Comparison of Feature Selection Methods for Cross-Laboratory Microarray Analysis

Hsi-Che Liu, Pei-Chen Peng, Tzung-Chien Hsieh, Ting-Chi Yeh, Chih-Jen Lin,
Chien-Yu Chen, Jen-Yin Hou, Lee-Yung Shih, and Der-Cherng Liang

Abstract—The amount of gene expression data of microarray has grown exponentially. To apply them for extensive studies, integrated analysis of cross-laboratory (cross-lab) data becomes a trend, and thus, choosing an appropriate feature selection method is an essential issue. This paper focuses on feature selection for Affymetrix (Affy) microarray studies across different labs. We investigate four feature selection methods: t-test, significance analysis of microarrays (SAM), rank products (RP), and random forest (RF). The four methods are applied to acute lymphoblastic leukemia, acute myeloid leukemia, breast cancer, and lung cancer Affy data which consist of three cross-lab data sets each. We utilize a rank-based normalization method to reduce the bias from cross-lab data sets. Training on one data set or two combined data sets to test the remaining data set(s) are both considered. Balanced accuracy is used for prediction evaluation. This study provides comprehensive comparisons of the four feature selection methods in cross-lab microarray analysis. Results show that SAM has the best classification performance. RF also gets high classification accuracy, but it is not as stable as SAM. The most naive method is t-test, but its performance is the worst among the four methods. In this study, we further discuss the influence from the number of training samples, the number of selected genes, and the issue of unbalanced data sets.

Index Terms—Microarray data analysis, feature selection, cancer, cross-laboratory experiment

1 Introduction

The gene expression profiling techniques by DNA microarrays provide the simultaneous analysis of thousands of genes [1], [2]. Microarray technology, therefore, has become a revolutionary tool for understanding human diseases. Golub et al. [3] demonstrated a generic approach of classifying cancer by gene expression profiling, for example, to distinguish between acute lymphoblastic leukemia and acute myeloid leukemia. Since then, disease classification has been one of the primary issues of microarray research. Those papers on disease classification mainly focus on the following three issues:

1. Identify relevant genes (features) for certain diseases/subtypes.
2. Build classifiers based on samples with known disease class labels.
3. Classify the unknown samples into known disease classes.

The high cost of microarray experiments causes the insufficient study samples of disease subtypes and deteriorates the problem of imbalance between the small sample size and large number of features for further studies. To alleviate this situation and enhance the confidence of microarray data analysis, researchers expect to integrate data from different laboratories into a larger database. Hence, cross-laboratory (cross-lab) analysis has become a common trend [4], [5], [6], [7], [8], [9], [10]. However, it is known [11] that the so-called batch effects, i.e., the different experimental conditions between laboratories, such as sample preparation, may affect the quality of cross-laboratory analysis.

Several approaches are available to minimize the bias caused by cross-lab microarray data; most of them focus on issues 2 and 3. Bloom et al. [4] collected samples of 21 tumor types across different laboratories and applied an artificial neural network model for classification. High prediction accuracy of 88 percent on average was reported. Our previous work [5] collected eight public microarray data sets of three cancer types and demonstrated that a simple rank-based approach can classify data from different sources. We also addressed a potential pitfall when evaluating cross-lab predictions. Liu et al. [6] designed a classification method ManiSVM for cross-platform microarray data. They achieved the overall accuracy of 70.7 percent. Some other related work has been published [7], [8], [9].

In microarray studies, correct prediction of disease subtypes relies on the robust selection of relevant genes. Many feature selection methods have been applied for discovering differential genes. But, it brings up a great challenge in computational analysis due to the characteristics of large amounts of irrelevant features and the paucity of samples. Compared to issues 2 and 3, recent studies on issue 1 concern more about the feature selection methods for microarray data within the same laboratory. A review

- *H.-C. Liu, T.-C. Yeh, J.-Y. Hou, and D.-C. Liang are with Mackay Medical College and Division of Pediatric Hematology-Oncology, Mackay Memorial Hospital, New Taipei, Taiwan. E-mail: dcliang@ms1.mmh.org.tw.*
- *P.-C. Peng, T.-C. Hsieh, and C.-J. Lin are with the Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan.*
- *C.-Y. Chen is with the Department of Bio-Industrial Mechatronics Engineering, National Taiwan University, Taipei, Taiwan.*
- *L.-Y. Shih is with the Division of Hematology-Oncology, Department of Internal Medicine, Chang Gung Memorial Hospital, Taipei, Taiwan, and the School of Medicine, Chang Gung University, Taoyuan, Taiwan.*

Manuscript received 27 Aug. 2012; revised 21 May 2013; accepted 7 June 2013; published online 12 June 2013.
For information on obtaining reprints of this article, please send e-mail to: tcbb@computer.org, and reference IEEECS Log Number TCBB-2012-08-0227.
Digital Object Identifier no. 10.1109/TCBB.2013.70.

Published by the IEEE CS, CI, and EMB Societies & the ACM

by Saeys et al. [12] has surveyed many recent developments. Generally, there are two major classes of feature selection methods: filter and embedded. Both classes have their own advantages. Filter methods analyze the intrinsic properties of the data. Therefore, they are more intuitive and can be efficiently applied to high-dimensional data. A simple filtering method is to set a threshold on the fold-change, which is a value describing how much the expression levels changes to find differentiated genes. However, fold-change may not be accurate because the scale of the fold varies according to expression levels. Some statistical methods are, therefore, proposed. ANOVA [13], t-test [13], and Bayesian frameworks [14] are widely used parametric filtering methods in microarray studies. However, due to the insufficient study samples, unclearness about the underlying distribution of data sets arises. Some nonparametric or model-free filtering methods such as significance analysis of microarrays (SAM) [15], Wilcoxon rank-sum [16], and rank products (RP) [17] are brought about. These methods estimate the reference distribution via random permutation which mitigates the problem of small sample sizes in microarray studies and provide a more reliable result. Apart from filtering class, embedded class offers an alternative way to choose the relevant features. These embedded methods typically employ a classifier to evaluate the performance of feature combinations and propose a list of discriminative features. Methods including random forest (RF) [18], weight vector of support vector machines [19], and weight of logistic regression [20] calculate the importance of features in a multivariate way. Most of the embedded methods also take the gene correlation into consideration.

Although many feature selection methods for microarray studies are brought out, their performances on cross-lab microarray data have not been carefully evaluated. In this paper, we collect Affymetrix (Affy) microarray data in public repositories. There are 12 data sets composed of four cancer types: acute lymphoblastic leukemia (ALL), acute myeloid leukemia (AML), breast cancer, and lung cancer. The four feature selection methods (t-test, SAM, RP, and RF) are then used to identify relevant genes from cross-lab data sets. The performance of classification is evaluated by balanced accuracy which is a suitable criterion for unbalanced data sets.

We recognize that each part of the microarray analysis, including normalization, feature selection, and classification, will have different impacts on the prediction performance. But, many efforts are required to study the influence of every part in the workflow. A recent study of MicroArray Quality Control (MAQC) [21] suggested that the model performance depends largely on the difficulty of the problem, and that the impact of feature selection methods seems less important. However, apart from achieving high classification performance, finding important features related to the diseases of interest is still an essential issue in microarray data studies. Here, we emphasize on microarray data from cross-lab and investigate the impact owing to the feature selection method selected. For this purpose, we fix on one normalization method and one classifier in the workflow to make the evaluation fair.

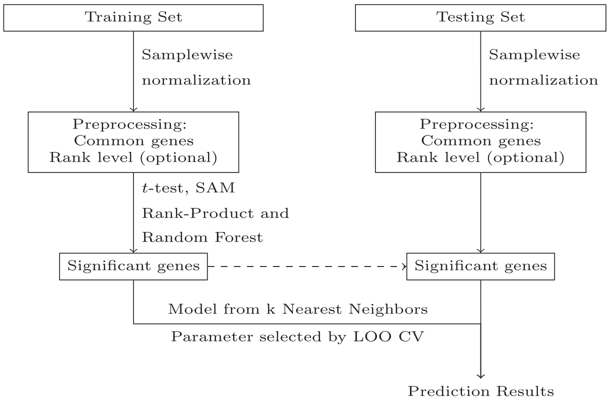
Training Set
Testing Set
Samplewise normalization
Samplewise normalization
Preprocessing: Common genes Rank level (optional)
Preprocessing: Common genes Rank level (optional)
t-test, SAM Rank-Product and Random Forest
Significant genes
Significant genes
Model from k Nearest Neighbors
Parameter selected by LOO CV
Prediction Results

Fig. 1. Flowchart of the whole experiment. The width of each box reflects the number of genes. Details are in the beginning of Section 2. Transforming expression values to rank levels is performed for comparing the results between using expression values and using rank levels.

2 Materials and Methods

The whole workflow is described in Fig. 1. Data sets from different laboratories are downloaded from public websites; most of them are from gene expression omnibus (GEO), which is a public functional genomics data repository [22]. To avoid array-specific biases from differential experimental conditions, all samples are consistently normalized and common genes of different generations of Affy microarray chips are identified. We may further apply sample-wise rank-based normalization on the data.

For each experiment, we analyze data sets from the same cancer type. Two ways to define the training/testing data sets are conducted. First, two data sets of a particular cancer type are chosen: one as training and the other as testing. Second, two data sets are combined as training and the remaining one as testing. Both training and test sets are constructed to separate samples associated with a subtype from those not with this subtype. Every pair of training set(s) and testing set of the same cancer type undergoes the experimental procedure.

