Analyst

PAPER

View Article Online
View Journal | View Issue

Published on 11 November 2013. Downloaded by Université Laval on 28/06/2014 13:07:50.

Cite this: *Analyst*, 2014, **139**, 455

Evaluation of Raman spectroscopy for diagnosing EGFR mutation status in lung adenocarcinoma

Lei Wang,[a] Zhipei Zhang,[a] Lijun Huang,[a] Weimiao Li,[a] Qiang Lu,[a] Miaomiao Wen,[a] Ting Guo,[b] Jinhai Fan,[c] Xuejiao Wang,[a] Xinwei Zhang,[a] Jixiang Fang,[c] Xiaolong Yan,[a] Yunfeng Ni*[a] and Xiaofei Li*[a]

Somatic mutations in the epidermal growth factor receptor (EGFR) gene were associated with sensitivity to small molecule tyrosine kinase inhibitors for patients with lung adenocarcinomas. In this research, EGFR mutation status was analyzed by DNA sequencing in 153 lung adenocarcinoma tissues. Of these, 75 samples carried EGFR mutations, including 29 with E19del mutation, 33 with L858R mutation, 7 with T790M mutation, and 6 with multiple mutations. Then, 30 samples including 10 with wild type (wt)-EGFR, 10 with L858R and 10 with E19del mutations were selected for Raman and immunohistochemistry (IHC) analyses. After removing the spectra from normal and non-mutated regions, 441 spectra were found appropriate for Raman analysis: 149 from wt-EGFR, 135 from L858R and 157 from E19del mutations. The Raman peaks at 675, 1107, 1127 and 1582 cm^{-1} were significantly increased in wt-EGFR tissues which can be attributed to specific amino acids and DNA. The Raman peaks at 1085, 1175 and 1632 cm^{-1} assigned to arginine were slightly increased in L858R tissues. The overall intensity of E19del tissues was weaker than others due to exon 19 deletion that removes residues 746–750 of the expressed protein. Principal component analysis (PCA) and support vector machine (SVM) were applied for final prediction. The PCA/SVM algorithm yielded an overall accuracy of 87.8% for diagnosing L858R or E19del from wt-EGFR tissues. Finally, RS provides a simple, rapid and low-cost procedure based upon the molecular signatures for predicting EGFR mutation status.

Received 19th July 2013
Accepted 27th October 2013

DOI: 10.1039/c3an01381b

www.rsc.org/analyst

1. Introduction

Lung cancer is the leading cause of cancer deaths in the world, with a five year survival rate of 15%.[1] Non-small-cell lung cancer (NSCLC) accounted for 80% of the lung cancer cases and is further categorized into the specific sub-types: adenocarcinoma, squamous cell carcinoma, and large cell carcinoma.[2] The epidermal growth factor receptor (EGFR) is a member of the receptor tyrosine kinase family, which includes the erbB family. Somatic mutations in the EGFR gene were detectable in 50% of Asian patients with lung adenocarcinomas and were associated with sensitivity to the small molecule tyrosine kinase inhibitors (TKIs), gefitinib and erlotinib.[3,4] The most common NSCLC-associated EGFR mutations are deletions in exon 19 (E19del in 45% of patients) and the point mutation in exon 21 (L858R in 40% of patients). The presence of an EGFR mutation is a robust predictor of improved progression-free survival with TKIs.[4]

Based on these clinical findings, evaluating the EGFR mutation status is thought to be highly important to guide treatment decisions in lung adenocarcinoma. Direct DNA sequencing of PCR-amplified genomic DNA is the preferred method to detect EGFR mutation status in patient tumor tissues although fluorescence *in situ* hybridization (FISH) and immunohistochemistry (IHC) have been used.[3,5,6] However, the adoption of these techniques as clinical tests suffered from high costs of equipment and reagents, technical difficulties in performing the assay, and length of the procedure.[5] In addition, the DNA obtained from paraffin specimens of standard biopsies is generally not sufficient or is of poor quality for DNA sequencing.[5,7]

Raman spectroscopy (RS) is a real-time optical technique for probing molecular vibrations to provide specific information about changes in the biomolecular structure and conformation within diseased tissues. A great deal of research has revealed the superiority of RS for detecting pathological sub-types of NSCLC, evaluating chemotherapeutic or radiotherapeutic response of lung cancer cells, assessing the margin of lung tissue, and

[a]*Department of Thoracic Surgery, Tangdu Hospital, The Fourth Military Medical University, No 1, Xinsi Road, Xi'an, Shaanxi Province, 710038, China. E-mail: lxfchest@fmmu.edu.cn (X.L.); niyunfng@fmmu.edu.cn (Y.N.); Fax: +86-29-83325811; Tel: +86-13909111010*

[b]*Department of Pain Management, Tangdu Hospital, The Fourth Military Medical University, Xi'an, Shaanxi, 710038, China*

[c]*Department of Urology, First Affiliated Hospital of Medical College, Xi'an Jiaotong University, Xi'an, Shaanxi, 710061, China*

[d]*State Key Laboratory for Mechanical Behavior of Materials, MOEKey Laboratory for Nonequilibrium Synthesis and Modulation of Condensed Matter, School of Science, Xi'an Jiaotong University, Shaan Xi, 710049, China*

Published on 11 November 2013. Downloaded by Université Laval on 28/06/2014 13:07:50.

increasing the detection rate of bronchoscopy *in vivo*.[8–18] RS was combined with some diagnostic algorithm, for example principal component analysis or linear discriminant analysis, to identify the pathological types of NSCLC with accuracy above 80%,[8–10] even to predict early postoperative cancer recurrence with 73% sensitivity and 74% specificity.[11] Magee *et al.*[12] utilized RS to analyze induced sputum for detection of molecular profiles associated with lung cancer, differentiating the sputum of lung cancer patients from "at-risk" subjects with 90% sensitivity and 60% specificity, while "at-risk" subjects were differentiated from healthy control subjects with 90% sensitivity and 93% specificity. Some researchers[13,14] were able to elucidate the Raman spectral changes of lung cancer cells (A549 or Calu-1) after exposing to cisplatin or gemcitabine, predicting the chemotherapeutic response at a molecular level in these cells. Others irradiated the lung cancer cells (H460) with a single fraction of 6 MV photons, demonstrating radiation-induced Raman spectral changes due to changes in cellular concentration of aromatic amino acids, conformational protein structure and certain nucleic acids.[15] Raman spectroscopic mapping analysis can provide a spectral pseudocolor map with precise linking of histological structure, and this method has the potential to assess the positive margin of lung cancer during surgery.[16,17] Short *et al.*[18] have shown that adding RS to current white light bronchoscopy/autofluorescence bronchoscopy improves sensitivity and specificity for the detection of preneoplastic lesions *in vivo*.

