J Pathol Inform

OPEN ACCESS
HTML format

Research Article

Automated quantification of aligned collagen for human breast carcinoma prognosis

Jeremy S. Bredfeldt[1,2], Yuming Liu[1], Matthew W. Conklin[1,3], Patricia J. Keely[1,3], Thomas R. Mackie[1,2], Kevin W. Eliceiri[1,2,3]

[1]Laboratory for Optical and Computational Instrumentation, [2]Morgridge Institute for Research, Madison, WI 53715, USA, [3]Laboratory for Cell and Molecular Biology, University of Wisconsin at Madison, Madison, WI 53706, USA

E-mail: *Kevin W. Eliceiri - eliceiri@wisc.edu
*Corresponding author

Received: 09 January 2014 Accepted: 08 May 2014 Published: 28 August 2014

This article may be cited as:
Bredfeldt JS, Liu Y, Conklin MW, Keely PJ, Mackie TR, Eliceiri KW. Automated quantification of aligned collagen for human breast carcinoma prognosis. J Pathol Inform 2014;5:28.

Available FREE in open access from: http://www.jpathinformatics.org/text.asp?2014/5/1/28/139707

Abstract

Background: Mortality in cancer patients is directly attributable to the ability of cancer cells to metastasize to distant sites from the primary tumor. This migration of tumor cells begins with a remodeling of the local tumor microenvironment, including changes to the extracellular matrix and the recruitment of stromal cells, both of which facilitate invasion of tumor cells into the bloodstream. In breast cancer, it has been proposed that the alignment of collagen fibers surrounding tumor epithelial cells can serve as a quantitative image-based biomarker for survival of invasive ductal carcinoma patients. Specific types of collagen alignment have been identified for their prognostic value and now these tumor associated collagen signatures (TACS) are central to several clinical specimen imaging trials. Here, we implement the semi-automated acquisition and analysis of this TACS candidate biomarker and demonstrate a protocol that will allow consistent scoring to be performed throughout large patient cohorts. **Methods:** Using large field of view high resolution microscopy techniques, image processing and supervised learning methods, we are able to quantify and score features of collagen fiber alignment with respect to adjacent tumor-stromal boundaries. **Results:** Our semi-automated technique produced scores that have statistically significant correlation with scores generated by a panel of three human observers. In addition, our system generated classification scores that accurately predicted survival in a cohort of 196 breast cancer patients. Feature rank analysis reveals that TACS positive fibers are more well-aligned with each other, are of generally lower density, and terminate within or near groups of epithelial cells at larger angles of interaction. **Conclusion:** These results demonstrate the utility of a supervised learning protocol for streamlining the analysis of collagen alignment with respect to tumor stromal boundaries.

Key words: Breast cancer, collagen, image processing, machine learning

Access this article online
Website:
www.jpathinformatics.org
DOI: 10.4103/2153-3539.139707
Quick Response Code:

BACKGROUND

Breast cancer diagnosis and staging have been revolutionized by new molecular screening assays based on immunohistochemistry,[1] fluorescence *in situ* hybridization,[2] and reverse transcription polymerase chain reaction,[3] which are all used to personalize care. These tools are helping patients live longer and receive better treatment than ever before. However, there remains a significant group of breast cancer patients for whom these new techniques ultimately fail, due to several factors including varying patient genotype and primary or acquired resistance to drugs such as the HER2/neu receptor targeting drug trastuzumab (trade name Herceptin).[4] In addition, molecular screens are confounded by the high-degree of intratumor genetic diversity and often require extra tissue sections to be cut, stained and evaluated on top of the standard hematoxylin and eosin (H&E) preparation. New assays that predict patient outcome and response to treatment are therefore critically needed if we are to continue improving breast cancer treatment and prevention. One promising area of development is image based assays, which leverage high content imaging hardware and image analysis software to classify biological samples.[5-7] In many cases, image based analysis does not require more than the standard histopathology H&E stained slides prepared as part of the normal clinical workflow. In this paper, we demonstrate the use of a new image-based assay for predicting patient outcome using information about tumor-stromal interactions from standard H&E stained histopathology specimens.

Aberrant tumor-stromal interactions have been shown to accelerate tumorigenesis in breast cancer.[8-10] The importance of stromal collagen in breast cancer is highlighted by the link between breast cancer, breast density, and the increased deposition of stromal collagen.[11-15] Interestingly, although mammographic density, which is attributable to collagen content, is one of the largest risk factors for the development of breast tumors, there is currently no clinical intervention based on mammographic density alone. This is due in part to the lack of a clear correlation observed between increased mammographic density and patient outcome. Most of the work to date[16-19] has defined mammographic density as a etiological factor and not as a prognostic factor. Recently, Cil *et al*.[20] explored mammographic density as predictor of local breast cancer recurrence. They reported that women with intermediate and high breast density had a significantly elevated risk to develop a local breast cancer recurrence. However, follow-up clinical trials that incorporate additional risk factors such as obesity are needed to examine the possible prognostic value of mammographic density in large and diverse patient cohorts before using density as a possible clinical target. As well recently there has been an effort to investigate the underlying contributor to mammographic density, focusing on one of the largest components present in the dense stroma, collagen. Several studies have shown a link between collagen remodeling and the invasion and progression of mammary cancer in mouse models.[21-23] Furthermore, there was a link observed between collagen morphology, particularly collagen alignment, and breast cancer patient outcome.[24] Provenzano *et al*.[21] first introduced the so called tumor associated collagen signature (TACS) nomenclature to describe collagen alignment patterns. The TACS phenotypes are currently classified into three groups. TACS-1 describes the standard desmoplastic response of increased collagen deposition surrounding initiating tumor cells. TACS-2 is observed as straightened fibers aligned tangentially around developing tumors, while TAC-3 is seen as radially aligned fibers that facilitate local invasion.[25] Conklin *et al*.[24] qualitatively searched for these patterns in human breast cancer samples and through extensive manual analysis found that the presence of the TACS-3 alignment phenotype was a prognostic indicator for disease free and disease specific survival (DFS and DSS respectively) for invasive breast cancer patients. Our quantitative study, presented here, computationally builds on this previous work by defining an algorithmic model for TACS-3 and applying this model to the same cohort of patients.[24]

Previous collagen alignment studies have largely been facilitated by the development of second harmonic generation (SHG) microscopy techniques, which have the ability to capture high contrast images of the collagen fiber extracellular matrix without the need for exogenous stains.[26-29] The application of SHG imaging in cancer research is growing rapidly. For example, changes in the ratio of the forward SHG (FSHG) to backward propagating SHG signal have been recently linked to breast tumor progression[30] and positive lymph node status.[31] SHG directionality was also used by Ajeti *et al*. to quantify the collagen composition in breast cancer models,[32] while Ambekar *et al*. used Fourier transform and polarization-resolved SHG imaging to differentiate malignant from benign tissues in breast biopsies.[33]

In addition, many new computational techniques are being developed to quantify patterns observed in SHG images. For example, a directional gradient method developed by Altendorf *et al*.[34] provides three-dimensional orientation and radius information about fibers in SHG images. Due to the fibrous nature of the collagen matrix, SHG images are particularly well-suited for the curvelet transform (CT), which is a multiscale, orientation sensitive version of the wavelet transform. The CT[35] and combined fiber tracking methods[36] have been applied to extract fiber orientation, length, curvature and radius from SHG images of collagen. One key feature however that is missing from all of the available image analysis techniques is the ability to incorporate cellular information into the analysis. The interaction between tumor cells and collagen fibers cannot be fully assessed without integration of information about cellular morphology and associated collagen morphology. As well this information is critical for finding regions of interest (ROI) with TACS, an essential task for any type of high-throughput screening where manual searching is not practical. Herein, we describe a computational protocol that achieves this goal by integrating information about collagen fibers from SHG images with information about cells captured through bright field imaging of standard H&E stained slides to perform highly automated, prognostic TACS-3 scoring.

