PATHOBIOLOGY IN FOCUS

Classification of malignant and benign tumors of the lung by infrared spectral histopathology (SHP)

Ali Akalin[1], Xinying Mu[2,3], Mark A Kon[2], Ayşegül Ergin[3], Stan H Remiszewski[3], Clay M Thompson[3,4], Dan J Raz[5] and Max Diem[3,6]

We report results of a study utilizing a novel tissue classification method, based on label-free spectral techniques, for the classification of lung cancer histopathological samples on a tissue microarray. The spectral diagnostic method allows reproducible and objective classification of unstained tissue sections. This is accomplished by acquiring infrared data sets containing thousands of spectra, each collected from tissue pixels $\sim 6\,\mu$m 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 that 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 is presented, suggesting that spectral histopathology can achieve levels of diagnostic accuracy that is comparable to that of multipanel immunohistochemistry.

This paper introduces optical techniques developed during the past 15 years that represent novel approaches and paradigms for the diagnosis of cells and tissues, and are referred to as 'spectral cytopathology' (SCP) and 'spectral histopathology' (SHP), respectively. Early reports in these fields, particularly in SCP, were fraud with instrumental artifacts and misinterpretations[1–3] of data and led to some exaggerated claims. However, in 1998 (ref. 4) and 1999,[5] detailed correlation with classical cytopathology and histopathology led to the establishment of these techniques as viable methods to aid in medical diagnostics.

SHP and SCP rely on objective spectroscopic measurements, in conjunction with machine learning algorithms (MLAs) for the classification of spectral data, to render a medical diagnosis. The sensitivity and specificity of SHP equal or surpass those of other recent improvements to classical techniques, such as whole slide imaging,[6] immunohistochemistry,[7] imaging mass spectrometry, or gene-based methodologies. Among the major advantages of SHP is the ease with which it can be incorporated into the present pathology workflow, its nondestructive nature, the minimal sample preparation, the traceability and high spatial resolution

afforded by this technique, its sensitivity to tumor heterogeneity, and the fact that it is a completely reproducible, quantitative, and objective method.

In this paper, we expand on the results of an SHP pilot study that was reported in 2012 in this journal.[8] This pilot study, based on 80 patient samples in tissue micro-array (TMA) format, represented the first successful attempt to classify, by SHP, different cancer types, namely the most frequently encountered lung cancers, small-cell lung carcinoma (SCLC), squamous cell carcinomas (SqCCs), and adenocarcinomas (ADCs). TMAs were first used in SHP by the group of Levin[9] at the NIH. Here, we present results of a follow-up study that included 80 samples that were judged cancer free, 308 patient biopsies from cancer patients, and 61 samples of benign lung tumors. The classification goals were extended to include the classes enumerated above, in addition to necrotic tissues from these three cancer classes, as well as ADC subtypes and grades of SqCC, and four classes of nonmalignant lung lesions.

Since the publication of the pilot study in 2012, the experimental methods have been refined substantially and the methods of analysis have been augmented to include

[1]Department of Pathology, University of Massachusetts Medical School, Worcester, MA, USA; [2]Department of Mathematics and Statistics and Program in Bioinformatics, Boston University, Boston, MA, USA; [3]Cireca Theranostics, LLC, Cambridge, MA, USA; [4]Creative Creek, LLC, Camano Island, WA, USA; [5]Division of Thoracic Surgery, City of Hope Medical Center, Duarte, CA, USA and [6]Laboratory for Spectral Diagnosis, Department of Chemistry and Chemical Biology, Northeastern University, Boston, MA, USA
Correspondence: Professor M Diem, PhD, Department of Chemistry and Chemical Biology, Northeastern University, 360 Huntington Avenue, 316 Hurtig Hall, Boston, MA 02115, USA.
E-mail: m.diem@neu.edu

more sophisticated statistical tools. Some of the (spectroscopic) experimental procedures and the statistical methods and results have been summarized in two recent papers.[10,11] Furthermore, a detailed analysis of the biostatistical significance of the results was carried out, and their confidence intervals have been established.[12]

Next, a short introduction to methods referred to in the past as 'optical diagnosis' will be presented. Optical and spectrometric methods are commonplace in histology and pathology. After all, staining tissues by hematoxylin/eosin (H&E), followed by (visual) microscopic examination is an optical method: different compartments of the cell respond differently to basophilic and eosinophilic stains and thus allow a 'spectral analysis' using the eye as a detector. Whole slide imaging, introduced in other contributions in this issue of *Laboratory Investigations*, is another optical method, but is not based on 'spectral analysis' *per se* but on morphometric criteria. Immunohistochemistry, to date the most advanced optical method to detect the presence of certain cancer signatures or markers,[7,13] uses optical detection of specific antibodies labeled with easily observable stains. However, neither of these techniques use an inherent, spectral property or 'spectral fingerprint signature' of the tissue.

Imaging mass spectrometry, reviewed in this issue of *Laboratory Investigations* as well, is more akin to SHP in that a 'vector' of information is collected from each pixel of tissue. In mass spectrometric methods, this vector contains intensities (abundances) of molecules as a function of mass to charge (m/e) ratios, whereas in SHP a 'spectral vector' fingerprint of infrared absorption intensities *vs* wavelength of absorbed radiation of all biochemical components in the pixel is observed. Mass spectrometric measurements offer more specific information on protein up- and downregulation, whereas the infrared spectrum offers a survey of all changes (including metabolomics and genomic variations) in biochemical composition between individual pixels from which spectra are collected and which are smaller than the size of an individual cell. Furthermore, SHP can be performed on standard, untreated histological sections, either from formalin-fixed, paraffin-embedded (FFPF) tissue blocks or from flash-frozen tissue sections.

SHP 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.[14] The paradigm for SHP is that the transition from normal tissue 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. These changes in biochemical composition are encoded and observed via changes in the infrared spectra. These changes are manifested mostly in the complicated envelope of the amide I vibration and, therefore, reflect changes in the proteome of cancerous cells. In addition, changes in metabolic activity of cancerous cells are readily observed in the infrared spectra and manifest themselves mostly by a marked decrease in intercellular glycogen. Furthermore, increased RNA abundance in the cellular cytoplasm is observed for actively dividing cell.[15] The combination of all spectral changes because of variations in the genome, proteome, and metabolome is analyzed by self-learning algorithms (MLAs) similar to those used in other approaches to automatic analysis of large data sets.

We believe that SHP can aid in the accurate classification of cancers that are difficult to distinguish on a morphological basis alone, and whose accurate diagnoses determine therapeutic options. This paper follows similar studies reported in this journal in which equivalent methods for the analysis of exfoliated cells were reported, using a methodology referred to as SCP. SCP has been proven to be capable of reliably distinguishing dysplastic from normal cells and can detect abnormality even in cells that still present normal morphology.[16–18] This could have clinical utility in identifying precancerous states within current models of cancer evolution[19] and may also differentiate between synchronous and metachronous cancer development, specifically in lung cancer.[20,21]

MECHANISM OF ACTION OF SHP

The mechanism of action of SHP can be understood from the biophysics and spectroscopic properties of the biochemical components of cells. All molecules—whether small inorganic or very large biochemical ones—respond to infrared radiation in a predictable and thoroughly understood manner: infrared radiation is absorbed by molecular vibrations at specific infrared 'colors' (wavelengths) to produce relatively complicated 'infrared spectra' that are specific 'fingerprints' of the molecules. Infrared spectra collected form an individual cell, or a tissue pixel, thus are a superposition of all molecules' individual fingerprints spectra.

Infrared spectral features of biomolecules are affected by the conformation (shape), hydration, oxidation state, and many other physical effects. For proteins, for example, it is well known that the same protein can exhibit different infrared spectral patterns when the protein is found in different secondary or tertiary structures, that is, when it is denatured, precipitated, or hydrated/dehydrated. In apoptosis and necrosis, proteins unfold from their native conformation; the partially unfolded proteins tend to form aggregates of mostly β-sheet structures and tend to precipitate, or become insoluble. The change from native protein structure to the unfolded aggregates causes a large change in the infrared spectral patterns. Consequently, SHP is a highly sensitive tool to detect necrosis.

Other spectral markers of disease that can be detected by SHP may have to do with changes in the metabolome, or changes in lipids or mucous, and so forth. Such changes are observed in several regions of the infrared spectra: for example, carbohydrate metabolites and glycoproteins such as mucin have vibrations well separated from those of the protein backbone. Of all these signatures, the changes in the protein spectral region are the most significant for SHP.

This is, in part, because of the fact that proteins are by far the most abundant cellular components and comprise ∼65% of the dry weight of cells and tissues. Furthermore, changes in protein composition tend to alter the spectral pattern observed in the protein-specific bands in the infrared spectra.

In particular, the 'amide I' vibration of proteins, observed at ∼1650 cm^{-1} (6.06 μm) in the infrared spectrum, is the most sensitive indicator of protein structure. In the spectral plots shown in Figure 1, the amount of light absorbed (the 'absorbance') is plotted on the ordinate axis against the inverse of the wavelength (or 'wavenumber'), measured in reciprocal units of length, such as cm^{-1}. The spectral region, marked in Figure 1 as 'protein amide I', exhibits similar spectral features shown for the three tissue types shown in Figure 1, but quite different features for the four major protein conformations, α-helical, β-sheet, turns, disordered, and other helical structures. This conformational sensitivity is due to long-range dipolar coupling of individual peptide linkages (mostly the C=O bond stretching motion of the peptide linkage) that produces highly delocalized vibrational states (known as 'exciton' states) that are sensitive to changes in the geometry of proteins.

Figure 1 depicts typical infrared spectra of several different tissue types. The top trace is from the superficial layer of squamous tissue that is known to accumulate glycogen that

Figure 1 Examples of mid-infrared spectra of different tissue classes. Top: superficial squamous tissue; middle: fibroconnective tissue; bottom: B lymphocytes. The three spectra are offset along the absorbance (Y) axis for clarity. From Bird et al.[8]

exhibits three sharp absorption peaks between 1000 and 1200 cm^{-1} because of C–O stretching and C–O–H deformation motions. These peaks are superimposed on the protein spectral signatures, consisting of the 'amide I' and 'amide II' regions, and a few peaks due to protein side groups.

