Introduction
Cancer is the second most common cause of death in the USA, which accounts for nearly one out of four deaths. In 2014, about 585,720 Americans are projected to die of cancer. A key challenge of cancer treatment is the classification of cancer to its correct subtype. Applying cancer subtype-specific treatment increases efficacy and reduces toxicity. However, classification of cancer is a challenging task. As a result, developments in cancer classification have been central to the advancements in medical treatment. Traditional classification techniques are mainly based on biological insights and morphological appearances of the tumor. Existing approaches in this category, however, have serious limitations, and they yield low prediction accuracies. Cancers with alike morphological appearances can follow significantly different clinical courses and show different responses to therapy.

Expression levels of the human genes show cancer-type specific variations. Because of such variations, the gene expression levels collected from patients provide great potential to improve the accuracy of cancer classification. Currently, expression levels of thousands of genes can be measured in parallel using experimental techniques such as microarrays or RNA-seq. Microarray technology in comparison to measurement of other cancer markers such as chromatin states is usually experimentally easier and slightly cheaper. Because of these reasons, microarrays measuring expression levels of genes at the genomic scale are often preferred for cancer classification studies.
Computational methods for predicting cancer type using gene expression levels are relatively new strategies with a promise of significantly better accuracy compared to the classical methods.\textsuperscript{1,8,9} They provide an alternative, cheaper, and more efficient approach to the low-efficiency traditional cancer classification techniques. They are not meant to replace the existing morphology-based approaches. However, with the advances in technology for collecting data and the algorithms for studying this data, we believe that the computational cancer classification methods will provide an additional and useful resource in clinical practice. Computational classification methods often build a classifier from a dataset called the training dataset. The class labels of all the samples in the training dataset are known in advance. Once the classifier is built, it assigns each new sample that is not in the training dataset to one of the possible classes to that sample.

Many supervised statistical learning algorithms such as decision trees, k-nearest neighbor (kNN), naïve Bayes (NB), support vector machines (SVM), and random forest (RF) have been used for the classification of cancer using gene expression datasets.\textsuperscript{10,11} Most of these traditional classification methods depend on the expression levels of individual genes. For example, kNN classification method uses expression levels of hundreds of genes to classify the samples to distinct cancer classes. However, usually gene expression levels show high variations in many subtypes of cancer. Leukemia subtypes, for instance, belong to this category. Furthermore, cancer can alter the gene expression because of primary and secondary effects. Primary effects indicate the transcriptional changes as a result of genetic and epigenetic mutations. Secondary effects indicate the indirect transcriptional changes arising from regulatory interactions of genes with other primarily or secondarily altered genes. As a result, only considering the gene expression levels of individual genes is not very informative and thus they can mislead in classification of complicated cancer types. New techniques that can summarize the collective aberrations in gene expressions of sets of genes are needed.

Our contributions are as follows:

i. In this study, we propose a new network-based classifier, called NBC. Briefly, our method works in two phases: learning and prediction. In the learning phase, for a given gene expression dataset of samples, first, we extract the most relevant features for classification of the training samples into their correct classes. Note that, features in our study are actually the genes. For this reason, in this paper, from now on we will use the terms feature and gene interchangeably depending on the underlying context. In the next step, using the gene expression levels, we create a gene association network for each class describing the dependency between these selected genes. Each node in an association network denotes a gene. Each edge between a pair of nodes indicates the correlation between the expression levels of the two corresponding genes in that cancer class. In the final step of the learning phase, for each gene, we create a predictor function using its immediate neighbors in the network model we built for each class. In the prediction phase, for each class, we use these functions to predict the expression levels of a given test sample and compare the prediction to the given test sample. We assign the given test sample to the class, which yields the least prediction error.

ii. We compare NBC to five traditional classification methods using two- and multi-class cancer microarray datasets. More specifically, we compare our network-based classifier, NBC, to SVM, NB, C4.5, kNN, and RF using five recently published large-scale cancer microarray datasets. The datasets we used in our experiments cover a wide spectrum of scenarios; they include gene expression levels from cancer to normal patients, different cancer cell lines, or cancer subtypes.

iii. One issue that affects the outcome of the classification analysis is the number of genes in the microarray datasets. Many of the genes are irrelevant to the classification of the cancer types. Thus, selecting the relevant genes improves the accuracy of the classification algorithms. Here we also compared the accuracy of five feature selection methods, namely support vector machine feature selection (SVM-FS), chi-square ($\chi^2$),\textsuperscript{12} symmetrical uncertainty (SU),\textsuperscript{13} information gain (IG),\textsuperscript{14} and prediction analysis of microarrays (PAM).\textsuperscript{15} We evaluated the class prediction efficiency of each classifier using these five-feature selection methods.

iv. In our experiments, we also studied the correlation-based co-expression network topology (degree, clustering coefficient, and closeness centrality distributions) of different cancer classes. For this purpose, we compared the distinctive networks that were suggested by the NBC method. Our analysis shows that different cancer classes lead to drastic changes in the network properties, which suggests that cancer leads to major changes in the gene-to-gene interactions in different cancer classes.

\textbf{Methods}

In this section, first we provide a short description of the datasets used in our study. Then, we present a detailed description of the NBC classifier and five state-of-the-art classifiers, namely C4.5, kNN, NB, SVM, and RF. We also provide the descriptions of five-feature selection methods: SVM-FS, $\chi^2$, IG, SU, and PAM.

\textbf{A. Datasets.} To observe the performance of our methods under a broad spectrum of scenarios, we have used five cancer microarray datasets in our experiments with varying characteristics. The datasets are summarized below.

i. Lung cancer dataset\textsuperscript{16} The lung cancer dataset consists of 120 samples (60 paired samples from tumor (class 1) and normal (class 2) tissues) with...
54,675 probe sets from the Affymetrix chip. The dataset can be obtained from NCBI (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE19804).

ii. Breast cancer dataset

The breast cancer dataset consists of 162 samples with 54,675 probe sets from the Affymetrix chip. The samples contain 57 women with breast cancer diagnosis, 37 women with benign diagnosis, 31 women with normal initial mammogram, 15 breast cancer patients following surgery, 15 patients with gastrointestinal cancer, and 7 patients with brain tumor. We excluded the 15 breast cancer patients following surgery since we do not know whether any of these patients had recurring diagnosis. We have also excluded 15 gastrointestinal cancer and 7 brain tumor patients as we focus on breast cancer. In the final dataset, we had 125 samples belonging to three different classes (class 1 = PBMC_Normal (31 samples), class 2 = PBMC_Malignant (57 samples), and class 3 = PBMC_Benign (37 samples)). The complete dataset can be obtained from NCBI (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE27562).

iii. NCI60 dataset

This dataset contains 174 samples spanning nine tumor types with 54,675 probe sets from the Affymetrix chip. Nine cancer tissue origins consist of class 1 = leukemia (18 samples), class 2 = breast (15 samples), class 3 = ovarian (21 samples), class 4 = melanoma (26 samples), class 5 = central nervous system (18 samples), class 6 = colon (21 samples), class 7 = renal (23 samples), class 8 = non-small cell lung (26 samples), and class 9 = prostate (6 samples). Among these, we excluded six prostate samples (class 9), since six samples are too few for classification studies. The NCI60 dataset can be obtained from NCBI (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE32474).

iv. Leukemia dataset

The leukemia dataset consists of 574 samples in 10 classes with 22,283 probe sets from the Affymetrix chip. Four CD34 and four CD10CD19 samples are excluded from the dataset since four samples are too few for classification studies. We also excluded the 153 samples, which do not have a known karyotype to focus on leukemia. The pruned dataset contains 413 samples belonging to seven different leukemia types: class 1 = hyperdiploid (115 samples), class 2 = TCF3-PBX1 (40 samples), class 3 = ETV6_RUNX1 (99 samples), class 4 = MLL (30 samples), class 5 = PH (23 samples), class 6 = hypodiploid (23 samples), and class 7 = T-ALL (83 samples). The complete leukemia dataset can be obtained from NCBI (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE33315).

v. Colon cancer dataset

The colon cancer dataset consists of 566 samples in six classes with 54,675 probe sets from the Affymetrix chip. The dataset contains six colon cancer subtypes: class 1 = CINImmune-Down (116 samples), class 2 = dMMR (104 samples), class 3 = KRASm (75 samples), class 4 = CSC (59 samples), class 5 = CINWntUp (152 samples), and class 6 = CINnormal (60 samples). The complete colon cancer dataset can be obtained from NCBI (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE39582).

