The American Journal of Pathology, Vol. 191, No. 7, July 2021

ELSEVIER

The American Journal of
PATHOLOGY
ajp.amjpathol.org

MACHINE LEARNING, COMPUTATIONAL PATHOLOGY, AND BIOPHYSICAL IMAGING

# Quantum Cascade Laser-Based Infrared Imaging as a Label-Free and Automated Approach to Determine Mutations in Lung Adenocarcinoma



Nina Goertzen,*† Roberto Pappesch,‡ Jana Fassunke,‡ Thomas Brüning,§ Yon-Dschun Ko,¶ Joachim Schmidt,‖ Frederik Großerueschkamp,*† Reinhard Buettner,‡ and Klaus Gerwert*†

*From the Center for Protein Diagnostics,* Biospectroscopy, and the Department of Biophysics,† Faculty of Biology and Biotechnology, Ruhr University Bochum, Bochum, Germany; the Institut für Pathologie,‡ Universitätsklinikum Köln, Cologne, Germany; the Institute for Prevention and Occupational Medicine of the German Social Accident Insurance,§ Institute of the Ruhr University Bochum, Bochum, Germany; the Department of Internal Medicine,¶ Johanniter-Kliniken Bonn GmbH, Johanniter Krankenhaus, Bonn, Germany; and the Lung Cancer Center Bonn,‖ Department of Thoracic Surgery, Helios Klinikum Bonn/Rhein-Sieg and Department of Surgery, Division of Thoracic Surgery, Universitätsklinikum Bonn, Germany*

Accepted for publication
April 22, 2021.

Address correspondence to Klaus Gerwert, Dr.rer.nat., Center for Protein Diagnostics, Ruhr University Bochum, Gesundheitscampus 4, 44801 Bochum, Germany. E-mail: klaus.gerwert@ruhr-uni-bochum.de.

Therapeutic decisions in lung cancer critically depend on the determination of histologic types and oncogene mutations. Therefore, tumor samples are subjected to standard histologic and immunohistochemical analyses and examined for relevant mutations using comprehensive molecular diagnostics. In this study, an alternative diagnostic approach for automatic and label-free detection of mutations in lung adenocarcinoma tissue using quantum cascade laser–based infrared imaging is presented. For this purpose, a five-step supervised classification algorithm was developed, which was not only able to detect tissue types and tumor lesions, but also the tumor type and mutation status of adenocarcinomas. Tumor detection was verified on a data set of 214 patient samples with a specificity of 97% and a sensitivity of 95%. Furthermore, histology typing was verified on samples from 203 of the 214 patients with a specificity of 97% and a sensitivity of 94% for adenocarcinoma. The most frequently occurring mutations in adenocarcinoma (*KRAS*, *EGFR*, and *TP53*) were differentiated by this technique. Detection of mutations was verified in 60 patient samples from the data set with a sensitivity and specificity of 95% for each mutation. This demonstrates that quantum cascade laser infrared imaging can be used to analyze morphologic differences as well as molecular changes. Therefore, this single, one-step measurement provides comprehensive diagnostics of lung cancer histology types and most frequent mutations. *(Am J Pathol 2021, 191: 1269–1280; https://doi.org/10.1016/j.ajpath.2021.04.013)*

Cancer is the most frequent cause of death, second only to cardiovascular diseases, in industrialized countries. In 2018, approximately 9.6 million deaths were attributed to cancer. Worldwide, 2.1 million people are confronted with lung cancer diagnosis, and 1.76 million die from it every year (World Health Organization, *https://gco.iarc.fr/today/home*, last accessed January 22, 2021). Lung tumors are characterized by a high degree of heterogeneity and are divided into numerous types [eg, small-cell lung carcinoma (SCLC), adenocarcinoma, squamous cell carcinoma, neuroendocrine carcinoma, carcinoids, and many rare histologies], which are linked to different prognoses and therapeutic approaches.[1–3] This complicates the process of diagnosis by the pathologist and may lead to intraobserver and interobserver variability.[4–7] If lung cancer is suspected, an X-ray and a subsequent computed tomographic examination of the thorax are performed, followed by tissue sample collection by bronchoscopy, fine-needle aspiration, transthoracic needle aspiration, or surgery. In addition to histologic typing, tumor samples are examined comprehensively for oncogene alterations by ultra-deep next-generation sequencing (NGS). The three most common

Supported by the German Social Accident Insurance project FP-0259 and by the Center for Protein Diagnostics.
Disclosures: None declared.


https://doi.org/10.1016/j.ajpath.2021.04.013

mutations in adenocarcinomas of the lung, which are by far the most common histologic type, are found in the genes encoding tumor protein 53 (TP53), KRAS proto-oncogene, GTPase (KRAS), and epidermal growth factor receptor EGFR). The presence of one of these mutations may influence both the patient's prognosis as well as further therapeutic decisions.[8] Tumors with activating mutations in the *EGFR* gene often elicit a good response to tyrosine kinase inhibitors, whereas non-*EGFR*-mutated tumors do not respond at all to these tyrosine kinase inhibitors.[9] Similarly, patients with *KRASG12C* mutated tumors respond to G12C-targeted GTPase inhibitors.[10] Furthermore, lung cancers with *TP53* mutations have a poor prognosis, and patients show limited response to targeted therapies, similar to *ALK* (ALK receptor tyrosine kinase) fusion-positive tumors.[11] On the basis of different genomic vulnerabilities, patients with lung cancer clearly benefit from prior characterization of mutations.[8]

In the last two decades, several reports by different groups have presented Fourier-transform infrared (FTIR)–based microscopy for tissue diagnostics.[12–16] FTIR imaging was used for the automated and label-free classification of cancerous tissues from lung,[17] colon,[18] bladder,[19,20] kidney,[21] and prostate cancer,[22] as well as from melanoma[23] patients, with sensitivity and specificity of >90% compared with diagnostics by pathologists using histologic staining and immunohistochemical methods. In addition to the tumor identification in tissue samples, FTIR imaging may be employed for further analyses, such as glioma grading[24] for distinction of tumors in both colon[25] and bladder[20,26] cancer samples, as well as for mutation analysis in gliomas.[27] A workflow to distinguish thoracic and lung tumors not only by tumor type, but also to classify the subtypes of diffuse malignant mesothelioma (sarcomatoid and epithelioid, with 88% accuracy)[28] and the five World Health Organization–defined lung adenocarcinoma histologic types (acinary, solid, papillary, micropapillary, and lepidic, with 96% accuracy)[29] was presented. The combination of FTIR imaging and laser capture microdissection (LCM) with subsequent proteomics adds molecular resolution to the spatial resolution provided by hyperspectral data sets. As demonstrated previously, this approach can also be used for biomarker identification.[26,28] These results indicate that a single index color image can provide the same biochemical information as several immunohistochemistry (IHC) stains.

Extending the previous work, new infrared (IR) microscopes with tunable quantum cascade lasers (QCLs) as IR sources, instead of globar and uncooled microbolometer detectors, were used to perform FTIR imaging feasible for routine diagnostic applications. The reduction of the measurement time using QCL-based IR microscopes on breast cancer tissue microarrays,[30] samples of liver fibrosis,[31] and colon tissues,[32] as well as the identification of goblet cells in the colon mucosa, has been reported previously.[33] The differentiation of malignant and nonmalignant structures in breast tissue has also been reported by this technique.[34] Using two Spero-QT IR microscopes (Daylight Solutions, San Diego, CA), tumor lesions and healthy tissue types on whole slices were identified with 96% sensitivity and 100% specificity compared with histopathology.[35] A comparison of the Spero-QT with the previously used Cary-FTIR system (Agilent, Santa Clara, CA) showed a reduction of measurement time at the same wave number (inverse wavelength, $1/\lambda$) resolution by a factor of 160. Therefore, <30 minutes was required for a whole slice measurement. This corresponds to the time required for histologic staining of fresh-frozen tissue and evaluation by a pathologist. A previous study revealed that QCL-IR imaging can classify changes at the molecular level in colorectal cancer tissue. The recognition of microsatellite stability and instability of cancerous tissue was verified with 100% sensitivity and 93% specificity compared with immunohistochemistry and fragment length analysis.[36]

In this study, a label-free, automated, spatially resolved, and observer-/operator-independent approach using QCL-based IR imaging is presented. This technique was verified on thin sections of 536 formalin-fixed, paraffin-embedded (FFPE) tumor and nontumor lung tissues from 214 patients. Cancerous regions were identified with a sensitivity of 95% and a specificity of 97% compared with histopathology. Furthermore, the tumor type (94% sensitivity and 97% specificity for adenocarcinoma) and adenocarcinoma mutation status (*KRAS*, *EGFR*, or *TP53* mutation) were determined with a sensitivity and specificity of 95% compared with the NGS gene panel result.

## Materials and Methods

### Sample Sets

Two different sample sets were used in this study (Table 1). The first ($N = 21$) set was used for training the random forest (RF) classifiers. It included tumor and normal tissue samples from patients diagnosed with adenocarcinoma ($n = 10$), squamous cell carcinoma ($n = 5$), neuroendocrine carcinoma ($n = 1$), SCLC ($n = 1$), carcinoid ($n = 1$), pulmonary chondroid hamartoma ($n = 1$), or other lung diseases ($n = 2$). A *KRAS* mutation occurred in four adenocarcinomas, and three adenocarcinomas harbored a *TP53* mutation and an *EGFR* mutation. The patients were 50 to 85 years old at specimen collection and had an average age of 68 years. Eleven patients were female, and 10 were male. The second sample set ($N = 214$) was used to verify the RF classifiers. Among them were tumor and normal tissue samples from patients diagnosed with adenocarcinoma ($n = 170$), squamous cell carcinoma ($n = 23$), neuroendocrine carcinoma ($n = 3$), SCLC ($n = 3$), carcinoid ($n = 4$), pulmonary chondroid hamartoma ($n = 5$), or other lung diseases ($n = 6$). *KRAS*, *EGFR*, or *TP53* mutations were detected in 20 adenocarcinomas. The remaining tumors with adenocarcinoma ($n = 110$) contained other or no

**Table 1** Sample Sets Included in This Study for Training and Verification

| | | Sex | | | Smoking state | | | Ex-smoker, years | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Group | | Male | Female | Unknown | Nonsmoker | Smoker | Unknown | >10 | 5–10 | <5 |
| Verification | 214 | 105 | 101 | 8 | 39 | 100 | 8 | 51 | 8 | 8 |
| Training | 21 | 10 | 10 | 1 | 6 | 7 | 1 | 4 | 1 | 2 |

mutations. In the verification samples, patients from 41 to 84 years with an average age of 68 years were included. A total of 104 patients were female, and 110 were male.

## Ethical Statement

The study was approved by the University of Cologne Ethics Committee (registration number 15-116). General informed consent for research was obtained from the patients. All procedures are in accordance with the approved guidelines and regulations for human experimental research.

## Sample Preparation

FFPE lung tissue sections were obtained from the Institute of Pathology, University Hospital Cologne (Cologne, Germany). The samples were collected during surgery and prepared following standardized protocols. Fresh-frozen or FFPE tissue blocks were cut into section (10 μm thin) and floated onto polyethylene terephtalat membrane frame slides. The management and distribution of the samples were performed by the Institute for Prevention and Occupational Medicine of the German Social Accident Insurance (Ruhr University Bochum, Bochum, Germany). Before the spectral data acquisition, the FFPE samples were dewaxed using established protocols.

## Data Acquisition

For spectral data acquisition, two Spero QT QCL-based microscopes and Chemical Vision software version 3.2 (Daylight Solutions) were used. In addition to the original setup, a purge air diffuser was connected to the sample chamber. Furthermore, the stage was modified so that two slides could be analyzed in a row to reduce the equilibrium time before measurements. The tissue samples were measured with a 4× objective (0.3 numerical aperture), which covers a field of view of 2 × 2 $mm^2$ in a spectral range of 1800 to 948 $cm^{-1}$ with a spectral resolution of 2 $cm^{-1}$ in transmission mode. Spero QT operates with an uncooled microbolometer focal plane array detector with 480 × 480 pixels and a pixel size of 4.25 × 4.25 μm.

## Data Processing and Analysis

Spectral artifacts from folds and cracks in the tissue were eliminated by quality control based on the integral of the amide I band. Disturbing bands caused by the polyethylene terephtalat membrane or embedding medium (Tissue-Tek, Sakura Finetek, Staufen, Germany) were removed on the basis of the relations of the integral of the amide I band and the integral of the regions between and 1135 to 1064 $cm^{-1}$ and 1800 to 1700 $cm^{-1}$. After quality control, Mie scattering was corrected using the resonant Mie scattering-extended multiplicative signal correction (RMieS-EMSC) algorithm by Bassan[37] (RMie_EMSC_v2) with one iteration. Unsupervised classification was performed using k-means or hierarchical cluster analysis (HCA). Supervised and unsupervised classification was performed on unsmoothed data on the fingerprint region from 1760 to 998 $cm^{-1}$.

## Classifier Setup and Spectral Database Generation

The workflow with the RF classifier used for this work was established and described in previous publications.[29,35,36] The RF classifier was shown to be robust and reliable for tissue classification using IR imaging.[25,38–40] In this study, five consecutive RF classifiers were generated. Therefore, a spectral database with tissue-specific spectral information for pathologic regions, infiltrated inflammatory cells, necrosis, muscle, connective tissue, alveoli, blood, calcification, pulmonary chondroid hamartoma, and mucus was set up. The databases for the other RF levels contained the spectral signatures for cancerous, necrotic, and inflammatory tissue (second-level RF), adenocarcinoma, squamous cell carcinoma, small-cell lung cancer, carcinoids, and neuroendocrine carcinoma (third-level RF), and adenocarcinoma with *KRAS*, *EGRF*, and *TP53* mutations (fourth- and fifth-level RF). The pathologic findings per sample were used as ground truth for morphologic detection (first- and second-level RF) as well as for the tumor type identification (third-level RF). For mutation analysis (fourth- and fifth-level RF), the NGS result of the whole tumor per sample was used as ground truth. The first- and second-level classifiers were set up with 50 decision trees and 16 spectral features randomly chosen per decision in the trees. For the other levels, 500 decision trees and 16 spectral features were used. The exact class composition and number of spectra of all five RF classifiers can be seen in Supplemental Tables S1 through S4. Because of the lower signal/noise ratios and baseline effects, the spectral data range was reduced to 1760 to 998 $cm^{-1}$, so that 382 wave numbers were used for RF training. The RF for lung tissue classification was built from samples of 21 patients. A total of 536 samples from 214 patients for the lung tissue classifier were available for verification. RF classifiers perform implicit feature selection

using a small subset of variables. The visualization of this feature selection can be accomplished using the Gini importance, which can be considered as an indicator for the relevance of the features in terms of a relative ranking. The Gini importance thus provides a relative value for the frequency of use of a certain feature for the split at a node within the decision trees of a model as well as for the overall discrimination value of a feature. The Gini importance plots for each of the trained classifiers detailed by individual training classes are illustrated in Supplemental Figures S1 through S5. All computations were performed using MATLAB R2019a (MathWorks, Natick, MA). The final annotation was provided as index color images and compared with that of the corresponding hematoxylin and eosin (H&E)–stained tissue images. Pathologists at the Pathology Institute, University Hospital Cologne, supplied their histologic reports.

### IR Imaging-Guided LCM Workflow

The workflow is based on the one previously described by Großerueschkamp et al.[28] The respective lung tissue samples were measured using a Spero-QT as usual, and the spectral data were classified during this process. The resulting index color image was used to determine the region of interest. For this analysis, only tumor regions that were incorrectly classified by the fourth or fifth RF classifier (mutation status) were selected as the region of interest. The coordinate transfer was performed using a two-dimensional Helmert transformation based on three reference points. Because the chemical vision software does not allow collection of coordinates, Helmert transformation was done using reference points taken from the respective false color image. The sample was transferred to an LCM microscope (PALM MicroBeam; Zeiss, Jena, Germany), and the coordinates of the reference points were taken. The coordinate transformation was performed in MATLAB. As only tissue pieces of certain shapes and sizes can be lifted and collected by the PALM Zeiss instrument, the region of interest was further subdivided. This resulted in shapes with areas in the range 100 to 50,000 $\mu m^2$. The coordinates of these shapes were imported to the PALM Robo software version 4.6 and cut using the 5× objective of the instrument. The tissue was collected in NGS incubation buffer for FFPE tissues and stored at −80°C until analysis.

