

Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

Mohamed Amgad Tageldin

Date

**Computational discovery of interpretable histopathologic prognostic biomarkers in
invasive carcinomas of the breast**

By

Mohamed Amgad Tageldin
Doctor of Philosophy

Computer Science and Informatics

Lee A.D. Cooper, Ph.D.
Advisor, Co-Chair

David A. Gutman, M.D., Ph.D.
Co-Chair

Mia M. Gaudet, Ph.D.
Committee Member

Carlos S. Moreno, Ph.D.
Committee Member

Accepted:

Kimberly Jacob Arriola, Ph.D.
Dean of the James T. Laney School of Graduate Studies

Date

**Computational discovery of interpretable histopathologic prognostic biomarkers in
invasive carcinomas of the breast**

By

Mohamed Amgad Tageldin
M.B.B.Ch., Cairo University, Cairo, Egypt, 2016
M.Sc., Emory University, GA, USA, 2019

Advisor: Lee A.D. Cooper, Ph.D.

An abstract of
A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Computer Science and Informatics
2021

Abstract

Computational discovery of interpretable histopathologic prognostic biomarkers in invasive carcinomas of the breast

By Mohamed Amgad Tageldin

While microscopic examination of tumor resections and biopsies has been a cornerstone in breast cancer grading for decades, it suffers from considerable inter-rater variability due to perceptual limitations and high clinical caseloads. Computational analysis of whole-slide image scans using convolutional neural networks (CNN) can help address this challenge. Unfortunately, CNNs can be difficult to interpret, which motivates our adoption of an approach called concept bottlenecking, where models first detect various tissue structures then use them to make their prediction. Concept bottleneck models require a large set of manual annotation data to train. Unfortunately, manual delineation of histopathologic structures is very demanding and impractical given pathologists' time constraints. This dissertation describes contributions that fall under the themes of scalable data collection, deep learning-based tissue detection, and the discovery of novel histopathologic biomarkers and associations.

First, we examine crowdsourcing approaches that engage medical students to collect manual annotation data. Our results show that a structured, collaborative approach with pathologist supervision is scalable; the resultant publicly-released BCSS and NuCLS datasets contain 20,000 and 200,000 annotations of tissue regions and nuclei, respectively. We show that medical students produce accurate annotations for predominant, visually distinctive structures and that algorithmic suggestions help scale and improve the accuracy of annotations.

Second, we describe a set of CNN modeling approaches for the accurate delineation of histopathologic structures. We describe various improvements to enhance the performance of nucleus detection CNN models and introduce a technique called Decision Tree Approximation of Learned Embeddings, which helps explain CNN nucleus classifications without compromising prediction accuracy. Additionally, we offer consensus recommendations from the International Immuno-Oncology Working Group surrounding the computational detection of tumor-infiltrating lymphocytes, a critical emerging biomarker. Following these recommendations, we develop and validate a multi-scale CNN model that jointly detects tissue regions and nuclei, employing pre-defined biological constraints to improve accuracy.

Finally, we describe the development of a morphologic signature based on quantitative features extracted from computationally-delineated histopathologic regions and cells. This morphologic signature relies partly on a set of stromal features not captured by clinical guidelines for breast cancer grading, and has a stronger prognostic value.

**Computational discovery of interpretable histopathologic prognostic biomarkers in
invasive carcinomas of the breast**

By

Mohamed Amgad Tageldin
M.B.B.Ch., Cairo University, Cairo, Egypt, 2016
M.Sc., Emory University, GA, USA, 2019

Advisor: Lee A.D. Cooper, Ph.D.

A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Computer Science and Informatics
2021

Acknowledgments

Upon acceptance of this dissertation, I will have permission to add three letters after my name, Ph.D. These three magical letters signify 5+ years of scholarly work, and are expected to have broad positive implications on my future career. But I didn't do this alone, not even close! I wish I could append these letters after the names of each and every person who made this possible; a list that is impossible to enumerate, but I will try anyway.

First and foremost, I wish to express my gratitude to the cancer patients whose biological samples were at the heart of all research presented here. Many of those patients have died, but their memory will always live on in the hearts of their loved ones. This research is part of their legacy, and I sincerely hope that it honors their memory and reduces future suffering by improving diagnostic strategies.

I sincerely thank my Ph.D. advisor, Dr. Lee A.D. Cooper, who took a risk by mentoring a fresh medical graduate with no computer science background and very limited programming experience. I owe him almost everything I learned during this journey about machine learning, computational pathology and how to navigate the academic landscape. I also wish to thank all my other academic mentors, especially my Ph.D. committee members: Dr. David A. Gutman, Dr. Mia M. Gaudet, and Dr. Carlos S. Moreno.

This work would not be possible without the love and support my dear wife, Maha, has provided. Our marriage is only slightly older than this Ph.D. journey; six years of laughter, friendly nitpicking, low-key competition, and silly song lyrics. Through thick and thin, we were always there for each other, clearing milestone after milestone and pushing each other to be better and to do better. We sometimes forget how much of a blessing that is.

No words can ever express my gratitude to my wonderful and incredibly loving family. Thank you, mom and dad, for providing me with an affectionate and adventurous childhood, for investing heavily in my happiness and education, for listening to my endless rants, and for always believing in me. Thank you Omar and Mostafa for being such incredible brothers, and for sparing no effort to help me whenever the need arose. Dearest Mariam, you've grown so fast! I am proud of you, I love you, and I hope we can always remain a large part of each other's lives. Thank you to all member of my extended family for always believing in me.

It is safe to say that without a robust network of supportive friends and colleagues, any academic endeavour ceases to be a source of joy and, instead, becomes a source of stress and despair. This is especially true for data-focused research like the one presented here, which tends to offer delayed gratification. For that, I am grateful to each one of my friends who stood by my side through these five years, and heard my repetitive complaints and boasting about works of no practical relevance to their lives. Thank you Ahmed Elmanzalawi and Heba Salah, my long-term friends and advocates. Thank you Mohamed Saleh, Reham Elfawal, Ahmad Elkashash, Toka Ibrahim, Kareem Hosny, Marco Tsui, Mandy Tin, Omar Mokhtar, Megan Keaveney, Kareem Emara, Ashraf Abdelghany, and Nizar Saleh for the fun times. I hope we always stay in touch.

Needless to say, no academic work happens in a vacuum, and I would like to acknowledge the contributions of volunteers who annotated the BCSS and NuCLS datasets, my friends and lab mates Dr. Pooya Mobadersany, Dr. Safoora Yousefi, Dr. Sanghoon Lee, and Dr. Saumya Gurbani for their insights, Dr. Jeff Goldstein for his helpful suggestions, and Dr. Kalliopi P Siziopikou for her expert opinion. I thank each and every professor who taught me a class, wrote me a recommendation letter, or provided me with direct or indirect mentorship and advice, especially Dr. Emad Shash, Dr. Roberto Salgado, Dr. Marco Tsui, Dr. Mary Ann Price, Dr. Hoda Z. Amer, Dr. Sherif Nasr, Dr. Abigail M. Brown, Dr. Melvyn M. Jones, and Dr. Amr S. Soliman. I would also like to acknowledge all academic collaborators from the American Cancer Society (Dr. Samantha Puvanesarajah, Dr. Lauren Teras, James Hodge, and Elizabeth Bain), Kitware (David Manthey), Roche Tissue Diagnostics (Dr. Srinivas Chukka, Dr. Anindya Sarkar, Dr. Uday Kurkure, Dr. Michael Barnes, Dr. Jim Martin), and the U.S. Food and Drug Administration (Dr. Brandon D. Gallas, Sarah N. Dudgeon).

Last but not least, I would like to acknowledge all funding organizations, public and private, who made this work possible. A sincere thank you to the amazing individuals, professional societies, and private companies that donated money to fund academic awards to encourage early-career scholars like myself, including Mr. Chris Schoettle, Dr. Edward Klatt, Association for Pathology Informatics, and the Digital Pathology Association.

Contents

1	Background and significance	1
1.1	Histopathology and Cancer Biology	2
1.2	Computational Pathology	8
1.3	Machine learning in pathology	17
1.4	Integrative computational analysis	24
1.5	Organization and summary of contributions	29
1.6	List of publications	36
 2	 Crowdsourcing strategies for scalable curation of annotation datasets	 40
2.1	Structured crowdsourcing enables convolutional segmentation of histology images . .	41
2.2	NuCLS: A scalable crowdsourcing approach and dataset for nucleus classification and segmentation in breast cancer	49
2.3	A pathologist-annotated dataset for validating artificial intelligence: a project de- scription and pilot study	60
 3	 Deep-learning methods for automatic detection of histopathology structures	 73
3.1	Report on computational assessment of tumor infiltrating lymphocytes from the International Immuno-Oncology Biomarker Working Group	75
3.2	Explainable nucleus classification using Decision Tree Approximation of Learned Embeddings	89
3.3	Joint region and nucleus segmentation for characterization of tumor infiltrating lym- phocytes in breast cancer	97
3.4	MuTILs: explainable, multiresolution computational scoring of Tumor-Infiltrating Lymphocytes in breast carcinomas using clinical guidelines	106

4	Histopathologic correlates of clinical and genomic phenotypes	117
4.1	Histomic Prognostic Score: a computational morphologic signature with independent prognostic value in invasive carcinomas of the breast	118
4.2	High expression of MKK3 is associated with worse clinical outcomes in African American breast cancer patients	144
5	HistomicsTK: an open-source software for computational pathology	164
5.1	Creation, parsing and expert review of WSI annotations	165
5.2	Image processing operations for computational pathology	170
5.3	Simple workflows for detection of salient tissue	171
6	Summary of conclusions and future directions	176
7	List of recurrent abbreviations	182
Appendix A : Supplementary methods and results		182
	Supplement for Section 2.1	183
	Supplement for Section 2.2	201
	Supplement for Section 3.2	217

List of Figures

Figure 1.1	Histological staining	3
Figure 1.2	Histologic characteristics of breast carcinomas	6
Figure 1.3	Gene Set Enrichment Analysis	28
Figure 2.1.1	Study overview	41
Figure 2.1.2	Screenshot of the DSA and HistomicsTK web interface	41
Figure 2.1.3	Evaluation slide set concordance and model accuracy	41
Figure 2.1.4	Model performance over the testing set.	41
Figure 2.2.1	Dataset annotation and quality control procedure	49
Figure 2.2.2	Inference from multi-rater datasets	49
Figure 2.2.3	Accuracy of participant annotations	49
Figure 2.2.4	Effect of algorithmic sugges. on annotation abundance and accuracy	49
Figure 2.2.5	Effect of clustering on detection and interrater agreement	49
Figure 2.3.1	Study design and validation workflow	60
Figure 2.3.2	Screenshots from GUIs of three platforms used in data collection.	60
Figure 2.3.3	The distribution of stromal TILs densities in three slides	60
Figure 2.3.4	Scatter plot of stromal TILs densities from two pathologists.	60
Figure 3.1.1	Outline of the visual and computational procedure for scoring TILs	75
Figure 3.1.2	Conceptual pathology report for computational TILs assessment	75
Figure 3.2.1	Example hybrid bounding box and segmentation data	89
Figure 3.2.2	Comparison of the NuCLS dataset with canonical datasets	89
Figure 3.2.3	DTALE provides falsifiable explanations for nucleus detection models	89
Figure 3.2.4	NuCLS model architecture	89

Figure 3.2.5	Qualitative performance of NuCLS model on testing sets	89
Figure 3.2.6	Explaining NuCLS model decisions using DTALE	89
Figure 3.2.7	DTALE enables fine-grained approximation of model decisions	89
Figure 3.3.1	Problem setting	97
Figure 3.3.2	Overall workflow used to obtain region and nucleus classification	97
Figure 3.3.3	Effect of architecture on training categorical accuracy	97
Figure 3.3.4	Qualitative examination of segmentation results on the testing set	97
Figure 3.3.5	Accuracy of segmentation and classification	97
Figure 3.3.6	Calculating sTIL scores and correlating with manual scores	97
Figure 3.3.7	Kaplan-Meier curves for human and computational scores	97
Figure 3.4.1	Components of variants of the computational TILs score	106
Figure 3.4.2	MuTILs model architecture	107
Figure 3.4.3	Reconciliation of manual region and nucleus ground truth	108
Figure 3.4.4	Sample whole-slide predictions from trained MuTILs models	111
Figure 3.4.5	Correlation between visual and computational TILs scores	112
Figure 3.4.6	Kaplan-Meier analysis of visual and computational TILs assessment	113
Figure 3.4.7	Qualitative examination of sample testing set predictions	116
Figure 4.1.1	Concept bottleneck modeling in invasive carcinomas of the breast	119
Figure 4.1.2	Visualization of two prognostically important histomic features	120
Figure 4.1.3	Features capturing stromal matrix and collagen entropy	123
Figure 4.1.4	Correlation between histomic features	125
Figure 4.1.5	Model fitting to obtain the Histomic Prognostic Score and groups	126
Figure 4.1.6	Model fitting to obtain the baseline model	127
Figure 4.1.7	Epithelial architecture measurements are associated with Nottingham grades	129
Figure 4.1.8	Epithelial nuclear measurements are associated with Nottingham grades	130
Figure 4.1.9	Association between histomic and pathologist TILs scores	131
Figure 4.1.10	Univariable coefficients for histomic features for BCSS on CPS-II	132
Figure 4.1.11	Differences in feature distributions for the All-grade model	134
Figure 4.1.12	Differences in feature distributions for the High-grade model	134

Figure 4.1.13	KM curve of Histomic Prognostic Groups on the CPS-II cohort	135
Figure 4.1.14	Risk group changes of Histomic Prognostic Grades are adopted	136
Figure 4.1.15	KM curve of Histomic Prognostic Groups on the TCGA cohort	137
Figure 4.1.16	Multivariable Cox PH of Histomic Prognostic Scores and Groups (v1)	138
Figure 4.1.17	Multivariable Cox PH of Histomic Prognostic Scores and Groups (v2)	139
Figure 4.2.1	Workflow to uncover potentially druggable genes	144
Figure 4.2.3	Differential expression and breast cancer patient survival analysis	144
Figure 4.2.5	Survival curves for the top-32 genes that contribute to survival disparity . .	144
Figure 4.2.7	MKK3 upregulation correlates with poor survival of Black patients	144
Figure 4.2.9	Black patients with high MKK3 are enriched in MYC targets	144
Figure 4.2.11	MKK3 increases the tumor aggressiveness in Black TNBC patients	144
Figure 4.2.13	MKK3 overexpression contributes in MYC induction of EMT	144
Figure 5.1.1	Use of review galleries for rapid expert review of annotations	165
Figure 5.1.2	Back and forth conversion of the annotation database formats	166
Figure 5.1.3	Compact format for encoding object and panoptic segmentation masks	168
Figure 5.1.4	Handling annotations and masks for semantic segmentation tasks	169
Figure 5.2.1	Masked color normalization and augmentation	170
Figure 5.3.1	Simple thresholding-based tissue detection workflow	172
Figure 5.3.2	Color thresholding-based semantic segmentation (methodology)	173
Figure 5.3.3	Color thresholding-based semantic segmentation (results)	174
Figure 5.3.4	Semantic segmentation of cellular regions using superpixels	175
Figure S 2.1.1	Processing annotations and integrating corrections	183
Figure S 2.1.2	Evaluation slide set discordance and model training process	183
Figure S 2.1.3	Two-stage review and correction process	183
Figure S 2.1.4	Sources of errors and the annotation correction process	183
Figure S 2.1.5	Effect of training dataset size on model generalization	183
Figure S 2.1.6	Semantic segmentation accuracy	183
Figure S 2.1.7	Semantic segmentation visualization over testing set ROI's (1)	183
Figure S 2.1.8	Semantic segmentation visualization over testing set ROI's (2)	183

Figure S 2.1.9	Semantic segmentation visualization over testing set ROI's (3)	183
Figure S 2.1.10	Patterns where segmentation algorithm departs from truth	183
Figure S 2.2.1	Use of galleries for scalable review of single-rater annotations	201
Figure S 2.2.2	Process for obtaining algorithmic suggestions for assisted annotation	201
Figure S 2.2.3	Super-class accuracy of annotations and inferred NP-labels	201
Figure S 2.2.4	Accuracy of algorithmic suggestions (single-rater dataset)	201
Figure S 2.2.5	Abundance and segmentation accuracy of clicked suggestions	201
Figure S 2.2.6	Annotation procedure on HistomicsUI	201
Figure S 2.2.7	Confusion matrix of participant annotations (Evaluation dataset)	201
Figure S 2.2.8	Ease of detection of various nucleus classes (Evaluation dataset)	201
Figure S 2.2.9	Sample poor annotation data excluded during correction	201
Figure S 3.2.1	Internal-external cross-validation procedure	217
Figure S 3.2.2	Progression of NuCLS model training and convergence on fold 1	217
Figure S 3.2.3	Qualitative performance of NuCLS model	217
Figure S 3.2.4	Confusion matrix of NuCLS model predictions	217
Figure S 3.2.5	Representative vs discriminative DTALE approximations	217

List of Tables

Table 1.1	The hallmarks of cancer	7
Table 1.2	A small sample of commonly used CNN architectures	21
Table 1.3	Organization of this dissertation	30
Table 2.1.1	Testing accuracy of the full semantic segmentation model	41
Table 2.3.1	Region of interest types	60
Table 2.3.2	Proposed context of use for a stromal TILs density annotated dataset	60
Table 3.1.1	Sample CTA algorithms from the published literature	75
Table 3.2.1	Generalization accuracy of NuCLS models (single-rater)	89
Table 3.3.1	Combined mask code and corresponding region and cell encoding	97
Table 3.4.1	Generalization accuracy for region and nucleus classification	110
Table 3.4.2	Cox regression analysis of the visual and computational TILs scores	114
Table 4.2.1	Frequency of tumor driver gene alterations	144
Table S 2.1.1	Guiding instructions given and tips learned from our experience	183
Table S 2.1.2	Number of annotations in final dataset, broken down by region class	183
Table S 2.1.3	Patch classification accuracy improves with larger training datasets	183
Table S 2.2.1	Definitions and abbreviations used	201
Table S 2.2.2	Accuracy of algorithmic suggestions	201
Table S 2.2.3	Hyperparameters used for MaskRCNN model training	201
Table S 3.2.1	NuCLS model tuning for the nucleus detection task	217
Table S 3.2.2	NuCLS model tuning for the nucleus classification task	217
Table S 3.2.3	Generalization accuracy of the NuCLS models (multi-rater)	217

Table S 3.2.4	Generalization accuracy of the NuCLS models by superclass	217
Table S 3.2.5	List of interpretable features used as input for DTALE	217

List of Algorithms

Algorithm S 2.2.1 Obtaining anchor proposals by constrained agglom. clustering 201

Chapter 1

Background and significance

Approximately 40% of men and women will be diagnosed with cancer at some point during their lifetimes [124]. In the U.S. in 2019, the patient economic burden associated with cancer care exceeded \$16 billion, with the average adult cancer survivor having incurred \$300 in time cost per year [81]. Research in oncology diagnostics and therapeutics is key to addressing this mortality and financial burden. Advances in slide scanners, machine-learning, and computational efficiency have increased interest in histology as a source of data in cancer studies [2, 184]. Tissue morphology contains essential prognostic and diagnostic information and reflects underlying molecular and biological processes. This body of work presents approaches for the computational discovery of interpretable predictive histologic biomarkers, focusing on invasive breast carcinomas as a model disease.

Section 1.1

Histopathology and Cancer Biology

Histopathology is a medical field where medical experts (i.e., pathologists) examine stained microscopic tissue sections to make diagnostic decisions, most often from tumor biopsies. While much of clinical medicine relies on the clinical examination of patients, histopathology is an imaging-focused field, like radiology, where much of the focus is on visual pattern recognition.

