The Journal of Pathology: Clinical Research
J Pathol Clin Res July 2026; **12:** e70110
Published online in Wiley Online Library
(wileyonlinelibrary.com). DOI: 10.1002/2056-4538.70110

ORIGINAL ARTICLE

Clinical and molecular correlates of digitally quantified anthracosis in lung tissue slides: utility of an automated quantitative workflow

Lucas C Fung[1,2,3,4], Brenden Joseph[2], Kenneth Ofori-Amanfo[3,4], Theodora Lo[3,4], Peiyao Wang[5], Mark Trinder[2], Georgina Barnabas[2], Taysia Nikaido-Landry[3,4], Clinton H Durney[6], Aadetri Tawara[6], Sharon Ong[1], Stephen Lam[5], Rafael Meza[6,7,8], Renelle Myers[5,9], William W Lockwood[5], Anna McGuire[1], Julia R Naso[2,10*] and Emilia L Lim[3,4*]

[1]*Department of Surgery, Division of Thoracic Surgery, University of British Columbia, Vancouver, Canada*
[2]*Department of Pathology and Laboratory Medicine, Vancouver General Hospital, Vancouver, Canada*
[3]*Department of Biochemistry and Molecular Biology, The University of British Columbia, Vancouver, Canada*
[4]*Edwin SH Leong Centre for Healthy Aging, The University of British Columbia, Vancouver, Canada*
[5]*Department of Basic and Translational Science, BC Cancer Research Institute, Vancouver, Canada*
[6]*Department of Population Health Sciences, BC Cancer Research Institute, Vancouver, Canada*
[7]*School of Population and Public Health, The University of British Columbia, Vancouver, Canada*
[8]*Rutgers Cancer Institute, Rutgers University, New Brunswick, NJ, USA*
[9]*Department of Medicine, Division of Respirology, The University of British Columbia, Vancouver, Canada*
[10]*Department of Pathology and Laboratory Medicine, The University of British Columbia, Vancouver, Canada*

**Correspondence to: Emilia L Lim, Department of Biochemistry and Molecular Biology, Vancouver, British Columbia, Canada. E-mail: emilia.lim@ubc.ca; Julia R Naso, Vancouver General Hospital, Vancouver, British Columbia, Canada. E-mail: julia.naso@vch.ca*

Abstract

Air pollution, such as particulate matter with a diameter of ≤2.5 μm ($PM_{2.5}$) is a major contributor to lung cancer in the never-smoking population. Anthracotic pigments (black deposits in the lungs) are physical evidence of environmental exposures. While links between anthracosis and lung disease have been established, anthracosis has not been routinely used as a measure of environmental exposure in research, because there is neither a standard quantitative measure of anthracosis nor an efficient method for quantifying it. We developed 'Slide-based methods for High-throughput Anthracosis Detection and Estimation' (SHADE). SHADE is an automated workflow that quantifies anthracotic pigments on scanned images of whole H&E slides. SHADE was optimized to identify anthracosis while avoiding detection of artefacts in images. SHADE scores and manual pathologist rankings of 10-image patches demonstrated high concordance ($R = 0.864$). Application of SHADE to background lung sections from 140 never-smoking lung cancer patients demonstrated significant associations between high anthracosis and older age, male sex, 30-year residential $PM_{2.5}$ estimates, and birthplace in Asia ($p < 0.05$). Low anthracosis was associated with less aggressive adenocarcinomas ($p = 0.052$). No significant relationship was identified between anthracosis and 3-year residential $PM_{2.5}$ estimates, lobe sampled, ethnicity, *EGFR* mutation status or pathologic tumor stage. Anthracosis levels in the never-smoking cohort were similar to that of background lung sections from 59 ever-smoking patients. Exploratory analysis of background lung sections of 15 individuals who had bronchoalveolar lavage specimen gene expression data available, SHADE scores were associated with the expression of several immunoregulatory genes (*CEACAM6*, *CXCL6*, *IL5RA*, *LCN2*, *SPA17*), the activation of inflammatory response and cytokine gene sets, and suppression of the antigen processing and presentation gene set (false discovery rate < 0.1). SHADE provides an automated workflow for anthracosis quantification and has demonstrated utility in uncovering insight into anthracosis-associated molecular changes in lung cancer.

Keywords: anthracosis; lung cancer; environmental exposures; H&E slides; normal lung; air pollution; digital pathology; $PM_{2.5}$ exposure; QuPath; gene expression

Received 27 February 2026; Revised 3 June 2026; Accepted 3 July 2026

No conflicts of interest were declared.

Introduction

While lung cancers are primarily associated with smoking, the proportion of lung cancer patients who have never smoked is on the rise [1,2]. In most parts of the world, particulate matter with an aerodynamic diameter of less than 2.5 μm ($PM_{2.5}$) air pollution is the second most common cause of lung cancer after tobacco smoking [3]. However, the mechanism by which it promotes lung cancer remains unclear. To effectively direct prevention strategies, it is crucial to understand the impact of air pollution on lung health and its role in the development of lung cancer.

In lung cancer research, the prevailing method for estimating an individual's ambient pollution exposure relies on geographic estimates, largely based on patient addresses in survey data [4]. These estimates are based on averaged readouts from monitoring stations with sparse networks, and studies that utilize them may underestimate the health effects of the measured air pollutant [4]. In addition, obtaining survey data can be resource-intensive, limiting the datasets to patients who are alive, reachable, and willing to participate.

By contrast, anthracosis, defined by black carbon deposits in the lungs, provides a visual indication of a person's environmental exposures [5]. Anthracotic pigment is thought to be representative of inhaled exposures that can originate from various sources, including environmental and occupational exposures, cooking, and drug use [6]. Anthracosis is more prevalent in older adults, suggesting that anthracotic pigment accumulates over a patient's lifespan [6]. While anthracosis is typically asymptomatic, elevated levels of anthracosis extracted from lung tissue from adenocarcinoma patients have been associated with shorter overall survival after tumor detection [7] and poorly differentiated carcinoma morphology [7,8].

Anthracotic pigments are easily identifiable in hematoxylin and eosin (H&E) stained tissue slides, which are routinely produced from both tumor and tumor-adjacent background normal lung as part of the diagnostic assessment of resected tumors. The presence of anthracotic pigment, detected in H&E slides of normal lung tissues from lung cancer patients, was associated with the clonal expansion of cells harboring oncogenic *EGFR* mutations [9], suggesting that anthracosis may promote lung cancer through molecular dysregulation. However, while anthracotic pigment can be detected on H&E slides, H&E slide anthracotic read-outs are not commonly used to estimate pollution exposure in research. This is because there is currently neither a standard quantitative measure of anthracosis nor a comprehensive characterization of samples harboring high levels of anthracotic pigment [10].

Here, we developed an automated workflow [Slide-based methods for High-throughput Anthracosis Detection and Estimation (SHADE)], to identify and quantify anthracotic pigments. To demonstrate the utility of SHADE, we assess whether anthracotic pigment quantified by SHADE in background normal lung sections is associated with pollution exposure, lung cancer pathology, patient demographics, or expression of immune-related genes. The identified associations of anthracosis with patient demographics and gene expression patterns in our analysis support the utility of SHADE in exploring the biology of anthracosis-related molecular dysregulation.

Materials and methods

Ethics statement and cohort characteristics

This study was approved by the University of British Columbia Research Ethics Board (H24-00375, H24-00628, and H25-00384). Patient and tumor characteristics were obtained through retrospective review of electronic medical records (never-smoking cohort: $N = 140$; ever-smoking cohort: $N = 59$; gene expression cohort: $N = 15$). H&E slides were obtained from clinical pathology archives. For each patient, one representative background non-neoplastic lung slide from the lung carcinoma resection specimen (i.e., benign lung tissue from the same lobe as the cancer) was scanned on an APERIO AT2 scanner (Leica, Wetzlar Germany, for the never-smoking and ever-smoking cohort) or NanoZoomer S360 scanner (Hamamatsu, Shizuoka Japan, for cases matched to gene expression data) using a 40× objective. Which lobes were sampled is indicated in Table 1. Background sections were from areas well-separated from the tumor, and were confirmed to have no cancer on review of the microscope slides by a pathologist (JRN). If more than one background lung tissue slide was present, the slide estimated to have more tissue area was used. In this study, the term never-smoking refers to individuals who have smoked fewer than 100 cigarettes in their lifetime.

