April 1, 2019

JOURNAL OF CLINICAL ONCOLOGY

HistoQC: An Open-Source Quality Control Tool for Digital Pathology Slides

ASCO_JCO

Disclaimer

The ideas and opinions expressed in the *Journal of Clinical Oncology* (JCO) do not necessarily reflect those of the American Society of Clinical Oncology (ASCO). The mention of any product, service, or therapy in any JCO article should not be construed as an endorsement of the products mentioned.

It is the responsibility of the treating physician or other health care provider, relying on independent experience and knowledge of the patient, to determine drug dosages and the best treatment for the patient. Readers are advised to check the appropriate medical literature and the product information currently provided by the manufacturer of each drug to be administered to verify the dosage, method, and duration of administration, or contraindications. Readers are also encouraged to contact the manufacturer with questions about the features or limitations of any products. ASCO assumes no responsibility for any injury or damage to persons or property arising out of or related to any use of the material contained in JCO or to any errors or omissions.

ASCO®

American Society of Clinical Oncology

Making a world of difference in cancer care

original report

HistoQC: An Open-Source Quality Control Tool for Digital Pathology Slides

Andrew Janowczyk, PhD[1]; Ren Zuo, MS[1]; Hannah Gilmore, MD[2]; Michael Feldman, MD, PhD[3]; and Anant Madabhushi, PhD[1,4]

abstract

PURPOSE Digital pathology (DP), referring to the digitization of tissue slides, is beginning to change the landscape of clinical diagnostic workflows and has engendered active research within the area of computational pathology. One of the challenges in DP is the presence of artefacts and batch effects, unintentionally introduced during both routine slide preparation (eg, staining, tissue folding) and digitization (eg, blurriness, variations in contrast and hue). Manual review of glass and digital slides is laborious, qualitative, and subject to intra- and inter-reader variability. Therefore, there is a critical need for a reproducible automated approach of precisely localizing artefacts to identify slides that need to be reproduced or regions that should be avoided during computational analysis.

METHODS Here we present HistoQC, a tool for rapidly performing quality control to not only identify and delineate artefacts but also discover cohort-level outliers (eg, slides stained darker or lighter than others in the cohort). This open-source tool employs a combination of image metrics (eg, color histograms, brightness, contrast), features (eg, edge detectors), and supervised classifiers (eg, pen detection) to identify artefact-free regions on digitized slides. These regions and metrics are presented to the user via an interactive graphical user interface, facilitating artefact detection through real-time visualization and filtering. These same metrics afford users the opportunity to explicitly define acceptable tolerances for their workflows.

RESULTS The output of HistoQC on 450 slides from The Cancer Genome Atlas was reviewed by two pathologists and found to be suitable for computational analysis more than 95% of the time.

CONCLUSION These results suggest that HistoQC could provide an automated, quantifiable, quality control process for identifying artefacts and measuring slide quality, in turn helping to improve both the repeatability and robustness of DP workflows.

INTRODUCTION

Definitive disease diagnosis routinely takes place via visual inspection of a tissue slide by a pathologist under a microscope. Before this can take place, the tissue slide itself must be created. This process, which involves gross organ dissection, selection and preparation of tissue blocks for slide creation, microtomy (cutting and tissue placement on the slide), staining, and cover slipping, is fraught with multiple preanalytic opportunities for the introduction of artefacts and batch effects.[1-3] These artefacts may include improper tissue placement (eg, folding, compressing, tearing, air bubbles), improper reagents (eg, over- or understaining, stain concentration differences, stain batch variation), and poor microtomy (eg, knife chatter, thickness variances). The increasingly popular digitization of these same slides, to take advantage of computational-aided diagnostic approaches,[4-6] for example, introduces yet another potential source of artefacts. This digitization process sees the same glass slides routinely used for microscope-based pathology being placed on the equivalent of a digital camera so that digital representations of the slide may be constructed. Scanner manufacturers may employ different approaches for slide digitization, including different hardware (eg, bulbs for lighting, charge-coupled device chips for digitization), algorithms for image manipulation (eg, stitching, compression), and file formats. Therefore, the choice of slide scanner could influence the image appearance, which in turn might have implications for any subsequent image analysis procedure.[7] These digital pathology (DP) slides may additionally include digitization artefacts such as blurriness, lighting, and contrast issues. Taken together, there are a number of different combinations of sources of preanalytic variance that may result in substantial differences in appearance and quality of a tissue slide.

These same artefacts and variances could negatively affect downstream clinical and research workflows.[8] In

ASSOCIATED CONTENT
Data Supplement
Author affiliations and support information (if applicable) appear at the end of this article.
Accepted on March 1, 2019 and published at ascopubs.org/journal/cci on April 16, 2019: DOI https://doi.org/10.1200/CCI.18.00157

the analog workflow in use today, continuous quality control (QC) processes are limited, unlike in laboratory medicine, which relies on continuous statistical process control. Clinically, slides rejected on quality grounds represent a drag on the clinical pathology workflow, because these slides need to be recut or rescanned, in turn causing additional delays and unnecessary costs. From a research standpoint, artefacts represent sources of noise that can adversely affect the development and validation of analytic classifiers for tasks such as disease detection, diagnosis, and prognosis.[9,10] This is especially important for increasingly popular deep learning– and machine learning–based approaches,[11-16] which rely on well-annotated and relatively artefact-free images to learn underlying disease-specific representations.

Currently, most QC processes for clinical and research applications are performed manually, making the process subjective, laborious, and error prone. For instance, a wide spectrum of artefacts and image qualities (Fig 1) can be seen in many of the 30,000 digitized tissue slides hosted by The Cancer Genome Atlas (TCGA).[17] This occurs despite the fact that these slide images undergo manual QC before being introduced into TCGA. In addition, more subtle artefacts such as variations in stain that may not affect a pathologist's diagnostic interpretation may still have implications for subsequent computational image analysis and machine-learning algorithms. For instance, technical artefacts resulting from the preparation facility (ie, batch effects) may be confounded with the biologic signal under investigation (Data Supplement). Although batch effects remain a well-known issue in the bioinformatics field,[18] they have received less attention in the DP domain.

Recently, other groups[19-23] have begun to develop DP algorithms for QC tasks, such as blurriness and stain assessment. Unfortunately, there has not been a single, unified user-friendly platform that has included these and other QC approaches for a comprehensive and integrated QC review of DP slide images.

Recognizing the need for a modular, user-friendly QC tool, we present here an open-source QC application, HistoQC, for automated assessment of slide quality alongside a public repository of slides containing artefacts. HistoQC employs a combination of image metrics (eg, color histograms, brightness, contrast), features (eg, edge and smoothness detectors), and supervised classifiers (eg, pen detection) to aid users in identifying slides with gross technical artefacts, artefact-affected regions that may not be suitable for computational analysis (Data Supplemental provides current list of classifiers and metrics), and samples potentially affected by batch effects. The modular nature of HistoQC allows for the facile embedding of additional metrics and artefact-detection algorithms as they become available in the literature.

