



OPEN Staining normalization in histopathology: Method benchmarking using multicenter dataset

Umair Khan¹, Jouni Härkönen², Marjukka Friman², Hesam Hakimnejad¹, Leena Latonen^{1,3}, Teijo Kuopio² & Pekka Ruusuvoori^{1,4,5}✉

Hematoxylin and Eosin (H&E) has been the gold standard in tissue analysis for decades, however, tissue specimens stained in different laboratories vary, often significantly, in appearance. This variation poses a challenge for both pathologists' and AI-based downstream analysis. Minimizing stain variation computationally is an active area of research. To further investigate this problem, we collected a unique multi-center tissue image dataset, wherein tissue samples from colon, kidney, and skin tissue blocks were distributed to 66 different labs for routine H&E staining. To isolate staining variation, other factors affecting the tissue appearance were kept constant. Further, we used this tissue image dataset to compare the performance of eight different stain normalization methods, including four traditional methods, namely, histogram matching, Macenko, Vahadane, and Reinhard normalization, and two deep learning-based methods namely CycleGAN and Pixp2pix, both with two variants each. We used both quantitative and qualitative evaluation to assess the performance of these methods. The dataset's inter-laboratory staining variation could also guide strategies to improve model generalizability through varied training data.

Tissue slide digitization has helped histological analysis tremendously over the decades. It has expedited the whole process by reducing the need of manual slide handling during observation and also allowed us to archive tissue specimens in a more efficient manner. Once the tissue specimen is scanned, the digital image is stored and kept safe from any physical changes or deterioration. Tissue processing and staining are complex processes that have a significant effect on tissue appearance. Typically, tissue specimens go through fixation, embedding, and sectioning before staining. Factors such as over- or under-fixation, choice of embedding medium and temperature, and thickness of the tissue section contribute to the varying appearance of tissue images^{1,2}. Once processed, the tissue sections are then stained. For instance, one of the most commonly used stain combinations is Hematoxylin and Eosin (H&E) that reveals the intricate morphology of the tissue specimen which is otherwise indiscernible prior to staining. Hematoxylin imbues the cell nuclei with shades of a purple-blue color and Eosin stains cytoplasm and extracellular matrix with shades of a pink-red color³. Even though H&E is the most commonly used stain combination in routine diagnostics, differences in stain manufacturer, composition and staining protocol manifest in varying appearance of the stained tissue across different laboratories⁴. The appearance of the stained tissue is prone to further variation during the imaging phase, as differences in microscope/scanner hardware, image acquisition and post processing techniques from different manufacturers add to the variation^{5,6}. The combined effect of these factors pose challenges for post digitalization tasks. For example, we have previously shown that differences in appearance due to fixatives contribute towards differences in performance of image analysis, specifically, in nucleus detection⁷.

Digitization of histopathological tissue samples into high resolution whole slide images has enabled the development and use of computer assisted tools based on artificial intelligence (AI) for diagnostics and decision support^{8,9}. These diagnostic tools have the potential to improve efficiency in healthcare as well as provide access to diagnostic services for regions that have been suffering from a chronic shortage of pathologists¹⁰. Studies have shown that use of AI in digital pathology can be highly effective and accurate on tasks such as cancer grading,

¹Institute of Biomedicine, University of Turku, Turku FI-20014, Finland. ²Department of Pathology, Hospital Nova of Central Finland, Jyväskylä, Finland. ³Institute of Biomedicine, University of Eastern Finland, Kuopio, Finland. ⁴Faculty of Medicine and Health Technology, Tampere University, Tampere 33100, Finland. ⁵InFlames Research Flagship, University of Turku, Turku, Finland. ✉email: pekka.ruusuvoori@utu.fi

tumor cell segmentation, prognostication, and biomarkers prediction^{11–18}, but gaining regulatory approval for such digital tools is a long-drawn and arduous journey. Only a handful of these efforts have successfully translated into AI-assisted diagnostic tools approved by regulatory authorities for clinical use¹⁹. One of the biggest challenges in developing tools that can be widely applied is their generalization capability, which is often significantly hindered by the high variation of tissue appearance stemming from the sample preparation, staining, and imaging^{20,21}.

Several approaches have been suggested to overcome the challenge set by the unwanted color variation. A simple yet logistically non-trivial solution is to source training data from a large number of different centers, which has shown to improve the generalizability of the AI models²². However, ensuring a representative range of variation is a big challenge. Multi-center setups are typically limited in breadth, with source data commonly originating from 2 to 6 sites. Such setups hardly capture the full range of variation across laboratories. Another approach to tackle staining variation and other causes for domain shift in histopathology is to aim at domain generalization through increasing variation in training data using synthetic data augmentation - commonly used in machine learning as a way to mimic the variations by computational means²³. It helps in reducing generalization errors²⁴. The augmentation techniques, however, still fall short of simulating real world color and stain variations because of their highly non-linear and dynamic nature and partially owing to the lack of datasets revealing the full range of variation, hindering the possibility to ground the augmentation in actual observed variation.

An alternative approach to stain augmentation is to reduce the variation in the input data originating from different data domains, such as the staining differences between source data centers. This approach is called stain normalization, where tissue images from one center are normalized with respect to a representative image or set of images from another center²⁵. Like data normalization in general, it can help to improve model generalizability, convergence pace, and robustness against overfitting to the extreme ends of the color spectrum. Stain normalization not only helps AI models to perform well with data originating from different sources, but a recent study has shown that the color uniformity of stain normalized tissue images can even help pathologists to speed up diagnoses with increased diagnostic confidence²⁶. Stain normalization methods are mainly divided into two categories: traditional methods typically based on mathematical frameworks and AI-based methods²¹ mostly using generative adversarial network (GAN)²⁷.

Traditional stain normalization methods based on mathematical frameworks often employ techniques such as moments (statistical) matching in different color spaces or color deconvolution whereby stains are first decoupled and then normalized separately. Typically, traditional methods rely on a reference (representative) patch to learn the target staining template and the choice of the reference patch is subjective depending on a specialist's assessment. In some cases, a composite patch is generated by combining several patches containing different morphological elements of the tissue. Some of the most commonly used of such approaches are Reinhard²⁸, Macenko²⁹, Vahadane³⁰, and the color histogram matching³¹. Prior to the widespread adoption of deep learning, although not commonplace, traditional stain normalization techniques were mostly used as a standardization step in tissue analysis^{32–34}.

AI-based stain normalization predominantly employs GANs²⁷ as an image-to-image translation method to transform source or unnormalized tissue images to match the color of the target or reference tissue image or images. Generative models are typically used to generate synthetic data. A typical GAN consists of a generator model, responsible for data synthesis, and a discriminator model, responsible for quality assurance feedback of the synthesized data. Both components are trained in a game theoretic way until a performance equilibrium is reached. In AI-based stain normalization, both unsupervised and supervised learning approaches are employed. CycleGAN³⁵, a very commonly used unsupervised image-to-image translation technique that does not rely on aligned image pairs, has been used for stain-to-stain translation^{36,37} and virtual histopathology staining^{38–40}. Similarly, it has been very effective for stain normalization as well, de Bel et al. used CycleGAN, with a UNet-like⁴¹ generator architecture, to normalize cross-center whole slide image (WSI) data consisting of Periodic Acid-Schiff (PAS) stained renal tissue. They further tested the effectiveness of normalization on a downstream segmentation task and observed a significant increase in the Dice coefficient of normalized WSIs⁴². De Bel et al. again used CycleGAN, with enhanced configuration using residual learning, on WSI data sourced from six different centers with a mix of colon and kidney samples stained with H&E and PAS. They demonstrated that the modified CycleGAN with residual learning outperforms traditional methods and vanilla CycleGAN on the downstream segmentation task keeping the tissue structure intact⁴³. A supervised image-to-image translation method called Pix2pix⁴⁴ has been widely adopted in domains where aligned image data is available. Despite aligned image pairs not being a possibility in the case of stain normalization, for supervised learning methods, typically, the input-output image pair consists of the grayscale version of the stained tissue patch as the input, and the RGB version of the same patch as the output which are inherently aligned. Salehi et al. used Pix2pix with such experimental setup and using quantitative and perceptual evaluation demonstrated the effectiveness of this method over traditional stain normalization approaches⁴⁵.

In recent years, many similar GAN-based methods have been proposed for stain normalization^{21,46–49}. While these studies provide valuable information about the strengths of these methods, their scope is limited by the number of centers involved. Most studies sourced tissue image data from two sites with a few exceptions, such as De Bal et al.⁴³, wherein six different sites provided tissue images for the study. In this study, we addressed the issue by analyzing the breadth of staining variation and compared different traditional and GAN-based stain normalization methods using a tissue image dataset with an unprecedented extent of variation, composed of H&E-stained slides collected from 66 different laboratories across 11 countries. We further provide the dataset as a resource for further stain normalization testing and development.

Results

The image data used in this study consists of H&E-stained skin, kidney, and colon tissue sections originating from the same blocks but stained in different laboratories. As expected, the visual appearance of the stained tissue samples varied quite drastically due to color variation originating from different stain compositions and staining protocols in use at different sites (Fig. 1). To examine the differences in color appearance quantitatively, we computed the mean intensity of red and blue channels and plotted the ratio of all the samples (Fig. 2). We then performed a comparative analysis of four traditional stain normalization methods, i.e., histogram matching³¹, Macenko²⁹, Reinhard²⁸, and Vahadane³⁰ and two GAN-based methods CycleGAN³⁵ and Pix2pix⁴⁴ with two variants each. For CycleGAN two different generators were used, one with a UNet-based⁴¹ generator and the other with a ResNet-based⁵⁰ generator. For Pix2pix, one with UNet-based generator and the other with DenseUNet-based generator which has shown to significantly reduce hallucination artifacts in a relatively similar histological image-to-image translation problem, i.e., virtual H&E staining of unstained tissue images³⁹.