Designing and implementing the SHADE workflow

Identification and quantification of anthracotic pigments were achieved using SHADE. SHADE is an automated workflow employing the functions of the

Table 1. Never-smoking cohort

Variable	Frequency (n= 140) Mean ± SD (for age & $PM_{2.5}$)	Median (for age and $PM_{2.5}$)	Range (for age and $PM_{2.5}$)	n (for $PM_{2.5}$)
Age (years)	68.1 ± 9.4	70	35–84	
$PM_{2.5}$				
3-Year $PM_{2.5}$	5.83 ± 1.06	5.6	5.1–15.7	111
30-Year $PM_{2.5}$	9.25 ± 5.90	6.93	6.1–32.4	85
Sex				
F	102			
M	38			
Lobe sampled				
LUL	33			
LLL	28			
RUL	37			
RML	10			
RLL	32			
Type				
Adenocarcinoma	139			
Squamous	1			
Predominant histologic pattern				
Acinar	79			
Cribriform	1			
Lepidic	16			
Micropapillary	13			
Mucinous	13			
Papillary	9			
Solid	8			
Squamous	1			
EGFR mutation status				
Positive	61			
Negative	38			
Not tested	41			
Tumor staging				
T1	80			
T2	46			
T3	13			
T4	1			
Node staging				
N0	102			
N1	21			
N2	9			
Unavailable	8			
Birth continent				
Africa	1			
Asia	72			
Europe	10			
North America	14			
South America	2			
Unavailable	41			

Clinicopathological characteristics of the never-smoking cohort, including number of patients, mean, standard deviation, median age, and age range of the cohort. Also shown is the mean, standard deviation, median, range, and number of cases with 3-year $PM_{2.5}$ and 30-year $PM_{2.5}$ data. $PM_{2.5}$: particulate matter with a diameter of ≤ 2.5μm

QuPath software (version 0.5.1-arm64) [11]. A detailed step-by-step protocol of the creation and implementation of SHADE can be found in Supplementary Materials and Methods. The necessary files and scripts for SHADE can be found on GitHub (see Data availability statement section).

The SHADE package relies on a combination of multiple scripts that together enable full automation of the analysis (Figure 1A). The SHADE package requires the tissue image file (a .svs or .ndpi file) as input. The first (and optional) step involves generating a bounding polygon using the preprocess_he_otsu.py Python script, which was implemented to remove artifacts around the edges of the tissue image that may appear during scanning. SHADE will then create a new QuPath project for each image in the directory it was launched in, using create_project.groovy. Once each image has been added to its QuPath project, create_bounding_polygon.groovy will generate the bounding polygon made by the Python script in QuPath based on the bounding polygon coordinates previously generated. SHADE will then begin analysis using the three filter module (TFM) implemented within tfm_script.groovy, which applies three thresholds to each pixel as described below (Figure 1B). After the TFM step, five types of output files can be created based on the TFM analysis using five scripts, where each script generates one output. Ometiffexporter.groovy will export the tissue image in a .tiff format with the anthracosis annotations overlaid on top, while tiff_to_jpeg_process.py will export the tissue image as a .jpeg or .png file. Geojsonexporter.groovy will export all the annotations created in the TFM in a .geojson format, allowing for import back into QuPath. Measure_exp_am.groovy will export all measurements made by each threshold during analysis into a designated .csv file. Finally, the SHADE_output.Rmd script produces an HTML document that summarizes the output and includes the % Anthracosis (SHADE score).

SHADE uses functions within QuPath. These functions have been designed to use various 'Channels' of data to classify each pixel in an image. The 'Max Channel' indicates the maximum RGB value of a pixel. The 'Hematoxylin Channel' indicates pixels that stain well for hematoxylin. The 'Residual Channel' indicates the portion of the image's signal that cannot be explained by hematoxylin or eosin stains.

To correctly identify anthracotic pigment, SHADE requires three types of pixel thresholds, which we calibrated based on training data: (1) Two RGB thresholds on the Max Channel, (2) one Residual threshold on the Residual Channel, and (3) one Hematoxylin threshold on the Hematoxylin Channel.

First, an RGB threshold was calibrated to identify and quantify the total tissue area (Figure 1Bi,Ci),

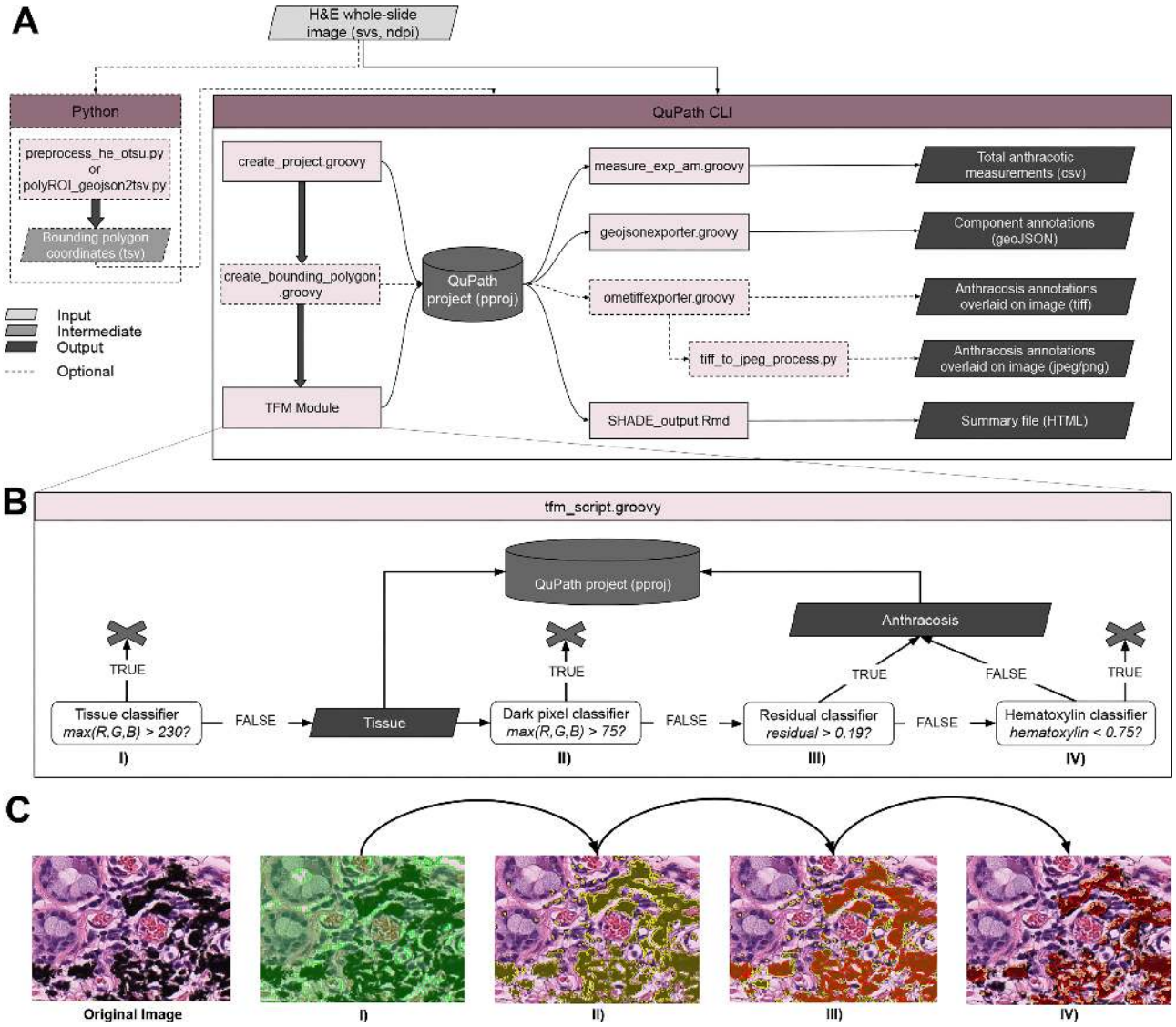


Figure 1. SHADE modules and workflow. (A) The Command Line Interface (CLI) integrated workflow of the SHADE package and shows all scripts involved. (B) Flowchart of the TFM, which performs the SHADE analysis, demonstrating how anthracotic pigment is identified. (C) Visual representation of each step of SHADE analysis. The overlays represent the regions detected at each step. Green corresponds to tissue area, yellow to dark-colored pixels, orange to dark tissue features, and red to anthracotic pigment. Each step of the SHADE analysis is numbered in Roman numerals (I, II, III, and IV) in (B) and (C).

