

Infrared spectral histopathology (SHP): a novel diagnostic tool for the accurate classification of lung cancer

Benjamin Bird^{1,2,*}, Miloš Miljković^{1,*}, Stan Remiszewski², Ali Akalin³, Mark Kon⁴ and Max Diem^{1,*}

We report results of a study utilizing a recently developed tissue diagnostic method, based on label-free spectral techniques, for the classification of lung cancer histopathological samples from a tissue microarray. The spectral diagnostic method allows reproducible and objective diagnosis of unstained tissue sections. This is accomplished by acquiring infrared hyperspectral data sets containing thousands of spectra, each collected from tissue pixels about 6 μm on edge; these pixel spectra contain an encoded snapshot of the entire biochemical composition of the pixel area. The hyperspectral data sets are subsequently decoded by methods of multivariate analysis, which reveal changes in the biochemical composition between tissue types, and between various stages and states of disease. In this study, a detailed comparison between classical and spectral histopathology (SHP) is presented, which suggests SHP can achieve levels of diagnostic accuracy that is comparable to that of multi-panel immunohistochemistry.

Laboratory Investigation (2012) 92, 1358–1373; doi:10.1038/labinvest.2012.101; published online 2 July 2012

KEYWORDS: artificial neural network analysis; histopathology; immunohistochemistry; lung cancer; spectral histopathology

This paper reports a large-scale study of a new technology to classify four common forms of lung cancers, and distinguish them from normal tissues. The new methodology introduced here utilizes optical measurements on unstained tissue¹ for spectral data acquisition, and does not utilize any immunohistochemical or other stains or labels for classification. As the diagnostic procedure is instrument based and utilizes trained computer algorithms for classification, this method offers reproducibility, complete objectivity and improved accuracy over present methodology.

Optical methods have been used in histology and pathology ever since these methods were first described. After all, staining tissues or cells by hematoxylin/eosin (H&E), followed by (visual) microscopic examination is a form of spectral analysis: different compartments of the cell respond differently to basophilic and eosinophilic stains and thus, allow a 'spectral analysis' using the eye as a detector. This method can reveal an amazing amount of information but is inherently subjective. More recent optical methods have used

image capture at a few selected wavelengths, and computer analysis of the image planes, for tissue analysis.² Immunohistochemistry (IHC), to date the most advanced optical method to detect the presence of certain cancer signatures or markers^{3,4} uses detection of specific antibodies labeled with easily observable stains.

The new approach reported here is based on the observation of inherent spectral signatures (as opposed to any external stains or labels used to treat the sample) of cellular components to aid classical cytopathology and histopathology.⁵ The paradigm for the spectral approach is that the transition from normal tissue or cells to diseased states is accompanied by changes in the overall biochemical composition of the tissue, along with well-known changes in cellular morphology and tissue architecture, which are particularly pronounced in advanced stages of cancer. These changes in biochemical composition are encoded and observed via changes in the infrared (IR) spectra. Other label-free spectral methods have been used to aid classical histopathology, in

¹Laboratory for Spectral Diagnosis, Department of Chemistry and Chemical Biology, Northeastern University, Boston, MA, USA; ²Cireca Theranostics, Parsippany, NJ, USA; ³Department of Pathology, University of Massachusetts Medical School, Worcester, MA, USA and ⁴Department of Mathematics and Statistics, and Program in Bioinformatics, Boston University, Boston, MA, USA

Correspondence: Professor M Diem, PhD, Laboratory for Spectral Diagnosis, Department of Chemistry and Chemical Biology, Northeastern University, 360 Huntington Avenue, 316 Hurtig Hall, Boston, MA 02115, USA.

E-mail: m.diem@neu.edu

*Principal investigators, contributed equally to research.

Received 14 March 2012; revised 27 April 2012; accepted 8 May 2012

particular for *in-vivo* applications: fluorescence spectroscopy-based imaging, for example, has been used for the diagnosis of hollow organs.^{6,7} In this method, the different chemical composition of a few tissue components is exploited. However, this method lacks fingerprint specificity toward specific changes in tissue composition, and only few constituents of human tissue actually exhibit detectable fluorescence.

Over the past two decades, other spectral techniques have gained attention for medical diagnostic imaging. These spectral methods are based on vibrational spectroscopy (either IR absorption or Raman scattering spectroscopy) and offer the advantage that all (bio)molecules exhibit distinct spectral signatures, in contrast to the fluorescence-based methods alluded to above in which only few select molecules respond to the excitation by light. This paper will deal exclusively with the application of IR absorption spectroscopy to the field of medical diagnostics.

As all molecules respond to the exciting IR radiation to produce relatively complicated 'IR spectra', the response observed for an individual dried cell or a tissue section used in classical histopathology is a complex superposition of all spectral features of all biomolecules in the sample. Although IR spectroscopy is usually referred to as a 'fingerprint' spectroscopic technique, which implies that every molecule known exhibits a distinct spectrum that identifies it, the superposition of such fingerprints leads to relatively broad spectral features that need to be decoded, or de-convolved, to enable an interpretation or diagnosis. Nevertheless, typical IR spectra observed for three different tissue types (see Figure 1) allow a coarse assessment of the biochemical composition: in Figure 1, the top trace is from the superficial layer of squamous tissue, which is known to accumulate glycogen, and which can be detected readily by IR spectroscopic methods.⁸ The spectral features of glycogen consist of a number of sharp peaks superimposed on protein spectral signatures. The middle trace of Figure 1 depicts an IR spectrum of connective tissue, which is dominated by the spectral features of collagen.⁹ Finally, the bottom trace shows the IR signature of metabolically highly active cells such as B-lymphocytes, which exhibit distinct nucleic acid features in addition to the protein peaks observed in the other traces.¹⁰ In general, the spectral differences between tissue types are much smaller than the ones shown in Figure 1, and require mathematical procedures for detection and interpretation. The concepts of these multivariate methods of analysis will be introduced later. The combination of IR microspectral data acquisition from a tissue sample, followed by multivariate data analysis, has been referred to as spectral histopathology (SHP).

In the results presented here, a commercial tissue microarray (TMA) containing 80 tissue spots (10 with normal and 70 with cancerous diagnosis) was analyzed by SHP. For each tissue spot, a consensus medical diagnosis was available from the array manufacturer, and the TNM (tumor-node-metastasis) classification was known. In addition, the diagnosis provided by the medical collaborator in this study, who is an

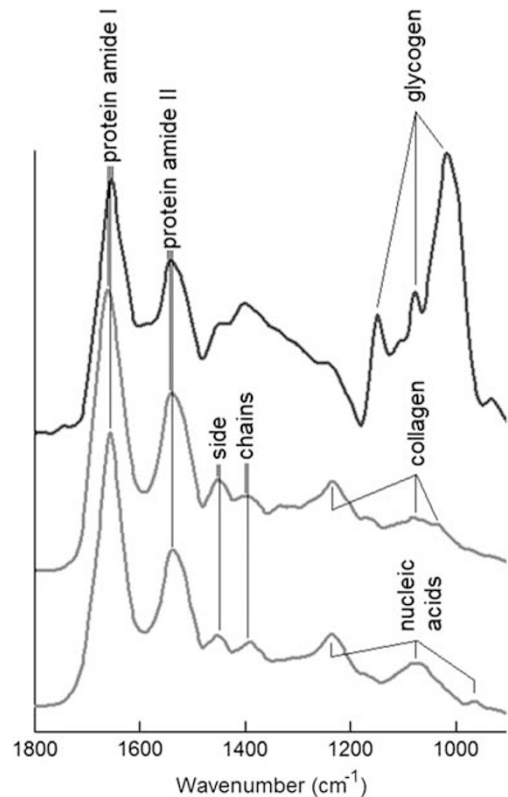


Figure 1 Examples of mid-IR spectra of different tissue classes. Top: superficial squamous tissue; middle: fibro-connective tissue; bottom: B-lymphocytes. The three spectra are offset along the absorbance (γ) axis for clarity.

expert pulmonary pathologist, was used to train and verify the spectral methodology. Using more than one hundred thousand individual IR spectra extracted from tissue spots, which were rigorously separated into training and test sets, artificial neural net (ANN)-based diagnostic algorithms¹¹ were constructed that could distinguish normal from cancerous tissue, and classify small cell lung carcinomas (SCLCs), squamous cell carcinomas (SqCCs) and adenocarcinomas/bronchiolo-alveolar carcinomas (ADCs/BACs) with an average accuracy of 95%. It cannot be over-emphasized that these results were obtained for a blinded test set, and that these results exceed the accuracy of routine IHC.

These studies were preceded by earlier work from the authors' laboratory on the detection of micrometastases in lymph nodes,^{12–14} cervical adeno- and SqCCs,¹⁵ and from other laboratories on the detection of colon cancer,^{11,16–18} prostate and breast cancer,^{19,20} brain cancers and brain metastases,^{21–23} as well as a few other organs.^{9,24} In addition, similar efforts using Raman spectroscopy, another form of vibrational spectroscopy, have yielded analogous results.^{25–28} These studies mostly were aimed at demonstrating that vibrational (IR and Raman) spectroscopy can detect spectral differences between normal tissue types, and between normal and diseased tissue. However, the data sets were generally restricted in size such that rigorous statistical validation was

impossible, or used data analysis procedures that were not completely objective. The study reported here used patient numbers that allowed for a rigorous separation of training and test data sets, and incorporates novel methods of correlating spectral and classical morphological features for the training data set. The requirement for a large number of patients in each of the training and test data sets for all diagnostic classes precluded a finer graduation of disease (such as papillary, acinar and solid tumor ADCs). Furthermore, inclusion of other disease states, such as certain rarer lung cancers (eg, large cell lung cancer) and non-neoplastic conditions, was not practical at this point. Rather, all non-cancerous tissue types were treated as one class (NOT cancer), but non-neoplastic diseases (eg, granulomas) are included in presently ongoing expanded studies, which involve over 300 patients. The aim of this study was to demonstrate that SHP can detect different tumor types for the diagnosis of lung disease, and that the sensitivity and specificity of this very first attempt rivals that of IHC.

We believe that this technology can aid in the accurate diagnosis of cancers that are difficult to distinguish on a morphological basis alone, and whose accurate diagnoses determine therapeutic options. This report follows similar studies from the same laboratory in which equivalent methods for the analysis of exfoliated cells were reported,^{29–31} using a methodology referred to as spectral cytopathology (SCP). SCP proved to be more sensitive than classical cytopathology in that morphologically normal cells from abnormal (dysplastic) samples exhibited spectral changes that could be associated with a transition from normal to pre-cancerous states.

