Computers in Biology and Medicine 77 (2016) 102–115

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/cbm

Identification of informative genes and pathways using an improved penalized support vector machine with a weighting scheme

Weng Howe Chan [a], Mohd Saberi Mohamad [a,*], Safaai Deris [b], Nazar Zaki [c], Shahreen Kasim [d], Sigeru Omatu [e], Juan Manuel Corchado [f], Hany Al Ashwal [c]

[a] Artificial Intelligence and Bioinformatics Research Group, Faculty of Computing, Universiti Teknologi Malaysia, 81310 Skudai, Johor, Malaysia
[b] Faculty of Creative Technology & Heritage, Universiti Malaysia Kelantan, Locked Bag 01, Bachok, 16300 Kota Bharu, Kelantan, Malaysia
[c] College of Information Technology, United Arab Emirate University, Al Ain 15551, United Arab Emirates
[d] Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, 86400 Batu Pahat, Malaysia
[e] Department of Electronics, Information and Communication Engineering, Osaka Institute of Technology, Osaka 535-8585, Japan
[f] Biomedical Research Institute of Salamanca/BISITE Research Group, University of Salamanca, Salamanca, Spain

ARTICLE INFO

Article history:
Received 21 March 2016
Received in revised form
3 August 2016
Accepted 3 August 2016

Keywords:
Artificial intelligence
Bioinformatics
Informative genes
Pathway-based microarray analysis
Penalized support vector machine
Weighting scheme
Penalty function

ABSTRACT

Incorporation of pathway knowledge into microarray analysis has brought better biological interpretation of the analysis outcome. However, most pathway data are manually curated without specific biological context. Non-informative genes could be included when the pathway data is used for analysis of context specific data like cancer microarray data. Therefore, efficient identification of informative genes is inevitable. Embedded methods like penalized classifiers have been used for microarray analysis due to their embedded gene selection. This paper proposes an improved penalized support vector machine with absolute *t*-test weighting scheme to identify informative genes and pathways. Experiments are done on four microarray data sets. The results are compared with previous methods using 10-fold cross validation in terms of accuracy, sensitivity, specificity and F-score. Our method shows consistent improvement over the previous methods and biological validation has been done to elucidate the relation of the selected genes and pathway with the phenotype under study.

1. Introduction

The advent of microarray technology has enabled researches to measure expression of thousands of genes with multiple samples. This has spurred the development of various sophisticated analytical methods to interpret and integrate the enormous data sets produced by high-throughput technology. Moreover, advancements in omics research has made available various kinds of data such as, pathway and network data [1,2]. Incorporation of pathway knowledge into microarray analysis has gained increasing attention owing to the improved biological interpretation of the analysis outcome. This is known as pathway-based microarray analysis. In contrast to most of the early methods of microarray analysis, which are based on analysis of individual genes, pathway-based methods analyse genes in groups or pathways [3]. Analysis of genes in terms of pathway allows the detection of subtle changes of expression which single-gene based methods are unable to detect [4]. This enables the identification of differentially expressed pathways and genes in the pathways that are related to the phenotype under study instead of only a list of differentially expressed genes. The identified active pathways allow us to gain better understanding and functional insights regarding the biology of the phenotype under study [5]. However, most of the pathway data are curated manually based on public domain databases, domain experts and laboratory study of cultured cells, which are not based on specific biological contexts like lung cancer disease [6]. The use of the pathway data in context specific data like lung cancer microarray data could possess the risk of inclusion of non-informative genes in the analysis, which could affect the analysis outcome. It is known that inclusion of these non-informative genes in classifier construction could lead to poor classification performance [7]. Therefore, methods with effective identification of informative genes in the pathways are needed in order to ensure the efficient utilization of pathway data in aiding the analysis of microarray data.

Generally, there are two types of approaches in pathway-based microarray analysis; the enrichment analysis approach and the machine learning approach [7]. In enrichment analysis, genes are grouped into pathways and scored using statistical tests such as

* Corresponding author.
E-mail address: saberi@utm.my (M.S. Mohamad).

http://dx.doi.org/10.1016/j.compbiomed.2016.08.004
0010-4825/

Kolmogorov-Smirnov test [8], Fischer's exact test [9] and gene randomisation [10,11]. Most of the enrichment analysis methods treat all genes equally in the analysis, which is incorrect because some of the genes may have higher relevancy to the phenotype under study which presumably possess better predictive power [12]. On the other hand, machine learning methods often use gene classification in the analysis such as penalized support vector machine [13], random forest [14], partial least square [15] and logistic regression [16]. Furthermore, machine learning methods can be integrated with the gene selection process in order to select informative genes for each pathway before the construction of classifier because not all genes in the pathway contribute to the development of the disease [17,15,18,19]. In order to improve the classification performance, integration of an efficient gene selection method is important to ensure only informative genes are selected for the analysis.

There are three types of gene selection methods commonly known as filter, wrapper and embedded. Embedded methods have been favoured by the researchers owing to their better complexity compared with wrapper methods and interaction with classifier [20]. These embedded methods include random forest, SVM with recursive feature elimination (SVM-RFE) as well as penalized classifier [21]. Among these embedded methods, penalized classifiers have been widely used in bioinformatics research [22]. One of the penalized classifiers is penalized support vector machine (SVM), which is a combination of SVM classifier with penalty functions for simultaneous gene selection. There are several penalized SVM with different penalty functions such as SVM-SCAD, which was proposed by Zhang et al. [23] that embeds smoothly clipped absolute deviation (SCAD) penalty function [24] and L1-SVM [25], which embeds L1 penalty function in standard SVM for feature selection. In a recent published research, a modified penalized SVM with SCAD penalty function, namely gSVM-SCAD has been proposed by Misman et al. [13] and used for pathway-based microarray analysis. gSVM-SCAD was inspired by SVM-SCAD. In contrast to SVM-SCAD, Misman et al. [13] has introduced specific tuning parameters for each pathway to achieve near optimal gene selection. However, performance of the pathway-based methods is often affected by the biological context-free pathway data. Despite the good performance shown by gSVM-SCAD, the gene selection efficiency can still be improved by integrating information derived from the measurements associated with the genes in the pathway. Previous investigation has shown that the use of non-uniform weights calculated from the measurements associated with the genes can improve the identification of informative genes in the pathway [26]. Efficient identification of informative genes in the pathway is crucial especially in the analysis of cancer gene expression data. Therefore, this paper proposes an improved gSVM-SCAD with the integration of *abs*T weighting scheme [26] to improve the efficiency of the identification of informative genes and pathways. The proposed method is referred to as wgSVM-SCAD. As the analysis is done in terms of pathways, the pathway with the most informative genes is expected to produce the best classifier and is identified as the informative pathway related to the phenotype under study.

2. Methods and materials

2.1. gSVM-SCAD

gSVM-SCAD was proposed by Misman et al. [13] for pathway-based microarray analysis. Group-specific tuning parameter, λ_k was introduced for each pathway in order to provide flexibility to SVM-SCAD for maintaining efficient identification of informative genes in every pathway. Given a gene expression data set with m samples and d genes (x_i, y_i), $i = 1, \ldots, m$, $y_i \in \{-1,1\} y$ represents the tissue samples with two classes where $y_i = -1$ and $y_i = 1$. While $x_i = (x_{i,1}, \ldots, x_{i,d}) \in \mathbb{R}^d$ represents the input vector of expression values of d genes of i-th sample. SVM is a classifier which separates the classes of interest by maximising the margin between them using a kernel function. Generally, in gSVM-SCAD, the input variables are classified into corresponding classes by the margin of

$$min_{\beta,c} \sum \left[1 - y_i f(x_i)\right]_+ + pen_{\lambda_k}(\beta) \tag{1}$$

where $\left[1 - y_i f(x_i)\right]_+$ is the SVM convex hinge loss function, $pen_{\lambda}(\beta)$ refers to the SCAD penalty function with parameter λ_k, which is the group-specific parameter for k pathway, $\beta = (\beta_1, \ldots \beta_i)$ represents the coefficients of the hyper-plane and c is the intercept of the hyper-plane. The penalty function shrinks the small coefficient to zero, thus, gene selection is achieved because SVM only uses non-zero variables. The SCAD penalty in gSVM-SCAD is calculated based on the equation below.

$$pen_{\lambda_k} = \begin{cases} \lambda_k |\beta| & \text{if } |\beta| \le \lambda_k \\ -\dfrac{|\beta|^2 - 2\alpha\lambda_k|\beta| + \lambda_k^2}{2(\alpha - 1)} & \text{if } \lambda_k < |\beta| \le \alpha\lambda_k \\ \dfrac{(\alpha+1)\lambda_k}{2} & \text{if } |\beta| > \alpha\lambda_k \end{cases} \tag{2}$$

where a and λ are the parameters with $a = 3.7$ and $\lambda > 0$ according to Fan and Li [24]. In gSVM-SCAD, grid search was used to search for the near-optimal parameter, λ_k for k pathway from a set of predefined values in the range of 0.001–0.009, 0.01–0.09, and 0.1–1. The penalization is done on an initial linear SVM model based on each λ, generating a list of enhanced models with selected genes based on corresponding λ. These models are then evaluated using Generalized Approximate Cross-Validation (GACV) [27] to obtain the best fit model and the identified genes in that model. The performance of the best fit model is then calculated based on 10-fold cross validation. This process iterates for all the pathways in the pathway data.

2.2. The proposed method (wgSVM-SCAD)

Despite the good performance of gSVM-SCAD, the gene selection efficiency can still be improved by integrating information derived from the measurements associated with the genes in the pathway. Meanwhile, the use of non-uniform weighting schemes in pathway analysis has been reported by Ha et al. [26] to improve the performance of the analysis. Therefore, this paper proposes the integration of *abs*T weighting scheme into gSVM-SCAD to improve the performance of the identification of the informative genes and pathways. The weights are calculated based on the measurements associated to the genes, which are more preferable so that they can efficiently represents the differential expression between genes in the data set.

In this paper, the main purpose of the integration of the non-uniform weights is to emphasize the differential expression of the genes and it allows better detection of these differentially expressed genes during the gene selection process using penalized support vector machine. The *abs*T weights are calculated based on two-sample t-test as shown in the equation below [26].

$$W_j = \frac{|T_j|}{\sum_{j=1}^{d} |T_j|} \tag{3}$$

In this equation, $|T_j|$ represents the absolute value of the two-sample t-test calculated for the j-th gene. The weight of the j-th gene, W_j is calculated by dividing $|T_j|$ with the sum of all $|T|$ for d genes in the pathway. In this paper, as the weights are calculated

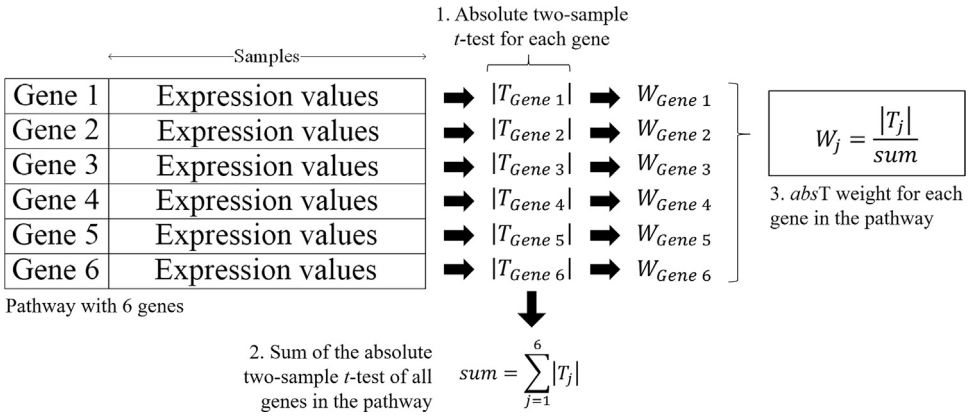

Fig. 1. Example of *abs*T weights calculation for a pathway with 6 genes.

based on the genes in a pathway, the weights indicate how different a gene is expressed between two groups of samples in a pathway. Therefore, the gene that is most differentially expressed in the pathway will obtain the largest weight. The pathway with the most differentially expressed genes will obtain highest classification accuracy. The calculation of the *abs*T weights is as illustrated in Fig. 1.

As shown in Fig. 1, for each gene, the weight is calculated by dividing the absolute two-sample *t*-test of the gene by the same *sum*, which is the sum of the absolute two-sample *t*-test of all genes in the pathway. During the calculation process, the *p*-value for each gene is recorded. After the calculation of the *abs*T weights, an initial filtration is done, which is aimed to filter out genes that do not differentially express between two groups of samples based on the *p*-values generated (larger than 0.5) during the two-sample *t*-test. Then, the corresponding weights are assigned into the remaining genes and used for further analysis. One weight is given to each of the gene in the pathway. The weight is assigned into the gene expression values of the gene across all the samples as shown in the example in Fig. 2. This produces a weighted gene expression for the pathway.