ultimately differentiating between colored (tissue) and white (non-tissue) regions of the slide. Next, another RGB threshold was calibrated to identify dark-colored pixel features within the tissue area (Figure 1Bii,Cii). A residual threshold was then calibrated to identify pixel colors not accounted for by H&E staining, and only areas identified by both the residual threshold and second RGB threshold were considered candidate regions of anthracosis (Figure 1Biii,Ciii). Lastly, a hematoxylin threshold was calibrated to detect non-anthracotic, darkly stained tissue features, such as tissue folds and blood, and exclude these from being misclassified as anthracosis (Figure 1Biv,Civ). Thus, anthracosis is differentiated from other pigments based on its color (black) and staining intensity of the residual channel (higher intensity relative to other tissue features). The final area of anthracosis was defined as below:

$$\text{Area of anthracosis} = \text{Residual threshold output} - \text{Hematoxylin threshold output}$$

With the area of anthracosis, the %Anthracosis was determined as the SHADE score using the following equation:

$$\%\text{Anthracosis} = \left(\frac{\text{Area of anthracosis}}{\text{Area of tissue}}\right) \times 100\%$$

The SHADE score (i.e., %Anthracosis) was the measurement used in subsequent data analysis.

SHADE works optimally on high-quality scanned images. Presence of artefacts such as dust, scratches, bubbles, and dark borders that appear from poor scanning may negatively impact SHADE's performance. For example, depending on the location of the artefacts, the bounding polygon that is generated may include the artefact regions, leading to their contribution in determining the SHADE score. As a result of these issues, 19 slides in the never-smoking cohort required a bounding polygon to be manually defined prior to application of SHADE.

Statistical analysis was performed in the R Project for Statistical Computing (version 4.5.0; R Foundation for Statistical Computing, Vienna, Austria). Ambient $PM_{2.5}$ exposure estimates were generated using a Python-based tool for residential $PM_{2.5}$ exposure assessment, APEX [12]. Participants' lifetime residential addresses were collected at enrollment and geocoded using APEX. Annual satellite-derived air pollution estimates at $0.01^\circ \times 0.01^\circ$ resolution come from the Atmospheric Composition Analysis Group [13,14]. For each geocoded address, all $PM_{2.5}$ grid cells within an 11-km radius were averaged to generate the final annual exposure estimate. When participants had multiple residential addresses in a given year, annual exposure was calculated using a duration-weighted average.

Univariable and multivariable analyses with SHADE scores

Univariable analysis using Spearman correlations for continuous variables and Wilcoxon rank-sum tests for discrete variables was first used to individually characterize the relationship between SHADE scores (%Anthracosis) and several clinical variables, including: Lobe sampled, age, sex, birth continent, ethnicity, predominant morphologic cancer subtype, pathological tumor stage, presence of *EGFR* mutation, presence of pleura, 3-year $PM_{2.5}$ (representing the short-term), 30-year $PM_{2.5}$ estimates (representing the longer term), smoking history, and pack-years of smoking. To obtain false discovery rate (FDR) values, p values were multiple tests corrected using the Benjamini and Hochberg method. After multiple test correction, variables that met a threshold of FDR <0.1 were then selected for a multivariable logistic regression model to test for the relationship after adjustment for other variables. Binary categorical variables were required for the logistic regression model. As such, 'lobe sampled' was dichotomized into the 'upper' (left or right upper lobes) and 'non-upper' (right middle lobe, right lower lobe, or left lower lobe) lobe categories. Birth continent was grouped by 'Asia' and 'Other', as Asia had the highest median SHADE score considering all birth continents. Multivariable regression results with $p \le 0.05$ were considered statistically significant.

Gene expression analysis of bronchoalveolar lavage samples

Gene expression profiling was performed on 15 archival clinical bronchoalveolar lavage (BAL) formalin fixed cell blocks. The BALs had been clinically performed to characterize lung nodules in these patients, but were negative for definitively malignant cells and instead provided background non-neoplastic lung cells for assessment. Per pathologist assessment, all the BAL samples contained predominantly airspace immune cells and smaller numbers of benign bronchial epithelial cells. Nodules in 13 of the 15 patients were subsequently diagnosed as adenocarcinoma by other means, and the others were fibrous nodules negative for malignancy.

RNA was extracted from six 20 μm scrolls with an elution volume of 25 μl using the RNeasy FFPE kit (Qiagen, Hilden, Germany). RNA was quantified using a Qubit 4 fluorometer (Thermo Fisher Scientific, Waltham, MA, USA). As close to 300 ng of input RNA as possible (minimum 36 ng) was input into NanoString nCounter gene expression profiling using the Human PanCancer Immune Profiling Panel (770 genes; Bruker, Billerica, MA, USA). Raw NanoString RCC data files were imported into nSolver Analysis Software (version 4.0; Bruker, https://brukerspatialbiology.com/products/ncounter-analysis-system/nsolver-advanced-analysis-software/, last accessed: 23 June 2026) for quality control and normalization. Background thresholding was performed on raw counts with a threshold count of 20 (i.e., raw counts ≤20 set to the threshold count of 20). Data were normalized using the geometric mean for both positive controls and reference genes that had an average count >100 and percent coefficient of variation <100. For each sample, the background of negative control genes was subtracted; and all values were

scaled according to the expression of housekeeping genes *GUSB*, *HPRT1*, *SDHA*, and *TBP*. To identify genes with expression levels that correlated with anthracosis as estimated by SHADE (on background normal lung sections from a resection of or near the lung region with prior BAL sampling), for each gene, Log2 normalized counts were used in Spearman correlation tests with SHADE scores. A Benjamini-Hochberg FDR <0.05 was considered significant. Pre-ranked gene set enrichment analysis was performed using fgsea and NanoString-curated immune gene set annotations. Genes were ranked by Spearman correlation coefficients as determined above, and enrichment was assessed using gene-label permutations. A Benjamini-Hochberg FDR <0.1 was considered significant. All enrichment results are interpreted within the gene universe represented by the NanoString Immune Expression Panel, reflecting relative gene set activity among the measured immune genes rather than genome-wide gene set activation.

Results

SHADE optimization

Thresholds for tissue and anthracosis detection in SHADE were manually calibrated using 10 whole-slide-images from the never-smoking cohort to maximize anthracosis detection while minimizing misclassification of artefacts as anthracosis. Our approach emphasized excluding artefacts, ensuring that the anthracotic pigment calls are of high confidence, acknowledging