The middle trace of Figure 1 depicts an infrared spectrum of connective tissue that is dominated by the spectral features of collagen,[22] a triple-helical protein of repeating Pro-Pro-Xxx sequences. Collagen exhibits a very characteristic infrared absorption pattern in the 1000–1250 cm^{-1} spectral region that can further be enhanced by converting the spectra to second derivatives (see below). Finally, the bottom trace shows the infrared signature of metabolically highly active cells such as B lymphocytes that exhibit distinct nucleic features in addition to the protein peaks observed in the other traces.[23] In general, the spectral differences between tissue types are much smaller than the extreme cases shown in Figure 1, and require mathematical procedures for detection and interpretation. With the exception of necrosis mentioned above, it is unlikely that large conformational changes occur within a cell when it transitions from normal to cancerous states. However, the abundance of proteins with different structural motifs certainly will change, as proteomic studies have revealed, and this change in overall protein composition is sampled by infrared spectroscopy.

In order to enhance the sensitivity of the spectral measurements, the observed spectra are converted to their second derivatives; this process collapses the width of the peaks and enhances the appearance of shoulders and deflections (see 'Spectral preprocessing and segmentation' section). In this manner, minute changes in protein composition can be detected. These compositional changes, in general, cannot be linked to the up- and down-regulation of a single protein by infrared spectroscopic method, but they rather manifest themselves by changes in the spectral envelopes that contain a snapshot of the total cellular composition, and changes therein.

Thus, SHP provides an integrated view of the averaged compositional change in a cell or tissue spot, in separate spectral 'bands' for proteomic, metabolomics, and genomic changes. The spectral changes between normal tissue types and between normal and cancerous tissues need to be decoded in order to utilize spectral information as class indicators. In principle, this can be accomplished by several approaches. One could define, for example, cancer reference spectra by selecting regions with unambiguously defined cancer types, and classifying any unknown sample spectrum by comparing it successively against all cancer reference spectra. The reference spectrum with the agreement would determine the classification result. Another approach used in earlier reports decomposed the observed spectra into combinations of a few 'basis spectra', such as several proteins, nucleic acids, sugars, etc. This method has the disadvantage that the omission of a reference compound spectrum in the spectral decomposition can produce unpredictable and erroneous results. In another approach, favored in the research
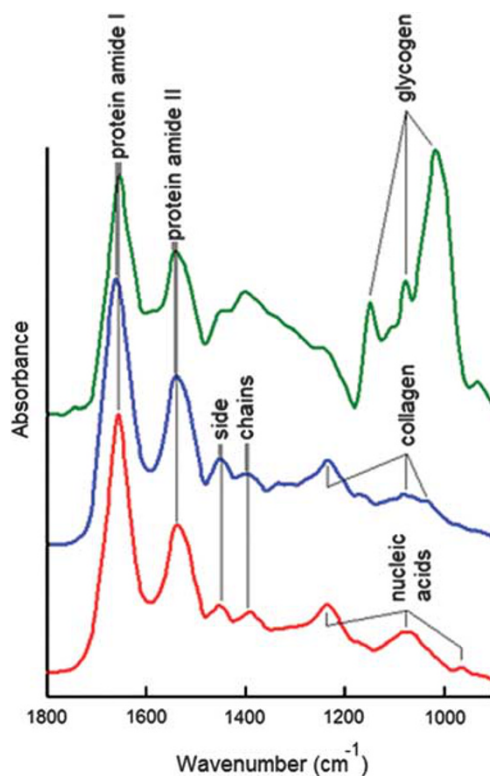
reported here, self-learning mathematical algorithms are trained to find recurring spectral changes and associate them with disease. This approach will be described in the section below. This method has the advantage that these algorithms scan the training spectra for significant spectral differences that are correlated with the desired outcome. Spectral regions of low correlation with the outcome are ignored. By carefully training these algorithms, and applying established rules of machine learning and bioinformatics, highly reliable and reproducible algorithms can be established.

EXPERIMENTAL ASPECTS (MATERIALS AND METHODS)
Sample Selection
The samples for this study were derived from commercial TMAs especially prepared for this work. In addition, samples of benign lung tumors were derived from the archives of the Department of Pathology, University of Massachusetts Medical School, under a local IRB. All samples were from FFPE tissue blocks. The TMAs were assembled to accommodate the goals of this study that include:

(1) Distinction of normal (NOR) from cancerous and necrotic lung tissue,
(2) Classification of lung cancers into SCLC and non-small-cell lung cancer (NSCLC),
(3) Further classification of NSCLC into ADC and SqCC,
(4) Classification of ADC into several subclasses of clinical relevance, and
(5) Distinction of benign lung lesions from normal tissue and from cancerous lesions.

For the first four of these goals, six TMAs were assembled that contained the samples summarized in Table 1.

From each TMA, three tissue sections were purchased, referred to as section A001, A002, and A003. Section A002 was mounted at Biomax on a standard microscope slide, de-paraffinized, stained, and coverslipped. The other two

sections were mounted on 'low-e' slides (see the section 'Infrared and visual image acquisition') and delivered as paraffin-embedded samples. Each tissue spot measured ∼1.8 mm in diameter, and will be referred to as 'patients' later in this report.

The benign lung lesions were not in a TMA format, but were standard excised tissue specimens presented as FFPE tissue sections and often measuring in excess of 1 cm². The diagnosis of these samples was based on pathology reports, and confirmed by the study pathologist. With these benign samples, the entire patient number was 449.

Infrared and Visual Image Acquisition
The methods of SHP, including data acquisition and data preprocessing, have been described in detail in the literature.[24,25] All spectroscopic studies reported here were carried out on 'low emissivity' (low-e) slides (Kevley Technologies, Chesterfield, OH, USA) that are reflective toward infrared radiation, but are nearly totally transparent to visible light. The use of these sample substrates has been discouraged by some authors,[26] citing the distortion of spectral intensities by the standing electromagnetic wave that forms when radiation is reflected from a metallic surface. Subsequently, the effect of the standing wave on the observed spectra was analyzed by Wrobel *et al.*[27] and found to be much smaller than originally reported when microscope objectives with large numeric apertures are used; in addition, using the second derivative, rather than the absorption intensities, can further reduce the intensity distortions.[10]

As the same tissue section was used for both infrared and white light imaging (after appropriate staining), the visible
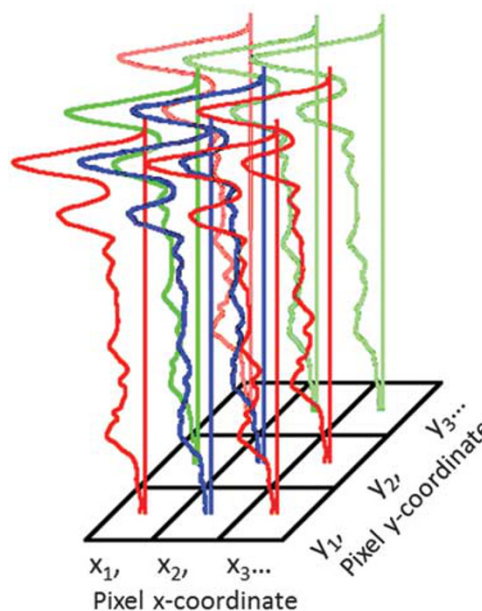
Table 1 Composition of the data set used in this study

Biomax	Number description	No. of patients	Comments
LC 701	Normal	80	Cancer-adjacent normal, biopsy free of cancerous cells
LC702	Necrotic	29	From SCLC, SqCC, and ADC
LC703	SCLC	61	Some patients from LC811
LC704	SqCC	89	Some patients from LC811
LC705/6	ADC	129	Mucinous and nonmucinous
Organizing pneumonia		18	Benign tumor
Sarcoidosis (granuloma)		18	Benign tumor
Hamartoma		18	Benign tumor
Meningothelial nodules		7	Benign tumor

Figure 2 Schematic of the hyperspectral data set collected for each tissue sample (from Diem *et al.*[11]).

and infrared images could be accurately registered. This is necessary for annotation (see the section 'Annotation and data traceability') of spectral features. Slides for spectral data acquisition were deparaffinized using standard procedure[28] and kept in a desiccator when not used.

Infrared spectral images were acquired as 'hyperspectral data sets' or 'spectral hypercubes' as shown in Figure 2.[24] Conceptually, each tissue sample is divided into thousands of individual pixels, measuring $6.25\,\mu m$ on edge. From each pixel, an entire infrared spectrum is collected in a wavelength range between $2.5\,\mu m$ ($4000\,cm^{-1}$) and $14.28\,\mu m$ ($700\,cm^{-1}$). Thus, the raw spectral hypercube for a $1.8\,mm$ diameter tissue spot consists of nearly $100\,000$ pixel spectra, each containing 1650 intensity data points at constant abscissa spacing of $2\,cm^{-1}$. Of these spectral vectors, the 'fingerprint' region between 800 and $1800\,cm^{-1}$ was used for tissue classification. The region below $800\,cm^{-1}$ contains only weak transitions in the infrared spectra and is difficult to access because of the detector wavelength cutoff. The region between 1800 and $2800\,cm^{-1}$ is devoid of any useful transitions. The C–H and N–H stretching region between 2800 and $3500\,cm^{-1}$ can be used for analysis as well but provides less information on the proteome of cells and tissue than the amid I manifold in the fingerprint region.

Infrared spectral hypercubes for each tissue spot were collected using a PerkinElmer (Shelton, CT, USA) model SpectrumOne/Spotlight 400 imaging infrared micro-spectrometer. This instrument incorporates a 16-element cryogenically cooled infrared HgCdTe detector array; thus, spectra from 16 pixels were collected simultaneously. Data acquisition and storage required $\sim 1\,h$ 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.