B. Classifiers

We have compared the NBC classifier to five state-of-the-art classifiers in our experiments. We believe that these methods collectively constitute a significant portion of the key literature on this topic. We provide a short description of all of these five classifiers as well as our method below. We also provide a detailed description of our method in the Supplementary Files.

i. kNN

This method is a non-parametric similarity-based classification algorithm. In this method, a sample is classified by majority vote of its k nearest neighbors (kNNs). More specifically, each testing sample is assigned to the class most common among its kNNs. We say that a sample in a training set is a neighbor of a given test sample if that training sample is one of the k closest samples to the test sample among all training samples. We have used k = 1 in this study and computed the distance between a pair of samples as the Euclidean distance between their gene expression values.

ii. C4.5

This method builds a decision tree, which consists of a set of internal and leaf nodes. The internal nodes are associated with a splitting criterion, which consists of a splitting feature and one or more splitting rules defined on this feature. The leaf nodes are labeled with a single class label. C4.5 employs a two-step algorithm to generate decision trees from a dataset, using information entropy. In the first step, C4.5 builds decision trees from a set of training data, using the concept of information entropy. The dataset is a set of already classified samples $S = \{S_1,S_2,...,S_M\}$. Each sample $S_i$ consists of an n-dimensional vector $(X[i,1],X[i,2],...,X[i,n])$, where the $X[i,j]$ represents the expression level of the jth gene in the ith sample $S_i$. At each node of the tree, C4.5 chooses the gene that most effectively splits its set of samples into subsets enriched in one class or the other. The splitting criterion is the normalized information gain (IG) (see Feature Selection Methods section for the IG formula). The feature with the highest normalized IG is chosen to make the decision. The C4.5 algorithm then iterates on the smaller sub-datasets. In the second step of the algorithm, the tree is pruned to avoid overfitting to the data.

iii. NB

This method uses probabilistic induction to assign class labels to test samples, assuming independence among the features. Briefly, a naive Bayes (NB) classifier creates rules based on Bayes’ theorem. Bayes’ theorem is a result in probability theory, which relates the conditional and marginal probability distributions of random events:
where \( P(U|V) \) and \( P(V|U) \) represent the probability of an event \( U \) conditional on event \( V \) and probability of \( V \) conditional on \( U \), respectively. Similarly, \( P(U) \) and \( P(V) \) represent the marginal probability of events \( U \) and \( V \), respectively. The NB method uses all features, and thus, allows them to contribute to the decision as if they were all equally important and independent of one another.

In short, the NB method works as follows. Given a sample \( S_j \) which consists of an \( n \)-dimensional feature vector, \( \{X[i,1],X[i,2],...,X[i,n]\} \), we construct the posterior probability for the class \( C^i \) among a set of possible classes \( C = \{C^1,C^2,...,C^K\} \) using Bayes’ rule:

\[
P(C^i | X[i,1], X[i,2],..., X[i,n]) = \frac{P(X[i,1], X[i,2],..., X[i,n] | C^i) P(C^i)}{P(X[i,1], X[i,2],..., X[i,n])}
\]

where \( P(C^i | X[i,1], X[i,2],..., X[i,n]) \) is the probability that sample \( S_j \) belongs to class \( C^i \). Since naïve Bayes theorem assumes that the conditional probabilities of the independent variables are independent, we can decompose \( P(S_j | C^i) \) to \( P(S_j | C^i) = \prod_{j=1}^{n} P(X[i,j] | C^i) \). So we can rewrite

\[
P(C^i | S_j) = P(C^i) \prod_{j=1}^{n} P(X[i,j] | C^i).
\]

Using Bayes’ rule above, we label a new case \( S_i \) with the class \( C^i \) that achieves the highest probability.

iv. SVM

This method is one of the fundamental supervised machine learning algorithms for binary classification.\(^{25,26}\) The most commonly used SVMs in the biological data classification are the linear SVMs because of their simplicity of implementation. For a given training dataset,

\[
S = \{(S_j, c_j) | S_j \in \mathbb{R}^n, c_j \in \{-1,1\} \text{ and } i = 1,...,l | S_j \}
\]

where \( c_j \) denotes the class of sample \( S_j \) (\( c_j = -1 \) (class \( C^1 \)) and \( c_j = 1 \) (class \( C^2 \)). The linear SVM separates the given data points into their correct classes with a hyperplane. The SVM learning algorithm constructs this hyperplane with the maximum margin that separates the positive samples from the negative samples. The points that lie closest to this max-margin hyperplane are called the support vectors. The hyperplane can be defined using these points alone, and the classifier only makes use of these support vectors to classify test samples. This hyperplane represents the largest separation between the samples \( S_j \) belonging to the two classes. Any such hyperplane can be written as the set of points \( S \) satisfying \( w \cdot S - b = 0 \), where “\( \cdot \)” denotes the dot product and the vector \( w = (w_1, w_2, ... , w_n) \), and the constant \( b \) represents the coefficients of the SVM. Linear SVM algorithm computes \( w \) and \( b \) to maximize the separation between the samples belonging to the two classes. In this approach, if a given sample \( S_j \) satisfies \( w \cdot S_j - b \leq -1 \), then \( S_j \) is assigned to the class \( C^1 \), and if it satisfies \( w \cdot S_j - b \geq 1 \), then \( S_j \) is assigned to the class \( C^2 \). If the samples are not linearly separable in the feature space, to allow for error tolerance, a limited fraction of training samples are allowed to fall to the wrong side of the hyperplane.\(^{27,28}\)

Linear SVM is a binary classifier; however, it can also be used for multi-class datasets in the same way as any multi-class problem can be reduced to binary classification problems. There are several strategies for this purpose. Here, we use one of the most common strategies known as the one-versus-all approach. This strategy transforms the single multi-class problem into \( K \) binary classification problems, one for each class (ie, \( S_j \) vs \( S \neq S_j \)). It then decides the class for a given new sample using the winner takes all strategy ie, the classifier with the highest output function assigns the class.

v. RF

This method is an ensemble approach that builds multiple decision trees (described above in the C4.5 section) using the training dataset to achieve a better classifier performance.\(^{29}\) The test samples are classified by assigning them to the classes that take the majority vote over all the decision trees. The method works as follows. In the first step, \( k \) sets of \( M \) samples \((S^k = \{S_{j_1}^k, S_{j_2}^k, ..., S_{j_M}^k\})\) are chosen from the original training dataset \((S = \{S_1, S_2, ... , S_M\})\) by replacement. In the second step, each of these \( k \) sets of \( M \) sample datasets is used to train a decision tree classifier. After the training step, prediction for each test sample is made using the \( k \) decision trees. The final class prediction for the test sample is made by majority voting. The sample is assigned to the class that gets the most votes over all \( k \) decision trees.

vi. NBC method

This method works in two phases: (A) learning and (B) prediction. In this section, we will summarize this method. A detailed description of this method is available in the Supplementary File.

The learning phase of the NBC method works in three steps: (i) feature selection, (ii) correlation network creation, and (iii) prediction function learning. In this phase, for a given gene expression training dataset \( S \) ([8]) samples in each disease class \( C^i \) containing the expression levels of \( N \) genes for \( t \) training samples belonging to \( K \) classes, we extract the \( n \) most informative genes for the classification of the samples into the correct classes. For each class \( C^i \), we build a correlation network \( G^i \) to describe the dependency of the \( n \) selected genes to each other using their expression levels. In the final step, using these networks, for each gene \( g_j \), we learn a predictor function \( f_{g_j}^i \) (classifiers) for each class \( C^i \) using the expression levels of the gene \( g_j \)’s adjacent neighbors in that network.

The prediction phase of the NBC method works in two steps: (i) gene expression prediction and (ii) class assignment. For a given testing sample set \( T \) ([9]) samples in each disease class \( C^i \) instead of predicting the class of each sample \( T_j \in T \) directly, we use our trained classifiers \( f_{g_j}^i \) to predict the gene expression levels \( Y^T[i, \cdot] \). We compare the predicted
expression levels $Y^T[i,T]$ to the expression level of the given testing sample $X^T[i,T]$. We assign each testing sample $T_i$ to the class $[C_1,C_2,…,C^K]$ that gives the smallest difference between $X^T[i,T]$ and $Y^T[i,T]$ in the $L^2$ norm.

