Molecular BioSystems

PAPER

View Article Online
View Journal | View Issue

Cite this: *Mol. BioSyst.*, 2015, 11, 791

Classification of lung cancer using ensemble-based feature selection and machine learning methods†

Zhihua Cai,[a] Dong Xu,[b] Qing Zhang,[c] Jiexia Zhang,*[d] Sai-Ming Ngai*[c] and Jianlin Shao*[e]

Lung cancer is one of the leading causes of death worldwide. There are three major types of lung cancers, non-small cell lung cancer (NSCLC), small cell lung cancer (SCLC) and carcinoid. NSCLC is further classified into lung adenocarcinoma (LADC), squamous cell lung cancer (SQCLC) as well as large cell lung cancer. Many previous studies demonstrated that DNA methylation has emerged as potential lung cancer-specific biomarkers. However, whether there exists a set of DNA methylation markers simultaneously distinguishing such three types of lung cancers remains elusive. In the present study, ROC (Receiving Operating Curve), RFs (Random Forests) and mRMR (Maximum Relevancy and Minimum Redundancy) were proposed to capture the unbiased, informative as well as compact molecular signatures followed by machine learning methods to classify LADC, SQCLC and SCLC. As a result, a panel of 16 DNA methylation markers exhibits an ideal classification power with an accuracy of 86.54%, 84.6% and a recall 84.37%, 85.5% in the leave-one-out cross-validation (LOOCV) and independent data set test experiments, respectively. Besides, comparison results indicate that ensemble-based feature selection methods outperform individual ones when combined with the incremental feature selection (IFS) strategy in terms of the informative and compact property of features. Taken together, results obtained suggest the effectiveness of the ensemble-based feature selection approach and the possible existence of a common panel of DNA methylation markers among such three types of lung cancer tissue, which would facilitate clinical diagnosis and treatment.

Received 9th November 2014,
Accepted 5th December 2014

DOI: 10.1039/c4mb00659c

www.rsc.org/molecularbiosystems

Introduction

Lung cancer is one of the leading causes of cancer-related deaths worldwide, developing in more than a million new patients annually.[1] There are three major types of lung cancers, non-small cell lung cancer (NSCLC), small cell lung cancer (SCLC) and carcinoid. NSCLC is further classified into lung adenocarcinoma (LADC), squamous cell lung cancer (SQCLC) as well as large cell lung cancer (LCLC) (http://www.cancer.org/Cancer/LungCancer-Non-SmallCell/DetailedGuide/non-small-cell-lung-cancer-what-is-non-small-cell-lung-cancer). NSCLC accounts for about 85% of all lung cancers with LADC and SQCLC representing almost 50% and 35% of NSCLC cases, respectively. About 10–15% of lung cancers are small cell lung cancers.[2–4] Accurate classification of lung cancer is the initial and significant step for the targeting therapy and clinical management since different treatment modalities exist. For example, Bevacizumab is not only less effective in the treatment of SQCLCs than LADCs, but also tends to contribute to mortality due to fatal hemoptysis.[5,6] Therefore, distinct subtypes of NSCLC should not be deemed as a single group clinically and it is increasingly acknowledged that such subtypes should be tackled as different diseases.[7]

Traditionally, the diagnosis of lung cancer is primarily based on the histology, with the use of immunohistochemical assays to confirm difficult cases. Whereas immunohistochemistry has exhibited an improved accuracy in the subclassification of lung cancers to some degree, it still presents a challenge in terms of effective treatment and prognosis due to interobserver variability among pathologists[8] or the variable sensitivity and specificity of

[a] *Affiliated Cancer Hospital of Guangzhou Medical University, Guangzhou, Guangdong Province, China*
[b] *Department of Mathematics & Scientific Computing Key Laboratory of Shanghai Universities, Shanghai Normal University, Shanghai, China*
[c] *School of Life Sciences and State (China) Key Laboratory of Agrobiotechnology, The Chinese University of Hong Kong, Hong Kong, China. E-mail: smngai@cuhk.edu.hk*
[d] *China State Key Laboratory of Respiratory Disease and Guangzhou Institute of Respiratory Disease, The First Affiliated Hospital of Guangzhou Medical University, Guangzhou, Guangdong Province, China. E-mail: drzjxcn@126.com*
[e] *First Affiliated Hospital, School of Medicine, Zhejiang University, Hangzhou, Zhejiang Province, China. E-mail: jianlin.shao@gmail.com*

† Electronic supplementary information (ESI) available. See DOI: 10.1039/c4mb00659c

individual markers.[9,10] Thus, more robust, specific molecular signatures are required for the purpose of clearer discrimination among lung cancer cases. Recently, a few studies have been performed to uncover molecular signatures/biomarkers for classification of lung cancers into specific subtypes such as specific microRNAs, genomic mutations and copy number alternations (CNAs). Lebanony *et al.*[11] identified has-miR-205 expression as a specific marker to distinguish SQCLCs from non-squamous NSCLCs. Bishop *et al.*[9] developed a score system based on hsa-miR-205, has-miR-21 and U6snR to classify NSCLCs cases into SQCLCs or LADCs. Another study made attempt to identify signature genes of three subhistological types of NSCLC, *i.e.* LADC, SQCLC as well as LCLC[12] based on the gene expression profiling data sets. Most recently, the information of 266 CNA probes was utilized to distinguish the LADC from SQCLC through Maximum Relevance Minimum Redundancy (mRMR) feature selection combined with the nearest neighbor algorithm.[13] The above-mentioned studies have been done with emphasis on the classification of subtypes in NSCLCs. In addition, two studies have been carried out regarding the molecular classification of subtypes of NSCLC and SCLC. Seidel and coauthors presented data from two cohorts totaling more than 6000 lung cancer patients, characterizing genome alternations in 1255 clinically annotated lung tumors of all histological subgroups to identify genetically defined and clinically relevant subtypes.[14] Due to the fact that marked distinction of genomic alterations existed between and within histological subtypes, they devised a statistic model for robust prediction of lung cancer subtypes based on such alterations including SQCLC, LADC, LCLC, SCLC and carcinoid. By the extension of the investigation done by Lebanony *et al.*,[11] Gilad and coworkers employed K-nearest neighbor (KNN) classifier with Pearson correlation distance metric to discriminate four types of lung cancers, including squamous NSCLC, nonsquamous NSCLC, SCLC as well as carcinoid based on selected eight-microRNA diagnosis assay from 110 array probes(109 microRNA probes and a probe for the small RNA U6).[15] Whereas these two studies made an effort to differentiate major types of NSCLC and SCLC based on either whole genomic alterations or a panel of microRNAs, the discrimination of different types of lung cancers still reaches incompletion.

Many previous studies demonstrated that DNA methylation has emerged as a potential lung cancer-specific biomarker.[16,17] A Prognostic DNA Methylation Signature for Stage I NSCLC was identified based on methylation profiling of a large cohort of NSCLC patients with normal lung tissues as control.[18] Two other studies endeavored to experimentally discover sensitive and specific DNA methylation markers to distinguish LADC/SQCLC from normal lung tissue, respectively.[19,20] Another study[21] proposed to use artificial neural networks (ANN) and linear discriminant analysis (LDA) to classify the cell lines into SCLC or into NSCLC, concluding that ANN models based on DNA methylation profiles can objectively classify SCLC and NSCLC cells lines with substantial to perfect concordance. As a part of The Cancer Genome Atlas Research Network, promoter methylome for 178 histopathologically reviewed SQCLCs was characterized.[22] However, given the methylome availability of lung cancers, whether there exists a panel of DNA methylation markers to simultaneously discriminate LADC, SQCLC and SCLC remains unknown.

