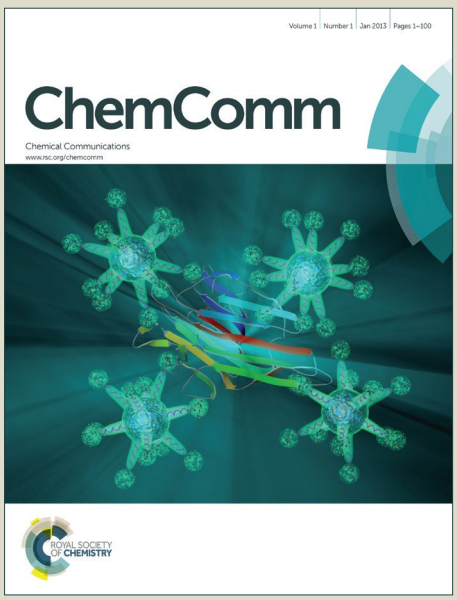
ChemComm

Accepted Manuscript



ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/chemcomm

ARTICLE TYPE www.rsc.org/xxxxxx | XXXXXXXX

Multi-dimensional On-particle Detection Technology for Multi-category Disease Classification

Jie Tan[a], Xiaomin Chen[a], Guansheng Du[a], Qiaohui Luo[a], Xiao Li[a], Yaqing Liu[a], Xiao Liang[b]* and Jianmin Wu[a]*

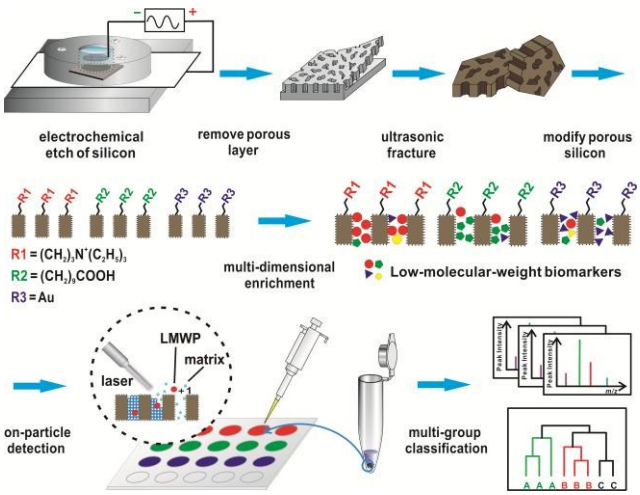
Received (in XXX, XXX) Xth XXXXXXXXX 200X, Accepted Xth XXXXXXXXX 200X
First published on the web Xth XXXXXXXXX 200X
DOI: 10.1039/b00000000x

Serum peptide profile contains important bio-information which may help disease classification. The motivation of this work is to take advantage of porous silicon microparticles with multiple surface chemistries to reduce the loss of peptide information and simplify the sample pretreatment. We developed a multi-dimensional on-particle MALDI-TOF technology to acquire high fidelity and cross-reactive molecular fingerprints for mining disease information. Peptide fingerprint of serum samples from colorectal cancer patients, liver cancer patients and healthy volunteers were measured with this technology. The featured mass spectral peaks can successfully discriminate and predict the multi-category disease. Data visualization for future clinical application was also demonstrated.