In order for TACS to become a useful and fully validated biomarker, it must be screened for in several large studies containing many patients and diverse populations. In addition, besides screening in heterogeneous populations, it ideally needs to be screened in diverse sample types to account for possible subtle differences in surgery, pathology or sample preparation that could negatively impact sample consistency. This ability to rapidly screen in many sample types of large diverse populations would also open the door for TACS to be explored in other cancer types such as pancreatic and renal cancer. Heretofore, there has not been a method that automates enough of the process to enable such large scale adaptation. In previous studies, collagen fiber angles were measured by hand, one at a time, using ImageJ ROI marking tools.[21,23] These experiments used information gathered a priori or from autofluorescence to identify tumor-stromal boundaries. In addition, imaging locations were chosen manually. Conklin *et al*. manually captured each individual image, used bright field images to manually identify tumor-stromal boundaries, and manually estimated collagen fiber angles.[24] In each of these cases, many subjective decisions were made while identifying which areas to image, which fibers to measure and what should be considered a tumor-stromal boundary. There has been progress made in automating the fiber angle analysis steps of this task.[35-39] However, none of these methods can automate all four steps of the TACS analysis process, which are: (1) Image capture, (2) Fiber angle measurement, (3) tumor-stromal boundary identification, and (4) relative angle measurement between fiber and boundary. In this paper, we use image analysis and supervised learning techniques to enable the automation of each of these tasks. The block diagram of our imaging and analysis protocol is shown in Figure 1. Starting with the previously-imaged invasive breast cancer tissue microarray (TMA), we captured registered, whole-slide SHG and bright field images, extracted fibers from the SHG images, identified tumor-stromal boundaries from the bright field images, and measured relative angles, all in a scripted pipelined process that requires little human intervention. We believe that this method will allow significantly larger scale studies to be performed in order to validate TACS-3 as a prognostic biomarker in breast cancer and potentially other cancer types, and to investigate if TACS-3 can be used to predict patient response to targeted therapies.

METHODS

Human Breast Carcinoma Tissue Microarray

The TMA used here was the same as that used by Conklin *et al*.[24] for the manual collagen alignment

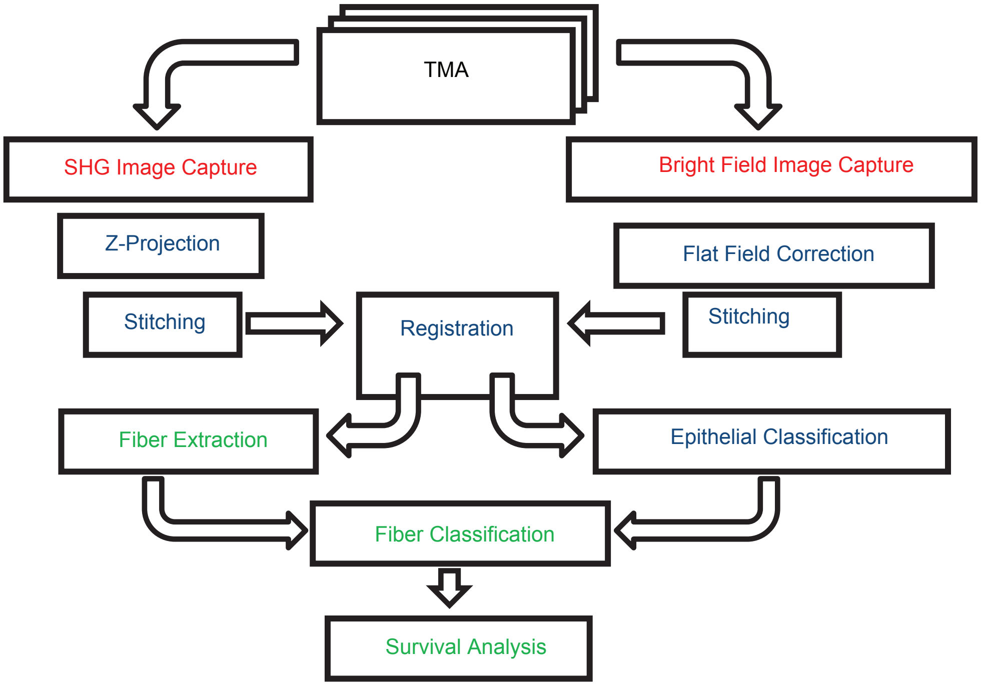

Figure 1: Block diagram of the tumor associated collagen signatures scoring system. The red steps are performed with WiscScan, the blue steps are performed with ImageJ/FIJI, and the green steps are performed with MATLAB based tools. The left side shows the steps performed on the second harmonic generation images, while the right side shows the steps performed on the bright field images. The middle column combines information from both modalities

analysis. The clinical profiles of all patients whose tissue was included in this TMA have been described in a previous study.[39] All tissue and patient information used in this study were acquired following Institutional Review Board approval. Tumor tissues from 353 patients diagnosed with invasive carcinoma were resected by the same surgeon between 1981 and 1995. Pieces of each resected tumor were embedded in paraffin according to standard histopathology protocols. After tumors smaller than 5 mm and severely damaged samples were excluded, 196 patients remained for analysis. Sections of 4 μm thickness were cut from archived TMA blocks containing 1.0 mm diameter tissue cores, placed on glass slides, stained with H&E and mounted under a glass coverslip. Patients were followed for a median of 6.2 years, ranging from 1 month to 18.6 years.

Imaging System

All samples in this study were imaged with the custom built integrated FSHG/bright field imaging system shown in Figure 2. A MIRA 900 Ti: Sapphire laser (Coherent, Santa Clara, CA) tuned to 780 nm, with a pulse length of approximately 100 fs, was directed through a Pockel's cell (ConOptics, Danbury, CT, USA), half and quarter waveplates (ThorLabs, Newton, NJ, USA), beam expander (ThorLabs), a 3 mm galvanometer driven mirror pair (Cambridge, Bedford, MA), a scan/tube lens pair (ThorLabs), through a dichroic beam splitter (Semrock, Rochester, NY) and focused by a 20X/0.75NA objective (Nikon, Melville, NY). SHG light was collected in the forward direction with a 0.54 NA condenser (ThorLabs) and filtered with an interference filter centered at 390 nm with a full width at half maximum bandwidth of 22.4 nm (Semrock). The back aperture of the condenser lens was imaged onto the 5 mm aperture of a 7422-40P photomultiplier tube (Hamamatsu, Hamamatsu, Japan) the signal from which was amplified with a C7319 integrating amplifier (Hamamatsu) and sampled with an analog to digital converter (Innovative Integration, Simi Valley, CA). Timing between the galvo scanners, signal acquisition, and motorized stage positioning was achieved using our custom software called WiscScan.[41] The Rapid Automated Modular Microscope system (Applied Scientific Instrumentation, Eugene, OR) served as our microscope base and we used ASI motorized translation stages for x, y, and z motion control. The SHG light source was verified to be circularly polarized at the sample using the protocol of Chen *et al*.[29] SHG images were captured as stacks of three images spaced 3 μm apart, then z-projected to improve field flatness. Bright field images were captured with the same system using a MCWHL2 white LED lamp (ThorLabs) set up for Kohler illumination. White light from this lamp was separated from SHG light traveling through the condenser assembly using a short pass dichroic mirror with a cutoff at 670 nm (Semrock). An RGB camera (QImaging, Surrey, BC, Canada) was used to capture bright field images through WiscScan to allow for acquisition within a single application. Both SHG and white light images were tiled with 10% overlap using automation provided by WiscScan. Stage positions for individual images and pixel size data were stored in Bio-Formats image metadata[42] and this was then used by the grid/collection stitching ImageJ plugin[43] to reassemble a high-resolution large field of view image of the entire TMA. When capturing large field of view images, the sample plane often