1.1.1 The anatomical pathology workflow

The typical workflow in anatomical pathology involves fixation of biopsy tissue in formalin, followed by cutting into sections (a.k.a. *grossing*) and embedding in paraffin blocks. The result is known as *FFPE* (formalin-fixed, paraffin-embedded) tissue blocks. These blocks are cut into thin sections and stained for examination under a microscope. The most widely used stain in histopathology is also one of the oldest and is a combination of two stains:

- *Hematoxylin*, which is alkaline, mainly binds to and stains nucleic acids like DNA (nuclei) and RNA (ribosomes). It has a bluish-purple color.
- *Eosin*, which is acidic, binds to and stains a variety of cytoplasmic elements. It is pinkish-red in color.

The contrast in color between the two stains, coupled with the fact that they have opposite pH and stain different cellular components, made them a prevalent choice for examining tissue structure and general morphology [134, 119]. The combined stain is known as *Hematoxylin & Eosin* or H&E for short (Figure 1.1). Besides H&E, pathologists order ‘special’ biochemical stains in different contexts [135]. These typically react chemically with specific normal or pathologic tissue components. Famous special stains include Congo Red (which stains amyloid aggregates), Reticulin (which stains some types of collagen), Prussian Blue (which stains iron), and many others.

Whereas H&E (and some special stains) primarily enable morphology examination, another set of techniques is used for the specific visualization of individual proteins. Surveying the expression of individual proteins in different cells and regions within a slide is done by the enzymatic or

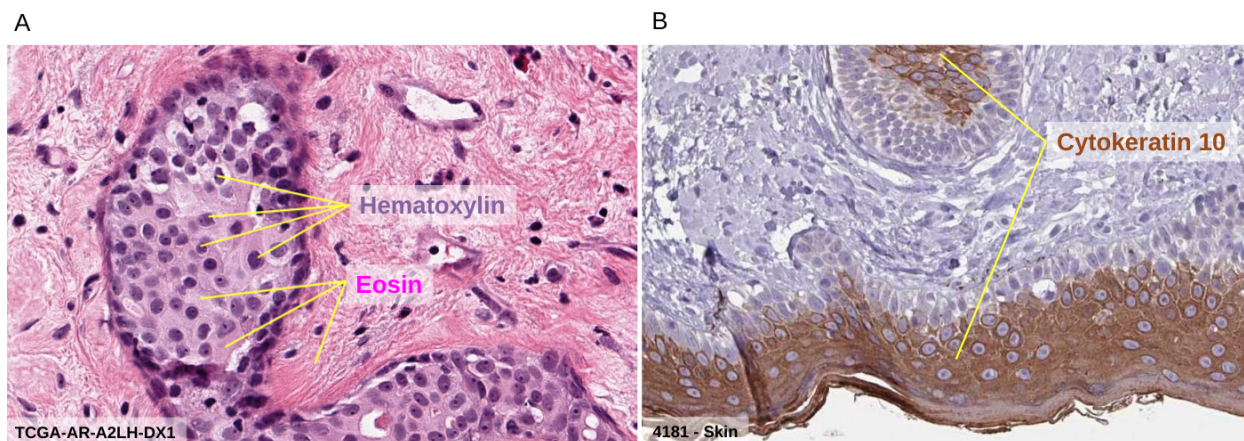


Figure 1.1: **Histological staining.** A. Hematoxylin and Eosin staining. This is a scanned image from one of the breast cancer slides from The Cancer Genome Atlas open-access dataset. B. Immunohistochemical staining. The figure shows a stained skin sample from the open-access Human Protein Atlas dataset.

fluorescent highlighting of individual antibody-tagged proteins using *immunohistochemistry* (IHC) or *In-Situ Hybridization* (FISH). IHC staining is ordered whenever a concrete diagnostic cannot be reached using H&E, as it offers high specificity and contrast (Figure 1.1) [119, 137].

Multiplex IHC and Dual/Multiplex ISH allow simultaneous visualization of multiple markers (typically less than five) and have gained much interest and adoption in recent years, especially in research applications like mapping the tumor microenvironment [176]. These techniques provide direct visualization of protein expression. They are, nonetheless, limited by experimental and financial constraints, including challenges related to antibody sensitivity and specificity [147].

Other visualization techniques are sometimes used in histopathology, depending on context. For example, *immunofluorescence microscopy* (IFM) enables high-contrast visualization of specific tissues. In different contexts where the resolution of the light microscope is insufficient, electron microscopy (EM) is used. Both IFM and EM are commonly used in the context of renal pathology [60].

1.1.2 Histologic grading and histologic biomarkers

Histologic grading categorizes cancer tissue into prognostic groups based on various morphologic criteria [12]. There are typically four grades, with the lower grades corresponding to *well-differentiated* while higher grades correspond to *poorly differentiated* tumors. *Differentiation* is a term used to

describe how much a neoplastic tissue resembles its ‘parent.’ As tumors progress, the tissue architecture becomes deranged, and cells become more chaotic, displaying morphological traits like high nucleus-to-cytoplasmic ratio, nuclear *hyperchromasia* (dense staining), prominent nucleoli, *pleomorphism* (high variability in size and shape), giant multinucleated cells, and frequent *mitosis* (cell division) (Figure 1.2) [119]. Poorly differentiated tumors are typically correlated with worse patient survival outcomes [12].

The term *biomarker* refers to a biological feature that we can use to indicate a clinical outcome. For example, *prognostic biomarkers* are biological features associated with good (or bad) prognosis, while *predictive biomarkers* predict response to therapy in randomized controlled trials [17]. Typically, when a histologic trait is related to outcomes in cancer, it is incorporated into the grading criteria, though this is not always the case. For example, there has been a strong focus on *tumor-infiltrating lymphocytes* (TILs) as a prognostic and predictive biomarker in breast cancer and other solid tumors in recent years [158]. This is because TILs infiltration can be a somewhat direct visualization of how well the host (patient) body can respond to the growing tumor by immune cells. Section 3.1 discusses TILs in more detail and outlines the benefits and challenges in the computational assessment of TILs; many of those issues are broadly applicable for the computational evaluation of other histologic biomarkers.

1.1.3 Key principles in neoplasia

Neoplasia occurs when a single cell divides uncontrollably, forming a clone of cells; hence neoplasia is *monoclonal*. Neoplasia starts locally, i.e., *benign*, and eventually invades surrounding tissue, i.e., becomes *malignant* (cancer). While some cancers have well-known causative agents (e.g., HPV virus-induced cervical cancer), most arise from the accumulation of various ‘hits’ or mutations that affect several cellular characteristics or phenotypes that are famously referred to as the *Hallmarks of Cancer* (Table 1.1) [62, 63].

Cancers behave differently depending on the cell of origin. *Carcinomas* originate from epithelial tissues and are the most common group of cancers. They tend to spread via lymphatics. On the other hand, *sarcomas* develop from connective tissue, are much less common, and tend to spread via blood (*hematogenous* spread). Other major cancer categories include myelomas, lymphomas, leukemias, and mixed. When a tumor spreads to distant tissues, the process is called *metastasis*

and is one of the most important indicators of poor prognosis. Different cancers have different metastatic potentials depending on site and tissue of origin, with the most common metastasis destinations being the lung, liver, bone, and brain [119, 62].

1.1.4 Breast carcinomas

More than three million women live with breast cancer in the U.S [174]. The majority of breast cancers are carcinomas. Based on morphology, breast carcinomas include many variants; the most common are *infiltrating ductal carcinoma* (which originates from breast duct epithelium) and *infiltrating lobular carcinoma* (from breast acini/glands). Before the invasion of the basement membrane (the collagen layer on which epithelium rests), carcinoma is said to be *in-situ*. Other morphological variants include *comedo* pattern (necrosis), *mucinous* change (mucous secretion), dense lymphocytic infiltration (*medullary* subtype), and many others [119, 12]. There are numerous morphological elements within a single breast cancer slide, some of which are illustrated in Figure 1.2.

Integrative genomic analysis of breast cancer identified four main subtypes, including *Luminal-A*, *Luminal-B*, *Her2-Enriched*, and *Basal* [29]. These subtypes have distinct alterations and are associated with distinct patient survival prospects [49]. Most basal breast cancers lack hormonal (estrogen and progesterone) and HER2 receptors [29] and hence do not respond to traditional hormonal therapy or targeted anti-HER2 therapy. This phenotype is called triple-negative breast cancer. *Triple-negative breast cancer* (TNBC) comprises 15% of these breast cancer cases with an average annual incidence rate of 15.5 per 100,000 women and has the poorest prognosis of all subtypes [49, 93]. One of the few available prognostic biomarkers in TNBC is TIL's; TNBC patients with dense lymphocytic infiltrates have significantly better survival outcomes [170].

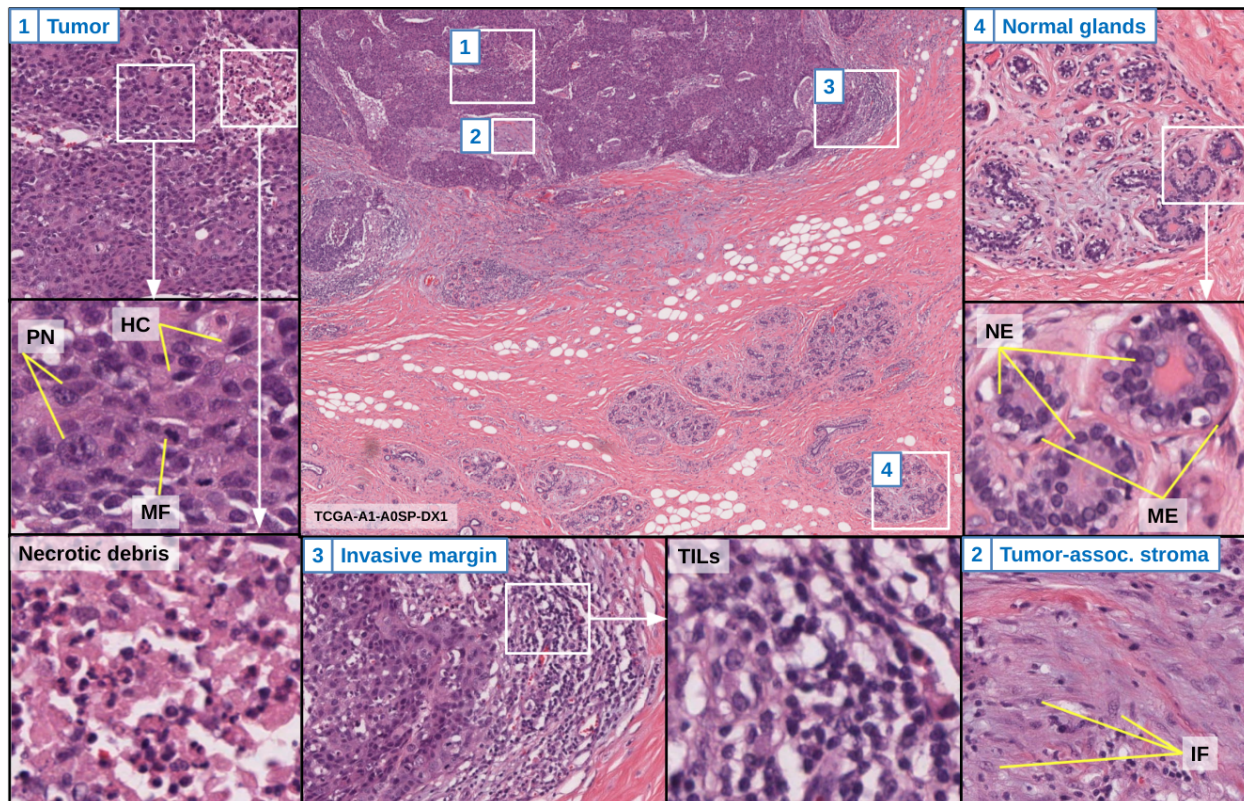


Figure 1.2: **Histologic characteristics of breast carcinomas.** This slide scan was obtained from an infiltrating ductal carcinoma of the breast from the open-access TCGA dataset. Note the differences between the tumor region in the upper corner and compare it to the central and lower portions of the field. Disturbed tissue architecture and poor differentiation are characteristic of high-grade tumors, which are, in turn, typically associated with poor patient survival outcomes. Region 1 shows a tumor region with evident architectural disruption, cell crowding, and cellular death (necrotic debris). Neoplastic nuclei show a significant degree of pleomorphism (note the variability in sizes in the left central box), hyperchromasia (HC), and prominent nucleoli (PN). A mitotic figure (MF) is also seen. Region 2 shows tumor-associated stroma, with immature fibroblasts (IF), which are plumper and less dense than normal fibroblasts. Immature fibroblasts are increasingly being recognized as a prognostic indicator in some cancers. Region 3 shows the invasive margin of the tumor, with tumor-infiltrating lymphocytes (TILs). Region 4 shows a normal breast acinus. Note how orderly the glands are, showing diverse but uniform cellular elements, including normal epithelium (NE) and myoepithelium (ME).

Table 1.1: **The hallmarks of cancer.** Based on [62, 63].

Hallmark	Explanation	Alterations
Sustaining proliferative signaling	Proteins involved in growth factor signaling can mutate to become constitutively active, sending a continuous proliferation signal. These mutated genes are called <i>oncogenes</i> .	BRAF RAS MYC Cyclins/CDK
Evading growth suppressors	Genes coding for proteins that usually prevent proliferation, <i>tumor-suppressor genes</i> , can get mutated and become dysfunctional. Usually, two mutations (maternal and paternal copies) are needed.	TP53 RB
Resisting apoptosis (cell death)	Mutations in genes which activate apoptosis when there is extensive DNA damage. Dysregulation favoring pro-apoptotic signals and direct mutations in members of the apoptotic pathway.	BCL2 TP53 Caspases
Enabling replicative immortality	Mutations can activate the enzyme <i>telomerase</i> , which prolongs telomeres (the ends of DNA) as cells divide indefinitely.	Telomerase
Inducing angiogenesis (blood vessel formation)	Like any tissue, cancer needs a blood supply to get oxygen and metabolic building blocks. Acquired mutations activate pro-angiogenic growth factors.	VEGF PDGF
Activating invasion and metastasis	Mutations that break down basement membrane (invasion) and blood vessel collagen (metastasis).	MMP Collagenase
Deregulating cellular energetics	Cancers often outgrow their blood supply and rely on <i>anaerobic respiration</i> for energy. Acquired IDH mutations alter histone and DNA methylation and block cellular differentiation.	IDH LDH
Avoiding immune destruction	Cancer cells may upregulate PD-L1 ligands, which are normally used to prevent autoimmunity.	PD-L1
Genome instability and mutation	Mutations in DNA repair genes result in higher DNA replication error rates (<i>mutation load</i>).	BRCA MSH
Tumor-promoting inflammation	Even though inflammation is critical in fighting cancer, it can have the opposite role by promoting mutations, angiogenesis, and invasion.	NFKB EGF

Section 1.2

Computational Pathology

Computational pathology is an emerging field in which imaging and data science augment and extend histopathology by automating diagnostic workflows and identifying visual elements correlated with genomic or clinical endpoints, including therapeutic response and patient survival outcomes [2]. There are ripe opportunities to incorporate state-of-the-art machine learning (ML) into the histopathology diagnostic workflow. Deep learning (DL) is a particular type of machine learning and has been a critical driving force behind many revolutionary advances in medical image analysis [104].

1.2.1 Whole-Slide Imaging

Digital Pathology, the process of acquisition, storage, retrieval, and management of scanned histology slides, is a critical component of computational pathology workflows [2, 181]. Histology slides are digitized using a scanning microscope (whole-slide scanner), and the resultant images are known as *Whole-Slide Images* (WSI), which are Gigapixels in size (typically 108x108 pixels). Most commercial scanners use a 20x or 40x magnification. WSIs are multi-resolution ‘pyramidal’ images, where each level corresponds to a particular magnification. These images are stored in the form of fixed-size tiles [2, 133, 154]. Unfortunately, various WSI formats have different standards for tile storage and metadata; this is a significant obstacle to interoperability. To tackle this issue, various groups have developed software tools like *openslide* and *large image*, and there are ongoing efforts to develop a standard DICOM format for encoding WSIs [55, 112, 72]. WSI scanners have been used in digital pathology research as early as the 1990s, but only recently has a WSI scanner been approved by the FDA for primary clinical diagnosis [48].

1.2.2 Types of analyses performed

There is a wide range of problems that computational pathology can address. These analyses differ in terms of context and ground truth requirement and include, among others:

- *Classification*: for example, a WSI from a breast cancer patient can be classified into one of

several classes like normal, benign, *in situ* lesion, and invasive carcinoma [14].

- *Semantic segmentation*: This is the process of mapping each pixel in an image to one of the distinct classes. For example, a WSI may be segmented into tumor, stromal, and necrotic regions [8, 186].
- *Object detection, classification, and segmentation*: for example, detecting various cells in bone marrow aspirates or WSI scans [36, 6, 191].
- *Regression*: For example, prediction of the expression of a specific gene or probability of a particular gene mutation given the WSI image [67, 159].
- *Survival analysis*: This is a special type of regression problem where not all outcomes are observed, and some patients may be censored (e.g., lost to follow-up). For example, predicting progression-free survival outcomes using WSI images [118, 167, 37].
- *Unsupervised clustering*: For example, the stratification of patients into distinct groups without access to ground truth labels [29, 33, 42].

The purpose of these analyses is either practical or explorative.

1.2.3 The promise and challenges of computational pathology

The incorporation of computational analysis into histopathology is potentially transformative, if not revolutionary [184, 181]. Algorithms could direct pathologists' attention to salient areas, improving efficiency in repetitive diagnostic tasks like 1. finding metastatic lesions in lymph nodes [47]; 2. detecting malignant glands [28]; 3. Narrowing down the differential diagnosis for metastases of unknown primary [109]; 4. Assessing basement membrane invasion to differentiate *in situ* and cancerous lesions [46]. Moreover, algorithms potentially increase accuracy, sensitivity, and specificity in tasks that exhibit a high degree of intra-rater and inter-rater variability, including 1. the quantification of IHC markers like Ki67 [151]; 2. counting mitotic figures [152, 16]; 3. quantitative assessment of TILs in H&E slides [10, 173]; 4. counting cells in blood smears or bone marrow aspirates [36, 191]. Other practical but less widely explored tasks include predicting clinical outcomes like survival and response to therapy [118, 167, 37, 199] and even facilitating the archival

and search of WSIs [86, 87]. Finally, the capacity of computational models to extract abstract spatial and morphological features potentially enables the discovery of histologic associations with genomic and clinical correlates and provides a path for scientific discovery unbound by the perceptual limitations of visual assessment of WSIs [44, 19]. However, despite the enormous promise, computational pathology has been relatively slow in gaining popularity and wide-scale adoption. This slow adoption is partly because computational models face unique challenges when deployed in histopathology and require unique ML solutions to address them [184, 154, 11].

1.2.4 The computational pathology workflow

The typical WSI, at scan resolution, is orders of magnitude larger than the maximum image size that can fit into standard CPU and GPU memory. Therefore, the analysis of WSIs must either be done at low resolution, with significant information loss, or using multiple tiles. The most common practice is to use a grid of non-overlapping tiles and process each tile independently, followed by an aggregation method to accumulate the results [181, 28, 77].

One of the main difficulties in computational analysis of WSI data is the significant variability in staining and color [2, 181]. This variability stems from several ‘preanalytic’ factors like staining routine and WSI scanner hardware and settings and can adversely impact downstream model performance [38, 83, 160]. Therefore, several image preprocessing steps are often used to account for this variability and train accurate and robust models. Of note are three steps: color deconvolution, color normalization, and color augmentation [178]. In addition, image processing operations like smoothing and thresholding are also commonly used, as discussed below.

