A method for normalizing pathology images to improve feature extraction for quantitative pathology

Allison Tam[a)]
Stanford Institutes of Medical Research Program, Stanford University School of Medicine, Stanford, California 94305

Jocelyn Barker[a)]
Department of Radiology, Stanford University School of Medicine, Stanford, California 94305

Daniel Rubin
Department of Radiology, Stanford University School of Medicine, Stanford, California 94305 and Department of Medicine (Biomedical Informatics Research), Stanford University School of Medicine, Stanford, California 94305

(Received 28 August 2015; revised 14 December 2015; accepted for publication 16 December 2015; published 5 January 2016)

Purpose: With the advent of digital slide scanning technologies and the potential proliferation of large repositories of digital pathology images, many research studies can leverage these data for biomedical discovery and to develop clinical applications. However, quantitative analysis of digital pathology images is impeded by batch effects generated by varied staining protocols and staining conditions of pathological slides.
Methods: To overcome this problem, this paper proposes a novel, fully automated stain normalization method to reduce batch effects and thus aid research in digital pathology applications. Their method, intensity centering and histogram equalization (ICHE), normalizes a diverse set of pathology images by first scaling the centroids of the intensity histograms to a common point and then applying a modified version of contrast-limited adaptive histogram equalization. Normalization was performed on two datasets of digitized hematoxylin and eosin (H&E) slides of different tissue slices from the same lung tumor, and one immunohistochemistry dataset of digitized slides created by restaining one of the H&E datasets.
Results: The ICHE method was evaluated based on image intensity values, quantitative features, and the effect on downstream applications, such as a computer aided diagnosis. For comparison, three methods from the literature were reimplemented and evaluated using the same criteria. The authors found that ICHE not only improved performance compared with un-normalized images, but in most cases showed improvement compared with previous methods for correcting batch effects in the literature.
Conclusions: ICHE may be a useful preprocessing step a digital pathology image processing pipeline. © *2016 American Association of Physicists in Medicine.* [http://dx.doi.org/10.1118/1.4939130]

Key words: image processing, digital pathology, normalization

1. INTRODUCTION

Normalization of medical images is a key step in quantitative analysis. In quantitative radiology, many protocols have been developed to normalize both intersubject and interscan variability.[1–5] While conceptually the need for image normalization is similar in both quantitative radiology and digital pathology, there are many practical differences which require the development of new methodologies.

Pathology images are created from slides on which a section of tissue has been stained by one or more stains and digitized at high magnification. The most common stains used in analyzing pathology images are hematoxylin, a bluish stain that binds nucleotides, and eosin, a pink stain that binds proteins (H&E). These can be used to identify nuclei and cytoplasm, respectively. Another commonly used stain is 3,3′-diaminobenzidine (DAB), which is most often used in immunohistochemistry (IHC) applications, to identify the localization of proteins of interest.

Normalizing a pathology image without distorting the signal from the stains can be difficult. When the slides are scanned, the stain values are represented as separate red, green, and blue (RGB) values. Since information from the stains is stored as a combination of the three RGB channels, normalization that does not take this into account may distort the signal from the stains.

There are two key factors that make pathology image normalization important. First, different stains and staining protocols introduce different colors and different stain ratios between images, or even within a single image.[6] The overlap of stains frequently affects multiple colors in the RGB color space, such that the signal from one stain may affect the signal from another. Second, batch effects are common in pathology, in which the same protocol produces different colors and stain ratios, making it difficult to compare images that have the same stain.[7]

In this paper, we propose intensity centering and histogram equalization (ICHE), a novel, fully automated method for stain normalization. The method normalizes both H&E as well as IHC images and for the first time shows that the normalization can improve comparisons between the two staining protocols. In Sec. 2, we discuss related work and discuss the advantages and disadvantages of previous methods. In Sec. 3, we describe our proposal and the methods used for further comparison. In Sec. 4, we report results of our evaluation. Finally, in Sec. 5, we discuss what we feel can be learned about normalization methodology from our evaluation of the ICHE method.

2. RELATED WORK

A variety of methods have been used to normalize pathology images. One key consideration in normalization is the color space to normalize. The original RGB color space can be used, however care must be taken to normalize all three channels at the same time to avoid distorting the image staining properties.[8] Other normalization methods[9] have worked in the L^*A^*B color space as described first by Reinhard.[10] Objections to this color space have been raised because the luminance value "L" assumes that color in the image is additive, and there is evidence that in pathology images, it is subtractive.[11] The most conceptually satisfying and most commonly used approach[12–14] is to create a new color space, the H&E color space. Creating this color space involves deconvoluting the RGB color space into a separate channel for each of the stains present in the pathology image.[11] The resulting color space consists of an independent channel for each stain, allowing for normalization of the source of the color differences. Additionally, since these stains are being used as proxies to measure biological entities, normalizations based on this color space are less likely to distort the original biological signal.

Another aspect to consider in choice of normalization technique is the need for a reference image. Some normalization methods use reference images as a template, which other images are matched to in the normalization process.[8,10,12,14] This can be advantageous as it ensures that all images have similar values, without forcing images to fit an arbitrary distribution. However, this adds subjectivity to the method based on the choice of reference image, so if the same method is used with two different reference images, the resulting normalized images can no longer be compared. Additionally, if the reference image has different characteristics than the image to be normalized, such as in the case in a highly positive IHC image vs a negative IHC image, these methods can introduce artifacts.

A final aspect to consider is the need to input information about the image into the method before use. Some methods require either the identification of regions with only one stain[8,14] or the segmentation of nuclei within the slide before application of the normalization method.[10,14] If this is done manually, it requires a time commitment from an individual with domain knowledge and introduces subjectivity that may make the normalization unstable.[14] If nuclei segmentation is automatic, this can cause problems, as one of the reasons for image normalization is to produce stable automatic segmentations. Since the segmentation would have to be performed on the un-normalized image, the segmentation, and therefore the normalization, may be unstable.

Evaluating image normalization is usually done by visually examining the images for similarity. One paper, Khan *et al.*,[12] used downstream applications as their metric for image normalization effectiveness. They compared automatic tumor tissue segmentation results after normalization with a variety of methods. Since the majority of image normalization methods are used as a preprocessing step for further image analysis applications, using downstream applications as a metric for normalization success can greatly improve image normalization evaluation.