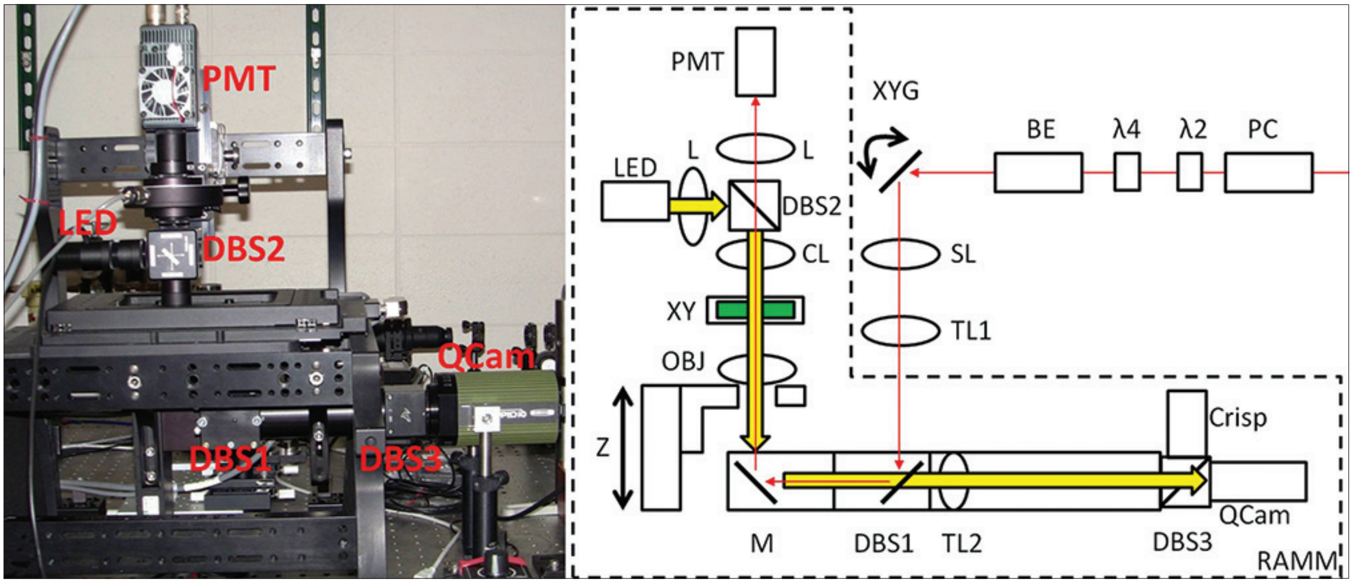

Figure 2: Photo of imaging system (a) and optical block diagram panel (b). PMT = Photomultiplier tube, QCam = QImaging RGB camera, LED = Bright field lamp, DBS = Dichroic beamsplitter, TL = Tube lens, CL = Condenser lens, L = Lens, M = Mirror, BE = Beam expander, Z = Z-direction translation, XY = xy translation, XYG = xy galvonometer driven mirrors, PC = Pockel's cell, RAMM = Rapid Automated Modular Microscope (ASI), λ2, λ4 = half and quarter waveplates

walks out of the in-focus imaging plane as the stage is translated over large distances in x or y. We alleviated this issue using the Continuous Reflection Interface Sampling and Positioning autofocus system (Applied Scientific Instrumentation), which maintained an accurate distance between the coverslip and the objective throughout the whole slide stitched image capture. This allowed for a single bright field image to be captured at each location rather than a z-stack, improving capture speed, reconstruction speed, and reducing production of unnecessary data. After SHG and white light images were captured and stitched, the two modalities were manually registered with the landmark correspondences ImageJ plugin using five control points per image. The image of the entire TMA was registered in a single step, and then each individual TMA core was cropped out of the full TMA image, producing 196 images. The resulting TMA core images were each 2048 × 2048 pixels, consisting of four eight-bit channels. The first three channels represented the red, green and blue planes of the white light image, while the fourth channel contained the SHG information.

Tumor Associated Collagen Signatures-3 Model

Our TACS-3 model was based on previously published observations relating collagen structure to breast cancer progression and survival. In these studies, the first step in the TACS-3 scoring process was the identification of groups of straightened, aligned collagen fibers. The second step was to determine if those fibers terminate at or near regions of epithelial cells at steep angles. If a fiber met both of these criteria, then it was considered TACS-3 positive. If one or more TACS-3 positive fibers were found in a sub-region of an image, then that region was scored TACS-3 positive. The number of regions with TACS-3 positive scores was then used to score the entire image. There were many details in these steps and defining parameters to account for each step would have produced a potentially fragile model. Instead, we have implemented a supervised learning approach that allows the data to most appropriately define the model. We performed this task computationally using a series of cascaded classifiers. The first classifier was trained to find epithelial regions in the images using a small training set of annotated ROIs. The resultant epithelial cell model was then used to segment epithelial cell regions within the entire cohort of images. Features describing the epithelial regions were then combined with features derived from our fiber extraction algorithm and were fed into a second classifier, which was trained to score each image as being TACS-3 positive or negative based on a training set of annotated images. TACS-3 scores were then fed into a cox proportional hazard model to regress to censored survival data.

Our entire image cohort can be represented by a set S of registered SHG and bright field (RGB) images. Each image $I_i(\vec{u},g)\hat{I}S$ was composed of pixels $\vec{u}=(x_i,y_i)\in R^2, 0<x_i,y_i<N$ and the function $g(\vec{u})$ which mapped each pixel to a quartet of intensity values corresponding to the R, G, B, and SHG intensity channels of the image. The first step of our protocol involved the extraction of collagen fiber objects from the SHG channel. Next, epithelial cell clusters were segmented from the R, G, and B channels. Each fiber from the first step was then associated with epithelial cell clusters from the second step to create a feature set for each fiber. Average feature values for all fibers in the image were then used to classify images based on a training set of images.

Fiber Extraction

We applied a technique called CT-FIRE[36] to the SHG images to enhance, trace and extract a network of collagen fibers for each SHG image $I(\vec{u},g_{shg})$. CT-FIRE combines the advantage of the CT[44] for denoising the image and enhancing the fiber ridge features with the advantage of a fiber tracing algorithm[45] for automatic fiber extraction, being capable of extracting fiber geometric information such as length, angle, width, and curvature of each fiber. We applied the fast discrete CT (FDCT) to capture a collection of coefficients C^D in curvelet space, which are defined as the inner product of the input SHG image channel with each of the curvelet basis functions.

$$C^D(j,l,k)=\sum_{0<x_i,y_i<n} I(\vec{u},g_{shg})\,\gamma^D_{jkl}(\vec{u})$$

where $\gamma^D_{jkl}(\vec{u})$ is the digital curvelet waveform and *jkl* represent the scale, orientation, and location indices, respectively. We used the open source FDCT MATLAB (The Mathworks, Natick, MA, USA)[44] library and specifically the "wrapping" version of the FDCT due to its simplicity. To denoise the image, we set all curvelet coefficients to zero that fall below a user defined threshold T as shown below

$$C^D_T(j,l,k)=\left\{C^D\ \underset{0,}{(j,l,k)},\left|C^D\ \underset{\mathit{otherwise}}{(j,l,k)}\right|>T\right.$$

This threshold was determined empirically on a small subset of SHG images to determine the appropriate level of noise reduction. The inverse FDCT was then applied to reconstruct an edge enhanced, noise reduced version of the SHG image. After reconstruction, CT-FIRE traced fibers, using the method of Stein *et al.*,[45] by first finding local maxima in the result of the smoothed distance transform. The distance transform computed the distance from each foreground pixel to the nearest background pixel. Fiber branches were formed by creating regions surrounding each local maxima, the size of which were defined by the result of the distance transform at the location of the local maximum point. The edges of this region were then searched for further local maxima. This process was repeated until no new local maxima were found indicating the end of a fiber branch. Short branches were then pruned from the network and closely spaced, similarly oriented fibers were merged. Fiber width (*FW*) was quantified for each extracted fiber by averaging the fiber widths ($2R_i$) at n points that were used to form the fiber

$$FW = \frac{1}{n}\sum_{i=1}^{n} 2R_i$$

Where R_i is the fiber radius at the i^{th} point, estimated by the result of the distance transform at that location. Fiber straightness (*FS*) was quantified for each extracted fiber by dividing the distance between the end points of the fiber (d_n) by the distance along the path of the fiber (d_0).

$$FS = \frac{d_n}{d_0}$$

Thus for perfectly straight fibers FS=1.0 and wavy fibers FS<1.0. After fiber objects have been extracted from each of the images, we next segment epithelial cell regions.

Epithelial Cell Segmentation

The TACS-3 phenotype consists of straightened aligned collagen fibers that terminate near regions of epithelial cells such that the angles of the collagen fibers appear perpendicular to the epithelial stromal boundary. Detecting this TACS-3 phenotype requires knowledge of the locations of epithelial cells within the sample. We must then identify regions of epithelial cell clusters and identify a boundary between the epithelial cells and surrounding stroma. This task was performed in two steps outlined here. Step 1 used the Trainable Weka Segmentation ImageJ plugin[46] to find epithelial cell nuclei and step 2 applied a cascaded matched filter, threshold operation to identify clusters and boundaries. The details of these steps are given below.

For step 1, a training set of 15 cropped 256 × 256 pixel images denoted as $t_i(\overrightarrow{u_t}, g) \in S$ was created that contains representative features from five classes: Epithelial cell nuclei, other cell nuclei (including lymphocytes and fibroblasts), cytoplasm, collagen, and background. A further subgroup of pixels within the training images $\overrightarrow{u_a} \in \overrightarrow{u_t}$ were annotated as belonging to each class $w_k(\overrightarrow{u_a}), k \in (1,\ldots,5)$. A feature vector $p_i(\overrightarrow{u_t}, g)$ was computed for each pixel and each channel of the training image where $i \in (1\ldots.d)$ was the feature index and d was the dimensionality of the feature subspace. The feature set we used is listed in Table 1 and incorporates features at five scales for a total of 80 feature planes for each image channel. The detailed implementations for each of these features are given in the online documentation for the Weka Segmentation plugin.[47] We then used a multithreaded implementation of the random forests classifier[48,49] with a forest of 200 trees and two random features per node to build a model based on $\overrightarrow{u_a}$ and w_k.