### NGS Data

Mutational analysis of low-input DNA NGS was performed using an Ion AmpliSeq Custom DNA Panel (Thermo Fisher Scientific, Waltham, MA) and the Ion AmpliSeq Library Kit 2.0 (Thermo Fisher Scientific), according to the Ion AmpliSeq Library Preparation User Guide (Thermo Fisher Scientific). After multiplex PCR and adapter ligation, libraries were generated by target enrichment using the Gene Read DNA Library I Core Kit, the Gene Read DNA I Amp Kit (Qiagen, Hilden, Germany), and the NEXTflex DNA Barcodes (Bio Scientific, Phoenix, AZ). For sequencing, 12 pmol/L of the constructed libraries was processed on the MiSeq platform (Illumina, San Diego, CA) with a MiSeq reagent kit V2 (Illumina) with 300 cycles following the manufacturer's recommendations. Data analysis and mutation calling were performed as previously described.[41,42] The genes that were evaluated for mutations are shown in Supplemental Table S5.

The QIAseq-targeted DNA panel for human lung cancer (NGHS-005X-96) with the GeneRead DNAseq Panel PCR Kit V2 (Qiagen) was used for a subset of samples by preparing libraries using the Gene Read DNA Library I Core Kit and the Gene Read DNA I Amp Kit (Qiagen), according to the manufacturer's protocol. Final library products were quantified, diluted, and pooled in equal amounts. A total of 1.2 pmol/L of the pooled final libraries was sequenced on a NextSeq Sequencer (Illumina) with the NextSeq 500 Mid Output Kit v2 following the manufacturer's recommendations. Refer to Supplemental Table S6 for details of the analyzed regions.

## Results

### Tumor Identification and Tumor Type Determination in Lung Tissues

FFPE tumor and nontumor lung tissue sections from 235 patients were used for this study. Within this cohort, 180 patients were diagnosed with adenocarcinoma of the lungs and 28 patients were diagnosed with squamous cell carcinoma. The remaining patients had other lung tumors (SCLC, neuroendocrine carcinoma, carcinoid, and pulmonary chondroid hamartoma), metastasis (eg, from colorectal carcinoma), or other lung diseases (pneumonia or chronic obstructive lung disease). The established IR imaging workflow[25,38] used for this study is shown in Figure 1. Data acquisition was performed with QCL-based infrared microscopes on unstained, unmodified thin sections of lung tissues. Each pixel of the image is represented by one IR spectrum, which shows an integral of the information of the biochemical composition of the tissue. Therefore, the IR spectrum serves as a fingerprint for morphologic or molecular changes in the tissues. Thus, machine-learning algorithms, such as supervised classifiers, can be used to distinguish between spectra of different tissue types or molecular conditions. The results are presented as index color images, where each color represents a different tissue type or molecular condition. For diagnosis, the pathologist uses histologic methods, such as H&E and IHC staining, as well as NGS gene panels. The results of the mentioned analyses per sample were used as ground truth for the construction of the spectral database for the classifiers. RF supervised classifier was used in this study, which provides reliable and robust results for the annotation of tissue samples.

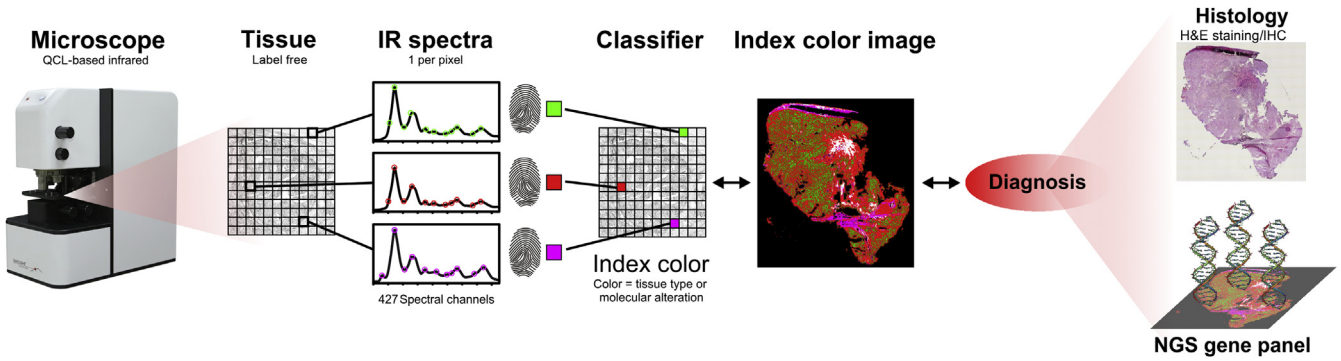



**Figure 1** Infrared (IR) imaging process. The spectral data are acquired with a quantum cascade laser (QCL)—based IR microscope on unmodified lung tissue samples. Characteristic spectra for each tissue type are extracted and used to train a supervised classifier. The classifier results are presented as index color images. Pathologic diagnosis by histologic methods [hematoxylin and eosin (H&E) and immunohistochemical (IHC) staining] and a next-generation sequencing (NGS) gene panel to determine the mutation status.

To analyze lung tissues, three hierarchical RF classifiers (Figure 2A) were elucidated using the spectral data of 21 patients. Tumor, chronic obstructive lung disease, pneumonia, and nontumor or nondiseased tissue samples of these patients were included in the training data set. The first RF classifier was used to (Figure 2A) differentiate between different tissue types, such as connective tissue, muscle, and blood, as well as calcification, necrotic tissue, pulmonary chondroid hamartoma (cartilage tumors), and pathologic regions. Subsequently, spectra classified as pathologic were

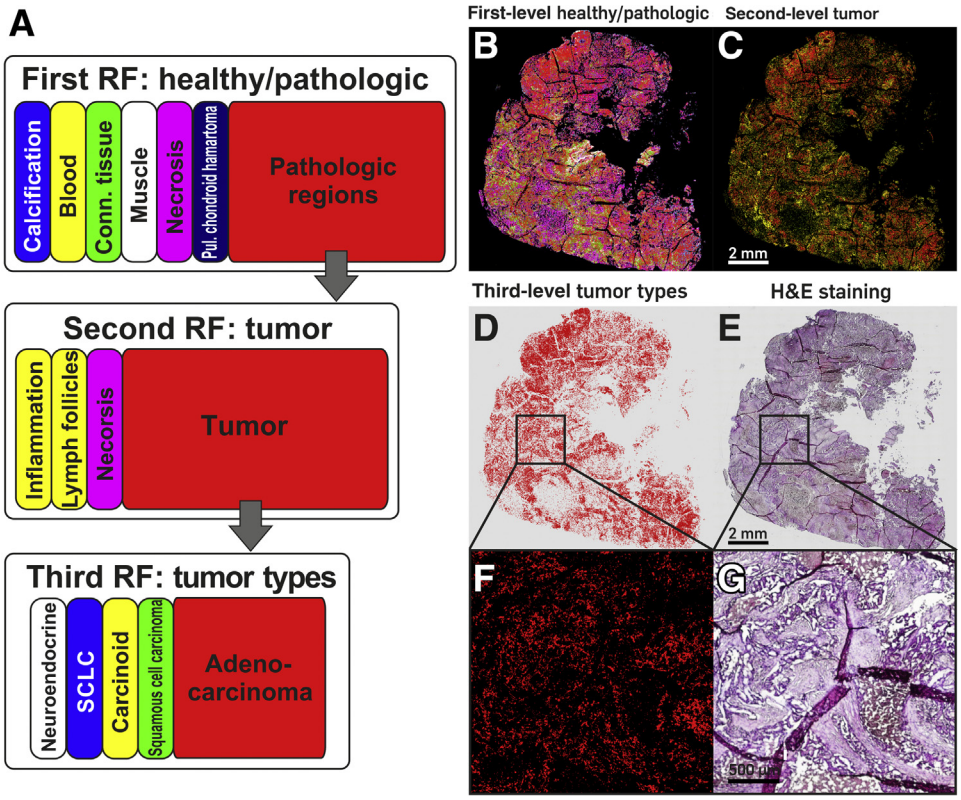



**Figure 2** **A:** Schematic representation of random forest (RF) classifier structure and color code. **Top row:** The first-level RF differentiates between healthy and pathologic tissue. **Middle and bottom rows:** The second-level RF (**middle row**) subdivides the pathologic region into inflammatory infiltrates, lymph follicles, and the tumor region, which is used in the third-level RF (**bottom row**) to determine the tumor type. **Right side:** Quantum cascade laser—based infrared imaging of the thin section of a whole slice lung adenocarcinoma tissue. **B:** Index color image of the result of the first RF classifier, which identifies pathologic regions. **C:** Subdivision of the pathologic regions using the second RF classifier to identify the tumor. **D:** Analysis of the tumor region by the third RF classifier to determine the tumor type. **E:** Hematoxylin and eosin (H&E)—stained image of a thin section of adenocarcinoma tissue. The tumor lesion is identifiable because of the purple hematoxylin staining. **F** and **G:** Enlarged sections of **boxed areas** in **D** and **E** show that the red pixels of the index color image (tumor region) match precisely with tumor lesions of the H&E-stained image. Scale bars: 2 mm (**B**—**E**); 500 µm (**F** and **G**). SCLC, small-cell lung carcinoma.

analyzed by a second RF classifier (Figure 2A), which was used to identify inflammatory infiltrates, lymph follicles, slightly necrotic tissue, and tumor regions. A detailed illustration of the separation of lymph follicles and inflammatory infiltrates is shown in Supplementary Figure S6. A third RF classifier (Figure 2A) used the tumor spectra to determine the tumor type. This RF classifier identified the five most common tumor types (adenocarcinoma, squamous cell carcinoma, SCLC, neuroendocrine carcinoma, and carcinoid). The results of the tree classifiers are presented in Figure 2 on a lung tissue section with adenocarcinoma. The different tissue types as well as the pathologic region (Figure 2B) were identified more precisely compared with histopathology (Figure 2E). The same applied to the tumor region (Figure 2C) determined by the second RF. Figure 2, D and F (detail of D with matching H&E image, G), illustrates that the tumor type (adenocarcinoma) was determined correctly and homogeneously within the tumor lesions.

For verification, tumor and nontumor samples from 214 patients were available. Of these patients, 208 were diagnosed with a lung tumor (Table 2). Because of the large size of the whole lung tissue, thin sections, and alterations that occur during the staining process, pixel-based analysis was not performed. Pixel-based analysis is not relevant for clinical diagnosis, but can be used for an overall annotation of the tissue section. Therefore, statistical analyses were performed using the overall diagnosis of the sections. For tumor identification, all sections with $\geq$5% pathologic-classified spectra were classified and recognized as tumor samples. Sections with <5% tumor-classified spectra were rated as nontumor samples. On considering these parameters, tumor identification achieved a sensitivity of 95% and a specificity of 97% compared with histopathology. For verification of tumor type (third RF), tumor samples from 203 patients were accessible. Most of these samples were diagnosed with lung adenocarcinoma (170 patients), with approximately 50% clinical incidence being the most common lung tumor type.[43] Twenty-three patients were diagnosed with squamous cell carcinoma. Only four patients had carcinoids, three had neuroendocrine carcinoma, and three had SCLC. The evaluation of the third classifier was performed using a simple majority vote (Table 3). Therefore, the tumor type could be determined with a sensitivity of 94% and a specificity of 97% for adenocarcinoma and a sensitivity of 96% and a specificity of 96% for squamous cell carcinoma. For carcinoids, neuroendocrine carcinomas, and SCLC tissue samples, a sensitivity and specificity of 100% for tumor type identification were achieved. Because of the low number of samples for verification, the reliability of these values remains questionable for this cohort and should be further addressed.

**Table 2** Lung Tissue Sample Data Set for Verification and Training, According to the Tumor Types of the Patients

| Group | ADC | SCC | CD | NEC | SCLC | H | Other* | $\sum$ |
|---|---|---|---|---|---|---|---|---|
| Training | 10 | 5 | 1 | 1 | 1 | 1 | 2 | 21 |
| Verification | 170 | 23 | 4 | 3 | 3 | 5 | 6 | 214 |

*Other indicates pneumonia, chronic obstructive lung disease, and metastasis.

ADC, adenocarcinoma; CD, carcinoid; H, pulmonary chondroid hamartoma; NEC, neuroendocrine carcinoma; SCC, squamous cell carcinoma; SCLC, small-cell lung cancer.

**Table 3** Results of the Tumor Typing RF Classifier (Third Level) for FFPE Tissue

| Variable | ADC | SCC | CD | NEC | SCLC | H |
|---|---|---|---|---|---|---|
| Sensitivity, % | 94 | 96 | 100 | 100 | 100 | 100 |
| Specificity, % | 97 | 96 | 100 | 100 | 100 | 100 |

ADC, adenocarcinoma; CD, carcinoid; FFPE, formalin fixed, paraffin embedded; H, pulmonary chondroid hamartoma; NEC, neuroendocrine carcinoma; RF, random forest; SCC, squamous cell carcinoma; SCLC, small-cell lung cancer.

## Analysis of Mutations in Lung Adenocarcinoma

In addition to the histochemical and immunohistochemical staining for subtyping lung cancer, the Institute of Pathology, University Hospital Cologne, sequenced an NGS gene panel to identify relevant mutations in lung tumor tissues. Previous studies showed that IR imaging can be used for biomarker identification,[26,28,36,44] otherwise performed by several IHC stainings. Therefore, to add a molecular dimension to the spatial IR resolution, herein, two additional RF classifiers were trained to identify mutations in lung cancer tissues (Figure 3). The most frequent mutations in lung adenocarcinomas within this data set were mutations in *KRAS*, *EGFR*, and *TP53*. Therefore, these three mutations were chosen to build the RF classifier. To train the spectral data of four patients with *KRAS*, three with *EGFR* mutations and three with *TP53* mutations were required. The structure of this RF is shown in Figure 3. In the first step, the spectra previously classified as adenocarcinoma on the third level (tumor type identification) were classified as *TP53* or spectra, which could be either *KRAS* or *EGFR*. The fifth RF subdivides the spectra further as *EGFR*- or *KRAS*-classified spectra. This is necessary because the spectra (Figure 4) of the tissues with these mutations are similar to each other, indicating that both mutated genes activate mitogen-activated protein kinase signaling. The most noticeable differences between the spectral data of lung tissues with these mutations occur within the fingerprint region between 1350 and 1000 $cm^{-1}$. This is probably based on the fact that the *EGFR* mutation causes a constant activation of this receptor, which also triggers the KRAS signaling cascade. For more detailed spectral information on the training data set, see Supplemental Figures S7 through S12.

The results of these two RF classifiers to determine the mutation status of lung adenocarcinoma are presented in Figure 5. The index color images of sections of tissue samples with *TP53* (Figure 5A), *KRAS* (Figure 5C), and

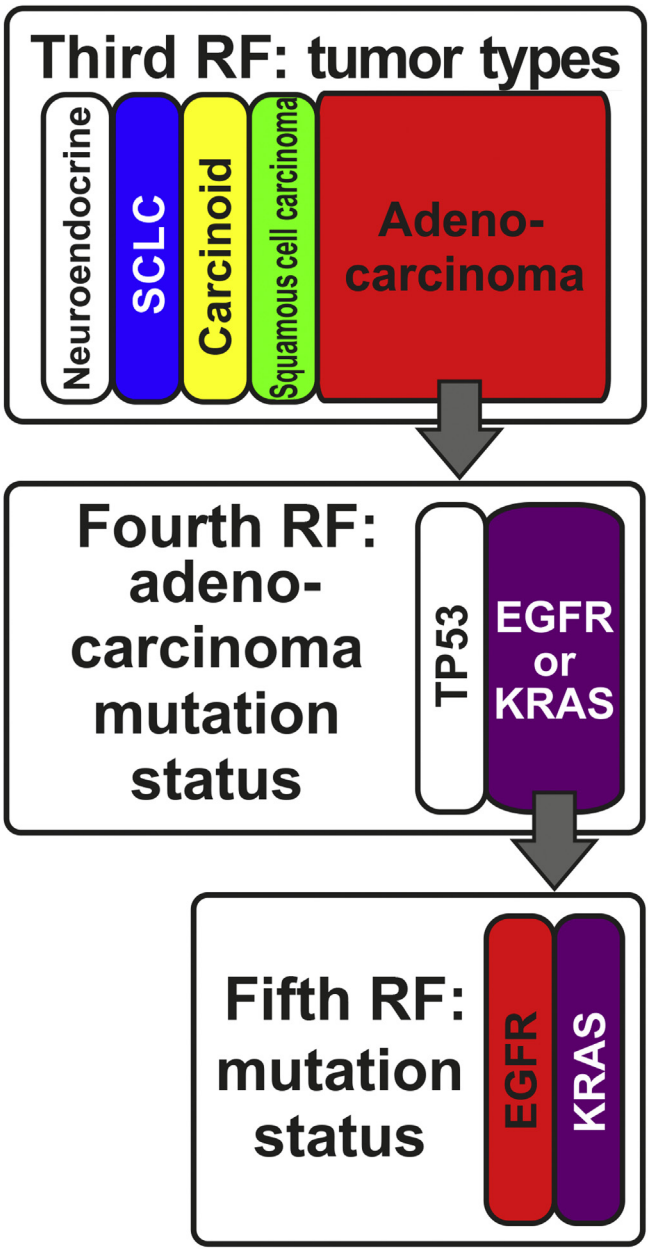



**Figure 3** Schematic representation of the structure and color code of the third to fifth level of the random forest (RF) classifier for mutation analysis of adenocarcinomas. **Top panel:** The third RF identifies the tumor type of the regions previously classified as tumor. **Middle** and **bottom panels:** Subsequently, the mutation status of the adenocarcinomas is determined on a fourth (*TP53* or *EGFR/KRAS*) and fifth level (*KRAS* or *EGFR*). SCLC, small-cell lung carcinoma.