3. MATERIALS AND METHODS

3.A. Dataset and preprocessing

3.A.1. Datasets

Three datasets were created from a single collection of 260 adenocarcinomas, 32 nonsmall cell lung cancers, 68 squamous cell carcinomas, and 74 other lung cancer tumors taken from 434 patients in the Stanford tissue Microarray Database TA-369.[15] Microarrays were digitized at 20× using a Bacus Labs BLISS microscope. Two datasets (heA and heB) contained two different cross sections of the same H&E-stained tumors (Fig. 1). The third dataset (ihcB) was prepared by stripping the H&E stain from heB slides and restaining them with hematoxylin and DAB on a keratin antibody (Fig. 1). Staining artifacts were present due to the slides being stained using different staining protocols (H&E versus IHC), and batch effects were present due to the slides having been stained on different dates (heA: 2012-02-09, heB: 2014-04-09 ihcB: 2014-04-10).

3.A.2. Segmenting regions of interest

Segmentations of the nuclear regions and tissue were performed on the original images. Tissue was segmented from background via the following steps: (1) convert the lowest-resolution scan of the image to grayscale, (2) apply automatic contrast enhancement using a histogram-stretching method,[16] (3) take the 8-bit depth complement, and (4) apply hysteresis thresholding with an experimentally chosen high threshold of 100 and a low threshold of 50. Nuclei are segmented using a method derived from the work of Gurcan.[17] In our modification of their work, the hematoxylin stain was transformed using morphological top-hat reconstruction.[18] An iterative series of hysteresis thresholds was used, where the upper threshold ranged from 150 to 50 and the lower threshold was 0.2× the upper threshold with each iteration reducing the upper threshold by 5. At each iteration, objects in the image meeting size specifications (30–200 pixels) were identified as nuclei. This allowed the identification of a large variety of nuclei with different staining properties. Superpixels (regions of the image

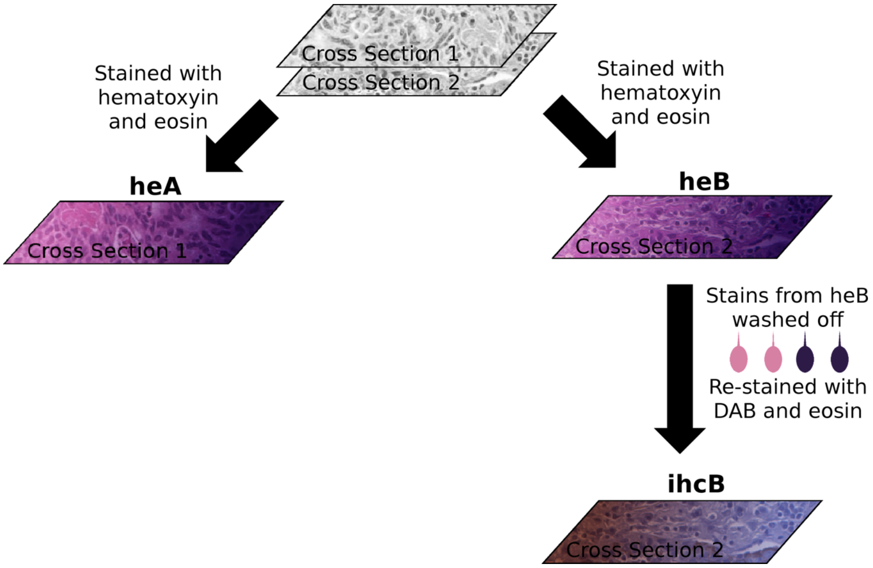
Cross Section 1
Cross Section 2
Stained with hematoxyin and eosin
Stained with hematoxyin and eosin
heA
Cross Section 1
heB
Cross Section 2
Stains from heB washed off
Re-stained with DAB and eosin
ihcB
Cross Section 2

FIG. 1. Protocol for production of the three datasets; heA, heB, and ihcB. Two cross sections are cut from a tumor and stained with hematoxylin and eosin to create the heA and heB datasets. The stain is then removed from heB and restained to create ihcB.

having similar pixel values) were created using the vl_feat SLIC extraction method[19] with region size of 28 and regularizer value of 10 000. These were used to achieve superpixels of roughly 450–650 pixels.

3.A.3. Unmixing the stains

In order to better represent the biological information of a sample, the hematoxylin and eosin stains were separated from the original image using the color deconvolution method used in CellProfiler.[20] This method for color deconvolution uses pseudoinverse matrices as described by Ruifrok and Johnston.[11] The RGB values used for hematoxylin were [0.644, 0.717, 0.267], for eosin were [0.093, 0.954, 0.283], and for DAB were [0.268, 0.570, 0.776]. To obtain the hematoxylin and eosin stain values, we find

$$e^{\ln(\mathrm{RGB}_{i,j})\begin{bmatrix}0.644,0.171,0.267\\0.093,0.954,0.283\end{bmatrix}^{+T}} = \mathrm{HE}_{i,j}$$

or

$$e^{\ln(\mathrm{RGB}_{i,j})\begin{bmatrix}0.644,0.171,0.267\\0.268,0.570,0.776\end{bmatrix}^{+T}} = \mathrm{IHC}_{i,j}, \tag{1}$$

where $\mathrm{RGB}_{i,j}$ are the red, green, and blue channel values for pixel i, j and $\mathrm{HE}_{i,j}$ are the deconvoluted hematoxylin and eosin values for pixel i, j. The symbol $+T$ indicates the transpose of the pseudoinverse of the matrix. Since hematoxylin binds to nucleotides and eosin to proteins, unmixing the stains allows the method to better probe the roles of these biologically important molecules.

3.B. Proposed method: ICHE

The ICHE method uses a two-phase approach to normalize pathology images. The first phase, centroid alignment, seeks to create a common intensity range in a localized manner. The second phase, tissue restricted contrast-limited adaptive histogram equalization (CLAHE), smooths the edges of the localized regions from the centroid alignment method.

3.B.1. First stage: Centroid alignment

Centroid alignment brings the image into a normalized intensity range without affecting the modality of the distribution. Preserving modality is desirable, since peaks in the distribution are likely to represent biological entities. In centroid alignment, linear scaling is used separately on each half of the image intensity histogram. The value of the centroid of the image is the weighted average of the intensity distribution of the image defined as

$$c = \frac{\sum_{n=0}^{255} n^{*}\mathrm{hist}(n)}{\sum_{n=1}^{254} \mathrm{hist}(n)}, \tag{2}$$

where c is the centroid of the distribution, n is an image intensity value, and $\mathrm{hist}(n)$ is the frequency of the value n in the image region being normalized. The intensity values above the intensity centroid are scaled to fit between half the maximum intensity value (128 for an uint8 encoded image) and the maximum intensity value (255 for an uint8 encoded image). Similarly, values below the intensity centroid are scaled between 0 and half the maximum intensity value. Succinctly, for the pixel intensities i within the tissue region of the image with intensity centroid c, we create the new set of pixel intensities I' such that

$$I' = \left\{\frac{i-c}{255-c}\times 128 + 128 : i \geq c\right\} \cup \left\{\frac{i}{c}\times 128 : i < c\right\}. \tag{3}$$

In order to scale the whole image, the intensity values for each superpixel is independently scaled iteratively until the centroid of each histogram reaches half the maximum intensity value (128 ± 0.0001) (Fig. 2). This target point was chosen because it places equal weight on light and dark intensity values.