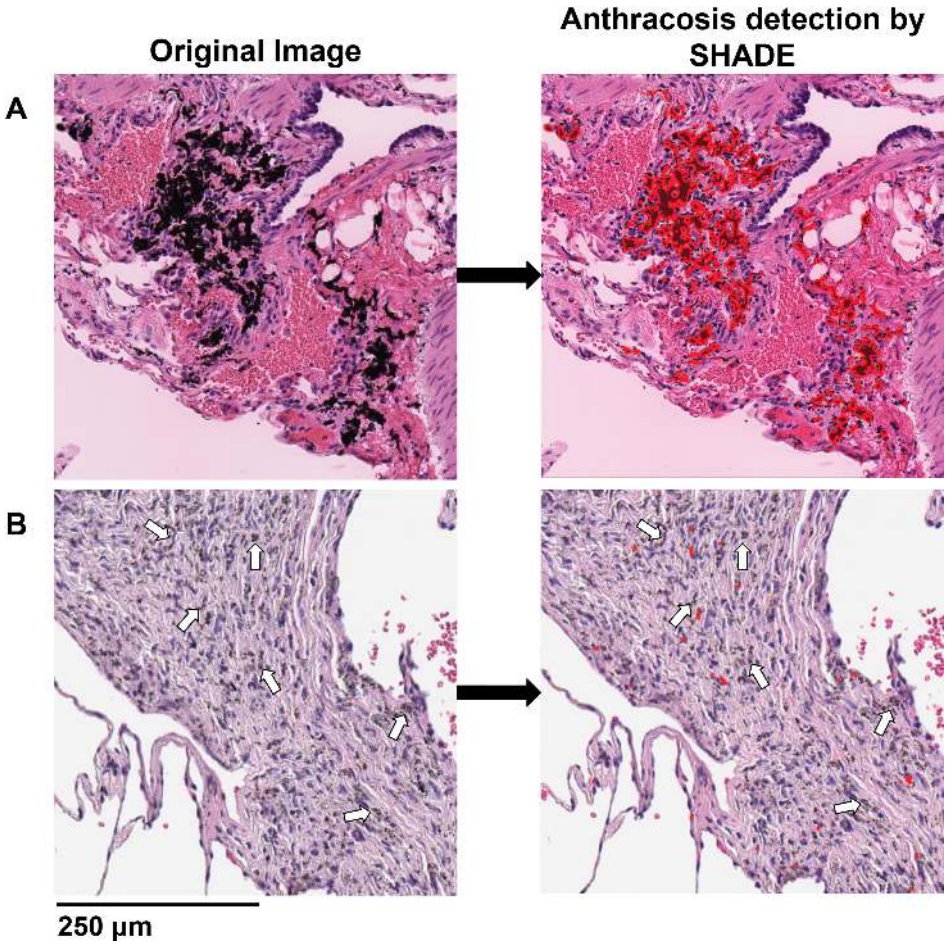


Figure 2. SHADE performs conservatively in anthracosis detection. (A) A region of thick and confluent anthracosis. (B) A region with scattered specks of anthracosis. The left panels show the H&E slide image and the right panels show a red overlay on areas of anthracosis detected by SHADE. The white arrows indicate minute foci of anthracosis that were not detected by the conservative metrics of SHADE.

that the actual extent of anthracosis may be slightly underestimated. Highly confluent areas of anthracosis were more consistently detected by SHADE than scattered specks of anthracosis and intracellular anthracotic pigments (Figure 2). To understand possible errors in SHADE detection of anthracosis, we tested SHADE on image patches showing potential artefacts, including darkly stained nuclei, alveolar macrophages with hemosiderin or smoking-related pigment, tissue folds and darkly stained cartilage (supplementary material, Figure S1A). We observed extremely minimal false positive SHADE calls in these test images (supplementary material, Figure S1B,C).

SHADE validation

To assess the accuracy of SHADE in detecting prespecified quantities of black pigments, SHADE was applied to 10 synthetically generated test images containing known values of 'tissue' area (pink with dark purple circles as 'nuclei') and 'anthracotic' areas (black), with white background space (supplementary material, Figures S2 and S3A). SHADE detected on average 92.5% of the true area of anthracosis, with no false positive classification (supplementary material, Figure S3B,C). The underestimation of anthracosis is likely due to the reduction of image resolution when input into SHADE, which was done for computational efficiency and to reduce false positive anthracosis calls on minuscule artifacts.

To assess whether SHADE could simulate human impressions of relative anthracotic pigment abundance in lung tissue, we assessed SHADE's performance on 10-image patches of normal lung with varying degrees of anthracosis (Figure 3A). Regions were selected to represent a wide range of amounts of anthracosis, including very fine and more confluent areas of pigmentation, and to include a mix of alveolar parenchyma and small airway wall tissue. These image patches were ranked from least to most anthracosis relative to total tissue area by SHADE and four human observers, including two pathologists (Figure 3B). Pairwise Pearson correlation coefficients between SHADE and each human observer produced an average correlation of $R = 0.864$ (Figure 3C). This was comparable to the average pairwise correlation between human observers ($R = 0.853$). When considering images with the most and least anthracosis, SHADE was consistent with the human rankings.

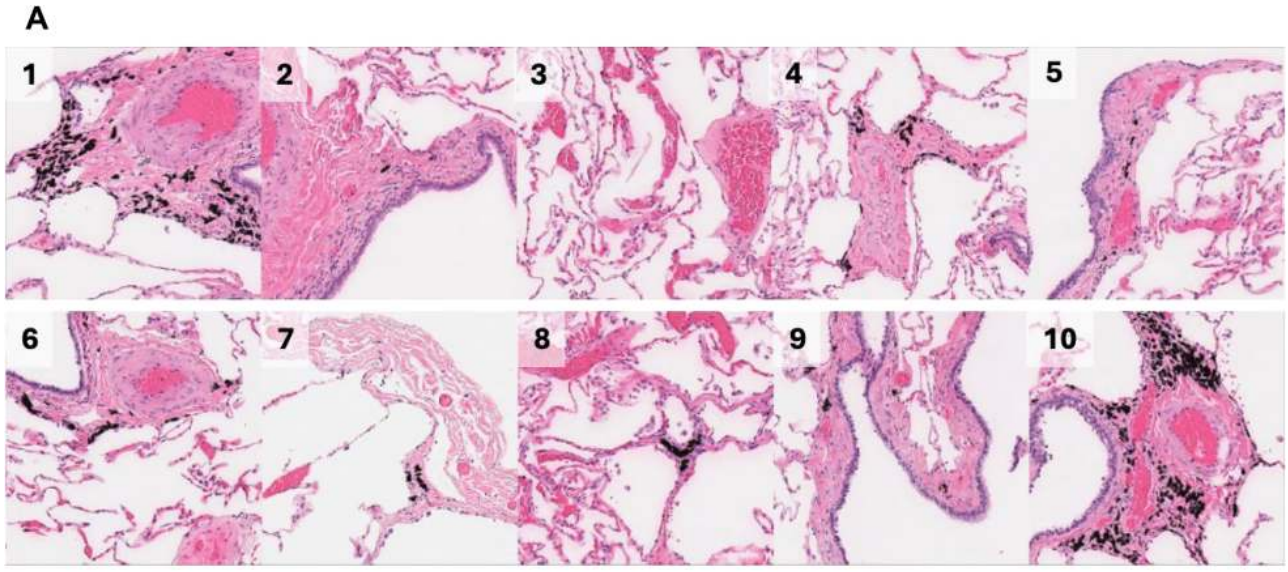
Associations with anthracosis in never-smoking lung cancer patients

We obtained scanned H&E whole slide images of archival background normal lung sections from 140 never-smoking patients who underwent excision of lung carcinoma (Table 1). SHADE scores from these slides ranged between 0.0029% and 0.37% with a median of 0.033%. We used the SHADE scores derived from these whole slide images to assess for associations between anthracotic pigment abundance and (1) patient demographics, (2) tumor morphology, (3) anatomical position of the lung section (to estimate biases in anthracosis that may result from the site sampled), and (4) $PM_{2.5}$ pollution exposure per residential history and satellite image data (Figure 4A).

Higher anthracosis was associated with older age at time of resection ($R = 0.18$, $p = 0.03$; Figure 4B), male sex ($p = 0.0053$; Figure 4C), non-lepidic predominant or mucinous tumors ($p = 0.052$; Figure 4D), and 30-year $PM_{2.5}$ exposure estimates ($R = 0.22$; $p = 0.042$; Figure 4E). Birth continent (Asia versus not Asia) also demonstrated significant associations with increased levels of anthracosis ($p < 0.05$; Figure 4F). The effect of birth continent on levels of anthracosis may be related to environmental rather than inherited factors, as the two patients of East Asian ethnicity who were born in North America had the lowest levels of anthracosis among patients whose ethnicity was reported as Asian (supplementary material, Figure S4C). When comparing morphologic adenocarcinoma types, the lowest median anthracosis scores were in patients with lepidic predominant and mucinous adenocarcinomas (supplementary material, Figure S4F).

Cumulative $PM_{2.5}$ exposures over the 3 years prior to lung cancer diagnosis was not significantly associated with anthracotic pigment levels ($R = -0.082$, $p = 0.39$; supplementary material, Figure S4A). Anthracosis was not significantly associated with *EGFR* mutation in the tumor ($p = 0.48$; supplementary material, Figure S4B), ethnicity ($p > 0.05$; supplementary material, Figure S4C), pathologic tumor stage ($p > 0.05$; supplementary material, Figure S4D), upper versus non-upper lobe sampled ($p = 0.15$; supplementary material, Figure S4E), tumor subtype ($p > 0.05$; supplementary material, Figure S4F), and whether visceral pleura was present in the section (i.e., subpleural lung sampled) or not (i.e., central lung sampled) ($p = 0.19$; supplementary material, Figure S4G).