Background: SHP for Medical Imaging

SHP^{16,19,20,32} is a new method to aid pathology with the diagnostic interpretation of a histopathological specimen. It offers the advantage that the diagnosis is based on spectral measurements, which determine a snapshot of the biochemical composition, and is not based on the morphology of the cells that make up the tissue, or its architecture. In order to perform SHP, IR absorption spectra are collected from tens of thousands of individual pixels, which are 6.25 μm on edge in the study reported here. Thus, for a 1 mm \times 1 mm tissue section, 25 600 individual IR spectra are collected, and stored as a 'hyperspectral data cube', a construct that contains the pixel coordinates and the associated spectrum.¹ This hyperspectral data cube contains the spatial variation of the sample composition, and hereby the sample diagnosis, encoded in the IR spectra.

IR spectroscopy is a well-established technique that measures fingerprint signatures (spectra) of all compounds contained in a sample via their molecular vibrations,³³ which interact with electromagnetic radiation ('light') in the IR region with wavelength between *ca* 2.5 to 25 μm . This interaction results in absorption of light at specific wavelengths, causing 'absorption peaks' (see Figure 1) in the light transmitted or reflected by the sample. Since the mid-1950s,

IR spectra for many biological molecules have been established, and spectra specific for proteins, nucleic acids, lipids and other molecules found in cells and tissue have been recorded.⁹ These spectra exhibit exquisite sensitivity toward subtle differences or changes in molecular structures: for example, IR spectroscopy can distinguish secondary and tertiary structural features (for a recent review, see Barth³⁴) and dynamics in proteins,^{35,36} structural differences between DNA shapes (A, B and Z-form DNA),³⁷ the degree of hydration of these molecules and their interactions with the solvent,³⁷ and the packing and structures found in phospholipid membranes.³⁸

Since the early years of the previous decade, easy-to-use and relatively fast instruments have been available to collect such IR spectra microscopically. This spawned a new research field, known as IR microspectroscopy or IR microscopy, which, in turn, enabled researchers to record the IR microspectra of cells and various tissue types. A small tissue section of 1 mm² may yield thousands of spectra (see above), which show small spectral changes depending on the chemical composition of the pixel area from where each spectrum was collected. Thus, pseudo-color images can be created that depict these spectral changes in relation to the location from which the spectra were collected.³⁹ As thousands of pixel spectra need to be analyzed, and because spectral changes between pixel spectra may be quite small, this technique lends itself to analysis by computer. The typical workflow in SHP thus involves the collection of the IR hyperspectral data cube from an unstained tissue section, followed by computer analysis of the data set. Stains, being molecular compounds, need to be avoided because they exhibit their own spectral patterns, which would interfere with the spectra of the tissue sample. However, subsequent to IR data acquisition, the tissue section may be stained by H&E or immunohistochemical agents to allow a detailed comparison between classical and SHP.

Computer analysis of the spectral hypercube produces a pseudo-color image based on the spectral information. Such an image may be obtained completely independently of other data sets by computing and appropriately color coding the similarity of the spectra within a data set. This procedure is called 'unsupervised' because the resulting image is based merely on a spectral correlation, and does not involve the input of a pathologist. Therefore, these images are valid only for the data set from which they were collected. Figure 2 shows a photo-micrograph of an H&E-stained lung tissue section, and the corresponding unsupervised IR pseudo-color image. The similarity of the structural features obvious in these images suggests that the spectral information parallels the variations in tissue composition visible in the H&E image. This discriminatory sensitivity of IR spectra toward different tissue and disease types suggests that it is possible to construct diagnostic computer algorithms for the analysis of the IR data sets. Such algorithms are trained by recurring spectral patterns associated with disease states or tissue types,

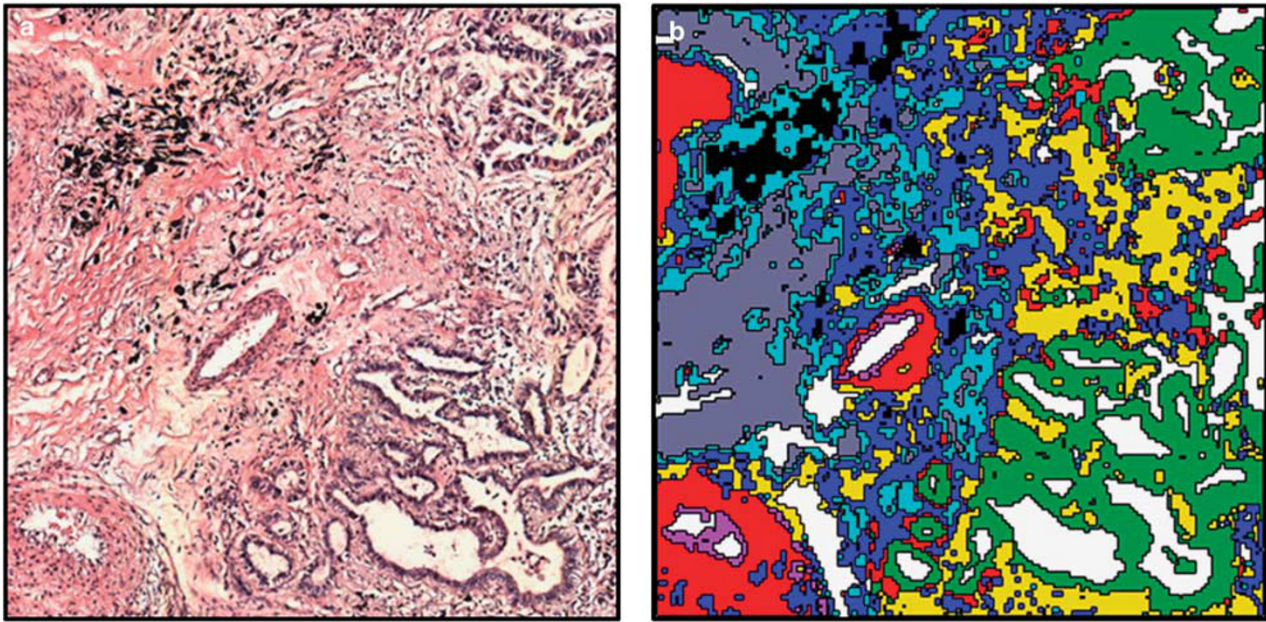


Figure 2 Panel **a**: photomicrograph of an H&E-stained 1 mm \times 1 mm lung ADC tissue section at \times 10 magnification. Panel **b**: HCA-based pseudo-color spectral image of the section shown in panel **a**. Note that the image shown in panel **b** is based on spectral similarity only, and does not require any histopathological diagnostic input. Green areas correlate well with the cancerous regions.

and are subsequently used to classify pixels of IR images collected from the test set. Such a ‘supervised’ diagnostic algorithm is based on associating spectral patterns with a diagnostic annotation obtained from a pathologist. Of course, this step assumes that the same diagnostic classes identified by a pathologist produce the same spectral changes from patient to patient, which has been verified by a number of research groups. The possibility for achieving SHP-based diagnostics was first demonstrated in the PhD dissertation of Lasch.⁴⁰ Owing to some unexpected experimental difficulties, computational restrictions and the lack of theoretical foundations at the time, it took another 10 years to refine experimental and computational methods for wide-spread applications of SHP in medicine.

MATERIALS AND METHODS

Samples

To demonstrate the ability of SHP to detect and differentiate different types of lung cancers, a tissue microarray (TMA, BiomaxUS (Rockville, MD), catalogue number LC811) was used that contained 80 tissue cores, or spots, each measuring between 1.5 and 2.0 mm in diameter. For simplicity, the discussion in the remainder of this section assumes 1.5 mm diameter spots. Although patient information and the TNM diagnosis were available for this microarray, no immunohistochemical information was available, and pathological diagnoses were strictly based on classical histopathology, carried out by one of the co-authors. The tissue section was microtomed at BiomaxUS to a thickness of 5 μ m, and mounted on a substrate suitable for IR microscopy (see below). A parallel

section was mounted on a standard microscope slide, de-paraffinized, H&E stained and cover slipped. This slide was imaged at \times 40 magnification via a visual microscope (Olympus, BX51), resulting in 256 tiles (fields of view), each covering *ca* 100 \times 100 μ m and occupying *ca* 4 MB. These tiles can be stitched together for a \times 40 view of the entire tissue spot. This feature is extremely important for the correlation between classical and SHP. The use of TMAs for SHP was first reported by Fernandez, *et al.*¹⁹

For SHP, samples need to be mounted on special substrates because glass completely absorbs IR radiation. The substrates used here are so-called ‘low emissivity’ (low-e for short) slides (Kevley Technologies, Chesterland, OH, USA) that consist of a normal glass substrate coated with a very thin silver layer, and a tin oxide overcoating. These slides are nearly completely transparent to visible light and therefore, can be used for classical histopathology. However, they are completely reflective in the IR spectral range, and can be used in reflection microscopy, as follows: the IR beam passes through the tissue section, is reflected by the silver layer, and passes the sample a second time. In both passes, the IR beam is attenuated by the molecular interactions with the light. This measurement method has been referred to as transmission-reflection, or ‘transflection’ measurement. Subsequent to IR data acquisition, the section may be stained and cover slipped for visual histopathology.

IR Data Acquisition

IR hyperspectral data sets were collected for each of the tissue spots of the TMA using a Perkin Elmer (Shelton, CT, USA)

Fourier transform IR imaging microspectrometer, model Spectrum One/Spotlight 400, henceforth referred to as the PE400. This instrument allows data acquisition from visually selected sample regions of arbitrary size, which is determined only by the available memory of the instrument computer to store the data. Spectra for each of the 1.5 mm diameter tissue spots were collected from pixels $6.25\ \mu\text{m}$ on edge. Thus, the raw imaging data set consisted of $(1500\ \mu\text{m}/6.25\ \mu\text{m})^2 = 57\,600$ individual pixel spectra for each tissue spot, and correspondingly larger for the larger diameter spots.

The PE400 incorporates a 16 element cryogenically cooled IR HgCdTe detector array; thus, spectra from 16 pixels were collected simultaneously. The acquisition of 16 pixel spectra took about 0.85 s. Subsequently, the sample was moved by $6.25\ \mu\text{m}$ in the focus of the IR beam, and another set of 16 pixel spectra was acquired until the entire sample area was scanned. For each pixel, four interferograms, collected at $4\ \text{cm}^{-1}$ spectral resolution were co-added, and stored after Fourier transform as 1626 point intensity vectors with $2\ \text{cm}^{-1}$ data spacing from 750 to $4000\ \text{cm}^{-1}$ in native PE 400 imaging format (.fsm files). Data acquisition, Fourier transform and storage required *ca* 45 min for each tissue spot. The entire instrument, including the optical path of the microscope, was purged with dry (-40° dew point) air to reduce atmospheric water vapor interferences.