*EGFR* (Figure 5E) mutation are shown in comparison to their corresponding H&E staining (Figure 5, B, D, and F; refer to Supplemental Figure S13 for whole tissue slices). The identified tumor regions and the tumor lesions visible based on H&E staining matched well. Furthermore, the three mutations were determined correctly as well as homogeneously within the tumor lesions.

For verification, samples of 60 patients with lung adenocarcinoma (20 tumor samples with *KRAS*, *EGFR*, and *TP53* mutations each) were used. The evaluation of the fourth and fifth RF classifiers was performed using a simple majority vote. Slices with >50% of the previously classified tumor spectra assigned to one mutation class were rated as positive for this mutation. The threshold was confirmed using receiver operating characteristic curves (Supplemental Figure S14) for both classifiers. The mutation status was determined with sensitivities and specificities of 95% for each mutation compared with the NGS gene panel.

### Clarification of Mutation Classifier Results via IR-Guided LCM

In total, 87% (52 of 60) of the verification data set for the mutation status classifier was identified, with >65% of the tumor spectra assigned to the correct mutation class (Supplemental Table S7). The tumors of three patients (one with *KRAS*, *EGFR*, and *TP53* mutations) were identified correctly, but the respective mutations were misclassified. The cases with *EGFR* and *KRAS* mutations were assigned to other mutations. The incorrect *TP53* case was classified as *KRAS* or *EGFR* (fourth RF). Five cases were correctly classified (50% to 65% spectra assigned to the correct mutation) but showed heterogeneity with large contributions of other mutations. After a renewed control of the NGS gene panel results, one of these cases (63.71% as *KRAS* and 36.29% as *TP53* mutated classified tumor spectra) showed not only a *KRAS* mutation (as formerly assumed), but also a co-occurring mutation within the *TP53* gene. The index color image as a result of the fourth RF classifier is shown in Figure 6A in comparison with the corresponding H&E staining of the thin section of lung tissues (Figure 6B). Both mutation classes (*KRAS* and *TP53*) were distributed relatively homogeneously within the tumor lesions. This may indicate that the mutations were not locally confined as heterogeneous mutations, but rather co-occurred homogeneously throughout the tumor tissue.

Three of the remaining five noticeable cases still had sufficient lung tissue to repeat the genetic analysis. In this regard, the combined IR imaging LCM workflow presented by Großerueschkamp et al[28] was used to collect homogeneous tissue samples without prior labeling of the tissue slices. Only areas that were assigned to the incorrect mutation classes were collected. The subsequent performance of the NGS gene panel showed that all previously detected mutations could be confirmed. This led to the conclusion that the incorrectly classified spectra in the three examined patient samples are a false detection of the classifier or may result from further undetected mutations that co-activate both mitogen-activated protein kinase and TP53 signaling.

## Discussion

This study used a label-free and operator-independent approach to identify tumor regions and to determine the

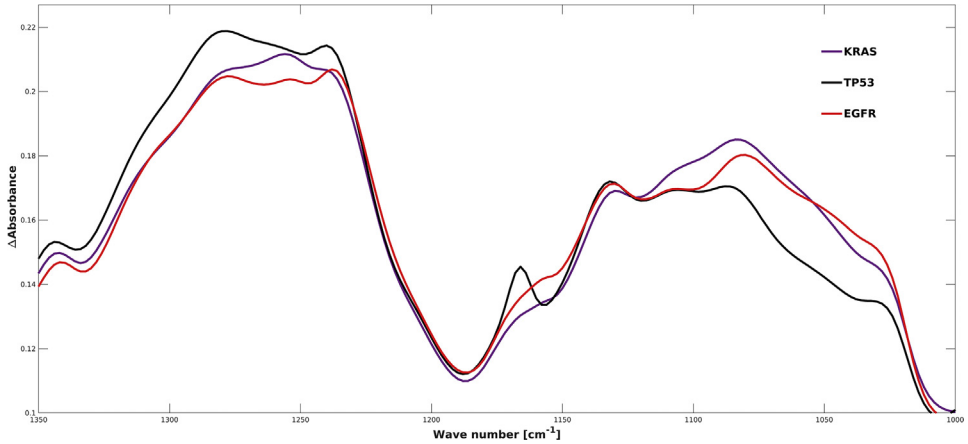



**Figure 4** Mean value infrared spectra (1350 to 1000 cm$^{-1}$) of the classes used to set up the fourth and fifth random forest classifiers to detect mutation status (KRAS, purple; TP53, gray; EGFR, red).

tumor type as well as the mutation status of lung tumor and nontumor tissue samples with high sensitivity and specificity. A QCL-based IR imaging workflow for whole-slice lung tissue sections, and a decrease in the measuring time by 160 times in comparison to the previously used Agilent Cary-FTIR system (Supplemental Figure S15) were used.

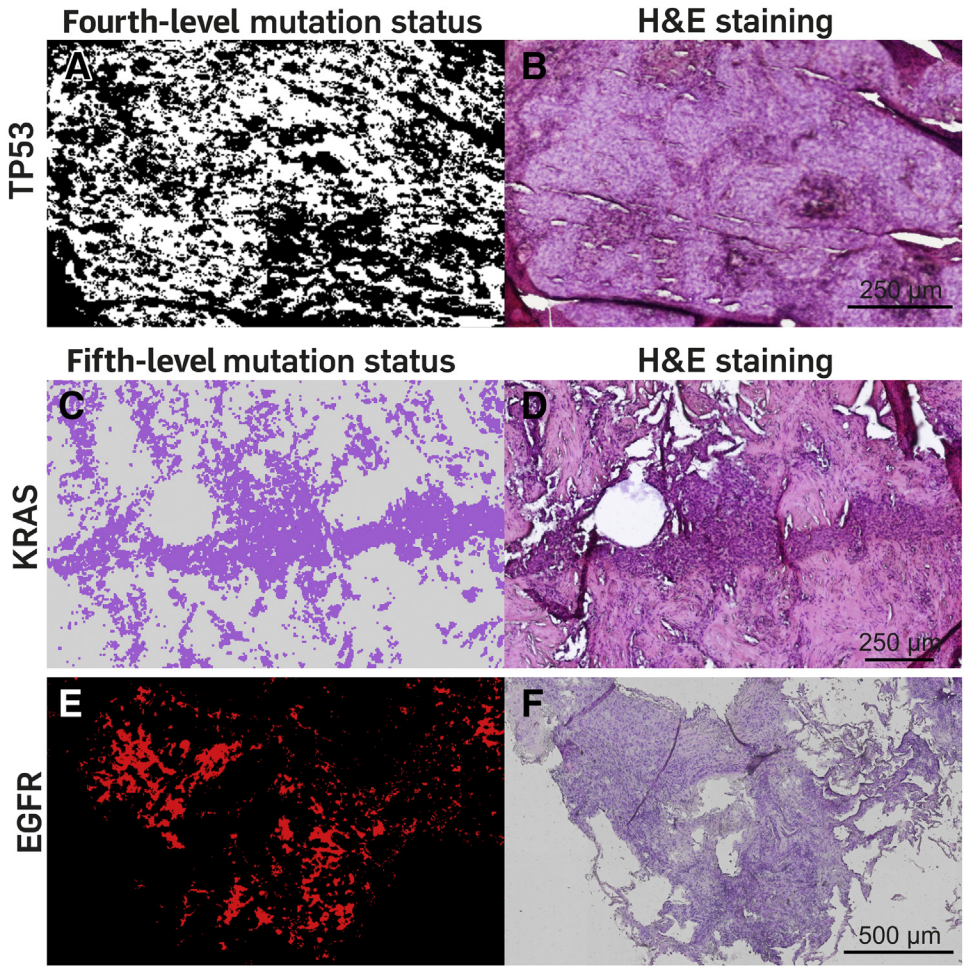



**Figure 5** Quantum cascade laser—based infrared imaging of lung tissues to determine the mutation status of adenocarcinomas. Results of the fourth and fifth random forest classifier on sections of lung adenocarcinoma tissues with *TP53* (**A**), *KRAS* (**C**), and *EGFR* mutation (**E**) and the corresponding hematoxylin and eosin (H&E) staining (**B**, **D**, and **F**) of the sections for comparison. Scale bars: 250 μm (**A**—**D**); 500 μm (**E** and **F**).

The classification of a large tissue sample could be performed in <30 minutes, which is within the same time range required for histopathology. Compared with previous studies,[30,32,34,44] a high number of whole slice samples (578 FFPE tissue samples from 235 patients) could be analyzed because of the high sample throughput. In addition, the compact design of the uncooled microbolometer detector of the Spero-QT can make its routine use in clinical settings feasible. Furthermore, the data processing time could be reduced by performing data correction and classification parallel to data acquisition for each field of view individually (480 × 480 pixels at 427 data points). Therefore, computing can be performed on ordinary personal computers, and no additional and expensive high-performance hardware would be needed.

Another important prerequisite for the clinical translation of QCL-based IR imaging is the validation of the classifiers on an independent data set. This not only includes the use of several instruments at different locations but also different operators and samples from different clinics. This ensures that the algorithm is not fitted to artifacts caused, for instance, by a certain preparation method. In the present study, two Spero-QT instruments were used, and data acquisition was performed by seven different operators. In addition, the measurements were performed at two different locations. The results of the presented RF classification and the determination of tumor type and mutation status of lung tissues are independent of the device, of the operator who performs the measurement, and of the device location. To increase the reliability of these results, a larger number of tissue samples from different clinics would be required. Furthermore, the RF classifier could be replaced by deep learning algorithms. This would add spatial information to the existing spectral information so that morphologic aspects of the tissue could also be included in the classification process. As reported by Schuhmacher et al[45] in regard to analyzing colon cancer samples using a neural network, this is a promising approach for the annotation of tissue samples based on IR spectral data. A further application of deep learning, as demonstrated by several groups, is the evaluation of H&E images. Recently, Kather et al[46] reported the identification of microsatellite stability or instability in colorectal cancer samples based on H&E images using deep residual learning. Weakly supervised multiple instance learning—based deep learning on whole slide images was performed by Lu et al[47] on H&E images of renal cell carcinoma as well as non—small-cell lung cancer and by Campanella et al[48] on basal cell carcinoma and prostate and breast cancer. Therefore, evaluating these modern weakly supervised classifiers on infrared hyperspectral data sets in advanced studies may be a promising approach.

To further reduce the measuring time, the number of recorded wave numbers can be reduced, or only discrete frequencies can be measured.[49] The latter, however, can be problematic with regard to the RMie-EMSC correction, as this requires the complete spectrum.

This study showed for the first time that QCL-based IR imaging can be used not only to identify different tissue types and tumor regions with a sensitivity of 95% and a specificity of 97% compared with histopathology, but also to identify spectral markers that allow differentiation of different molecular states. This illustrates that a single IR measurement can be used to obtain information about a sample that would otherwise require several methods and time-consuming procedures (IHC or NGS). Mayerich et al[50] presented the possibility of mimicking several IHC stains on breast tissue using FTIR imaging. The RF classifier for the determination of mutation status introduced in this study could be verified with a sensitivity and specificity of 95% for adenocarcinoma tissue samples from 60 patients. Only one patient per mutation type (*KRAS*, *EGFR*, or *TP53*) was found to be incorrectly detected compared with the results of the NGS gene panel. In one case where there was heterogeneity in the recognition of the classifier (Figure 6), the

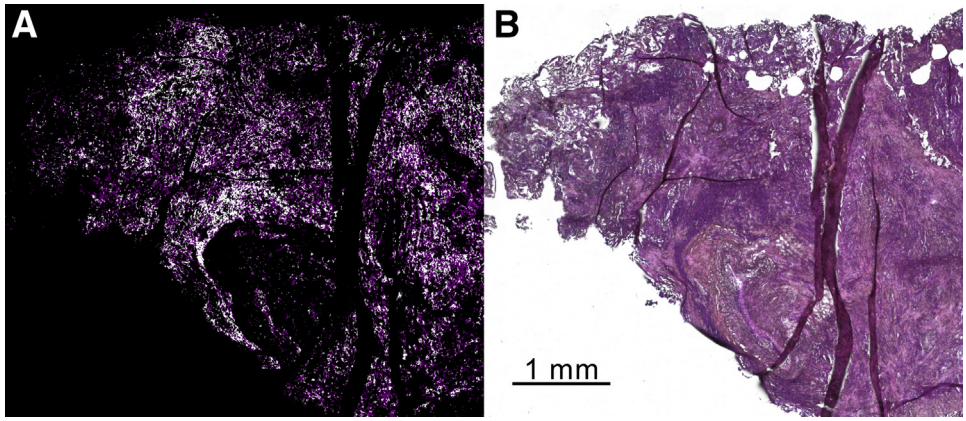



**Figure 6** Quantum cascade laser—based infrared imaging of lung tissue to determine the mutation status of adenocarcinomas. Results of the random forest (RF) on sections of lung adenocarcinoma tissue with *KRAS* and *TP53* mutation (**A**) and the corresponding hematoxylin and eosin staining of the section for comparison (**B**). This section was assumed to contain the *KRAS* mutation. The RF result classified 63.71% of the tumor spectra as *KRAS* mutated and 36.29% as *TP53* mutated. Scale bar = 1 mm (**A** and **B**).

presence of both *KRAS* and *TP53* mutations was confirmed using the gene panel. A different approach for the automated determination of lung tumor types and mutations was shown by Coudray et al,[51] who used a deep learning algorithm on H&E-stained images of a comparable number of patients. Kather et al[52] performed a pan-cancer analysis based on H&E staining using a deep learning algorithm. In contrast to the method presented in this study, these approaches only provide probabilities for each image tile, not a spatially resolved assignment. Furthermore, it is not possible to identify different tissue types or spatially resolve tumor regions or mutation patterns.

In addition, the method introduced with this study facilitates the use of an LCM to isolate precisely defined tissue types from untreated and unstained samples. The regions of individual tissue types can be isolated precisely, the material contains little unwanted (contaminating) tissue, and the amount of material required for further analyses can be reduced significantly. Furthermore, this approach can also be used to examine samples that have been obtained by minimally invasive methods and provide only a small amount of material, such as endobronchial ultrasound-guided transbronchial needle aspiration. These samples can be used not only for genome analyses, as shown in this study, but also for proteomic and transcriptomic studies. Cumulative data on a single sample can contribute to a better understanding of the molecular changes occurring in different lung cancer types and thus improve diagnostic and therapeutic approaches to the disease.

Treatment procedures on FFPE tissues lead to changes within the tissue, and thus, influence the spectral data compared with spectra from fresh-frozen tissue. These differences can be seen in bands, which are caused or influenced by lipids. Therefore, two different classifiers must be trained for FFPE and fresh-frozen tissue. The first three RF levels were built similar to the FFPE tissue classifier. Because of the low number of patients with certain mutations, no classifier could be trained to determine the mutation status. This can be addressed in the future by obtaining more fresh-frozen patient samples.

In summary, this study presents a new application for QCL-based IR imaging by showing that both morphologic and molecular alterations can be detected reproducibly by this automatic and label-free method. To increase the reliability of IR imaging, the next step is to conduct studies with larger patient numbers adapted to clinical needs, which will augment acceptability of this method in the medical community.

## Acknowledgments

We thank Thomas Brüning and Thomas Behrens (Institute for Prevention and Occupational Medicine of the German Social Accident Insurance, Ruhr University Bochum) for management and distribution of the samples; and Catharina Vaerst for great efforts in enrolling patients for this study.

## Author Contributions

N.G. performed the experiments, analyzed spectral data, and wrote the manuscript. F.G. supervised the experiments edited. R.P. and J.F. performed next-generation sequencing analysis edited. R.B. supplied the histologic reports and contributed as clinical pathologist. All authors edited the manuscript. T.B. supervised the management and distribution of the samples. J.S. and Y.K. supervised the clinical study. K.G. was the senior author for this study.