After infrared data acquisition, the tissue sections were stained at the Department of Pathology at the University of Massachusetts Medical School, using H&E and following standardized and validated methods. After coverslipping, the tissue sections were imaged using an Olympus (Center Valley, PA, USA) BX51 microscope equipped with a computer-controlled microscope stage with linear stepping motors ($0.1\,\mu m$ resolution). Images were taken via a Qimaging (Surrey, BC, Canada) model QICAM high resolution digital camera. The microscope was operated using Media Cybernetics (Rockville, MD, USA) Image Pro Plus software. The tissue spots were imaged at $\times 20$ magnification, producing large mosaic visual image data files at sufficiently high spatial resolution for pathological interpretation. Registration of the slide position for visual and infrared microscopy was aided by mounting the slides in a specially designed and manufactured slide holder that was equipped with three reticles whose positions in the particular microscope table were read and recorded at $0.1\,\mu m$ accuracy. Summaries of data acquisition procedures and protocols have recently been published.[10,29]

Spectral Preprocessing and Segmentation

Each tissue spot produced $\sim 10^5$ individual pixel spectra that were preprocessed as follows. First, the size of hyperspectral data cubes was reduced by a factor of four by co-adding four individual pixel spectra into a new spectrum with better signal-to-noise ratio, but larger pixel size, $12.5\,\mu m$ on edge. This step was deemed necessary when these studies were carried to reduce computation time of the hierarchical cluster analysis (see below). Furthermore, the spatial resolution of the microspectrometers was found to be $\sim 12\,\mu m$ at $1000\,cm^{-1}$; thus, the averaging process did not result in a loss of spatial resolution.

The resulting set of $\sim 25\,000$ pixels per tissue spot was corrected for confounding contributions such as noise, water vapor, and resonance Mie (R-Mie) scattering (via a phase correction algorithm[30]) using procedures developed and reported previously in the literature.[10] In order to enhance the sensitivity of spectral methods toward specific changes of protein abundance, the broad and often unstructured raw spectra were converted to second derivatives. This process is known to reduce the half width of spectral bands, thereby providing better discriminatory power that provides for the ability to classify different tumor types. Second derivative spectra are also less susceptible to the standing wave artifact as they depend on the curvature of a peak, rather than its intensity. The second derivative spectra are the primary information obtained in an SHP experiment, and the task at hand is the decoding and correlation of the spectral information with the pathological diagnosis. Details of these preprocessing procedures have been reported previously.[10]

The preprocessed hyperspectral data sets for each of the tissue spots were subsequently converted to pseudocolor images by hierarchical cluster analysis (HCA). This is a well-known method to extract recurring patterns in data sets;[31] in this particular application, HCA was used to segment the data set into groups of high spectral similarity and homogeneity, and to present these groups as pseudocolor displays. Typical HCA-based pseudocolor images of tissue spots are shown in Figure 3 (middle column), whereas the left column shows the corresponding visual image of the stained tissue spot. In the HCA images in the middle column, regions of the same color represent similar spectra. Visual inspection of Figure 3 immediately reveals a spatial correspondence between the IR pseudocolor image and the H&E-stained image. This correspondence becomes even more obvious at higher magnification of the visual image: Figure 4 shows a magnified view of the ADC sample shown in the top column of Figure 3. Here, the red regions in the infrared pseudocolor map represent tissue areas richer in connective tissue that appear pinkish in the H&E-stained visual image. These red areas were excluded from the ADC tissue class in the annotation step, see below.

It should be noticed that the presegmentation step described above is completely unsupervised in the sense that it does not requires any input from a pathologist. The HCA
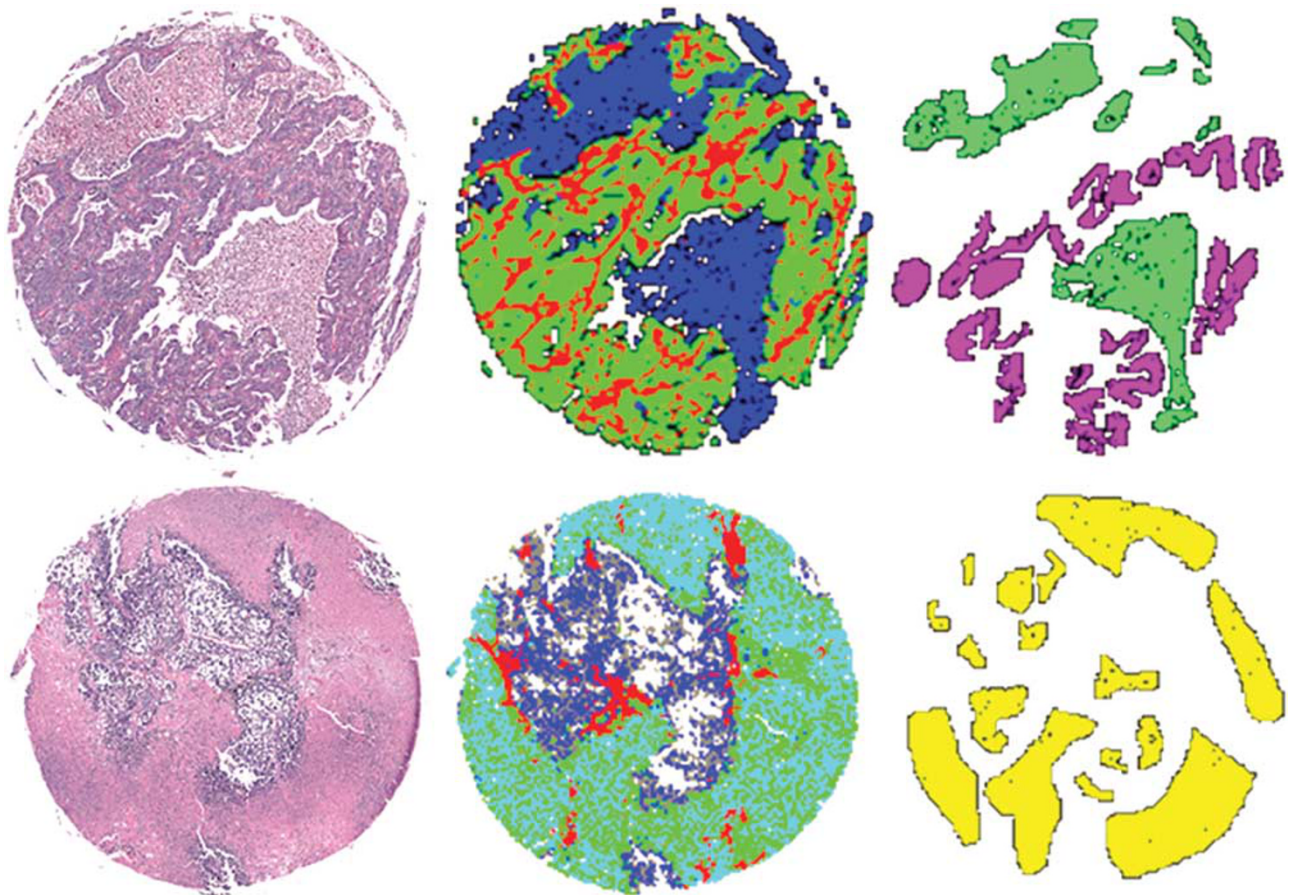
Figure 3 Images of H&E-stained tissue spots (left column), 3-cluster infrared pseudocolor images (middle column), and annotated regions (right column, see text) of tissue spots with ADC and necrosis (top row) and SCLC and necrosis (bottom row).
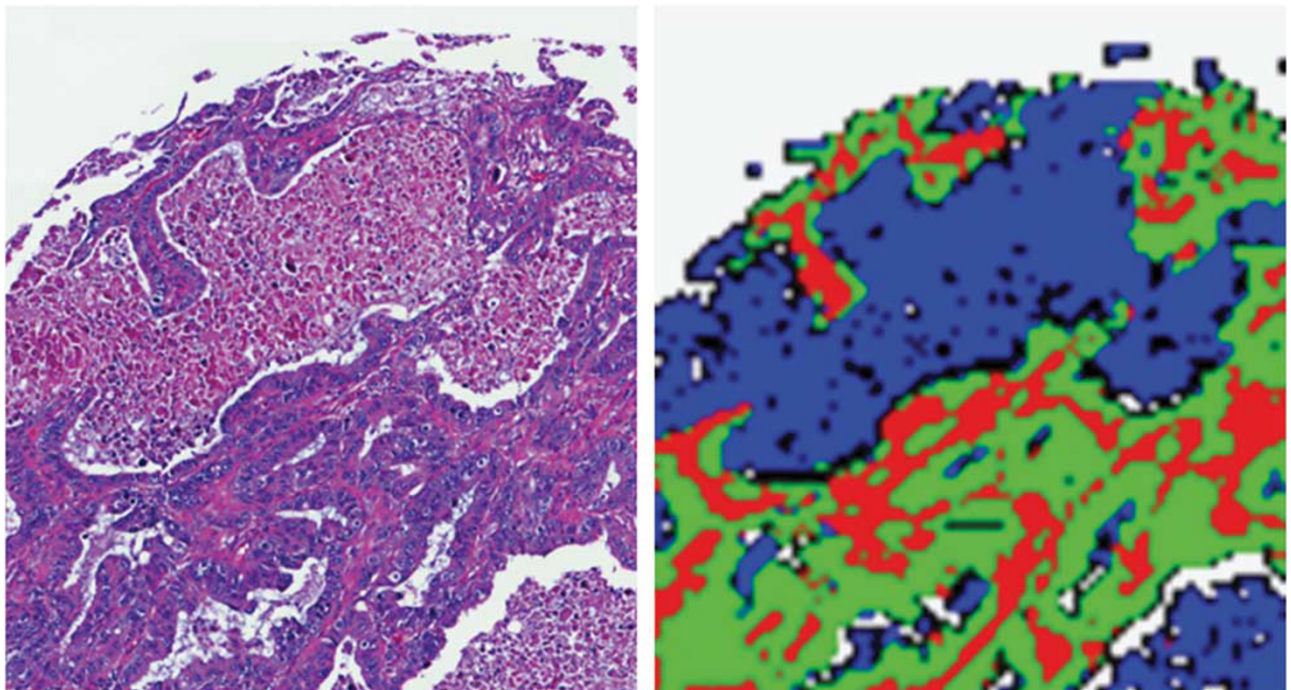

Figure 4 Higher detail images of ADC sample shown in Figure 3. (Left) Microscopic image of H&E-stained tissue; (right) infrared image from HCA.