To this end, we attempted to discover a panel of DNA methylation markers through constructing a multiclass classification models for accurate characterization of the above-mentioned three types of lung cancers. Feature selection is one of the important steps for classification modeling. Many types of feature selection methods were proposed based on machine learning framework or information theory framework. For example, Fernandez-Lozano *et al.* employed Support Vector Machine Recursive Feature Elimination(SVM-RFE) to classify enzyme regulatory proteins or predict transport proteins.[23,24] Li *et al.* proposed the mRMR feature selection approach to predict protein cleavage sites or protein domains.[25,26] In the present work, methylome-wide ranking and screening of DNA methylation markers (probe sets from array experiments) was performed using ensemble-based feature extraction methods, which incorporates Multi-category Receiver Operating Characteristic (Multi-ROC), Random Forests (RFs) as well as Maximum Relevance and Minimum Redundancy (mRMR) methods. The final panel of DNA methylation markers was further determined by comprehensive performance evaluation of multiclass support vector machine classifier trained with the Incremental Feature Selection (IFS) strategy. The resulting classification model demonstrates its ability to accurately differentiate LADC, SQCLC as well as SCLC, suggesting the existence of a common panel of DNA methylation markers among such three types of lung cancers.

Materials and methods

Data collection and preprocessing

The DNA methylome data sets for construction of the classification model used in the present study originated from two sources, the Cancer Genome Atlas (TCGC) and Gene Expression Omnibus (GEO),[27] all of which were produced based on Illumina HumanMethylation27 array and represented as beta values. The methylomes of LADC and SQCLC were downloaded from TCGC, which are composed of 141 and 162 samples, respectively. The methylomes of 28 SCLC cases were retrieved from GEO with accession number GSE50412. Due to the presence of missing values in the DNA methylomes from TCGC, we first performed the removal of probes with missing values for methylation array data sets of each sample. Common probe sets for each sample among LADC, SQCLC and SCLC were then retained for further analysis. Thus, the final numbers of LADC, SQCLC and SCLC used for classification model training are 126, 134 and 28 (Table 1), respectively, the ratio of which from LADC, SQCLC and SCLC is approximately the same as the clinical estimation ($\sim$80–85% as NSCLC and $\sim$10–15% as SCLC). To evaluate the performance of the classification model, we obtained an independent cohort only consisting of 454 LADCs, 401 SQCLCs from TCGA (due to the scarcity of SCLC data sets), the methylome of which was assayed on the Illumina HumanMethylation450 array platform (including all of probes from Illumina HumanMethylation27 array).

Table 1 Summary of high-throughput data sets used in this study

Platform	Roles of datasets	Sample types	Number of samples
Illumina HumanMethylation27	Training/cross validation data sets	LADC	126
		SQCLC	134
		SCLC	28
Illumina HumanMethylation450	Independent test data sets	LADC	452
		SQCLC	359

We extracted DNA methylation data of probe sets common to those utilized in model construction for the purpose of independent validation of the final classification model. After removal of probes and samples with missing values, the final numbers of LADCs and SQCLCs are 452 and 359, respectively (Table 1). All of the data sets were processed by in-house Java scripts.

Multi-category receiver operating characteristic (Multi-ROC)

Receiver Operating Characteristic (ROC) analysis has been extensively employed in diagnostic, prognostic and predictive biomarker research and the area under ROC curve (AUC) is typically calculated as measurement for the assessment of the differentiating ability of some biomarker(s) for binary classification problems.[28] With respect to multi-category classification problems, Scurfield proposed the concept of Multi-ROC analysis and hypervolume under the manifold (HUM) as measurement to evaluate the discriminative ability of corresponding biomarker(s).[29,30] To select a panel of DNA methylation markers (*i.e.* probes, which we hereafter refer to features), the discriminative ability of features with respect to classification of LADC, SQCLC, and SCLC was ranked through Multi-ROC analysis and the HUM measurement.

As we know, in the case of the binary classification problem, AUC can be represented as:

$$\mathrm{AUC} = \int_0^1 \mathrm{ROC}(u)\mathrm{d}u$$

where ROC(u) refers to the function expressing ROC curve, $u \in [0,1]^1$.

With respect to M-class classification problems, Li *et al.*[31,32] made the extension of the ROC curve as ROC "surface", which is an $(M-1)$-dimensional manifold and the definition of HUM can be expressed as an $(M-1)$ integral of ROC "surface":

$$\mathrm{HUM} = \int_0^1 \cdots \int_0^1 \mathrm{ROC}(u)\mathrm{d}u_1 \ldots \mathrm{d}u_{M-1}$$

where ROC(u) denotes the function expressing ROC surface, $u \in [0,1]^{M-1}$ and M corresponds to the number of category. HUM = 1 means the perfect discrimination ability of the classifier and $\mathrm{HUM} = \frac{1}{M!}$ suggests that the distinguishing ability of the corresponding classifier is equivalent to that by chance. For instance, in the case of $M = 2$, the AUC = 1/2! = 0.5 is non-informative and for $M = 3$ (in the present study), the non-informative HUM is 1/3! = 0.1667.The larger the HUM value is, the more accurate classification probability of the classifier is.

In the present study, the HUM value for each feature was estimated by the approach proposed by Li *et al.*[33] R package HUM was used for the implementation.

Maximum relevance and minimum redundancy (mRMR)

In addition to rank the discriminative ability of each feature by Multi-ROC, we also ranked the features according to their relevance to target classes through the mRMR strategy,[34] which has been successfully applied to many classification problems.[25,26,35–37] Briefly, the dependency $I(x, y)$ between any two random variables x and y is defined as:

$$I(x,y) = \iint p(x,y)\log\frac{p(x,y)}{p(x)p(y)}\mathrm{d}x\mathrm{d}y$$

where $p(x)$ and $p(y)$ are the probability density functions of x and y, respectively. $p(x, y)$ is the joint probability density function of x and y. Thus, the relevance of feature x and target class $c = (c_1,c_2,\ldots,c_k)$ can be represented as $D = I(x, c)$, the maximum of which is utilized to rank features for classification. To reduce the redundancy due to the correlations among ranked features by the maximum relevance strategy, the average mutual information $R = \frac{1}{m}\sum_{j=1}^{m} I(x, x_j)$ between candidate feature x and ranked features x_j ($i = 1,\ldots,m$) needs to be minimized (maximum Relevance Minimum Redundancy (mRMR)). In practice, incremental search methods can be used to rank the features through the following optimization problem based on the mRMR principle:

$$\max_{x\in C}\left(I(x,c) - \frac{1}{m}\sum_{j=1}^{m} I(x,x_j)\right)$$

where C is a set of unranked/candidate features, $\{x_1,\ldots,x_m\}$ ($m \geq 1$) refers to a set of ranked features. mRMR was implemented by the software downloaded at http://penglab.janelia.org/proj/mRMR/.

Random forests

Random Forests (RFs) is an ensemble learning method for classification and regression,[38] which removes the features with less contribution to classification accuracy by introducing random variables for the competition. RFs use an ensemble of unpruned decision trees, each of which is built from a bootstrap sample of the training data using a randomly selected subset of variables. The trained random forest classifier provides an importance estimate for all features,[39] the merit of which is suitable to perform feature selection for classification modeling.

The R package implementation of Boruta was employed in the study.

Multi-class support vector machines (multi-SVMs) and performance assessment

In the present study, multi-class support vector machines were employed to construct the classifier for the differentiation of LADC, SQCLC and SCLC. The classifier was implemented through LIBSVM[40] with a one-against-one strategy.[41] To evaluate the performance of the classifier, we adopted the LOOCV training scheme, which takes out one sample from the entire training data sets for test and the remaining samples for training in each of N rounds (N is the number of entire training data sets). The measurements of performance assessment for multi-class classifier proposed by Sokolova *et al.*[42] were used and defined as follows:

$$\text{average Accuracy (aAcc)} = \frac{\sum_{i=1}^{l} \frac{tp_i + tn_i}{tp_i + fn_i + fp_i + tn_i}}{l}$$

$$\text{precision} = \frac{\sum_{i=1}^{l} \frac{tp_i}{tp_i + fp_i}}{l}$$

$$\text{recall} = \frac{\sum_{i=1}^{l} \frac{tp_i}{tp_i + fn_i}}{l}$$

$$F\text{-score} = \frac{2 \times \text{precison} \times \text{recall}}{\text{precision} + \text{recall}}$$

where tp_i, fn_i, fp_i, tn_i are counts of true positive, false negative, false positive and true negative for class $i(i = 1,\ldots,l)$, respectively.