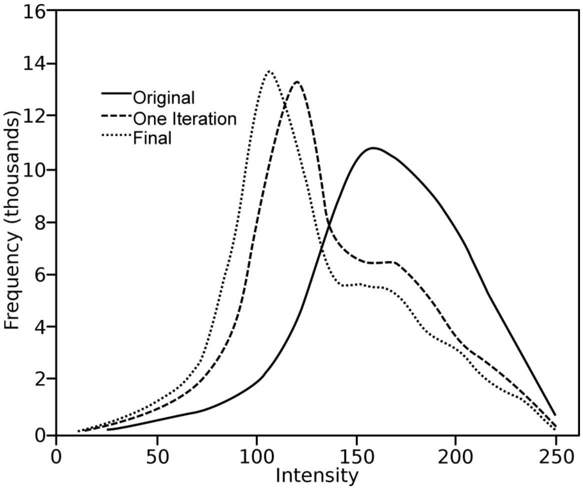

FIG. 2. Intensity histograms at different points of centroid normalization. The solid histogram is derived from the original with a centroid at 163. The dashed histogram with a centroid at 137 is derived from the image after one iteration of the centroid scaling. The section of the solid histogram to the right of its centroid is stretched out to form the right side of the dashed histogram, while the left section of the solid histogram is compressed to form the left side of the dashed histogram. After eleven iterations, the dotted histogram is formed with a centroid at 128.

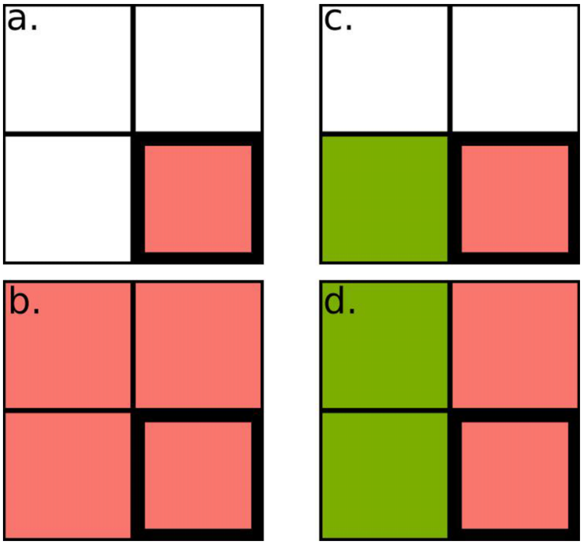

FIG. 3. CLAHE's scheme for replacing the CDFs of tiles with missing information. The tile in the lower right with a bold outline represents the target tile being normalized, white tiles indicate tiles with missing data, and unique colors represent a unique CDF. [(a) and (c)] State of tiles before replacement. [(b) and (d)] The CDFs used to replace missing tiles in (a) and (c), respectively.

3.B.2. Second stage: Tissue restricted CLAHE

This second stage smooths the superpixel boundary artifacts created in the first portion of the technique, since the superpixels are normalized independently in that stage. An all-generative/tissue-restrictive modification of CLAHE was used. CLAHE functions by dividing up an image into regularly sized tiles and maps each tile's intensity histogram to a target distribution. We experimentally found that 32×32 pixel tiles and a Rayleigh distribution mapping function were the best, based on the Pearson's correlations between image pair features. Interpolation is applied between tiles to smooth out inconsistencies. In CLAHE, the interpolated pixel values for the lower right tile in a 2×2 grid are generated by taking the weighted average of the tiles in the 2×2 grid. The weighted averages depend on the distance between pixels (farther pixels influence the interpolated values less) and the cumulative relative frequency intensity histogram of the tiles in the 2×2 grid (hereafter referred to at the CDF).

In pathology images, the border between tissue and the background slide may generate artifacts in the normalization. Many tissue structures, such as glands, have a large amount of diagnostically important tissue on these boundaries. In order to avoid these artifacts impacting analysis, it is important to avoid including background slide in the interpolation. In ICHE's modification of CLAHE, interpolation is only applied between tiles that are at least 75% covered by tissue to avoid artifacts from the nontissue regions. In order to do this, the CDF for the nontissue tiles is replaced with the CDF for one of the tissue covered regions. Two special cases where this must happen have previously been addressed by CLAHE's handling of image edges, and here the same solution is used for tissue edges (Fig. 3).

There remain three special cases: where there is no established method for replacing the empty tiles; where the nontissue tiles are one neighboring tile, both neighboring tiles, and one diagonal tile; and where either one or both neighboring tiles are in need of replacement, the tiles are replaced with the target tissue tile. Large difference in tissue composition can occur over short distances. To reduce the impact of these artifacts, when a nontissue tile is a neighboring tile, the tissue restricted version of CLAHE replaces the tile with the CDF of the target tile [Figs. 4(a)–4(d)]. However, when the nontissue tile is diagonal, the target tile replacing the data with the target may be a poor representative. In this case, the missing tile is replaced with the average of the CDFs for the three present tiles [Figs. 4(e) and 4(f)].

Three alternative tile replacement schemes and traditional CLAHE with no consideration of tissue borders were also tested (Figure S1[28]) to determine the impact of tile selection in this method. All alternative replacement schemes which took tissue boarders into account show high similarity to the replacement scheme described here, showing the method's robustness. Images normalized using traditional CLAHE without concern for tissue borders showed high variation compared to the method described here, demonstrating the importance of considering tissue borders in CLAHE normalization.

