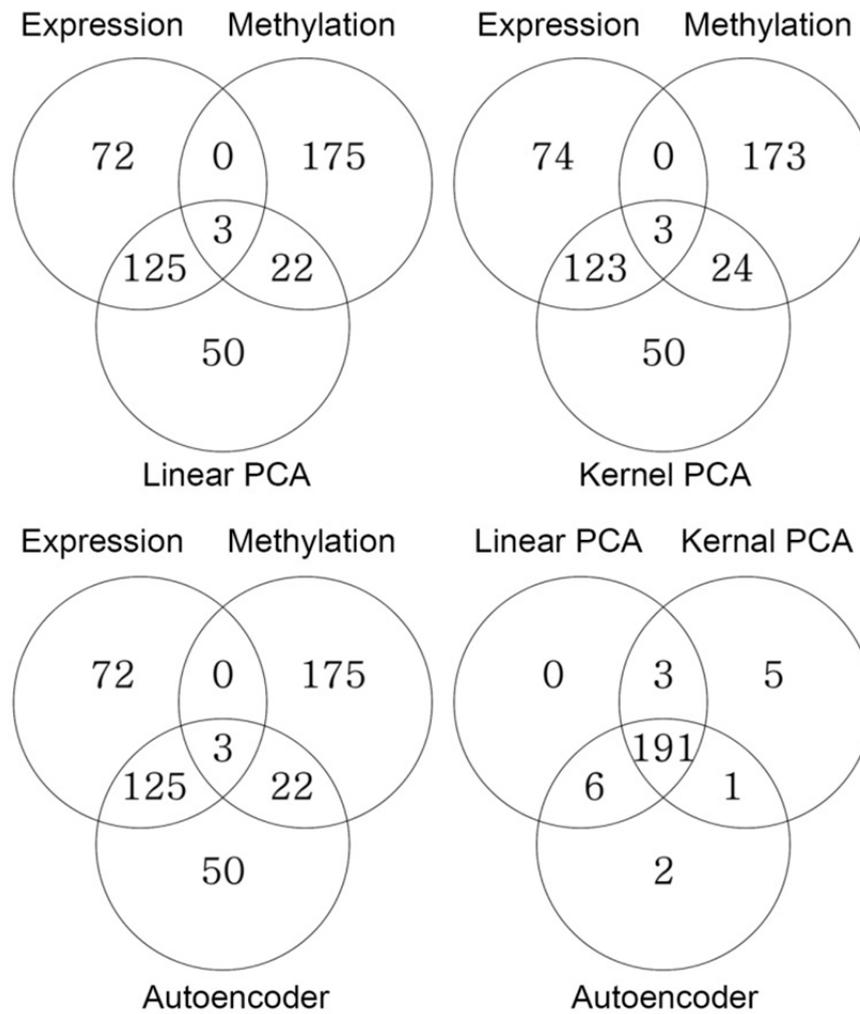


Figure 13 - Venn diagrams of the overlapped features

The Venn diagrams demonstrate the number of genes that overlap between the gene expression, DNA methylation, and three integrated datasets.

