Journal of Pathology
J Pathol March 2021; **253:** 268–278
Published online 5 January 2021 in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1002/path.5590

ORIGINAL PAPER

Assessment of a computerized quantitative quality control tool for whole slide images of kidney biopsies

Yijiang Chen[1†*], Jarcy Zee[2†], Abigail Smith[2], Catherine Jayapandian[1], Jeffrey Hodgin[3], David Howell[4], Matthew Palmer[5], David Thomas[4,6], Clarissa Cassol[7,8], Alton B Farris III[9], Kathryn Perkinson[4], Anant Madabhushi[1,10], Laura Barisoni[4,11‡] and Andrew Janowczyk[1,12‡]

[1] *Department of Biomedical Engineering, Case Western Reserve University, Cleveland, OH, USA*
[2] *Arbor Research Collaborative for Health, Ann Arbor, MI, USA*
[3] *Department of Pathology, University of Michigan, Ann Arbor, MI, USA*
[4] *Department of Pathology, Duke University, Durham, NC, USA*
[5] *Department of Pathology, University of Pennsylvania, Philadelphia, PA, USA*
[6] *Nephrocor, Memphis, TN, USA*
[7] *Renal Pathology Division, Arkana Laboratories, Little Rock, AK, USA*
[8] *Department of Pathology - Renal Pathology Division, Ohio State University Medical Center, Columbus, OH, USA*
[9] *Department of Pathology and Laboratory Medicine, Emory University, Atlanta, GA, USA*
[10] *Louis Stokes VA Medical Center, Cleveland, OH, USA*
[11] *Department of Medicine, Division of Nephrology, Duke University, Durham, NC, USA*
[12] *Precision Oncology Center, University of Lausanne, Lausanne, Switzerland*

**Correspondence to: Y Chen, Department of Biomedical Engineering, Case Western Reserve University, Wickenden Building, 2071 Martin Luther King Jr Dr, Rm 525, Cleveland, OH 44106, USA. E-mail: yxc627@case.edu*

[†]Equal first authors.
[‡]Co-senior authors.

Abstract

Inconsistencies in the preparation of histology slides and whole-slide images (WSIs) may lead to challenges with subsequent image analysis and machine learning approaches for interrogating the WSI. These variabilities are especially pronounced in multicenter cohorts, where batch effects (i.e. systematic technical artifacts unrelated to biological variability) may introduce biases to machine learning algorithms. To date, manual quality control (QC) has been the *de facto* standard for dataset curation, but remains highly subjective and is too laborious in light of the increasing scale of tissue slide digitization efforts. This study aimed to evaluate a computer-aided QC pipeline for facilitating a reproducible QC process of WSI datasets. An open source tool, HistoQC, was employed to identify image artifacts and compute quantitative metrics describing visual attributes of WSIs to the Nephrotic Syndrome Study Network (NEPTUNE) digital pathology repository. A comparison in inter-reader concordance between HistoQC aided and unaided curation was performed to quantify improvements in curation reproducibility. HistoQC metrics were additionally employed to quantify the presence of batch effects within NEPTUNE WSIs. Of the 1814 WSIs (458 H&E, 470 PAS, 438 silver, 448 trichrome) from $n = 512$ cases considered in this study, approximately 9% (163) were identified as unsuitable for subsequent computational analysis. The concordance in the identification of these WSIs among computational pathologists rose from moderate (Gwet's AC1 range 0.43 to 0.59 across stains) to excellent (Gwet's AC1 range 0.79 to 0.93 across stains) agreement when aided by HistoQC. Furthermore, statistically significant batch effects ($p < 0.001$) in the NEPTUNE WSI dataset were discovered. Taken together, our findings strongly suggest that quantitative QC is a necessary step in the curation of digital pathology cohorts.

Keywords: digital pathology; kidney biopsies; quality control; computational pathology; computer vision; machine learning; whole-slide image; inter-reader variability; batch effects; NEPTUNE

Received 10 August 2020; Revised 30 October 2020; Accepted 11 November 2020

Conflict of interest statement: AM is an equity holder in Elucid Bioimaging and in Inspirata Inc. In addition, he has served as a scientific advisory board member for Inspirata Inc, Astrazeneca, Bristol Meyers-Squibb and Merck. Currently he serves on the advisory board of Aiforia Inc. He also has sponsored research agreements with Philips, AstraZeneca and Bristol Meyers-Squibb. His technology has been licensed to Elucid Bioimaging. He is also involved in a NIH U24 grant with PathCore Inc, and three different R01 grants with Inspirata Inc. No other conflicts of interest were declared.
The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health, the U.S. Department of Veterans Affairs, the Department of Defense, or the United States Government.

Introduction

Large multi-institutional consortia [1] and digital pathology repositories (DPRs) of renal biopsies have exposed the lack of standardization in tissue preparation and the presence of variabilities in the presentation and quality of whole-slide images (WSIs). These variabilities are often due to differences in either slide preparation (e.g. tissue fixation, processing, cutting, staining) or scanning (e.g. brightness, contrast, saturation, compression) across and within laboratories. Multicenter DPRs are especially likely to contain these variabilities due to the lack of cross-site standardized protocols, particularly in the context of special stains.

In a fully digital pathology laboratory setting, poor-quality tissue presentation may result in delays in pathology reporting. This delay is usually caused by poor-quality glass slides needing to be reproduced or WSIs rescanned and adds unnecessarily to storage overheads if the slides are not diagnostically usable. Evidence suggests that although clinical pathologists' interpretations are not impacted by differences in WSI quality [2], these differences may negatively affect the performance of digital pathology-based computational tools, including machine or deep learning algorithms [3,4]. Current best practices in digital pathology include the manual quality control (QC) of WSIs before experimental execution. This involves the subjective removal of poor-quality slides or avoidance of tissue regions containing artifacts as determined by the experimenter. However, this manual process of identifying poor-quality slides can vary substantially between experts and raises concerns from a scientific reproducibility standpoint.

Another issue that has been well recognized in fields [5] other than digital pathology regards the presence of batch effects, i.e. systematic technical differences when samples are processed and measured in different batches, that are unrelated to the biological variation of the tissue samples. Batch effects are especially likely to be introduced into multisite DPRs, where samples originate from different laboratories, as previously illustrated in The Cancer Genome Atlas (TCGA) dataset [6]. However, the importance of the identification and compensation for batch effects in DPRs remains still widely unrecognized. Without special care and attention to the problem, machine learning algorithms such as deep learning [7] may attempt to model batch effects, thus introducing significant biases. Notably, the identification and management of subtle batch effects remains an open challenge in the field of computational pathology.

We hypothesized that a quantitative QC approach, driven by algorithmically defined metrics, may facilitate an efficient and reproducible QC paradigm, as well as aid in the identification of both obvious and subtle batch effects. Janowczyk *et al* [6] introduced HistoQC (Version 1.0, Center of Computational Imaging and Personalized Diagnostics, Case Western Reserve University, Cleveland, OH, USA), an open-source digital QC tool, which was employed in the context of identifying suboptimal breast cancer WSIs. HistoQC has been shown to: (1) quantitatively measure and capture WSI-level metrics (e.g. color, brightness, contrast) and (2) localize WSI regions affected by artifacts (e.g. coverslip edges, bubbles). This information facilitates the discovery of poor-quality WSI, ultimately helping users to select WSIs for subsequent image analysis and DPR storage. However, HistoQC has not been extensively evaluated in the context of kidney pathology WSIs or within non-H&E-stained WSIs.