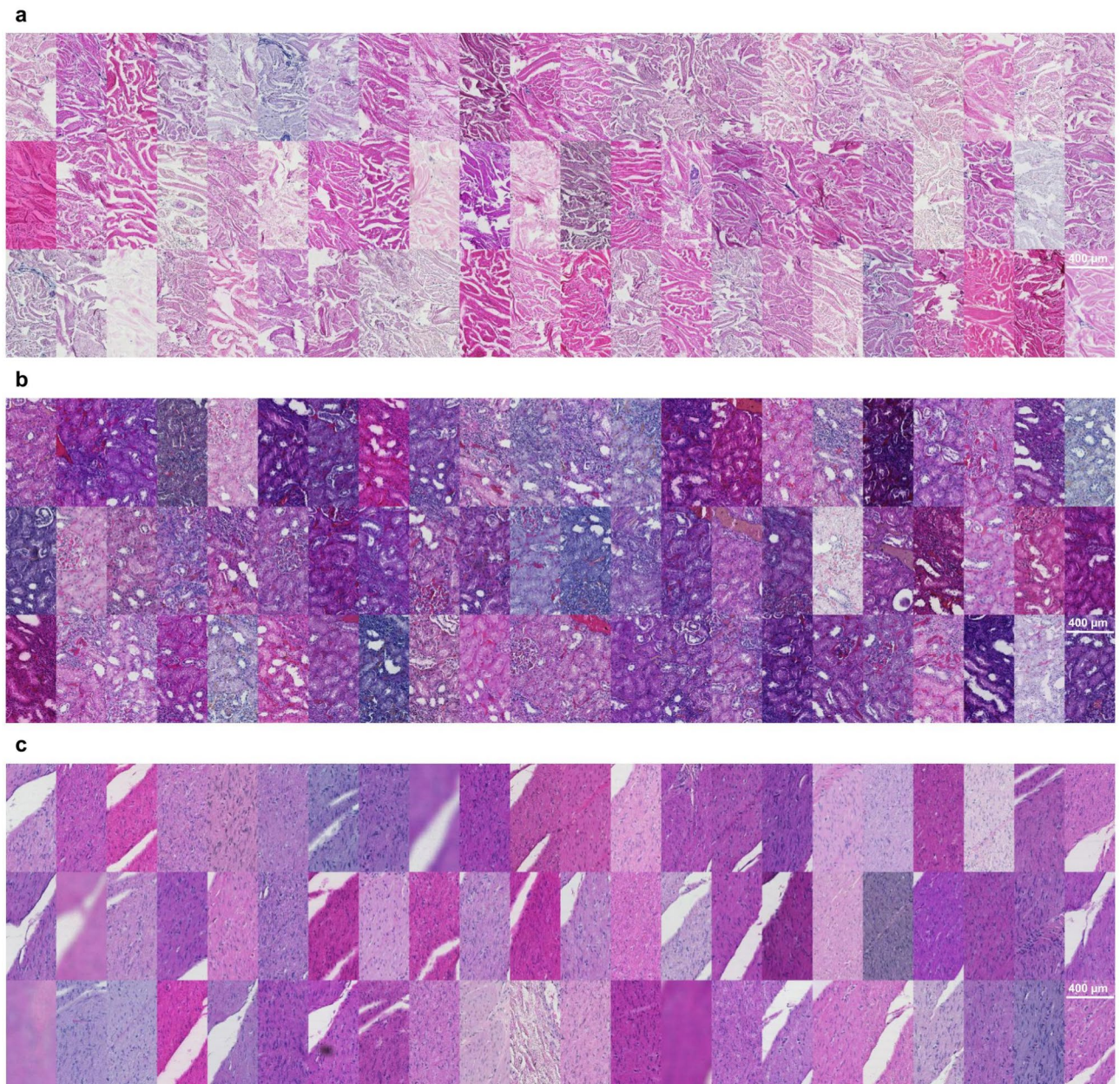


Fig. 1. A collage of patches representing variance in the appearance of tissue sections extracted from the same tissue block. **(a)** Image patches extracted from the dermis layer of skin tissue sections. **(b)** Image patches extracted from kidney tissue sections. **(c)** Image patches primarily extracted from the smooth muscle layer of colon tissue sections with a single exception of submucosal connective tissue (row 3 column 11).

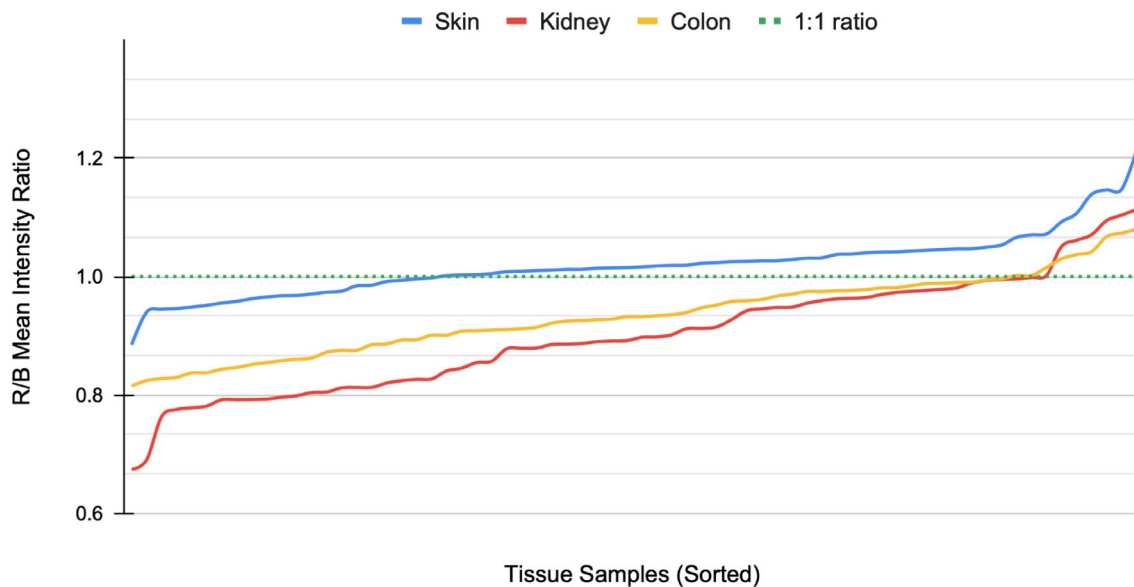


Fig. 2. Pre-normalization red and blue channel mean intensity plot. The plot shows the spread of color variation based on the ratio of red and blue channel mean intensities of the tissue images. All three tissues, i.e., skin, kidney and colon WSIs deviate drastically from the color-balanced (dotted) line.

Methods	Intersection		PCC		Euclidean Distance		JS Divergence	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std
CycleGAN (Resnet)	0.871	0.042	0.926	0.050	0.321	0.097	0.140	0.050
CycleGAN (UNet)	0.840	0.050	0.885	0.063	0.398	0.111	0.167	0.062
Pix2pix (UNet)	0.816	0.071	0.851	0.099	0.436	0.150	0.202	0.077
Pix2pix (DenseUNet)	0.822	0.077	0.856	0.104	0.425	0.161	0.191	0.080
Histogram Matching	0.891	0.034	0.938	0.050	0.279	0.101	0.119	0.041
Macenko	0.822	0.055	0.748	0.095	0.612	0.119	0.240	0.058
Reinhard	0.829	0.062	0.819	0.102	0.499	0.147	0.207	0.079
Vahadane	0.706	0.056	0.638	0.092	0.732	0.102	0.345	0.054
Original	0.779	0.088	0.803	0.119	0.503	0.170	0.220	0.081

Table 1. Skin tissue samples stain normalization comparative evaluation stats. Table shows mean intersection, Pearson correlation coefficient (PCC), Euclidean distance, Jensen-Shannon (JS) divergence scores of color-normalized WSIs, generated by the eight methods and the original WSIs compared with the reference WSI. Scores were computed on channel histograms in the $l\alpha\beta$ color space. Bold digits represent the best performing method.

Quantitative evaluation

We used three different types of quantitative evaluation methods to assess the performance of stain normalization methods. First, to assess the performance on color transfer task, the normalized images were transformed to the $l\alpha\beta$ color space⁵¹ and then the normalized histogram of each channel was compared with the reference image histograms using intersection, Pearson correlation coefficient (PCC), euclidean distance, and Jensen-Shannon (JS) divergence. Then, using an InceptionV3 model⁵², bottleneck features were extracted for all normalized images and were compared against reference image features using Frachet Inception Distance (FID)⁵³, a distance measure that works on high-level abstracted features taking into account both style and structure. Finally, to evaluate the structural integrity of the normalized tissue images with respect to the original ones, we used structural similarity index measure (SSIM)⁵⁴.

Skin

Skin samples were from tissue microarray punch biopsy containing epidermis and dermis (Fig. 1a). Histogram matching emerged as the most efficient normalization method and achieved mean scores of 0.891, 0.938, 0.279, and 0.119 for intersection, PCC, Euclidean distance and JS divergence, respectively (Table 1). As far as high-level feature-based similarity is concerned, histogram matching again, outperformed the rest of the methods with a mean FID score of 61.67 as compared to the original images with a mean FID score of 69.49 (Table 4). The

structural similarity was best preserved by Vahadane normalization with a mean SSIM score of 0.995, however, it was the worst performing method of all as far as the other metrics are concerned. Methods like Reinhard and histogram matching also had high mean SSIM scores of 0.977 and 0.955, respectively (Table 5). Overall, mean SSIM scores were over 0.92 which shows that tissue structure was mostly preserved well by all the methods (Table 5). Figure 3 shows an example of stain-normalized skin tissue from each method.

Kidney

Kidney samples were from tissue microarray punch biopsy, representing the cortical layer of the kidney, comprising tubules and renal corpuscles and vasculature (Fig. 1b). Similar to skin, histogram matching achieved the best mean scores in all the metrics for kidney tissue as well, 0.944, 0.985, 0.144, and 0.101 for intersection, PCC, Euclidean distance and JS divergence, respectively (Table 2). Histogram normalization performed exceptionally well as compared to other methods likely due to the morphological uniformity of kidney tissue images. In high-level feature comparison, histogram normalization achieved a mean FID score of 55.38 as compared to 69.54 of the original images (Table 5). For SSIM scores we observed a similar trend as skin tissue images, i.e., Vahadane preserving the structural similarity best with a mean SSIM score of 0.967 and all mean SSIM scores being over 0.945 indicating that all the methods performed quite well in preserving the tissue structure (Table 5). Figure 4 shows an example of stain-normalized kidney tissue from each method.

Colon

Colon tissue samples comprehensively represented the large intestine, containing mucosa, muscularis mucosae, submucosa and the outer muscle layers, with occasional mucosa-associated lymphoid structures (Fig. 1c). Intra-tissue morphological heterogeneity peaked in colon tissue as compared to the skin and kidney tissues, however, since the tissue specimens were sectioned from the same tissue block there still existed inter-tissue structural uniformity. We noticed that even for colon tissues images, histogram matching outperformed other methods with a mean scores of 0.906, 0.935, 0.295, and 0.143 for intersection, PCC, Euclidean distance and JS divergence, respectively (Table 3). In high-level feature comparison, however, Reinhard and both CycleGAN (Resnet-based and Unet-based) variants performed better than histogram matching with the mean FID scores of 92.37, 96.12, 96.50, and 99.92, respectively (Table 4). Similar to skin and kidney, Vahadane performed the best in preserving structural similarity with a mean SSIM score of 0.989 (Table 5). It was closely followed by Reinhard