## Supplemental Data

Supplemental material for this article can be found at *http://doi.org/10.1016/j.ajpath.2021.04.013*.

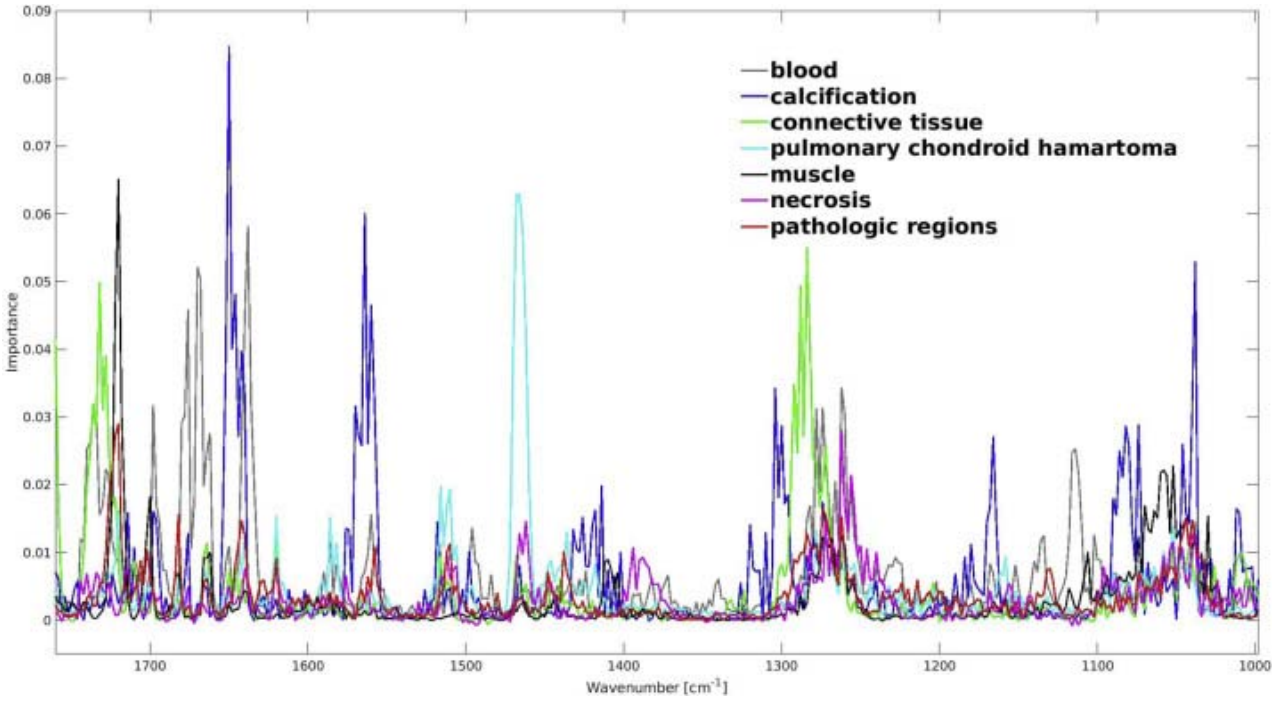
## References

1. Travis WD, Rekhtman N, Riley GJ, Geisinger KR, Asamura H, Brambilla E, Garg K, Hirsch FR, Noguchi M, Powell CA, Rusch VW, Scagliotti G, Yatabe Y: Pathologic diagnosis of advanced lung cancer based on small biopsies and cytology. J Thorac Oncol 2010, 5:411−414
2. Travis WD, Brambilla E, Riely GJ: New pathologic classification of lung cancer: relevance for clinical practice and clinical trials. J Clin Oncol 2013, 31:992−1001
3. Yoshizawa A, Motoi N, Riely GJ, Sima CS, Gerald WL, Kris MG, Park BJ, Rusch VW, Travis WD: Impact of proposed IASLC/ATS/ERS classification of lung adenocarcinoma: prognostic subgroups and implications for further revision of staging based on analysis of 514 stage I cases. Mod Pathol 2011, 24:653−664
4. Loo PS, Thomas SC, Nicolson MC, Fyfe MN, Kerr KM: Subtyping of undifferentiated non-small cell carcinomas in bronchial biopsy specimens. J Thorac Oncol 2010, 5:442−447
5. Roggli VL, Vollmer RT, Greenberg SD, McGavran MH, Spjut HJ, Yesner R: Lung cancer heterogeneity: a blinded and randomized study of 100 consecutive cases. Hum Pathol 1985, 16:569−579
6. Sparrow JM, Ayliffe W, Bron AJ, Brown NP, Hill AR: Inter-observer and intra-observer variability of the Oxford clinical cataract classification and grading system. Int Ophthalmol 1988, 11:151−157
7. Grilley-Olson JE, Hayes DN, Moore DT, Leslie KO, Wilkerson MD, Qaqish BF, Hayward MC, Cabanski CR, Yin X, Socinski MA, Stinchcombe TE, Thorne LB, Allen TC, Banks PM, Beasley MB, Borczuk AC, Cagle PT, Christensen R, Colby TV, Deblois GG, Elmberger G, Graziano P, Hart CF, Jones KD, Maia DM, Miller CR, Nance KV, Travis WD, Funkhouser WK: Validation of interobserver agreement in lung cancer assessment: hematoxylin-eosin diagnostic reproducibility for non-small cell lung cancer: the 2004 World Health Organization classification and therapeutically relevant subsets. Arch Pathol Lab Med 2013, 137:32−40
8. König K, Peifer M, Fassunke J, Ihle MA, Künstlinger H, Heydt C, Stamm K, Ueckeroth F, Vollbrecht C, Bos M, Gardizi M, Scheffler M, Nogova L, Leenders F, Albus K, Meder L, Becker K, Florin A, Rommerscheidt-Fuss U, Altmüller J, Kloth M, Nürnberg P, Henkel T, Bikár S-E, Sos ML, Geese WJ, Strauss L, Ko Y-D, Gerigk U, Odenthal M, Zander T, Wolf J, Merkelbach-Bruse S, Buettner R, Heukamp LC: Implementation of amplicon parallel sequencing leads to improvement of diagnosis and therapy of lung cancer patients. J Thorac Oncol 2015, 10:1049−1057

9. Giaccone G: Epidermal growth factor receptor inhibitors in the treatment of non-small-cell lung cancer. J Clin Oncol 2005, 23: 3235−3242
10. Ostrem JM, Peters U, Sos ML, Wells JA, Shokat KM: K-Ras(G12C) inhibitors allosterically control GTP affinity and effector interactions. Nature 2013, 503:548−551
11. Kron A, Alidousty C, Scheffler M, Merkelbach-Bruse S, Seidel D, Riedel R, Ihle MA, Michels S, Nogova L, Fassunke J, Heydt C, Kron F, Ueckeroth F, Serke M, Krüger S, Grohe C, Koschel D, Benedikter J, Kaminsky B, Schaaf B, Braess J, Sebastian M, Kambartel K-O, Thomas R, Zander T, Schultheis AM, Büttner R, Wolf J: Impact of TP53 mutation status on systemic treatment outcome in ALK-rearranged non-small-cell lung cancer. Ann Oncol 2018, 29:2068−2075
12. Diem M, Mazur A, Lenau K, Schubert J, Bird B, Miljković M, Krafft C, Popp J: Molecular pathology via IR and Raman spectral imaging. J Biophoton 2013, 6:855−886
13. Fernandez DC, Bhargava R, Hewitt SM, Levin IW: Infrared spectroscopic imaging for histopathologic recognition. Nat Biotechnol 2005, 23:469−474
14. Krafft C, Codrich D, Pelizzo G, Sergo V: Raman and FTIR microscopic imaging of colon tissue: a comparative study. J Biophotonics 2008, 1:154−169
15. Pilling MJ, Henderson A, Shanks JH, Brown MD, Clarke NW, Gardner P: Infrared spectral histopathology using haematoxylin and eosin (H&E) stained glass slides: a major step forward towards clinical translation. Analyst 2017, 142:1258−1268
16. Pilling M, Gardner P: Fundamental developments in infrared spectroscopic imaging for biomedical applications. Chem Soc Rev 2016, 45:1935−1957
17. Akalin A, Mu X, Kon MA, Ergin A, Remiszewski SH, Thompson CM, Raz DJ, Diem M, Bird B, Miljković M: Classification of malignant and benign tumors of the lung by infrared spectral histopathology (SHP): laboratory investigation. A J Tech Methods Pathol 2015, 95:406−421
18. Lasch P, Haensch W, Naumann D, Diem M: Imaging of colorectal adenocarcinoma using FT-IR microspectroscopy and cluster analysis. Biochim Biophys Acta 2004, 1688:176−186
19. Hughes C, Iqbal-Wahid J, Brown M, Shanks JH, Eustace A, Denley H, Hoskin PJ, West C, Clarke NW, Gardner P: FTIR microspectroscopy of selected rare diverse sub-variants of carcinoma of the urinary bladder. J Biophoton 2013, 6:73−87
20. Demos SG, Gandour-Edwards R, Ramsamooj R, deVere White R: Spectroscopic detection of bladder cancer using near-infrared imaging techniques. J Biomed Optics 2004, 9:767−771
21. Sablinskas V, Urboniene V, Ceponkus J, Laurinavicius A, Dasevicius D, Jankevicius F, Hendrixson V, Koch E, Steiner G: Infrared spectroscopic imaging of renal tumor tissue. J Biomed Optics 2011, 16:96006
22. Baker MJ, Gazi E, Brown MD, Shanks JH, Clarke NW, Gardner P: Investigating FTIR based histopathology for the diagnosis of prostate cancer. J Biophotonics 2009, 2:104−113
23. Wald N, Goormaghtigh E: Infrared imaging of primary melanomas reveals hints of regional and distant metastases. Analyst 2015, 140: 2144−2155
24. Steiner G, Shaw A, Choo-Smith L-P'i, Abuid MH, Schackert G, Sobottka S, Steller W, Salzer R, Mantsch HH: Distinguishing and grading human gliomas by IR spectroscopy. Biopolymers 2003, 72: 464−471
25. Kuepper C, Großerueschkamp F, Kallenbach-Thieltges A, Mosig A, Tannapfel A, Gerwert K: Label-free classification of colon cancer grading using infrared spectral histopathology. Faraday Discuss 2016, 187:105−118
26. Witzke KE, Großerueschkamp F, Jütte H, Horn M, Roghmann F, Landenberg Nvon, Bracht T, Kallenbach-Thieltges A, Käfferlein H, Brüning T, Schork K, Eisenacher M, Marcus K, Noldus J, Tannapfel A, Sitek B, Gerwert K: Integrated Fourier transform infrared imaging and proteomics for identification of a candidate histochemical biomarker in bladder cancer. Am J Pathol 2019, 189: 619−631
27. Cameron JM, Conn JJA, Rinaldi C, Sala A, Brennan PM, Jenkinson MD, Caldwell H, Cinque G, Syed K, Butler HJ, Hegarty MG, Palmer DS, Baker MJ: Interrogation of IDH1 status in gliomas by Fourier transform infrared spectroscopy. Cancers 2020, 12:3682
28. Großerueschkamp F, Bracht T, Diehl HC, Kuepper C, Ahrens M, Kallenbach-Thieltges A, Mosig A, Eisenacher M, Marcus K, Behrens T, Brüning T, Theegarten D, Sitek B, Gerwert K: Spatial and molecular resolution of diffuse malignant mesothelioma heterogeneity by integrating label-free FTIR imaging, laser capture microdissection and proteomics. Sci Rep 2017, 7:44829
29. Großerueschkamp F, Kallenbach-Thieltges A, Behrens T, Brüning T, Altmayer M, Stamatis G, Theegarten D, Gerwert K: Marker-free automated histopathological annotation of lung tumour subtypes by FTIR imaging. Analyst 2015, 140:2114−2120
30. Bassan P, Weida MJ, Rowlette J, Gardner P: Large scale infrared imaging of tissue micro arrays (TMAs) using a tunable quantum cascade laser (QCL) based microscope. Analyst 2014, 139: 3856−3859
31. Bird B, Rowlette J: A protocol for rapid, label-free histochemical imaging of fibrotic liver. Analyst 2017, 142:1179−1184
32. Bird B, Rowlette J: High definition infrared chemical imaging of colorectal tissue using a Spero QCL microscope. Analyst 2017, 142: 1381−1386
33. Kröger-Lui N, Gretz N, Haase K, Kränzlin B, Neudecker S, Pucci A, Regenscheit A, Schönhals A, Petrich W: Rapid identification of goblet cells in unstained colon thin sections by means of quantum cascade laser-based infrared microspectroscopy. Analyst 2015, 140: 2086−2092
34. Pilling MJ, Henderson A, Gardner P: Quantum cascade laser spectral histopathology: breast cancer diagnostics using high throughput chemical imaging. Anal Chem 2017, 89:7348−7355
35. Kuepper C, Kallenbach-Thieltges A, Juette H, Tannapfel A, Großerueschkamp F, Gerwert K: Quantum cascade laser-based infrared microscopy for label-free and automated cancer classification in tissue sections. Sci Rep 2018, 8:7717
36. Kallenbach-Thieltges A, Großerueschkamp F, Jütte H, Kuepper C, Reinacher-Schick A, Tannapfel A, Gerwert K: Label-free, automated classification of microsatellite status in colorectal cancer by infrared imaging. Sci Rep 2020, 10:10161
37. Bassan P, Kohler A, Martens H, Lee J, Byrne HJ, Dumas P, Gazi E, Brown M, Clarke N, Gardner P: Resonant Mie Scattering (RMieS) correction of infrared spectra from highly scattering biological samples. Analyst 2010, 135:268−277
38. Kallenbach-Thieltges A, Großerüschkamp F, Mosig A, Diem M, Tannapfel A, Gerwert K: Immunohistochemistry, histopathology and infrared spectral histopathology of colon cancer tissue sections. J Biophoton 2013, 6:88−100
39. Byrne HJ, Baranska M, Puppels GJ, Stone N, Wood B, Gough KM, Lasch P, Heraud P, Sulé-Suso J, Sockalingum GD: Spectropathology for the next generation: quo vadis? Analyst 2015, 140:2066−2073
40. Goormaghtigh E: Infrared imaging in histopathology: is a unified approach possible? BSI 2017, 5:325−346
41. Peifer M, Fernández-Cuesta L, Sos ML, George J, Seidel D, Kasper LH, et al: Integrative genome analyses identify key somatic driver mutations of small-cell lung cancer. Nat Genet 2012, 44: 1104−1110
42. Wittersheim M, Heydt C, Hoffmann F, Büttner R: KRAS mutation in papillary fibroelastoma: a true cardiac neoplasm? the journal of pathology. Clin Res 2017, 3:100−104
43. Koch-Institut R, Gesellschaft Der Epidemiologischen Krebsregister In Deutschland E.V.: Krebs in Deutschland 2015/2016. Berlin, Germany: Robert Koch-Institut, 2019