In this study, we sought to identify WSIs unsuitable for computational analysis, confirm the reproducibility of this assessment among computational pathologists and evaluate the presence of batch effects within the Nephrotic Syndrome Study Network (NEPTUNE) DPR using HistoQC-derived metrics.

Materials and methods

The overall study design, including dataset curation, QC and experiments, is illustrated in Figure 1.

Dataset

NEPTUNE is a multisite observational cohort study of children and adults with glomerular disease, enrolled at the time of a clinically indicated kidney biopsy [8]. The renal biopsies were processed in 38 different pathology laboratories, collected and shipped to the NEPTUNE image coordinating center, where glass slides were centrally scanned by two scanners (Aperio Scanscope AT2, Leica Biosystems Inc., Buffalo Groove, IL, USA and Hamamatsu Nanozoomer 2.0 HT, Hamamatsu Corporation, Hamamatsu City, Japan; both with an Olympus UPlan-SApo 20X objective, with a 0.75 NA, and image doubler) and subsequently uploaded into the NEPTUNE DPR [9]. In total, 1814 WSIs from 512 digital renal biopsies, including 458 stained with H&E, 470 with PAS, 438 with silver (SIL) and 448 with trichrome (TRI), were included in this study (Figure 1A). WSIs were chosen such that each patient contributed up to one randomly selected WSI per stain, resulting in a minimum of one to a maximum of four WSIs per case.

HistoQC functionality

HistoQC is designed to aid users in the completion of quantitative QC. HistoQC consists of a pipeline of modules sequentially applied to a WSI. Each module acts on the image to either: (1) quantify visual characteristics associated with a digital pathology image, allowing for identification of heterogeneity within a population of images (e.g. color, brightness and contrast) or (2) detect various artifacts that may be present on a WSI (e.g. pen markings and folded tissue). The modules used in HistoQC analysis for this study are summarized in Table 1. The HistoQC output consists of a quantitative report corresponding to the visual characteristics mentioned above as well as images delineating regions identified as artifact-free in each WSI; together these HistoQC

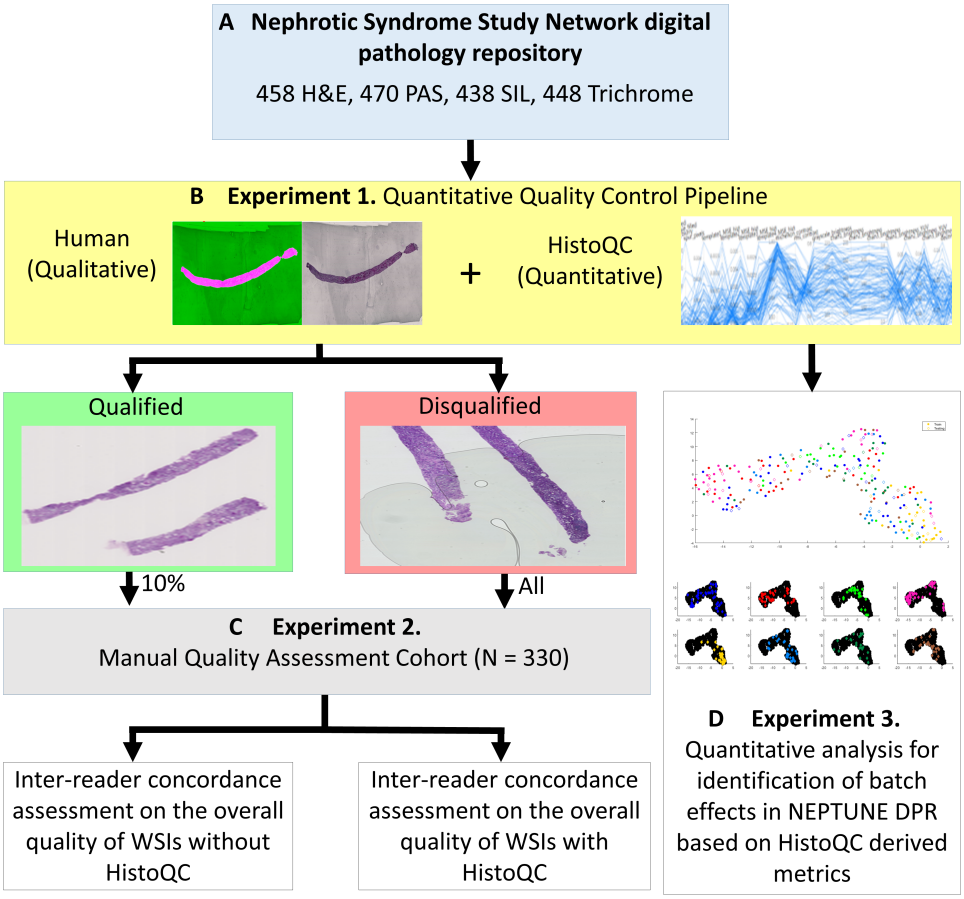

Figure 1. Experimental pipeline. Using over 1800 WSIs stained with H&E, PAS, SIL and TRI, from the NEPTUNE DPR, a HistoQC-aided QC pipeline was applied to each stain independently. WSIs were assessed for qualification of computational analysis as determined by the presence of artifacts and whether the WSI was an outlier within the stain population. Ten percent of the qualified WSIs and all of the disqualified WSIs were reviewed and scored by three reviewers R_1, R_2 and R_3, for evaluation of inter-reader concordance with and without using HistoQC. HistoQC quantitative quality metrics were later used to assess the presence of batch effects in the NEPTUNE data.

outputs facilitate the interactive interrogation of the WSI via its user interface.

Experiment 1: quantitative QC pipeline for computational analysis qualification of WSIs

For each stain, HistoQC was used to compute quality metrics reflecting visual properties of the WSIs (Table 2). HistoQC's quantitative metrics were visualized in a parallel coordinate plot (PCP) supplied by the user interface (Figure 1B). As each line in the plot represents a WSI, the clustering of lines visually indicates WSIs with similar properties. In contrast, WSI outliers with distinct visual properties compared with the rest of the cohort can be identified when the WSI's corresponding line is highly divergent from other WSIs. Metrics presenting with a large standard deviation indicate strongly diverse presentations of visual features in the dataset. From the HistoQC PCP and user interface, a computational pathologist identified outlying WSIs in each quality metric and visually assessed the HistoQC artifact identification results to identify suboptimal WSIs (Figure 2). WSIs that are especially poor in one metric, could probably be removed without much additional scrutiny at higher magnification. The extreme outliers shown in Figure 3 are some of these easily eliminated cases. Once a WSI passes each of the individual metrics, a global review of the cohort takes place to identify if there are any combinational metrics that should result in WSI removal.

Table 1. HistoQC modules.