The weighted gene expression data of the pathway is then used to train an initial linear SVM model and undergoes the penalization based on the predefined values of λ. Since the input is the weighted gene expression data, the classification of the proposed method is represented by following equation in contrast to Eq. (1).

$$min_{\beta,c} \sum \left[1-y_i f\left(Wx_i\right)\right]_+ + pen_{\lambda_k}(\beta) \tag{4}$$

where Wx_i represents the weighted input based on the calculated *abs*T weights. Similar to the workflow in gSVM-SCAD, the models

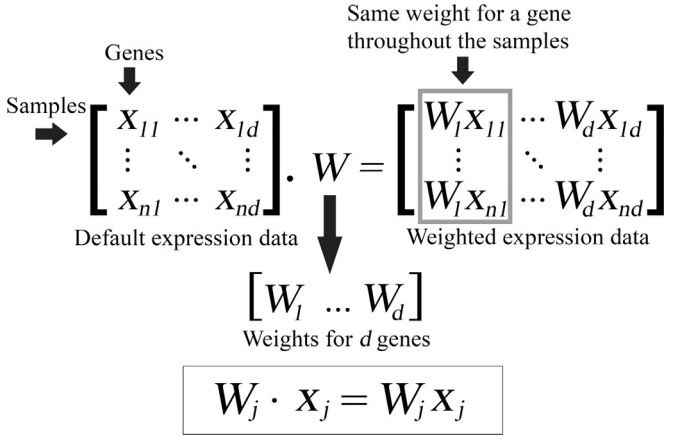

Fig. 2. Assignment of the *abs*T weight in gene expression values.

generated based on different λ in the predefined list are evaluated through GACV and the best fit model is used for performance measurement based on 10-fold CV. The detailed flowchart of the proposed method is shown in Fig. 3.

Compared to gSVM-SCAD by Misman et al. [13], the proposed method integrates non-uniform weights that measure the magnitude of the differential expression of each gene within a pathway based on the concept that the most differentially expressed gene within the pathway have the largest weight in order to improve the efficiency of the identification of informative genes within the pathway. The calculated weights are then integrated into the gene expression data for every gene throughout the samples. Efficient identification of these informative genes improves classification performance, which also enables the identification of informative pathways that are relevant to the phenotype under study. The proposed method is aimed to surmount the limitation caused by the existence of non-informative genes in the pathway data and at the same time provides better biological interpretation regarding the phenotype under study.

2.3. Data sets

This paper uses two types of data, which are gene expression data and pathway data. For gene expression data, four gene expression data sets are used as shown in Table 1. Generally, gene expression data consists of *m* samples and *n* genes. The data is in the form of matrix where the rows represent the genes while the columns represent the samples. These data sets have been used in pathway-based microarray analysis and can be downloaded at Gene Set Enrichment Analysis (GSEA) website (http://www.broadinstitute.org/gsea/datasets.jsp). According to Table 1, there are two gene expression data sets for lung cancers, Lung Michigan and Lung Boston, which was published by both Beer et al. [28] and Bhattacharjee et al. [29] respectively. While for the gender data set, it consists of transcriptional profiles from male and female lymphoblastoid cell lines. This unpublished data has been used in previous studies in pathway-based microarray analysis [11,13,14]. Lastly, the p53 data set used in this research consists of mutational status of p53 gene in the expression patterns from the NCI-60 cancer cell lines and has been used in several studies in pathway-based analysis [11,30,14].

Meanwhile, for pathway data a total of 480 pathways or gene sets are used similar to the previous works [13,14], which consist of 168 pathways from Kyoto Encyclopedia of Genes and Genomes (KEGG) [31] where majority of the pathways were related to metabolism, degradation, biosynthesis and signal processing, and 312 Biocarta pathways [32] where the pathways are mostly related to metabolism and signal transduction. The pathway data are

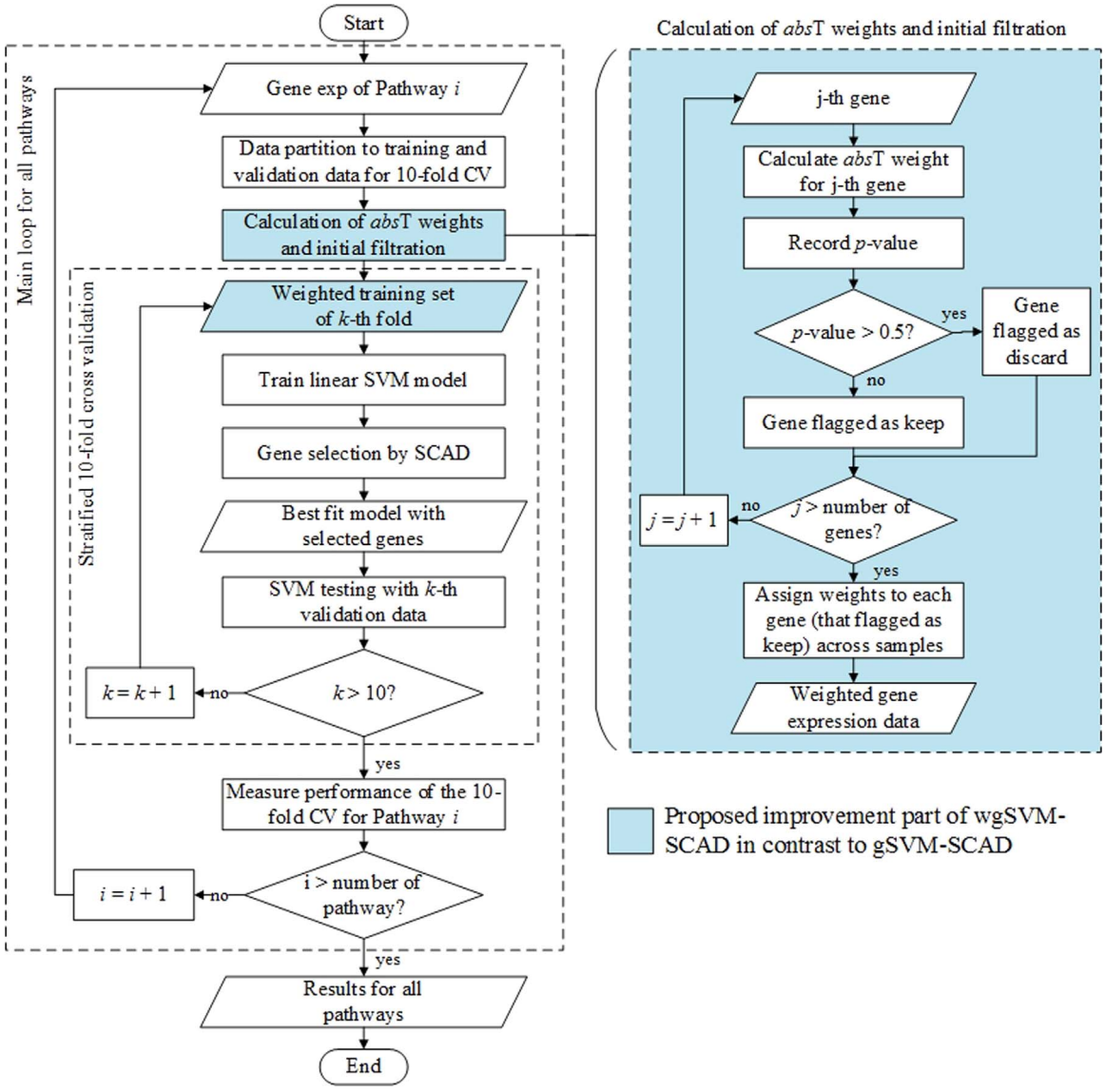

Fig. 3. Flowchart of the proposed wgSVM-SCAD where the blue shaded area represents the improvement over gSVM-SCAD.

Table 1
Summary of gene expression data sets used.

Name	Samples	Genes	Class	Reference
Lung Michigan	86	7129	2 (normal/tumour)	Beer et al. [28]
Lung Boston	62	12,600	2 (normal/tumour)	Bhattacharjee et al. [29]
Gender	32	22,283	2 (male/female)	Unpublished
p53	50	12,625	2 (p53+/p53 mutant)	Unpublished

available for download at MSigDB (url: www.broadinstitute.org/gsea/download.js) [11].

3. Results and discussion

3.1. 10-fold cross validation

The results of wgSVM-SCAD from the 10-fold CV for all the data sets are compared with four embedded methods, which are gSVM-SCAD, L1-SVM, SVM-SCAD, and PathwayRF. L1-SVM is a penalized

Table 2
Comparisons of average 10-fold CV classification accuracy of top ten pathways with highest accuracy between wgSVM-SCAD and four embedded methods.

Method	Average accuracy (%)			
	Lung Michigan[a]	Lung Boston	Gender[a]	p53
wgSVM-SCAD	**80.86** (79.66, 80.48)	**71.83** (69.68, 70.81)	**88.75** (87.69, 88.32)	**82.22** (80.43, 81.45)
gSVM-SCAD	73.77 (—)	68.92 (66.48, 67.76)	87.33 (—)	81.17 (78.43, 80.23)
PathwayRF	71.00 (—)	65.16 (64.66, 65.06)	81.75 (—)	81.60 (80.73, 81.26)
L_1-SVM	55.14 (—)	67.13 (66.96, 67.11)	80.76 (—)	79.80 (79.03, 79.47)
SVM-SCAD	53.50 (—)	65.58 (64.99, 65.33)	77.96 (—)	73.93 (73.04, 73.49)

Note:
–Values in **bold** are highest accuracy.
–Values in the parenthesis are the 95% CI of the accuracy for each method for 10 experiment samples.
[a] Accuracy values are based on the published results from [13].

SVM classifier using L1 penalty function proposed by Zhu et al. [25]. While PathwayRF is the embedded method based on random forest developed for pathway analysis [14]. Table 2 shows the comparison of the performance in terms of average 10-fold CV accuracy of the ten pathways with highest classification accuracy. Comparison of average accuracy for Lung Michigan and gender data sets are based on the results in [13]. Whereas the comparison of average accuracy for Lung Boston and p53 data sets are based on self-experiments. The 95% CI of the accuracy are listed for the results from self-experiments. Based on the data shown in Table 2, wgSVM-SCAD obtains consistent improvements throughout the data sets except in p53 data set, where the improvement of accuracy is not that obvious compared to gSVM-SCAD and PathwayRF.

First, compared with gSVM-SCAD, the better results obtained by the proposed wgSVM-SCAD implies the effectiveness of the introduced *abs*T weights and the initial filter based on the *p*-value of the two-sample *t*-test in improving the identification of informative genes in the pathway. Meanwhile, in comparison with PathwayRF, which is an embedded method developed for pathway-based microarray analysis, the proposed wgSVM-SCAD obtains 9.86%, 6.67%, 7.00%, 0.62% higher accuracy in Lung Michigan, Lung Boston, Gender and p53 data sets respectively. This is due to the better gene selection behaviour in wgSVM-SCAD that consists of a group-specific parameter for every pathway and in the same time enhanced by the integrated *abs*T weighting process. As for L_1-SVM, better result from wgSVM-SCAD shows the potency of the SCAD penalty function in providing nearly unbiased estimation for large coefficients compared to LASSO penalty function, which leads to more consistent gene selection [24]. Lastly, compared with SVM-SCAD, flexibility from the group-specific parameter, λ_k and the assigned *abs*T weights have provided better identification of subset of informative genes that are closely related to the phenotype under study. Identification of these informative genes helps to build better classifier and thus better classification accuracy. Furthermore, according to Table 2, the improvement of results from wgSVM-SCAD on Lung Michigan data set is generally higher (for an average difference of 17.51%) in contrast to the results on Lung Boston, Gender and p53 data sets where the corresponding average differences of accuracy are 5.13%, 5.76%, and 3.10% respectively. The main difference between Lung Michigan data and other data sets is the distribution of classes. In Lung Michigan, the sample class distribution is more imbalanced (62 normal samples vs. 24 tumour samples). This also suggests that the proposed wgSVM-SCAD also performs well on data sets with imbalanced classes.

This research aims to develop an improved gSVM-SCAD with better identification of informative genes and pathways. Therefore, in order to analyse and justify the performance of the classifier built, a more comprehensive comparison is done between the proposed wgSVM-SCAD and gSVM-SCAD in terms of sensitivity, specificity and F-score. Both sensitivity and specificity measure how accurate a positive and negative sample is classified respectively. While F-Score is the harmonic measure between sensitivity and precision. A total of 10 runs of full experiment are done for both wgSVM-SCAD and gSVM-SCAD in order to show the consistency of the results of the proposed method. Table 3 shows the comparison of the average sensitivity, specificity and F-Score between wgSVM-SCAD and gSVM-SCAD for all data sets. The complete data and details for the experiments can be found in the supplementary files (Supp. Tables A1 and A4).

