PLOS ONE

RESEARCH ARTICLE

Pixelwise H-score: A novel digital image analysis-based metric to quantify membrane biomarker expression from immunohistochemistry images

Sripad Ram[1]*, Pamela Vizcarra[2], Pamela Whalen[2], Shibing Deng[3], C. L. Painter[2], Amy Jackson-Fisher[2], Steven Pirie-Shepherd[2], Xiaoling Xia[2¤], Eric L. Powell[2]

1 Drug-Safety Research and Development, Pfizer Inc., San Diego, California, United States of America, **2** Tumor Morphology Group, Oncology Research and Development, Pfizer Inc., San Diego, California, United States of America, **3** Biostatistics Unit, Oncology Research and Development, Pfizer Inc., San Diego, California, United States of America

¤ Current address: Ventana Medical Systems, Tucson, Arizona, United States of America
* Sripad.ram@pfizer.com

OPEN ACCESS

Citation: Ram S, Vizcarra P, Whalen P, Deng S, Painter CL, Jackson-Fisher A, et al. (2021) Pixelwise H-score: A novel digital image analysis-based metric to quantify membrane biomarker expression from immunohistochemistry images. PLoS ONE 16(9): e0245638. https://doi.org/10.1371/journal.pone.0245638

Editor: Konradin Metze, University of Campinas, BRAZIL

Received: December 28, 2020

Accepted: September 2, 2021

Published: September 27, 2021

Data Availability Statement: Whole slide images for PD-L1, P-cadherin and 5T4 are available from the Zenodo repository: https://zenodo.org/record/4900103#.YMkMR6hKiUk. All other relevant data are within the manuscript and its Supporting information files.

Funding: This work was funded by Pfizer Inc. The funder provided support in the form of salaries for authors and research materials but did not have any additional role in the study design, data

Abstract

Immunohistochemistry (IHC) assays play a central role in evaluating biomarker expression in tissue sections for diagnostic and research applications. Manual scoring of IHC images, which is the current standard of practice, is known to have several shortcomings in terms of reproducibility and scalability to large scale studies. Here, by using a digital image analysis-based approach, we introduce a new metric called the pixelwise H-score (pix H-score) that quantifies biomarker expression from whole-slide scanned IHC images. The pix H-score is an unsupervised algorithm that only requires the specification of intensity thresholds for the biomarker and the nuclear-counterstain channels. We present the detailed implementation of the pix H-score in two different whole-slide image analysis software packages Visiopharm and HALO. We consider three biomarkers P-cadherin, PD-L1, and 5T4, and show how the pix H-score exhibits tight concordance to multiple orthogonal measurements of biomarker abundance such as the biomarker mRNA transcript and the pathologist H-score. We also compare the pix H-score to existing automated image analysis algorithms and demonstrate that the pix H-score provides either comparable or significantly better performance over these methodologies. We also present results of an empirical resampling approach to assess the performance of the pix H-score in estimating biomarker abundance from select regions within the tumor tissue relative to the whole tumor resection. We anticipate that the new metric will be broadly applicable to quantify biomarker expression from a wide variety of IHC images. Moreover, these results underscore the benefit of digital image analysis-based approaches which offer an objective, reproducible, and highly scalable strategy to quantitatively analyze IHC images.

collection and analysis, decision to publish, or preparation of the manuscript. The specific roles of these authors are articulated in the 'author contributions' section".

Competing interests: SR, PV, PW, SD, CLP, AJF, SPS, XX and ELP were all employees of Pfizer when the work reported in this manuscript was conducted. XX is now an employee of Ventana Medical Systems. We can confirm that this does not alter our adherence to PLOS One policies on sharing data and materials".

Introduction

Immunohistochemistry (IHC) is a core technology that is used to evaluate the spatial distribution and abundance of biomarkers at the protein level in tissue samples. In oncology clinical diagnosis and research applications, IHC assays play a central role in tumor characterization and biomarker assessment. Typically, IHC images are qualitatively evaluated by a trained expert, such as a pathologist, and in some cases this is complemented by a semi-quantitative score [1]. However, visual quantitative scoring of IHC images is not routinely performed due to several shortcomings. On the one hand, visual quantitative scoring is time consuming and is often not feasible to perform on a routine basis especially for large studies. On the other hand, visual quantitative scores are subjective and often have a limited dynamic range due to their categorical nature (e.g. manual scores of 0, 1+, 2+, and 3+). Consequently, they may not have the granularity to adequately capture biomarker expression from an IHC slide [2, 3]. The subjectivity of the scoring process, in turn, can manifest as poor inter- and intra-observer concordance, and this has been the subject of numerous studies [4–8]. While concordance in visual quantitative scoring can be improved by the development of standardized scoring guidelines and extensive training [9, 10], the labor-intensive aspect and the limited dynamic range still remain as major impediments to the widespread use of visual quantitative scoring of IHC images.

Digital image analysis (DIA) based tools overcome some of these limitations of visual quantitative scoring by enabling fast, objective, and highly reproducible quantification of biomarkers from whole-slide IHC images [1, 11]. DIA endpoints are typically continuous variables (e.g. cell density and % positive cells) and offer adequate dynamic range to represent biomarker expression in the IHC image. One of the widely used endpoints to quantify biomarker expression is the H-score [2, 12]. In the H-score algorithm (Fig 1A) individual cells and their sub-cellular compartments (i.e. nucleus, cytoplasm, and cell membrane) are first detected, and based on the relative expression of the biomarker of interest in one or more sub-cellular compartments the cells are classified as either positive or negative. The positive cells are further classified into high (3+), medium (2+), or low (1+) based on the biomarker signal intensity. The H-score is given by the ratio of the weighted sum of the number of positive cells to the total number of detected cells. The H-score captures both the intensity and the proportion of the biomarker of interest from the IHC image and comprises values between 0 and 300, thereby offering a dynamic range to quantify biomarker abundance. A different scoring method developed to quantify estrogen and progesterone receptors in breast cancers, the Allred score [2, 12], assigns separate categorical scores for the intensity (0–3) and the proportion (0–5) of the biomarkers in immunolabeled specimens, and the final score is the sum of these two scores. Compared to the H-score, the Allred score has a limited dynamic range (0–8) and is not extensively used for purposes other than ER/PR quantification in breast cancer. From a digital image analysis standpoint, both the H-score and the Allred score require the detection of individual cells, and this requires robust nucleus and cell segmentation algorithms for individual nucleus detection and delineation of individual cell boundaries.

Another scoring methodology, the average threshold method (ATM), adopts a pixelwise approach for quantifying biomarker abundance [13]. The ATM score does not require the detection of individual nuclei or cells and is solely based on the pixel intensities of the DAB chromogen in the spectrally deconvolved image. Consequently, the calculation of the ATM score is relatively straightforward but at the expense of decreased dynamic range as compared to the H-score.

The AQUA score [14] also makes use of a pixelwise strategy for quantifying biomarker expression. Here, the tissue is fluorescently labeled for the biomarker of interest along with a

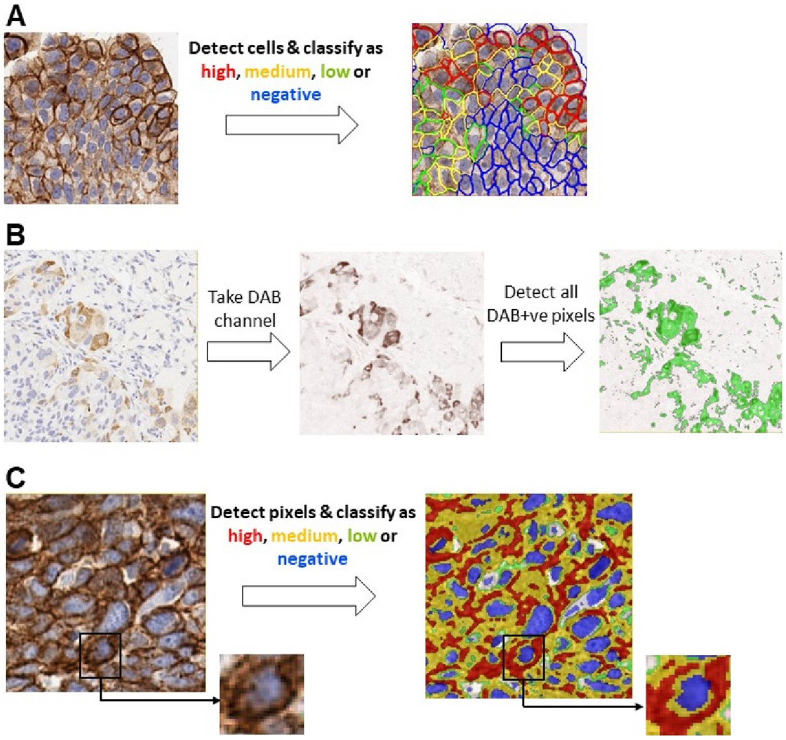
A
Detect cells & classify as high, medium, low or negative
B
Take DAB channel
Detect all DAB+ve pixels
C
Detect pixels & classify as high, medium, low or negative

Fig 1. Overview of the different scoring algorithms. Panel A shows the traditional cell-based H-score, panel B shows the average threshold method (ATM) score, and panel C shows the pix H-score.

https://doi.org/10.1371/journal.pone.0245638.g001

nuclear stain and a cell membrane marker. This in turn allows the generation of pixel masks pertaining to different subcellular compartments (e.g., cell membrane, nucleus, or cytosolic mask). The AQUA score is then calculated by taking the total fluorescence signal of the biomarker of interest for a given subcellular mask (e.g. the cell-membrane mask) and normalizing it by the total area of the mask [14]. The advantage of the AQUA score is that it offers a broad dynamic range. However, the calculation of the AQUA score requires the development of a fluorescence-based multiplex assay which can be time consuming and technically challenging. Moreover, the use of fluorescence readout masks anatomic and morphological information (e.g. necrotic regions, stroma, etc.) that are readily detectable from a brightfield IHC image.