C. Feature selection methods. Feature selection methods rank the features of the given dataset based on their importance. They remove noisy features (genes in our study) from the dataset, with the goal of increasing the classification accuracy while reducing the running time.14,30–33 The total number of genes available in the datasets used in our study is very large (i.e., more than 10,000). So, it is critical to reduce the number of genes to a small subset when performing classification of the samples. Typically, we observe that 50–150 genes have been selected in the literature for binary classification studies.1,14,35 Herein, we use a subset of the top 50–300 genes based on the underlying feature selection algorithm employed. In this study, we compare five-feature selection methods to select a small set of genes. These methods are SVM-FS, $\chi^2$, IG, SU, and PAM. These methods are summarized below.

i. SVM-FS

This feature selection method is based on the SVM classifier algorithm described above. SVM classifier is trained as $w_0+w_0+b$, where $w$ is the vector of weights for each gene $x_i$ and $b$ is the bias. In this approach, the entire set of genes is used to train the SVM classifier. We pick the top $n$ genes with the largest weights. The rationale is that, for trained SVM classifier, the weights of the genes are proportional to their importance in the classification. More specifically, as the weight of a gene increases, so does its contribution to the separation of the two classes.

ii. $\chi^2$ feature selection method

This method evaluates each gene's importance for classification individually by measuring the $\chi^2$ statistics with respect to the classes.12 First, the gene expression values are discretized into several intervals. Then using these discrete expression values, $\chi^2$ value of each gene is calculated as described below. Let us denote the number of samples in the $i$th interval and $j$th class with $A_{ij}$, and the expected frequency of $A_{ij}$ with $E_{ij}$. $E_{ij}$ is equal to $(R_i \times C_j)/N$, where $N$ is the total number of samples, and $R_i$ and $C_j$ are the number of samples in the $i$th interval and the $j$th class, respectively. Also, let us denote the number of intervals and the number of classes with $k$ and $l$, respectively. We compute the $\chi^2$ value of each gene as

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{(A_{ij} - E_{ij})^2}{E_{ij}}$$

We then pick the top $n$ genes with the highest $\chi^2$ statistic values as significant genes for classification.

iii. IG feature selection method

This method evaluates the worth of a gene by measuring the information gain (IG) with respect to the class.14 Let us denote the total entropy of the class with $H$(class) and the conditional entropy of the class given the gene with $H$(class|gene). We compute the IG as

$$IG = H$(class) − $H$(class|gene)$$

We then pick the top $n$ genes with the highest IG score as significant genes for classification.

iv. SU feature selection method

This method evaluates the worth of a gene by measuring the SU with respect to the class.13,36 Symmetric uncertainty (SU) is measured by

$$SU = \frac{2 \times IG}{H$(class) + H$(gene)}$$

where $H$(class) is the total entropy of the class, and $H$(gene) is the entropy of the gene. We then pick the top $n$ genes with the highest SU score as significant genes for classification.

v. PAM feature selection method

This method is based on the nearest shrunken centroid method.15 The method works as follows. First, the method computes a centroid for each class and an overall centroid. Centroid for a class ($\bar{x}_k$) and the overall centroid ($\bar{x}$) are defined as the average gene expression in the class $k$ and in the entire dataset, respectively. In the next step, the method finds the $t$-statistic for each gene $j$ by comparing class $k$ centroid to the overall centroid as follows: $d_{jk} = (\bar{x}_k − \bar{x})(s_j + s_0)$, where $\bar{x}_k$ represents the $j$th index of the centroid for class $k$, $\bar{x}$ the $j$th index of the overall centroid, $s_j$ the pooled within-class standard deviation, $s_0$ the median value of $s_j$ over the set of genes. $m_j = \sqrt{n_j + 1}/n$, in which $n_j$ is the number of samples in class $k$ and $n$ is the total number of samples in the entire dataset. The method shrinks the class centroids toward the overall centroid ($\bar{x}_k = \bar{x}$) by ($m_j s_j + s_0$) $d_{jk}$ by reducing $d_{jk}$ using the soft thresholding method, which is defined by $d_{jk} = \text{sign}(d_{jk}) (|d_{jk}| - \Delta)$, where $\Delta$ is the positive part of the number. This method provides a list of significant genes whose expression characterizes each cancer class. In order to obtain a unified list of 50–300 genes, we combined the genes provided by the PAM method for different classes.

D. Cross-validation. Cross-validation is a model validation method for assessing generalizability of the classifiers into other independent datasets. It is a key step in classifier construction to assess the performance of the classifiers. $k$-fold cross-validation is a commonly used cross-validation technique for small-size datasets to assess the classifier performance. In this approach, the original dataset is randomly partitioned into $k$ subsets. In each fold, $k - 1$ subsets are used to train the classifier, and the remaining one subset is used to test the accuracy of the trained classifier. The average of the $k$ accuracies is reported as the performance of the classifier. In this paper, we use $k = 10$ in our experiments.
E. Network measures. Analysis of the differences and the similarities between the correlation-based co-expression networks in different cancer classes is key to understanding cancer. In this study, we used three network measures – namely, degree distribution, clustering coefficient, and closeness centrality – to compare different cancer classes. These network measures have been calculated as described below.

i. Degree distribution
The degree of a node (gene in our case) is the number of connections it has to other genes. The degree distribution $P(k)$ is defined as the probability distribution of these degrees over the whole network, i.e., fraction of genes in the network with degree $k$. The degree distribution has been used frequently to study the topological characteristics of networks in the literature.37,38

$$d_i = \sum_{j \in S} d_{ij}$$

where $d_{ij}$ represents the distance between the two genes $i$ and $j$. If two genes, $i$ and $j$, are not reachable from each other in the underlying network, then $d_{ij} = \infty$. In the closeness centrality computation above, $1/\infty$ is taken as 0.

ii. Closeness centrality
The closeness centrality of a gene $g_i$ shows the importance of that gene in the network it belongs to. It is defined as the sum of the reciprocal of its distances to all other genes. Thus, the more central a gene, the higher is its closeness centrality. A gene with a high closeness centrality generally has a quick access to other genes in a network. The closeness centrality in our study is defined as

$$C_i = \frac{1}{\sum_{g \in S} d(g_i,g_j)}$$

where $d(g_i,g_j)$ represents the distance between the two genes $g_i$ and $g_j$. If two genes, $g_i$ and $g_j$, are not reachable from each other in the underlying network, then $d(g_i,g_j) = \infty$. In the closeness centrality computation above, $1/\infty$ is taken as 0.

iii. Clustering coefficient
Clustering coefficient of a gene measures the degree to which the adjacent genes of that gene in a graph tend to connect together. More specifically, the clustering coefficient of a gene is defined as

$$C(g_i) = \frac{\text{Number of edges among } N(g_i)}{\text{Max possible number of edges among } N(g_i)}$$

where $N(g_i)$ is the set of genes adjacent to $g_i$.

Results
In this section, we evaluated the performance of our network-based classifier NBC. Many supervised classification algorithms have been proposed for predicting cancer types.10,11 Among them, we compared NBC to five traditional classifiers (SVM, NB, kNN, C4.5, and RF). Collectively, these five methods covered a broad spectrum of alternative methods. We selected a subset of available genes in the transcriptome using five alternative feature selection methods, namely, SVM-FS, IG, SU, $\chi^2$, and PAM. We implemented SVM, NB, kNN, C4.5, and RF classifiers and PAM feature selection method in MATLAB software, and NBC classifier and SVM-FS feature selection method in the C programming language. We used the Weka platform39 for the other feature selection methods.

We used 10-fold cross-validation to calculate the prediction accuracy of each of the classifiers. More specifically, we kept one fold (one-tenth of the set of all samples) as the test samples, and selected relevant genes and trained classifiers on these genes using the remaining nine folds as the training data. This way, we avoided any positive bias toward the test samples. We repeated this 10 times by using each fold as the test data. We reported the average accuracy we observed in all the 10 folds.

We tested the classifiers and feature selection methods using five cancer microarray datasets with varying characteristics. Descriptions of these datasets are provided in the Methods section. Diversely selected microarray datasets in this study include gene expression levels from cancer vs normal patients, different cancer cell lines, or cancer subtypes. In order to ensure that the noise arising from combining different experimental techniques does not give unfair disadvantage to any of the methods we compared, we focused on large-scale Affymetrix cancer microarray datasets in the gene expression omnibus (GEO) database, rather than combining datasets from varying experimental sources. We, first, normalized each of these datasets, so that each gene expression value has a mean of 0 and a variance of 1. Next, we applied feature selection methods on these normalized datasets. These methods provide the ranking of all the genes in respect to their importance for cancer classification. We choose the top $k$ genes with $k \in [50,75,100,\ldots,300]$ from these rankings and used them as the features for the classifiers.