Next, from training data set(s) we can obtain a relevant gene list through one of the four feature selection methods (t-test, SAM, RP, and RF). We changed the number of selected relevant genes from 5 to 250 by an interval of 5. We then utilize the k-nearest neighbors [23] for the classification analysis, where the parameter k is obtained from the procedure of leave-one-out cross-validation (LOO CV).

Then the balanced accuracy is used to evaluate the results. Hereafter, we define the term "experiment" as the procedures conducted in Fig. 1, including choosing training/testing data sets for a particular cancer subtype, separating samples from a specific cancer subtype, optionally doing rank-based normalization, applying four feature selection methods to get a relevant gene list, and evaluating classification performance. The details of our methods is described in the following sections.

2.1 Microarray Data Collection and Preprocessing

All the data sets we used are from the public websites. Data quality is an issue of concern because it is difficult to ensure that all collected data sets are reliable. For example, when we were searching for lung cancer data in GEO, we

TABLE 1
Key Characteristics of the Analyzed Data

Study reference	Microarray generation	No. of samples	Cancer subtypes for analysis {No. of samples in the subtype}
Acute lymphoblastic leukemia (ALL)			
ALL-1 [24]	HG-U133A	130	t(12;21) {20}; t(1;19) {18}; HD>50 {17}
ALL-2 [25]	HG-U133A	107	t(12;21) {23}; t(1;19) {6}; HD>50 {23}
ALL-3 [26]	HG-U133A	188	t(12;21) {43}; t(1;19) {13}; HD>50 {44}
Acute myeloid leukemia (AML)			
AML-1 [27]	HG-U133A	285	t(15;17) {18}; inv(16) {19}; t(8;21) {22}
AML-2 [29]	HG-U133A	130	t(15;17) {15}; inv(16) {14}; t(8;21) {21}
AML-3 [28]	HG-U133A	43	t(15;17) {10}; inv(16) {4}; t(8;21) {0}
Breast cancer			
Breast-1 [30]	HuGeneFL	49	ER+ {25}; ER− {24}
Breast-2 [31]	HG-U95Av2	89	ER+ {74}; ER− {15}
Breast-3 [32]	HG-U133A	286	ER+ {209}; ER− {77}
Lung cancer			
Lung-1 [33]	HG-U133 Plus 2.0	111	AD {58}; SQ {53}
Lung-2 [34]	HG-U133 Plus 2.0	138	AD {63}; SQ {75}
Lung-3 [35]	HG-U133 Plus 2.0	58	AD {40}; SQ {18}

excluded some data for analysis because their corresponding papers had been retracted.

We have collected the data sets of four cancer types including ALL, AML, breast cancer, and lung cancer. For the cancer type of ALL, the three data sets from two laboratories all adopted Affy HG-U133A array. The subtypes to be studied are ALLs with defined recurrent chromosomal aberrations: t(12;21), t(1;19), and hyperdiploid with more than 50 chromosomes (HD > 50). The first data set, ALL-1, is from Ross et al. [24]. It provides 132 samples originally. Two cases with hyperdiploid are excluded because they also had another chromosomal aberration t(9;22). The remaining 130 cases are the target for further analysis. The second data set, denoted as ALL-2, is from Mullighan et al. [25]. ALL-2 contains 175 samples, including 68 samples of ALL-1. After removing the 68 overlapped samples, 107 samples are used for analysis. The third data set is ALL-3 from Den Boer et al. [26]. In ALL-3, samples present with more than one ALL subtype are excluded. After the removal of one sample of t(9;22) and hyperdiploid and one sample of t(12;21) and hyperdiploid, 188 samples are remained for the study.

The second cancer type is AML. Its predicted subtypes are AMLs with t(8;21), inv(16), and t(15;17). Three data sets of HG-U133A chips are enrolled. AML-1 from Valk et al. [27] and AML-3 from Gutiérrez et al. [28] are adult studies, but AML-2 from Ross et al. [29] is a childhood study.

The third cancer type is breast cancer. There are three data sets across three generations of Affy chips: Breast-1 [30] uses HuGeneFL, Breast-2 [31] uses HG-U95Av2, and Breast-3 [32] uses HG-U133A. The probe sets differ between generations of Affy microarrays, so it is crucial to identify the common genes for comparative analysis. We adopted Affy's matching tables, which are based on the similarity of sequence information of probe sets (http://www.affymetrix.com/support/technical/comparison_spreadsheets.affx). The gene intersection among the three generations contains 5,045 common genes, which we then used for the further analysis. The estrogen receptor (ER) status, which is either positive (ER+) or negative (ER−), is the subtype to be predicted.

The last cancer type is lung cancer. Three data sets of HG-U133 Plus 2.0 arrays are Lung-1 [33], Lung-2 [34], and Lung-3 [35]. The subtypes chosen to be predicted are squamous cell carcinoma (SQ) and adenocarcinoma (AD). The details of the above data sets are summarized in Table 1 and their downloading URLs are in supplemental Table S1, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TCBB.2013.70.

We follow the Affymetrix Microarray Suite 5.0 (Affymetrix statistical algorithms description document, 2002 Rev 3) program to rescale gene expression in each array by setting the 2 percent trimmed mean of all the genes to be 500. Exceptions are lung cancer data sets, which have no raw data available on GEO. These sets were prenormalized by different methods in individual studies and cannot be used without rank-based normalization (see Section 2.1.1). Since the main objective of this study is mostly about the comparison between feature selection methods, we only take care of the array specific biases but not of the between batches biases (i.e., different experimental conditions between laboratories). To take full consideration of batches biases, further investigation using mean-centering, standardization, ratio-based, or Extended Johnson-Li-Rabinovic methods [11] may be performed.

2.1.1 Rank-Based Normalization

Many studies have shown that the classification accuracy with rank-based normalization is more stable than only using expression values [5], [36], [37], [38], [39]. Expression values may be biased because the scale of each gene may vary among different experimental environments. The added procedure rank-based normalization can adjust and improve the situation. In this study, we use both expression values and rank levels for feature selections. The only exception is lung cancer, for which only rank levels are considered. We mentioned that raw data are not available on GEO. However, because normalization conducted by earlier studies preserve the order of expression values. Therefore, rank levels can still be obtained.

2.2 Feature Selection Methods

Feature selection methods in microarray studies can be divided into two major classes: filtering and embedded [12]. We choose methods from both classes for analysis. In the filtering class with parametric assumptions, t-test is considered. Adapted from t-test, SAM is a representative nonparametric method in the filtering class. Just as SAM, RP applies random permutations to alleviate the problem of unknown reference distribution. With regard to the embedded class, RF is investigated.

Another famous feature selection method, moderated t-test from limma's package (Bioconductor) [40], is also popular in microarray studies. We observed that the performance of moderated t-test is similar to that of SAM in some pilot experiments (data not shown). In this regard, only SAM is used for the comparison in this study.

2.2.1 t-Test

Student's t-test was first used in Golub et al. [3] to evaluate the importance of genes. Since then, student's t-test has become a widely used technique in microarray data analysis. The t-test used here is unpaired two sample t-test. Next, t-test assumes that two groups are normally distributed and their variance are equal. With the assumption, we have the null hypothesis that the means of two groups are the same. The following equation:

$$t = \frac{\overline{X}_1 - \overline{X}_2}{S_{X_1X_2} \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \tag{1}$$

$$S_{X_1X_2} = \sqrt{\frac{(n_1 - 1)S^2_{X_1} + (n_2 - 1)S^2_{X_2}}{n_1 + n_2 - 2}}, \tag{2}$$

calculates the t statistic, where $\overline{X}_1$ and $\overline{X}_2$ are averages of group one and group two, $S_{X_1X_2}$ is the common standard deviation of the two groups. The n_1 and n_2 are the number of samples of predicted subtype and others, respectively. Once the value of t is determined, we can find the p-value by checking a table of values from Student's t-distribution. At last, the list of relevant genes is obtained by sorting their p-values in ascending order.

2.2.2 Significant Analysis of Microarrays

Significant analysis of microarrays (SAM) is a general approach of detecting changes in gene expression in DNA microarrays. We use the R package "samr" [15] for computation. At first, we calculate the relative difference $d(i)$ for each gene:

$$d(i) = \frac{\overline{x}_I(i) - \overline{x}_U(i)}{s(i) + s_0}, \tag{3}$$

where $\overline{x}_I(i)$ and $\overline{x}_U(i)$ are the means of gene expression for gene i in the two groups I and U, $s(i)$ is the standard deviation of repeated expression measurement, and parameter s_0 is added to ensure the distribution of $d(i)$ is not largely affected by small variation of gene expression. $s(i)$ can be computed by the following formula:

$$s(i) = \sqrt{a\left\{\sum_{m\in I}[x_m(i) - \overline{x}_I(i)]^2 + \sum_{n\in U}[x_n(i) - \overline{x}_U(i)]^2\right\}}. \tag{4}$$

In (4), we use

$$a = \frac{\frac{1}{n_1} + \frac{1}{n_2}}{n_1 + n_2 - 2}, n_1 = |I|, \text{ and } n_2 = |U|.$$

Because the underlying distribution is unknown, all samples are permutated 300 times following the setting in the "samr" package. In each permutation, the values $d(i)$ are sorted descendantly. We then give a new notation, $d_p(i)$, which denotes the ith largest relative difference values in pth permutation. After 300 iterations of permutation, we calculate the expected relative difference,

$$d_E(i) = \frac{\sum_p d_p(i)}{300}.$$

To find the relevant genes in the expression data, we drew the scatter plot of the sorted relative difference $d(i)$ versus the expected relative difference $d_E(i)$ obtained above. For the vast majority of genes, $d(i) \cong d_E(i)$, but some genes are deviated from the $d(i) = d_E(i)$ line by a distance greater than a threshold Δ. The value Δ is obtained from false discovery rate (FDR), which is set as 5 percent.