In this manuscript, three different scoring methods are compared, which are illustrated in Fig 1. We introduce a new DIA method, the pixelwise H-score (pix H-score), for quantifying biomarker abundance from brightfield IHC images by making use of individual pixel intensities in DAB and hematoxylin channels and leveraging weighted intensity averages. Our motivation behind developing the pix H-score is to create a simple, yet robust metric to accurately quantify biomarker expression without relying on the detection and delineation of individual cells and their sub-cellular compartments. The latter makes the implementation of the pix H-score to be relatively straightforward. The pix H-score can be thought of as an equivalent of the traditional H-score that is applied to pixels rather than to cells. The pix H-score takes values between 0 and 300 thereby providing a dynamic range similar to that of the H-score.

We evaluated the performance of pix H-score using IHC images of three different membrane biomarkers P-cadherin, PD-L1, and 5T4. For comparison, we also calculated the ATM score and the DIA H-score for these images, where the latter is a DIA implementation of the traditional H-score. Using the pathologist H-score and biomarker mRNA transcript level (measured using qRT-PCR or NanoString analysis of mRNA in adjacent serial sections) as orthogonal measurements of biomarker abundance, we demonstrate that the pix H-score is either comparable or superior to other DIA endpoints in quantifying biomarker abundance in IHC images. We present the detailed implementation of the pix H-score in two commercial, whole-slide, image analysis software packages, Visiopharm and HALO. We also present an empirical resampling approach to quantitatively assess the ability of the pix H-score to estimate biomarker abundance when it is calculated from select regions within the tumor resection when compared to the whole slide pix H-score. We note that a subset of the results reported here was previously disclosed in a scientific poster at the 34th annual meeting of the Society for Immunotherapy of Cancer [15]. We anticipate that the new metric will have broad applicability and pave the way towards establishing an objective, reproducible strategy to quantify biomarker abundance in IHC images.

Materials and methods

Previously-developed IHC assays for P-cadherin, PD-L1, and 5T4 were used to immunolabel three cohorts of human tumors. Serial sections from these cohorts were also evaluated for target mRNA via NanoString (P-cadherin and PD-L1) or qRT-PCR (5T4). Following H-scoring of the immunolabeled tumor sections by a pathologist, the concordance between the H-score and mRNA values was evaluated by Spearman correlation. To automate the scoring process through digital image analysis, we implemented several DIA strategies using different software tools. Specifically, we implemented digital H-scoring using QuPath and HALO software packages, the ATM score using Visiopharm software, and the pix H-score, the new digital scoring method, using HALO and Visopharm software packages. To assess the performance of the various DIA algorithms, we calculated the Spearman's correlation coefficient between each DIA endpoint and two different measurements of biomarker abundance, i.e. the pathologist H-score and the target transcript level as assessed using either NanoString technology or qRT-PCR.

Immunohistochemistry

All human tissue biospecimens used in the study were anonymized specimens that were acquired by Pfizer from multiple collaboration partners and commercial vendors. These specimens were acquired and used in compliance with Pfizer's policy on the Use of Human Biological Specimens [16]. Specifically, these biospecimens were collected with written patient consent, processed, and distributed in full ethical and regulatory compliance with the sites from which they were collected. This includes independent ethical review, Institutional Review Board approval (where appropriate), and independent regulatory review. Head and neck tumor resections were procured from Flagship Biosciences (Broomfield, CO), and lung cancer resections were procured from the following vendors: Indivumed (Hamburg, Germany), ProteogeneX (Inglewood, CA), Weill Medical College at Cornell (New York, NY) and University of Michigan (Ann Arbor, MI).

For PD-L1, we used twenty-four cases of routinely collected non-small cell lung carcinoma surgical resections. The SP142 clone of anti-PD-L1 antibody was used as per the manufacturer-recommended protocol. For P-cadherin, we used thirty cases of routinely collected head and neck tumor resections. The P-cadherin IHC assay was developed and optimized on the Dako Autostainer system using a custom anti-P-cadherin antibody that was generated as an analyte specific reagent for use in a clinical diagnostic assay. For 5T4, we used twenty-one cases of routinely collected non-small cell lung tumor resections. The development and validation of the 5T4 IHC assay was reported previously [17]. In all three IHC assays hematoxylin was used as the nuclear counterstain and diaminobenzidine (DAB) was the chromogen that was used to detect the biomarker of interest. P-cadherin and PD-L1 slides were scanned using a Leica Aperio AT2 whole-slide scanner at 20x magnification, whereas 5T4 slides were scanned using a Hamamatsu Nanozoomer whole-slide scanner at 20x magnification. All IHC images were subjected to visual quality assessment which verified that the data was devoid of out of focus artefacts, gross variation in background level due to white-balancing errors, and significant variation in hematoxylin staining among the images for a given biomarker.

NanoString assay

Messenger RNA (mRNA) was isolated from two 4-micron FFPE slide sections using FormaPure® nucleic acid isolation kit according to manufacturer's instructions with the addition of a DNA digestion step. NanoString technology was used to measure RNA transcript levels using the nCounter assay according to manufacturer's recommended protocols. Custom nCounter CodeSet containing either the CDH3 probe (for P-cadherin) or the CD274 probe (for PD-L1) was used. One hundred nanograms of total RNA was hybridized to the custom panel for 16 to 20 hours at 65˚C. Samples were processed using an automated nCounter sample prep station. Cartridges containing immobilized and aligned reporter complex were subsequently imaged and counted on an nCounter Digital Analyzer set for maximum fields of view. Reporter counts were analyzed and normalized using NanoString nSolver Analysis Software. Briefly, raw counts were multiplied by scaling factors proportional to the sum of counts for spiked in positive control probes to account for individual assay efficiency variation, and to the geometric average of the housekeeping gene probes to account for variability in the mRNA content. FFPE sample sets were normalized to the following housekeeping genes; for P-cadherin: FTL, GAPDH, GUSB, HMBS, HPRT1, OAZ1, PCBP1, PFN1, PPIA, PSAP and TBP; and for PD-L1: AMMECR1L, CNOT10, CNOT4, COG7, DDX50, EDC3, EIF2B4, ERCC3, FCF1, FTL, GPATCH3, GUSB, HDAC3, HPRT1, MTMR14, PPIA, SAP130, TBP, TMUB2, and ZNF143.

qRT-PCR assay

The qRT-PCR reaction was performed using the TaqMan Probe-Based Gene Expression Analysis and ABI ViiA7 Real-Time PCR Systems (Life Technologies) as described previously [17]. Target gene and endogenous controls were run in quadruplicate for each probe set on prefabricated TaqMan low density array cards. For each tumor sample 1000 ng of cDNA was diluted to 55 uL with nuclease-free water and 55 uL of TaqMan gene expression master mix was added (Life Technologies, cat # 4352042). A total of 100 uL of sample was added to each of the 8 ports on a single card, after which the plate was sealed and centrifuged two times in Sorvall/Heraeus buckets based on manufacturer's directions. TaqMan array cards were then sealed and loaded into the ABI ViiA7 thermal cycler and run. Default thermal cycling conditions were as follows; the RT-PCR reaction was run on the thermal cycler in three stages; 2 minutes at 50˚C, 10 minutes at 90˚C and 40 cycles of 15 seconds at 90˚C followed by 1 minute at 60˚C.

ExpressionSuite Software v1.0.3 (Life Technologies) was used to generate automated threshold values for signal amplification for a majority of samples. Rarely were automated thresholds adjusted manually. Amplification plots resulting in Ct values >35 were discarded, as were those plots that generated a Ct value but did not display a trend of logarithmic amplification. All Ct values were exported from the ExpressionSuite software and relative quantification calculations were performed in Microsoft Excel 2010.

Digital image analysis

IHC images of P-cadherin, PD-L1, and 5T4 were analyzed at 20x magnification using multiple software packages. The detailed implementation in each software package is described below. Briefly, the traditional cell-based H-score was implemented in HALO (Version 2.3) and QuPath (Version 0.2.0-m2) and was calculated based on the cell-membrane localized biomarker signal. The ATM score was implemented in Visiopharm (Version 2017.7.3.469) and the pix H-score was implemented in Visiopharm and HALO. For each biomarker, the results of the DIA algorithm for every image along with the orthogonal measurements of biomarker abundance (path H-score an mRNA transcript) are provided in S1 Table.

HALO implementation of H-score (H-score (HALO))

The Membrane module (v1.4) in HALO was used to detect cells and calculate the H-score. The algorithm first deconvolves the IHC image into hematoxylin and DAB channels, then detects individual cells and their subcellular compartments, i.e. nucleus and cell membrane, in the image, and scores the cells as high, medium, and low based on the average DAB signal associated with the cell membrane. The thresholds for high, medium, and low were determined separately for each biomarker by examining the membrane-associated DAB signal across multiple images pertaining to that biomarker. A separate algorithm was implemented for each biomarker in order to optimize the detection and segmentation of the nucleus and cell membrane specific to that biomarker. The App outputs the number of negative, high, medium and low cells, which is then used to calculate the H-score that is given by [18–20]

$$\mathrm{H}-\mathrm{score}=100\frac{3H+2M+L}{H+N+L+N}. \tag{1}$$

In the above equation, H M, L and N denote the number of high, medium, low and negative cells, respectively. The H-score quantifies biomarker expression by taking in account the proportion and the intensity of the biomarker in positive cells. Specifically, the numerator in Eq 1 considers the proportion of positive cells and the weighting factors, i.e., 3, 2 and 1 for high, medium and low cells, respectively, account for the intensities of the positive cells. It should be pointed out that the choice of weighting factors is empirical and does not always imply a linear relationship (i.e., intensity of medium cells is not always equal to two times the intensity of low cells) [2].