Color spaces

Image preprocessing often involves the conversion of images between different color spaces [54]. Color space is a term that loosely describes an arrangement of colors. In the context of this dissertation, it is also used to describe a mapping function for the different ways in which an image can be expressed to disentangle various properties like color, intensity, and even specific histological stains. For example, we can convert the same image into:

- *Grayscale*: an $m \times n$ matrix with brightness but not color information.

- *Red-Green-Blue* (RGB) color space: an $m \times n \times 3$ matrix where each channel corresponds to variations in a different wavelength. Capture devices like WSI scanners and CCD cameras obtain images in the RGB space.
- *Hue-Saturation-Intensity* (HSI) space: An $m \times n \times 3$ matrix where channels correspond to the color hue, saturation, and luminance/intensity. It is sometimes considered a more natural representation than RGB, as it matches visual perceptual notions of color and brightness.
- *Hematoxylin and Eosin* (H&E) space: This is an $m \times n \times 2$ matrix where channels correspond to the estimated amount of H&E stains bound by tissue.

Color deconvolution

Color deconvolution, also known as *stain unmixing* in this context, is the process by which the color image from a WSI, typically in the RGB color space, is mapped to another color space whereby each channel represents one of the stains used for the slide [110, 185]. Thus, for example, the original RGB image can be converted to a two-channel H&E image through a series of algorithmic and matrix algebraic steps. Several algorithms have been proposed for color deconvolution, including Macenko's method, which is based on *principal component analysis*, Xu's method, Vahadane's method, and others [110, 195, 183].

Color normalization

Color normalization is when RGB images are corrected and mapped to a prespecified standard, showing less variability and being more suitable for computational analysis [178]. One of the simplest and most widely used color normalization methods is *Reinhard's color normalization* technique [143]. Reinhard's method simply maps the mean and standard deviation of the image histogram to a target mean and standard deviation values derived from a target image with favorable color properties. More sophisticated methods have been developed for color normalization, most commonly based on color deconvolution [110, 183]. Recently, Tellez et al. introduced a robust, data-driven color normalization method that relies on color augmentation and convolutional neural networks [178].

Color augmentation

While color normalization aims to homogenize and standardize the colors of images used in computational analysis, *color augmentation* aims to increase variability during model training to increase robustness [36, 178, 164]. These two concepts are not contradictory; if one wants to train robust models that can handle histological images with variable color properties, it makes sense to use color augmentation during training and color normalization at inference (i.e., deployment) time. A systematic analysis of the effect of color normalization and augmentation on the accuracy of deep learning models found that augmentation is more critical than normalization, but that color normalization still helps [178].

Other preprocessing and postprocessing routines

Several image processing procedures are also commonly used for storage, preprocessing, and postprocessing of pathology imaging data, most notably [72, 54]:

- *Compression*: Image compression is most often used to reduce storage space requirements and access speed. Compression can be *lossy* (introducing some artifacts and acceptable reduction in quality) or *lossless* (original image can be fully retrieved). *Discrete Cosine Transform* (DCT) is the most commonly used lossy compression method (as part of the JPEG format) and involves modeling the image as a sum of cosine functions. *Run Length Encoding* (RLE) is a standard lossless compression method representing consecutive runs of same-value pixels by the value and count. Deep learning-based compression is also used in histopathology, often using architectures called *autoencoders* [179, 100].
- *Resizing*: Frequently, image analysis models require a specific size to work with, and images are resized to match it. When the original image is larger than the desired size, *downsampling* is used. If instead, the image needs to be increased in size, interpolation is used. Commonly used methods include *bilinear interpolation* (for color or grayscale images) and *nearest-neighbor interpolation* (for images where pixel values have intrinsic meaning like ground truth masks).
- *Padding*: A set of pixels is added to the edges to allow kernels to convolve with the edge pixels

or meet a specific image size requirement without resizing. A *kernel/filter* is a small matrix of weights that encode specific image processing operations. Different padding strategies are available, depending on the use-case; *constant padding* with zeros is most commonly used, but *mirror padding* (reflecting the edge pixels) is sometimes used for a more 'natural' result.

- *Edge detection*: Most commonly, a kernel of weights is passed over (i.e., *convolved with*) the image to yield a new image where edges are highlighted. Each pixel in the new image is the weighted average of the corresponding neighborhood in the original image. Thus, the kernel size and weights determine which pixels are emphasized, and the weights kernel can be designed to highlight edges in a specific direction (e.g., horizontal edges) [131].
- *Smoothing*: Smoothing can help improve the image's visual perceptual quality, but it also improves image analysis by de-noising the image. In *mean smoothing*, a kernel is convolved with the image such that the corresponding resultant pixel is the average of the original pixel's neighborhood. *Gaussian smoothing* is also commonly used and uses a kernel with radially decreasing weights, quantized from a gaussian distribution with a predetermined standard deviation (*sigma*).
- *Thresholding*: Pixels below a specific value are set to zero, leaving only foreground pixels. *Global thresholding* is when the cutoff value is the same for all pixels in the image. In contrast, *Local (adaptive) thresholding* is when the cutoff value is determined based on each pixel's neighborhood [162]. *Otsu's method* is one of the most commonly used thresholding techniques. It clusters pixels into foreground and background by iteratively optimizing a cutoff threshold that minimizes the weighted sum of within-class variance in the foreground and background pixel histograms [132].
- *Connected component analysis*: It is often helpful to identify contiguous pixels that belong to distinct regions or objects in a binary image. Pixels are modeled as *nodes* in a *graph*, and if they are spatially contiguous, their corresponding nodes share an edge. *4-connectivity* assumes pixels are contiguous if they are immediately above, below, or beside each other (north, south, east, west), while *8-connectivity* also includes diagonal connections (e.g., north-east). Pixel values in the resultant *labeled image* denote distinct regions/objects [70].

- *Dilation and erosion*: These techniques fall under the umbrella of *morphological operations* and are used as preprocessing steps in many image processing routines. Like edge detection and smoothing, morphological operations rely on convolving the image with a pre-set filter. Dilation adds pixels to object boundaries, while erosion removes pixels. *Binary opening* involves binary erosion followed by a dilation, resulting in the removal of small objects (denoising). *Binary closing* is the opposite process and is used to remove small holes [54, 115].

Image segmentation

Image segmentation is the delineation of boundaries of various regions and objects based on properties like texture, edges, contours, and color [54, 79]. For example, we might delineate the boundaries of epithelial glands and nuclei in prostatic carcinoma lesions. Segmentation is one of the central tasks in computer vision, and there is a vast body of literature describing techniques that vary widely by application. Classical image segmentation techniques rely on various image processing operations, graph modeling, and other mathematical and algorithmic modeling techniques [54, 79]. Over the last decade, DL methods have become the *de facto* state-of-the-art for almost all biomedical image segmentation applications [104, 82]. We describe the application of DL for breast cancer WSI segmentation in Chapter 3. Classical image segmentation and ML remain highly relevant in some applications, e.g., in active learning, where inference speed is critical [100, 123]. Moreover, there is a growing interest in utilizing classical image segmentation techniques to overcome some of the limitations of data-hungry DL models [89, 7, 141, 76].

‘Handcrafted’ feature extraction

ML models typically require the manual extraction of meaningful feature vectors from images, which serve as the model input. DL models are a notable exception and can learn directly from the raw images. In the context of histopathology, handcrafted features fall into three broad categories [6, 44, 194]:

- *Shape features*: For example, nuclear area, perimeter, circularity, and boundary complexity.
- *Texture features*: Texture is a repeating pattern of pixel intensities. It is commonly calculated by analyzing the *Gray Level Colocalization Matrix* (GLCM). Each entry in a GLCM is the

probability that a pixel with a specific value i is adjacent to other pixels with a value j . The adjacency distance and direction are tunable parameters. *Haralick texture features* are calculated from the GLCM at various scales, and the resultant feature vector includes contrast, entropy, and variance, among other features [64]. Texture features are sometimes considered ‘abstract’ because they do not typically require the delineation of object boundaries or locations. Consequently, texture features are especially useful in applications like fluorescence microscopy [61, 9].

- *Spatial features*: These are features that encode contextual relationships between neighboring entities. In the context of histopathology, spatial features include 1. Relationship of cells to each other within a prespecified neighborhood radius [192, 200]; 2. The hierarchical relationship between cells and tissue compartments [19, 85]; 3. Relationship of tissue compartments to their neighbors [19]; 4. Global cell clustering patterns [156, 1].

1.2.5 Concept bottleneck models

One of the difficulties facing widespread adoption of state-of-the-art DL in medical domains is their opacity. There is a broad consensus that explainability is critical to trustworthiness, especially in clinical applications — this topic is discussed further in Sections 3.1 and 3.2 [2, 6, 11, 98, 148].

The standard application of DL models in histopathology involves the direct prediction of targets from the raw images. For example, we may predict patient survival given a WSI scan [118]. However, an alternative paradigm is beginning to emerge that combines the strong predictive power of opaque DL models and the interpretable nature of handcrafted features — a technique called Concept bottleneck modeling [92]. The fundamental idea is simple: 1. Use DL to delineate various tissue compartments and cells; 2. Extract handcrafted features that make sense to a pathologist; 3. Learn to predict the target variable, say patient survival, using an interpretable ML model that takes handcrafted features as its input. Hence, the most challenging task is handled using powerful DL models, while the terminal prediction task uses highly interpretable models. This methodology was used to discover novel biological associations, including, for example, the prognostic role of collagen entropy in breast cancer [44, 102].

Much of the work presented in this dissertation relies on concept bottleneck models. These

models require a large set of manual annotation data to train to predict the intermediate concepts. Unfortunately, manual delineation of histopathologic structures is very demanding and impractical given pathologists' time constraints. To handle this issue, we also describe the development of scalable approaches to alleviate the burden of annotation using crowdsourcing approaches.

Section 1.3

Machine learning in pathology

1.3.1 Machine learning

Definitions

Machine learning is the use of statistical models and computational algorithms to learn patterns from data. While there is a significant overlap between ‘traditional’ statistical modeling and ML, ML is almost purely data-driven [66]. When the *ground truth* (target variable, i.e., what we want to achieve) is available, this is called *supervised learning*. When ground truth is unavailable, *unsupervised learning* (e.g., clustering) is used. The critical goal of ML models is to maximize *generalization performance*, i.e., the ability of patterns learned on available data to generalize well and be applicable to future, unseen real-world data. To that end, ML models are designed to fit the training data well (i.e., *reduce bias*) while avoiding poor model performance on unseen data — i.e., reduce variance in accuracy when the same model is applied to different samples from the population. Unfortunately, the model would learn the noise in training data if bias is reduced beyond a certain point. As a result, it would not generalize well to unseen data, a phenomenon known as *overfitting*. This observation is known as the *bias-variance trade-off* and is one of the fundamental principles of ML [66]. Complex ML models are usually *regularized* to reduce the chance of overfitting, often by encouraging their learned parameters to have small values [97].

During supervised learning model development, the available data is divided into a *training set* for parameter fitting, a *validation set* for hyperparameter tuning, and a *testing set* for assessing generalization performance. Because these subdivisions are random, they can be overly optimistic in estimating generalization performance or overly pessimistic. *Cross-validation* is often performed to mitigate this, a procedure that involves cycling the training and testing sets. For example, in 10-fold cross-validation, the dataset is divided into ten parts, and each part becomes the testing set during one training cycle. Commonly used supervised ML algorithms include naive Bayes, linear regression, decision trees, random forests, support vector machines (SVM), boosting, k-nearest neighbors (KNN), and artificial neural networks [66]. Different supervised learning algorithms make various assumptions about the data and are hence suitable in other contexts. For example, lin-

ear classification models assume there is a linear classification boundary that reasonably separates different classes. Different algorithms also vary in their learning capacity and ability to generalize; decision trees are prone to overfitting, while random forests and neural networks have better generalization ability (broadly speaking).

Practical considerations

There are also practical considerations that influence the development and performance of supervised learning models, including:

- *Target variable*: Models designed for classification tasks are different from those intended for regression tasks. Regression tasks may be converted to classification tasks in specific scenarios.
- *Meaningfulness of features*: If the raw/extracted features are not meaningful, *feature engineering* combines available features. Sometimes it is difficult to extract meaningful features from the data manually. Moreover, the process often involves manually tuning many hyperparameters, with higher chances of poorer generalization. This is especially true for image data, in which feature learning with neural networks is the current state-of-the-art [104].
- *Sample size*: Simpler ML models like linear classification require fewer data points to learn, while complex neural networks are often considered *data-hungry* [8, 28, 65].
- *Dimensionality*: When the number of features relative to the sample size is high, this is a more complex learning problem known as *high-P-low-N*. *K-nearest neighbors* (KNN), in particular, suffers when applied to high dimensionality datasets.
- *Missing data*: In real-world applications, and especially in the context of clinical medicine, data is often missing. If the data are *Missing At Random* (MAR), a process called *imputation* can replace the missing values based on some criteria. For example, *mean imputation* replaces missing values with the mean, while *KNN imputation* uses data from similar samples. Sometimes, however, the data is *Missing Not At Random* (MNAR). Handling MNAR issues depends on the application, as it is often vital to know why data points are missing before naively imputing them [136, 103].

- *Inference speed*: Real-time applications require resource-efficient, time-critical prediction of new samples. This requirement often necessitates the use of less complex models [34].
- *Interpretability*: Simple algorithms like decision trees and linear models are readily interpretable. On the other hand, neural networks are typically more accurate but less interpretable [113].
- *Development cycle*: *Active learning* is used to quickly provide feedback on model performance and alter the training process [100, 123].

1.3.2 Deep learning

Deep learning refers to the practice of ML using artificial neural networks [66, 56]. *Artificial Neural Networks* are a class of ML models that can learn to extract features from raw data, i.e., *feature learning*, avoiding the need for handcrafted feature extraction and engineering [82]. This process is less biased and is partly why DL models are very popular; virtually all state-of-the-art medical applications rely on DL models [104].

Neural networks are composed of *layers* of connected functional units, often called neurons, in reference to biological nervous tissue. The inputs to each neuron are outputs from neurons from previous layers, while its output is input to the next layer. In the most basic formulation, each neuron in one layer receives input from all neurons in the previous layer, an architecture termed *fully connected feed-forward neural network* (FFN). The output of a neuron is the linear combination of its inputs (where the weight parameters are learned), followed by a nonlinear function like sigmoid, tanh, or rectifying linear units (ReLU) [122]. The nonlinearity is essential; without it, neural networks can be simplified by a linear regression model. Some nonlinearities are better than others, and a key consideration in choosing the type of nonlinearity to use is its range — e.g., sigmoid maps everything to the range [0, 1]. The final output is compared to the desired value, and a *loss value* is calculated. Next, the derivative of the loss value with respect to the weights and biases is calculated, which are 'nudged' to reduce the loss value, a process called *backpropagation* or *gradient descent* [149]. This process is repeated until *convergence*, i.e., minimal change in loss value with iterations. The amount by which parameters are nudged is called the *learning rate*. The learning rate could be fixed, but often it is adaptively reduced as epoch count (number of

times the network has seen the data) increases. An *optimizer* is an algorithm that controls other hyperparameters, most notably the learning rate, to reach convergence quickly and with a lower risk of falling into *local minima* (suboptimal but stable models). The most basic optimizer is the *gradient descent* optimizer; others like *Adam* are commonly used for better learning rate adaptation [90].

Deeper neural networks have a higher risk of *gradient diminution* (gradual reduction to zero), especially when certain nonlinearities like sigmoid are used. *Gradient explosion*, uncontrolled increase in gradient values during learning, is another issue encountered [56]. These issues, along with model learning capacity and generalizability, are all directly impacted by design choices, a.k.a. *Hyperparameters*. Tunable hyperparameters include network architecture, depth, width, nonlinearity type, and optimizer. Therefore, strategies have been devised to improve deep neural networks' learning behavior and generalization, including stochastic gradient descent, dropout, batch normalization, and many others [97, 56, 78]. In *Stochastic gradient descent*, a limited subset of data points (termed the *batch size*) is used to update the gradient. Hence, backpropagation is done multiple times during each epoch. This procedure helps avoid falling to local minima, hence increasing generalization. *Dropout* is a regularization method specific to DL, whereby weight parameters are randomly set to zero during training [169]. Optimization of neural network learning is a very active area of research but is not the focus of this dissertation.

1.3.3 Convolutional Neural Networks

Convolutional Neural Networks (CNN) are DL architectures that work with features that have some spatial dependency, like audio waves (1-dimensional signal) or images (2- or 3-dimensional signal). CNNs have revolutionized computer vision and are almost always the top-performing algorithms in pattern recognition challenges like ImageNet, CIFAR, COCO, and in the context of pathology, CAMELYON [47, 150]. While many CNN architectures include fully connected layers, most CNN layers are convolutional [96, 165, 69, 146]. In the context of CNNs, *convolutional filters* are learned weight kernels that are convolved with the image (first layer) or *feature map* (subsequent layers) to yield a new feature map. The convolutional filters may be passed over each pixel in the image, but often a *stride* is used to skip some pixels. At the edge of the image or feature map, *padding* is often used to prevent reducing size with consecutive convolution operations. The size of the

Table 1.2: A small sample of commonly used CNN architectures

Architecture	Main characteristics	Task
VGG	A very deep, uniform CNN architecture, typically composed of 16 or 19 convolutional layers.	Image classification
ResNet	'Residual' connections bypass layers and enable using very deep networks with less likelihood of gradient diminishment.	Image classification
VGG-FCN	Upconvolutions were added on top of the VGG architecture to enable semantic segmentation.	Semantic segmentation
U-Net	These contain upconvolutions and symmetric skip connections (earlier layers connected with later ones) to enable semantic segmentation with preservation of feature map detail.	Semantic segmentation
Faster R-CNN	A CNN model that is composed of 1. A region proposal network last proposes potential object locations; 2. A regression head that learns to improve object boundary predictions; 3. An object classification head.	Object detection and classification
Mask-R CNN	Same as Faster-R CNN, but with an added segmentation head that delineates object boundaries.	Faster-R CNN + object segmentation

convolutional filter determines how much of each pixel's neighborhood is taken into consideration. *Pooling* operations commonly follow convolutional filters, most often *summation pooling*, which 'summarize' the feature map and help reduce overfitting [56].

The first few layers of a CNN (and DL models) typically extract lower-level features like edges and contours. In contrast, deeper layers extract higher-level concepts like, in the context of face recognition tasks, eyes and ears [73]. This realization has led to the standard practices of transfer learning and pretraining to reduce the data volume requirements and increase the accuracy of CNNs. *Transfer learning* is the practice of learning some everyday tasks, like edge extraction, using datasets other than ours to increase accuracy in tasks where data is not abundant. For example, a CNN trained to classify images of dogs and cats can be adapted to detect metastases in lymph node biopsies. Transfer learning is frequently achieved through *pretraining*; the CNN is first trained on the non-target task, then the trained weights are copied to the new CNN to be adapted for the new task. Some weights may be 'locked' (i.e., only the last layers are tuned), in which case the first part of the CNN functions as a complex but non-adaptable feature extractor [100, 189, 175]. Alternatively, all weights may be fully tunable, such that the main benefit from pretraining is favorable weight initialization values [34, 73].

Many image recognition datasets involve image classification tasks, which is why 'classical' CNN architectures end with fully connected layers and predict a vector of probabilities for each

of the classes in the dataset. A relatively recent development has been the adaptation of CNNs to perform end-to-end segmentation. They are called *fully-convolutional networks* (FCN) as they lack fully connected layers [107]. FCNs use *up-convolutions* to obtain feature maps that match the original image size, with some loss at the edges unless some padding is used. Table 1.2 highlights some of the commonly used CNN architectures in the literature [165, 69, 146, 107, 68, 144].