2.2.3 Rank Products

Rank products (RP), proposed by Hong et al. [17], is a simple, yet powerful method to detect differentially expressed genes in microarray data analysis. Unlike t-test or other parametric feature selection methods, rank products has no distributional assumptions and is often referred as nonparametric or model-free feature selection method. In this study, we utilize RankProd, a bioconductor package which modifies and extends the rank products method proposed by Breitling et al. [41], to find the relevant genes. The algorithm implemented in RankProd is described as follows [17]:

1. For each gene in the samples within the two groups, compute their pair-wise fold-change ratios. Here, we calculated fold-change ratio by dividing the gene expression value of the training data set by its counterpart of the testing data set. Suppose, there are m samples in the training data set and n samples in the testing data set. There will be mn comparisons in this step.
2. Rank the ratios within each comparison, i.e., $\text{rank} = 1$ for the gene with the largest fold-change ratio.
3. For each gene, multiply all its rank values of comparisons. We call the result of multiplication by rank product.
4. Shuffle the gene expression within each sample independently and repeat steps 1 to 3 300 times and get 300 "rank products." The permutation partly alleviates the problem of sample sparsity in microarray data analysis, improving the robustness against outliers.
5. With the 300 rank products values derived from step 4, a reference distribution of rank products is formed which determines the p-value and FDR of

each gene. The list of relevant genes is accessed according to the ascending order of the p-value associated with each gene.

2.2.4 Random Forest

Random forest (RF), proposed by Breiman [42], is an embedded feature selection method which interacts with the classifier. RF is based on the construction of a large number of classification trees using bootstrapped samples from the training data set. And then majority voting determines the prediction class of each sample. Given a training data set of size N, bootstrap aggregating (bagging) produces new training data sets by sampling N cases with replacement from the original data. And at each node of the classification tree, a gene is specified as the candidate. In this study, we built 200 classification trees on every training data set. After generating each classification tree, "out-of-bag (oob)" samples which are not included in the bootstrapped samples are applied in getting variable importance. In prediction, we put each sample down on every classification tree in the forest, and each tree decides a class for the sample. For each sample the class with the most votes is considered as predicted class. An oob error rate is calculated internally during the process and treated as "importance measure" of each variable. We then obtain the list of relevant genes according to the features with higher importance measures.

2.3 Classification

We evaluate feature selection methods by checking cross-lab classification accuracy. For every cancer subtype, one data set or a combination of two data sets from Table 1 is used for feature selection and considered as the training set, and another data set (across generations or laboratories) is selected as testing. All pairs of training and testing data sets undergo this procedure. An exception is the t(8;21) subtype of AML because AML-3 has no instances in this subtype and cannot be used for training/testing. The k-nearest neighbors (kNN) [23] is applied in the classification task. For any sample in the testing data set, kNN predicts the majority class of the sample's k closest neighbors. Meanwhile, the distance between two data samples is defined by Euclidean metric. The best choice of parameter k required in kNN depends on the training data. The larger value of k can reduce the noise of classification, but makes the boundaries between classes less distinct. To select a good k, we utilize leave-one-out cross-validation (LOO CV). In the training data set, LOO CV sequentially chooses one instance out for validation upon any given k. The value k with the best LOO CV balanced accuracy is employed in the prediction of independent testing data; see the definition of balanced accuracy in Section 2.4. We choose the minimum value of k if more than one k have the best balanced LOO CV accuracy. In addition, to avoid tied voting in classification, k is chosen from odd integers of 1 to 17. Values larger than 17 do not give better LOO CV.

2.4 Evaluation

The most common prediction measurement is accuracy. However, for unbalanced data, a high accuracy value by predicting all data to the major class may be misleading. Note that some data sets used are unbalanced. For example, Breast-2 had 74 ER+, but only 15 ER−. Therefore, we applied an alternative measurement:

$$\text{Balanced accuracy} = \frac{1}{2} \cdot \left(\frac{\text{TP}}{\text{TP} + \text{FN}} + \frac{\text{TN}}{\text{TN} + \text{FP}} \right),$$

where TP, TN, FP, and FN indicate numbers of true positive, true negative, false positive, and false negative predictions, respectively. It is generally believed that the balanced accuracy better handle the data imbalance and can reveal the performance on identifying the specific cancer subtype.

3 Results and Discussion

To begin, we briefly summarize all experimental results and leave details in sections. The main result is the comparison on the classification performance of four feature selection methods. It is observed that SAM, RP, and RF outperform t-test in most of the experiments.

Then we compare the results of data with gene expression values and the ones with sample-wise rank-based normalization. When rank-based normalization is applied, all four feature selection methods, especially SAM, have better performance on almost all data sets.

The above results are similar both in training on one data set or two combined data sets. For subsequent analysis of each individual feature selection method, we use one single data set for training and apply rank-based normalization.

Choosing an optimal number of genes is an essential issue in feature selection. We find that the performance (balanced accuracy) is more stable when the number of selected relevant genes is above 100.

We discuss the influence of unbalanced data sets on the performance. We further point out how to properly select the parameter k of kNN in the LOO CV procedure.

We then specifically compare these feature selection methods. SAM and t-test have similar statistical principles, so we investigate why t-test performs much worse than SAM. Stability of feature selection methods is also considered and RF is demonstrated to be less robust.

Based on comprehensive comparisons, we conclude that SAM performs the best among the four methods. RP and RF have their individual shortcomings in cross-lab microarray analysis. The method that gives the worst performance is t-test.

3.1 Performance of Four Feature Selection Methods

Four feature selection methods, t-test, SAM, RP, and RF are analyzed. In Fig. 2, the predicted subtype is ALL with t(12;21). The training data sets are ALL-3 alone or ALL-3 combined with either ALL-1 or ALL-2, while the testing data sets are ALL-1 and ALL-2, respectively. When rank-based normalization is applied, SAM, RP, and RF can approach 100 percent of balanced accuracy in some numbers of relevant genes. They give much better performance than t-test. In the online supplementary Fig. S1, the predicted subtype is AML with t(15;17). The training data sets are AML-1 alone or AML-1 combined with either AML-2 or AML-3, and the testing data sets are AML-2 and AML-3, respectively. The balanced accuracy obtained by SAM, RP, and RF approaches 100 percent regardless of using expression values or rank-based normalization. Again, t-test has inferior performance. The results of balanced accuracy