To assess which variables maintained associations with the SHADE score after adjusting for other variables, a multivariable logistic regression was performed on all variables associated with anthracosis at an FDR <0.1 in single variable analysis: age, sex, tumor morphology, 30-year $PM_{2.5}$ exposure estimates, and Asia as birth continent (Figure 4G). Birthplace in Asia and age remained significantly associated with



B

SHADE vs. rater scores

	Person 1	Person 2	Person 3	Person 4	SHADE
Most	10	10	10	10	10
	1	1	1	1	1
	6	6	4	4	6
	4	8	6	8	4
	8	9	8	6	8
	9	7	7	9	7
	7	4	9	7	9
	5	5	5	5	5
	2	2	2	2	2
Least	3	3	3	3	3

C

Pairwise Pearson correlation between ratings

	Person 1	Person 2	Person 3	Person 4	SHADE
Person 1	X	0.818	0.903	0.855	0.952
Person 2	X	X	0.794	0.842	0.745
Person 3	X	X	X	0.903	0.952
Person 4	X	X	X	X	0.806
SHADE	X	X	X	X	X

Figure 3. Validation of SHADE on synthetic images. Panel (A) shows 10-image patches used in the validation of SHADE by pathologists. Panel (B) displays the rankings of each image patch from least to most anthracosis relative to the total tissue area. Panel (C) illustrates the pairwise Pearson correlation matrix comparing all raters and SHADE.

anthracosis in the multivariable regression model (birthplace: $p < 0.05$; age: $p < 0.01$; Figure 4G).

Comparison of never and ever-smoking patient anthracosis levels

As cohorts of interest for SHADE assessment may include patients who have smoked, we assessed whether levels of anthracosis show an association with smoking history that may confound the detection of other associations with anthracosis. SHADE was applied to background lung sections from lung adenocarcinoma (LUAD) resections of 52 patients who formerly smoked and 7 patients with ongoing smoking around the time of resection (Figure 5A, Table 2). SHADE scores were not significantly different between the ever-smoking cohort and the above described never-smoking cohort ($p = 0.058$; Figure 5B), and showed no association with pack-year smoking history among patients who had ever smoked ($R = 0.15$, $p = 0.26$; Figure 5C).

Associations between anthracosis and gene expression

To demonstrate the utility of SHADE scores in exploring the biology of anthracosis-related molecular

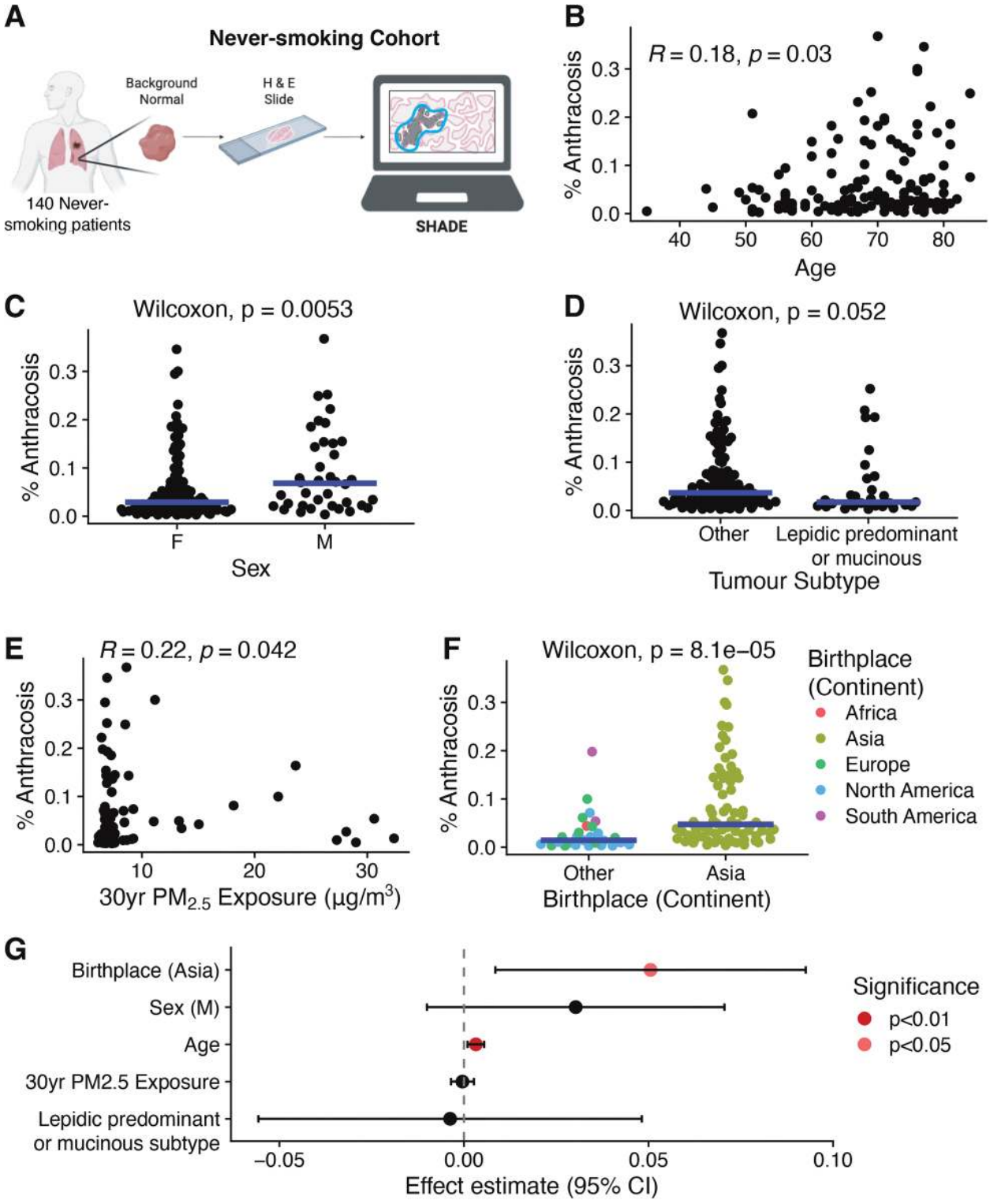


Figure 4. SHADE applied to background normal lung H&E slides of never-smoking patients who underwent lung carcinoma resection (never-smoking cohort; $N = 140$) enables identification of clinical and molecular features associated with anthracosis. (A) Illustration of the workflow. (B–E) SHADE score associations across the never-smoking cohort with (B) age (Spearman correlation R and p are reported), (C) sex, (D) tumor subtype group, (E) 30-year cumulative $PM_{2.5}$ exposure, and (F) birthplace (continent). (G) Multivariable logistic regression for SHADE association with clinical features. CI, confidence interval. Panel A was created with BioRender.com.

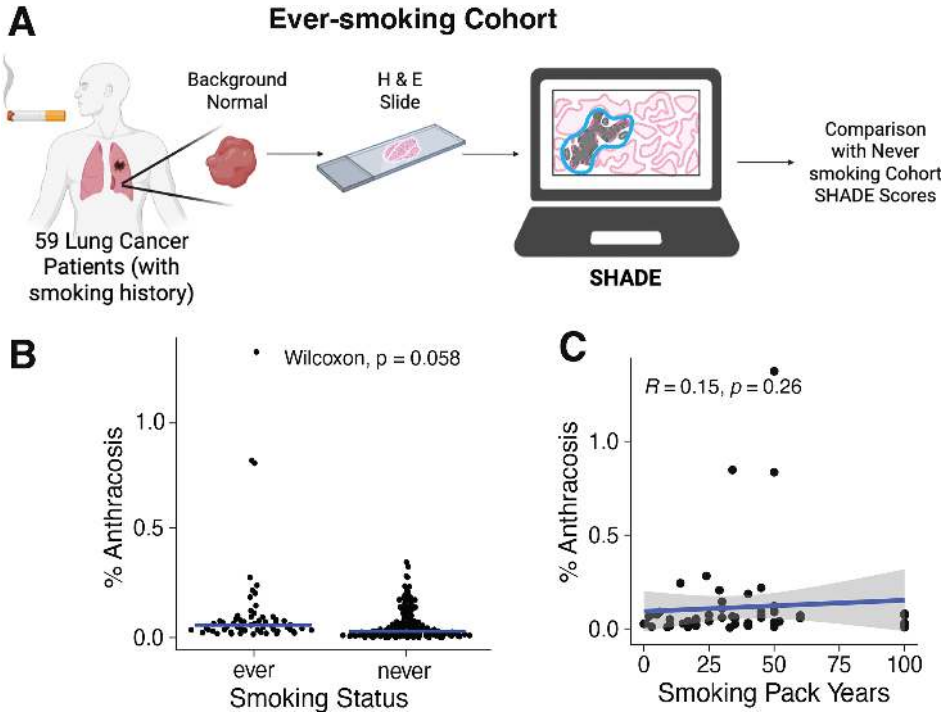


Figure 5. SHADE applied to background normal lung H&E slides of patients with a smoking history and who underwent lung carcinoma resection (ever-smoking cohort; $N = 59$). (A) Illustration of the workflow. (B) Comparison of SHADE scores based on smoking status. (C) SHADE score association with pack-year history from the ever-smoking cohort. Panel A was created with BioRender.com.

