Journal of Bioinformatics and Computational Biology

MULTI-RESOLUTION-TEST FOR CONSISTENT PHENOTYPE DISCRIMINATION AND BIOMARKER DISCOVERY IN TRANSLATIONAL BIOINFORMATICS

HENRY HAN

Department of Computer and Information Sciences
Fordham University, New York, New York 48105, USA
Quantitative Proteomics Center, Columbia University,
New York, New York 10027, USA
xhan9@fordham.edu

XIAO-LI LI

Institute for Infocomm Research, Agency for Science, Technology and Research,
Singapore, 138632, Singapore
xlli@i2r.a-star.edu.sg

SEE-KIONG NG

Institute for Infocomm Research, Agency for Science, Technology and Research,
Singapore, 138632, Singapore
skng@i2r.a-star.edu.sg

ZHOU JI

Center for Computational Biology and Bioinformatics
Columbia University, New York, NY 10032 USA
zji@c2b2.columbia.edu

While high-throughput technologies are expected to play a critical role in clinical translational research for complex disease diagnosis, the ability to accurately and consistently discriminate disease phenotypes by determining the gene and protein expression patterns as signatures of different clinical conditions remains a challenge in translational bioinformatics. In this study, we propose a novel feature selection algorithm: Multi-Resolution-Test (*MRT-test*) that can produce significantly accurate and consistent phenotype discrimination across a series of omics data. Our algorithm can capture those features contributing to subtle data behaviors instead of selecting the features contributing to global data behaviors, which seems to be essential in achieving clinical level diagnosis for different expression data. Furthermore, as an effective biomarker discovery algorithm, it can achieve linear separation for high-dimensional omics data with few biomarkers. We apply our *MRT-test* to complex disease phenotype diagnosis by combining it with state-of-the-art classifiers and attain exceptional diagnostic results, which suggests that our method's advantage in molecular diagnostics. Experimental evaluation showed that *MRT-test* based diagnosis is able to generate consistent and robust clinical-level phenotype separation for various diseases.

In addition, based on the seed biomarkers detected by the *MRT-test,* we design a novel network marker synthesis (NMS) algorithm to decipher the underlying molecular

mechanisms of tumorigenesis from a systems viewpoint. Unlike existing top-down gene network building approaches, our network marker synthesis method has a less dependence on the global network and enables it to capture the gene regulators for different subnetwork markers, which will provide biologically meaningful insights for understanding the genetic basis of complex diseases.

Keywords: Biomarker; feature selection; subnetwork.

1. Introduction

As high-throughput methods in experimental biotechnology reach technological maturity, genomic technologies such as gene expression array are expected to play a critical role in clinical translational research for complex disease diagnosis. However, there are stringent requirements in genomic medicine that have to be met. The ability to accurately and consistently discriminate disease phenotypes by determining gene expression patterns as signatures of different clinical conditions (e.g., three pathological stages of a tumor) is of particular importance. This has remained a challenge in translational bioinformatics due to the special characteristics of gene expression data. A gene expression dataset can be represented as a $p \times n$ matrix, where each column represents a gene (variable) and each row represents a sample (observation)[a]. The number of variables (genes) is typically much greater than the number of observations, i.e. $n \gg p$, even though only a small portion of the genes actually contribute biologically meaningful data variations for disease detection. Traditional pattern recognition methods, designed for low or medium dimensionalities of input data (smaller n that is more compatible with p), fail to achieve the requirements of clinical-level diagnosis.

Furthermore, gene expression data are notoriously noisy due to profiling variations, experimental design artifacts, and technical biases. In fact, normalization techniques are even ineffective in removing the 'systems noise' completely. They may also introduce false positives into the data[1]. Thus, translational bioinformatics methods for effective disease biomarker discovery and robust phenotype diagnosis from gene expression data must overcome the high dimensionality, information redundancy, and built-in noises, because they usually have caused traditional classification methods (e.g., discriminant analysis) to lose discriminative power.

The existing methods of biomarker discovery can be categorized into two groups, namely, gene-set methods and network methods. The gene-set methods identify a set of significant genes, i.e., gene markers from gene expression profiles using complicated biomarker capturing approaches (e.g., filter-wrapper methods) or feature selection methods to discriminate different disease phenotypes. Although widely employed, as a result of the gene expression data challenges mentioned above, the gene lists obtained from different gene expression studies can differ widely and share few common genes[2]. As such, there is no guarantee that good diagnosis achieved

[a]It is different from the traditional notation where each row is a variable (e.g., a gene) and each column is a sample for the convenience of algorithm description.

from one dataset can be generalized to another. In fact, the phenotype diagnostic performance from almost all gene-set methods lack stability and demonstrates strong dependence on input data. A method may work quite well for one dataset but fail badly on another.

Alternatively, the network methods[2] don't view a biomarker as an individual genes. Instead, they view each biomarker as a small network, i.e., a subnetwork by incorporating protein-protein interaction (PPI) or even pathway information with the expression profiles. They employ subnetwork identification algorithms such as jActiveModule[3], PNA[4], and COSINE[5] to look for a set of subnetwork markers, which are usually orthogonal to each other, from expression data.

Although such a systems approach claimed in-depth insights into the underlying biological processes of complex diseases, they still have hard time to improve disease diagnostic accuracy not to mention to attain rivaling clinical phenotype discrimination. Especially, a noisy global protein-protein interaction (PPI) network may lead to subnetwork markers to include some unrelated "false positive" genes, which will decrease the discriminative power of the network markers and lead to poor diagnosis, because these methods usually search subnetwork markers from the global PPI network in a top-down fashion. Moreover, the subnetwork markers are usually orthogonal to each other and no information is available about gene regulators connecting them. However, those information can be essential to fully understand the dynamics of a disease from a systems viewpoint.

In this work, we develop a *de novo* feature selection method: Multi-Resolution-Test (*MRT-test*) to identify meaningful biomarkers from expression data. The *MRT-test* is based on our previous studies in local and global feature selection[6,7]. We have found that our *MRT-test* is an exceptional way to extract subtle data characteristics in high dimensional omics data by capturing the local features of data in addition to removing system noise. Alternatively, as a novel disease biomarker discovery algorithm, our proposed *MRT-test* is particularly useful (as compared to existing methods) as it is able to identify meaningful biomarkers and demonstrate the linear separability of high-dimensional expression data.

We demonstrate that, by integrating our *MRT-test* with the state-of-the-art classifiers (e.g., SVM), we are able to achieve consistent clinical-level diagnosis accuracy across a series of gene expression data, whereas existing methods typically can only work well in one or two omics data. It is noted that we mainly use gene expression data in this study, though we also demonstrate our method's effectiveness in proteomics data. Furthermore, we propose a *MRT-test* based network marker discovery method: '*network marker synthesis*' (*NMS*) to seek network markers by 'growing' the *'seed'* gene markers from the *MRT-test.* As a bottom-up network approach, it has less dependence on the global network for its more targeted search and enables the identification of gene regulators for the subnetwork markers to reveal underlying tumorigenesis, in addition to demonstrate rivaling clinical phenotype diagnosis, because of its more targeted search initialized by our *MRT-test* identified gene markers.

2. Methods

We provide some background on feature selection in section 2.1 before we introduce our Multi-Resolution-Test *(MRT-test)* method in section 2.2.

2.1. *Background on feature selection*

The proposed *MRT-test* extends our previous work in local and global feature selection to input-space feature selection[6,7]. We first categorize feature selection as either subspace or input-space feature selection before introducing the *MRT-test*. The former seeks to find meaningful feature combinations in a low dimensional subspace induced by a linear or nonlinear transform (e.g., PCA). The latter selects a feature subset in the same space as input data through conducting a statistical test (e.g., *t-test*) or building different feature filtering models according to different selection criteria (e.g., filter-wrapper methods). Due to the high dimensionality of gene expression data, input-space feature selection methods are often used for its efficiency and simplicity. The two-sample *t-test* and its variants are probably the most popular input-space methods in gene expression analysis. However, similar to other input-space approaches, they have the following limitations.

First, these methods usually assume input data '*clean*' or '*nearly clean*' and ignore necessary de-noising in feature selection. Such an assumption may not be appropriate because the noises in gene expression data can be non-linear. This means that they are more than white noises or non-differential ones, and they can cause statistical tests (e.g., the *t-test* and even its more sophisticated extensions) to lose robustness. For example, the noises would behave as 'outliers' by participating in *t-score* calculations for each gene, and cause the sample means from the two groups of a gene (e.g., cancer vs. control) to be very close, which usually results in a zero or an approximately zero *t-score*. As such, the *t-test*'s robustness may be significantly decreased, leading to biased feature selection, i.e., "*pseudo-significant*" genes are selected while "*truly-important*" genes are filtered. This may inevitably limit the subsequent classifier's phenotype discrimination, generalization, and stability.

Second, these methods are single resolution approaches, where all features are indistinguishably displayed in one resolution despite the nature of their frequencies. As such, the high frequency features will have a high likelihood to be picked up in feature selection than those low frequency ones. The high (low) frequency features refer to those features which appears the more (less) often in the input space respectively. Furthermore, we refer to the high frequency and low frequency features in the input space as global features and local features, which describe global/local data characteristics and capture general/subtle data behaviors, respectively. The local features can be viewed as signals occurring at a short time interval (e.g., transient signals with sharp peaks), while the global features can be viewed as signals observed often in a long time interval.

For example, a gene whose expression plot curve is similar to those of most other genes is a global feature. On the other hand, a gene with several exceptional

local "high peaks" on control, which are rarely found on most other genes, would be a local feature. The local features are key to discriminating samples sharing similar global characteristics but with different local characteristics to achieve high diagnosis accuracy in translational bioinformatics. For example, Different subtypes of tumors may share similar global characteristics but different local characteristics in their expression data according to their different pathological states.

However, the local features are hard to extract by traditional feature selection methods, where each feature is an indecomposable information unit, because the global features may always have a high likelihood to be selected for their high-frequency. As a result, there are redundant global characteristics involved in disease phenotype discrimination because global feature dominated data will enter the subsequent classifier (e.g., SVM) used for diagnosis. However, the redundant global features may overshadow the subsequent classifier's training phrase and lead to a decision rule only favoring those global features. As such, the resulting classifier would fail to diagnose the samples with similar global characteristics but different local characteristics.

To some degree, the redundant global features act as '*external noise*' along with the built-in noise to affect the classifier's discrimination and generalization capabilities, and increase the risk of mis-diagnosis and over-fitting. For this reason, such classifiers may demonstrate large oscillations in performance for different data, i.e., it may fit some data well but poorly on others, due to the global features' unpredictable contributions to phenotype discrimination. In this study, we present a novel feature selection model: Multi-Resolution-Test (*MRT-test)* designed to extract local features, conduct de-noising, and avoid redundant global features so as to enhance biomarker discovery by detecting subtle changes in expression of the few genes causing disease, and achieve rivaling clinical phenotype diagnosis.