Table 5 - List of the top 200 genes from each dataset

Gene Expression		DNA_Methylation		PCA_Linear		PCA_Kernel		Autoencoder	
Gene	P-value	Gene	P-value	Gene	P-value	Gene	P-value	Gene	P-value
KIF20A	1.53E-28	NKX6-2	4.89E-44	NKX6-2	6.77E-43	NKX6-2	1.88E-43	NKX6-2	6.77E-43
UBE2C	2.01E-28	BARHL2	3.12E-42	FERD3L	5.94E-40	FERD3L	2.27E-40	FERD3L	5.94E-40
PTTG1	8.08E-27	TLX3	6.58E-42	BHLHE23	2.36E-33	BHLHE23	1.14E-33	ONECUT3	1.08E-31
HJURP	2.33E-26	SOX1	4.93E-42	ONECUT3	1.08E-31	SIX6	4.71E-33	SOX14	5.33E-31
CENPA	3.28E-26	POU4F2	2.74E-41	SOX14	5.33E-31	ONECUT3	4.52E-32	NKX2-6	7.78E-29
GTSE1	3.57E-26	FERD3L	5.53E-40	NKX2-6	7.78E-29	SOX14	5.43E-31	RRM2	8.61E-29
NCAPG	5.24E-26	PCDH8	2.68E-40	RRM2	8.61E-29	RRM2	1.63E-29	NCAPG	7.96E-29
CDC25C	1.47E-25	TBX20	1.06E-39	NCAPG	7.96E-29	ZNF177	6.48E-29	KIF20A	1.78E-28
NEIL3	2.56E-25	DLX6AS	1.16E-39	KIF20A	1.78E-28	NKX2-6	9.40E-29	BHLHE23	5.25E-28
BIRC5	2.25E-25	NPBWR1	1.74E-38	SIX6	1.48E-27	NCAPG	6.21E-29	SIX6	1.48E-27
SKA1	3.25E-25	ZIC4	1.30E-37	FEZF2	2.26E-27	KIF20A	1.17E-28	FEZF2	2.26E-27
TPX2	6.12E-25	UTF1	1.93E-37	ZNF132	3.11E-27	FEZF2	1.14E-27	ZNF132	3.11E-27
BUB1	1.28E-24	DLX5	3.87E-37	PTTG1	2.07E-27	ZNF132	2.00E-27	PTTG1	2.07E-27
ASPM	1.66E-24	KCNQ1DN	3.82E-37	UBE2C	3.72E-27	PTTG1	8.91E-28	UBE2C	3.72E-27
KIF18B	5.90E-24	POU4F3	9.78E-37	ZNF177	9.85E-27	CSDAP1	2.47E-27	ZNF177	9.85E-27
CEP55	9.14E-24	DIO3	1.90E-36	CSDAP1	1.03E-26	UBE2C	1.07E-27	CSDAP1	1.04E-26
TROAP	2.50E-23	HTR1A	2.57E-36	NEIL3	3.63E-26	RAX	7.66E-27	NEIL3	3.63E-26
CCNB2	2.32E-23	PRLHR	9.57E-37	CENPA	3.06E-26	NEIL3	7.33E-27	CENPA	3.06E-26
OIP5	2.79E-23	INSM1	1.65E-35	NEUROG1	4.94E-26	LBX1	1.66E-26	NEUROG1	4.94E-26
CENPF	9.21E-23	TWIST1	9.33E-36	HJURP	4.65E-26	HJURP	1.80E-26	HJURP	4.64E-26
SKA3	9.66E-23	SPAG6	1.30E-35	GTSE1	3.56E-26	NEUROG1	2.71E-26	GTSE1	3.56E-26
NUF2	1.49E-22	GBX2	6.18E-35	LBX1	1.39E-25	CENPA	2.01E-26	LBX1	1.36E-25
KIF14	2.57E-22	HMX3	3.83E-34	CDC25C	9.36E-26	GTSE1	2.89E-26	CDC25C	9.36E-26
DEPDC1	2.53E-22	BHLHE23	3.93E-33	WDR8	3.81E-25	WDR8	8.63E-26	WDR8	3.81E-25
CDC20	3.58E-22	SIX6	4.55E-34	BIRC5	4.16E-25	CDC25C	7.38E-26	BIRC5	4.16E-25
KIF2C	5.09E-22	ZIC1	4.19E-34	SKA1	5.29E-25	BIRC5	2.00E-25	SKA1	5.30E-25
CDCA5	9.95E-22	CACNG8	5.19E-34	TPX2	8.05E-25	SKA1	3.19E-25	TPX2	8.05E-25
NDC80	7.86E-22	SLC32A1	1.81E-33	PUS3	1.33E-24	TPX2	5.03E-25	PUS3	1.33E-24
AURKB	1.55E-21	DSC3	2.09E-33	ASPM	1.72E-24	PUS3	7.95E-25	BUB1	1.76E-24
CDC42BPG	2.07E-21	CBLN1	1.94E-33	BUB1	1.76E-24	ASPM	1.08E-24	KIF18B	3.11E-24
DLGAP5	3.09E-21	HAND2	2.19E-33	KIF18B	3.11E-24	KIF18B	2.59E-24	CEP55	7.15E-24
KIFC1	3.36E-21	PENK	1.71E-33	CEP55	7.22E-24	CEP55	1.92E-24	TROAP	1.03E-23
FOXM1	4.99E-21	SLC18A3	4.13E-33	TROAP	1.03E-23	BUB1	1.83E-24	KIFC1	8.68E-24
CDK1	3.99E-21	DMRTA2	1.06E-32	KIFC1	8.68E-24	TROAP	5.14E-24	OIP5	9.58E-24
BUB1B	4.51E-21	TLX2	1.23E-32	OIP5	9.58E-24	KIFC1	4.46E-24	CDC20	5.29E-23
NEK2	6.49E-21	GDF6	1.13E-32	CDC20	5.29E-23	FGF4	1.13E-23	SHOX2	2.08E-22
RRM2	9.30E-21	INA	3.81E-32	SHOX2	2.08E-22	GPR149	1.62E-23	CDK1	6.31E-23
MKI67	7.85E-21	ONECUT3	5.84E-32	CDK1	6.31E-23	OIP5	7.92E-24	CENPF	8.69E-23
TTK	1.46E-20	RIPPLY2	1.09E-31	CENPF	8.69E-23	CDC20	1.93E-23	SCGB3A1	1.24E-22
KIF18A	1.42E-20	TBX5	1.24E-31	SCGB3A1	1.23E-22	CDK1	3.50E-23	NUF2	1.28E-22
PBK	1.45E-20	DLX6	1.35E-31	NUF2	1.28E-22	SHOX2	1.58E-22	SKA3	1.13E-22
TOP2A	2.06E-20	NBLA00301	1.53E-31	SKA3	1.12E-22	SCGB3A1	6.09E-23	CCNB2	1.49E-22
KIF11	2.31E-20	PTF1A	2.29E-31	CCNB2	1.49E-22	CENPF	7.20E-23	KIF14	2.59E-22
CKAP2L	4.96E-20	LHX5	4.31E-31	KIF14	2.59E-22	SKA3	7.56E-23	KIF2C	3.63E-22
CDCA8	7.32E-20	DBX1	3.22E-31	KIF2C	3.63E-22	NUF2	1.09E-22	KRT7	1.78E-22
PLK1	1.02E-19	TRH	5.65E-31	KRT7	1.78E-22	CCNB2	1.41E-22	DEPDC1	5.57E-22
CCNA2	8.64E-20	GSC	4.59E-31	DEPDC1	5.57E-22	KIF14	2.28E-22	CDCA5	9.45E-22
HMMR	1.29E-19	SOX14	5.17E-31	CDCA5	9.45E-22	DEPDC1	2.48E-22	AURKB	1.91E-21
PRC1	1.50E-19	PRDM13	4.15E-31	AURKB	1.91E-21	KIF2C	2.62E-22	ASPM	3.10E-21
SHOX2	6.79E-19	PCSK1	3.18E-30	AIF1L	9.83E-22	KRT7	1.30E-22	C16orf86	2.44E-21
POLQ	5.39E-19	FOXD3	3.26E-30	C16orf86	2.44E-21	C16orf86	4.57E-22	CDC42BPG	1.75E-21
GPR19	5.27E-19	CDX2	4.17E-30	CDC42BPG	1.75E-21	CDCA5	5.77E-22	DLGAP5	3.66E-21
CDKN3	6.43E-19	CIDEA	3.91E-30	DLGAP5	3.66E-21	AURKB	1.25E-21	BUB1B	4.76E-21
MACC1	3.55E-19	gg	4.60E-30	BUB1B	4.76E-21	AIF1L	6.71E-22	FOXM1	5.96E-21
MYBL2	9.16E-19	FBLN1	4.32E-30	FOXM1	5.96E-21	CDC42BPG	1.25E-21	NEK2	7.10E-21
C15orf42	1.63E-18	PAX1	8.34E-30	NEK2	7.10E-21	TRIM58	1.35E-21	MKI67	7.60E-21
NUSAP1	9.97E-19	NHLH2	2.98E-29	MKI67	7.60E-21	DLGAP5	2.43E-21	MYBL2	1.56E-20
C1orf210	1.33E-18	LOC100192379	1.96E-29	MYBL2	1.56E-20	TMEM25	4.81E-21	CDCA8	1.22E-20
KIAA0101	1.32E-18	HMX2	2.48E-29	CDCA8	1.22E-20	FOXM1	4.38E-21	PBK	1.11E-20
ZNF132	2.13E-18	EVX2	1.73E-29	PBK	1.11E-20	BUB1B	3.97E-21	TMEM25	2.45E-20
FAM83D	1.52E-18	HBM	6.99E-29	TMEM25	2.45E-20	PAX6	5.83E-21	PLK1	2.41E-20
CDCA3	3.07E-18	ADCYAP1	3.31E-29	PLK1	2.41E-20	NEK2	5.36E-21	RAX	4.35E-20
KIF15	2.40E-18	NEFH	3.02E-29	RAX	4.35E-20	CDCA8	7.09E-21	TWIST1	1.32E-20
IQGAP3	2.80E-18	ALX1	4.55E-29	TWIST1	1.32E-20	MYBL2	1.25E-20	KIF18A	1.90E-20
CYB5D2	1.72E-18	PITX1	5.05E-29	KIF18A	1.90E-20	MKI67	7.53E-21	TOP2A	1.71E-20
ASF1B	3.28E-18	LBXCOR1	8.37E-29	TOP2A	1.71E-20	PBK	8.95E-21	CCNA2	1.96E-20
C16orf86	2.88E-18	NKX2-6	1.19E-28	CCNA2	1.96E-20	PLK1	2.09E-20	TTK	3.09E-20
EXO1	5.24E-18	MKX	2.34E-28	TTK	3.09E-20	KIF18A	1.45E-20	TRIM58	3.01E-20
TRIP13	1.23E-17	GPR83	1.74E-28	TRIM58	3.01E-20	TOP2A	1.27E-20	PAX6	4.41E-20
TMEM25	1.59E-17	CSDAP1	1.55E-28	PAX6	4.41E-20	TNFSF13	1.36E-20	TNFSF13	3.41E-20
KRT7	9.04E-18	LHX8	2.04E-28	ATOH1	3.80E-20	CCNA2	1.49E-20	MACC1	5.36E-20
IGF2BP3	4.02E-17	IRF4	5.97E-28	TNFSF13	3.41E-20	ATOH1	2.86E-20	SNRPF	1.38E-19
DEPDC1B	1.94E-17	FZD10	2.64E-28	MACC1	5.37E-20	PLCD1	1.91E-20	CPD	1.24E-19
SPC25	2.29E-17	RAX	3.66E-28	SNRPF	1.38E-19	TTK	3.40E-20	GPR19	2.02E-19
CCNB1	3.22E-17	NKX2-4	3.18E-28	CPD	1.24E-19	TWIST1	2.67E-20	ZNF582	2.74E-19
PRR15L	2.21E-17	EVX1	5.54E-28	GPR19	2.02E-19	SNRPF	1.02E-19	HMMR	3.03E-19
PLCD1	2.90E-17	ADRA1D	4.83E-28	ZNF582	2.74E-19	MACC1	6.33E-20	PLCD1	2.09E-19
DNAJC28	3.87E-17	GPR149	6.34E-28	HMMR	3.03E-19	ZNF582	1.52E-19	PRC1	2.74E-19
ANLN	6.30E-17	ZNF132	4.18E-28	PLCD1	2.09E-19	GPR19	1.58E-19	POLQ	5.68E-19
IL20RB	1.21E-16	OTP	4.83E-28	PRC1	2.74E-19	CPD	1.29E-19	BSX	7.75E-19
E2F2	2.20E-16	POU3F1	9.86E-28	POLQ	5.68E-19	HMMR	2.21E-19	NUSAP1	9.82E-19
NUDT6	2.17E-16	IRX2	5.18E-28	BSX	7.75E-19	PRC1	1.83E-19	PRR15L	8.12E-19