QuPath implementation of H-score (H-score (QuPath))

QuPath (verion 0.2.0-m2) is an open-source software for whole-slide image analysis of histopathology data [21]. A custom script was written in the Groovy programming language to detect cells and score them as high, medium, and low based on the average DAB signal in the cell membrane (see Supporting information). The script first deconvolves the IHC image into hematoxylin and DAB channels. A watershed-based cell and membrane detection algorithm (Analyze -> Cell Analysis -> Cell + membrane detection) was used to detect individual cells and identify their subcellular compartments, i.e. nucleus and cell membrane. The cell detection algorithm includes a pre-processing step that involves a local background subtraction by using the minimum filter. The optional median filtering step was not used. Cells that were devoid of a nucleus (due to weak or missing hematoxylin staining) were excluded and the remaining cells were scored as high, medium, and low based on the mean DAB signal associated with the membrane compartment. The thresholds for high, medium, and low were determined separately for each biomarker. A separate script was implemented for each biomarker in order to optimize the detection and segmentation of the nucleus and cell membrane specific to that biomarker. The script outputs the total number detected cells along with the number of high, medium, and low cells, which is then used to calculate the H-score that is given in Eq 1.

ATM score

The motivation behind the ATM score is discussed elsewhere [13]. Briefly, the idea is to use all the intensity values in the DAB channel so that the final metric is independent of the choice of thresholds. Further, the ATM score is a pixel-based metric that does not depend on the detection of individual cells and/or its subcellular components. Assuming 8-bit resolution for the color-deconvolved biomarker channel, the ATM score is given by (see Ref. [13] for derivation)

$$\begin{aligned}\text{ATM score} &= \frac{1}{255}\sum\nolimits_{k=1}^{255} PS(k)\\ &= \frac{1}{255}\,(\text{average value of all the pixels in the DAB channel}), \quad (2)\end{aligned}$$

where PS(k) is the proportion score which denotes the proportion of pixels with intensity greater than or equal to k, where k takes values from 1 to 255 (i.e. 2^8–1 grey levels). From the above equation, we see that the ATM score is a weighted average of all the pixels in the DAB channel. The ATM score was implemented in Visiopharm software. The IHC image was color deconvolved into hematoxylin and DAB channels. Therefore, the ATM score was calculated by taking the average intensity of all DAB positive pixels and then dividing this by 255 (see Supporting information for the App).

Visiopharm implementation of pix H-score (pix H-score (VIS))

A threshold-based detection App was used to implement the pix H-score in Visiopharm (see Supporting information). The App does not require any specific add-on module and was implemented using the default functionality of the software. The App first deconvolves the IHC image into hematoxylin and DAB channels. The App then detects and classifies DAB positive pixels as high, medium, and low, and then detects the hematoxylin positive pixels. The thresholds for DAB and hematoxylin were separately selected for each biomarker. The App then outputs the total area of the DAB high, DAB medium, and DAB low pixels and the hematoxylin positive pixels. These values are then used to calculate the pix H-score which is given by

$$\text{pix H} - \text{score} = 100\frac{3H_P + 2M_P + L_P}{H_P + M_P + L_P + N_P}, \quad (3)$$

where H_p, M_p, L_p and N_p denote the area of DAB high, DAB medium, DAB low and hematoxylin positive pixels, respectively. The pix H-score is analogous to the traditional cell-based H-score (Eq 1) but is applied to pixels as opposed to individual cells. Specifically, it is a weighted sum of the relative proportion of the DAB high, DAB medium and DAB low pixels, where we have used the same weighting factors (i.e., 3 for DAB high, 2 for DAB medium and 1 for DAB low) as that of the traditional H-score. This ensures that the pix H-score takes the same range of values as that of the traditional H-score (i.e. 0–300).

In Visiopharm, the output of the intensity-based thresholding algorithm depends on the order in which the different color-deconvolved channels are used. For instance, if a pixel contains both hematoxylin and DAB signal that are above their respective threshold values for positivity and the DAB channel is first analyzed followed by the hematoxylin channel, then that pixel will be labeled as positive only for the DAB channel. In other words, if a pixel is found to be positive for one of the color-deconvolved channels then it is excluded from any subsequent classification for the other color-deconvolved channels.

HALO implementation of pix H-score (pix H-score (HALO))

The area quantification module (v2.1.3), which is a default module available as part of the basic HALO software package, was used to calculate the pix H-score with the number of phenotypes set to 1 (see Supporting information for the settings files). The algorithm deconvolves the IHC image into hematoxylin and DAB channels and can detect and classify hematoxylin and DAB positive pixels as high, medium, and low based on a user defined threshold. For the calculation of pix H-score, a single threshold was used to detect all hematoxylin positive pixels and three separate thresholds were used to detect and classify the DAB positive pixels. In HALO, these thresholds take values between 0 and 1. In order to keep the thresholds implemented in Visiopharm and HALO identical, the threshold values used in Visiopharm, which take values between 1–255, were rescaled to take values between 0 and 1 and these were then used in HALO. Unlike Visiopharm, HALO keeps track of the detected pixels in the DAB and hematoxylin channels separately. Consequently, pixels that contain both DAB and hematoxylin signal that are above the thresholds will be accounted for in both the hematoxylin and DAB channels. In order to mimic the Visiopharm implementation of pix H-score, we define a third channel, which is denoted as phenotype 1 channel in HALO that pertains to pixels that are positive for hematoxylin but negative for DAB. This phenotype 1 channel will contain pixels that are analogous to the hematoxylin positive pixels detected in the Visiopharm implementation of pix H-score algorithm. The algorithm outputs the area of high, medium, and low pixels in the DAB channel, and the area of positive pixels in the phenotype 1 channel, which is used as an estimate of the total area of pixels containing only the hematoxylin signal. These values are then used in Eq 3 to calculate the pix H-score.

Statistical analysis

Spearman's rank correlation coefficient was calculated to assess the correlation between different pairs of variables of interest. Our choice of correlation analysis was based on the nature of relationship between the variables of interest. Unlike Pearson's correlation, Spearman's correlation analysis is a rank-based metric that provides a robust estimate of correlation when there is a non-linear, monotone relationship between the variables of interest [22], which is typically the case for the different pairs of variables considered in this manuscript. The William's t test was used to test for significant difference between a pair of dependent correlation coefficients [23, 24].

Spatial resampling analysis

For each biomarker, an empirical resampling procedure was performed on every whole-slide IHC image. The viable tissue region was sampled by non-overlapping circular regions of radius 0.8 mm (Fig 6A). For each region, the area of DAB high, DAB medium, DAB low, and hematoxylin positive pixels were determined using Visiopharm. The results were exported to MATLAB (Mathworks, Natick, MA) for subsequent analysis. For every IHC image, N different circular regions were randomly selected (N = 1–50), and a regional pix H-score was calculated using the area of DAB high pixels, DAB medium pixels, DAB low pixels, and hematoxylin positive pixels that were summed from the N circular regions. This procedure is repeated N_{iter} times with replacement ($N_{iter} = 100$ for all the biomarkers). Then for each iteration k = 1,. . ., N_{iter}, the Spearman correlation coefficient C(N,k) is computed between the regional pix H-score and the corresponding pathologist H-scores (or the corresponding mRNA levels). The average Spearman correlation coefficient for each value of N is computed using the formula

$$C_{av}(N) = \frac{1}{N_{iter}} \sum_{k=1}^{N_{iter}} C(N, k).$$

Results

DIA algorithms for P-cadherin quantification

IHC images for P-cadherin (Fig 2A) showed strong immunoreactivity at the cell membrane and in the cytoplasm, which was consistent with prior reports [25, 26]. Spearman's correlation analysis of the membrane H-scores of the 30 cases immunolabeled for P-cadherin, as assessed by a board-certified pathologist (see S1 Table), and NanoString nCounter values for P-cadherin mRNA transcript from serial sections of the same cases had a correlation coefficient of 0.81, $p<0.0001$ (Fig 2B). Throughout this manuscript, we have used Spearman's correlation analysis as it is a more appropriate measure of correlation when the variables of interest exhibit a non-linear, monotone relationship [22] which is typically the case in our data.