Table 2 Number of pixel spectra, processed spectra, and annotated spectra in entire TMA data set

Total spectra:	$\sim 39 \times 10^6$ (388 spots, $\sim 100\,000$ spectra per spot)
Processed spectra:	4×10^6 pixel spectra (2×2 pixel averaging and elimination of blank pixels)
Annotated spectra:	5.5×10^5 spectra
Annotation regions:	9.3×10^3 annotation regions
Annotation regions/spot:	24 Regions/spot (average)
Pixels /annotation region:	60 Pixels (average)
Main tissue types:	168 (54 malignant, 114 normal classes)

images represent the inherent ability of infrared spectral imaging to detect differences in tissue composition by the spectral signatures. At this point, no diagnostic information is available from the infrared images; in order to achieve diagnostic and/or prognostic capabilities, regions of clearly defined normal tissue types and disease states need to be identified from the images shown in Figure 3. Subsequently, spectra associated with these regions are extracted and entered into a database from which algorithms are trained to associate spectral features with pathological diagnosis. This step is referred to as 'annotation', described in the section 'Annotation and data traceability' below. Subsequently, unknown data sets can be analyzed for the occurrence of the disease-specific spectral signatures.

Annotation and Data Traceability

The annotation process correlates unambiguously assignable tissue areas from the H&E-stained visual images with corresponding regions of the pseudocolor infrared image, and permits extraction of the spectra from the selected regions in the infrared images (Table 2). To this end, a semitransparent overlay of the visible and infrared images was created automatically, using image registration methods, such that the tissue features still can be perceived, but are displayed on a color background that corresponds to the HCA clusters. This can be performed equally well for the tissue microarray spots or the large biopsy tissue sections. Within an HCA cluster, a pathologist selected areas that represented, in his opinion, the most typical histological regions of a diseased or normal tissue type. Each area was tagged with a code that identified disease state or tissue types. This is shown in the right column of Figure 3. Here, the regions that represent ADC and were selected by the pathologist are shown in purple, whereas the selected necrotic tissue regions are shown in green. In the SCLC sample, the annotated SCLC regions are shown in yellow. The annotation software ascertains that each contiguous area selected by the pathologist corresponded to one HCA cluster only, and eliminates pixels that did not conform

to the majority assignment within one selected area. The number of selected areas in Figure 3 is typical for an annotated tissue spot, and yielded, on average, ~ 1400 pixel spectra for each tissue spot, corresponding to ~ 350 cells. This later assessment was based on an estimate of a cell's size ($\sim 25\,\mu m$ in diameter) and the aggregated pixel size (12.5 μm on edge).

In the data set of $\sim 550\,000$ annotated pixels (see Table 2), each pixel spectrum is uniquely defined and traceable to the tissue microarray name (eg, LC706), the particular section, the individual tissue spot identified by row and column (eg, C3), and the coordinate of the pixel spectrum. This coordinate is uniquely defined by the pixel X, Y address, and the pixel size. The pixel X, Y address was referenced against the reticle positions in the slide holder. Each annotated pixel spectrum, in addition, was tagged with a code that identified the pathology diagnosis. Thus, any pixel spectrum can be relocated and traced, and may be compared with the corresponding region of the visual image that was used for annotation.

The predominant diagnosis from the annotating pathologist, in general, agreed with the diagnosis obtained from Biomax. The annotation yielded pixel spectra in over 160 classes that were subsequently combined into 26 major tissue types/disease classes by similarity analysis, using graph partitioning methodology. These graphs clearly demonstrate which of the classes were so closely related that they could be combined into larger groups by minimizing the connection between classes. These groups were referred to as group A (10 normal tissue types); group B (necrotic ADC, necrotic SqCC, and keratin pearl); group C (SCLC and necrotic SCLC); group D SqCC (3 grades) and group E ADC (8 subtypes).

Computational Aspects

All computations were carried out on a Dell workstation equipped with a 12-core Intel processor and 56-GByte memory, running 64-bit Windows operating system. This workstation was connected to a server with 20 TByte hard drive space and cloud backup. All computations were carried out in MATLAB version R2013.b (The Mathworks, Natick, MA, USA) using scripts developed in-house. The scripts and input data sets are archived for each of the 'studies' carried out, where each study was defined by a distinct aim, result and contributing data subsets.

The statistical and computational procedures developed and utilized were described in a previous paper.[12] In the work described therein, efforts were reported that detailed the composition, in terms of patients and pixel spectra, of the training and validation data sets, the metrics used for the evaluation of the quality of the diagnostic results, the type of MLAs, the number of pixels per patient to be included in the training phase, the number of spectral features (data points per spectrum) included in the analysis, and a measure of the power and confidence interval of the classification. This

paper also reported the reasons why support vector machines (SVMs)[31,32] were eventually chosen for MLAs. The SVM-based classification was carried out in a 'pixel-based', a 'patient-based', or 'image-based' manner, to be discussed later. Furthermore, a 'multi-classifier' vs a step-wise hierarchical decision tree approach was investigated, and the choice was made in favor of the multiclassifier algorithm that separated the samples into normal, necrotic, SCLC, SqCC, and ADC classes in one step.

The most salient features of final algorithm and procedure utilized are summarized below:

- SVMs were used for the classification task because of their high reproducibility and more easily understood mode of action.
- The entire spectral vector of 501 2nd derivative data points (spanning the range from 800 to 1800 cm^{-1}) was used for classification. The range below 800 cm^{-1} did not contribute to the overall accuracy and was ignored. Feature selection (ie, eliminating certain spectral features in the 800 to 1800 cm^{-1} range) reduced classification accuracy and was not implemented.
- A total of 2000 pixel spectra per disease or tissue type class were selected randomly for algorithm training. The classification accuracy did not improve by using more pixel spectra, but the computation time increased significantly. If there were <2000 spectra in one of the classes, 'oversampling' was applied: in oversampling, spectra were repeatedly used in the training set, rather than reducing the number of spectra in the larger classes ('undersampling').[33]
- The final SVM utilized a radial kernel or basis function (rbf). Two parameters, 'c' (penalty weight on misclassification error) and 'γ' (width of the radial basis kernel) were optimized by varying them independently from 0.000061 (2^{-14}) to 0.031 (2^{-5}) for γ and 0.0625 (2^{-4}) to 32 (2^5) for c.[32] This resulted in an optimized SVM that yielded an accuracy of 92.4 ± 0.85% for a benchmark data set consisting of ~190 500 training spectra and 48 600 test spectra in the five major classes listed above. As indicated in the bullet point above, 10 000 spectra were randomly selected from these data sets; repeated training/test processes yielded results that were found to lie within the expected confidence interval limits (see below).

The 95% confidence intervals (CIs) were computed as follows. Ten independent SVM training and test runs were carried out by randomly selecting 10 000 training spectra from the entire training data set. These 10 000 training spectra were selected from the five classes of interest (NOR, NECR, SCLC, SqCC, and ADC) with 2000 spectra per class. Increasing the number of spectra per class did not improve the classification accuracy.[12] The number of patients contributing to the 2000 spectra was varied from 30 to 135 but the size of the training set was held constant at 10 000. The results of this simulation are shown in Figure 5. The overall accuracy increases, as expected, as the number of patients in
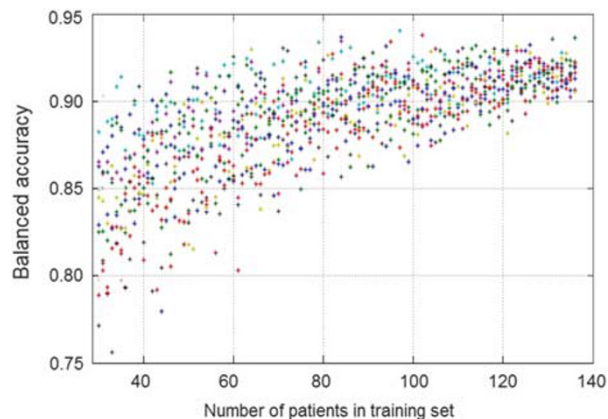

Figure 5 Simulation of accuracy and confidence interval as a function of patient number in the training set. Each symbol in the graph represents one training/test result for 10 000 pixel spectra randomly selected from the number of patients indicated.

the training set increases, from ~85 to over 90%, and the scatter in the accuracy for 10 independent runs decreased by a factor of ~5. The CIs were also obtained by analytical methods and agree very well with the simulations.[12]

The results of this simulation also suggest that the annotation method described earlier that often yields hundreds or thousands of individual pixel spectra for each annotated spot produces a representative sampling of tissue homogeneity and patient-to-patient variance. This is in contrast to other cancer diagnostic methods that yield one data point per patient, whereas in SHP thousands of data points are created for each patient.

Finally, the data were analyzed using both standard SVMs and probability-based Platt-SVMs.[34] In the latter, the classifier reports probability limits (eg, 0.1 and 0.9) that a pixel belongs with 90% probability to class A, whereas a pixel with <10% probability most likely belongs to class B (with 90% probability). Pixels between 0.1 and 0.9 probability are considered 'unclassifiable'. This approach reduces the total number of classified pixel spectra, but increases the probability that those remaining were classified correctly.

RESULTS AND DISCUSSION
Overall Goals
The overall goals of the study presented here were the reliable distinction between normal (NOR) and diseased tissue, the distinction between necrotic (NECR) and cancerous tissue and between SCLC and NSCLC. Furthermore, the latter category was to be distinguished into SqCC and ADC that themselves had several grades and/or subclasses.