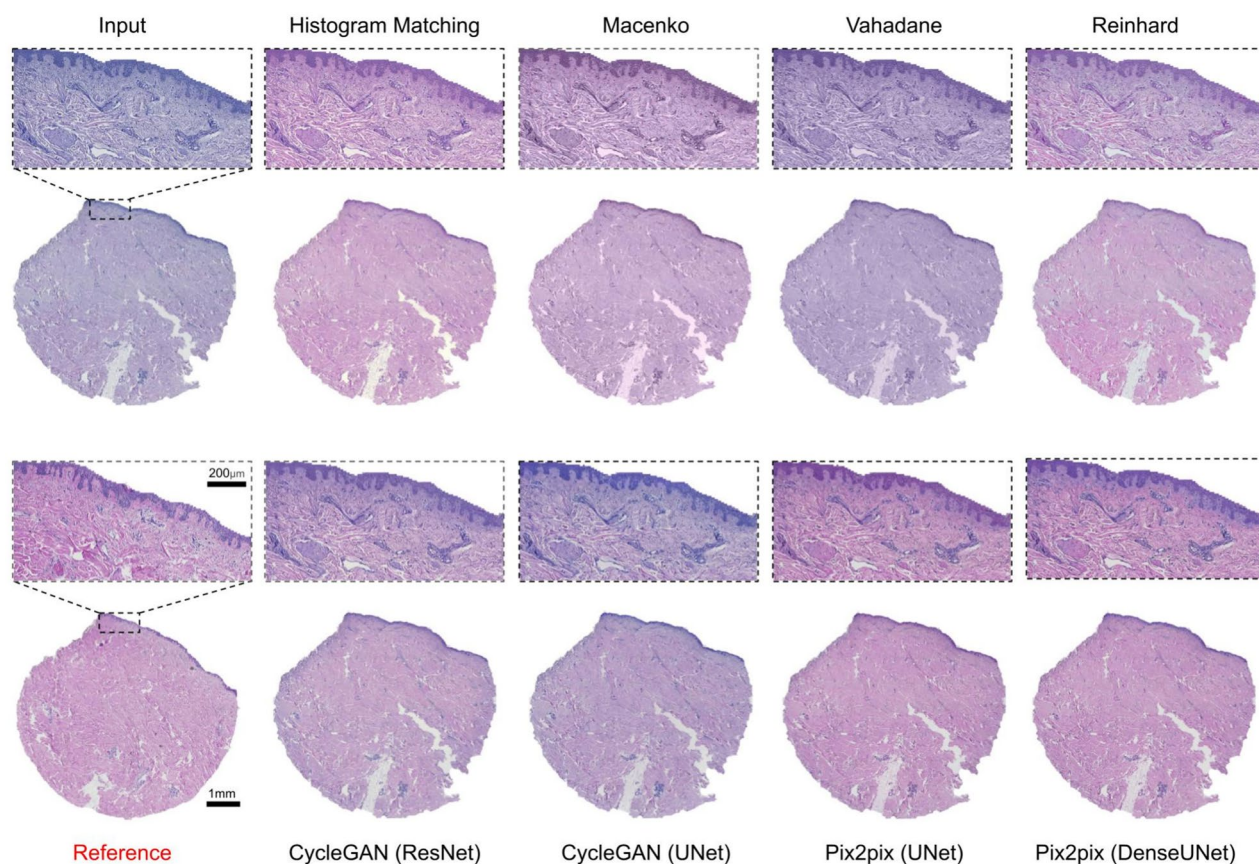


Fig. 3. Skin tissue stain normalization. Stain-normalized tissue images from eight different methods, i.e., histogram matching, Macenko, Vahadane, Reinhard, CycleGAN (ResNet), CycleGAN (UNet), Pix2pix (UNet), Pix2pix (DenseUNet) along with input and reference WSI.

Methods	Intersection		PCC		Euclidean Distance		JS Divergence	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std
CycleGAN (Resnet)	0.895	0.040	0.927	0.040	0.310	0.070	0.159	0.027
CycleGAN (UNet)	0.857	0.062	0.890	0.089	0.370	0.108	0.187	0.042
Pix2pix (UNet)	0.782	0.139	0.764	0.217	0.515	0.214	0.246	0.101
Pix2pix (DenseUNet)	0.794	0.121	0.795	0.188	0.472	0.191	0.227	0.085
Histogram Matching	0.944	0.022	0.985	0.008	0.144	0.037	0.101	0.030
Macenko	0.784	0.065	0.823	0.093	0.476	0.120	0.242	0.045
Reinhard	0.854	0.072	0.885	0.075	0.389	0.106	0.195	0.053
Vahadane	0.676	0.078	0.578	0.135	0.770	0.140	0.378	0.056
Original	0.721	0.115	0.722	0.189	0.549	0.200	0.251	0.093

Table 2. Kidney tissue samples stain normalization comparative evaluation stats. Table shows mean intersection, Pearson correlation coefficient (PCC), Euclidean distance, Jensen-Shannon (JS) divergence scores of color-normalized WSIs, generated by the eight methods and the original WSIs compared with the reference WSI. Scores were computed on channel histograms in the $l\alpha\beta$ color space. Bold digits represent the best performing method.

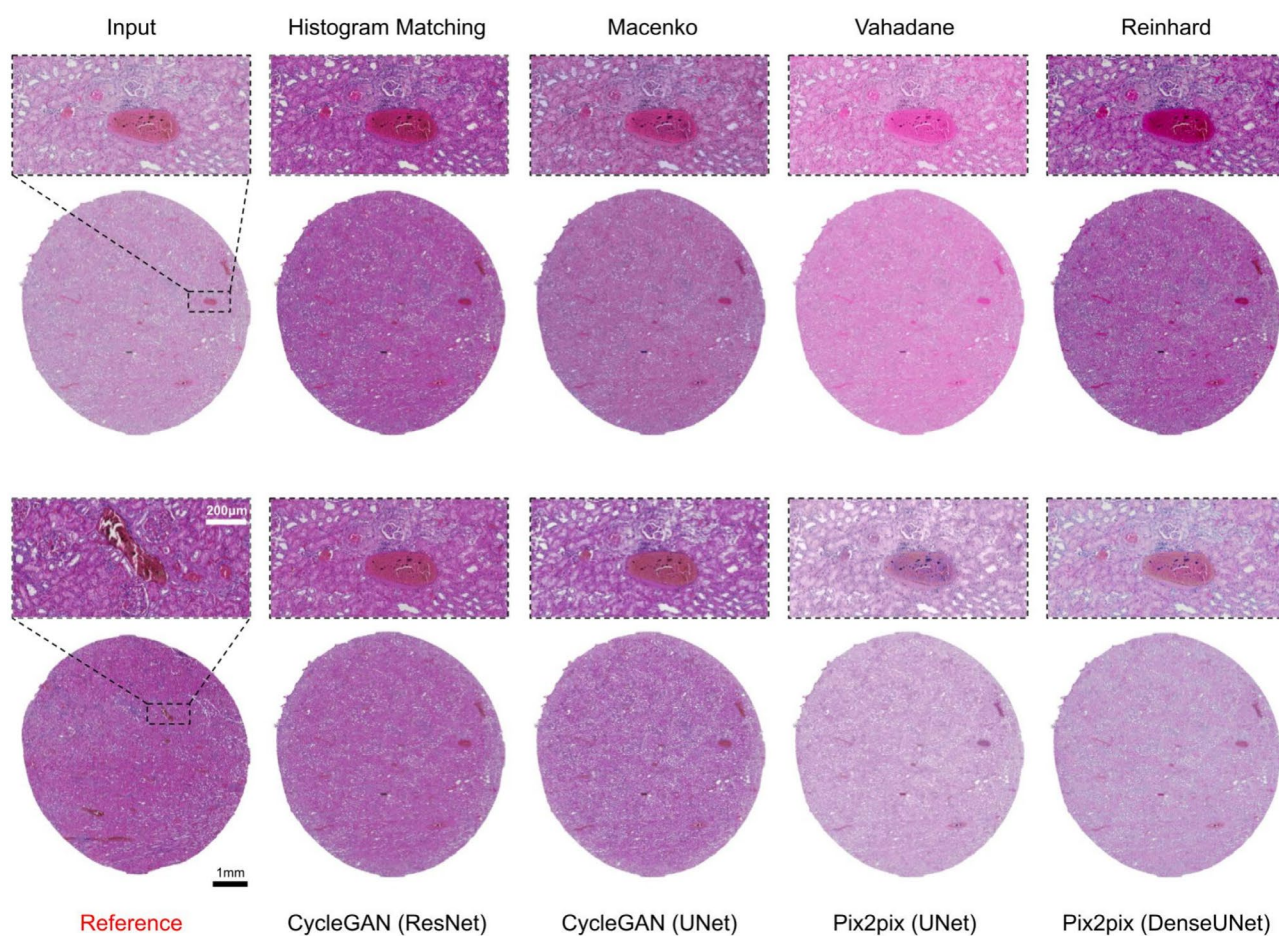


Fig. 4. Kidney tissue stain normalization. Stain-normalized tissue images from eight different methods, i.e., histogram matching, Macenko, Vahadane, Reinhard, CycleGAN (ResNet), CycleGAN (UNet), Pix2pix (UNet), Pix2pix (DenseUNet) along with input and reference WSI.

Methods	Intersection		PCC		Euclidean Distance		JS Divergence	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std
CycleGAN (Resnet)	0.869	0.045	0.875	0.045	0.422	0.079	0.186	0.035
CycleGAN (UNet)	0.803	0.091	0.800	0.113	0.518	0.149	0.231	0.070
Pix2pix (UNet)	0.727	0.100	0.667	0.178	0.630	0.178	0.279	0.078
Pix2pix (DenseUNet)	0.748	0.094	0.681	0.158	0.633	0.164	0.270	0.067
Histogram Matching	0.906	0.022	0.935	0.026	0.295	0.058	0.143	0.035
Macenko	0.817	0.049	0.810	0.072	0.522	0.094	0.246	0.042
Reinhard	0.839	0.061	0.839	0.074	0.477	0.108	0.223	0.063
Vahadane	0.726	0.066	0.662	0.120	0.695	0.127	0.336	0.055
Original	0.739	0.082	0.705	0.131	0.595	0.153	0.260	0.069

Table 3. Colon tissue samples stain normalization comparative evaluation stats. Table shows mean intersection, Pearson correlation coefficient (PCC), Euclidean distance, Jensen-Shannon (JS) divergence scores of color-normalized WSIs, generated by the eight methods and the original WSIs compared with the reference WSI. Scores were computed on channel histograms in the $l\alpha\beta$ color space. Bold digits represent the best performing method.

Methods	Skin		Kidney		Colon	
	Mean	Std	Mean	Std	Mean	Std
CycleGAN (Resnet)	92.54	20.57	96.27	21.38	96.12	40.40
CycleGAN (UNet)	101.63	21.57	91.68	24.22	96.50	18.54
Pix2pix (UNet)	109.00	20.49	118.17	35.51	123.10	54.75
Pix2pix (DenseUNet)	109.06	22.67	117.72	36.21	119.05	55.92
Histogram Matching	61.67	24.29	55.38	38.62	99.92	61.76
Macenko	89.65	40.54	106.51	52.78	134.63	52.69
Reinhard	62.08	31.30	64.43	39.68	92.37	65.58
Vahadane	134.57	17.99	153.15	49.35	150.15	55.01
Original	69.49	37.35	69.54	42.26	96.88	64.28

Table 4. Frachet inception distance (FID) for skin, kidney and colon tissue samples. Color-normalized WSIs, generated by the eight methods and the original WSIs were compared using FID, a metric that uses high-level features extracted through the bottleneck layer of models like InceptionV3. Bold digits represent the best performing method.

Methods	Skin		Kidney		Colon	
	Mean	Std	Mean	Std	Mean	Std
CycleGAN (Resnet)	0.940	0.019	0.946	0.023	0.860	0.117
CycleGAN (UNet)	0.920	0.024	0.949	0.015	0.822	0.118
Pix2pix (UNet)	0.955	0.010	0.957	0.019	0.933	0.015
Pix2pix (DenseUNet)	0.958	0.010	0.963	0.008	0.945	0.011
Histogram Matching	0.955	0.048	0.949	0.050	0.951	0.040
Macenko	0.937	0.067	0.945	0.034	0.891	0.039
Reinhard	0.977	0.027	0.956	0.045	0.968	0.033
Vahadane	0.995	0.005	0.967	0.027	0.989	0.011

Table 5. Structural similarity index measure (SSIM) for skin, kidney and colon tissue samples. Color-normalized WSIs, generated by the eight methods, were compared against the original WSIs to test the (tissue) structural preservation of the methods, using structural similarity index measure (SSIM). Bold digits represent the best performing method.

and histogram matching with mean SSIM scores of 0.968 and 0.951 (Table 5). Figure 4 shows an example of stain-normalized colon tissue from each method.