1.3.4 Interpretability in deep learning

DL models are often criticized as black boxes that use complex nonlinear transformations that are not amenable to human interpretation. While there is some truth to this assertion, there have been many advances in explaining DL model decisions. Some investigators like to use the following definitions: *interpretability* is the ability to understand the model’s inner workings, while *explainability* is the ability to understand why the model makes its decision, without necessarily intuiting its inner workings. For example, many post-hoc explainability methods highlight regions within the input image that the model “looks at” when making its predictions [113]. Model accuracy and interpretability are, at least traditionally, at odds. Complex patterns need complex models to recognize, and complex models are harder to understand [148, 113]. That being said, it is essential to realize that at least some explainability is critical for the clinical adoption of CNN-based diagnostic algorithms. Explainability builds trust in model predictions, especially those directly affecting patient care and therapeutic decisions [2, 11].

We can visualize some of the inner workings of CNNs by visualizing convolutional filters directly or the *activation maps* when a specific image is used as an input. In well-trained CNNs, early activations often correspond to edges and contours (non-specific) while later filters are task-specific [73]. Alternatively, we can backpropagate the gradient when a particular image is used as the model input. This process allows us to visualize what pixels are more important for predicting that specific image. This method gives noisy results but can be improved by setting negative gradients to zero); this is called *guided backpropagation* [168]. Unfortunately, guided backpropagation is poorly discriminative; visualization tends to be similar for different predicted classes. An improved technique called *Grad-CAM* (gradient guided class activation maps) offers better class discriminative ability [161].

Recently, there has been a lot of discussion about the need for explainability methods that are

falsifiable. One of the limitations of post-hoc saliency approaches like GRAD-CAM is that they are very liable to confirmation bias: investigators can always point to some examples where the model is looking at the expected region when making its decision [148, 52]. Falsifiability is a one of the bedrocks of scientific methodology, as noted by the philosopher Karl Popper, in his landmark book *The Logic of Science* [140]:

“These considerations suggest that not the verifiability but the falsifiability of a system is to be taken as a criterion of demarcation. In other words: [...] it must be possible for an empirical scientific system to be refuted by experience.”

This critique is the reason why some investigators and thought leaders are advocating for less reliance on qualitative explainability techniques, and more reliance on any of: 1. Models that are inherently interpretable [148, 113]; 2. Explainability techniques that are quantitative and falsifiable [85]; 3. Building trust through robust validation on large patient cohorts or hospital settings [52].

In Section 3.2, we further discuss this topic, and introduce a CNN explainability technique that is quantitative, falsifiable, and better suited to the task of nucleus classification than existing approaches.

Section 1.4

Integrative computational analysis

Recent advances in biomedical data acquisition, storage, retrieval, and regulation enable the integration of multiple clinical, laboratory, and imaging modalities to improve healthcare and discover novel biological associations. The relatively recent widespread use of structured *Electronic Medical Records* (EMR), DICOM radiological imaging standard, and WSI acquisition systems enable the correlation of clinical imaging and genomic features for the same patient [33, 42, 41]. These technologies allow an in-depth holistic assessment of biomarkers and help drive better clinical decisions, paving the way for precision medicine. *Precision medicine* refers to the concept of stratification of patient cohorts into smaller and smaller subsets based on evidence-driven clinical/laboratory/imaging criteria, such that more targeted therapeutic strategies can be used to improve clinical outcomes [15].

1.4.1 Clinical outcomes and surrogate endpoints

Prognosis is at the heart of medical decision-making and is used to guide life-altering interventions, including surgery, radiation treatment, and chemotherapy. Accurate stratification of outcomes is also critical in conducting clinical trials and identifying respondents to specific treatments. The most commonly used predictive system used in the clinical setting is the *TNM staging system*, which uses a combination of tumor size, lymph node spread, and distant metastasis to stratify patients by expected prognosis [12]. *Pathologic staging*, often determined after tumor excision, is a more robust prognostic indicator than histologic grading (see Section 1.1). However, it should be noted that in central nervous system tumors like glioma, staging is not applicable; histologic grading and genomics are used instead for prognostic stratification [108].

There are several ways by which we can determine if a particular clinical, imaging or genomic feature is relevant in terms of patient care. In the context of oncology, association with patient survival outcomes is typically used, although other outcome measures like quality of life and morbidity are being increasingly recognized [22, 138, 105]. Sometimes, it is impractical to rely on *clinical endpoints* (i.e., overall survival outcomes) to identify biomarkers or determine therapy response. In

these situations, *surrogate endpoints* are used, biomarkers with a strong association with clinical endpoints. For example, *progression-free survival*—how long the patient lives without worsening the disease—is a commonly used surrogate endpoint in oncology [138].

1.4.2 Survival analysis

In the medical context, *survival analysis* refers to the modeling and prediction of adverse cancer outcomes, such as patient death, cancer progression, recurrence, or dissemination to other sites (metastasis). One key issue that complicates this task is the patient loss to follow-up, also known as *right-censoring*. *Left-censoring* (e.g., unknown exact follow-up start date) is also encountered but is less common and is not discussed further here. When patients are lost to follow-up, there is no observed event time, e.g., no known time of death, and investigators need to account for this missing data in their models.

While it may be tempting to discard patients with missing data from the dataset, this introduces bias since the loss to follow-up is usually not random; e.g., the patient got better and did not return for a follow-up visit [39, 88]. Another simple strategy is to convert the outcomes into binary form at clinically meaningful time intervals. For example, if we were primarily interested if the patient was alive at five years, we would retain patients censored after (but not before) the 5-year timepoint.

To use all data points, however, special techniques in statistical modeling have been developed to use partially observed outcomes data [24]. The most commonly-used survival analysis models, called *Cox proportional-hazards* regression, assume that the relationship between features and survival outcomes is linear [43]. Like other linear models, Cox regression is less likely to succeed as the number of features increases relative to the number of patients. Nonlinear ML techniques have been applied to survival analysis to overcome this issue, including random survival forests and neural networks [199, 39, 80].

1.4.3 Genomics

An organism's *genome* is its total deoxyribonucleic nucleic acid material (DNA), while its *transcriptome* is its ribonucleic acid material (RNA). The nuclear DNA genome encodes most cellular RNA and proteins in humans, while mitochondrial DNA encodes a few but critical sets of enzymes. Thus, in the broad sense of the term, *Genomics* studies cellular DNA or RNA sequence, expression,

and epigenetic modifications. *Proteomics* is the study of protein expression. Arguably, the most significant recent development in genomics is the (near) completion of the Human Genome Project in 2000; most of the human genome was sequenced and annotated at a total cost of \$300 million. The cost of genomic sequencing has since been dramatically reduced thanks in large part to *Next Generation Sequencing* (NGS) methods, which enable very high-throughput analysis [130, 114, 116]. The current cost is well below \$1000 per genome.

Genomic platforms and features

Genomic platforms could be classified according to what they detect into three groups: those that detect/sequence DNA, those that detect/sequence RNA, and those that detect proteins [190, 23]. In addition, a particular class of DNA sequencing platforms can also detect epigenetics modifications like methylation. The genomic techniques/modalities relevant to our discussion include:

- *Whole-genome sequencing*: Every nucleotide in the genome is sequenced. Classically, DNA sequencing was performed by ‘chain termination’ assays (*Sanger sequencing*), but most non-trivial sequencing nowadays is performed using NGS platforms [157, 57, 163]. NGS platforms sequence numerous small fragments of DNA at random sites along the genome, which are then *aligned* based on their overlap regions and their position along a known *reference genome*. Because the process is random, many overlapping fragments have to be sequenced to avoid gaps in the computationally reconstructed genome (i.e., to achieve higher *coverage*); this is known as the sequencing *depth* [166]. To perform deeper sequencing, we need to perform multiple reads to identify the same nucleotide at the same position in multiple reads. Depth is critical when looking for rare mutations or *single-nucleotide polymorphisms* (SNPs).
- *Whole-exome sequencing*: Only the exomic (protein-coding) regions are sequenced [142]. This method is an economical alternative to full-genome sequencing.
- *DNA Microarrays*: A set of probes are used to detect the presence of specific genes [71]. The genes are not sequenced, and only a pre-specified set of genes are examined. Microarrays were the standard practice before NGS and are still considered a lower-cost alternative in some settings.

- *RNA-Seq*: The total RNA (or optionally, only messenger RNA) is sequenced [193]. RNA-Seq is useful in quantifying gene expression, including multiple proteins expressed from the same DNA sequence (*alternative splice variants*).
- *Reverse Phase Protein Arrays* (RPPA): Expressed proteins are quantified using microarrays [23]. Unlike RNA-Seq, this method of surveying gene expression is less mature, and the set of proteins needs to be determined *a priori*. Nevertheless, RPPA can detect the final protein expression, including post-translational modifications that RNA-Seq does not survey.

To obtain relatively pure samples for genomic analysis, a pathologist may microscopically dissect the tumor, a process called *Laser Capture Microdissection* (LCM) [50]. The percentage of the obtained sample composed of tumor cells (instead of stromal or immune elements) is known as *tumor purity*. Tumor purity is usually inferred *a posteriori* by statistical *deconvolution* of gene expression data [197, 13].

The Cancer Genome Atlas

The Cancer Genome Atlas (TCGA) is a large-scale multi-institutional collaborative initiative to collect and analyze cancer genomic data. The TCGA dataset is one of the most extensive resources available for integrative analysis of clinical, genomic, epigenomic, proteomic, radiologic, and histopathologic data [182, 40]. The TCGA project started in 2005 and includes data from 10,000 patients and 33 tumor sites, with matched normal tissue controls. The TCGA is collected from designated medical institutions across the U.S., following standardized data collection and quality criteria.

When the project started, all TCGA samples underwent whole-genome sequencing; currently, a small fraction of samples is fully sequenced, while the rest undergo whole-exome sequencing. Many TCGA cases also have associated WSIs, pathology and radiology reports, radiology scans, proteomic data, and epigenomic methylation data. Since its publication, the TCGA has been analyzed by numerous research groups worldwide and has resulted in radical changes in how cancers are classified and, consequently, on precision medicine clinical guidelines [129, 127, 29, 128, 126]. Integrated analysis of TCGA data resulted in molecular characterization and updated clinical classification of most cancers [29, 33, 188, 30, 31]. Many groups have also performed large-scale *pan-cancer*

analysis, finding genomic themes common to multiple cancer types to understand fundamental drivers of cancer and identify cancer subtypes that are likely to respond to similar therapeutic strategies [40, 21, 74, 32].

Molecular pathway analysis

One of the goals of high-throughput genomic analysis studies is to identify biological pathways and processes responsible for observed phenotypes. A biological pathway is a collection of biological molecules that interact in specific ways to achieve a target. Thus, understanding biological processes at the level of gene sets enables a deeper understanding of biology. Genes can be grouped functionally or by their spatial proximity at the chromosome or cellular compartment. There are three generations of pathway analysis techniques: Over-

Representation Analysis, Functional Class Scoring [172, 18], and Pathway Topology [111]. These techniques typically find differences in gene expression between experimental control groups. Gene Set Enrichment Analysis (GSEA) is among the most popular approaches used (Figure 1.3) [172]. Single-sample GSEA (ssGSEA) is a GSEA adaptation that can estimate pathway activity for each sample, independent of phenotype, acting as a biologically-inspired dimension reduction technique [18].

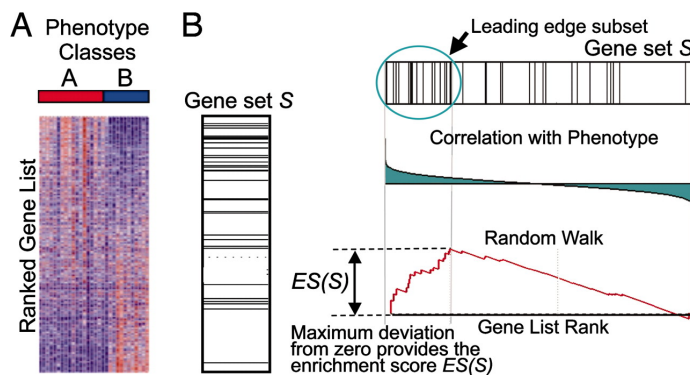


Figure 1.3: Gene Set Enrichment Analysis. In the most classical implementation of GSEA, genes from an RNA-Seq or microarray experiment are ordered by their correlation with the observed phenotype. The algorithm walks down the ordered list, keeping track of a running sum that increases when a gene in the pathway is encountered and decreases otherwise. The maximum deviation from zero when the complete gene list has been traversed is called the Enrichment Score (ES), which is usually normalized to the size of the gene set (NES). The phenotype is shuffled multiple times to obtain the empirical null distribution, which is used to calculate p-values for the NES values, then adjusted for multiple hypothesis testing. *This figure is reproduced from Subramanian et al., 2005. Copyright 2005, The National Academy of Sciences. PNAS open-access policy does not require permission for noncommercial and educational use.*

Section 1.5

Organization and summary of contributions

In this dissertation, we present a series of approaches for discovering prognostic morphologic biomarkers in invasive carcinomas of the breast. The approach we take is based on concept bottleneck modeling and is specifically designed to ensure the biomarkers are interpretable by pathologists and to reduce the chance of learning spurious correlations. The inputs to our pipeline are standard formalin-fixed paraffin-embedded H&E stained slides obtained from resections of invasive carcinomas of the breast. The slides are then digitized using commercial whole-slide image (WSI) scanners and analyzed using our computational algorithms. These convolutional neural networks (CNNs) use the input image to delineate the boundaries of all tissue region compartments and cell nuclei. A set of features describing the shape, staining, texture, and spatial context of these regions and cells are extracted and then used for modeling and predicting patient survival outcomes.

This dissertation explores four related themes that contribute to the end goal of discovering interpretable histopathology biomarkers (Table 1.3). These themes are: scalable annotation data acquisition (Chapter 2), training interpretable CNN models fit for histopathology applications (Chapter 3), discovering novel histomic associations with clinical outcomes (Chapter 4), and developing open-source software and visualization capabilities to support data collection and model development (Chapter 5).

Chapter 2 focuses on the scalable acquisition of annotation data using crowdsourcing

One of the difficulties in adopting concept bottleneck modeling is the requirement for training supervised CNNs to delineate histopathologic regions and cells as an intermediate step. This supervision necessarily consumes a large number of annotation training data, including tissue region boundary delineation and cell localization and classification. This training data is not readily available and is difficult to obtain because it requires specialized medical knowledge and is very time-consuming. Pathologists have heavy clinical demands; asking pathologists to create annotation data requires a significant financial and time investment. Hence, Chapter 2 is dedicated to the process of data generation and systematically investigates the use and adaptation of crowdsourcing

Table 1.3: **Organization of this dissertation.**

	Problem	Goal
Chapter 2: Crowdsourcing strategies for scalable curation of annotation datasets	Deep-learning models are data-hungry, yet pathologists have heavy clinical demands to create annotation data	Explore crowdsourcing strategies for scalable curation of annotation datasets
Chapter 3: Deep-learning methods for automatic detection of histopathology structures	Existing deep-learning models are not well-suited for histopathology applications, and are often uninterpretable.	Develop interpretable deep-learning models to detect histopathologic regions and cells
Chapter 4: Histopathologic correlates of clinical and genomic phenotypes	Many morphological patterns in histopathology slides are difficult to capture visually in a reliable and consistent way.	Develop computational tools to discover and quantify patterns that have prognostic value.
Chapter 5: HistomicsTK: an open-source software for computational pathology	There is a need for scalable software solutions to support annotation and WSI analysis projects	Develop open-source tools for managing distributed annotation projects and visualizing algorithmic results

approaches for the scalable generation of annotation data in histopathology. Chapter 2 is composed of the following sections:

- *Amgad, M., Elfandy, H., Hussein, H., Atteya, L. A., Elsebaie, M. A., Abo Elnasr, L. S., Sakr, R. A., Salem, H. S., Ismail, A. F., Saad, A. M., et al. (2019). Structured crowdsourcing enables convolutional segmentation of histology images. Bioinformatics, 35(18):3461–3467.*

Section 2.1 describes the development and validation of a *structured* crowdsourcing approach for delineating histopathologic tissue regions by a crowd of 25 medical students and graduates under the supervision of practicing pathologists. We propose that two modes of data collection are used synergistically: a high-volume single-rater dataset and a substantially smaller multi-rater dataset for estimation of interrater variability and participant reliability. We were able to collect 20,000 annotations of tissue regions, which were made public for use by the scientific community. We showed that non-pathologists are reliable annotators of visually-distinctive patterns and that their annotations enable training highly-accurate supervised CNN models.

- *Amgad, M., Atteya, L. A., Hussein, H., Mohammed, K. H., Hafiz, E., Elsebaie, M. A.,*

Alhusseiny, A. M., AlMoslemany, M. A., Elmatboly, A. M., Pappalardo, P. A., et al. (2021b). NuCLS: A scalable crowdsourcing, deep learning approach and dataset for nucleus classification, localization and segmentation. arXiv preprint arXiv:2102.09099.

Section 2.2 describes an extension of the approach used for tissue region crowdsourcing to the context of nucleus annotation. We developed a technique for producing a set of *bootstrapped* algorithmic suggestions using existing region labels and heuristic nucleus segmentation algorithms. These suggestions were then displayed to non-pathologists, who were asked to annotate multiple sets of data under different conditions. We collected and published over 200,000 annotations of nuclear detection, segmentation, and classification data. We found that these algorithmic suggestions scale the annotation process and improve the classification labels produced by non-pathologists. Additionally, we introduced a framework for estimating participant reliability and inferring a single truth from multi-rater data and showed that the number of annotators per image needed for accurate aggregate data is class-dependent.

- *Dudgeon, S.N., Wen, S., Hanna, M.G., Gupta, R., Amgad, M., Sheth, M., Marble, H., Huang, R., Herrmann, M.D., Szu, C.H. and Tong, D., 2020. A pathologist-annotated dataset for validating artificial intelligence: a project description and pilot study. J Pathol Inform, 2021;12:45.*

Section 2.3 describes a collaborative pilot study that we participated in, led by the U.S. Food and Drug Administration (FDA). The study investigated the use of crowdsourcing approaches for the decentralized collection of annotation data for regulatory purposes. Unlike the previous two projects, the annotations produced here were only produced by pathologists, and the workflow was specifically tailored to enable *validation* of algorithms that detect Tumor-Infiltrating Lymphocytes (TILs) in invasive carcinomas of the breast. Annotations produced using this approach are kept internally by the FDA for standardized assessment of algorithms.

Chapter 3 discusses automated detection of histopathologic structures using custom CNNs

Once we had the annotation data from Chapter 2, we were able to use it to train supervised CNN models to automatically delineate all histopathologic tissue regions and nuclei in scanned

WSIs from invasive breast cancer. Our CNN modeling focused on two issues. First, we adapted general-purpose computer vision models to be well-suited for histopathology applications. Second, we wanted to ensure that the models produce outputs and explanations that map well to concepts that are used or understood by practicing pathologists. Chapter 3 is divided into the following sections:

- **Amgad, M., Stovgaard, E. S., Balslev, E., Thagaard, J., Chen, W., Dudgeon, S., Sharma, A., Kerner, J. K., Denkert, C., Yuan, Y., et al. (2020).** *Report on computational assessment of tumor-infiltrating lymphocytes from the International Immuno-oncology Biomarker Working Group. NPJ breast cancer, 6(1):1–13.*

Section 3.1 is a set of recommendations that we published with the International Immuno-Oncology Working Group (also known as the TILs Working Group). The publication includes an in-depth discussion of various considerations for computational assessment of TILs in a manner that is consistent with visual scoring guidelines. The TILs score is defined as the fraction of stroma within the tumor bed that is occupied by TILs cells. Hence, computational scoring requires delineation of cancer regions, stromal regions, and TILs nuclei. In addition, a number of confounders like necrotic regions and normal ducts and acini need to be delineated and excluded from the calculation. This makes these recommendations relevant not only for the specific application of TILs scoring but to our broad endeavor.