For Lung Michigan data, the average difference of sensitivity for the 10 runs is significantly lower compared to gSVM-SCAD at 1.06%. Despite of the low sensitivity, result from the proposed wgSVM-SCAD obtains a significant boost in specificity with an average of 16.28%, which supported by the small *p*-value as well as the 95% confidence interval (CI). As mentioned previously, the Lung Michigan data set consists of imbalanced class distribution, the result suggests that the introduced *abs*T weights have improved the identification of tumour samples because specificity is the measurement of how accurate a negative sample (in this case refers to tumour samples) is being classified. Furthermore, genes discarded during the filtration process could have increase the prediction power of the remaining genes in the pathway. At the same time, the discarded genes might cause the drop in sensitivity, where information from the discarded genes could have better prediction power on normal samples or positive samples. Meanwhile, F-score from the proposed wgSVM-SCAD has shown significant and consistent improvement (average difference of 1.98%) over the previous method. According to the 95% CI, there is 95% chance that improvement in F-Score of the proposed method over the previous method is within 1.60% and 2.36%.

For Lung Boston data set, the proposed wgSVM-SCAD shows consistent and significant improvement in both sensitivity and

Table 3
Average sensitivity, specificity and F-Score from the ten runs of experiments for both wgSVM-SCAD and gSVM-SCAD.

	Average sensitivity (%)		Difference (%)	*p*-value	95% CI
	wgSVM-SCAD	gSVM-SCAD			
Lung Michigan	89.59	**90.65**	-1.06	1.71E-02	(−1.90, −0.21)
Lung Boston	**70.82**	65.21	5.61	3.51E-07	(4.07, 7.14)
Gender	**93.10**	89.40	3.70	1.98E-03	(1.55, 5.85)
p53	**70.95**	68.05	2.90	0.145	(−1.10, 6.90)
	Average specificity (%)				
	wgSVM-SCAD	gSVM-SCAD			
Lung Michigan	**55.90**	39.62	16.28	3.68E-10	(13.49, 19.08)
Lung Boston	**69.53**	69.03	0.50	0.500	(−1.03, 2.04)
Gender	**86.10**	84.95	1.15	0.277	(−1.06, 3.46)
p53	84.50	**85.35**	-0.85	0.276	(−2.45, 0.75)
	Average F-Score (%)				
	wgSVM-SCAD	gSVM-SCAD			
Lung Michigan	**87.26**	85.28	1.98	1.73E-09	(1.60, 2.36)
Lung Boston	**71.95**	67.93	4.02	1.13E-07	(3.00, 5.05)
Gender	**92.17**	89.72	2.45	1.55E-03	(1.06, 3.84)
p53	**74.62**	72.78	1.84	9.19E-02	(−0.34, 4.01)

Note: *p*-value < 0.05 for significance.

F-Score based on small *p*-value obtained and the 95% CI. In sensitivity, the proposed wgSVM-SCAD has brought a significant increase of an average of 5.61% compared to gSVM-SCAD. In contrast to lung Michigan, lung Boston data set is well-balanced in the classes distribution (31 normal samples vs. 31 tumour samples). The increase of the performance is due to the better gene selection of the proposed method due to the integrated *abs*T weighting scheme and the *p*-value filter. In Lung Boston data set, the efficient gene selection was able to select better genes subsets that have better prediction power on normal samples without compromising too much in the prediction performance on tumour samples. Thus, better classification. Meanwhile, F-score obtained by the proposed wgSVM-SCAD in Lung Boston is steadily higher than gSVM-SCAD in an average of 4.03%. The 95% CI shows that in 95% of the time, the proposed method could obtain3–5.05% higher F-Score. This indicates that the classifier constructed by the proposed wgSVM-SCAD performs better than gSVM-SCAD in identifying informative genes in pathways.

For the gender data set, the result from wgSVM-SCAD shows increment of an average of 3.70% and 1.15% compared to the results from gSVM-SCAD for both sensitivity and specificity, respectively. Similar with the case of lung Boston data set, the more efficient gene selection provided by the proposed method has led to a significant and consistent improvement on identification of a certain class without compromising the identification of other class. The performance is depending on the class distribution in the data sets. While the average F-score of wgSVM-SCAD is 2.45% higher than the average F-score of gSVM-SCAD in the gender data set. wgSVM-SCAD achieves significant improvement in both sensitivity and F-Score, which supported with the small *p*-value obtained and the 95% CI.

For p53 data set, the average sensitivity of wgSVM-SCAD is 2.90% higher than gSVM-SCAD but unfortunately, the specificity of wgSVM-SCAD suffers a drop of 0.85% compared to gSVM-SCAD. This could be due to the initial filtration, which has filtered out the marginal informative genes that have better prediction power on negative samples or p53 mutant samples. Meanwhile, the average F-score of wgSVM-SCAD is 1.84% higher than gSVM-SCAD. However, compared with previous method, the proposed wgSVM-SCAD does not shows significant improvement in p53 data set. According to the statistics, overall performance of the proposed method in p53 data set is on par with the previous method. Overall, this implies that the better classifier is constructed in wgSVM-SCAD due to the efficient identification of informative genes in the pathways through the introduced *abs*T weighting scheme and filtration process.

3.2. Biological validation

In the proposed method, gene selection and classification is performed in each pathway. Therefore, the results consist of the list of pathways ranked based on the average 10-fold cv accuracy. Within each pathway, there is a list of selected informative genes which used to build the classifier and achieve the corresponding accuracy.

As shown in Fig. 4, the identified informative genes from the top five pathways with highest average 10-fold CV accuracy are selected and validated through online biological literature and databases in order to show the biological relevance of these genes to the phenotype under study. Meanwhile, Fig. 5 shows the flow of the process of biological validation. Based on Fig. 5, the identified informative genes and pathways by the proposed method in this research are check through biological literatures and databases like Genecards (url: www.genecards.com) [33] for the related publications in biological research that support the relevance of the genes or pathways to the phenotype of study. Genecards is a multifunctional database that contains numerous types of information regarding a single gene. This process iterates throughout the identified genes for the top five ranked pathways in this research. The validated informative genes or pathways will be further discussed in order to get more biological interpretation related to the phenotype of study.

Table 4 shows the top five pathways identified by the proposed method in Lung Michigan data set. The top ranked pathway is a NFAT signalling pathway, known as NFAT and Hypertrophy of the Heart. Although there is no proof for the direct relationship of this pathway to lung cancer, cardiac involvement in lung cancer is common [34]. Furthermore, expression of the nuclear factor of activated T-cell (NFAT) often involve in cancer cells progression including lung cancer [35]. The proposed wgSVM-SCAD has selected 9 informative genes for the classification and 8 of those genes (*CALR*, *CREBBP*, *HRAS*, *MYH2*, *PIK3R1*, *PRKACB*, *RAF1*, *RPS6KB1*) are found related to the development of lung cancer. Among the validated genes, there are several marker genes such as *CALR* gene, which its expression is associated with tumour pathological grade and previous research has shown that it can be used as a biomarker for lung cancer prediction and diagnosis [36], *RAF1* gene, which is associated with tumorigenesis and it is an independent prognostic marker for poor survival rates of lung cancer [37], *RPS6KB1* gene, which is a prognostic marker for tumour development in lung cancer [38], and *CREBBP* gene, where mutation or deletion of this gene contributes to the genesis and the progression of lung cancer cells [39]. The second pathway with highest 10-fold CV accuracy is a Vitamin D pathway known as Control of Gene Expression by Vitamin D Receptor (VDR). The nuclear *VDR* status has been suggested as a prognostic marker in lung cancer [40]. Investigation also shows that deficiency of *VDR* occurs in patients with lung cancer [41]. Based on the validation through biological literatures, 8 of the 10 identified informative genes (*CREBBP*, *EP300*, *VDR*, *MED1*, *SMARCC1*, *SMARCC2*, *TOP2B*, *TSC2*) are

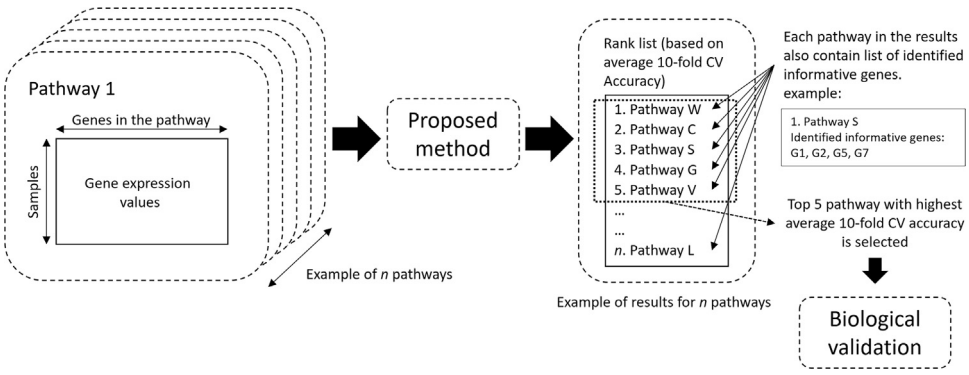

Fig. 4. Illustration of selection of the top five pathways from the results for biological validation.

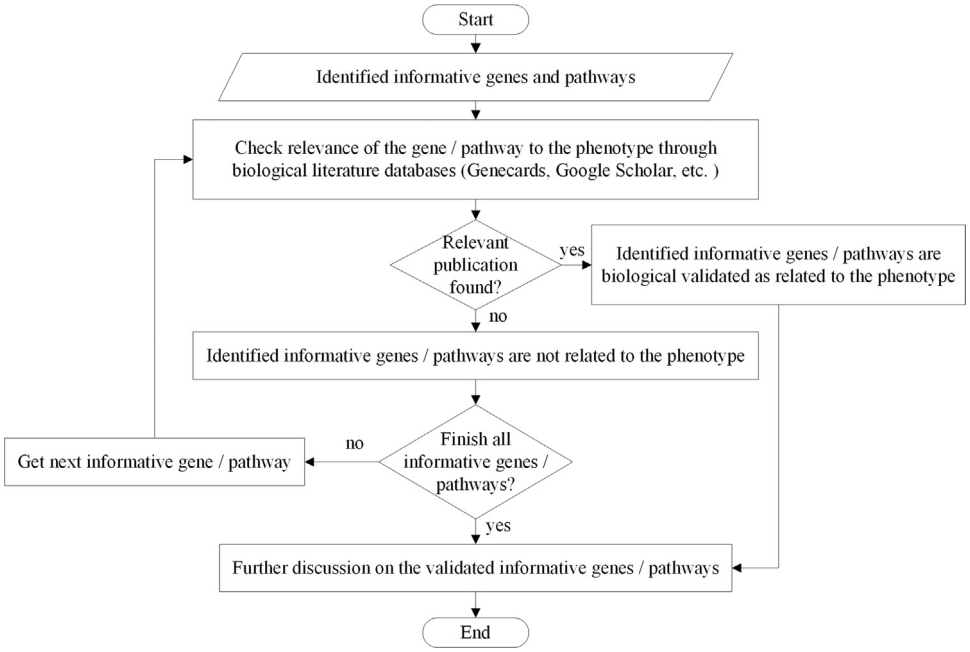

Fig. 5. Flow of the process of biological validation.

Table 4
Selected genes from top five pathways in Lung Michigan data set.

Pathways	No. of genes in the pathway	No. of selected genes	Selected genes
NFAT and hypertrophy of the heart	60	9	***CALR*** [36], ***CREBBP*** [39], ***HRAS*** [42], ***MYH2*** [43], ***PIK3R1*** [44], ***PRKACB*** [45], ***RAF1*** [37], ***RPS6KB1*** [38], *CAMK4*
Control of gene expression by vitamin D receptor	21	10	***CREBBP*** [39], ***EP300*** [46], ***VDR*** [47], ***MED1*** [48], ***SMARCC1*** [49], ***SMARCC2*** [50], ***TOP2B*** [51], ***TSC2*** [52], *NCOA2*, *SMARCD1*
Limonene and pinene degradation	20	9	***ALDH1A1*** [53], ***ALDH2*** [54], ***CYP24A1*** [55], ***HADHA*** [56], ***NAT6*** [29], *EHHADH*, *ALDH1B1*, *ALDH3A2*, *ALDH9A1*
Butanoate metabolism	43	9	***PRDX6*** [57], ***ALDH2*** [54], ***HADHA*** [56], *ABAT*, *ALDH3A2*, *ALDH9A1*, *GAD1*, *GAD2*, *EHHADH*
IL 17 signalling pathway	20	11	***CD34*** [58], ***CD3D,CD3E,CD3G*** [59], ***IL6*** [60], ***IL8*** [61], *PRSS1*, *TRBV21–1*, *CD2*, *CD247*, *CD58*

Note: The genes in **bold** are genes directly related to lung cancer.

found related to the development of lung cancer. The third ranked pathway is known as limonene and pinene degradation pathway. Previous investigations show that pinene has antitumor effect against lung cancer cells [62], while limonene is capable in inhibiting metastatic progression of cancer cells [63]. 9 informative genes have been selected by the proposed wgSVM-SCAD and 5 of those genes (*ALDH1A1*, *ALDH2*, *CYP24A1*, *HADHA*, *NAT6*) are validated for their involvement in lung cancer. Among these genes, *CYP24A1* is a known independent prognostic marker associated with the survival of patients in lung cancer, where overexpression of *CYP24A1* abrogates the antiproliferative effects of 1α,25-dihydroxyvitamin D (3), an active form of vitamin D [55]. The fourth ranked pathway is butanoate metabolism pathway. Butanoate also known as butyrate, which is proven its potential therapeutic implications for diseases such as cancers [64]. Furthermore, investigation has revealed that recognition of alteration in metabolism could have prognostic impact in lung cancer [65]. 9 genes have been selected by the proposed wgSVM-SCAD for classification and 3 of these genes (*PRDX6*, *ALDH2*, *HADHA*) are found related to the lung cancer based on the biological literature. Lastly, the fifth ranked pathway is IL-17 signalling pathway. Previous research has shown that Interleukin 17 (IL-17) is associated with the risk of progressive cancers including lung cancer [66]. In this pathway, the proposed wgSVM-SCAD has selected 11 genes for classification and 6 of these selected genes (*CD34*, *CD3D*, *CD3E*, *CD3G*, *IL6*, *IL8*) are found related to the lung cancer. Among the validated genes, *CD34* gene has been identified as potential marker of lung cancer naïve cells [58]. It encodes the protein that plays a role in attachment of stem cells to stromal cells and previous investigation has shown that CD34-positive stromal cells are specific in the stoma of lung adenocarcinomas and may play a supportive role in primary lung cancer [67].