2.1.1. *Multi-Resolution-Test (*MRT-test)

Unlike traditional feature selection methods, the *MRT-test* makes local feature extraction and de-nosing possible by incorporating a screening mechanism to separate the global and local features via discrete wavelet transform (DWT)[8] based multi-resolution analysis. It is noted that we view each entry of a gene in the DWT as a sampling point at a specific time point, and the entries belonging to different types (e.g., cancer type) are points sampled at different time intervals.

For example, given a gene from an omics data set with 40 cancer and 50 control samples, we view the 40 entries from the cancer type and 50 entries from the control type as the sampling points from time intervals 1 and 2 respectively. The reason for this assumption is that we implicitly view gene expression data as a kind of special time series data, where different types of observations are samples from different biological time intervals. For example, three pathological stages of a certain tumor correspond to three different biological time intervals.

The *MRT-test* can be sketched as following steps. At first, a discrete wavelet

transform (DWT) with a transform level L is applied to input data X column-wise to decompose it hierarchically as L detail coefficient matrices: $D_1, D_2, \cdots D_L$, and an approximation matrix A_L. It is noted that we assume input expression data X have n columns of genes and p rows of samples in our context. Since DWT is done on a set of dyadic grid points hierarchically, the dimensionalities of the approximation and detail coefficient matrices shrink dyadically level by level.

For example, given input data with 100 samples across 1024 genes under the DWT with a transform level $L = 4$, D_1 is a 100×512 matrix and D_2 is 100×256 matrix. Similarly, D_4 and A_4 both are 100×64 matrices. That is, given input expression data with p samples and n genes $X = [x_1, x_2, \ldots x_n]$,$x_j \in \Re^p$, a L-level DWT is applied to such data set gene by gene to obtain L number of detailed coefficient matrices $D_j \in \Re^{p_j \times n} (j = 1, 2, \cdots L)$ and an approximation matrix $A_L \in \Re^{p_L \times n}$, i.e., such a transform can be summarized as $T = \{D_1, D_2, \cdots D_L, A_L\}$, where $p_j \sim p/2^j$, $j = 1, 2, \cdots, L$.

The approximation matrix and coarse level detail coefficient matrices (e.g., D_{L-1}, D_L) capture the global data characteristics, because they contain contributions from the features disclose data behaviors usually often observed in 'long-time windows'. Similarly, the fine-level detail coefficient matrices (e.g., D_1, D_2), capture the subtle (latent) data characteristics, because they contain contributions from those features describing less-often data behaviors observed in 'short-time windows'. In fact, these fine-level detail matrices are components to reflect data derivatives in different time windows. Furthermore, most system noise is hidden in these components for its heterogeneity with respect to the 'true' signals. In summary, the first step separates the global characteristics, subtle data characteristics, and noise in different resolutions.

Second, we retrieve the most important subtle data behaviors and remove noise by reconstructing the fine-level detail coefficient matrices before or at a presetting cutoff level τ (e.g., τ=3). Such construction consist of two steps:

1) We apply principal component analysis (PCA) to each of the detail coefficient matrices $D_1, D_2 \cdots D_\tau$ to obtain its principal component (PC) matrix U and score matrix S.

2) We employ the first loading vector (1^{st} PC), the 1^{st} column in the PC matrix U, to reconstruct each coefficient matrix D_j, $j = 1, 2, \cdots \tau$, to retrieve the most important subtle data behaviors described by this fine-level detail coefficient matrix. We have to point it out that such the first principal component based reconstruction is actually a de-nosing process also. This is because only the maximum variance direction (the 1^{st} PC direction) is employed in reconstructing each targeted fine-level coefficient matrix and those less important and noise-contained principal components are dropped in reconstruction.

In fact, we have found that the first PC direction counts quite a large percentage of data variance (e.g., 60%) for each fine-level detail coefficient matrix in a gene/protein expression array data. Moreover, as we pointed out before, the system

noise is usually transformed to those fine-level detail coefficient matrices (e.g., D_1) for its heterogeneity by the DWT. Thus, using the 1^{st} PC in the fine-level coefficient matrix reconstruction would guarantee the maximum de-noising and the most significant subtle data behavior retrieval.

Alternatively, the coarse level detail coefficient matrices after the cutoff τ: $D_{\tau+1}, D_{\tau+1} \cdots D_L$ and approximation coefficient matrix A_L are kept intact to retrieve these global data characteristics. Since the local data characteristics extraction and system noise removal are achieved through decomposing features in the multi-resolution data approach, we avoid the global feature dominance problem faced by the traditional feature selection methods, which would contribute to the following phenotype diagnosis. It is worth pointing out that, although theoretically feasible, using several leading loading vectors instead of the 1^{st} PC in the detail coefficient matrix reconstruction may not achieve desirable de-noising, especially under a 'large' τ value (e.g., $\tau = 16$).

Third, a corresponding inverse DWT is applied to the current coefficient matrices and the approximation matrix to calculate a meta-profile X^*, which is a same dimensional synthesis of the original data but catch subtle data characteristics and has less memory storage because less important principal components are dropped in our detail coefficient matrix reconstruction. Moreover, the meta-feature of each feature (e.g., gene), i.e., the expression values of the feature in X^*, has values in a relatively smaller range than that of the original feature due to the leading PC based detail coefficient matrix reconstruction. For the same reason, the data variances of the meta-profile X^* are smaller than those of the original omics data X.

Finally, we employ *t-statistic* and *F-statistic* to score each feature, i.e., gene, for the binary class and multi-class cases respectively. The genes with the most statistically significant scores will be chosen as gene markers. It is noteworthy that de-noising and local feature extraction process in steps 2 and 3 make the phenotype distributions in the meta-profile X^* fit better into the normal distribution than those of the original data X, which theoretically contributes to the effectiveness of the *MRT-test.* The detailed *MRT-test* algorithm is presented as follows.

Algorithm 1 Multi-Resolution-Test (*MRT-test*)

(1) *Wavelet multi-resolution analysis.* Given an expression dataset with p observations and n variables $X = [x_1, x_2, \ldots x_n]$, $x_j \in \Re^p$, $n \gg p$, a L-level DWT is applied each column to obtain L detailed coefficient matrices $D_j \in \Re^{p_j \times n}$, $(j = 1, 2 \cdots L)$ and an approximation matrix $A_L \in \Re^{p_L \times n}$, i.e., $T = \{D_1, D_2, \ldots D_L, A_L\}$, where $p_j \sim p/2^j$, $j = 1, 2 \cdots L$.

(2) *De-noising and local feature selection.* A level threshold $1 \leq \tau < L - 1$, is selected to conduct de-noising and local feature extraction.

   (a) Case 1: $1 \leq j \leq \tau$

      i. Apply principal component analysis for each matrix D_j to obtain its principal component (PC) matrix $U = [u_1, u_2, \ldots u_p]$, $u_i \in D_j \in \Re^{p \times 1}$, and cor-

responding score matrix $S = [s_1, s_2, \ldots, s_p]^T$, $s_k \in \Re^{n_j}$, $(k = 1, 2, \ldots, p)$.

ii. Reconstruct each matrix D_j by using the first loading vector u_1 in the PC matrix as $D_j \leftarrow (1/p_j)D_j(\overrightarrow{1})_{p_j} + u_1 \times s_1^T$, where $(\overrightarrow{1})_{p_j} \in \Re^{p_j \times 1}$, with all entries being '1's.

(b) Case 2 : $j > \tau$, keep all matrices $D_{\tau+1}$,$D_{\tau+2}$,$\cdots D_L$ intact.

(3) *Meta-profile synthesis.* Conduct the corresponding inverse DWT using the updated coefficient matrices $T = \{D_1, D_2, \ldots D_L, A_L\}$ to reconstruct $X^* \in \Re^{p \times n}$, which is a same-dimensional de-noised data with local feature extracted.

(4) *Meta-profile hypothesis testing.*

(a) For a binary class input data, calculate a t-statistic ($t = |\bar{x} - \bar{y}|/\sqrt{s_x^2/n_x + s_y^2/n_y}$) for each gene/feature in the meta-profile X^*, where $\bar{x}$ and $\bar{y}$ are the sample means, s_x^2 and s_y^2 are the sample variances, and n_x and n_y are the numbers of samples in the control and disease classes respectively.

(b) For the multiclass input data with $k > 2$ types, calculate an F-statistic for each gene as $F = (\sum_j^k n_j(\overline{x_j^*} - \overline{\overline{x^*}})^2/(k-1))/(\sum_{j=1}^k (n_j - 1)s_j^2/(n_T - k))$,where n_j is the sample size, parameters, and $\overline{x_j^*}$ and s_j^2 are the sample mean and sample variance for the *j-th* class. $\overline{\overline{x^*}}$=$\sum_{j=1}^k \sum_{i=1}^{n_j} x_{ij}^*/n_T$ is the overall sample mean where x_{ij}^* is the expression value of *i-th* observation for the class j and $n_T = \sum_{j=1}^T n_j$ is the total sample size for the k groups.

(c) Select a feature set S that includes statistically significant genes by picking genes with the smallest *p-values* (or the largest testing statistic) from t-distribution or F-distribution with a pre-specified significant value α (e.g., $\alpha = 0.05$). These genes in S are different from the original ones because they went through de-noising and local feature extraction.

We uniformly conduct a 12-level DWT with a '*db8*' wavelet for input data by setting the level threshold $\tau = 4$ in the *MRT-test* due to its good performance. It is noted that we require the wavelet ψ employed in the DWT to be orthogonal and have a compact support such as *Daubechies* wavelets for the sake of subtle data behavior capturing.

In fact, we suggest an empirical threshold: $3 \leq \tau \leq \lceil L/2 \rceil$. A threshold (say, $\tau = 1$) that is too low may affect the quality of de-noising as some noise may still enter the meta-profile through a fine coefficient matrix (e.g., D_2) without reconstruction, which can cause our method to fail to extract some local features. On the other hand, using a threshold (say, $\tau > \lceil L/2 \rceil$) that is too high may miss some global features for the first load vector based coefficient reconstruction. In this paper, we choose the wavelet '*db8*' as it is not only orthogonal but also smooth for capturing subtle data behavior in DWT.

3. Results

We have performed comprehensive experiments to evaluate our proposed *MRT-test* based techniques for four different tasks, namely, biomarker discovery (Section 3.2), phenotype diagnosis for binary and multiclass data (Section 3.3), comparison with the state-of-the-art diagnostic methods (Section 3.4), as well as a *MRT-test* induced network marker identification (Section 3.4), in addition to applying the *MRT-test* to proteomics data (section 3.5).

3.1. *Data sets*

We employ six binary and two multiclass benchmark gene expression array data in our experiment. We say these data are '*heterogeneous*' because they were produced from different profiling technologies, experiments, and processed by different normalization methods. We summarize the data sets in a little detailed way.

Stroma data consist of 47 samples from two subtypes of breast cancers: 13 inflammatory breast cancers ('*ibc*') and 34 non-inflammatory breast cancers ('*non-ibc*') across 18995 selected genes[9]. Inflammatory breast cancer is a relatively rare but very aggressive disease that cancer cells block lymph vessels in the skin of breast. Compared with other types of breast cancer, it progresses rather rapidly and is very hard to diagnosis, though a PET or CT scan may help diagnosis. It has a low prognosis ratio (e.g., 34% 5-year relative survival). A molecular diagnosis via gene expression array is essential to early discovery of such cancer and its prognosis.