3.C. Established normalization methods

Two methods from the digital pathology literature, as well as one standard image normalization technique, were evaluated in order to compare our ICHE method to standard methods in the field. We reimplemented Macenko's technique based on descriptions in the literature, and we

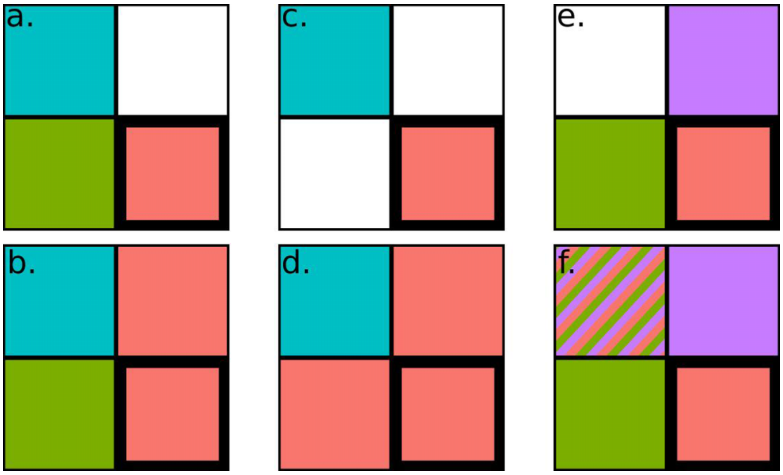

FIG. 4. Modified, tissue restricted CLAHE's scheme for replacing the CDFs of tiles with missing information. The tile in the lower right with a bold outline represents the target tile being normalized, white tiles indicate tiles with missing data, and unique colors represent a unique CDF. The multicolor tile in (f) represents the averaging of the red, green, and purple tiles. [(a), (c), and (e)] State of tiles before replacement. [(b), (d), and (f)] The CDFs used to replace missing tiles in (a), (c), and (e), respectively.

obtained code for Reinhard's method, available from the MathWorks library.[21] CLAHE normalization was produced using the adapthisteq method in MATLAB, which is an implementation of the original paper describing the method.[22]

3.C.1. Autocontrast

This technique[13] is an intensity histogram-stretching method in which the extremes in the intensity histogram in the image (1st and 99th percentile) become the new pseudomaxia for the intensity histogram in the normalized image. This method is applied to each individual deconvolved stain component.

3.C.2. Lab

This normalization technique[10] operates in the L^*A^*B color space. A single target image is selected to act as the standard for all normalized images. Although other papers that have used a Reinhard variation recommend that a pathologist select the target image, random selection is suitable, since the original Reinhard paper states that images can even be normalized to a dissimilar target image. The method matches the channel histograms of the normalized image to a single target image through a linear transform. It is worth noting that this is the only technique that does not work directly in the unmixed stain color space, since it normalizes within L^*A^*B.

3.C.3. CLAHE

In order to compare the method to the most similar method from the literature, the CLAHE normalization method was applied to the images. We found no record of this method being previously applied to pathology; however, it is common in other image normalization applications The same 32×32 pixel tiles and a Rayleigh distribution mapping function were used as in the ICHE implementation, but the tissue restriction was not applied. For more information on CLAHE, see its description in Sec. 3.B.2.

3.D. Evaluation

In order to study the effects of the various normalization techniques, all methods were evaluated by multiple criteria. These criteria included the similarity of the color values after normalization, the consistency of quantitative features extracted from the images, and the effect of the methods on downstream application. All criteria were evaluated in the context of comparing images from the same tumors in the heA, heB, and ihcB datasets.

3.D.1. Feature extraction

In order to evaluate the impact of various normalization methods on an application, commonly used pathology image features were extracted for all images: shape, color,[23] and Haralick texture features.[24] These features were collected from the tissue segmentation, from the nuclear segmentation, and from the non-nuclear regions. Additionally, the features were collected independently from images of both the H&E and DAB stains. The features were collected for the aggregate tissue as a whole, as well as individually for each superpixel. Haralick features were derived by using a one-pixel displacement, with symmetric gray level co-occurrence matrix at 0° and 90°, between each pair of pixels. When evaluations were made between the two H&E datasets (heA and heB), all image features are used. When evaluations are made between heB and ihcB, the eosin and DAB derived features are not used, since they have different staining properties and cannot be compared.

3.D.2. Evaluation of intensity distribution

The color values after normalization are first evaluated by visually looking at the similarities between the normalized images. In order to more quantitatively compare the intensities, intensity histograms were created for the heA–heB images before and after normalization (Fig. 5). Correlations were