Module name	Targeted artifact or metric
Basic Module	WSI magnification, file pyramid levels and microns per pixel
Light Dark Module	Identify tissue location and folded tissue
Classification Module	Identify pen marking, cover slip and cracks
Bubble Region By Region	Demarcate contours of air bubbles on WSIs
Bright Contrast Module	Overall and per channel tissue brightness, indicating stain/scan variations
Blur Detection Module	Identifies out of focus WSI regions

Experiment 2: evaluation of quantitative QC on inter-reader concordance of WSI curation

To determine if WSI cohort curation was more reproducible when computational pathologists employed HistoQC, the difference between their aided and unaided curation efforts was assessed. For further clinical comparison, these results were juxtaposed with an

examination of pathologists' unaided quality assessment as a baseline. All suboptimal WSIs (n = 163) and a 10% random sample of qualified WSIs (n = 167) from experiment 1 were combined to create experiment 2's (Figure 1C) dataset, resulting in 330 WSIs (78 H&E, 92 PAS, 77 SIL and 83 TRI). Inter-reader concordance of the QC process among two groups of investigators was assessed: (a) three computational pathologist readers ($R_{1–3}$) having extensive experience in WSI quality needed for computational digital pathology; and (b) seven clinical renal pathologist readers ($P_{1–7}$), who possess field expertise associated with renal pathology and historically have determined the suitability of histology preparations for human interpretation.

A comprehensive scoring with a detailed explanation of the scoring process and example images of each type of artifact was provided to help scorers (computational pathologist readers and clinical pathologist readers) minimize subjectivity within the experiment. The protocol (available in supplementary material, Section I) was reviewed for consensus during a webinar to facilitate cross-training between all scorers. A scoring sheet was designed for scorers to indicate their subjective perception on the quality/adequacy of a WSI for clinical assessment and computer analysis. Four choices were available for the subjective assessment of WSI adequacy: (1) good for feature extraction at the cellular level, (2) good for histologic primitive segmentation analysis but not cellular-level feature extraction, (3) good for conventional disease diagnosis but not for machine learning, or (4) not good for either machine learning or clinical diagnostic tasks. For the purpose of image analysis QC, choices (1) and (2) were merged to indicate good quality and (3) and (4) to indicate poor quality. Each scorer received a file containing an individual scoring sheet for each WSI, with an associated link to the WSI and identifying information (WSI ID, stain, disease category) so scorers could confirm they were scoring the correct images. WSIs were randomly assigned to $P_{1–7}$ such that each image was scored by two different clinical pathologists with every possible pair of clinical pathologists evenly distributed across the sample. The same WSIs were also randomly assigned to $R_{1–3}$ such that any pair of computational pathologists could be compared. After an 8-week washout period from the initial concordance assessment, $R_{1–3}$ re-scored the same WSIs with the aid of the HistoQC user interface. $R_{1–3}$ were blinded to their original scores and cases were randomly re-ordered. They referred to both qualitative output (artifact-free mask generated) and quantitative output (PCP of quality metrics) from HistoQC to evaluate each WSI and compare it against the rest of the dataset. For each WSI, $R_{1–3}$ indicated whether the WSI had good or poor quality as defined above.

Concordance was assessed among $R_{1–3}$ on their agreement in good versus poor quality of WSIs within each stain without (C_{unaid}) and with (C_{aid}) the use of HistoQC. Similarly, concordance (C_p) was also assessed among $P_{1–7}$ in their manual assessment of WSIs quality. Concordance is measured using proportion of agreement, Cohen's kappa and Gwet's Agreement Coefficient 1 (AC1), with the latter two agreement statistics supplied due to their corrections for chance agreement [10]. The statistical analysis in this study was conducted using SAS (SAS Institute Inc., Version 9.4, Cary, NC, USA).

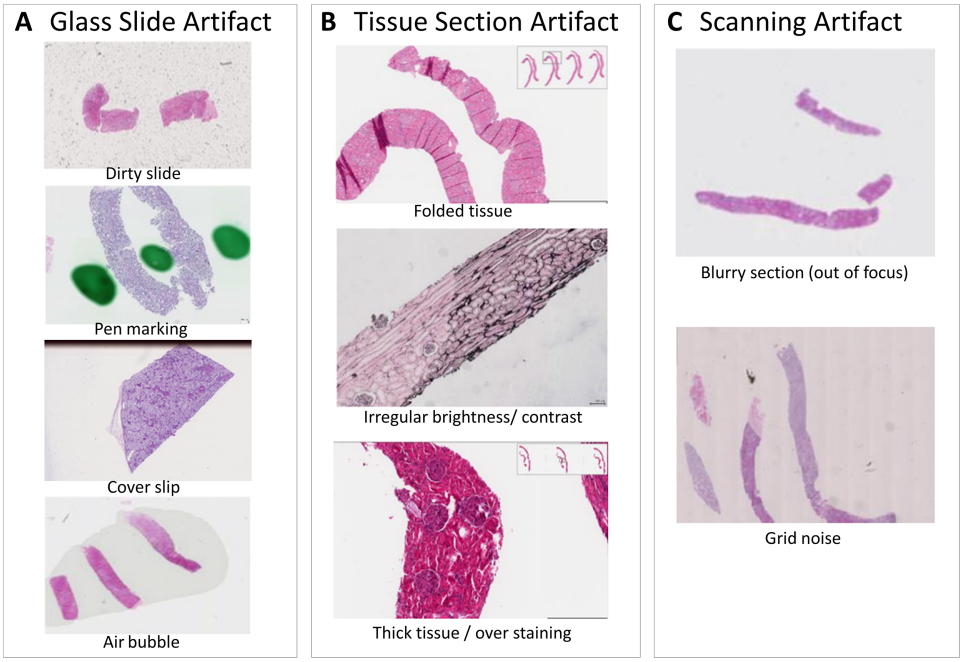

Figure 2. Example artifacts that frequently present on digital renal pathology images. In general, common artifacts found in digital renal pathology images can be divided into: (A) glass slide artifacts, (B) tissue section artifacts and (C) scanning artifacts.