Human diseases are biological states caused by multiple components of perturbed pathways and regulatory networks rather than individual failing components.[1] The constituents of serum sample are derived from diverse biological processes and many of them are logically linked with the complex physiological changes associated with disease.[2, 3] Therefore, to reveal the complexity of molecules variation occurred in disease progress, innovations in analytical technologies are urgently needed. Serum peptide profile has been known as the "fingerprint library" which containing important bio-information.[4] By analyzing and comparing differences in the expression of serum peptides between target disease and normal healthy population, multiple different-expressed serum peptides can be therefore mapped out.[5] However, acquiring serum peptide fingerprints with high fidelity is quite difficult, as the concentration and nature of serum peptides have a large dynamic range. Numerous technologies to capture and enrich serum peptides have been developed in recent years. For example, magnetic beads,[6-8] porous particles,[9-13] and carbon-based nanomaterial have been used for peptide enrichment.[14-17] Following by the sample pretreatment steps, mass spectroscopic technologies are usually required to profile or identify those captured peptides. Multi-dimensional liquid chromatography coupled with electrospray ionization (ESI) mass spectrometry (MS) is undoubtedly the most powerful technology to acquire high authentic peptide information owing to its strong separation capability. However, this technology is not appropriate for large scale clinical screening. Matrix-assisted laser desorption/ ionization time-of-flight mass spectrometry (MALDI-TOF-MS) is a fast and high throughput technology for analyzing metabolites/ peptides in clinical samples. But the resolution of MALDI-TOF is significantly lower than that of LC-MS due to the lack of chromatographic separation step. The complexity of biological samples will impose severe interference and thereby decrease the authenticity of molecular pattern. Therefore, innovation in sample pretreatment and ionization technology is urgently needed for the MALDI-TOF analysis. Surface-enhanced laser desorption/ionization (SELDI) chip has been commercialized as a suitable tool for serum peptide profiling, since a SELDI chip can capture a fraction of peptides based on their specific surface chemistry.[18, 19] However, the surface of SELDI chip is none porous, which cannot exclude large proteins. Meanwhile, the efficiency of laser desorption/ ionization of SELDI chip was not improved. In the present work, we developed an "on-particle" detection technology to profiling peptide fingerprint. Porous silicon microparticles (PSMPs) with ordered pore channel and accurate cutoff pore size can selectively capture peptides below a specific molecular weight. Furthermore, peptides captured on PSMPs do not need to be eluted from the particles. The captured molecules can be ionized on the PSMPs by MALDI technology with enhanced peak intensity . In order to increase the capacity of molecular information, PSMPs were functionalized with different surface chemistry, which will have a cross-reactive affinity for peptides with different pI, hydrophobicity, and ligand complexity. As a proof of concept work, we use three types of surface chemistry to demonstrate the advantage of the multidimensional on-particle (MDOP) detection technology. The porous silicon with quaternary ammonium groups (Q-pSi) and carboxylation porous silicon (C-pSi) were designed to capture negative and positive charged peptides, respectively, while the Au coated PSMPs (Au-pSi) target to peptides with rich cysteine residues or other metal affinity group. The goal of this work is to take the advantage of MDOP detection technology to rapidly acquire overlaid peptide profiles for reliable disease prediction. Serum samples from colorectal cancer (CRC) patients, liver cancer patients and healthy volunteers were tested with the MDOP detection technology. The results demonstrated the three groups of clinical samples can be successfully discriminated based on the featured mass spectral peaks obtained from different surface chemistry.

The whole procedure for peptide profiling based on the MDOP technology is illustrated in Scheme 1. The detail method for the particle preparation and surface modification is described in supporting information. SEM images of the PSMPs surface before (see Fig. S1A, ESI†) and after surface modification (see Fig. S1B~D, ESI†) indicate that the uniform pore structure retained after the particles was modified. The particle size was 30 - 50 μm measured by optical microscopy (see Fig. S2, ESI†). The surface chemistry of PSMPs was confirmed by X-ray

photoelectron spectroscopy (see Fig. S3, ESI†). The fresh etched PSMPs sample shows Si 2p, Si 2s, and a low-intensity C 1s peak, which is mainly due to adventitious carbon contamination. The Si 2p peak was observed at about 103 eV due to the partial oxidation of PSMPs. For the carboxyl surface, the relative intensities of the C 1s and O 1s lines increase concomitantly with a decrease in the intensity of Si 2p and Si 2s lines. As for the Q-pSi sample, a low-intensity of N 1s peak at 402.0 eV was observed. For the Au-pSi, a high-intensity of Au 4f double-humped peak is shown at 83.8 and 87.5 eV. The surface chemistry of Q-pSi and C-pSi were also proved by diffuse reflectance infrared Fourier transform spectral (see Fig. S4, ESI†).