These studies demonstrate that RS is capable of detecting and distinguishing the unique molecular signatures of lung cancer, which could prove beneficial as a noninvasive, rapid screening modality. Thus, in this study, we analyzed the mutation type of EGFR using both RS and DNA sequencing of PCR-amplified exon sequences in fresh lung adenocarcinoma tissues, and then constructed the principal component analysis/support vector mechine (PCA/SVM) algorithm for further diagnosis.

2. Materials and methods

2.1 Cell culture and sample preparation

Human lung adenocarcinoma cell lines A549, H1975 and H1650 were obtained from ATCC (Manassaa, VA, USA), and cultured in RPMI-1640 (Gibco, USA) medium with 10% fetal bovine serum (Hyclone, USA) at 37 °C with 5% CO_2 in humidified incubators. Cells were seeded in a 10 cm culture dish at a density of 2×10^4 cells per cm^2 and grown to 80% confluence in growth medium. The growth medium was then aspirated and the remaining cells washed with 4 °C phosphate buffered saline (PBS) 3 times. The cells were dispersed from their culture dish using 0.25% trypsin–EDTA (Sigma-Aldrich, USA), transferred to a 15 mL universal tube and centrifuged at $92 \times g$ (4 °C) for 5 min. After centrifugation again, the sediment was used to extract the DNA and RNA as per standard protocols (E.Z.N.A.® Tissue DNA Kit and Total RNA Kit, Omega, USA). DNA and RNA were dispersed in water at a concentration of approximately 1 mg mL^{-1}. The whole cell lysate was prepared using RIPA buffer (Sigma-Aldrich, USA) containing proteinase inhibitors. After washing with PBS, 300 μL RIPA buffer was added to the10 cm culture dish, then incubated on ice for 5 minutes and the cells were scraped. The supernatants were centrifuged for 30 minutes at $13\,000 \times g$ (4 °C). Supernatants were collected as total protein at a concentration of approximately 50 mg mL^{-1}. All three compounds were deposited on CaF_2 substrates for drying (4 °C) before spectral acquisition.

2.2 Specimen collection and preparation

153 NSCLC patients who underwent surgery with a histopathological diagnosis of adenocarcinoma were included in this study. All of these patients were admitted in the Tangdu Hospital of the Fourth Military Medical University from January 2011 to December 2012. The study protocol was approved by the Institutional Review Board and Research Ethical Board, Fourth Military Medical University, and a written informed consent was obtained from each participant before the initiation of any study-related procedures. No patients received any anticancer therapies before surgery. Once retrieved from the patient the sample was placed in a cryogenic vial, which was then immediately snap-frozen in liquid nitrogen. Between collection and sectioning, the sample was stored at −80 °C. An EGFR mutation analysis was performed on a portion from each sample, and the remainder of the sample was stored at −80 °C until scanning on the Raman system. After spectral acquisition, specimens were marked with ink to indicate the region sampled, fixed in formalin, routinely processed, paraffin-embedded, consecutively cut through the marked locations into three parallel 4 μm thick sections, and respectively stained with hematoxylin & eosin (H&E) and IHC based EGFR exon 19 and exon 21 mutation analysis was performed. The histological slides were examined by an experienced pathologist who was blinded to the outcome of the RS analysis. At least 25 spectra were collected from one sample. The histology was based on the criteria of the World Health Organization and the TNM (Tumor, Node, and Metastasis) stage was determined according to version 7 of the International Association for the Study of Lung Cancer (IASLC) staging system.

2.3 Mutation analysis

Genomic DNA was extracted from lung adenocarcinoma tissues as per standard protocols (E.Z.N.A.® Tissue DNA Kit, Omega, USA). Genomic DNA was used for polymerase chain reaction (PCR) amplification and sequencing. Mutations of the EGFR gene in exons 18 (G719A, G719S, G719C), 19 (delE746_A750), 20 (T790M, S768I) and 21 (L858R, L861Q) were PCR amplified. Cycle sequencing of the purified PCR products was carried out with PCR primers using the commercially available ADx Mutation Detection Kits (Amory, Xiamen, China). The assay was carried out according to the manufacturer's protocol with the MX3000P (Stratagene, La Jolla, USA) real-time PCR system. A positive or negative result could be obtained if it met the criterion that was defined by the manufacturer's instruction. Firstly, the PCR amplification curves of positive and negative controls were confirmed. Then, amplification curves of tissue samples were plotted out and the Ct value was calculated when

 This journal is © The Royal Society of Chemistry 2014

Published on 11 November 2013. Downloaded by Université Laval on 28/06/2014 13:07:50.

the curve starts ascending from the cut-off point (Fig. 1). The Ct values that we used to determine whether a sample was positive or negative were based on extensive validation as follows: strong positive, if Ct value < 26; weak positive, if 26 < Ct value < 29; negative, if Ct value > 29.