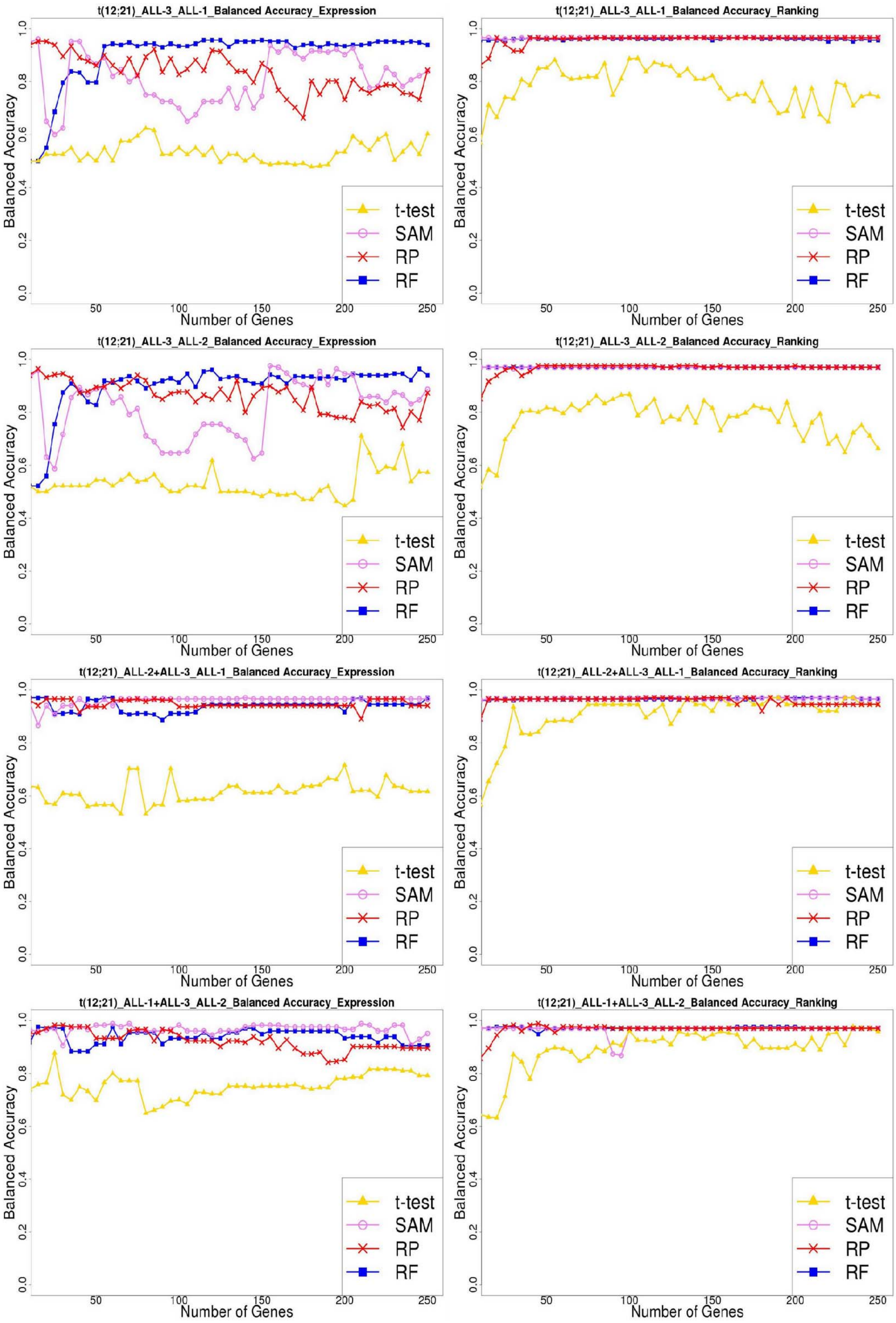

Fig. 2. The balanced accuracy for predicting t(12;21) subtype of ALL. The left figures are using expression value, and the right ones are using rank-based normalization. The main title of each figure stands for "predicted subtype_traning data set(s)_testing data set_evaluation method_ranking or expression value."

for all cancer subtypes are shown in the online supplemental Figs. S2-S19.

As shown in Fig. 2, the number of selected genes affects the resulting balanced accuracy. To compare the four methods by summarizing their results, for each experiment we show the highest balanced accuracy and boldface the value which gives the best balanced accuracy among the four feature selection methods (see Tables 2 and 3). In Table 2, the training data set is from a single data set. It is observed that SAM, RP, and RF achieve almost perfect classification for ALL and AML. On the other hand, *t*-test is the worst in most of the experiments. Similar results are observed in Table 3,

TABLE 2
A Comparison of Different Feature Selection Methods by Training on One Data Set and Testing on Either One of the Remaining Two: Using Expression Values and Rank Levels

Training→Testing	Cancer	Expression values				Rank levels			
	subtype	*t*-test	SAM	RP	RF	*t*-test	SAM	RP	RF
Acute lymphoblastic leukemia (ALL)									
ALL-1 → ALL-2	t(12;21)	87.5	97.6	**98.8**	97.6	93.3	**97**	96	94.8
ALL-1 → ALL-3	t(12;21)	76.9	54.7	64.6	**89.7**	97	**98.3**	97.7	92
ALL-2 → ALL-1	t(12;21)	90	**97**	96.6	**97**	**97**	**97**	**97**	**97**
ALL-2 → ALL-3	t(12;21)	**73.5**	60.5	69.8	64	97.7	**100**	97.7	**100**
ALL-3 → ALL-1	t(12;21)	62.5	**96.1**	95.2	95.7	88.9	**96.6**	**96.6**	**96.6**
ALL-3 → ALL-2	t(12;21)	71.1	**97.6**	96.4	96.4	86.7	97	**97.6**	97
ALL-1 → ALL-2	t(1;19)	96.5	**100**	**100**	**100**	**100**	**100**	**100**	**100**
ALL-1 → ALL-3	t(1;19)	61.2	73.1	**84.6**	80.8	**95**	84.6	84.6	84.6
ALL-2 → ALL-1	t(1;19)	93.6	**96.8**	**96.8**	**96.8**	**96.8**	**96.8**	**96.8**	**96.8**
ALL-2 → ALL-3	t(1;19)	71.6	65.4	**76.9**	57.7	**88.5**	84.6	84.6	84.6
ALL-3 → ALL-1	t(1;19)	78.5	94	95.9	**96.8**	85.7	**96.8**	**96.8**	**96.8**
ALL-3 → ALL-2	t(1;19)	82.2	99.5	**100**	**100**	99.5	**100**	**100**	**100**
ALL-1 → ALL-2	HD>50	73.5	**89.9**	86.7	86.5	93.7	94.8	**95.2**	92.7
ALL-1 → ALL-3	HD>50	77.7	**88.6**	84.2	79.5	87.6	**94.3**	88.3	80.7
ALL-2 → ALL-1	HD>50	82.9	**95.3**	94.4	**95.3**	**96.6**	**96.6**	95.7	**96.6**
ALL-2 → ALL-3	HD>50	82.4	**89.4**	84.6	83.1	85.4	88.6	**93**	92
ALL-3 → ALL-1	HD>50	78.2	93.2	91.5	**93.5**	88.2	94	92.6	**94.8**
ALL-3 → ALL-2	HD>50	72.9	84.6	79.8	**86.2**	85.3	**92.9**	89.5	**92.9**
Acute myeloid leukemia (AML)									
AML-1 → AML-2	t(15;17)	76.1	**100**	**100**	**100**	**100**	**100**	**100**	**100**
AML-1 → AML-3	t(15;17)	75	**100**	**100**	**100**	93.5	**100**	**100**	**100**
AML-2 → AML-1	t(15;17)	97.4	94.4	**99.8**	97	**99.8**	**99.8**	**99.8**	97
AML-2 → AML-3	t(15;17)	95	95	**100**	95	**100**	**100**	**100**	**100**
AML-3 → AML-1	t(15;17)	92.9	**100**	**100**	99.8	99.6	**99.8**	**99.8**	**99.8**
AML-3 → AML-2	t(15;17)	91.6	**100**	**100**	98.3	99.6	**100**	**100**	**100**
AML-1 → AML-2	inv(16)	82.7	92.4	88	**92.9**	92.9	**100**	**100**	**100**
AML-1 → AML-3	inv(16)	84.6	**100**	98.7	**100**	87.5	**100**	**100**	**100**
AML-2 → AML-1	inv(16)	90	**99.1**	80.1	98.1	96.8	**99.2**	**99.2**	**99.2**
AML-2 → AML-3	inv(16)	96.2	**100**	87.5	**100**	**100**	**100**	**100**	**100**
AML-3 → AML-1	inv(16)	**73.9**	61.5	52.4	68	96.6	**99.2**	**99.2**	93.8
AML-3 → AML-2	inv(16)	**72.4**	62.1	53.6	62.3	71.4	**100**	**92.9**	81.7
AML-1 → AML-2	t(8;21)	93.4	95.2	97.6	**100**	97.2	**100**	**100**	**100**
AML-2 → AML-1	t(8;21)	96.6	**99.8**	99.4	**99.8**	**100**	**100**	**100**	**100**
Breast cancer									
Breast-1 → Breast-2	ER-	83.9	**89.2**	74.5	86.6	**87.9**	85.9	75	83.2
Breast-1 → Breast-3	ER-	82.9	83.2	81	**83.4**	86.1	**88.7**	79.8	82.8
Breast-2 → Breast-1	ER-	60.4	56.3	70.9	**77.3**	81.6	**85.6**	81.4	72.9
Breast-2 → Breast-3	ER-	61	61.7	**76.1**	71.2	76.3	**83.5**	79.4	71.9
Breast-3 → Breast-1	ER-	79.3	**87.8**	**87.8**	**87.8**	81.7	93.8	91.8	**93.9**
Breast-3 → Breast-2	ER-	65.3	66.6	79.2	**80.6**	67.3	87.3	**87.9**	85.2
Lung cancer									
Lung-1 → Lung-2	AD	N/A[a]	N/A	N/A	N/A	84.2	91	**93**	92.2
Lung-1 → Lung-3	AD	N/A	N/A	N/A	N/A	76	93.5	93.2	**96.2**
Lung-2 → Lung-1	AD	N/A	N/A	N/A	N/A	83.4	88.3	88.3	**89.3**
Lung-2 → Lung-3	AD	N/A	N/A	N/A	N/A	87.2	**94.7**	93.5	93.2
Lung-3 → Lung-1	AD	N/A	N/A	N/A	N/A	88.1	88.2	85.4	**89.1**
Lung-3 → Lung-2	AD	N/A	N/A	N/A	N/A	92.3	**95**	91.6	92.2
Lung-1 → Lung-2	SQ	N/A	N/A	N/A	N/A	84.2	91.7	**93**	92.7
Lung-1 → Lung-3	SQ	N/A	N/A	N/A	N/A	78.5	92.2	91.9	**94.7**
Lung-2 → Lung-1	SQ	N/A	N/A	N/A	N/A	81.6	87.2	90	**91.2**
Lung-2 → Lung-3	SQ	N/A	N/A	N/A	N/A	84.7	93.5	94.7	**96**
Lung-3 → Lung-1	SQ	N/A	N/A	N/A	N/A	85.2	88.3	87.2	**89.1**
Lung-3 → Lung-2	SQ	N/A	N/A	N/A	N/A	91.6	**95**	91.6	93

In each row (an experiment), we boldface the value which gives the best balanced accuracy for expression values and rank levels, respectively.
(a) N/A, not available.

which used a combination of two data sets as the training data set. An exception occurs when the training data sets are the combination of ALL-1 and ALL-2 and the testing data set is ALL-3. The balanced accuracy is not as high as other ALL experiments. By investigating the number of overlapping genes in relevant gene lists selected by the four different feature selection methods (see the online supplemental Table S2), we find out that ALL-1 and ALL-2 are intrinsically more similar. More specifically, we consider the following three pairs of training sets:

1. ALL-1 versus (ALL-2 + ALL-3).
2. ALL-2 versus (ALL-1 + ALL-3).
3. ALL-3 versus (ALL-1 + ALL-2).

For example, if the first pair is used, we obtained top 100 relevant genes using ALL-1 and (ALL-2 + ALL-3), respectively. The number of overlapping genes of the third pair is much smaller than those of the other two. This result indicates that ALL-3 is very different from ALL-1 and ALL-2. We further confirm from Table 1 that ALL-1 and ALL-2 are from the same lab which is different from the source of ALL-3. Hence, while combining more data sets for training is often useful, sometimes the combined set becomes more noisy or biased toward certain distributions. Then the results of feature selection and classification may be less stable.