Scheme 1 Schematic procedures for the preparation and application of PSMPs in serum peptide profiling. 1) Porous Si film was prepared by an electrochemical etch of a crystalline silicon wafer. 2) The porous layer was removed and fractured into microparticles in an ultrasonic bath. 3) The particles were chemically modified with different functional group to impart affinity for peptides and to stabilize the porous nanostructure. 4) After incubation and washing, the microparticles laden with peptides were isolated and spotted on a MALDI plate and then subjected to on-particle detection with MALDI-TOF MS.

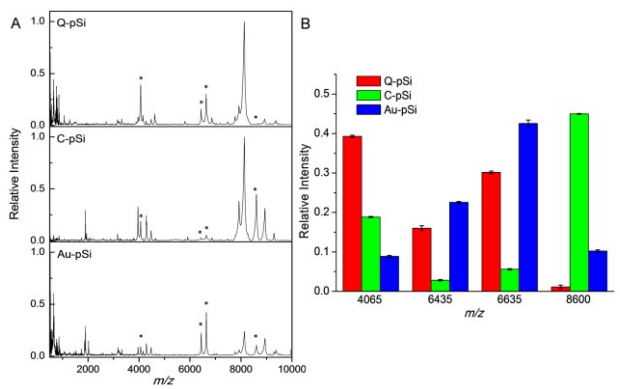


Figure 1 An example that shows the on-particle cross-reactive affinity of PSMPs with different surface chemistry. (A) The normalized MS spectrum from the same serum sample enriched and detected on PSMPs with three different surface functionalities. (B) Relative peak intensity of selected four peaks (*m/z* ratio: 4065, 6435, 6635, and 8600) marked in A.

To confirm that PSMPs with different surface chemistry could have a cross-reactive affinity for different peptides, the mixture of RGDC (pI = 6.2, MW = 450 Da), TP2 (pI = 10.8, MW = 1.49 kDa), and insulin (pI = 5.3, MW = 5.8 kDa) were detected with the MDOP technology. RGDC peptide was chosen for its cysteine group that can be strongly bonding with Au-pSi. While TP2 and insulin was selected as positive and negative charged peptide in physiological condition, respectively. The mass spectra indicate that the relative peak intensity of TP2 and insulin is apparently different on C-pSi and Q-pSi, respectively(see Fig. S5B, ESI†). The results are complied with the discipline of electrostatic interaction mechanism. When Au-pSi was used, both the TP2 and insulin peaks increased in relative to the control experiment due to the localized surface plasmon resonance effect of Au layer. However, the intensity of RGDC peak on Au-pSi was lower than that on the other materials (see Figure S6-S7, ESI†). The special phenomena observed on Au-pSi could be ascribed to the strong interaction between Au and sulfhydryl group in RGDC. The too strong Au-S bonds may hinder the peptides from ionization on the Au-pSi, resulting in the reduction of signal intensity. Nevertheless, these model peptides confirmed that cross-reactive affinity exists on different surfaces, which is sensitive to the chemical nature of peptides.

For real serum sample analysis, the advantage of on-particle detection methods is significant. The PSMPs with different pore size display excellent size exclusion effect in serum analysis due the well-controlled pore size (see Fig. S8, ESI†). The PSMPs can eliminate the interference from larger proteins (see Fig. S8 , ESI†). On the contrary, when the serum sample was directly spotted on MALDI plate without PSMPs pretreatment, only few peaks can be detected (see Fig. S9A, ESI†). The on-particle detection method can also avoid the dilution of analytes during the elution step. Meanwhile, the semiconductor nano-structure may help to enhance the energy transfer efficiency within the particles, leading to the improved signal intensity. To clarify the advantage of on-particle method, serum peptides enriched by PSMPs were eluted using 5 μL of the solution (50 % ACN aqueous solution + 0.1 % TFA), which was subsequently subjected to conventional MALDI detection. Compared with on-particle detection method (see Fig. S9B, D, F, ESI†), the peak number and signal to noise ratio found in the eluted sample decreased significantly (see Fig. S9C, E, G, ESI†). The results may not only be caused by the dilution of peptides during the elution step,[20] but also due to the irreversible adsorption and low ionization efficiency of peptides on a conventional MALDI plate.