Colon data may be one of earliest benchmark data in translational bioinformatics that consist of 22 control and 40 cancer samples across 2000 selected genes[10]. Since it is a data published in early years (1999), we can view the oligonucleotide array technology used by this data set is different from current oligonucleotide array technologies. However, it can be a good candidate to validate the effectiveness of our *MRT-test* algorithm across different types of data.

Medulloblastoma data consist of 25 *classic* and 9 *desmoplastic* tumor samples across 5893 genes that represent two subtypes of medulloblastomas, an early childhood carcinoma whose pathogenesis is not well understood yet[11]. Similar to the *Colon* data, both *Prostate*[12] and *Breast*[13] data are widely used benchmark data but generated by early oligonucleotide array technologies. The former consists of 59 normal controls and 77 prostate cancers across 12625 selected genes, and the latter consists of 46 tumor samples with metastasis in 5 years and another 51 tumor samples without metastasis in 5 years across 24188 probes. It is noted that the *Breast* data set is also well-known for its low diagnosis ratio (classification accuracy) for different classifiers[2,7].

Different from the previous carcinoma expression data, *Smoke* data consist of the gene expression of 34 current smoking subjects and 23 never-smoking subjects by high-density gene expression arrays[14]. Alternatively, as its three-class extension, *SMOKE* data includes the 18 former smokers in addition to 34 current smokers and 23 non-smokers[14]. Although such data demonstrated how smoking altered the

transcriptome, and why former smokers had risk to develop lung cancer even long after they quitted smoking, it was important but still unexplored yet to distinguish current smokers, former smokers and non-smokers from their expression data from a translational bioinformatics approach.

As a widely used tumor data, *CNS (*central nervous system) data set consists of 10 medulloblastomas ('*Md*'), 10 malignant gliomas ('*Mg*'), 10 rhabdoids ('*Rhab*'), 4 normal human cerebells ('*Ncer*'), and 8 supratentorial PNETs ('*PNET*') across 5669 selected genes[15]. Although previous work showed medulloblastomas were molecularly distinct with the other brain tumors[15], it is quite challenging to separate all the five brain tumors with each other completely. Tables 1 sketch information about these '*heterogeneous*' expression data. In addition to these eight major omics data sets, we also include a gene expression array data: *wang-breast,* which consist of 163 breast cancer patients with *ER≥10 fmol per mg* protein and 53 breast cancer patients with $ER < 10$ *fmol per mg protein* across 2000 selected genes in this study[16].

Table 1. Benchmark gene expression array data

Data	**#Gene**	**#Sample**
Stroma	18995	13 *inflammatory breast cancer ('ibc')* + 34 *non-inflammatory breast cancer ('non-ibc')*
Colon	2000	22 controls+ 40 colon cancers
Medulloblastoma	5893	25 *classic* + 9 *desmoplastic*
Prostate	12625	59 controls + 77 prostate cancers
Smoke	7939	34 *'smoking'* +23 *'non-smoking'*
Breast	24188	46 patients *with metastasis in 5 years* + 51 *without metastasis in 5 years*
SMOKE	7939	18 *'smoked'* + 34 *'smoking'* + 23 '*non-smoking'*
CNS	5669	10 *Md* + 10 *Mg* + 10 *Rhab* + 4 *Ncer* + 8 *PNET*

3.2. Identify biomarkers by Multi-Resolution-Test (MRT-test)

As a local feature selection algorithm with a de-noising scheme, the *MRT-test* is also a biomarker discovery algorithm that is effective in discovering specific genes as biomarkers that are able to disclose subtle data behaviors in addition to general data behaviors. We systemically demonstrate effective biomarker identification by the *MRT-test* for binary-class and multiclass data. Unlike other biomarker discovery methods, the proposed *MRT-test* can completely separate phenotypes of high-dimensional array with few top-ranked genes. In contrast with general assumption that high-dimensional gene expression arrays are non-separable nonlinear

data, Our results strongly demonstrate they are actually linearly separable data through concrete examples.

3.2.1. *Biomarker discovery for binary-class data*

We firstly compare 3 top-ranked genes (biomarkers) from the *MRT-test* with those from the *t-test* as well as 3 randomly picked genes for *Stroma* data set, a binary-class array data with 13 '*ibc*' and 34 '*non-ibc*' samples. Figure 1 separates *Stroma* data's 47 samples with 3 randomly selected genes, the top 3 ranked genes from *t-test* and from *MRT-test*. We choose two different cutoff values $\tau = 3, 4$ for the *MRT-test* in our experiment to fully explore its biomarker discovery mechanism.

Each blue (yellow) dot in Figure 1 represents an '*ibc*' ('*non-ibc*') sample. The x, y, and z axes represent expression values of the 1^{st}, 2^{nd}, and 3^{rd} gene marker respectively. The top 3 genes from the *t-test* showed some degree of phenotypic separation than those randomly selected ones, which indicates the *t-test* demonstrating some degree phenotype discrimination ability. However, it is still unable to separate the two types of phenotype samples clearly. This implies that while the *t-test* can be used for feature selection, it is ineffective as a biomarker discovery algorithm for separating phenotypes.

In contrast, the two sets of top-ranked three genes from the *MRT-test* both demonstrated clear spatial separations for the 13 '*ibc*' and 34 '*non-ibc*' samples. It is noted that the samples sharing the same phenotype are clustered automatically in an independent group with a well-built spatial boundary that separates the two different phenotypes.

On closer investigation, three sub-clusters can be further identified for the '*non-ibc*' samples, which may indicate different pathological stages of '*non-ibc*' breast cancer. Such '*self-clustering*' of the samples from the different phenotypic classes (or even subclasses) strongly suggests that the *MRT-test* is a much more effective biomarker discovery algorithm that can detect discriminative gene markers.

It is worth noting that the top-ranked gene markers detected by our *MRT-test* also demonstrated significant biological relevance with breast cancer. '*GOLGB1*', the top-ranked gene both with $\tau = 3$ and 4, has been reported as a gene that is strongly correlated to the inflammatory carcinoma of breast[9 17]. It also interacts with gene '*BRMS1*', a breast cancer metastasis suppressor, and gene *'ACBD3'* that plays an important role in asymmetric cell division and breast cancer[18].

The second and third top-ranked gene markers from our *MRT-test* ($\tau = 3$) are '*CCNT2*' and '*CBY1*', both of which were also reported to have tissue expression in breast tumors. In addition, the '*CCNT2*' gene regulates '*CDK9*', which was reported to be closely related to the breast cancer in Johnston et al 's work[19]. '*CBY1*', the 3rd top-ranked gene, can inhibit wingless pathway by binding to beta-catenin, which is a transcriptional activator and oncoprotein involved in the development of breast tumor and other cancers[20].

In addition to sharing the first top-ranked gene, the third top-ranked gene *'CBY'*

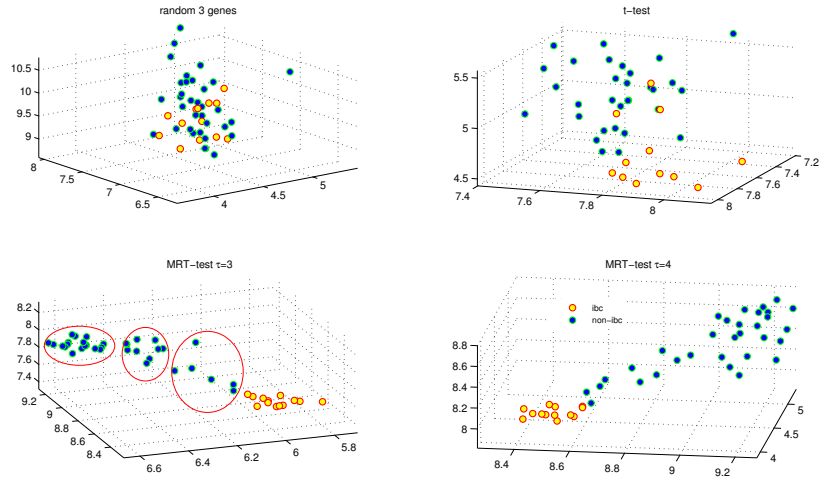

Fig. 1. Separating 13 '*ibc*' and 34 '*non-ibc*' samples with three genes selected by *random gene selection, t-test,* and *MRT-test. The MRT-test* demonstrates clear phenotype separation

from the *MRT-test* at τ=4 is also the third top-ranked gene at $\tau = 3$, which suggests the reproducibility of our *MRT-test.* The second top-ranked from the *MRT-test* at $\tau = 4$ is *'TOX3'*. As a previously identified biomarker for breast cancers, Easton et al. suggested that mutations from the '*TOX3*' gene are associated with an increased risk of breast cancers[21].

It is noted that we also compared our *MRT-test* with other feature selection methods such as *permutation t-test* (*pt-test*), *Bayesian t-test* (*bt-test*), and Mann-Whitney tests (*u-test*) by finding the three top-ranked genes according to their *p-values*. We have found all these methods have achieved almost same level performance as the *t-test.* That is, they are unable to achieve a perfect phenotype separation like the *MRT-test*, though they demonstrate better phenotype separation than the random gene selection. Such a result suggests these methods are just general feature selection methods instead of a biomarker discovery algorithm like our *MRT-test.* Furthermore, we have to point out that the reason we use the three top ranked genes for biomarker discovery is just to demonstrate the gene markers identified by our *MRT-test* can separate the phenotype linearly from a 3D visualization viewpoint. In fact, more top ranked genes (e.g., top 20 scored genes) can be invited in biomarker discovery.

3.2.2. *Compare top-ten ranked genes from the* MRT-test *and* t-test.

Table 2 shows the top 10 genes ranked by the *MRT-test* ($\tau = 3$) *and t-test* . Interestingly, the *p-values* obtained from the *MRT-test* for each gene are much

Table 2. Top 10 genes ranked by the *MRT-test* ($\tau = 3$) and *t-test*

MRT-test			*t-test*	
Rank	**Gene**	***p-value***	**Gene**	***p-value***
1	'GOLGB1'	4.4183e-017	'USP46'	4.8681e-005
2	'CCNT2'	4.9947e-015	'ARFRP1'	5.0277e-005
3	'CBY1'	1.4475e-014	'INPP5E'	5.0920e-005
4	'EIF4A2'	7.9884e-014	'GOLGB1'	1.1635e-004
5	'TBL1X'	8.0746e-013	'MAGED2'	1.1651e-004
6	'DIDO1'	8.4294e-013	'DKFZP686A01247'	1.6098e-004
7	'ABAT'	1.3722e-012	'DNAJB9'	1.7203e-004
8	'LOC400451'	1.4843e-012	'TTC3'	1.7507e-004
9	'COPB1'	1.5404e-012	'DZIP3'	1.7597e-004
10	'SPI1'	3.7805e-012	'DNAJB9'	1.9878e-004

smaller than those of in the *t-test,* which statistically indicates the better sensitivity of our *MRT-test,* which will be further validated in the next section.