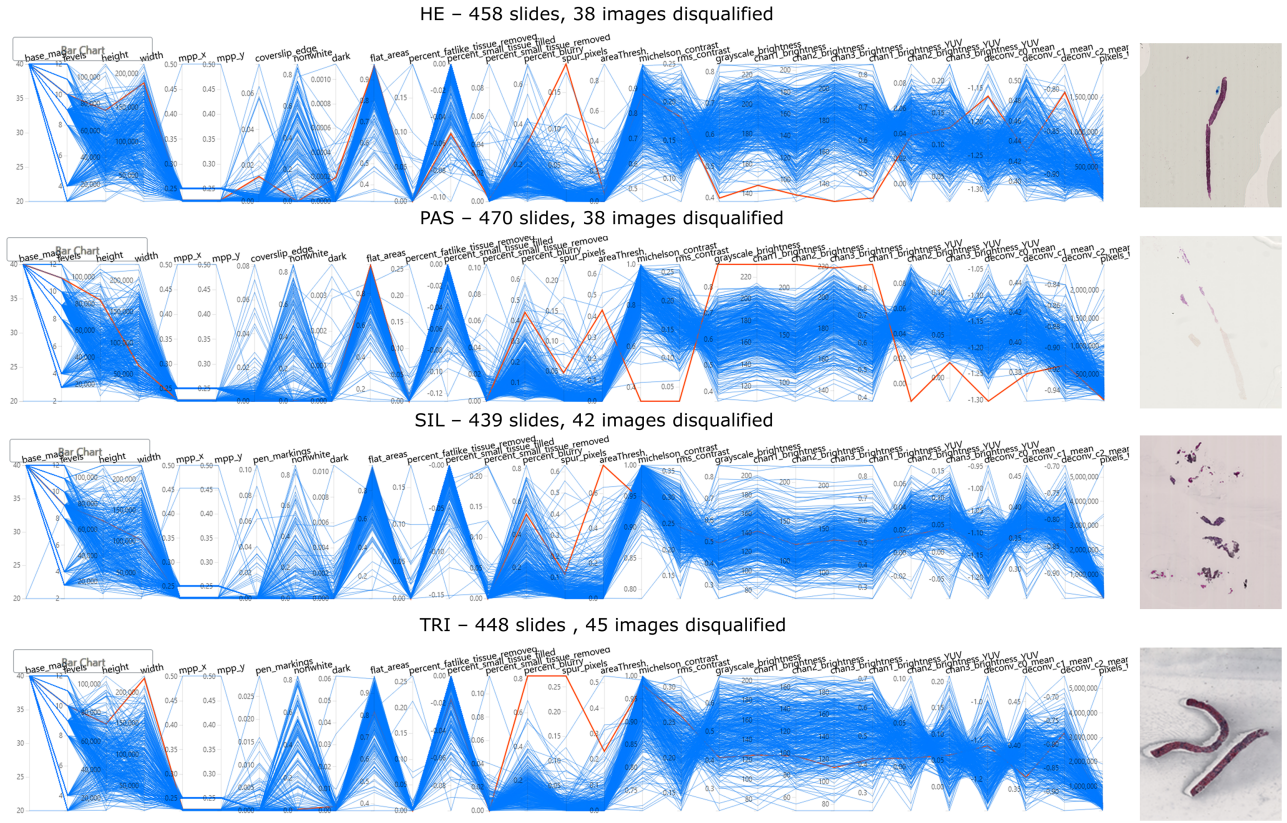

Figure 3. HistoQC user interface demonstrating selected metrics across four stain types. The PCPs provide an overview of the distribution of WSIs, with each blue line representing a single WSI and each *y*-axis representing the metric plotted on a normalized axis. Each vertical axis corresponds to a distinct image metric computed by HistoQC. Each horizontally orientated line represents a WSI analyzed by HistoQC. Examples of disqualified WSIs are shown and are highlighted in red in the plot. Disqualified images (red lines) are examples of outliers in certain metrics, indicating potential preparation artifacts. For example, in the first outlier in the H&E-stained image cohort, the WSI highlighted in red deviates from the majority of WSIs (blue lines) in metrics such as 'Spur_pixels', and the brightness of all color channels. This indicates that this WSI has many more spur pixels compared with the rest of the H&E-stained WSI, and the tissue itself is probably too dark compared with other H&E WSIs, as the brightness is low. These outlying metrics indicate that a more thorough manual quality assessment is warranted for this particular WSI: it was discovered that the image was dark because of thick cutting, over-staining and a large air bubble covering the entire core. The collection of these artifacts resulted in the disqualification for computational analysis of this WSI.

Experiment 3: heterogeneity and batch effect assessment of NEPTUNE WSIs

As HistoQC facilitates efficient, quantitative QC for digital pathology on large WSI cohorts, it was hypothesized that the HistoQC pipeline can also be applied to identify batch effects. In this experiment (Figure 1D), HistoQC quantitative metrics from experiment 1 were used to identify preparation artifacts associated with individual histology laboratories at the enrolling centers. These types of artifact are quantified by the brightness, contrast and color channel intensity measurements produced by the HistoQC pipeline. In this analysis, nine metrics associated with the chromatic appearance of WSI tissue were employed (Table 2). These features (denoted by F_1–F_9) were used to train a random forest (RF) classification model with 50 trees to match each WSI with its originating site S. To eliminate possible biases introduced by an unbalanced number of WSIs from each histology laboratory, only WSIs from sites supplying >20 WSIs were employed. These criteria reduced the 38 enrolling centers down to eight, which were then labeled as S_1–S_8. The resulting 250 WSIs were split into training (G_{train}) and testing (G_{test}) sets by a ratio of 8:2, where the training set (n = 200) was used to train the RF classifier and the testing cohort (n = 50) was used to determine classification accuracy.

To identify batch effects, a permutation test was conducted to assess whether the RF predictions significantly differed from predictions based on randomized labels. The null hypothesis for this test thus implies that HistoQC metrics cannot predict WSI origination sites any better than random assignment. For the 1000 iteration permutation test, site labels were randomly assigned to WSIs, and RF models with 50 trees were trained and evaluated using the same training (G_{train}) and testing (G_{test}) cohorts as with the original RF classifier. These permutations resulted in a distribution of 1000 accuracy measures based on random site labels. A P-value was generated by calculating the proportion of permuted accuracy measures that were at least as extreme as the accuracy from our original RF model. Features were then ranked by their predictor importance to identify factors driving the identification of batch effects.

Table 2. Quantitative metrics used to identify site-based batch effects in NEPTUNE.

Quality feature (F)	Description
F_1 rms_contrast	Root mean square (RMS) contrast, defined as the standard deviation of the pixel intensities across the pixels of interests
F_2 michelson_contrast	Measurement of image contrast defined by luminance difference over average luminance
F_3 grayscale_brightness	Mean pixel intensity of the image after converting the image to grayscale
F_4 chan1_brightness	Mean pixel intensity of the red color channel of the image
F_5 chan2_brightness	Mean pixel intensity of the green color channel of the image
F_6 chan3_brightness	Mean pixel intensity of the blue color channel of the image
F_7 chan1_brightness_YUV	Mean channel brightness of red color channel of image after converting to YUV color space
F_8 chan2_brightness_YUV	Mean channel brightness of green color channel of image after converting to YUV color space
F_9 chan3_brightness_YUV	Mean channel brightness of blue color channel of image after converting to YUV color space

These nine HistoQC features were selected to identify batch effects as they quantify chromatic artifacts imparted during the staining and cutting of the tissue samples – steps conducted at individual laboratories before central scanning in NEPTUNE.

Finally, all the features (F_1–F_9) were used in an unsupervised Uniform Manifold Approximation and Projection (UMAP) algorithm to reduce their dimensionality for 2D plotting to visualize the distribution of features among each site ($S_{1–8}$). For plotting, each site S was assigned a color, allowing for a visual representation of metrics according to site S. WSIs in the G_{train} of RF classification are shown as circles, whereas those from the G_{test} are represented by hollow diamonds. To allow for easier visualization, each histology laboratory was shown individually overlaid with the other institutions/laboratories in black.