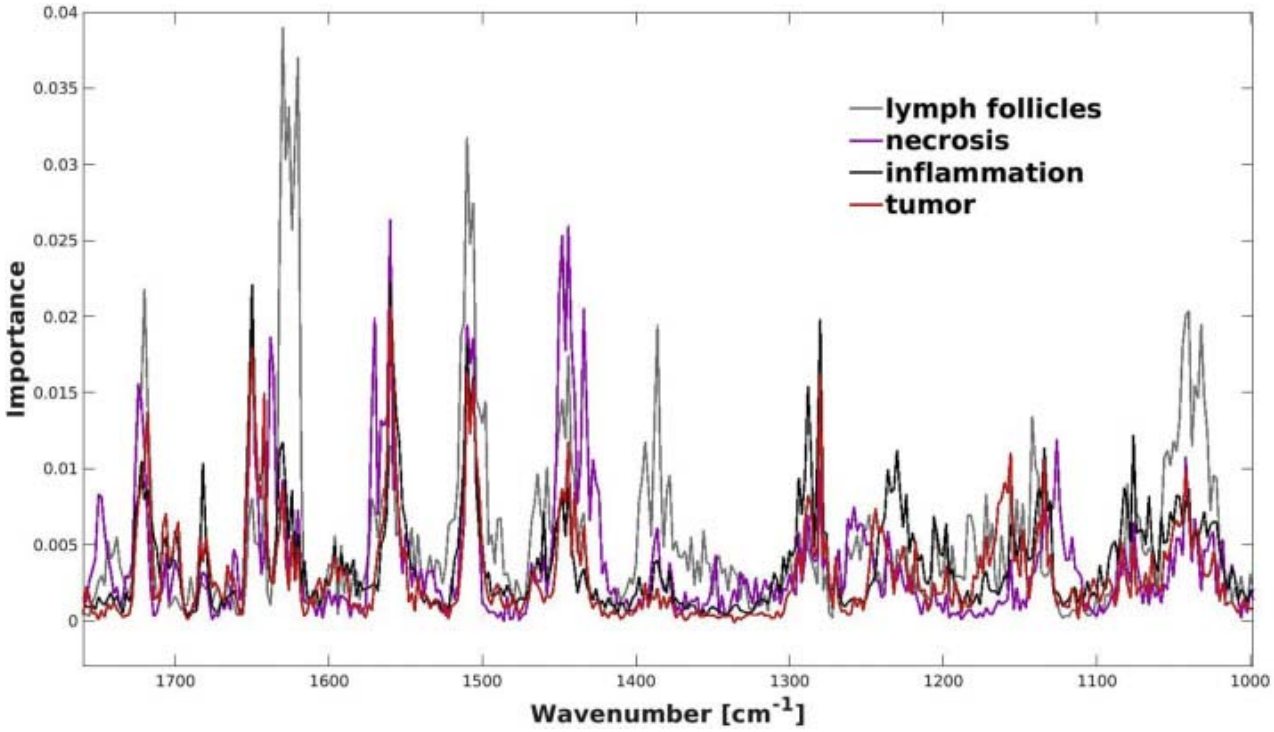
44. Nallala J, Diebold M-D, Gobinet C, Bouché O, Sockalingum GD, Piot O, Manfait M: Infrared spectral histopathology for cancer diagnosis: a novel approach for automated pattern recognition of colon adenocarcinoma. Analyst 2014, 139:4005–4015
45. Schuhmacher D, Gerwert K, Mosig A: A generic neural network approach to infer segmenting classifiers for disease-associated regions in medical image data. medRxiv 2020:20028845
46. Kather JN, Pearson AT, Halama N, Jäger D, Krause J, Loosen SH, Marx A, Boor P, Tacke F, Neumann UP, Grabsch HI, Yoshikawa T, Brenner H, Chang-Claude J, Hoffmeister M, Trautwein C, Luedde T: Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. Nat Med 2019, 25:1054–1056
47. Lu MY, Williamson DFK, Chen TY, Chen RJ, Barbieri M, Mahmood F: Data efficient and weakly supervised computational pathology on whole slide images. Nat Biomed Eng 2020, 5:1–16
48. Campanella G, Hanna MG, Geneslaw L, Miraflor A, Werneck Krauss Silva V, Busam KJ, Brogi E, Reuter VE, Klimstra DS, Fuchs TJ: Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nat Med 2019, 25:1301–1309
49. Yeh K, Lee D, Bhargava R: Multicolor discrete frequency infrared spectroscopic imaging. Anal Chem 2019, 91:2177–2185
50. Mayerich D, Walsh MJ, Kadjacsy-Balla A, Ray PS, Hewitt SM, Bhargava R: Stain-less staining for computed histopathology. Technology 2015, 3:27–31
51. Coudray N, Ocampo PS, Sakellaropoulos T, Narula N, Snuderl M, Fenyö D, Moreira AL, Razavian N, Tsirigos A: Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. Nat Med 2018, 24:1559–1567
52. Kather JN, Heij LR, Grabsch HI, Loeffler C, Echle A, Muti HS, Krause J, Niehues JM, Sommer KAJ, Bankhead P, Kooreman LFS, Schulte JJ, Cipriani NA, Buelow RD, Boor P, Ortiz-Brüchle N, Hanby AM, Speirs V, Kochanny S, Patnaik A, Srisuwananukorn A, Brenner H, Hoffmeister M, van den Brandt PA, Jäger D, Trautwein C, Pearson AT, Luedde T: Pan-cancer image-based detection of clinically actionable genetic alterations. Nat Cancer 2020, 1:789–799

## Supplemental Data



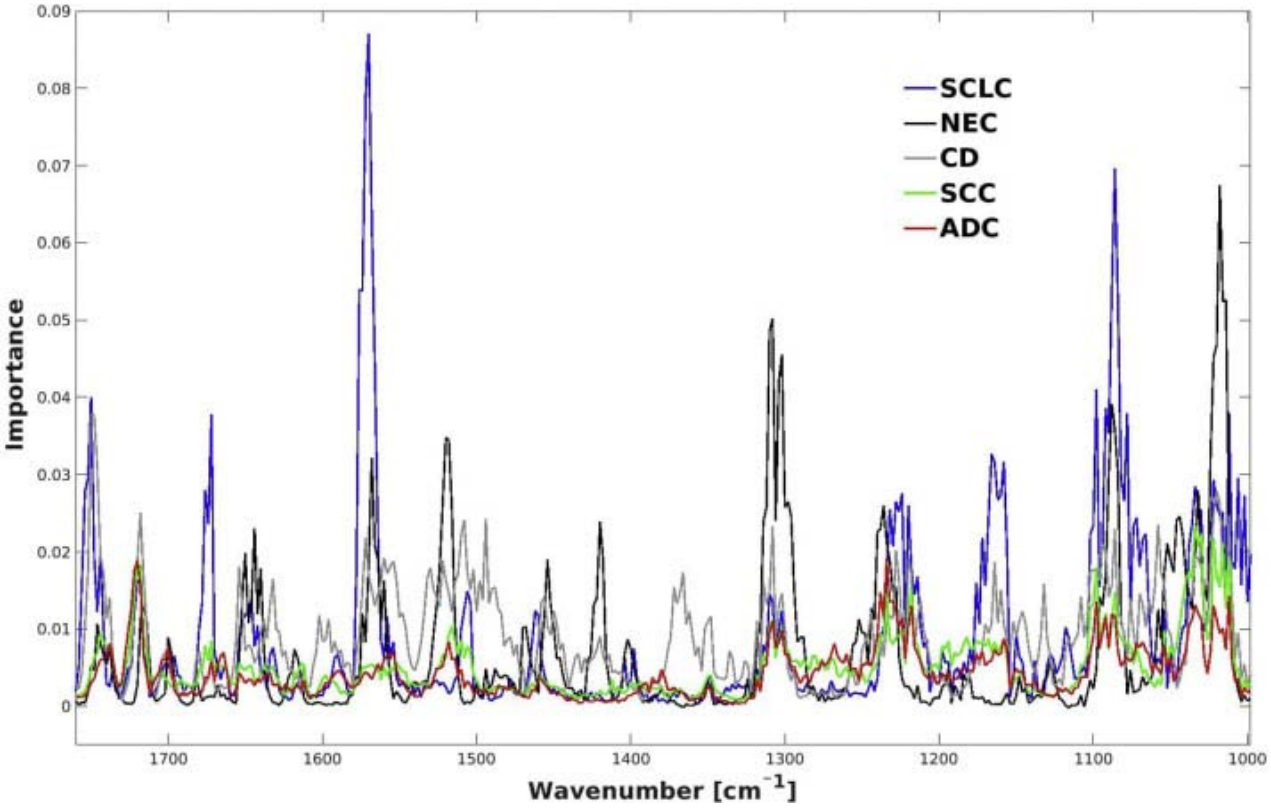

Download : Download high-res image (363KB)  Download : Download full-size image

Supplemental Figure S1. Gini importance plot of the first random forest (RF) classifier for lung tissue differentiation: blood (gray), calcification (dark blue), connective tissue (green), pulmonary chondroid hamartoma (light blue), muscle (black), necrosis (magenta), and pathologic regions (red). For the calcification class, the Gini importance shows the highest values at 1650, 1564, and 1038 $cm^{-1}$. These values are 1676, 1670, and 1638 $cm^{-1}$ for the blood class and 1468 $cm^{-1}$ for the pulmonary chondroid hamartoma class. The importance value at 1720 $cm^{-1}$ is noticeable for the muscle class as well as those at 1760, 1732, and 1284 $cm^{-1}$ for the connective tissue class. The Gini importance values for the necrosis and pathologic region classes, in contrast to the other tissue type classes, indicate that there are no particularly characteristic bands in the spectrum for the detection of these tissue types, so therefore many features that are less frequently used and distributed over the entire fingerprint region are needed. This corresponds with the number of training spectra used for this RF classifier, which are significantly lower for the calcification, pulmonary chondroid hamartoma, and blood classes than especially for the pathologic regions class (Supplemental Table S1).



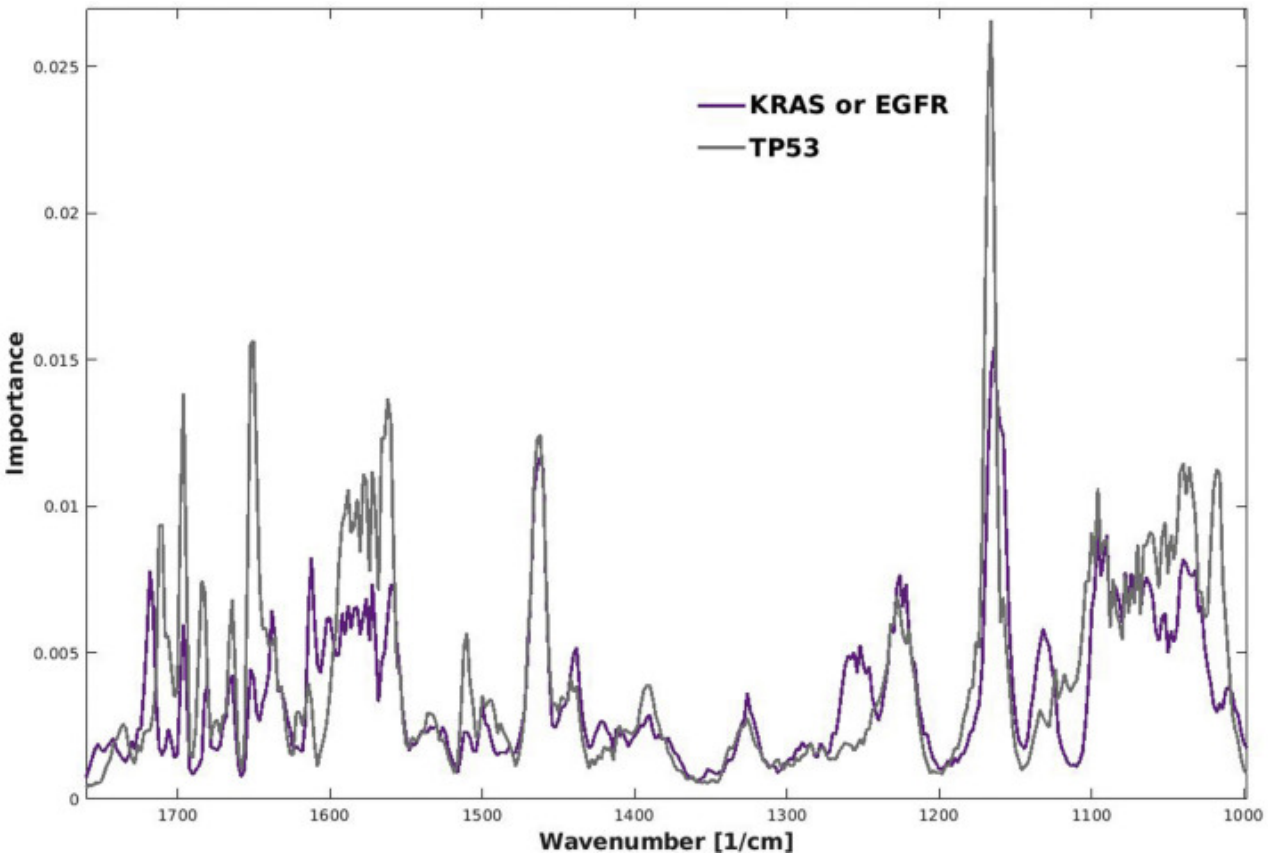

Download : Download high-res image (327KB)  Download : Download full-size image

Supplemental Figure S2. Gini importance plot of the second random forest classifier for tumor identification: lymph follicles (gray), necrosis (magenta), inflammation (black), and tumor (red). The lymphoid follicle class shows the most prominent importance values at 1720, 1630, 1620, 1510, 1386, and 1040 $cm^{-1}$. The necrosis class manifests the highest importance at 1560 as well as around 1448 $cm^{-1}$. The Gini importance plots for the inflammation and tumor classes show many comparatively less prominent features. This also corresponds to the number of spectra necessary for the training of this classifier (Supplemental Table S2).



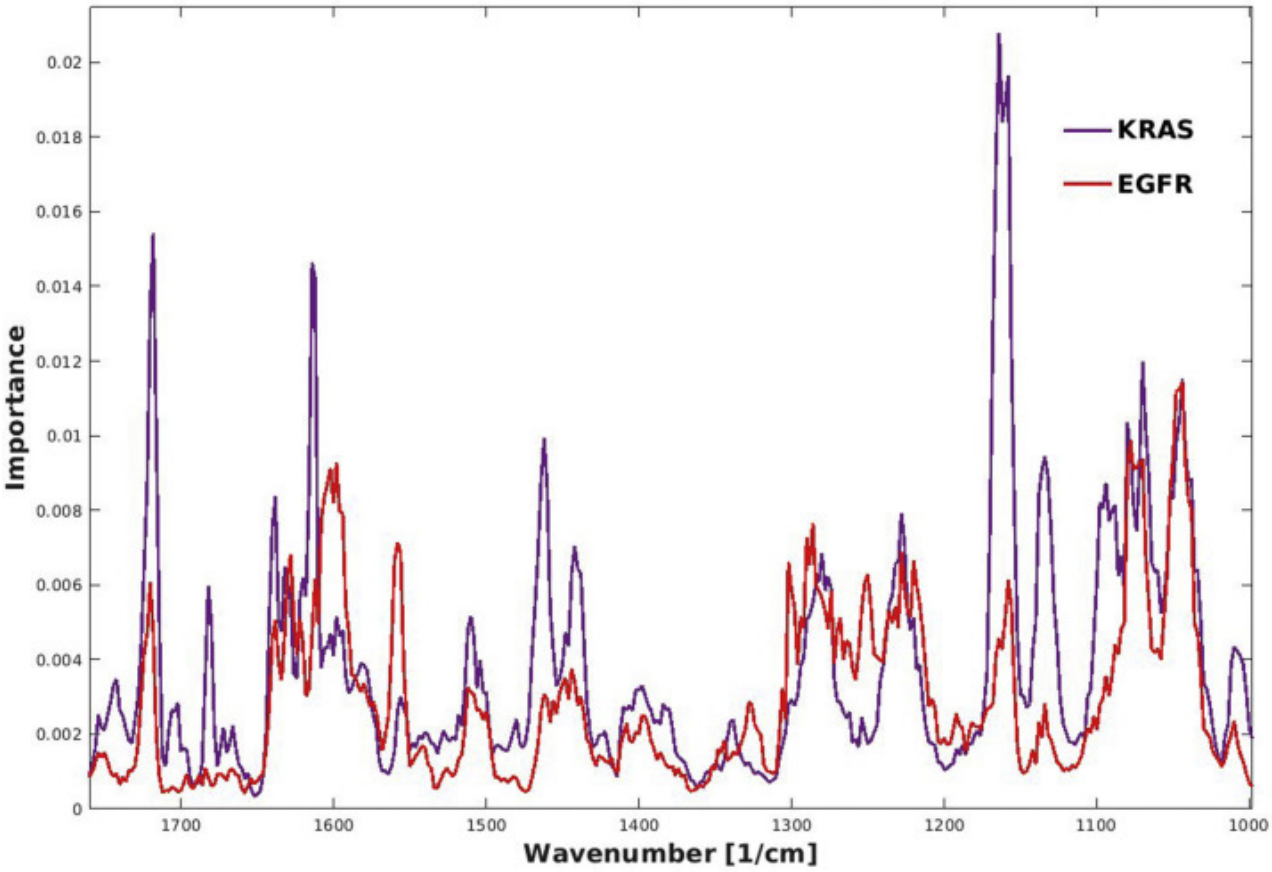

Download : Download high-res image (331KB)  Download : Download full-size image

Supplemental Figure S3. Gini importance plot of the third random forest classifier for lung tumor typing: small-cell lung cancer (SCLC; blue), neuroendocrine carcinoma (NEC; black), carcinoid (CD; gray), squamous cell carcinoma (SCC; green), and adenocarcinoma (ADC; red). The highest importance values show up for the SCLC class at 1570 and 1086 $cm^{-1}$. The most prominent features of the model for the NEC class can be seen at 1308, 1302, and 1018 $cm^{-1}$. For the CD class, the highest importance can be found at 1748 $cm^{-1}$. In contrast, the Gini importance values of the ADC as well as those of the SCC class are significantly lower than the values of the other tumor types. This suggests that the SCLC, NEC, and CD data sets each possess characteristic spectral features within specific wave number ranges that are sufficient for tumor type identification, whereas ADC and SCC detection must be accomplished via less frequently selected features of the entire fingerprint range. This corresponds with the number of training spectra used, which are significantly lower for the SCLC, NEC, and CD classes than for the ADC and SCC classes (Supplemental Table S3).