In addition to the top three genes, other genes identified by the *MRT-test* were reported to be related to breast cancer. For example, EIF4A2 was inversely expressed in breast cancer[22]; TBL1X was reported over-expressed in breast cancer[23]; DIDO1 (Death inducer-obliterator) was related to expression level control of breast cancer[24]; ABAT and SPI1 were associated with the prognosis of breast cancer[25]. However, there are 5 genes among the total 9 genes (one is duplicate) in the *t-test*'s gene list are not reported to be related to breast cancer in current available literature such as USP46, ARFRP1, TTC3, DKFZP686A01247, and DZIP3, which may suggest the ad-hoc of the *t-test.* The top-10 genes ranked by the *MRT-test at* $\tau = 4$ can be found in Table S1 in the supplemental materials, which can be found at `https://sites.google.com/site/heyaumsystembiology/`.

3.2.3. *MRT-test for Colon, Breast, Medulloblastoma, and Wang-breast data.*

Gene expression array technologies have seen great progresses in recent years, which may contribute to enhancing biomarker's capability in phenotype separation. However, the biomarkers identified by our *MRT-test* ($\tau = 4$) demonstrate exceptional phenotype separation capability for different array data produced by different time and experiments. Figure 2 demonstrates phenotype separation by using the three gene markers identified by the *MRT-test* for four data sets: *Colon*, *Breast*, *Medulloblastoma,* and *Wang-breast data,* published in 1999, 2002, 2003, and 2005 respectively[10,13,11,16].

It is noted that the *Colon* data has been widely used in secondary data analysis. However, there was no method able to separate 22 controls and 40 cancers completely. Our *MRT-test* firstly achieved this by only using three gene markers, which

may indicate that our *MRT-test is independent of data sets*. It not only works for the *Stroma* data, which is a relatively new array data published in 2008, but also old array data like the *Colon* data.

Although the *Breast* data set that consisting of 46 '*metastasis*' and 51 '*without metastasis*' samples is well-known for its poor diagnostic accuracy from previous studies[2], all 97 samples are well separated spatially by our method: there were no mixed samples and the resulting two distinct groups are clearly separated by an easily identified boundary (NW plot in Figure 2). This indicates that this previously thought of a '*hard*' high-dimensional data set is actually *linearly separable* using our biomarkers detected by the *MRT-test*. The similar *linearly separable* scenario is demonstrated for the *Wang-breast* data with 216 breast cancer samples by the SE plot in Figure 2.

Furthermore, the SW plot in Figure 2 shows the complete separation of 25 classic and 9 desmoplastic samples in the *Medulloblastoma* data by three *MRT-test* identified gene markers, where two obvious clusters can be easily identified. To our best knowledge, there were no clustering or biomarker discovery methods able to group the 25 classic and 9 desmoplastic samples into two well-formed independent clusters. This again illustrates that the *MRT-test* is able to accurately detect novel biomarkers which are difficult to be identified by existing methods.

In fact, previous studies showed that a distinct *desmoplastic* class could not be discovered by using traditional clustering methods such as hierarchical clustering or self-organizing maps[11]. In other words, there was no way to group all 9 *desmoplastic* samples into a cluster which is also an "evidence" to support that such data are nonlinear data with the least hope to achieve rival clinical diagnosis. Brunet et al employed a quite complicate nonnegative matrix factorization (NMF) based consensus clustering, which is a high-complexity algorithm, and only grouped seven desmoplastic samples into a cluster, where two desmoplastic samples were still scattered into the group of the *classic* samples[11].

3.2.4. *Biomarker discovery for multiclass array data*

Our *MRT-test* also demonstrates a similar phenotype separation for multiclass array data *SMOKE* and *CNS*. It is noted that relatively few multiclass biomarker discovery algorithms are available for multi-class data due to the complexity of multiclass classification itself and corresponding low detection ratios. However, multiclass array data are quite common due to several subtypes of a tumor or different pathological states of a disease. Since gene expression data usually follow or approximately follow a normal distribution after normalization, it is reasonable to use a one-way ANOVA to identify the most significant genes as potential biomarkers for a multiclass data set. Thus, we compare our *MRT-test* with one-way ANOVA, which can be viewed as an extension of the *t-test* in multiclass scenarios, for multiclass expression data biomarker discovery.

Figure 3 employs the 3 top-ranked genes from ANOVA (one-way ANOVA) and

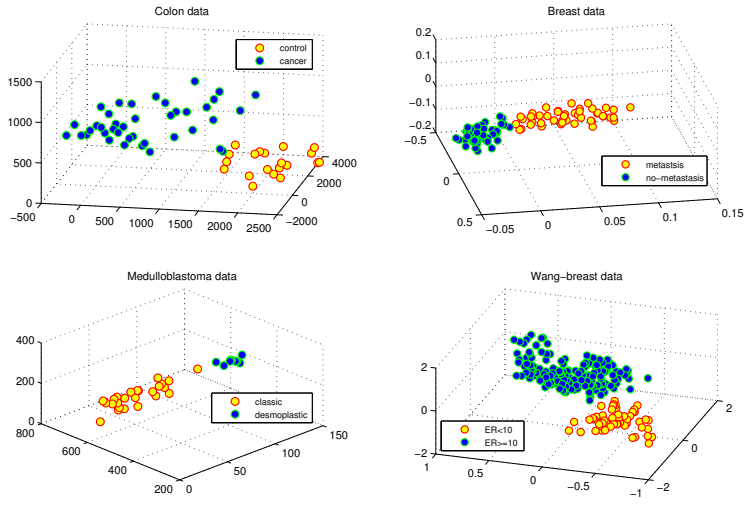

Fig. 2. Separating phenotypes for four benchmark high-dimensional arrays by the *MRT-test*.

the *MRT-test* (τ=4) to separate *SMOKE* data that consists of 18 previous smokers ('*smoked*'), 34 current smokers ('*smoking*'), and 23 non-smokers ('*non-smoking*')[14], as well as the *CNS* data that consists of 10 medulloblastomas ('*md*'), 10 malignant gliomas ('*mg*'), 10 rhabdoids ('*Rhab*'), 4 normal human cere-bells ('*Ncer*'), and 8 supratentorial PNETS ('*PNET*') across 5669 genes[15].

The NW plot in Figure 3 shows there are no obvious separations for samples from different phenotypes for the *SMOKE* data under *ANOVA*. However, the NE plot in Figure 3 demonstrates that three independent clusters: *'smoked'*, '*smoking',* and *'non-smoking',* can be easily identified from our *MRT-test.* It is interesting to see that the *'smoked'* cluster seems to have much closer distance to the '*smoking*' cluster than the *'non-smoking'* cluster, which seems to be consistent to the previous results that former smokers had risk to develop lung cancer as current smokers[14].

Similarly, the SW plot in Figure 3 shows that only four '*Ncer*' samples were separated from the other 38 *CNS* samples using gene markers from the *ANOVA*. However, once again, the SE plot in Figure 3 demonstrates clear phenotypic separations were achieved by using the gene markers: 'APC', DCTN1', 'APBA2' from our *MRT-test* in these multiclass scenarios.

The three selected gene markers are well-known oncogenes related to medulloblastoma, glioma, and neuron diseases, where 'APC' and 'APBA2' mutations were reported to play an important role in medulloblastoma[26]; DCTN1 was reported associated with neuron diseases and glioma respectively [27]. It is also interesting to see that these medulloblastoma samples are spatially much closer to the those malignant

glioma samples than others (See SE plot in Figure 3). Actually, these two diseases have very similar or same symptoms such that medulloblastoma was confused as glioma for years before their different pathologies were found in[28].

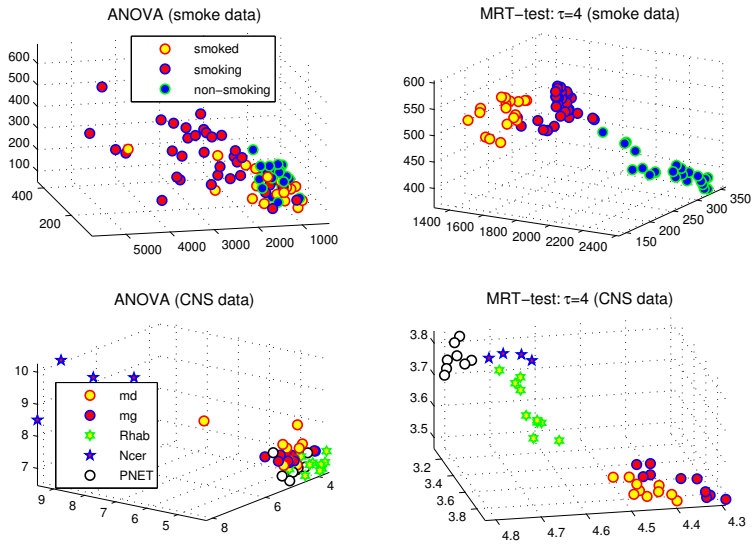

Fig. 3. Separating multiclass *SMOKE* data (18 previous smokers, 34 current smokers and 23 nonsmokers) and *CNS* data (10 '*md*', 10 '*mg*', 10 '*Rhab*', 4 '*Ncer*', and 8 '*PNET*' samples) by using the top 3 ranked genes from *ANOVA* and *MRT-test*.

It is worth pointing out that our phenotypic separation on the *CNS* data is superior to that by Pomeroy et al's approach[15], which employed PCA analysis for the 50 genes ranked by *SNR* (*signal-to-noise* ratios. However, compared with our complete separation of five classes by using only three gene markers, Pomeroy et al's approach has one '*Mg*' out of its group and makes eight '*PET*' samples spread too much such that they can be somewhat mixed with '*Md*' and '*Mg*' samples. In summary, the experimental results across six data sets clearly demonstrate that our *MRT-test* is a powerful biomarker discovery algorithm that is able to demonstrate linear separability in gene expression array data that other methods have failed to do so.

3.3. MRT-test phenotype diagnosis

It is desirable to validate these biomarkers identified by the *MRT-test* from a diagnostic viewpoint under data cross-validation and further explore the *MRT-test*'s discriminative power for a number of genes identified by the *MRT-test*. Thus, we ex-

amine the *MRT-test* phenotype diagnosis by combining it with the state-of-the-art classifiers such as support vector machines (SVM) and linear discriminant analysis (LDA), in addition to comparing the *MRT-test* with different variants of the *t-test* and *one-way ANOVA*. We skip the detail descriptions about SVM and LDA. More detailed information about these two classifiers can be found in[7,29].