Table 1: Features used in epithelial cell segmentation and TACS-3 fiber classification tasks

Task	Feature description	Total number of features
Epithelial cell classification	Gaussian blur	5
	Sobel filter	5
	Hessian	40
	Difference of Gaussians	24
	Membrane projections	6
TACS-3 fiber classification	Fiber curvature	1
	Fiber width	1
	Fiber length	2
	Fiber density	9
	Fiber alignment	9
	Epithelial proximity	3
	Relative epithelial angle	2

TACS: Tumor associated collagen signatures

The trained model was then applied to every pixel $\vec{u}$ in the cohort producing a probability map for each class and each image using a scripted version of the plugin.

For the second step in the segmentation process, the epithelial class probability map was filtered with a Gaussian filter matched to the average width of the epithelial cell nuclei (three microns) and thresholded such that the top 80% of resulting pixels were retained. The resulting image was then filtered with a Gaussian filter matched to the width of the average sized epithelial cell cluster (25 microns), then finally thresholded such that, again, the top 80% of resulting pixels were retained. Following the final threshold step, regions smaller than 50 pixels in area were discarded and a mask was generated with epithelial cell clusters in the foreground and all else in the background. Epithelial mask pixels are represented here as e_i while epithelial region boundary pixels were created using an eight-connected neighborhood and are denoted as b_i.

Mask images were saved as tiff files and read, along with the extracted fiber data, into the custom, open source CurveAlign software, described more below, for fiber/epithelial region feature extraction. Outlines of the resulting mask files were overlaid onto the original white light images to qualitatively validate the segmentation accuracy of the applied epithelial region model.

Combined Fiber-Epithelial Features and Fiber Classification

In the sections above, we described our methods for epithelial cluster segmentation and collagen fiber extraction. With these two pieces of information, we associated fibers with epithelial cell clusters and measured the interaction between the two using the features described here. This task was performed by an open source, MATLAB based tool called CurveAlign.[35] This tool started by reading in a fiber database file (generated by CT-FIRE) and an epithelial mask file (generated by our epithelial segmentation script). A feature vector p_i was then built for each fiber endpoint $v_i \in R^2$ in the image. The feature vector was populated directly with features derived above in the fiber extraction section including fiber length, curvature, radius and grey level. Both endpoints were given the same values for these single fiber derived features. The rest of the features were unique to each fiber end point. All features used in TACS-3 fiber classification are listed in Table 1. Many of the features in this section rely heavily on the nearest neighbor search routine which is formulated here as

$$\varphi^n(X,Q) = {}^{\arg\min}_{X \in D}\, \rho(X,Q)$$

Where $D = \{X_1,\ldots,X_n\}$ is a set of vectors in R^2,Q is a query vector, $\rho(X,Q)$ is the Euclidean norm $\|X-Q\|_2$, and the result $\varphi^n(X,Q)$ is a vector of n points in X that are nearest to each point in Q. Given a collection of points on a two-dimensional plane, if we select a query point, this algorithm will return the nearest neighbors within our collection of points. For example, a fiber end point can be used as the query point and we can use this algorithm to search for the nearest point in the list of epithelial cell boundary points, as described below. In addition, some of the features compute a metric for alignment using vector addition according to the following algorithm

$$\sigma(Q_X) = \frac{1}{n}\left|\sum_X \exp(i \times 2\theta)\right|$$

where $\theta \in \{0,\ldots,\pi\}$ is a vector of orientations associated with the vector of n positions in X. The factor of two is included since we used fiber orientations supported from $0 to \pi$ rather than full $0 to 2\pi$ oriented direction vectors. In words, the alignment metric is calculated as the normalized vector sum of orientation vectors. The larger the vector sum, the more aligned the group of fibers. On the other hand, if the vector sum is small, then the group of fibers is more randomly oriented. Fiber density features were computed as the average distance from the current fiber endpoint to the n =2, 4, 8, and 16 nearest neighbors. The density features for fiber endpoint v are therefore given by:

$$fd_{n,v} = \frac{1}{n}\sum_j \rho(v - \vartheta_j^n(v_i, v_j))$$

If fiber density is higher, then this result will be lower, since it is measuring the average distance between fibers. Fiber alignment features were computed as the absolute values of the vector sum of the nearest neighbor fiber endpoints and are given by $fa_{n,v} = \sigma([\theta_v, \theta_{\vartheta_j^n(v_{i,v})}])$ for fiber endpoint . In the equation for $fa_{n,v}$,$[..]$ indicates vector concatenation and n =2, 4, 8, and 16. Features that incorporate epithelial cell information included distance to nearest epithelial mask point $de_v = \rho(v - \varphi^1(e_i, v))$ and distance to nearest epithelial region boundary $db_v = \rho(v - \varphi^1(b_i, v))$. These features had the same value if the fiber end point was outside an epithelial cell region; however, they were different if the end point was colocal with an epithelial region. Next, we extracted relative angle features. Angle with respect to nearest epithelial region boundary point was computed as $ab_v = \sigma([\theta_v, \theta_{\varphi(b_{i,v})}])$, and angle with respect to nearest "extension boundary intersection point" was given by $ae_v^r = \sigma([\theta_v, \theta_{\varphi(l_i^r, v)}])$, where the set of points in l^r was computed by taking the intersection of all epithelial boundary points b_i and a line of length $2r$ extending from the fiber endpoint at an angle θ and is formulated below

$$l^r = b_i \cap bres(v-q, v+q)$$

where *bres* indicates a modified Bresenham algorithm[50] which is used to find all pixels along a line between two points. The term $q = [r\exp(\theta v)]$ is the offset from v in the x and y directions. For this last feature, three values of r (50, 100 and 200 μm) were calculated. These three lengths corresponded to 5, 10, and 20 times the diameter of a typical epithelial cell and were selected based on estimates of intercellular signaling distances.[51] If no intersection was found, then the ae_v^r feature value was set to zero. The angle of the tumor stromal boundary line $\theta_{\varphi(l_i^r, v)}$ was estimated by fitting a quadratic to nine contiguous points on the boundary surrounding the intersection point (or nearest boundary point in the previous feature) and computing the tangent angle of the line fit at the midpoint. The steps in the process of relative angle feature extraction are diagrammed in Figure 3.

Each of these fiber level features p_i were calculated for every fiber endpoint $v_i$in the cohort. Fiber level

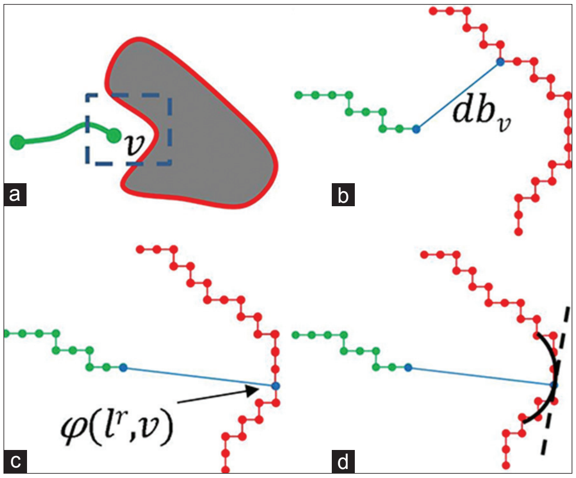

Figure 3: Integrated fiber angle and epithelial boundary feature algorithms. Panel A shows a single fiber (green) and epithelial region boundary (red) with one highlighted fiber endpoint. Zoomed versions of panel a are shown in panels b, c and d where individual image pixels are represented as filled circles. The nearest distance from to the boundary is indicated in panel B, intersection between the endpoint extension line and the boundary is shown in panel C and the quadratic curve fit to the boundary at the intersection point and tangent line are shown in panel D

features were then averaged among all fibers in a given image and training was performed with a subset of 16 images $I_t \in I_i$ that had been manually annotated as being TACS-3 positive or negative. A linear support vector machine (SVM) was used to build a model, which was then applied to all 196 images in the cohort for classifying each image as being TACS-3 positive or negative.

Classification and Survival Analysis

The TACS-3 scores were correlated with DFS and DSS data using the Cox-proportional-hazards regression method.[52] DFS was defined as the time from date of diagnosis to the first date of recurrence and DSS was defined as the time from diagnosis to death from breast cancer or date of last follow-up evaluation. In both cases, all other events were censored. The Kaplan-Meier method was used to compare DFS and DSS between TACS-3 negative and TACS-3 positive patients. Hazard ratios were computed using a log-rank test. Correlations between manual and computationally generated TACS-3 scores were made using the Pearson's linear correlation coefficient.