In the following sections, we first report our findings on the comparison of feature selection and classification methods (Sections A, B, and C). Then, we analyze the NBC method in detail (Sections D, E, and F).

A. Evaluation of feature selection methods. The choice of feature selection method and the underlying dataset impact the classification accuracy. The first question we seek to answer is how NBC compares to existing classification methods. Here, we present an extensive comparison of all the six methods on five different cancer datasets by using five alternative feature selection methods. Notice that each method can have a peak performance (i.e., highest accuracy) at different number of features. As a result, setting the number of genes to a fixed value would give unfair advantage to some of the methods we evaluate. To avoid this, we report the highest accuracy of each method along with the number of features at which it is capable of achieving that accuracy. Table 1 presents the detailed results (see the last two columns).

On the leukemia, colon, and breast cancer datasets, if only the expression levels of up to 100 genes are used, then the gap between the best and worst mean accuracy levels for feature selection methods is large (62.43 vs 86.04%, 52.53 vs
| LUNG CANCER <100 GENES | CLASSIFIER | Mean | Std |
|-------------------------|------------|------|-----|
| NBC SVM NB kNN C4.5 RF |            |      |     |
| SVM-FS (50,92.50) (75,86.67) (100,85.00) (75,86.67) (100,83.33) (75,93.33) | 87.92 | 4.07 |
| SU (75,95.83) (75,95.83) (50,91.67) (75,95.83) | 90.97 | 4.10 |
| IG (100,95.00) (100,96.67) (50,91.67) | 93.47 | 3.59 |
| PAM (50,93.33) (100,92.50) | 92.64 | 2.13 |
| Mean 94.50 93.17 92.67 | 91.00 86.50 95.16 |
| Std 1.51 3.97 4.50 | 2.46 2.16 1.09 |

| BREAST CANCER <100 GENES | CLASSIFIER | Mean | Std |
|--------------------------|------------|------|-----|
| NBC SVM NB kNN C4.5 RF |            |      |     |
| SVM-FS (100,54.40) (75,56.80) (50,55.20) (75,44.00) (100,58.40) (50,56.80) | 54.27 | 5.22 |
| SU (50,59.20) (75,56.80) (50,53.60) (75,49.60) (100,56.00) | 55.33 | 3.33 |
| IG (100,54.40) (75,52.00) (50,52.00) (100,55.20) | 53.87 | 1.65 |
| PAM (100,50.40) (75,47.20) (50,38.40) (75,44.00) | 46.00 | 4.26 |
| Mean 54.88 52.80 50.88 48.64 | 54.40 54.08 |
| Std 3.18 4.08 7.21 | 4.47 3.88 3.65 |

| NCI60 <100 GENES | CLASSIFIER | Mean | Std |
|------------------|------------|------|-----|
| NBC SVM NB kNN C4.5 RF |            |      |     |
| SVM-FS (75,98.21) (100,95.83) (100,92.26) (100,98.21) (100,77.38) (75,93.45) | 92.56 | 7.82 |
| SU (75,100) (75,99.40) (100,89.88) (50,99.40) (100,87.50) (100,96.43) | 95.44 | 5.42 |
| IG (50,100) (100,98.81) (100,90.48) (50,99.40) (50,83.33) (100,96.43) | 94.74 | 6.59 |
| PAM (100,97.62) (100,97.02) (100,88.69) (100,98.81) (100,72.02) (100,93.45) | 91.27 | 10.1 |
| Mean 99.05 98.09 90.71 99.04 | 80.71 95.24 |
| Std 1.08 1.60 1.55 | 0.53 6.05 1.63 |

| LEUKEMIA <100 GENES | CLASSIFIER | Mean | Std |
|---------------------|------------|------|-----|
| NBC SVM NB kNN C4.5 RF |            |      |     |
| SVM-FS (100,60.05) (100,74.33) (100,44.79) (100,58.35) (100,83.68) (100,73.37) | 62.43 | 10.9 |
| SU (100,85.71) (100,89.35) (100,80.39) (75,89.59) (75,80.15) (100,90.07) | 85.88 | 4.61 |
| IG (75,85.71) (100,89.83) (100,79.90) (100,89.59) (75,81.11) | 85.96 | 4.51 |
| PAM (100,71.43) (100,81.36) (100,63.92) (100,77.97) (100,67.31) (100,78.93) | 73.49 | 7.01 |
| Mean 77.48 84.75 70.07 81.26 | 74.62 84.36 |
| Std 11.46 6.78 15.87 | 13.83 8.44 7.75 |

| COLON CANCER <100 GENES | CLASSIFIER | Mean | Std |
|-------------------------|------------|------|-----|
| NBC SVM NB kNN C4.5 RF |            |      |     |
| SVM-FS (100,72.44) (100,58.48) (100,56.36) (100,54.77) (100,49.47) (100,62.54) | 59.01 | 7.86 |
| SU (100,73.85) (100,59.01) (50,69.79) (75,66.61) (100,64.31) | 60.41 | 10.9 |
| IG (100,73.14) (100,63.07) (100,64.66) (100,62.01) | 66.02 | 5.80 |
| PAM (100,59.72) (100,51.24) (100,50.00) (75,52.83) (100,46.82) | 52.53 | 4.12 |
| Mean 70.21 58.09 61.52 | 61.20 57.17 68.76 |
| Std 7.81 4.04 8.32 | 6.85 8.33 7.97 |
68.49%, and 46.00 vs 55.33%, respectively). These gaps are significant, particularly, given the fact that the standard deviations of different feature selection methods vary from 4.38 to 10.90, 4.12 to 7.86, and 1.65 to 5.22, respectively. Thus, even if we consider the largest standard deviations for these datasets, the gaps are close to or more than two standard deviations. As the number of maximum admissible genes used for the classification increases from 100 to 300, for the leukemia, colon, and breast cancer datasets, the gap between the best and worst mean accuracy levels reduces (70.18 vs 87.69%, 56.36 vs 69.17%, and 51.07 vs 56.67%, respectively) (Supplementary Tables 4 and 5). Similar but less severe behavior is observed for lung and NCI60 cancer datasets. Overall, we observe that the SVM-FS and PAM feature selection methods are not as efficient as other feature selection methods for all of the cancer datasets (Table 1). Only exception to this behavior is observed in the breast cancer dataset. SVM-FS performs comparable to other feature selection methods for this dataset. Thus, we observe that SVM-FS and PAM feature selection methods perform worse than other feature selection methods on an average especially when small number of genes is selected. From the remaining three feature selection methods, there is no significant winner. However, SU method shows slightly better accuracy results in most of the cancer datasets. Although, relevancy of the selected genes by the feature selection methods for the specific cancer types is not a focus of our study, we checked the appropriateness of the selected genes by the SU method for different cancer types. The genes selected by this method are usually relevant to the specific cancer type. As an example, the gene set selected for the colon cancer classification includes the colon cancer-related genes such as MLH1,\textsuperscript{40,41} AXIN2,\textsuperscript{42,43} ASC1L2,\textsuperscript{44} and LGALS4.\textsuperscript{45}

We observe reduction in the variation of the accuracy levels between different feature selection methods as the number of maximum permissible genes increases from 100 to 300. This suggests that although there might be variations in the top 100 genes, this variation decreases as the number of genes increases. However, the accuracy levels of different classifiers show more variation for the SVM-FS and PAM feature selection methods in comparison to other feature selection methods. This suggests that SVM-FS and PAM feature selection methods are more sensitive to the underlying classification algorithm. The variation in classification accuracy is generally comparable and lower for the genes selected by the other methods.

Finally, we observe that as the number of genes used in the classification increases, although we see slight increases in the accuracy levels, above observations for the feature selection methods remain unchanged.

**B. Evaluation of classification algorithms.** The choice of appropriate classification method is of utmost importance for correctly classifying samples to correct cancer types. Here, we compare the performance of the NBC method to five traditional classifiers described above by considering the average performance across all the five feature selection methods considered (see the last two rows in Table 1 for each dataset.)

Our results suggest that the NBC method performs at comparable or better accuracy levels than the traditional classification methods for nearly all datasets (Table 1, and Supplementary Tables 4 and 5). The NBC classifier on average achieves 54.88–99.05% accuracy on the five cancer microarray datasets using up to only 100 genes. The classification accuracies for the other classifiers such as SVM in these five datasets are usually lower than that of the NBC method. This suggests that the performance of our network-based cancer classifier is very promising.

