https://doi.org/10.1016/j.compbiomed.2018.09.005

Multi-scale Representation of Proteomic Data Exhibits Distinct MicroRNA Regulatory Modules in Non-smoking Female Patients with Lung Adenocarcinoma

Lawrence W. Chan[1]*, Fengfeng Wang[1], Fei Meng[1], S.C. Cesar Wong[1], Joseph S. Au[2], Sijun Yang[3], William C. Cho[2]*

[1]Department of Health Technology and Informatics, The Hong Kong Polytechnic University, Hong Kong. E-mail: wing.chi.chan@polyu.edu.hk

[2]Department of Clinical Oncology, Queen Elizabeth Hospital, Kowloon, Hong Kong. E-mail: williamcscho@gmail.com

[3]Institute of Animal Model for Human Diseases, ABSL-3 Laboratory and state key lab of virology, Wuhan University, Wuhan, Hubei, China

Abstract

Adenocarcinoma in female non-smokers is an under-explored subgroup of non-small cell lung cancer (NSCLC) where the molecular mechanism and genetic risk factors remain unclear. We analyzed the protein profiles of plasma samples of 45 patients in this subgroup and 60 non-cancer subjects using surface-enhanced laser desorption/ionization time-of-fight mass spectrometry. Among 85 peaks of mass spectra, the differential expression analysis identified 15 markers based on False Discovery Rate control and the Digital Wavelet Transforms further selected a cluster of 6 markers that were consistently observed at multiple scales of mass-charge ratios. This marker cluster, corresponding to 7 unique proteins, was able to distinguish the

female non-smokers with adenocarcinoma from non-cancer subjects with a very high accuracy, 87.6%. We also predicted the role of competing endogenous RNAs (ceRNAs) in 3 out of these 7 proteins. It was found in many other studies that these ceRNAs and their targeting microRNAs (miRNAs), miR-206 and miR-613, were significantly associated with NSCLC. This study paves a crucial path for further investigating the genetic markers and molecular mechanism of this special NSCLC subgroup.

Keywords: Lung adenocarcinoma, mass spectrometry, multi-scale representation, marker cluster, microRNA, regulatory modules

Introduction

With high mortality rate, lung cancer is the leading cause of cancer deaths worldwide and over 80% of cases are non-small cell lung cancer (NSCLC) [1]. According to World Health Organization (WHO), the global cancer incidence and mortality rate of lung cancer in men are three times of that in women and smoking is a key risk factor [2]. Adenocarcinoma is classified as one of several NSCLC, which constitutes about 40% of lung cancers [3]. Interestingly, adenocarcinoma is more frequently observed in female and non-smokers [4,5]. In East Asia, about half of the female patients with adenocarcinoma never smoke, forming an under-explored lung cancer subgroup [6]. To explore this interesting subgroup, a recent study found that the protein profile of lung tumor from female non-smoker with adenocarcinoma was differentially expressed when compared with the adjacent normal tissue [6].

Proteins act as a better proxy for biomolecular activity than RNAs because they are the actual effectors directly interacting with the other proteins, RNAs and DNAs. Moreover, proteins could be extracellular, intracellular or transmembrane so that they can be detected in various body fluids, like blood plasma, whose collection is non-invasive [7]. Proteomics is the study of hundreds or thousands of proteins and their interactions in biological sample through massive detection and quantification methods [8]. The protein characteristics and quantity profiles of patients' samples are very useful for the early detection, diagnosis, monitoring and theranosis of disease, and the drug target identification [9]. Mass spectrometry (MS) is a high throughput technology that makes use of both advanced analytical methods and bioinformatics to study the biological role of proteins in diseases [6]. The matrix-assisted laser desorption/ionization (MALDI) MS and its variants, surface-enhanced laser desorption/ionization time-of-fight (SELDI-TOF) MS and matrix assisted laser desorption and ionization, time-of-flight (MALDI-TOF) MS, have been widely used for generating the protein profiles of tissue, serum and urine samples from patients. Among these platforms, SELDI-TOF MS is the most suitable to analyze body fluids [10]. Proteinchip arrays represent a powerful tool of MS that reduces sample complexity using selective capture strategies and increases the sensitivity in detecting the low-abundance proteins.