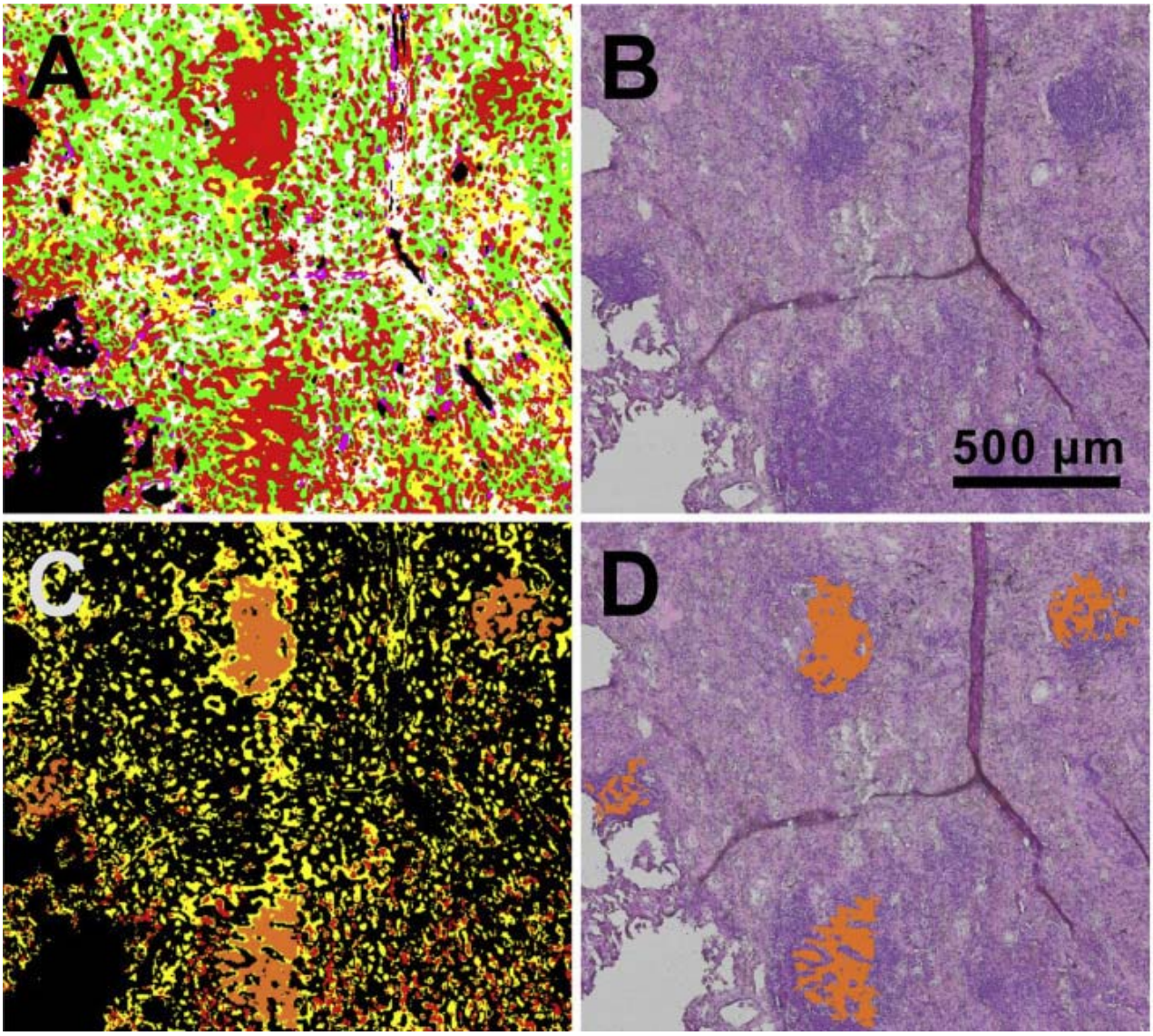

Download : Download high-res image (229KB)  Download : Download full-size image

Supplemental Figure S4. Gini importance plot of the fourth random forest classifier for lung tissue: KRAS or EGFR (purple) and TP53 (gray) mutations. The highest Gini importance is shown for the TP53 class at wave numbers 1696, 1650, and 1160 $cm^{-1}$. The most important features for the detection of the KRAS/EGFR class are found to be at 1440 and 1162 $cm^{-1}$.



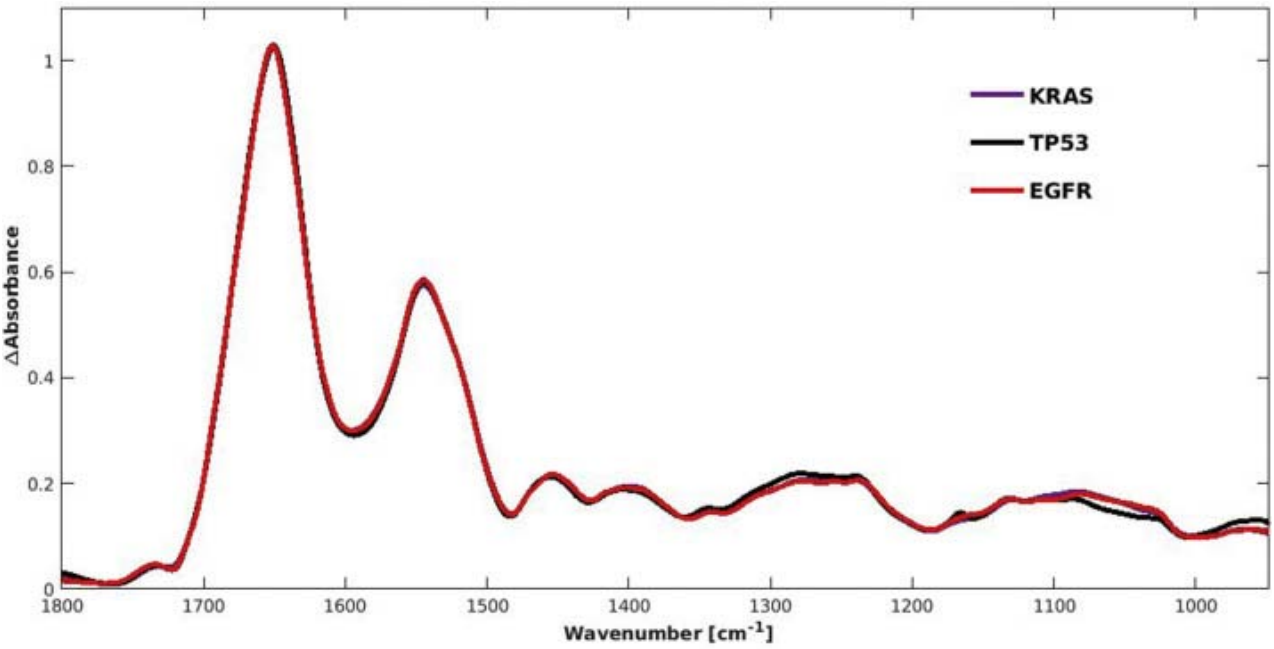

Download : Download high-res image (328KB)  Download : Download full-size image

Supplemental Figure S5. Gini importance plot of the fifth random forest classifier for lung tissue: KRAS (purple) and EGFR (red) mutations. The most prominent features display for the KRAS class at 1718, 1640, 1164, and 1158 $cm^{-1}$. For the EGFR class, the wave numbers of 1598, 1078, and 1044 $cm^{-1}$ show the highest importance values.



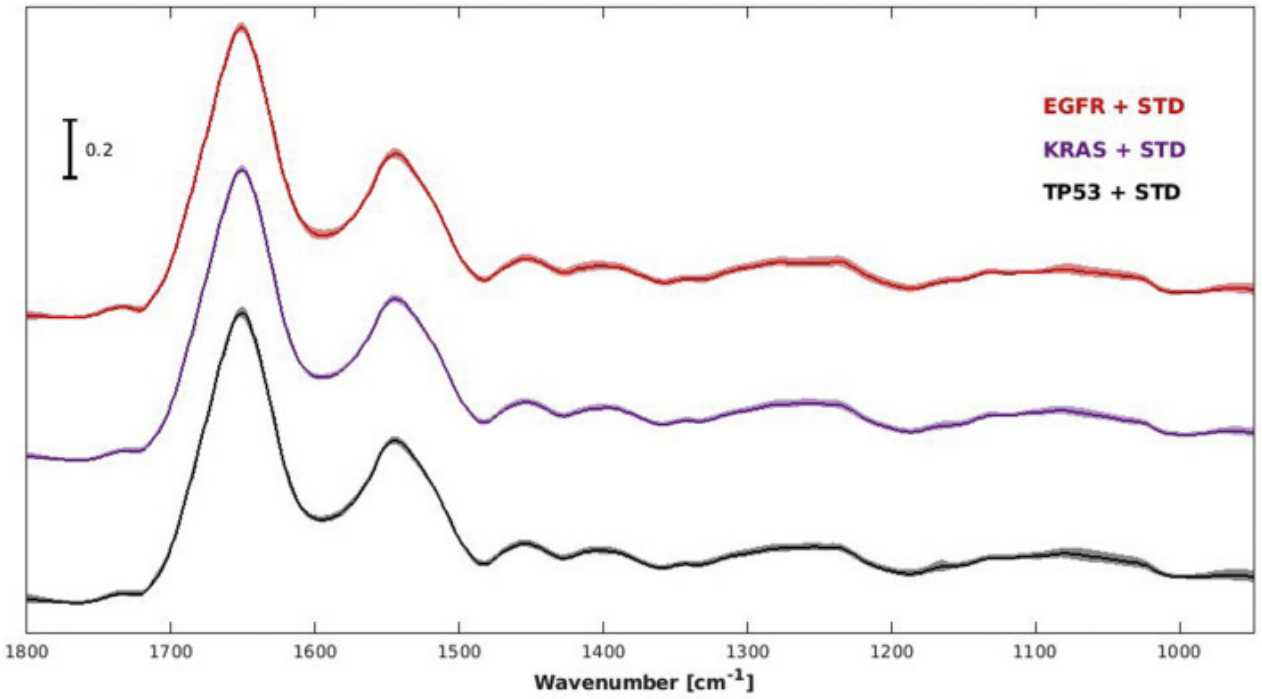

Download : Download high-res image (2MB)  Download : Download full-size image

Supplemental Figure S6. Separation of infiltrated inflammatory cells and lymph follicles using quantum cascade laser–based infrared imaging. **A:** Index color image of the result of the first random forest (RF) classifier to identify pathologic regions. **B:** Hematoxylin and eosin (H&E)–stained tissue. **C:** The second RF classifier for subdivision of the pathologic regions: infiltrated inflammatory cells (yellow), lymph follicles (orange), and tumor (red). **D:** Overlay of H&E staining and as lymph follicle classified pixels. Scale bar = 500 µm (**A–D**).




Download : Download high-res image (147KB) Download : Download full-size image

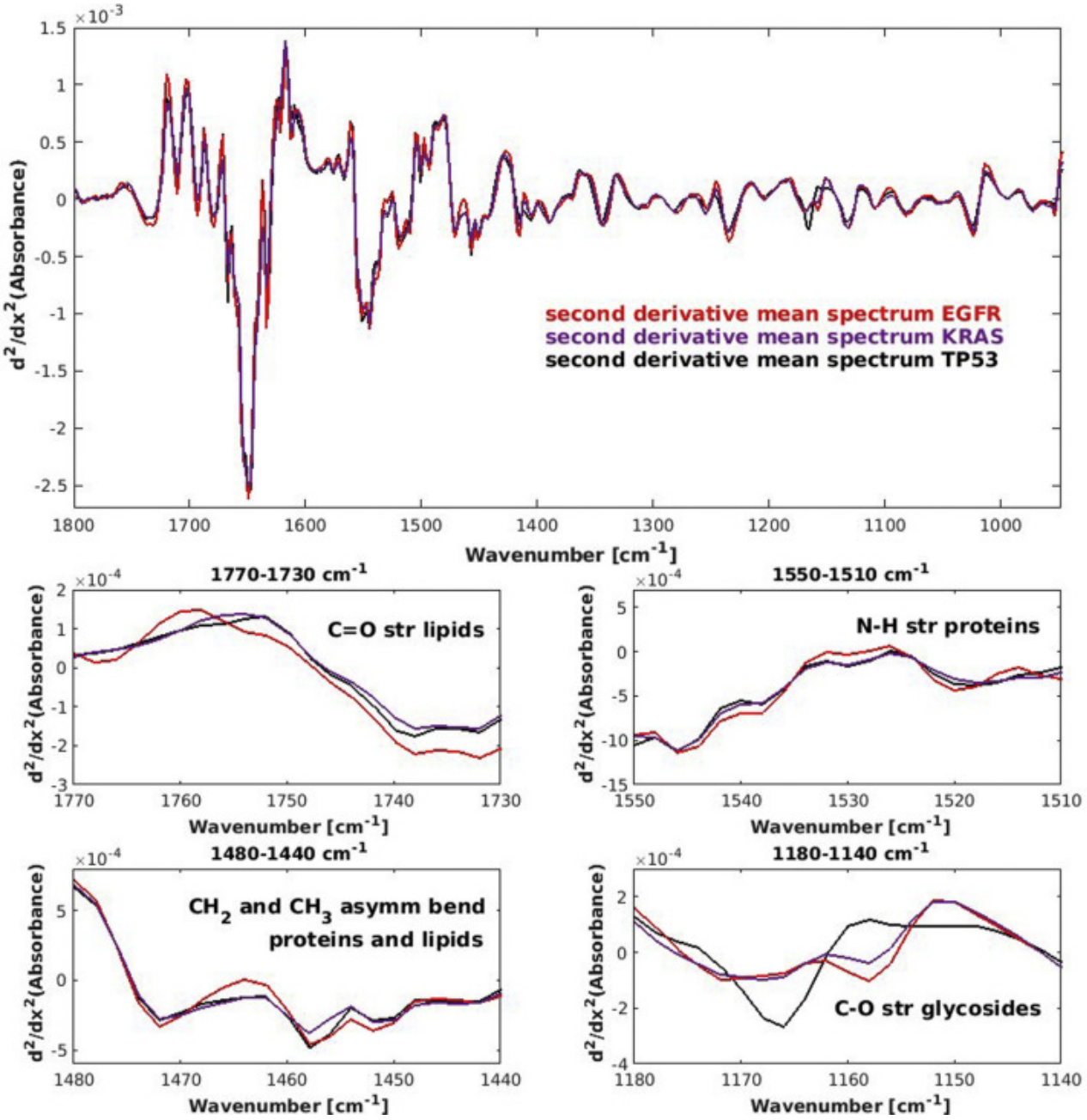
Supplemental Figure S7. Mean value infrared spectra of the classes used to set up the fourth and fifth random forest classifiers to detect mutation status (KRAS, purple; TP53, gray; EGFR, red).




Download : Download high-res image (186KB) Download : Download full-size image

Supplemental Figure S8. Mean value infrared spectra and SD of the classes used to set up

the fourth and fifth random forest classifiers to detect the mutation status of lung adenocarcinoma: KRAS (purple), TP53 (gray), and EGFR mutation (red). The shading in the respective color shows the SD of the classes.