METHODS

HistoQC functions in the following way. The user supplies a configuration file that defines the parameters of the QC pipeline, such as which modules to execute and in what

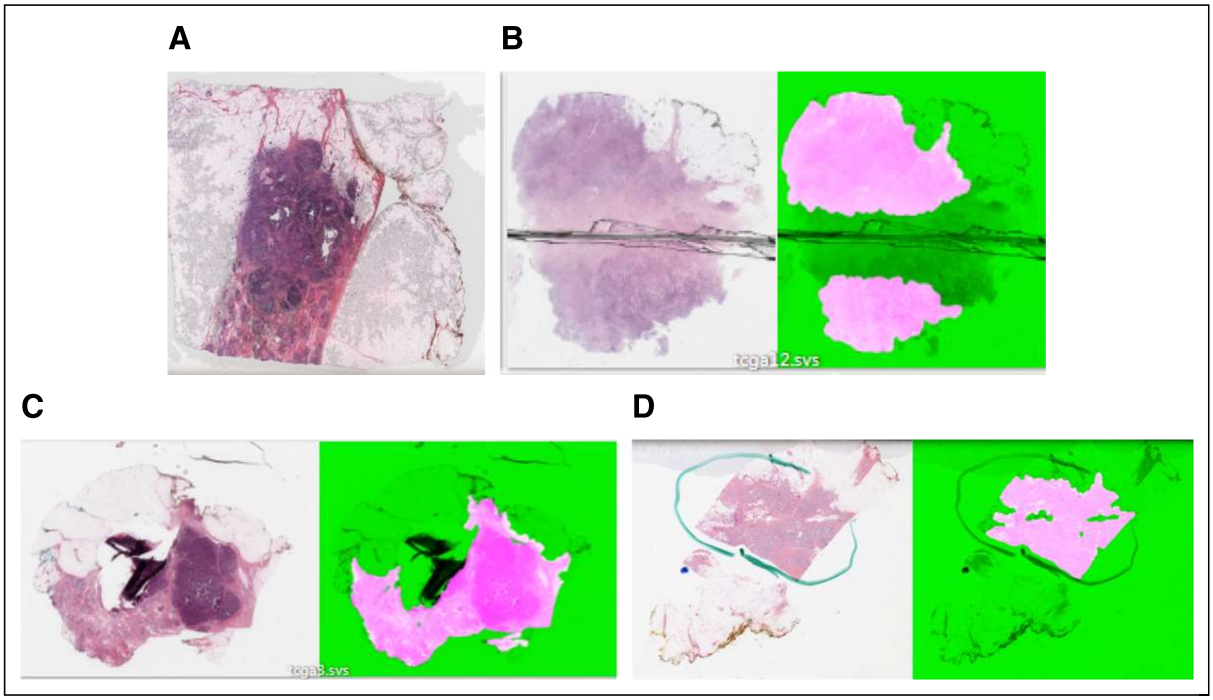


FIG 1. Original images juxtaposed with corresponding results from HistoQC (fuchsia indicating acceptable tissue): images identified as having (A) a significant air bubble artefact requiring removal from experimental cohort, (B) blurry tissue near a coverslip crack, (C) folded tissue, and (D) pen markings correctly identified as regions to be avoided.

order. As the python-based pipeline is executed on the slide, relevant output images are created (eg, thumbnail images indicating regions of potential blurriness), with metadata (eg, scanner type, magnification, microns per pixel) and metrics being saved in a tab-separated value file. As designed, image metrics can be computed on the entire slide (ie, including the background) or limited solely to regions containing detected tissue. Although any common data analytic tool may be used to review the tab-separated value output (eg, Matlab [MathWorks, Natick, MA], Excel [Microsoft, Redmond, WA], R [R Foundation, Vienna, Austria]), we have developed an HTML5-based user interface (Fig 2) that seamlessly allows for real-time visualization and filtering of the data. This approach helps identify those specific slides that might require additional scrutiny.

Slides requiring additional scrutiny can be discovered in multiple ways: sorting various columns to view outliers with unexpectedly high or low values (Fig 2 green arrow), viewing interactive parallel coordinate plots[24] (Fig 2 red box; Fig 3) of metrics that help visualize potential batch effects and outliers, or viewing the original slides juxtaposed with various output masks (Fig 2 blue arrow). For improved user experience and efficiency, clicking a row or image shows the various masks produced by the pipeline (Fig 2 middle), with a subsequent click taking the user to a higher-magnification version of the mask of interest for more detailed review (Fig 2 bottom). After the user has either annotated rows using the comments field or removed rows from the table, the resulting table can be saved and used as a list of samples suitable for downstream experiments. We note that postinstallation, no Internet connection is required, making our approach suitable for nonanonymized clinical data.

We evaluated the ability of HistoQC to identify regions of artefact-free tissue on a total of 450 randomly selected slides from the TCGA breast cancer cohort at a magnification of 40 . Because of the efficient implementation of HistoQC, the analysis took 130 minutes using a four-hyperthreaded core processor. Representative slides identified by HistoQC as containing artefacts such as air bubbles and slide cracks are illustrated in Figures 1A and 1B. HistoQC successfully identified regions with tissue folding and pen markings, removing them from the outputted masks (Figs 1C and 1D). As derived from HistoQC output, the Data Supplement shows the potential presence of batch effects in microns per pixel and considerable heterogeneity in tissue brightness for the TCGA breast cancer data set, important considerations for downstream experimental design.