In MS analysis, proteins are identified by time-of-flight (TOF), which can be converted to the corresponding mass to charge ratio (m/z), and their concentrations in the processed sample are reflected by the signal intensities. MS generates a mass spectrum as output showing the signal intensities across different m/z values [7]. The analysis of mass spectrum focuses on fitting peaks that represent meaningful proteomic characteristics. The most straightforward strategy, proposed by Yasui et al., is to seek for the local maxima of a spectrum, each of which represents an m/z

point with higher intensity than the neighboring points [11]. Making use of some filtering criteria, such as signal-to-noise ratio, intensity and area under the fitted curve, these maxima could be filtered to exclude the noise and increase the likelihood for corresponding to peptides [12]. Some other studies proposed to transform the mass spectrum to the wavelet coefficient space and look for clusters of wavelet coefficients with high values on similar position for different scales [13]. All the above-mentioned methods emphasize on pre-processing the spectrum using mathematical approach without considering the relatedness of peaks with disease and molecular regulation. The identified peaks would be more meaningful if they represent clusters of m/z points that differentiate disease and non-disease cases effectively.

Although many studies have analyzed the tumor cells for lung cancer diagnosis, the proteomics data of plasma is rarely considered and the analytical approach has not been extensively explored and validated. This study is aimed to propose the analytical method that transforms a SELDI-TOF spectrum to wavelet coefficient profiles at multiple scales using digital wavelet transform (DWT), filters the wavelet coefficients based on the differential expression analysis and identifies clusters, which could accurately distinguish the plasma samples from non-smoking women with lung adenocarcinoma and non-cancer controls.

As a class of endogenous and non-coding RNA short fragments of 21-25 nucleotides long in mature forms, miRNAs could post-transcriptionally regulate the expression of their target genes (including oncogenes or tumor suppressor genes) through the binding to the miRNA response elements (MREs) of their target messenger RNAs (mRNAs). MREs represent the sequence motifs in 3’-untranslated region (3’-UTR) of mRNAs that complement with the targeting miRNAs [14].

Competing endogenous RNAs (ceRNAs) are RNA transcripts that interact indirectly by sharing the same MREs and competing for shared miRNAs. An up-regulated ceRNA attracts the targeting miRNAs and keeps them away from another ceRNA whose expression is indirectly promoted [14]. The co-regulation of ceRNAs paves the undiscovered crosstalk cascades in the cancer signaling pathway. It was shown that PTEN and its putative ceRNAs are co-regulated in prostate cancer and glioblastoma [15]. The ceRNA crosstalk depends heavily on the miRNA:target concentration ratio and is thus mediated by the targeting miRNAs [16]. Identification of miRNA regulatory modules consisting of miRNAs and the proteins of their target ceRNAs plays an important role for discovering the new molecular mechanism in cancer.

Methods

Proteomic Analysis of Blood Samples

Before the enrolment, informed written consent was obtained from all subjects. We followed the principles of the Declaration of Helsinki. We randomly selected 45 non-smoking females with adenocarcinoma (cancer group) from pretreatment patients of QEH and recruited 60 non-cancer controls (non-cancer group). Plasma samples were collected from these two groups of subjects. The non-cancer group consists of 30 lung disease patients without known neoplastic tumor and 30 healthy volunteers free of any known acute or chronic illness. We fractionated and profiled the plasma samples with the SELDI-TOF-MS. The array was analyzed on a Proteinchip PCS4000 Reader (Ciphergen Biosystems) with acquisition up to 200 kDa and the m/z spectrum was generated by averaging a total of 338 laser shots at an intensity of 195. The proteomic profile data from fractionation with pH 9 and cationic CM10 chip was considered in this study.

Digital Wavelet Transform

The m/z values of mass spectrum were rounded to the nearest integers and fitted to a discrete integer domain from 1 to 2^k where k was the lowest integer such that all the m/z values are bounded by 2^k. For each plasma sample, the level 0 spectrum was constructed by filling the intensities for available m/z values and zeros for the remaining. Digital wavelet transforms (DWTs) were performed on the level 0 spectrum for each plasma sample. Daubechies wavelet, db1, which is orthogonal, biorthogonal and symmetry filter with length 2, was used for DWTs. A profile of approximation coefficients was collected for each time of DWT. The approximation profile collected after the i^{th} DWT is regarded as the level i spectrum.

Marker Identification

To identify the markers (m/z values) that differentiate the female non-smokers with adenocarcinoma from the non-cancer controls, t-test was performed on the intensities between two groups for each peak of the spectrum. Among the peaks of mass spectrum, p-values of multiple t-tests were sorted in ascending order and False Discovery Rates (FDRs) were calculated. The markers are regarded as significant if they satisfy the criterion FDR<0.05. The same analysis was performed on each wavelet coefficient of the approximation profiles at each level.