Enrichment analysis

GO and pathway enrichment analysis of genes corresponding to features/probes were performed through DAVID.[43] Statistical significance was determined by hypergeometric analysis followed by Bonferroni and Benjamini correction.[44]

Results and discussion

Feature selection results

The training data sets and independent test data sets used in this study were summarized and are listed in Table 1. The systematic batch effects among different microarray data sets used in the present study were adjusted by the DWD method.[45] To obtain an unbiased and compact set of features, three selection methods were proposed to rank the features. mRMR ranked all the features related to the types of lung cancers according to the criteria of maximum relevance and minimum redundancy. Multi-ROC ranked all the features based on the HUM value of each feature. RFs ranked all the features through assessment of their importance for the classification of different types of lung cancers. The overlapped features chosen by three approaches were utilized to develop the final classification model. Due to the fact that only 140 top features were evaluated as important features to discriminate three types of lung cancers through RFs method, we chose 200 top features obtained by mRMR and Multi-ROC methods as a candidate feature list, respectively (the reason why we selected a little larger arbitrary number 200 is that it would facilitate to make a fair performance comparison among individual methods with the IFS strategy). Thus, the intersection of three top-ranked feature lists is composed of 45 features. The detailed feature selected is listed in ESI,† Table S1A.

Performance of the classification model and comparison results between ensemble-based and individual feature selection methods

To initially check the discriminative ability of selectively overlapped features, we performed the unsupervised hierarchical clustering of all the 288 samples to make an attempt to discover the methylation pattern of 45 features. As the heatmap shown in Fig. 1, most of the three types of lung cancer samples can be separated into their correct types. The final optimal set of features was selected from ranked 45 features by the Incremental Feature Selection (IFS) strategy.[25,26,34] With regard to the IFS strategy, ranked features $(f_1,f_2,\ldots,f_N)$ are added to new data sets one by one from the higher to lower rank. A new feature subset is produced when one feature is added. Therefore, there would be N feature sets produced from the ranked feature list $(f_1,f_2,\ldots,f_N)$ and the i-th feature set is $s_i = \{f_1,f_2,\ldots,f_i\}$ $(1 \le i \le N)$. For each of the N feature sets, multi-class SVMs predictor was developed and the total number of individual predictors is N. To achieve an unbiased assessment of the performance of the prediction model, the following three cross-validation methods are often utilized because of their effectiveness in practical applications: independent dataset test, subsampling test, and leave-one-out cross-validation (LOOCV). However, of the three test methods, the LOOCV is deemed as the least arbitrary that can always yield a unique result for a given benchmark dataset as elaborated in a previous work.[46] Accordingly, LOOCV has been increasingly used by investigators to examine the quality of various predictors.[47–50] Therefore, LOOCV was employed to evaluate each classification model on each subset of features for the purpose of determination of the optimal feature set as well as the final classification model. The IFS performance curve for ranked 45 features was plotted (Fig. 2A). As shown in Fig. 2A, the best performance of IFS prediction models is achieved with a maximum F-score of 0.7445 when the first top 16 features were selected (the IFS classification performance for each subset of features is listed in ESI,† Table S1A). Therefore, such 16 top ranked features were utilized as the final optimal set of features for prediction model development. To further assess the robustness of the classification model based on the optimal set of features, the final model was tested on independent test datasets (Table 1), which achieves similar performance to that in the LOOCV experiment with 84.6% accuracy. The detailed LOOCV performance information on training data sets and the performance of the final prediction model on independent test data sets are shown in Table 2 and the confusion matrices of training datasets and independent datasets are listed in ESI,† Table S2. Whereas the

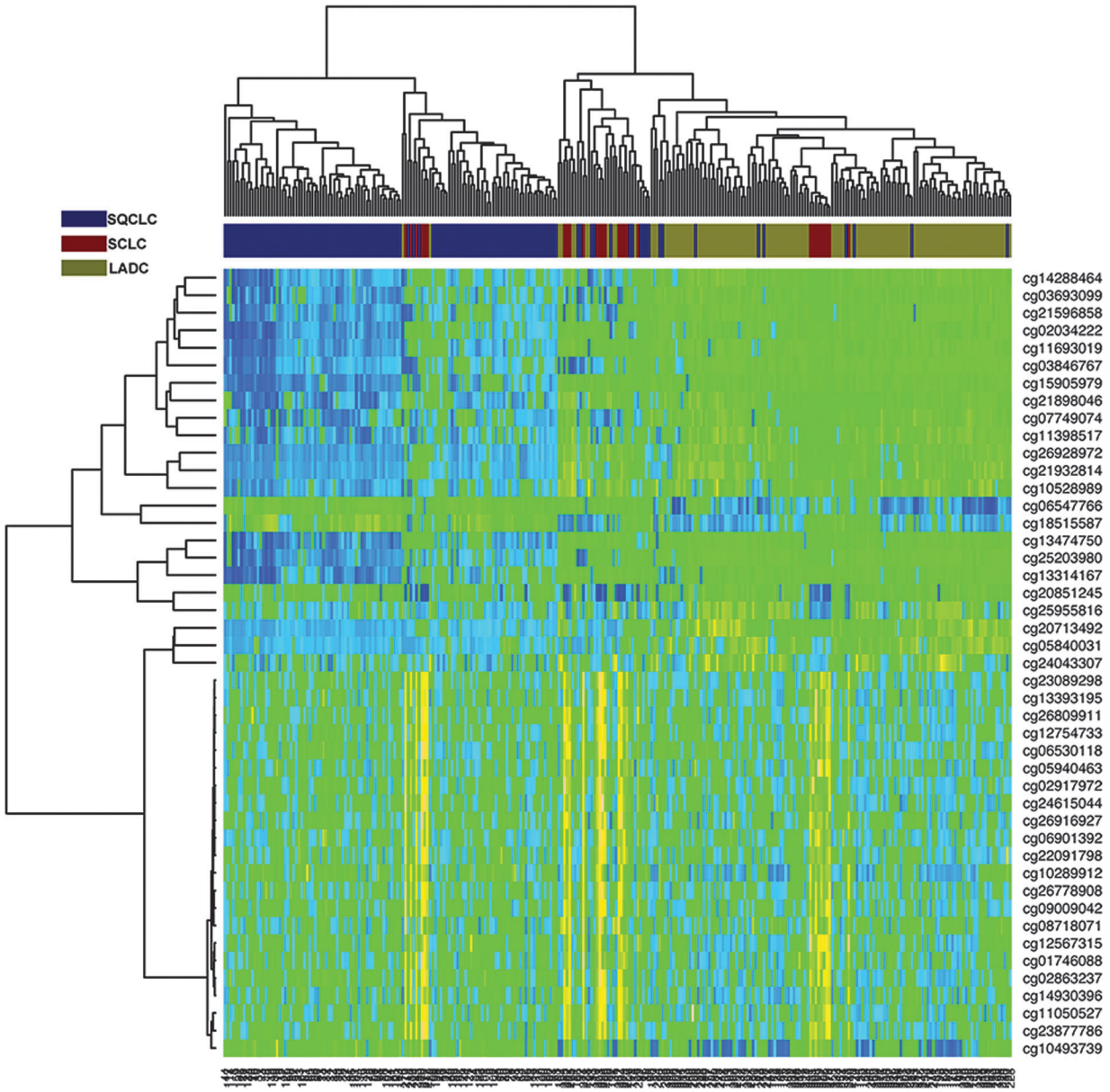

Fig. 1 Hierarchical clustering heatmap of 288 samples in terms of DNA methylation levels of 45 overlapped features by mRMR, HUM and Multi-ROC methods. Each row stands for each feature (probes) and samples are arranged along columns. The three types of lung cancer (SQCLC, SCLC and LADC) are colored in navy blue, maroon and olive on the bar on top, respectively.

test achieved satisfactory results in terms of accuracy and recall metrics (similar to LOOCV results), such test biased independent test performance in terms of precision and *F*-score metrics since our prediction model was developed based on an imbalanced number of three types of lung cancer samples with only 28 SCLC samples, which contributes to a larger independent test precision (on LADC and SQCLC data sets only) compared to that in LOOCV experiments (on LADC, SQCLC and SCLC data sets).