Results

Experiment 1: quantitative QC pipeline for computational analysis qualification of WSIs

The quantitative quality metrics generated by HistoQC for the four stains evaluated can be seen in the PCPs of Figure 3. These plots exhibit dramatic variability in most HistoQC metrics, indicating notable tissue presentation differences (heterogeneity) in the WSIs considered. Employing the HistoQC user interface, $R_{1–3}$ identified 163 poor-quality WSIs (9.0% of 1814 WSIs). Disqualified WSI were distributed across all four stains: 38 from H&E, 38 PAS, 42 SIL and 45 TRI. These WSIs either had a variety of pronounced artifacts, including dirty slides, pen markings, tissue folding, thick tissue and blurriness (Figure 4), or appeared to be outlying from the majority of the WSIs with the same type of staining in the dataset. The remaining 1651 WSI were considered to be of good quality. Although minor artifacts may still remain in these images, they can be identified and masked by HistoQC, potentially facilitating computational image analysis (Figure 4).

Experiment 2: evaluation of quantitative QC on inter-reader concordance of WSI curation

The inter-reader concordance C_{unaid} across $R_{1–3}$ for the manual scoring of WSI quality/adequacy for each stain is shown in Table 3. For comparison, Table 4 gives the results of concordance among $P_{1–7}$. Moderate concordance was observed on WSIs stained with H&E (Gwet's AC1 = 0.59). By contrast there was slightly poorer concordance for the scoring of TRI-stained WSI quality (Gwet's AC1 = 0.43). Moderate concordance was obtained on SIL (Gwet's AC1 = 0.52) and PAS stains (Gwet's AC1 = 0.56). These results indicate only moderate reproducibility among $R_{1–3}$ on the selection of WSIs for computational analysis. C_p among $P_{1–7}$ varied widely across stains, with the highest agreement among WSIs stained with TRI (Gwet's AC1 = 0.68). By contrast, there was poor agreement on image quality of SIL-stained WSIs (Gwet's AC1 = 0.23). Moderate concordance was obtained on H&E (Gwet's AC1 = 0.58) and PAS stains (Gwet's AC1 = 0.52) (Table 4).

HistoQC-aided reader concordance C_{aid} was notably higher compared with C_{unaid}, the inter-reader concordance among $R_{1–3}$ without use of HistoQC (Table 3). Of note, in H&E staining, C_{aid} compared with C_{unaid} improved the most, by 0.33 in Gwet's AC1. In addition, computational pathologists reported that less effort was needed in identifying cohort-level outliers, whereas the time needed for artifact quality assessment was also noticeably decreased (about 90% decrease in time). WSIs that remained discordant were individually discussed by $R_{1–3}$ and found to be borderline cases. As such, these cases could not be consistently called without explicitly defining additional quantitative guidelines (examples are shown in supplementary material, Section II).

Experiment 3: heterogeneity and batch effect assessment of the NEPTUNE DPR

In the absence of batch effects, the RF classifier predicting the sites that produced the WSIs based on HistoQC metrics should not be able to perform better than random guessing. This null hypothesis coincides with a classification accuracy of between 0.1 and 0.15 (Figure 5A). In comparison, our RF classifier yielded a 0.52 overall accuracy, with a P-value <0.001, implying the existence of batch effects. The corresponding sites driving this appeared to be sites with high recall values: S_2 (recall = 66.7%), S_3 (recall = 100%), S_4 (recall = 80%), S_5 (recall = 75%), S_6 (recall = 75%) (Figure 5B). Features associated with WSI contrast and color intensity of red, green and blue channels appeared to contribute the most to the classification results, with the brightness of the blue channel reflecting the highest feature

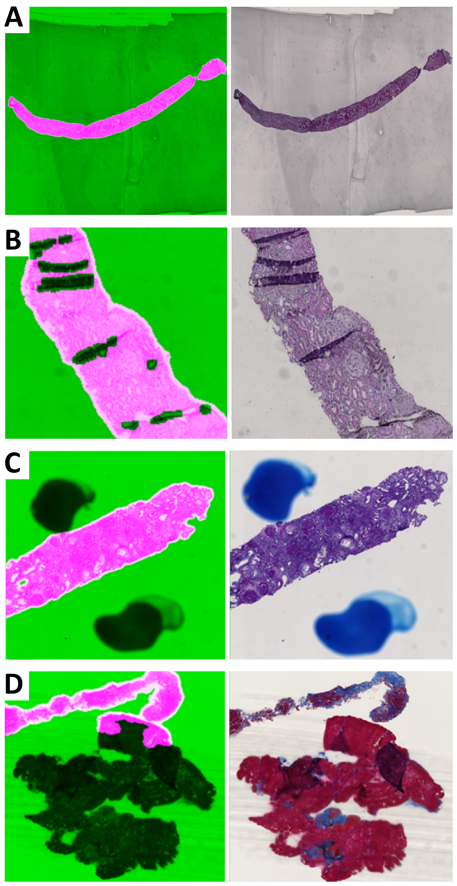

Figure 4. Examples of artifacts identified and associated HistoQC overlay image. For each example, the mask of computationally acceptable tissue overlaid on the WSI is presented on the left, where acceptable tissue areas are highlighted in pink, whereas background and noisy tissue areas are shown in green. The raw thumbnail for each WSI is presented on the right. For each panel, different artifact detection results are shown: (A) glass artifact: stain residue on a glass slide of a SIL WSI; (B) glass artifact: pen marking outside the core of a PAS WSI; (C) tissue artifact: tissue folding on a PAS WSI; (D) tissue and scanning artifact: thick tissue, tissue folding and blurriness on a TRI WSI.

Table 3. Agreement between computational pathologists R_1 and R_2/R_3 with and without the help of HistoQC.

Stain	Agreement	Kappa	Gwet's AC1
Without HistoQC			
H&E	0.73	0.26	0.59
PAS	0.73	0.31	0.56
SIL	0.75	0.50	0.52
TRI	0.69	0.36	0.43
With HistoQC			
H&E	0.96	0.91	0.92
PAS	0.89	0.75	0.79
SIL	0.96	0.93	0.93
TRI	0.90	0.77	0.81

importance (supplementary material, Section III). This seems to suggest that the batch effects in NEPTUNE DPR may have been imparted during the tissue staining process. The UMAP plot of the HistoQC metrics appears to confirm these quantitative results (Figure 6), as clustering was observed in sites S_2, S_3, S_4 and S_5.

Table 4. Agreement among human-assessed image quality.