Qualitative evaluation

We examined the quality of the normalized staining by visually comparing a subset of original and stain-normalized colon specimens, with respect to distinguishability of tissue structures. We evaluated the visual quality of hematoxylin-eosin hue and contrast, as well as structural or other digital artifacts, as these factors contribute to the interpretability of histopathological H&E images. Since the colon contains several different tissue structures and types, it serves as a representative example that we especially focus here on. The colon consists of four distinct layers: a) the mucosa, comprising glandular epithelia with interconnected connective tissue (lamina propria) and underlying smooth muscle (muscularis mucosae); b) the submucosa, comprising dense connective tissue with blood and lymphatic vessels; c) muscularis externa composed of smooth muscle layers; as well as d) serosa or adventitia comprising thin connective tissue and an outer layer of squamous epithelia or mesothelium. As the inner two layers include prominent variation in structure, harboring epithelial cells, fibroblasts, vessels, erythrocytes, and smooth muscle, we primarily focused on these layers, while asking how well the normalization preserves fine detail. Because over-staining, under-staining as well as uneven staining were the most notable features of the unprocessed tissue images, we focused on the mitigation of these issues when evaluating the effectiveness of normalization methods. In addition, we compared the methods to a reference section, to evaluate how well the methods achieved the sought-out result.

Mitigation of staining quality issues

Visually separable contrast between blue/purple (hematoxylin) and pink/red (eosin), as well as the dynamic range of eosin staining, are central for high-quality H&E staining suitable for histological interpretation. Thus, the hue of the H&E-image is an important determinant of quality. A reference image of a high-quality H&E is shown in Fig. 6a. With the applied methodology, in general, if normalization resulted in a global shift in hue, it was towards pink/red. A hue-shift of varying strength was present in all methods (Fig. 5). Most notably, the Vahadane method transformed the white background to pink in most instances, and completely disposed of hematoxylin's blue/purple. The shift towards pink/red may compromise the separability of veins, fibrous components, and actin bundles. With histogram matching, in some instances, the hue was improved with heavily over-stained images, resulting in a more conventional coloration and pronounced contrast, albeit under-stained

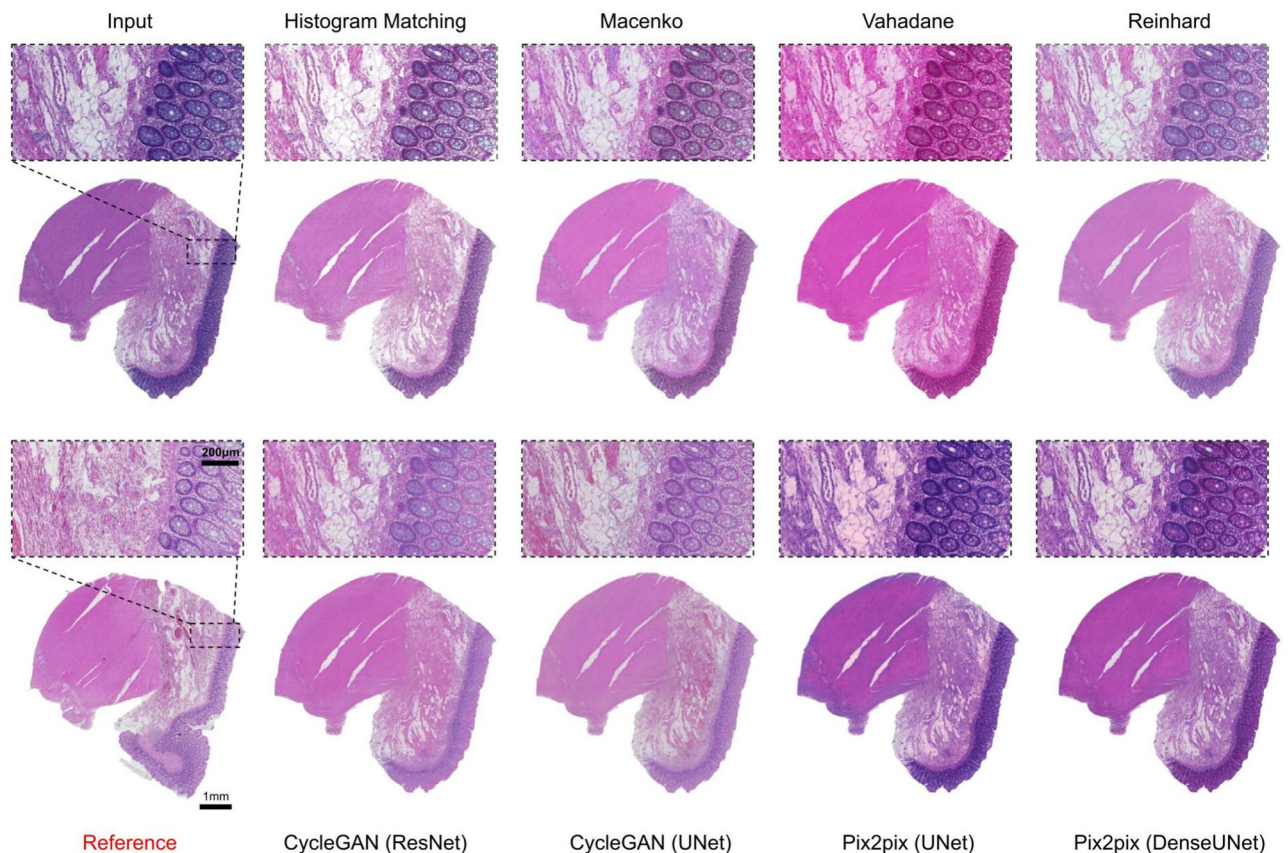


Fig. 5. Colon tissue stain normalization. Stain-normalized tissue images from eight different methods, i.e., histogram matching, Macenko, Vahadane, Reinhard, CycleGAN (ResNet), CycleGAN (UNet), Pix2pix (UNet), Pix2pix (DenseUNet) along with input and reference WSI.

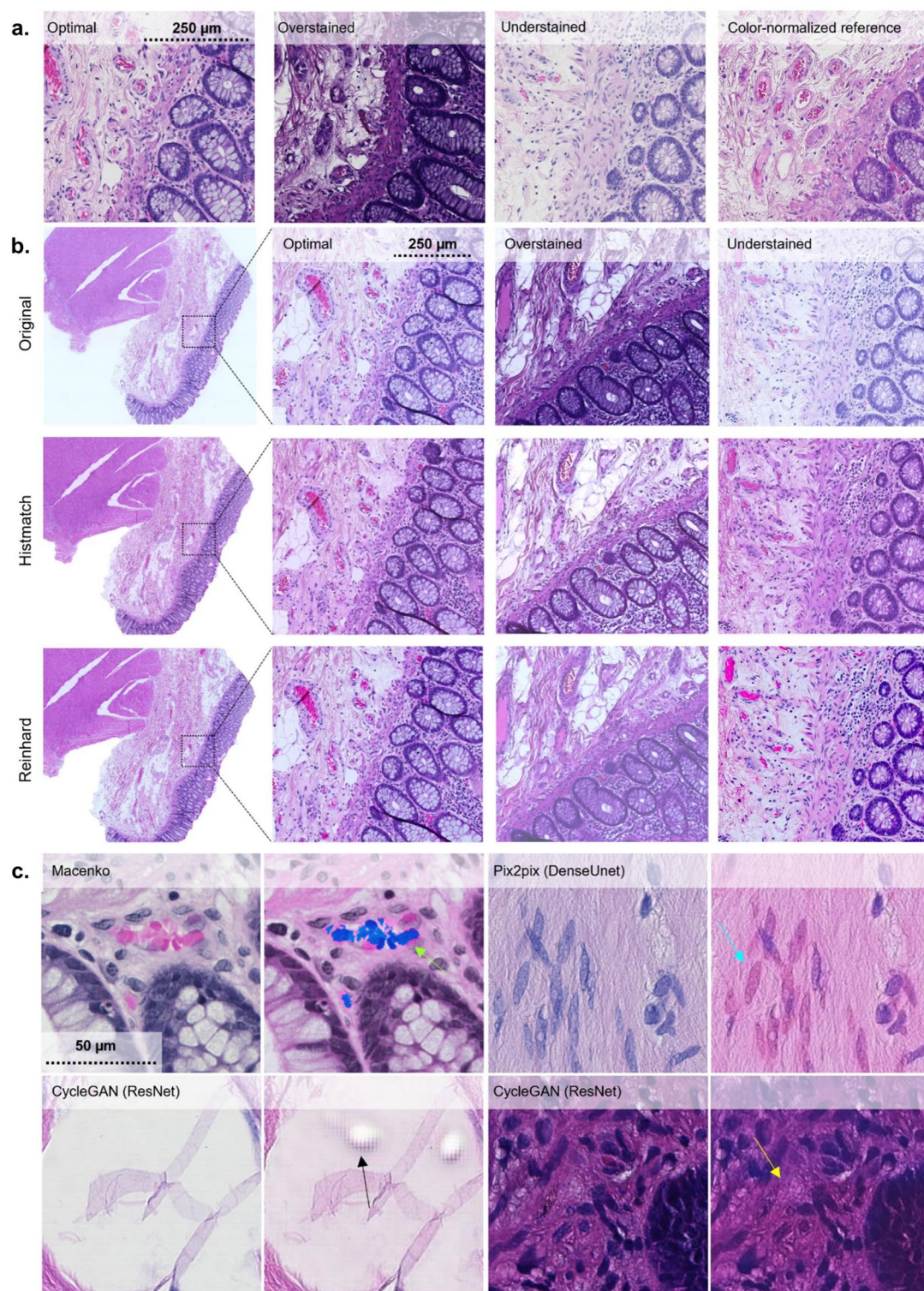


Fig. 6. Colon tissue qualitative analysis. **(a)** Representative images of H&E-slides that are stained either optimally in terms of staining intensity for pathological interpretation, or suboptimally (over- or understained). **(b)** Comparisons between original, Histmatch, and Reinhard normalized images from colonic mucosa-submucosa border in optimally and suboptimally stained slides. **(c)** Artifacts emerging from image processing.

images tended to shift heavily towards pink/red (Fig. 6b). With under-stained images, Macenko and Reinhard performed best in correcting substandard contrast, while not performing well with overstaining (Fig. 6b).

Some of the utilized normalization methods generated digital artifacts. From the conventional methods, Macenko transformed the color of erythrocytes and parts of the lamina propria to blue. With deep-learning models, generative artifacts were observed with CycleGAN (ResNet) in empty (fat containing) regions of adipose tissue, as well as, with Pix2pix (DenseUNet), some spindly nuclei from smooth-muscle were transformed red. In addition, CycleGAN (ResNet) incorporated some context-dependent tile artifacts at luminal epithelium – smooth muscle borders (Fig. 6c). We did not observe artifacts with histogram matching and Reinhard.