To investigate whether the ensemble-based feature approach can capture the more informative, stable and compact set of features than those captured by individual ones, we performed IFS experiments based on 200, 200 and 140 top ranked features selected by mRMR, Multi-ROC and RFs, respectively. The IFS curves are plotted in Fig. 2 (data shown in ESI,† Table S1). As shown in Fig. 2, the much larger number of optimal top features was required to obtain similar performance for individual feature selection methods as compared to ensemble-based feature selection methods. Theoretically, each of the three feature selection methods would provide the same set of features as an overlapped set of features and performance by the forward feature selection or backward feature selection strategy. However, the time cost of such procedure would increase exponentially. The efficiency and effectiveness of the ensemble-based method might be due to the fact that different methods measure the relationship between features and target in a different manner.

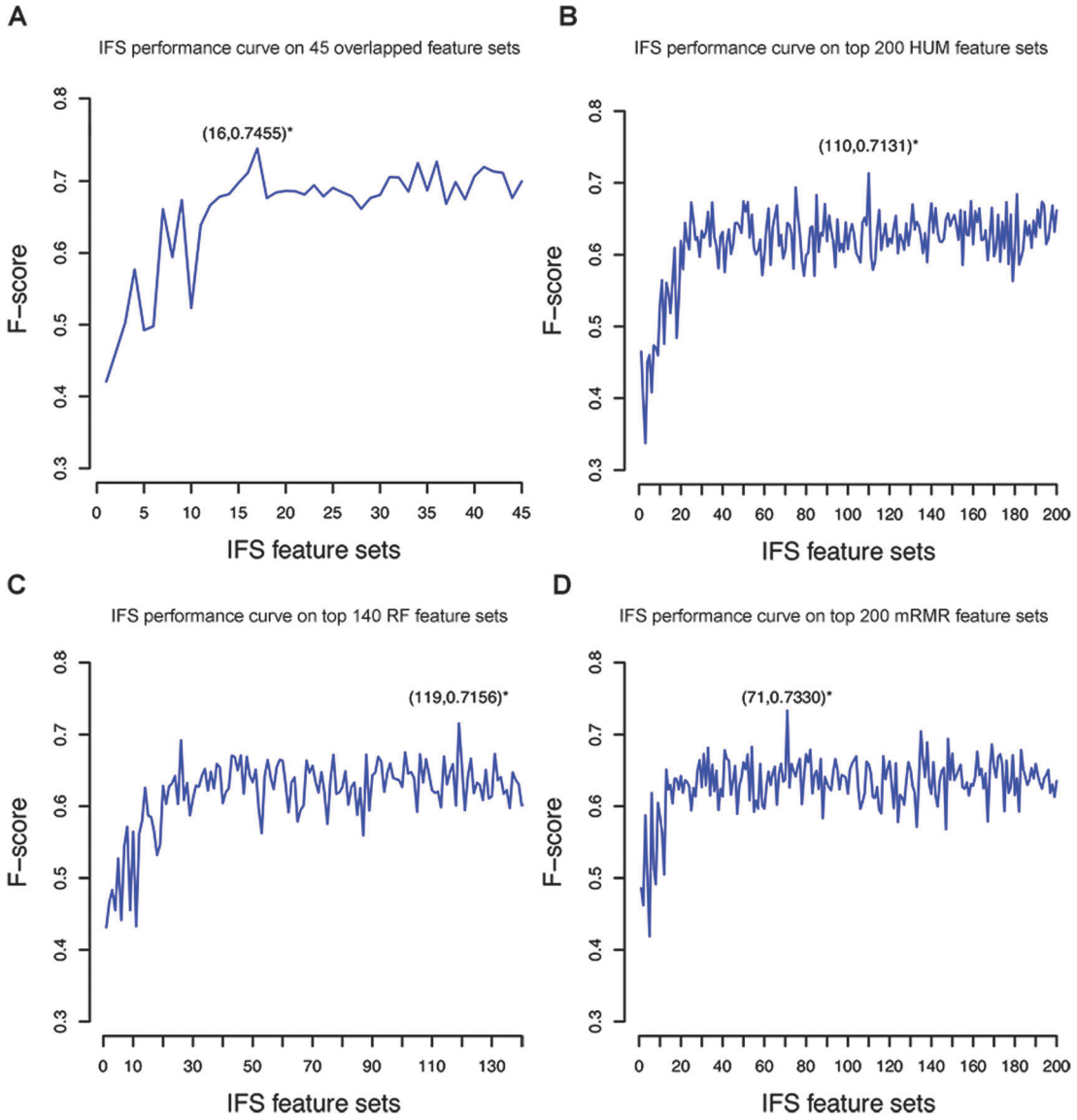

Fig. 2 IFS (incremental feature selection) performance comparison among ensemble-based and individual methods (the *x*-axis corresponds to IFS data sets and the *y*-axis refers to *F*-score values, all of which were averagely measured based on the LOOCV experiments). (A) IFS performance curve on 45 common feature sets shared by mRMR, Multi-ROC and RFs feature selection approaches (ensemble-based method) with a maximum *F*-score of 0.7455 when the first 16 features were employed. The ranking of 45 common features was kept the same as that in the mRMR feature ranking list; (B) IFS performance curve on top 200 HUM feature sets with a maximum *F*-score of 0.7455 when the first 110 features were used; (C) IFS performance curve on top 140 RF feature sets with a maximum *F*-score of 0.7156 when the first 119 features were utilized; (D) IFS performance curve on top 200 mRMR feature sets with a maximum *F*-score of 0.7330 when the first 71 features were applied.

Table 2 Performance in LOOCV experiments on independent data sets of the final classification model constructed with 16 optimal set of IFS features

	Performance assessment metrics			
Data sets	Accuracy (%)	Precision (%)	Recall (%)	*F*-score (%)
Training/cross validation data sets	86.54 ± 2.2	66.79 ± 1.9	84.37 ± 2.5	74.55 ± 2.2
Independent data sets	84.60	85.94	85.52	85.04

The corresponding assessment metrics were represented by average performance ± standard error in the case of LOOCV experiments. Independent test experiments were just conducted on LADC and SQCLC.

Thus, common features selected by ensemble-based methods would provide much more information than those only obtained by individual methods. Taken together, these observations indicate that the ensemble-based approach followed by the IFS procedure might be able to provide a more stable, compact as well as informative set of features than individual feature

Table 3 KEGG pathway and GO analysis (BP: biological process) indicated that several KEGG pathways and GO terms are significantly enriched in a common set of features obtained by three methods ($p \leq 0.05$)

Term	*P* value	Bonferroni	Benjamini
KEGG:hsa05223:non-small cell lung cancer	2.57×10^{-06}	1.75×10^{-04}	1.75×10^{-04}
KEGG:hsa04012:ErbB signaling pathway	2.73×10^{-05}	0.001858	9.29×10^{-04}
KEGG:hsa05200:pathways in cancer	3.19×10^{-04}	0.021481	0.007212
KEGG:hsa04510:focal adhesion	0.001395	0.090535	0.023446
KEGG:hsa04630:Jak-STAT signaling pathway	0.003956	0.236282	0.037776
BP:GO:0042981:regulation of apoptosis	4.16×10^{-05}	0.042106	0.008567
BP:GO:0010941:regulation of cell death	4.68×10^{-05}	0.047153	0.006876

selection approaches, which would facilitate the discovery of stable DNA methylation biomarkers for the diagnosis of different subtypes of lung cancers.