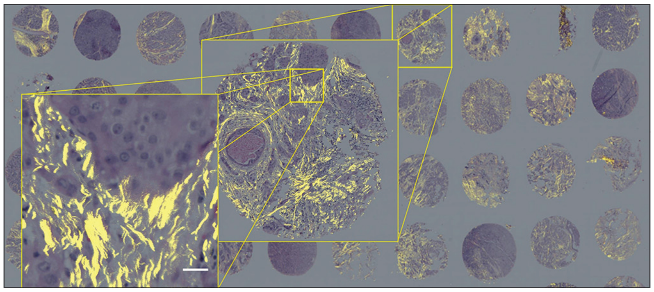

Figure 4: The forward second harmonic generation image was overlaid upon the bright field image for an entire H&E stained tissue microarray slide. Scale bar = 25 μm

RESULTS

Registered SHG and bright field images of a subsample of the TMA are shown in Figure 4 along with two zoomed versions of regions within the image. SHG information is added as an alpha channel on top of the raw RGB bright field image and pseudo colored yellow. The fully zoomed panel shows the detail available in the full resolution images captured with the 20X, 0.75 NA lens and shows a region with a positive TACS-3 signature. A collection of three more TACS-3 positive and three TACS-3 negative regions were cropped out of the TMA images and shown in Figure 5. These images illustrate the features that are common to the TACS-3 signature including straightened, aligned fibers terminating in or near regions of epithelial cells at near perpendicular angles with respect to the epithelial region border. In addition, the TACS-3 negative cases show wavy fibers, fibers that terminate at adipose tissue, and a curved fiber encapsulating an epithelial cell cluster [Figure 5d-f, respectively].

A sample of our fiber extraction and epithelial region segmentation results are shown in Figures 6 and 7, respectively. In both cases, epithelial region segmentation and fiber extraction were observed to accurately represent the data. The orientations of the epithelial cell region boundaries were compared to collagen fiber angles derived from the results of a fiber object extraction algorithm CT-FIRE which has been shown to perform well in

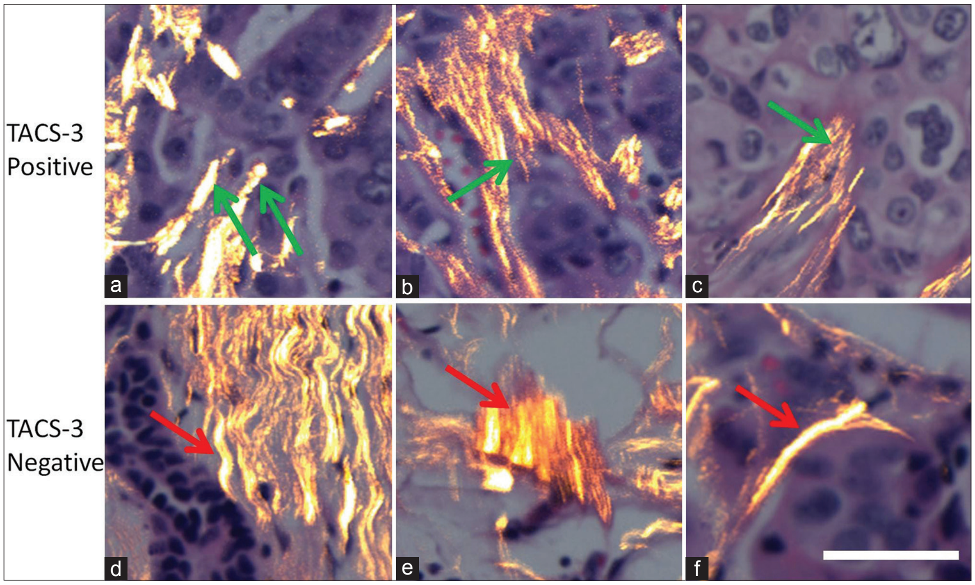

Figure 5: Examples of tumor associated collagen signatures (TACS-3) positive and negative regions within our cohort of breast cancer samples. Collagen fibers (yellow) registered and overlaid onto bright field images of an H&E stained tissues. These images illustrate the features of the TACS-3 signature (green arrows, top row), particularly, straightened aligned collagen fibers that terminate at steep angles relative to epithelial region boundaries (a-c). Three TACS-3 negative cases are shown for comparison (red arrows, bottom row), where collagen fibers are wavy (d), terminate at adipose tissue (e), and wrap around epithelium (f) scale bar = 25 μm

comparison to other techniques.[36] A representative sample of the results produced by this algorithm are shown in Figure 6. The intermediate product after the CT denoising step is shown in Figure 6a, while the extracted fiber network is shown overlaid on the original SHG image as shown in Figure 6b. Although some fibers are over-or under-segmented (annotated by green arrows), most of the extracted fibers properly represent the data. Figure 7 clearly demonstrates the ability of our epithelial cell segmentation algorithm to properly classify many of the regions of epithelial cells as positive. However, a few small regions of stromal fibroblasts and endothelial cells are included in the epithelial cell regions (annotated by green arrows). Although these errors occurred occasionally throughout the cohort, the noise they generated did not overcome the TACS-3 signal. Another feature evident in Figure 7d is the smoothness of the epithelial region boundaries. The boundary smoothness was dependent on the selection of our filter widths and binary mask thresholds. These parameters were selected to accurately represent the boundary orientation at the spatial scale of the epithelial cell regions.

Although correlation with survival is our ultimate goal, automated TACS-3 scores should also correlate with manual scores for each of the images. The Pearson linear correlation coefficient was used to determine this correlation, the results of which are tabulated in Table 2. The manual analysis performed by Conklin *et al*. produced three scores. Score 1 was the number of TACS-3 positive regions divided by the total number of regions analyzed, score 2 was the average number of TACS-3 positive votes per region among three observers, and score 3 indicated if one or more region received a TACS-3 positive rating. Table 2 shows positive correlation between all manual scoring methods and our computational scoring system presented here, with the highest correlation observed to be with manual score 2.

The Kaplan-Meier curves in Figure 8 demonstrate the prognostic potential of our TACS-3 scoring system. TACS-3 negative patients showed significantly better disease-free and disease-specific survival compared to TACS-3 positive patients. In addition, Cox proportional hazard regression showed significant correlation between our computationally generated TACS-3 scores and survival as listed in Table 3. We also correlated scores created by individual fiber feature metrics with survival. Although fiber features alone were correlated with survival, the highest correlation was observed when the TACS-3 scores were composed of multiple integrated fiber/epithelial features. This result shows that a multimodality imaging and analysis approach that

Table 2: Correlation between manual and supervised learning approaches

Scoring method	Correlation coefficient	*P* value
Manual score 1	0.295	2.7E-5
Manual score 2	0.311	0.9E-5
Manual score 3	0.271	1.2E-5

There is statistically significant positive correlation between the supervised learning approach and all manual scoring approaches

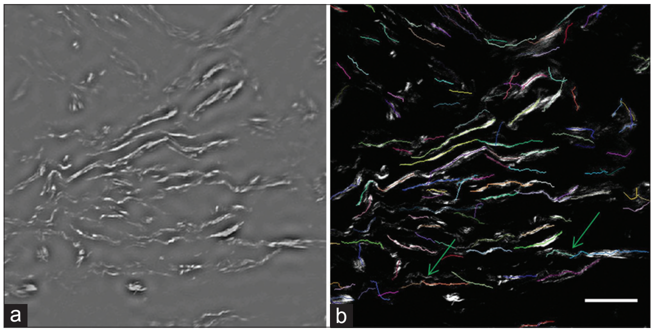

Figure 6: Sample fiber extraction results. The resulting image after curvelet denoising (a) shows likely fiber pixels in white and likely background pixels in grey. The extracted fiber network is overlaid on the original second harmonic generation image (b) with many appropriate segmentations and a few under-and over-segmentations (green arrows). Scale bar = 50 μm

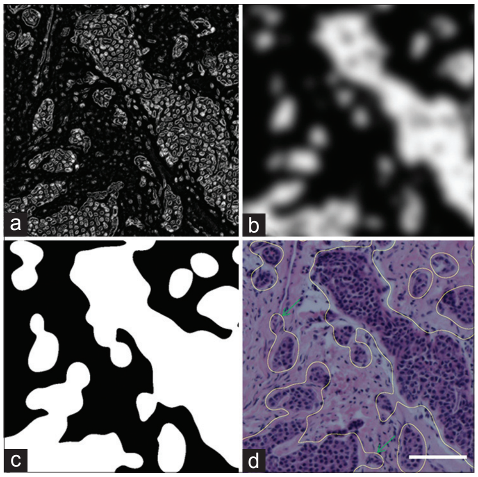

Figure 7: Sample epithelial cell segmentation results. The raw probability map produced by the trainable Weka Segmentation ImageJ plugin (a) is filtered to estimate epithelial cluster density (b) and thresholded (c) to produce epithelial region boundaries which are overlaid onto original bright field images to validate the segmentation (d) scale bar = 100 μm