All in all, no method improved visual quality in all instances of suboptimal staining. Histogram matching resulted in improved quality in heavily over-stained images, albeit not performing equally well with under-stained images. Reinhard improved under-stained images, albeit producing below par results in over-stained images.

Performance comparison to reference

We next visually evaluated if the methods match the utilized standard, i.e., the color-normalized reference image. The worst performance was observed in Vahadane, which completely discarded hematoxylin. Reinhard increased overall color intensities, resulting in more perceived contrast, thus deviating from the reference. In addition, the hue of hematoxylin varied, and in overstained specimens, the normalized image appeared cloudy. Macenko varied, depending on the original image, most notably in hematoxylin and background hue, as well as blue artifacts in eosin, that deviated from the reference. Macenko also had significant variance and deviation from the reference in both contrast and hue. Reinhard harbored variance in contrast and in the color of the epithelial lining of the lumen. Pix2pix (DenseUNet) was close to reference in most instances, albeit in uneven staining the normalization did deviate. In addition, the other generative AI-models, CycleGAN (UNet) and CycleGAN (ResNet) achieved reasonable similarity with the reference. Histogram matching was the best performing method in terms of evenness and comparability to the reference. In histogram matching, the most notable deviation from the reference was the background, which in some instances had a pink hue. However, when considering only the ROI, histogram matching harbored the most even performance compared to a reference image, both with smaller magnification and when evaluating the finer details in different tissue compartments. An overall comparison of the best performing methods is shown in Fig. 7.

Effect of staining normalization to downstream tasks

To complement the quantitative analysis of staining normalization performance and qualitative histological inspection, we perform indirect analysis of the performance of normalization methods through their effect on downstream analysis tasks. To assess how the stain normalization methods affect the results, we use two common computational pathology downstream analysis tasks; nucleus detection and feature extraction using histopathology foundation model.

Nucleus-level quantification

First, we quantified the effect of staining normalization methods on the number of segmented nuclei. Segmentation was implemented using state-of-the-art nucleus instance segmentation method Cellpose-SAM⁵⁵, with hyperparameters tuned for the reference sample. The same reference sample was used as the reference target for staining normalization methods. Using the same parameter settings, we then proceeded to segment the nuclei for all other images in the original set as well as for the staining normalized images produced by all 8 methods. The results are presented as boxplots in Fig. 8, showing the average cell counts, as well as the variation between images. The cell count of the reference image is illustrated as a blue dashed line.

The nucleus quantification experiment confirms that normalization has a potential major effect on downstream tasks, but that the effect is not systematic across methods in terms of the average number of detected nuclei in the normalized samples or in terms of the range of deviation in the numbers of detected nuclei. The deep learning

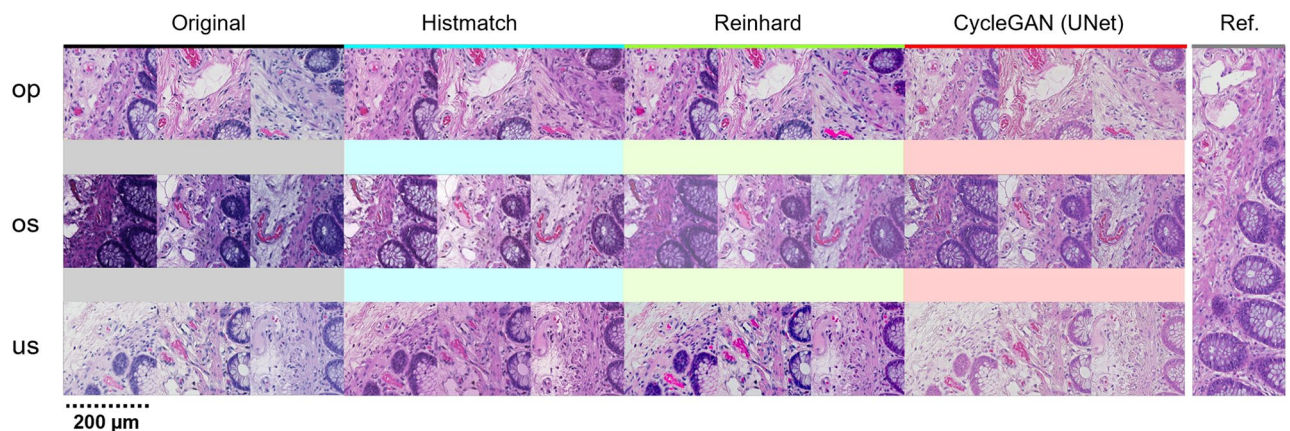


Fig. 7. Overview of the best performing methods (columns) with varying input staining quality (rows).

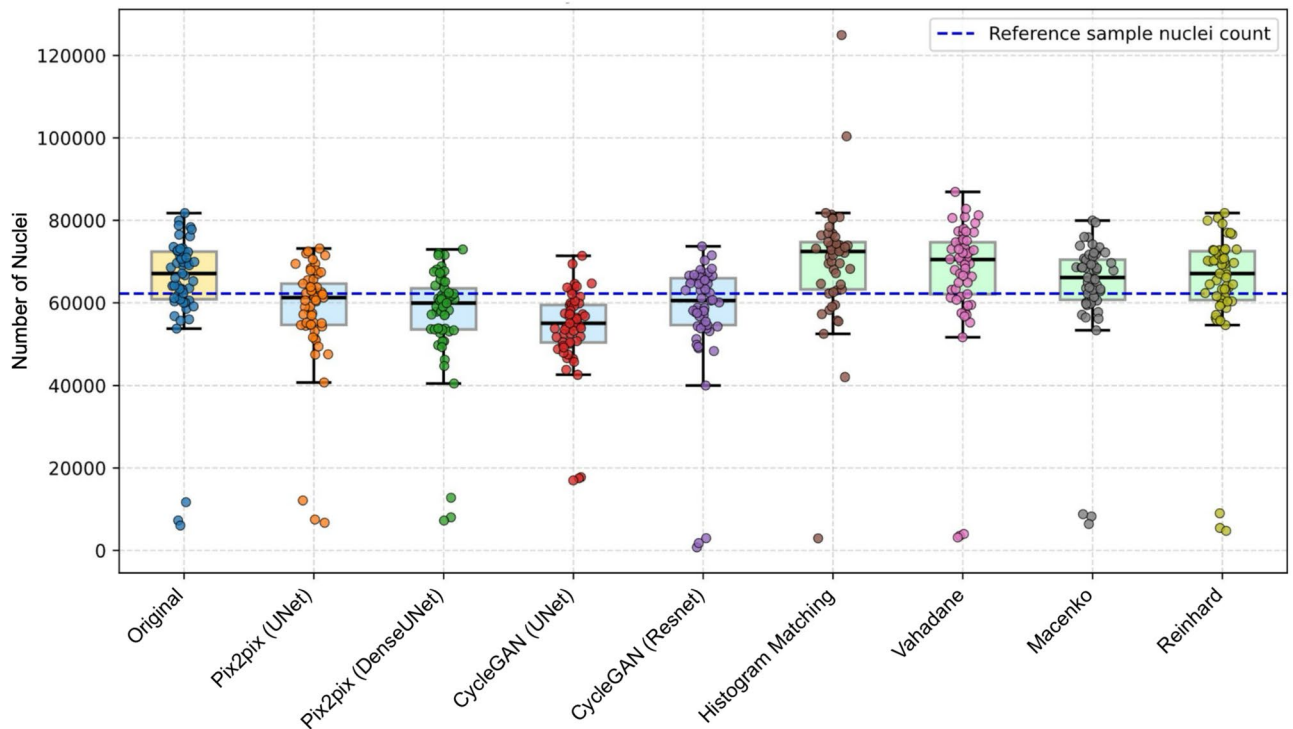


Fig. 8. Comparison of nuclei count for the stain normalization methods. The boxplot displays the nuclei counts achieved by Cellpose-SAM cell segmentation model for images by each normalization method, showing how close each method's results are to the reference sample's nuclei count.

based methods tend to lead to a lower number of detections, and especially Pix2pix (UNet), CycleGAN (Resnet), and Pix2pix (DenseUNet) yield counts closest to that of the reference sample. Among traditional methods, histogram matching gives detection counts closest to the reference count. It should be noted that the assumption that the samples would have an equal number of nuclei despite being adjacent sections of the same tissue block does not hold and the nucleus count in the reference section should not be considered as the ground truth for the expected number of nuclei.

Foundation model feature extraction

Second, we examine the effect of normalization on feature extraction by a histopathology foundation model. The colon tissue images are converted into features using UNI-2 foundation model⁵⁶ by first tiling the tissue, which are then converted into feature embeddings. The resulting multi-dimensional feature vectors are used to characterize colon tissue sections. In Fig. 9, the feature embeddings for original WSIs (coral red) and the normalized WSIs by the 8 methods (blue) are shown in 2D using t-SNE for dimensionality reduction. For clarity, each normalization method is given its separate plot where the corresponding data as well as the unnormalized data are presented with color, and the rest of the datapoints in background with gray.

The t-SNE plots suggest that normalization has a clear effect on the feature representation obtained with a foundation model, and that the methods lead to different patterns in terms of how compact the point cloud is in t-SNE embedded feature space. Here, CycleGAN-based deep learning methods and Macenko normalization yield compact point clouds, whereas Vahadane and Pix2pix-based methods lead to scattered point clouds. Considering the nature of the dataset, the compact cluster can be considered a preferred outcome, suggesting the CycleGAN-based methods and Macenko method to be worth consideration in combination with the Uni-2 foundation model. Overall, the foundation model appears not to be robust against the intensity variation/normalization as all of the normalization methods had a substantial effect on the feature space representation.