RESULTS

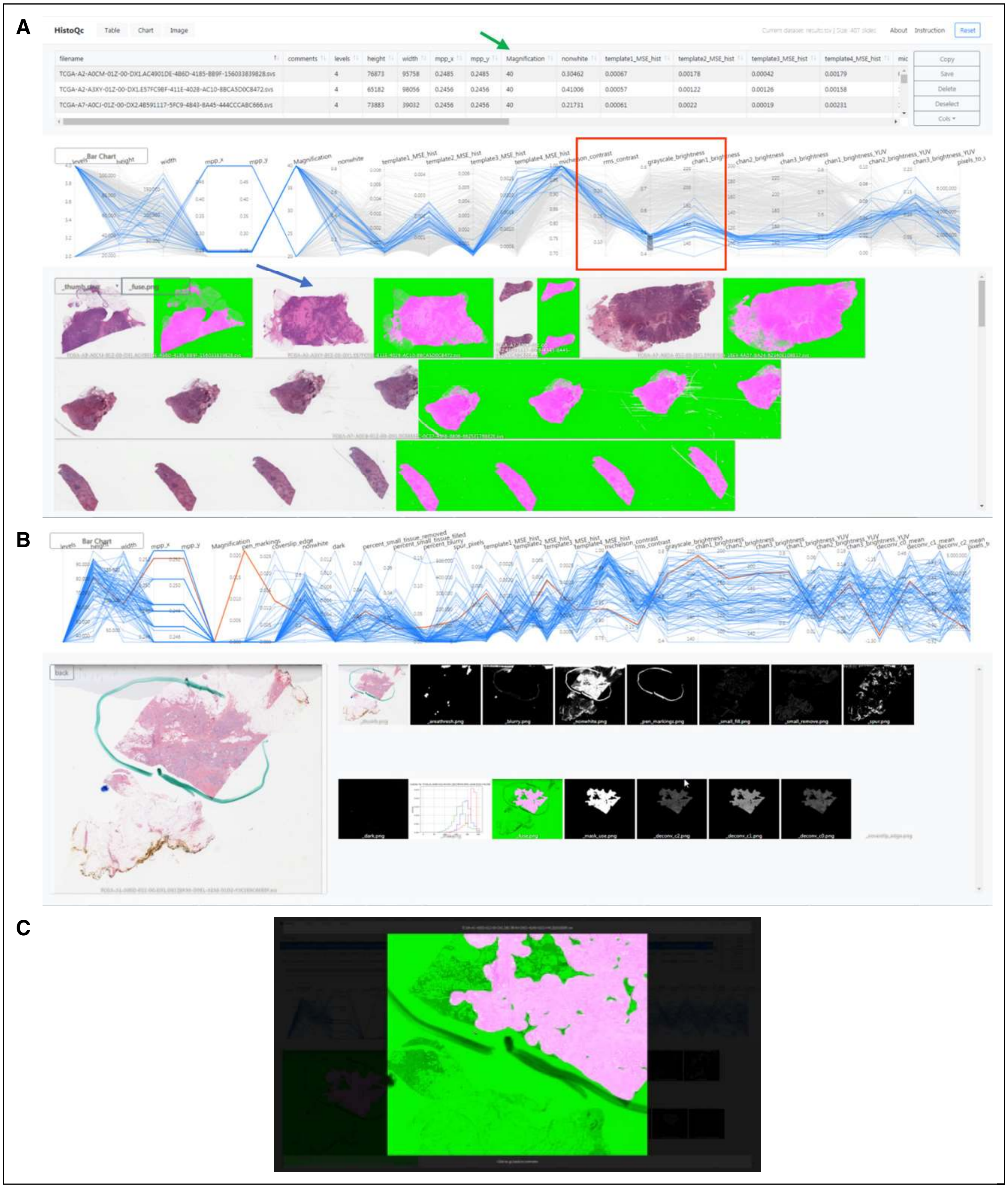
To validate the results generated by HistoQC, two pathologists with experience in DP were asked to assign a value of either acceptable or not acceptable to each of the masks produced by HistoQC. Acceptability was defined by at least an 85% area overlap between the pathologists' visual assessment and the computational assessment by HistoQC of artefact-free tissue. Each pathologist independently reviewed 250 samples. In addition, a total of 50 images from TCGA were evaluated by both pathologists and HistoQC to determine interexpert agreement on HistoQC output. Overall, the agreement between HistoQC and the experts was 94% (235 of 250) for expert 1 and 97% (242 of 250) for expert 2. For the 50 slides evaluated by both experts, interobserver agreement was 96% (48 of 50), comparable to that of HistoQC with the individual readers. The main reasons for the disagreement were faintly stained slides resulting in tissue detection failures and a few regions of predominantly stromal-rich areas being incorrectly identified as blurry (Data Supplement). These failures appeared on the HistoQC user interface as outliers, primarily because of metrics (eg, estimated tissue area) being a number of standard deviations away from those associated with the remainder of the analyzed slides. Slides identified by HistoQC as containing artefacts were uploaded to the Histology Quality Control Repository[25] for community review.

The pathologists also provided qualitative feedback regarding patterns of cases that HistoQC seemed to incorrectly identify as being compromised or not. These cases generally fell into three categories: poorly fixed tissue, necrotic tissue, or subtle adipose tissue infiltrate with scant tissue reaction. HistoQC also sometimes struggled to fully identify parenchyma in mucinous tumors. We are working on further improving HistoQC to address these limitations in the next version.

DISCUSSION

To summarize, we presented and have released an open-source QC tool for DP slides called HistoQC. Initial results suggest HistoQC is suitable for delineation of slide level artefacts. Comparison of HistoQC against manual QC by two pathologists on 450 images yielded an average agreement greater than 95%, comparable in range of agreement to that between the two individual human readers. In addition, the image metrics computed by HistoQC could be used by researchers and analytic pipeline developers to precisely define the input image characteristics with which their algorithms have been both trained and validated. Stringent specification of these image characteristic ranges allows for algorithms to be selectively invoked only on the appropriate images, likely improving algorithm confidence.

Taken together, the clinical pathology and DP vendor communities seem to be beginning to appreciate the importance of quantifiable QC processes for engendering DP workflows.[26] Before DP systems can be used within a clinical setting, slide scanners themselves must receive regulatory approval. Scanner manufacturers have been attempting to quantitatively assess the reproducibility of



filename	comments	levels	height	width	mpp_x	mpp_y	Magnification	nonwhite	template1_MSE_hist	template2_MSE_hist	template3_MSE_hist	template4_MSE_hist	mic
TCGA-A2-A0CM-01Z-00-DX1.AC4901DE-4B6D-4185-BB9F-156033839828.svs		4	76873	95758	0.2485	0.2485	40	0.30462	0.00067	0.00178	0.00042	0.00179	
TCGA-A2-A3XY-01Z-00-DX1.E57FC9BF-411E-4028-AC10-8BCA5D0C8472.svs		4	65182	98056	0.2456	0.2456	40	0.41006	0.00057	0.00122	0.00126	0.00158	
TCGA-A7-A0CJ-01Z-00-DX2.4B591117-5FC9-4843-8A45-444CCCABC666.svs		4	73883	39032	0.2456	0.2456	40	0.21731	0.00061	0.0022	0.00019	0.00231	

FIG 2. (A) HistoQC user interface showing table of HistoQC-produced metrics with sortable columns (green arrow), parallel coordinate plot (red box; additional details in Fig 3), and thumbnail images of the cohort alongside HistoQC overlay output indicating artefact-free regions (blue arrow). (B) Selecting a single image highlights the appropriate line in the parallel coordinate graph and shows the series of outputs produced by the modules of the pipeline, allowing for more detailed subsequent review. (C) Double clicking on any image brings up a higher-resolution version with dynamic zoom, allowing for fine-tuned inspection of potential artefacts. Hist, histogram; MPP, microns per pixel; MSE, mean squared error.