3.3.1. *Binary phenotype diagnosis.*

We employ a support vector machine (SVM) with a '*linear*' kernel under standard 5-fold cross-validation (5-fold CV) to evaluate several sets of top-ranked genes from the *MRT-test for* binary class data. It is noted that each gene set identified by the *MRT-test* are de-noised and local feature extracted. Figure 4 shows the SVM's diagnostic performance in terms of accuracies (diagnostic ratios), sensitivities, specificities, and positive prediction rates for six binary data sets using the 3, 10, 50, 100, 200, 500, and 1000 top-ranked genes from the *MRT-test*. Each data set in Figure 4 is represented by its first letter, where 'S' and 'S1' represent the *Stroma* and *Smok*e data respectively. The reason for us to select a relatively large gene set (e.g., a set with 100, 200, or even 500 top-ranked genes) for phenotype diagnosis is that we plan to demonstrate the proposed *MRT-test* is not only a good biomarker discovery algorithm but also an exceptional feature selection algorithm, which will bring rivaling clinical performances for high dimensional omics data. We briefly sketch these diagnostic performance measures for the convenience of description as follows.

The diagnostic accuracy is the ratio of the correctly classified test samples over total test samples. The sensitivity, specificity, and positive predication ratio are defined as the ratios: TP/(TP+FN), TN/(TN+FP), and TP/(FP+TP) respectively. The TP(TN) is the number of positive (negative) targets correctly diagnosed, and FP (FN) is the number of negative (positive) targets incorrectly diagnosed by the classifier (e.g., SVM). A positive (negative) target is a sample with '+1' ('-1') label, which usually represents 'cancer' ('control') or a subtype of tumors respectively.

Figure 4 shows that clinical level prediction ratios with only a few genes are consistently achieved for most data sets. For example, the three top-ranked gene markers for the *Smoke*, *Colon*, *Medulloblastoma*, and *Breast* data achieve 100%, 96.93%, 96.68%, and 99% diagnostic accuracies respectively under the 5-fold CV. Although the three top-ranked gene markers for the *Stroma* data only achieve ˜92% diagnostic accuracy, it is probably because the relatively too close distance between two boundary samples, i.e. support vectors, and corresponding numerical artifacts in the optimal SVM hyper-plane construction cause the classifier not to be able to determine a specific sample's class type. This can be used to explain why the diagnostic accuracies of some data are not perfect 100%, even if their three selected gene markers can separate two different phenotypes spatially. However, the SVM classifier attains 100% accuracy on the *Stroma* data by using the 10 top-ranked genes. The same level performance remains for the 50,100, and 200 top-ranked gene scenarios.

Interestingly, the SVM classifier demonstrates an '*early-arrival*' in phenotype diagnosis for almost all data, i.e., only a few genes are needed to achieve nearly perfect sensitivity and specificity. It suggests that using more genes as biomarkers may not necessarily enhance diagnostic performance, because more less-discriminative genes may be included in training data that leads to lower the classifier 's performance. For example, the SVM attained slightly low 97.78% accuracy (sensitivity: 100%, specificity: 90%) for the 500 and 1000 top-ranked genes. However, such *'early-arrival'* in diagnosis is consistent with the biomarker separation results presented in the previous section.

Although our SVM did not achieve as a 'good' performance on the *Prostate* data under the 3, 10, and 50 top-ranked genes, it still achieved a 97.78% (sensitivity: 98.65%, specificity: 96.52%) accuracy for the 100 top-ranked genes and 100% for the 200 or more genes, which indicates that more biomarkers are needed to distinguish different phenotypes for this dataset. It can be viewed as a special case of the *'early arrival'* mechanism. Although we did not include *Wang-breast* data *in* our experiment for it is not the original data but a data set already going through a previous feature selection, it is worthwhile to report that the SVM classifier can reach 98.60% (sensitivity: 98.77%, specificity:98.00), 99.07% (sensitivity: 98.77%, specificity:100.00%), and 99.53% (sensitivity: 99.38%, specificity: 100%) accuracies under the 3, 10, and 50 top-ranked genes from our *MRT-test.* However, the SVM classifier can only achieve ˜85% accuracy for this data set under the same cross validation without feature selection.

3.3.2. *Compare the* MRT-test *with the other feature selection methods.*

Just as we pointed out that our *MRT-test* is a novel feature selection algorithm besides a biomarker discovery method. It is necessary to compare the *MRT-test* with other similar feature selection methods to further demonstrate its superiority. We select five input-space feature selection methods as the comparison peers, which include *t-test*, Bayesian *t-test* (*'bt-test'*)[34], and permuted *t-test* (*'pt-test'*)[35], ROC curve area ('*roc*')[36], and Mann-Whitney tests (*'u-test'*)[37]. We employ the *MRT-test* and the five peers to select two gene sets: one with with 100 top-ranked genes, another with 200 top-ranked genes for each data. Then, we run the SVM classifier with a linear kernel for all these data sets under 5-fold cross validation to compare their diagnostic accuracies.

Figure 5 shows the SVM classifier 's diagnostic accuracies for the two gene sets selected by six feature selection methods respectively under the 5-fold CV for six binary data sets. It is clear that the SVM classifier demonstrates consistently leading performance for all gene sets from the *MRT-test* than those from the others. Interestingly, the SVM classifier shows relatively low accuracies and high-level oscillations for the gene sets from the five comparison peers, which seems to indicate the unpredictable impacts of these methods on diagnosis.

For example, the SVM classifier achieves 98.46% diagnostic accuracy for the

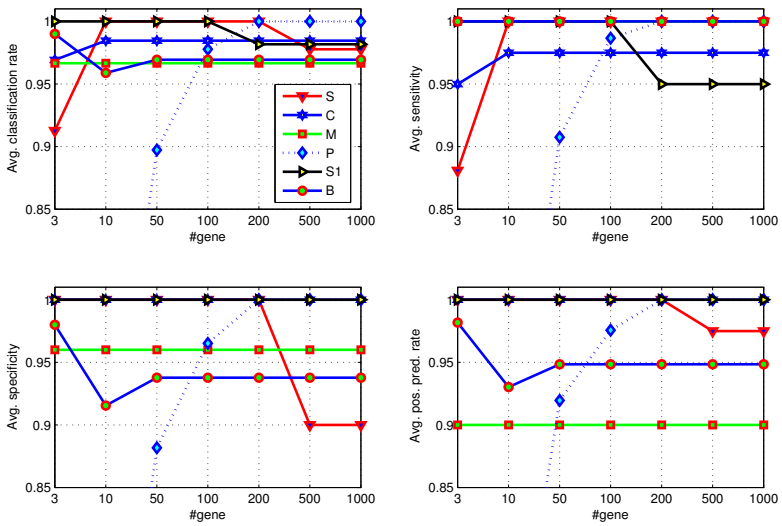

Fig. 4. The SVM diagnostic performance under the 3, 10, 50, 100, 200, 500, and 1000 genes selected by the *MRT-test* in terms of diagnostic accuracies, sensitivities, specificities, and positive prediction rates under the 5-fold CV. each data set is represented by its first letter, where '*S*' and '*S1*' represent the *Stroma* and *Smoke* data respectively

Colon data for the gene set with 100 top-ranked genes from the *MRT-test.* But it only attains 82.82%, 87.56%, 72.56%, 79.23%, and 79.74% for the same-size gene sets from *t-test, bt-test, roc, and pt-test* respectively. Furthermore, the same classifier only attains 85.64% diagnostic accuracy on the 200 top-ranked genes from the *bt-test,* which is lower than the 87.56% accuracy obtained on the 100 top-ranked genes. In fact, relatively low and oscillated accuracies are observed for those genes from the other methods on different data sets. It is also worthwhile to point out that similar scenarios are observed even when more genes are selected (e.g., 500 and 1000 genes).

It is clear that the *MRT-test* is corresponding to the rival clinical diagnosis on these binary data, where Bayesian t-test (*'bt-test'*) and *t-test* are corresponding second-level diagnosis, the *roc* and *u-test* are corresponding to the third-level diagnosis, and the permutation t-test (*pt-test)* seems to have the worst performance from a diagnostic viewpoint. The results suggest that the *MRT-test*'s capability in selecting unique and discriminative local features enhances robust and accurate binary phenotype diagnosis.

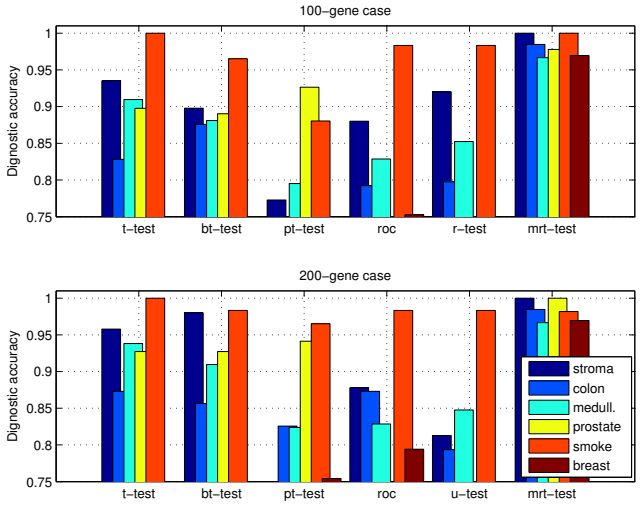

Fig. 5. Comparisons of the *MRT-test* and other five feature selection methods on six binary data by using 100 and 200 top-ranked genes through a SVM classifier under 5-fold CV

3.3.3. *Multiclass phenotype diagnosis*

Multiclass phenotype diagnosis has been a challenging problem in translational bioinformatics for its usually low accuracies due to the complexity of multiclass data classification itself. We employ the two multiclass data sets: S*MOKE (three-class)* and *CNS (five-class)* data in our experiment to demonstrate the *MRT-test*'s superiority in multiclass phenotype diagnosis.

We use '*one-against-one*' SVM to handle multiclass phenotype diagnosis because its proven advantage over '*one-against-all*' and '*directed acyclic*' SVM[30], in addition to taking advantage of linear discriminant analysis (LDA)'s build-in multiclass handle mechanism. Moreover, we combine the *MRT-test* with the '*one-against-one*' SVM and LDA to get corresponding *MRT-SVM* and *MRT-LDA* algorithms, where input data consist of the top-ranked genes from t*he MRT-test*. Similarly, we integrate one-way ANOVA with the two classifiers to get corresponding *ANOVA-SVM* and *ANOVA-LDA* algorithms, where input data consist of top-ranked genes from the one-way ANOVA. They act as the comparison algorithms for our *MRT-SVM* and *MRT-LDA* in multiclass phenotype diagnosis. To be consistent with the previous section, we still employ the same *'linear'* kernel for the '*one-against-one*' SVM and conduct 5-fold cross validation for each data.

Figure 6 compares the performance of the *MRT-SVM* and *MRT-LDA* with the *ANOVA-SVM* and *ANOVA-LDA* classifiers in terms of diagnostic accuracy, sensitivity, and specificity for a sequence of 3, 10, 50, 100, 200, 500, 1000 and 2000

top-ranked genes from the *ANOVA* and *MRT-test* respectively for the *SMOKE* and *CNS* data. It is interesting to see that the '*early-arrival*' phenomenon can still be obervsed for the *MRT-SVM* and *MRT-LDA* classifiers in addition to their rival clinical diagnosis, For example, the *MRT-LDA* achieved 93.06% accuracy with the 3 top-ranked genes from the *MRT-test* on the *CNS* data, whereas *ANOVA-LDA* only manages a low 64.72% accuracy with the 3 top-ranked genes from the *ANOVA*.