Using Tables 2 and 3, we give an overall comparison of the four methods. We let the balanced accuracy in Tables 2

TABLE 3
A Comparison of Different Feature Selection Methods by Training on Two Data Sets and Testing on the Remaining One: Using Expression Values and Rank Levels

Training→Testing	Cancer subtype	Expression values				Rank levels			
		t-test	SAM	RP	RF	*t*-test	SAM	RP	RF
Acute lymphoblastic leukemia (ALL)									
ALL-2+ALL-3 → ALL-1	t(12;21)	71.6	**97**	96.6	**97**	**97**	**97**	**97**	**97**
ALL-1+ALL-3 → ALL-2	t(12;21)	87.7	**98.8**	98.2	97.6	97.6	97.6	**98.8**	97.6
ALL-1+ALL-2 → ALL-3	t(12;21)	**56.3**	52.3	51.2	55.8	52.3	81.4	**87.4**	83.7
ALL-2+ALL-3 → ALL-1	t(1;19)	90.8	**96.8**	**96.8**	**96.8**	96.3	**96.8**	**96.8**	**96.8**
ALL-1+ALL-3 → ALL-2	t(1;19)	90.2	**100**	**100**	**100**	**100**	**100**	**100**	**100**
ALL-1+ALL-2 → ALL-3	t(1;19)	58.5	50	**73.1**	69.2	65.4	**84.6**	**84.6**	80.8
ALL-2+ALL-3 → ALL-1	HD>50	86.5	**93.7**	93.2	**93.7**	96.2	**96.6**	**96.6**	**96.6**
ALL-1+ALL-3 → ALL-2	HD>50	68.2	86.2	84.6	**88.9**	93.3	93.9	**94.5**	91.5
ALL-1+ALL-2 → ALL-3	HD>50	61.4	58	**63.6**	56.8	60.5	**80.7**	63.6	58
Acute myeloid leukemia (AML)									
AML-2+AML-3 → AML-1	t(15;17)	98.1	99.8	**100**	99.8	97	**99.8**	**99.8**	**99.8**
AML-1+AML-3 → AML-2	t(15;17)	91	**100**	**100**	**100**	**100**	**100**	**100**	**100**
AML-1+AML-2 → AML-3	t(15;17)	98.5	**100**	**100**	**100**	95	**100**	**100**	**100**
AML-2+AML-3 → AML-1	inv(16)	87.8	98.9	93.2	**99.4**	98.9	**99.2**	**99.2**	**99.2**
AML-1+AML-3 → AML-2	inv(16)	80.8	**92.9**	87.1	**92.9**	92.9	**100**	**100**	**100**
AML-1+AML-2 → AML-3	inv(16)	94.9	**100**	**100**	**100**	**100**	**100**	**100**	**100**
AML-2+AML-3 → AML-1	t(8;21)	82.4	99.8	99.6	**100**	**100**	**100**	**100**	**100**
AML-1+AML-3 → AML-2	t(8;21)	90	**97.6**	**97.6**	**97.6**	94.8	**100**	**100**	**100**
Breast cancer									
Breast-2+Breast-3 → Breast-1	ER-	70.9	83.4	**87.6**	**87.6**	81.2	**89.7**	83.4	87.7
Breast-1+Breast-3 → Breast-2	ER-	75.1	**90**	83.2	74	79.1	**89.9**	82.6	87.9
Breast-1+Breast-2 → Breast-3	ER-	66.6	76	61.4	**81**	68.2	85.6	80.4	**86.3**
Lung cancer									
Lung-2+Lung-3 → Lung-1	AD	N/A[a]	N/A	N/A	N/A	75.3	**90.1**	89.2	90
Lung-1+Lung-3 → Lung-2	AD	N/A	N/A	N/A	N/A	88.8	92.2	91.7	**92.3**
Lung-1+Lung-2 → Lung-3	AD	N/A	N/A	N/A	N/A	**93.5**	93.2	90.4	93.2
Lung-2+Lung-3 → Lung-1	SQ	N/A	N/A	N/A	N/A	76.2	**90.1**	89.2	90
Lung-1+Lung-3 → Lung-2	SQ	N/A	N/A	N/A	N/A	88.8	92.2	91.7	**93**
Lung-1+Lung-2 → Lung-3	SQ	N/A	N/A	N/A	N/A	93.5	93.2	90.4	**96**

In each row (an experiment), we boldface the value which gives the best balanced accuracy for expression values and rank levels, respectively. (a) N/A, not available.

and 3 be subtracted by the best one in each experiment. Then in a particular cancer type, the values for each feature selection method are averaged and presented in Table 4. In Table 4, RF has the best average performance of balanced accuracy in using expression values, but SAM has better results after rank-based normalization.

3.2 The Influence of Rank-Based Normalization

In Section 2.1.1, we introduced the rank-based normalization. In our previous paper [5], we indicated that the classification accuracy with rank-based normalization was better than using expression values when SAM was employed for feature selection. In this study, we further demonstrate that rank-based normalization is also better when *t*-test, RP, and RF are employed.

Fig. 2 shows that when expression values are used, balanced accuracy of all four methods fluctuate. In contrast, the curves are more stable if rank-based normalization is applied. SAM, RP, and RF even approach 100 percent of balanced accuracy. In the online supplemental Figure S1, SAM, RP, and RF have achieved nearly 100 percent of balanced accuracy when expression values are used, but these methods also give very high accuracy after applying rank-based normalization. For *t*-test, the classification performance has also improved after rank-based normalization in Fig. 2 and the online supplemental Fig. S1.

Although almost all figures and tables show that the performance improves with rank-based normalization, no improvement is observed for breast cancer (see the online supplemental Fig. S8). The reason may be that the three data sets of breast cancer came from different generations of Affy arrays. The number of common features of the three generations is only 5,045, which is much lower than 22,283 of other cancer types.

In Table 4, we find that *t*-test and SAM have greater improvement in the four methods after rank-based normalization. In contrast, RP and RF have only minor improvement after rank-based normalization. The reason for RP might be that its algorithm ranks the fold-change ratio. Applying rank-based normalization will rescale the gene expression so that the effect of RP decreases. Therefore, it does not benefit from rank-based normalization. Further discussion of RF will be in Section 3.6.

TABLE 4
The Average of Performance Difference to the Best Balanced Accuracy of Each Row for a Particular Feature Selection Method in Tables 2 and 3

Expression values				Rank levels			
t-test	SAM	RP	RF	*t*-test	SAM	RP	RF
Average of acute lymphoblastic leukemia (ALL)							
-12.41	-4.21	-2.40	-2.47	-5.38	-1.05	-1.77	-2.77
Average of acute myeloid leukemia (AML)							
-8.48	-1.78	-4.24	-1.22	-3.80	0	-0.32	-1.20
Average of breast cancer							
-7.87	-10.98	-4.94	-1.86	-9.34	-0.38	-5.75	-4.62
Average of lung cancer							
N/A[a]	N/A	N/A	N/A	-7.68	-1.20	-1.96	-0.58
Weighted average[b]							
-10.21	-4.34	-3.49	-1.90	-5.94	-0.70	-1.87	-2.02

(a) N/A, not available. (b) The average of all experiments for a particular feature selection method.

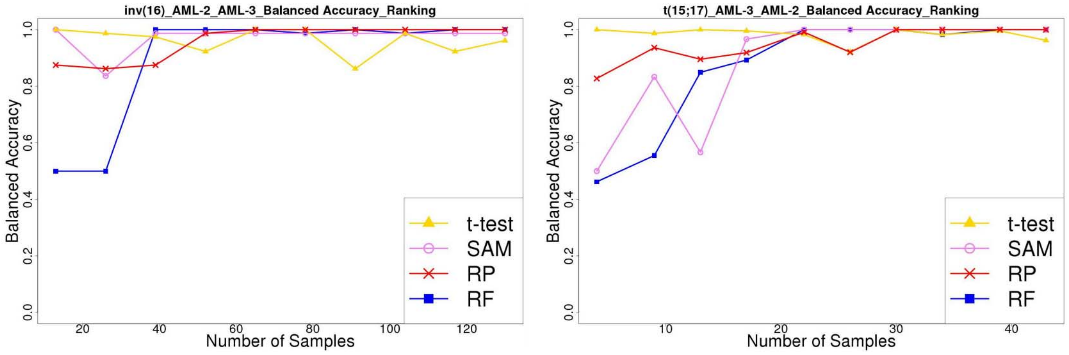

Fig. 3. The plots of balanced accuracy versus training sample size. The main title of each figure stands for "predicted subtype_training data set_testing data set_evaluation method_ranking or expression value."