	Concordance among clinical pathologists		
Stain	Agreement	Kappa	Gwet's AC1
HE	0.77	0.48	0.58
PAS	0.71	0.27	0.52
SIL	0.60	0.19	0.23
TRI	0.79	0.41	0.68

Discussion

Multicenter DPRs of renal biopsy WSIs have been established to facilitate precision medicine but are inundated with heterogeneity in tissue processing and scanning differences between centers. National guidelines for standardizing tissue processing were instituted in the pre-digital pathology era and may no longer be adequate in ensuring the tissue presentation quality needed by modern machine learning-enabled digital workflows. Currently, most clinical and research digital pathology workflows rely on manual QC of WSIs, a subjective, laborious and error prone process. In comparison, other technologies that have transitioned from analog to digital signals (e.g. RNA sequencing) now operate under rigorous QC processes, and as such disciplines like genomics [11,12], proteomics [13] and radiomics [14] have adopted systematic and reproducible digital QC [15]. Thus, during a similar transition in pathology, digital QC tools are expected to be similarly adopted. To this end, there have already been efforts to develop algorithms to improve QC in the DPR space, such as those detecting blurriness [16] and assessing slide quality [17–19]. HistoQC is not designed to replace these methods, but instead to provide a singular open-sourced pipeline for them to be embedded in, thus providing a means to visually examine their outputs via a singular user interface.

This study set out to begin evaluation of the effects of quantitative QC, via the integration of HistoQC, into routine QC procedures typically undertaken by computational researchers. The experiments included evaluating the multi-institutional NEPTUNE WSI DPR to identify WSIs with artifacts or cohort-level abnormal presentation, together with the presence of DPR-level batch effects.

Our study found that 9% of the NEPTUNE DPR WSIs have a suboptimal quality and should be considered for disqualification from subsequent computational image analysis. Poor-quality images were not limited to specific stain types and had a variety of artifacts contributing to their disqualification, thus implying a wide range of issues during WSI creation. In contrast to centers uploading WSIs directly to DPRs, the majority of NEPTUNE WSIs were centrally scanned by two scanners, probably minimizing differences associated with digitization. In settings where the WSIs are scanned by multiple scanners, scanner variabilities are anticipated to contribute additional quality inconsistency. These results highlight the importance of detailed, standardized protocols for histology preparation along with

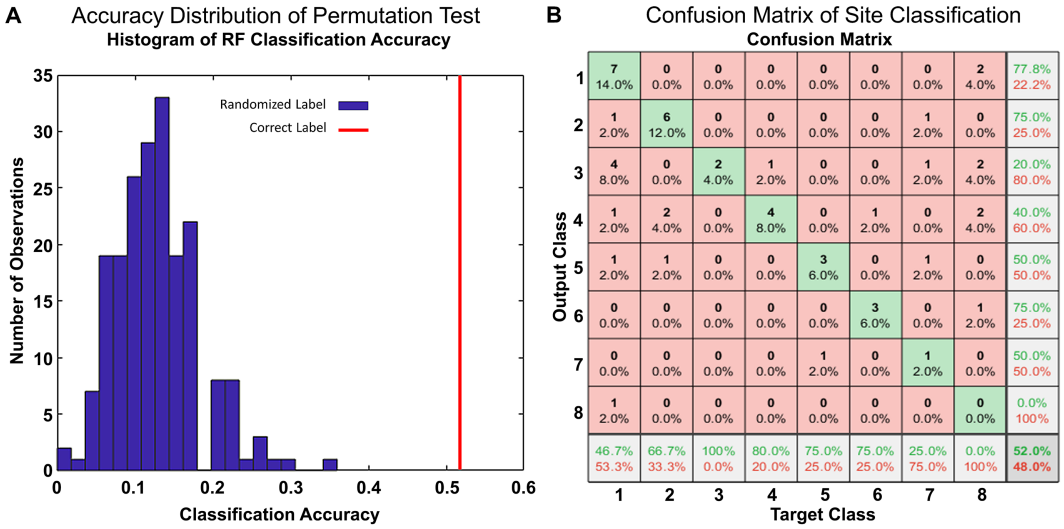

Figure 5. Statistical analysis of batch effect presence. (A) Histogram showing accuracy distribution of RF classifiers trained with randomized site labels (blue bins) from a permutation test. The accuracy of a RF classifier trained with correct labels is highlighted on the figure in red. (B) Confusion matrix illustrating RF predicted sites of the $n = 50$ testing cohort; rows correspond to the predicted class (output class) and columns to true class (target class). Diagonal cells correspond to observations that are correctly classified. Both the number of observations and the percentage of the total number of observations are shown in each cell. The last column shows the precision, or positive predictive value, in green. The bottom row shows the recall, or true positive rate, in green. The bottom right cell shows the overall accuracy. Sites S_2 (recall = 66.7%), S_3 (recall = 100%), S_4 (recall = 80%), S_5 (recall = 75%), S_6 (recall = 75%) can be seen to have high recall values, driving the overall accuracy of the classifier, and demonstrating the presence of detectable batch effects.

quality monitoring during the image acquisition and DPR curation.

Our results further demonstrate that despite detailed guidelines provided to readers, manual quality assessment of WSIs, whether performed by computational pathologists or clinical pathologists, remains subjective and showed limited reproducibility. On the other hand, a substantially higher concordance was witnessed when the scoring computational pathologists were aided by HistoQC. It is interesting to note that although a WSI may appear artifact free in isolation, when placed among its peers, it may manifest as an outlier due to presentation differences. Although stark infrequent differences are typically easily identified, non-systemic subtle differences are harder to manually identify at scale. For example, in images with large contrast and staining differences from the rest of the cohort (Figure 6, as seen by the WSI thumbnails in red and yellow boxes), both visual and quantitative metrics show high divergence, making them easier to identify. On the other hand, those with subtle contrast or staining differences within the dataset (see supplementary material, Section II Case 1) may not be readily appreciated without comparison to other slides, thus necessitating quantitative metrics. Our results indicate that employing HistoQC helps to address this issue, evidenced by the increase in computational pathologist concordance during cohort curation. On the other hand, systemic differences are also of concern, as these batch effects may potentially confound clinical variables of interest with WSI preparation artifacts. In our NEPTUNE DPR, HistoQC metrics were able to statistically identify batch effects in at least three pathology laboratories. These batch effects were associated with stain and tissue thickness heterogeneity. Although obvious batch effects (see examples of WSIs with heterogeneous visual characteristics in Figure 6) are likely to be identifiable via visual inspection, our results show that subtle batch effects are more difficult to identify in isolation, and yet may still affect image analysis and machine learning algorithms. Although HistoQC now includes a real-time UMAP plot to facilitate similar investigations, improved awareness of batch effects will be needed in the future to help limit their negative impact.

This study does however have limitations worth noting. Artifacts of interest were constrained to those presenting within the NEPTUNE DPR and thus may not represent the entire spectrum of possible slide generation and scanning issues. Grading of slides was grouped into a two-tier system (qualified/disqualified), which may obfuscate subtle patterns only visible within more granularly refined tiers. This decision was made due to significant reader discordance in our preliminary experiments involving the aforementioned four tiers, thus a refinement of the experimental design for improving concordance in the unaided case was deemed necessary. Furthermore, two key areas for future investigations not addressed in this study remain: first, associating specific processes in WSI preparation with artifacts and batch effects, in addition to suggesting protocol adjustments to ameliorate them; and second, assessing concordance implications after training clinical personnel in the use of HistoQC, and how that knowledge impacts future WSI generation.