To acquire more peptides information, a serum sample was incubated with the three types of PSMPs, which were subsequently detected with the on-particle method. PSMPs with a specific surface chemistry can generate high quality and distinguishable peptide profile (Fig. 1A), since the PSMPs can effectively exclude the interference from large molecular weight protein and sequester a subset of peptide according to their surface chemistry. To display the cross-reactive characteristics of peptide fingerprint obtained with the MDOP technology, four peaks was selected and their relative signal intensity detected on three types of PSMPs was compared. On the Q-pSi, a high signal intensity at *m/z* = 4065 was observed, whereas on the C-pSi surface, with the peak at *m/z* = 8600 increased significantly. The Au-pSi produced a larger signal at the *m/z* = 6435 and 6635 (Fig. 1B). The results suggest that surface chemistry exert a great effect on the signal intensity,

inferring that the complementary peaks found on different surface chemistry can significantly increase the capacity of molecular information. Theoretically, the three-dimensional surface can produce triple peak information compared to that on the one-dimensional surface. The expanded peak information may help to enhance the ability of disease prediction and diagnosis based on molecular fingerprint.

Table 1 The number of m/z values that obtained after student's t test ($p < 10^{-7}$).

Groups	Q-pSi	C-pSi	Au-pSi
CRC vs health	N/A	7 peaks	9 peaks
Liver cancer vs health	12 peaks	N/A	11 peaks
CRC vs liver cancer	25 peaks	5 peaks	N/A

Table 2 Biomarker candidates those calculated using ClinProTools and matched with the peaks screened by student's t test.

Groups	Q-pSi	C-pSi	Au-pSi
CRC vs health	3444	***535*** ***2210*** 3161	***8640*** ***8820*** 8700
Liver cancer vs health	***4065***[a] ***8127*** 2710	3242 4286	***1530*** 8820 3170
CRC vs liver cancer	8127 4065 ***8600***	***8129*** 3242 3161	8700
Total peaks	**5 peaks**	**6 peaks**	**5 peaks**

[a] The peaks highlighted in bold italic type are selected to be the final feature subset.

Although, in some binary classification cases, one type of disease sample can be successfully discriminated from health control just based on a subset of peptides enriched with solo surface chemistry.[21, 22] Nevertheless, in practical clinical diagnosis, multi-category disease classification is always encountered. To deal with the multi-category problem in cancer prediction, a total of 72 serum samples were analyzed with MDOP detection technology. Among them, 24 were from patients with CRC, 24 from liver cancer patients, and 24 from healthy volunteers. Patients were diagnosed according to the standard diagnosis criteria[23,24]. Demographic features of all the patients and healthy volunteers are provided in **Table S1**. The raw spectra were processed using FlexAnalysis software (Bruker Daltonics Corp.) to eliminate the impact of background and remove noise interference (see Fig. S10, ESI†). With these data in hand, disease information can be mined with statistical method.

The data analysis here involved three stages: (i) peak detection and alignment using student's t test; (ii) selection of differently expressed peaks among the three groups and validate the classification by cluster analysis; (iii) sensitivity and accuracy examination for the predictive model generated from serum profiling data with neural network software (see ESI†). In the first stage, for each pair of classes, the number of the selected mass spectral peaks (p value $< 10^{-7}$) was listed in **Table 1**. We found that three groups (health, CRC and liver cancer) could not be simultaneously distinguished in one type of surface chemistry. For example, when using Q-pSi only, the CRC patients and health volunteers could not be discriminated. Therefore, for the discrimination of multi-category disease, multi-dimensional detection was so necessary that the disease information hidden in the serum profile can be revealed. We subsequently used ClinProTools to find the features with the best accuracy from the obtained feature subset. The spectral overlays of the 16 most distinct mass spectral peaks listed in **Table 2** were shown in Figure S11. Combined with the selected peaks screened by student's t test, a final feature subset was sorted out (shown in bold italic in **Table 2**) according to the order of p value given by ClinProTools.