Logistic Regression and Receiver-Operating Characteristics (ROC) Analyses

The following logistic regression model generates the value of logit based on the intensities of n identified markers.

$$logit = a_1T_1 + a_2T_2 + \cdots + a_nT_n + a_0$$

where T_i and a_i represent the intensity of the n^{th} marker, M_n and the associated coefficient; a_0 is a constant; logit>0 (≤0) indicates the probability of the cancer is higher (lower) than that of the non-cancer. The cancer group is referred to the female non-smokers with adenocarcinoma and the non-cancer group, the non-cancer controls.

The values of logit/marker were sorted in ascending order and cut-off levels were set between any two consecutive values. For each cut-off level, sensitivity and specificity were calculated by checking the values of logit/marker against the actual outcome. Fitted ROC curves were plotted using the pairs of calculated sensitivity and specificity. The discriminatory abilities of logit and individual markers were evaluated and compared using the areas under the curves (AUC).

Protein Prediction

For *n* significant markers, incremental lists of markers, $\{M_1\}$, $\{M_1, M_2\}$, $\{M_1, M_2, M_3\}$, …, $\{M_1, \ldots, M_n\}$, were generated and entered to Mascot MS/MS Ions Search (Matrix Science, publicly available at http://www.matrixscience.com/) one-by-one. The search results showed the protein scores of predicted proteins, corresponding to the markers. Based on $p = 10^{-S/10}$, a protein score, S, can be converted to p-value, which indicates the significance of prediction with $p<0.05$. Only significant prediction was considered for further analysis.

MiRNA Target Prediction

We searched for the miRNAs targeting the mRNAs of the predicted proteins using three representative resources: TargetScan, miRDB, and MicroCosm-Targets [17-21]. The candidate miRNAs were selected based on the support of at least two out of these three databases, in order to improve the prediction accuracy.

Results

Plasma Protein Profiling

On the CM10 proteinchips, 82 peaks of m/z values from 1008.94 to 229279 were detected in the pH 9 fraction. A domain of integer m/z values from 1 to 262144 (=2^{18}) was formed. After rounding the m/z values to the nearest integer and fitting the intensities of these 82 peaks to this domain and the rest with zero intensity, the level 0 spectrum was constructed for each plasma sample. In Figure 1(a-d), the points of the lowest panel show the level 0 spectra averaged over cancer and non-cancer groups in the m/z ranges of 2,000-33,000 and 7,600-9,800 respectively.

Digital Wavelet Transform

Digital wavelet transforms (DWTs) were performed on the level 0 spectrum for each plasma sample using Daubechies wavelet, db1. DWTs were performed for 10 times and the approximation profiles were collected after each DWT. The level i spectrum is referred to the approximation profile collected after the i^{th} DWT. In Figure 1(a-d), the curves show the levels 7-10 spectra averaged over cancer and non-cancer groups in the m/z ranges of 2000-33000 and 7600-9800 respectively.

Peaks Associated with Adenocarcinoma in Female Non-smokers

To compare the intensities between cancer and non-cancer groups, t-test was performed for each peak and the significance was indicated by the value of p. After sorting p in ascending order among 82 peaks, FDR was calculated for each m/z value as shown in Table 1. With the criterion FDR<0.05, we identified 15 m/z values, whose intensities were significantly different between

female non-smokers with adenocarcinoma and healthy controls. We also performed t-test on each wavelet coefficient of the approximation profiles at each level.

Cluster of Approximation Coefficients

The t-test pvalues for the markers, given by the original peaks or wavelet coefficients at levels 0-10, were transformed to the values, log(1/p). The magnitude of log(1/p) directly reflects the significance level of marker's association with adenocarcinoma in female non-smokers. In Figure 1(e,f), a cluster of six markers was consistently observed in the m/z range, 7600-9300, at levels 0, 7-9. Profiles at levels 1-6 exhibiting the same cluster were not shown here for simplicity. These six markers, namely M_1, …, M_6, are highlighted and in bold font in Table 1. At level 10, the clusters started to merge in a global scale and the newly merged cluster centered at m/z value, 7600, covering only one of the 15 significant peaks identified by FDR criterion.