We next investigated whether the differences in the Spearman correlation coefficients for the various DIA endpoints are statistically significant. Table 1 shows the results of our statistical analysis where we carried out pairwise comparisons of the correlation coefficients for different DIA endpoints obtained from P-cadherin IHC images. Our analysis shows that the correlation coefficient between the pix H-score and either of the biomarker abundance endpoints (pathologist H-score and P-cadherin transcript) is significantly higher than the correlation coefficient between DIA based H-scores and biomarker abundance endpoints. This

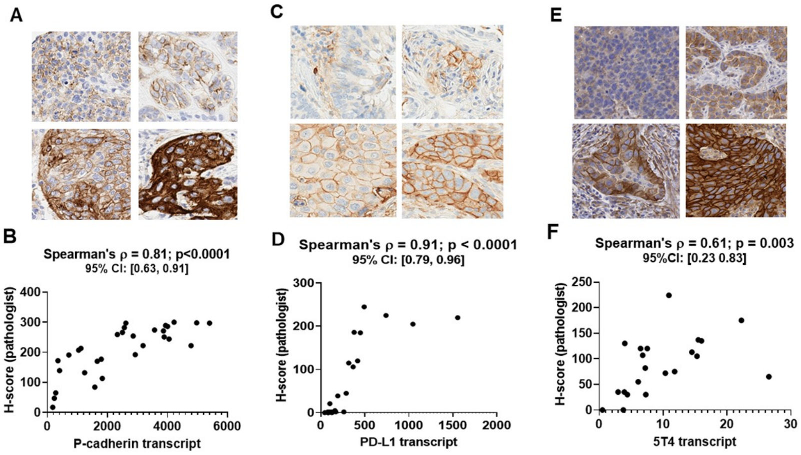

Fig 2. P-cadherin, PD-L1 and 5T4 IHC datasets. Panels A, C and E show representative images at 20x magnification with varying levels of P-cadherin, PD-L1 and 5T4 expression, respectively, in tumor resections. Panels B, D and F show the plot of the pathologist H-score versus mRNA transcript level for P-cadherin (n = 30 cases), PD-L1 (n = 24 cases) and 5T4 (n = 21 cases), respectively. The panels also show Spearman's correlation coefficient along with the p-value and 95% confidence interval. When compared to the P-cadherin pathologist H-score, all P-cadherin DIA endpoints (see S1 Table) yielded positive correlations (Fig 3A–3E). The correlation with the ATM score (Fig 3C) and pix H-score (Fig 3D and 3E) were higher than the correlations with the DIA based H-scores (Fig 3A and 3B). More specifically, the Spearman's correlation coefficient for HALO and QuPath DIA H-scores were 0.5 (p = 0.005) and 0.39 (p = 0.03), respectively, whereas the Spearman's correlation coefficient for the ATM score, the VIS pix H-score and the HALO pix H-score were 0.78 ($p<0.001$), 0.77 ($p<0.0001$) and 0.88 ($p<0.0001$), respectively. When compared to the P-cadherin transcript, all DIA endpoints similarly yielded positive correlations (Fig 3F–3J), with the pix H-score exhibiting the highest Spearman's correlation coefficient (Fig 3I and 3J; ρ = 0.83 and ρ = 0.81, respectively, for VIS and HALO pix H-score; $p < 0.0001$) followed by the ATM score (Fig 3H; ρ = 0.62, $p < 0.0001$) and the DIA H-scores (Fig 3F and 3G; ρ = 0.5, p = 0.005 for HALO and ρ = 0.45, p = 0.01 for QuPath).

https://doi.org/10.1371/journal.pone.0245638.g002

Table 1. Table lists the results of William's t-test to test for significant difference in the Spearman correlation coefficients between P-cadherin transcript or pathologist H-score and different DIA endpoints.

Statistical analysis of correlation coefficients for P-cadherin	ρ_{12}	ρ_{13}	z-score	p-value	Result
Comparing pairwise correlations between DIA endpoint and Pathologist H-score					
ρ(Path H-score, H-score HALO) vs ρ(Path H-score, H-score QuPath)	0.50	0.39	1.48	0.15	N.S.D.
ρ(Path H-score, pix H-score VIS) vs ρ(Path Hscore, pix H-score HALO)	0.77	0.88	-2.5	0.01	S.D.
ρ(Path H-score, pix H-score VIS) vs ρ(Path H-score, H-score HALO)	0.77	0.50	3.01	0.005	S.D.
ρ(Path H-score, pix H-score VIS) vs ρ(Path H-score, H-score QuPath)	0.77	0.39	3.87	0.0006	S.D.
ρ(Path H-score, pix H-score VIS) vs ρ(Path H-score, ATM score)	0.77	0.78	-0.12	0.90	N.S.D.
ρ(Path H-score, pix H-score HALO) vs ρ(Path H-score, H-score HALO)	0.88	0.50	5.21	1.7e-05	S.D.
ρ(Path H-score, pix H-score HALO) vs ρ(Path H-score, H-score QuPath)	0.88	0.39	5.34	1.2E-05	S.D.
ρ(Path H-score, pix H-score HALO) vs ρ(Path H-score, ATM score)	0.88	0.78	1.64	0.11	N.S.D.
Comparing pairwise correlations between DIA endpoint and P-cadherin transcript					
ρ(Pcad transcript, H-score HALO) vs ρ(Pcad transcript, H-score QuPath)	0.50	0.45	0.79	0.43	N.S.D.
ρ(Pcad transcript, pix H-score VIS) vs ρ(Pcad transcript, pix H-score HALO)	0.83	0.81	0.50	0.62	N.S.D.
ρ(Pcad transcript, pix H-score VIS) vs ρ(Pcad transcript, H-score HALO)	0.83	0.5	4.19	0.0002	S.D.
ρ(Pcad transcript, pix H-score VIS) vs ρ(Pcad transcript, H-score QuPath)	0.83	0.45	4.49	0.0001	S.D.
ρ(Pcad transcript, pix H-score VIS) vs ρ(Pcad transcript, ATM score)	0.83	0.62	2.45	0.02	S.D.
ρ(Pcad transcript, pix H-score HALO) vs ρ(Pcad transcript, H-score HALO)	0.81	0.5	3.31	0.002	S.D.
ρ(Pcad transcript, pix H-score HALO) vs ρ(Pcad transcript, H-score QuPath)	0.81	0.45	3.22	0.003	S.D.
ρ(Pcad transcript, pix H-score HALO) vs ρ(Pcad transcript, ATM score)	0.81	0.62	2.37	0.025	S.D.

S.D. significant difference, N.S.D–no significant difference.

https://doi.org/10.1371/journal.pone.0245638.t001

suggests that for the P-cadherin dataset, the pix H-score is a better DIA metric to quantify biomarker abundance over traditional DIA based H-score. In the case of the ATM score, we observe a mixed result in that the correlation coefficient between pix H-score and P-cadherin transcript is significantly higher than the correlation coefficient between ATM score and P-

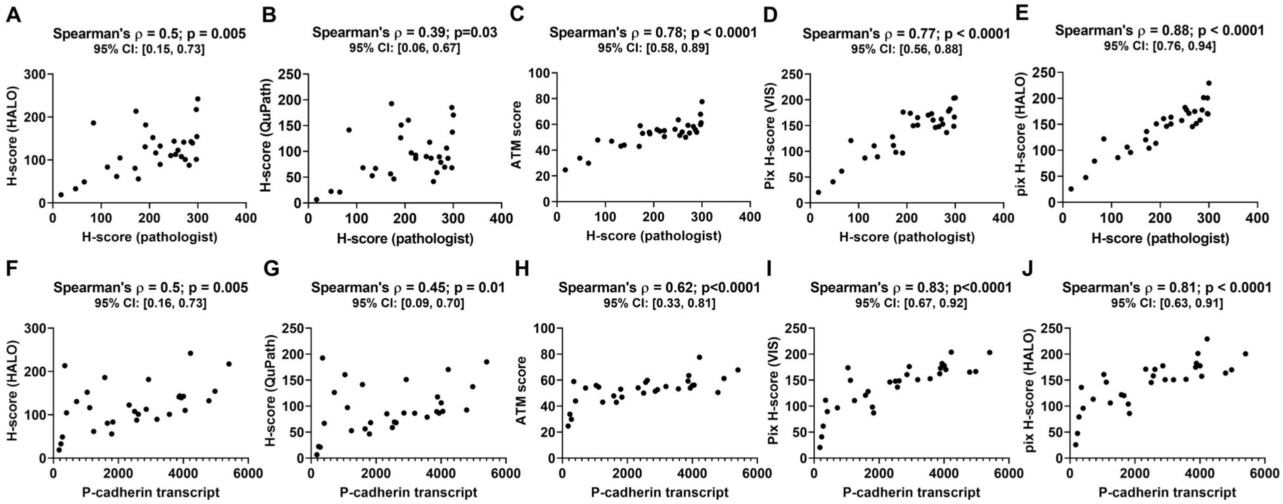

Fig 3. Performance of DIA endpoints obtained from P-cadherin IHC images. Panels A through E show the plots of different DIA endpoints versus pathologist H-score for a cohort of 30 head and neck cancer resections. Panels F through J show the plots of different DIA endpoints versus P-cadherin mRNA transcript for the same 30 cases. Each panel also shows the Spearman's correlation coefficient between the two quantities plotted in that panel along with the p-value and the 95% confidence interval.

https://doi.org/10.1371/journal.pone.0245638.g003

cadherin transcript, whereas statistical significance is lost when we consider the pathologist H-score as the reference for biomarker abundance (Table 1). We also compared the two DIA based H-scores. We found no significant difference in the Spearman's correlation coefficient between QuPath H-score and biomarker abundance endpoints versus HALO H-score and biomarker abundance endpoints (Table 1). Similarly, we found no significant difference in the correlation coefficients for the HALO and VIS implementations of the pix H-score for P-cadherin.

DIA algorithms for PD-L1 quantification

IHC images for PD-L1 (Fig 2C) showed strong immunoreactivity at the cell membrane and minimal to no cytoplasmic staining, which was consistent with prior reports [25, 26]. Spearman's correlation analysis of the membrane H-scores of the 24 cases immunolabeled for PD-L1, as assessed by a board-certified pathologist (see S1 Table), and NanoString nCounter values for PD-L1 mRNA transcript from serial sections of the same cases had a correlation coefficient of 0.91, $p<0.0001$ (Fig 2D).

When compared to the pathologist H-score, all DIA endpoints (see S1 Table) yielded positive correlations (Fig 4A–4E). The Spearman's correlation coefficient for the HALO H-score, QuPath H-score, ATM score, VIS pix H-score and HALO pix H-score with respect to the pathologist H-score were 0.69 ($p = 0.0002$), 0.74 ($p<0.0001$), 0.55 ($p = 0.005$), 0.76 ($p < 0.0001$) and 0.71 ($p < 0.0001$), respectively. When compared to the PD-L1 transcript, all DIA endpoints similarly yielded positive correlations (Fig 4F–4J). The Spearman's correlation coefficient for the HALO H-score, QuPath H-score, ATM score, VIS pix H-score and HALO pix H-score with respect to PD-L1 transcript were 0.73 ($p<0.0001$), 0.75 ($p<0.0001$), 0.55 ($p = 0.005$), 0.79 ($p<0.0001$) and 0.79 ($p<0.0001$), respectively.