Data Pre-Processing

Raw data sets were imported into a data manipulation software package written in house in the MATLAB (Natick, MA, USA) environment. The data pre-processing included the following steps:

- Noise reduction via noise adjusted principal component analysis (PCA).⁴¹
- Spectral quality test to remove spectra from areas not occupied by tissue (cracks and voids), and from tissue edges.
- Region-of-interest selection ($1\ \text{mm} \times 1\ \text{mm}$), see below.
- Truncation of spectra to include the 'fingerprint' region only (512 data points at $2\ \text{cm}^{-1}$ data spacing between 778 and $1800\ \text{cm}^{-1}$).
- Conversion of spectral vectors to second derivatives.⁴²
- Removal of scattering effects by a phase correction method.^{43,44}
- Vector normalization of individual spectral vectors.

The pre-processing steps increase the quality of spectra in a data set by reducing regions of low diagnostic value. Furthermore, selection of a $1\ \text{mm} \times 1\ \text{mm}$ square region within the 1.5 mm diameter tissue spot reduced the number of raw spectra to a computationally manageable data size for the ensuing steps (see below) thereby concentrating on areas which have the most pronounced and diagnostic features. Decisions on the selection of the $1\ \text{mm}^2$ region were made based on inspection of the parallel stained tissue section. After these pre-processing steps, each data set from one tissue spot was stored in MATLAB format and subject to pre-

classification by unsupervised (agglomerative) hierarchical cluster analysis (HCA), which grouped image pixels based on their spectral similarities.

Pre-Segmentation of Data by HCA

The hyperspectral data sets for each of the tissue spots in the training set (see below) were subsequently converted to pseudo-color images by HCA with colors based on the discovered clusters. HCA is a well-known method to extract patterns in data sets;⁴⁵ in this particular application, HCA is used to detect spectral similarities.⁴⁶ To this end, the similarity between all pairs of spectra in a $1\ \text{mm} \times 1\ \text{mm}$ section of each tissue spot was computed by a metric known as *Euclidean distance*,⁴⁵ which results in a similarity (correlation) coefficient that ranges from 1.0 for perfectly identical spectra to 0.0 for completely dissimilar spectra. This is a computationally highly intensive step, because a $1\ \text{mm} \times 1\ \text{mm}$ square region of the tissue containing 25 600 individual spectra requires the computation of $25\,600^2/2$ or about 300 million similarity coefficients (for each data set). Subsequently, spectra from each data set are segmented into clusters according to their similarity coefficients. Each cluster is associated with a color, and the positions from which a spectrum was collected are displayed in the color corresponding to the cluster membership. In Figure 2b, all the areas shown in green are due to similar spectra from areas diagnosed subsequently as ADC, which show significant and reproducible differences from the spectra collected from the regions shown in red. The areas delineated in the HCA map correspond well with regions visible in the H&E-stained image. Increasing the number of clusters displayed increases the detail available from the spectral images; previous results have indicated that the best agreement with H&E images is revealed by HCA images containing between 4 and 10 clusters, corresponding to 4 to 10 diagnostic classes. Next, the high-resolution H&E images and overlays of the HCA images were used in a step referred to as *annotation* (see next section) to associate spectral features contained in a given cluster with pathological features visible on the H&E images.

Annotation

For the annotation procedure, the tissue spots of the TMA were separated into training and test sets, as shown in Table 1. The annotation, that is, the correlation between the HCA segmentation and the pathological diagnosis, represents an important step, because the spectra associated with specific tissue or disease features are subsequently used to train diagnostic algorithms. The annotation step was carried out by one of the authors, who is an expert pulmonary pathologist, using a software package referred to as *CirecaAnnotate* (CMAT). CMAT imports the high-resolution visual imaging data sets along with the 2–15 cluster images from HCA, and displays H&E and HCA panels side-by-side. Image registration is performed based on visually selected 'landmarks' that appear on both panels. The selected landmark points usually

Table 1 Separation of TMA tissue spots into training and test data sets

	Training	Test
Normal lung tissue (NOR)	H02, H04, H06, H08, H10	H01, H03, H05, H07, H09
Small cell lung carcinoma (SCLC)	F01, F03, F05, F09, F10, G01, G03, G06, G09	E08, E09, E10, F02, F04, F06, F08, G02, G07, G10
Squamous cell carcinoma (SqCC)	C03, C06, C07, C09, D01, D03	C02, C05, C08, C10, D02, D04
Adenocarcinoma (ADC)	A03, A06, A07, A10, B02, B04, B06, B09, C01, E05	A01, A04, A08, A09, B03, B05, B07, B10, D06, E02
Bronchio-alveolar carcinoma (BAC)	A02, D05, D09, D10, E3	D07, D08, E01, E04, E06, E07

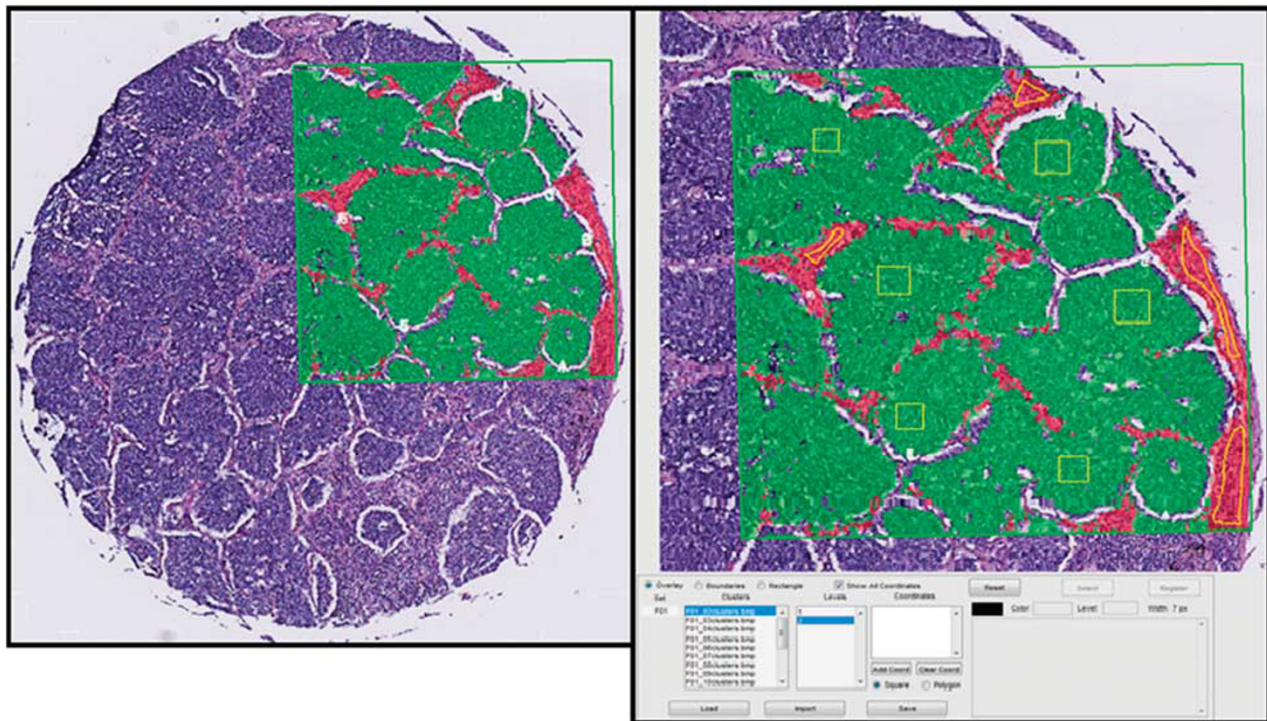


Figure 3 Screenshots of the CirecaAnnotate software. Left panel: H&E-stained tissue section with overlay of 1 mm × 1 mm IR map derived by HCA. Right panel: expanded area of image shown in left panel. This image was used by the pathologist to draw polygons, or freeform shapes, shown in yellow that select typical regions of a particular disease stage or tissue type. The yellow squares in the green overlay depict areas of SCLC, whereas the freeform areas in the red overlay depict fibro-connective tissue regions.

are based on voids or cracks in the tissue, which are readily identifiable. A minimum of three points must be selected on both panels, but more points are preferred. Figure 3 shows a typical screen of the registration result, where a 2-cluster HCA image is superimposed on the H&E image. Once the two images are registered, a magnified view is displayed, shown in Figure 3, right panel. Both images can now be zoomed without losing registration; furthermore, the number of clusters in the HCA overlay can be increased/decreased without losing registration. Once sufficiently good spatial resolution has been achieved by zooming in on the picture, the pathologist draws, by cursor action, free-form areas on the display that unambiguously define a region of homogeneous pathological diagnosis, and enters a diagnostic code for this region. Such regions are shown in Figure 3, right

panel for areas selected as being typical for SCLC (yellow squares in green areas) and fibro-connective tissue (yellow free-form shapes in red areas).

Annotation was carried out on the 1 mm × 1 mm areas for which HCA images were computed, as shown in Figure 3. Each region selected by the pathologist may contain between 50 and 200 pixel spectra, and up to 20 regions—albeit of different diagnostic codes—were typically selected from each tissue spot. Thus, the annotation procedure for each 1 mm × 1 mm tissue area will yield between 1000 and 2000 pixel spectra of high homogeneity and well-defined pathology. The only condition for this annotation step is that the region selected lies within one HCA cluster.

This procedure produced an annotated data set that was based on the pathology input available from the manu-

facturer of the TMA (BiomaxUS) as well as from the lung pathologist collaborating in this research project. The tissue areas selected for annotation are documented, and can be re-diagnosed if needed. A typical selection of regions used for annotation is shown in Figure 4. In this figure, the panels shown of cancerous spots measure $120\ \mu\text{m} \times 120\ \mu\text{m}$, whereas the panels for normal lung tissue measure $240\ \mu\text{m} \times 240\ \mu\text{m}$. The number of pixel spectra contained in each panel is about 380 for the cancerous spots, and about 1440 for the normal lung tissue. The areas in each panel marked in yellow designate the regions from which training spectra were extracted. The regions typically encompass an area between $\frac{1}{4}$ to $\frac{1}{2}$ of the panels depicted; thus, these regions define between 200 and 400 spectra to be included in the annotated data sets.

Post-Processing

The files produced by the CMAT routines were subsequently processed to extract all selected spectral files at the individual pixel level, and combine them into training and test databases. These databases were strictly separated by patient, as shown in Table 1. The total number of annotated pixel spectra was 106 000; thus, both the training and the test database contained 53 000 spectra in five classes (normal, SCLC, SqCC, ADC and BAC), or between *ca* 6000 and 15 000 spectra in each class of the training and test set. This ensured that no interdependent data were used in the training and test sets.