2.4 EGFR mutation specific IHC staining

Serial 4 μm thick tissue sections were cut from the tissue microarrays for IHC based EGFR exon 19 and exon 21 mutation analyses. The slides were baked at 55 °C overnight, then deparaffinized in xylene and rehydrated through a graded series of ethanol concentrations. Antigen retrieval was performed by microwaving these sections in 10 mM citrate buffer (pH 6.0). To reduce nonspecific binding, slides were blocked with goat serum (5%, sigma) for 30 min. Then, the sections were incubated in a humidified chamber at 4 °C overnight with primary antibodies, EGFR E19del (1 : 100, rabbit IgG; catalog number 2085, Cell Signaling Technologies (CST), Danvers, MA) and EGFR L858R (1 : 100, rabbit IgG; catalog number 3197, CST) which were diluted in 1% BSA. After the sections were washed, they were incubated with the corresponding secondary antibodies for 1 h. Peroxidase activity was visualized with the DAB Elite kit (K3465, Dako), and the brown coloration of tissues represented positive staining. Finally, the sample sections were viewed under a light microscope (Zeiss Axioplan 2, Berlin, Germany).

IHC expression of the monoclonal antibody against EGFR was evaluated using the following scoring, as previously described.[5] Staining intensity was scored from 0 to 3+; the intensity score was established as follows: 0, if tumor cells had complete absence of staining or faint staining intensity in <10% cells; 1+, if >10% of tumor cells had faint staining; 2+, if tumor cells had moderate staining; and 3+, if tumor cells had strong staining. Tumors with 1+, 2+, and 3+ expression were interpreted as positive for E19del or L858R EGFR antibody expression, and tumors with no expression (0 score) were interpreted as negative. IHC staining overview was performed by a pathologist (Zhipei Zhang) and a trained reader (Yunfeng Ni) at Fourth Military Medical University. The final score per patient was calculated by the two readers using the core with the maximum value for each patient.

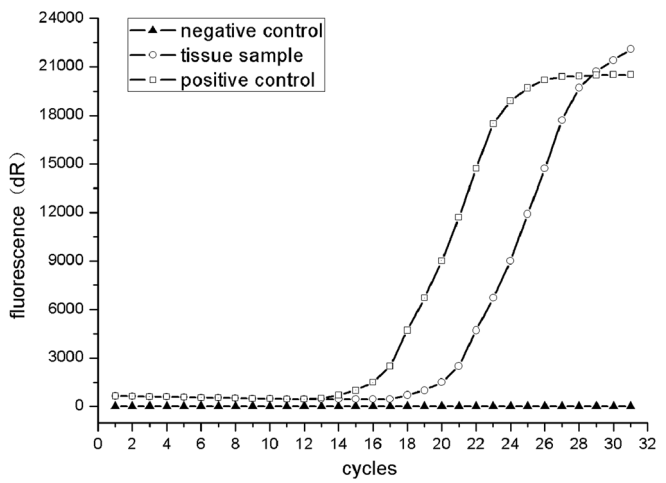

Fig. 1 PCR amplification curve of mutation positive (box) and negative (triangle) controls for DNA sequencing. Tissue sample means the mutation tissue sample (L858R or E19del) which was detected in the research.

2.5 Raman instrumentation and data preprocessing

Raman spectra data were recorded on a Labram HR 800 (Horiba-Jobin-Yvon) spectrophotometer. The laser beam was set at 17 mW with a 632.8 nm He–Ne laser radiation, and accurately focused on a 1 square μm spot on the surface of the sample. The tissues were centered and photographed using 10 objective magnifications, and measured using 100 objectives. The acquisition period was 20 seconds, with a 1 cm^{-1}spectral resolution over a 700–1800 cm^{-1} Raman shift range. The raw spectra were preprocessed by a first-order Savitsky–Golay filter (7 points) for noise smoothing, and then a fifth-order polynomial was found to be optimal for fitting the autofluorescence background in the noise-smoothed spectrum. This polynomial was then subtracted from the noise-smoothed spectrum to yield the tissue Raman spectrum alone.

2.6 Statistical analysis

PCA combined with SVM in Matlab7.1 (The Mathworks, Inc, Natick, MA) was used to develop a diagnostic algorithm for predicting the mutation type within the data set. PCA is a data compression procedure which identifies major trends within the spectral data set and redefines the data set using a small set of component spectra or principal components (PC) and scores.[19] The order of PC denoted the importance to a data set. First three PCs represented the highest variance of the data set. PCA was employed in this study to highlight the variability existing in the spectral data set recorded during the different experiments. Otherwise, analysis of loading of PCs could give the information about the source of variability inside a data set, derived from variations in the molecular components contributing to the spectra.[20,21]

SVM with a radial basis function kernel and a particle swarm optimization solver were applied to discriminate different groups in the form of their PC scores. Leave one-spectrum out cross-validation (LOOCV) was used to train and test PCA/SVM. In this procedure, an SVM model was initially built from all the spectra except one. Then, the algorithm predicted the classification of the omitted spectrum and stored the result. This procedure was repeated with each omitted spectrum in turn, leading to independent predictions per spectrum. For each predicted spectrum, the probability of prediction was calculated and expressed as the sensitivity and specificity for each mutation type.

The diagnostic accuracy of the predictions was quantified by constructing a Receive Operating Characteristic (ROC) curve. This represented the relationship between sensitivity and specificity at different cutoff values of the probability. When the features of a single spectrum completely fitted the characteristics of an E19del spectrum as described by the PCA/SVM model and was in complete disagreement with the characteristics of

Published on 11 November 2013. Downloaded by Université Laval on 28/06/2014 13:07:50.

wild type (wt)-EGFR, the spectrum was classified as E19del with a 100% probability. For the ROC curve, this cutoff value was varied from 0% probability, where every spectrum was predicted as E19del (no false negative results), to 100% probability, where every spectrum was predicted as wt-EGFR (no false positive results). The same methods were used to construct the ROC curve of L858R. The integration area under the ROC curves for PCA/SVM was calculated, illustrating the efficacy of RS together with PCA/SVM algorithms for mutation type prediction.

3. Results and discussion

3.1 Assembly of tumor samples

From January 2011 to December 2012, samples from 153 patients with lung adenocarcinoma were collected consecutively in this study. All patients were ethnic Chinese (Han). Of these, patients were enrolled in this study based on the following criteria: pathologic diagnosis of lung adenocarcinoma, the tumor specimen contained a minimum of 50% tumor cells, enough tissue was available for DNA sequencing and Raman analysis.