Enrichment analysis of genes corresponding to features

KEGG and GO analysis was carried out through system DAVID.[43] Only KEGG and GO categories with Bonferroni and Benjamini corrected *p*-values ≤ 0.05 were kept for further analysis. As shown in Table 3, the genes corresponding to overlapped 45 features were enriched in the non-small cell lung cancer pathway, the ErbB signaling pathway, the Jak-STAT signaling pathway, Focal adhesion as well as other types of cancers pathways. Current research supports that the ErbB signaling pathway is involved in non-small cell lung cancer through genetic and epigenetic regulations.[51,52] The focal adhesion pathway plays important roles in cell proliferation, survival and metastasis in cancer cells.[53,54] Focal adhesion kinase (FAK), one of the central genes represented in this pathway, plays a significant role in cell survival signaling in NSCLCs and SCLCs.[55–57] The JAK-STAT pathway is vital in cytokine-mediated immune responses. Research studies in the JAK-STAT field have elucidated its roles in various cellular processes such as proliferation, apoptosis and migration, and have found frequent dysregulation of the JAK-STAT pathway in diverse types of cancers, including in NSCLCs.[58–60] With respect to GO analysis, the genes corresponding to 45 features were categorized into a few significant biological processes (BPs) such as regulation of apoptosis and cell death, which are associated with lung cancers.

Methylation marker genes identified in this study

In the present study, we obtained 45 overlapped DNA methylation markers through three independent feature selection methods. The methylation levels of some of the genes corresponding to the identified markers such as ERBB2, FAM19A4 and RASSF1 have been shown to be related to NSCLCs and may be utilized to distinguish NSCLCs and normal lung cells in previous studies.[61–63] Eventually, we identified a panel of 16 DNA methylation markers to distinguish SCLCs, SQCLCs and LADCs simultaneously based on the IFS strategy combined with the SVMs method (the annotation of 16 DNA methylation marker genes is listed in ESI,† Table S3). Of the identified 16 DNA methylation marker genes, the expression or DNA methylation levels of five of them (ST18, PKP1, HOXA1, CDKN2A and ZCCHC11) has been demonstrated to be associated with lung cancer.

The methylation level of ST18 ranks the first in our optimal set of features, which is a repressor that binds to DNA sequences containing a bipartite element consisting of a direct repeat of the sequence 5′-AAAGTTT-3′ separated by 2–9 nucleotides. Job *et al.*[64] performed high-resolution array comparative genomic hybridization analysis of lung adenocarcinoma in sixty never smokers and identified fourteen new minimal common regions (MCR) of gain or loss, of which five contained a single gene (MOCS2, NSUN3, KHDRBS2, SNTG1 and ST18). ST18 was found lost, hypermethylated and its mRNA down-regulated in breast cancer,[65] which might be the case for ST18 in lung adenocarcinoma. Therefore, the methylation of ST18 might present a different level among SCLCs, SQCLCs and LADCs and be a biomarker for lung cancer subtype diagnosis.

The PKP1 (plakophilin-1) gene is involved in many biological processes such as cell adhesion, cell–cell signaling and signal transduction. A previous study[66] reported that PKP1 exhibits aberrant promoter DNA methylation in NSCLCs for the first time. Another study[67] demonstrated differential expression of PKP1 between LADC and SQCLC, which might be attributed to the differential DNA methylation status of PKP1 between LADC and SQCLC. Taken together, these results indicated that the methylation level of PKP1 might be different between SQCLCs and LADCs, even among SCLCs, SQCLCs and LADCs and thus could be a potential marker for the subtyping of lung cancers.

The HOXA1 gene encodes homeobox transcription factor 1. Selamat *et al.*[68] used MethyLight, a sensitive real-time PCR-based quantitative method, to analyze DNA methylation levels at the HOXA1 gene, showing that significant DNA hypermethylation of HOXA1 presents in lung adenocarcinoma (LADC). Abe *et al.*[69] demonstrated that expression levels of HOXA1 in lung squamous cell carcinoma tissues were significantly higher than those in the normal tissues, which indicates that HOXA1 might be DNA hypomethylated in lung squamous cell carcinoma and overexpression of HOXA1 is likely to play a vital role in human carcinogenesis.

DNA hypermethylation is frequent for the CDKN2A/p16 gene,[70] which should be a useful biomarker for diagnosis of NSCLC.[71] The expression suppression of the CDKN2A/p16 gene was reported as a major causative event in LADC and promoter hyper-methylation of CDKN2A(P16) could be as a biomarker in lung cancer.[72] Besides, another study showed that methylation of CDKN2A is more common in SCLC compared to LADC and differential gene hypermethylation frequencies in tumor tissues from patients with adenocarcinoma or squamous cell cancers.[73] Thus, DNA methylation of CDKN21A/p16 would be able to

classify LADC and SQCLC, even for the differentiation among LADC, SQCLC and SCLC.

Another gene ZCCHC11 ranks the eighth in our optimal feature list and is an uridylyltransferase that acts as a suppressor of miRNA biogenesis by specifically mediating the terminal uridylation of some miRNAs. A most recent study[74] demonstrated that Zcchc11 promoted tumor growth and metastasis, and it was prominently overexpressed due to hypo-methylation in human cancers including NSCLCs. Therefore, the methylation of Zcchc11 might play different roles in different types of lung cancers and be able to act as a potential maker among LADC, SQCLC and SCLC.

Conclusion

In conclusion, the results obtained in this study show that ensemble-based feature selection followed by the IFS method presents the merit of acquisition of more informative and compact features than those obtained by individual methods, which have been demonstrated by the present study.

Acknowledgements

We would like to thank the editor and reviewers' helpful suggestions. This study was supported by the National Key Scientific & Technology Support Program (NO. 2013BAI09B09.): Collaborative innovation of Clinical Research for chronic obstructive pulmonary disease and lung cancer.