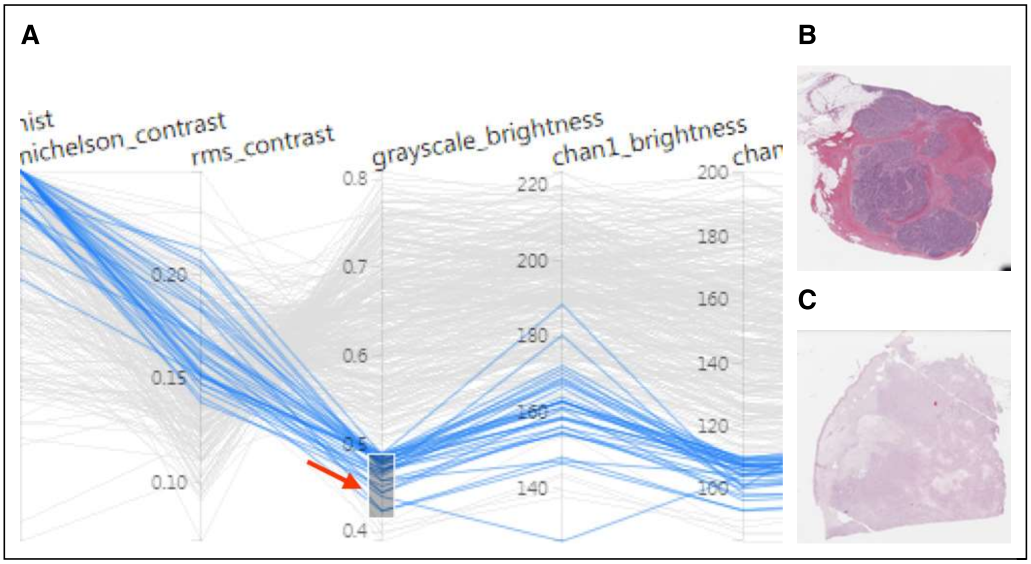


FIG 3. (A) Higher magnification of the parallel coordinate plot bounded by the red box in Figure 2. The different -axes correspond to different image metrics determined by HistoQC and may have their own ranges and scales. Each horizontally oriented line, in either gray or blue, represents a whole-slide image (WSI) analyzed by HistoQC. By examining the convergence or divergence of each horizontal WSI line with respect to the rest of the images in the cohort, batch effects and outliers can be more easily visually identified. In the example illustrated, the user has interactively drawn a gray box (red arrow) to select images with a grayscale intensity value between 0.4 and 0.5, resulting in all lines that do not meet this criterion turning gray. This image selection mechanism alters the visibility of slides in both the table (Fig 2 green arrow) and the thumbnails (Fig 2 blue arrow). Furthermore, the user can dynamically drag or extend the gray box upward and downward to update the visible slides in real time. As a result, images with (B) low and (C) high grayscale brightness values are easier to identify and review.

scanner-generated images, especially to ensure consistency of image quality over time.[27,28] With penetration of these scanners into clinical workflows, these types of validation studies may become enshrined as part of the routine quality assurance and maintenance of the scanners.[26] As such, from both regulatory and maintenance standpoints, a single automated QC pipeline like HistoQC can provide quantitative metrics for benchmarking the quality and consistency of scanner-generated DP images.

DP workflows are on the verge of leveraging powerful computer-aided diagnostic (CAD) support algorithms, potentially helping to greatly reduce inter- and intraobserver diagnostic variability. As revealed in a number of recent publications, many CAD and artificial intelligence (AI) algorithms appear to not generalize very well when evaluated on a cohort distinct from the set of images on which they were initially trained.[14,29] Consequently, these AI and CAD algorithms must be robustly validated on a large collection of heterogeneous inputs.[9,10] Approaches like HistoQC could allow for pre-evaluation of test sets to ensure that the CAD and AI algorithms are evaluated on a sufficiently diverse set of test images.

Although HistoQC is ready for research applications, there remain areas for additional improvement beyond addressing the comments of our pathologists. For example, due to the heterogeneity in compression levels typically present between DP scanners and the evidence that the resulting compression artefacts affect performance of deep-learning and AI algorithms,[30] HistoQC could be extended to detect and measure compression effects. Building further on the need to incorporate additional features, we envision HistoQC evolving into a collection of community-driven reference implementations of sophisticated detectors and metrics. For example, Senaras et al[22] presented a deep learning–based blur detector, and Avanaki et al[19] proposed texture-based image quality metrics. We hope that these types of algorithms will in the future be embedded into HistoQC to enable the comparison of results across different sites and laboratories. The work presented here focused on the evaluation of HistoQC in the context of hematoxylin and eosin bright-field microscopy images. Clearly, there is also a need for the application of QC metrics in other types of multimodal microscopy images, such as immunohistochemical staining and quantitative immunofluorescence.

Last, we hope to aggregate unique artefacts identified by the user community during its use of HistoQC. We have stood up an image quality repository to allow end users to upload slides that contain artefacts.[25] This repository will help provide training and validation material needed for the benchmarking of future CAD approaches. The source code of HistoQC (Data Supplement) is freely available for use, modification, and contribution (http://github.com/choosehappy/HistoQC).

AFFILIATIONS

[1]Case Western Reserve University, Cleveland, OH
[2]University Hospitals Cleveland Medical Center, Cleveland, OH
[3]University of Pennsylvania Perelman School of Medicine, Philadelphia, PA
[4]Louis Stokes Cleveland Veterans Administration Medical Center, Cleveland, OH

The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health, US Department of Veterans Affairs, US Department of Defense, or US Government.

CORRESPONDING AUTHOR

Andrew Janowczyk, PhD, Case Western Reserve University, 2071 Martin Luther King Drive, Wickenden 523, Cleveland, OH 44106-7207; e-mail: andrew.janowczyk@case.edu.

SUPPORT

Supported by National Cancer Institute, National Institutes of Health, Awards No. 1U24CA199374-01, R01CA202752-01A1, R01CA208236-01A1, R01 CA216579-01A1, and R01 CA220581-01A1; National Center for Research Resources Award No. 1 C06 RR12463-01; Veterans Affairs (VA) Merit Review Award No. IBX004121A from the US Department of VA Biomedical Laboratory Research and Development Service; Nephrology Training Grant No. 5T32DK007470 CWRU; US Department of Defense Lung Cancer Investigator-Initiated Translational Research Award No. W81XWH-18-1-0440; the Ohio Third Frontier Technology Validation Fund, the Wallace H. Coulter Foundation Program in the Department of Biomedical Engineering; and the Clinical and Translational Science Award Program, Case Western Reserve University.

AUTHOR CONTRIBUTIONS

Conception and design: All authors
Financial support: Anant Madabhushi
Administrative support: Anant Madabhushi
Provision of study material or patients: Michael Feldman, Hanna Gilmore, Anant Madabhushi
Collection and assembly of data: Andrew Janowczyk, Michael Feldman, Hanna Gilmore, Anant Madabhushi
Data analysis and interpretation: All authors
Manuscript writing: All authors
Final approval of manuscript: All authors
Accountable for all aspects of the work: All authors

AUTHORS' DISCLOSURES OF POTENTIAL CONFLICTS OF INTEREST

The following represents disclosure information provided by authors of this manuscript. All relationships are considered compensated. Relationships are self-held unless noted. I = Immediate Family Member, Inst = My Institution. Relationships may not relate to the subject matter of this manuscript. For more information about ASCO's conflict of interest policy, please refer to www.asco.org/rwc or ascopubs.org/jco/site/ifc.

Andrew R. Janowczyk
Consulting or Advisory Role: Merck

Hannah Gilmore
Travel, Accommodations, Expenses: Sectra

Michael Feldman
Consulting or Advisory Role: Philips Healthcare
Travel, Accommodations, Expenses: Philips Healthcare

Anant Madabhushi
Leadership: Inspirata
Stock and Other Ownership Interests: Inspirata, Elucid Bioimaging
Honoraria: AstraZeneca, Inspirata
Consulting or Advisory Role: Inspirata, AstraZeneca, Merck
Research Funding: Inspirata (Inst), Philips Healthcare (Inst)
Patents, Royalties, Other Intellectual Property: Intellectual property licensed by Inspirata (Inst); intellectual property licensed by Elucid Bioimaging (Inst)

No other potential conflicts of interest were reported.