Discriminating Power of Marker Cluster

Logistic regression analysis of the markers, M_1, …, M_6, identified the following model.

$$logit = -0.0147T_1 - 0.134T_2 - 0.0549T_3 + 0.436T_4 - 0.0105T_5 + 0.168T_6 - 1.35$$

where T_1, …, T_6 represent the intensities of markers, M_1, …, M_6 respectively. In Figure 2, the ROC curves of logit, M_1 and M_4 are plotted and compared. It was found that the AUC of logit, 0.876, was substantially higher than that of M_1, 0.787, and M_4, 0.714.

Predicted Protein and Targeting MiRNAs

The incremental lists of six markers were entered to Mascot MS/MS Ions Search one-by-one. The scores and p-values of the search results are shown in Table 2. Only the entries of M_1 and

M_2 gave significant prediction results (p = 0.019953 and 0.039811 respectively), corresponding to 7 unique proteins. The predicted proteins' IDs (gene symbols) are EAW79013.1 (SLC25A36), NP_001305698.1 (MRPL14), CAD62325.1 (EFCAB11), NP_001310608.1 (NRSN2), XP_016885033.1 (SPANXN4), EAW95336.1 (LIMS2), and AFI99088.1 (TRIM77). Through database search, we found 4 miRNAs concurrently targeting mRNAs of 3 predicted proteins. The regulatory modules are illustrated in Figure 3.

Discussion

Mass spectrometry generates high-dimensional data where the sample size is relatively small when it is used for disease detection and biomarker identification. A study proposed an approach for extracting the "common" peaks approximated by Gaussian kernels before the classification using machine learning method, AdaBoost [22]. Such approach reduced the dimension of feature space substantially but it cannot guarantee that the eliminated features are disease-irrelevant. On the other hand, False Discovery Rate (FDR) is widely used for selecting multiple disease-relevant features whilst controlling the inflation of false positives due to multiple comparisons. This study applied FDR control to shortlist 15 peaks of mass spectrum that significantly differentiate the female non-smokers with adenocarcinoma and the non-cancer subjects. Some of the significant peaks may be statistical artifacts caused by m/z axis shift of peaks [11]. DWT was used by Randolph and Yasui to represent the mass spectra and align the peaks in multiple scales. However, it is questionable whether the aligned peaks is biologically meaningful [23]. Therefore, we further performed the Digital Wavelet Transforms (DWTs) to obtain the approximation profiles and analyze the differential expression at multiple scales of m/z domain. The merit of the

proposed method is to determine the “common” peaks based on the consistent pattern of log(1/p) profiles resulting from the differential expression analysis of multi-scale approximation profiles. We observed consistently across 10 scales, from level 0 (original spectrum) to level 9, a pattern covering 6 peaks of the 15 shortlisted. Compared with individual marker features, the logistic regression model combining these 6 marker features exhibited superior ability, 0.876, for discriminating the female non-smokers with adenocarcinoma from the non-cancer subjects.

Protein search of these 6 markers resulted in 7 unique proteins. The potential ceRNA role was found in 3 out of these 7 proteins, whose gene IDs are SLC25A36, EFCAB11 and NRSN2. A study developed an algorithm for exploring the impact of copy number alterations on gene expression and found SLC25A36 as one of the genes with recurrent copy number gains in NSCLC [24]. Another study conducted a genome-wide analysis of gene copy number gains and corresponding gene expression levels in NSCLC patients. SLC25A36 is one of the genes exhibiting significant association ($r>0.7$) between the gene copy number gain and expression level [25]. A genetic factor investigation was conducted on 17 members of a three-generation family with lung cancer susceptibility using whole-exome sequencing. EFCAB11 hosts one of 71 germline mutations that were found in three affected family members but not in the unaffected [26]. It was found in the search of Oncomine Database that NRSN2 was highly expressed in NSCLC compared to normal lung tissues. The overexpression of NRSN2 was also shown in 18 tumor tissues compared with adjacent tissues in NSCLC patients. It was also found in the cell line experiment that the NSCLC cell growth was promoted by NRSN2 through PI3K/Akt/mTOR pathway [27].