Table 5 - List of the top 200 genes from each dataset (Continued)

Gene Expression		DNA_Methylation		PCA_Linear		PCA_Kernel		Autoencoder	
Gene	P-value	Gene	P-value	Gene	P-value	Gene	P-value	Gene	P-value
FAM47E	2.71E-16	FOXE1	9.04E-28	NUSAP1	9.82E-19	POLQ	3.69E-19	KIAA0101	1.25E-18
NOP16	3.24E-16	MYOD1	1.34E-27	PRR15L	8.12E-19	RAB6C	1.05E-18	RAB6C	2.44E-18
SGOL1	5.93E-16	PRDM12	1.23E-27	KIAA0101	1.25E-18	NUSAP1	4.40E-19	CDCA3	2.13E-18
AMT	5.17E-16	FEZF2	1.30E-27	RAB6C	2.44E-18	KIAA0101	6.05E-19	FAM83D	1.47E-18
C1orf172	4.40E-16	GHSR	7.49E-28	CDCA3	2.13E-18	SLC6A5	1.20E-18	IQGAP3	2.53E-18
ZIC2	1.99E-15	PITX2	1.28E-27	FAM83D	1.48E-18	CDCA3	1.39E-18	GASS	1.23E-18
NCAPH	7.81E-16	IRX4	1.06E-27	IQGAP3	2.53E-18	BSX	1.14E-18	ASF1B	2.93E-18
ESCO2	1.07E-15	USP44	6.90E-28	GAS5	1.23E-18	GAS5	6.24E-19	C15orf42	3.37E-18
FAM64A	1.18E-15	NKAPL	8.70E-28	ASF1B	2.93E-18	IQGAP3	1.85E-18	ZIC2	1.47E-17
NMNAT1	9.09E-16	ACTA1	2.12E-27	C15orf42	3.37E-18	FAM83D	1.17E-18	CYB5D2	2.17E-18
UHRF1	1.13E-15	ZNF177	2.23E-27	ZIC2	1.47E-17	CYB5D2	1.19E-18	PHB	3.59E-18
ZSCAN18	1.21E-15	SOX17	3.78E-27	CYB5D2	2.17E-18	PRR15L	1.25E-18	EXO1	5.30E-18
CENPE	1.83E-15	CCDC140	2.11E-27	PHB	3.59E-18	ASF1B	2.17E-18	MIAT	8.87E-18
MCM10	2.25E-15	ALX4	4.61E-27	EXO1	5.30E-18	ZIC2	1.06E-17	ZNF471	5.64E-18
E2F7	2.36E-15	CPXM1	5.64E-27	MIAT	8.88E-18	MIAT	3.83E-18	C1orf210	7.08E-18
MELK	2.33E-15	ADAMTS20	6.52E-27	ZNF471	5.64E-18	PHB	2.66E-18	KIF15	9.10E-18
CXCL13	2.42E-15	TLX1	7.15E-27	C1orf210	7.08E-18	C15orf42	3.48E-18	FAM64A	1.25E-17
LMNB1	2.77E-15	POU4F1	9.18E-27	KIF15	9.10E-18	CENPM	3.12E-18	SLC6A5	1.73E-17
C19orf46	1.83E-15	COMP	1.01E-26	FAM64A	1.25E-17	EXO1	4.24E-18	ZSCAN18	8.26E-18
PDCD5	2.91E-15	VSX1	1.12E-26	SLC6A5	1.73E-17	C1orf210	5.10E-18	ZNF542	1.41E-17
MYL3	2.00E-15	LBX1	1.27E-26	ANLN	1.11E-17	CDKN3	5.12E-18	CENPM	1.22E-17
FGF5	6.32E-15	C1QL2	1.30E-26	ZSCAN18	8.26E-18	ZSCAN18	4.37E-18	CDKN3	1.57E-17
KCNGB1	4.83E-15	FOXE3	8.70E-27	ZNF542	1.41E-17	ZNF542	7.99E-18	KIAA1755	1.68E-17
PKMYT1	4.88E-15	C5orf38	6.94E-27	CENPM	1.23E-17	KIF15	8.83E-18	SPC25	1.68E-17
TNFSF13	2.94E-15	EOMES	8.26E-27	CDKN3	1.57E-17	ZNF471	8.23E-18	CCNB1	2.32E-17
PLEKHA7	4.50E-15	TBX1	5.80E-27	KIAA1755	1.68E-17	FAM64A	1.38E-17	EMX2	3.08E-17
ESPL1	5.35E-15	ISL1	1.49E-26	SPC25	1.68E-17	ANLN	9.82E-18	DEPDC1B	2.94E-17
FAM72B	6.37E-15	C14orf23	8.94E-27	CCNB1	2.32E-17	EMX2	1.35E-17	NDC80	2.83E-17
MMP15	4.40E-15	C20orf56	1.36E-26	EMX2	3.08E-17	CCNB1	1.54E-17	IGF2BP3	8.30E-17
NICN1	4.85E-15	CR1	2.05E-26	DEPDC1B	2.94E-17	SPC25	1.47E-17	OR56A3	5.98E-17
PRRG2	4.26E-15	SHOX2	2.28E-26	NDC80	2.83E-17	NDC80	1.66E-17	SMC4	4.78E-17
FAM54A	6.88E-15	ZNF560	1.85E-26	IGF2BP3	8.30E-17	DEPDC1B	1.96E-17	HBA1	6.34E-17
HSD17B8	8.32E-15	PRDM8	2.56E-26	OR56A3	6.00E-17	SMC4	2.38E-17	CKAP2L	4.33E-17
FAM54B	7.97E-15	FOXL2	5.60E-26	SMC4	4.78E-17	HBA1	3.66E-17	NKX1-2	1.64E-16
C10orf81	1.33E-14	SLITRK1	3.73E-26	HBA1	6.34E-17	IGF2BP3	8.36E-17	NCAPH	1.25E-16
CAMK2N2	9.75E-15	NEUROG1	4.71E-26	CKAP2L	4.33E-17	OR56A3	4.99E-17	CASC4	8.82E-17
C9orf117	5.35E-15	C17orf104	3.99E-26	NKX1-2	1.64E-16	KIAA1755	4.96E-17	FGF5	4.11E-16
ARHGAP11A	9.25E-15	ALX3	4.72E-26	NCAPH	1.25E-16	CKAP2L	4.17E-17	E2F2	2.40E-16
CIQL1	1.06E-14	NKX2-5	1.10E-25	CASC4	8.82E-17	NKX1-2	1.13E-16	C19orf46	1.61E-16
EMX2	1.33E-14	ZAR1	7.12E-26	FGF5	4.11E-16	EPHA5	1.08E-16	IL20RB	4.05E-16