Similarly, the *MRT-SVM* achieved 98.75% accuracy on the *SMOKE* data with only 10 *MRT-test* selected genes, which seems to be consistent with the previous biomarker discovery results, whereas the *ANOVA*-SVM only achieved a low 63.83% accuracy with the 10 *ANOVA* selected genes. The results show that the *MRT-test* is also more effective in choosing distinguishing biomarkers for multiclass phenotype diagnosis. Furthermore, we observe that the *MRT-test* well demonstrates the advantages of effective feature-selection in multiclass phenotype classification, since the original '*one-against-one*' SVM and LDA algorithms without any feature selection can only achieve 63.83% and 54.67% on the *SMOKE* data, and 90.56% and 88.06% accuracies on the CNS data respectively.

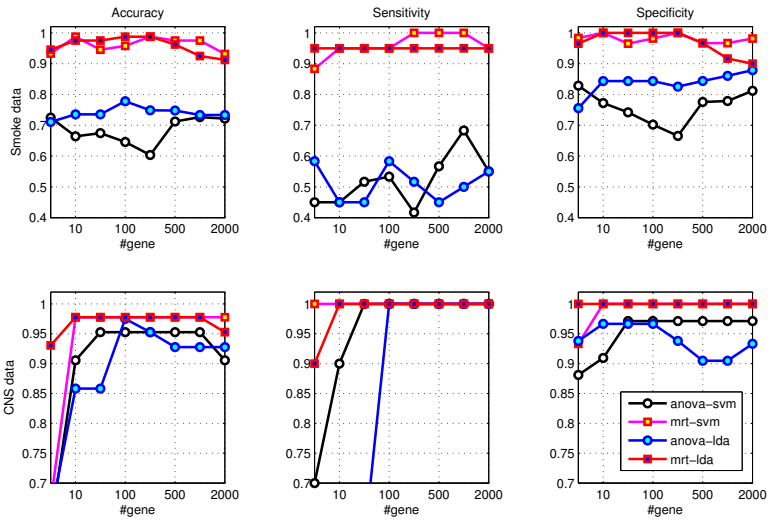

Fig. 6. The diagnostic performance of classifiers on a sequence of top-ranked genes from the *MRT-test* and *ANOVA* from the *SMOKE* and *CNS* data

3.3.4. *Comparison with other state-of-the-art algorithms*

We further compared our *MRT-test* based methods: *MRT-SVM* and *MRT-LDA* with six other state-of-the-art classifiers for total eight gene expression array data,

where the *MRT-SVM* employs the standard SVM and '*one-against-one*' SVM for each binary and multiclass data respectively.

The six comparison classifiers consist of two PCA-induced classifiers: PCA-induced linear discriminant analysis (PCA-LDA) and PCA-induced support vector machines (PCA-SVM)[6], two partial least square (PLS) based regression methods: PLS-based linear logistic regression (*PLS-LLD*)[31], a PLS-based ridge penalized logistic regression (*RPLS*)[32], and an ensemble learning method: a decision tree with bootstrap aggregation (*DS-TREE*)[33]. It is worthwhile to point out that, like PCA, PLS is also a subspace feature selection algorithms that seek the meaningful feature combinations in a low dimensional space. Due to different '*early-arrivals*' of the *MRT-SVM* and *MRT-LDA* for different data sets, we compared the performance for the top-ten ranked genes for all the seven datasets and the top-100 ranked genes for the *Prostate* data.

Figure 7 compares the performance of the *MRT-test* based methods with the other methods in terms of diagnostic accuracies under the 5-fold CV. It indicated that our *MRT-test* based methods outperformed the others in terms of accuracies and stability. For example, *MRT-SVM* and *MRT-LDA* achieved both 100% accuracies for the binary smoke data, and 97.50% and 98.75% accuracies for the three-class *SMOKE* data respectively, whereas the best performance from the other peers are only 91.67% and 82.70%. It seems that the *MRT-SVM* achieved the best diagnostic performance on all data, and the *MRT-LDA* algorithm achieved second-best performance on all seven datasets except *Prostate* data, on which it appeared to encounter down-fitting. Moreover, the SVM, PCA-SVM and PCA-LDA classifiers appeared to achieve same level diagnostics that may suggest PCA may not contribute to enhancement of diagnostics statistically for SVM. It also seems that the *PLS-LLD* algorithm outperformances all the other five classifiers except the *MRT-SVM* and *MRT-LDA* for the eight data sets, in addition that the decision tree method has the worst performance.

Unlike the other methods demonstrating large oscillations in performance, our *MRT-test* based methods again demonstrated consistency in attaining high-level phenotype discrimination in *'hetergeneous*' binary and multi-class expression data. This suggests that our proposed techniques may be robust enough for clinical tests.

3.4. *Network marker synthesis (NMS): seeking meaningful network markers from the* MRT-test

It is natural to extend the biomarkers obtained from the *MRT-test* to corresponding network markers to enhance its reproducibility by capturing more meaningful gene markers, because some statistically expressed genes in array data may not *be the* "real" gene markers in clinical tests and vice versa. We propose a network marker synthesis (NMS), a bottom-up network marker construction algorithm, by growing from the *MRT-test* identified *seed* gene markers into network-based biomarkers through protein-protein networks (PPI).

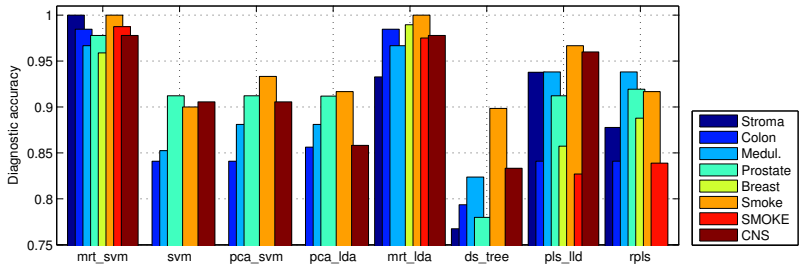

Fig. 7. Comparing *MRT-SVM/LDA* with the other state-of-the-arts on eight data sets.

Compared with the existing top-down subnetwork marker construction methods (e.g., jActiveModule, PNA)[34], our NMS method has a more targeted search and less dependence on the global network because it starts from meaningful gene markers identified by the *MRT-test.* Moreover, *NMS* outputs a small scale network consisting of several subnetwork markers connected by bridge genes instead of orthogonal subnetworks. Thus, it can identify the gene regulators for subnetwork markers, which can be a key issue to understand the network dynamics related to a complex disease. The basic idea of our NMS can be sketched as following steps.

(1) Starting from few gene markers from our *MRT-test* to induce corresponding subnetworks with differentially expressed genes by checking a global PPI network, which is the human PPI network from BioGrid 3.1.77, which has 39,331 protein-protein interactions between 10,271 proteins (genes)[38].
(2) Identifying top bridge genes linking the subnetworks and top connection genes that are elements in each subnetwork with direct-interaction with the selected bridge genes by some thresholds.
(3) Collecting semi-gene cliques containing the selected bridge genes and connection genes by searching the global PPI network.
(4) Union all subnetworks by removing overlapped genes and interactions.

We define the five key concepts in our algorithm as follows before we present our network marker synthesis (NMS) method in algorithm 2.

(1). *Global PPI networks* and *total gene lists*. The global PPI network $G_{ppi} = (V_{ppi}, E_{ppi})$ is a graph where V_{ppi} denotes a set of proteins/genes (nodes) and E_{ppi} represent a set of edges (interactions) between the proteins in V_{ppi}. G_{ppi} in our context refers to the human PPI network from BioGrid 3.1.77[38]. The total gene list T ($T \subseteq V_{ppi}$) includes all genes in the input expression profile X.

(2). *Gene cliques* and *semi-gene cliques*. A gene clique C is a fully connected subnetwork $C = (V, E)$, where all the gene $g_i \in V$, $g_j \in V$, $(g_i, g_j) \in E$, i.e., each node in a gene clique C has a maximal degree $|V| - 1$. A semi-gene clique is a subnetwork where each gene has a degree at least $|V| - 2$. Clearly, a gene clique is

always a semi-gene clique but not vice versa. It is noted that we require the smallest clique at least has three nodes, i.e., $|V| \geq 3$, in our algorithm.

(3). *Bridging genes.* A bridging/bridge gene $b \in T$ is a gene which connects two or more subnetworks by interacting with at least one gene in each subnetwork. For example, the bridging gene b connecting two subnetworks $S_1 = (V_1, E_1)$ and $S_2 = (V_2, E_2)$ means $\exists\ v_1 \in V_1$, $v_2 \in V_2$ such that $(b, v_1) \notin E_1$ and $(b, v_2) \notin E_2$. It is noted a bridging gene can be a gene that does not belong to any subnetworks according to our definition. A bridging gene usually has a large degrees (e.g., 355), i.e., it interacts with quite a lot genes in the whole network. As such, the few top bridging genes are actually regulators of the network marker inferred by our NMS method.

(4). *Connection genes.* Given a bridging gene b and a subnetwork $S_1 = (V_1, E_1)$, a connection gene c is the gene in S_1 (or V_1) that interacts with a bridging gene b directly, i.e., $(b, c) \in E_1$. As an interface in each subnetwork that interacts with the bridging gene, a connection gene plays an essential role along the bridging genes to create other subnetworks that interact with the subnetwork markers induced by the original gene markers.

(5). *Bridge hubs.* Given a bridge gene b, a bridge hub H_b is the union of the genes in all the semi-gene cliques containing b, i.e., $H_b = \cup_e C_{(b,e)}$, where $C_{(b,e)}$ is a semi-gene clique created with a bridge gene b and connection gene e.

Note that the top bridging genes are the regulators of the identified subnetwork markers, while the connection genes act as a local interface to process signals from a regulator. A bridge hub collects almost all qualified genes in the total gene list T interacting with the regulator and related connection genes, and works as an inferred subnetwork marker.

To grow gene markers into subnetwork markers, we use the *MRT-test* identified gene markers $G_M = \{g_1, g_2, \ldots g_k\}$ to induce their corresponding subnetworks $S_1, S_2, \ldots, S_k$ by collecting all the genes in T interacting with the gene markers in G_M by querying the global network G_{ppi}, i.e., $S_i = \cup v_j, (v_j, g_i) \in E_{ppi}$, $v_j \in T$, $g_i \in G_M$. Genes with *p-values* (from *MRT-test)* less than a cutoff α (e.g., 0.001) are kept for each subnetwork.