dysregulation, we performed an exploratory, hypothesis-generating assessment of SHADE scores for patients with gene expression data also available. We obtained background benign lung H&E whole-slide images from 15 individuals who underwent resection of a lung nodule (Table 3). For each patient, gene expression profiling was performed on a BAL sample collected prior to surgery (Figure 6A). Only BAL samples negative for malignant cells were used to reflect background lung biology. The BAL samples contained predominantly inflammatory cells from alveolar spaces and gene expression profiling assessed 770 immune-related genes, enabling assessment of the association between anthracosis and immune cell function. The expression of five genes (*CEACAM6*, *CXCL6*, *IL5RA*, *LCN2*, *SPA17*) was significantly positively correlated with anthracosis levels (FDR < 0.05) (Figure 6B,C). To determine whether these associations remain significant after adjusting for other variables, we performed a multi-variable regression analysis on gene expression, including as variables SHADE score, age, sex, smoking status, and upper versus non-upper lobe sampled. The expression of four genes was still significantly associated with SHADE scores (*CEACAM6*, *CXCL6*, *IL5RA*, *SPA17*, $p < 0.05$). Gene set enrichment analysis based on gene sets provided by NanoString annotations showed activation of 'regulation of inflammatory response', and 'cytokines and receptors' gene sets ($p < 0.05$; FDR < 0.1; Figure 6D). The 'antigen processing and presentation' gene set was suppressed ($p < 0.05$; FDR < 0.1; Figure 6D).

Discussion

Environmental exposures are increasingly recognized as critical contributors to the global burden of lung cancer. Studying anthracotic pigmentation as an indicator for exposures and as a potential mediator of tissue dysregulation may provide insights into lung cancer pathogenesis. However, such work has been limited by the lack of an automated workflow for detecting and quantifying anthracosis in the large clinical archives of H&E slides.

Based on free, open-source software [11], we developed the SHADE method to identify and quantify anthracotic pigments in a fully automated workflow on H&E whole-slide scanned images. We demonstrate

Table 2. Ever-smoking cohort

Variable	Frequency ($n = 59$) Mean ± SD (for age)	Median (for age)	Range (for age)
Age (years)	69.1 ± 9.4	70	43–84
Sex			
Female	29		
Male	30		
Lobe sampled			
LUL	16		
LLL	6		
RUL	25		
RML	3		
RLL	9		
Type			
Adenocarcinoma	49		
Squamous	10		
Predominant histologic pattern			
Acinar	20		
Pleomorphic	1		
Lepidic	3		
Micropapillary	13		
Mucinous	2		
Papillary	3		
Solid	7		
Squamous	10		
EGFR mutation status			
Positive	8		
Negative	23		
No data	28		
Tumor staging			
T1	27		
T2	24		
T3	5		
T4	3		
Node staging			
N0	45		
N1	7		
N2	6		
Unavailable	1		
Smoking status			
Current	7		
Former	52		
Pack years			
0–20	18		
21–40	18		
41–60	16		
61–80	0		
81–100	5		
No data	2		

Clinicopathological characteristics of the ever-smoking cohort, including number, mean, standard deviation, median age, and age range of the cohort.

Table 3. Gene expression cohort

Variable	Frequency ($n = 15$) Mean ± SD (for age)	Median (for age)	Range (for age)
Age (years)	69 ± 8.7	70	54–82
Sex			
F	9		
M	6		
Lobe sampled (both slide and lavage)			
LUL	4		
LLL	3		
RUL	5		
RML	0		
RLL	3		
Type			
Adenocarcinoma	13		
Benign	2		
Smoking status			
Never	6		
Former	9		
Pack years			
0–20	4		
21–40	1		
41–50	2		
No data	2		

Clinicopathological characteristics of the gene-expression cohort, including number, mean, standard deviation, median age, and age range of the cohort.

that SHADE measures relative levels of anthracotic pigments between specimens comparably to human observers with limited miscalling of artifacts. Prior efforts to quantify anthracosis have involved either manual 'counting of dots' on H&E slides [9], densitometry assessment of black dust matter extracted from tissue sections and blotted on nitrocellulose membranes [7,8], macroscopic photographs of the pleural surface of autopsy lung specimens [5], or mass spectrometry based analysis of carbon-bound exogenous compounds extracted from tissue microarray sections [15]. To our knowledge, SHADE is the first automated anthracosis quantification approach for archival H&E slides, which are routinely prepared after each lung cancer surgery.

Our analysis of the never-smoking cohort demonstrated relationships between SHADE scores and age, males, tumor morphology, 30-year $PM_{2.5}$ exposure estimates, and the birth continent in Asia. Prior publications have reported positive associations between levels of anthracosis and age, in keeping with possible accumulation of anthracotic pigment over time through additive exposures [5,6,16]. Previous reports have shown slightly more prevalent diagnosis of anthracosis in females than males [6], yet we observed greater SHADE scores in males. This may be due to gender-associated occupational, recreational, or environmental exposures of individuals in our cohort, which were unknown at the time of this study. The tumor morphologies with lower anthracosis were those typically associated with more favorable outcomes of early stage disease, in keeping with the notion that anthracosis may be associated with the development of more aggressive adenocarcinomas [17,18].