3.2 EGFR mutation status measurement

Mutations of L858R and E19del, resulted in the activation of the tyrosine kinase domain, which were associated with sensitivity to TKIs. Other drug-sensitive mutations, for instance point mutations at exon 21 (L861Q) and exon 18 (G719X), were detected infrequently. Apart from lung adenocarcinoma, mutation of EGFR was also found in 10% squamous cell carcinoma cases.[22] Patients with EGFR mutation would benefit most from TKIs, with a significantly extended progression-free survival and overall survival.[4] Thus, the assessment of EGFR mutation had an extraordinary role in lung cancer therapy. Mutation analyses were first performed by DNA sequencing. The EGFR kinase domain mutations were detected in 75 patients (49.0%), and the remaining 78 cases were regarded as wt-EGFR. Among these 153 patients, 29 had deletions in exon 19 (E19del), 33 had L858R mutation, and 7 had T790M mutation. 6 patients had multiple mutations, including 2 patients with E19del and L858R mutation; 2 patients with L858R and T790M mutations; 1 patient with E19del and T790M mutations; 1 patient with G719S, S768I and L861Q mutations.

3.3 Raman and IHC analyses of EGFR status in NSCLC patients

Among the 75 patients with EGFR mutations, 30 were selected for Raman analysis, including 10 with E19del, 10 with L858R and 10 with wt-EGFR, and the detailed clinical characteristics are listed in Table 1. After spectral acquisition, specimens were marked with ink to indicate the region sampled, and respectively stained with H&E and IHC based E19del and L858R mutation analysis was performed. For these consecutively parallel sections, tumor samples expressing E19del or L858R mutations were positive with the respective mutation-specific antibodies and samples with wt-EGFR were never stained with these two mutation-specific antibodies (Fig. 2). In total, as compared with DNA sequencing, the sensitivity of the IHC assay for E19del and L858R mutation was 90% (9/10) and 80% (8/10) respectively, with a specificity of 100% (10/10). Table 2 shows the detailed correspondence between the IHC score and PCR Ct value stage.

Table 1 Clinical characteristics of patients with lung adenocarcinoma

Characteristics	Non-mutation	L858R	E19del
No. of patients	10	10	10
Age (mean, years)	57.4	61	65.3
Sex (no., male/female)	8/2	7/3	3/7
Clinical stage (%)			
I	1	1	1
II	5	3	2
III	4	6	7
IV	0	0	0
Differentiation			
Poor	3	4	1
Good	2	3	3
Moderate	5	3	6

In total, 672 spectra were collected from these 30 patients. Spectra taken from the following regions could be selected for the ultimate Raman analysis: the regions marked with ink were cancerous epithelial cells on the H&E stained tissue section, and positive with the respective mutation-specific antibodies on the corresponding parallel IHC section. Thus, after removing the spectra from normal and non-mutated regions, 441 spectra were appropriate for Raman analysis: 149 from wt-EGFR adenocarcinoma, 135 from L858R mutation and 157 from E19del mutation.

Fig. 3 illustrates the average spectrum of fresh adenocarcinoma tissues of wt-EGFR, L858R and E19del. It could be seen that the general Raman spectral shape of these three mutation types was very similar in most regions but had slight differences in the relative regions. The peaks at 675, 1107, 1127 and 1582 cm^{-1} were significantly increased in wt-EGFR tissues which can be attributed to specific amino acids, such as tryptophan and DNA.[17,19,23] The strong bands of wt-EGFR tissues at 1127 and 1582 cm^{-1} were assigned to cytochrome c (cyt-c) which is associated with electron transfer in oxidative phosphorylation.[24] Cyt-c is usually localized at the mitochondrial inner membrane in the cytoplasm, thus, the intensity of these Raman signals seemed to be associated with the amount of the mitochondrial contents and the distribution of mitochondria in the cell.

L858R is the exon 21 point substitution that replaced leucine 858 with arginine, and the peaks at 1085, 1175 and 1632 cm^{-1} assigned to arginine were slightly increased in L858R tissues (Fig. 3A).[25,26] The structure of the wild-type EGFR kinase domain has been previously determined in both active and inactive conformations.[27,28] The L858R mutation lies in the N-terminal portion of the activation loop. In the inactive state, the N-terminal portion of the activation loop formed a short helix that displaced the regulatory C-helix from the active site to stabilize the inactive conformation. Substitution of the leucine side chain with the arginine (Arg) side chain was expected to

 This journal is © The Royal Society of Chemistry 2014

Published on 11 November 2013. Downloaded by Université Laval on 28/06/2014 13:07:50.

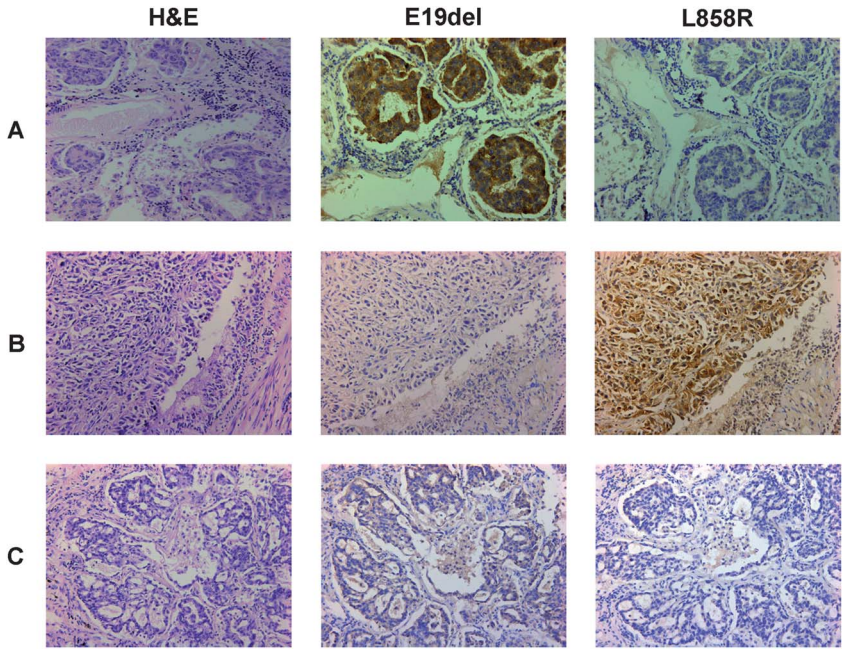

Fig. 2 H&E and IHC staining of NSCLC tumor samples (20× magnification). (A) Samples with E19del mutation were stained with anti-E19del-specific antibody (middle, score 3+). (B) Samples with L858R mutation were stained with anti-L858R-specific antibody (right, score 3+). (C) Samples with wt-EGFR were never stained with these two mutation-specific antibodies.