The subnetworks are further sorted by their cardinalities–the '*orphan*' subnetworks with only one entry are removed. Next, the bridging genes for the top K subnetworks with the largest cardinalities ($K \leq k$) are identified and further sorted according to their PPI numbers. We choose a set of top bridging genes $b_1, b_2, \ldots, b_M$ with the largest PPI numbers as '*real*' bridging genes, and identify their corresponding connection genes $c_1^{(i)}, c_2^{(i)}, \ldots c_l^{(i)}$ with respect to each subnetwork S_i,$i = 1, 2, \ldots K$. We also compute the bridge hub H_b for each bridge gene b_i collecting all semi-gene cliques via searching G_{ppi}. Finally, we collect all subnetwork markers by taking the set-union for all bridge hubs and the K subnetworks. Algorithm 2 below provides the detailed procedure for our NMS algorithm.

Algorithm 2 network marker synthesis (NMS)

Input:

1. Identified k gene markers: $G_M = \{g_1, g_2, \ldots, g_k\}$, from the *MRT-test*
2. Global network $G_{ppi} = (V_{ppi}, E_{ppi})$, gene expression data X, and p-value cutoff: α

Output: Identified network markers

(1) *for* each gene marker g_i, $i = 1, 2 \cdots k$
   (a) Create subnetwork S_i for g_i by collecting its PPIs, i.e., $(v_j,\ g_i) \in E_{ppi}$, $v_j \in T$, $g_i \in G_M$, where each subnetwork S_i is a subgraph induced by genes g_i, $i = 1, 2 \cdots k$.
   (b) Remove the gene set S_i^α where each gene has *p-value* $\geq \alpha : S_i = S_i - S_i^\alpha$.

(2) Sort all subnetworks by its cardinality such that S_i be the subnetwork with the i-th largest cardinality and drop the subnetworks with cardinality 1 to get K major subnetworks, $K \leq k$. $S = \{S_1, S_2, \ldots, S_K\}$, $K \leq k$.

(3) Identify the bridging gene list $B = \{b_1, b_2, \ldots, b_n\}$ for the top N subnetworks in S $(N < K \leq k)$ by searching the global network $G_{ppi} = (V_{ppi}, E_{ppi})$. The bridging gene list is sorted according to each gene's PPI number, which is the degree of the gene node in the global network, such that has the *i-th* larget PPI numbers among all genes in the bridging gene list $(i = 1, 2 \cdots n)$. Update the list by the top M $(M < n)$ bridging genes $B = \{b_1, b_2, \cdots b_M\}$.

(4) *for* each bridging gene b_j, $j = 1, 2, \ldots, M$.
   (a) Identify its connection gene set for S_i, $i = 1, 2, \ldots, K$.
   (b) Compute its bridge hub by collecting all semi-gene cliques.

5) Union all bridge hubs and major subnetworks $(\cup_j H_{b_j}) \cup_i S_{i_j})$

To avoid the runtime complexities caused by the possible need to traverse the entire global PPI to seek semi-gene cliques given that there are up to $K(K-1)/2$ bridge hubs to be calculated, we suggest keeping the number of major subnetworks $K \leq 5$ to avoid overheads, which can be achieved by using few gene markers or dropping the induced subnetworks with few genes. Furthermore, the *p-value* cutoff in *NMS* can be adapted for different needs. A large *p-value* cutoff or even no *p-value* restriction $(\alpha = 1.0)$ on subnetwork genes are acceptable, if we aim to decrease or remove the potential correlation between expression data and the final subnetwork markers. On the other hand, a small *p-value* cutoff (e.g. a=0.001) can also be set to seek out differentially expressed genes in network marker identification.

In our implementation, we compute the bridging gene list from the top 2 $(N = 2)$ induced subnetworks and select the top 3 bridging genes $(M = 3)$ to avoid high runtime complexities and obtain more '*general*' bridging genes with large PPIs. Similarly, we select at most two connection genes for each bridging gene in a subnetwork by dropping those with fewer PPIs. For computing the bridge hubs, we

greedily collect the semi-gene cliques with the richest interactions for each bridging gene.

3.4.1. *Network markers inference for* Stroma *data.*

Figure 8 shows the network markers inferred from our NMS algorithm, where TP53, HDAC1, and BRCA1 (within rectangles) are the bridging genes identified, all of which are famous tumor genes related to general cancer (e.g., TP53) or widely known breast cancer marker (e.g., BRCA1)[39,40]. As highly connected genes interacting with other genes, these bridging genes have 322, 297, and 172 unique interactions in the global PPI network respectively.

The bridging genes act as network regulators that enable and facilitate communication between different subnetworks. It is interesting to note that induced subnetworks (from our three seed gene markers) are orthogonal, and they and the other genes located in subnetworks (induced by semi-gene cliques) communicate with each other via the bridging genes.

In fact, all the three bridge genes are well-known breast cancer associated genes or oncogenes: HDAC1 has been reported as an important indicator in malignant human breast prognosis in previous RT-PCR analysis[39], BRCA1 is a key gene associated with breast cancer and its mutation usually increases breast cancer risk, and TP53 is a well-known oncogene that mutates in most types of human cancers[2,39,40]. Furthermore, UBE2I & RB1, UBE2I & CDK9, and BRMS1 & RB1 are identified as the connection genes between the bridging genes BRCA1, TP53, and HDAC1, and the two subnetworks induced by GOLGB1 and CCNT2 respectively (e.g., UBE2I and RB1 are two different connection genes for BRCA1 with respect to the two induced subnetworks). CTNNB1 is identified as the connection gene between HDAC1 and the subnetwork induced by CBY1.

There are seven bridge hubs generated by collecting the semi-gene cliques for the bridging and connection genes. For example, the bridge hub induced by BRCA1 and the connection gene UBE2I has five other genes: AR, TP53, JUN, PIAS1, and SUMO1. Some genes in these bridge hubs demonstrate high relevance with respect to breast cancer. For instance, Smith et al have reported that JUN over-expressed in some breast cancer cells may result in the production of tumorigenic, invasive, and hormone-resistant phenotypes[39]. Park et al also reported that SUMO1 negatively regulated BRCA1-mediated transcription[40,41]. Gonzalez et al reported that AR-positive tumors had a significant longer overall survival than those with AR-negative breast carcinomas in their tumor analysis of androgen receptors (AR)[42,43].

Finally, we employed the network markers consisting of 32 genes to conduct classification under a 5-fold CV using a SVM classifier with a linear kernel and achieved diagnostic accuracy 97.78% with 100% sensitivity and 90% specificity. It is noted that the best diagnostic result among other classifiers' performance under the same cross validation condition is achieved by the *PLS-LLD* algorithm with diagnostic accuracy 93.78% by using the whole data sets, which still fall behind

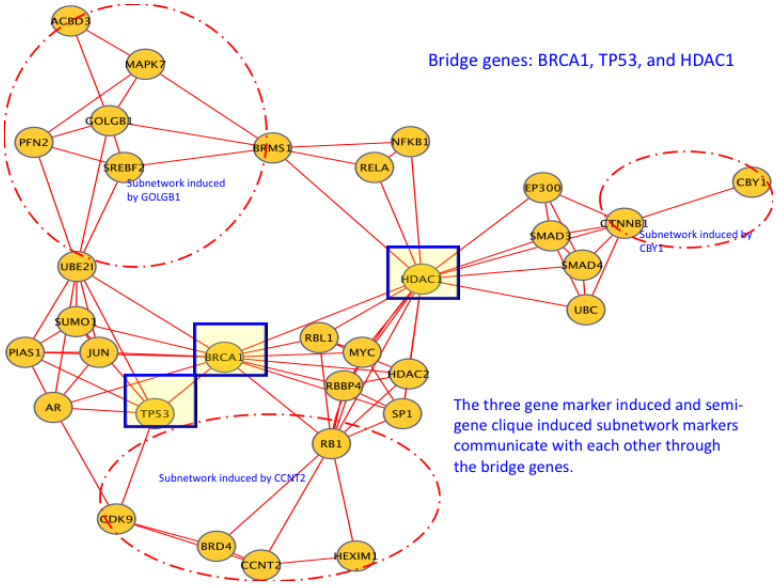

Fig. 8. Network markers identified for *Stroma* data, where the 'BRCA1', 'TP53', and 'HDAC1' are bridging genes.

our network marker methods' result. The superiority demonstrated by our network markers is due to the good gene markers identified by the *MRT-test* and more targeted search based on the gene markers in the network marker synthesis (NMS).

Similarly, we identified the network markers for the *CNS* data[15] by using the top 10 gene markers from the *MRT-test*. Among them, four gene markers APC, DCTN1, APBA2, and JMJD1C led to non-trivial subnetworks with 46, 9, 6, and 3 genes respectively, after collecting interactions among the total 5569 genes in the input data using the *p-value* cutoff a=*0.001*. Again, APC and APBA2 mutations have been reported to play an important role in medulloblastoma[26]; DCTN1 and JMJD1C were reported associated with neuron diseases and glioma respectively[28]. The details of the bridging gene information of this network marker can be found in the supplemental materials.

4. Discussion

In this work, we have presented a novel feature selection and biomarker discovery algorithm *MRT-test* and investigate its application in effective phenotype diagnosis. The *MRT-test* is designed to be able to detect subtle data behaviors so that phenotype samples with similar global features but different local features can be effectively detected, which is essential for the rival-clinical diagnosis in translational bioinformatics. We demonstrated that our *MRT-test* can separate different phenotypes from high-dimensional omics profiles, including those challenging gene

expression datasets that were previously thought to be non-separable based on other existing methods. Our proposed *MRT-test* not only provides an effective biomarker discovery approach, but also demonstrated that high-dimensional gene/protein expression data's linear separability, which provides a strong support for employing effective biomarker patterns in complex disease diagnosis in translational bioinformatics.

Although our *MRT-test* can achieve almost perfect phenotype separation by using only three top selected gene markers for almost all data in our experiment, the three top-ranked gene markers from the *MRT-test* for the *Prostate* data, an old gene expression array data published in 2002[12], can not achieve a perfect phenotype separation as the other data when we only use the first-PC in *MRT-test* to conduct the detail coefficient matrix reconstruction, which seems to be consistent to its late arrival performance in classification. However, such case may suggest us that more PCs should be used in the detail coefficient matrix reconstruction process for the *MRT-test* for some omics data.

By integrating our *MRT-test* with the state-of-the-art classifiers, we showed that we can achieve consistent clinical-level diagnosis accuracy across a series of omics data in disease phenotype discrimination by comparing our methods with the state-of-the-arts. It is noted that the rivaling-clinical diagnosis performance is due to the novelty of our techniques proposed in our unique *MRT-test*. Our results are obtained from different omics data under rigorous cross validation, which prevents any possibility of overfitting because overfitting may only produce deceptive diagnostic performance for one or two data but it has no way to generalize the similar performance to the other data, not to mention a series of heterogeneous gene/protein expression array data.

We also proposed a bottom-up network marker identification (NMS) algorithm by starting from the *MRT-test* identified biomarkers to identify corresponding network markers. Unlike the existing top-down subnetwork marker identification algorithms, our NMS algorithm has less dependence on the global network due to its more targeted search, and enables to identify regulatory genes for its special network marker building technique, which can lead to useful insights about the complex diseases. Alternatively, we have to point out that our NMS is an algorithm that relies on the *MRT-test,* which is theoretically not as rigorous or independent as the existing subnetwork marker identification algorithms such as jActiveModule, COSINE, and PNA[3,4,5].