REFERENCES

1. Chatterjee S: Artefacts in histopathology. J Oral Maxillofac Pathol 18:S111-S116, 2014 (suppl 1)
2. Rastogi V, Puri N, Arora S, et al: Artefacts: A diagnostic dilemma—A review. J Clin Diagn Res 7:2408-2413, 2013
3. Taqi SA, Sami SA, Sami LB, et al: A review of artifacts in histopathology. J Oral Maxillofac Pathol 22:279, 2018
4. Doyle S, Rodriguez C, Madabhushi A, et al: Detecting prostatic adenocarcinoma from digitized histology using a multi-scale hierarchical classification approach. Conf Proc IEEE Eng Med Biol Soc 1:4759-4762, 2006
5. Doyle S, Monaco J, Feldman M, et al: An active learning based classification strategy for the minority class problem: Application to histopathology annotation. BMC Bioinformatics 12:424, 2011
6. Lee G, Singanamalli A, Wang H, et al: Supervised multi-view canonical correlation analysis (sMVCCA): Integrating histologic and proteomic features for predicting recurrent prostate cancer. IEEE Trans Med Imaging 34:284-297, 2015
7. Janowczyk A, Basavanhally A, Madabhushi A: Stain Normalization using Sparse AutoEncoders (StaNoSA): Application to digital pathology. Comput Med Imaging Graph 57:50-61, 2017
8. Leo P, Elliott R, Shih NNC, et al: Stable and discriminating features are predictive of cancer presence and Gleason grade in radical prostatectomy specimens: A multi-site study. Sci Rep 8:14918, 2018
9. Madabhushi A, Lee G: Image analysis and machine learning in digital pathology: Challenges and opportunities. Med Image Anal 33:170-175, 2016
10. Bhargava R, Madabhushi A: A review of emerging themes in image informatics and molecular analysis for digital pathology. Annu Rev Biomed Eng 18:387-412, 2016
11. Janowczyk, Madabhushi A: Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use case. J Pathol Inform 7:29, 2016
12. Ehteshami Bejnordi B, Veta M, Johannes van Diest P, et al: Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. JAMA 318:2199-2210, 2017
13. Saltz J, Gupta R, Hou L, et al: Spatial organization and molecular correlation of tumor-infiltrating lymphocytes using deep learning on pathology images. Cell Reports 23:181-193.e7, 2018
14. Coudray N, Ocampo PS, Sakellaropoulos T, et al: Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. Nat Med 24:1559-1567, 2018
15. Bychkov D, Linder N, Turkki R, et al: Deep learning based tissue analysis predicts outcome in colorectal cancer. Sci Rep 8:3395, 2018

16. Wang H, Cruz-Roa A, Basavanhally A, et al: Cascaded ensemble of convolutional neural networks and handcrafted features for mitosis detection. Presented at the International Society for Optics and Photonics, San Diego, CA, February 15-20, 2014
17. Weinstein J. N., Collisson EA, Mills GB, et al: The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet 45:1113-1120, 2013
18. Goh WWB, Wang W, Wong L: Why batch effects matter in omics data, and how to avoid them. Trends Biotechnol 35:498-507, 2017
19. Avanaki AR, Espig KS, Xthona A, et al: Automatic image quality assessment for digital pathology. Presented at the 13th International Workshop on Breast Imaging, Malmö, Sweden, June 19-22, 2016
20. Ameisen D, Deroulers C, Perrier V, et al: Towards better digital pathology workflows: Programming libraries for high-speed sharpness assessment of whole slide images. Diagn Pathol 9:S3, 2014 (suppl 1)
21. Ameisen D, Deroulers C, Perrier V, et al: Stack or trash? Quality assessment of virtual slides. Diagn Pathol 8:S23, 2013 (suppl 1)
22. Senaras C, Niazi MKK, Lozanski G, et al: DeepFocus: Detection of out-of-focus regions in whole slide digital images using deep learning. PLoS One 13:e0205387, 2018
23. Wen S, Kurc TM, Gao Y, et al: A methodology for texture feature-based quality assessment in nucleus segmentation of histopathology image. J Pathol Inform 8:38, 2017
24. Edsall RM: The parallel coordinate plot in action: design and use for geographic visualization. Comput Stat Data Anal 43:605-619, 2003. https://www.sciencedirect.com/science/article/pii/S0167947302002955
25. Janowczyk A: HistoQCRepo. http://histoqcrepo.com/
26. Bui MM, Riben MW, Allison KH, et al: Quantitative image analysis of human epidermal growth factor receptor 2 immunohistochemistry for breast cancer: Guideline from the College of American Pathologists. Arch Pathol Lab Med [epub ahead of print on January 15, 2019]
27. Shrestha P, Hulsken B: Color accuracy and reproducibility in whole slide imaging scanners. J Med Imaging (Bellingham) 1:027501, 2014
28. Shrestha P, Kneepkens R, Vrijnsen J, et al: A quantitative approach to evaluate image quality of whole slide imaging scanners. J Pathol Inform 7:56, 2016
29. Zech JR, Badgeley MA, Liu M, et al: Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. PLoS Med 15:e1002683, 2018
30. Dodge S, Karam L: Understanding how image quality affects deep neural networks Presented at the Eighth International Conference on Quality of Multimedia Experience, Lisbon, Portugal, June 6-8, 2016

Supplemental Material

1. TCGA-BRCA Batch Effects

Technical artifacts are defined as non-biological factors which affect presentation or measurements in an experimental study (e.g., compression algorithm used in digital pathology slides). Batch effects are the systematic occurrence of a technical artifact within a subset of a cohort. For example, the preferred tissue thickness may be different between two institutions, creating a batch effect if those slides were evaluated in unison. Batch effects require special management in experimental designs (Goh WWB et al: Trends Biotechnol 35:498-507, 2017) since they are known to impede the success of bioinformatics experiments (e.g., mutation-calling and gene-expression). This is also true in the case of digital pathology. Consider the following example. Facility A produces thick slices which also belong to a positive target class, while Facility B produces thin slices which also belong to a negative target class. When studied together, one must be very careful to make sure that the learned model is in fact focusing on the biological signal of the target class, as opposed to solely the tissue thickness.

We sought to determine if there was evidence that similar batch effects are present in popular public digital pathology repositories, such as The Cancer Genome Atlas (TCGA; Weinstein JN, et al: Nat Genet 45:1113–1120, 2013) as a result of being curated from multiple laboratories. To evaluate the presence of batch artifacts, HistoQC was used to compute 12 color metrics (see below) on n=518 breast cancer (BRCA) tissue images which were plotted in a 2 dimensional space using T-SNE (a popular dimensionality reduction technique; van der Maaten L, et al: J Mach Learn Res 9:2579-2605, 2008). We made the assumption that if no batch effects were present, the images would be randomly clustered, as opposed to co-clustering with other images corresponding to the laboratory of origin. In other words, the stronger the clustering between multiple slide images from the same lab, the stronger this suggests the presence of batch effects within the cohort.