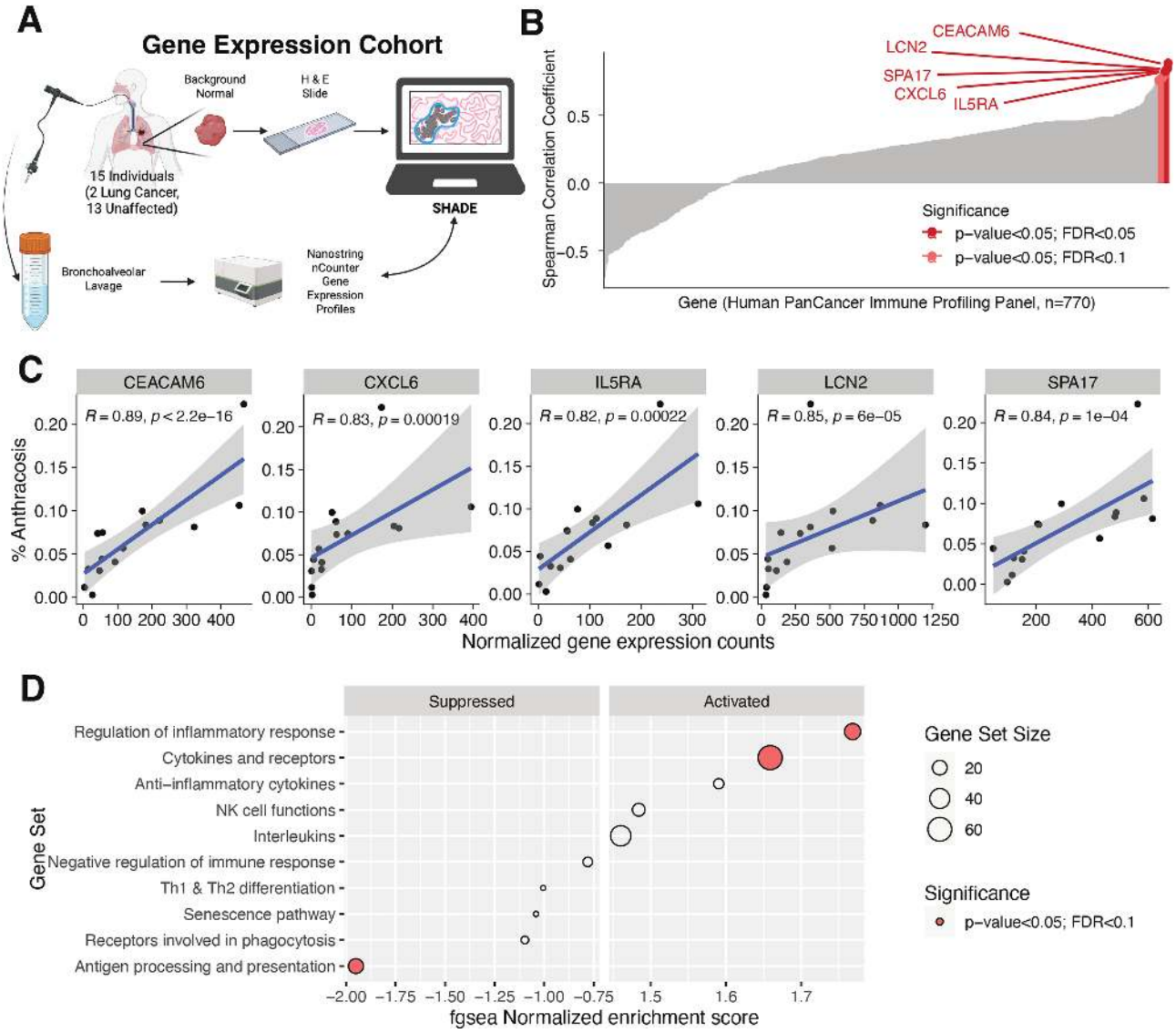


Figure 6. Analysis of SHADE anthracosis scores with gene expression data. (A) Illustration of the workflow. (B) Waterfall plot displaying Spearman correlation coefficients for the expression of each gene and SHADE scores across the gene expression cohort. Genes with expression patterns that correlated significantly with SHADE scores are indicated in red. (C) Scatter plots displaying the positive associations between SHADE scores and normalized gene expression counts. (D) Normalized enrichment scores from gene set enrichment using fgsea. The top 10 activated and suppressed gene sets are displayed. Panel A was created with BioRender.com.

Although the relationship between greater quantities of anthracosis and cumulative 30-year $PM_{2.5}$ levels has not been previously studied, animal studies have indicated both a higher prevalence and greater intensity of anthracosis in animals living in urban areas with higher annual atmospheric PM levels compared to urban areas with lower annual atmospheric PM levels [19]. Our finding further extends this observation to humans and demonstrates the possibility of anthracosis as an indicator of long-term $PM_{2.5}$ exposure. We did not identify a significant association between 3-year $PM_{2.5}$ exposure and levels of anthracosis. This could be due to the limited number of patients with very high $PM_{2.5}$ within our cohort, reducing power for detection of associations with high-level $PM_{2.5}$ exposure (only 2 patients with >10 μg/m^3 3-year $PM_{2.5}$, and 13 patients with >10 μg/m^3 30-year $PM_{2.5}$). The lack of association may also highlight a limitation of using residential addresses as an estimate for pollution exposure, as most individuals

tend to spend a portion of their day away from their residential area and may have other occupational, recreational or indoor exposures contributing to their burden of anthracotic pigment. In addition, exposures beyond the 30-year window of $PM_{2.5}$ data collection would not have been captured in the $PM_{2.5}$ estimates. In contrast, SHADE scores potentially represent the accumulation of anthracotic pigments over the lifespan, independent of exact source.

The association between higher quantities of anthracotic pigment and birth in Asia has not been previously described, though bronchoscopic studies indicate a generally higher prevalence of anthracotic disease in rural areas of Asia compared to elsewhere [20].

Comparison of SHADE scores between the never-smoking and ever-smoking cohorts revealed no significant difference, in agreement with previous reports [9]. Likewise, there was no significant association between smoking pack-year and SHADE scores. These findings are in keeping with the notion that cigarette smoking is not associated with the presence of anthracotic pigments [6]. Statistical power for detecting small differences between smoking and never-smoking cohorts is limited by sample size.

Our gene set enrichment analysis indicated that SHADE scores were associated with the activation of inflammation and cytokine genes and suppression of genes involved in antigen presentation, in line with previous reports of the effects of pollution on lung cells [21,22]. Gene-level analysis further revealed specific genes that were dysregulated with increasing SHADE scores. The genes that had greater expression in high-anthracosis cases (*CEACAM6, CXCL6, IL5RA, LCN2, SPA17*) have been putatively linked to oncogenesis [23–30]. *CEACAM6* marks proliferative epithelial cells after lung injury [23]. *CEACAM6* overexpression in LUAD samples has been associated with shorter survival, non-smoking history, and activating *EGFR* mutations [24]. Accordingly, knockdown of *CEACAM6* in *EGFR*-mutated lung cancer cells induced apoptosis [25]. There have been several reports of the inflammatory signaling gene *CXCL6* being upregulated in many tumors [26], and one study in particular found that it was upregulated in tumors of individuals living in the highly polluted Xuanwei region of China [27]. *LCN2* (lipocalin-2) was increased in normal airways from mice that are prone to developing inflammation and LUAD. LCN2 protein in BAL fluid was also significantly and progressively increased with time during lung tumorigenesis [28]. While not specific to lung cancer, *IL5RA* has been shown to be involved in anti-tumor immunity [29], while *SPA17* predicts the immunotherapy response in cancer patients [30]. Higher expression of these genes in lungs of patients with greater anthracosis supports the notion that dysregulation of these genes may play a role in pollution-promoted lung cancer initiation and progression, potentially targetable for lung cancer prevention. We also note that our analysis of gene expression data in relation to anthracosis is limited by small sample size. While the particular gene expression associations identified require validation, this aspect of our analysis demonstrates the capacity of SHADE to generate hypotheses for further investigation; our findings for gene expression features associated with anthracosis should not be considered definitive.

A limitation of SHADE is poor sensitivity for very small particles of anthracosis, as SHADE is limited by the resolution of its pixel thresholds. SHADE's pixel thresholds used the 'High' resolution (1.01 μm/px) setting rather than utilizing the full resolution of the image available (using 40× objective). However, use of the full resolution setting increased false positive detection of artefacts. It is possible that a case with a high volume of very tiny dispersed black particles may incorrectly score lower than a case with an equivalent area of large black particles. A brief manual review of slides for any cases with a very high number of fine particles could allow such cases to be flagged; in our study set there were no such cases, with particle size appearing on manual inspection to increase with the total area of anthracosis (i.e., all high anthracosis cases had large confluent deposits). Despite this limitation in absolute quantification, the relative quantification of anthracosis by SHADE was aligned with human observers, in keeping with utility for comparisons of relative anthracosis within a cohort.

Our use of a single H&E slide from one lobe of each patient's lung may not be fully representative of a patient's lungs, however tissue from more than one lobe is rarely available, and we observed no significant differences between upper versus lower lobes, or central versus subpleural areas. Thus, the section used may be a reasonable approximation of overall lung anthracosis, balancing representativeness with specimen availability. Previous attempts to quantify anthracosis in the lungs also used similar or smaller size tissue samples (e.g., tissue microarray cores) to represent overall anthracotic burden [7–9,15]. For slides in our study, the lung tissue typically filled most of the slide area, and variability in the exact area of the tissue on each slide was controlled for by dividing the quantities of anthracosis by total tissue area. The total tissue area is calculated from the H&E colored pixels and therefore excludes airspaces, such that variable inflation of the tissue is also adjusted for.

In summary, we have developed a free open-access tool capable of leveraging extensive archives of lung pathology slides for the quantification of anthracotic pigment deposition in an automated and high-throughput fashion. We observed an association of anthracosis with age, in accordance with prior literature. We also demonstrate relationships between anthracosis and male sex, 30-year $PM_{2.5}$ exposure estimates, tumor morphology, birth continent, and expression of immune-related genes that have not previously been discussed in the literature. These findings underscore the relevance of anthracotic pigment quantification to uncovering possible relationships between inhalational exposures and lung cancer pathobiology. SHADE scores potentially reflect one's lifetime exposures, which may be attributable to multiple factors including pollution, occupational, recreational and domestic exposures, and migration history. Future application of SHADE to large archival cohorts would provide unprecedented statistical power for identifying associations between anthracosis and the pathologic and molecular features of lung tissue and tumors. Examining the tissues harboring these pigments may uncover how they may contribute to dysregulation of normal lung tissues and potentially promote carcinogenesis. Understanding the role and impact of anthracotic pigments in the lungs may help us better comprehend mechanisms driving the development of lung cancer and lead to improved treatments, diagnosis, and prevention strategies.