Discussion

In this work, we present a tissue image dataset consisting of H&E-stained skin, kidney and colon tissues by staining sections of the same samples in 66 different laboratories to study the variation in the appearance of tissue images and assess the performance of different stain normalization techniques. The tissue sections were cut from the same block for each tissue type, representing a setup where a relatively high level of content similarity can be expected across the 66 sections. Tissue staining is one of the most critical steps in sample preparation and subtle variations in implementation of the staining protocol could contribute to the fluctuation in the appearance of the dyed tissue, commonly observed as variations in the intensity and hue of the staining. Previous studies have mostly employed tissue images sourced from a handful of sites^{42,43}. This study, however, presents a dataset from numerous laboratories that captures a wide range of variation in H&E staining (Fig. 1), while excluding other

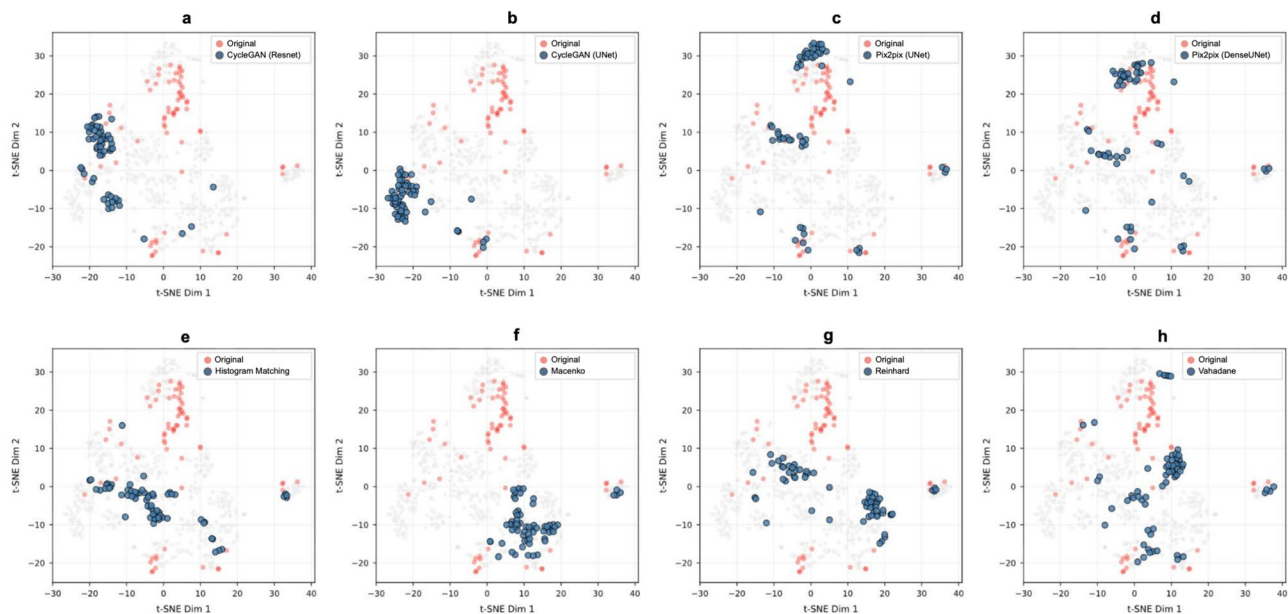


Fig. 9. A two dimensional visualization of foundation model extracted features. Each subplot illustrates the feature distribution in two-dimensional t-SNE embedding from the original, unnormalized WSIs together with those from images normalized by one of eight different stain normalization methods.

sources of technical and biological variation. The data is released as an open access resource for the community to enable benchmarking of normalization methods as well as other further investigation of staining variation in histopathology.

The substantial variation in staining offers an opportunity to assess the effectiveness of different strategies to normalize the color heterogeneity across the tissue samples and also evaluate their ability to maintain morphological interpretability of the tissue content. To that end, we employed four traditional and four GAN-based methods to normalize the staining in tissue images. The traditional methods included histogram matching, Macenko, Reinhard, and Vahadane stain normalization. Recently, methods based on generative AI have gained popularity for out-performing traditional methods^{42,43}. Here, we included GAN-based methods in addition to the traditional normalization methods. The GAN-based methods included two variants of Pix2pix, one with UNet-based generator, the other with DenseUNet-inspired generator, and two variants of CycleGAN, one using UNet-based generator and the other using ResNet-based generator.

In this study, both quantitative and qualitative evaluation revealed that WSIs normalized by variants of CycleGAN and Pix2pix bore reasonable resemblance with the reference sample and largely maintained structural integrity, however, the methods could not conclusively outperform traditional methods across all tissue types and also produced unwanted hallucination artifacts in some instances. The reason for such performance is potentially rooted in the fact that one WSI per laboratory is insufficient for yielding optimal results from the data-hungry deep learning models. For instance, only the reference WSI and its grayscale version, as a single source-target pair, was used for training both Pix2pix and its DenseUNet variant. Although more than one WSI were used in CycleGAN (both ResNet and UNet variants) trainings, only one representative WSI per cluster (representing smaller subsets of the tissue image dataset) was chosen for the training set for each tissue type.

On the other hand, histogram matching as a traditional stain normalization method performed relatively well across all three tissue types compared to all the other methods (Figs. 10 and 11). This performance can be attributed to at least two different factors. 1) For traditional methods, typically a small representative patch is selected as reference. We, instead, used the whole tissue image as the reference sample, and as opposed to applying normalization patch by patch which typically leads to tiling artifacts, we applied normalization to the WSIs at once. Reinhard and Macenko normalized WSIs also seem to benefit from this strategy. 2) Although the dataset captures a broad spectrum of staining variation, the tissue content across the sections does not change much by virtue of originating from the same tissue block. Further explanations for the better performance of histogram matching and Reinhard normalization could be the fact that both methods adjust the statistical attributes of the image (histogram, mean, and standard deviation) and the adjustments are applied globally, this enables the methods to leverage the morphological similarity factor. However, it is important to note that not all traditional methods yielded satisfactory results. For instance, Macenko sporadically produced blue artifacts in eosin regions, whereas Vahadane had the worst performance as it appeared to infuse a pink hue in the whole tissue image, overriding the hematoxylin component completely. This was clearly reflected in the quantitative evaluation as well. Overall, we emphasize the role of this study as an example and benchmark for the range of stain variation rather than a conclusive comparison of the most accurate normalization method.

This study highlights a broad spectrum of variation observed in the inter-laboratory staining manifestation, reinforcing similar observations made in⁵⁷, which also indicates that stain normalization is an indispensable

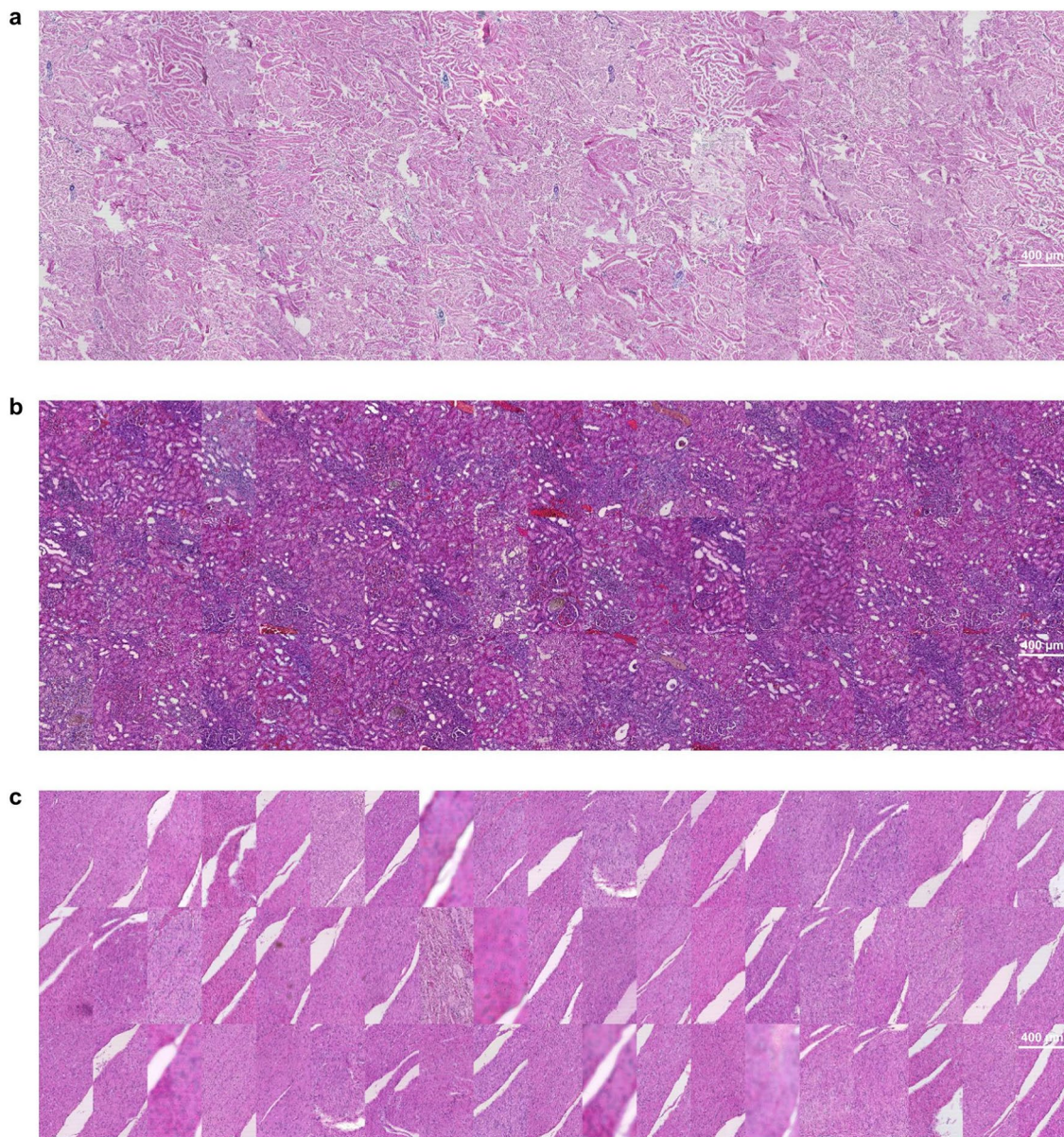


Fig. 10. A collage of post-normalization tissue patches showing improved color uniformity across samples extracted from the same tissue block. Tissue images stain normalized using histogram normalization. (a) Image patches extracted from skin tissue sections. (b) Image patches extracted from kidney tissue sections. (c) Image patches primarily extracted from the smooth muscle layer of colon tissue sections with a single exception of submucosal connective tissue (row 2 column 8).

step to homogenize the appearance of the tissue images whenever inter-laboratory comparisons or general tools are used (Figs. 6 and 7). Furthermore, it shows that while deep learning-based methods are effective for stain normalization, they have some shortcomings as well, and traditional methods may remain valuable in specific scenarios, such as the present case, where tissue sample morphology exhibits a high degree of uniformity. The dataset is shared as a freely available public resource. In addition to serving as a benchmark for staining normalization methods, it can also benefit the community in determining the extent of staining variation⁵⁸ which could be leveraged to enhance data augmentation techniques and domain generalization methods²³, thereby contributing to the development of more robust AI systems. While the dataset presented here is unique in terms of the variety it showcases, only a single section per laboratory does not facilitate evaluating intra-laboratory variation or quantifying the performance of, e.g., typical deep learning-based stain normalization methods wherein several samples per laboratory are required to effectively train a method and validate its results. Such setup would enable using downstream task performance, such as diagnostic accuracy as a measure of normalization method performance^{59,60}. A similar dataset containing more samples per laboratory and featuring more morphological variation across tissue samples could thus provide a stronger foundation for a more comprehensive study in the future.