We predicted that the ceRNAs of SLC25A36, EFCAB11 and NRSN2 were targeted by miR-206, miR-613, miR-6766-5p and miR-6756-5p. It was found in cell line experiment that miR-206 and

miR-613 as single agents or in combination could sensitize the cisplatin-resistant lung cancer cells to cisplatin treatment through the suppression of 6-phosphogluconate dehydrogenase [28]. In 76.8 % of 56 primary NSCLC tissues, down-regulation of miR-613 was found when compared to the adjacent tissues. It was also observed that the miR-613 mimic induced cell cycle arrest and reduced cell viability and colony formation in NSCLC cell culture, and inhibited tumor growth in xenograft model [29]. To the best of our knowledge, the association of miR-6766-5p or miR-6756-5p with cancer has not been reported in the other studies. A meta-analysis was performed on 2.83 billion raw reads in 737 mouse and human small RNA data sets. The researchers confidently annotated 240 human splicing-derived miRNAs, the vast majority of which are novel genes, including two hosting miR-6766-5p or miR-6756-5p [30]. In a study differentiating 429 breast cancer patients from 895 healthy controls, miR-6756-5p is combined with miR-1246 and miR-8073 to form a neural network cascade model, which detected breast cancer with accuracy, 97.1% [31].

Contemporary lung cancer research has distinguished itself from the traditional one with the unprecedentedly large amount of data and tremendous diagnostic and therapeutic innovations. Data are currently generated in high-throughput fashion with the integration and application of genomics, proteomics, metabolomics and bioinformatics, each of which plays an essential role for molecular biomarker discovery. High-throughput mass spectrometry facilitates better understanding of the disease, including its diagnosis, monitoring, treatment and prognostics. In the era of molecular targeted therapy, specific treatment to the potential target using technologies, such as immunotherapy and RNAi, has been translated from bench to bedside application and thus makes molecular biomarker discovery more meaningful for lung cancer management [32]. The findings of this study could help identifying the novel genetic risk factors, including

ceRNAs and non-coding miRNAs, of an under-explored subgroup of non-small cell lung cancer representing adenocarcinoma in female non-smokers.

Acknowledgments: This project was supported by the Health and Medical Research Fund (HMRF, Project No. 02131026).

References

1. Al-Saleh K, Quinton C, Ellis P. Role of pemetrexed in advanced non-small-cell lung cancer: meta-analysis of randomized controlled trials, with histology subgroup analysis. Current Oncology 2012;19:e9.

2. WHO Department of Gender, Women and Health. Gender in lung cancer and smoking research. World Health Organization 2004. ISBN 92-4-159252-4.

3. Travis WD, Brambilla E, Müller-Hermelink HK, Harris CC. Pathology and Genetics of Tumours of the Lung, Pleura, Thymus and Heart. World Health Organization Classification of Tumours. Lyon: IARC Press 2004. ISBN 92-832-2418-3.

4. Siegfried JM. Women and lung cancer: does oestrogen play a role? Lancet Oncolology 2001, 2:506-513.

5. Sy SM et al. Genetic alterations of lung adenocarcinoma in relation to smoking and ethnicity. Lung Cancer 2003, 41:91-99.

6. Au JS, Cho WC, Yip TT, Law SC. Proteomic approach to biomarker discovery in cancer tissue from lung adenocarcinoma among nonsmoking Chinese women in Hong Kong. Cancer Invest 2008, 26(2):128-135.

7. Roy P, Truntzer C, Maucort-Boulch D, Jouve T, Molinari N. Protein mass spectra data analysis for clinical biomarker discovery: a global review. Briefings in bioinformatics 2011, 12(2), 176-186.

8. Morris JS, Brown PJ, Herrick RC, Baggerly KA, Coombes KR. Bayesian Analysis of Mass Spectrometry Proteomic Data Using Wavelet-Based Functional Mixed Models. Biometrics 2008, 64(2), 479-489.

9. Cho WC. Contribution of oncoproteomics to cancer biomarker discovery. Molecular cancer 2007, 6(1), 1.

10. Cho WC. Research progress in SELDI-TOF MS and its clinical applications. Sheng Wu Gong Cheng Xue Bao. 2006;22(6):871-6.11. Yasui Y, Pepe M, Thompson ML, et al. A data analytic strategy for protein biomarker discovery: profiling of high-dimensional proteomic data for cancer detection. Biostatistics 2003, 4:449–63.

12. Coombes KR, Fritsche J, Clarke C, et al. Quality control and peak finding for proteomics data collected from nipple aspirate fluid by surface-enhanced laser desorption and ionization. Clin Chem 2003, 49:1615–23.

13. Antoniadis A, Bigot J, Lambert-Lacroix S, et al. Nonparametric pre-processing methods and inference tools for analyzing time-of-flight mass spectrometry data. Curr Anal Chem 2007, 3:127–47.