A total of n=518 samples from the TCGA-BRCA cohort were randomly selected under the following constraints:

1) The image was scanned at 40x
2) The slide came from a laboratory which contributed *at least* 20 samples to the TCGA. The minimum of 20 image samples was selected to ensure appropriate representation of potential intra and inter laboratory preparation and scanning variabilities.

This resulted in the following slide counts:

Facility Code	**Count**	**Facility Code**	**Count**
A2	46	BH	81
A7	30	C8	25
A8	48	D8	64
AC	28	E2	48
AO	22	E9	36
AR	33	EW	25
B6	32		

We note that in this cohort, all slides appear to have been scanned using an Aperio scanner (as indicated by the metadata). However, it appears unlikely that the same model or physical scanner was universally employed.

The HistoQC standard config.ini pipeline (see https://github.com/choosehappy/HistoQC/blob/master/config.ini) was subsequently applied on these samples. From the tab separated results file, 12 metrics associated with color presentation were extracted (see online documentation for feature explanations):

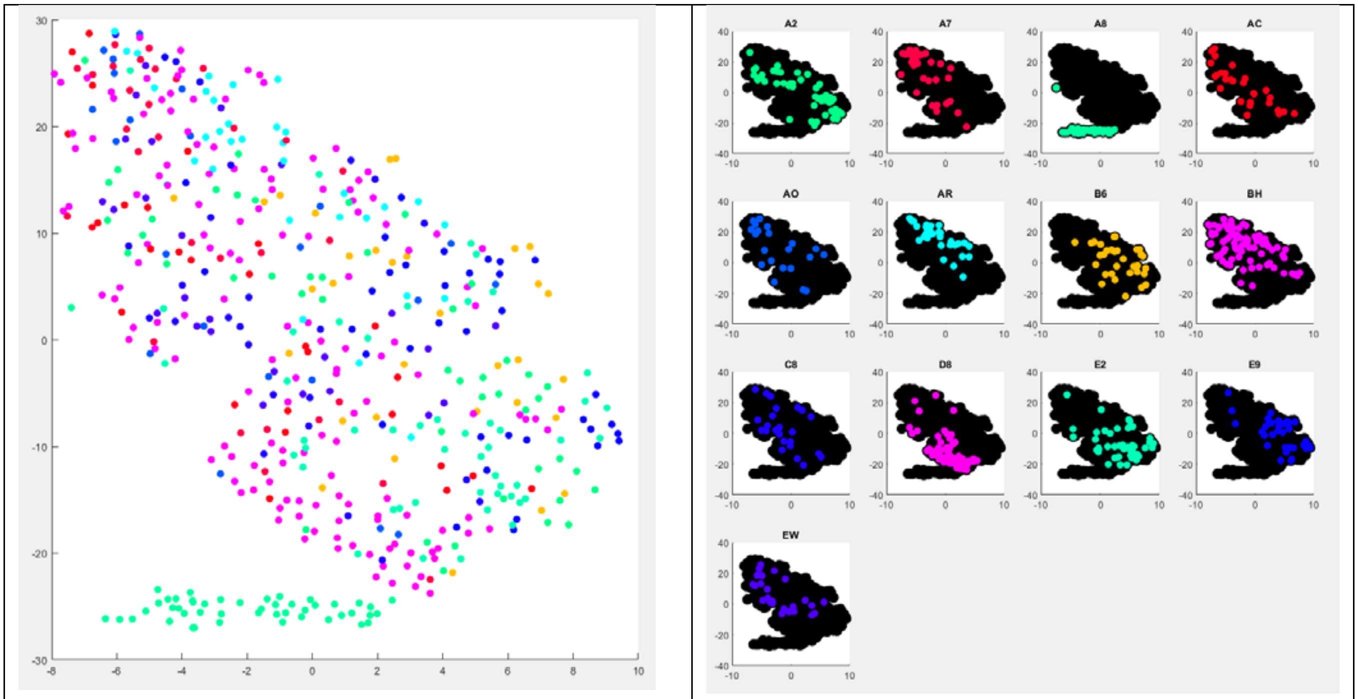
template1_MSE_hist	michelson_contrast	chan1_brightness	chan1_brightness_YUV
template2_MSE_hist	rms_contrast	chan2_brightness	chan2_brightness_YUV
template3_MSE_hist	grayscale_brightness	chan3_brightness	chan3_brightness_YUV
template4_MSE_hist			

This set of 12 metrics was then supplied to T-SNE to perform dimensionality reduction to enable 2D plotting. Note that this is an unsupervised process. During plotting, each laboratory was assigned a color code, allowing for a visual representation of metrics according to their lab of origin (See Figure S1a). Additionally, to allow for easier visualization, we show each lab individually overlaid with the other labs in black (See Figure S1b).

In this experiment, there are two arrangements of the points which provide evidence for a lack of batch effects in the samples. In the first arrangement, all points would converge to a single location thus indicating no difference in metrics across not only labs but samples. We note that this configuration is unlikely as each slide will have a different composition of biological material. For example, different proportions of stroma to epithelium will affect the 12 metrics used in this study. In the second more desired arrangement, the colors of the points should present with high spatial entropy (such as subplot BH). That is to say, they should not form localized clusters by color, as this would imply that some metrics are highly correlated with the underlying lab label. Examples of this include grouping of certain labs (e.g., A8, D8) and preference of spatial location in others (e.g., E9 favoring the right side), together suggesting batch type effects in sample preparation and/or scanning.

The slides used in this experiment can be downloaded following this tutorial:

http://www.andrewjanowczyk.com/download-tcga-digital-pathology-images-ffpe/



(a)

(b)

Figure S1. (a) Samples from 13 laboratories plotted in the 2 dimensional embedded space produced by T-SNE, with (b) the same labs shown in individual plots (with other laboratories in black) to highlight their locations. The high dimensional color space features computed by HistoQC appear to allow for the appreciation of visual clusters based on the originating facility of the sample, suggesting the presence of potential batch effects.

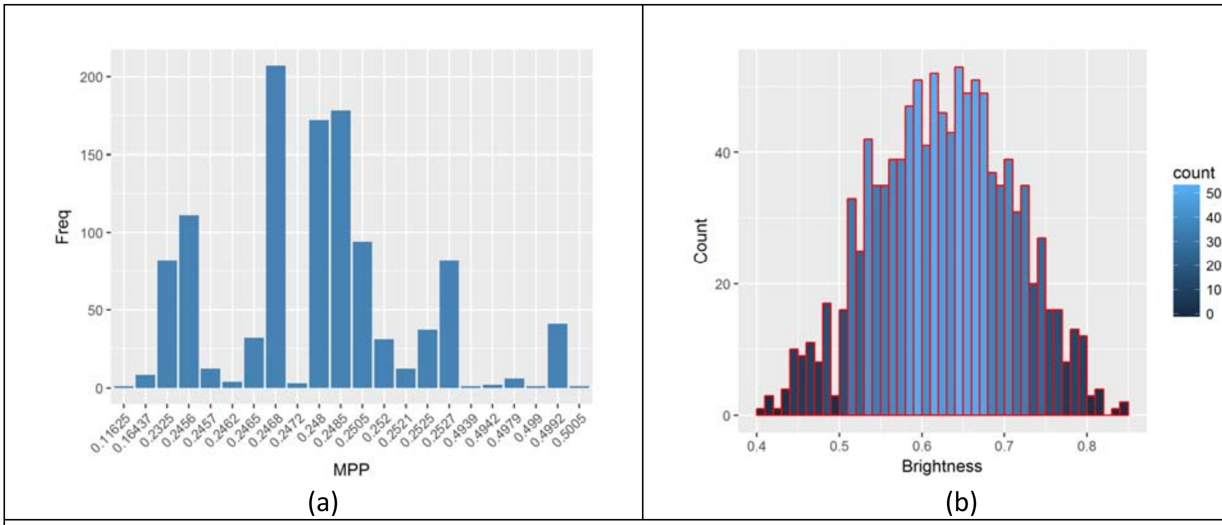
2. Currently implemented HistoQC modules