Diagnostic Algorithms

MATLAB-based implementations of feature selection and diagnostic algorithms were used in this study. Feature selection refers to a step in which all 512 data points of each spectral vector are searched for maximally discriminating features; that is, for the wavelengths at which the spectra differed maximally for the classes of cancers to be differentiated. The features to be used were determined using the MATLAB functions *rankfeatures*, employing *t-test* based feature selection.⁴⁷ Typically, the 60 most significant spectral features were utilized. The diagnostic classification algorithms (which differentiated individual tissue types on a per-pixel basis) were based on ANNs, invoked via the MATLAB function *feedforwardnet*, using the features selected previously as activation levels of the input neurons, a single hidden layer, and two output neurons (because all individual classifications were binary, forming nodes of a full binary decision tree). Details of the selected features will be discussed in more detail in section 'Analysis of the spectral features used by the ANNs for classification'.

The tree-based binary classification of all cancer and normal tissue types followed a scheme suggested by cluster analysis of the mean annotated spectra. To this end, mean spectra for 15 tissue classes from the annotation were calculated, and subject to unsupervised HCA. The results are represented in the form of a dendrogram,⁴⁵ shown in

Figure 5. Such dendrograms are frequently used in biology to study genetic similarities of, for example, bacterial species, or to express similarity in gene analysis.

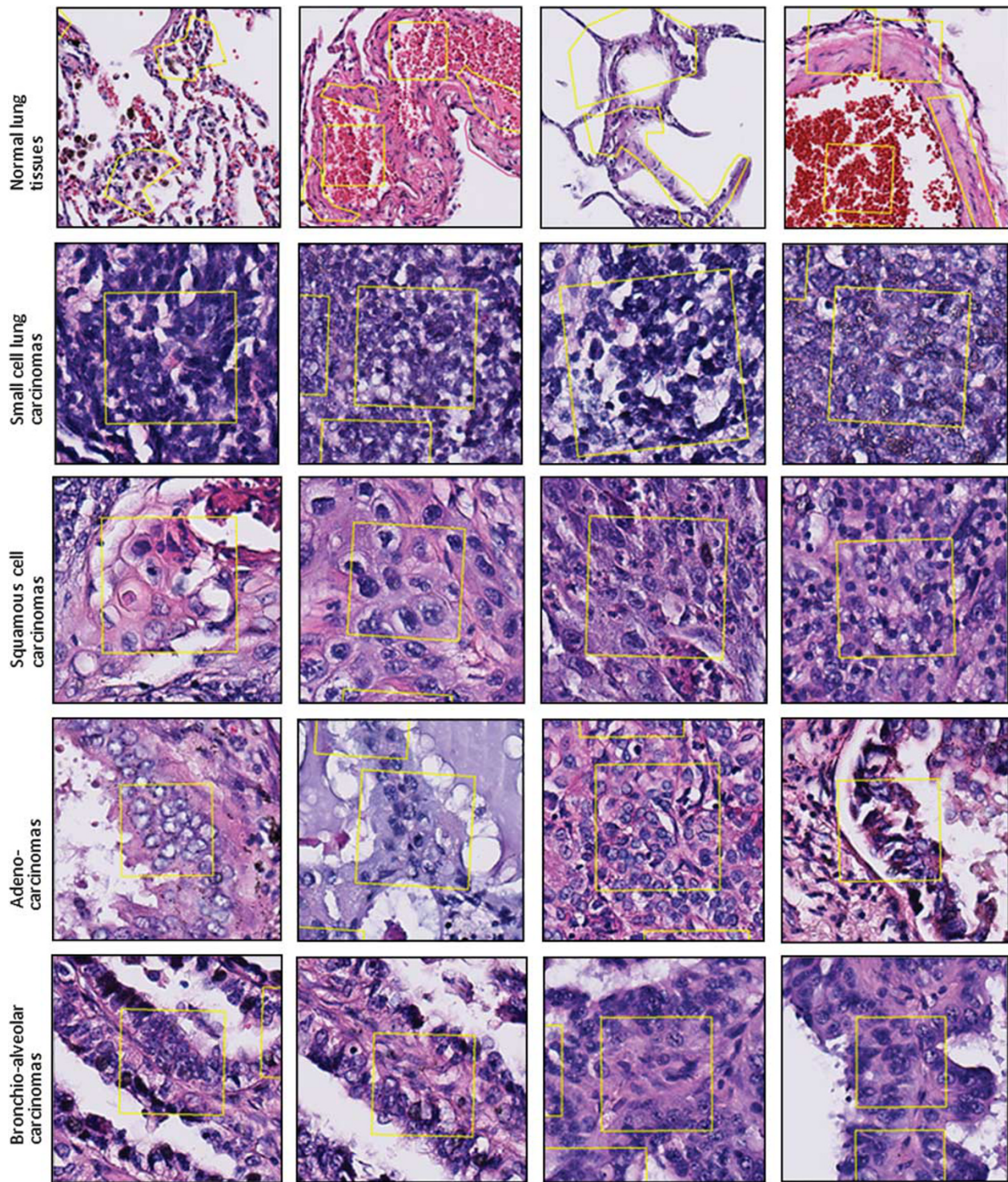
These HCA results were used to indicate an order for the binary classification steps to be carried out. The dendrogram shown in Figure 5 suggested that all cancerous spectra are grossly different from the spectra of normal tissue classes, and from necrotic tissues and tissues containing mucins. However, it also suggests that the mean spectra of lung ADC and BACs are very similar, but different from those of squamous cell and even more different than the spectra from SCLC. It should be noted that the method used to compute the dendrogram shown in Figure 5 is identical to that used to construct the HCA image shown in Figure 2, panel (b), which displays the similarity of spectral features in one entire data set in the form of a pseudo-color image, except that the results are displayed differently. Based on the dendrogram displayed in Figure 5, the binary classification scheme shown in Figure 6 was developed and followed for the analysis of the tissue microarray data set.

RESULTS AND DISCUSSION

Two different approaches were used in the diagnosis of the tissue microarray data set. The first approach dealt strictly with a scenario where both the training and test data sets consisted of annotated spectra, whereas the second approach applied the trained algorithms to entire tissue spots of the test set, subjecting them to the same binary classification scheme discussed above. Both types of analysis depend critically on two factors: the quality and the size of the annotated data set.

In order to define the annotation data set as rigorously as possible, the tissue spots on the BiomaxUS LC811 lung TMA, shown in Figure 7, were divided into training and test sets, as listed in Table 1. A selection of annotated tissue spot areas is shown in Figure 4; however, the reader is reminded that many areas were selected from each tissue spots, and that only a selection of the tissue spots is shown. Annotated spectra were also collected for the tissue spots in the test set to allow a diagnostic algorithm on a pixel spectrum basis. Figures 3 and 4 also demonstrate that each area marked by a yellow square yielded on average between 100 and 200 pixel spectra; thus, the total yield of annotated spectra varied between 1000 and 2000 pixel spectra for each tissue spot. In total, 106 000 annotated tissue spectra were available for normal tissues and the four cancer types. This number, and the number of cancer classes, exceeds the number of spectra annotated in detail in any previous study on SHP.

The size of the data sets for training and testing is of huge importance for a study, which tests new methodology for medical diagnostics. We found that typical diagnostic algorithms require about a thousand spectra of all classes for reliable prediction of unknown data sets. Furthermore, we found that the classification into five classes is best accomplished in a binary hierarchical fashion, as indicated



Normal lung, (L) to (R): spots H01, H03, H05, H07

Small cell lung carcinomas: spots E08, E10, F08, G10

Squamous cell carcinomas: spots C02, C05, C10, D04

Adenocarcinomas: spots A01, A09, B10, E02

Bronchio-alveolar carcinomas: spots D07, D07, E04, E04

Panels: 240 x 240 μm^2

Panels: 120 x 120 μm^2

Panels: 120 x 120 μm^2

Panels: 120 x 120 μm^2

Panels: 120 x 120 μm^2

Figure 4 Photomicrographs of selected tissue spots used for construction of the data sets. Each panel depicts areas selected (yellow squares) by the pathologist to indicate typical disease or tissue type. Spectra extracted from the identified areas of the same disease or tissue type from different patients were combined into the training and test databases.

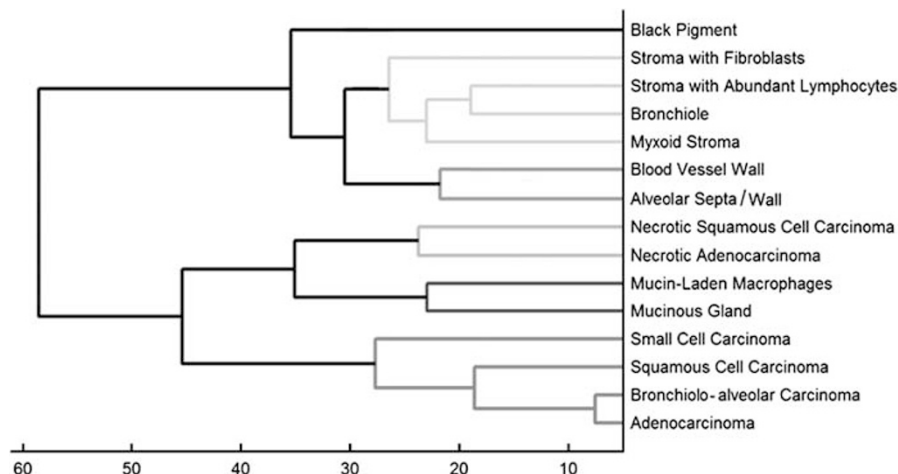


Figure 5 Display of the similarity of mean lung tissue spectra. This dendrogram was constructed by performing HCA on the mean spectra from each of the 15 diagnostic classes assigned by the pathologist. In this dendrogram, the spectral dissimilarity (increasing from right to left) is plotted along the abscissa. This graph demonstrates the high similarity between ADC and BAC, and the dissimilarity of cancer/necrosis/mucin-rich spectra from connective-tissue rich spectra.

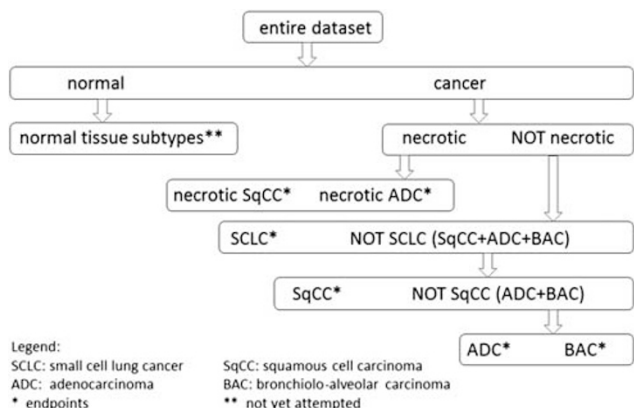


Figure 6 Binary classification scheme of entire annotated spectral data set of the lung tissue microarray. This scheme determined the order (hierarchy) of the ANN analysis. Endpoints of the present classification algorithms are marked by an asterisk.

in Figure 6. It is possible to train a single ANN for classifications involving more than two classes; however, it is advantageous to perform the classification such that the pairs of classes with the largest differences among them are classified in order from largest to smallest differences. This is suggested by inspection of the dendrogram presented in Figure 5, which indicates that the spectra of cancerous regions, while distinguishable, exhibit smaller spectral variances than other tissue classes. Accordingly, the ANNs were trained to distinguish the spectral classes in a hierarchical manner.