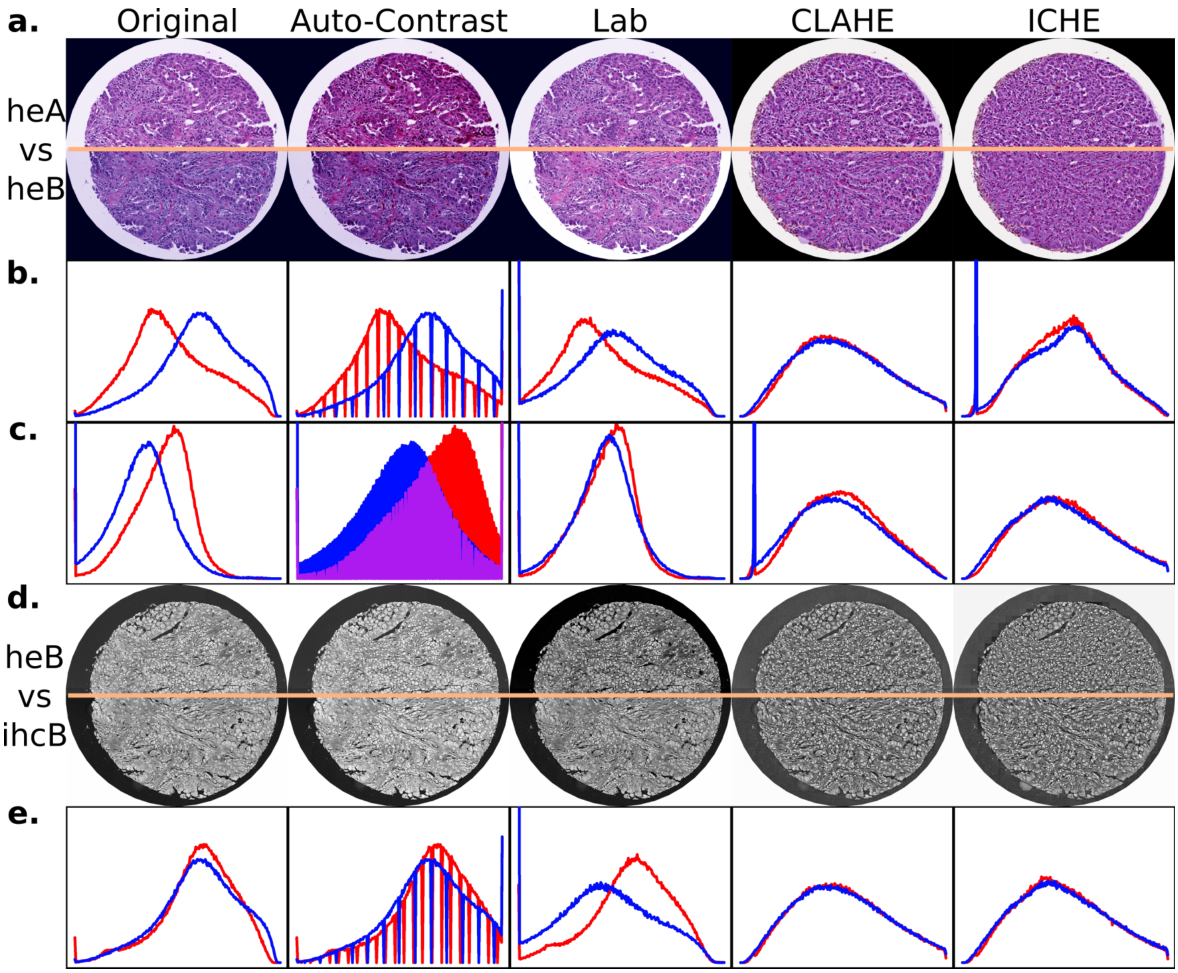

FIG. 5. Staining intensities of the normalized images from the same tumor in the [(a)–(c)] heA–heB and [(d)–(e)] heB–ihcB comparisons. (a) Shows sample images with heA above the horizontal line and heB below it. (d) Shows the hematoxylin stain only with heB above the horizontal line and ihcB below it. [(b), (c), and (e)] Intensity histograms for the stains of the normalized images (x-axis is the intensity value and the y-axis is the frequency). [(b) and (c)] Red lines represent the heA image and blue the heB image. (e) Red lines represent the ihcB image and blue lines the heB image. In all cases, the first column shows the un-normalized images, and the rest show results from applying the various normalization techniques. ICHE and CLAHE are the only methods able to remove visible color intensity differences between the datasets.

found between the intensity histograms of each image pair (Table S1, Figure S1[28]).

3.D.3. *Correlation of extracted features*

The consistency of the features was evaluated using Pearson's and canonical correlations. Standard pathological features (see Sec. 3.D.1) were collected for all images before normalization and after normalization with each method. Pearson's correlation coefficients between features of datasets are evaluated to gain a more quantitative relationship of the similarity between the image pairs. Pearson's correlations were statistically evaluated through a one-sided two-sample t-test for the null hypothesis that the Pearson's correlations in the normalized images are greater than those in the original images. In order to examine how the data sets as a whole were affected by normalization, canonical correlations between the features in the two datasets were also calculated.

3.D.4. *Modeling disease subtypes*

A computer aided diagnosis classifier was built to distinguish two lung cancer subtypes: adenocarcinoma and squamous cell carcinoma. Features were normalized by subtracting the mean and dividing by the standard deviation, and then input as predictors into a LASSO regression model.[25] The glmnet[26] implementation was used on a binomial model with a binary output (adenocarcinoma vs squamous cell carcinoma). The free parameter λ, which is the penalty on the number of features, was selected using 10-fold cross validation. Two classification models were created. One, the heA–heB comparison, evaluates the effect of normalization between two same-stained sets, H&E. The other, heB–ihcB, looks at normalization between differently stained sets, H&E and IHC. The heA–heB comparison was trained on heA and tested on heB. The heB–ihcB comparison was trained on ihcB and tested on heB. The training sets consisted of a randomly selected set of 220 adenocarcinomas and 34 squamous cell tumors from the training dataset. The testing set consisted of the remaining 34 adenocarcinomas and 38 squamous cell tumors from the other dataset. To evaluate the accuracy of the classifier, a one-sided one-proportion z-test is conducted with a no-information rate (NIR) of 0.53. The no-information rate is the closest possible proportion to an equally weighted testing set.

4. RESULTS

4.A. Similarity for color values

One of the simplest metrics for evaluating image similarity is examining how similar intensity values in one image are to another. This criterion has also been used in numerous other pathology normalization papers, such as the works of Macenko *et al.*[13] and Kothari *et al.*,[8] In the original images, batch effects are visibly present [Figs. 5(a) and 5(d)] with greater intensity of the hematoxylin stain in the heB dataset than the heA dataset [Fig. 5(b)], and greater intensity of the eosin stain in the heA dataset than the heB dataset [Fig. 5(c)]. The over all intensity of the hematoxylin stain was similar in the heB and ihcB datasets, but the shape differed due to the secondary stain properties.

The two methods from the literature, autocontrast and Lab, did not improve the intensity correlation as much as would be desired. Differences in the images and intensity histograms can easily be seen in the sample image for the autocontrast method (Fig. 5), and statistical analysis shows a decrease in intensity histogram correlation for all three stain comparisons in the datasets as a whole (Figure S1, Table S1[28]). Lab showed a statistically significant increase in intensity histogram correlation for the heA vs heB correlation of the hematoxylin stain, but is showed a drastic decrease in correlation for the eosin stain in the same comparison.

CLAHE and ICHE on the other hand showed improvement in image similarity in all comparisons (Fig. 5). While the most drastic improvement was seen in the heA and heB hematoxylin stain, where the initial correlation was lowest, improvement was seen in all three comparisons (Figure S1, Table S1[28]). This increase in correlation is likely due to the histogram equalization element of both methods, where the image's original intensity histogram is mapped to a target Rayleigh distribution.

One difference to note between CLAHE and ICHE is the shape of the final histograms. While CLAHE largely produces the same final shape no matter the input, ICHE's final shape varies more depending on the initial input histogram. This gives ICHE the potential for feature extraction that is more dependent on the original input image.

4.B. Correlation of pathological features

While visual inspection and intensity distributions can be informative, most image normalization in pathology is done for the purpose of feature extraction and quantitative modeling. In order to determine the impact of normalization on the features extracted from the images, correlations were examined between the same tumors in each dataset.

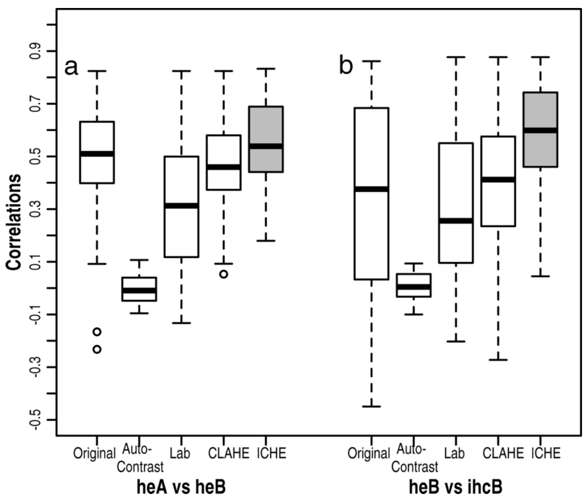

Fig. 6. Boxplots of the Pearson's correlations of the pathological features after normalization with different techniques in heA–heB (a) and heB–ihcB (b) comparison. The gray boxplots indicate a statistically significant improvement over the original image ($p < 0.05$).