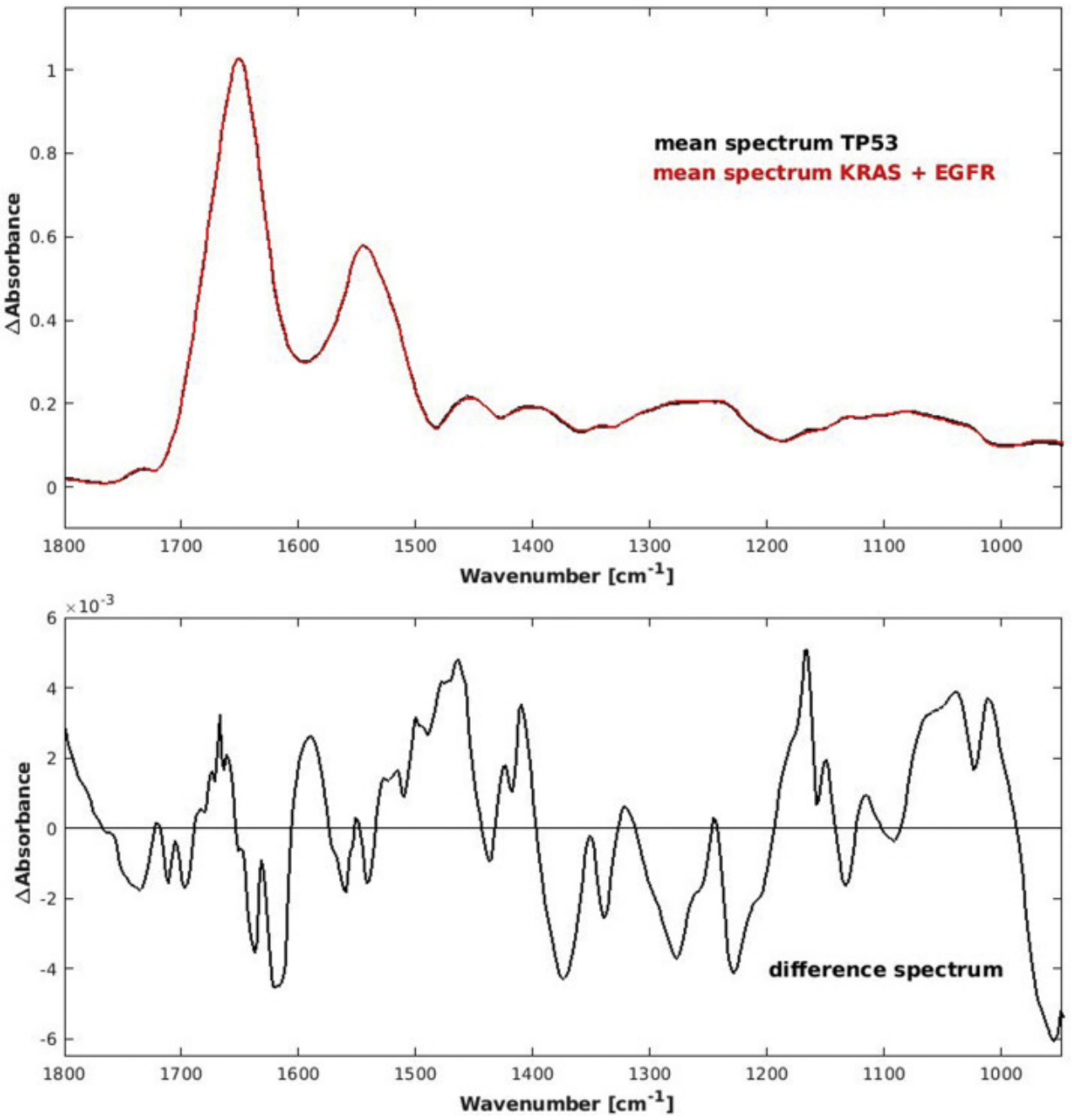



Download : Download high-res image (507KB)

Download : Download full-size image

Supplemental Figure S9. **Top panel:** Second derivative mean value infrared spectra of the mutation random forest classifier training classes: EGFR (red), KRAS (purple), and TP53 mutation (black). **Bottom panels:** Detailed images at specific wave numbers. At 1770 to 1730 $cm^{-1}$, C=O stretching (str) vibration of lipids. At 1550 to 1510 $cm^{-1}$, C-N stretching vibration (amide II band) of the peptide bond and C=C bending vibration of nucleic acids and aromatic amino acids. At 1480 to 1440 $cm^{-1}$, $CH_2$ and $CH_3$ asymmetrical (asymm) bending vibration of proteins and lipids. At 1180 to 1140 $cm^{-1}$, C-O stretching vibration is mainly

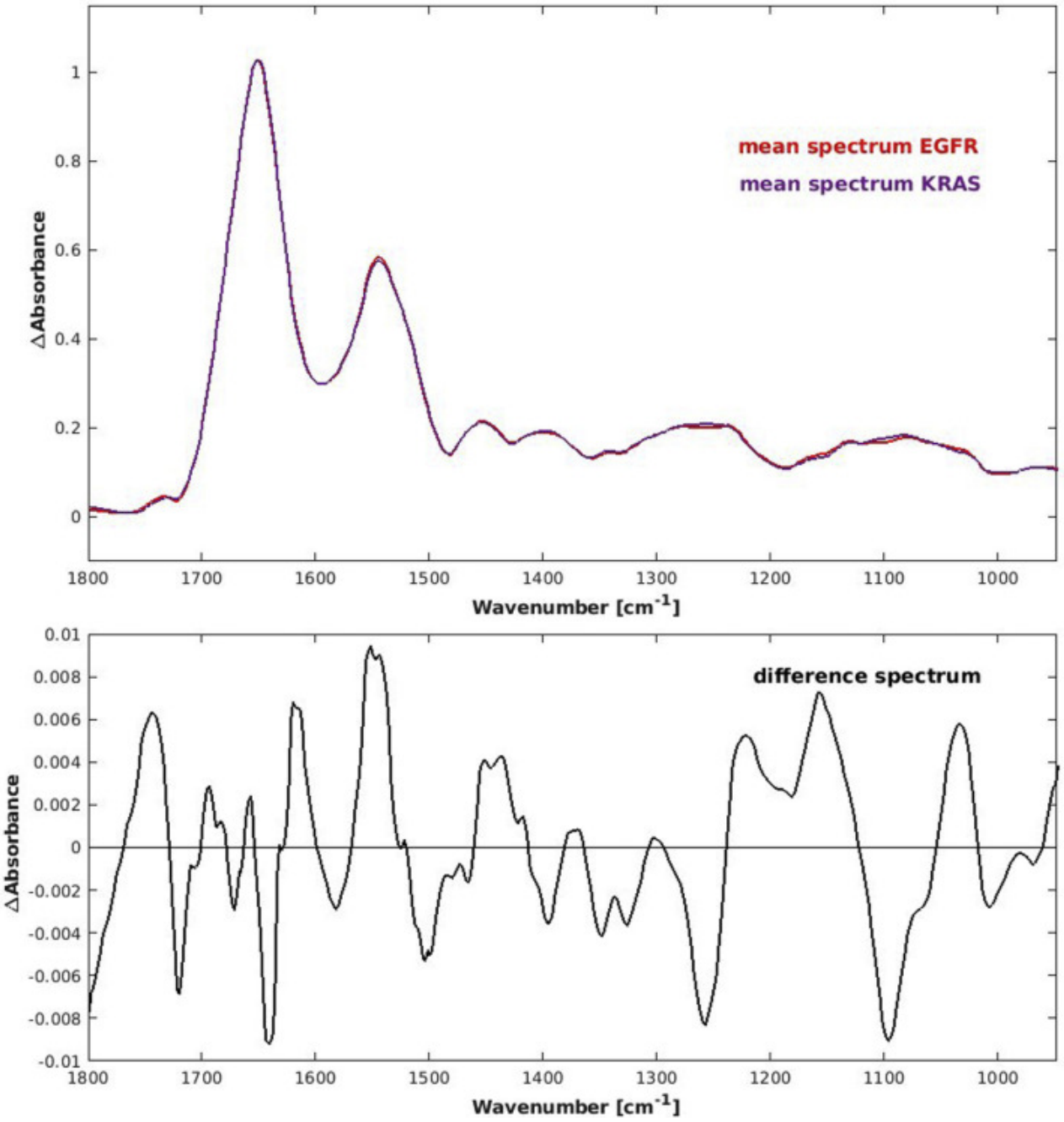
caused by glycosides. The largest differences were found in the range 1180 to 1140 $cm^{-1}$ between the TP53 spectrum and those representing KRAS or EGFR mutations. In addition, small differences between the EGFR spectrum and the other two spectra can be seen at 1760, 1505, and 1450 $cm^{-1}$, among others.




Download : Download high-res image (298KB)  Download : Download full-size image

Supplemental Figure S10. **Top panel:** Mean value infrared (IR) spectra of the classes used to set up the fourth random forest classifier (TP53, gray; EGFR/KRAS, red). **Bottom panel:** Difference spectrum of the mean value IR spectra of the TP53 and EGFR/KRAS classes. The largest differences were observed at 1622 $cm^{-1}$ (amide I band, C=O stretching vibration of the peptide bond), 1464 $cm^{-1}$ ($CH_2$ and $CH_3$ asymmetrical bending vibration in proteins

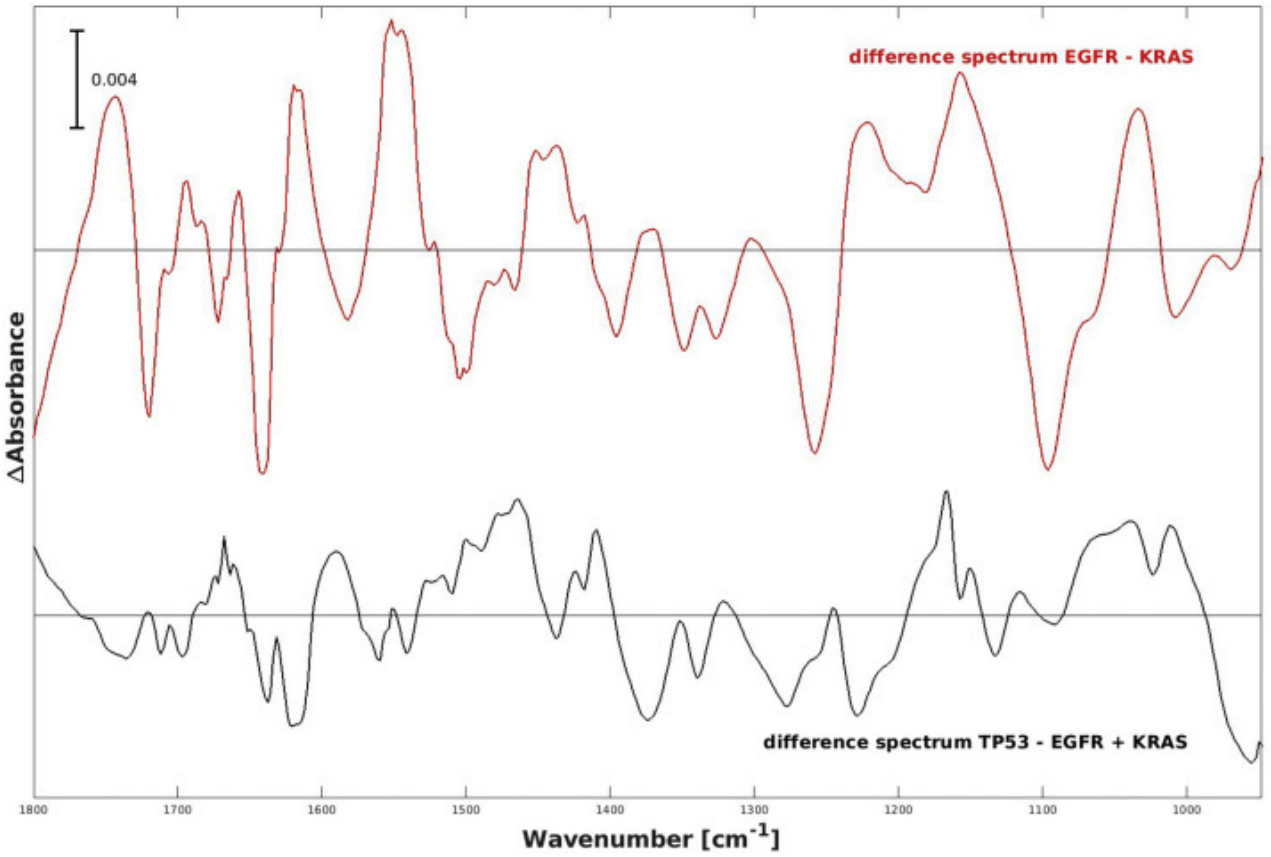
and lipids), 1376 $cm^{-1}$ ($CH_2$ and $CH_3$ symmetrical-bending vibration in proteins and lipids), 1228 $cm^{-1}$ (asymmetrical $PO_2$ stretching vibration of nucleic acids and phospholipids), and 1168 $cm^{-1}$ (C-O stretching vibration of glycosides).




Download : Download high-res image (305KB)  Download : Download full-size image

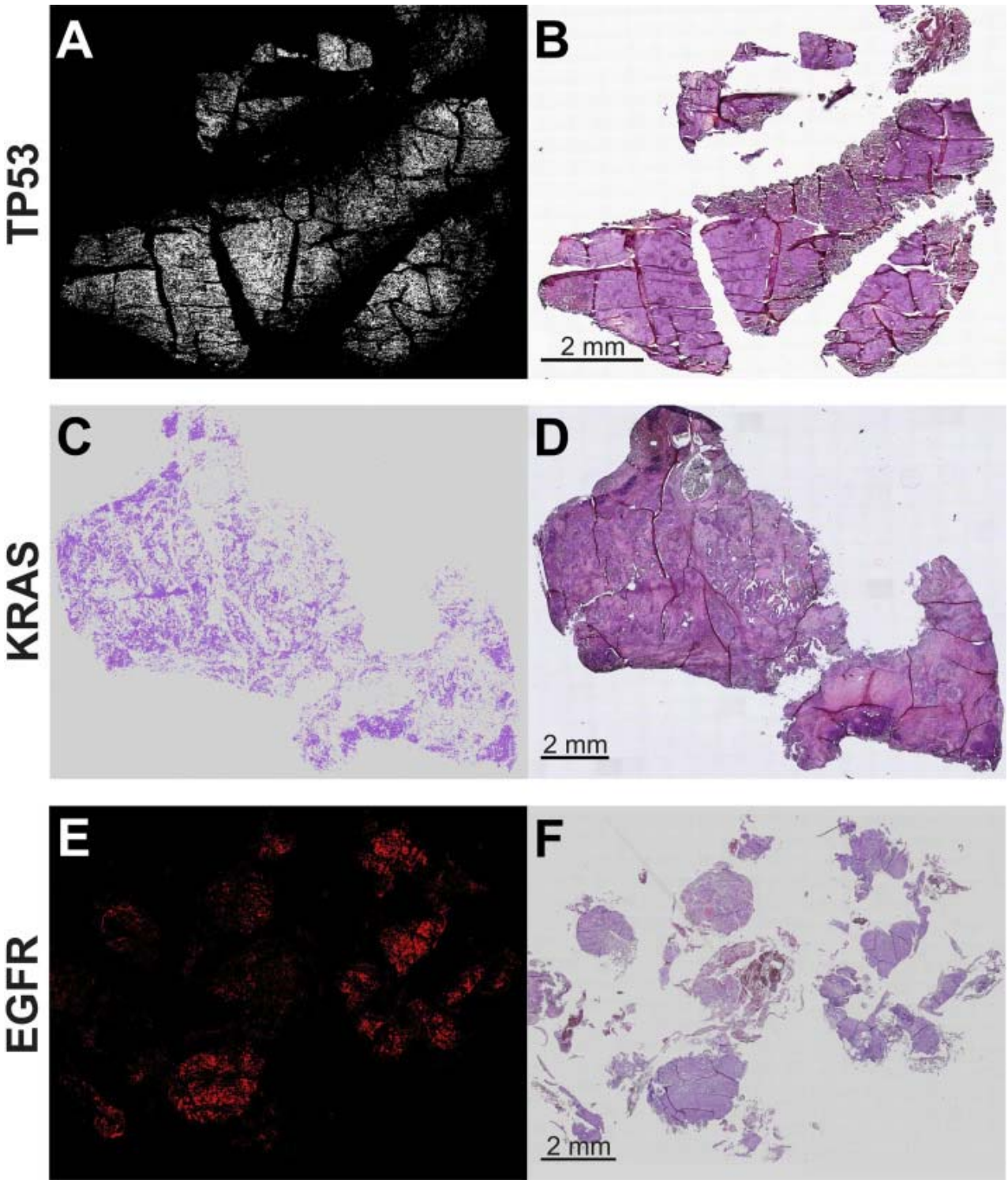
Supplemental Figure S11. **Top panel:** Mean value infrared (IR) spectra of the classes used to set up the fifth random forest classifier (EGFR, red; KRAS, purple). **Bottom panel:** Difference spectrum of the mean value IR spectra of the EGFR and KRAS classes. The largest differences were observed at 1740/1720 $cm^{-1}$ (C=O stretching vibration in lipids), 1642 $cm^{-1}$ (amide I band, C=O stretching vibration of the peptide bond), 1550 $cm^{-1}$ [C-N stretching vibration (amide II band) of the peptide bond], 1258/1096 $cm^{-1}$

(asymmetrical/symmetrical $PO_2$ stretching vibration of nucleic acids and phospholipids), and 1160 $cm^{-1}$ (C-O stretching vibration of glycosides).