NBC method outperforms the other classifiers for the NCI60, breast, and colon cancer datasets. For the leukemia and lung cancer datasets, SVM and RF classifiers provide the best mean performance, respectively. The C4.5 decision tree classifier shows the worst mean performance for the NCI60, lung, and colon cancer datasets. NB and kNN classifiers show the worst mean performance for the leukemia and breast cancer datasets, respectively. These behaviors are generally independent of the number of genes used for classification (Supplementary Tables 4 and 5).

Classifiers show different variation levels in their accuracies. While NBC classifier exhibits the least standard deviation for breast cancer dataset; RF, kNN, and SVM classifiers show the least variation for lung, NCI60, leukemia, and colon cancer datasets. Even for the four datasets (lung, NCI60, leukemia, and colon cancer datasets) for which NBC does not have the smallest standard deviation, its standard deviation remains to be very small. These observations suggest that the NBC method while providing comparative or better accuracy than other methods, its classification accuracy is also robust to choice of the feature selection method.

Finally, we observe that as the number of genes used in the classification increase, above observations for the classification methods are mostly conserved (see Supplementary Tables 4 and 5).

**C. Feature selection and classification method combination.** So far, we have discussed the average behavior of different feature selection or classification methods. Here, we focus on the specific combination of feature selection and classification methods. An appropriate combination of feature selection and classification methods leads to the best performance on different multi-class cancer datasets.\textsuperscript{46} Here we will focus on the accuracy ranking of feature selection and classification methods using up to 100 genes selected from the five datasets (Table 2). However, our main results are mostly conserved as the number of genes used for classification increased (Supplementary Tables 6 and 7). Following from the results in Tables 1 and 2, we observe that NBC and RF classifiers are the two best classifiers when we adopt SU as the feature selection method. While NBC classifier achieves the highest prediction accuracy in NCI60 (accuracy: 100%) and breast cancer (accuracy: 59.2%) datasets, its performance in the lung...
Table 2. A summary of accuracy rankings for lung, breast, NCI60, leukemia, and colon cancer datasets for different feature selection and classification methods when the maximum number of allowable genes is set to 100. In the first five tables, we report the best accuracy obtained by each combination of classifier and feature selection method, and their ranking. In these tables, entries containing a pair of numbers \( V:W \) indicate the following: \( V \) refers to the best classification accuracy and \( W \) refers to its ranking. We use bold face to highlight the top five highest accuracies. In the last table, we report the ranking of each combination of classifier and feature selection method based on the average accuracy obtained over all the five datasets.

| Cancer Type | Selection | Classifier | NBC | SVM | NB | kNN | C4.5 | RF |
|-------------|-----------|------------|-----|-----|----|-----|------|----|
| **LUNG CANCER** | SVM-fS | 92.50 : 16 | 86.67 : 25 | 85.00 : 29 | 86.67 : 25 | 83.33 : 30 | 93.33 : 14 |
| | SU | 95.83 : 2 | 95.83 : 2 | 95.83 : 2 | 91.67 : 21 | 85.83 : 28 | 95.83 : 2 |
| | \( \chi^2 \) | 95.83 : 2 | 94.17 : 12 | 95.83 : 2 | 92.50 : 16 | 86.67 : 25 | 95.83 : 2 |
| | IG | 95.00 : 10 | 96.67 : 1 | 94.17 : 12 | 91.67 : 21 | 87.50 : 24 | 95.00 : 10 |
| | PAM | 93.33 : 14 | 92.50 : 16 | 92.50 : 16 | 92.50 : 16 | 89.17 : 23 | 95.83 : 2 |
| **BREAST CANCER** | SVM-fS | 54.40 : 13 | 56.80 : 3 | 55.20 : 10 | 44.00 : 28 | 58.40 : 2 | 56.80 : 3 |
| | SU | 59.20 : 1 | 56.80 : 3 | 53.60 : 16 | 49.60 : 24 | 56.00 : 7 | 56.80 : 3 |
| | \( \chi^2 \) | 56.00 : 7 | 51.20 : 21 | 51.20 : 21 | 53.60 : 16 | 54.40 : 13 | 55.20 : 10 |
| | IG | 54.40 : 13 | 50.40 : 23 | 52.00 : 19 | 52.00 : 19 | 55.20 : 10 | 53.60 : 16 |
| | PAM | 50.40 : 23 | 47.20 : 27 | 38.40 : 30 | 44.00 : 28 | 48.00 : 25 | 48.00 : 25 |
| **NCI60** | SVM-fS | 98.21 : 9 | 95.83 : 18 | 92.26 : 21 | 98.21 : 9 | 77.38 : 29 | 93.45 : 19 |
| | SU | 100.00 : 1 | 99.40 : 3 | 89.88 : 24 | 99.40 : 3 | 87.50 : 26 | 96.43 : 15 |
| | \( \chi^2 \) | 99.40 : 3 | 99.40 : 3 | 92.26 : 21 | 99.40 : 3 | 83.33 : 27 | 96.43 : 15 |
| | IG | 100.00 : 1 | 98.81 : 9 | 90.48 : 23 | 99.40 : 3 | 83.33 : 27 | 96.43 : 15 |
| | PAM | 97.62 : 13 | 97.02 : 14 | 88.69 : 25 | 98.81 : 9 | 72.02 : 30 | 93.45 : 19 |
| **LEUKEMIA** | SVM-fS | 60.05 : 28 | 74.33 : 22 | 44.79 : 30 | 58.35 : 29 | 63.68 : 27 | 73.37 : 23 |
| | SU | 85.71 : 10 | 89.35 : 8 | 80.39 : 17 | 89.59 : 5 | 80.15 : 18 | 90.07 : 2 |
| | \( \chi^2 \) | 85.71 : 10 | 89.83 : 3 | 79.90 : 19 | 89.59 : 5 | 81.11 : 15 | 89.59 : 5 |
| | IG | 84.50 : 12 | 88.66 : 9 | 81.36 : 13 | 90.80 : 1 | 80.87 : 16 | 89.83 : 3 |
| | PAM | 71.43 : 24 | 81.36 : 13 | 63.92 : 26 | 77.97 : 21 | 67.31 : 25 | 78.93 : 20 |
| **COLON CANCER** | SVM-fS | 72.44 : 7 | 58.48 : 21 | 56.36 : 23 | 54.77 : 25 | 49.47 : 29 | 62.54 : 16 |
| | SU | 73.85 : 4 | 59.01 : 20 | 69.79 : 8 | 66.61 : 11 | 64.31 : 13 | 74.91 : 3 |
| | \( \chi^2 \) | 73.14 : 6 | 59.89 : 19 | 63.07 : 15 | 64.66 : 12 | 62.01 : 17 | 75.09 : 2 |
| | IG | 75.27 : 1 | 61.84 : 18 | 68.37 : 9 | 67.14 : 10 | 63.25 : 14 | 75.09 : 2 |
| | PAM | 56.36 : 23 | 51.24 : 27 | 50.00 : 28 | 52.83 : 26 | 46.82 : 30 | 57.95 : 22 |

| **FINAL RANKING** | Classifier | NBC | SVM | NB | kNN | C4.5 | RF |
|-------------------|------------|-----|-----|----|-----|------|----|
| | SVM-fS | 15 | 19 | 26 | 27 | 28 | 16 |
| | SU | 1 | 4 | 14 | 12 | 21 | 2 |
| | \( \chi^2 \) | 3 | 11 | 17 | 8 | 22 | 5 |
| | IG | 5 | 10 | 12 | 9 | 20 | 7 |
| | PAM | 22 | 22 | 29 | 25 | 30 | 18 |
of nodes and edges in the network models generated by

These are (i) the number of genes selected and (ii) the Pearson
distance scores characterize the predictive models generated by NBC.

To understand these differences, let’s briefly describe two para-
drase: Two predictor combination with SU as the feature selection method.

In the other datasets is also poor (ranking in lung cancer: 16, breast: 23, NCI60: 25, leukemia: 26, and colon cancer: 28). In the lung cancer and leukemia datasets, C4.5 and NB algorithms show poor performance in the lung (accuracy: 89.17%; ranking: 23), breast (accuracy: 48.00%; ranking: 25), and leukemia (accuracy: 67.31%; ranking: 25) datasets is also disappointing. Similarly, NB combined with PAM feature selection method shows the worst performance in the breast cancer dataset (accuracy: 38.40%). In addition, this method combination’s performance in the other datasets is also poor (ranking in lung cancer: 16, NCI60: 25, leukemia: 26, and colon cancer: 28). In the lung cancer and leukemia datasets, C4.5 and NB algorithms show the worst performance when they are used in conjunction with SVM-FS feature selection method (83.33 and 44.79%, respectively). These observations suggest that kNN, NB, and C4.5 algorithms are not competitive against NBC, RF, and SVM classifiers especially when these classifiers are used in conjunction with PAM feature selection method.