For Lung Boston data set, the top five pathways are shown in Table 5. The top ranked pathway is the polycomb repressive complexes (PRC2) pathway. PRC2 has been frequently implicated in human cancer, acting either as oncogenes or tumour suppressors. High expression of PRC2 components in most of the small cell lung cancer (SCLC) has been reported by Murai et al. [68] and it may play a role in genesis in SCLC as well [69]. Recent investigation also shows that *PRC2* is a critical regulator of KRAS-driven non-small cell lung carcinoma progression [70]. 5 genes have been selected by the proposed method for classification and 3 of the selected genes (*SUZ12*, *RBBP7*, *YY1*) are found related to the lung cancer. *SUZ12* gene has been reported as an oncogene and a potential diagnostic marker in non-small-cell carcinoma (NSCLC), which associated with the promote of lung tumour growth, migration and invasion [71]. While for the *RBBP7* gene, it is found that this gene promotes migration ability of lung cancer cells and

Table 5
Selected genes from top five pathways in Lung Boston data set.

Pathways	No. of genes in the pathway	No. of selected genes	Selected genes
The PRC2 complex sets long-term gene silencing through modification of histone tails	14	5	***SUZ12*** [71], ***RBBP7*** [72], ***YY1*** [73], *EZH1*, *RBBP4*
Visual signal transduction	24	5	*PDE6B*, *RHO*, *SAG*, *SAFB*, *PDE6A*
Glutamate metabolism	30	12	***GCLC*** [74], ***GCLM*** [75], ***GLS*** [76], ***GSS*** [77], *GAD2*, *GLS2*, *GLUD1*, *GFPT1*, *TMEM11*, *GMPS*, *ALDH4A1*, *ALDH5A1*
Estrogen responsive protein efp controls cell cycle and breast tumours growth	15	9	***CCNB1*** [78], ***CDK4*** [79], ***CDK5*** [80], ***CDK6*** [81], ***SFN*** [82], ***SMURF2*** [83], ***TP53*** [84], ***TRIM25*** [85], *ABCB7*
Regulation of EIF2	11	6	***EIF2AK2*** [86], ***EIF2S1*** [87], ***GSK3B*** [88], *EIF2S2*, *EIF2S3*, *EIF5*

Note: The genes in **bold** are genes directly related to lung cancer.

it is a novel biomarker and prognostic marker for distant metastasis in NSCLC [72]. The second ranked pathway is a visual signal transduction pathway. However, there is no support from literature regarding the relationship between the visual signal transduction pathway and development of lung cancer. The third ranked pathway is glutamate metabolism pathway. Cellular metabolism has been targeted by researchers to improve cancer therapeutics because metabolism in cancer cells is significantly different from normal cells [89]. Glutamate regulates proliferation of neurons and previous research suggests that it is also associated with the motility and the invasive growth of tumour cells in lung cancer [90]. 12 genes have been selected by the proposed wgSVM-SCAD and 4 of these genes (*GCLC*, *GCLM*, *GLS*, *GSS*) are found related to lung cancers according to the biological literature. The fourth ranked pathway is an estrogen signalling pathway. Estrogen is involved in the biology of NSCLC and inhibition of estrogen synthesis have been shown to prevent lung tumorigenesis and inhibit the growth of lung tumour [91]. The proposed method in this research has selected 9 genes for classification and 8 of these genes (*CCNB1*, *CDK4*, *CDK5*, *CDK6*, *SFN*, *SMURF2*, *TP53*, *TRIM25*) are identified as lung cancer related. *CCNB1* and *CDK4* genes are among the potential prognostic marker in NSCLC, where both *CCNB1* and *CDK4* genes are associated with pathogenesis of lung cancer [78,79]. Lastly, the fifth ranked pathway is *eIF2* regulation pathway. The protein synthesis initiation factor, *eIF2* is important for translation initiation and protein synthesis [87]. 6 genes have been selected by the proposed method and 3 of the selected genes (*EIF2AK2*, *EIF2S1*, *GSK3B*) are validated as related to lung cancer based on the biological literatures. Expression of *EIF2AK2* gene is associated with cell growth and it has identified as an independent prognostic variable in NSCLC patients, where low expression of EIF2AK2 leads to aggressive behaviour of lung cancer cells [86], while *GSK3B* gene is associated with tumorigenesis and aberrant expression of *GSK3B* gene is known as an independent marker of poor prognosis in NSCLC [88].

Table 6 shows the top five pathways identified by wgSVM-SCAD. The top ranked pathway is GNF_FEMALE_GENES pathway. In this pathway, 1 gene, XIST has been selected by the proposed method. *XIST* gene is a non-protein coding gene exclusively expressed in female that is involved in the X chromosome silencing in female cells and allow X chromosome equilibration with males [92]. A recent study also found that overexpression of *XIST* is involved in psychiatric disorders in females [93]. The second ranked pathway is TESTIS_GENES_FROM_XHX_AND_ NETAFFX pathway. Similar to the top ranked pathway, 1 gene has been selected by the proposed method, which is *RPS4Y1* and it is found differentially expressed between male and female [94]. *RPS4Y1* gene is located in the Y chromosome in males, which encodes for ribosomal protein S4 [95]. While the third ranked pathway is the *XINACT* pathway, which is an inactivation process of X chromosome to achieve equal dosage of X chromosome genes in males and females [96]. 2 genes have been selected by the proposed method and both (*DDX3X* and *RPS4X*) are validated through biological literature. The fourth ranked pathway is RAP_DOWN, which is related to the rapamycin down regulation. Previous studies found that rapamycin can extend the lifespan in mice, especially for female mice [97,98]. 3 genes have been selected by the proposed method and 2 of these (*DDX3X* and *HDHD1A*) are found to be related in the differentiation between genders. Lastly, the fifth ranked pathway is INSULIN_2F_UP pathway. A previous study has shown the role of insulin in cell proliferation [99]. 3 genes have been selected by the proposed method for classification and 1 gene is validated through the biological literature, which is *HDHD1A*.

For p53 data set, the top five pathways identified by the proposed method is shown in Table 7. The top ranked pathway is known to influence RAS and RHO proteins on G1 to S transition pathway. P53 is a well-known tumour suppressor protein. Mutation or loss of p53 have been a critical event in many human cancers. Previous investigation has shown that mutation of p53 if coupled with the activated RAS can induces RhoA activity and promotes cancer progression [100]. 10 genes have been selected by wgSVM-SCAD for classification and 9 of the genes (*AKT1*, *RHOA*, *CDK4*, *CDKN1A*, *E2F1*, *IKBKB*, *RAC1*, *RELA*, *TFDP1*) are found related to the p53 mutation in the cancer cell lines. The second ranked pathway, is known as Cell Cycle: G1/S Check Point pathway. Previous studies have reported the involvement of p53 in regulation of G1 checkpoint to control cell proliferation by triggering apoptosis in normal cells as well as cancer cells [101,102]. Thus, mutation in p53 leads to the differential expression of this pathway. In

Table 6
Selected genes from top five pathways in Gender data set.

Pathways	No. of genes in pathway	No. of selected genes	Selected genes
GNF_FEMALE_GENES	116	1	***XIST*** [93]
TESTIS_GENES_FROM_XHX_AND_NETAFFX	111	1	***RPS4Y1*** [94]
XINACT	34	2	***DDX3X***, ***RPS4X*** [94]
RAP_DOWN	434	3	***HDHD1A***, ***DDX3X*** [94], *RTN4*
INSULIN_2F_UP	405	3	***HDHD1A*** [94], *RPS20*, *RPL41*

Note: The genes in **bold** are biological validated differentially expressed genes between male and female in lymphoblastoid cell lines.

Table 7
Selected genes from top five pathways in p53 data set.

Pathways	No. of genes in the pathway	No. of selected genes	Selected genes
Influence of RAS and RHO proteins on G1 to S Transition	28	10	***AKT1*** [103], ***RHOA*** [104], ***CDK4*** [105], ***CDKN1A*** [106], ***E2F1*** [107], ***IKBKB*** [108], ***RAC1*** [109], ***RELA*** [110], ***TFDP1*** [111], *PAK1*
Cell cycle: G1/S check point	26	11	***ABL1*** [112], ***ATR*** [113], ***CDK2*** [114], ***CDKN1A*** [106], ***CDKN2A*** [115], ***CDKN2B*** [116], ***DHFR*** [117], ***E2F1*** [107], ***HDAC1*** [118], *SMAD3*, *TGFB3*
Apoptosis	81	7	***BAX*** [119], ***BCL2*** [120], ***CASP8*** [121], ***RELA*** [110], ***TNFRSF10B*** [122], *DFFA*, *IL1R1*
Nuclear receptors in lipid metabolism and toxicity	53	8	***PPARG*** [123], ***RARB*** [124], ***VDR*** [125], *CYP1A2*, *CYP2B7P*, *CYP2E1*, *CYP2C9*, *SFTPB*
Role of BRCA1 and BRCA2 and ATR in cancer susceptibility	22	10	***MRE11A***, ***ATM*** [126], ***ATR*** [113], ***BRCA2*** [127], ***CHEK1*** [128], ***NBN*** [129], ***RAD50*** [130], *BRDT*, *RAD1*, *FANCG*

Note: The genes in **bold** are genes related to p53 mutations in cancer cell lines.

this pathway, the proposed method has selected 11 genes for classification and 9 of these genes (*ABL1*, *ATR*, *CDK2*, *CDKN1A*, *CDKN2A*, *CDKN2B*, *DHFR*, *E2F1*, *HDAC1*) are identified as closely related to the p53 mutation in cancer cell lines. The third ranked pathway is the apoptosis pathway. *P53* has been known for its influence in cell cycle control and apoptosis. Mutation of p53 or loss of p53 could deregulate apoptosis and promotes tumour cell migration and invasion [131]. 7 genes have been selected in this pathway by wgSVM-SCAD for cancer classification. 5 of the selected genes (*BAX*, *BCL2*, *CASP8*, *RELA*, *TNFRSF10B*) are found related to the *p53* mutation in cancer cell lines. The fourth ranked pathway is the Nuclear Receptors in Lipid Metabolism and Toxicity pathway. *P53* has been reported as a novel regulator in lipid metabolism [132], therefore, mutation on P53 could lead to deregulation of the lipid metabolism. Furthermore, several lipid ligands for some of the orphan receptors have been identified, where upon binding to these ligands, several proteins are synthesized including some cytochrome P450 member proteins which catalyse lipid metabolism, storage, elimination as well as the activation of procarcinogens [133]. The proposed method has selected 8 genes in this pathway for classification. 3 of these genes (*PPARG*, *RARB*, *VDR*) are found related to the p53 mutation in cancer cell lines.

Lastly, the fifth ranked pathway is the *BRCA* related pathway known as Role of *BRCA1* and *BRCA2* and *ATR* in cancer susceptibility. Similar to *p53*, *BRCA1* gene is a tumour suppressor gene involved in deoxyribonucleic acid (DNA) damage response pathway and mutation of *BRCA1* gene is common in a variety of cancers. A previous study has reported that expression of *BRCA1* is controlled by the presence of *p53* [134]. Moreover, it was reported that *p53* deficiency is highly cooperative with both *BRCA1* and *BRCA2* in promoting tumorigenesis [135]. 10 genes have been selected from this pathway by the proposed method for classification and 7 of these genes (*MRE11A*, *ATM*, *ATR*, *BRCA2*, *CHEK1*, *NBN*, *RAD50*) are found related to the *p53* mutation in cancer cell lines.