- **Amgad, M., Atteya, L., Hussein, H., Mohammed, K. H., Hafiz, E., Elsebaie, M. A., Mobadersany, P., Manthey, D., Gutman, D. A., Elfandy, H., et al. (2021).** *Explainable nucleus classification using decision tree approximation of learned embeddings. Bioinformatics.*

Section 3.2 describes our CNN modeling for detection, segmentation, and classification of nuclei using crowdsourced annotation data. We focus on nuclear, not cellular, detection since nuclei are well-defined and much easier to detect visually than cell boundaries in H&E stained slides. Nuclear detection datasets differ from "standard" object detection datasets in many ways; most importantly, each image contains numerous objects, and they tend to be very small compared to the size of the image. Additionally, the features that differentiate

various nucleus classes tend to be global object-level characteristics (such as nuclear size, shape, and chromatin distribution), as opposed to the local characteristics typically used in standard datasets (like the distinguishing shape of a dog’s nose). We adapt object detection modeling approaches for optimal nucleus detection and classification, and we introduce a new explainability technique that provides explanations referencing nucleus-level characteristics to provide sensible and quantitative explanations for model decisions.

- **Amgad, M., Sarkar, A., Srinivas, C., Redman, R., Ratra, S., Bechert, C. J., Calhoun, B. C., Mrazek, K., Kurkure, U., Cooper, L. A., et al. (2019).** *Joint region and nucleus segmentation for characterization of tumor-infiltrating lymphocytes in breast cancer. In Medical Imaging 2019: Digital Pathology, volume 10956, page 109560M. International Society for Optics and Photonics.*

Section 3.3 describes a CNN modeling approach that we developed to enforce compatibility between tissue region and cell type predictions. There are two disadvantages to using independent CNN models for delineating tissue regions and for detecting/classifying nuclei. First, the approach is not efficient since the two independent models do not share parameters and redundantly extract many of the same visual building blocks like edges. Second, they can produce incompatible predictions, such as a fibroblast cell within an epithelial nest. In this section, we propose an approach based on ground truth engineering to overcome this limitation. We validate the results using a cohort of 120 patients from the Cleveland Clinic.

- *MuTILs: explainable, multiresolution computational scoring of Tumor-Infiltrating Lymphocytes in breast carcinomas using clinical guidelines*

Section 3.4 is an extension of the premise from Section 3.3, with an improved and more disciplined modeling strategy. We used a multi-task, multiresolution learning approach to simultaneously delineate tissue regions and nuclei in WSIs. Our CNN model passes down information from the low- to high-resolution branch and imposes a set of predefined biological constraints to ensure the predictions are sensible. We show that this approach results in accurate predictions, and we validate its ability to computationally score TILs in infiltrating ductal carcinomas (in general) and in Her2-enriched breast carcinomas from The Cancer

Genome Atlas (TCGA) dataset.

Chapter 4 explores histopathologic correlates of clinical and genomic phenotypes

After having developed the capability to automatically detect all salient tissue regions and cells in scanned WSIs, we were now able to extract a set of global, morphological, and spatial contextual features to summarize pertinent information in the tumor microenvironment. These features were, by design, interpretable. For example, we described the mean of- and variance in size, shape, and staining of tissue regions and cells, cellular clustering, and co-localization, the composition of the neighborhood at the margin of epithelial nests, among other features. These features were then analyzed to discover associations with clinical and other phenotypes.

- *Histomic Prognostic Score: a computational morphologic signature with independent prognostic value in invasive carcinomas of the breast*

Section 4.1 systematically explores morphological features that can predict patient survival in invasive carcinomas of the breast. These *histomic* features were categorized into biological themes and sub-themes, then entered into a regularized Cox proportional hazards regression model to produce the *Histomic Prognostic Score* (HPS), which was discretized to produce three *Histomic Prognostic Groups* (HPG). Akin to commonly-used gene panels like *PAM-50* and *OncoTypeDx*, this scoring system can help identify high- and low- risk patients. HPS only requires an H&E WSI scan and the three standard IHC breast panel markers: Estrogen Receptor expression (ER), Progesterone Receptor expression (PR), and Her2 growth receptor overexpression (Her2). We show that HPS has a stronger prognostic value than a baseline model using manual grading and IHC markers, independent of pathologic stage, tumor size, patient age, race, cancer detection method, expression of basal markers, and treatment. The results are validated using a large cohort of 1,655 patients obtained from the Cancer Prevention Study II (CPS-II), a long-term prospective cohort study organized by the American Cancer Society. Additionally, we validated the results using an independent cohort of 971 patients with invasive breast cancer from the TCGA dataset.

- *Yang, X., Amgad, M., Cooper, L. A., Du, Y., Fu, H., and Ivanov, A. A. (2020). High expression of MKK3 is associated with worse clinical outcomes in African American breast*

cancer patients. Journal of translational medicine, 18(1):1–19.

Section 4.2 describes an integrative study in which we examined the association between histomic features with gene expression data from African American patients with invasive breast cancer from the TCGA dataset. Our contribution to this collaborative work was to show that the overexpression of Mitogen-Activated Protein Kinase Kinase 3 (MKK3), and its binding partner MYC, is associated with features of cancer aggression in WSI scans, including a high tumor-to-stroma ratio and confluent tumor nests.

Chapter 5 introduces software tools for data collection, visualization and model development

Finally, we describe open-source software tools that we developed along the way to support all of the aforementioned analyses. We developed multiple software tools for this purpose, including direct and indirect contributions to [HistomicsTK](#), [HistomicsUI](#), [large_image](#), and [Histolab](#).

In Chapter 5 we describe our contributions to one software library that was critical to the work presented here: HistomicsTK. Over the span of 2.5 years, we worked closely with developers from the company Kitware, who maintain HistomicsTK, and we provided many indirect contributions in the form of beta testing, feature suggestions, and bug reports. However, this chapter only focuses on our *direct* contributions to the software library.

Chapter 6 provides a set of overall conclusions and future research directions

Throughout this dissertation, we describe the conclusions and limitations of each set of experimental results independently. In addition, at the end of the dissertation, we provide a set of overall conclusions, limitations, and suggestions for future research. These conclusions and recommendations are based on the aggregate impression and expertise we gained as we conducted various experiments.

Section 1.6

List of publications

We completed a wide variety of projects in computational pathology throughout this 5+ year PhD journey. A number of these projects were collaborative in nature, and only the most important contributions were presented in this document. The full list of publications is provided below.

1.6.1 Peer reviewed publications

- **Amgad, M.**, Atteya, L.A., Hussein, H., Mohammed, K.H., Hafiz, E., Elsebaie, M.A., Mobadersany, P., Manthey, D., Gutman, D.A., Elfandy, H. and Cooper, L.A., 2021. Explainable nucleus classification using Decision Tree Approximation of Learned Embeddings. *Bioinformatics*.
- Dudgeon, S.N., Wen, S., Hanna, M.G., Gupta, R., **Amgad, M.**, Sheth, M., Marble, H., Huang, R., Herrmann, M.D., Szu, C.H. and Tong, D., 2020. A pathologist-annotated dataset for validating artificial intelligence: a project description and pilot study. *J Pathol Inform*, 2021;12:45.
- López-Pérez, M., **Amgad, M.**, Morales-Álvarez, P., Ruiz, P., Cooper, L.A., Molina, R. and Katsaggelos, A.K., 2021. Learning from crowds in digital pathology using scalable variational Gaussian processes. *Scientific reports*, 11(1), pp.1-9.
- Farris, A.B., Vizcarra, J., **Amgad, M.**, Cooper, L.A., Gutman, D. and Hogan, J., 2021. Artificial intelligence and algorithmic computational pathology: an introduction with renal allograft examples. *Histopathology*, 78(6), pp.791-804.
- Farris, A.B., Vizcarra, J., **Amgad, M.**, Cooper, L.A.D., Gutman, D. and Hogan, J., 2021. Image Analysis Pipeline for Renal Allograft Evaluation and Fibrosis Quantification. *Kidney International Reports*.
- Lee, S., **Amgad, M.**, Mobadersany, P., McCormick, M., Pollack, B.P., Elfandy, H., Hussein, H., Gutman, D.A. and Cooper, L.A., 2021. Interactive classification of whole-slide imaging data for cancer researchers. *Cancer Research*, 81(4), pp.1171-1177.

- Yang, X., **Amgad, M.**, Cooper, L.A., Du, Y., Fu, H. and Ivanov, A.A., 2020. High expression of MKK3 is associated with worse clinical outcomes in African American breast cancer patients. *Journal of translational medicine*, 18(1), pp.1-19.
- Hudeček, J., Voorwerk, L., van Seijen, M., Nederlof, I., de Maaker, M., van den Berg, J., van de Vijver, K.K., Sikorska, K., Adams, S., Demaria, S., Viale, G., et al., 2020. Application of a risk-management framework for integration of stromal tumor-infiltrating lymphocytes in clinical trials. *NPJ breast cancer*, 6(1), pp.1-8.
- Kos, Z., Roblin, E., Kim, R.S., Michiels, S., Gallas, B.D., Chen, W., van de Vijver, K.K., Goel, S., Adams, S., Demaria, S. and Viale, G., et al., 2020. Pitfalls in assessing stromal tumor infiltrating lymphocytes (sTILs) in breast cancer. *NPJ Breast Cancer*, 6(1), pp.1-16.
- **Amgad, M.**, Stovgaard, E.S., Balslev, E., Thagaard, J., Chen, W., Dudgeon, S., Sharma, A., Kerner, J.K., Denkert, C., Yuan, Y. and AbdulJabbar, K., 2020. Report on computational assessment of tumor infiltrating lymphocytes from the International Immuno-Oncology Biomarker Working Group. *NPJ breast cancer*, 6(1), pp.1-13.
- Chandradevan, R., Aljudi, A.A., Drumheller, B.R., Kunananthaseelan, N., **Amgad, M.**, Gutman, D.A., Cooper, L.A. and Jaye, D.L., 2020. Machine-based detection and classification for bone marrow aspirate differential counts: initial development focusing on nonneoplastic cells. *Laboratory Investigation*, 100(1), pp.98-109.
- Lee, S., **Amgad, M.**, Masoud, M., Subramanian, R., Gutman, D. and Cooper, L., 2019, November. An Ensemble-based Active Learning for Breast Cancer Classification. In 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (pp. 2549-2553). IEEE.
- **Amgad, M.**, Elfandy, H., Hussein, H., Atteya, L.A., Elsebaie, M.A., Abo Elnasr, L.S., Sakr, R.A., Salem, H.S., Ismail, A.F., Saad, A.M. and Ahmed, J., 2019. Structured crowdsourcing enables convolutional segmentation of histology images. *Bioinformatics*, 35(18), pp.3461-3467.
- Yousefi, S., Shaban, A., **Amgad, M.** and Cooper, L., 2019. Learning Cancer Outcomes from Heterogeneous Genomic Data Sources: An Adversarial Multi-task Learning Approach. *ICML*

2019 Workshop AMTL (Open Review).

- **Amgad, M.**, Sarkar, A., Srinivas, C., Redman, R., Ratra, S., Bechert, C.J., Calhoun, B.C., Mrazek, K., Kurkure, U., Cooper, L.A. and Barnes, M., 2019, March. Joint region and nucleus segmentation for characterization of tumor infiltrating lymphocytes in breast cancer. In *Medical Imaging 2019: Digital Pathology* (Vol. 10956, p. 109560M). International Society for Optics and Photonics.
- Elsebaie, M.A., **Amgad, M.**, Elkashash, A., Elgebaly, A.S., Shash, E. and Elsayed, Z., 2018. Management of low and intermediate risk adult rhabdomyosarcoma: a pooled survival analysis of 553 patients. *Scientific reports*, 8(1), pp.1-12.
- Mobadersany, P., Yousefi, S., **Amgad, M.**, Gutman, D.A., Barnholtz-Sloan, J.S., Vega, J.E.V., Brat, D.J. and Cooper, L.A., 2018. Predicting cancer outcomes from histology and genomics using convolutional networks. *Proceedings of the National Academy of Sciences*, 115(13), pp.E2970-E2979.
- Nalisnik, M., **Amgad, M.**, Lee, S., Halani, S.H., Vega, J.E.V., Brat, D.J., Gutman, D.A. and Cooper, L.A., 2017. Interactive phenotyping of large-scale histology imaging data with HistomicsML. *Scientific reports*, 7(1), pp.1-12.
- Yousefi, S., Amrollahi, F., **Amgad, M.**, Dong, C., Lewis, J.E., Song, C., Gutman, D.A., Halani, S.H., Vega, J.E.V., Brat, D.J. and Cooper, L.A., 2017. Predicting clinical outcomes from large scale cancer genomic profiles with deep survival models. *Scientific reports*, 7(1), pp.1-11.

1.6.2 Preprints

- Zhang, L., **Amgad, M.** and Cooper, L.A., 2021. A Histopathology Study Comparing Contrastive Semi-Supervised and Fully Supervised Learning. arXiv preprint arXiv:2111.05882.
- **Amgad, M.**, Atteya, L.A., Hussein, H., Mohammed, K.H., Hafiz, E., Elsebaie, M.A., Al-husseiny, A.M., AlMoslemany, M.A., Elmatboly, A.M., Pappalardo, P.A. and Sakr, R.A., 2021. NuCLS: A scalable crowdsourcing, deep learning approach and dataset for nucleus classification, localization and segmentation. arXiv preprint arXiv:2102.09099.

-
- Yousefi, S., Shaban, A., **Amgad, M.**, Chandradevan, R. and Cooper, L.A., 2019. Learning clinical outcomes from heterogeneous genomic data sources. arXiv preprint arXiv:1904.01637.
 - **Amgad, M.**, Fouad, Y.A., Elsebaie, M.A, 2019. Approach to a highly-virulent emerging viral epidemic: A thought experiment and literature review. PeerJ Preprints. 7:e27518v1.

Chapter 2

Crowdsourcing strategies for scalable curation of annotation datasets

This chapter summarizes crowdsourcing approaches for scalable curation of accurate annotation data for training and validating computer vision models in histopathology. The work is divided into three sections, composed of the following publications:

- **Amgad, M., Elfandy, H., Hussein, H., Atteya, L. A., Elsebaie, M. A., Abo Elnasr, L. S., Sakr, R. A., Salem, H. S., Ismail, A. F., Saad, A. M., et al. (2019).** *Structured crowdsourcing enables convolutional segmentation of histology images. Bioinformatics, 35(18):3461–3467.*
- **Amgad, M., Atteya, L. A., Hussein, H., Mohammed, K. H., Hafiz, E., Elsebaie, M. A., Alhusseiny, A. M., AlMoslemany, M. A., Elmatboly, A. M., Pappalardo, P. A., et al. (2021b).** *NuCLS: A scalable crowdsourcing, deep learning approach and dataset for nucleus classification, localization and segmentation. arXiv preprint arXiv:2102.09099.*
- **Dudgeon, S.N., Wen, S., Hanna, M.G., Gupta, R., Amgad, M., Sheth, M., Marble, H., Huang, R., Herrmann, M.D., Szu, C.H. and Tong, D., 2020.** *A pathologist-annotated dataset for validating artificial intelligence: a project description and pilot study. J Pathol Inform, 2021;12:45.*

Section 2.1

Structured crowdsourcing enables convolutional segmentation of histology images

This section is an exact reproduction of the following open-access journal paper:

Amgad, M., Elfandy, H., Hussein, H., Atteya, L. A., Elsebaie, M. A., Abo Elnasr, L. S., Sakr, R. A., Salem, H. S., Ismail, A. F., Saad, A. M., et al. (2019). Structured crowdsourcing enables convolutional segmentation of histology images. Bioinformatics, 35(18):3461–3467.


The open-access dataset is downloadable from the [BCSS repository](#) on Github.

An abstract version was also presented at the 2018 Pathology Visions Conference (San Diego, CA):

Amgad, M., Elfandy, H., Khallaf, H. H., Beezley, J., Chittajallu, D. R., Manthey, D., et al. (2019). Hierarchical Crowdsourcing for Generating Large-Scale Annotations of Histopathology. Journal of Pathology Informatics, 10(10).

Bioimage informatics

Structured crowdsourcing enables convolutional segmentation of histology images

Mohamed Amgad ¹, Habiba Elfandy², Hagar Hussein³,
Lamees A. Atteya⁴, Mai A. T. Elsebaie⁵, Lamia S. Abo Elnasr⁶,
Rokia A. Sakr⁶, Hazem S. E. Salem⁵, Ahmed F. Ismail⁷, Anas M. Saad⁵,
Joumana Ahmed³, Maha A. T. Elsebaie⁵, Mustafijur Rahman⁸,
Inas A. Ruhban⁹, Nada M. Elgazar¹⁰, Yahya Alagha³,
Mohamed H. Osman¹¹, Ahmed M. Alhusseiny¹⁰, Mariam M. Khalaf¹²,
Abo-Alela F. Younes⁵, Ali Abdulkarim³, Duaa M. Younes⁵,
Ahmed M. Gadallah⁵, Ahmad M. Elkashash³, Salma Y. Fala¹³,
Basma M. Zaki¹³, Jonathan Beezley¹⁴, Deepak R. Chittajallu¹⁴,
David Manthey¹⁴, David A. Gutman¹⁵ and Lee A. D. Cooper ^{1,16,*}

¹Department of Biomedical Informatics, Emory University School of Medicine, Atlanta, GA 12065, USA,

²Department of Pathology, National Cancer Institute and ³Department of Medicine, Cairo University, Cairo 12613,

Egypt, ⁴Egyptian Ministry of Health, Cairo 11514, Egypt, ⁵Department of Medicine, Ain Shams University, Cairo

11566, Egypt, ⁶Department of Medicine, Menoufia University, Menoufia 32511, Egypt, ⁷Department of Pathology,

Medical Research Institute, Alexandria University, Alexandria 21131, Egypt, ⁸Department of Medicine, Chittagong

University, Chittagong 4331, Bangladesh, ⁹Department of Medicine, Damascus University, Damascus 97089, Syria,

¹⁰Department of Medicine, Mansoura University, Mansoura 35511, Egypt, ¹¹Department of Medicine, Zagazig

University, Zagazig 44511, Egypt, ¹²Department of Medicine, Batterjee Medical College, Jeddah 21442, Saudi

Arabia, ¹³Department of Medicine, Suez Canal University, Ismailia 41523, Egypt, ¹⁴Kitware Inc., Clifton Park, NY

12065, USA, ¹⁵Department of Neurology, Emory University School of Medicine and ¹⁶Department of Biomedical

Engineering, Emory University, Atlanta, GA 30322, USA

*To whom correspondence should be addressed.

Associate Editor: Robert Murphy

Received on July 18, 2018; revised on December 30, 2018; editorial decision on January 31, 2019; accepted on February 5, 2019

Abstract

Motivation: While deep-learning algorithms have demonstrated outstanding performance in semantic image segmentation tasks, large annotation datasets are needed to create accurate models. Annotation of histology images is challenging due to the effort and experience required to carefully delineate tissue structures, and difficulties related to sharing and markup of whole-slide images.

Results: We recruited 25 participants, ranging in experience from senior pathologists to medical students, to delineate tissue regions in 151 breast cancer slides using the Digital Slide Archive. Inter-participant discordance was systematically evaluated, revealing low discordance for tumor and stroma, and higher discordance for more subjectively defined or rare tissue classes. Feedback provided by senior participants enabled the generation and curation of 20 000+ annotated tissue regions. Fully

convolutional networks trained using these annotations were highly accurate (mean AUC=0.945), and the scale of annotation data provided notable improvements in image classification accuracy.

Availability and Implementation: Dataset is freely available at: <https://goo.gl/cNM4EL>.

Contact: lee.cooper@emory.edu

Supplementary information: [Supplementary data](#) are available at *Bioinformatics* online.