4.B.1. *Pearson's correlations*

The Pearson's correlations for standard pathology features show the heA–heB comparisons generally have lower correlations than the heB–ihcB comparisons (Fig. 6, Table I). This is likely due to the features being derived from the same tissue and therefore having less potential variability between datasets.

The median value of the correlation before normalization was 0.509 for the heA–heB comparison and 0.573 for the heB–ihcB comparison. The median correlation value after normalization decreased for the three methods we studied from the literature: autocontrast, Lab, and CLAHE (Table I, Fig. 6). The autocontrast method decreases the correlation such that the image correlation is near zero for all features. This indicates the importance of quantitative comparison when developing image normalization techniques, as there is the potential to negatively impact feature collection, even if the images appear more similar.

ICHE, on the other hand, shows a statistically significant increase in median correlation for both the heA–heB (0.538, $p = 0.027$) and the heB–ihcB (0.607, $p = 0.017$) comparison.

Table I. Statistics for Pearson's Correlations for features extracted from normalized images. Medians of the distribution between the image pairs for each feature. A one-sided t-test is performed on a sample size of 59 (heA–heB) or 39 (heB–ihcB). The p-value of this test is displayed for the null hypothesis that the mean correlation found for a given technique is greater than that for the un-normalized images.

	heA–heB		heB–ihcB	
Technique	Median	p-value	Median	p-value
Original	0.509	n/a	0.573	n/a
Autocontrast	−0.010	1.000	0.007	1.000
Lab	0.313	0.999	0.497	0.565
CLAHE	0.459	0.427	0.470	0.461
ICHE	0.538	0.027	0.607	0.017

This indicates that the features collected from this normalization method more closely match between the datasets, increasing the potential for accurate quantitative comparisons, compensating for batch effects.

4.B.2. Canonical correlations

Canonical correlations measure how well one dataset's features linearly correlate or predict the other dataset's features. Because corresponding images from different datasets are extracted from the same tumor, canonical correlations should be high. As when evaluating Pearson correlations, heA vs heB comparisons used all features, but heB vs ihcB excluded eosin and DAB-derived features. Correlations were calculated for the whole group of features as well as independently for the texture features, and for the nontexture features.

As opposed to the Pearson's correlations, the canonical correlation behaved similarly in the heA vs heB datasets as in the heB vs ihcB datasets (Fig. 7). When an increase in canonical correlations was seen, it seemed largely due to an increase in texture correlations, as the nontexture correlations produced a less dramatic difference.

Once again, the autocontrast method resulted in a poorer canonical correlation of the features overall, as well as in each of the subgroups. Lab produced higher canonical correlation using this metric, as opposed to the decrease seen in Pearson's correlation metric. This indicates that Lab improves feature correlation more for the features as a group than for the individual features.

CLAHE increased the canonical correlation of the features in both the heA–heB and heB–ihcB comparisons, with a correlation near unity in the heB–ihcB comparison. The ICHE method also showed increased in canonical correlation, with a correlation near unity in both heA–heB and heB–ihcB comparisons. Taken as a whole, the correlation data show that the CLAHE and ICHE methods produce feature extraction that appears to be affected less by staining artifacts than the other normalization methods we tested.

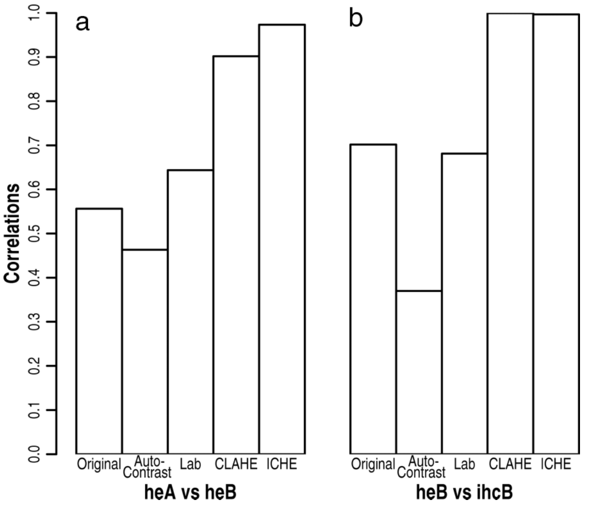

Fig. 7. heA–heB (a) and heB–ihcB comparison (b). First canonical correlation of the normalization techniques after 1000 permutations, showing the impact of normalization on features extracted from the images.

Table II. Performance of the computer aided diagnosis classifier for the heA–heB and heB–ihcB comparisons for various methods of normalization. Accuracy is the proportion of correctly diagnosed images. The p-value is for the null hypothesis that accuracy in using images normalized by a given normalization technique is not greater than when using original images.

	heA–heB		heB–ihcB	
Technique	Accuracy	p-value	Accuracy	p-value
Original	0.568	n/a	0.514	n/a
Autocontrast	0.581	0.434	0.554	0.311
Lab	0.635	0.201	0.635	0.067
CLAHE	0.635	0.201	0.635	0.067
ICHE	0.716	0.030	0.662	0.033

4.C. Impact on applications

The effect of normalization was evaluated in the application of computer aided diagnosis to determine how normalization might affect downstream applications. A LASSO binomial regression model was used to create a binary classifier to distinguish between two lung cancer subtypes, adenocarcinoma and squamous cell carcinoma, based on quantitative image features.

The results of using original images, and those normalized using autocontrast, Lab, and CLAHE methods produced classification results that were not significantly better than random chance in both the heA vs heB and heB vs ihcB comparisons (Table II). Only ICHE performed significantly better than chance (heA–heB, $p = 0.002$, heB–ihcB, $p = 0.047$). ICHE also showed significant improvement over original images in both the heA–heB and the heB–ihcB comparison (Fig. 8).