Although our methods achieved exceptional diagnostic performance and biomarker discovery for benchmark omics data, we are collaborating pathologists to apply our algorithms to different individual data sets because the current results are from secondary data analysis for public data, in addition to further extending our NMS method to proteomics data. Although it is still not theoretically clear how multi-resolution corresponds to underling biological or phenotype behaviors, our work suggests the multi-resolution analysis approach is an effective way to separate the true signals from red herrings, which obviously contributes to high performance

diagnosis in translational bioinformatics. As for future work, we are interested in integrating gene expression, protein expression, microRNA expression, and RNA-Seq data[44] by using TCGA data[45] to infer consensus network markers by extending our *MRT-test* and network marker synthesis algorithms, in addition to investigating how the ordering of samples in an omics data set will our *MRT-test* results.

Acknowledgments

Authors sincerely thank all three anonymous reviewers for their valuable suggestions and critiques, which contribute to improving this paper. The first author also wants to thank faculty research fellowship support from Fordham University in 2013.

References

1. Pepper, SD et al, The utility of MAS5 expression summary and detection call algorithms. *BMC Bioinformatics*, **8**:273., 2007
2. Chuang, H et al., Network-based classification of breast cancer metastasis. *Mol. Syst. Biol.* **3**: 140, 2007
3. Ideker, T., et al., Discovering regulatory and signalling circuits in molecu-lar interaction networks, *Bioinformatics,* **18** Suppl 1, S233-240, 2002
4. Kim, Y et al., Principal network analysis: Identification of subnetworks representing major dynamics using gene expression data, *Bioinfomratics*, **27(**3):391-8*,* 2011
5. Ma, H et al, COSINE: COndition-SpecIfic sub-NEtwork identification using a global optimization method, *Bioinformatics*, doi: 10.1093, 2011
6. Han, X, Nonnegative Principal component Analysis for Cancer Molecular Pattern Discovery, *ACM Transaction of Computational Biology and Bioinformatics* **7**(3), 537-549, 2010
7. Han, H, Li, X Multi-resolution independent component analysis for high-performance tumor classification and biomarker discovery, B*MC Bioinformatics*, **12(**S1):S7, 2011
8. Mallat, S, A wavelet tour of signal processing*, Acad. Press.* CA, 1999
9. Boersma BJ et al., A stromal gene signature associated with inflammatory breast cancer. *Int J Cancer*, **122**(6):1324-32, 2008
10. Alon, U et al., Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays, *Proc Natl Acad Sci U S A*. 96(12): 6745–6750, 1999
11. Brunet JP et al., Metagenes and molecular pattern discovery using matrix factorization, *Proc Natl Acad Sci* , **101** (12), 4164-4169, 2004
12. Singh, D et al., Gene expression correlates of clinical prostate cancer behavior, *Cancer Cell*, **1**(2),203-209, 2002
13. van 't Veer, L et al., Gene Expression Profiling Predicts Clinical Outcome of Breast Cancer, *Nature*, **415**, 530-536, 2002
14. Spira A et al., Effects of cigarette smoke on the human airway epithelial cell transcriptome. *Proc Natl Acad Sci USA,* **101**:10143-8, 2004
15. Pomeroy SL et al., Prediction of central nervous system embryonal tumour outcome basedon gene expression, *Nature,* **415**(24):436-442, 2002
16. Wang Y et al Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet* **365:** 671–679, 2005
17. Gu, Y., et al., Systematic Interpretation of Comutated Genes in Large-Scale Cancer Mutation Profiles, *Molecular Cancer Therapeutics*, **9**:2155 2010

18. Fan J et al., Acyl-coenzyme A binding domain containing 3 (ACBD3; PAP7; GCP60): an emerging signaling molecule, *Prog Lipid Res* **49**(3): 218–234, 2010
19. Johnston CN et al., Parvin-beta inhibits breast cancer tumorigenicity and pro-motes CDK9-mediated peroxisome proliferator-activated receptor gamma 1 phosphorylation, *Mol Cell Biol.* **28**(2):687-704 , 2008
20. Takemaru, K et al. Chibby, a nuclear β-catenin-associated antagonist of the Wnt/Wingless pathway, *Nature* **422**, 905-909, 2003
21. Easton DF, et al., Genome-wide association study identifies novel breast cancer, *Nature,* **447**(7148):1087-109, 2009
22. Yan, LX et al., Knockdown of miR-21 in human breast cancer cell lines inhibits proliferation, in vitro migration and in vivo tumor growth, *Breast Cancer,* **13***:R2,* 2011
23. Expression Atlas, http://www.ebi.ac.uk/gxa/, 2013
24. Lose F, et al., BCoR-L1 variation and breast cancer, *Breast Cancer Res.* **9**(4):R54, 2007
25. Ronnberg, JA et al., Methylation profiling with a panel of cancer related genes: association with estrogen receptor, TP53 mutation status and expression subtypes in sporadic breast cancer, *Mol Oncol.* **5**(1):61-76, 2011
26. Parsons DW et al, The genetic landscape of the childhood cancer medulloblastoma, *Science*, **331**;435-9, 2011
27. Huang H et al , APC mutations in sporadic medulloblastomas. *Am J Pathol.,* **156**(2):433-7, 2000
28. Mendez et al, Knock down of HIF-1 a in glioma cells reduces migration in vitro and invasion in vivo and impairs their ability to form tumor spheres, *Mol Cancer,* 9: 133, 2011
29. Sampson et al, A Comparison of Methods for Classifying Clinical Samples Based on Proteomics Data: A Case Study for Statistical and Machine Learning Approaches, *PLoS ONE,* 2011
30. Hus, C, Lin, C, A Comparison of Methods for Multi-class Support Vector Machines, *IEEE Transactions on Neural Networks*, **13**(2): 415-425, 2002
31. Nguyen, D, Rocke, D, Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics* **18**:39–50 33, 2002
32. Fort, G, and Lambert-Lacroix S, Classification using partial least squares with penalized logistic regression, *Bioinformatics* **21**(7):1104-1111, 2005
33. Che D et al, Decision tree and ensemble learning algorithms with their applications in bioinformatics, *Adv Exp Med Biol.* 2011;696:191-9, 2011
34. Gï¿œnen M, The Bayesian t-test and beyond, *Methods Mol Biol.* **620**:179-99, 2010
35. Collingridge, DS, A Primer on Quantitized Data Analysis and Permutation Testing, *Journal of Mixed Methods Research*, **7**(1), 79-95, 2013
36. Wang Z, and Chang Y, Marker selection via maximizing the partial area under the ROC curve of linear risk scores, *Biostatistics*, 2010
37. Fay MP et al, Wilcoxon–Mann–Whitney or t-test? On assumptions for hypothesis tests and multiple interpretations of decision rules. *Statistics Surveys* **4**: 1–39. 2010
38. BioGrid, http://thebiogrid.org/, 2013
39. Zhang Z et al, Quantitation of HDAC1 mRNA expression in invasive carci-noma of the breast, *Breast Cancer Res Treat.* 94(1):11-6, 2005
39. Smith LM et al, Jun overexpression in MCF-7 breast cancer cells produces a tumorigenic, invasive and hormone resistant phenotype. *Oncogene*, **18**, 6063-70, 1999
40. Park MA et al., SUMO1 negatively regulates BRCA1-mediated transcription via modulation of promoter occupancy, *Nucleic Acids Research*, **36**(1) 263-283, 2008
41. Morris, JR et al, The SUMO modification pathway is involved in the BRCA1 response

to genotoxic stress, *Nature* **462**, 886-890 17 2009

42. Gonzalez LO et al, Androgen receptor expresion in breast cancer: Relationship with clinicopathological characteristics of the tumors, prognosis, and expression of metalloproteases and their inhibitors, *BMC Cancer*, **8**:149, 2008
43. Thike, AA et al. Loss of androgen receptor expression predicts early recurrence in triple-negative and basal-like breast cancer, *Modern Pathology*, — doi:10.1038/modpathol.1**45**, 2013
44. Wang et al, RNA-Seq: a revolutionary tool for transcriptomics, *Nat Rev Genet.* **10**(1): 57–63, 2009
45. Braun R, Finney R, Yan C, Chen Q-R, Hu Y, et al, Discovery Analysis of TCGA Data Reveals Association between Germline Genotype and Survival in Ovarian Cancer Patients. *PLoS ONE* **8**(3): e55037, 2013

Henry Han is currently an Associate Professor in the Department of Computer and Information Science at Fordham University, New York, NY USA. He is the program coordinator of Master degree program in Cyber Security at Fordham University. He is also an affiliated faculty in Quantitative Proteomics Center at Columbia University. He received his Ph.D. in Computational Sciences and M.S. in Computer Science from The University of Iowa in 2001 and 2004 respectively. He published approximately forty papers in leading journals and conferences in bioinformatics, data mining, computational finance, and computer graphics. His current interests include bioinformatics, big data analytics, cyber security, and financial informatics.

Xiao-Li Li is currently a machine learning lab head and senior scientist at the Institute for Infocomm Research, A*STAR Singapore. He also holds an adjunct appointment in the School of Computer Engineering, Nanyang Technological University. His research interests include bioinformatics, data mining, and machine learning. He has been serving as a PC member in the leading data mining conferences KDD, ICDM, SDM, PKDD/ECML, PAKDD, WWW, AAAI, and CIKM as well as co-Editor-in-Chief of International Journal of Knowledge Discovery in Bioinformatics (IJKDB). Some of his reprehensive research include PU learning (900+ citations), biological/social network mining (400+ citations) etc. He received the Best Paper Awards in GIW 2005 and DASFAA 2011 as well as the Best Poster Award in the RECOMB 2008. He also won two Best Performance Awards in international benchmarking competitions, i.e. Dialogue for Reverse Engineering Assessments and Methods (DREAM) 2007 USA, EU activity recognition challenge 2011.

See-Kiong Ng is currently the Programme Director of the Urban Systems Initiative of the Agency of Science, Technology and Research (A*STAR) of Singapore. He is also a Principal Scientist and the Advisor to the Data Analytics Department at A*STAR's Institute for Infocomm Research. See-Kiong obtained his PhD in Computer Science from Carnegie Mellon University. He has a long-standing interest in cross-disciplinary computer science research and has published more than 100 papers in leading peer-reviewed journals and conferences. His primary research is in machine learning and data mining, with applications in text mining, bioinformatics, privacy-preserving data mining, and social network mining.

Zhou Ji Zhou Ji received his Ph.D. degree in Computer Science in 2006, and another Ph.D. in Mechanical Engineering in 2000, both from the University of Memphis. He also had a Master degree in Computer Science from the University of Memphis in 2000, and a Master degree (1992) and a Bachelor degree (1989) in Fluid Mechanics from University of Science and Technology of China. His research interests include bio-inspired computing, machine learning, and bioinformatics. He is currently a Lead Programmer Analyst in the Department of Systems Biology, Columbia University in the City of New York.