In a previous lung cancer pilot study, we used hierarchical binary classifiers based on artificial neural networks to consecutively classify and remove the most different spectral classes in the data set and arrived at good classification accuracies.[8] The much larger data set reported here, analyzed
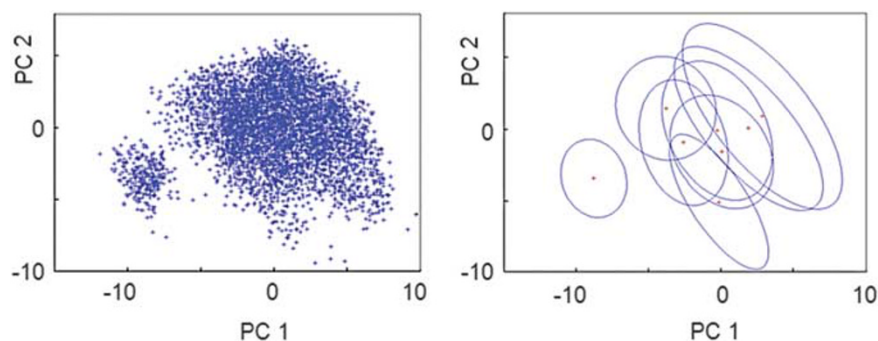
Figure 6 (Left) PCA scores plot of all spectral vectors in one tissue class. Each symbol represents on pixel spectrum. (Right) Same plot, represented by 95% membership contour and ellipsoid center for each patient.

Table 3 Summary of training and testing data sets

	Number of patients	Number of annotated pixel spectra	Number of annotated pixels in 26 major classes
Training	173	219 422	216 767
Test	196	256 729	246 725

by SVM, yielded accuracies somewhat lower than those reported earlier, mostly because the larger data set increased the heterogeneity of the data substantially, and because a new class—necrosis—was added to the classification. Necrosis does present a problem, in particular in necrotic SCLC samples, as the occurrence of the two disease stages appears to be coupled, and the samples often contained mixtures of the two classes.

The heterogeneity within the spectral classes can be appreciated by inspection of Figure 6 that represents a principal component analysis (PCA) 'scores plot' for all spectral vectors in a given spectral class (left panel), and the same information represented by the 95% confidence ellipsoids for each patient, and the center of the ellipsoid. Clearly, for this class, one patient exhibits quite different spectral features as compared with the other seven patients. It is not clear, at this point, whether the one outlying patient represents a slightly different disease profile that is detected by SHP or whether the annotations placed two different disease stages into the same class.

Aside from eliminating pixel spectra with poor signal-to-noise ratio (shown as the black regions in the HCA plot of Figure 4), all spectra were included in training and validation subsets, with the following caveat: we required that at least three patients and 400 pixel spectra were represented in a tissue class before it was allowed to be analyzed.

In order to report sensitivity and specificity, the SHP results needed to be correlated against a gold standard, in this case, classical histopathology. Thus, only annotated pixel spectra could be used for the evaluation of the accuracy of SHP. This reduced the number of pixel spectra that could be included in this study to ∼550 000, as shown in Table 2. Of these 550 000 annotated pixel spectra, 219 422 from 173 patients were in the training set and 256 729 from 196 patients were in the test set. Of these pixel spectra, 216 767 were in the 26 subtypes for which sufficient patients/pixels were represented in the classes (see the criteria above) and 246 725 in the test set (see Table 3). It should be noted that spectra in both the training and test set were from annotated regions; that is, data from the blinded test set had been processed and annotated as described before. However, these data were never used in the training process of the classification algorithm to avoid overfitting of the data. The blinded data sets needed to be processed and annotated as well in order to allow the accuracy of the classifier to be assessed.

Three different approaches were used to analyze this data set. One approach is referred to as a 'pixel-based' test in which the algorithm was trained, as indicated above, with 2000 pixel spectra from each of the 26 subgroups that fall into the five major classes of tissue types (groups A–E) listed at the end of the section 'Annotation and data traceability'. If the number of spectra in one of the 26 tissue classes was <2000, an oversampling strategy was applied as described in the discussion of the final algorithm above. This test samples the global agreement of all annotated areas from patients in the test set with the diagnosis rendered by the annotating physician.

The second approach of analysis is referred to as 'full spot test'. In this test, the agreement between the SHP diagnosis and the whole spot diagnosis from Biomax was determined. Here, the criteria were set such that the predominant SHP cancer prediction had to agree with the Biomax pathology diagnosis, and that at least 400 pixels (100 cells) had to be detected by SHP to conform to this diagnosis. Thus, this test ascertains that SHP does not miss a cancerous sample, and properly diagnoses the major cancer type.

The final approach was an image-based representation of the results from the 'full spot test', in which the agreement

between the annotated regions for each patient sample was compared with the SHP prediction via a 'label image', discussed in the section 'Label images'.

Pixel-Level Test

This test was carried out 10 times, with different random selection of 52 000 training spectra in the 26 subgroups. The balanced accuracy, established by using each training model to classify the spectra in the test set, varied by <1% for the consecutive runs, indicating that the training spectra sampled the variance in the data set adequately. The results of the 26 subgroups were subsequently combined into the major five classes shown in Table 4.

The average of the balanced accuracy classification is 87.2%. As pointed out before, the major source of disagreement is the classification of necrosis that has a low sensitivity, particularly in the case of samples with SCLC and necrosis, where SHP diagnosed many necrotic pixels as SCLC. This brings up the question of how sharp the distinction between necrosis on one hand and SCLC with necrosis on the other can be drawn in solid tumors for which the interior can exhibit serious hypoxia.

The average balanced accuracy could be improved substantially by implementing the Platt-SVM approach introduced earlier. By eliminating low probability pixels (by raising the Platt threshold to 0.75), the balanced accuracy could be improved from 87.2% to over 91%, while reducing the number of classified pixels from 256 729 to 202 579 (an 18% reduction of the pixel number). When the Platt

Table 4 Pixel-based sensitivity, specificity, and balanced accuracy for test data set

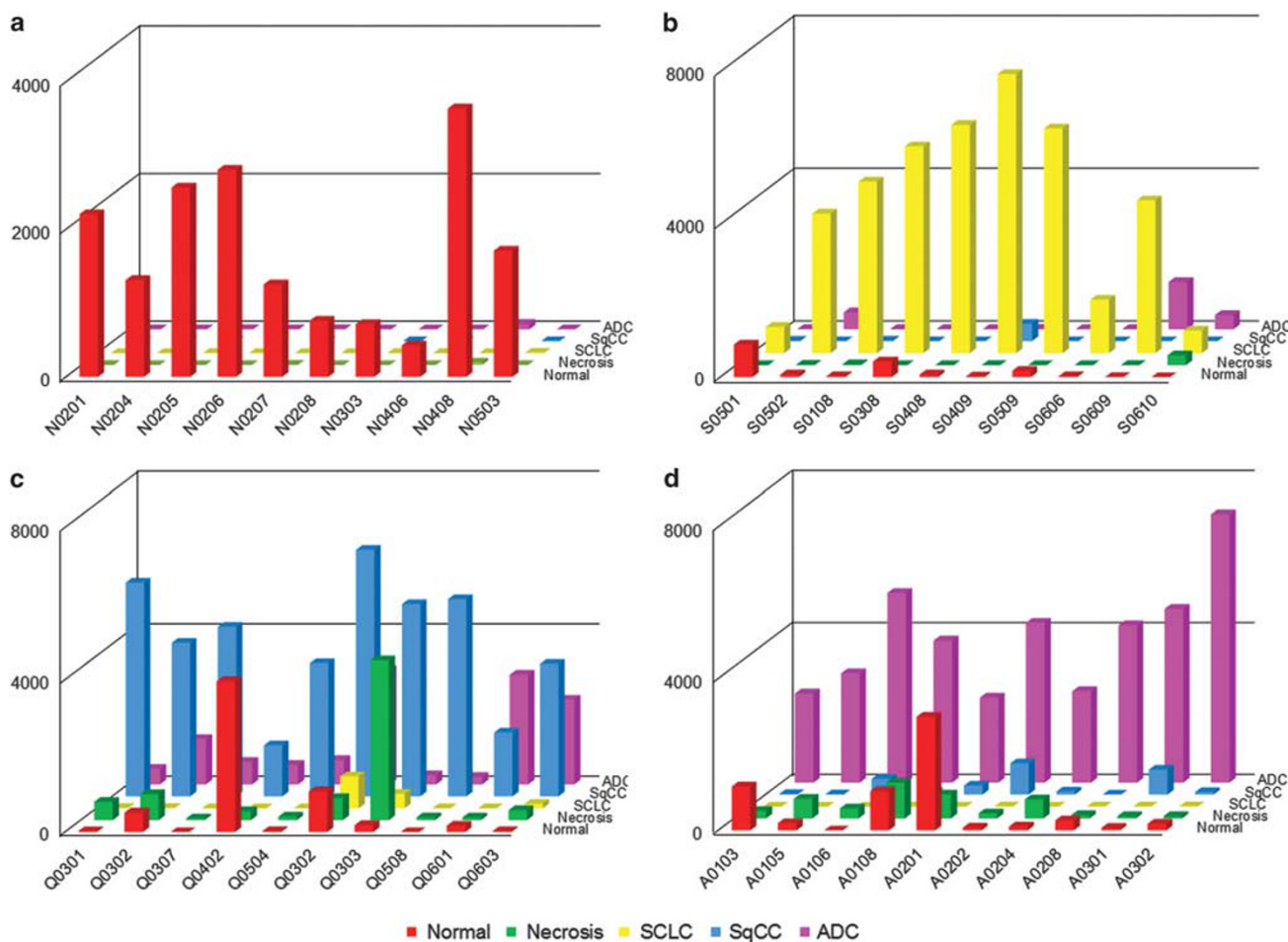| Class	Sensitivity	Specificity	Balanced accuracy
Normal	98.0%	97.2%	97.6%
Necrosis	78.8%	83.9%	81.4%
SCLC	95.9%	82.8%	89.4%
SqCC	83.2%	86.0%	84.6%
ADC	82.2%	84.0%	83.1%

Figure 7 Partial results of the whole spot analyses described in the section 'Full spot test results' (see text for detail). (**a**) Normal samples, (**b**) small cell carcinomas, (**c**) squamous cell carcinomas, and (**d**) adenocarcinomas.

threshold was raised to 0.9, the balanced accuracy improved to 93.2%, with a concomitant reduction of the diagnosed pixels to 191 608 (a 22.4% reduction of pixel number). The lower probability pixels are mainly because of two causes: low signal-to-noise data at the edges of tissue and low patient number in some of the tissue subclasses that reduced the statistical significance of the spectral analysis.