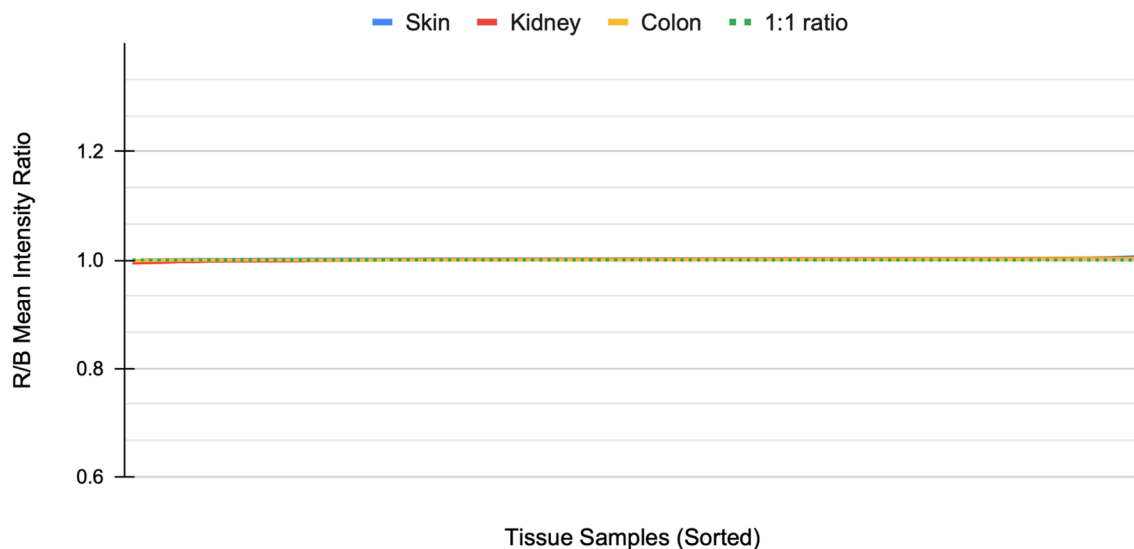


Fig. 11. Post-normalization red and blue channel mean intensity plot. The plot shows the convergence of red and blue channel mean intensity ratio of the tissue images after histogram matching stain normalization. All three tissue types, i.e., skin, kidney and colon WSIs are virtually on the color-balanced (dotted) line.

Material & methods

Tissue material

The H&E-stained tissue image data was obtained as part of an external quality assessment (EQA) initiative coordinated by Labquality, a company specializing in external quality assessment programs for clinical laboratories located in Helsinki, Finland. These slides featured a tissue microarray section comprising three 6 mm punch biopsies, each extracted from normal human skin, kidney, and colon. These tissue samples were obtained from anonymized, formalin-fixed, and paraffin-embedded histological specimens from a reference pathology laboratory. During the EQA round, slides with 3-micron thick unstained tissue sections were dispatched to EQA participant laboratories. These laboratories were instructed to apply their routine H&E staining methodology, typically used in their daily diagnostic practices. In total, 66 laboratories from 11 different countries participated in the evaluation process. Subsequently, these slides were digitized using a Hamamatsu Photonics NanoZoomer-XR slide scanner, employing a 20 \times objective lens, resulting in a scanning resolution of 0.46 μm per pixel.

The tissue images were resampled to 10 \times to expedite the normalization process, since 10 \times is sufficient for clinical diagnosis. Another reason for resampling to 10 \times was to aid traditional normalization method by enabling the use of whole tissue sample as the reference and also apply color normalization to the source tissue at once instead of using just a small tissue patch from the reference tissue sample and applying color normalization patch-wise to the source tissue sample which is the norm with traditional stain normalization methods. The rationale behind this approach was to achieve more uniform results and avoid tiling effects.

Reference sample selection

The reference sample was selected computationally using red-blue channel mean intensity ratio as the criterion. For each tissue type, the sample that had the red-to-blue ratio closest to one was selected as the reference sample (Fig. 2). The rationale behind selecting the blue and red channels stems from blue and red color expression being characteristic of Hematoxylin and Eosin stained tissue samples.

Traditional stain normalization methods

Histogram matching

It is an image processing technique that modifies pixel values in an image to transform its color distributions represented by its histogram so that it matches the color distribution of a reference image³¹. By normalizing the color distribution of the source image with respect to the reference image, the process aims to achieve color consistency between the two images, enabling effective image analysis and visualization across varying imaging conditions. While histogram matching is a general-purpose color normalization technique that is not specifically designed for histopathology images, it has been effectively used in the domain^{61,62}.

Macenko

In Macenko stain normalization, separate H&E vectors are computed using singular value decomposition (SVD) in the optical density space which is a logarithmic version of the RGB color space. Optical density space makes stain manipulation easier by enabling linear coupling and decoupling of stains. By scaling the stain vectors of the source image to match those of the reference image, Macenko normalizes staining variations, enabling consistent color representation across images²⁹.

Reinhard

Like histogram normalization, Reinhard normalization²⁸ is also a generic image normalization technique that was not specifically designed for histopathology images. In this method, the images are first converted to the $l\alpha\beta$ color space⁵¹. Then statistical standardization (subtracting a mean value and scaling by standard deviation) is applied to source images and then the images are linearly scaled using mean and standard deviation of the reference image, this is done separately for each channel. This normalization is effective because unlike RGB, $l\alpha\beta$ color space has low correlation between color channels.

Vahadane

It is another stain normalization approach that employs color deconvolution in optical density space, but unlike Macenko it uses sparse non-negative matrix factorization which is more effective than SVD. The method works by optimizing the deconvolution approach over successive iterations to generate two matrices, one for color appearance and the other for stain density. The source stain density is scaled, a step similar to Macenko's scaling step, and then by combining with the color appearance of the reference sample a normalized image is reconstructed³⁰.

We used a publicly available implementation of Macenko, Reinhard, and Vahadane stain normalization methods: https://github.com/wanghao14/Stain_Normalization, whereas histogram matching was used from the Python package *skimage*.

GAN-based stain normalization methods

CycleGAN

One of the most commonly used unsupervised image-to-image translation methods is called CycleGAN. The biggest advantage of using a method like CycleGAN is that it doesn't necessitate the source and target images to be aligned or have pixel-to-pixel correspondence. The method learns from a mechanism called distribution matching loss, which, in practice, is achieved by using the cycle-consistency loss function in case of CycleGAN. A typical CycleGAN model consists of two generator and two discriminator models. It is capable of bi-directional translation from source to target and vice versa. We used two different variants of CycleGAN generators, one based on ResNet⁵⁰ and the other based on UNet⁴¹ with skip connections.

Pix2pix

Domains wherein it is possible to generate paired image data, Pix2pix and its variants are the most widely used GAN-based supervised image-to-image translation methods. It requires the training (image) data to be well registered with pixel-level correspondence because it relies on direct pixel loss functions for its learning. It is based on conditional GAN⁴⁴. A typical Pix2pix network consists of a pair of generator and discriminator models that are trained in a game-theoretic way until a performance equilibrium is achieved. We used two different variants of Pix2pix generator, one based on UNet⁴¹ and the other based on DenseUNet³⁹.

GAN-based methods experiment setup

While we made it possible to use the whole tissue sample as the so called reference patch for the traditional method, a WSI, as is, can not be fed directly to the GPU for GAN-based model training, therefore, the tissue images were first split into 512px x 512px (433.36 μ m x 433.36 μ m) tiles before the training.

CycleGAN training required a representative dataset, since each laboratory provided one tissue per organ, the idea was to treat the whole dataset as one and choose a small representative dataset. We chose a semi-heuristic approach of clustering the tissue images. First, the tissue images were transformed to the $l\alpha\beta$ color space⁵¹ for a more independent and distinguishable representation in each channel. Normalized histograms of each channel were used as features. Then dimensionality reduction was applied using principal component analysis, the first two components were used to plot the samples in two dimensions. The components were then used to cluster the samples using k-means clustering⁶³. We used the Within-Cluster Sum of Square approach to find the optimal number of clusters. In order to keep the experiments uniform across different organ tissues, we chose 8 clusters and samples closest to the cluster mean were chosen as the representative samples of that cluster. Altogether, 8 samples were chosen to represent the dataset that were used in both CycleGAN trainings.

For Pix2pix trainings, each image tile in the reference WSI was first converted to grayscale and then the model was trained on the grayscale and colored tile pairs of the same tissue patch. The objective was to thoroughly train the model to precisely map grayscale tiles to their corresponding color representations. Once the model was trained, the rest of the WSI tiles were fed to the model as grayscale input, to be converted to a similar color as the reference WSI.

Computational cost

While computational cost may be an important factor in real-life usage, here we did not include it as a criterion for method comparison, neither did we optimize any of the implementations for computational efficiency. However, notes on the computational cost are provided, which may be helpful for potential practical deployment of the methods.

Deep learning methods incur a substantial computational cost during the training phase. CycleGAN variants training could take up to several days on a single GPU like 32GB NVIDIA V100 primarily because 4-network architecture and bigger training dataset compared to Pix2pix. Pix2pix (UNet), requiring only several hours, was the fastest to train because of the simplicity of architecture and single pair WSI in training data, however, the DenseUNet variant took approximately 2.5X compared to the UNet variant. In contrast, none of the four traditional methods (histogram matching, Macenko, Reinhard, Vahadane) requires the so-called the training phase.

For inference, all deep learning methods except Pix2pix (DenseUNet) took approximately 4–5 minutes per WSI, attributable to the similarity of their generators' architecture (generator being the only network used during inference). Pix2pix (DenseUNet) approximately took 2–2.5X per WSI compared to the UNet variant. Traditional methods like histogram normalization and Reinhard took approximately 30 seconds to 2 minutes per WSI, whereas Macenko and Vahadane took approximately 2–7 minutes per WSI. Traditional methods were executed on Intel Xeon 2.1 GHz processors.

Data availability

The tissue WSI dataset used in this study is freely available on a FAIR-compliant server under <https://doi.org/10.23729/4cb0b2ed-b074-4e5e-8cd6-ff5242c77fd0>. The implementation of all the stain normalization methods, used in this study, has been deposited at Zenodo under <https://doi.org/10.5281/zenodo.12344369> and are publicly available as of the date of publication. Refer to `readme.md` for detailed instructions on how to use the code and `requirements.txt` to install the dependencies.