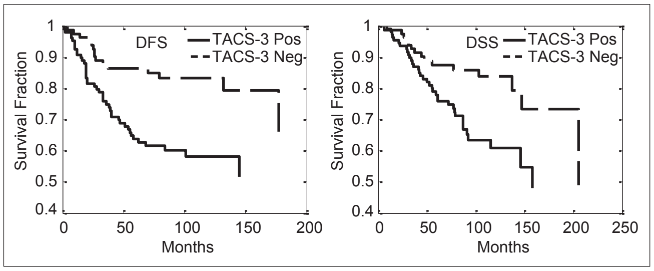

Figure 8: KM curves for disease free survival and disease specific survival showing the prognostic classification produced by our supervised learning based tumor associated collagen signatures-3 scoring approach

combines features of not only collagen fibers, but both collagen fibers and cellular structures is most likely to succeed in predicting survival.

Table 4 lists the 14 most informative features in the TACS-3 scoring process ranked according to their weight produced by the linear SVM algorithm. The SVM weight was used to assess, which features were more or less informative in the classification. Of particular interest are the features labeled as "nearest distance to boundary" and "inside epithelial region". These features indicate the proximity between fibers and epithelial cell regions and were highly important in the TACS-3 classification. In addition, the difference in mean feature scores $d_f = f_p - f_n$ for the training set is shown in Table 4 for each of the ranked features. If d_f is >0, then the TACS-3 positive images had larger values for those features and if d_f is < 0, then the TACS-3 negative images had larger values. For example, the density features resulted in lower d_f values in the TACS-3 positive cases indicating that the TACS-3 positive images had lower density collagen fibers. On the other hand, d_f was positive for the alignment features indicating that TACS-3 positive images tended to have more aligned fibers. Interestingly, relative boundary angle was not as highly informative as many other features; however, still was ranked within the top 14 of 27 features.

Table 3: Univariate Cox proportional hazard analysis results for various feature combinations

Feature type	DFS		DSS	
	Hazard ratio	*P* value	Hazard ratio	*P* value
Fiber curvature	1.432	0.179	1.657	0.077
Fiber density	2.195	0.003	1.831	0.032
Fiber alignment	1.958	0.011	1.588	0.100
TACS-3 score	2.588	0.002	2.250	0.008

The TACS-3 scoring method that includes both fiber and cellular information produces the more significantly prognostic scores compared to fiber information alone. TACS: Tumor associated collagen signatures, DFS: Disease free survival, DSS: Disease specific survival

Table 4: Feature ranking based on SVM feature weight for a 16 patient (8 TACS-3 positive and 8 TACS-3 negative) training set including the average feature value difference between the positive and negative training cases f_p-f_n

Feature description	*w*	f_p-f_n
Standard nearest align	0.716	–0.004
Nearest distance to bound	0.491	–0.086
Inside epithelial region	0.393	0.572
Standard nearest distance	0.348	0.211
Mean nearest distance	0.238	0.188
Curvature	0.219	0.010
Box density 128	0.190	–0.311
Width	0.183	–0.050
Box alignment 64	0.155	0.136
Box alignment 32	0.150	0.079
Box alignment 128	0.111	0.145
Box density 64	0.109	–0.249
Nearest relative boundary angle	0.087	0.004
Total length	0.079	0.004

SVM: Support vector machine, TACS: Tumor associated collagen signatures. These values show the trend of the feature values between TACS-3 positive and TACS-3 negative

DISCUSSION AND CONCLUSIONS

The search for new prognostic and predictive breast cancer biomarkers is motivated by the need to improve patient outcome. A significant number of patients present with none of the currently available markers. In addition, survival and treatment response is often heterogeneous among patients within current biomarker classifications. The discovery and validation of new biomarkers will help to further improve breast cancer diagnosis and treatment planning. These new biomarkers need to be quantifiable, scalable and ideally correlate with both disease outcome and treatment specific response. The candidate biomarker we are focused on in this study (TACS-3) measures collagen alignment relative to tumor-stromal boundaries and has been associated with progression in mouse models and has been shown to predict disease recurrence and survival in human patients. Here, we demonstrate a protocol for using large field of view imaging techniques, image analysis and supervised learning to automate and quantify all of the steps in the process of TACS-3 scoring. These advances provide the tools for increasing the scale of TACS-3 investigations and applying TACS-3 scoring to cancers in other tissues such as ovarian[53] and pancreatic cancer[54,55] where collagen fiber characteristics are predicted to correlate with prognosis. These techniques could also be used to characterize other TACS both current and yet to be identified to see if they have research value in animal models or prognostic value in clinical specimens.

Tumor associated collagen signatures (TACS) analysis requires the simultaneous analysis of information about epithelial cells and extracellular collagen. The interactions between collagen and cells can only be assessed computationally if the cellular information is carefully registered with images of the collagen. We have therefore optimized our imaging system for highly automated capture of large fields of view, registered SHG and bright field images of stained microscope slides with the purpose of analyzing collagen angle with respect to cell cluster boundaries. For this paper, we originally planned to use the same SHG and bright field images captured by Conklin *et al.*[24] since these were already manually annotated. Unfortunately, these images contained artifacts, which, although trivial for the human visual perception system to overcome, were extremely difficult for our computational systems to handle effectively. For example, SHG images were originally captured in the backwards direction with elliptically polarized light, causing two artifacts. The first was simply a low signal to noise level due to few SHG photons traveling in the backward direction from the thin tissue sections.[56,57] The second artifact was observed as a larger relative SHG signal from fibers in the direction parallel to the long axis of the laser polarization ellipse.[58] Artifacts in the bright field images included significant vignetting at field edges and low signal to noise due to short exposure times. These artifacts were easily hurdled by the human observers making TACS-3 assessments in a previous study.[24] However, they are particularly difficult to handle by a computer vision based approach. We therefore decided to develop an optimized imaging system and protocol that would fix many of these artifacts and allow for more consistent automated imaging. Similar image quality and consistency can be achieved with other SHG microscopes including commercial systems with the appropriate hardware, but our analysis protocol did identify a necessary rigorous acquisition protocol that is best achieved with our new automated SHG microscope. In general, the system should allow for FSHG and bright field imaging with a field of view as large and flat as possible, numerical aperture of at least 0.75, automated xyz motion control with appropriate position logging, circular polarization at the sample for SHG imaging, autofocus and automated switching between SHG and bright field imaging.

Our system of imaging and analysis to produce prognostic TACS-3 scores uses standard histopathology H and E slide preparations. The technique is therefore completely compatible with routine clinical protocols and is intended to augment currently available diagnostic tests. The current process requires no changes to current clinical protocol and the sample is returned to the clinician unmodified. We present a system that uses SHG imaging to capture collagen fiber images; however, wide field polarization sensitive techniques[59] such as LC-PolScope[60] or Picrosirius red staining[61,62] might be used to alternatively capture images of collagen fibers. One advantage of using SHG is that it does not require additional stains and can capture three-dimensional fiber information in thick, unstained tissue samples. Unfortunately, when imaging in thick unstained tissue, the identification of epithelial regions can be difficult; however, techniques using autofluoresence and fluorescent lifetime imaging have been shown to be capable of this task.[63,64] As implemented here, our TACS-3 scoring algorithm is necessarily two-dimensional, since we are relying on H&E stained slides for our epithelial cell information. However, fiber extraction, epithelial region segmentation and relative angle measurements can be extended to three-dimensions without significant alteration of our general protocol. In addition, although our current TACS-3 scoring protocol is able to process standard H&E stains, staining for epithelial cells, with, for example, pan-cytokeratin conjugated stains, may simplify and improve epithelial cell segmentation. Future methods may also be adapted to segment clusters of fibroblasts, macrophages and other stromal cells, whose proximity and relative morphological structure with respect to surrounding collagen fibers may further improve correlation with survival or metastatic potential.