References

1 C. K. Toh, *Methods Mol. Biol.*, 2009, **472**, 397–411.
2 R. Govindan, N. Page, D. Morgensztern, W. Read, R. Tierney, A. Vlahiotis, E. L. Spitznagel and J. Piccirillo, *J. Clin. Oncol.*, 2006, **24**, 4539–4544.
3 D. M. Jackman and B. E. Johnson, *Lancet*, 2005, **366**, 1385–1396.
4 T. Sher, G. K. Dy and A. A. Adjei, *Mayo Clin. Proc.*, 2008, **83**, 355–367.
5 R. S. Herbst, V. J. O'Neill, L. Fehrenbacher, C. P. Belani, P. D. Bonomi, L. Hart, O. Melnyk, D. Ramies, M. Lin and A. Sandler, *J. Clin. Oncol.*, 2007, **25**, 4743–4750.
6 A. Sandler, *Clin. Cancer Res.*, 2007, **13**, 4613s–4616s.
7 A. F. Gazdar, *Ann. Oncol.*, 2010, **21**, 225–229.
8 A. Stang, H. Pohlabeln, K. M. Muller, I. Jahn, K. Giersiepen and K. H. Jockel, *Lung Cancer*, 2006, **52**, 29–36.
9 J. A. Bishop, H. Benjamin, H. Cholakh, A. Chajut, D. P. Clark and W. H. Westra, *Clin. Cancer Res.*, 2010, **16**, 610–619.
10 A. M. Maeshima, M. Omatsu, K. Tsuta, H. Asamura and Y. Matsuno, *Pathol. Int.*, 2008, **58**, 31–37.
11 D. Lebanony, H. Benjamin, S. Gilad, M. Ezagouri, A. Dov, K. Ashkenazi, N. Gefen, S. Izraeli, G. Rechavi, H. Pass, D. Nonaka, J. J. Li, Y. Spector, N. Rosenfeld, A. Chajut, D. Cohen, R. Aharonov and M. Mansukhani, *J. Clin. Oncol.*, 2009, **27**, 2030–2037.
12 J. Hou, J. Aerts, B. den Hamer, W. van Ijcken, M. den Bakker, P. Riegman, C. van der Leest, P. van der Spek, J. A. Foekens, H. C. Hoogsteden, F. Grosveld and S. Philipsen, *PLoS One*, 2010, **5**, e10312.
13 B. Q. Li, J. You, T. Huang and Y. D. Cai, *PLoS One*, 2014, **9**, e88300.
14 D. Seidel, T. Zander, L. C. Heukamp, M. Peifer, M. Bos, L. Fernandez-Cuesta, F. Leenders, X. Lu, S. Ansen, M. Gardizi, C. Nguyen, J. Berg, P. Russell, Z. Wainer, H. U. Schildhaus, T. M. Rogers, B. Solomon, W. Pao, S. L. Carter, G. Getz, D. N. Hayes, M. D. Wilkerson, E. Thunnissen, W. D. Travis, S. Perner, G. Wright, E. Brambilla, R. Buttner, J. Wolf, R. K. Thomas, F. Gabler, I. Wilkening, C. Muller, I. Dahmen, R. Menon, K. Konig, K. Albus, S. Merkelbach-Bruse, J. Fassunke, K. Schmitz, H. Kuenstlinger, M. A. Kleine, E. Binot, S. Querings, J. Altmuller, I. Bossmann, P. Numberg, P. M. Schneider, M. Bogus, R. Buttner, S. Perner, P. Russell, E. Thunnissen, W. D. Travis, E. Brambilla, A. Soltermann, H. Moch, O. T. Brustugun, S. Solberg, M. Lund-Iversen, A. Helland, T. Muley, H. Hoffmann, P. A. Schnabel, Y. Chen, H. Groen, W. Timens, H. Sietsma, J. H. Clement, W. Weder, J. Sanger, E. Stoelben, C. Ludwig, W. Engel-Riedel, E. Smit, D. A. M. Heideman, P. J. F. Snijders, L. Nogova, M. L. Sos, C. Mattonet, K. Topelt, M. Scheffler, E. Goekkurt, R. Kappes, S. Kruger, K. Kambartel, D. Behringer, W. Schulte, W. Galetke, W. Randerath, M. Heldwein, A. Schlesinger, M. Serke, K. Hekmat, K. F. Frank, R. Schnell, M. Reiser, A. N. Hunerliturkoglu, S. Schmitz, L. Meffert, Y. D. Ko, M. Litt-Lampe, U. Gerigk, R. Fricke, B. Besse, C. Brambilla, S. Lantuejoul, P. Lorimier, D. Moro-Sibilot, F. Cappuzzo, C. Ligorio, S. Damiani, J. K. Field, R. Hyde, P. Validire, P. Girard, L. A. Muscarella, V. M. Fazio, M. Hallek, J. C. Soria, S. L. Carter, G. Getz, D. N. Hayes, M. D. Wilkerson, V. Achter, U. Lang, D. Seidel, T. Zander, L. C. Heukamp, M. Peifer, M. Bos, W. Pao, W. D. Travis, E. Brambilla, R. Buttner, J. Wolf, R. K. Thomas, Clcgp and Ngm, *Sci. Transl. Med.*, 2013, **5**, 209ra153.
15 S. Gilad, G. Lithwick-Yanai, I. Barshack, S. Benjamin, I. Krivitsky, T. B. Edmonston, M. Bibbo, C. Thurm, L. Horowitz, Y. J. Huang, M. Feinmesser, J. S. Hou, B. St Cyr, I. Burnstein, H. Gibori, N. Dromi, M. Sanden, M. Kushnir and R. Aharonov, *J. Mol. Diagn.*, 2012, **14**, 510–517.
16 G. Nikolaidis, O. Y. Raji, S. Markopoulou, J. R. Gosney, J. Bryan, C. Warburton, M. Walshaw, J. Sheard, J. K. Field and T. Liloglou, *Cancer Res.*, 2012, **72**, 5692–5701.
17 W. D. Travis, E. Brambilla and G. J. Riely, *J. Clin. Oncol.*, 2013, **31**, 992–1001.
18 J. Sandoval, J. Mendez-Gonzalez, E. Nadal, G. A. Chen, F. J. Carmona, S. Sayols, S. Moran, H. Heyn, M. Vizoso, A. Gomez, M. Sanchez-Cespedes, Y. Assenov, F. Muller, C. Bock, M. Taron, J. Mora, L. A. Muscarella, T. Liloglou, M. Davies, M. Pollan, M. J. Pajares, W. Torre, L. M. Montuenga, E. Brambilla, J. K. Field, L. Roz, M. Lo Iacono, G. V. Scagliotti, R. Rosell, D. G. Beer and M. Esteller, *J. Clin. Oncol.*, 2013, **31**, 4140–4147.

19 J. A. Tsou, J. S. Galler, K. D. Siegmund, P. W. Laird, S. Turla, W. Cozen, J. A. Hagen, M. N. Koss and I. A. Laird-Offringa, *Mol. Cancer*, 2007, **6**, 70.

20 P. P. Anglim, J. S. Galler, M. N. Koss, J. A. Hagen, S. Turla, M. Campan, D. J. Weisenberger, P. W. Laird, K. D. Siegmund and I. A. Laird-Offringa, *Mol. Cancer*, 2008, **7**, 62.