TCTA	1.08E-14	SCGB3A1	1.28E-25	E2F2	2.40E-16	IL20RB	1.07E-16	RPS6	1.91E-16
SHCBP1	1.29E-14	PHOX2B	8.25E-26	C19orf46	1.61E-16	CASC4	6.46E-17	NUDT6	2.74E-16
SPRYD3	1.34E-14	VSTM2B	2.04E-25	IL20RB	4.05E-16	NCAPH	1.30E-16	MED1	3.47E-16
FAM72D	1.95E-14	WDR8	1.26E-25	RPS6	1.91E-16	HSD17B8	1.79E-16	HSD17B8	3.83E-16
TRIOBP	1.34E-14	ASCL2	1.22E-25	NUDT6	2.74E-16	NUDT6	1.40E-16	TBX18	4.40E-16
SPC24	2.25E-14	NKX2-1	2.67E-25	MED1	3.47E-16	E2F2	1.97E-16	SGOL1	6.00E-16
RAD51	1.88E-14	PHOX2A	1.71E-25	HSD17B8	3.83E-16	FGF5	4.89E-16	CAMK2N2	6.05E-16
HDAC11	1.69E-14	VAX1	2.80E-25	TBX18	4.48E-16	RPS6	1.40E-16	C1orf172	4.17E-16
CLDN7	2.04E-14	PUS3	2.61E-25	SGOL1	6.00E-16	C19orf46	1.53E-16	ZNF586	7.63E-16
PHYHD1	3.03E-14	RSP02	2.76E-25	CAMK2N2	6.05E-16	MED1	3.55E-16	MYO5B	9.01E-16
PAEP	6.77E-14	EN2	2.51E-25	C1orf172	4.17E-16	CAMK2N2	4.62E-16	CXCL13	9.57E-16
TTC9	2.86E-14	ICAM5	3.76E-25	ZNF586	7.63E-16	ZNF586	4.80E-16	NICN1	7.11E-16
FKBP10	4.11E-14	CYP26A1	3.58E-25	MYO5B	9.01E-16	TBX18	6.07E-16	FAM72B	1.02E-15
PRELID1	4.99E-14	NEUROD1	3.13E-25	CXCL13	9.57E-16	C1orf172	4.27E-16	LMNB1	9.79E-16
GNP7	5.29E-14	SLFN12L	3.63E-25	NICN1	7.11E-16	LMNB1	5.83E-16	FOXA1	2.24E-15
ICA1	4.96E-14	VENTX	4.36E-25	FAM72B	1.02E-15	SGOL1	7.68E-16	ZSCAN21	9.39E-16
EPR1	6.92E-14	ZIC5	4.42E-25	LMNB1	9.79E-16	NICN1	5.29E-16	ESCO2	1.16E-15
EIF4EBP1	5.70E-14	ISLR2	4.15E-25	FOXA1	2.24E-15	ZSCAN21	6.19E-16	MELK	1.58E-15
KIF1C	4.50E-14	NLXPH1	1.07E-24	ZSCAN21	9.39E-16	NMNAT1	7.51E-16	OVOL1	1.17E-15
CCNF	8.20E-14	KCNA7	1.10E-24	ESCO2	1.16E-15	TRIP13	9.97E-16	PLEKHA7	1.44E-15
WDR62	7.83E-14	ERN2	6.52E-25	MELK	1.58E-15	CCDC71	6.59E-16	DUT	1.31E-15
ATAD2	7.33E-14	EN1	1.17E-24	OVOL1	1.17E-15	MAL	8.54E-16	UHRF1	1.71E-15
DTL	6.89E-14	GPR6	5.09E-25	PLEKHA7	1.44E-15	MYO5B	1.03E-15	MAL	1.70E-15
C17orf108	7.09E-14	FOXA2	9.51E-25	DUT	1.31E-15	MELK	1.16E-15	E2F7	2.08E-15
BAG1	1.08E-13	psiTPTE22	1.13E-24	UHRF1	1.71E-15	PHYHD1	1.07E-15	RAB20	1.69E-15
CENPM	9.16E-14	ALDH1A3	7.00E-25	MAL	1.70E-15	FAM72B	1.25E-15	CENPE	1.99E-15
FLJ32063	5.42E-14	WNT1	1.47E-24	E2F7	2.08E-15	DUT	9.96E-16	NMNAT1	1.79E-15
PMCH	1.14E-13	STAG3	1.25E-24	RAB20	1.69E-15	C16orf48	1.30E-15	KCNGB1	2.79E-15
FOXA1	2.32E-13	ZNF542	1.31E-24	CENPE	1.99E-15	RAB20	1.19E-15	CCDC71	1.65E-15
PRRX2	1.28E-13	HOXD13	1.85E-24	NMNAT1	1.79E-15	UHRF1	1.46E-15	NOP16	1.80E-15
KIAA1324	1.54E-13	GALR1	2.64E-24	KCNGB1	2.79E-15	ESCO2	1.54E-15	PHYHD1	2.27E-15
SLC44A4	1.09E-13	SHE	2.15E-24	CCDC71	1.65E-15	PLEKHA7	1.39E-15	OTOL1	2.32E-15
HMGCL	1.42E-13	HOXC13	1.57E-24	NOP16	1.80E-15	NOP16	1.22E-15	MCM10	2.60E-15
WSP2	1.88E-13	PPP1R14A	3.16E-24	PHYHD1	2.27E-15	KCNGB1	2.04E-15	TRIP13	2.83E-15
PAFAH2	1.55E-13	FOXF1	1.24E-24	OTOL1	2.32E-15	FOXA1	3.71E-15	NCRNA00175	1.50E-15
TRIB3	1.87E-13	BARX1	2.15E-24	MCM10	2.60E-15	CXCL13	1.86E-15	CIQL1	3.04E-15
RAD54L	2.02E-13	LOC100128811	3.24E-24	TRIP13	2.83E-15	CENPE	1.83E-15	ICA1	3.19E-15
C1orf190	1.80E-13	CHAT	4.76E-24	NCRNA00175	1.50E-15	OVOL1	1.63E-15	PRRG2	2.12E-15
GAPDH	2.04E-13	KRT7	2.87E-24	CIQL1	3.04E-15	E2F7	2.52E-15	TOX3	2.33E-15
OVOL1	1.72E-13	NKX3-2	3.90E-24	ICA1	3.19E-15	ICA1	2.36E-15	MMP15	2.67E-15
OLFM2B	2.03E-13	GALR2	5.91E-24	PRRG2	2.12E-15	MMP15	1.84E-15	PRDM8	4.60E-15
E2F8	2.58E-13	EYA4	6.37E-24	TOX3	2.33E-15	PTH2	3.77E-15	FAM54A	4.00E-15

Table 5 - List of the top 200 genes from each dataset (continued)