Statistical analysis of the Spearman's correlation coefficients revealed that there is no significant difference in the correlation coefficient between DIA based H-scores and PD-L1 biomarker abundance endpoints versus the correlation coefficient between pix H-score and PD-L1 biomarker abundance endpoints (Table 2). This shows that the performance of pix H-score is analogous to that of the DIA based H-score which is in contrast with our observations for P-cadherin. Also, there was no significant difference in Spearman's correlation coefficient

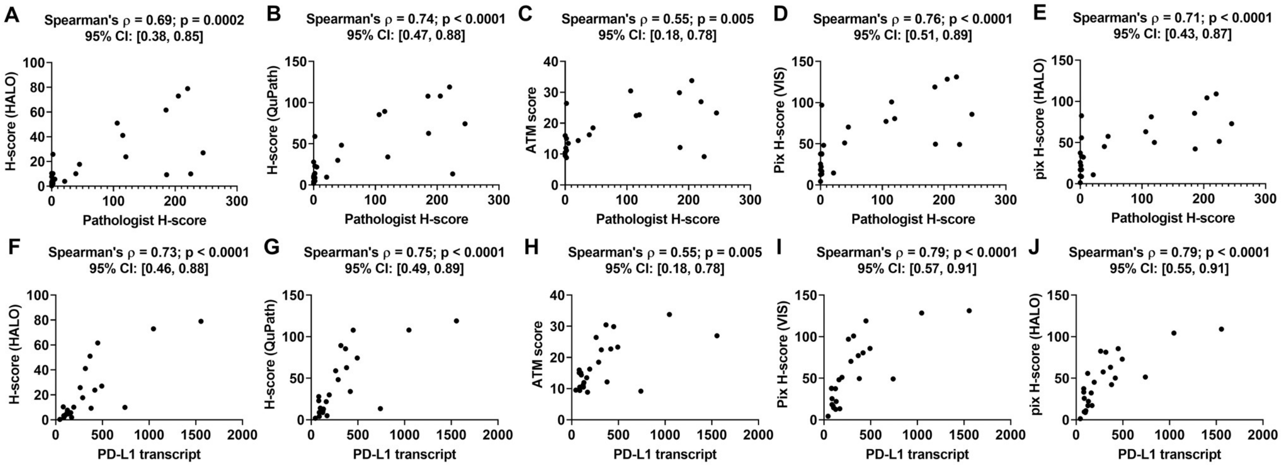

Fig 4. Performance of DIA endpoints obtained from PD-L1 IHC images. Panels A-E show plots of the different DIA endpoints as a function of the pathologist H-score, while panels F-J show the same as a function of PD-L1 mRNA transcript for a cohort of 24 lung cancer resections. All panels show the Spearman's correlation coefficient between the two quantities plotted in that panel along with the p-value and the 95% confidence interval.

https://doi.org/10.1371/journal.pone.0245638.g004

Table 2. Table lists the results of William's t-test to test for significant difference in the Spearman correlation coefficients between PD-L1 mRNA transcript or pathologist H-score and different DIA endpoints.

Statistical analysis of correlation coefficients for PD-L1	ρ_{12}	ρ_{13}	z-score	p-value	Result
Comparing pairwise correlations between DIA endpoint and Pathologist H-score					
ρ(Path H-score, H-score HALO) vs ρ(Path H-score, H-score QuPath)	0.69	0.74	-1.04	0.31	N.S.D.
ρ(Path H-score, pix H-score VIS) vs ρ(Path H-score, pix H-score HALO)	0.76	0.72	0.93	0.36	N.S.D.
ρ(Path H-score, pix H-score VIS) vs ρ(Path H-score, H-score HALO)	0.76	0.69	1.23	0.23	N.S.D.
ρ(Path H-score, pix H-score VIS) vs ρ(Path H-score, H-score QuPath)	0.76	0.74	0.34	0.74	N.S.D.
ρ(Path H-score, pix H-score VIS) vs ρ(Path H-score, ATM score)	0.76	0.55	2.58	0.02	S.D.
ρ(Path H-score, pix H-score HALO) vs ρ(Path H-score, H-score HALO)	0.72	0.69	0.44	0.67	N.S.D.
ρ(Path H-score, pix H-score HALO) vs ρ(Path H-score, H-score QuPath)	0.72	0.74	-0.33	0.74	N.S.D.
ρ(Path H-score, pix H-score HALO) vs ρ(Path H-score, ATM score)	0.72	0.55	1.54	0.14	N.S.D.
Comparing pairwise correlations between DIA endpoint and PDL1 transcript					
ρ(PD-L1 transcript, H-score HALO) vs ρ(PD-L1 transcript, H-score QuPath)	0.74	0.76	-0.43	0.67	N.S.D.
ρ(PD-L1 transcript, pix H-score VIS) vs ρ(PD-L1 transcript, pix H-score HALO)	0.80	0.79	0.19	0.85	N.S.D.
ρ(PD-L1 transcript, pix H-score VIS) vs ρ(PD-L1 transcript, H-score HALO)	0.80	0.74	1.16	0.26	N.S.D.
ρ(PD-L1 transcript, pix H-score VIS) vs ρ(PD-L1 transcript, H-score QuPath)	0.80	0.76	0.80	0.43	N.S.D.
ρ(PD-L1 transcript, pix H-score VIS) vs ρ(PD-L1 transcript, ATM score)	0.80	0.56	3.29	0.003	S.D.
ρ(PD-L1 transcript, pix H-score HALO) vs ρ(PD-L1 transcript, H-score HALO)	0.79	0.74	1.00	0.33	N.S.D.
ρ(PD-L1 transcript, pix H-score HALO) vs ρ(PD-L1 transcript, H-score QuPath)	0.79	0.76	0.47	0.64	N.S.D.
ρ(PD-L1 transcript, pix H-score HALO) vs ρ(PD-L1 transcript, ATM score)	0.79	0.56	2.48	0.02	S.D.

S.D. significant difference, N.S.D–no significant difference.

https://doi.org/10.1371/journal.pone.0245638.t002

between HALO and QuPath implementations of the H-score, which is analogous to what we observed for P-cadherin. In addition, we observed that there was no significant difference between the HALO and Visiopharm implementations of the pix H-score for PD-L1. Spearman's correlation coefficients between the pix H-score and PD-L1 biomarker abundance endpoints were mostly significantly higher than Spearman's correlation coefficients between ATM score and PD-L1 biomarker abundance endpoints (Table 2). Although both the pix H-score and the ATM score are pixel-based algorithms, the higher Spearman's correlation coefficient for the pix H-score suggests that this algorithm is superior to the ATM score in estimating biomarker abundance for PD-L1.

DIA algorithms for 5T4 quantification

IHC images for 5T4 (Fig 2E) showed strong immunoreactivity at the cell membrane with limited cytoplasmic staining, which was consistent with prior reports [17]. Spearman's correlation of the membrane H-scores of the 21 cases immunolabeled for 5T4, as assessed by a board-certified pathologist (see S1 Table), and qRT-PCR values for 5T4 mRNA transcript from serial sections of the same cases had a ρ value of 0.61, p = 0.003 (Fig 2F).

When compared to the pathologist H-score, all DIA endpoints (see S1 Table) yielded positive correlations (Fig 5A–5E). The Spearman's correlation coefficient for the HALO H-score, QuPath H-score, ATM score, VIS pix H-score and HALO pix H-score with respect to the pathologist H-score were 0.75 ($p<0.0001$), 0.79 ($p<0.0001$), 0.76 ($p < 0.0001$), 0.83 ($p < 0.0001$) and 0.82 ($p < 0.0001$), respectively. When compared to the 5T4 transcript, all DIA endpoints similarly yielded positive correlations (Fig 5F–5J). The Spearman's correlation coefficient for the HALO H-score, Qupath H-score, ATM score, VIS pix H-score and HALO

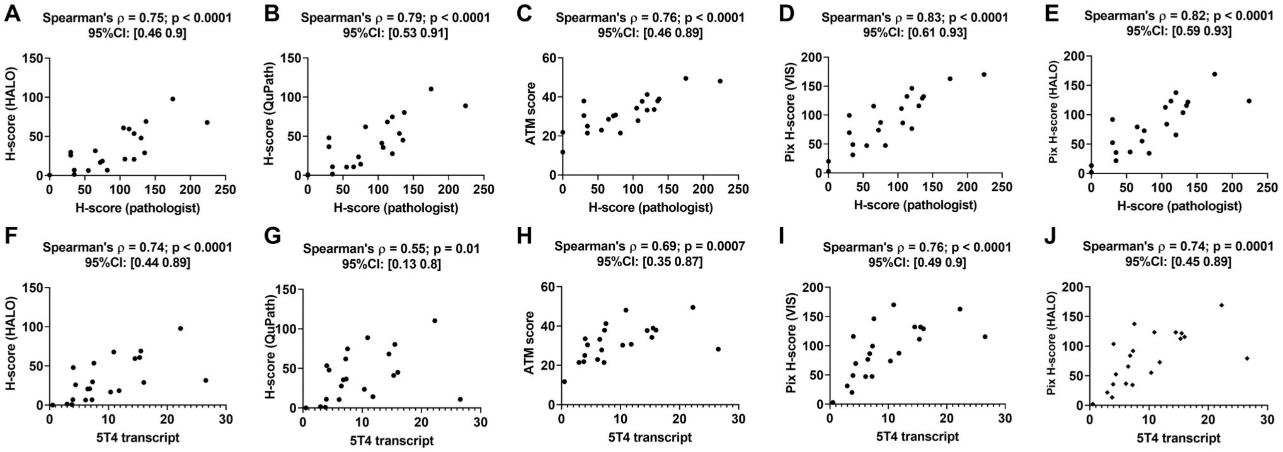

Fig 5. Performance of DIA endpoints obtained from 5T4 IHC images. Panels A-E show plots of the different DIA endpoints as a function of the pathologist H-score, while panels F-J show the same as a function of 5T4 mRNA transcript for a cohort of 21 lung cancer resections. All panels show the Spearman's correlation coefficient between the two quantities plotted in that panel along with the p-value and the 95% confidence interval.

https://doi.org/10.1371/journal.pone.0245638.g005

pix H-score with respect to 5T4 transcript were 0.74 (p<0.0001), 0.55 (p = 0.01), 0.69 (p = 0.0007), 0.76 (p<0.0001) and 0.74 (p = 0.0001), respectively.