Received: 5 June 2025; Accepted: 17 February 2026

Published online: 26 February 2026

References

1. Troiano, N. W., Ciovacco, W. A. & Kacena, M. A. The effects of fixation and dehydration on the histological quality of undecalcified murine bone specimens embedded in methylmethacrylate. *J. Histotechnol.* **32**, 27–31 (2009).
2. Chlipala, E. A. et al. Impact of preanalytical factors during histology processing on section suitability for digital image analysis. *Toxicol. Pathol.* **49**, 755–772 (2021).
3. Chan, J. K. The wonderful colors of the hematoxylin-eosin stain in diagnostic surgical pathology. *Int. J. Surg. Pathol.* **22**, 12–32 (2014).
4. Bancroft, J. D. & Gamble, M. *Theory and practice of histological techniques* (Elsevier health sciences, 2008).
5. Weinstein, R. S. et al. Overview of telepathology, virtual microscopy, and whole slide imaging: Prospects for the future. *Hum. Pathol.* **40**, 1057–1069 (2009).
6. Ji, X. et al. Physical color calibration of digital pathology scanners for robust artificial intelligence assisted cancer diagnosis. arXiv preprint [arXiv:2307.05519](https://arxiv.org/abs/2307.05519) (2023).
7. Valkonen, M., Högnäs, G., Bova, G. S. & Ruusuvoori, P. Generalized fixation invariant nuclei detection through domain adaptation based deep learning. *IEEE J. Biomed. Health Inform.* **25**, 1747–1757 (2020).
8. Rakha, E. A. et al. Current and future applications of artificial intelligence in pathology: A clinical perspective. *J. Clin. Pathol.* **74**, 409–414 (2021).
9. Moscalu, M. et al. Histopathological images analysis and predictive modeling implemented in digital pathology—current affairs and perspectives. *Diagnostics* **13**, 2379 (2023).
10. Wilson, M. L. et al. Access to pathology and laboratory medicine services: A crucial gap. *Lancet* **391**, 1927–1938 (2018).
11. Ström, P. et al. Artificial intelligence for diagnosis and grading of prostate cancer in biopsies: a population-based, diagnostic study. *The Lancet Oncology* **21**, 222–232 (2020).
12. Balkenhol, M. C. et al. Deep learning assisted mitotic counting for breast cancer. *Lab. Invest.* **99**, 1596–1606 (2019).
13. Zadeh Shirazi, A. et al. A deep convolutional neural network for segmentation of whole-slide pathology images identifies novel tumour cell-perivascular niche interactions that are associated with poor survival in glioblastoma. *Br. J. Cancer* **125**, 337–350 (2021).
14. Vu, Q. D. et al. Methods for segmentation and classification of digital microscopy tissue images. *Front. Bioeng. Biotechnol.* <https://doi.org/10.3389/fbioe.2019.00053> (2019).
15. Wulczyn, E. et al. Deep learning-based survival prediction for multiple cancer types using histopathology images. *PLoS One* **15**, e0233678 (2020).
16. Saillard, C. et al. Predicting survival after hepatocellular carcinoma resection using deep learning on histological slides. *Hepatology* **72**, 2000–2013 (2020).
17. Fu, Y. et al. Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis. *Nat. Cancer* **1**, 800–810 (2020).
18. Ahmed, A. A., Abouzid, M. & Kaczmarek, E. Deep learning approaches in histopathology. *Cancers* **14**, 5264 (2022).
19. Paige AI. Paige ai - transforming cancer diagnostics with ai (2024). Accessed: 2024–12–21.
20. Dunn, C. et al. Quantitative assessment of h&e staining for pathology: Development and clinical evaluation of a novel system. *Diagn. Pathol.* **19**, 42 (2024).
21. Breen, J., Zucker, K., Allen, K., Ravikumar, N. & Orsi, N. M. Generative adversarial networks for stain normalisation in histopathology. In *Applications of Generative AI* 227–247 (Springer, 2024).
22. Therrien, R. & Doyle, S. Role of training data variability on classifier performance and generalizability. In *Medical Imaging 2018: Digital Pathology*, **10581**, 58–70 (SPIE, 2018).
23. Jahanifar, M. et al. Domain generalization in computational pathology: Survey and guidelines. *ACM Comput. Surv.* **57**, 1–37 (2025).
24. Tellez, D. et al. Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. *Med. Image Anal.* **58**, 101544 (2019).
25. Hoque, M. Z., Keskinarkaus, A., Nyberg, P. & Seppänen, T. Stain normalization methods for histopathology image analysis: A comprehensive review and experimental comparison. *Inf. Fusion* **102**, 101997 (2024).
26. Salvi, M. et al. Impact of stain normalization on pathologist assessment of prostate cancer: A comparative study. *Cancers* **15**, 1503 (2023).
27. Goodfellow, I. et al. Generative adversarial nets. *Advances in neural information processing systems* **27** (2014).
28. Reinhard, E., Adhikhmin, M., Gooch, B. & Shirley, P. Color transfer between images. *IEEE Comput. Graph. Appl.* **21**, 34–41 (2001).
29. Macenko, M. et al. A method for normalizing histology slides for quantitative analysis. In *2009 IEEE international symposium on biomedical imaging: from nano to macro*, 1107–1110 (IEEE, 2009).
30. Vahadane, A. et al. Structure-preserving color normalization and sparse stain separation for histological images. *IEEE Trans. Med. Imaging* **35**, 1962–1971 (2016).
31. Gonzalez, R. C. & Woods, R. E. *Digital image processing*, prentice hall (Upper Saddle River, 2008).
32. Magee, D. et al. Colour normalisation in digital histopathology images. In *Proc Optical Tissue Image analysis in Microscopy, Histopathology and Endoscopy (MICCAI Workshop)*, **100**, 100–111 (Daniel Elson London, 2009).
33. Niethammer, M., Borland, D., Marron, J., Woosley, J. & Thomas, N. E. Appearance normalization of histology slides. In *Machine Learning in Medical Imaging: First International Workshop, MLMI 2010, Held in Conjunction with MICCAI 2010, Beijing, China, September 20, 2010. Proceedings* 1, 58–66 (Springer, 2010).

34. Fuchs, T. J. & Buhmann, J. M. Computational pathology: Challenges and promises for tissue analysis. *Comput. Med. Imaging Graph.* **35**, 515–530 (2011).
35. Zhu, J.-Y., Park, T., Isola, P. & Efros, A. A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, 2223–2232 (2017).
36. Levy, J. J., Jackson, C. R., Sriharan, A., Christensen, B. C. & Vaickus, L. J. Preliminary evaluation of the utility of deep generative histopathology image translation at a mid-sized nci cancer center. *bioRxiv* 2020–01 (2020).
37. Gadermayr, M., Appel, V., Klinkhammer, B. M., Boor, P. & Merhof, D. Which way round? a study on the performance of stain-translation for segmenting arbitrarily dyed histological images. In *Medical Image Computing and Computer Assisted Intervention-MICCAI 2018: 21st International Conference, Granada, Spain, September 16–20, 2018, Proceedings, Part II* **11**, 165–173 (Springer, 2018).
38. Koivukoski, S., Khan, U., Ruusuvoori, P. & Latonen, L. Unstained tissue imaging and virtual hematoxylin and eosin staining of histologic whole slide images. *Lab. Invest.* **103**, 100070 (2023).
39. Khan, U., Koivukoski, S., Valkonen, M., Latonen, L. & Ruusuvoori, P. The effect of neural network architecture on virtual H&E staining: Systematic assessment of histological feasibility. *Patterns* <https://doi.org/10.1016/j.patter.2023.100725> (2023).
40. Salido, J., Vallez, N., González-López, L., Deniz, O. & Bueno, G. Comparison of deep learning models for digital h&e staining from unpaired label-free multispectral microscopy images. *Comput. Methods Programs Biomed.* **235**, 107528 (2023).
41. Ronneberger, O., Fischer, P. & Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention-MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III* **18**, 234–241 (Springer, 2015).
42. de Bel, T., Hermesen, M., Kers, J., van der Laak, J. & Litjens, G. Stain-transforming cycle-consistent generative adversarial networks for improved segmentation of renal histopathology. In *International Conference on Medical Imaging with Deep Learning*, 151–163 (PMLR, 2019).
43. de Bel, T., Bokhorst, J.-M., van der Laak, J. & Litjens, G. Residual cyclegan for robust domain transformation of histopathological tissue slides. *Med. Image Anal.* **70**, 102004 (2021).
44. Isola, P., Zhu, J.-Y., Zhou, T. & Efros, A. A. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1125–1134 (2017).
45. Salehi, P. & Chalechale, A. Pix2pix-based stain-to-stain translation: A solution for robust stain normalization in histopathology images analysis. In *2020 International conference on machine vision and image processing (MVIP)*, 1–7 (IEEE, 2020).
46. Shaban, M. T., Baur, C., Navab, N. & Albarqouni, S. Staingan: Stain style transfer for digital histological images. In *2019 IEEE 16th international symposium on biomedical imaging (Isbi 2019)*, 953–956 (IEEE, 2019).
47. Cho, H., Lim, S., Choi, G. & Min, H. Neural stain-style transfer learning using gan for histopathological images. *arXiv* 2017. *arXiv preprint arXiv:1710.08543* (2017).
48. BeniTaieb, A. & Hamarneh, G. Adversarial stain transfer for histopathology image analysis. *IEEE Trans. Med. Imaging* **37**, 792–802 (2017).
49. Cong, C. et al. Semi-supervised adversarial learning for stain normalisation in histopathology images. In *Medical Image Computing and Computer Assisted Intervention-MICCAI 2021: 24th International Conference, Strasbourg, France, September 27-October 1, 2021, Proceedings, Part VIII* **24**, 581–591 (Springer, 2021).
50. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778 (2016).
51. Consortium, I. C. et al. Image technology colour management-architecture, profile format, and data structure. *Specification ICC. 1: 2004–10 (Profile version 4.2. 0.0)* (2004).
52. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. & Wojna, Z. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2818–2826 (2016).
53. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B. & Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems* **30**, (2017).
54. Zhang, L., Zhang, L., Mou, X. & Zhang, D. Fsim: A feature similarity index for image quality assessment. *IEEE Trans. Image Process.* **20**, 2378–2386 (2011).
55. Pachitariu, M., Rariden, M. & Stringer, C. Cellpose-sam: superhuman generalization for cellular segmentation. *bioRxiv* 2025–04 (2025).
56. Ding, T. et al. A multimodal whole-slide foundation model for pathology. *Nature medicine* 1–13 (2025).
57. Dunn, C., Brettle, D., Hodgson, C., Hughes, R. & Treanor, D. An international study of stain variability in histopathology using qualitative and quantitative analysis. *J. Pathol. Inform.* **17**, 100423 (2025).
58. Prezja, F., Pölönen, I., Äyrämö, S., Ruusuvoori, P. & Kuopio, T. H&e multi-laboratory staining variance exploration with machine learning. *Appl. Sci.* **12**, 7511 (2022).
59. Lin, S. et al. Impact of stain variation and color normalization for prognostic predictions in pathology. *Sci. Rep.* **15**, 2369 (2025).
60. Ji, X. et al. Physical color calibration of digital pathology scanners for robust artificial intelligence-assisted cancer diagnosis. *Mod. Pathol.* **38**, 100715 (2025).
61. Zewdie, E. T., Tessema, A. W. & Simegn, G. L. Classification of breast cancer types, sub-types and grade from histopathological images using deep learning technique. *Health Technol.* **11**, 1277–1290 (2021).
62. Habtemariam, L. W., Zewde, E. T. & Simegn, G. L. Cervix type and cervical cancer classification system using deep learning techniques. *Med. Devices Evid. Res.* <https://doi.org/10.2147/MDER.S366303> (2022).
63. MacQueen, J. et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, 14, 281–297 (Oakland, CA, USA, 1967).

Acknowledgements

The authors would like to thank Zuhair Iftikhar and Robin Ekman for their skillful technical assistance in this study. Computational experiments were conducted on high-performance computing environment provided by CSC - IT Centre for Science (Finland), the authors are grateful for the resources and support by CSC.

Author contributions

Conceptualization, P.R., L.L., T.K.; Methodology, U.K., J.H., M.F, H.H., L.L., T.K., and P.R.; Software, U.K. and H.H.; Validation, U.K., J.H., and M.F; Formal analysis, U.K.; Investigation, U.K., J.H., and M.F; Resources, T.K. and P.R.; Data Curation, U.K., J.H., and P.R.; Writing - Original Draft, U.K. and J.H.; Writing - Review and Editing, U.K., J.H., L.L., and P.R.; Visualization, U.K., J.H., and H.H.; Supervision, T.K. and P.R.; Project Administration, T.K. and P.R.; Funding Acquisition, T.K. and P.R.

Funding

Financial support from Research Council of Finland (PR, LL), Cancer Foundation Finland (PR, LL), Sigrid

Juselius Foundation (PR), University of Turku Graduate School (UK) is gratefully acknowledged.

Declarations

Competing interests

PR is a co-founder and shareholder of Louhi Health Data.

Additional information

Correspondence and requests for materials should be addressed to P.R.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2026