Owing to the restricted size of the total data sets, two steps in the classification scheme shown in Figure 6—the classification of normal tissue subtypes, and the automatic detection of necrosis (see below)—could not be carried out. In the case of normal tissues, there were nine normal diagnostic classes

(stroma with fibroblasts, stroma with abundant lymphocytes, bronchiole, myxoid stroma, blood vessel wall, alveolar wall, alveolar septa, mucinous gland and mucin-laden macrophages) for which insufficient number of pixel spectra could be annotated. As the main emphasis of this study was the distinction of the four non-necrotic cancer types, all the normal classes were combined into one. The spectra due to necrosis were removed from the analysis; however, they could be separated into originating from SqCCs or ADCs, based on their spectral patterns. The necrotic tissue spectra could be distinguished from non-necrotic cancer locations because necrosis induces a significant spectral change that has been detected and described very early in research efforts to detect disease by IR spectral methods; see below.⁴⁸

Classification Results/Pixel Spectrum Based

The first classification scheme used the 106 000 annotated spectra, which were separated into completely independent training and test sets; see Table 1. According to the discussion in the previous paragraph, the top level ANN (ANN level 1, see Table 2) was trained to distinguish NORMAL from NOT NORMAL spectra (equivalent to CANCER vs NOT CANCER). To this end, the training set for NORMAL tissue included various non-cancerous tissue types, for example, normal fibro-connective tissue spectra from cancerous tissue spots, as well as several other normal tissue features (endothelium, connective tissue) from normal tissue spots. The NOT NORMAL spectra were randomly selected from the patient-separated tissue spots representing the four cancer types. A total of 1848 NOT NORMAL spectra and 1840 NORMAL spectra were used in the training set, where the number of spectra used was determined by the smallest number of patient spectra in one of the cancer data sets. The top level ANN distinguished the NORMAL from the NOT

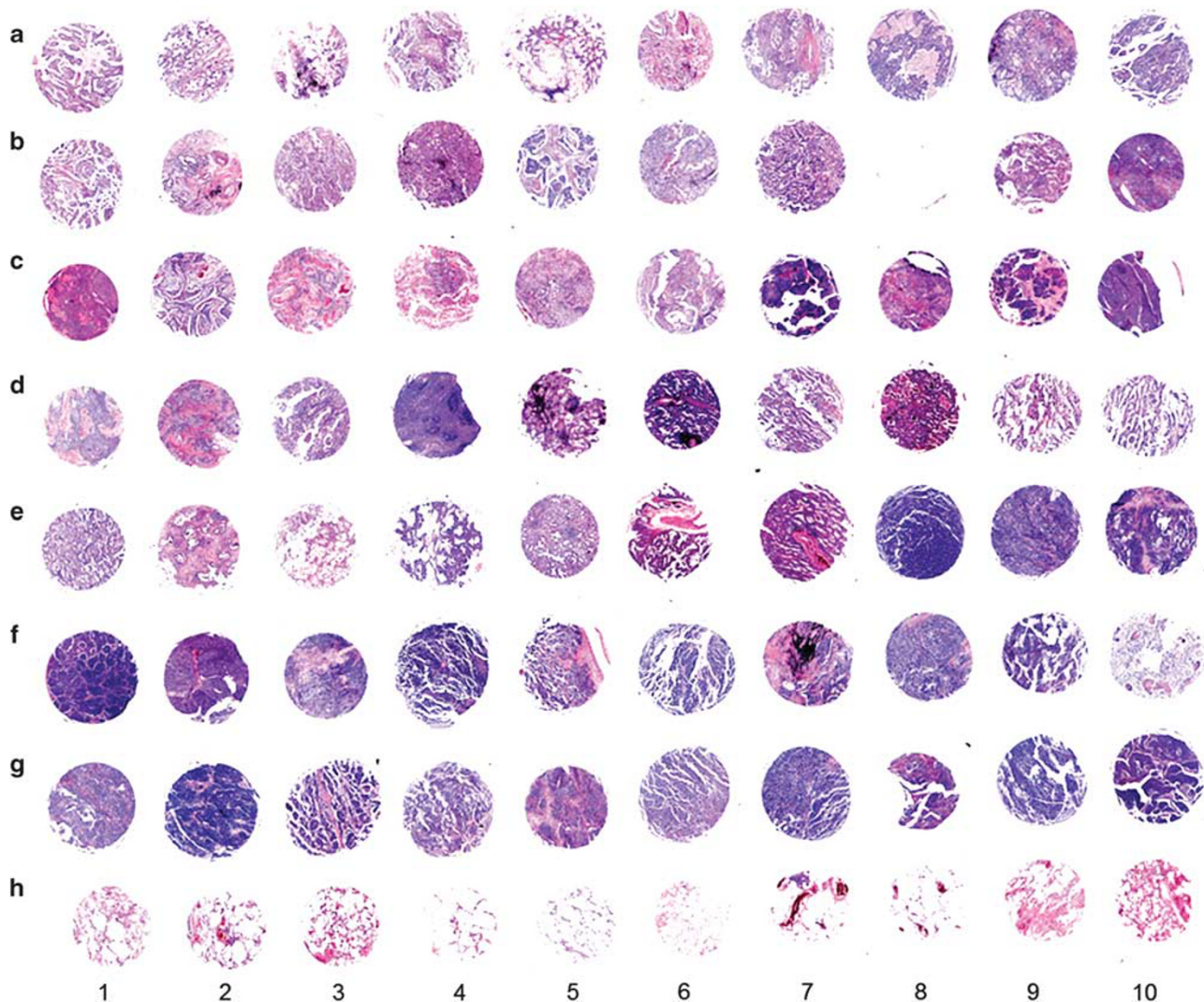


Figure 7 Low-resolution image of Biomax LC811 TMA. The letters a–h on the left, and the numbers 1–10 on the bottom define each tissue spot.

Table 2 Sensitivity, specificity and accuracy of hierarchical ANNs for classification of lung cancers

ANN level	Classification	Sensitivity	Specificity	Accuracy
Level 1 ANN	NOR vs NOT NORMAL	99.3%	94.4%	96.8%
Level 2 ANN	SCLC vs NOT SCLC	91.2%	98.0%	94.6%
Level 3A ANN	SqCC vs NOT SqCC (1)	90.4%	95.0%	92.7%
Level 3B ANN	SqCC vs NOT SqCC (2)	97.3%	99.6%	98.4%
Level 4 ANN	ADC vs BAC	88.8%	47.2%	68.0%

NORMAL spectra with a pixel-level sensitivity of 99.3% and a specificity of 94.4%, when applied to the entire test set, see Table 2. This ANN used a feature selection⁴⁷ of the 65 most significant (second derivative) intensity points. The neural network topography (number of hidden layers, nodes in the hidden layer and the number of input features) affected the

network performance only minimally; the values of sensitivity and specificity changed by <1%. Thus, the reported results are for ANN structures with one hidden layer that contained five nodes.

As there were insufficient spectra in the data set to train an ANN for necrosis, the necrotic spectra were removed from the data sets. The effect of necrosis on the observed vibrational spectra was first reported by Jamin, *et al*,⁴⁸ in these spectra, a strong shoulder of the amide I peak at *ca* 1630 cm⁻¹ was reported (in fact, this shoulder was reported to have higher intensity than the ‘main’ amide I peak at 1655 cm⁻¹). The large spectral changes observed for necrosis indicate major changes in the protein composition of necrotic cells, as the 1630 cm⁻¹ peak is associated with unfolded and precipitated proteins.^{34,49}

The second derivative spectra used in our study displayed the ‘necrosis signal’ as a sharp, very intense peak at *ca* 1630 cm⁻¹ in the protein amide I manifold, which was so

prominent that it confounded the ANNs. Thus, spectra exhibiting the 'necrosis signal' were removed from the data set. Subsequently, the pixels in the tissue spots exhibiting the necrotic spectral patterns were identified as being necrotic by the pathologist. Furthermore, the spectra in the 'necrosis' data subset could be distinguished, by SHP, as originating from either adeno- or SqCC.

The remaining data set was subject to a second level ANN (level 2 ANN, see Table 2) to distinguish between SCLC and NOT SCLC, which includes SqCC, ADC and BAC. To this end, 5280 spectra from the SCLC training set were selected randomly. The NOT SCLC training set consisted of 1760 spectra each, selected randomly from the three cancer classes, SqCC, ADC and BAC. This second level ANN distinguished SCLC from NOT SCLC spectra from the test set (again at an individual pixel-level) with a sensitivity of 91.2% and a specificity of 98.0%. The neural network structure was the same as discussed above.

A third level ANN (level 3 ANN, see Table 2) in the tree of binary classification was trained for the distinction of SqCC from NOT SqCC (ADC and BAC). This distinction is clinically highly significant, because treatment options are quite different between adeno- and SqCCs.⁵⁰ Training of this algorithm was accomplished using 6640 SqCC spectra from the training set and 6640 NOT SqCC (ADC and BAC) spectra. For this ANN, two different feature selection methods were used: the feature selection introduced above resulted in 90.4% sensitivity and 95.0% specificity when applied to the test set, whereas a feature selection based on PCA scores gave significantly better results, 97.3% sensitivity and 99.6% specificity. We attribute this increased sensitivity and specificity of the PCA scores-based ANN to the fact that PCA scores contain the entire spectral variance, and the correlation between different intensity values, whereas a simple feature selection utilizes the spectral variance at a pre-determined number of intensity points.

Finally, a fourth level ANN was trained to differentiate ADC from BAC. As BAC originates in the alveolar lining of lung tissue, it is considered a precursor of ADC. The pathological distinction between the two is based mainly on whether or not the neoplasm has penetrated through the basement membrane (ADC); if not, the disease is classified as BAC. From the viewpoint of the biochemical composition of the cancer cells, the two diseases differ mostly in the stage of disease; thus, SHP was not able to distinguish these two states reliably (with an acceptable sensitivity of 88.8%, but a low specificity of 47.2%). Larger data sets and even more careful annotation may increase the accuracy of this diagnosis as well. We note that—once the classifications are made on the TMA spot level rather than the pixel level (so that individual pixel scores can be aggregated)—the performance on sensitivity and specificity are expected to improve significantly.