D. Analysis of the NBC method in depth. So far, in our experiments we have demonstrated that NBC yields better or similar accuracy as compared to state-of-the-art methods. Next, we focus further on the NBC method to understand its characteristics, strengths, and limitations. Briefly, two parameters characterize the predictive models generated by NBC. These are (i) the number of genes selected and (ii) the Pearson correlation threshold. These two parameters control the number of nodes and edges in the network models generated by NBC, respectively. We vary the values of these two parameters and report the accuracy of NBC for each parameter setting. More specifically, we vary the number of genes in the [50:300] interval and the Pearson correlation threshold in the [0.6:0.95] interval. Figure 1 presents the results.

The dependency of NBC on these two variables is also governed by underlying feature selection method and the data-
set. While the SU, χ², and IG algorithms show qualitatively similar behavior in different datasets, SVM-FS and PAM feature selection methods show very distinct behaviors. For example, classification accuracy mostly decreases as the number of genes increases for low correlation levels in the SVM-FS feature selection method in contrast to SU, χ², and IG feature selection methods. The NBC classifier accuracy behavior is very distinct in different cancer datasets also, possibly because of different network structures (discussed below).

Despite the feature selection method and dataset depen-
dencies, there are clear patterns with regard to the number of genes used in the classification and correlation threshold. For example, in general as the gene numbers increase, accuracy of the NBC method does not increase. We observe that our network-based classifier can predict at high accuracy levels while using only up to 75 genes in lung, breast, and NCI60 cancer datasets. The other algorithms, in particular SVM, usually need more genes to reach similar accuracy levels in these three datasets (Supplementary Table 5). For example, in the breast cancer dataset, while NBC can reach the accuracy level of 59.20% using only 50 genes, SVM and C4.5 classifiers use 300 genes to reach approximately the same accuracy level. The leukemia cancer dataset was generally difficult for all the algorithms, and they needed higher gene numbers for high accuracy levels (Supplementary Table 5). Since measuring gene expression levels is expensive, these observations suggest that our method is probably more relevant for biological applications since it can function at high accuracy levels using smaller number of genes in comparison to traditional classifiers.

With regard to Pearson correlation threshold cutoff, NBC method shows nonlinear behavior. When the threshold cutoff is small, we get lots of false-positive connections in the network that should not exist. When the threshold cutoff is large, probably we miss lots of gene-to-gene associations that should actually exist in the network. For this reason as the correlation threshold increases, first accuracy levels increase for the NBC method and then decrease dramatically. Despite this general behavior, there is no single threshold level that works best for all the cancer datasets. While a correlation cutoff of ~0.7 works best for lung cancer, we need a threshold cut off score of ~0.8—0.9 for the NCI60 dataset.

NBC method constructs a different and unique network for each cancer class and uses these networks and predictor functions constructed by linear regression to predict expres-
sion levels for the selected genes in each sample. In the next step, for each sample, it compares these class-specific pre-
dictions to actual gene expression levels in the sample. The
Figure 1. Accuracy dependency of the NBC method to the number of genes and gene-to-gene associations in the network. Heat maps depicting the accuracy levels for varying number of genes and gene-to-gene interaction density are shown. In the figure, columns refer to the cancer datasets: leukemia, breast, lung, NCI60, and colon. Similarly, rows correspond to the feature selection methods: SVM-FS, symmetrical uncertainty, $\chi^2$, information gain, and PAM. The x-axis in each heat map refers to the Pearson correlation cutoff used to determine the gene-to-gene associations. The y-axis denotes the number of genes used in the NBC method.

Method assigns the sample to the class that gives the minimum distance between the predicted and actual gene expression levels in the $L^2$ norm. To see how distinctive our method is in separating different classes, we computed the prediction errors for inter- and intra-subclasses using each class-specific predictor function of the NBC method. More specifically, we computed the error using the relative $L^2$ norm. Relative $L^2$ norm is defined as $\| Y - f(X) \|_2 / \| Y \|_2$, where $Y$ represents the actual gene expression levels for the test sample and $f(X)$ represents the predicted gene expression levels for the same test sample. Figure 2 presents the results for all of the five cancer datasets we used. We make two important observations from these results. First, the prediction errors explain the classification accuracies of our method. As an example, for NCI60 dataset NBC classifier provides perfect accuracy levels, and in Figure 2, we observe that the models created
by the NBC classifier yield the least prediction error for the samples in the same class (i.e., the diagonal entries have the lowest values). However, the same cannot be observed for the breast cancer dataset, which provides the lowest classification accuracies out of five cancer datasets (see Table 1). Second, our results suggest that models for different classes have different prediction errors. For example, for the breast cancer, the model for class 1 produces significantly lower prediction error for the test samples in class 1 as compared to the samples in other classes. However, the model for class 2 fails to predict the samples from its own class, since it gives lower prediction errors for other classes. Similar observations can be seen in the model for class 3. These results suggest that the low classification accuracy (see Table 1, and Supplementary Tables 4 and 5) for breast cancer is because of the inaccurate predictions of cancer patients in classes 2 and 3.

E. Network density. Next, we focus on one of the most fundamental characteristics of the network models constructed by the NBC method, namely, we study the density of the resulting networks (i.e., average number of gene-to-gene associations) formed by the NBC method for different cancer datasets. Figure 3 plots the results for varying number of genes and Pearson correlation threshold values. We observe that network density depends on the number of genes and the correlation threshold. In general, as the number of genes increases and correlation threshold decreases, the number of associations in the network increases. While this qualitative behavior is dataset independent, it shows slight quantitative differences. For example, some of the networks formed by the NBC method for the leukemia, NCI60, and lung cancer datasets are very dense networks (up to ~99, ~52, and ~102 average gene-to-gene associations, respectively). However, the breast and colon cancer datasets show significantly lower density levels.

For the leukemia, lung, and NCI60 cancer datasets, network densities are specific to feature selection methods. For the leukemia, while the genes selected by the PAM method show a very sparse network (up to one average gene-to-gene association), the other feature selection methods have up to ~99 average adjacent genes. Similar behavior is observed for the NCI60 and lung cancer datasets. For the lung cancer, genes selected by the PAM feature selection method give a dense network (up to ~102 average adjacent genes) in comparison to other feature selection methods (up to ~3 average gene-to-gene associations). Similarly, for the NCI60 dataset, SU and $\chi^2$ feature selection methods give a dense network (up to ~52 average gene-to-gene associations) in comparison to other feature selection methods (up to ~5 average gene-to-gene associations).

F. Network measure analysis. Despite the vast amount of experimental and computational studies, we still have limited knowledge about the mechanisms of different cancer types. In order to understand cancer-dependent changes in the correlation-based co-expression networks, here we give a brief analysis of the network measures for the networks created by the NBC method for different cancer classes in leukemia and NCI60 cancer datasets (Figs. 3–6). As suggested above, the best feature selection method for the NBC classifier is the SU feature selection method. Because of that, in this section we focused on the association networks created by the genes selected by the SU feature selection method. Owing to sparse network structures, we omitted the lung, breast, and colon cancer datasets in this experiment (see Figure 3). We compared the networks created for different cancer classes with respect to three network measures, namely degree, clustering coefficient, and closeness centrality distributions of the nodes of the network models generated by NBC. For both datasets, we have used the network, which leads to the best classification of the datasets if up to 100 genes are used (Table 1 and Fig. 1). For the NCI60 dataset, the best accuracy is achieved at 75 genes with a correlation threshold of 0.725, and for the leukemia dataset, 100 genes with a correlation threshold of 0.75.

i. NCI60

In this cancer dataset, all of the correlation-based co-expression networks formed in different cancer classes show scale-free behavior. However, the frequency of isolated genes and the highest degree in the networks show slight variations in different cancer classes. While more than half of the genes in the networks formed by the NBC method for five classes (classes 3, 4, 6, 7, and 8) is isolated (Fig. 4A), in the remaining three classes (classes 1, 2, and 5) the frequency of the genes that are isolated is 50% or slightly less than 50%.
while the maximum degree in the classes 3, 4, and 7 ranges between 3 and 4, in classes 1, 5, 6, and 8, it ranges between 5 and 7. Class 2 shows the highest degree (10) in this cancer type.