If the features selected from one type of PSMPs were used only, a misclassified result was observed on the cluster map (see Fig. S12, ESI†). In contrast, by using all the 9 features selected from 3 types of surface chemistry, a high accurate classification can be obtained (Fig. 2). The cluster tree indicates that serum samples from the CRC patients, liver cancer patients and those from the healthy controls appear in separate clusters with only two exception (healthy volunteers No. 27 and No. 35 with black branches were misclassified as liver cancer patient and CRC patient, respectively). It is important to note that the feature selection was based on prediction accuracy, not on the order of index as shown in **Table S2-S4**. The 9 peaks were further set as the input layer of Artificial Neural Networks to generate a prediction model. Validation using training data was performed to estimate the error rates of this model (see Fig. S13A, ESI†). The percentage of correctly classified training and test data indicated that, for CRC group, the specificity is 96 %, the sensitivity is 100 %, and the false positive predictive value is 4.2 %, respectively (Fig. S13B, ESI†). Furthermore, the serum profiles allow us to visualize differences in the molecular fingerprints of multi-category diseases. The radar chart shows that the traces of

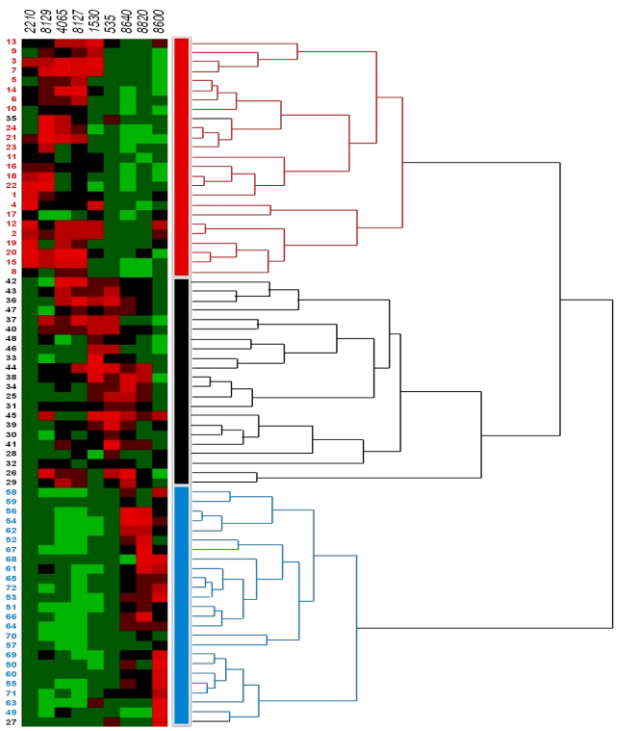


Figure 2 Cluster analysis and heat map view of mass spectra data selected from NPMPs with different surface chemistry. Serum samples are from CRC patients (No. 1~24), healthy volunteers (No. 25~48) and liver cancer patients (No. 49~72). The red, black and green color in the heat map represents high, medium, and low peak intensity, respectively. Branches and clusters in the cluster tree are color-coded: CRC patients in red; healthy person controls in black and liver cancer patients in blue.

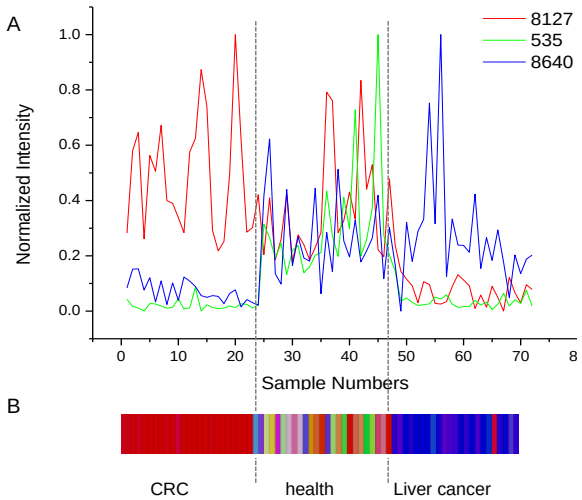


Figure 3 Digital color maps of 3 most significantly different peaks. (A) The normalized intensities of the selected peaks, 8127 (red line), 535 (green line), and 8640 (blue line); (B) Digital color map created by the selected peaks in RGB color mode, with the R (8127), G (535), and B (8640) values set as the normalized peak intensity of each patients using MATLAB software.