3.3 The Influence of the Number of Training Samples and the Number of Selected Relevant Genes

An essential issue in feature selection is the minimal number of training sample size required to achieve reliable performance. We chose two data sets from the same cancer type: one as training and the other as testing, to study the effect of the training sample size on the balanced accuracy. For a specific number of training samples, we randomly selected a subset, identified the corresponding gene list of the top 250 relevant genes from one of the four feature selection methods (t-test, SAM, RP, and RF), and obtained the balanced accuracy on the test set. Fig. 3 are the plots of the balanced accuracy versus the sample size for inv(16) and t(15;17) subtypes of AML, respectively. The experimental results show a common trend that the balanced accuracy rises up until a specific number of samples [40 for inv(16) and 20 for t(15;17)]. It is concluded here that the minimal required number of sample size to achieve the best balanced accuracy varies in different experiments. Besides, according to Fig. 2, when using a bigger training data set (i.e., ALL-1+ALL-3 or ALL-2+ALL-3), both expression level and rank-based data yield higher balanced accuracy. This might explain the poor performance of only adopting ALL-3 as the training data set when using expression level data since the number of samples is only 43. The sample size may not severely impact the balanced accuracy when data are applied with rank-based normalization, because rank-based normalization has already largely improved the performance of each feature selection method.

Another important issue in feature selection is to select an optimal number of relevant genes. Li et al. [43] suggested 150 genes in ALL data set when t-test was applied. Beer et al. [44] and Bhattacharjee et al. [45] reported 50 and 175 genes, respectively, in lung cancer data. In Fig. 2 and the online supplemental Figs. S2-S19, we have conducted experiments by varying the number of selected relevant genes from 5 to 250. There is no obvious number of selected relevant genes whose performance is the best. In contrast, with rank-based normalization, when the number of selected relevant genes is less than 100, balanced accuracy increases with the number of genes. However, when the number of selected relevant genes is larger than 100, the performance is similar. Therefore, if rank levels are used, selecting 100 relevant genes may be appropriate in cross-lab microarray data analysis.

3.4 The Issue of Unbalanced Data Sets

The unbalanced number of samples in different subtypes can easily mislead experimental results. In this study, the application of balanced accuracy is the first step to address this issue, but additional cares may be needed.

Data imbalance affects the prediction performance more if rank-based normalization is not applied. In Table 2, when the training set is AML-3 and expression values are used, the balanced accuracy for predicting inv(16) subtype of AML is very low. The data set AML-3 has only four inv(16) samples out of 43. The result is better with rank-based normalization. Another unbalanced set is Breast-2, which has only 15 ER$-$ samples out of 89. When Breast-2 is trained to predict Breast-3, the balanced accuracy is the worst among the breast cancer data sets.

Besides, we found that the selection of k in kNN is related to the issue of an unbalanced number of samples. When considering the LOO CV balanced accuracy to select k, more than one k values may lead to the same best balanced accuracy. In this situation, we choose the minimum k. If we choose a larger k rather than the minimum one, the prediction performance may become much worse. The reason is that for unbalanced data, the number of training samples of the predicting subtype may be less than k. Then, most of the nearest neighbors of a testing sample may be in the majority class, so testing samples may never be predicted as in the minor class. This explains our decision to select the minimum k when more than one value has the same LOO CV best balanced accuracy.

3.5 The Comparison Between t-Test and SAM

The two methods, t-test and SAM, follow similar statistical principles, where t-test calculates the t statistic for each gene, but SAM calculates the relative difference. However, the performance of t-test is much worse than that of SAM. In particular, when the number of selected relevant genes is small, t-test has even inferior balanced accuracy (see the online supplemental Figs. S2-S19). A crucial difference between these two methods is that SAM has the important step of permutation, which calculates the expected relative difference to guess the underlying distribution. To confirm the importance of permutation, we run SAM with a smaller number of permutation steps than the default 300. The results become worse (data not shown).

3.6 Stability of Feature Selection Methods

Both SAM and RF have good performance in most cancer subtypes. However, RF has considerably lower balanced accuracy than SAM in the two experiments in Table 2: 1) The training/testing data sets are ALL-1/ALL-3 and the predicted subtype is $\mathrm{HD} > 50$. 2) The training/testing data sets are Breast-2/Breast-1 and the predicted subtype is ER−. It is observed that the top five relevant genes selected by SAM is not on the list of 250 relevant genes generated by RF. We replace the top five relevant genes of RF with the top five relevant genes of SAM. The balanced accuracy is increased from 80.7 to 89.32 percent in $\mathrm{HD} > 50$ prediction and 72.9 to 89.83 percent in ER− prediction, respectively. The results seem to indicate that SAM performs better in choosing relevant genes than RF.

To compare the stability of these feature selection methods, we run the whole procedure five times. For every two relevant gene lists, we analyze the overlapping percentage. The average of the 10 pairs of relevant gene lists is in the online suppelental Table S3. The ratio of RF is only about 30 percent, but those of other methods are close to 100 percent for almost all cancer subtypes. Therefore, RF may not be as stable as other three feature selection methods in cross-lab microarray studies.

We further investigate why the performance of RF is unstable. Contrast with SAM and RP, which put all genes into consideration on every permutation, RF only analyzes a group of genes in a tree. The default number is the square root of the number of features. After we increase the number of considered features, RF has more stable accuracy (data not shown).

4 Conclusions

In this paper, we have proposed a procedure to evaluate feature selection methods for cross-lab Affy microarray data. We demonstrated that rank-based normalization lead to a better selection of relevant genes and higher balanced accuracy. Among the four feature selection methods experimented in this study, SAM performs the best. RF also gets high classification accuracy, but it is not as stable as SAM. RP comes close to RF but does not benefit from rank-based normalization. The worst method is t-test, whose performance is far behind the other three methods.

We conclude that with a careful selection of data sets, a cross-lab analysis like ours can give valuable results. Our data and codes are available upon request for reproducing experimental results.

Acknowledgments

This work was supported by grant MMH-E-101009 from the Mackay Memorial Hospital. The authors would like to thank Mei-Ju May Chen, Hao-Han Liaw, and Yun-Chieh Sung for their great help in collecting the required information and data.

References

[1] A. Richard and Young, "Biomedical Discovery with DNA Arrays," *Cell*, vol. 102, pp. 9-15, 2000.

[2] J. Quackenbush, "Computational Analysis of Microarray Data," *Nature Rev. Genetics*, vol. 2, pp. 418-427, 2001.

[3] T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield, and E.S. Lander, "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring," *Science*, vol. 286, pp. 531-537, 1999.

[4] G. Bloom, I.V. Yang, D. Boulware, K.Y. Kwong, D. Coppola, S. Eschrich, J. Quackenbush, and T.J. Yeatman, "Multi-Platform, Multi-Site, Microarray-Based Human Tumor Classification," *Am. J. Pathology*, vol. 164, pp. 9-16, 2004.

[5] H.-C. Liu, C.-Y. Chen, Y.-T. Liu, C.-B. Chu, D.-C. Liang, L.-Y. Shih, and C.-J. Lin, "Cross-Generation and Cross-Laboratory Predictions of Affymetrix Microarrays by Rank-Based Methods," *J. Biomedical Informatics*, vol. 41, pp. 510-519, 2008.

[6] C.-C.C. Liu, J. Hu, M. Kalakrishnan, H. Huang, and X.J.J. Zhou, "Integrative Disease Classification Based on Cross-Platform Microarray Data," *BMC Bioinformatics*, vol. 10, article S25, 2009.

[7] H. Jiang, Y. Deng, H.-S. Chen, L. Tao, Q. Sha, J. Chen, C.-J. Tsai, and S. Zhang, "Joint Analysis of Two Microarray Gene-Expression Data Sets to Select Lung Adenocarcinoma Marker Genes," *BMC Bioinformatics*, vol. 5, article 81, 2004.

[8] L. Xu, A.C. Tan, D.Q. Naiman, D. Geman, and R.L. Winslow, "Robust Prostate Cancer Marker Genes Emerge from Direct Integration of Inter-Study Microarray Data," *Bioinformatics*, vol. 21, pp. 3905-3911, 2005.

[9] A.C. Tan, D.Q. Naiman, L. Xu, R.L. Winslow, and D. Geman, "Simple Decision Rules for Classifying Human Cancers from Gene Expression Profiles," *Bioinformatics*, vol. 21, pp. 3896-3904, 2005.

[10] J.T. Leek, R.B. Scharpf, H.A.C. Bravo, D. Simcha, B. Langmead, W.E. Johnson, D. Geman, K. Baggerly, and R.A. Irizarry, "Tackling the Widespread and Critical Impact of Batch Effects in High-Throughput Data," *Nature Rev. Genetics*, vol. 11, pp. 733-739, Sept. 2010.

[11] J. Luo, M. Schumacher, A. Scherer, D. Sanoudou, D. Megherbi, T. Davison, T. Shi, W. Tong, L. Shi, H. Hong, C. Zhao, F. Elloumi, W. Shi, R. Thomas, S. Lin, G. Tillinghast, G. Liu, Y. Zhou, D. Herman, Y. Li, Y. Deng, H. Fang, P. Bushel, M. Woods, and J. Zhang, "A Comparison of Batch Effect Removal Methods for Enhancement of Prediction Performance Using MAQC-II Microarray Gene Expression Data," *Pharmacogenomics J.*, vol. 10, no. 4, pp. 278-291, Aug. 2010.

[12] Y. Saeys, I. Inza, and P. Larrañaga, "A Review of Feature Selection Techniques in Bioinformatics," *Bioinformatics*, vol. 23, pp. 2507-2517, 2007.