1 Introduction

Accurate segmentation of tissue regions in histology images is a challenging problem with important applications in computational pathology. The ability to accurately delineate tissue regions can provide important information for computational diagnosis, prognosis, assessments of treatment response and investigations of cancer biology. The problem of *semantic segmentation*, or exhaustive pixel-level classification of tissues, is particularly challenging. While deep-learning methods have demonstrated promising results in general semantic image segmentation problems, these encoder-decoder convolutional architectures require large training datasets to generalize well. Generating annotated histology datasets with adequate scale presents significant challenges, especially when careful delineation of regions or structures is required, and the lack of annotated histology remains a significant barrier in the growth of computational pathology. Semantic segmentation is particularly challenging, as complete labeling of the scene is required. Generating a meaningful number of annotations requires engaging with multiple experts, and even experienced pathologists will exhibit some inter-rater discordance. Annotations need to be captured on many images, as remarkable histologic variations can be observed even within a single lab, and variations in tissue processing (fixation, staining, mounting) and imaging have a strong influence on image texture and color. Data augmentation techniques are often used when training networks to simulate this variation by artificially manipulating the color and contrast of images with some success, reducing the annotation burden. Still, a pathologist with significant clinical demands often cannot produce enough annotations on their own to adequately train deep-learning models for challenging applications like semantic segmentation. Interfaces for viewing and annotating whole-slide histology images, collaborative review and data management are also a critical element in engaging pathologists to scale the production of accurate ground truth.

Crowdsourcing has been extensively used in general non-medical tasks, and has been shown to markedly speed and scale the process of image annotation (Su, 2012). In the life sciences, crowd-based approaches based on gamification or micropayments enabled successful scaling of biological annotations (Hughes et al., 2018). In pathology, however, the value of crowdsourcing is not immediately apparent due to the complexity and subjectivity of tasks, scale of whole-slide scans (necessitating custom large-scale viewers and annotation platforms), and domain expertise needed; a recent systematic review found that almost all crowdsourcing articles in the pathology literature focused on malaria diagnosis and relatively simple scoring of immunohistochemical biomarkers (Alialy et al., 2018). Recent work has established the value of crowdsourcing for nucleus detection and segmentation microtasks in hematoxylin and eosin stained images, and that research fellows and some non-pathologists (NPs) are able to reach acceptable concordance with senior pathologists (SPs) (Irshad et al., 2015, 2017). This work was based on a limited number of slides (10), focused on small regions of interest (400×400 up to 800×800 pixels), and did not explore more challenging tasks such as semantic segmentation. Moreover, this work did not investigate how to organize participants and leverage their various experience levels, and how technology can facilitate feedback and

collaboration between more and less experienced participants to improve crowdsourcing efficiency and accuracy.

To address some of these issues, we investigate the use of crowdsourcing in the context of semantic segmentation of breast cancer images. This task is widely regarded as the most laborious and challenging type of ground truth generation (Kovashka et al., 2016). We describe our experience using web-based technology to facilitate an international crowdsourcing effort, and in expanding this effort to include junior pathology residents (JPs) and medical students. We also describe how training and directed feedback using web-based tools like the Digital Slide Archive (DSA) can be instrumental in streamlining the annotation and review process. Our annotation efforts focus on triple-negative breast cancer (TNBC), an aggressive genomic subtype that comprises ~15% of breast cancer cases (Plasilova et al., 2016).

2 Materials and methods

2.1 Dataset description

The dataset used in this study consists of 151 hematoxylin and eosin stained whole-slide images (WSIs) corresponding to 151 histologically-confirmed breast cancer cases. These images of formalin-fixed paraffin-embedded tissues were acquired from the Cancer Genome Atlas, with triple-negative status determined from clinical data files. A representative region of interest (ROI) was selected within each slide by the study coordinator, a medical doctor, and approved by a SP. The mean ROI size was 1.18 mm² (SD = 0.80 mm²). ROIs were selected to be representative of predominant region classes and textures within each slide. Very large ROIs were avoided to prevent degradation in quality due to participant fatigue (Irshad et al., 2015). Regions with high tumor density were selected whenever possible, for two reasons: (i) to maximize the proportion of ROI occupied by tumor; (ii) to minimize the need to distinguish normal/inflammatory cells and cancerous tissue, an exhaustive process requiring expertise not expected from NPs.

2.2 Participant recruitment and training

The study workflow is illustrated in Figure 1. Research interest groups on social media (including Facebook and LinkedIn) were used to recruit participants, who were asked to submit a resume and brief motivation statement to the study coordinator. A total of 25 participants, including 20 medical students, 3 JPs and 2 SPs were selected during recruitment. Throughout this manuscript, we use the following notation to denote the various participant classes: SP (senior resident or faculty); JP; NP. JPs were defined as pathology residents who have not finished their second year of residency training. Participants underwent a training session, composed of introductory videos and a detailed document describing guidelines, histological patterns to annotate, common pitfalls as well as instructions for using the DSA interface. [Supplementary Table S1](#) illustrates sample instructions provided to participants to help improve and standardize the annotation process. Slack, an online team communication tool, was used for NPs to ask questions and to receive feedback from other participants and pathologists. Extensive feedback was

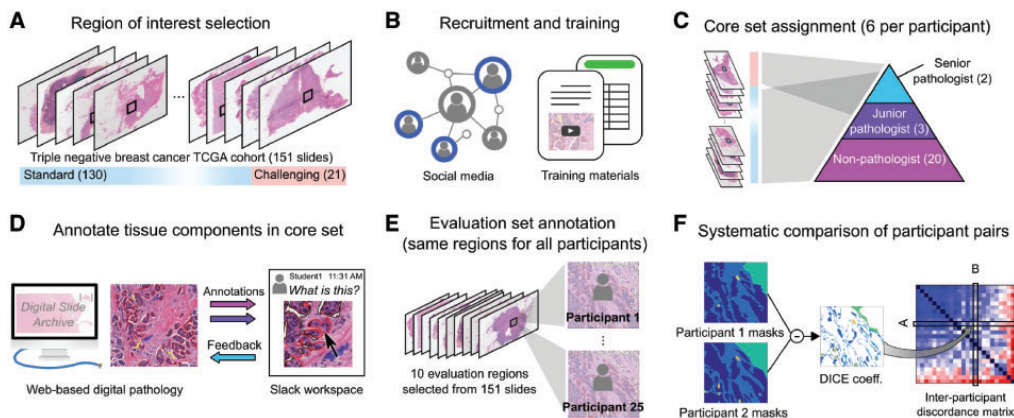


Fig. 1. Study overview. (A) Slides from the TNBC cohort were reviewed for difficulty and the study coordinator selected a single representative ROI in each slide. (B) Participants were recruited on social media from medical student interest groups. Documentation and instructional videos were developed to train participants in breast cancer pathology and the use of DSA annotation tools. A spreadsheet lists slide-level descriptions of histologic features for each of the 151 images to aid in training. (C) Participants were each assigned six slides based on experience. Challenging slides were assigned to faculty/pathology residents, while standard slides were distributed among all participants. (D) The DSA was used by participants to draw the outlines of tissue regions in their assigned slides/ROIs. A Slack workspace enabled less experienced users to ask questions and receive guidance from the more experienced users. (E) Ten evaluation ROIs were identified in the slides and were annotated by all participants in an unsupervised manner to enable inter-participant comparisons. (F) Agreement between each pair of participants was evaluated using the Dice coefficient to generate an inter-participant discordance matrix

provided on the first slide annotated by a participant, serving as a *de-facto* practical component of their training.

2.3 Structured crowdsourcing

We use the term *structured crowdsourcing* to refer to systematic assignment of tasks based on participant experience and expertise. SPs assisted in mentoring and correcting annotations made by NPs. Two types of ROIs were annotated: (i) a ‘core’ set comprising 151 large ROIs to be used for training and validating algorithms (ALs) and (ii) an ‘evaluation’ set comprising 10 smaller ROIs used to evaluate inter-participant concordance.

Each participant was asked to annotate 5–6 ROIs from the core set (uniquely assigned to the participant) and all 10 evaluation ROIs. ROIs from challenging slides (21 total) were assigned to SPs and JPs, whereas all other ROIs (130 total) were evenly distributed among the participants. Slides were considered challenging if a considerable fraction of the ROI was occupied by uncommon features like extensive tumor cell vacuolation, stromal epithelialization or stromal hyalinization. Throughout the study, SPs provided feedback and made corrections to the core slides annotated by the participants. Feedback was not provided on evaluation set ROIs to avoid biasing the analysis of inter-participant concordance.

2.4 The DSA annotation interface

The DSA is an open-source web-based digital pathology platform for the management, visualization and annotation of WSI datasets (Gutman et al., 2013, 2017). A Docker software container along with instructions for creating a DSA instance is available at: <https://github.com/DigitalSlideArchive/digitalslidearchive.info>. Figure 2 shows a screenshot of the DSA interface used for annotation. This interface organizes annotations by region class (e.g. tumor, necrosis). Each class has a style that defines the rendering properties for its annotations including class names and boundary colors. These styles were pre-defined in a template by the study coordinator in consultation with an SP, and serve to improve the consistency of

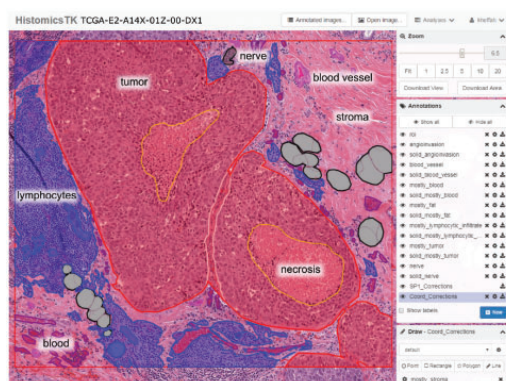


Fig. 2. Screenshot of the DSA and HistomicsTK web interface. The main viewport allows panning and zooming within the slide. Annotations are grouped by class into layers (middle right panel) whose style properties like color and fill can be adjusted (bottom right panel). Other features include: controlling annotation transparency, an interactive mode to highlight individual annotations, and ability to download the WSI, regions of interest or annotations. Annotation properties can also be programmatically manipulated using the DSA API

annotations across participants and to facilitate review. The DSA also provides a REST API for programmatic management of slide and annotation data that was used throughout the study to enable management of users, slide assignments, annotations, review and inter-participant concordance analyses.

2.5 Annotation review process

The following regions were annotated during crowdsourcing: (i) predominant classes including tumor, stroma, lymphocyte-rich regions and necrosis. (ii) Non-predominant classes including artifacts, adipose tissue, blood vessels, blood (intravascular or extravasated red

blood cells), glandular secretions and extracellular mucoid material and (iii) challenging classes including plasma cells, mixed inflammatory infiltrates (e.g. neutrophils), normal ducts or acini, metaplastic changes (osteoid matrix, cartilaginous metaplasia, etc.), lymph vessels, skin adnexa, angiogenesis and nerves. Since stroma is the most prevalent component in many slides, it was considered to be the ‘default’ class and defined by absence of annotations. JPs and NPs were directed to focus their effort on annotating predominant classes and to ask for feedback on Slack when annotating non-predominant or challenging classes. Providing feedbacks publicly on the Slack channel allowed all participants to access and learn from each other’s questions and responses. The study coordinator and SPs reviewed all annotations for mistakes using two mechanisms: (i) providing feedback to the participants on the Slack channel and (ii) generating new correction overlay annotations that are patched on top of the original annotations (Supplementary Fig. S1). Two phases of review and corrections were used (Supplementary Methods and Supplementary Fig. S3).

2.6 Measuring annotation discordance

The polygonal coordinates are queried using the DSA server REST API and are converted to a mask image format offline, where pixel values encode region class (Supplementary Fig. S1). The Inter-participant discordance was assessed for the 300 unique pairs of participants using their annotations on the evaluation set images. Discordance was measured using the Dice coefficient:

$$\Delta_{ij} = 1 - 2 \cdot \frac{\sum_{c=1}^{N_c} |I_c \cap J_c|}{\sum_{c=1}^{N_c} (|I_c| + |J_c|)} \quad (1)$$

where i and j are two participants, with corresponding masks I, J , composed of c binary channels, where N_c is the number of classes being considered. Δ_{ij} lies in the range $[0, 1]$, where 0 indicates no discordance. Our analysis makes comparisons on the effect of experience level and feedback on annotation quality. We used two techniques to visualize discordance between participants. The first is a bi-clustered heatmap of the inter-participant discordance matrix that groups participants based on discordance profiles. The second is a multidimensional scaling (MDS) analysis of the discordance matrix that depicts participants as points in two-dimensional space and where proximity indicates concordance.

2.7 Semantic segmentation and classification models

A pre-trained fully convolutional VGG-16, FCN-8 network was trained to segment histology images into five classes: tumor, stroma, inflammatory infiltrates, necrosis and other classes (Long et al., 2015). Shift and crop data augmentation was used to improve model robustness—see Supplementary Methods for details. Focusing on the 125 ROIs from infiltrating ductal carcinomas [the majority of TNBCs (Plasilova et al., 2016)], we first applied color normalization to the RGB images of the ROIs (Reinhard et al., 2001). Several different types of models were trained to evaluate different aspects of crowdsourcing:

Firstly, to investigate the effects of using crowdsourced versus single-expert annotations for training, we trained ‘comparison’ models for semantic segmentation. These models used annotations from evaluation set ROIs for training, and were evaluated on the post-correction core-set annotations (see Supplementary Fig. S2C).

Second, to evaluate peak accuracy, we trained ‘full’ models for semantic segmentation using the largest amounts of crowdsourced annotations possible. The full models were trained using annotations from core-set ROIs, assigning the ROIs from 82 slides (from 11 institutes) to the training set, and the ROIs from 43 slides (from seven institutes) to the testing set. Strict separation of ROIs by institute into either training or testing provides a better measure of how

models developed with our data will generalize to slides from new institutions and multi-institute studies.

Finally, to evaluate the effect of training set size on the accuracy of predictive models, we developed ‘scale-dependent’ image classification models using varying amounts of our crowdsourced annotation data (Supplementary Fig. S5). Since training hundreds of semantic segmentation models is time prohibitive, we instead trained classification models based on the pre-trained VGG-16 network to classify 224×224 pixel patches from the three predominant classes: tumor, stroma and inflammatory infiltration, using the same train/test assignment used in the semantic segmentation model (see details in Supplementary Methods).

3 Results

Our study produced a total of 50 057 polygonal annotations, including 3988 corrections. Following integration of the corrections (Supplementary Fig. S1B), a total of 20 340 polygonal annotations were extracted from the final mask images. The number of annotations within each ROI ranged from 11 to 541. Supplementary Table S2 describes the number of annotations by class, with the predominant classes representing more than 71% of the annotations. This data can be visualized in a public instance of the DSA at <https://goo.gl/cNM4EL>. Mask images derived from this data are used in training and validation are available at: goo.gl/UoUm9w. Further details can be found in the Supplementary Materials.

3.1 Annotation concordance is class-dependent

Discordance varies significantly by class, and reflects the difficulty and subjectivity inherent in the classes (Fig. 3 and Supplementary Fig. S2). Tumor annotations were the least discordant, with 0.13 (SP–SP), 0.16 (NP–NP) and 0.15 (SP–NP). These results indicate that both the bias (SP–NP) and variance (NP–NP) of annotations made by NPs are lower when only the predominant class is considered (Mann–Whitney $P = 3.66e-30$ and $P = 1.99e-168$, for SP–NP and NP–NP, respectively).

SPs had low median discordance for tumor (0.13), stroma (0.19) and necrosis (0.09), and had relatively higher discordance for lymphocytic infiltration (0.48). The median discordance between NPs and SPs was 0.14, 0.27, 0.54 and 1.0 for tumor, stroma, lymphocyte infiltration and necrosis/debris, respectively. Similarly, the median discordance among NPs was 0.14, 0.33, 0.50 and 1.0 for tumor, stroma, lymphocytic infiltration and necrosis/debris, respectively. The high discordance for necrosis/debris reflects the fact that many participants either missed this class when it was truly present or misclassified stroma as necrosis.

Supplementary Figure S4A shows the pixel-wise average SP–NP discordance between NPs and pathologists for two typical regions. Most of the discordance for tumor occurs around the region boundary. On the other hand, discordance for lymphocytic infiltration and, to a lesser extent, necrosis/debris follows a more diffuse pattern, and is not limited to the region boundary.

3.2 Feedback improves annotation quality

There was some clustering of participants by experience level, with three of the more experienced participants (two SPs and one JP) being highly mutually concordant as seen in the MDS plot (Fig. 3). The median SP–SP discordance was 0.24, compared to 0.30 for NP–NP. Discordance for SP–NP comparisons lies in the middle of this range at 0.27. Predictions of a semantic segmentation AL trained on corrected annotations from independent institutions results in discordance

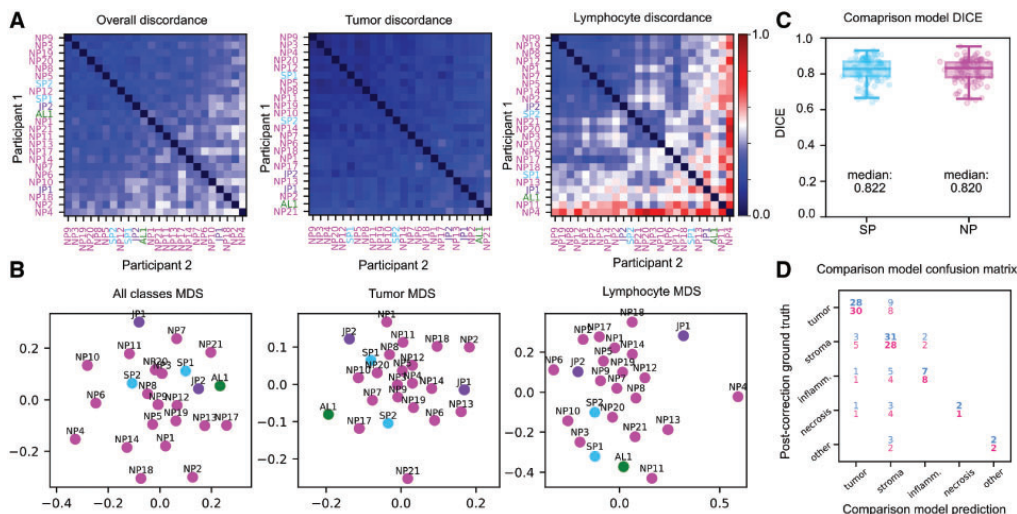


Fig. 3. Evaluation slide set concordance and model accuracy. **(A)** Inter-participant discordance matrices for SP, JP, NP and AL. **(B)** 2-D MDS plots of participant discordance. **(C, D)** Testing accuracy and confusion of comparison models trained on evaluation set ROIs from SPs (cyan) and NPs (magenta), measured against post-correction masks from the core set. Confusion matrix values are percentages relative to total pixel count. (Color version of this figure is available at *Bioinformatics online*.)

values similar to those of SPs (Fig. 3); the median SP-AL discordance is 0.22 overall and 0.15 for tumor. Three primary mistakes observed during the correction of core-set annotations were: (i) imprecise region boundaries, (ii) region misclassification and (iii) missing annotations for non-predominant classes. Examples of common mistakes are presented in [Supplementary Figure S4B](#). Discordance analysis results of the pre- and post-correction core-set annotations were consistent with trends observed in the evaluation set, but were notably lower. The median discordance between pre- and post-correction masks is 0.08 for all classes (SD = 0.30). This was significantly lower (Wilcoxon $P = 8.77e-14$) for binary tumor classification (0.01, SD = 0.11). Predominant classes had relatively low discordance, non-predominant classes had higher discordance, while challenging classes were almost always missing in NP annotations and were added by SPs in corrections.

The following are pre- and post-correction discordance values for other region classes (median \pm SD): stroma (0.09 \pm 0.15), lymphocytic infiltrate (0.08 \pm 0.31), necrosis (0.25 \pm 0.42), blood (0.02 \pm 0.32), exclude (0.01 \pm 0.42), fat ($3.98e-4 \pm 0.29$), extracellular mucoid material (0.50 \pm 0.50), glandular secretions (0.87 \pm 0.42) and blood vessel (1.00 \pm 0.28).