Full Spot Test Results

There was a total of 188 patient samples in the full spot test set after removing some duplicate tissue spots from the same patient. As pointed out before, this test required that the major diagnostic category from the Biomax diagnosis was properly reproduced, with at least 400 pixels agreeing with this diagnosis. Figure 7 summarizes results from some of these analyses: because of the large size of the overall data set, only 10 representative spots from each class are reported. Figure 7 displays the results as follows. Each panel consists of five rows: the front row, shown in red, depicts the correctly predicted normal tissue pixels, whereas the next row (green) shows correctly predicted necrotic pixels, followed by SCLC (yellow), SqCC (blue), and ADC (pink). This color scheme is maintained in all four panels of Figure 7. The columns

designate individual tissue spots, where the labels at the bottom designates the slide and spot number (N, normal; S, small cell carcinoma; Q, squamous cell carcinoma; and A, adenocarcinoma), and the four-digit number the tissue spot location (0102 corresponds to a spot in row A, column 2, and so forth).

Correct predication, in the context of Figure 7, implies that the major pathological diagnosis, as reported by Biomax, agrees with the SHP classification. Figure 7a indicates that normal tissue types were classified highly accurately in this whole spot analysis of the test data set. The few pixels mis-classified as ADC in spot N0408 are most likely because of low signal quality. Similarly, Figure 7b shows the results for 10 SCLC spots. Spots S0502 and S0609 show some pixels misclassified as ADC; however, the normal pixels (first row)

Table 5 False-positive (FP) and false-negative (FN) classifications

	Normal	Necrosis	SCLC	SqCC	ADC	Diagnosis	
N0105	3768	0	512	5	178	NOR	FP
A0401	773	113	5091	9339	42	ADC/NECR	FN
A0404	538	166	8695	2156	42	ADC/NECR	FN
A0204	2707	65	5523	1552	3646	ADC/NECR	FN
Q0307	0	48	374	73	471	SqCC	FN
Q0304	58	4622	1	374	428	SqCC/NECR	FN
S0106	123	2	8116	8	2	SCLC/NECR	FN
S0303	905	0	5264	0	1	SCLC/NECR	FN
S0503	1236	32	10 360	6	10	SCLC/NECR	FN

Table 6 Sensitivities and specificities of the full spot analyses including (A) and excluding (B), respectively, the samples for which SCLC was classified correctly, but necrosis was missed

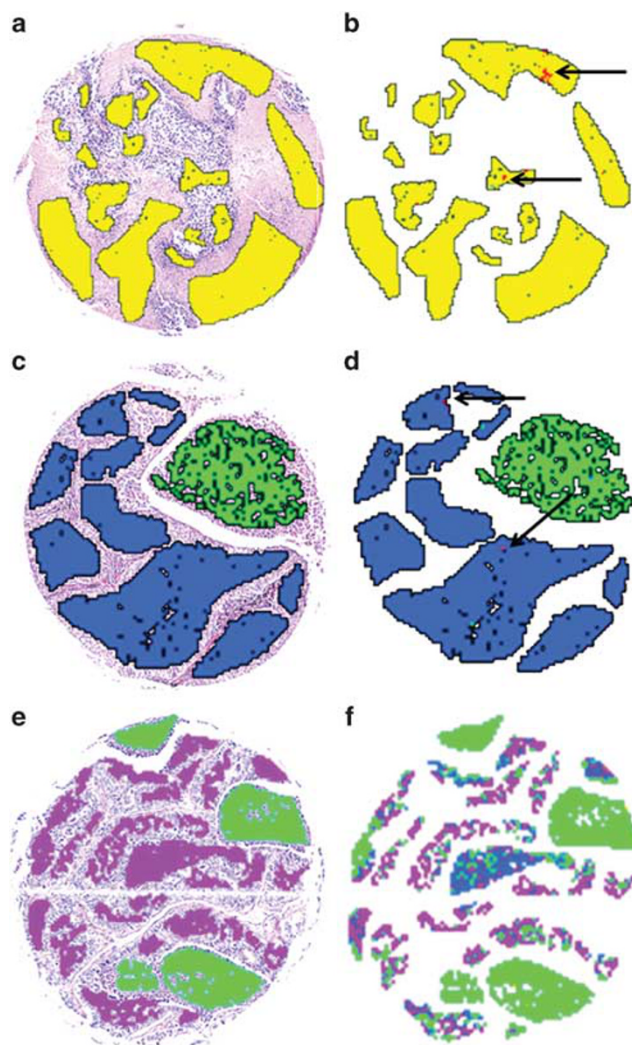| (A)				
Cancer	148	8	94.9%	Sensitivity
Normal	1	31	96.9%	Specificity
(B)				
Cancer	151	5	96.8%	Sensitivity
Normal	1	31	96.9%	Specificity

Figure 8 (**a**, **c**, **f**) Semitransparent photomicrographs of H&E-stained tissue spots with SCLC (**a**), SqCC and necrosis (**c**), and ADC and necrosis (**e**) with the annotated areas superimposed ('true' diagnoses). (**b**, **d**, **f**) SHP predictions. The black arrows in (**b**) and (**d**) mark misclassifications (red). In (**f**), an area of ADC was misclassified by SHP as SqCC.
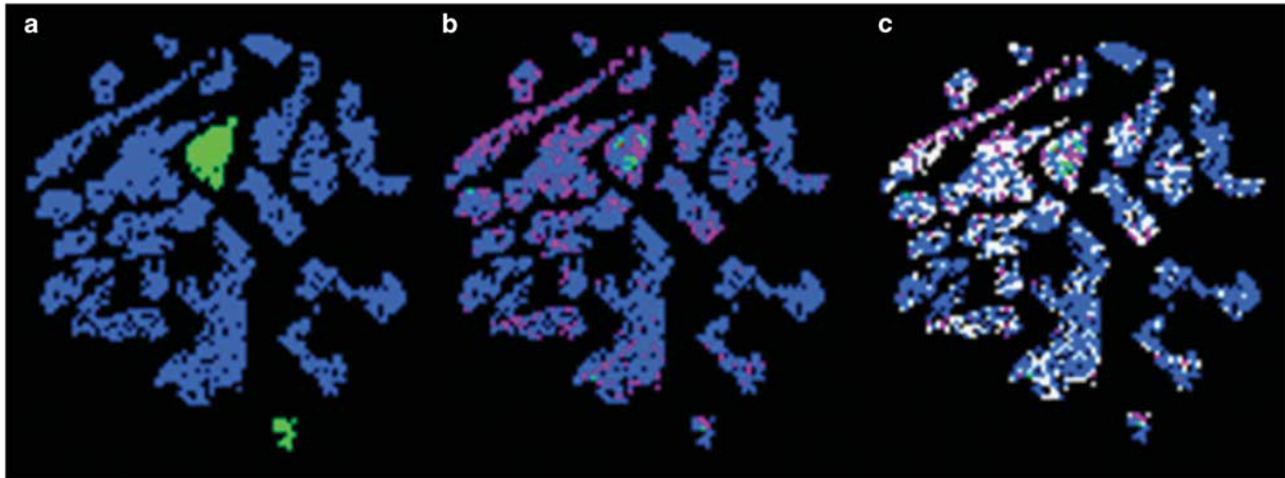
Figure 9 (**a**) Annotated region of SqCC (blue) and necrosis (green) of a tissue spot. (**b**) SHP prediction. Notice the misclassifications (ADC, purple) and the (near) absence of necrosis. (**c**) Platt-SVM result, indicting pixels of low certainty (probability) in white.

are not misclassifications but in fact normal tissue areas in the spot. In Figure 7c, the normal pixels again are because of normal tissue areas. These squamous cell cancer tissue spots do exhibit necrotic pixels (green), but a few show significant contributions from ADC. It is not clear, at this point, whether these pixels are misclassifications, or represent a mixed adenosquamous carcinoma. The identification of ADC (Figure 7d) is very good, with a few spots showing normal tissue areas.

Table 5 lists the misclassified (FP and FN) tissue spots. The total number of misclassifications, 9/188 or 4.8%, is quite small; however, of these, a number are borderline cases. The FP diagnosis (eg, N0105) properly identified the majority of pixels as normal, but the number of SCLC misclassifications (512) exceeded the threshold. Similarly, the three last entries in Table 5 identified the major cancer (SCLC) correctly, but failed to detect a sizeable fraction of necrotic pixels. This is in line with the previous discussion in the section 'Pixel-level test' that indicated that the SHP tends to underestimate necrosis in mixed NECR/SCLC samples. Furthermore, Table 5 indicates that necrosis presents some difficulties in the case of mixed NECR/ADC as well (spots A0401, A0404, and A0204).

Table 6 summarizes the full spot results in terms of sensitivity and specificity. Table 6A presents the sensitivity and specificity of the full spot test for all nine misclassifications listed in Table 5, whereas Table 6B lists sensitivity and specificity when samples LC702_A001_0106, LC702_A001_0303, and LC702_A001_0503 (see Table 5) were included as true positives as the cancer was identified correctly.

From spectroscopic and biochemical viewpoints, necrosis presents a major change in cellular structure and biochemical composition: the changes in protein conformation alone distorts the amide I spectral region (see the section 'Mechanism of action of SHP') so severely that smaller changes, such as the ones distinguishing ADC and SqCC, may get overwhelmed.