Although the spectral differences among the cancer types are small, they can be perceived by inspection of the mean (second derivative) spectra of the different cancer types (see

section 'Analysis of the spectral features used by the ANNs for classification'). For example, SCLC spectra, in general, display stronger DNA/RNA spectral peaks than the other cancerous types. At this point, it is possible to qualitatively interpret some of the spectral differences between the cancer classes in terms of biochemical changes within the tissue, similar to the approach taken in the past⁵¹ for SCP. There, the distinction between cancerous oral mucosal cells from normal ones differed by the spectral signature of keratin, among other changes, in the case of keratinizing SqCC.

The discussion in the next two sections concentrates on the classification of four cancer types, and 'normal' (ie, NOT cancerous) tissues only, and does not include a discussion of future possibilities of detecting individual cancer markers by spectral means. However, there are preliminary data from another laboratory that suggest that data mining techniques of IR imaging data sets against IHC can reveal the presence of certain marker proteins. This could, in the future, lead to label-free methods of detecting markers of disease, and prognostic information.

Classification Results/Tissue Spot Based

To test how well the diagnostic algorithms discussed in the previous section performed on test spectra that were not annotated at the pixel level, entire tissue spots from the test set were subjected to ANN analysis. This was accomplished as follows. The raw data set from entire spots, consisting of *ca* 57600 pixel spectra each, were preprocessed as discussed above (section 'Data pre-processing'). However, rather than pre-segmenting the data set by HCA, as was done for the annotation and pixel-based diagnostic tests, all pixel spectra of each spot were analyzed directly by the trained ANNs, and the pixel-level output of the ANN was converted to a graphical binary format. Thus, in this format, each pixel spectrum analyzed by the various level ANN's can have a binary output, 'YES' coded in red, or 'NO', coded in green. Depending on the ANN, the red and green areas can have different diagnostic meanings, as shown in Figure 8.

When the 'Level 1 ANN' algorithm (CANCER *vs* NOT CANCER) was applied to tissue spot H9 from the 'normal' tissue spot test set (top row of Figure 8), the green areas imply 'NO' or NOT CANCER, which is the correct diagnosis for this normal tissue spot. When the same Level 1 ANN was applied to spot E8 from the SCLC test database (Figure 8, second row), the answer was 'YES', or positive for cancer. When the same spot was analyzed by the level 2 ANN, SCLC *vs* NOT SCLC, the results again was YES (positive for SCLC) as indicated by the red display.

The third row depicts applications of the level 1 ANN, level 2 ANN and level 3 ANN (left to right) to tissue spot D2 diagnosed as SqCC by classical methods. Application of the level 1 ANN reported the fibro-connective tissue areas as 'NO' (NOT CANCER) in green, and the cancerous regions in red. The level 2 ANN subsequently analyzed the CANCER areas, but determined that they were NOT SCLC, hence they

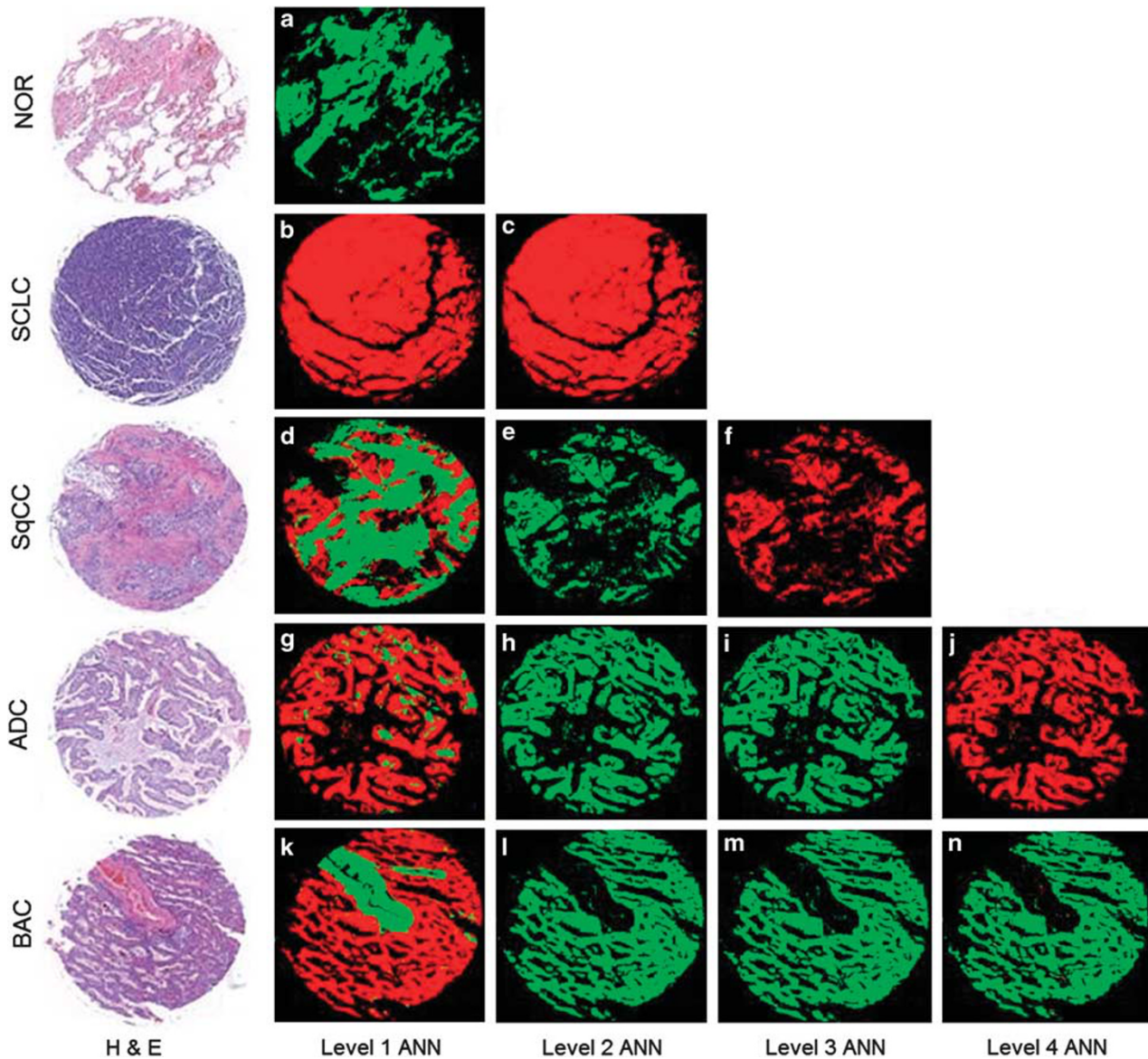


Figure 8 Examples of results from full tissue spot based analyses (see text for details). Top row (a): analysis of normal tissue spot H9 by level 1 ANN (cancer vs NOT cancer, see Table 2), which properly assigned the vast majority of spectra as normal. Second row (b): analysis of SCLC spot E8 by level 1 ANN, which correctly assigned the vast majority of spectra as cancer. (c) Analysis of same spot by level 2 ANN (SCLC vs NOT SCLC, see Table 2), which correctly assigned the vast majority of spectra as SCLC. Third row (d): analysis of SqCC spot D2 by level 1 ANN, which correctly classified regions of SqCC and fibro-connective tissue. (e) Analysis of same spot by level 2 ANN, which depicted fibro-connective tissue in black, and cancerous regions in green (because they were NOT SCLC). (f) Analysis of same spot by level 3 ANN (SqCC vs NOT SqCC, see Table 2), which correctly assigned the vast majority of spectra as SqCC. Fourth row (g): analysis of ADC spot A1 by level 1 ANN, which correctly classified regions of cancer (red), and a few areas of NOT cancer. (h) Analysis of same spot by level 2 ANN, which depicted all cancerous regions as NOT SCLC (green), whereas the non-cancerous regions appeared in black. (i) Analysis of same spot by level 3 ANN, which depicted all cancerous regions as NOT SqCC (green), whereas the non-cancerous regions appeared in black. (j) Analysis of same spot by level 4 ANN (ADC vs BAC, see Table 2), which depicted all cancerous regions as ADC (red). Fifth row (k): analysis of BAC spot E7 by level 1 ANN, which correctly classified regions of cancer (red), and areas of NOT cancer in green. (l) Analysis of same spot by level 2 ANN that depicted fibro-connective tissue in black, and cancerous regions in green (because they were NOT SCLC). (m) Analysis of same spot by level 3 ANN that depicted all cancerous regions as NOT SqCC (green), whereas the non-cancerous regions appeared in black. (n) Analysis of same spot by level 4 ANN that depicted all cancerous regions as NOT ADC (green).

were displayed in green. Level 3 ANN analyzed the cancerous regions, and found them to be positive (red) for SqCC.

The fourth row in Figure 8 depicts analysis of tissue spot A1, which was diagnosed as ADC by classical methods. The

level 1 ANN properly defined most areas as cancerous (red) with just a few normal regions. Application of the level 2 ANN and the level 3 ANN both revealed negative results (green) when analyzing the cancerous regions, because the