As clinical pathology is transitioning to a more digital practice, approaches for QC will have to evolve accordingly. This process will naturally require increased interdisciplinary collaboration and cross-specialization

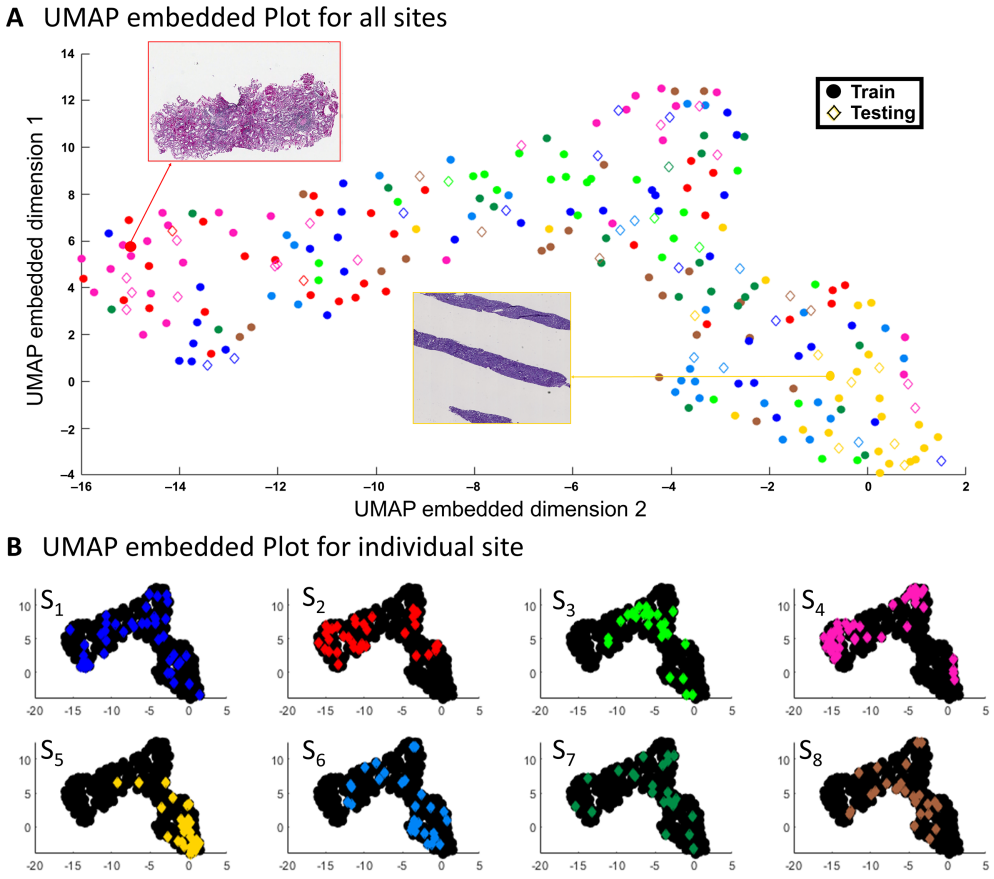

Figure 6. UMAP embedded plot for assessment of batch effects in NEPTUNE DPR. (A) Samples from eight sites plotted in the 2D embedded space produced by UMAP, where examples from left (red arrow, site S_2) and right (yellow arrow, site S_5) are shown, where circles represent cases from the training set and diamonds represent cases from the test set for all color pairs, with (B) the same sites shown in individual plots (with other laboratories in black) to highlight their distributions. The UMAP embedding was generated in an unsupervised manner, with the training and testing cases used in the RF experiment shown as circles and diamonds, respectively. These labels appear to cluster by originating WSI site well, indicating that training and testing samples are near each other in the high dimensional color space features computed by HistoQC. As can be observed, sites S_2, S_3, S_4, S_5 are demonstrating concise clusters, indicating the potential presence of batch effects. These findings are in line with observations from the confusion matrix in Figure 5B. Panel (A) further demonstrates that notable presentation differences are driving divergent locations on the plot, with the left WSI showing a higher red and lower blue channel intensity versus the WSI on the right having heavy contrast and a high intensity blue channel.

education between pathology and computational personnel. For example, pathology personnel will probably require specialized training to employ HistoQC, as it is currently geared toward those with computational expertise. These cross-pollination training activities will help to better define the vocabulary needed for each group to work toward their clinical targets as well as computational algorithm development. Armed with more precise laboratory procedures and appropriate monitoring tools, workflows for consistent slide creation can be designed, starting from tissue collection to WSI generation. The constant monitoring of WSI production can alert laboratory staff of potential equipment malfunctions sooner, reducing the number of slides not suitable for computational analysis. On a broader scale, when a similar process is undertaken across multiple sites, the process of homogenous DPR creation is greatly eased.

In conclusion, our results strongly suggest that a quantitative human–machine interactive process is needed for the robust and reproducible QC of digital pathology slides. Quantitative QC not only substantially improved overall concordance but enabled the identification of batch effects in our digital pathology cohort. As we leverage DPRs more fully for the creation of WSI-based tools and biomarkers, having pristine input data will be critical for both the development and deployment of these applications. Only through a concerted effort of improved laboratory standards, cross-discipline training and collaboration, will a suitable environment be primed for the employment of novel precision medicine tools.

Acknowledgements

Research reported in this publication was supported by: The National Cancer Institute of the National Institutes of Health under award numbers: 1U24CA199374-01, R01CA202752-01A1, R01CA208236-01A1, R01 CA216579-01A1, R01 CA220581-01A1, 1U01 CA239055-01, 1U01CA248226-01; National Institute for Biomedical Imaging and Bioengineering, 1R43EB028736-01; National Center for Research

Resources under award number 1 C06 RR12463-01; VA Merit Review Award IBX004121A from the United States Department of Veterans Affairs Biomedical Laboratory Research and Development Service; The DoD Breast Cancer Research Program Breakthrough Level 1 Award W81XWH-19-1-0668; The DOD Prostate Cancer Idea Development Award (W81XWH-15-1-0558); The DOD Lung Cancer Investigator-Initiated Translational Research Award (W81XWH-18-1-0440); The DOD Peer Reviewed Cancer Research Program (W81XWH-16-1-0329); The Kidney Precision Medicine Project (KPMP) Glue Grant; 5T32DK747033 CWRU Nephrology Training Grant; Neptune Career Development Award; 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 (CTSA) at Case Western Reserve University; KidneyCure (ASN Foundation). National Heart, Lung and Blood Institute 1R01HL15127701A1.

Author contributions statement

YC, JZ, AM, LB and AJ conceived the study design and experiments. YC and JZ carried out experiments and analysed data. All authors carried out experiments. All authors were involved in writing the paper and had final approval of the submitted and published versions.