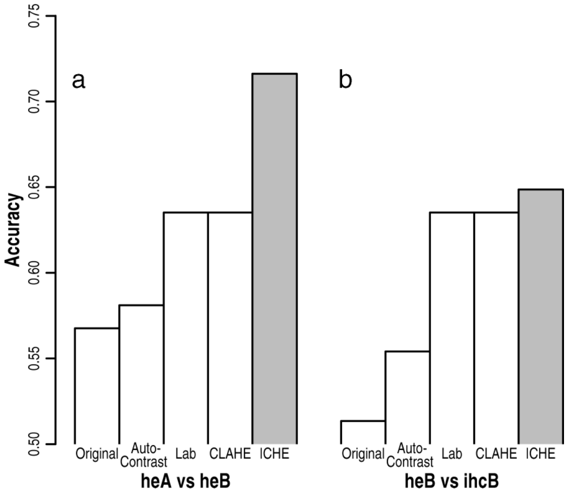

Fig. 8. Accuracy of the computer aided diagnosis classifier for the heA–heB (a) and heB–ihcB (b) comparisons for various methods of image normalization. The white bars represent the techniques whose accuracies are not significantly greater than the NIR at the 0.05 alpha-level. The gray represents the techniques whose accuracies are significantly greater than the NIR. The ICHE model accuracy is also a significant improvement over the accuracy of the un-normalized images model at the 0.05 alpha-level.

5. DISCUSSION AND CONCLUSIONS

We have developed a novel method for pathology image normalization using a technique that combines shifting the intensity distribution and a generative/tissue-restrictive modification of CLAHE. Performing a centroid shift in the intensity distributions makes all pixels in the image have a common center and thus will be in about the same intensity range. The variation of the CLAHE method smooths out uneven staining among superpixels which were normalized independently of one another in the first stage of ICHE. Together, these steps create a normalized image that is not only visually appealing but makes a reliable first step for pathology feature collection.

It is important when evaluating normalization methods to look at the impact on feature extraction and not only color similarity. While ICHE did show improvement over most other methods in color similarity, examining extracted features showed a statistically significant increase in feature correlation not seen in the other methods. This increase similarity of the ICHE normalized images allowed ICHE to be the only model with a statistically significant increase in accuracy of the computer aided diagnosis models. Since most image normalization applications are used as a preprocessing step before feature extraction for further analysis, ICHE provides a more stable platform on which to build an image analysis pipeline.

Only one other group has attempted to evaluate normalization in such a way as to take downstream applications into account.[12] However, this paper only evaluates the ability of the method to normalize scanner artifacts, not the more common staining artifacts. Additionally, they never directly evaluate the features extracted, only the impact of normalization on tumor segmentation. By evaluating commonly used features in histopathology image analysis as well as downstream applications, it may be possible to get a more general view of how well the normalization may generalize to other applications.

To our knowledge, ICHE is the first method to attempt to evaluate image normalization between two different staining protocols. Images produced from these protocols show great differences in color and staining patterns, making normalization between them difficult. However, H&E and IHC can provide different information about a tissue sample, so being able to easily compare these two data types for the same tissue expands the potential for synergistic analysis. ICHE's ability to improve comparisons between these two staining methods improves the potential for discovery of new relationships between H&E imaging and protein expression patterns.

Before normalization takes place, the image is segmented into superpixels with similar properties that are used in the first phase of normalization. These superpixels are regions that are similar in terms of rudimentary texture and color properties. By normalizing first within superpixels, we avoid having abnormally dark areas or overly bright areas that can bias the normalization of the whole slide. This is particularly important because some of the causes of these extreme staining regions may not be present in every image. For instance, large groups of invading lymphocytes cause a bright region in the hematoxylin stain. If the whole slide were to be considered at once, these bright regions would force the rest of the slide to have a lower value than if the same tissue were normalized without that region present.

Rather than normalizing a traditional color space such as RGB or Lab, ICHE deconvolves the image into the H&E color space. Since hematoxylin binds to nucleic acids and eosin binds to protein, unmixing the stains allows ICHE to directly normalize the signal these important biological molecules. Methods which work in other color spaces may distort this signal and cause bias in downstream applications.

In the first stage of normalization, ICHE uses centroid alignment to create a common mean intensity for all areas of the images. This favorably removes the batch effects that create darker or lighter images, while preserving the modality of the distribution. Peaks in the staining distribution may represent distinct biological entities, so if the peaks in an image's intensity histogram is not preserved, the signal from these entities may be difficult to distinguish from one another.

In the second stage of normalization, ICHE uses a variation of CLAHE to smooth any abnormalities at the border of superpixels created in the centroid alignment. CLAHE allows for localized normalization without creating an overly peaked distribution in homogeneous regions, as may occur in other adaptive histogram equalization methods. Additionally, our modification of CLAHE takes into account known information about tissue edges, preventing the background of the slide from impacting tissue normalization.

The combination of both of these processing steps creates normalized images that reduce the batch effects not only in appearance but also quantitatively in terms of the correlation metrics. We also showed potential benefits of our normalization method in applications as well.

In addition to the efficacy of normalization, ICHE has a number of advantages which lower the level of subjectivity and resources required to normalize sets of pathology image. ICHE does not require the selection of a target image, as is common in prior methods.[8,10,12,14] It is fully automatic and does not require a manual selection of regions with only one stain, which other methods do require.[8,14] This means that images normalized by ICHE are deterministic and can easily be compared without need for renormalization. Additionally, it does not require the segmentation of nuclei as an input,[10,14] which has been called the "Achilles heel of pathology."[27]

Our work has some limitations. First, while we tested our method in a dataset with multiple forms of batch effects, all images were from lung cancers, so generalizability has not yet been tested. However, we believe our results will hold in other cancers. Second, in evaluating normalization by comparing H&E with IHC, we only used one antibody with a cytoplasmic antigen. It is possible that other antibodies for proteins with different localizations (e.g., nuclear or membrane localized proteins) might be impacted by normalization in different ways. This would be need to be studied in future work. Third, some datasets may not need normalization at all. When all slides are stained in a single batch, such as is the case in a single tissue microarray or when analyzing the data from a single whole slide image, batch effects may not be present, and so analysis may be possible without normalization. It is even possible that in these images with minimal staining artifacts normalization may result in a loss of signal. However, when comparing between datasets where batch effects are present, our data indicate that the increase in feature similarity outweighs any potential loss in signal from normalization.

As digital pathology grows as a field and more data from varied sources becomes available, comparing datasets will become more important for pathology image analysis. We have shown that the ICHE method of image normalization improves our ability to find images with similar properties, through visual analysis, quantitative comparison, and through implementation of a downstream image classification application. We thus believe that ICHE could be useful as a first step in a pathology image processing pipeline.

ACKNOWLEDGMENTS

The authors would like to thank Rob West for providing the TMA samples. This work was supported in part by the Li Ka Shing Foundation Big Data for Human Health Seed Grant and the Stanford Institutes of Medicine Summer Research Program.

[a] A. Tam and J. Barker contributed equally to this work.

[1] B. M. Ellingson, T. Zaw, T. Cloughesy, K. M. Naeini, S. Lalezari, S. Mong, A. Lai, P. Nghiemphu, and W. B. Pope, "Comparison between intensity normalization techniques for dynamic susceptibility contrast (DSC)-MRI estimates of cerebral blood volume (CBV) in human gliomas," J. Magn. Reson. Imaging **35**(6), 1472–1477 (2012).