File module	*Operations*	*Description*
MorphologyModule.py	removeSmallObjects	Remove small items from the image. This is typically done for reducing small pixel noise, dust, etc
	fillSmallHoles	Fill in small/medium sized ""holes"" in images. For example, lumen spaces in tubules often are detected as background and removed from the final mask. This module will fill them in.
LightDarkModule.py	getIntensityThresholdOtsu	Thresholds the image based on dynamic Otsu threshold
	getIntensityThresholdPercent	Thresholds the image based on user supplied values. This is good for detecting where the tissue is on the slide (non-white) and where folded tissue may be (very dark)
HistogramModule.py	getHistogram	Makes a histogram image in RGB space
	compareToTemplates	Compares the image's histogram to template images provided by the user
DeconvolutionModule.py	seperateStains	Performs stain deconvolution using skimage's built in matrices
ClassificationModule.py	pixelWise	Applies an RGB based classifier to the image whose values come from a user inputted TSV
	byExampleWithFeatures	Computes features of template images provided by the user which have associated binary masks indicating positive and negative classes. Trained classifier is then used on images. Excellent for, e.g., pen detection (with texture), cracks, etc
BubbleRegionByRegion.py	roiWise	Detect contours of lines of air bubbles on slide. Contains exemplar of how to use HistoQC to iteratively loop over very large images at high mag. (work in progress)
BrightContrastModule.py	getBrightnessGray	Computes the average value of the image in gray color space, which ultimately represents how bright the image is perceived
	getBrightnessByChannelinColorSpace	Computes a triplet (one per color channel) in the desired color space. Useful for detecting outliers
	getContrast	Computes both RMS and Michelson contrast metrics (Kukkonen H, et al: *Vision Res* 33:1431-1436, 1993)
PenMarkingModule.py	identifyPenMarking	Identities pen markings on a pixel by pixel basis by using user supplied TSV file of color values. This is usually suitable when the marking is very different from the staining (e.g., green/blue marker on pink tissue).
BlurDetectionModule.py	identifyBlurryRegions	Uses a Laplace matrix to determine which regions in the image are likely blurry
BasicModule.py	getBasicStats	Pulls out metadata from image header
	getMag	Pulls out base magnification. This is required by HistoQC. In the future we'll add ability to predict magnification
	finalComputations	Computes the final number of pixels available in the output image. Too high or low of a number often indicate incorrect processing or image outliers
	finalProcessingSpur	Removes spurious morphology from the final mask. Essentially small "arms" of tissue are rounded off and removed

	finalProcessingArea	Removes larger islands from the output mask, e.g., isolated pieces of tissue
SaveModule.py	saveFinalMask	Saves both the output mask from HistoQC but also the overlay on the original image
	saveThumbnails	Save thumbnails for easier viewing. This needs to be completed for the UI to work

Table S1 Currently implemented HistoQC modules. Additionally planned modules and their statuses can be found in the Github Issues section.

3. Heterogeneity in TCGA-BRCA cohort



(a) (b)

Figure S2 Histograms illustrating potential batch effects and heterogeneity in (a) microns per pixel (MPP) and (b) brightness in TCGA-BRCA slides. We can see the MPP heterogeneity (a) does not produce a Gaussian distribution, while the brightness metric (b) does. Regardless, one can appreciate that when comparing images at the extremes of either spectrum, one would anticipate significant differences in presentation and algorithm performance.

Being able to automatically identify the extent of heterogeneity with respect to slide preparation and color variance in a batch is important, so that appropriate steps can be taken to harmonize slide image appearance, thus likely improving algorithmic performance. For example, as evidenced by Figure S2(b) the likelihood that a single brightness threshold value would provide a consistent result across the visualized cohort is low. As such, the heterogeneity reflects implicitly the fact that the cohort is actually comprised of smaller and distinct slide clusters with unique stain and preparation induced attributes. Hence, ideally one would want to define optimal parameters for subsequent processing on a subset basis. Tuning the parameters for downstream applications (e.g., segmentation) per cluster should result in a more robust algorithmic performance as opposed to a global set of parameters attempting to describe the entire cohort.