the three groups are significantly different from each other (see Fig. S14, ESI†). By picking out the feature peak ranking first on each type of surface chemistry (see **Table S2-S4**, ESI†), three most significantly different peaks were obtained. The *m/z* of the three candidates were 8127 (from Q-pSi), 535 (from C-pSi), and 8640 (from Au-pSi). The normalized intensities of the three peaks (Fig. 3A) were set as R (8127), G (535), and B (8640) values in RGB color mode using MATLAB software. A digital color map could allow to directly "see" the disease information hidden in serum (see Fig. 3B). We can suggest that one shown in "red" may be a CRC patient, while, the one shown in "blue" have a great chance to be a liver cancer patient. Thus, this data visualization approach could be conveniently used in diagnosis and distinguish multi-category diseases in the future.

In summary, we have developed a MDOP detection technology to acquire high fidelity and expanded molecular information for disease mining. With the assistance of artificial intelligence learning algorithms, a combination of the selected mass spectral peaks from different surface chemistry can be selected. These candidate biomarkers display high accuracy for the multi-category disease classification. This method does not care about identities of each peak candidate, but rely on the differences in the pattern of selected peaks, which can be either statistically classified or visualized with chart and color. By integrating molecular fingerprints data with other clinical data, early screening, diagnosis, and management of cancer disease will be potentially realized. We believe this is a promising area in translational medicine.

This work is supported by National Science Foundation of China (Grant No. 21275218) and Ministry of Education of China (Grant No. 20130101110048)

Notes and references

[a] *Institute of Microanalytical System(IMAS), department of chemistry, Zhejiang University, Hangzhou, China.. Fax: 86 571 88273572; Tel: 86 57188273496; E-mail: wjm-st1@zju.edu.cn*
[b] *Department of General Surgery, Sir Run Run Shaw Hospital School of Medicine, Zhejiang University,Hangzhou,310058; E-mail: srrshlx@163.com*

† Electronic Supplementary Information (ESI) available:. See DOI: 10.1039/b000000x/