Gene Expression		DNA_Methylation		PCA_Linear		PCA_Kernel		Autoencoder	
Gene	P-value	Gene	P-value	Gene	P-value	Gene	P-value	Gene	P-value
EMX2OS	3.00E-13	NKX6-1	8.00E-24	MMP15	2.67E-15	CD47	3.20E-15	PTH2	5.56E-15
FBXO43	3.56E-13	TMEM155	8.44E-24	PRDM8	4.59E-15	TOX3	2.14E-15	KIAA1024	4.49E-15
CCNE2	2.84E-13	HAND1	7.86E-24	FAM54A	4.00E-15	NCRNA00175	1.85E-15	PITX2	6.55E-15
SPAG7	2.32E-13	FOXB2	6.98E-24	PTH2	5.56E-15	C1QL1	3.12E-15	C16orf48	5.00E-15
MLF1IP	3.32E-13	TCF15	1.10E-23	KIAA1024	4.49E-15	ESPL1	3.31E-15	CD47	5.28E-15
LMO7	3.09E-13	SOX8	1.26E-23	PITX2	6.47E-15	RPL12	2.53E-15	KRTAP25-1	5.14E-15
MTHFD2	3.68E-13	RXFP3	6.19E-24	C16orf48	5.00E-15	MCM10	4.09E-15	C19orf71	4.16E-15
ZNF433	3.67E-13	LOC440040	8.09E-24	CD47	5.28E-15	FAM54A	4.14E-15	OR51B6	5.81E-15
AIF1L	3.91E-13	PAX3	1.03E-23	KRTAP25-1	5.14E-15	CENPJ	4.83E-15	ESPL1	5.74E-15
PRRX1	4.06E-13	SCR1	1.74E-23	C19orf71	4.16E-15	PRRG2	2.92E-15	PAEP	1.34E-14
ZNF629	4.22E-13	ISL2	1.24E-23	OR51B6	5.82E-15	C7orf41	4.69E-15	TRIM36	6.94E-15
GLS2	4.95E-13	OLIG2	2.51E-23	ESPL1	5.74E-15	KIAA1024	5.44E-15	LOC643387	8.76E-15
TBX18	5.48E-13	ONECUT1	2.47E-23	PAEP	1.34E-14	KRTAP25-1	5.43E-15	ZIC5	2.07E-14
CDC6	5.24E-13	S1PR5	2.68E-23	TRIM36	6.95E-15	PRDM8	6.51E-15	EEF1B2	7.90E-15
FAM128A	5.50E-13	PDX1	2.94E-23	LOC643387	8.76E-15	C19orf71	4.43E-15	RPL12	6.13E-15
LOC388955	5.63E-13	DLL3	4.55E-23	ZIC5	2.08E-14	EEF1B2	5.52E-15	ARHGAP11A	8.31E-15
TJP2	4.71E-13	LHX6	4.20E-23	EEF1B2	7.91E-15	TRIM36	6.81E-15	ZNF577	8.19E-15
C5orf46	6.20E-13	SLC12A5	3.97E-23	RPL12	6.13E-15	OTOL1	5.85E-15	KIAA1324	1.06E-14
OXSM	5.64E-13	ECEL1	6.45E-23	ARHGAP11A	8.31E-15	PITX2	9.02E-15	MFSD4	9.80E-15
IGSF22	5.95E-13	CENPV	3.61E-23	ZNF577	8.19E-15	PAEP	1.53E-14	CENPJ	1.28E-14
CHL1	4.97E-13	GJD2	8.82E-23	KIAA1324	1.06E-14	ZFP28	7.92E-15	CMA1	6.73E-15
C17orf53	7.30E-13	PAX6	4.73E-23	MFSD4	9.80E-15	KIAA1949	7.13E-15	KIAA1949	1.22E-14
ST7OT1	6.90E-13	DDX25	5.35E-23	CENPJ	1.28E-14	ARHGAP11A	6.37E-15	ZFP28	1.49E-14
FN1	6.66E-13	FEZF1	5.09E-23	CMA1	6.73E-15	LOC643387	8.84E-15	C7orf41	1.31E-14
MARVELD2	8.15E-13	VGLL2	8.28E-23	KIAA1949	1.22E-14	MFSD4	8.06E-15	CDS1	1.34E-14
FUCA1	8.65E-13	SLC8A2	9.71E-23	ZFP28	1.49E-14	FAM54B	9.50E-15	TACSTD2	1.14E-14
ACAD8	9.02E-13	CYP26C1	4.56E-23	C7orf41	1.31E-14	KIAA1324	9.25E-15	PMCH	1.60E-14
SMCR7	7.04E-13	FGF4	1.25E-22	CDS1	1.34E-14	OR51B6	9.85E-15	ITPKA	1.87E-14
KIAA0649	8.91E-13	SLC2A14	7.33E-23	TACSTD2	1.14E-14	ZIC5	2.52E-14	FAM72D	1.92E-14
RAET1K	1.15E-12	LYPD5	6.71E-23	PMCH	1.60E-14	MYL3	7.62E-15	AMT	1.70E-14
TYRO3	9.39E-13	SLC22A16	7.43E-23	ITPKA	1.86E-14	CDS1	9.27E-15	FAM54B	2.01E-14
GLOD4	7.22E-13	IGF2	1.13E-22	FAM72D	1.92E-14	CMA1	7.22E-15	SPC24	2.29E-14
DNAAJ4	1.02E-12	HS3ST3B1	7.30E-23	AMT	1.69E-14	ZNF577	1.12E-14	RAD51	1.99E-14
DYNC2LI1	7.67E-13	KCNC2	2.62E-22	FAM54B	2.02E-14	SHCBP1	1.30E-14	SHCBP1	2.09E-14
MAL	1.03E-12	PRDM6	1.30E-22	SPC24	2.29E-14	FNDC3A	1.54E-14	GNB2L1	1.83E-14
CTHRC1	1.11E-12	SPTBN4	1.47E-22	RAD51	1.99E-14	PMCH	1.51E-14	C9orf117	1.29E-14

Performance evaluation

After feature selection, a 10-fold cross-validation test was conducted for performance evaluation. First, I applied a forward feature selection approach for the test, as classification performance can differ according to the number of genes used. Starting from one gene, I continuously conducted 10-fold cross-validation by successively adding genes in the rank order acquired from the Welch's t-test results. Here, SVM was used as a classifier and the AUC was taken as the measurement for the cross-validation score. The test results are presented in Figure 14. Integrative analysis using feature extraction generally exhibited superior performance over the approach using gene expression or DNA methylation data alone regardless of the number of selected genes. This implies that additional genes of interest can be obtained from integrated datasets. In particular, use of gene expression data only for the analysis, as is the case for most conventional methods, resulted in considerably lower classification performance than the other approaches. In addition, I considered the top 20 genes from each dataset for the test. I utilized four classifiers to calculate four performance measures, as described in the previous section. The detailed results of the performance evaluation are listed in Table 6. As is apparent from the table, the integrated datasets showed an overall better performance as compared to the single dataset for all four classifiers and most performance indices. Further, use of the gene expression or DNA methylation dataset alone failed to yield cancer stage classification when logistic regression and SVM were employed, whereas the integrated datasets demonstrated overall consistency in their scores regardless of classifier. The differences among the integrated datasets were minute, as the genes included in each dataset were very

similar and differed only slightly in their ranking orders. These findings imply that the proposed integrative analysis can be utilized for better discovery of candidate cancer biomarkers.