[2] K. J. Friston, C. D. Frith, P. F. Liddle, R. J. Dolan, A. A. Lammertsma, and R. S. Frackowiak, "The relationship between global and local changes in PET scans," J. Cereb. Blood Flow Metab. **10**(4), 458–466 (1990).

[3] M. Gavrilescu, M. E. Shaw, G. W. Stuart, P. Eckersley, I. D. Svalbe, and G. F. Egan, "Simulation of the effects of global normalization procedures in functional MRI," NeuroImage **17**(2), 532–542 (2002).

[4] S. S. Kim, J. B. Seo, N. Kim, E. J. Chae, Y. K. Lee, Y. M. Oh, and S. D. Lee, "Improved correlation between CT emphysema quantification and pulmonary function test by density correction of volumetric CT data based on air and aortic density," Eur. J. Radiol. **83**(1), 57–63 (2014).

[5] N. I. Weisenfeld and S. K. Warfteld, "Normalization of joint image-intensity statistics in MRI using the Kullback–Leibler divergence," in *IEEE International Symposium on Biomedical Imaging: Nano to Macro*, *Washington, D.C.* (IEEE, 2004), Vol. 1, pp. 101–104.

[6] A. Ljungberg and O. Johansson, "Methodological aspects on immunohistochemistry in dermatology with special reference to neuronal markers," Histochem. J. **25**(10), 735–745 (1993).

[7] H. O. Lyon, A. P. Leenheer, R. W. Horobin, W. E. Lambert, E. K. W. Schulte, B. Van Liedekerke, and D. H. Wittekind, "Standardization of reagents and methods used in cytological and histological practice with emphasis on dyes, stains and chromogenic reagents," Histochem. J. **26**(7), 533–544 (1994).

[8] S. Kothari, J. H. Phan, R. A. Moffitt, T. H. Stokes, S. E. Hassberger, Q. Chaudry, A. N. Young, and M. D. Wang, "Automatic batch-invariant color segmentation of histological cancer images," in *IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, *Chicago, IL* (IEEE, 2011), pp. 657–660.

[9] Y.-Y. Wang, S.-C. Chang, L.-W. Wu, S.-T. Tsai, and Y.-N. Sun, "A color-based approach for automated segmentation in tumor tissue classification," in *29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. EMBS 2007*, *Lyon, France* (IEEE, 2007), pp. 6576–6579.

[10] E. Reinhard, M. Adhikhmin, B. Gooch, and P. Shirley, "Color transfer between images," IEEE Comput. Graphics Appl. **21**(5), 34–41 (2001).

[11] A. C. Ruifrok and D. A. Johnston, "Quantification of histochemical staining by color deconvolution," Anal. Quant. Cytol. Histol. **23**(4), 291–299 (2001).

[12] A. Khan, N. Rajpoot, D. Treanor, and D. Magee, "A nonlinear mapping approach to stain normalization in digital histopathology images using image-specific color deconvolution," IEEE Trans. Biomed. Eng. **61**(6), 1729–1738 (2014).

[13] M. Macenko, M. Niethammer, J. S. Marron, D. Borland, J. T. Woosley, X. Guan, C. Schmitt, and N. E. Thomas, "A method for normalizing histology slides for quantitative analysis," in *IEEE International Symposium on Biomedical Imaging: From Nano to Macro. ISBI '09*, *Boston, MA* (IEEE, 2009), pp. 1107–1110.

[14] D. Magee, D. Treanor, D. Crellin, M. Shires, K. Mohee, and P. Quirke, "Colour normalisation in digital histopathology images," in *Proceedings of Optical Tissue Analysis in Microscopy, Histopathology, and Endoscopy*, 2009.

[15] R. J. Marinelli, K. Montgomery, C. L. Liu, N. H. Shah, W. Prapong, M. Nitzberg, Z. Zachariah, G. J. Sherlock, Y. Natkunnam, R. B. West, M. van be Rijn, P. O. Brown, and C. A. Ball, "The Stanford tissue microarray database," Nucleic Acids Res. **36**(Suppl. 1), D871–D877 (2008).

[16] R. Divakar, Auto Conrast, MATLAB Central File Exchange, 2009.

[17] M. N. Gurcan, T. Pan, H. Shimada, and J. Saltz, "Image analysis for neuroblastoma classification: Segmentation of cell nuclei," in *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, *New York, NY* (IEEE, 2006), Vol. 1, pp. 4844–4847.

[18] F. Meyer, "Iterative image transformations for an automatic screening of cervical smears," J. Histochem. Cytochem. **27**(1), 128–135 (1979).

[19] A. Vedaldi and B. Fulkerson, *VLFeat: An Open and Portable Library of Computer Vision Algorithms*, *Proceedings of the 18th ACM International Conference on Multimedia*, *Firenze, Italy*, *2010* (2008).

[20] L. Kamentsky, T. R. Jones, A. Fraser, M. A. Bray, D. J. Logan, K. Madden, V. Ljosa, C. Rueden, K. W. Eliceiri, and A. E. Carpenter, "Improved structure, function and compatibility for CellProfiler: Modular high-throughput image analysis software," Bioinformatics **27**(8), 1179–1180 (2011).

[21] R. Manohar, *Stain Normalization—File Exchange,* MATLAB *Central* (MathWorks, Natick, MA, 2013).

[22] K. Zuiderveld, "Contrast limited adaptive histograph equalization," in *Graph. Gems IV*, edited by P. Heckbert (Academic, Inc., San Diego, CA, 1994), pp. 474–485.

[23] M. N. Gurcan, L. Boucheron, A. Can, A. Madabhushi, N. Rajpoot, and B. Yener, "Histopathological image analysis: A review," IEEE Rev. Biomed. Eng. **2**, 147–171 (2009).

[24] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," IEEE Trans. Syst., Man, Cybern. **SMC-3**(6), 610–621 (1973).

[25] R. Tibshirani, "Regression Shrinkage and selection via the lasso," J. R. Stat. Soc. Ser. B Methodol. **58**(1), 267–288 (1996).

[26] J. H. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," J. Stat. Software **33**(1), 1–22 (2010).

[27] J. Gil and H.-S. Wu, "Applications of image analysis to anatomic pathology: Realities and promises," Cancer Invest. **21**(6), 950–959 (2003).

[28] See supplementary material at http://dx.doi.org/10.1118/1.4939130 for impact of alternate ICHE tile replacement schemes (Figure S1) and Pearson's correlations of intensity histograms after normalization (Table S1, Figure S2).