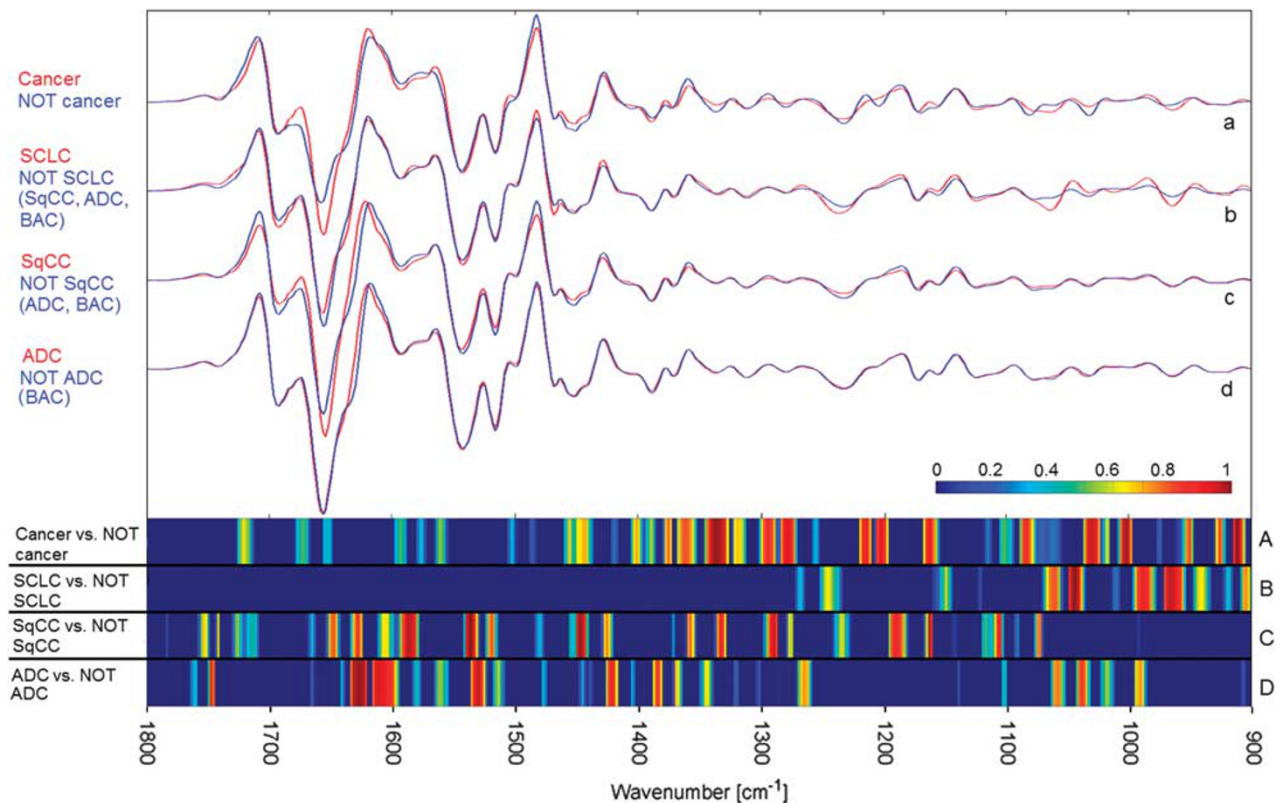


Figure 9 Comparison of mean second derivative spectral features (top, traces a–d) and heat map of features selected by the ANNs for the binary discrimination (panels A–D) of tumor classes. See text for details.

cancer was NOT SCLC and NOT SqCC. However, level 4 ANN recognized the cancer as ADC. Notice that the areas shown in green in the leftmost picture of row 4 appear blank in the output of the level 4 ANN, because the algorithm properly detects them as ‘NOT ADC’.

Finally, in row 5 of Figure 8, results are shown for spot E7 diagnosed as BAC by classical histopathology. The majority of the spot was identified as cancerous except a region of blood vessel with red blood cells within its lumen in the upper left quadrant that were clearly diagnosed as normal by ‘level 1 ANN’. Level 2 ANN, level 3 ANN and level 4 ANN found ‘NOT SCLC’, ‘NOT SqCC’ and ‘NOT ADC’, respectively, which were the correct diagnoses for a tissue spot diagnosed with BAC.

In summary, Figure 8 demonstrates that a decision tree of hierarchical, binary ANNs can be used to analyze for the presence of various cancers. The consecutive applications of these algorithm requires <1 min once the training of the algorithms is accomplished.

Analysis of the Spectral Features Used by the ANNs for Classification

At this point, it is instructive to investigate the spectral features used by the individual neural networks for the classification of the cancer types. This analysis can be presented in

the form of a ‘heat map’. Heat maps are commonly used in IHC or gene array studies to visually demonstrate that markers or features are responsible for a classification decision. Figure 9 shows a heat map of the spectral features used by each of the four ANNs, along with the mean class spectra. The details of this figure will be discussed next.

The binary neural networks were trained on data sets containing large numbers (thousands) of spectra, see section ‘Classification results/pixel spectrum based’. *T*-test-based feature selection was carried out, as discussed above, to select the spectral features that created the best discrimination of spectra in this binary, two-class approach. The top pair of traces in Figure 9, marked ‘a’, shows the mean spectra of the cancer (red) and NOT cancer spectral classes (blue). Note that all normal tissue classes described in Figure 5 were contained in this latter data set. Similarly, all cancers (SCLC, SqCC, ADC and BAC) were contained in the former set. Visual inspection of this pair of spectral traces reveals systematic differences between the two classes. The importance of these spectral differences is emphasized in the ‘Cancer vs NOT cancer’ heat map display, marked ‘a’. In this trace, dark blue hues indicates no significance, whereas red and reddish brown colors indicate high significance (see color scale in Figure 9). This heat map demonstrates that the algorithm did not utilize all wavenumber ranges where different spectral

features can be observed, but selected ranges of most significance for the desired discrimination. On the other hand, it is important to ascertain that regions of high significance, indicated in the heat map actually exhibit spectral intensities or intensity differences between the two classes. Combinations of individual features selected by the algorithm may be viewed as 'spectral marker sets' for a specific condition or disease.

The differentiation of the SCLC and NOT SCLC (SqCC, ADC and BAC) classes occurred in the low wavenumber region of the spectra (900–1500 cm^{-1} , see trace pair 'b'). This result agrees well with results from visual inspection of SCLC spectra, which always exhibit distinct DNA-associated peaks at *ca.* 1235, 1090, 1065 and 965 cm^{-1} .⁹ These regions are indicated in the heat map 'B' as most significant in distinguishing SCLC from the other cancers. The distinction between SqCC and NOT SqCC (ADC and BAC) uses spectral features spread over the entire spectral range, as shown in C. The spectral differences found in the 1000–1250 and 1580–1620 cm^{-1} regions, shown in trace pair 'c', are most likely associated with glycoproteins. ADC and BAC originate in mucus-producing alveolar cells, in contrast to SqCC. Mucus, a glycosylated protein, exhibits spectroscopic features of the carbohydrates between 1000 and 1250 cm^{-1} ,⁵² and generally slightly different protein conformational features, which show up in the region between the amide I and amide II manifolds.

Similarly, the features differentiating ADC and BAC are spread over the entire spectral range, however, the mean spectra shown in traces marked 'd' are very similar. This is in line with the histopathological view that BAC (also known as Tis, *in situ*) represents an early stage of ADC where the cancerous cells are confined to the top layer of alveolar tissue. As these tissues—and consequently, their spectra—are so similar, the algorithm had to rely on a number of small features for the differentiation of these classes.

It should be pointed out that the heat map diagram shown in Figure 9a–d are the output of the *t*-test-based feature selection algorithms, not the actual weights assigned to each of the features by the ANN. These latter weights can be displayed, but are complicated by the fact that the hidden layer in the ANN consists of several individual nodes, and the weights triggering each node need to be displayed. Thus, the output of these weights produces a less intuitive display than the feature selection plots shown here. Nevertheless, the heat plots in Figure 9 provide information on the selection of distinguishing markers that can be interpreted on the basis of biochemical composition.

Conclusions

This study demonstrates that SHP can classify tissues that present different cancer types, with overall accuracies comparable to that of multi-panel IHC. This is one of the largest scale SHP studies to date, employing an annotated data set totaling over 100 000 spectra, with an overall data set

comprising over 3 million tissue spectra from over 70 patients. In this study, the locations from which training and test spectra were collected are documented, and the diagnoses of these spectra are verified by classical histopathology. The experimental methods utilized in this study represent the most automated procedures for data acquisition and pre-processing. The pre-processing steps, in particular, are based on most recent research and understanding of confounding effects in IR micro-spectroscopy,^{53,54} and use procedures (such as segmenting of second derivative data by HCA, and subsequent analysis by ANN) that have been shown by several research groups^{11,49,55,56} to produce the most reliable data and most reproducible diagnostic algorithms. In addition, the software and procedure developed during this study for pathology-based annotation are novel, and allow a more reliable and reproducible classification of the training spectra.

In this pilot study, we demonstrated the diagnostic value of SHP in classifying various cancerous and normal states in lung histopathology. Once sufficiently large training sets for a particular organ have been established, and machine learning algorithms have been trained, the SHP methodology can easily be incorporated into standard pathology workflow, because it requires nothing but an unstained tissue section to be cut when the samples for classical pathology and IHC are prepared. This slide can be analyzed whereas the standard slides are stained and cover slipped, and the SHP results can be available, for example as an overlay with a standard H&E-stained image as shown in Figure 2 or 3, to direct a pathologist toward areas of high interest or to remove areas of low or unambiguous diagnostic value. Furthermore, recent results have indicated that the signatures related to individual cancer markers⁵⁷ and disease progression and prognosis⁵⁸ can be derived by vibrational spectroscopic means, thereby adding prognostic and therapeutic value. This, in turn, can lead to a platform that combines diagnosis, prognosis and therapeutic information in one comprehensive laboratory procedure. Furthermore, SHP has a distinct advantage over IHC: in the latter technique, only the markers, which are pre-selected and included in the test panel can produce a signal; that is, it is impossible to detect any unselected markers. However, the spectral signatures of such markers will be included in the IR spectral information.

ACKNOWLEDGEMENTS

Early support of this research to establish the methodology was provided under grant CA111330 (to MD) by the NIH/NCI. We are grateful for the present financial support of this work under a license agreement between Northeastern University and Cireca Thernostics, LLC.

DISCLOSURE/CONFLICT OF INTEREST

Cireca Thernostics, LLC provided financial support for this study.

1. Romeo MJ, Dukor RK, Diem M. Introduction to spectral imaging, and applications to diagnosis of lymph nodes. In: Diem M, Griffiths PR, Chalmers JM (eds). *Vibrational Spectroscopy for Medical Diagnosis*. John Wiley and Sons: Chichester, UK, 2008, pp 1–26.