4. Discussion

The proposed wgSVM-SCAD is aimed to provide effective gene selection in order to surmount the performance limitation caused by the existence of non-informative genes in the pathways. Based on the 10-fold CV performance shown in the Table 2, the proposed method obtains improved classification accuracy compared to gSVM-SCAD as well as previous studies throughout all the data sets. Furthermore, performance of both wgSVM-SCAD and gSVM-SCAD have been further evaluated in terms of specificity and sensitivity as shown in Tables 3 and 4. The performance of wgSVM-SCAD is consistently improved compared to the results of gSVM-SCAD throughout all the data sets. This indicates that the gene selection efficiency of the proposed method has been increased because the use of informative genes in classifier construction leads to better classification performance. This shows that the introduced *abs*T weights for every gene in the pathway can improve the gene selection efficiency by emphasizing the differential expression of the genes in the pathway. Moreover, the initial filtration step also helps to filter out non-informative genes based on the *p*-value. While the remaining weighted genes become better candidates in the selection process by the SCAD penalty function. Besides, notable improvement in the specificity in lung Michigan data set implies that the introduced *abs*T weights and initial filtration is able to improve the classification performance on data set with imbalanced class distribution (62 normal samples and 24 tumour samples). Thus, the proposed method, wgSVM-SCAD is a better classifier compared to gSVM-SCAD, which is further supported by the improved F-scores shown in Table 5.

As for biological validation, top five pathways with highest classification accuracy have been validated and shown in previous section. It's worth mentioning that there are several genes that have been selected more than once in different pathways and even different data sets. These genes have been validated through

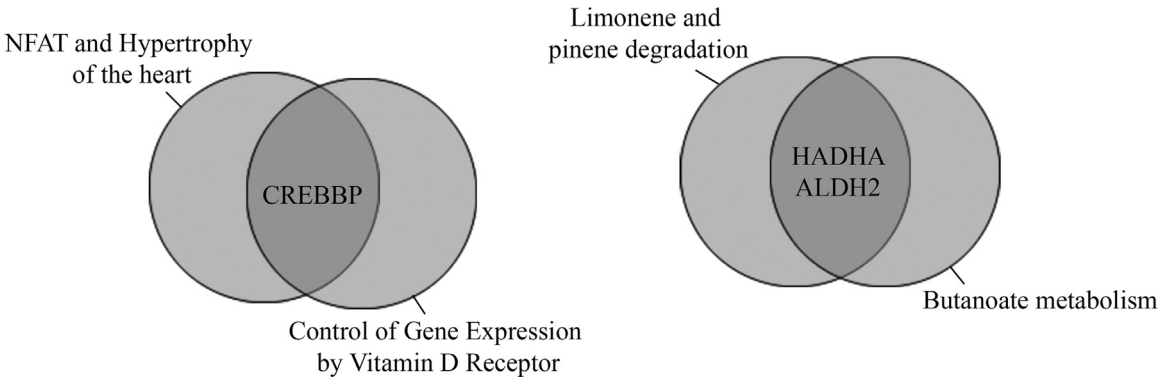

Fig. 6. Overlapped genes identified in the top five pathways in Lung Michigan data set.

biological literature. This indicates that some of the selected genes could be the potential driver genes between two or more processes involved in the development of cancers. Fig. 6 shows the genes that have been identified more than once in the top five pathways in lung Michigan data set. As discussed in previous section, the NFAT pathway and Vitamin D Receptor pathway are associated with the cancer cells progression [40,41,35,34]. While *CREBBP* gene has been reported for its role in development of lung cancer, where mutation or deletion of this gene has resulted the genesis and progression of lung cancer cells [39]. Meanwhile, *HADHA* and *ALDH2* also have been selected in both limonene and pinene degradation pathway and butanoate metabolism pathway. These two pathways are associated with antitumor effect and the therapeutic implications in lung cancer.

Meanwhile, Fig. 7 shows the overlapped genes that have been selected by the proposed method in different pathways. *HDHD1A* gene has been selected in both XINACT and RAP_DOWN pathway while *DDX3X* gene has been selected in both RAP_DOWN and INSULIN_2F_UP pathway. Higher expression of both *HDHD1A* and *DDX3X* gene are found in female compared to male according to the previous investigations [94,136]. Recent investigation has shown the role of X inactivation and rapamycin downregulation pathway in gender-specific functions [137,96].

For p53 data set, several genes have been selected in more than one pathway as shown in Fig. 8. *RELA* gene, where its function is associated with cancer cells progression, has been identified in both apoptosis pathway and RAS and RHO pathway. While *CDKN1A* and *E2F1* have been selected in both RAS and RHO pathway and G1/S check point pathway, which are closely related to the cell cycle control. *CDKN1A* is involved in the G1 progression control, which associated negative regulation of p53 tumour suppressor [106] while *E2F1* plays an important role in making vital cellular decision to regulate the cell cycle progression and defect on the *E2F1* and *P53* is associated with tumour development [107]. Meanwhile, *ATR* gene has been selected in both G1/S check point pathway and the *BRCA* pathway. *ATR* plays an important role in regulation of cancer cells proliferation.

The weighting scheme used in this paper allows more efficient detection of informative genes and pathways that are differentially expressed between two condition. With the rise of the next-generation sequencing such as RNA-seq, which has shown strong potential to replace microarrays for whole-genome transcriptome profiling. This is because RNA-seq offers better specificity and sensitivity in examining transcriptome fine structure. In this case, the weighting scheme could be beneficial in improving the detection of differential expressed transcript between different conditions. Meanwhile, in comparison with some of the widely used pathway analysis software suite, such as Ingenuity Pathway analysis and GeneSpring, these software suites offer very comprehensive analysis capabilities such as comparative analysis,

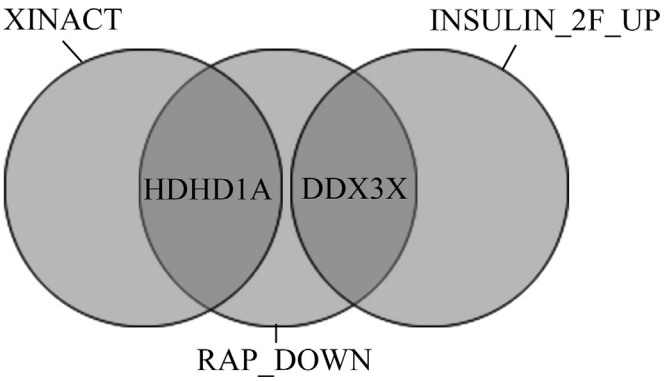

Fig. 7. Overlapped genes identified in the top five pathways in Gender data set.

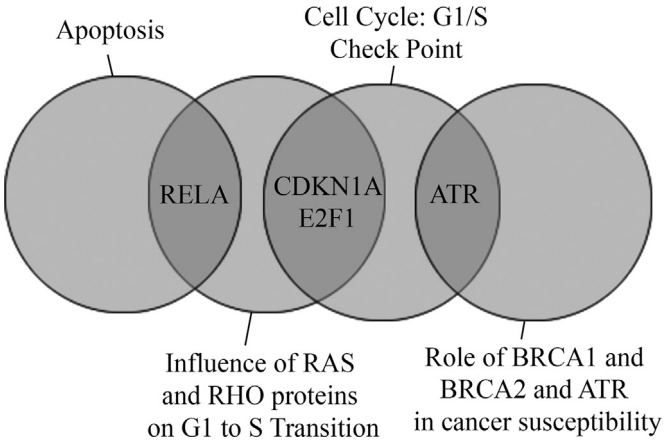

Fig. 8. Overlapped genes identified in the top five pathways in p53 data set.

causal network analysis, multi-omics analysis, visualization and etc. In terms of the specific usage in analysis of pathway differential expression, our method executes gene selection and classification for each pathway and these pathways are ranked based on the classification accuracy. This provides not only the informative pathways but also the informative genes that have been detected within each pathway.

5. Conclusion

The proposed wgSVM-SCAD has shown overall improvements over gSVM-SCAD in terms of classification accuracy, sensitivity, specificity and F-Score. Integration of the *abs*T weights improves the performance of the selection, allowing better identification of the informative genes in the pathways related to the phenotype under study and leads to better classification performance. This provides a potential solution to effectively identify informative genes related to the phenotype under study from pathway data that are often not based on specific biological contexts. Furthermore, biological validation of the top five pathways shows that our proposed method is able to identify some of the marker genes as well as genes that play important roles in the development of the phenotype under study. Despite the overall improvements of the proposed wgSVM-SCAD over gSVM-SCAD, the introduction of *abs*T weights and the initial filtration have led to the slight performance drop on sensitivity in lung Michigan data set, and specificity in lung Boston, gender and p53 data sets. This could be due to the sub-optimal choice of parameter λ in SCAD because SCAD penalty is parametric. Therefore, its performance of gene selection is relying on the choice of λ. When the λ is too big, underfitting of data may occur and produces a very sparse classifier. When the λ is too small, there is risk of overfitting the data and produces a less sparse classifier. As the proposed method uses a grid search of predefined set of λ for SCAD, improvement could be done by integrating more robust penalty with better λ selection using stochastic search such as metaheuristics.

Conflict of interest

None declared.

Acknowledgement

This research is supported by Malaysian Ministry of Education through several Fundamental Research Grant Schemes (Grant

numbers: 1559, R.J130000.7828.4F720, and S.J130000.7828.4X115). We also would like to thank the Office of Deputy Vice Chancellor for Research and Graduate Studies, United Arab Emirates University (UAEU) for supporting this research work through a UAEU Grant (Grant number: 31T052-UPA).

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.compbiomed.2016.08.004.