14. Marques, A.C., Tan, J. & Ponting, C.P. Wrangling for microRNAs provokes much crosstalk. Genome Biology 12:132 (2011).

15. Tay, Y. et al. Coding-independent regulation of the tumor suppressor PTEN by competing endogenous mRNAs. Cell 147, 344-357 (2011).

16. Bosson, A.D., Zamudio, J.R. & Sharp, P.A. Endogenous miRNA and Target Concentrations Determine Susceptibility to Potential ceRNA Competition. Molecular Cell 56, 347-359 (2014).

17. Wang F, Chan LW, Law HK, Cho WC, Tang P, Yu J, Shyu CR, Wong SC, Yip SP, Yung BY. Exploring microRNA-mediated alteration of EGFR signaling pathway in non-small cell lung cancer using an mRNA:miRNA regression model supported by target prediction databases. Genomics 104 (2014) 504-11.

18. Wang F, Wong SC, Chan LW, Cho WC, Yip SP, Yung BY. Multiple regression analysis of mRNA-miRNA associations in colorectal cancer pathway. Biomed Res Int 2014 (2014) 676724.

19. Agarwal, V.; Bell, G. W.; Nam, J.-W.; Bartel, D. P., Predicting effective microRNA target sites in mammalian mRNAs. elife 2015, 4.

20. Wong, N.; Wang, X., miRDB: an online resource for microRNA target prediction and functional annotations. Nucleic acids research 2014, 43, (D1), D146-D152.

21. Betel, D.; Koppal, A.; Agius, P.; Sander, C.; Leslie, C., Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. Genome biology 2010, 11, (8), R90.

22. Fushiki T, Fujisawa H, Eguchi S. Identification of biomarkers from mass spectrometry data using a “common” peak approach. BMC Bioinformatics 2006, 7:358. https://doi.org/10.1186/1471-2105-7-358

23. Randolph TW, Yasui Y. Multiscale processing of mass spectrometry data. Biometrics. 2006 Jun;62(2):589-97.

24. Lazar V, et al. Integrated molecular portrait of non-small cell lung cancers. BMC Medical Genomics 2013 6:53.

25. Jabs V, Edlund K, KoÈnig H, Grinberg M, Madjar K, RahnenfuÈhrer J, et al. (2017) Integrative analysis of genome-wide gene copy number changes and gene expression in non-small cell lung cancer. PLoS ONE 12(11): e0187246.

26. Tomoshige K, et al. Germline mutations causing familial lung cancer. Journal of Human Genetics (2015), 1-7.

27. Zhang XY, Kuang JL, Yan CS, Tu XY, Zhao JH, Cheng XS, Ye XQ. NRSN2 promotes non-small cell lung cancer cell growth through PI3K/Akt/mTOR pathway. Int J Clin Exp Pathol. 2015 Mar 1;8(3):2574-81. eCollection 2015.

28. Zheng W, Feng Q, Liu J, Guo Y, Gao L, Li R, Xu M, Yan G, Yin Z, Zhang S, Liu S and Shan C (2017) Inhibition of 6-phosphogluconate Dehydrogenase Reverses Cisplatin Resistance in Ovarian and Lung Cancer. Front. Pharmacol. 8:421. doi: 10.3389/fphar.2017.00421

29. Li D, Li DQ, Liu D and Tang XJ. MiR-613 induces cell cycle arrest by targeting CDK4 in non-small cell lung cancer. Cell Oncol (Dordr) 2016; 39: 139-147.

30. Ladewig E, Okamura K, Flynt AS, Westholm JO, Lai EC. Discovery of hundreds of mirtrons in mouse and human small RNA data. Genome Res. 2012 Sep;22(9):1634-45. doi: 10.1101/gr.133553.111.

31. Cui X, Li Z, Zhao Y, Song A, Shi Y, Hai X, Zhu W. Breast cancer identification via modeling of peripherally circulating miRNAs. PeerJ. 2018 Mar 26;6:e4551. doi: 10.7717/peerj.4551. eCollection 2018.

32. Cho WC, Yip TT, Cheng WW, Au JS. Serum amyloid A is elevated in the serum of lung cancer patients with poor prognosis. Br J Cancer. 2010 Jun 8;102(12):1731-5. doi: 10.1038/sj.bjc.6605700. Epub 2010 May 25.