21 A. M. Marchevsky, J. A. Tsou and I. A. Laird-Offringa, *J. Mol. Diagn.*, 2004, **6**, 28–36.

22 P. S. Hammerman, M. S. Lawrence, D. Voet, R. Jing, K. Cibulskis, A. Sivachenko, P. Stojanov, A. McKenna, E. S. Lander, S. Gabriel, G. Getz, C. Sougnez, M. Imielinski, E. Helman, B. Hernandez, N. H. Pho, M. Meyerson, A. Chu, H. J. E. Chun, A. J. Mungall, E. Pleasance, A. G. Robertson, P. Sipahimalani, D. Stoll, M. Balasundaram, I. Birol, Y. S. N. Butterfield, E. Chuah, R. J. N. Coope, R. Corbett, N. Dhalla, R. Guin, A. C. Hirst, M. Hirst, R. A. Holt, D. Lee, H. I. Li, M. Mayo, R. A. Moore, K. Mungall, K. M. Nip, A. Olshen, J. E. Schein, J. R. Slobodan, A. Tam, N. Thiessen, R. Varhol, T. Zeng, Y. Zhao, S. J. M. Jones, M. A. Marra, G. Saksena, A. D. Cherniack, S. E. Schumacher, B. Tabak, S. L. Carter, N. H. Pho, H. Nguyen, R. C. Onofrio, A. Crenshaw, K. Ardlie, R. Beroukhim, W. Winckler, P. S. Hammerman, G. Getz, M. Meyerson, A. Protopopov, J. H. Zhang, A. Hadjipanayis, S. Lee, R. B. Xi, L. X. Yang, X. J. Ren, H. L. Zhang, S. Shukla, P. C. Chen, P. Haseley, E. Lee, L. Chin, P. J. Park, R. Kucherlapati, N. D. Socci, Y. P. Liang, N. Schultz, L. Borsu, A. E. Lash, A. Viale, C. Sander, M. Ladanyi, J. T. Auman, K. A. Hoadley, M. D. Wilkerson, Y. Shi, C. Liquori, S. W. Meng, L. Li, Y. J. Turman, M. D. Topal, D. H. Tan, S. Waring, E. Buda, J. Walsh, C. D. Jones, P. A. Mieczkowski, D. Singh, J. Wu, A. Gulabani, P. Dolina, T. Bodenheimer, A. P. Hoyle, J. V. Simons, M. G. Soloway, L. E. Mose, S. R. Jefferys, S. Balu, B. D. O'Connor, J. F. Prins, J. Liu, D. Y. Chiang, D. N. Hayes, C. M. Perou, L. Cope, L. Danilova, D. J. Weisenberger, D. T. Maglinte, F. Pan, D. J. den Berg, T. Triche, J. G. Herman, S. B. Baylin, P. W. Laird, G. Getz, M. Noble, D. Voet, G. Saksena, N. Gehlenborg, D. DiCara, J. H. Zhang, H. L. Zhang, C. J. Wu, S. Y. Liu, M. S. Lawrence, L. H. Zou, A. Sivachenko, P. Lin, P. Stojanov, R. Jing, J. Cho, M. D. Nazaire, J. Robinson, H. Thorvaldsdottir, J. Mesirov, P. J. Park, L. Chin, N. Schultz, R. Sinha, G. Ciriello, E. Cerami, B. Gross, A. Jacobsen, J. Gao, B. A. Aksoy, N. Weinhold, R. Ramirez, B. S. Taylor, Y. Antipin, B. Reva, R. L. Shen, Q. Mo, V. Seshan, P. K. Paik, M. Ladanyi, C. Sander, R. Akbani, N. X. Zhang, B. M. Broom, T. Casasent, A. Unruh, C. Wakefield, R. C. Cason, K. A. Baggerly, J. N. Weinstein, D. Haussler, C. C. Benz, J. M. Stuart, J. C. Zhu, C. Szeto, G. K. Scott, C. Yau, S. Ng, T. Goldstein, P. Waltman, A. Sokolov, K. Ellrott, E. A. Collisson, D. Zerbino, C. Wilks, S. Ma, B. Craft, M. D. Wilkerson, J. T. Auman, K. A. Hoadley, Y. Du, C. Cabanski, V. Walter, D. Singh, J. Y. Wu, A. Gulabani, T. Bodenheimer, A. P. Hoyle, J. V. Simons, M. G. Soloway, L. E. Mose, S. R. Jefferys, S. Balu, J. S. Marron, Y. Liu, K. Wang, J. Liu, J. F. Prins, D. N. Hayes, C. M. Perou, C. J. Creighton, Y. Q. Zhang, W. D. Travis, N. Rekhtman, J. Yi, M. C. Aubry, R. Cheney, S. Dacic, D. Flieder, W. Funkhouser, P. Illei, J. Myers, M. S. Tsao, R. Penny, D. Mallery, T. Shelton, M. Hatfield, S. Morris, P. Yena, C. Shelton, M. Sherman, J. Paulauskis, M. Meyerson, S. B. Baylin, R. Govindan, R. Akbani, I. Azodo, D. Beer, R. Bose, L. A. Byers, D. Carbone, L. W. Chang, D. Chiang, A. Chu, E. Chun, E. Collisson, L. Cope, C. J. Creighton, L. Danilova, L. Ding, G. Getz, P. S. Hammerman, D. N. Hayes, B. Hernandez, J. G. Herman, J. Heymach, C. Ida, M. Imielinski, B. Johnson, I. Jurisica, J. Kaufman, F. Kosari, R. Kucherlapati, D. Kwiatkowski, M. Ladanyi, M. S. Lawrence, C. A. Maher, A. Mungall, S. Ng, W. Pao, M. Peifer, R. Penny, G. Robertson, V. Rusch, C. Sander, N. Schultz, R. L. Shen, J. Siegfried, R. Sinha, A. Sivachenko, C. Sougnez, D. Stoll, J. Stuart, R. K. Thomas, S. Tomaszek, M. S. Tsao, W. D. Travis, C. Vaske, J. N. Weinstein, D. Weisenberger, D. Wheeler, D. A. Wigle, M. D. Wilkerson, C. Wilks, P. Yang, J. J. Zhang, M. A. Jensen, R. Sfeir, A. B. Kahn, A. L. Chu, P. Kothiyal, Z. Wang, E. E. Snyder, J. Pontius, T. D. Pihl, B. Ayala, M. Backus, J. Walton, J. Baboud, D. L. Berton, M. C. Nicholls, D. Srinivasan, R. Raman, S. Girshik, P. A. Kigonya, S. Alonso, R. N. Sanbhadti, S. P. Barletta, J. M. Greene, D. A. Pot, M. S. Tsao, B. Bandarchi-Chamkhaleh, J. Boyd, J. Weaver, D. A. Wigle, I. A. Azodo, S. C. Tomaszek, M. C. Aubry, C. M. Ida, P. Yang, F. Kosari, M. V. Brock, K. Rogers, M. Rutledge, T. Brown, B. Lee, J. Shin, D. Trusty, R. Dhir, J. M. Siegfried, O. Potapova, K. V. Fedosenko, E. Nemirovich-Danchenko, V. Rusch, M. Zakowski, M. V. Iacocca, J. Brown, B. Rabeno, C. Czerwinski, N. Petrelli, Z. Fan, N. Todaro, J. Eckman, J. Myers, W. K. Rathmell, L. B. Thorne, M. Huang, L. Boice, A. Hill, R. Penny, D. Mallery, E. Curley, C. Shelton, P. Yena, C. Morrison, C. Gaudioso, J. S. Bartlett, S. Kodeeswaran, B. Zanke, H. Sekhon, K. David, H. Juhl, X. Van Le, B. Kohl, R. Thorp, N. V. Tien, N. Van Bang, H. Sussman, B. D. Phu, R. Hajek, N. PhiHung, K. Z. Khan, T. Muley, K. R. M. Shaw, M. Sheth, L. Yang, K. Buetow, T. Davidsen, J. A. Demchok, G. Eley, M. Ferguson, L. A. L. Dillon, C. Schaefer, M. S. Guyer, B. A. Ozenberger, J. D. Palchik, J. Peterson, H. J. Sofia, E. Thomson, M. Meyerson and C. G. A. R. Network, *Nature*, 2012, **489**, 519–525.

23 C. Fernandez-Lozano, E. Fernandez-Blanco, K. Dave, N. Pedreira, M. Gestal, J. Dorado and C. R. Munteanu, *Mol. BioSyst.*, 2014, **10**, 1063–1071.

24 C. Fernandez-Lozano, M. Gestal, N. Pedreira-Souto, L. Postelnicu, J. Dorado and C. R. Munteanu, *Curr. Top. Med. Chem.*, 2013, **13**, 1681–1691.

25 B. Q. Li, Y. D. Cai, K. Y. Feng and G. J. Zhao, *PLoS One*, 2012, **7**, e45854.

26 B. Q. Li, L. L. Hu, L. Chen, K. Y. Feng, Y. D. Cai and K. C. Chou, *PLoS One*, 2012, **7**, e39308.

27 T. Barrett, S. E. Wilhite, P. Ledoux, C. Evangelista, I. F. Kim, M. Tomashevsky, K. A. Marshall, K. H. Phillippy, P. M. Sherman, M. Holko, A. Yefanov, H. Lee, N. G. Zhang, C. L. Robertson, N. Serova, S. Davis and A. Soboleva, *Nucleic Acids Res.*, 2013, **41**, D991–D995.