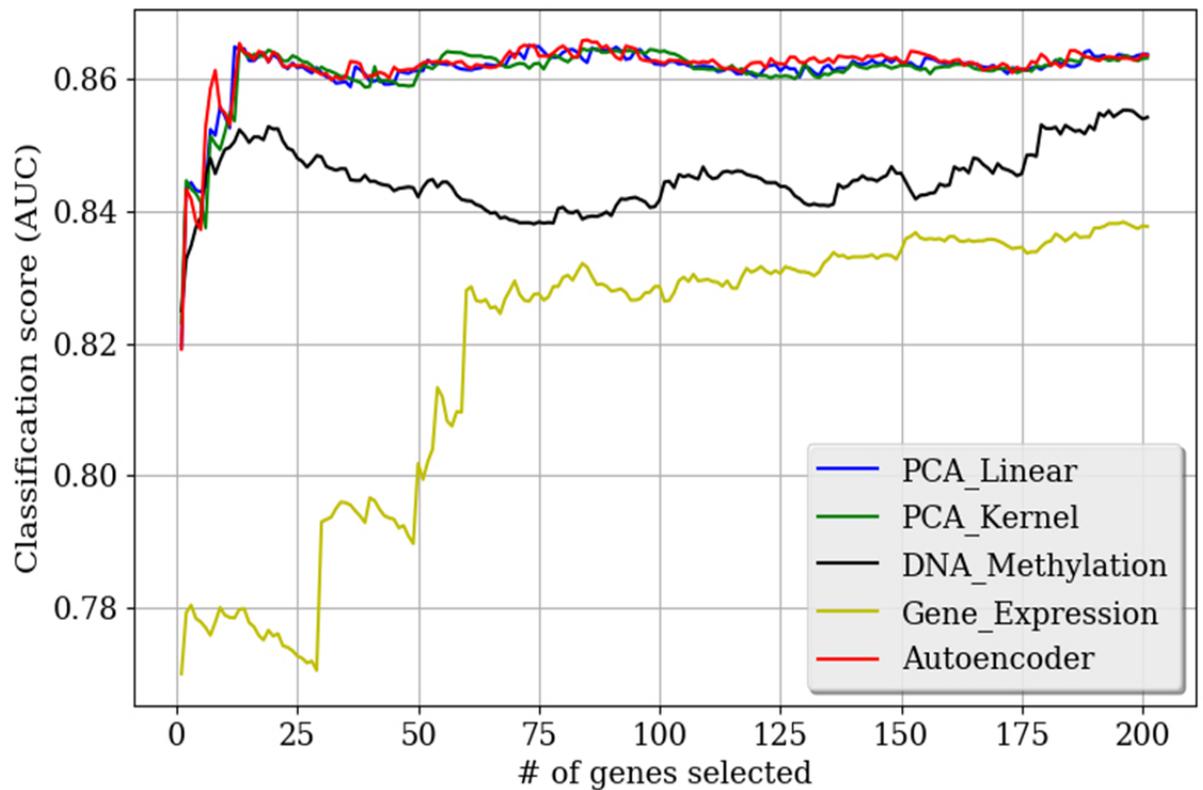


Figure 14 - Classification performance according to 10-fold cross-validation
The results of 10-fold cross-validation using an SVM. The X- and Y-axes denote the number of genes selected through forward feature selection and 10-fold cross-validation classification score (AUC), respectively. The overall best classification score was achieved by the autoencoder using the top 12 genes, with an AUC of 0.866.

Table 6 – Result of 10-fold cross-validation using top 20 genes

The results of performance tests using the top 20 genes only. The items yielding the best scores are highlighted in bold.

Classifier	Performance Measure	Gene Expression	DNA Methylation	Linear PCA	Kernel PCA	Autoencoder
Logistic Regression	Accuracy	0.341	0.341	0.786	0.789	0.781
	F1-score	0.59	0.509	0.707	0.712	0.702
	MCC	0	0	0.568	0.573	0.56
	AUC	0.776	0.853	0.864	0.867	0.864
Naïve Bayes	Accuracy	0.692	0.784	0.79	0.786	0.792
	F1-score	0.588	0.701	0.696	0.687	0.7
	MCC	0.373	0.552	0.559	0.545	0.563
	AUC	0.774	0.852	0.86	0.859	0.857
Random Forest	Accuracy	0.72	0.773	0.787	0.783	0.784
	F1-score	0.5	0.637	0.653	0.658	0.662
	MCC	0.346	0.486	0.523	0.513	0.512
	AUC	0.737	0.829	0.846	0.841	0.845
SVM	Accuracy	0.562	0.562	0.797	0.789	0.803
	F1-score	0.152	0.152	0.709	0.708	0.72
	MCC	0	0	0.576	0.571	0.588
	AUC	0.776	0.852	0.864	0.865	0.864

The top-ranked genes selected by Welch's *t*-test in the original gene expression (*KIF20A*), DNA methylation, and integrated datasets (*NKX6-2*) showed high likelihoods of being cancer biomarkers [51,52]. Similarly, the three genes that were differentially expressed (*SHOX2*, *ZNF132*, and *KRT7*) were also previously reported as cancer biomarkers [53-55]. Further, low-ranked genes in the gene expression and DNA methylation datasets that were highly ranked in the integrated dataset should be considered as potential novel candidate biomarkers. For example, *LBX1* (ranked 200+ in the gene expression dataset, 103 in the DNA methylation dataset, and 21 in the integrated datasets) was previously identified as a biomarker highly related to cancer progression [56]. Therefore, I focused on the top 100 genes from the integrated datasets with ranks of 200+ in both the gene expression and DNA methylation datasets, finding that their ranks improved in the integrated datasets [*TRIM58* (64), *SNRPF* (74), *CPD* (76), *ZNF582* (77), *BSX* (84), *RAB6C* (85), *GAS5* (89), *PHB* (95), *MIAT* (96), *SLC6A5* (96), and *ZNF471* (100); the numbers in parentheses denote the average ranks in the three integrated datasets]. Among these 11 genes, *TRIM58*, *ZNF582*, *RAB6C*, *GAS5*, *PHB*, and *MIAT* were previously reported as cancer-related genes [57-62]. These findings support the efficacy of the proposed method by utilizing integrative analysis of gene expression and DNA methylation data. Further, these results imply that differentially expressed genes found in the integrated dataset and not previously reported as being related to cancer progression (*SNRPF*, *CPD*, *BSX*, *SLC6A5*, *ZNF471*, and other top genes in the integrated datasets) may constitute strong potential candidates as cancer-progression biomarkers.

Conclusion

In this thesis, I first presented a novel feature selection method based on the application of the l_1 -norm SVM over data perturbation, which was shown to be efficient for biomarker discovery. The nature of the l_1 -norm, which generates a sparse solution, renders it a reasonably efficient method of feature selection for high-dimensional data. The l_1 -norm SVM is also suitable for biomarker selection, as it delivers high performance in terms of classification and is applicable to diverse situations. However, application of the l_1 -norm SVM to a single dataset generates difficulty with regard to the detection of closely correlated factors, which commonly appear in biomarker detection. In addition, a result that is subordinate to a certain dataset may be produced. In the experiments, the feature stability was successfully improved because the l_1 -norm SVM was applied to several bootstrap samples considering instance perturbation. Instead of using the general SVM ranking criteria, I considered only the number of bootstrap samples that contained a given selected feature as the stability measure of that feature. By applying backward feature elimination based on the proposed stability score, I could then determine the optimal subset of features for which good classification performance was obtained. I applied the approach to RNA-seq data of renal clear cell carcinoma to find candidate biomarkers related to stage progress, which may be closely associated with tumor advancement and the metastasis issue. Through comparison with established feature selection methods, the good performance of the proposed algorithm in terms of classification performance and stability was established. The stability of feature selection is a significant issue and its importance has been underestimated for a long time; indeed, many research efforts aimed at feature

selection have focused solely on the performance of the examined methods. However, as is apparent from the non-ensemble methods examined in the experiment, feature selection algorithms designed without considering stability may yield many different subsets of features if the data changes even slightly. This causes low reproducibility for high-dimensional datasets such as microarray data or RNA-seq, and renders the analysis result less meaningful. Thus, stable feature selection is an essential issue in biomarker discovery. Of course, the feature performance should not be neglected because feature stability does not guarantee true biomarker detection. Although, based on a simple idea, the proposed approach was shown to be moderately successful when applied to datasets consisting of a very large number of features and much smaller samples.