3.3 Accuracy of semantic segmentation models

The comparison semantic segmentation models had similar accuracy whether they were trained with NP annotations or SP annotations (median DICE = 0.820 versus 0.822, Fig. 3C and D). This result is consistent with the concordance results presented in Section 3.1.

The segmentations generated by the full semantic segmentation model were highly accurate and concordant with human annotations of ROIs in the testing set (see [Table 1](#) and [Supplementary Fig. S6](#)). The model predictions correspond well to region boundaries and are often more granular than human annotations (see [Figure 4](#) and [Supplementary Figs S7–S9](#)). When misclassifications occur, they are generally due to the composition of the training set.

Errors were found in uncommon or mixed patterns including: dense pure plasma cell infiltrates (classified as tumor), acellular

Table 1. Testing accuracy of full semantic segmentation model

	Mean AUC (SD)	DICE	Accuracy
Overall	0.945 (0.042) (micro)	0.888	0.799
Tumor	0.941 (0.058)	0.851	0.804
Stroma	0.881 (0.056)	0.800	0.824
Inflammatory	0.917 (0.150)	0.712	0.743
Necrosis	0.864 (0.237)	0.723	0.872
Other	0.885 (0.129)	0.666	0.670

hyaline stroma (classified as tumor) and necrotic regions containing dense inflammatory infiltrates (classified as infiltrates). Examples of these errors are shown in [Supplementary Figure S10](#).

3.4 Increasing scale improves image classification accuracy

The accuracy of scale-dependent models for patch classification are presented in [Figure 4C](#) (extended results [Supplementary Table S3](#)). A peak AUC above 0.95 was observed when all training data were used. With training data from only 2–4 randomly selected slides, AUCs of 0.78–0.9 are observed. Average AUC increases rapidly from 0.88 for two slides to 0.94 for eight slides. Average AUC continues to increase from 8 to 49 slides but with much slower growth. Beyond 49 slides growth in average AUC continues but is modest. This asymptotic trend is often observed in machine learning experiments where orders of magnitude more data are needed to significantly improve performance near the asymptote.

4 Discussion

The success of convolutional networks in analyzing histology images has increased interest in strategies for producing annotation data. While ALs are demonstrating diagnostically meaningful performance in many applications, large amounts of annotations are required to develop and validate these models. This necessitates

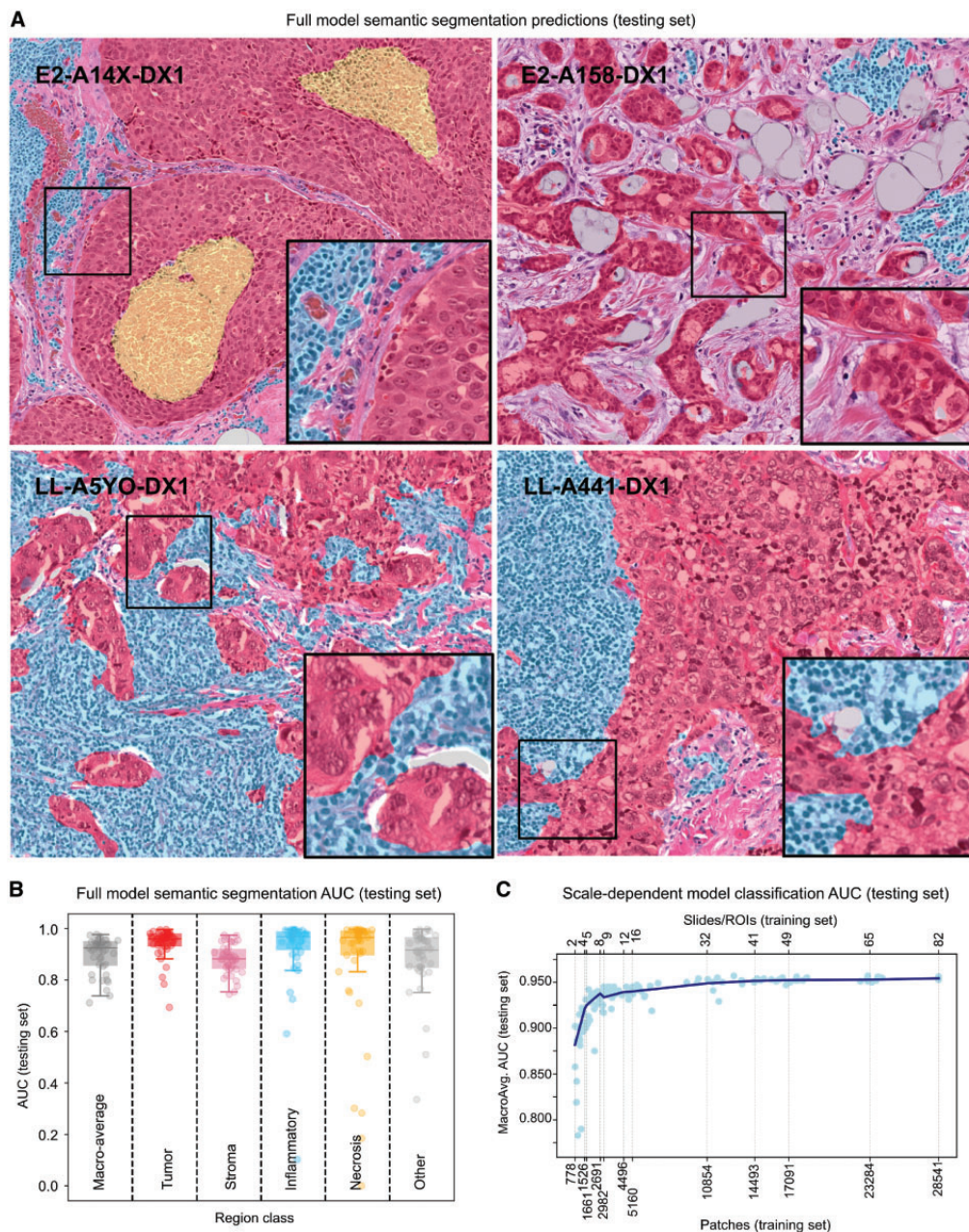


Fig. 4. Model performance over the testing set. **(A)** Visualization of full semantic segmentation model predictions on testing set regions of interest. Color codes used: red (tumor); transparent (stroma); cyan (inflammatory infiltrates); yellow (necrosis). **(B)** Area under ROC curve for semantic segmentation algorithm, broken down by region class. **(C)** Effect of training sample size on scale-dependent patch classification models. Each point represents the macro-average AUC of a single model, trained on different sets of randomly selected slides. (Color version of this figure is available at *Bioinformatics online*.)

engaging multiple participants in annotation studies, and the ability to efficiently organize participants with a range of experience levels is one approach to scaling the annotation process. While significant expertise is needed for accurate semantic annotation of histology,

our study provides an example application where non-experts can be trained to effectively perform much of the time-consuming work.

While non-experts cannot be expected to recognize rare patterns or to accurately annotate difficult cases, a large majority of the

work in delineating tissue boundaries does not fall in these categories. By utilizing expertise where it is needed, in the annotation of rare or difficult classes, and in reviewing and correcting the annotations of non-experts, we were able to produce a large dataset containing over 20 000 annotated tissue regions. This resource can be used to train semantic segmentation models for breast cancer histology to characterize the tumor microenvironment and inflammatory infiltration, both of which are known to strongly correlate with cancer progression and patient outcomes (Fouad and Aanei, 2017).

The annotations in our study were produced using the DSA, a web-based digital pathology platform. While DSA provides a wide array of annotation tools, future development will enhance review and collaboration capabilities by formalizing these processes in specialized interfaces. These enhancements will increase the utility of DSA for annotation studies, education and diagnostic review including tumor boards.

Although concordance among participants was generally strong, important sources of discordance between SPs and NPs were observed (Supplementary Fig. S4). In the predominant classes, discordance was often observed in cases where judgment was either difficult or subjectively defined (e.g. a region is lymphocyte-rich if at least 80% of its area was occupied by lymphocytes). Less frequently occurring non-predominant classes were also often missed by NP participants, likely due to difficulty in recognizing these classes and a lack of training. Examples of annotation errors include stromal regions being mislabeled as necrosis or vice versa, hyalinized or acellular stroma misclassified as mucinous change, plasma cells being mislabeled as lymphocytes, and endothelial cells, activated fibroblasts or activated histiocytes being mislabeled as tumor. Missed classes include blood vessels, glandular secretions, as well as rare metaplastic changes and non-lymphocytic inflammatory infiltrates (it should be noted that much of the discordance arises from non-predominant classes added by the SPs during correction). We provide further evidence of the utility of NP annotations, showing that comparison models derived from NP annotations had similar accuracy to models derived from SP annotations. These comparison experiments were based on the limited set of ROIs for which both SP and NP annotations were available, and hence we still recommend supervision and feedback by SPs following the initial training of NP participants.

Semantic segmentation models derived from our annotation dataset were highly accurate, and provide new opportunities for feature extraction from breast cancer histology and tissue based studies. These models have a high macro-average AUC (0.897) and class-wise AUCs ranging from 0.881 (stroma) to 0.941 (inflammatory). Visualization shows that many of the areas where the human and computational prediction disagree is due to increased sensitivity of the models to granular regions that are not annotated by human participants.

While our study presents important findings on annotating histology images, there are a number of research questions that were not addressed. Our study relied on medical students and graduates, with the rationale being that basic familiarity with histology and general biology may reduce error rates. Future studies may investigate whether this assumption is correct, and if it is possible to engage a broader pool of participants that lack this training to further scale annotation efforts. Our study also did not evaluate intra-participant discordance,

an issue that is known to be significant in pathology. Measuring intra-participant discordance would provide a baseline to evaluate inter-participant discordances against, and would provide better context for the differences in discordance observed among and between participants with different experience levels. The time participants spent in making annotations was also not recorded, nor was the time that more experienced users spent correcting annotations. This information, while difficult to acquire reliably, could provide further insights on how to best allocate resources in structured crowdsourcing studies. Finally, we would point out that the value of crowdsourcing likely varies by application. The amount of data required varies with the difficulty of the prediction task, whether the model is expected to generalize to specimens from different institutions, expectations for prediction model accuracy and the availability of experts to produce annotations. In some tasks, a well-resourced organization may be able to engage their pathologists to produce sufficient annotations for 'in-house' models not intended to generalize to specimens generated at other labs.

Funding

This work was supported by the U.S. National Institutes of Health; and National Cancer Institute grants [U01CA220401, U24CA194362].

Conflict of Interest: Ventana did not fund this study, but they are funding others.

References

- Alialy, R. *et al.* (2018) A review on the applications of crowdsourcing in human pathology. *J. Pathol. Inform.*, 9, 2.
- Fouad, Y.A. and Aanei, C. (2017) Revisiting the hallmarks of cancer. *Am. J. Cancer Res.*, 7, 1016–1036.
- Gutman, D.A. *et al.* (2013) Cancer Digital Slide Archive: an informatics resource to support integrated in silico analysis of TCGA pathology data. *J. Am. Med. Inform. Assoc.*, 20, 1091–1098.
- Gutman, D.A. *et al.* (2017) The digital slide archive: a software platform for management, integration, and analysis of histology for cancer research. *Cancer Res.*, 77, e75–e78.
- Hughes, H. *et al.* (2018) Quanti.us: a tool for rapid, flexible, crowd-based annotation of images. *Nat. Methods*, 15, 587.
- Irshad, H. *et al.* (2015) Crowdsourcing image annotation for nucleus detection and segmentation in computational pathology: evaluating experts, automated methods, and the crowd. *Pac. Symp. Biocomput.*, 294–305.
- Irshad, H. *et al.* (2017) Crowdsourcing scoring of immunohistochemistry images: evaluating Performance of the Crowd and an Automated Computational Method. *Sci. Rep.*, 7, 43286.
- Kovashka, A. *et al.* (2016) Crowdsourcing in computer vision. *Found. Trends Comput. Graph. Vis.*, 10, 177–243.
- Long, J. *et al.* (2015) Fully convolutional networks for semantic segmentation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 3431–3440.
- Plasilova, M.L. *et al.* (2016) Features of triple-negative breast cancer: analysis of 38, 813 cases from the national cancer database. *Medicine*, 95, e4614.
- Reinhard, E. *et al.* (2001) Color transfer between images. *IEEE Comput. Graph.*, 21, 34–41.
- Su, H. *et al.* (2012) Crowdsourcing annotations for visual object detection. In: *Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence*. Vol. 1. Toronto, Ontario, Canada.

Section 2.2

NuCLS: A scalable crowdsourcing approach and dataset for nucleus classification and segmentation in breast cancer

This section is a modified partial reproduction of the following open-access preprint:

Amgad, M., Atteya, L. A., Hussein, H., Mohammed, K. H., Hafiz, E., Elsebaie, M. A., Alhusseiny, A. M., AlMoslemany, M. A., Elmatboly, A. M., Pappalardo, P. A., et al.(2021b). NuCLS: A scalable crowdsourcing, deep learning approach and dataset for nucleus classification, localization and segmentation. arXiv preprint arXiv:2102.09099.

The open-access dataset is downloadable from the [NuCLS website](#).

PAPER

NuCLS: A scalable crowdsourcing approach & dataset for nucleus classification and segmentation in breast cancer

Mohamed Amgad¹, Lamees A. Atteya^{2,†}, Hagar Hussein^{3,†}, Kareem Hosny Mohammed^{4,†}, Ehab Hafiz^{5,†}, Maha A.T. Elsebaie^{6,†}, Ahmed M. Alhusseiny⁷, Mohamed Atef AlMoslemany⁸, Abdelmagid M. Elmatboly⁹, Philip A. Pappalardo¹⁰, Rokia Adel Sakr¹¹, Pooya Mobadersany¹, Ahmad Rachid¹², Anas M. Saad¹³, Ahmad M. Alkashash¹⁴, Inas A. Ruhban¹⁵, Anas Alrefai¹², Nada M. Elgazar¹⁶, Ali Abdulkarim¹⁷, Abo-Alela Farag¹², Amira Etman⁸, Ahmed G. Elsaeed¹⁶, Yahya Alagha¹⁷, Yomna A. Amer⁸, Ahmed M. Raslan¹⁸, Menatalla K. Nadim¹⁹, Mai A.T. Elsebaie¹², Ahmed Ayad²⁰, Liza E. Hanna³, Ahmed Gadallah¹², Mohamed Elkady²¹, Bradley Drumheller²², David Jaye²², David Manthey²³, David A. Gutman²⁴, Habiba Elfandy^{25,26} and Lee A.D. Cooper^{1,27,28,*},[†]

¹Department of Pathology, Northwestern University, Chicago, IL, USA

*Address correspondence to: lee.cooper@northwestern.edu

[†]Contributed equally.

[‡]See full list of author affiliations at the end.

Abstract

Background: Deep learning enables accurate high-resolution mapping of cells and tissue structures that can serve as the foundation of interpretable machine-learning models for computational pathology. However, generating adequate labels for these structures is a critical barrier, given the time and effort required from pathologists. **Results:** This paper describes a novel collaborative framework for engaging crowds of medical students and pathologists to produce quality labels for cell nuclei. We used this approach to produce the NuCLS dataset, containing over 220,000 annotations of cell nuclei in breast cancers. This builds on prior work labeling tissue regions to produce an integrated tissue region- and cell-level annotation dataset for training that is the largest such resource for multi-scale analysis of breast cancer histology. This paper presents data and analysis results for single and multi-rater annotations from both non-experts and pathologists. We present a novel method for suggesting annotations that allows us to collect accurate segmentation data without the need for laborious manual tracing of cells. Our results indicate that even noisy algorithmic suggestions do not adversely affect pathologist accuracy, and can help non-experts improve annotation quality. We also present a new approach for inferring truth from multiple raters, and show that non-experts can produce accurate annotations for visually distinctive classes. **Conclusions:** This study is the most extensive systematic exploration of the large-scale use of wisdom-of-the-crowd approaches to generate data for computational pathology applications.

Key words: Crowdsourcing; Deep learning; Nucleus segmentation; Nucleus classification; Breast cancer.

Background

Motivation

Convolutional neural networks (CNN) and other deep learning methods have been at the heart of recent advances in medicine (see Table S1 for terminology) [1]. A key challenge in computational pathology is the scarcity of large-scale labeled datasets for model training and validation [2, 3, 4]. Specifically, there is a shortage of annotation data for delineating tissue regions and cellular structures in histopathology. This information is critical for training interpretable deep-learning models, as they allow the detection of entities that are understood by pathologists and map to known diagnostic criteria [4, 5, 6, 7]. These entities can then be used to construct higher-order relational graphs that encode complex spatial and hierarchical relationships within the tumor microenvironment, paving the way for the computationally-driven discovery of histopathologic biomarkers and biological associations [4, 8, 9, 10, 11, 12, 13]. Data shortage is often attributed to the domain expertise required to produce annotation labels, with pathologists spending years in residency and fellowship training [2, 14]. This problem is exacerbated by the time constraints of clinical practice and the repetitive nature of annotation work. Manual tracing of object boundaries is an incredibly demanding task, and there is a pressing need to obtain this data using facilitated or assisted annotation strategies [15]. By comparison, traditional annotation problems like detecting people in natural images require almost no training and typically engage the general public [15]. Moreover, unique problems often require new annotation data, underscoring the need for scalable and reproducible annotation workflows [16].

We address these issues using an assisted annotation method that leverages the participation of non-pathologists (NPs), including medical students and graduates. Medical students typically have strong incentives to participate in annotation studies, with increased reliance on research participation in residency selection [17]. We describe adaptations to the data collection to improve scalability and reduce effort. This work focuses on nucleus classification, localization, and segmentation (NuCLS, for short) in whole-slide scans of hematoxylin and eosin-stained (H&E) slides of breast carcinoma from 18 institutions from The Cancer Genome Atlas (TCGA). Our annotation pipeline enables low-effort collection of nucleus segmentation and classification data, paving the way for systematic discovery of histopathologic-genomic associations and morphological biomarkers of disease progression [4, 5, 8, 10, 11].

Related work

There has been growing interest in addressing data scarcity in histopathology by either 1. scaling data generation or 2. reducing reliance on manually labeled data using data synthesis techniques like Generative Adversarial Networks [18, 19, 20, 21, 22, 23, 24, 25]. While there is a pressing need for both approaches, this work is meant to fit into the broad context of scalable assisted manual data generation when expert annotation is expensive or difficult. Crowdsourcing, the process of engaging a “crowd” of individuals to annotate data, is critical to solving this problem. There exists a large body of relevant work in crowdsourcing for medical image analysis [15, 26, 27]. Previously, we published a study and dataset using crowdsourcing of NPs for annotation of low-power regions in breast cancer [28]. Our approach was structured because we assigned different tasks depending on the level of expertise and leveraged collaborative annotation to obtain data that is large in scale and high in quality. Here, we significantly expand this idea by focusing

on the challenging problems of nucleus classification, localization, and segmentation. This computer vision problem is a subject of significant interest in computational pathology [29, 30, 31].

While the public release of data is only one aspect of our study, it is essential to acknowledge related nucleus classification datasets. Some of these datasets can be used in conjunction with ours and include MoNuSAC, CoNSEp, PanNuke, and Lizard [29, 30, 32, 33, 34, 35, 36, 37, 38]. Lizard, in particular, is a highly related dataset that was recently published after we released NuCLS but focuses on colon cancer instead [37]. Additionally, the US Food and Drug Administration is leading an ongoing study to collect regulatory-grade annotations of stromal tumor-infiltrating lymphocytes (sTILs) [39]. Unfortunately, with few exceptions, most public computational pathology datasets are either limited in scale, were generated through exhaustive annotation efforts by practicing pathologists, or do not disclose or discuss data generation [2, 26, 30, 40]. Additionally, to the best of our knowledge, most other works do not explore crowdsourcing as a data generation approach or systematically explore interrater agreement for experts vs. non-experts.