Table 2 The correspondence between IHC scores and PCR Ct value stage

		DNA sequencing			
		Negative (Ct > 29)	Weak positive (26 < Ct < 29)	Strong positive (Ct < 26)	Total
IHC	0	10	2	1	13
	1+	0	2	1	3
	2+	0	1	2	3
	3+	0	1	10	11
Total		10	6	14	30

destabilize this conformation, as arginine cannot be favorably accommodated in the hydrophobic pocket occupied by leucine. In the active conformation, the side chain of Arg858 was well-ordered, and formed a hydrogen bond with the mainchain carbonyl of Arg836. The activation loop was reorganized and the C-helix rotated into its active position. There was no shift in the protein backbone around Arg858. In our research, we observed an increase of arginine in L858R tissues and no Raman shift of the protein backbone, such as amide I at 1660 and amide III at 1220–1300. In addition, we observed no Raman spectral changes of leucine at 964, 1342 and 1459 cm^{-1} that might be overlapped by some adjacent strong bonds.[29]

E19del is the exon 19 deletion that removes residues 746–750 of the expressed protein.[27,28] We observed an overall decrease in the Raman signal except at 1155 and 1372 cm^{-1} which were assigned to ribodesose and the basic group of DNA in E19del tissues (Fig. 3B).[17,19,23] This indicated that there was a decrease in the percentage of a certain type of biomolecules relative to the total Raman-active constituents from wt-EGFR to E19del transformation. However, these changes were just a spectral phenomenon of EGFR mutation. Raman spectra should be further compared between the wild-type and the synthetic peptide of mutant EGFR.

3.4 Principal component analysis

Fig. 4 shows scatter plots of first principal component *versus* second principal component (PC) of Raman spectra to demonstrate distinctive spectral clustering, in which each of the spectra has been represented by a different color and a different shape. It showed a significant classification between three mutation types of lung adenocarcinoma in the 2-dimensional coordinate system. The mutation status of lung adenocarcinoma A549, H1975 and H1650 cell lines was detected using DNA sequencing. It showed that H1975 carried the L858R mutation gene, H1650 carried the E19del mutation gene, and A549 carried no mutation genes. The pure biomolecular compounds (DNA/RNA/protein) of different mutation types were extracted from these cells, and the spectra were compared with the spectra of different pure compounds.

The loading represented the variability described by a principal component as a function of the wavenumber of the spectra. The loading was a compilation of different peaks with different intensities, both positive and negative, and correspond to increased or decreased contributions of specific molecular components to Raman spectra.[20,21] Fig. 5(I and II) show the loading of PC1 and PC2 obtained from the PCA analysis based on the wt-EGFR cancerous tissues. The loadings have been compared with the spectra of DNA, RNA and protein extracted from A549 cells. The peaks with highest variability

Published on 11 November 2013. Downloaded by Université Laval on 28/06/2014 13:07:50.

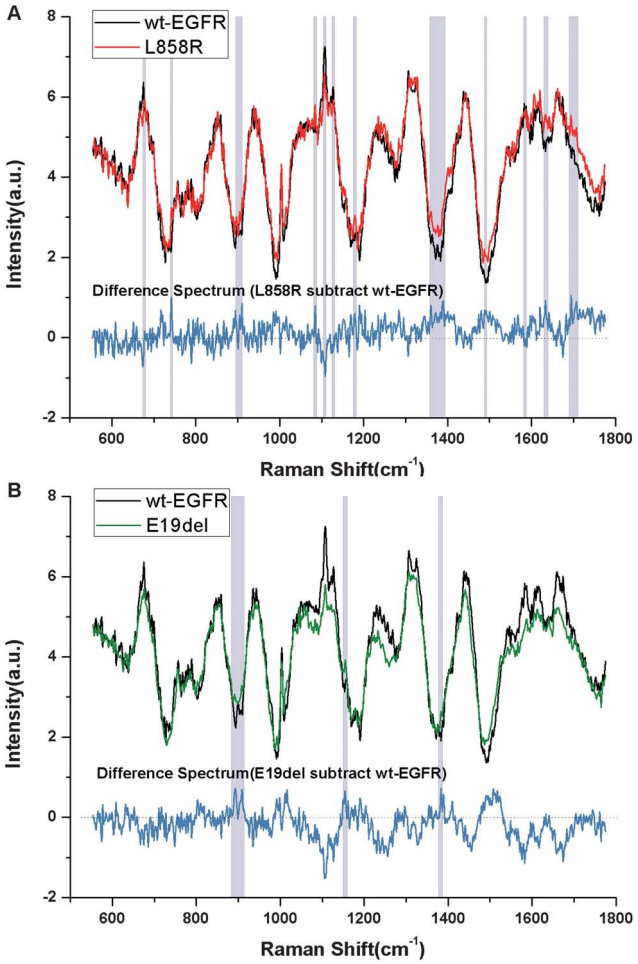

Fig. 3 (A) The average spectra obtained from wt-EGFR and L858R cancerous tissues and the difference spectrum. (B) The average spectra obtained from wt-EGFR and E19del cancerous tissues and the difference spectrum.