Since a general process for resolution of binary classification problems on high-dimensional data was proposed in this study, I expect the proposed method to be applicable to many other kinds of biomarker discovery. However, although the proposed method generally demonstrated improved performance compared to conventional techniques, it depends on gene expression data only. Therefore, it is necessary to integrate other datasets, such as the DNA methylation dataset. Thus, I also proposed an integrative-analysis method combining gene expression and DNA methylation data for cancer biomarker discovery. Gene expression analysis has been widely used for cancer analysis; however, use of a single dataset has limited effectiveness because of the amount of information and noise included. Here, simple integrative analysis with an additional dataset, i.e., DNA methylation dataset, along with implementation of normalization and unsupervised feature extraction, yielded improvement over the results obtained using a single gene

expression or DNA methylation dataset. Additionally, I showed that candidate biomarkers related to cancer progression can be acquired by analysis of integrated datasets. The proposed approaches are expected to be applied to various research studies aimed at candidate cancer biomarker discovery.

References

1. Moon M, Nakai K: **Stable feature selection based on the ensemble L_1 -norm support vector machine for biomarker discovery.** *BMC genomics* 2016, 17(13):1026.
2. He Z, Yu W: **Stable feature selection for biomarker discovery.** *Computational biology and chemistry* 2010, 34(4):215-225.
3. Saeys Y, Inza I, Larrañaga P: **A review of feature selection techniques in bioinformatics.** *Bioinformatics* 2007, 23(19):2507-2517.
4. Guyon I, Elisseeff A: **An introduction to variable and feature selection.** *The Journal of Machine Learning Research* 2003, 3:1157-1182.
5. Hall MA: **Correlation-based feature selection for machine learning.** The University of Waikato; 1999.
6. Saeys Y, Abeel T, Van de Peer Y: **Robust feature selection using ensemble feature selection techniques.** In: *Machine learning and knowledge discovery in databases.* Springer; 2008: 313-325.
7. Abeel T, Helleputte T, Van de Peer Y, Dupont P, Saeys Y: **Robust biomarker identification for cancer diagnosis with ensemble feature selection methods.** *Bioinformatics* 2010, 26(3):392-398.
8. Dernoncourt D, Hanczar B, Zucker J-D: **Stability of Ensemble Feature Selection on High-Dimension and Low-Sample Size Data-Influence of the Aggregation Method.** In: *ICPRAM: 2014.* 325-330.
9. Bach FR: **Bolasso: model consistent lasso estimation through the bootstrap.** In: *Proceedings of the 25th international conference on Machine learning: 2008.* ACM: 33-40.
10. Meinshausen N, Bühlmann P: **Stability selection.** *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2010, 72(4):417-473.

11. Yu L, Liu H: **Feature selection for high-dimensional data: A fast correlation-based filter solution.** In: *ICML: 2003*. 856-863.
12. Díaz-Uriarte R, De Andres SA: **Gene selection and classification of microarray data using random forest.** *BMC bioinformatics* 2006, **7**(1):1.
13. Guyon I, Weston J, Barnhill S, Vapnik V: **Gene selection for cancer classification using support vector machines.** *Machine learning* 2002, **46**(1-3):389-422.
14. Broad Institute TCGA Genome Data Analysis Center: Broad Institute of MIT and Harvard; 2015
15. Wagner GP, Kin K, Lynch VJ: Measurement of mRNA abundance using RNA-seq data: **RPKM measure is inconsistent among samples.** *Theory in biosciences* 2012, **131**(4):281-285.
16. Li B, Dewey CN: **RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome.** *BMC bioinformatics* 2011, **12**(1):1.
17. Guinan P, Sobin LH, Algaba F, Badellino F, Kameyama S, MacLennan G, Novick A: **TNM staging of renal cell carcinoma.** *Cancer* 1997, **80**(5):992-993.
18. Noble WS: **What is a support vector machine?** *Nature biotechnology* 2006, **24**(12):1565-1567.
19. Bradley PS, Mangasarian OL: **Feature selection via concave minimization and support vector machines.** In: *ICML: 1998*. 82-90.
20. Zhu J, Rosset S, Hastie T, Tibshirani R: **1-norm support vector machines.** *Advances in neural information processing systems* 2004, **16**(1):49-56.
21. Efron B, Hastie T, Johnstone I, Tibshirani R: **Least angle regression.** *The Annals of statistics* 2004, **32**(2):407-499.
22. Friedman J, Hastie T, Tibshirani R: **Regularization paths for generalized linear models via coordinate descent.** *Journal of statistical software* 2010, **33**(1):1.
23. Efron B, Tibshirani RJ: **An introduction to the bootstrap:** CRC press; 1994.

24. Breiman L: **Bagging predictors**. *Machine learning* 1996, **24**(2):123-140.
25. Hsu C-W, Chang C-C, Lin C-J: **A practical guide to support vector classification**. 2003.
26. Kalousis A, Prados J, Hilario M: **Stability of feature selection algorithms: a study on high-dimensional spaces**. *Knowledge and information systems* 2007, **12**(1):95-116.
27. Jiang Z, Xu R: **A novel feature extraction approach for microarray data based on multi-algorithm fusion**. *Bioinformatics* 2015, **11**(1):27.
28. Braga-Neto UM, Dougherty ER: **Is cross-validation valid for small-sample microarray classification?** *Bioinformatics* 2004, **20**(3):374-380.
29. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH: **The WEKA data mining software: an update**. *ACM SIGKDD explorations newsletter* 2009, **11**(1):10-18.
30. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V: **Scikit-learn: Machine learning in Python**. *The Journal of Machine Learning Research* 2011, **12**:2825-2830.
31. Jagga Z, Gupta D: **Classification models for clear cell renal carcinoma stage progression, based on tumor RNAseq expression trained supervised machine learning algorithms**. In: *BMC proceedings: 2014*. BioMed Central Ltd: S2.
32. Haury A-C, Gestraud P, Vert J-P: **The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures**. *PloS one* 2011, **6**(12):e28210.
33. Welsh JB, Sapinoso LM, Su AI, Kern SG, Wang-Rodriguez J, Moskaluk CA, Frierson HF, Hampton GM: **Analysis of gene expression identifies candidate markers and pharmacological targets in prostate cancer**. *Cancer research* 2001, **61**(16):5974-5978.
34. Van De Vijver MJ, He YD, Van't Veer LJ, Dai H, Hart AA, Voskuil DW, Schreiber