Statistical analysis of the Spearman's correlation coefficients revealed that there is no significant difference in the correlation coefficient between each of the DIA based endpoints and pathologist H-score (Table 3). An analogous behavior was also observed for the correlation coefficient between each of the DIA based endpoints and 5T4 transcript except for the QuPath H-score. Specifically, the correlation between QuPath H-score and 5T4 transcript was

Table 3. Table lists the results of William's t-test to test for significant difference in the Spearman correlation coefficients between 5T4 mRNA transcript or pathologist H-score and different DIA endpoints.

Statistical analysis of correlation coefficients for 5T4	ρ_{12}	ρ_{13}	z-score	p-value	Result
Comparing pairwise correlations between DIA endpoint and Pathologist H-score					
ρ(Path H-score, H-score HALO) vs ρ(Path H-score, H-score QuPath)	0.75	0.79	-0.49	0.63	N.S.D.
ρ(Path H-score, pix H-score VIS) vs ρ(Path H-score, pix H-score HALO)	0.83	0.82	0.31	0.76	N.S.D.
ρ(Path H-score, pix H-score VIS) vs ρ(Path H-score, H-score HALO)	0.83	0.75	1.69	0.11	N.S.D.
ρ(Path H-score, pix H-score VIS) vs ρ(Path H-score, H-score QuPath)	0.83	0.79	0.49	0.63	N.S.D.
ρ(Path H-score, pix Hscore VIS) vs ρ(Path H-score, ATM score)	0.83	0.76	1.37	0.18	N.S.D.
ρ(Path H-score, pix H-score HALO) vs ρ(Path H-score, H-score HALO)	0.82	0.75	1.46	0.16	N.S.D.
ρ(Path H-score, pix H-score HALO) vs ρ(Path H-score, H-score QuPath)	0.82	0.79	0.37	0.71	N.S.D.
ρ(Path H-score, pix H-score HALO) vs ρ(Path H-score, ATM score)	0.82	0.76	1.32	0.20	N.S.D.
Comparing pairwise correlations between DIA endpoint and 5T4 transcript					
ρ(5T4 transcript, H-score HALO) vs ρ(5T4 transcript, H-score QuPath)	0.74	0.55	2.11	0.05	S.D.
ρ(5T4 transcript, pix H-score VIS) vs ρ(5T4 transcript, pix H-score HALO)	0.76	0.74	0.67	0.51	N.S.D.
ρ(5T4 transcript, pix H-score VIS) vs ρ(5T4 transcript, H-score HALO)	0.76	0.74	0.50	0.63	N.S.D.
ρ(5T4 transcript, pix H-score VIS) vs ρ(5T4 transcript, H-score QuPath)	0.76	0.55	2.40	0.03	S.D.
ρ(5T4 transcript, pix H-score VIS) vs ρ(5T4 transcript, ATM score)	0.76	0.69	1.40	0.18	N.S.D.
ρ(5T4 transcript, pix H-score HALO) vs ρ(5T4 transcript, H-score HALO)	0.74	0.74	0.10	0.92	N.S.D.
ρ(5T4 transcript, pix H-score HALO) vs ρ(5T4 transcript, H-score QuPath)	0.74	0.55	2.10	0.05	S.D.
ρ(5T4 transcript, pix H-score HALO) vs ρ(5T4 transcript, ATM score)	0.74	0.69	1.11	0.28	N.S.D.

S.D. significant difference, N.S.D–no significant difference.

https://doi.org/10.1371/journal.pone.0245638.t003 significantly lower than the correlation between the HALO H-score or the pix H-score endpoints and 5T4 transcript (Table 3). Finally, we note that there is no significant difference in the correlation coefficient between the HALO and Visiopharm implementations of the pix H-score and either of the biomarker abundance endpoints for 5T4. These results suggest that the pix H-score algorithm has comparable performance to the other DIA algorithms to quantify biomarker abundance for 5T4.

Effect of spatial sampling on pix H-score

We next investigated the robustness of the pix H-score when it is calculated from select regions within the tissue section as opposed to the entire tumor resection. For this purpose, a statistical sampling procedure known as bootstrapping needs to be performed. However, technical limitations in Visiopharm and HALO software packages precluded us from implementing a formal bootstrapping procedure. Therefore, we resorted to an empirical resampling approach (see Methods for details) wherein for a given biomarker each tumor resection was divided into non-overlapping circular regions (Fig 6A). N different circular regions (N ranging from 1 to

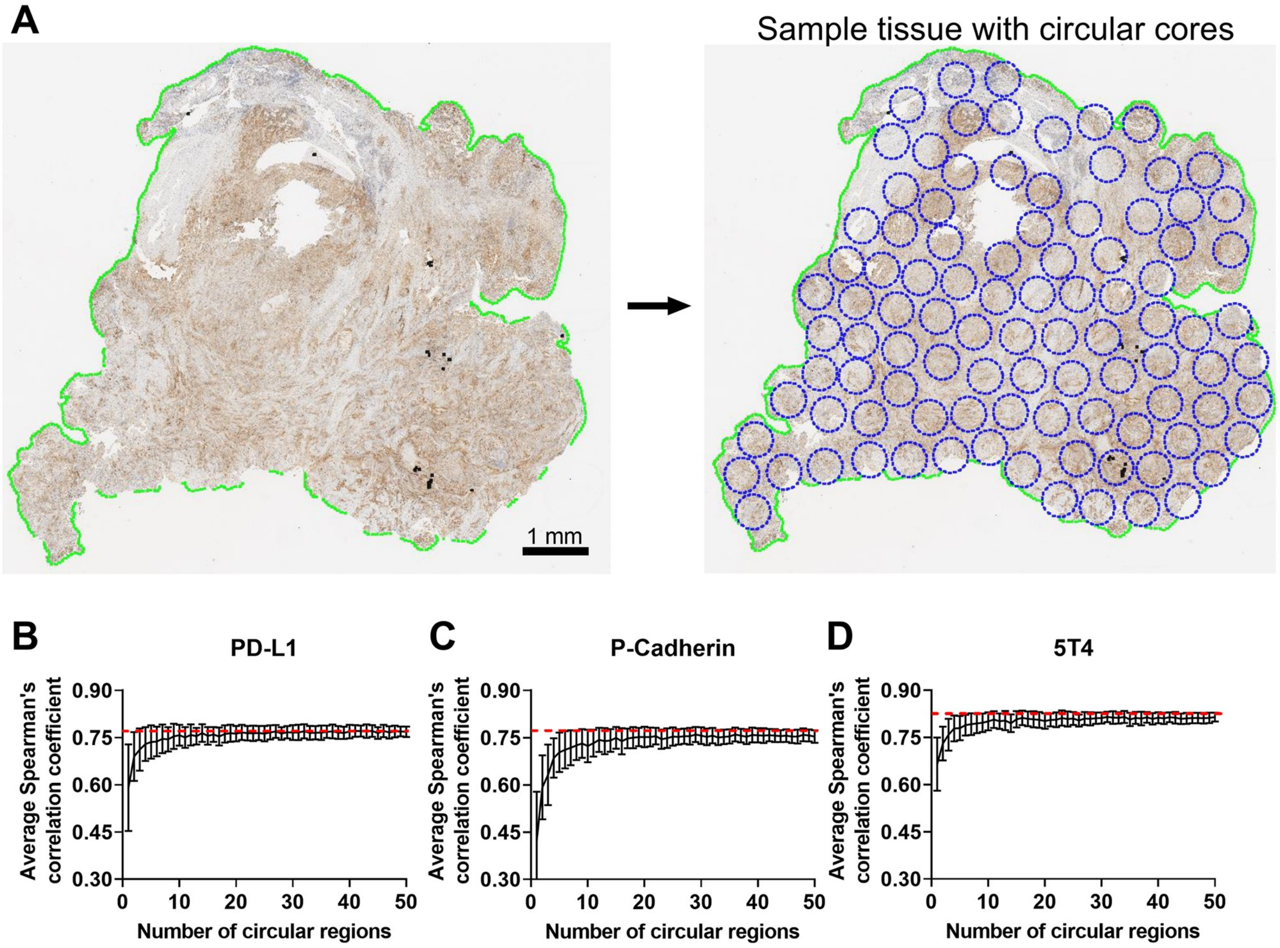

Fig 6. Empirical approach to assess robustness of pix H-score to spatial sampling. Panel A shows the breakup of the tumor resection into non overlapping circular regions. Panels B, C and D show the results of the bootstrap analysis for PD-L1, P-cadherin and 5T4, respectively, where the average Spearman's correlation coefficient between the regional pix H-score estimate from N circular regions and pathologist H-score is plotted as a function of the number of circular regions, where N varies from 1 to 50. The red dashed line shows the Spearman's correlation coefficient between whole-slide Pix H-score and pathologist H-score for that biomarker. Error bars indicate ± SEM.

https://doi.org/10.1371/journal.pone.0245638.g006

50) were randomly selected, and a regional pix H-score was computed from these circular regions. Then the Spearman's correlation coefficient between the pathologist H-score and the regional pix H-score was computed for that biomarker. This procedure was repeated 100 times for all the tumor resections pertaining to that biomarker, and the average Spearman correlation coefficient from 100 iterations was then plotted as a function of the number of circular regions N.