References

[1] D. Gomez-Cabrero, I. Abugessaisa, D. Maier, A. Teschendorff, M. Merkenschlager, A. Gisel, E. Ballestar, E. Bongcam-Rudloff, A. Conesa, J. Tegner, Data integration in the era of omics: current and future challenges, BMC Syst. Biol. 8 (Suppl. 2) (2014), I1.
[2] J.M. Villaveces, P. Koti, B.H. Habermann, Tools for visualization and analysis of molecular networks, pathways, and -omics data, Adv. Appl. Bioinform. Chem. 8 (2015) 11–22.
[3] D. Nam, S.Y. Kim, Gene-set approach for expression pattern analysis, Brief. Bioinform. 9 (3) (2008) 189–197.
[4] R.K. Curtis, M. Oresic, A. Vidal-Puig, Pathways to the analysis of microarray data, Trends Biotechnol. 23 (8) (2005) 429–435.
[5] P. Khatri, M. Sirota, A.J. Butte, Ten years of pathway analysis: current approaches and outstanding challenges, PLoS Comput. Biol. 8 (2) (2012) e1002375.
[6] M.E. Adriaens, M. Jaillard, A. Waagmeester, S.L.M. Coort, A.R. Pico, C.T. A. Evelo, The public road to high-quality curated biological pathways, Drug Discov. Today 13 (19–20) (2008) 856–862.
[7] X.W. Wang, E. Dalkic, M. Wu, C. Chan, Gene module level analysis: identification to networks and dynamics, Curr. Opin. Biotechnol. 19 (5) (2008) 482–491.
[8] M. Scheer, F. Klawonn, R. Munch, A. Grote, K. Hiller, C. Choi, I. Koch, M. Schobert, E. Hartig, U. Klages, D. Jahn, JProGO: a novel tool for the functional interpretation of prokaryotic microarray data using gene ontology information, Nucleic Acids Res. 34 (2006) W510–W515.
[9] C. Backes, A. Keller, J. Kuentzer, B. Kneissl, N. Comtesse, Y.A. Elnakady, R. Muller, E. Meese, H.P. Lenhof, GeneTrai – advanced gene set enrichment analysis, Nucleic Acids Res. 35 (2007), W186-192.
[10] V. Mootha, C. Lindgren, K. Eriksson, A. Subramanian, S. Sihag, J. Lehar, P. Puigserver, E. Carlsson, M. Ridderstrale, E. Laurila, N. Houstis, M. Daly, N. Patterson, J. Mesirov, T. Golub, P. Tamayo, B. Spiegelman, E. Lander, J. Hirschhorn, D. Altshuler, L. Groop, PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes, Nat. Genet 34 (2003) 267–273.
[11] A. Subramanian, P. Tamayo, V. Mootha, S. Mukherjee, B. Ebert, M. Gillette, A. Paulovich, S. Pomeroy, T. Golub, E. Lander, J. Mesirov, Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles, Proc. Natl. Acad. Sci. USA 102 (2005) 15545–15550.
[12] S. Draghici, P. Khatri, A. Tarca, K. Amin, A. Done, C. Voichita, C. Georgescu, R. Romero, A systems biology approach for pathway level analysis, Genome Res. 17 (2007) 1537–1545.
[13] M.F. Misman, M.S. Mohamad, S. Deris, S.Z. Hashim, A group-specific tuning parameter for hybrid of SVM and SCAD in identification of informative genes and pathways, Int J. Data Min. Bioinform 10 (2) (2014) 146–161.
[14] H. Pang, A. Lin, M. Holford, B.E. Enerson, B. Lu, M.P. Lawton, E. Floyd, H. Zhao, Pathway analysis using random forests classification and regression, Bioinformatics 22 (16) (2006) 2028–2036.
[15] F. Tai, W. Pan, Incorporating prior knowledge of predictors into penalized classifiers with multiple penalty terms, Bioinformatics 23 (14) (2007) 1775–1782.
[16] E. Lee, H.Y. Chuang, J.W. Kim, T. Ideker, D. Lee, Inferring pathway activity toward precise disease classification, PLoS Comput. Biol. 4 (11) (2008) e1000217.
[17] F. Tai, W. Pan, Incorporating prior knowledge of gene functional groups into regularized discriminant analysis of microarray data, Bioinformatics 23 (23) (2007) 3170–3177.
[18] X. Chen, L. Wang, J.D. Smith, B. Zhang, Supervised principal component analysis for gene set enrichment of microarray data with continuous or survival outcomes, Bioinformatics 24 (21) (2008) 2474–2481.
[19] P. Minguez, J. Dopazo, Functional genomics and networks: new approaches in the extraction of complex gene modules, Expert Rev. Proteom. 7 (1) (2010) 55–63.
[20] Y. Saeys, I. Inza, P. Larranaga, A review of feature selection techniques in bioinformatics, Bioinformatics 23 (19) (2007) 2507–2517.
[21] Z.M. Hira, D.F. Gillies, A review of feature selection and feature extraction methods applied on microarray data, Adv. Bioinform. 2015 (2015) 198363.
[22] S. Ma, J. Huang, Penalized feature selection and classification in bioinformatics, Brief. Bioinform. 9 (5) (2008) 392–403.
[23] H.H. Zhang, J. Ahn, X.D. Lin, C. Park, Gene selection using support vector machines with non-convex penalty, Bioinformatics 22 (1) (2006) 88–95.
[24] J. Fan, R. Li, Variable selection via nonconcave penalized likelihood and its oracle properties, J. Am. Stat. Assoc. 96 (2001) 1348–1360.
[25] J. Zhu, S. Rosset, T. Hastie, R. Tibshirani, 1-norm support vector machines, Adv. Neural Inf. Process. Syst. 16 (16) (2004) 49–56.
[26] S.S. Ha, I. Kim, Y. Wang, J. Xuan, Applications of different weighting schemes to improve pathway-based analysis, Comp. Funct. Genom. 2011 (2011) 463645.
[27] G. Wahba, Y. Lin, H. Zhang, Generalized Approximate Cross Validation for Support Vector Machines, or, Another Way to Look at Margin-like Quantities, 1999.
[28] D.G. Beer, S.L. Kardia, C.C. Huang, T.J. Giordano, A.M. Levin, D.E. Misek, L. Lin, G. Chen, T.G. Gharib, D.G. Thomas, M.L. Lizyness, R. Kuick, S. Hayasaka, J. M. Taylor, M.D. Iannettoni, M.B. Orringer, S. Hanash, Gene-expression profiles predict survival of patients with lung adenocarcinoma, Nat. Med. 8 (8) (2002) 816–824.
[29] A. Bhattacharjee, W.G. Richards, J. Staunton, C. Li, S. Monti, P. Vasa, C. Ladd, J. Beheshti, R. Bueno, M. Gillette, M. Loda, G. Weber, E.J. Mark, E.S. Lander, W. Wong, B.E. Johnson, T.R. Golub, D.J. Sugarbaker, M. Meyerson, Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses, Proc. Natl. Acad. Sci. USA 98 (24) (2001) 13790–13795.
[30] Z. Fang, W. Tian, H. Ji, A network-based gene-weighting approach for pathway analysis, Cell Res. 22 (3) (2012) 565–580.
[31] M. Kanehisa, S. Goto, KEGG: kyoto encyclopedia of genes and genomes, Nucleic Acids Res. 28 (1) (2000) 27–30.
[32] D. Nishimura, BioCarta, Biotechnol. Softw. Internet Rep.: Comput. Softw. J. Sci. 2 (3) (2001) 117–120.
[33] M. Rebhan, V. Chalifa-Caspi, J. Prilusky, D. Lancet, GeneCards: integrating information about genes, proteins and diseases, Trends Genet. 13 (4) (1997) 163.
[34] B. Cheruvu, P. Cheruvu, M. Boyars, An unusual case of metastasis to the left side of the heart: a case report, J. Med Case Rep. 5 (2011) 23.
[35] K. Zhang, N. Li, Z. Chen, K. Shao, F. Zhou, C. Zhang, X. Mu, J. Wan, B. Li, X. Feng, S. Shi, M. Xiong, K. Cao, X. Wang, C. Huang, J. He, High expression of nuclear factor of activated T cells in Chinese primary non-small cell lung cancer tissues, Int J. Biol. Markers 22 (3) (2007) 221–225.
[36] R. Liu, J. Gong, J. Chen, Q. Li, C. Song, J. Zhang, Y. Li, Z. Liu, Y. Dong, L. Chen, B. Jin, Calreticulin as a potential diagnostic biomarker for lung cancer, Cancer Immunol. Immunother. 61 (6) (2012) 855–864.
[37] Z.-X. Qiu, L. Wang, J. Han, D. Liu, W. Huang, K. Altaf, X.-S. Qiu, M.A. Javed, J. Zheng, B.-J. Chen, W.-M. Li, Prognostic impact of Raf-1 and p-Raf-1 expressions for poor survival rate in non-small cell lung cancer, Cancer Sci. 103 (10) (2012) 1774–1779.
[38] Y. Zhang, H.J. Ni, D.Y. Cheng, Prognostic value of phosphorylated mTOR/RPS6KB1 in non- small cell lung cancer, Asian Pac. J. Cancer Prev. 14 (6) (2013) 3725–3728.
[39] M. Kishimoto, T. Kohno, K. Okudela, A. Otsuka, H. Sasaki, C. Tanabe, T. Sakiyama, C. Hirama, I. Kitabayashi, J.D. Minna, S. Takenoshita, J. Yokota, Mutations and deletions of the CBP gene in human lung cancer, Clin. Cancer Res. 11 (2005) 512–519.
[40] M. Srinivasan, A.V. Parwani, P.A. Hershberger, D.E. Lenzner, J.L. Weissfeld, Nuclear vitamin D receptor expression is associated with improved survival in non-small cell lung cancer, J. Steroid Biochem. Mol. Biol. 123 (1–2) (2011) 30–36.
[41] L. Vuolo, C. Di Somma, A. Faggiano, A. Colao, Vitamin D and cancer, Front. Endocrinol. 3 (2012) 58.
[42] C. Mascaux, N. Iannino, B. Martin, M. Paesmans, T. Berghmans, M. Dusart, A. Haller, P. Lothaire, A.P. Meert, S. Noel, J.J. Lafitte, J.P. Sculier, The role of RAS oncogene in survival of patients with lung cancer: a systematic review of the literature with meta-analysis, Br. J. Cancer 92 (1) (2005) 131–139.
[43] Z.H. Li, A.R. Bresnick, The S100A4 metastasis factor regulates cellular motility via a direct interaction with myosin-IIA, Cancer Res. 66 (10) (2006) 5173–5180.
[44] C.R. Zito, L.B. Jilaveanu, V. Anagnostou, D. Rimm, G. Bepler, S.M. Maira, W. Hackl, R. Camp, H.M. Kluger, H.H. Chao, Multi-level targeting of the phosphatidylinositol-3-kinase pathway in non-small cell lung cancer cells, PloS One 7 (2) (2012) e31331.
[45] Y. Chen, Y. Gao, Y. Tian, D.L. Tian, PRKACB is downregulated in non-small cell lung cancer and exogenous PRKACB inhibits proliferation and invasion of LTEP-A2 cells, Oncol. Lett. 5 (6) (2013) 1803–1808.
[46] C.X. Zheng, Z.H. Gu, B. Han, R.X. Zhang, C.M. Pan, Y. Xiang, X.J. Rong, X. Chen, Q.Y. Li, H.Y. Wan, Whole-exome sequencing to identify novel somatic mutations in squamous cell lung cancers, Int J. Oncol. 43 (3) (2013) 755–764.
[47] S.H. Kim, G. Chen, A.N. King, C.K. Jeon, P.J. Christensen, L. Zhao, R.U. Simpson, D.G. Thomas, T.J. Giordano, D.E. Brenner, B. Hollis, D.G. Beer, N. Ramnath, Characterization of vitamin D receptor (VDR) in lung adenocarcinoma, Lung Cancer 77 (2) (2012) 265–271.
[48] H.J. Kim, M.S. Roh, C.H. Son, A.J. Kim, H.J. Jee, N. Song, M. Kim, S.Y. Seo, Y. H. Yoo, J. Yun, Loss of Med1/TRAP220 promotes the invasion and metastasis of human non-small-cell lung cancer cells by modulating the expression of metastasis-related genes, Cancer Lett. 321 (2) (2012) 195–202.
[49] D.N. Reisman, J. Sciarrotta, W. Wang, W.K. Funkhouser, B.E. Weissman, Loss