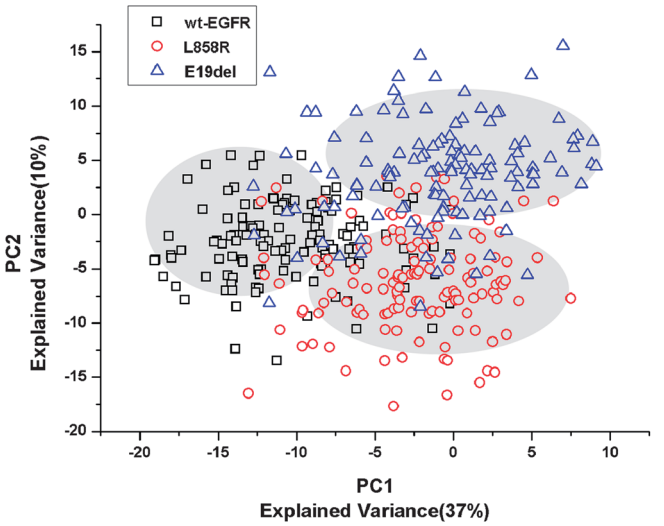

Fig. 4 A scatter plot of the spectra from wt-EGFR (black box), L858R (red round) and E19del (blue triangle) tissues used to develop the model projected onto a plane in PC. The gray elliptic shadow gave the outline of the scatter distribution between three mutation types.

have been highlighted by a gray frame in the figure and have a correspondence with the spectra from the compounds tested. The peaks at 1107, 1127 and 1582 cm^{-1} were prominent in the spectra of DNA and RNA, and the peaks at 675, 1307–1324 were attributable to protein. In Fig. 3A, these peaks were significantly stronger in the average spectra of wt-EGFR tissues, thus giving a strong contribution to PC1 and PC2. The main variation was the positive peak at 1582 cm^{-1} in PC1 and 1127 cm^{-1} in PC2. The peaks at 1127 cm^{-1} and 1582 cm^{-1} were assigned to cyt-c, the key molecule in the apoptosis pathway, and its intensity was on behalf of the content in the cytoplasm. In Fig. 5(III and IV), the loadings of PC1 and PC2 acquired from L858R cancerous tissues were compared with the spectra of DNA, RNA and protein extracted from H1975 cells which carried the L858R mutation gene. The positive peaks at 1085 and 1175 cm^{-1} in PC1 were attributed to the DNA, RNA and protein content, and that at 1632 cm^{-1} in PC2 was related to protein. The loading differences between wt-EGFR and L858R cancerous tissues could be matched with the average Raman spectra. The loading of PC1 and PC2 (Fig. 5(V and VI)) from E19del cancerous tissues represented the variability between wt-EGFR and E19del. The loadings were compared with the spectra of pure compounds from H1650 cells with E19del mutation gene. The spectra exhibited similar spectral features as revealed in the difference spectra shown in Fig. 3B. The intensity of many peaks in PC1 and PC2 related to DNA and RNA especially at 783, 1045, 1127, 1582, 1660 cm^{-1} decreased as compared with wt-EGFR. The peak at 1155 cm^{-1} in PC2 which is assigned to ribodesose of DNA increased. It showed that the main molecular difference between E19del and wt-EGFR cancerous tissues was derived from the DNA/RNA content.

However, a high intra-group variability which could interfere the information contained in the loading existed. The Raman spectra of cytoplasm and nucleus were obviously different in cells, and PCA analysis provided an effective identification of components from any region.[20,21] In this research, it was difficult to obtain the spectra based on different regions since the cell structure of cancerous tissues could not be recognized without stain when viewed under a microscope. Thus, we only compared the complex mixtures of cancerous tissues with the molecular components, such as DNA, RNA and protein. Otherwise, EGFR is a member of the receptor tyrosine kinase family, which is mainly expressed on the cell membrane. Mutation of EGFR might change the adhesion function of cancer cells, which could result in the changes of morphology from wt-EGFR to the mutated one. Previous research had demonstrated the Raman spectral differences in the molecular signature of different sub-cellular structures.[21] Therefore, it is possible that these observed spectral differences are caused by a change of tissue morphology induced by EGFR mutations.

3.5 Construction of diagnostic algorithms

The PCA/SVM diagnostic algorithm yielded an overall accuracy of 87.8% [*i.e.*, sensitivity of 89.6% (121/135), 88.5% (139/157) and specificity of 85.2% (127/149)] for identifying L858R or

This journal is © The Royal Society of Chemistry 2014

Published on 11 November 2013. Downloaded by Université Laval on 28/06/2014 13:07:50.