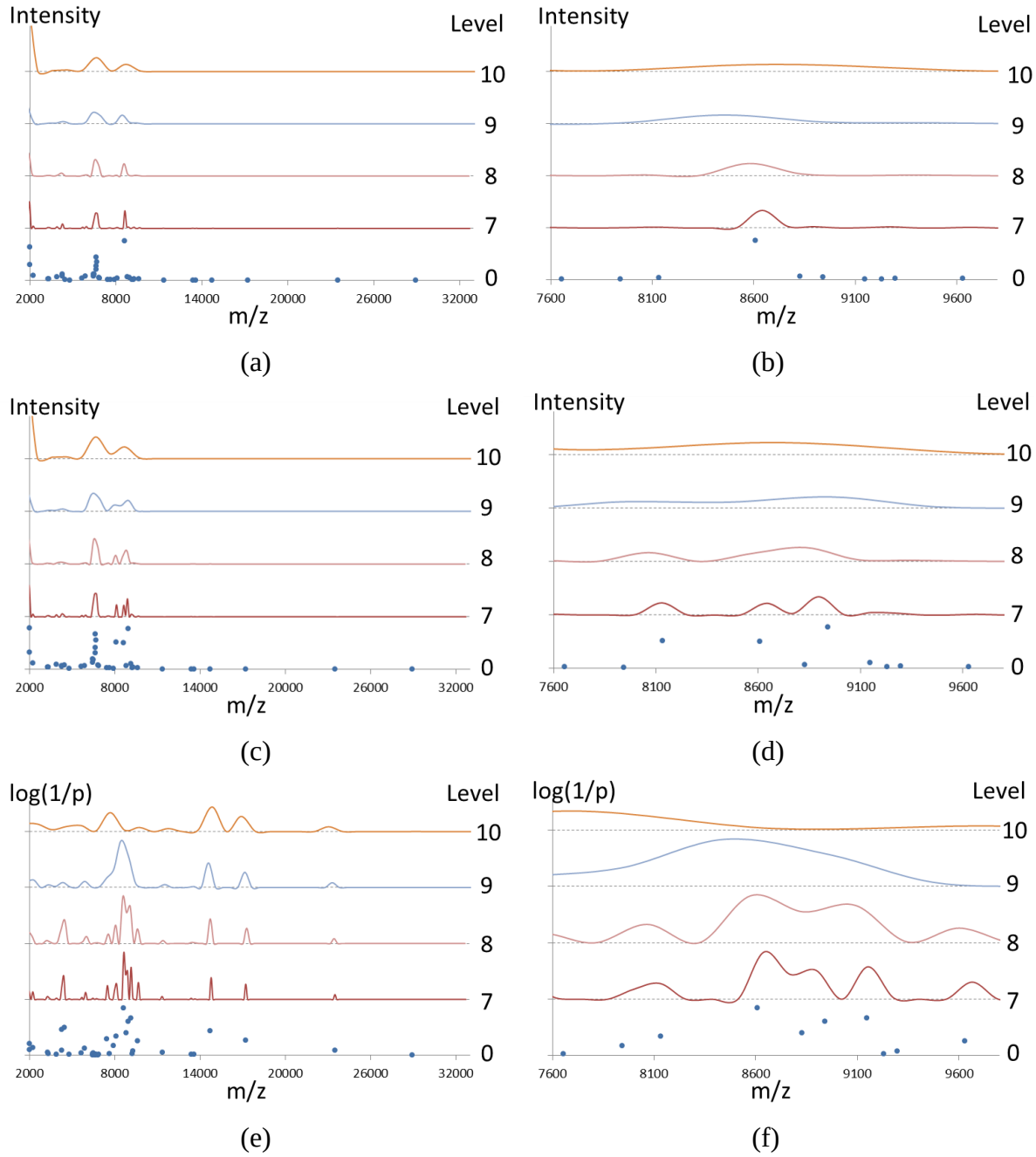

Figure 1 (a,c,e) Analytical results in the range of m/z ratio 2000-33000. (b,d,f) Analytical results in the range of m/z ratio 7600-9800. (a-d) Points at level 0 represent original mass spectrum and the curves at levels 7-10 represent the approximation profiles obtained by DWTs for 7-10 times respectively. (a,b) Intensities are averaged over plasma samples of female non-smokers with adenocarcinoma (cancer group). (c,d) Intensities are averaged over plasma samples of non-cancer controls (non-cancer group). (e,f) The value of log(1/p) against m/z value, where p represents the p-value of t-test comparing cancer and non-cancer groups.