[13] P. Jafari and F. Azuaje, "An Assessment of Recently Published Gene Expression Data Analyses: Reporting Experimental Design and Statistical Factors," *BMC Medical Informatics and Decision Making*, vol. 6, article 27, 2006.

[14] P. Baldi and A.D. Long, "A Bayesian Framework for the Analysis of Microarray Expression Data: Regularized *t*-Test and Statistical Inferences of Gene Changes," *Bioinformatics*, vol. 17, pp. 509-519, 2001.

[15] V.G. Tusher, R. Tibshirani, and G. Chu, "Significance Analysis of Microarrays Applied to the Ionizing Radiation Response," *Proc. Nat'l Academy of Science USA*, vol. 98, pp. 5116-5121, 2001.

[16] J.G. Thomas, J.M. Olson, S.J. Tapscott, and L.P. Zhao, "An Efficient and Robust Statistical Modeling Approach to Discover Differentially Expressed Genes Using Genomic Expression Profiles," *Genome Research*, vol. 11, pp. 1227-1236, July 2001.

[17] F. Hong, R. Breitling, C.W. McEntee, B.S. Wittner, J.L. Nemhauser, and J. Chory, "RankProd: A Bioconductor Package for Detecting Differentially Expressed Genes in Meta-Analysis," *Bioinformatics*, vol. 22, pp. 2825-2827, Nov. 2006.

[18] R. Diaz-Uriarte and S.A. de Andres, "Gene Selection and Classification of Microarray Data Using Random Forest," *BMC Bioinformatics*, vol. 7, article 3, 2006.

[19] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene Selection for Cancer Classification Using Support Vector Machines," *Machine Learning*, vol. 46, pp. 389-422, 2002.

[20] S. Ma and J. Huang, "Regularized ROC Method for Disease Classification and Biomarker Selection with Microarray Data," *Bioinformatics*, vol. 21, pp. 4356-4362, 2005.

[21] L. Shi, G. Campbell, W.D. Jones, F. Campagne, Z. Wen, S.J. Walker, Z. Su, T.-M.M. Chu, F.M. Goodsaid, L. Pusztai, J.D. Shaughnessy, A. Oberthuer, R.S. Thomas, R.S. Paules, M. Fielden, B. Barlogie, W. Chen, P. Du, M. Fischer, C. Furlanello, B.D. Gallas, X. Ge, D.B. Megherbi, W.F. Symmans, M.D. Wang, J. Zhang, H. Bitter, B. Brors, P.R. Bushel, M. Bylesjo, M. Chen, J. Cheng, J. Cheng, J. Chou, T.S. Davison, M. Delorenzi, Y. Deng, V. Devanarayan, D.J. Dix, J. Dopazo, K.C. Dorff, F. Elloumi, J. Fan, S. Fan, X. Fan, H. Fang, N. Gonzaludo, K.R. Hess, H. Hong, J. Huan, R.A. Irizarry, R. Judson, D. Juraeva, S. Lababidi, C.G. Lambert, L. Li, Y. Li, Z. Li, S.M. Lin, G. Liu, E.K. Lobenhofer, J. Luo, W. Luo, M.N. McCall, Y. Nikolsky, G.A. Pennello, R.G. Perkins, R. Philip, V. Popovici, N.D. Price, F. Qian, A. Scherer, T. Shi, W. Shi, J. Sung, D. Thierry-Mieg, J. Thierry-Mieg, V. Thodima, J. Trygg, L. Vishnuvajjala, S.J.J. Wang, J. Wu, Y. Wu, Q. Xie, W.A. Yousef, L. Zhang, X. Zhang, S. Zhong, Y. Zhou, S. Zhu, D. Arasappan, W. Bao, A.B.B. Lucas, F. Berthold, R.J. Brennan, A. Buness, J.G. Catalano, C. Chang, R. Chen, Y. Cheng, J. Cui, W. Czika, F. Demichelis, X. Deng, D. Dosymbekov, R. Eils, Y. Feng, J. Fostel, S. Fulmer-Smentek, J.C. Fuscoe, L. Gatto, W. Ge, D.R. Goldstein, L. Guo, D.N. Halbert, J. Han, S.C. Harris, C. Hatzis, D. Herman, J. Huang, R.V. Jensen, R. Jiang, C.D. Johnson, G. Jurman, Y. Kahlert, S.A. Khuder, M. Kohl, J. Li, L. Li, M. Li, Q.-Z. Z. Li, S. Li, Z. Li, J. Liu, Y. Liu, Z. Liu, L. Meng, M. Madera, F. Martinez-Murillo, I. Medina, J. Meehan, K. Miclaus, R.A. Moffitt, D. Montaner, P. Mukherjee, G.J. Mulligan, P. Neville, T. Nikolskaya, B. Ning, G.P. Page, J. Parker, R.M. Parry, X. Peng, R.L. Peterson, J.H. Phan, B. Quanz, Y. Ren, S. Riccadonna, A.H. Roter, F.W. Samuelson, M.M. Schumacher, J.D. Shambaugh, Q. Shi, R. Shippy, S. Si, A. Smalter, C. Sotiriou, M. Soukup, F. Staedtler, G. Steiner, T.H. Stokes, Q. Sun, P.-Y.Y. Tan, R. Tang, Z. Tezak, B. Thorn, M. Tsyganova, Y. Turpaz, S.C. Vega, R. Visintainer, J. von Frese, C. Wang, E. Wang, J. Wang, W. Wang, F. Westermann, J.C. Willey, M. Woods, S. Wu, N. Xiao, J. Xu, L. Xu, L. Yang, X. Zeng, J. Zhang, L. Zhang, M. Zhang, C. Zhao, R.K. Puri, U. Scherf, W. Tong, R.D. Wolfinger and MAQC Consortium, "The MicroArray Quality Control (MAQC)-II Study of Common Practices for the Development and Validation of Microarray-Based Predictive Models," *Nature Biotechnology,* vol. 28, no. 8, pp. 827-838, Aug. 2010.

[22] R. Edgar, M. Domrachev, and A.E. Lash, "Gene Expression Omnibus: NCBI Gene Expression and Hybridization Array Data Repository," *Nucleic Acids Research,* vol. 30, pp. 207-210, 2002.

[23] T. Cover and P. Hart, "Nearest Neighbor Pattern Classification," *IEEE Trans. Information Theory,* vol. 13, no. 1, pp. 21-27, Jan. 1967.

[24] M.E. Ross, X. Zhou, G. Song, S.A. Shurtleff, K. Girtman, W.K. Williams, H.-C. Liu, R. Mahfouz, S.C. Raimondi, N. Lenny, A. Patel, and J.R. Downing, "Classification of Pediatric Acute Lymphoblastic Leukemia by Gene Expression Profile," *Blood,* vol. 102, pp. 2951-2959, 2003.

[25] C.G. Mullighan, X. Su, J. Zhang, I. Radtke, L.A.A. Phillips, C.B. Miller, J. Ma, W. Liu, C. Cheng, B.A. Schulman, R.C. Harvey, I.-M. Chen, R.J. Clifford, W.L. Carroll, G. Reaman, W.P. Bowman, M. Devidas, D.S. Gerhard, W. Yang, M.V. Relling, S.A. Shurtleff, D. Campana, M.J. Borowitz, C.-H. Pui, M. Smith, S.P. Hunger, C.L. Willman, and J.R. Downing, "Deletion of IKZF1 and Prognosis in Acute Lymphoblastic Leukemia," *New England J. Medicine,* vol. 360, pp. 470-480, 2009.

[26] M. Den Boer, M. van Slegtenhorst, R. De Menezes, M. Cheok, J. Buijs-Gladdines, S. Peters, L. Van Zutven, H. Beverloo, P. Van der Spek, G. Escherich, M. Horstmann, G. Janka-Schaub, W. Kamps, W. Evans, and R. Pieters, "A Subtype of Childhood Acute Lymphoblastic Leukaemia with Poor Treatment Outcome: A Genome-Wide Classification Study," *Lancet Oncology,* vol. 10, pp. 125-134, 2009.

[27] P.J.M. Valk, R.G.W. Verhaak, M.A. Beijen, C.A.J. Erpelinck, S. Barjesteh van Waalwijk van Doorn-Khosrovani, J.M. Boer, H.B. Beverloo, M.J. Moorhouse, P.J. van der Spek, B. Löwenberg, and R. Delwel, "Prognostically Useful Gene-Expression Profiles in Acute Myeloid Leukemia," *New England J. Medicine,* vol. 350, pp. 1617-1628, 2004.

[28] N.C. Gutiérrez, R. López-Pérez, J.M. Hernández, I. Isidro, B. González, M. Delgado, E. Fermiñán, J.L. García, L. Vázquez, M. González, and J.F.S. Miguel, "Gene Expression Profile Reveals Deregulation of Genes with Relevant Functions in the Different Subclasses of Acute Myeloid Leukemia," *Leukemia,* vol. 19, pp. 402-409, 2005.

[29] M.E. Ross, R. Mahfouz, M. Onciu, H.-C. Liu, X. Zhou, G. Song, S.A. Shurtleff, S. Pounds, C. Cheng, J. Ma, R.C. Ribeiro, J.E. Rubnitz, K. Girtman, W.K. Williams, S.C. Raimondi, D.-C. Liang, L.-Y. Shih, C.-H. Pui, and J.R. Downing, "Gene Expression Profiling of Pediatric Acute Myelogenous Leukemia," *Blood,* vol. 104, pp. 3679-3687, 2004.