1. R. S. Wang, B. A. Maron and J. Loscalzo, *Wires. Syst. Biol. Med.*, 2015, **7**, 141-161.
2. R. R. Drake, L. Cazares and O. J. Semmes, *Proteomics Clin. Appl.*, 2007, **1**, 758-768.
3. S. Ray, P. J. Reddy, R. Jain, K. Gollapalli, A. Moiyadi and S. Srivastava, *Proteomics*, 2011, **11**, 2139-2161.
4. L. A. Liotta, M. Ferrari and E. Petricoin, *Nature*, 2003, **425**, 905-905.
5. R. Srinivasan, J. Daniels, V. Fusaro, A. Lundqvist, J. K. Killian, D. Geho, M. Quezado, D. Kleiner, S. Rucker, V. Espina, G. Whiteley, L. Liotta, E. Petricoin, S. Pittaluga, B. Hitt, A. J. Barrett, K. Rosenblatt and R. W. Childs, *Exp. Hematol.*, 2006, **34**, 796-801.
6. S. S. Liu, H. M. Chen, X. H. Lu, C. H. Deng, X. M. Zhang and P. Y. Yang, *Angew. Chem. Int. Ed.*, 2010, **49**, 7557-7561.
7. Q. Liu, J. Shi, M. Cheng, G. Li, D. Cao and G. Jiang, *Chem. Commun.*, 2012, **48**, 1874-1876.
8. S. Y. Huang and Y. C. Chen, *Anal. Chem.*, 2013, **85**, 3347-3354.
9. A. Luchini, D. H. Geho, B. Bishop, D. Tran, C. Xia, R. L. Dufour, C. D. Jones, V. Espina, A. Patanarut, W. Zhou, M. M. Ross, A. Tessitore, E. F. Petricoin and L. A. Liotta, *Nano Lett.*, 2008, **8**, 350-361.
10. D. Tamburro, C. Fredolini, V. Espina, T. A. Douglas, A. Ranganathan, L. Ilag, W. Zhou, P. Russo, B. H. Espina, G. Muto, E. F. Petricoin, 3rd, L. A. Liotta and A. Luchini, *J. Am. Chem. Soc.*, 2011, **133**, 19178-19188.
11. H. Qin, P. Gao, F. Wang, L. Zhao, J. Zhu, A. Wang, T. Zhang, R. Wu and H. Zou, *Angew. Chem. Int. Ed.*, 2011, **50**, 12218-12221.
12. T. M. Guinan, P. Kirkbride, C. B. Della Vedova, S. G. Kershaw, H. Kobus and N. H. Voelcker, *Analyst*, 2015, **140**, 7926-7933.
13. S. J. P. McInnes, C. T. Turner, S. A. Al-Bataineh, M. J. I. A. Leccardi, Y. Irani, K. A. Williams, A. J. Cowin and N. H. Voelcker, *J. Mater. Chem. B*, 2015, **3**, 4123-4133.
14. M. Najam-ul-Haq, M. Rainer, C. W. Huck, P. Hausberger, H. Kraushaar and G. K. Bonn, *Anal. Chem.*, 2008, **80**, 7467-7472.
15. K. Shrivas and H. F. Wu, *J. Mass. Spectrom.*, 2010, **45**, 1452-1460.
16. G. Fang, W. Gao, Q. Deng, K. Qian, H. Han and S. Wang, *Anal. Biochem.*, 2012, **423**, 210-217.
17. P. Yin, N. Sun, C. Deng, Y. Li, X. Zhang and P. Yang, *Proteomics*, 2013, **13**, 2243-2250.
18. E. F. Petricoin and L. A. Liotta, *Curr. Opin. Biotechnol.*, 2004, **15**, 24-30.
19. J. Koopmann, Z. Zhang, N. White, J. Rosenzweig, N. Fedarko, S. Jagannath, M. I. Canto, C. J. Yeo, D. W. Chan and M. Goggins, *Clin. Cancer Res.*, 2004, **10**, 860-868.
20. J. Tang, Y. Liu, D. Qi, G. Yao, C. Deng and X. Zhang, *Proteomics*, 2009, **9**, 5046-5055.
21. E. F. Petricoin, A. M. Ardekani, B. A. Hitt, P. J. Levine, V. A. Fusaro, S. M. Steinberg, G. B. Mills, C. Simone, D. A. Fishman, E. C. Kohn and L. A. Liotta, *Lancet*, 2002, **359**, 572-577.
22. R. Terracciano, M. Preiano, G. P. Palladino, G. E. Carpagnano, M. P. Barbaro, G. Pelaia, R. Savino and R. Maselli, *Proteomics*, 2011, **11**, 3402-3414.
23. R. W. Burt, J. A. Cannon, D. S. David, D. S. Early, J. M. Ford, F. M. Giardiello, A. L. Halverson, S. R. Hamilton, H. Hampel, M. K. Ismail, K. Jasperson, J. B. Klapman, A. J. Lazenby, P. M. Lynch, R. J. Mayer, R. M. Ness, D. Provenzale, M. S. Rao, M. Shike, G. Steinbach, J. P. Terdiman, D. Weinberg, M. Dwyer and D. Freedman-Cass, *J. Natl. Compr. Canc. Netw.*, 2013, **11**, 1538-1575.
24. A. B. Benson, T. A. Abrams, E. Ben-Josef, P. M. Bloomston, J. F. Botha, B. M. Clary, A. Covey, S. A. Curley, M. I. D'Angelica, R. Davila, W. D. Ensminger, J. F. Gibbs, D. Laheru, M. P. Malafa, J. Marrero, S. G. Meranze, S. J. Mulvihill, J. O. Park, J. A. Posey, J. Sachdev, R. Salem, E. R. Sigurdson and C. Sofocleous, *J. Natl. Compr. Canc. Netw.*, 2009, **7**, 350-391.