Table 1 The first 15 Peaks selected by the criterion, FDR<0.05, from the list in ascending order of p value. The cluster concurrently supported by levels 7-9 approximation profiles covers six of these peaks, M_1, …, M_6 (highlighted and in bold font).

Marker	m/z	p	FDR
M_1	**8608.438**	4.13×10^{-9}	1.78×10^{-7}
M_2	**9146.840**	2.12×10^{-7}	4.56×10^{-6}
M_3	**8942.507**	1.01×10^{-6}	1.45×10^{-5}
M_7	4468.357	1.07×10^{-5}	0.000115
M_8	4306.476	2.37×10^{-5}	0.000204
M_9	14733.68	3.89E-05	0.000279
M_4	**8826.615**	0.000102	0.000627
M_5	**8133.093**	0.000398	0.002142
M_{10}	7465.883	0.001361	0.006511
M_{11}	17230.83	0.001888	0.008129
M_{12}	9627.812	0.002423	0.009484
M_{13}	2028.677	0.008516	0.028204
M_{14}	1502.769	0.007884	0.028287
M_{15}	1900.81	0.010255	0.031537
M_6	**7945.342**	0.016744	0.04806

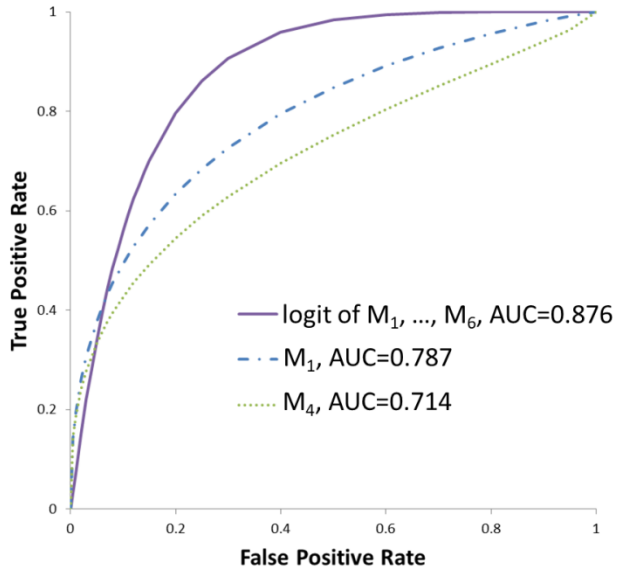

Figure 2 Fitted ROC curves and AUCs of M_1 (dash dotted), M_4 (dotted) and logit combining M_1, …, M_6 (solid).

Table 2 Summary of protein search of six markers

Entry	Number of Proteins	Score	p-value
M_1	12	17	0.019953
M_1, M_2	12	14	0.039811
M_1, M_2, M_3	12	12	0.063096
M_1, M_2, M_3, M_4	12	11	0.079433
M_1, M_2, M_3, M_4, M_5	13	10	0.100000
M_1, M_2, M_3, M_4, M_5, M_6	12	9	0.125893

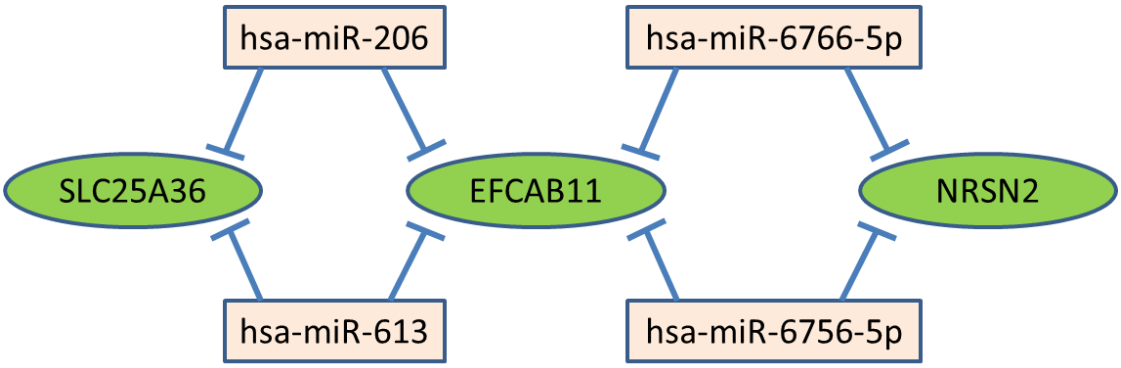

Figure 3 MiRNA regulatory modules