Fig 6B–6D show the behavior of the average Spearman's correlation coefficient for PD-L1, P-cadherin and 5T4, respectively, between pathologist H-score and the regional pix H-score as a function of the number of circular regions from which the regional pix H-score was calculated. For all the biomarkers, we see that for fewer than five circular regions the average Spearman correlation coefficient between the regional pix H-score and pathologist H-score is consistently smaller than the Spearman's correlation coefficient between the whole-slide pix H-score and pathologist H-score (shown by the red dashed line). When 10 or more circular regions are sampled the average Spearman's correlation coefficient for the regional pix H-score starts to plateau out and reaches a steady state. In the case of PD-L1, the plateau region converges with the Spearman's correlation coefficient between the whole-slide pix H-score and pathologist H-score (Fig 6B). In contrast, for P-cadherin and 5T4 the plateau region is slightly lower than the Spearman's correlation coefficient for the whole-slide pix H-score (Fig 6C and 6D). A similar behavior is also observed when biomarker mRNA levels are used as the reference ground truth data in the Spearman's correlation coefficient calculation (S1 Fig).

Discussion

Robust quantification of biomarker expression in tissue sections is a critical need in many diagnostic and investigative pathology workflows. Our motivation to develop a new digital image analysis metric was driven by the need to automate the process of manual scoring by a pathologist. Digital image analysis holds the promise to offer a fast, objective, and reproducible strategy to quantify biomarker expression from histopathology images. In this manuscript, we introduced an unsupervised algorithm, the pix H-score. With it we quantified P-cadherin, PD-L1, and 5T4 signals in immunolabeled FFPE sections of human tumors and found good correlation between the digitally-analyzed IHC signals and manual (visual) signal quantitation as performed by a board certified pathologist. As pathologist scoring is known to be susceptible to intra- and inter-observer variability, we also used biomarker mRNA level as an orthogonal measurement of biomarker abundance to validate the pix H-score. Our observation that there was good concordance between both digital and visual IHC signal quantitation and mRNA transcript abundance for each analyte not only demonstrated the robust nature of the pix H-score algorithm but also validated the pathologist scores.

There are two basic approaches to quantifying biomarker expression from histology images. One approach utilizes cell segmentation and quantifies markers per unit cell whereas a second approach avoids cell segmentation and quantifies markers per unit pixel. In this manuscript, we compared both approaches to quantify biomarker levels from immunohistochemistry images. Unlike the H-score and the Allred score, the pix H-score is a pixel-based algorithm that does not rely on the identification of individual cells and their subcellular compartments. This reduces the computational complexity of the pix H-score and renders its implementation in two different software packages as relatively straightforward.

In our case, the IHC assay for each biomarker was carried out using a different brand of instrument (PD-L1 –Ventana, P-cadherin–DAKO, and 5T4 –Leica Bond RX). Similarly, the slides were scanned at different times (2014–2018) using different slide scanners, although the same scanner was used for a given biomarker. These differences could introduce stain variation [27] and shading correction artefacts [28, 29] which in turn can affect the colorimetric composition of the IHC images [30], especially in large histopathology datasets that are generated in multiple batches. In our study, all the IHC images passed our visual image quality assessment (see Methods). Consequently, we did not carry out stain normalization or shading correction. Despite this, the pix H-score demonstrated robust performance when benchmarked against orthogonal measurements of biomarker abundance. This can be attributed in part to the small batch size of our datasets which likely did not exhibit significant variability that arise due to the above factors. Nevertheless, a systematic analysis on the impact of image quality variation on the pix H-score is warranted and is a topic for future investigation for validating the pix H-score in large, multi-batch image sets.

Our observation that the Visiopharm and the HALO versions of pix H-score exhibited similar performance suggests that the pix H-score is a robust algorithm for estimating IHC biomarker abundance in whole-slide images. This is especially relevant due to the proprietary nature of these software packages which precludes users from understanding several technical aspects of the image analysis workflow. For instance, the specific details regarding the color deconvolution algorithm, which is a key pre-processing step, are not accessible to the user in either Visiopharm or HALO. Consequently, while implementing the pix H-score we did not know how similar the output of the color deconvolution step (i.e. hematoxylin and DAB channels) would be in the two software packages.

An important question arises as to why the DIA based H-score exhibited very different performance for P-cadherin but not for PD-L1. The H-score algorithm relied on the detection of individual cells and their subcellular compartments to quantify biomarker levels. Although this task may seem relatively straightforward for a human observer, nucleus/cell-membrane detection and segmentation are challenging image processing problems especially when applied to whole-slide image analysis where there can be considerable variability in the intensity and the sub-cellular localization pattern of the biomarker of interest [31, 32]. In our case, the latter could be a contributing factor since in the P-cadherin IHC images the biomarker signal was localized to both the cell membrane and cytoplasm whereas in the PD-L1 IHC images the biomarker signal was predominantly localized to the cell membrane. Consequently, this may partly explain the reason why for P-cadherin the performance of the DIA H-score was consistently lower than that of the pix H-score whereas for PD-L1 the performance of the DIA H-score was comparable to that of the pix H-score. Not surprisingly others have also reported similar challenges in automated analysis of membrane-localized biomarker signal [33]. This may also partly explain our observation for 5T4 where the correlation between QuPath H-score and 5T4 transcript was lower than the correlation between pix H-score and 5T4 transcript. More specifically, while 5T4 immunoreactivity is predominantly membranous, there is still detectable cytoplasmic signal in the tumor cells which can affect the quantification of the DIA based H-score.

A similar question also arises for the ATM score which, unlike the H-score, is a pixel-based algorithm but also exhibited very different performance for P-cadherin but not for PD-L1 and 5T4. By definition. the ATM score is proportional to the average intensity of the biomarker in the DAB channel. This is calculated by taking all pixels in the DAB channel including pixels that are negative for the biomarker. When the averaging is performed on a whole-slide image, this can significantly dilute the contribution from pixels that are positive for the biomarker resulting in poor performance in predicting biomarker abundance from the IHC image. In contrast, the pix H-score only considers pixels with a valid biomarker signal as DAB positive pixels (based on a user defined threshold). As a result, the pix H-score can robustly estimate biomarker abundance the IHC image. These differences may also explain in part the reason for the limited range of values taken by the ATM score when compared to the pix H-score. Specifically, the ATM score for P-cadherin, PD-L1, and 5T4 took values in the range of 24 to 77, 8 to 33, and 11 to 49, respectively. In contrast the pix H-score for P-cadherin, PD-L1, and 5T4 took values in the range of 20 to 207, 1 to 131, and 3 to 170, respectively. The latter values are more comparable to the pathologist H-score, which for P-cadherin, PD-L1, and 5T4 ranged from 17 to 298, 0 to 225, and 0 to 224, respectively. In this context, we note that the pix H-score, analogous to traditional H-score, is based on binned intensity data (i.e., DAB high, DAB medium and DAB low pixels) whereas the ATM score is based on continuous intensity values (0–255).

The application of deep learning methodology for nucleus and cell membrane segmentation holds significant promise as it has been shown to have improved performance over traditional algorithms [34]. However, deep learning methods are supervised approaches that require a substantial amount of training data and extensive validation. In many practical applications, generating such large training datasets is not feasible and algorithm validation can be time consuming. In this regard, the pix H-score algorithm introduced here provides a simple yet robust strategy to quantify biomarker expression even from small datasets, as demonstrated here, and can be implemented within a very short timeframe. An interesting follow up study would be to compare the performance of the pix H-score algorithm with deep learning based, scoring approaches.

We note that while our results are encouraging and show the potential for the pix H-score in scoring membrane biomarkers, the algorithm can benefit from additional validation for other biomarkers. Also, the effect of pre-analytical variables (e.g., cold ischemia time, age of unstained cut slides, etc.) on the performance of the pix H-score needs to be investigated. Although not shown here, we expect the pix H-score to also be applicable to immunofluorescence images. In conclusion, we anticipate the pix H-score to be a useful addition to the digital image analysis toolbox for a fast, reproducible and objective strategy to quantify biomarker expression from immunolabeled tissue sections.

Supporting information

S1 Fig. Robustness of pix H-score to spatial sampling. Panels A, B and C show the results of the bootstrap analysis for PD-L1, P-cadherin and 5T4, respectively, where the average Spearman's Correlation coefficient between the regional pix H-score estimate from N circular regions and mRNA transcript is plotted as a function of the number of circular regions, where N varies from 1 to 50. The red dashed line shows the Spearman's correlation coefficient between whole-slide Pix H-score and mRNA transcript for that biomarker. Error bars indicate ± SEM.
(TIF)

S1 Table. Visual and digital scores for P-cadherin, PD-L1 and 5T4. The table lists the pathologist H-score, mRNA transcript level and the various DIA endpoints for every sample for a given biomarker.
(XLSX)

S1 File. Zip file that contains the QuPath scripts in Groovy scripting language to detect and score cells based on membrane signal. The scripts were written in Version 0.2.0-m2. The script has not been tested in subsequent releases of QuPath.
(ZIP)

S2 File. Zip file that contains the Visiopharm apps to calculate the ATM score and pix H-score for the different biomarkers.
(ZIP)

S3 File. Zip file that contains the settings file which can be loaded in HALO to implement the pix H-score algorithm for the different biomarkers.
(ZIP)

Acknowledgments

We thank Shawn O'Neil and Timothy Affolter for critical reading of the manuscript.