Label Images

The label images to be discussed next are graphic representations of the results introduced in the previous section. Examples of these representations are shown in Figure 8. Here, the images in the left column represent a highly transparent view of the H&E-stained tissue spots, with the annotated areas superimposed in the same color scheme introduced before (red, normal; green, necrosis, yellow, SCLC; blue, SqCC and purple, ADC). These annotations are considered the gold standard, or 'true' value. The images in the right column show the SHP predictions in the same color scheme. The first row (Figure 8a and b) show the true and predicted outcome for a tissue spot diagnosed with SCLC; obviously, the agreement is excellent, and only a small region (arrow) was misclassified. Similarly, a spot with necrosis and SqCC (Figure 8c and d) is correctly classified by SHP in its entirety. These two samples exemplify the overall quality of the SHP predictions, and the majority of the spots in the test set were predicted correctly by SHP.

Figure 8e and f shows a tissue spot with some misclassifications. Here, the regions diagnosed as necrosis were, in general, well predicted by SHP, but some of the ADC regions were predicted as mixtures of ADC and SqCC or ADC and necrosis. Again, the images shown here are typical for SHP misclassifications that occur mostly between cases of poorly differentiated and advanced ADC and SqCC, particularly when accompanied by necrosis. A few spots were misdiagnosed entirely between SqCC and ADC; however, no malignant sample was ever classified as normal.

Figure 9 demonstrates that areas of misclassification often exhibit low Platt probabilities. Figure 9a shows the 'true' annotation, whereas Figure 9b shows the SHP prediction. There are two regions where necrosis (green) was missed by SHP (although the central necrosis spot shows a few green pixels in the SHP image), and many SqCC pixels were misclassified as ADC. Figure 9c shows that many of the
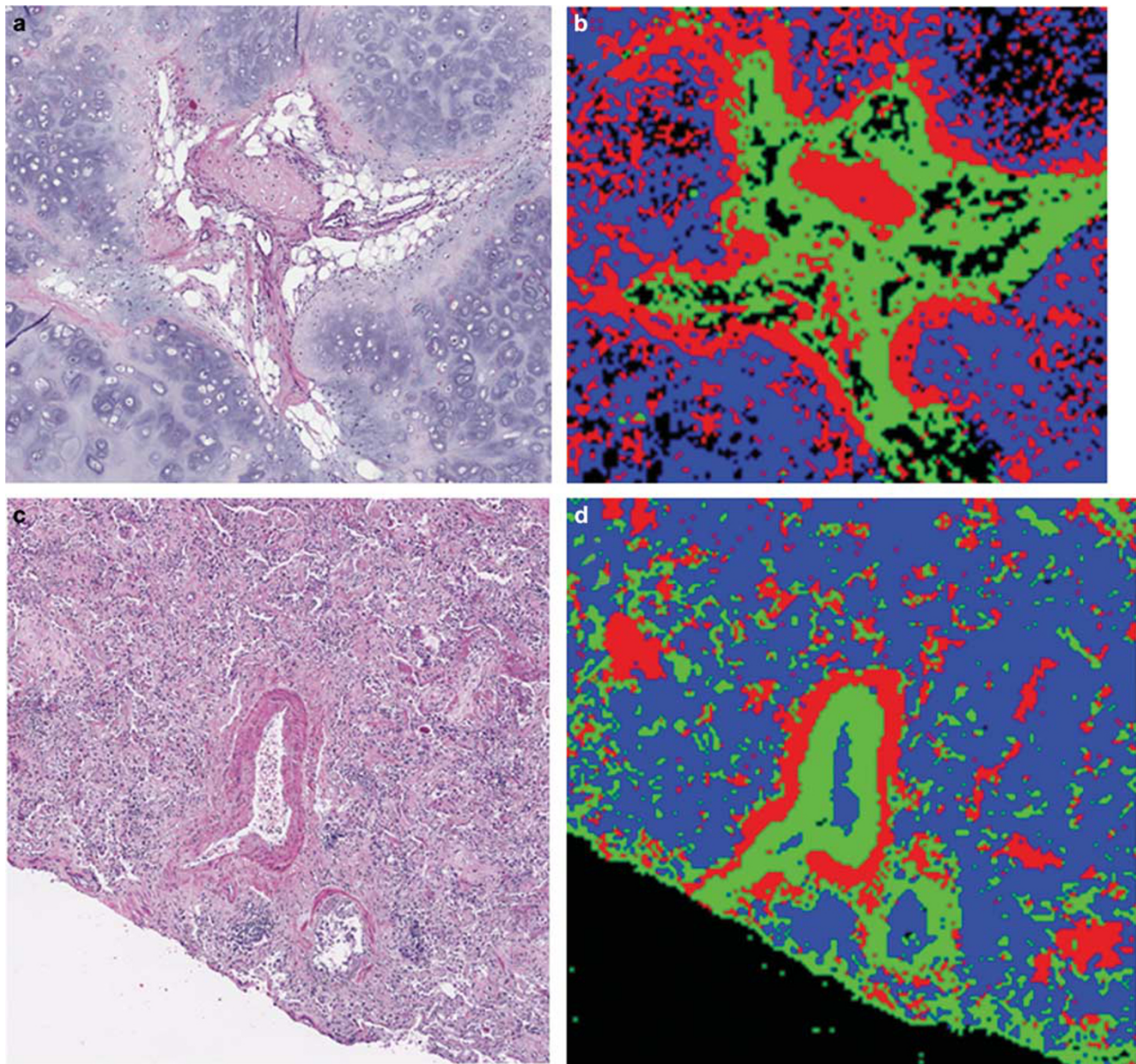
Figure 10 Visible microscopic images (left) and corresponding infrared pseudocolor images of a hamartoma (**a**, **b**) and organizing pneumonia (**c**, **d**). The cluster colors in panels B and D are arbitrary and not comparable.

misclassifications were made with low confidence, as indicated by the white pixels. The misclassifications and low probability classifications occurred in the vicinity of the necrotic region, and we believe that the presence of necrosis dominates spectral features such that the accuracy of the ADC vs SqCC classification is compromised.

Benign Lesions

The benign lesions were analyzed in a completely analogous manner. Images of H&E-stained sections and their corresponding HCA images were annotated as described in the section Annotation and data traceability'. Figure 10 shows

two examples of benign tumors, diagnosed as 'organizing pneumonia' and 'hamartoma'. The pixel spectra of annotated regions were incorporated into the tissue data bases of normal tissues, and contributed 20 new tissue types; in addition, many of the normal tissue types found in the noncancerous regions of cancer tissue spots were found in the benign lesions as well.

The tissue types from the nonmalignant tumors could be easily distinguished from those of malignancies. Figure 11 shows an HCA-based dendrogram that demonstrates that the class mean spectra from malignant and nonmalignant lesions can be easily distinguished, with all benign classes (blue) well
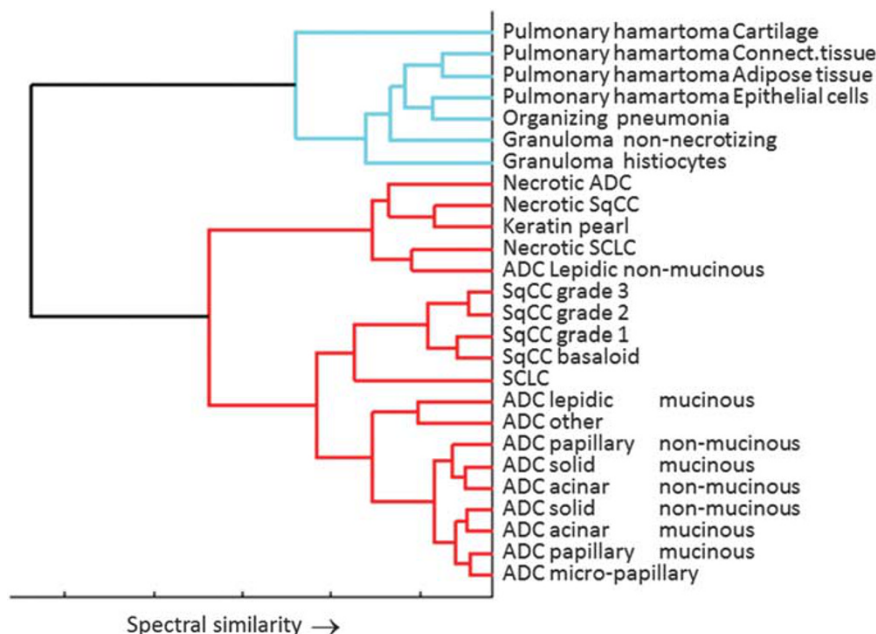
Figure 11 HCA-based dendrogram of class mean spectra of benign and malignant lung tumors, including the main cancer classes (SCLC, SqCC, and ADC).

Table 7 Distinction of cancer-adjacent and benign and inflammatory lesion-adjacent tissue types

	Blood vessel wall	Connective tissues
Cancer-adjacent tissues	13 356 (41)	128 485 (369)
Benign and inflammatory lesion-adjacent tissues	9825 (56)	3806 (14)
Accuracy	90.3%	88.9%

separated from the cancerous classes (red). This graph also depicts that unsupervised HCA clusters the cancerous spectra quite well by pathological criteria: all necrotic tissue classes are found together in one cluster along with keratin pearls, and all three SqCC grades and SCLC are differentiated from ADC. There is a clear distinction between the different ADC subtypes. A trained, LOOVC-based SVM classifier could separate the benign from malignant lesions with 99% accuracy.