2. Bahlmann C, Patel A, Johnson J, *et al*. Automated detection of diagnostically relevant regions in H&E stained digital pathology slides. Proc SPIE 'Medical Imaging 2012: Computer-Aided Diagnosis' 2012; 8315:831504.
3. Jagidar J. Application of immunohistochemistry to the diagnosis of primary and metastatic carcinoma to the lung. Arch Pathol Lab Med 2008;132:384–396.
4. Yoo J, Jung JH, Lee MA, *et al*. Immunohistochemical analysis of non-small cell lung cancer: correlation with clinical parameters and prognosis. J Korean Med Sci 2007;22:318–325.
5. Romeo MJ, Boydston-White S, Matthäus C, *et al*. Vibrational microspectroscopy of cells and tissues. In: Lasch P, Kneipp J (eds). Biomedical Vibrational Spectroscopy. Wiley-Interscience: Hoboken, NJ, 2008, pp 121–147.
6. Park SY, Follen M, Milbourne A, *et al*. Automated image analysis of digital colposcopy for the detection of cervical neoplasia. J Biomed Optics 2008;13.
7. Utzinger U, Heintzelmann DL, Mahadevan-Jansen A, *et al*. Near IR Raman spectroscopy for *in vivo* detection of cervical precancers. Appl Spectrosc 2001;55:955–959.
8. Chiriboga L, Xie P, Yee H, *et al*. IR spectroscopy of human tissue. IV. Detection of dysplastic and neoplastic changes of human cervical tissue via infrared microspectroscopy. Cell Mol Biol 1998;44:219–229.
9. Diem M, Griffiths PR, Chalmers JM. Vibrational Spectroscopy for Medical Diagnosis. In: Diem M, Griffiths PR, Chalmers JM (eds). John Wiley and Sons: Chichester, UK, 2008.
10. Bird B, Miljkovic M, Romeo MJ, *et al*. Infrared micro-spectral imaging: automatic distinction of tisuetypes in axillary lymph node histology'. BMC J Clin Pathol 2008;8:1–14.
11. Lasch P, Diem M, Hänsch W, *et al*. Artificial neural networks as supervised techniques for FT-IR microspectroscopic imaging. J Chemometrics 2007;20:209–220.
12. Bird B, Romeo M, Laver N, *et al*. Spectral detection of micro-metastases in lymph node histopathology. J Biophotonics 2009;2:37–46.
13. Bird B, kovæ M, Laver N, *et al*. Spectral detection of micro-metastases and individual metastatic cells in lymph node histology. Tech Cancer Res Treatment 2011;10:135–144.
14. Bird B, Bedrossian K, Laver N, *et al*. Detection of breast micro-metastases in axillary lymph nodes by infrared micro-spectral imaging. Analyst 2009;134:1067–1076.
15. Wood BR, Chiriboga L, Yee H, *et al*. FTIR mapping of the cervical transformation zone, squamous and glandular epithelium. Gynecol Oncol 2004;93:59–68.
16. Wolthuis R, Travo A, Nicolet C, *et al*. IR spectral imaging for histopathological characterization of xenografted human colon carcinomas. Anal Chem 2008;80:8461–8469.
17. Lasch P, Haensch W, Lewis EN, *et al*. Characterization of colorectal adenocarcinoma sections by spatially resolved FT-IR microspectroscopy. Appl Spectrosc 2002;56:1–9.
18. Lasch P, Naumann D. FT-IR microspectroscopic imaging of human carcinoma in thin sections based on pattern recognition techniques. Cell Mol Biol 1998;44:189–202.
19. Fernandez DC, Bhargava R, Hewitt SM, *et al*. Infrared spectroscopic imaging for histopathology recognition. Nat Biotech 2005;23:469–474.
20. Pounder FN, Bhargava R. Toward automated breast histopathology using mid-IR spectroscopic imaging. In: Srinivasa G (ed). Vibrational Spectroscopic Imaging for Biomedical Applications. McGraw Hill: New York, 2010.
21. Krafft C, Shapoval L, Sobottka SB, *et al*. Identification of primary tumors of brain metastases by SIMCA classification of IR spectroscopic images. Biochem Biophys Acta 2006;1758:883–891.
22. Amharref N, Beljebbar A, Dukic S, *et al*. Brain tissue characterization by infrared imaging in a rat glioma model. Biochem Biophys Acta 2006;1758:892–899.
23. Bamberg KR, Schültke E, Wood BR, *et al*. A Fourier transform infrared microspectroscopic image investigation into an animal model exhibiting glioblastoma multiforme. Biochem Biophys Acta 2006;1758:900–907.
24. Lasch P, Kneipp J. Biomedical vibrational spectroscopy. In: Lasch P, Kneipp J (eds). Biomedical Vibrational Spectroscopy. Wiley-Interscience: Hoboken, NJ, 2008, pp 121–147.
25. Krafft C, Diderhoshan MA, Recknagel P, *et al*. Crisp and soft multivariate methods visualize individual cell nuclei in Raman images of liver tissue sections. Vibr Spectrosc 2011;55:90–100.
26. Horsnell J, Stonelake P, Christie-Brown J, *et al*. Raman spectroscopy - a new method for the intra-operative assessment of axillary lymph nodes. Analyst 2010;135:3042–3047.
27. Hutchings J, Kendall C, Smith B, *et al*. The potential for histological screening using a combination of rapid Raman mapping and principal component analysis. J Biophoton 2009;2:91–103.
28. Barr H, Kendall C, Stone N. Raman spectroscopy as a potential tool for early diagnosis of malignancies in esophageal and bladder tissues. In: Diem M, Griffiths PR, Chalmers JM (eds). Vibrational Spectroscopy for Medical Diagnosis. John Wiley and Sons: Chichester, UK, 2008, pp 203–230.
29. Papamarkakis K, Bird B, Schubert JM, *et al*. Cytopathology by optical methods: spectral cytopathology of the oral mucosa. Lab Invest 2010;90:589–598.
30. Schubert JM, Bird B, Papamarkakis K, *et al*. Spectral cytopathology of cervical samples: detecting cellular abnormalities in cytologically normal cells. Lab Invest 2010;90:1068–1077.
31. Schubert JM. Spectral cytology of human oral and cervical samples. PhD Dissertation, Northeastern University: Boston, 2011.
32. Bhargava R, Levin IW. Infrared spectroscopic imaging protocols for high-throughput histopathology. In: Diem M, Griffiths PR, Chalmers JM (eds). Vibrational Spectroscopy for Medical Diagnosis. John Wiley and Sons: Chichester, UK, 2008, pp 155–186.
33. Diem M. Introduction to Modern Vibrational Spectroscopy. Wiley-Interscience: New York, 1993.
34. Barth A. Infrared spectroscopy of proteins. Biochim Biophys Acta 2007;1767:1073–1101.
35. Gerwert K. Molecular reaction mechanisms of proteins monitored by time-resolved FTIR spectroscopy. Biol Chem 1999;380:931–935.
36. Gerwert K. Molecular reaction mechanisms of proteins monitored by time-resolved FT-IR difference spectroscopy. In: Gremlich HU, Yan B (eds). Infrared and Raman Spectroscopy of Biological Materials. Marcel Dekker Inc.: New York, 2001, pp 193–230.
37. Whelan DR, Bamberg KR, Heraud P, *et al*. Monitoring the reversible B to A-like transition of DNA in eukaryotic cells using Fourier transform infrared spectroscopy. Nucleic Acids Res 2011;39:5439–5448.
38. Rerek ME, Van Wyck D, Mendelsohn R, *et al*. FTIR spectroscopic studies of lipid dynamics in phytosphingosine ceramide models of the stratum corneum lipid matrix. Chem Phy Lipids 2005;134:51–58.
39. Diem M, Chiriboga L, Yee H. Infrared spectroscopy of human cells and tissue VIII. Strategies for analysis of infrared tissue mapping data and applications to liver tissue. Biopolymers 2000;57:282–290.
40. Lasch P. Computergestuetzte Bildrekonstruktion auf Basis FTIR-mikrospektrometrischer Daten humaner Tumoren. PhD Dissertation, Freie Universitaet Berlin: Berlin, 1999.
41. Reddy RK, Bhargava R. Accurate histopathology from low signal-to-noise ratio spectroscopic imaging data. Analyst 2010;135: 2818–2815.
42. Savitzky A, Golay MJE. Smoothing and differentiation of data by simplified least-squares procedures. Anal Chem 1964;36:1627–1639.
43. Diem M, Bird B, Miljkovic M. Phase correction to compensate for reflective distortions of optical spectra. US Patent Office, USA, 2011.
44. Miljkovic M, Bird B, Diem M. Dispersive line shape effects in infrared spectroscopy. Analyst 2012 (submitted).
45. Adams MJ. Chemometrics in Analytical Spectroscopy, 2nd edn. In: Barnett NW (ed). Royal Society of Chemistry: Cambridge, 2004.
46. Wood BR, Chiriboga L, Yee H, *et al*. Fourier transform infrared (FTIR) spectral mapping of the cervical transformation zone, and dysplastic squamous epithelium. Gynecol Oncol 2004;93:59–68.
47. Lee H-M, Chen CM, Chen JM, *et al*. An efficient fuzzy classifier with feature selection based on fuzzy entropy. IEEE Trans Systems Man Cybernetics 2001;31:426–432.
48. Jamin N, Miller L, Moncuit J, *et al*. Chemical heterogeneity in cell death: combined synchrotron IR and fluorescence microscopy studies of single apoptotic and necrotic cells. Biopolymers (Biospectroscopy) 2003;72:366–373.
49. Lasch P, Beekes M, Fabian H, *et al*. Antemortem identification of transmissible spongiform encephalopathy (TSE) from serum by mid-infrared spectroscopy. In: Diem M, Griffiths PR, Chalmers JM (eds). Vibrational Spectroscopy for Medical Diagnosis. John Wiley & Sons: Chichester, UK, 2008, pp 97–122.
50. Jefferson T, Leung S, Laskin J, *et al*. Optimal immunohistochemical markers for distinguishing lung adenocarcinomas from squamous

- cell carcinomas in small tumor samples. *Am J Surg Path* 2010;34:1805–1811.
51. Diem M, Papamarkakis K, Schubert J, *et al*. The infrared spectral signatures of disease: extracting the distinguishing spectral features between normal and diseased states. *Appl Spectrosc* 2009;63:307A–318A.
 52. Chiriboga L, Xie P, Vigorita V, *et al*. Infrared spectroscopy of human tissue. II. A comparative study of spectra of biopsies of cervical squamous epithelium and of exfoliated cervical cells. *Biospectroscopy* 1998;4:55–59.
 53. Bassan P, Byrne HJ, Lee J, *et al*. Reflection contributions to the dispersion artefact in FTIR spectra of single biological cells. *Analyst* 2009;134:1171–1175.
 54. Bassan P, Kohler A, Martens H, *et al*. Resonant Mie Scattering (RMieS) correction of infrared spectra from highly scattering biological samples. *Analyst* 2010;135:268–277.
 55. Lasch P, Haensch W, Naumann D, *et al*. Imaging of colorectal adenocarcinoma using FT-IR microspectroscopy and cluster analysis. *Biochim Biophys Acta* 2004;1688:176–186.
 56. Krafft C, Salzer R. Neuro-oncological applications of infrared and Raman spectroscopy. In: Diem M, Griffiths PR, Chalmers JM (eds). *Vibrational Spectroscopy for Medical Diagnosis*. John Wiley and Sons: Chichester, UK, 2008.
 57. Hartsuiker L, Zeijen NJL, Terstappen LWMM, *et al*. A comparison of breast cancer tumor cells with varying expression of the Her2/neu receptor by Raman microspectroscopic imaging. *Analyst* 2010;135:3220–3226.
 58. Gazi E, Baker M, Dwyer J, *et al*. A correlation of FTIR spectra derived from prostate cancer tissue with Gleason grade, PSA and tumour stage. *Eur Urol* 2006;50:750–761.