Acknowledgements

This project was supported by funds from the University of British Columbia Diagnostic and Molecular Pathology Residency Program, Terry Fox Research Institute 'Victor Ling New Investigator Award', Terry Fox Research Institute, 'The Environment and Lung Cancer' New Frontiers Program Project Grant, and a University of British Columbia New Faculty Award.

Author contributions statement

LCF contributed to methodology, validation, software development, visualization, formal analysis and wrote the original draft of the manuscript. BJ contributed to data curation, validation, and investigation. KO-A contributed to software development. TL contributed to methodology, validation and software development. PW contributed to investigation and resources. MT contributed to investigation and data curation. GB contributed to investigation and data curation. TN-L contributed to investigation. CHD contributed to investigation, software development and formal analysis. AT contributed to investigation, software development and formal analysis. SO contributed to data curation, supervision and resources. SL, RMez, RM and WWL provided supervision. AM contributed to conceptualization, supervision and resources. JRN contributed to conceptualization, supervision, resources, funding acquisition, validation and manuscript review and editing. ELL contributed to conceptualization, supervision, resources, funding acquisition, validation and manuscript review and editing.

Data availability statement

The necessary scripts for SHADE to run can be found on our GitHub. Please contact Emilia L Lim for tissue slide information.

References

1. Corrales L, Rosell R, Cardona AF, *et al*. Lung cancer in never smokers: the role of different risk factors other than tobacco smoking. *Crit Rev Oncol Hematol* 2020; **148**: 102895.
2. Pelosof L, Ahn C, Gao A, *et al*. Proportion of never-smoker non–small cell lung cancer patients at three diverse institutions. *J Natl Cancer Inst* 2017; **109**: djw295.
3. Ebrahimi H, Aryan Z, Moghaddam SS, *et al*. Global, regional, and national burden of respiratory tract cancers and associated risk factors from 1990 to 2019: a systematic analysis for the Global Burden of Disease Study 2019. *Lancet Respir Med* 2021; **9**: 1030–1049.
4. de Preux L, Rizmie D, Fecht D, *et al*. Does it measure up? A comparison of pollution exposure assessment techniques applied across hospitals in England. *Int J Environ Res Public Health* 2023; **20**: 3852.
5. Takano APC, Justo LT, dos Santos NV, *et al*. Pleural anthracosis as an indicator of lifetime exposure to urban air pollution: an autopsy-based study in Sao Paulo. *Environ Res* 2019; **173**: 23–32.
6. Mirsadraee M. Anthracosis of the lungs: etiology, clinical manifestations and diagnosis: a review. *Tanaffos* 2014; **13**: 1–13.
7. Hou M, Morishita Y, Iijima T, *et al*. The implication of Anthracosis in the development of pulmonary adenocarcinoma. *Jpn J Cancer Res* 1998; **89**: 1251–1256.
8. Wang D, Minami Y, Shu Y, *et al*. The implication of background anthracosis in the development and progression of pulmonary adenocarcinoma. *Cancer Sci* 2003; **94**: 707–711.
9. Hill W, Lim EL, Weeden CE, *et al*. Lung adenocarcinoma promotion by air pollutants. *Nature* 2023; **616**: 159–167.

10. Hang W, Bu C, Cui Y, *et al*. Research progress on the pathogenesis and prediction of pneumoconiosis among coal miners. *Environ Geochem Health* 2024; **46**: 319.
11. Bankhead P, Loughrey MB, Fernández JA, *et al*. QuPath: open source software for digital pathology image analysis. *Sci Rep* 2017; **7**: 16878.
12. Durney CH, Tawara A, Brauer M, *et al*. APEX: a web-based tool for estimating long-term PM2.5 exposure from residential history – brief report. *JTO Clin Res Rep* 2026; **7**: 100975.
13. van Donkelaar A, Hammer MS, Bindle L, *et al*. Monthly global estimates of fine particulate matter and their uncertainty. *Environ Sci Technol* 2021; **55**: 15287–15300.
14. Hammer MS, van Donkelaar A, Bindle L, *et al*. Assessment of the impact of discontinuity in satellite instruments and retrievals on global PM2.5 estimates. *Remote Sens Environ* 2023; **294**: 113624.
15. Shen J, Sun N, Wang J, *et al*. Patterns of carbon-bound exogenous compounds impact disease pathophysiology in lung cancer subtypes in different ways. *ACS Nano* 2023; **17**: 16396–16411.
16. Balchum E, Sussman S, Toyama T, *et al*. The Nashville air pollution study: II. Pulmonary anthracosis as an index of air pollution. *Am J Public Health Nations Health* 1963; **53**: 185–199.
17. Boutros CS, Hu YL, Bassiri A, *et al*. Comparative analysis of overall survival in mucinous and nonmucinous adenocarcinoma of the lung: insights from a National Cancer Database Study. *J Surg Res* 2025; **315**: 151–158.
18. Lee JO, Lee GD, Choi S, *et al*. Surgical prognosis of lung invasive mucinous and non-mucinous adenocarcinoma: propensity score matched analysis. *Eur J Cardiothorac Surg* 2024; **66**: ezae316.
19. Ellwanger JH, Ziliotto M, Chies JAB. Lung deposition of particulate matter as a source of metal exposure: a threat to humans and animals. *Toxics* 2025; **13**: 788.
20. Jamaati H, Sharifi A, Mirenayat MS, *et al*. What do we know about anthracofibrosis? A literature review. *Tanaffos* 2017; **16**: 175–189.
21. Kim DI, Song MK, Yuk JE, *et al*. Immunosuppression induced by co-exposure to allergens and PM2.5 increases the risk of lung disease upon viral protein challenge. *Environ Int* 2025; **203**: 109722.
22. McCarrick S, Malmborg V, Gren L, *et al*. Pulmonary exposure to renewable diesel exhaust particles alters protein expression and toxicity profiles in bronchoalveolar lavage fluid and plasma of mice. *Arch Toxicol* 2025; **99**: 797–814.
23. Lin S, Barrette AM, Chapin C, *et al*. Expression of human carcinoembryonic antigen-related cell adhesion molecule 6 and alveolar progenitor cells in normal and injured lungs of transgenic mice. *Physiol Rep* 2015; **3**: e12657.
24. Kim EY, Cha YJ, Jeong S, *et al*. Overexpression of CEACAM6 activates Src-FAK signaling and inhibits anoikis, through homophilic interactions in lung adenocarcinomas. *Transl Oncol* 2022; **20**: 101402.
25. Zheng MY, Lin YT, Kuo YS, *et al*. Cytokine and epigenetic regulation of CEACAM6 mediates EGFR-driven signaling and drug response in lung adenocarcinoma. *NPJ Precis Oncol* 2025; **9**: 115.
26. Dai CL, Yang HX, Liu QP, *et al*. CXCL6: a potential therapeutic target for inflammation and cancer. *Clin Exp Med* 2023; **23**: 4413–4427.
27. Wang GZ, Cheng X, Zhou B, *et al*. The chemokine CXCL13 in lung cancers associated with environmental polycyclic aromatic hydrocarbons pollution. *Elife* 2015; **4**: e09419.
28. Treekitkarnmongkol W, Hassane M, Sinjab A, *et al*. Augmented lipocalin-2 is associated with chronic obstructive pulmonary disease and counteracts lung adenocarcinoma development. *Am J Respir Crit Care Med* 2021; **203**: 90–101.
29. Ikutani M, Yanagibashi T, Ogasawara M, *et al*. Identification of innate IL-5–producing cells and their role in lung eosinophil regulation and antitumor immunity. *J Immunol* 2012; **188**: 703–713.
30. Tu Z, Peng J, Long X, *et al*. Sperm autoantigenic protein 17 predicts the prognosis and the immunotherapy response of cancers: a Pan-cancer analysis. *Front Immunol* 2022; **13**: 844736.

SUPPLEMENTARY MATERIAL ONLINE

Supplementary materials and methods

Figure S1. Specificity of anthracotic pigment detection by SHADE. Each column shows a potential mimic of anthracosis, specified above the images

Figure S2. Ten synthetic images for validation of SHADE

Figure S3. Validation of SHADE on synthetic images

Figure S4. Comparison of SHADE scores to tumor characteristics