Next we measured the clustering coefficient and closeness centrality values for each gene. We observed that in classes 3, 4, 6, 7, and 8, networks have very small clustering coefficients (Fig. 5A), which suggests that in these cancer classes, most of the genes’ neighbors are not associated with each other. In regard to the closeness centrality, these five classes showed centrality scores less than or equal to 8 (Fig. 6A), which is possibly because of the fact that the networks formed in these cancer classes are small because of the many isolated genes in the networks. We observed slightly different behaviors in cancer classes 1, 2, and 5 probably because of the smaller number of isolated genes. In these classes, networks showed slightly more clustering between genes and higher centrality score (9–21) (Figs. 5A and 6A).
These observations suggest that in classes 1, 2, and 5, the expression levels of the genes are slightly more correlated. Because of that, we observed genes with high degree, clustering, and centrality scores.

**ii. Leukemia**

In this dataset, the networks show less number of isolated genes in comparison to the NCI60 dataset. There are 30–35% isolated genes in classes 1, 2, 3, and 5; approximately 20% in classes 6 and 7; and 5% in class 4 (Fig. 4B). The networks for the classes 1, 2, and 3 show scale-free behavior; however, they show slight variations in their distributions. The maximum degree in these three classes ranges from 35 to 55. In classes 4, 5, 6, and 7, the degree distributions show non-scale-free behavior with degrees ranging between 57 and 65. In these classes, the frequency increases as the degree increases. This observation suggests that in classes 4, 5, 6, and 7, probably many genes form a tight clique, leading to higher frequency values for higher degree genes.

Clustering coefficient and centrality measures for each gene in this cancer type show cancer class-dependent behavior also. While in all the seven classes the clustering coefficient distributions show Gaussian behavior, the variance of these
distributions is slightly different. In three of the seven classes (classes 1, 2, and 3), non-isolated genes have clustering coefficients between 0.3 and 1. In contrast, in classes 4, 5, 6, and 7, genes have slightly higher clustering coefficients (0.5–1) (Fig. 5B). In regard to closeness centrality, we observed Gaussian-like distribution for closeness centrality scores in classes 1, 2 and 3 (Fig. 6B), in which centrality scores range between 20 and 65. In contrast to this, in classes 4, 5, 6 and 7, the frequency of the genes with high closeness centrality score is larger. In these classes, we observe more central genes where centrality scores vary between 30 and 75.

This network analysis suggests that the NBC method does not only achieve high classification accuracies for different cancer types, but also reveal potentially important insights about the topological differences among gene associations in various cancer types.

### Discussion/Conclusion

Network-based approaches for cancer classification have been proposed previously, where a combination of protein–protein interaction (PPI) networks and gene expression levels is used in sub-network identification for cancer classification. In this study, we proposed a new network-based classifier, NBC method, and compared its classification accuracy to five traditional classifiers using five different cancer microarray datasets. To reduce the dependency of the classification accuracy results on the genes used, we have employed five alternative feature selection methods to choose up to 300 genes. Our experimental results demonstrated that the choice of feature selection method does not have a very big impact on the classification accuracy. We observed that the SU, $\chi^2$, and IG feature selection methods show comparable but better results than SVM-FS and PAM feature selection methods. Our results also exhibited that our network-based classifier NBC method shows similar or better accuracy levels in comparison to those of the traditional classifiers such as SVM and RF. We also showed that the correct combination of the feature selection and classification method is the key for successful cancer classification studies. In this regard, we observed that NBC and RF classifiers combined with SU feature selection method showed the best overall performances in five different cancer datasets. Our results also support that the selection of the best classifier and feature selection method is dataset specific.

In a recent study, Staiger and colleagues compared the network-based and single-gene-based methods to breast cancer datasets. They tested our method using five different cancer datasets. To reduce the dependency of the classification accuracy results on the genes used, we have employed five alternative feature selection methods to choose up to 300 genes. Our experimental results demonstrated that the choice of feature selection method does not have a very big impact on the classification accuracy. We observed that the SU, $\chi^2$, and IG feature selection methods show comparable but better results than SVM-FS and PAM feature selection methods. Our results also exhibited that our network-based classifier NBC method shows similar or better accuracy levels in comparison to those of the traditional classifiers such as SVM and RF. We also showed that the correct combination of the feature selection and classification method is the key for successful cancer classification studies. In this regard, we observed that NBC and RF classifiers combined with SU feature selection method showed the best overall performances in five different cancer datasets. Our results also support that the selection of the best classifier and feature selection method is dataset specific.

i. The network-based methods compared in Staiger et al. use PPI data in combination with microarray data. The PPI datasets are usually generated by high-throughput biological experiments that results in noisy datasets. Inclusion of these noisy datasets might be leading to the low cancer classification accuracy levels for network-based approaches. Our NBC method differs from these earlier network-based methods. Instead of using existing PPI networks, our method constructs correlation-based association networks. By this way, we believe that our method is not affected by experimental noise as much as the earlier network-based classifiers that use PPI datasets.

ii. Earlier network-based cancer classifiers overlay mRNA expression data with protein-level information. Since these two data types reflect events on different molecular levels, combining them is not trivial and might be resulting in inaccurate results. However, our NBC method uses only gene expression levels and results in better accuracy levels than the single-gene-based classifiers.

iii. The network-based classifiers used in Staiger et al. use a single network for all the cancer classes regardless of the key differences in the gene expression levels of different cancer classes. Then, these methods look for the network motifs that show significant expression-level changes between different cancer classes. Our method, however, constructs a different and unique network for each cancer class and uses these networks to model and classify different cancer types and subtypes. We believe that this significant technical difference between our approach and earlier network-based approaches gives our method an edge over earlier network-based and single-gene-based classifiers.

iv. Staiger et al. compared the network-based and single-gene-based methods only on breast cancer datasets. So, the results observed on this study might be breast cancer specific. In our study, to reduce the dataset specific effects, we tested our method using five different cancer datasets.

The aforementioned differences between the NBC method and earlier network-based approaches are notable, and suggest that our method in contrast to earlier network-based methods is more suitable for cancer classification. Owing to high microarray costs, supervised cancer classification methods are still not employed in many cancer diagnoses. In this sense, new classifiers that can produce accurate classification of different cancer types using small number of genes are needed. Detailed analysis of the NBC method showed that our method could reach to high classification accuracy levels using usually less than 100 genes. In contrast, in general the traditional classifiers require more genes than the NBC method to reach similar accuracy levels. This suggests that our new network-based classifier might be medically more relevant in comparison to the other traditional classifiers. Future work in the medical application of our method to diagnosis of different cancer types is needed to elucidate this strength of our new classifier.
In order to analyze the class-dependent topological differences in gene-to-gene associations in different cancer types, we have also analyzed the network measures (degree, clustering coefficient, and closeness centrality distributions) in leukemia and NCI60 cancer datasets. In-depth analysis of the networks suggested by the NBC method provided new insights into the class-to-class changes of gene-to-gene interactions in cancer. While in some cancer classes we observed scale-free behavior in degree distributions of the genes, this scale behavior was lost in other cancer classes. Similarly, clustering and centrality distributions of the genes show distinct behaviors in different cancer classes. These changes in the network properties suggest that different cancer classes will show distinct responses to similar drugs since their gene regulatory network topologies are different. It also suggests that the design of new cancer drugs should take into account the topological differences in the regulatory networks of different cancer classes. Finally, our study indicates the need for new network-based classification algorithms and analysis techniques to decipher cancer mechanisms and find new therapeutic treatments for cancer.

Acknowledgments
We thank the members of the Kahveci Lab, Haitham Gabr and Mahmudul Hasan, for thoughtful discussions.

Author Contributions
Developed the NBC method: AA, DG, TK. Conceived and designed the experiments: AA, TK. Analyzed the data: AA, TK. Wrote the first draft of the manuscript: AA. Contributed to the writing of the manuscript: AA, TK. Agreed with manuscript results and conclusions: AA, DG, TK. Jointly developed the structure and arguments for the paper: AA, TK. Made critical revisions and approved the final version: AA, TK. All authors reviewed and approved the final manuscript.

Supplementary Files
Supplementary File 1. This document, which includes Supplementary Tables 1–3 and Supplementary Figure 1, describes the Network Based Classifier (NBC) method in detail.

Supplementary File 2. This document, which includes Supplementary Tables 4–7, summarizes results if the maximum number of allowable genes is set to 200 or 300. The tables present 10-fold cross-validation prediction accuracy and accuracy rankings for different feature selection and classification method combinations on lung, breast, NCI60, leukemia, and colon cancer datasets.