28 K. Soreide, *J. Clin. Pathol.*, 2009, **62**, 1–5.

29 B. K. Scurfield, *J. Math. Psychol.*, 1996, **40**, 253–269.

30 B. K. Scurfield, *J. Math. Psychol.*, 1998, **42**, 5–31.
31 J. L. Li and J. P. Fine, *Biostatistics*, 2008, **9**, 566–576.
32 J. L. Li and X. H. Zhou, *J. Stat. Plan. Inference*, 2009, **139**, 4133–4142.
33 N. Novoselova, C. Della Beffa, J. Wang, J. Li, F. Pessler and F. Klawonn, *Bioinformatics*, 2014, **30**, 1635–1636.
34 H. Peng, F. Long and C. Ding, *IEEE Trans. Pattern Anal. Mach. Intell.*, 2005, **27**, 1226–1238.
35 B. Q. Li, T. Huang, L. Liu, Y. D. Cai and K. C. Chou, *PLoS One*, 2012, **7**, e33393.
36 T. Huang, J. Zhang, Z. P. Xu, L. L. Hu, L. Chen, J. L. Shao, L. Zhang, X. Y. Kong, Y. D. Cai and K. C. Chou, *Biochimie*, 2012, **94**, 1017–1025.
37 T. Huang, J. Wang, Y. D. Cai, H. Yu and K. C. Chou, *PLoS One*, 2012, **7**, e34460.
38 L. Breiman, *Mach. Learn.*, 2001, **45**, 5–32.
39 R. H. Chung and Y. E. Chen, *PLoS One*, 2012, **7**, e36662.
40 C.-C. Chang and C.-J. Lin, *ACM Trans. Intell. Syst. Technol.*, 2011, **2**, 1–27. Software available at http://www.csie.ntu.edu.tw/∼cjlin/libsvm.
41 C. W. Hsu and C. J. Lin, *IEEE Trans. Neural Networks*, 2002, 415–425.
42 M. Sokolova and G. Lapalme, *Inf. Proc. Manag.*, 2009, **45**, 427–437.
43 W. Huang da, B. T. Sherman and R. A. Lempicki, *Nat. Protoc.*, 2009, **4**, 44–57.
44 Y. Benjamini and Y. Hochberg, *J. R. Statist. Soc. B*, 1995, 289–300.
45 M. Benito, J. Parker, Q. Du, J. Wu, D. Xiang, C. M. Perou and J. S. Marron, *Bioinformatics*, 2004, **20**, 105–114.
46 K. C. Chou, *J. Theor. Biol.*, 2011, **273**, 236–247.
47 G. L. Fan and Q. Z. Li, *J. Theor. Biol.*, 2012, **304**, 88–95.
48 K. C. Chou, Z. C. Wu and X. Xiao, *Mol. BioSyst.*, 2012, **8**, 629–641.
49 X. Xiao, P. Wang and K. C. Chou, *Mol. BioSyst.*, 2011, **7**, 911–919.
50 Z. C. Wu, X. Xiao and K. C. Chou, *Mol. BioSyst.*, 2011, **7**, 3287–3297.
51 T. Yu, J. Li, M. Yan, L. Liu, H. Lin, F. Zhao, L. Sun, Y. Zhang, Y. Cui, F. Zhang, J. Li, X. He and M. Yao, *Oncogene*, 2014, DOI: 10.1038/onc.2013.574.
52 M. O. Hoque, M. Brait, E. Rosenbaum, M. L. Poeta, P. Pal, S. Begum, S. Dasgupta, A. L. Carvalho, S. A. Ahrendt, W. H. Westra and D. Sidransky, *J. Thorac. Oncol.*, 2010, **5**, 1887–1893.
53 T. Yamasaki, N. Seki, H. Yoshino, T. Itesako, H. Hidaka, Y. Yamada, S. Tatarano, T. Yonezawa, T. Kinoshita, M. Nakagawa and H. Enokida, *J. Urol.*, 2013, **190**, 1059–1068.
54 S. Ocak, H. Yamashita, A. R. Udyavar, A. N. Miller, A. L. Gonzalez, Y. Zou, A. Jiang, Y. Yi, Y. Shyr, L. Estrada, V. Quaranta and P. P. Massion, *Oncogene*, 2010, **29**, 6331–6342.
55 W. Y. Kim, J. Y. Jang, Y. K. Jeon, D. H. Chung, Y. G. Kim and C. W. Kim, *Exp. Mol. Med.*, 2014, **46**, e90.
56 G. K. Dy, L. Ylagan, S. Pokharel, A. Miller, E. Brese, W. Bshara, C. Morrison, W. G. Cance and V. M. Golubovskaya, *J. Thorac. Oncol.*, 2014, **9**, 1278–1284.
57 S. Ocak, H. Chen, C. Callison, A. L. Gonzalez and P. P. Massion, *Cancer*, 2012, **118**, 1293–1301.
58 Y. Hu, Y. Hong, Y. Xu, P. Liu, D. H. Guo and Y. Chen, *Apoptosis*, 2014, **19**, 1627–1636.
59 L. Song, B. Rawal, J. A. Nemeth and E. B. Haura, *Mol. Cancer Ther.*, 2011, **10**, 481–494.
60 P. J. Murray, *J. Immunol.*, 2007, **178**, 2623–2629.
61 A. J. Hubers, M. A. van der Drift, C. F. Prinsen, B. I. Witte, Y. Wang, N. Shivapurkar, V. Stastny, A. S. Bolijn, B. E. Hol, Z. Feng, P. N. Dekhuijzen, A. F. Gazdar and E. Thunnissen, *Lung Cancer*, 2014, **84**, 127–133.
62 K. Walter, T. Holcomb, T. Januario, P. Du, M. Evangelista, N. Kartha, L. Iniguez, R. Soriano, L. Huw, H. Stern, Z. Modrusan, S. Seshagiri, G. M. Hampton, L. C. Amler, R. Bourgon, R. L. Yauch and D. S. Shames, *Clin. Cancer Res.*, 2012, **18**, 2360–2373.
63 N. Shivapurkar, V. Stastny, M. Suzuki, I. I. Wistuba, L. Li, Y. Zheng, Z. Feng, B. Hol, C. Prinsen, F. B. Thunnissen and A. F. Gazdar, *Cancer Lett.*, 2007, **247**, 56–71.
64 B. Job, A. Bernheim, M. Beau-Faller, S. Camilleri-Broet, P. Girard, P. Hofman, J. Mazieres, S. Toujani, L. Lacroix, J. Laffaire, P. Dessen, P. Fouret and L. G. Investigators, *PLoS One*, 2010, **5**, e15145.
65 B. Jandrig, S. Seitz, B. Hinzmann, W. Arnold, B. Micheel, K. Koelble, R. Siebert, A. Schwartz, K. Ruecker, P. M. Schlag, S. Scherneck and A. Rosenthal, *Oncogene*, 2004, **23**, 9295–9302.
66 M. Kusakabe, T. Kutomi, K. Watanabe, N. Emoto, N. Aki, H. Kage, E. Hamano, H. Kitagawa, T. Nagase, A. Sano, Y. Yoshida, T. Fukami, T. Murakawa, J. Nakajima, S. Takamoto, S. Ota, M. Fukayama, Y. Yatomi, N. Ohishi and D. Takai, *Int. J. Cancer*, 2010, **126**, 1895–1902.
67 A. Sanchez-Palencia, M. Gomez-Morales, J. A. Gomez-Capilla, V. Pedraza, L. Boyero, R. Rosell and M. E. Farez-Vidal, *Int. J. Cancer*, 2011, **129**, 355–364.
68 S. A. Selamat, J. S. Galler, A. D. Joshi, M. N. Fyfe, M. Campan, K. D. Siegmund, K. M. Kerr and I. A. Laird-Offringa, *PLoS One*, 2011, **6**, e21443.
69 M. Abe, J. Hamada, O. Takahashi, Y. Takahashi, M. Tada, M. Miyamoto, T. Morikawa, S. Kondo and T. Moriuchi, *Oncol. Rep.*, 2006, **15**, 797–802.
70 A. Suzuki, H. Makinoshima, H. Wakaguri, H. Esumi, S. Sugano, T. Kohno, K. Tsuchihara and Y. Suzuki, *Nucleic Acids Res.*, 2014, DOI: 10.1093/nar/gku885.
71 P. Xiao, J. R. Chen, F. Zhou, C. X. Lu, Q. Yang, G. H. Tao, Y. J. Tao and J. L. Chen, *Lung Cancer*, 2014, **83**, 56–60.
72 S. A. Belinsky, *Nat. Rev. Cancer*, 2004, **4**, 707–717.
73 S. E. Hawes, J. E. Stern, Q. Feng, L. W. Wiens, J. S. Rasey, H. Lu, N. B. Kiviat and H. Vesselle, *Lung Cancer*, 2010, **69**, 172–179.
74 X. Fu, Z. Meng, W. Liang, Y. Tian, X. Wang, W. Han, G. Lou, X. Wang, F. Lou, Y. Yen, H. Yu, R. Jove and W. Huang, *Oncogene*, 2014, **33**, 4296–4306.