Normal Tissue Adjacent to Cancer *Vs* Normal Tissue Adjacent to Benign or Inflammatory Lesions

We also investigated spectral differences between normal tissues from cancer patients and normal tissue from patients with benign lesions. To this end, the same normal tissue types found in cancer, benign, and inflammatory samples were extracted: the tissue class 'blood vessel wall in cancer adjacent tissue' comprised 13 356 pixel spectra from 41 patients, whereas the corresponding class 'blood vessel wall in benign

samples' contained 9825 pixel spectra from 56 patients. Similarly, connective tissue classes contained 128 485 pixel spectra in cancer-adjacent tissue and 3806 pixel spectra from benign and inflammatory samples, see Table 7. (Notice that patients could contribute to more than one tissue class in the case of cancer adjacent connective tissue; hence, the number of patients listed exceeds the number of patients in the training and test set).

The calculations here were carried out on a patient-based leave-one-out cross validation (LOOCV) because it was felt that the discrepancy in the number of patients and pixel spectra (particularly in the case of the connective tissue) did not warrant a pixel-based classifier. The LOOCV produced very impressive accuracy for the discrimination of these tissue types.

These results suggest that the same tissue types in cancer-adjacent and benign and inflammatory lesion-adjacent tissues exhibit significant spectral differences. At present, the underlying mechanisms for these differences are unknown. However, several possible explanations exist:

1. Leakage and/or secretions of substances from cancer cells into the surrounding normal tissue.
2. Early molecular changes (as a result of field cancerization) that have not yet resulted in morphologic abnormalities detectable by histopathologic examination in the cells located in cancer-adjacent normal tissue.
3. Host response-related changes including chemotaxis of immune cells in cancer-adjacent normal tissue.

SHP appears to be particularly sensitive to such changes: a combination of X-ray and (synchrotron-based) infrared

microspectroscopic studies[35] has suggested that metallo-proteinases in the vicinity of cancers may be responsible for the subtle spectral changes encountered in the vicinity of cancers. This observation may have far-reaching consequences for the detection and definition of the margins of resection, and the mechanism of metastasis formation.

CONCLUSIONS/SUMMARY

This paper demonstrates for a large patient data set the potential of SHP, a combination of infrared spectral methods with SVMs, for the classification of benign and malignant lung tumors. The major findings of this paper are:

1. SHP discriminates with high accuracy between normal tissue types and cancer.
2. SHP can be used with high accuracy for the classification between the three major lung cancers, SCLC, ADC, and SqCC, and between cancerous and necrotic tissue.
3. Preliminary results indicate that subclassification of ADC into the clinically relevant subclasses (lepidic, acinar, solid, papillary, and micropapillary) via SHP appears feasible once training and test data sets of sufficient size are available. These efforts are presently underway.
4. Benign and malignant lung tumors could be distinguished with high accuracy.
5. The findings of previous studies[35,36] that indicated that SHP is sensitive to the microenvironment of a tumor were confirmed. These findings also confirm similar results from SCP, an equivalent spectral method applied to exfoliated cells.[16,17] In these SCP studies it was found that morphologically normal cells that were harvested in the vicinity of cancerous lesions already showed spectral abnormality.

The label-free, morphology-independent, and composition-sensitive spectral methods introduced here have inherent advantages over many other methods used presently in tissue diagnostics, and present a major innovation in the field of medical diagnostics.

DISCLOSURE/CONFLICT OF INTEREST
Ayşegül Ergin and Stan Remiszewski are employees of Cireca Theranostics, LLC. Ali Akalin, Xinying Mu, Mark Kon, Clay M Thompson, and Max Diem are consultants for Cireca Theranostics, LLC.

1. Wong PT, Cadrin M, French SW. Distinctive Infrared spectral features in liver tumor tissue of mice: evidence of structural modifications at the molecular level. Exp Mol Pathol 1991;55:269–284.
2. Wong PT, Wong R, Caputo T, et al. Infrared spectroscopy of exfoliated human cervical cells: evidence of extensive structural changes during carcinogenesis. Proc Natl Acad Sci USA 1991;88:10988–10992.
3. Wong PT, Wong R, Fung MFK. Pressure-tuning FT-IR study of human cervical tissues. Appl Spectrosc 1993;47:1058–1063.
4. Cohenford M, Rigas B. Cytologically normal cells from neoplastic cervical samples display extensive structural abnormalities on IR spectroscopy: implications for tumor biology. Proc Natl Acad Sci USA 1998;95:15327–15332.
5. Diem M, Boydston-White S, Chiriboga L. Infrared spectroscopy of cells and tissues: shining light onto a novel subject. Appl Spectrosc 1999;53, pp148A–161A.
6. Pantanowitz L, Valenstein PN, Evans AJ, et al. Review of the current state of whole slide imaging in pathology. J Pathol Inform 2011;2:36–45.
7. Jagirdar J. Application of immunohistochemistry to the diagnosis of primary and metastatic carcinoma to the lung. Arch Pathol Lab Med 2008;132:384–396.
8. Bird B, Miljković MS, Remiszewski S, et al. Infrared spectral histopathology (SHP): a novel diagnostic tool for the accurate classification of lung cancers. Lab Invest 2012;92:1358–1373.
9. Fernandez DC, Bhargava R, Hewitt SM, et al. Infrared spectroscopic imaging for histopathology recognition. Nat Biotechnol 2005;23, pp469–474.
10. Miljković M, Bird B, Lenau K. Spectral cytopathology: new aspects of data collection manipulation and confounding effects. Analyst 2013;138:3975–3982.
11. Diem M, Mazur AI, Lenau K, et al. Molecular pathology via Infrared and Raman spectral imaging in ex-vivo and in-vivo optical pathology. In: Schmitt M, Popp J (eds). Vibrational Spectroscopy for Medical Diagnosis. Wiley-VCH, 2013.
12. Mu X, Kon M, Ergin A, et al. Statistical analysis of a lung cancer spectral histopathology (SHP) data set. Analyst 2015; doi: 10.1039/C4AN01832J (in press).
13. 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.
14. Romeo MJ, Boydston-White S, Matthäus C, et al. Vibrational micro-spectroscopy of cells and tissues. In: Lasch P, Kneipp J (eds). Biomedical Vibrational Spectroscopy. Wiley-Interscience: Hoboken, NJ, 2008, pp121–147.
15. Romeo MJ, Bird B, Boydston-White S, et al. Infrared and Raman Microspectroscopic Studies of Individual Human Cells Vibrational Spectroscopy for Medical Diagnosis. In: Diem M, Griffiths PR, Chalmers JM (eds). Vibrational Spectroscopy for Medical Diagnosis. John Wiley & Sons: Chichester, UK, 2008, pp 27–70.
16. Papamarkakis K, Bird B, Schubert JM, et al. Cytopathology by optical methods: spectral cytopathology of the oral mucosa. Lab Invest 2010;90:589–598.
17. 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.
18. Schubert JM. Spectral Cytology of Human Oral and Cervical Samples in Chemistry & Chemical Biology. Northeastern University: Boston, 2011.
19. Bach PB. Is our natural-history model of lung cancer wrong? Lancet Oncol 2008;9:693–697.
20. Marusyk A, Almendro V, Polyak K. Intra-tumour heterogeneity: a looking glass for cancer? Nat Rev Cancer 2012;12:323–334.
21. Girard N, Deshpande C, Lau C, et al. Comprehensive histologic assessment helps to differentiate multiple lung primary nonsmall cell carcinomas from metastases. Am J Surg Pathol 2009;33:1752–1764.
22. Diem M, PR Griffiths PR, Chalmers JM. Vibrational spectroscopy for medical diagnosis. In: Diem M, Griffiths PR, Chalmers JM (eds). John Wiley & Sons: Chichester, UK, 2008.
23. Bird B, Miljkovic M, Romeo MJ, et al. Infrared micro-spectral imaging: automatic distinction of tissuetypes in axillary lymph node histology. BMC Clin Pathol 2008;8:1–14.
24. Diem M, Mazur A, Lenau K, et al. Molecular pathology via IR and Raman spectral imaging. J Biophotonics 2013;6, 855–886.
25. Mazur AI, Monahan JL, Miljković M, et al. Vibrational spectroscopic changes of B-lymphocytes upon activation. J Biophotonics 2013;6:101–109.
26. Bassan P, Lee J, Sachdeva A, et al. The inherent problem of transflection-mode infrared spectroscopic microscopy and the ramifications for biomedical single point and imaging applications. Analyst 2013;138:144–157.

27. Wrobel TP, Wajnchold B, Byrne HJ, *et al*. Electric field standing wave effects in FT-IR transflection spectra of biological sections: simulated models of experimental variability. Vib Spectrosc 2013;69: 84–92.

28. foliobiocom Deparaffin Protocol Available from http://foliobiocom/media/pdf/Deparaffin%20protocolpdf.

29. Baker MJ, Trevisan J, Bassan P, *et al*. Using Fourier transform IR spectroscopy to analyze biological materials. Nat Protoc 2014;9: 1771–1791.

30. Miljković M, B Bird B, Diem M. Dispersive line shape effects in infrared spectroscopy. Analyst 2012;137:3954–3964.

31. Adams MJ. Chemometrics in analytical spectroscopy. 2nd edn. In: Barnett NW (ed). RSC Analytical Spectroscopy Monographs; Royal Society of Chemistry: Cambridge, 2004.

32. Statnikov A, Aliferis CF, Hardin DP, *et al*. A Gentle Introduction to Support Vector Machines in Biomedicine, Vol 1: Theory and Methods. World Scientific Publishing: Hackensack, NJ, 2011, pp 1–181.

33. Rahman MM, Davis DN. Addressing the class imbalance problem in medical datasets. Int J Machine Learn Comp 2013;3:224–228.

34. Platt J. Probabilistic outputs for support vector machines and comparison to regularized likelihood. In: Smola A, *et al*. (eds). Methods in Advances in Large Margin Classifiers; MIT Press: Cambridge, MA, 2000.

35. Federman S, Miller LM, Sagi I. Following matrix metalloproteinases activity near the cell boundary by infrared micro-spectroscopy. Matrix Biol 2002;21:567–577.

36. Holton SE, Walsh MJ, Bhargava R. Subcellular localization of early biochemical transformations in cancer-activated fibroblasts using infrared spectroscopic imaging. Analyst 2011;136:2953–2958.