REFERENCES
1. Golub TR, Slonim DK, Tamayo P, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science. 1999;286(5439):531–7.
2. Sun Y, Goodison S, Li J, Liu L, Farmerie W. Improved breast cancer prognosis through the combination of clinical and genetic markers. Bioinformatics. 2007;23(1):30–7. doi: 10.1093/bioinformatics/btl543.
3. Ramaswamy S, Tamayo P, Rifkin R, et al. Multiclass cancer diagnosis using tumor gene expression signatures. Proc Natl Acad Sci USA. 2001;98(26):15149–54. doi: 10.1073/pnas.211566398.
4. Pomeroy SL, Tamayo P, Gaasenbeek M, et al. Prediction of central nervous system embryonal tumour outcome based on gene expression. Nature. 2002;415(6870):436–42. doi: 10.1038/415436a.
5. Stephenson CF, Bridge JA, Sandberg AA. Cytogenic and pathologic aspects of Ewing’s sarcoma and neuroectodermal tumors. Hum Pathol. 1992;23(11):1270–7.
6. Lakhani SR, Ashworth A. Microarray and histopathological analysis of tumours: the future and the past? Nat Rev Cancer. 2001;1(2):151–7. doi: 10.1038/3501087.
7. van ‘t Veer LJ, Dai H, van de Vijver MJ, et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature. 2002;415(6871):530–6. doi: 10.1038/415530a.
8. Shipp MA, Ross KN, Tamayo P, et al. Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. Nat Med. 2002;8(1):68–74.

Figure 6. Closeness centrality distributions of the networks in different cancer classes. The closeness centrality distributions of the networks that are created by the NBC method for NCI60 (A) and leukemia (B) datasets. In each graph, x-axis represents the closeness centrality score and y-axis represents the frequency.
9. Lu J, Getz G, Miska EA, et al. MicroRNA expression profiles classify human cancers. *Nature*. 2005;435(7043):834–8.
10. Lu Y, Han J. Cancer classification using gene expression data. *Inf Syst*. 2003;28(4):241–68. doi: 10.1016/S0304-3975(02)00672-8.
11. Pirsonoa M, Yang JY, Yang MQ, Deng Y. A comparative study of different m-RNA microarray background correction methods on microarray expression data. *BMC Genomics*. 2008;9(suppl 1):S13. doi: 10.1186/1471-2164-9-S1-S13.
12. Liu H, Seo S, Chi2: feature selection and discretization of numeric attributes. In: Proceedings of the 7th IEEE International Conference with Artificial Intelligence. Herndon, VA. IEEE: 1995;38-91.
13. Yu L, Liu H. Feature selection for high-dimensional data: a fast correlation-based filter solution. In: Proceedings of the Twentieth International Conference on Machine Learning. Washington, DC. The AAAI Press. 2003;856–63.
14. Wang Y, Tetko IV, Hall MA, et al. Gene selection from microarray data for cancer classification – a machine learning approach. *Comput Biol Chem*. 2005;29(3):37–46. doi: 10.1016/j.compbiolchem.2004.11.001.
15. Tibshirani R, Hastie T, Narasimhan B, Chu G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc Natl Acad Sci U S A*. 2002;99(10):6567–72. doi: 10.1073/pnas.0209929.
16. Lu TP, Tsai MH, Lee JM, et al. Identification of a novel biomarker, SEMA5A, for non-small-cell lung carcinoma in nonsmoking women. *Cancer Epidemiol Biomarkers Prev*. 2010;19(10):2590–7.
17. LaBreche HG, Nevinis JR, Huang E. Integrating factor analysis and a transgenic mouse model to reveal a peripheral blood predictor of breast tumors. *BMC Med Genomics*. 2011;4:61. doi: 10.1186/1755-8794-4-61.
18. Pfister TD, Reinhold WC, Agama K, et al. Topoisomerase I levels in the NCI-60 cancer cell line panel determined by validated ELISA and microarray analysis and correlation with indenoisoquinoline sensitivity. *Mol Cancer Ther*. 2009;8(7):1878–84. doi: 10.1158/1535-7163.MCT-09-016.
19. Zhang J, Ding L, Hoadley KA, et al. The genetic basis of early T-cell precursor acute lymphoblastic leukaemia. *Nature*. 2012;481(7380):157–63. doi: 10.1038/ nature10725.
20. Marisa L, de Reynièrs A, Duval A, et al. Gene expression classification of colon cancer into molecular subtypes: characterization, validation, and prognostic value. *PLoS Med*. 2010;7(5):e1001453. doi: 10.1371/journal.pmed.1001453.
21. Aha DW, Kibler D, Albert MK. Instance-based learning algorithms. *Mach Learn*. 1991;6(1):37–66. doi: 10.1007/BF00151579.
22. Quinlan JR. *C4.5: Programs for Machine Learning*. San Francisco, CA. Morgan Kaufmann Publishers Inc. 1993:302. doi: 10.1016/S0025-5472(62)90649-6.
23. Friedman N, Linial M, Nachman I, et al. Using Bayesian networks to analyze expression data. *J Comput Biol*. 2000;7(3–4):301–20. Available at: http://eprints.ecs.soton.ac.uk/9578/1.
24. Cortes C, Vapnik V. Support-vector networks. *Machine Learn*. 1995;20(3):163–37. doi: 10.1023/A:1008086819398.
25. Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. *Bioinformatics*. 2007;23(19):2507–17. doi: 10.1093/bioinformatics/btm204.
26. Jin J. Impossibility of successful classification when useful features are rare and weak. *Proc Natl Acad Sci U S A*. 2009;106(22):8859–64. doi: 10.1073/pnas.090331106.
27. Lu X, Peng S, Zhan X, Zhang J, Xu Y. Comparison of feature selection methods for microarray cancer classification based on microarray data. *Biomed Eng Inform*. 2011;4(3):1692–6. Available at: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6098612.
28. Li T, Zhang C, Oghara M. A comparative study of feature selection and multi-task classification methods for tissue classification based on gene expression. *Bioinformatics*. 2004;20(15):2429–37. doi: 10.1093/bioinformatics/bth267.
29. Ding C, Peng H. Minimum redundancy feature selection from microarray gene expression data. *J Bioinform Comput Biol*. 2003;5:185–205.
30. Hall MA. *Correlative-Feature Selection for Discrete and Numerical Class Machine Learning*. 2000;359–66. Available at: http://dl.acm.org.citation. dfn?id=465259.67793.
31. Breiman L. *Random forests*. *Mol Syst Biol*. 2008;4:120. doi: 10.1038/msb4100362.
32. Jin J, Li J, Wong L. A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns. *Genome Inform*. 2002;13:51–60.
33. Wang Y, Tetko IV, Hall MA, et al. Gene selection from microarray data for cancer classification – a machine learning approach. *Comput Biol Chem*. 2005;29(3):37–46. doi: 10.1016/j.compbiolchem.2004.11.001.
34. Vinaya V, Buhara N, Gadgil CJ, Gadgil M. Comparison of feature selection and classification combinations for cancer classification using microarray data. *Int J Bioinform Res Appl*. 2009;5(4):32170. doi: 10.1080/17455192.2010.5098612.
35. Segalinas S, Tomlinson I. Colorectal cancer and genetic alterations in the Wnt pathway. *Oncogene*. 2006;25(57):7371–7. doi: 10.1080/09609860500540979.
36. Choudhury SA, Koyutürk M. Identification of coordinately dysregulated subnetworks for multiclass cancer classification based on microarray data. *PLoS One*. 2012;7(2):e302170. doi: 10.1371/journal.pone.0030217.
37. Ideo H, Seko A, Yamashita K. Galectin-4 binds to sulfated glycosphingolipids and carcinoembryonic antigen in patches on the cell surface of human colon adenocarcinoma cells. *J Biol Chem*. 2005;280(6):6470–7. doi: 10.1074/jbc.M410362200.
38. Venkateswaran V, Buhara N, Gadgil CJ, Gadgil M. Comparison of feature selection and classification combinations for cancer classification using microarray data. *Int J Bioinform Res Appl*. 2009;5(4):32170. doi: 10.1080/17455192.2010.5098612.
39. Dowling SB, Swamy SN. Analysis and correlation with indenoisoquinoline sensitivity. *Mol Cancer Ther*. 2005;435(7043):834–8. doi: 10.1038/sj.onc.1210059.
40. Liu H, Li J, Wong L. A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns. *Genome Inform*. 2002;13:51–60.
41. Quinlan JR. *C4.5: Programs for Machine Learning*.