[30] M. West, C. Blanchette, H. Dressman, E. Huang, S. Ishida, R. Spang, H. Zuzan, J. John, A. Olson, J.R. Marks, and J.R. Nevins, "Predicting the Clinical Status of Human Breast Cancer by Using Gene Expression Profiles," *Proc. Nat'l Academy of Sciences USA,* vol. 98, pp. 11462-11467, 2001.

[31] E. Huang, S.H. Cheng, H. Dressman, J. Pittman, M.H. Tsou, C.F. Horng, A. Bild, E.S. Iversen, M. Liao, C.M. Chen, M. West, J.R. Nevins, and A.T. Huang, "Gene Expression Predictors of Breast Cancer Outcomes," *Lancet,* vol. 361, pp. 1590-1596, 2003.

[32] Y. Wang, J.G.M. Klijn, Y. Zhang, A.M. Sieuwerts, M.P. Look, F. Yang, D. Talantov, M. Timmermans, M.E.M. van Gelder, J. Yu, T. Jatkoe, E.M.J.J. Berns, D. Atkins, and J.A. Foekens, "Gene-Expression Profiles to Predict Distant Metastasis of Lymph-Node-Negative Primary Breast Cancer," *Lancet,* vol. 365, pp. 671-679, 2005.

[33] A.H. Bild, G. Yao, J.T. Chang, Q. Wang, A. Potti, D. Chasse, M.-B. Joshi, D. Harpole, J.M. Lancaster, A. Berchuck, J.A. Olson, J.R. Marks, H.K. Dressman, M. West, and J.R. Nevins, "Oncogenic Pathway Signatures in Human Cancers As a Guide to Targeted Therapies," *Nature,* vol. 439, pp. 353-357, 2006.

[34] E.-S. Lee, D.-S. Son, S.-H. Kim, J. Lee, J. Jo, J. Han, H. Kim, H.J. Lee, H.Y. Choi, Y. Jung, M. Park, Y.S. Lim, K. Kim, Y.M. Shim, B.C. Kim, K. Lee, N. Huh, C. Ko, K. Park, J.W. Lee, Y.S. Choi, and J. Kim, "Prediction of Recurrence-Free Survival in Postoperative Non-Small Cell Lung Cancer Patients by Using an Integrated Model of Clinical Information and Gene Expression," *Clinical Cancer Research,* vol. 14, pp. 7397-7404, 2008.

[35] R. Kuner, T. Muley, M. Meister, M. Ruschhaupt, A. Buness, E.C. Xu, P. Schnabel, A. Warth, A. Poustka, H. Sültmann, and H. Hoffman, "Global Gene Expression Analysis Reveals Specific Patterns of Cell Junctions in Non-Small Cell Lung Cancer Subtypes," *Lung Cancer,* vol. 63, pp. 32-38, 2009.

[36] R.W. Tothill, A. Kowalczyk, and D. Rischin, "An Expression-Based Site of Origin Diagnostic Method Designed for Clinical Application to Cancer of Unknown Origin," *Cancer Research,* vol. 65, pp. 4031-4040, 2005.

[37] P. Warnat, R. Eils, and B. Brors, "Cross-Platform Analysis of Cancer Microarray Data Improves Gene Expression Based Classification of Phenotypes," *BMC Bioinformatics,* vol. 6, article 265, 2005.

[38] X. Qiu, A.I. Brooks, L. Klebanov, and N. Yakovlev, "The Effects of Normalization on the Correlation Structure of Microarray Data," *BMC Bioinformatics,* vol. 6, article 120, 2005.

[39] A. Tsodikov, A. Szabo, and D. Jones, "Adjustments and Measures of Differential Expression for Microarray Data," *Bioinformatics,* vol. 18, pp. 251-260, Feb. 2002.

[40] G.K. Smyth, "Limma: Linear Models for Microarray Data," *Bioinformatics and Computational Biology Solutions using R and Bioconductor,* R. Gentleman, V. Carey, S. Dudoit, R. Irizarry, and W. Huber, eds., pp. 397-420, Springer, 2005.

[41] R. Breitling, P. Armengaud, A. Amtmann, and P. Herzyk, "Rank Products: A Simple, Yet Powerful, New Method to Detect Differentially Regulated Genes in Replicated Microarray Experiments," *FEBS Letters,* vol. 573, pp. 83-92, Aug. 2004.

[42] L. Breiman, "Random Forests," *Machine Learning,* vol. 45, pp. 5-32, 2001.

[43] T. Li, C. Zhang, and M. Ogihara, "A Comparative Study of Feature Selection and Multiclass Classification Methods for Tissue Classification Based on Gene Expression," *Bioinformatics,* vol. 20, pp. 2429-2437, 2004.

[44] D.G. Beer, S.L. Kardia, C.-C. Huang, T.J. Giordano, A.M. Levin, D.E. Misek, L. Lin, G. Chen, T.G. Gharib, D.G. Thomas, M.L. Lizyness, R. Kuick, S. Hayasaka, J.M. Taylor, M.D. Iannettoni, M.B. Orringer, and S. Hanash, "Gene-Expression Profiles Predict Survival of Patients with Lung Adenocarcinoma," *Nature Medicine,* vol. 8, pp. 816-824, 2002.

[45] A. Bhattacharjee, W.G. Richards, J. Staunton, C. Li, S. Monti, P. Vasa, C. Ladd, J. Beheshti, R. Bueno, M. Gillette, M. Loda, G. Weber, E.J. Mark, E.S. Lander, W. Wong, B.E. Johnson, T.R. Golub, D.J. Sugarbaker, and M. Meyerson, "Classification of Human Lung Carcinomas by mRNA Expression Profiling Reveals Distinct Adenocarcinoma Subclasses," *Proc. Nat'l Academy of Sciences USA*, vol. 98, pp. 13 790-13 795, Nov. 2001.

Hsi-Che Liu received the MD degree from Chung-Shan Medical University, Taichung, Taiwan, in 1991. He is currently the chief of the Division of Pediatric Hematology/Oncology, Mackay Memorial Hospital, Taipei, Taiwan. He is an assistant professor in the Department of Medicine, Mackay Medical College, New Taipei, Taiwan. His major research interests include pediatric oncology/hematology and gene expression analysis. He has published 50 peer-reviewed papers.

Pei-Chen Peng received the BS degree from the National Taiwan University, Taipei, in 2011. She is working toward the master's degree in the Department of Computer Science, National Taiwan University. Her research interests include gene expression analysis and gene regulatory network construction.

Tzung-Chien Hsieh received the BS degree from the National Taiwan University, Taipei, in 2011. He is currently working toward the master's degree in the Department of Computer Science, National Taiwan University. His research interests include gene expression analysis and next-generation sequence analysis.

Ting-Chi Yeh received the MD degree from the China Medical University, Taichung, Taiwan, in 1998. He is currently an attending physician in the Division of Pediatric Hematology/Oncology, Mackay Memorial Hospital, Taipei, Taiwan. His major research interests include pediatric oncology and pediatric hematology.

Chih-Jen Lin received the BS degree from the National Taiwan University, Taipei, in 1993 and the PhD degree from the University of Michigan, Ann Arbor, in 1998. He is currently a distinguished professor in the Department of Computer Science, National Taiwan University. His major research interests include machine learning, data mining, and numerical optimization. He is best known for his work on support vector machines (SVM) for data classification. His software LIBSVM is one of the most widely used and cited SVM packages. Nearly all major companies apply his software for classification and regression applications. He has received many awards for his research work, including the ACM KDD 2010 Best Paper Award. He is an ACM Distinguished Scientist for his contribution to machine learning algorithms and software design. He is a fellow of the IEEE.

Chien-Yu Chen received the BS degree in electrical engineering from the National Taiwan University, Taipei, in 1996, the MS degree in electrical engineering from Stanford University, California, in 1998, and the PhD degree in computer science and information engineering from the National Taiwan University, in 2003. She is currently an associate professor in the Department of Bio-Industrial Mechatronics Engineering, National Taiwan University. As a computational biologist, her research centers on two problems: protein sequence/structure analysis and gene regulatory network modeling. Several useful computational tools have been developed and published in the past years, and all of them are available on the web.

Jen-Yin Hou received the MD degree from the National Yang-Ming University, Taipei, Taiwan, in 2004. He is currently an attending physician in the Division of Pediatric Hematology/Oncology, Mackay Memorial Hospital, Taipei, Taiwan. His major research areas include pediatric oncology and pediatric hematology.

Lee-Yung Shih received the MD degree from the National Taiwan University in 1973. She is currently an attending physician in the Division of Hematology/Oncology, Department of Internal Medicine, Chang Gung Memorial Hospital, Taipei, Taiwan. She is a professor at the School of Medicine, Chang Gung University, Taoyuan, Taiwan. Her major research interests include medical oncology, hematology, and molecular genetics/diagnostics of hematologic malignancies. She published more than 200 peer-reviewed papers.

Der-Cherng Liang received the MD degree from the National Taiwan University, Taipei, in 1972. He is currently an attending physician in the Division of Pediatric Hematology/Oncology, Mackay Memorial Hospital, Taipei, Taiwan. He is a professor at the School of Medicine, National Taiwan University. His major research interests include pediatric oncology/hematology and molecular genetics of leukemia. He has published more than 150 peer-reviewed papers.