Collagen alignment related image features are interesting not only because they have been shown to be prognostic, but because they have been shown to be directly linked to cancer biology. Researchers have found that cells are more likely to invade along parallel, aligned collagen fibers,[25,65] features that are directly being measured by our system. Access to the breadth of fiber data available with our techniques could lead to advances in our understanding of these biological phenomena. Relevant feature sets are not always available with other machine vision systems developed for biological image classification. For example, although WNDCHRM[5] is an extremely powerful image classification tool, informative image features often do not relate to the biology at hand. In the case of our TACS-3 analysis system, biological observations have driven the image analysis model; therefore, features are more easily linked back to biological functions potentially revealing new insights.

High mammographic density is one of the largest risk factors for the development of breast cancer and has been associated with increased epithelial cellularity and increased collagen density.[12,14,19] Increased collagen density has been observed to promote tumor progression in a mouse tumor model[23] and in node positive breast cancer[31] leading one to potentially conclude that collagen density causes elevated risk. However, Maller *et al*.[66] observed that high density, nonfibrillar collagen protected against tumor progression and alternatively, that linearized collagen fibers induced invasive cellular behavior. In agreement with these recent findings, we observe here that TACS-3 fibers are more commonly present in regions of lower fiber density and are more likely to be thinner, more linearized fibers. Thick, curvy, and denser collagen fibers are unlikely to contain TACS-3 fibers and are observed to be associated with a better prognosis. These observations support the hypothesis that collagen fiber shape and organization is a key aspect of the invasive extracellular matrix (ECM) phenotype.

The imaging instrumentation presented here consists of a relatively compact and highly automated multiphoton microscope with an integrated bright field slide scanner. The system has been optimized to capture registered whole slide images of both bright field and SHG images of histopathology slides by imaging each small × 20 field of view and automatically aligning and stitching each image together. Capturing large fields of view in this manner allows for a more thorough and consistent data collection potentially reducing sampling bias and supporting pipelined computational image analysis. The registration of cellular with extracellular collagen information provided by our system allows for the quantitative analysis of key relative structural features between collagen fibers and cancer cell clusters. In addition to SHG, our multiphoton system is capable of imaging other endogenous fluorophores such as nicotinamide adenine dinucleotide (NADH) or flavin adenine dinucleotide (FAD) as well as any of the routine exogenous multiphoton probes used to stain tissue.

CONCLUSION

We present an imaging and analysis protocol that uses high content imaging techniques coupled with supervised learning to perform semi-automated TACS-3 scoring of slide mounted biopsy samples. We apply our technique to a previously annotated TMA containing tissue from 207 patients with invasive breast cancer. The resulting scores are shown to positively correlate with manual annotations and to predict patient outcome with good statistical significance. Future work will attempt to validate this technique on larger cohorts of breast cancer patients, to study ECM targeted drug responses in animal models, and to study collagen alignment in other cancers. As well, future work will focus on improving the clinical application of these techniques so they can be run by untrained clinical personnel and be run at the time of acquisition to find ROIs and TACS within those regions automatically. Together with more automation, TACS screening has great potential as a clinical diagnostic tool that can provide relevant prognostic information from large numbers of tissue samples.

ACKNOWLEDGMENTS

We would like to acknowledge assistance and valuable discussions with members of the Keely, Medical Devices and LOCI groups. We also acknowledge funding from NIH R01 Grants CA114462 and CA136590 to P.J.K and K.W.E, T32 CA009206 to J.S.B., and the Morgridge Institute for Research. The authors are grateful to Johannes Schindelin, Curtis Rueden, Andreas Velten, Ilya Goldberg, and Paolo Provenzano for their helpful technical suggestions and Andreas Friedl for providing access to the TMA. As well we wish to acknowledge the help of the WiscScan programmers at LOCI, Ajeet Vivekanandan, David Mayer and Mohit Chainani for their assistance in adding support for new functionality in WiscScan and debugging current features.

REFERENCES

1. Allred DC, Harvey JM, Berardo M, Clark GM. Prognostic and predictive factors in breast cancer by immunohistochemical analysis. Mod Pathol 1998;11:155-68.
2. Press M, Slamon D, Cobleigh M, Vogel C, Zhou JY, Anderson S, *et al.* Improved clinical outcomes for herceptin (R)-treated patients selected by fluorescence *in situ* hybridization (FISH). Mod Pathol 2002;15:47A.
3. Habel LA, Shak S, Jacobs MK, Capra A, Alexander C, Pho M, *et al.* A population-based study of tumor gene expression and risk of breast cancer death among lymph node-negative patients. Breast Cancer Res 2006;8:R25.
4. Nahta R, Esteva FJ. HER2 therapy: Molecular mechanisms of trastuzumab resistance. Breast Cancer Res 2006;8:215.
5. Shamir L, Delaney JD, Orlov N, Eckley DM, Goldberg IG. Pattern recognition software and techniques for biological image analysis. PLoS Comput Biol 2010;6:e1000974.
6. Madabhushi A, Agner S, Basavanhally A, Doyle S, Lee G. Computer-aided prognosis: Predicting patient and disease outcome via quantitative fusion of multi-scale, multi-modal data. Comput Med Imaging Graph 2011;35:506-14.
7. Myers G. Why bioimage informatics matters. Nat Methods 2012;9:659-60.
8. Rønnov-Jessen L, Petersen OW, Koteliansky VE, Bissell MJ. The origin of the myofibroblasts in breast cancer. Recapitulation of tumor environment in culture unravels diversity and implicates converted fibroblasts and recruited smooth muscle cells. J Clin Invest 1995;95:859-73.
9. Elenbaas B, Spirio L, Koerner F, Fleming MD, Zimonjic DB, Donaher JL, *et al.* Human breast cancer cells generated by oncogenic transformation of primary mammary epithelial cells. Genes Dev 2001;15:50-65.
10. Tlsty TD, Hein PW. Know thy neighbor: Stromal cells can contribute oncogenic signals. Curr Opin Genet Dev 2001;11:54-9.
11. Boyd NF, Martin LJ, Yaffe MJ, Minkin S. Mammographic density and breast cancer risk: Current understanding and future prospects. Breast Cancer Res 2011;13:223.
12. Guo YP, Martin LJ, Hanna W, Banerjee D, Miller N, Fishell E, *et al.* Growth factors and stromal matrix proteins associated with mammographic densities. Cancer Epidemiol Biomarkers Prev 2001;10:243-8.
13. Boyd NF, Martin LJ, Sun L, Guo H, Chiarelli A, Hislop G, *et al.* Body size, mammographic density, and breast cancer risk. Cancer Epidemiol Biomarkers Prev 2006;15:2086-92.
14. Boyd NF, Guo H, Martin LJ, Sun L, Stone J, Fishell E, *et al.* Mammographic density and the risk and detection of breast cancer. N Engl J Med 2007;356:227-36.
15. Boyd NF, Martin LJ, Bronskill M, Yaffe MJ, Duric N, Minkin S. Breast tissue composition and susceptibility to breast cancer. J Natl Cancer Inst 2010;102:1224-37.
16. Thurfjell E. Breast density and the risk of breast cancer. N Engl J Med 2002;347:866.
17. Habel LA, Dignam JJ, Land SR, Salane M, Capra AM, Julian TB. Mammographic density and breast cancer after ductal carcinoma in situ. J Natl Cancer Inst 2004;96:1467-72.
18. Boyd NF, Rommens JM, Vogt K, Lee V, Hopper JL, Yaffe MJ, *et al.* Mammographic breast density as an intermediate phenotype for breast cancer. Lancet Oncol 2005;6:798-80.
19. McCormack VA, dos Santos Silva I. Breast density and parenchymal patterns as markers of breast cancer risk: A meta-analysis. Cancer Epidemiol Biomarkers Prev 2006;15:1159-69.
20. Cil T, Fishell E, Hanna W, Sun P, Rawlinson E, Narod SA, *et al.* Mammographic density and the risk of breast cancer recurrence after breast-conserving surgery. Cancer 2009;115:5780-7.
21. Provenzano PP, Eliceiri KW, Campbell JM, Inman DR, White JG, Keely PJ. Collagen reorganization at the tumor-stromal interface facilitates local invasion. BMC Med 2006;4:38. 22. Provenzano PP, Eliceiri KW, Yan L, Ada-Nguema A, Conklin MW, Inman DR, *et al.* Nonlinear optical imaging of cellular processes in breast cancer. Microsc Microanal 2008;14:532-48.
23. Provenzano PP, Inman DR, Eliceiri KW, Knittel JG, Yan L, Rueden CT, *et al.* Collagen density promotes mammary tumor initiation and progression. BMC Med 2008;6:11.
24. Conklin MW, Eickhoff JC, Riching KM, Pehlke CA, Eliceiri KW, Provenzano PP, *et al.* Aligned collagen is a prognostic signature for survival in human breast carcinoma. Am J Pathol 2011;178:1221-32.
25. Provenzano PP, Inman DR, Eliceiri KW, Trier SM, Keely PJ. Contact guidance mediated three-dimensional cell migration is regulated by Rho/ROCK-dependent matrix reorganization. Biophys J 2008;95:5374-84.