Members of the Nephrotic Syndrome Study Network (NEPTUNE)

NEPTUNE Enrolling Centers

Cleveland Clinic, Cleveland, OH: K Dell*, J Sedor**, M Schachere#, J Negrey#
Children's Hospital, Los Angeles, CA: K Lemley*, E Lim#
Children's Mercy Hospital, Kansas City, MO: T Srivastava*, A Garrett#
Cohen Children's Hospital, New Hyde Park, NY: C Sethna*, K Laurent #
Columbia University, New York, NY: P Canetta*, A Pradhan#
Emory University, Atlanta, GA: L Greenbaum*, C Wang**, C Kang#
Harbor-University of California Los Angeles Medical Center: S Adler*, J LaPage#
John H. Stroger Jr. Hospital of Cook County, Chicago, IL: A Athavale*, M Itteera
Johns Hopkins Medicine, Baltimore, MD: M Atkinson*, S Boynton#
Mayo Clinic, Rochester, MN: F Fervenza*, M Hogan**, J Lieske*, V Chernitskiy#
Montefiore Medical Center, Bronx, NY: F Kaskel*, M Ross*, P Flynn#
NIDDK Intramural, Bethesda MD: J Kopp*, J Blake#
New York University Medical Center, New York, NY: H Trachtman*, O Zhdanova**, F Modersitzki#, S Vento#
Stanford University, Stanford, CA: R Lafayette*, K Mehta#
Temple University, Philadelphia, PA: C Gadegbeku*, S Quinn-Boyle#
University Health Network Toronto: M Hladunewich**, H Reich**, P Ling#, M Romano#
University of Miami, Miami, FL: A Fornoni*, C Bidot#
University of Michigan, Ann Arbor, MI: M Kretzler*, D Gipson*, A Williams#, J LaVigne#
University of North Carolina, Chapel Hill, NC: V Derebail*, K Gibson*, E Cole#, J Ormond-Foster#
University of Pennsylvania, Philadelphia, PA: L Holzman*, K Meyers**, K Kallem#, A Swenson#
University of Texas Southwestern, Dallas, TX: K Sambandam*, Z Wang#, M Rogers#
University of Washington, Seattle, WA: A Jefferson*, S Hingorani**, K Tuttle**§, M Bray #, M Kelton#, A Cooper#§
Wake Forest University Baptist Health, Winston-Salem, NC: JJ Lin*, Stefanie Baker#

Data Analysis and Coordinating Center

M Kretzler, L Barisoni, J Bixler, H Desmond, S Eddy, D Fermin, C Gadegbeku, B Gillespie, D Gipson, L Holzman, V Kurtz, M Larkina, J Lavigne, S Li, S Li, CC Lienczewski, J Liu, T Mainieri, L Mariani, M Sampson, J Sedor, A Smith, A Williams, J Zee.

Digital Pathology Committee

C Avila-Casado (University Health Network, Toronto), S Bagnasco (Johns Hopkins University), J Gaut (Washington University in St Louis), S Hewitt (National Cancer Institute), J Hodgin (University of Michigan), K Lemley (Children's Hospital of Los Angeles), L Mariani (University of Michigan), M Palmer (University of Pennsylvania), A Rosenberg (Johns Hopkins University), V Royal (University of Montreal), D Thomas (University of Miami), J Zee (University of Pennsylvania). Co-Chairs: L Barisoni (Duke University) and C Nast (Cedar Sinai).

*Principal Investigator; **Co-investigator; #Study Coordinator
§Providence Medical Research Center, Spokane, WA

References

1. Barisoni L, Gimpel C, Kain R, *et al*. Digital pathology imaging as a novel platform for standardization and globalization of quantitative nephropathology. *Clin Kidney J* 2017; **10:** 176–187.
2. Kiser PK, Löhr CV, Meritet D, *et al*. Histologic processing artifacts and inter-pathologist variation in measurement of inked margins of canine mast cell tumors. *J Vet Diagn Invest* 2018; **30:** 377–385.
3. Bhargava R, Madabhushi A. Emerging themes in image informatics and molecular analysis for digital pathology. *Annu Rev Biomed Eng* 2016; **18:** 387–412.

4. Madabhushi A, Lee G. Image analysis and machine learning in digital pathology: challenges and opportunities. *Med Image Anal* 2016; **33:** 170–175.
5. Sharma V, Sreedhar CM, Debnath J. Combat radiology: challenges and opportunities. *Med J Armed Forces India* 2017; **73:** 410–413.
6. Janowczyk A, Zuo R, Gilmore H, *et al*. HistoQC: an open-source quality control tool for digital pathology slides. *JCO Clin Cancer Inform* 2019; **3:** 1–7.
7. Fei T, Zhang T, Shi W, *et al*. Mitigating the adverse impact of batch effects in sample pattern detection. *Bioinformatics* 2018; **34:** 2634–2641.
8. Gadegbeku CA, Gipson DS, Holzman LB, *et al*. Design of the Nephrotic Syndrome Study Network (NEPTUNE) to evaluate primary glomerular nephropathy by a multidisciplinary approach. *Kidney Int* 2013; **83:** 749–756.
9. Barisoni L, Nast CC, Jennette JC, *et al*. Digital pathology evaluation in the multicenter Nephrotic Syndrome Study Network (NEPTUNE). *Clin J Am Soc Nephrol* 2013; **8:** 1449–1459.
10. Zee J, Hodgin JB, Mariani LH, *et al*. Reproducibility and feasibility of strategies for morphologic assessment of renal biopsies using the nephrotic syndrome study network digital pathology scoring system. *Arch Pathol Lab Med* 2018; **142:** 613–625.
11. Laurie CC, Doheny KF, Mirel DB, *et al*. Quality control and quality assurance in genotypic data for genome-wide association studies. *Genet Epidemiol* 2010; **34:** 591–602.
12. Turner S, Armstrong LL, Bradford Y, *et al*. Quality control procedures for genome wide association studies. *Curr Protoc Hum Genet* 2011; Chapter 1:Unit1.19.
13. Bielow C, Mastrobuoni G, Kempa S. Proteomics quality control: quality control software for MaxQuant results. *J Proteome Res* 2016; **15:** 777–787.
14. GroundAI. MRQy: an open-source tool for quality control of MR imaging data. [Accessed 8 July 2020]. Available from: https://www.groundai.com/project/mrqy-an-open-source-tool-for-quality-control-of-mr-imaging-data/2
15. Ewels P, Magnusson M, Lundin S, *et al*. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* 2016; **32:** 3047–3048.
16. Hosseini MS, Brawley-Hayes JAZ, Zhang Y, *et al*. Focus quality assessment of high-throughput whole slide imaging in digital pathology. *IEEE Trans Med Imaging* 2020; **39:** 62–74.
17. Ameisen D, Deroulers C, Perrier V, *et al*. Stack or trash? Quality assessment of virtual slides. *Diagn Pathol* 2013; **8:** S23.
18. Avanaki ARN, Espig KS, Xthona A, *et al*. Automatic image quality assessment for digital pathology. In *Breast Imaging* Lecture Notes in Computer Science, Tingberg A, Lång K, Timberg P (eds). Springer: New York, 2016; 431–438.
19. 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* 2014; **9** (suppl 1): S3.

SUPPLEMENTARY MATERIAL ONLINE

Section I. Manual WSI scoring protocol

Section II. Borderline cases in quality assessment

Section III. Feature (predictor) importance for batch effect identification in PAS-stained WSIs of NEPTUNE DPR

Figure S1. Example scoring form

Figure S2. Example WSI as it appears on your screen when you click on the 'link to image'

Appendix A. Examples of artifacts