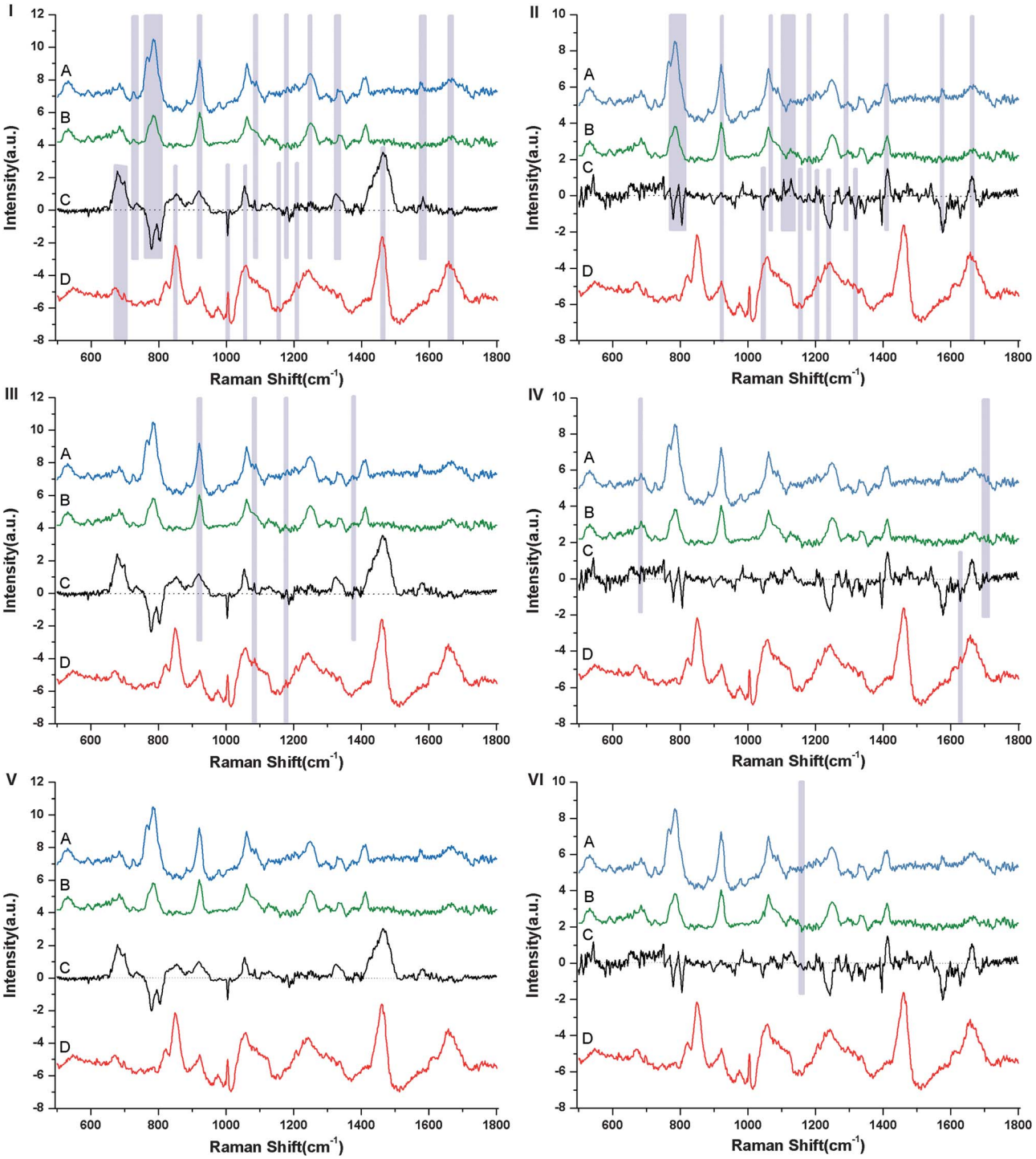

Fig. 5 (I) Plot of loading of PC1 of the PCA analysis (C). This loading has been compared with spectra recorded from different components of A549 cells: DNA (A), RNA (B) and protein (D). (II) Plot of loading of PC2 of the PCA analysis (C). This loading has been compared with spectra recorded from different components of A549 cells: DNA (A), RNA (B) and protein (D). (III) Plot of loading of PC1 of the PCA analysis (C). This loading has been compared with spectra recorded from different components of H1975 cells: DNA (A), RNA (B) and protein (D). (IV) Plot of loading of PC2 of the PCA analysis (C). This loading has been compared with spectra recorded from different components of H1975 cells: DNA (A), RNA (B) and protein (D). (V) Plot of loading of PC1 of the PCA analysis (C). This loading has been compared with spectra recorded from different components of H1650 cells: DNA (A), RNA (B) and protein (D). (VI) Plot of loading of PC2 of the PCA analysis (C). This loading has been compared with spectra recorded from different components of H1650 cells: DNA (A), RNA (B) and protein (D).

E19del from wt-EGFR tissues. As compared with IHC, RS provided approximate sensitivity, but lower specificity for detection of mutation status, which could be attributed to the numbered spectra of wt-EGFR. IHC could be assessed in several fields of a microscope, however, only an average of 14.9 spectra per wt-EGFR sample were involved in the Raman analysis.

Published on 11 November 2013. Downloaded by Université Laval on 28/06/2014 13:07:50.

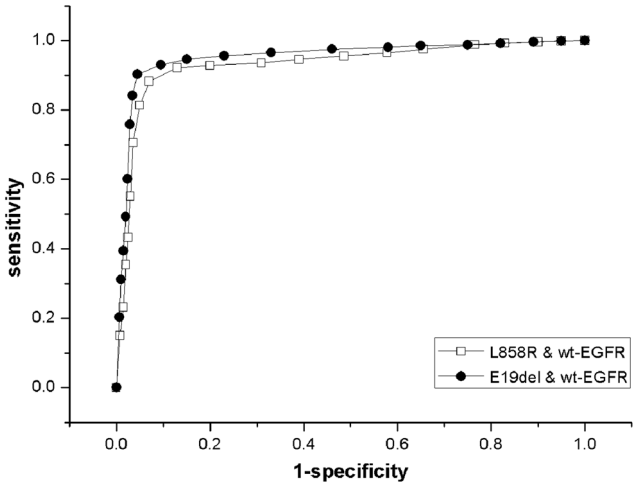

Fig. 6 ROC curve representing the accuracy of Raman spectroscopy discrimination with PCA/SVM. The areas under the curve were 0.945 (black box) and 0.953 (black dots).

Otherwise, a small sample size decreased the computational efficiency of the PCA/SVM algorithm. Thus, large scale screening should be done to verify the sensitivity and specificity of RS. To further evaluate the performance of the PCA/SVM algorithms together with the LOOCV method, the ROC curve (Fig. 6) was also generated. The areas under the ROC curve for L858R & wt-EGFR and E19del & wt-EGFR were 0.945 and 0.953, confirming that the PCA/SVM diagnostic model was powerful for clinical diagnosis at the molecular level. However, we only enrolled patients with E19del and L858R mutation in this research, thus, it might be inefficient for predicting the few patients with multiple mutations.

4. Conclusion

In this paper, we analyzed the EGFR mutation type by using DNA sequencing, IHC and RS, demonstrating the Raman spectral changes at a molecular level. It revealed that the concentration of some specific amino acids, such as arginine and tryptophan, and DNA increased or decreased in different mutation types. Furthermore, we constructed a PCA/SVM algorithm for diagnosing E19del and L858R mutations with an accuracy of 87.8%. The ROC curve showed the excellent accuracy of this method.

Conflicts of interest

None.

Acknowledgements

Lei Wang, Zhipei Zhang and Lijun Huang contributed equally to this paper. This research was supported by the National Natural Science Foundation of China (no. 81000938), co-funded by Science and Technology Innovation Development fund of Tangdu Hospital of The Fourth Military Medical University.