4. Selected HistoQC Failures

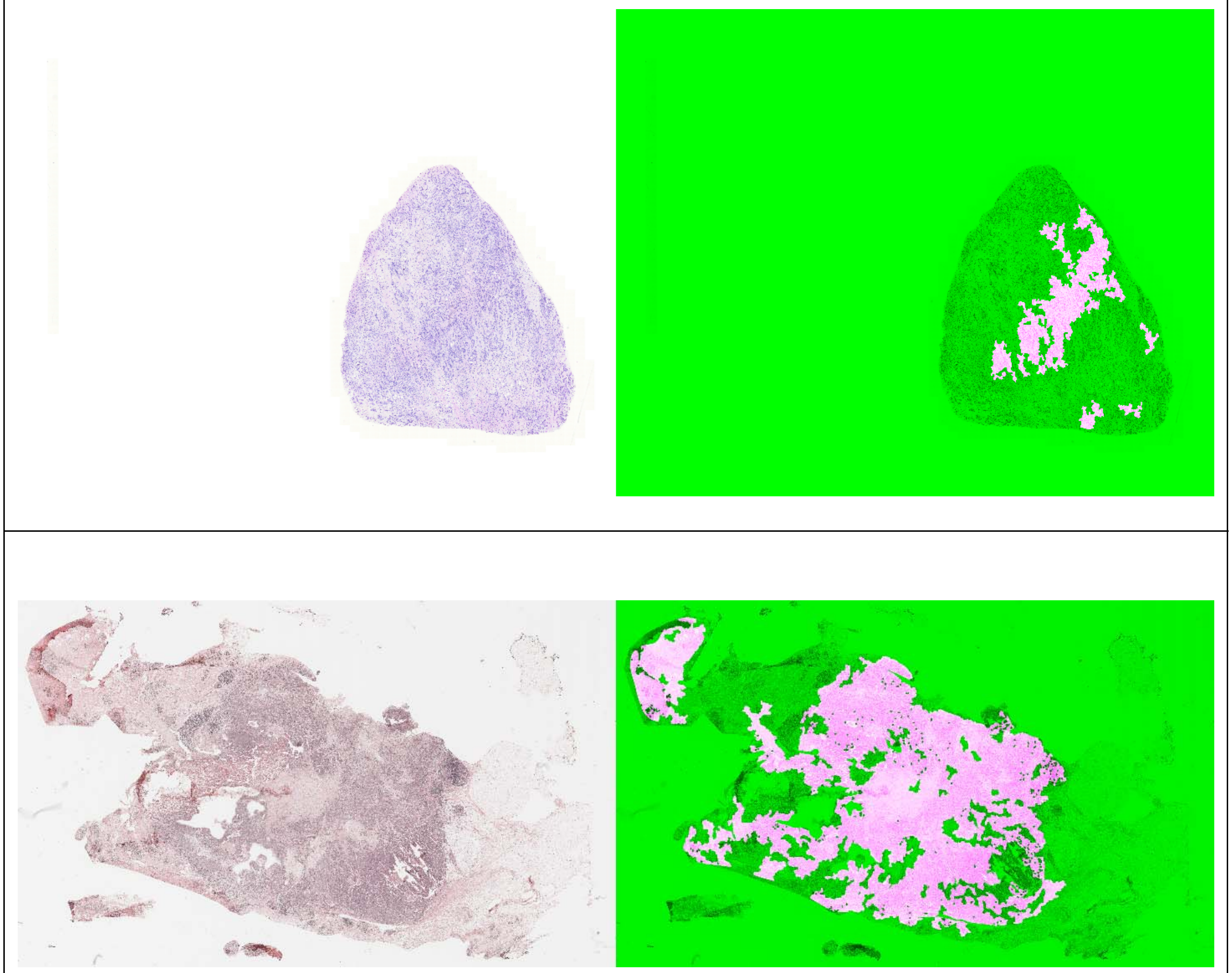


Figure S3a HistoQC faliures of tissue detection in regions where stroma are very lightly stained. The pathologists indicated that they would have expected more tissue to be covered by the masks.

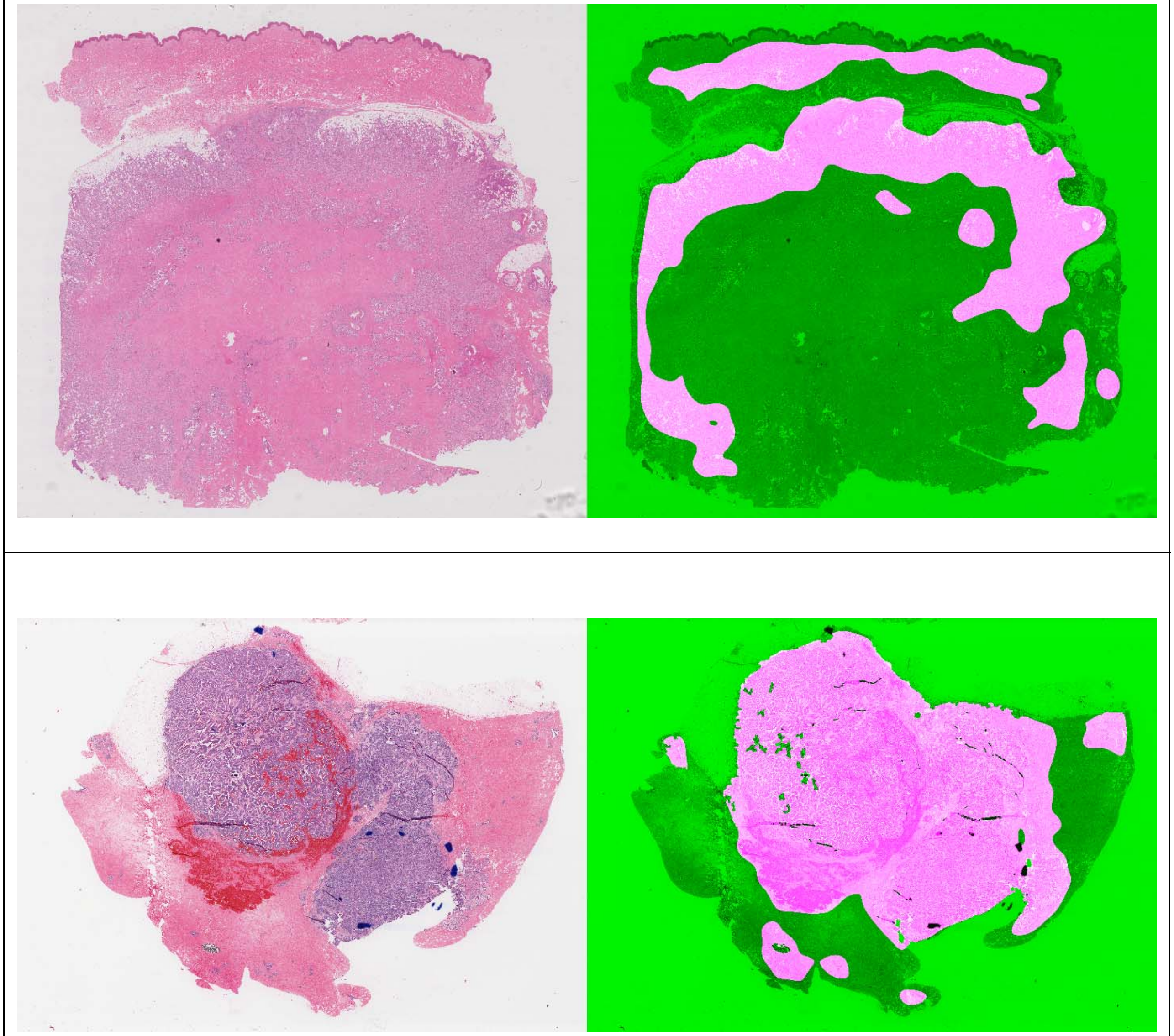


Figure S3b HistoQC faliures of tissue detection in large stroma rich regions, due to the lack of appreciable edges required by the blurriness detector to function. The experts reviewed the non-masked tissue and determined it to be suitable for computation.

Above are presented 4 cases where the experts were not in agreement with HistoQC's output masks. In the cases shown in of Figure S3a, the light stromal regions were incorrectly masked out, and subsequent pipeline steps (such as the spurious tissue removal module, see Github for documentation) further exacerbated the process of tissue masking. Figure S3b illustrates cases where HistoQC identified tissue as too blurry for computational usage. The bluriness detector uses an edge detection algorithm to determine the quantity and strength of image edges in a region. One can appreciate that in blurry regions, sharp edges are typically not present. In the case of these stromal regions, a lack of a sufficient number of tissue primitives (e.g. nuclei) at the chosen operating magnigication (2.5x) resulted in an insufficient number of image edges. This in turn incorrectly caused the algorithm to identify these regions in the image to be identified as blurry.

In both cases, metrics associated with the number of pixels masked by HistoQC would show up as outliers on HistoQC's user interface (Figure 2a and Figure 3, parallel coordinate plot). A review of the output masks (Figure 2b) from each of the modules would help identify which modules needed adjustment.