Author Contributions

Conceptualization: Sripad Ram.

Data curation: Sripad Ram, Pamela Whalen, Xiaoling Xia.

Formal analysis: Sripad Ram, Shibing Deng, Xiaoling Xia, Eric L. Powell.

Investigation: Sripad Ram.

Methodology: Pamela Vizcarra, Pamela Whalen, C. L. Painter, Amy Jackson-Fisher, Eric L. Powell.

Project administration: Sripad Ram.

Resources: Amy Jackson-Fisher, Steven Pirie-Shepherd, Eric L. Powell.

Software: Sripad Ram.

Validation: Sripad Ram.

Visualization: Sripad Ram.

Writing – original draft: Sripad Ram, Pamela Whalen, Eric L. Powell.

Writing – review & editing: Sripad Ram, Eric L. Powell.

References

1. Aeffner F., et al., Introduction to Digital Image Analysis in Whole-slide Imaging: A White Paper from the Digital Pathology Association. J Pathol Inform, 2019. 10: p. 9. https://doi.org/10.4103/jpi.jpi_82_18 PMID: 30984469
2. Meyerholz D.K. and Beck A.P., Principles and approaches for reproducible scoring of tissue stains in research. Lab Invest, 2018. 98(7): p. 844–855. https://doi.org/10.1038/s41374-018-0057-0 PMID: 29849125
3. Aeffner F., et al., Commentary: Roles for Pathologists in a High-throughput Image Analysis Team. Toxicol Pathol, 2016. 44(6): p. 825–34. https://doi.org/10.1177/0192623316653492 PMID: 27343178
4. Brunnstrom H., et al., PD-L1 immunohistochemistry in clinical diagnostics of lung cancer: inter-pathologist variability is higher than assay variability. Mod Pathol, 2017. 30(10): p. 1411–1421. https://doi.org/10.1038/modpathol.2017.59 PMID: 28664936
5. Gomes D.S., et al., Inter-observer variability between general pathologists and a specialist in breast pathology in the diagnosis of lobular neoplasia, columnar cell lesions, atypical ductal hyperplasia and ductal carcinoma in situ of the breast. Diagn Pathol, 2014. 9: p. 121. https://doi.org/10.1186/1746-1596-9-121 PMID: 24948027
6. Hirsch F.R., et al., PD-L1 Immunohistochemistry Assays for Lung Cancer: Results from Phase 1 of the Blueprint PD-L1 IHC Assay Comparison Project. J Thorac Oncol, 2017. 12(2): p. 208–222. https://doi.org/10.1016/j.jtho.2016.11.2228 PMID: 27913228
7. Rimm D.L., et al., A Prospective, Multi-institutional, Pathologist-Based Assessment of 4 Immunohistochemistry Assays for PD-L1 Expression in Non-Small Cell Lung Cancer. JAMA Oncol, 2017. 3(8): p. 1051–1058. https://doi.org/10.1001/jamaoncol.2017.0013 PMID: 28278348
8. Rizzardi A.E., et al., Quantitative comparison and reproducibility of pathologist scoring and digital image analysis of estrogen receptor beta2 immunohistochemistry in prostate cancer. Diagn Pathol, 2016. 11(1): p. 63. https://doi.org/10.1186/s13000-016-0511-5 PMID: 27401406

9. Barnes M., et al., Whole tumor section quantitative image analysis maximizes between-pathologists' reproducibility for clinical immunohistochemistry-based biomarkers. Lab Invest, 2017. 97(12): p. 1508–1515. https://doi.org/10.1038/labinvest.2017.82 PMID: 28805805

10. Tsao M.S., et al., PD-L1 Immunohistochemistry Comparability Study in Real-Life Clinical Samples: Results of Blueprint Phase 2 Project. J Thorac Oncol, 2018. 13(9): p. 1302–1311. https://doi.org/10.1016/j.jtho.2018.05.013 PMID: 29800747

11. Stalhammar G., et al., Digital image analysis outperforms manual biomarker assessment in breast cancer. Mod Pathol, 2016. 29(4): p. 318–29. https://doi.org/10.1038/modpathol.2016.34 PMID: 26916072

12. Aeffner F., et al., The Gold Standard Paradox in Digital Image Analysis: Manual Versus Automated Scoring as Ground Truth. Arch Pathol Lab Med, 2017. 141(9): p. 1267–1275. https://doi.org/10.5858/arpa.2016-0386-RA PMID: 28557614

13. Choudhury K.R., et al., A robust automated measure of average antibody staining in immunohistochemistry images. J Histochem Cytochem, 2010. 58(2): p. 95–107. https://doi.org/10.1369/jhc.2009.953554 PMID: 19687472

14. Camp R.L., Chung G.G., and Rimm D.L., Automated subcellular localization and quantification of protein expression in tissue microarrays. Nat Med, 2002. 8(11): p. 1323–7. https://doi.org/10.1038/nm791 PMID: 12389040

15. Fisher A.-J., et al. Pixelwise H-score: a novel digital image analysis-based metric to quantify membrane biomarker expression from IHC images. 2019. Journal for Immunotherapy of Cancer, 7(Suppl 1), Abstract #53.

16. Pfizer Inc. Use of Human tissue. https://www.pfizer.com/science/clinical-trials/integrity-transparency/policy-usehuman-%20tissue.

17. Pirie-Shepherd S.R., et al., Detecting expression of 5T4 in CTCs and tumor samples from NSCLC patients. PLoS One, 2017. 12(7): p. e0179561. https://doi.org/10.1371/journal.pone.0179561 PMID: 28727782

18. Dabbs, D.J., Diagnostic Immunohistochemistry. 3rd ed. 2013: Elsevier Health Sciences.

19. Detre S., Saclani Jotti G., and Dowsett M., A "quickscore" method for immunohistochemical semiquantitation: validation for oestrogen receptor in breast carcinomas. J Clin Pathol, 1995. 48(9): p. 876–8. https://doi.org/10.1136/jcp.48.9.876 PMID: 7490328

20. McClelland R.A., et al., Automated quantitation of immunocytochemically localized estrogen receptors in human breast cancer. Cancer Res, 1990. 50(12): p. 3545–50. PMID: 2187598

21. Bankhead P., et al., QuPath: Open source software for digital pathology image analysis. Sci Rep, 2017. 7(1): p. 16878. https://doi.org/10.1038/s41598-017-17204-5 PMID: 29203879

22. Schober P., Boer C., and Schwarte L.A., Correlation Coefficients: Appropriate Use and Interpretation. Anesth Analg, 2018. 126(5): p. 1763–1768. https://doi.org/10.1213/ANE.0000000000002864 PMID: 29481436

23. Steiger J.H., Tests for comparing elements of a correlation matrix. Psychological Bulletin, 1980. 87(2): p. 245–251.

24. Williams E.J., The Comparison of Regression Variables. Journal of the Royal Statistical Society: Series B (Methodological), 1959. 21(2): p. 396–399.

25. Kovacs A., Dhillon J., and Walker R.A., Expression of P-cadherin, but not E-cadherin or N-cadherin, relates to pathological and functional differentiation of breast carcinomas. Mol Pathol, 2003. 56(6): p. 318–22. https://doi.org/10.1136/mp.56.6.318 PMID: 14645693

26. Paredes J., et al., P-cadherin overexpression is an indicator of clinical outcome in invasive breast carcinomas and is associated with CDH3 promoter hypomethylation. Clin Cancer Res, 2005. 11(16): p. 5869–77. https://doi.org/10.1158/1078-0432.CCR-05-0059 PMID: 16115928

27. Macenko, M., et al., A method for normalizing histology slides for quantitative analysis. Processing of the IEEE International Symposium on Biomedical Imaging, 2009. 5: p. 1107–1110.

28. Kayser K., Quantification of virtual slides: Approaches to analysis of content-based image information. J Pathol Inform, 2011. 2: p. 2. https://doi.org/10.4103/2153-3539.74945 PMID: 21383926

29. Kayser K., et al., How to measure image quality in tissue-based diagnosis (diagnostic surgical pathology). Diagn Pathol, 2008. 3 Suppl 1: p. S11.

30. Rojo M.G., et al., Critical comparison of 31 commercially available digital slide systems in pathology. Int J Surg Pathol, 2006. 14(4): p. 285–305. https://doi.org/10.1177/1066896906292274 PMID: 17041192

31. Irshad H., et al., Methods for nuclei detection, segmentation, and classification in digital histopathology: a review-current status and future potential. IEEE Rev Biomed Eng, 2014. 7: p. 97–114. https://doi.org/10.1109/RBME.2013.2295804 PMID: 24802905

32. Xing F. and Yang L., Robust Nucleus/Cell Detection and Segmentation in Digital Pathology and Microscopy Images: A Comprehensive Review. IEEE Rev Biomed Eng, 2016. 9: p. 234–63. https://doi.org/10.1109/RBME.2016.2515127 PMID: 26742143

33. Lopes N., et al., Digital image analysis of multiplex fluorescence IHC in colorectal cancer recognizes the prognostic value of CDX2 and its negative correlation with SOX2. Lab Invest, 2020. 100(1): p. 120–134. https://doi.org/10.1038/s41374-019-0336-4 PMID: 31641225

34. Caicedo J.C., et al., Evaluation of Deep Learning Strategies for Nucleus Segmentation in Fluorescence Images. Cytometry A, 2019. 95(9): p. 952–965. https://doi.org/10.1002/cyto.a.23863 PMID: 31313519