References

1 A. Jemal, *et al.*, Cancer statistics, *Ca-Cancer J. Clin.*, 2010, **60**(5), 277–300.
2 C. J. Beadsmoore, *et al.*, Classification, staging and prognosis of lung cancer, *Eur. J. Radiol.*, 2003, **45**(1), 8–17.
3 F. R. Hirsch, *et al.*, EGFR testing in lung cancer is ready for prime time, *Lancet Oncol.*, 2009, **10**(5), 432–433.
4 T. S. Mok, *et al.*, Gefitinib or carboplatin–paclitaxel in pulmonary adenocarcinoma, *N. Engl. J. Med.*, 2009, **361**(10), 947–957.
5 J. Yu, *et al.*, Mutation-specific antibodies for the detection of EGFR mutations in non-small-cell lung cancer, *Clin. Cancer Res.*, 2009, **15**(9), 3023–3028.
6 L. M. Sholl, *et al.*, EGFR mutation is a better predictor of response to tyrosine kinase inhibitors in non-small cell lung carcinoma than FISH, CISH, and immunohistochemistry, *Am. J. Clin. Pathol.*, 2010, **133**(6), 922–934.
7 D. A. Eberhard, *et al.*, Biomarkers of response to epidermal growth factor receptor inhibitors in Non-Small-Cell Lung Cancer Working Group: standardization for use in the clinical trial setting, *J. Clin. Oncol.*, 2008, **26**(6), 983–994.
8 Z. Huang, *et al.*, Near infrared Raman spectroscopy for optical diagnosis of lung cancer, *Int. J. Cancer*, 2003, **107**(6), 1047–1052.
9 N. D. Magee, *et al.*, Ex vivo diagnosis of lung cancer using a Raman miniprobe, *J. Phys. Chem. B*, 2009, **113**(23), 8137–8141.
10 M. Young-Kun, *et al.*, 1064 nm near-infrared multichannel Raman spectroscopy of fresh human lung tissues, *J. Raman Spectrosc.*, 2005, **36**(1), 73–76.
11 N. D. Magee, *et al.*, Raman microscopy in the diagnosis and prognosis of surgically resected non small cell lung cancer, *J. Biomed. Opt.*, 2010, **15**(2), 026015.
12 N. D. Magee, *et al.*, Raman spectroscopy analysis of induced sputum in lung cancer, *Am. J. Respir. Crit. Care Med.*, 2010, **181**, A3492.
13 H. Nawaz, *et al.*, Evaluation of the potential of Raman microspectroscopy for prediction of chemotherapeutic response to cisplatin in lung adenocarcinoma, *Analyst*, 2010, **135**(12), 3070–3076.
14 F. Draux, *et al.*, Raman imaging of single living cells: probing effects of non-cytotoxic doses of an anti-cancer drug, *Analyst*, 2011, **136**(13), 2718–2725.
15 Q. Matthews, *et al.*, Biochemical signatures of in vitro radiation response in human lung, breast and prostate tumour cells observed with Raman spectroscopy, *Phys. Med. Biol.*, 2011, **56**(21), 6839–6855.
16 S. Koljenović, *et al.*, Raman microspectroscopic mapping studies of human bronchial tissue, *J. Biomed. Opt.*, 2004, **9**(6), 1187–1197.
17 C. Krafft, *et al.*, Raman mapping and FTIR imaging of lung tissue: congenital cystic adenomatoid malformation, *Analyst*, 2008, **133**(3), 361–371.
18 M. A. Short, *et al.*, Using laser Raman spectroscopy to reduce false positives of autofluorescence bronchoscopies: a pilot study, *J. Thorac. Oncol.*, 2011, **6**(7), 1206–1214.

 This journal is © The Royal Society of Chemistry 2014

Published on 11 November 2013. Downloaded by Université Laval on 28/06/2014 13:07:50.

19 L. Wang, *et al.*, Raman spectroscopy, a potential tool in diagnosis and prognosis of castration-resistant prostate cancer, *J. Biomed. Opt.*, 2013, **18**(8), 87001.
20 F. Bonnier, *et al.*, Understanding the molecular information contained in principal component analysis of vibrational spectra of biological systems, *Analyst*, 2012, **137**(2), 322–332.
21 F. Bonnier, *et al.*, Imaging live cells grown on a three dimensional collagen matrix using Raman microspectroscopy, *Analyst*, 2010, **135**(12), 3169–3177.
22 G. J. Riely, *et al.*, Update on epidermal growth factor receptor mutations in non-small cell lung cancer, *Clin. Cancer Res.*, 2006, **12**(24), 7232–7241.
23 Y. Oshima, *et al.*, Discrimination analysis of human lung cancer cells associated with histological type and malignancy using Raman spectroscopy, *J. Biomed. Opt.*, 2010, **15**(1), 017009.
24 K. Hamada, *et al.*, Raman microscopy for dynamic molecular imaging of living cells, *J. Biomed. Opt.*, 2008, **13**(4), 044027.
25 B. Sharma, *et al.*, UV resonance Raman finds peptide bond-Arg side chain electronic interactions, *J. Phys. Chem. B*, 2011, **115**(18), 5659–5664.
26 A. C. Fonseca, *et al.*, Study of *N*α-benzoyl-L-argininate ethyl ester chloride, a model compound for poly(ester amide) precursors: X-ray diffraction, infrared and Raman spectroscopies, and quantumchemistry calculations, *J. Chem. Phys.*, 2011, **134**(12), 124505.
27 J. G. Paez, *et al.*, EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy, *Science*, 2004, **304**(5676), 1497–1500.
28 C. H. Yun, *et al.*, Structures of lung cancer-derived EGFR mutants and inhibitor complexes: mechanism of activation and insights into differential inhibitor sensitivity, *Cancer Cell*, 2007, **11**(3), 217–227.
29 P. F. Façanha Filho, *et al.*, Structure–property relations in crystalline L-leucine obtained from calorimetry, X-rays, neutron and Raman scattering, *Phys. Chem. Chem. Phys.*, 2011, **13**(14), 6576–6583.