of BRG1/BRM in human lung cancer cell lines and primary lung cancers: correlation with poor prognosis, Cancer Res. 63 (3) (2003) 560–566.
[50] W. Sun, K. Zhang, X. Zhang, W. Lei, T. Xiao, J. Ma, S. Guo, S. Shao, H. Zhang, Y. Liu, J. Yuan, Z. Hu, Y. Ma, X. Feng, S. Hu, J. Zhou, S. Cheng, Y. Gao, Identification of differentially expressed genes in human lung squamous cell carcinoma using suppression subtractive hybridization, Cancer Lett. 212 (1) (2004) 83–93.
[51] S. Wachi, K. Yoneda, R. Wu, Interactome-transcriptome analysis reveals the high centrality of genes differentially expressed in lung cancer tissues, Bioinformatics 21 (23) (2005) 4205–4208.
[52] A. Fuchs, K. Konig, L.C. Heukamp, J. Fassunke, J. Kirfel, S. Huss, A.J. Becker, R. Buttner, M. Majores, Tuberous-sclerosis complex-related cell signaling in the pathogenesis of lung cancer, Diagn. Pathol. 9 (2014) 48.
[53] J.S. Moreb, H.V. Baker, L.J. Chang, M. Amaya, M.C. Lopez, B. Ostmark, W. Chou, ALDH isozymes downregulation affects cell growth, cell motility and gene expression in lung cancer cells, Mol. Cancer 7 (2008) 87.
[54] S.Y. Eom, Y.W. Zhang, S.H. Kim, K.H. Choe, K.Y. Lee, J.D. Park, Y.C. Hong, Y. D. Kim, J.W. Kang, H. Kim, Influence of NQO1, ALDH2, and CYP2E1 genetic polymorphisms, smoking, and alcohol drinking on the risk of lung cancer in Koreans, Cancer Causes Control 20 (2) (2009) 137–145.
[55] G. Chen, S.H. Kim, A.N. King, L. Zhao, R.U. Simpson, P.J. Christensen, Z. Wang, D.G. Thomas, T.J. Giordano, L. Lin, D.E. Brenner, D.G. Beer, N. Ramnath, CYP24A1 is an independent prognostic marker of survival in patients with lung adenocarcinoma, Clin. Cancer Res. 17 (4) (2011) 817–826.
[56] T. Kageyama, R. Nagashio, S. Ryuge, T. Matsumoto, A. Iyoda, Y. Satoh, N. Masuda, S.X. Jiang, M. Saegusa, Y. Sato, HADHA is a potential predictor of response to platinum-based chemotherapy for lung cancer, Asian Pac. J. Cancer Prev. 12 (12) (2011) 3457–3463.
[57] S.B. Lee, J.N. Ho, S.H. Yoon, G.Y. Kang, S.G. Hwang, H.D. Um, Peroxiredoxin 6 promotes lung cancer cell invasion by inducing urokinase-type plasminogen activator via p38 kinase, phosphoinositide 3-kinase, and Akt, Mol. Cells 28 (6) (2009) 583–588.
[58] D. Liu, W.M. Li, X.M. Mo, L.X. Liu, Y. Wang, G.W. Che, Z. Wu, J.L. Gou, [Multiparametric flow cytometry analyzes the expressions of immunophenotype CD133, CD34, CD44 in lung cancer naive cells], Sichuan Da Xue Xue Bao Yi Xue Ban. 39 (5) (2008) 827–831.
[59] X. Zhao, X. Yan, Y.F. Ju, H.M. Yu, S.C. Jiao, [Expression and clinical significance of CD3, CD4 and COX-2 in non-small cell lung cancer], Xi Bao Yu Fen Zi Mian Yi Xue Za Zhi 28 (4) (2012) 407–409.
[60] D. Campa, S. Zienolddiny, V. Maggini, V. Skaug, A. Haugen, F. Canzian, Association of a common polymorphism in the cyclooxygenase 2 gene with risk of non-small cell lung cancer, Carcinogenesis 25 (2) (2004) 229–235.
[61] M. Orditura, F. De Vita, G. Catalano, S. Infusino, E. Lieto, E. Martinelli, F. Morgillo, P. Castellano, C. Pignatelli, G. Galizia, Elevated serum levels of interleukin-8 in advanced non-small cell lung cancer patients: relationship with prognosis, J. Interferon Cytokine Res. 22 (11) (2002) 1129–1135.
[62] Z. Zhang, S. Guo, X. Liu, X. Gao, Synergistic antitumor effect of alpha-pinene and beta-pinene with paclitaxel against non-small-cell lung carcinoma (NSCLC), Drug Res. 65 (4) (2015) 214–218.
[63] T.J. Raphael, G. Kuttan, Effect of naturally occurring monoterpenes carvone, limonene and perillic acid in the inhibition of experimental lung metastasis induced by B16F-10 melanoma cells, J. Exp. Clin. Cancer Res. 22 (3) (2003) 419–424.
[64] R. Berni Canani, M. Di Costanzo, L. Leone, The epigenetic effects of butyrate: potential therapeutic implications for clinical practice, Clin. Epigenetics 4 (1) (2012) 4.
[65] L. Li, Y. Wei, C. To, C.Q. Zhu, J. Tong, N.A. Pham, P. Taylor, V. Ignatchenko, A. Ignatchenko, W. Zhang, D. Wang, N. Yanagawa, M. Li, M. Pintilie, G. Liu, L. Muthuswamy, F.A. Shepherd, M.S. Tsao, T. Kislinger, M.F. Moran, Integrated omic analysis of lung cancer reveals metabolism proteome signatures with prognostic impact, Nat. Commun. 5 (2014) 5469.
[66] K. Gu, M.M. Li, J. Shen, F. Liu, J.Y. Cao, S. Jin, Y. Yu, Interleukin-17-induced EMT promotes lung cancer cell migration and invasion via NF-kappaB/ZEB1 signal pathway, Am. J. Cancer Res. 5 (3) (2015) 1169–1179.
[67] H. Nakayama, H. Enzan, E. Miyazaki, N. Kuroda, K. Naruse, H. Kiyoku, M. Toi, M. Hiroi, CD34 positive stromal cells in gastric adenocarcinomas, J. Clin. Pathol. 54 (11) (2001) 846–848.
[68] F. Murai, D. Koinuma, A. Shinozaki-Ushiku, M. Fukayama, K. Miyaozono, S. Ehata, EZH2 promotes progression of small cell lung cancer by suppressing the TGF-β-Smad-ASCL1 pathway, Cell Discov. 1 (2015) 15026.
[69] T. Sato, A. Kaneda, S. Tsuji, T. Isagawa, S. Yamamoto, T. Fujita, R. Yamanaka, Y. Tanaka, T. Nukiwa, V.E. Marquez, Y. Ishikawa, M. Ichinose, H. Aburatani, PRC2 overexpression and PRC2-target gene repression relating to poorer prognosis in small cell lung cancer, Sci. Rep. 3 (2013) 1911.
[70] M. Serresi, G. Gargiulo, N. Proost, B. Siteur, M. Cesaroni, M. Koppens, H. Xie, Sutherland, D. Kate, D. Hulsman, E. Citterio, S. Orkin, A. Berns, M. van Lohuizen, Polycomb repressive complex 2 is a barrier to KRAS-driven inflammation and epithelial-mesenchymal transition in non-small-cell lung cancer, Cancer Cell 29 (1) (2016) 17–31.
[71] C. Liu, X. Shi, L. Wang, Y. Wu, F. Jin, C. Bai, Y. Song, SUZ12 is involved in progression of non-small cell lung cancer by promoting cell proliferation and metastasis, Tumour Biol. 35 (6) (2014) 6073–6082.
[72] C.L. Wang, C.I. Wang, P.C. Liao, C.D. Chen, Y. Liang, W.Y. Chuang, Y.H. Tsai, H. C. Chen, Y.S. Chang, J.S. Yu, C.C. Wu, C.J. Yu, Discovery of retinoblastoma-associated binding protein 46 as a novel prognostic marker for distant metastasis in nonsmall cell lung cancer by combined analysis of cancer cell secretome and pleural effusion proteome, J. Proteome Res. 8 (10) (2009) 4428–4440.
[73] S. Wu, V. Kasim, M.R. Kano, S. Tanaka, S. Ohba, Y. Miura, K. Miyata, X. Liu, A. Matsuhashi, U.I. Chung, L. Yang, K. Kataoka, N. Nishiyama, M. Miyagishi, Transcription factor YY1 contributes to tumor growth by stabilizing hypoxia factor HIF-1alpha in a p53-independent manner, Cancer Res. 73 (6) (2013) 1787–1799.
[74] P. Yang, S.J. Mandrekar, S.H. Hillman, K.L. Allen Ziegler, Z. Sun, J.A. Wampfler, J.M. Cunningham, J.A. Sloan, A.A. Adjei, E. Perez, J.R. Jett, Evaluation of glutathione metabolic genes on outcomes in advanced non-small cell lung cancer patients after initial treatment with platinum-based chemotherapy: an NCCTG-97-24-51 based study, J. Thorac. Oncol. 4 (4) (2009) 479–485.
[75] Y. Inoue, M. Tomisawa, H. Yamazaki, Y. Abe, H. Suemizu, H. Tsukamoto, Y. Tomii, M. Kawamura, H. Kijima, H. Hatanaka, Y. Ueyama, M. Nakamura, K. Kobayashi, The modifier subunit of glutamate cysteine ligase (GCLM) is a molecular target for amelioration of cisplatin resistance in lung cancer, Int J. Oncol. 23 (5) (2003) 1333–1339.
[76] A.P. van den Heuvel, J. Jing, R.F. Wooster, K.E. Bachman, Analysis of glutamine dependency in non-small cell lung cancer: GLS1 splice variant GAC is essential for cancer cell growth, Cancer Biol. Ther. 13 (12) (2012) 1185–1194.
[77] T.C. Allen, L.A. Granville, P.T. Cagle, A. Haque, D.S. Zander, R. Barrios, Expression of glutathione S-transferase pi and glutathione synthase correlates with survival in early stage non-small cell carcinomas of the lung, Hum. Pathol. 38 (2) (2007) 220–227.
[78] W.A. Cooper, M.R. Kohonen-Corish, B. McCaughan, C. Kennedy, R. L. Sutherland, C.S. Lee, Expression and prognostic significance of cyclin B1 and cyclin A in non-small cell lung cancer, Histopathology 55 (1) (2009) 28–36.
[79] A. Wu, B. Wu, J. Guo, W. Luo, D. Wu, H. Yang, Y. Zhen, X. Yu, H. Wang, Y. Zhou, Z. Liu, W. Fang, Z. Yang, Elevated expression of CDK4 in lung cancer, J. Transl. Med. 9 (2011) 38.
[80] J.L. Liu, X.Y. Wang, B.X. Huang, F. Zhu, R.G. Zhang, G. Wu, Expression of CDK5/p35 in resected patients with non-small cell lung cancer: relation to prognosis, Med. Oncol. 28 (3) (2011) 673–678.
[81] X. Zhu, Y. Li, H. Shen, H. Li, L. Long, L. Hui, W. Xu, miR-137 inhibits the proliferation of lung cancer cells by targeting Cdc42 and Cdk6, FEBS Lett. 587 (1) (2013) 73–81.
[82] D.J. Li, G. Deng, Z.Q. Xiao, H.X. Yao, C. Li, F. Peng, M.Y. Li, P.F. Zhang, Y.H. Chen, Z.C. Chen, Identificating 14-3-3 sigma as a lymph node metastasis-related protein in human lung squamous carcinoma, Cancer Lett. 279 (1) (2009) 65–73.
[83] M. Blank, Y. Tang, M. Yamashita, S.S. Burkett, S.Y. Cheng, Y.E. Zhang, A tumor suppressor function of Smurf2 associated with controlling chromatin landscape and genome stability through RNF20, Nat. Med. 18 (2) (2012) 227–234.
[84] A. Mogi, H. Kuwano, TP53 mutations in nonsmall cell lung cancer, J. Biomed. Biotechnol. 2011 (2011) 583929.
[85] Y. Qin, H. Cui, H. Zhang, Overexpression of TRIM25 in lung cancer regulates tumor cell progression, Technol. Cancer Res Treat. (2015).
[86] A. Pataer, M.G. Raso, A.M. Correa, C. Behrens, K. Tsuta, L. Solis, B. Fang, J. A. Roth, Wistuba II, S.G. Swisher, Prognostic significance of RNA-dependent protein kinase on non-small cell lung cancer patients, Clin. Cancer Res. 16 (22) (2010) 5522–5528.
[87] Y. He, A.M. Correa, M.G. Raso, W.L. Hofstetter, B. Fang, C. Behrens, J.A. Roth, Y. Zhou, L. Yu, Wistuba II, S.G. Swisher, A. Pataer, The role of PKR/eIF2alpha signaling pathway in prognosis of non-small cell lung cancer, PloS One 6 (11) (2011) e24855.
[88] J. Zeng, D. Liu, Z. Qiu, Y. Huang, B. Chen, L. Wang, H. Xu, N. Huang, L. Liu, W. Li, GSK3beta overexpression indicates poor prognosis and its inhibition reduces cell proliferation and survival of non-small cell lung cancer cells, PloS One 9 (3) (2014) e91231.
[89] Y. Zhao, E.B. Butler, M. Tan, Targeting cellular metabolism to improve cancer therapeutics, Cell Death Dis. 4 (2013) e532.
[90] W. Rzeski, C. Ikonomidou, L. Turski, Glutamate antagonists limit tumor growth, Biochem. Pharmacol. 64 (8) (2002) 1195–1200.
[91] T.F. Burns, L.P. Stabile, Targeting the estrogen pathway for the treatment and prevention of lung cancer, Lung Cancer Manag. 3 (1) (2014) 43–52.
[92] S.M. Weakley, H. Wang, Q. Yao, C. Chen, Expression and function of a large non-coding RNA gene XIST in human cancer, World J. Surg. 35 (8) (2011) 1751–1756.
[93] B. Ji, K.K. Higa, J.R. Kelsoe, X. Zhou, Over-expression of XIST, the master gene for X chromosome inactivation, in females with major affective disorders, EBioMedicine (2015).
[94] C.M. Johnston, F.L. Lovell, D.A. Leongamornlert, B.E. Stranger, E. T. Dermitzakis, M.T. Ross, Large-scale population study of human cell lines indicates that dosage compensation is virtually complete, PLoS Genet. 4 (1) (2008) e9.
[95] O. Andres, T. Kellermann, F. Lopez-Giraldez, J. Rozas, X. Domingo-Roura, M. Bosch, RPS4Y gene family evolution in primates, BMC Evol. Biol. 8 (2008) 142.
[96] A.I. Shevchenko, I.S. Zakharova, S.M. Zakian, The evolutionary pathway of x chromosome inactivation in mammals, Acta Nat. 5 (2) (2013) 40–53.
[97] O.V. Leontieva, G.M. Paszkiewicz, M.V. Blagosklonny, Weekly administration of rapamycin improves survival and biomarkers in obese male mice on high-fat diet, Aging Cell. 13 (4) (2014) 616–622.
[98] K.A. Rodriguez, S.G. Dodds, R. Strong, V. Galvan, Z.D. Sharp, R. Buffenstein, Divergent tissue and sex effects of rapamycin on the proteasome-chaperone