26. Zipfel WR, Williams RM, Christie R, Nikitin AY, Hyman BT, Webb WW. Live tissue intrinsic emission microscopy using multiphoton-excited native fluorescence and second harmonic generation. Proc Natl Acad Sci U S A 2003;100:7075-80.
27. Zipfel WR, Williams RM, Webb WW. Nonlinear magic: Multiphoton microscopy in the biosciences. Nat Biotechnol 2003;21:1369-77.
28. Williams RM, Zipfel WR, Webb WW. Interpreting second-harmonic generation images of collagen I fibrils. Biophys J 2005;88:1377-86.
29. Chen X, Nadiarynkh O, Plotnikov S, Campagnola PJ. Second harmonic generation microscopy for quantitative analysis of collagen fibrillar structure. Nat Protoc 2012;7:654-69.
30. Burke K, Tang P, Brown E. Second harmonic generation reveals matrix alterations during breast tumor progression. J Biomed Opt 2013;18:31106.
31. Kakkad SM, Solaiyappan M, Argani P, Sukumar S, Jacobs LK, Leibfritz D, *et al*. Collagen I fiber density increases in lymph node positive breast cancers: Pilot study. J Biomed Opt 2012;17:116017.
32. Ajeti V, Nadiarnykh O, Ponik SM, Keely PJ, Eliceiri KW, Campagnola PJ. Structural changes in mixed Col I/Col V collagen gels probed by SHG microscopy: Implications for probing stromal alterations in human breast cancer. Biomed Opt Express 2011;2:2307-16.
33. Ambekar R, Lau TY, Walsh M, Bhargava R, Toussaint KC Jr. Quantifying collagen structure in breast biopsies using second-harmonic generation imaging. Biomed Opt Express 2012;3:2021-35.
34. Altendorf H, Decencière E, Jeulin D, De sa Peixoto P, Deniset-Besseau A, Angelini E, *et al*. Imaging and 3D morphological analysis of collagen fibrils. J Microsc 2012;247:161-75.
35. Pehlke C, Bredfeldt JS, Doot J, Sung KE, Provenzano P, Riching K, *et al*. Quantification of collagen architecture using the curvelet transform. Integrative Biology, in Review; January 2014.
36. Bredfeldt JS, Liu Y, Pehlke CA, Conklin MW, Szulczewski JM, Inman DR, *et al*. Computational segmentation of collagen fibers from second-harmonic generation images of breast cancer. J Biomed Opt 2014;19:16007.
37. Falzon G, Pearson S, Murison R. Analysis of collagen fibre shape changes in breast cancer. Phys Med Biol 2008;53:6641-52.
38. Rubbens MP, Driessen-Mol A, Boerboom RA, Koppert MM, van Assen HC, TerHaar Romeny BM, *et al*. Quantification of the temporal evolution of collagen orientation in mechanically conditioned engineered cardiovascular tissues. Ann Biomed Eng 2009;37:1263-72.
39. Bayan C, Levitt JM, Miller E, Kaplan D, Georgakoudi I. Fully automated, quantitative, noninvasive assessment of collagen fiber content and organization in thick collagen gels. J Appl Phys 2009;105:102042.
40. Baba F, Swartz K, van Buren R, Eickhoff J, Zhang Y, Wolberg W, *et al*. Syndecan-1 and syndecan-4 are overexpressed in an estrogen receptor-negative, highly proliferative breast carcinoma subtype. Breast Cancer Res Treat 2006;98:91-8.
41. Eliceiri K, Nazir M. Wiscscan, 2012. Available: http://www.loci.wisc.edu/software/wiscscan.[2012 Apr 04]
42. Linkert M, Rueden CT, Allan C, Burel JM, Moore W, Patterson A, *et al*. Metadata matters: Access to image data in the real world. J Cell Biol 2010;189:777-82.
43. Preibisch S, Saalfeld S, Tomancak P. Globally optimal stitching of tiled 3D microscopic image acquisitions. Bioinformatics 2009;25:1463-5.
44. Candes E, Demanet L, Donoho D, Ying L×. Fast discrete curvelet transforms. Multiscale Model Simul 2006;5:861-99.
45. Stein AM, Vader DA, Jawerth LM, Weitz DA, Sander LM. An algorithm for extracting the network geometry of three-dimensional collagen gels. J Microsc 2008;232:463-75.
46. Schindelin J, Arganda-Carreras I, Frise E, Kaynig V, Longair M, Pietzsch T, *et al*. Fiji: An open-source platform for biological-image analysis. Nat Methods 2012;9:676-82.
47. Ignacio AC, Kaynig V, Schindelin J. Trainable Weka Segmentation. Available: http://www.fiji.sc/Trainable_Weka_Segmentation.[2013 Oct 25].
48. Breiman L. Random forests. Mach Learn 2001;45:5-32.
49. Criminisi A, Shotton J, Konukoglu E. Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning. Found Trends Comput Graph Vision 2011;7:81-227.
50. Bresenham JE. Algorithm for computer control of a digital plotter. Ibm Syst J 1965;4:25-30.
51. Francis K, Palsson BO. Effective intercellular communication distances are determined by the relative time constants for cyto/chemokine secretion and diffusion. Proc Natl Acad Sci U S A 1997;94:12258-62.
52. Cox DR. Regression models and life-tables. J R Stat Soc Series B Stat Methodol 1972;34:187.
53. Nadiarnykh O, LaComb RB, Brewer MA, Campagnola PJ. Alterations of the extracellular matrix in ovarian cancer studied by Second Harmonic Generation imaging microscopy. BMC Cancer 2010;10:94.
54. Drifka CR, Eliceiri KW, Weber SM, Kao WJ. A bioengineered heterotypic stroma-cancer microenvironment model to study pancreatic ductal adenocarcinoma. Lab Chip 2013;13:3965-75.
55. Hu W, Zhao G, Wang C, Zhang J, Fu L. Nonlinear optical microscopy for histology of fresh normal and cancerous pancreatic tissues. PLoS One 2012;7:e37962.
56. Cox G, Kable E, Jones A, Fraser I, Manconi F, Gorrell MD 3-dimensional imaging of collagen using second harmonic generation. J Struct Biol 2003;141:53-62.
57. Lacomb R, Nadiarnykh O, Townsend SS, Campagnola PJ. Phase Matching considerations in Second Harmonic Generation from tissues: Effects on emission directionality, conversion efficiency and observed morphology. Opt Commun 2008;281:1823-32.
58. Stoller P, Kim BM, Rubenchik AM, Reiser KM, Da Silva LB. Polarization-dependent optical second-harmonic imaging of a rat-tail tendon. J Biomed Opt 2002;7:205-14.
59. Kliger DS, Lewis JW, Randall CE. Polarized Light in Optics and Spectroscopy. New York: Adademic Press; 1990.
60. Oldenbourg R. Polarization Microscopy with the LC-PolScope. Cold Spring Harbor, New York: Cold Spring Harbor Laboratory Press; 2005.
61. Junqueira LC, Bignolas G, Brentani RR. Picrosirius staining plus polarization microscopy, a specific method for collagen detection in tissue sections. Histochem J 1979;11:447-55.
62. Whittaker P, Kloner RA, Boughner DR, Pickering JG. Quantitative assessment of myocardial collagen with picrosirius red staining and circularly polarized light. Basic Res Cardiol 1994;89:397-410.
63. Rueden CT, Conklin MW, Provenzano PP, Keely PJ, Eliceiri KW. Nonlinear optical microscopy and computational analysis of intrinsic signatures in breast cancer. Conf Proc IEEE Eng Med Biol Soc 2009;2009:4077-80.
64. Conklin MW, Provenzano PP, Eliceiri KW, Sullivan R, Keely PJ. Fluorescence lifetime imaging of endogenous fluorophores in histopathology sections reveals differences between normal and tumor epithelium in carcinoma *in situ* of the breast. Cell Biochem Biophys 2009;53:145-57.
65. Wang W, Wyckoff JB, Goswami S, Wang Y, Sidani M, Segall JE, *et al*. Coordinated regulation of pathways for enhanced cell motility and chemotaxis is conserved in rat and mouse mammary tumors. Cancer Res 2007;67:3505-11.
66. Maller O, Hansen KC, Lyons TR, Acerbi I, Weaver VM, Prekeris R, *et al*. Collagen architecture in pregnancy-induced protection from breast cancer. J Cell Sci 2013;126:4108-10.