- GJ, Peterse JL, Roberts C, Marton MJ: **A gene-expression signature as a predictor of survival in breast cancer.** New England Journal of Medicine 2002, 347(25):1999-2009.
35. Griffith OL, Melck A, Jones SJ, Wiseman SM: **Meta-analysis and meta-review of thyroid cancer gene expression profiling studies identifies important diagnostic biomarkers.** Journal of Clinical Oncology 2006, 24(31):5043-5051.
36. Van't Veer LJ, Dai H, Van De Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, Van Der Kooy K, Marton MJ, Witteveen AT: **Gene expression profiling predicts clinical outcome of breast cancer.** nature 2002, 415(6871):530-536.
37. Klahan S, Huang W-C, Chang C-M, Wong HS-C, Huang C-C, Wu M-S, Lin Y-C, Lu H-F, Hou M-F, Chang W-C: **Gene expression profiling combined with functional analysis identify integrin beta1 (ITGB1) as a potential prognosis biomarker in triple negative breast cancer.** Pharmacological research 2016, 104:31-37.
38. Sadahiro S, Suzuki T, Tanaka A, Okada K, Saito G, Kamijo A, Nagase H: **The gene expression levels of gamma-glutamyl hydrolase in tumor tissues may be a useful biomarker for proper use of S-1 and tegafur-uracil/leucovorin in preoperative chemoradiotherapy in patients with rectal cancer.** Annals of Oncology 2016, 27(suppl_6).
39. Jin H, Lee H-C, Park SS, Jeong Y-S, Kim S-Y: **Serum cancer biomarker discovery through analysis of gene expression data sets across multiple tumor and normal tissues.** Journal of biomedical informatics 2011, 44(6):1076-1085.
40. Sadler T, Bhasin JM, Xu Y, Barnholz-Sloan J, Chen Y, Ting AH, Stylianou E: **Genome-wide analysis of DNA methylation and gene expression defines molecular characteristics of Crohn's disease-associated fibrosis.** Clinical epigenetics 2016, 8(1):30.
41. Schübeler D: **Function and information content of DNA methylation.** Nature

- 2015, 517(7534):321.
42. Fleischer T, Edvardsen H, Solvang HK, Daviaud C, Naume B, Børresen-Dale AL, Kristensen VN, Tost J: **Integrated analysis of high-resolution DNA methylation profiles, gene expression, germline genotypes and clinical end points in breast cancer patients.** *International journal of cancer* 2014, 134(11):2615-2625.
 43. Yoo S, Takikawa S, Geraghty P, Argmann C, Campbell J, Lin L, Huang T, Tu Z, Feronjy R, Spira A: **Integrative analysis of DNA methylation and gene expression data identifies EPAS1 as a key regulator of COPD.** *PLoS genetics* 2015, 11(1):e1004898.
 44. Rhee J-K, Kim K, Chae H, Evans J, Yan P, Zhang B-T, Gray J, Spellman P, Huang TH-M, Nephew KP: **Integrated analysis of genome-wide DNA methylation and gene expression profiles in molecular subtypes of breast cancer.** *Nucleic acids research* 2013, 41(18):8464-8474.
 45. Low G, Huang G, Fu W, Moloo Z, Girgis S: **Review of renal cell carcinoma and its common subtypes in radiology.** *World journal of radiology* 2016, 8(5):484.
 46. Osborne JW: **Improving your data transformations: Applying the Box-Cox transformation.** *Practical Assessment, Research & Evaluation* 2010, 15(12):1-9.
 47. Abdi H, Williams LJ: **Principal component analysis.** *Wiley interdisciplinary reviews: computational statistics* 2010, 2(4):433-459.
 48. Schölkopf B, Smola A, Müller K-R: **Kernel principal component analysis.** In: *International Conference on Artificial Neural Networks: 1997.* Springer: 583-588.
 49. Ng A: **Sparse autoencoder.** *CS294A Lecture notes* 2011, 72(2011):1-19.
 50. Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, Corrado GS, Davis A, Dean J, Devin M: **Tensorflow: Large-scale machine learning on heterogeneous distributed systems.** *arXiv preprint arXiv:160304467* 2016.
 51. Imai K, Hirata S, Irie A, Senju S, Ikuta Y, Yokomine K, Harao M, Inoue M, Tomita Y, Tsunoda T: **Identification of HLA-A2-restricted CTL epitopes of a novel**

- tumour-associated antigen, KIF20A, overexpressed in pancreatic cancer.** British journal of cancer 2011, 104(2):300-307.
52. Chung W, Bondaruk J, Jelinek J, Lotan Y, Liang S, Czerniak B, Issa J-PJ: **Detection of bladder cancer using novel DNA methylation biomarkers in urine sediments.** Cancer Epidemiology and Prevention Biomarkers 2011, 20(7):1483-1491.
53. Ilse P, Biesterfeld S, Pomjanski N, Fink C, Schramm M: **SHOX2 DNA methylation is a tumour marker in pleural effusions.** Cancer Genomics-Proteomics 2013, 10(5):217-223.
54. Abildgaard MO, Borre M, Mortensen MM, Ulhøi BP, Tørring N, Wild P, Kristensen H, Mansilla F, Ottosen PD, Dyrskjød L: **Downregulation of zinc finger protein 132 in prostate cancer is associated with aberrant promoter hypermethylation and poor prognosis.** International journal of cancer 2012, 130(4):885-895.
55. Szponar A, Kovacs G: **Expression of KRT7 and WT1 differentiates precursor lesions of Wilms' tumours from those of papillary renal cell tumours and mucinous tubular and spindle cell carcinomas.** Virchows Archiv 2012, 460(4):423-427.
56. Yu M, Smolen GA, Zhang J, Wittner B, Schott BJ, Brachtel E, Ramaswamy S, Maheswaran S, Haber DA: **A developmentally regulated inducer of EMT, LBX1, contributes to breast cancer progression.** Genes & development 2009, 23(15):1737-1742.
57. Qiu X, Huang Y, Zhou Y, Zheng F: **Aberrant methylation of TRIM58 in hepatocellular carcinoma and its potential clinical implication.** Oncology reports 2016, 36(2):811-818.
58. Chang C-C, Huang R-L, Liao Y-P, Su P-H, Hsu Y-W, Wang H-C, Tien C-Y, Yu M-H, Lin Y-W, Lai H-C: **Concordance analysis of methylation biomarkers**

- detection in self-collected and physician-collected samples in cervical neoplasm.** BMC cancer 2015, 15(1):418.
59. Young J, Ménétrey J, Goud B: **RAB6C is a retrogene that encodes a centrosomal protein involved in cell cycle progression.** Journal of molecular biology 2010, 397(1):69-88.
60. Qiao H-P, Gao W-S, Huo J-X, Yang Z-S: **Long non-coding RNA GAS5 functions as a tumor suppressor in renal cell carcinoma.** Asian Pacific journal of cancer prevention 2013, 14(2):1077-1082.
61. Guo W, Xu H, Chen J, Yang Y, Jin JW, Fu R, Liu HM, Zha XL, Zhang ZG, Huang WY: **Prohibitin suppresses renal interstitial fibroblasts proliferation and phenotypic change induced by transforming growth factor- β 1.** Molecular and cellular biochemistry 2007, 295(1):167-177.
62. Crea F, Venalainen E, Ci X, Cheng H, Pikor L, Parolia A, Xue H, Saidy NRN, Lin D, Lam W: **The role of epigenetics and long noncoding RNA MIAT in neuroendocrine prostate cancer.** 2016.

Acknowledgements

First, I would like to express my gratitude to my supervisor, Prof. Kenta Nakai, who has provided me with instruction and valuable advice during my PhD studies. In addition, I would like to thank the members of our laboratory, the “Laboratory of Functional Analysis in Silico,” for sharing ideas and providing me with considerable assistance. I would also like to thank the Japanese Government (MEXT) for financial support in the form of a scholarship, and the University of Tokyo for providing an excellent environment and resources for this research project. Lastly, I would like to thank my beloved wife Yoojin, who has always supported and encouraged me.