network of old mice, Front. Mol. Neurosci. 7 (2014) 83.
[99] X. Wang, M. Huang, Y. Wang, The effect of insulin, TNFalpha and DHA on the proliferation, differentiation and lipolysis of preadipocytes isolated from large yellow croaker (*Pseudosciaena crocea* R.), PLoS One 7 (10) (2012) e48069.
[100] M. Xia, H. Land, Tumor suppressor p53 restricts Ras stimulation of RhoA and cancer cell motility, Nat. Struct. Mol. Biol. 14 (3) (2007) 215–223.
[101] S.Y. Hyun, Y.J. Jang, p53 activates G(1) checkpoint following DNA damage by doxorubicin during transient mitotic arrest, Oncotarget 6 (7) (2015) 4804–4815.
[102] N.S. Pellegata, R.J. Antoniono, J.L. Redpath, E.J. Stanbridge, DNA damage and p53-mediated cell cycle arrest: a reevaluation, Proc. Natl. Acad. Sci. USA 93 (26) (1996) 15209–15214.
[103] Y. Ogawara, S. Kishishita, T. Obata, Y. Isazawa, T. Suzuki, K. Tanaka, N. Masuyama, Y. Gotoh, Akt enhances Mdm2-mediated ubiquitination and degradation of p53, J. Biol. Chem. 277 (24) (2002) 21843–21850.
[104] J. Ma, Y. Xue, W. Cui, Y. Li, Q. Zhao, W. Ye, J. Zheng, Y. Cheng, Y. Ma, S. Li, T. Han, L. Miao, L. Yao, J. Zhang, W. Liu, Ras homolog gene family, member A promotes p53 degradation and vascular endothelial growth factor-dependent angiogenesis through an interaction with murine double minute 2 under hypoxic conditions, Cancer 118 (17) (2012) 4105–4116.
[105] M.E. Ewen, C.J. Oliver, H.K. Sluss, S.J. Miller, D.S. Peeper, p53-dependent repression of CDK4 translation in TGF-beta-induced G1 cell-cycle arrest, Genes Dev. 9 (2) (1995) 204–217.
[106] K. Lohr, C. Moritz, A. Contente, M. Dobbelstein, p21/CDKN1A mediates negative regulation of transcription by p53, J. Biol. Chem. 278 (35) (2003) 32507–32516.
[107] S. Polager, D. Ginsberg, p53 and E2f: partners in life and death, Nat. Rev. Cancer 9 (10) (2009) 738–748.
[108] P.M. Yang, W.C. Huang, Y.C. Lin, W.Y. Huang, H.A. Wu, W.L. Chen, Y.F. Chang, C. W. Chou, C.C. Tzeng, Y.L. Chen, C.C. Chen, Loss of IKKbeta activity increases p53 stability and p21 expression leading to cell cycle arrest and apoptosis, J. Cell Mol. Med. 14 (3) (2010) 687–698.
[109] E.E. Bosco, W. Ni, L. Wang, F. Guo, J.F. Johnson, Y. Zheng, Rac1 targeting suppresses p53 deficiency-mediated lymphomagenesis, Blood 115 (16) (2010) 3320–3328.
[110] G. Schneider, A. Henrich, G. Greiner, V. Wolf, A. Lovas, M. Wieczorek, T. Wagner, S. Reichardt, A. von Werder, R.M. Schmid, F. Weih, T. Heinzel, D. Saur, O.H. Kramer, Cross talk between stimulated NF-kappaB and the tumor suppressor p53, Oncogene 29 (19) (2010) 2795–2806.
[111] S.W. Hiebert, G. Packham, D.K. Strom, R. Haffner, M. Oren, G. Zambetti, J. L. Cleveland, E2F-1: DP-1 induces p53 and overrides survival factors to trigger apoptosis, Mol. Cell Biol. 15 (12) (1995) 6864–6874.
[112] Y.E. Whang, C. Tran, C. Henderson, R.G. Syljuasen, N. Rozengurt, W. H. McBride, C.L. Sawyers, c-Abl is required for development and optimal cell proliferation in the context of p53 deficiency, Proc. Natl. Acad. Sci. USA 97 (10) (2000) 5486–5491.
[113] L. Xie, C. Gazin, S.M. Park, L.J. Zhu, M.A. Debily, E.L. Kittler, M.L. Zapp, D. Lapointe, S. Gobeil, C.M. Virbasius, M.R. Green, A synthetic interaction screen identifies factors selectively required for proliferation and TERT transcription in p53-deficient human cancer cells, PLoS Genet. 8 (12) (2012) e1003151.
[114] K.R. Nevis, M. Cordeiro-Stone, J.G. Cook, Origin licensing and p53 status regulate Cdk2 activity during G(1), Cell Cycle 8 (12) (2009) 1952–1963.
[115] H.B. Rui, J.Z. Su, Co-transfection of p16(INK4a) and p53 genes into the K562 cell line inhibits cell proliferation, Haematologica 87 (2) (2002) 136–142.
[116] P. Kanellou, A. Zaravinos, M. Zioga, D.A. Spandidos, Deregulation of the tumour suppressor genes p14(ARF), p15(INK4b), p16(INK4a) and p53 in basal cell carcinoma, Br. J. Dermatol. 160 (6) (2009) 1215–1221.
[117] E. Goker, M. Waltham, A. Kheradpour, T. Trippett, M. Mazumdar, Y. Elisseyeff, B. Schnieders, P. Steinherz, C. Tan, E. Berman, et al., Amplification of the dihydrofolate reductase gene is a mechanism of acquired resistance to methotrexate in patients with acute lymphoblastic leukemia and is correlated with p53 gene mutations, Blood 86 (2) (1995) 677–684.
[118] L.J. Juan, W.J. Shia, M.H. Chen, W.M. Yang, E. Seto, Y.S. Lin, C.W. Wu, Histone deacetylases specifically down-regulate p53-dependent gene activation, J. Biol. Chem. 275 (27) (2000) 20436–20443.
[119] A. Mrozek, H. Petrowsky, I. Sturm, J. Kraus, S. Hermann, S. Hauptmann, M. Lorenz, B. Dorken, P.T. Daniel, Combined p53/Bax mutation results in extremely poor prognosis in gastric carcinoma with low microsatellite instability, Cell Death Differ. 10 (4) (2003) 461–467.
[120] M.T. Hemann, S.W. Lowe, The p53-Bcl-2 connection, Cell Death Differ. 13 (8) (2006) 1256–1259.
[121] J. Liu, H. Uematsu, N. Tsuchida, M.A. Ikeda, Essential role of caspase-8 in p53/p73-dependent apoptosis induced by etoposide in head and neck carcinoma cells, Mol. Cancer (2011) 10.
[122] R. Takimoto, W.S. El-Deiry, Wild-type p53 transactivates the KILLER/DR5 gene through an intronic sequence-specific DNA-binding site, Oncogene 19 (14) (2000) 1735–1743.
[123] M. Nagamine, T. Okumura, S. Tanno, M. Sawamukai, W. Motomura, N. Takahashi, Y. Kohgo, PPAR gamma ligand-induced apoptosis through a p53-dependent mechanism in human gastric cancer cells, Cancer Sci. 94 (4) (2003) 338–343.
[124] C. Tan, Y.T. Jin, H.Y. Xu, C.Y. Zhang, H. Zhang, W.M. Zhang, C.M. Chen, X.Y. Sun, [Correlation between RARbeta gene promoter methylation and P53 gene mutations in non-small cell lung cancer], Zhonghua Yi Xue Yi Chuan Xue Za Zhi 29 (2) (2012) 131–136.
[125] P. Stambolsky, Y. Tabach, G. Fontemaggi, L. Weisz, R. Maor-Aloni, Z. Siegfried, I. Shiff, I. Kogan, M. Shay, E. Kalo, G. Blandino, I. Simon, M. Oren, V. Rotter, Modulation of the vitamin D3 response by cancer-associated mutant p53, Cancer Cell. 17 (3) (2010) 273–285.
[126] H. Song, M. Hollstein, Y. Xu, p53 gain-of-function cancer mutants induce genetic instability by inactivating ATM, Nat. Cell Biol. 9 (5) (2007) 573–580.
[127] K. Wu, S.W. Jiang, F.J. Couch, p53 mediates repression of the BRCA2 promoter and down-regulation of BRCA2 mRNA and protein levels in response to DNA damage, J. Biol. Chem. 278 (18) (2003) 15652–15660.
[128] V. Gottifredi, O. Karni-Schmidt, S.S. Shieh, C. Prives, p53 down-regulates CHK1 through p21 and the retinoblastoma protein, Mol. Cell Biol. 21 (4) (2001) 1066–1076.
[129] T. Watanabe, S. Nobusawa, S. Lu, J. Huang, M. Mittelbronn, H. Ohgaki, Mutational inactivation of the nijmegen breakage syndrome gene (NBS1) in glioblastomas is associated with multiple TP53 mutations, J. Neuropathol. Exp. Neurol. 68 (2) (2009) 210–215.
[130] P.A. Muller, K.H. Vousden, p53 mutations in cancer, Nat. Cell Biol. 15 (1) (2013) 2–8.
[131] P.A. Muller, K.H. Vousden, J.C. Norman, p53 and its mutants in tumor cell migration and invasion, J. Cell Biol. 192 (2) (2011) 209–218.
[132] I. Goldstein, O. Ezra, N. Rivlin, A. Molchadsky, S. Madar, N. Goldfinger, V. Rotter, p53, a novel regulator of lipid metabolism pathways, J. Hepatol. 56 (3) (2012) 656–662.
[133] D.F. Romagnolo, J. Zempleni, O.I. Selmin, Nuclear receptors and epigenetic regulation: opportunities for nutritional targeting and disease prevention, Adv. Nutr. 5 (4) (2014) 373–385.
[134] P. Arizti, L. Fang, I. Park, Y. Yin, E. Solomon, T. Ouchi, S.A. Aaronson, S.W. Lee, Tumor suppressor p53 is required to modulate BRCA1 expression, Mol. Cell Biol. 20 (20) (2000) 7450–7459.
[135] M.E. Moynahan, The cancer connection: BRCA1 and BRCA2 tumor suppression in mice and humans, Oncogene 21 (58) (2002) 8994–9007.
[136] L. Snijders Blok, E. Madsen, J. Juusola, C. Gilissen, D. Baralle, M.R. Reijnders, H. Venselaar, C. Helsmoortel, M.T. Cho, A. Hoischen, L.E. Vissers, T.S. Koemans, W. Wissink-Lindhout, E.E. Eichler, C. Romano, H. Van Esch, C. Stumpel, M. Vreeburg, E. Smeets, K. Oberndorff, B.W. van Bon, M. Shaw, J. Gecz, E. Haan, M. Bienek, C. Jensen, B.L. Loeys, A. Van Dijck, A.M. Innes, H. Racher, S. Vermeer, N. Di Donato, A. Rump, K. Tatton-Brown, M.J. Parker, A. Henderson, S.A. Lynch, A. Fryer, A. Ross, P. Vasudevan, U. Kini, R. Newbury-Ecob, K. Chandler, A. Male, D.D.D. Study, S. Dijkstra, J. Schieving, J. Giltay, K.L. van Gassen, J. Schuurs-Hoeijmakers, P.L. Tan, I. Pediaditakis, S.A. Haas, K. Retterer, P. Reed, K. G. Monaghan, E. Haverfield, M. Natowicz, A. Myers, M.C. Kruer, Q. Stein, K. A. Strauss, K.W. Brigatti, K. Keating, B.K. Burton, K.H. Kim, J. Charrow, J. Norman, A. Foster-Barber, A.D. Kline, A. Kimball, E. Zackai, M. Harr, J. Fox, J. McLaughlin, K. Lindstrom, K.M. Haude, K. van Roozendaal, H. Brunner, W.K. Chung, R. F. Kooy, R. Pfundt, V. Kalscheuer, S.G. Mehta, N. Katsanis, T. Kleefstra, Mutations in DDX3X are a common cause of unexplained intellectual disability with gender-specific effects on Wnt signaling, Am. J. Hum. Genet. 97 (2) (2015) 343–352.
[137] D. Gurgen, A. Kusch, R. Klewitz, U. Hoff, R. Catar, B. Hegner, U. Kintscher, F. C. Luft, D. Dragun, Sex-specific mTOR signaling determines sexual dimorphism in myocardial adaptation in normotensive DOCA-salt model, Hypertension 61 (3) (2013) 730–736.

Weng Howe Chan is currently a Ph.D. student of Computer Science in Universiti Teknologi Malaysia. He received his B.Sc. in Computer Science from Universiti Teknologi Malaysia in 2011. His research interests include classification algorithm and computational biology.

Mohd Saberi Mohamad is currently an Associate Professor at the Faculty of Computing, Universiti Teknologi Malaysia. He received the B.Sc. and M.Sc. degrees in Computer Science from Universiti Teknologi Malaysia in 2002 and 2005, respectively. He received the Ph.D. degree in Intelligent Systems for Bioinformatics from Osaka Prefecture University in 2010. He has published more than 230 publications in the field of bioinformatics using computational intelligence approaches. His interests are computational methods such as particle swarm optimization, hybrid approaches, genetic algorithms, support vector machines and neural network.

Safaai Deris is currently a Professor at Faculty of Creative Technology and Heritage, Universiti Malaysia Kelantan. He received the ME degree in Industrial Engineering and the Doctor of Engineering degree in Computer and System Sciences, both from Osaka Prefecture University, Japan, in 1989 and 1997, respectively. His recent academic interests include the application of intelligent techniques in scheduling and bioinformatics.

Nazar Zaki is an Associate Professor and coordinator of two tracks namely Intelligent Systems and Software Development. His research interest is in the field of bioinformatics, data mining and machine learning. He mainly focuses on developing an intelligent data mining algorithms to solve specific biological problems, such as protein function/structure prediction, protein interaction network analysis and protein complex detection.

Shahreen Kasim is a Senior Lecturer at the Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia. She received the B.Sc., M.Sc. and Ph.D. degrees in Computer Science from the Universiti Teknologi Malaysia in 2003, 2005 and 2011, respectively. Her research interests focus on gene function prediction, clustering algorithm and computational biology.

Sigeru Omatu has been a Professor at Osaka Institute of Technology since 2010. He received his Ph.D. in Electronic Engineering from Osaka Prefecture University. He was a Professor at the University of Tokushima in 1988 and Osaka Prefecture University in 1995. His research area is intelligent signal processing based on neural networks.

Juan Manuel Corchado is a full Professor with chair at the University of Salamanca. He holds a Ph.D. in Computer Sciences from the University of Salamanca and a Ph.D. in Artificial Intelligence from the University of the West of Scotland. His interests are in areas of information fusion, intelligent distributed systems, ambient intelligence and sensors.

Hany Al Ashwal joined UAEU in 2105 as Assistant Professor in the College of Information Technology. His research interests focus on gene expression data analysis and bioinformatics applications in Biomedical and Pharmaceutical Sciences. Research interests also include artificial intelligence, machine learning, genome wide analysis, and computational biology.