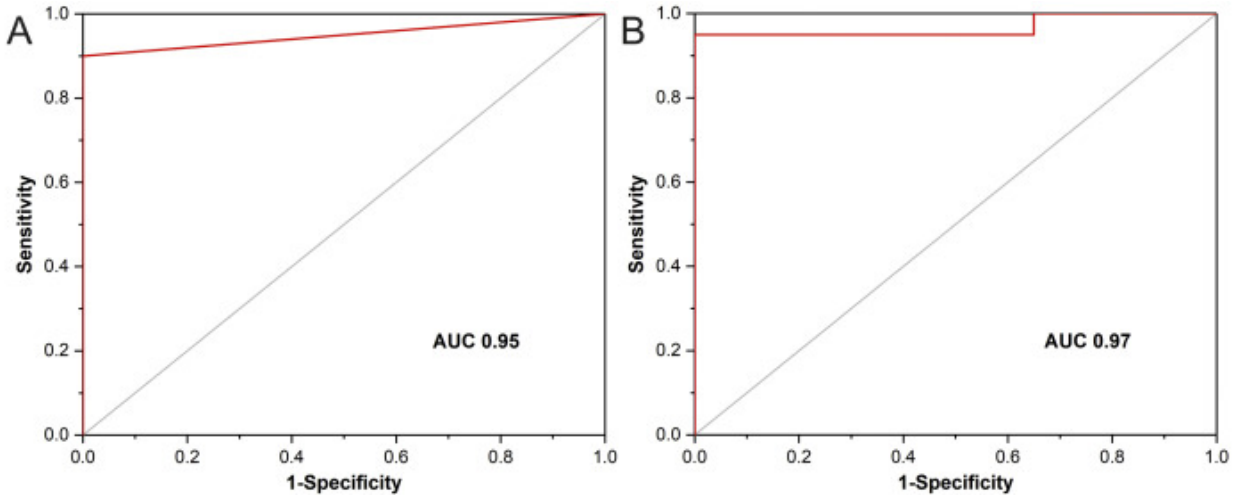

Download : Download high-res image (272KB)  Download : Download full-size image

Supplemental Figure S12. Comparison of the different spectra of the training classes of the fourth (EGFR and KRAS; red; **top panel**) and fifth random forest (RF; TP53 and EGFR/KRAS; black; **bottom panel**). It is noticeable that in the spectra of the fourth RF, there are greater differences in bands caused by aliphatic hydrocarbons (1464 and 1376 $cm^{-1}$) than in those of the fifth RF. In the fifth RF spectra, larger differences can be seen in the ranges around 1740 $cm^{-1}$ (lipids), 1550 $cm^{-1}$ (proteins), and 1096 $cm^{-1}$ (nucleic acids and phospholipids).



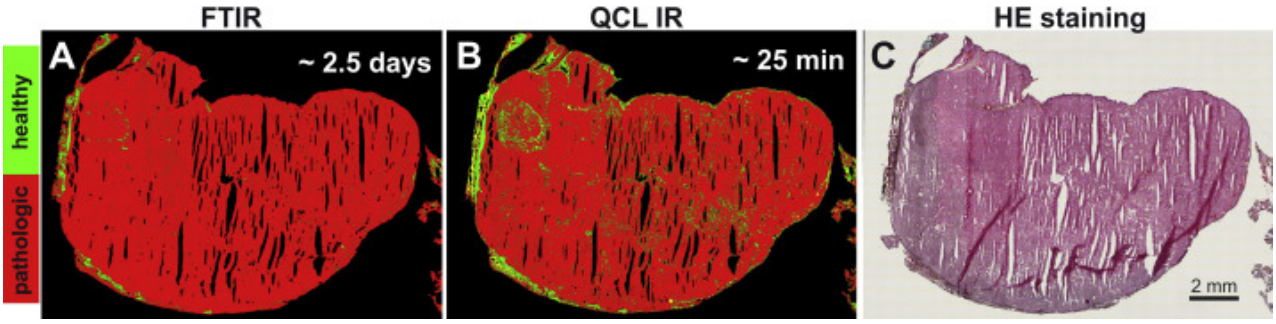

Download : Download high-res image (3MB)  Download : Download full-size image

Supplemental Figure S13. Quantum cascade laser–based infrared imaging of lung tissue to determine the mutation status of adenocarcinomas. Results of the fourth and fifth random forest on whole-slice lung adenocarcinoma tissue with KRAS (**A**), TP53 (**C**), and EGFR mutation (**E**) and the corresponding hematoxylin and eosin staining of the sections for comparison (**B**, **D**, and **F**). Scale bar = 2 mm (**A–F**).




Download : Download high-res image (216KB)  Download : Download full-size image

Supplemental Figure S14. Receiver operating characteristic curves of the fourth (**A**) and fifth (**B**) level random forest classifiers obtained by varying the threshold for TP53 or EGFR mutation positivity. The area under the curve (AUC) was determined to be 0.95 and 0.97 for the verification data set of 60 patients (20 patients per mutation).




Download : Download high-res image (971KB)  Download : Download full-size image

Supplemental Figure S15. Fourier-transform infrared (FTIR); **A**) versus quantum cascade laser (QCL)–based infrared (IR) imaging (**B**) of the same lung cancer tissue section. **C:** Hematoxylin and eosin (H&E)–stained tissue. Pathologic regions are shown in red, and healthy regions are shown in green. The time for data acquisition was approximately 2.5 days for the FTIR, but only 25 minutes for the QCL-based IR instrument. Scale bar = 2 mm (**A–C**).

Supplementary Table 1. Number of classes and spectra per class used to build up the 1st level RF for tissue type identification.

| | classes | Patho -logic | connective tissue | calcifi- cation | blood | necrosis | muscle | H | spectra total |
|---|---|---|---|---|---|---|---|---|---|
| **1st level RF** | 7 | 3808 | 1030 | 468 | 767 | 1009 | 3356 | 501 | 10939 |

Supplementary Table 2. Number of classes and spectra per class used to build up the RF classifier for tumor identification (2nd level).

| | classes | tumor | necrosis | inflammation | lymph follicles | spectra total |
|---|---|---|---|---|---|---|
| **2nd level RF** | 4 | 4822 | 3048 | 10542 | 3034 | 21446 |

Supplementary Table 3. Number of classes and spectra per class used to build up the RF classifier for tumor typing (3rd level).

| | classes | ADC | SCC | SCLC | CD | NEC | spectra total |
|---|---|---|---|---|---|---|---|
| **3rd level RF** | 5 | 7332 | 8220 | 4366 | 4657 | 4492 | 29067 |

Supplementary Table 4. Number of classes and spectra per class used to build up the RF classifier for mutation status analysis ($4^{th}$ and $5^{th}$ level).

| | classes | TP53 | KRAS/EGFR | spectra total |
|---|---|---|---|---|
| **$4^{th}$ level RF** | 2 | 5375 | 5715 | 11090 |
| | | | | |
| | **classes** | **EGFR** | **KRAS** | **spectra total** |
| **$5^{th}$ level RF** | 2 | 6104 | 7135 | 13239 |

Supplementary Table 5. Mutations were determined by sequencing using a MiSeq reagent kit V2 (Illumina, San Diego, CA, USA). Accesion numbers see www.ncbi.nlm.nih.gov/nuccore

| Gene | NM_Number | Exon |
|---|---|---|
| *AKT1* | NM_001014431 | 4 |
| *ALK* | NM_004304 | 21 - 25 |
| *BRAF* | NM_004333 | 11, 15 |
| *CTNNB1* | NM_001904 | 3 |
| *DDR2* | NM_006182 | 3 - 18 |
| *EGFR* | NM_005228 | 18, 19, 21 |
| *EGFR* | NM_005228 | 20 |
| *ERBB2* | NM_004448 | 19, 20 |
| *KRAS* | NM_033360 | 2, 3 |
| *MAP2K1* | NM_002755 | 2 |
| *MET* | NM_001127500 | 14 |
| *NRAS* | NM_002524 | 2, 3 |
| *PIK3CA* | NM_006218 | 10, 21 |
| *PTEN* | NM_000314 | 1 - 8 |
| *TP53* | NM_000546 | 5 - 8 |

Supplementary Table 6. Regions analyzed with the NextSeq 500 Mid Output Kit v2 (Illumina, San Diego, CA, USA). Accesion numbers see www.ncbi.nlm.nih.gov/nuccore

| Gene | NM_Number | Exon |
|---|---|---|
| *ALK* | NM_004304 | 22 - 25 |
| *BRAF* | NM_004333 | 11, 15 |
| *CTNNB1* | NM_001904 | 3 |
| *EGFR* | NM_005228 | 18 - 21 |
| *ERBB2* | NM_004448 | 8, 19, 20 |
| *FGFR1* | NM_023110 | 4 - 7, 10, 12 - 15 |
| *FGFR2* | NM_000141 | 6 - 15, 18 |
| *FGFR2* | NM_022970 | 8 |
| *FGFR3* | NM_000142 | 3, 7, 9, 10 (Codon 429 - 471), 12 (Codon 512-529), 14, 16, 18 (Codon 769 - 807) |
| *FGFR4* | NM_213647 | 3, 6, 9, 12, 13 (Codon 556-607), 15, 16 |
| *IDH1* | NM_005896 | 4 |
| *IDH2* | NM_002168 | 4 |
| *KRAS* | NM_033360 | 2 - 4 |
| *MAP2K1* | NM_002755 | 2, 3 |
| *MET* | NM_001127500 | 14, 16 - 19 |
| *NRAS* | NM_002524 | 2 - 4 |
| *PIK3CA* | NM_006218 | 10, 21 |
| *PTEN* | NM_000314 | 1 - 8 |
| *ROS1* | NM_002944 | 34 - 41 |
| *TP53* | NM_000546 | 4 (Codon 97 - 125), 5, 6, 7, 8 |

Supplementary Table 7 Verification of the mutation status analysis. Results of NGS gene panel and RF classifiers. Three patient samples were classified incorrectly (see bold entries).

| **Patient No.** | **Gene panel** | **4th RF *TP53* [%]** | **4th RF *KRAS/EGFR* [%]** | **5th RF *KRAS* [%]** | **5th RF EGFR [%]** | **RF result** |
|---|---|---|---|---|---|---|
| V01 | *KRAS* | 37.62 | 62.38 | 51.42 | 48.59 | *KRAS* |
| V02 | *KRAS* | 38.78 | 61.22 | 80.31 | 19.68 | *KRAS* |
| V03 | *KRAS* | 25.63 | 74.37 | 57.65 | 42.35 | *KRAS* |
| V04 | *KRAS* and *TP53* | 36.29 | 63.71 | 50.43 | 49.56 | *KRAS* |
| V05 | *KRAS* | 23.6 | 76.4 | 60.07 | 39.92 | *KRAS* |
| V06 | *KRAS* | 26.47 | 73.53 | 54.87 | 45.11 | *KRAS* |
| V07 | *KRAS* | 32.27 | 67.73 | 61.27 | 38.75 | *KRAS* |
| V08 | *KRAS* | 32.29 | 67.71 | 70.65 | 29.34 | *KRAS* |
| V09 | *KRAS* | 35.26 | 64.74 | 70.29 | 29.72 | *KRAS* |
| V10 | *KRAS* | 29.88 | 70.12 | 49.41 | 50.59 | ***EGFR*** |
| V11 | *KRAS* | 23.15 | 76.85 | 51.77 | 48.23 | *KRAS* |
| V12 | *KRAS* | 33.97 | 66.03 | 74.97 | 25.02 | *KRAS* |
| V13 | *KRAS* | 25.3 | 74.7 | 70.54 | 29.46 | *KRAS* |
| V14 | *KRAS* | 14.61 | 85.39 | 51.93 | 48.06 | *KRAS* |
| V15 | *KRAS* | 25.97 | 74.03 | 77.12 | 22.86 | *KRAS* |
| V16 | *KRAS* | 33.22 | 66.78 | 82.64 | 17.35 | *KRAS* |
| V17 | *KRAS* | 26.51 | 73.49 | 62.62 | 37.39 | *KRAS* |
| V18 | *KRAS* | 24.1 | 75.9 | 58.79 | 41.2 | *KRAS* |
| V19 | *KRAS* | 29.14 | 70.86 | 55.07 | 44.92 | *KRAS* |
| V20 | *KRAS* | 26.53 | 73.47 | 62.59 | 37.4 | *KRAS* |
| V21 | *EGFR* | 33.37 | 66.63 | 69.3 | 30.69 | ***KRAS*** |

| **Patient No.** | **Gene panel** | **4th RF *TP53* [%]** | **4th RF *KRAS/EGFR* [%]** | **5th RF *KRAS* [%]** | **5th RF EGFR [%]** | **RF result** |
|---|---|---|---|---|---|---|
| V22 | *EGFR* | 25.58 | 74.42 | 35.49 | 64.52 | *EGFR* |
| V23 | *EGFR* | 13.43 | 86.57 | 17.31 | 82.71 | *EGFR* |
| V24 | *EGFR* | 24.61 | 75.39 | 37.47 | 62.51 | *EGFR* |
| V25 | *EGFR* | 28.48 | 71.52 | 35.92 | 64.08 | *EGFR* |
| V26 | *EGFR* | 23.44 | 76.56 | 47.68 | 52.32 | *EGFR* |
| V27 | *EGFR* | 27.14 | 72.86 | 47.34 | 52.65 | *EGFR* |
| V28 | *EGFR* | 26.75 | 73.25 | 32.21 | 67.8 | *EGFR* |
| V29 | *EGFR* | 29.24 | 70.76 | 33.43 | 66.56 | *EGFR* |
| V30 | *EGFR* | 31.88 | 68.12 | 39.75 | 60.23 | *EGFR* |
| V31 | *EGFR* | 26.73 | 73.27 | 23.18 | 76.82 | *EGFR* |
| V32 | *EGFR* | 20.92 | 79.08 | 42.72 | 57.28 | *EGFR* |
| V33 | *EGFR* | 26.5 | 73.5 | 47.33 | 52.67 | *EGFR* |
| V34 | *EGFR* | 31.92 | 68.08 | 39.92 | 60.07 | *EGFR* |
| V35 | *EGFR* | 22.5 | 77.5 | 36.04 | 63.96 | *EGFR* |
| V36 | *EGFR* | 26.94 | 73.06 | 35.36 | 64.65 | *EGFR* |
| V37 | *EGFR* | 16.35 | 83.65 | 26.66 | 73.34 | *EGFR* |
| V38 | *EGFR* | 40.63 | 59.37 | 46.88 | 53.12 | *EGFR* |
| V39 | *EGFR* | 30.93 | 69.07 | 19.28 | 80.72 | *EGFR* |
| V40 | *EGFR* | 29.07 | 70.93 | 30.73 | 69.28 | *EGFR* |
| V41 | *TP53* | 74.76 | 25.24 | | | *TP53* |
| V42 | *TP53* | 80.12 | 19.88 | | | *TP53* |
| V43 | *TP53* | 80.80 | 19.20 | | | *TP53* |
| V44 | *TP53* | 60.50 | 39.50 | | | *TP53* |

| **Patient No.** | **Gene panel** | **4th RF *TP53* [%]** | **4th RF *KRAS/EGFR* [%]** | **5th RF *KRAS* [%]** | **5th RF EGFR [%]** | **RF result** |
|---|---|---|---|---|---|---|
| V45 | *TP53* | 73.98 | 26.02 | | | *TP53* |
| V46 | *TP53* | 58.17 | 41.83 | | | *TP53* |
| V47 | *TP53* | 94.16 | 5.84 | | | *TP53* |
| V48 | *TP53* | 67.61 | 32.39 | | | *TP53* |
| V49 | *TP53* | 75.48 | 24.52 | | | *TP53* |
| V50 | *TP53* | 74.83 | 25.17 | | | *TP53* |
| V51 | *TP53* | 36.47 | 63.53 | | | ***EGFR/KRAS*** |
| V52 | *TP53* | 65.77 | 34.23 | | | *TP53* |
| V53 | *TP53* | 70.28 | 29.72 | | | *TP53* |
| V54 | *TP53* | 69.77 | 30.23 | | | *TP53* |
| V55 | *TP53* | 54.59 | 45.41 | | | *TP53* |
| V56 | *TP53* | 74.63 | 25.37 | | | *TP53* |
| V57 | *TP53* | 57.77 | 42.23 | | | *TP53* |
| V58 | *TP53* | 67.78 | 32.22 | | | *TP53* |
| V59 | *TP53* | 64.20 | 35.80 | | | *TP53* |
| V60 | *TP53* | 71.27 | 28.73 | | | *TP53* |