A few studies are of particular relevance to this paper. A study by Irshad et al. showed that non-experts, recruited through the Figure Eight platform, can produce accurate nucleus detections and segmentations in renal clear cell cancer but was limited to 10 whole-slide images [20]. Hou et al. explored the use of synthetic data to produce nuclear segmentations [41]. While a significant contribution, their work did not address classification, relied on qualitative slide-level evaluations of results, and did not explore how algorithmic bias affects data quality [42, 22]. The approach we used involves click-based approval of annotations generated by a deep-learning algorithm. This methodological aspect is not the central focus of this paper; it is only one of many approaches for interactive segmentation and classification of nuclei explored in past studies like HistomicsML and NuClick [42, 22].

Our contributions

This work describes a scalable crowdsourcing approach that systematically engaged NPs and produced annotations for localization, segmentation, and classification of nuclei in breast cancer. Our workflow required minimal effort from pathologists and used algorithmic suggestions to scale the annotation process and obtain hybrid annotation datasets containing numerous segmentation boundaries without laborious manual tracing. We show that algorithmic suggestions can improve the accuracy of NP annotations and that NPs are reliable annotators of common cell types. In addition, we discuss a new constrained clustering method that we developed for reliable truth inference in multi-rater datasets. We also show how multi-rater data can ensure the quality of NP annotations or replace expert supervision in some contexts. Finally, we note that downstream deep-learning modeling using the NuCLS dataset is discussed in a related publication and is not the focus of this paper [43].

Data Description

NuCLS is a large-scale multi-class dataset generated by engaging crowds of medical students and pathologists. NuCLS is sourced from the same images as the Breast Cancer Semantic Segmentation (BCSS) dataset [28]. Together, these datasets contain region- and cell-level annotations and constitute the most extensive resource for multi-scale analysis of breast cancer slides. We obtained a total of 222,396 nucleus annotations, including over 125,000 single-rater

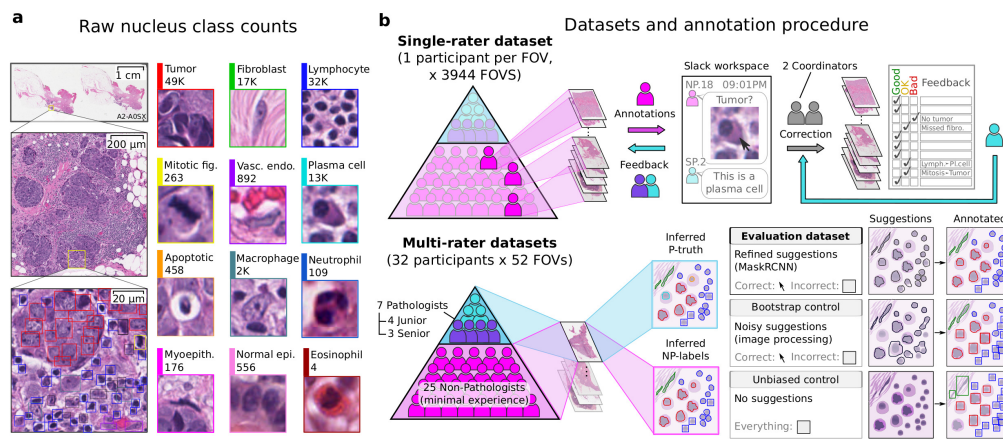


Figure 1. Dataset annotation and quality control procedure. a. Nucleus classes annotated. b. Annotation procedure and resulting datasets. Two approaches were used to obtain nucleus labels from non-pathologists (NPs). (Top) The first approach focused on breadth, collecting single-rater annotations over a large number of FOVs to obtain the majority of data in this study. NPs were given feedback on their annotations, and two study coordinators corrected and standardized all single-rater NP annotations based on input from a senior pathologist. (Bottom) The second approach evaluated interrater reliability and agreement, obtaining annotations from multiple NPs for a smaller set of shared FOVs. Annotations were also obtained from pathologists for these FOVs to measure NP reliability. The procedure for inferring a single set of labels from multiple participants is described in Figure 2. We distinguished between inferred NP-labels and inferred P-truth for clarity. Three multi-rater datasets were obtained: an Evaluation dataset, which is the primary multi-rater dataset, as well as Bootstrap and Unbiased experimental controls to measure the value of algorithmic suggestions. In all datasets except the Unbiased control, participants were shown algorithmic suggestions for nucleus boundaries and classes. They were directed to click nuclei with correct boundary suggestions and annotate other nuclei with bounding boxes. The pipeline to obtain algorithmic suggestions consisted of two steps: 1. Using image processing to obtain bootstrapped suggestions (Bootstrap control); 2. Training a Mask R-CNN model to refine the bootstrapped suggestions (single-rater and Evaluation datasets).

annotations and 97,000 multi-rater annotations. A detailed description of the dataset creation protocol is presented in the methods section.

Analyses and Discussion

Structured crowdsourcing enables scalable data collection

Pathologist time is limited and expensive, and relying solely on pathologists for generating annotations can hinder the development of state-of-the-art models based on CNNs. In this study, we show that NPs can perform most of the time-consuming annotation tasks and that pathologist involvement can be limited to low-effort tasks that include:

- Training NPs and answering their questions (Figure 1) [44].
- Qualitative scoring of NP annotations (Figure S1).
- Low-power annotation of histologic regions (Figure S2) [28].

We used a web-based annotation platform called HistomicsUI for annotation, feedback, and quality review [45]. HistomicsUI provides a user interface with annotation tools and an API for programmatic querying and manipulating the centralized annotation database. The NuCLS dataset includes annotations from 32 NPs and seven pathologists in the US, Egypt, Syria, Australia, and the Maldives. We obtained 128,000 nucleus annotations from 3,944 fields-of-view (FOV) and 125 triple-negative breast cancer patients. The annotations included bounding box placement, classification, and for a sizable fraction of nuclei, segmentation boundaries. Half of these annotations underwent quality control correction based on feedback by a practicing pathologist.

Additionally, we obtained three multi-rater datasets containing 97,300 annotations, where the same FOV was annotated by multiple participants (Figure 1b, Figure 2). The collection of multi-rater data enables quantitative evaluation of NP reliability, interrater variability, and the impact of algorithmic suggestions on NP accuracy.

Multi-rater annotations were not corrected by pathologists and enabled an unbiased assessment of NP performance. Pathologist annotations were also collected for a limited set of multi-rater FOVs to evaluate NP accuracy.

NPs can reliably classify common cell types

The detection accuracy of NPs was moderately high ($AP=0.68$) and was similar to the detection accuracy of pathologists. Classification accuracy of NPs, on the other hand, was only high for common nucleus classes (micro-average $AUROC=0.93[0.92,0.94]$ vs. macro-average $AUROC=0.75[0.74,0.76]$) and was higher when grouping by super-class (Figure 3, Figure S3). We reported the same phenomenon in our previous work on crowdsourcing annotation of tissue regions [28]. In addition, we observed moderate clustering by participant experience (Figure 3d) and variability in classification accuracy among NPs ($MCC=60.7-84.2$). This observation motivated our quality control procedures. Study coordinators manually corrected missing or misclassified cells for the single-rater dataset, and practicing pathologists supervised and approved annotations. For the multi-rater datasets, we inferred a singular label from pathologists (P-truth) and NPs (NP-label) using an Expectation-Maximization (EM) framework that estimates reliability values for each participant [46, 47].

When pathologist supervision is not an option, multi-rater datasets need to have annotations from a sufficient number of NPs to infer reliable data. We used the annotations we obtained to perform simulations to estimate the accuracy of inferred NP-labels with fewer numbers of participating NPs (Figure 3e). The inferred NP-label accuracy increased up to six NPs per FOV, after which there were diminishing returns. Our simulations also showed that stromal nuclei require more NPs per FOV than tumor nuclei or sTILs.

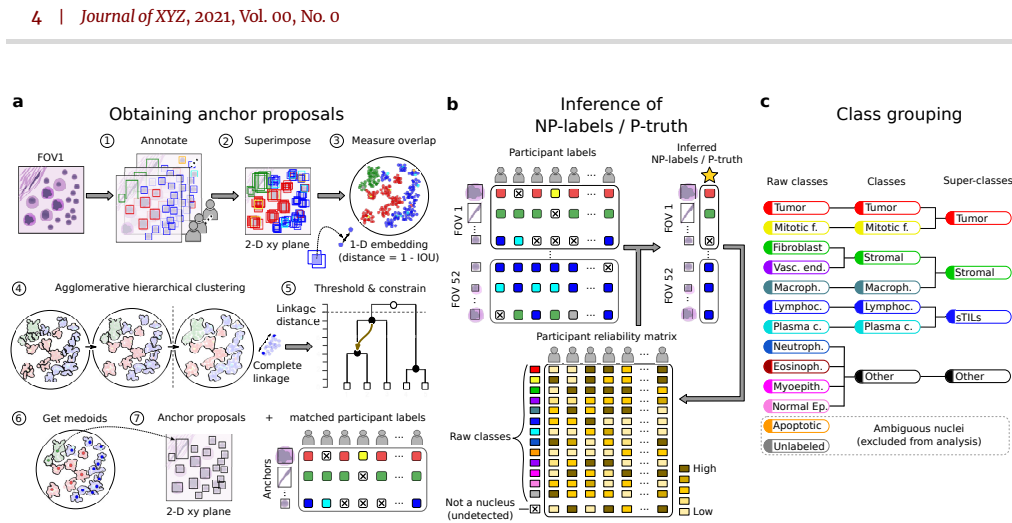


Figure 2. Inference from multi-rater datasets. The purpose of this step was to infer the nucleus locations and classifications from multi-rater data. a. The first step involved agglomerative hierarchical clustering of bounding boxes using Intersection-Over-Union (IOU) as a similarity measure. We imposed a constraint during clustering that prevents merging annotations where a single participant has annotated overlapping nuclei. Participant intention was preserved by demoting annotations from the same participant to the next node (step 5, arrow). After clustering was complete, a threshold IOU value was used to obtain the final clusters (step 5, black nodes). Within each cluster, the medoid bounding box was chosen as an anchor proposal. The result was a set of anchors with corresponding clustered annotations. When a participant did not match to an anchor, it was considered a conscious decision not to annotate a nucleus at that location. b. Once anchors were obtained, an expectation-maximization (EM) procedure was used to estimate: 1. which anchors represent actual nuclei, and 2. which classes to assign these anchors. The EM procedure estimates and accounts for the reliability of each participant for each classification. EM was performed separately for NPs and pathologists. c. Grouping of nucleus classes. Consistent with standard practice in object detection, nuclei were grouped, based on clinical reasoning, into five classes and three super-classes.

Minimal-effort collection of nucleus segmentation data

Many nucleus detection and segmentation algorithms were developed using conventional image analysis methods before the widespread adoption of CNNs. These algorithms have little or no dependence on annotations, and while they may not be as accurate as CNNs, they can correctly segment a significant fraction of nuclei. We used simple nucleus segmentation heuristics, combined with low-power region annotations from the BCSS dataset, to obtain bootstrapped annotation suggestions for nuclei (Figure S2) [28]. The suggestions were refined using a deep-learning model (Mask R-CNN) as a function approximator trained on the bootstrapped suggestions. This procedure allowed poor quality bootstrapped suggestions in one FOV to be smoothed by better suggestions in other FOVs (Figure S4, Table S2) and is analogous to fitting a regression line to noisy data [18, 48]. This model was applied to the FOVs to generate refined suggestions shown to participants when annotating the single-rater dataset and the Evaluation dataset (the primary multi-rater dataset) [44]. Two additional multi-rater datasets were obtained as controls:

- *Bootstrap control*: participants were shown unrefined bootstrapped suggestions.
- *Unbiased control*: participants were not shown any suggestions. This dataset was the first multi-rater dataset to be annotated.

Accurate suggestions can be confirmed during annotation with a single click, reducing effort and providing valuable nucleus boundaries that can aid the development of segmentation models. Participants can annotate other nuclei with bounding boxes that require more effort than click annotations but less effort than manual tracing [15]. We obtained a substantial proportion of nucleus boundaries through clicks: $41.7 \pm 17.3\%$ for the Evaluation dataset and 36.6% for the single-rater dataset (Figure 4, Figure S5). The resultant hybrid dataset contained a mixture of bounding boxes and accurate segmentation boundaries (Evaluation dataset DICE= 85.0 ± 5.9). We argue that it is easier to handle hybrid datasets at the level of algorithm development than to have participants trace missing

boundaries or correct imprecise ones. We evaluate the bias of using these suggestions in the following section.

Algorithmic suggestions improve classification accuracy

There was value in providing the participants with suggestions for nuclear class, which included suggestions directly inherited from BCSS region annotations, as well as high-power refined suggestions produced by Mask R-CNN (Figure 4). Pathologists had substantial self-agreement when annotating FOVs with or without refined suggestions (Kappa= 87.4 ± 7.9). NPs also had high self-agreement but were more impressionable when presented with suggestions (Kappa= 74.0 ± 12.6). This was, however, associated with a reduction in bias in their annotations; refined suggestions improved the classification accuracy of inferred NP-labels (AUROC= $0.95 [0.94, 0.96]$ vs. $0.92 [0.90, 0.93]$, $p < 0.001$). This observation is consistent with Marzahl et al., who reported similar findings in a crowdsourcing study using bovine cytology slides [27].

Region-based class suggestions for nuclei were, overall, more concordant with the corrected single-rater annotations compared to Mask R-CNN refined (high-power) nucleus suggestions (MCC= 67.6 vs. 52.7) (Figure S4, Table S2). Nonetheless, high-power nucleus suggestions were more accurate for 24.8% of FOVs and had a higher recall for sTILs (96.8 vs. 76.6) [4, 11]. This result makes sense since stromal regions often contain scattered sTILs, and a region-based approach to labeling would incorrectly mark these as stromal nuclei (e.g., see Figure S6) [28, 49]. Hence, the value of low and high-power classification suggestions is context-dependent.

Exploring nucleus detection and classification tradeoffs

Naturally, there is some variability in the judgments made by participants about nuclear locations and classes and the accuracy of suggested boundaries. We study the process of inferring a single truth from multi-rater datasets and discuss the effect of various parameters. There is a tradeoff between the number of nucleus anchor

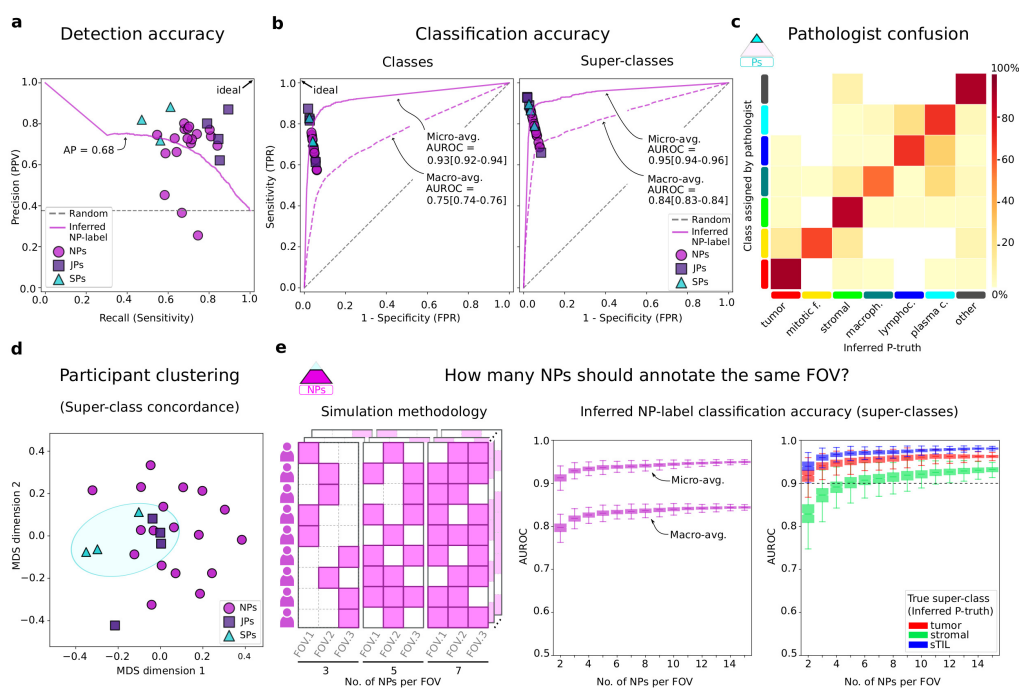


Figure 3. Accuracy of participant annotations. a. Detection precision–recall comparing annotations to inferred P-truth. Junior pathologists tend to have similar precision but higher recall than senior pathologists, possibly reflecting the time constraints of pathologists. b. Classification ROC for classes and super-classes. The overall classification accuracy of inferred NP-labels was high. However, class-balanced accuracy (macro-average) is notably lower since NPs are less reliable annotators of uncommon classes. c. Confusion between pathologist annotations and inferred P-truth. d. Multidimensional scaling (MDS) analysis of interrater classification agreement. Some clustering by participant experience (blue ellipse) highlights the importance of modeling reliability during label inference. e. A simulation was used to measure how redundancy impacts the classification accuracy of inferred NP-labels. While keeping the total number of NPs constant, we randomly kept annotations for a variable number of NPs per FOV. Accuracy in these simulations was class-dependent, with stromal nuclei requiring more redundancy for accurate inference.

proposals and interrater agreement (Figure 5). The clustering IOU threshold that defines the minimum acceptable overlap between any two annotations substantially impacted the number of anchor proposals. We found that an IOU threshold of 0.25 detects most nuclei with adequate pathologist classification agreement (1,238 nuclei, $\text{Alpha}=55.5$). We imposed a constraint to prevent annotations from the same participant from mapping to the same cluster—this improved detection of touching nuclei when the number of pathologists was limited (Figure 5b).

Nucleus detection was a more significant source of discordance among participants than nucleus classification (Figure 3, Figure S7, Figure S8). Some nucleus classes were easier to detect than others. sTILs were the easiest to detect, likely due to their hyperchromicity and tendency to aggregate; 53.3% of sTILs were detected by 16+ NPs (Figure S9). Fibroblasts were demonstrably harder to detect (only 21.4% were detected by 16+ NPs), likely because of their relative sparsity and lighter nuclear staining. Lymphocytes and plasma cells, which often co-aggregate in lymphoplasmacytic clusters, were a source of interrater discordance for pathologists and NPs [4, 50]. This discordance may stem from variable degrees of reliance on low-power vs. high-power morphologic features. Interrater agreement for nuclear classification was high and significantly improved when classes were grouped into clinically-salient super-classes ($\text{Alpha}=66.1$ (pathologists) and 60.3 (NPs); Figure 5).

Methods

Data sources

The scanned diagnostic slides we used were generated by the TCGA Research Network (<https://www.cancer.gov/tcga>). They were obtained from 125 patients with breast cancer (one slide per patient). Specifically, we chose to focus on all carcinoma of unspecified type cases that were triple-negative. The designation of histologic and genomic subtypes was based on public TCGA clinical records [28]. All slides were stained with Hematoxylin and Eosin (H&E) and were formalin-fixed and paraffin-embedded. The scanned slides were accessed using the Digital Slide Archive repository [45].

Region annotations were obtained from BCSS, a previous crowdsourcing study that we conducted [28]. Regions of Interest (ROIs), 1 mm² in size, were assigned to participants by difficulty level. All region annotations were corrected and approved by a practicing pathologist. These region annotations were used to obtain nucleus class suggestions as described below. Region classes included tumor, stroma, lymphocytic infiltrate, plasmacytic infiltrate, necrosis/debris, and other uncommon regions.

Algorithmic suggestions

The process for generating algorithmic suggestions is summarized in Figure S2 and involves the following steps:

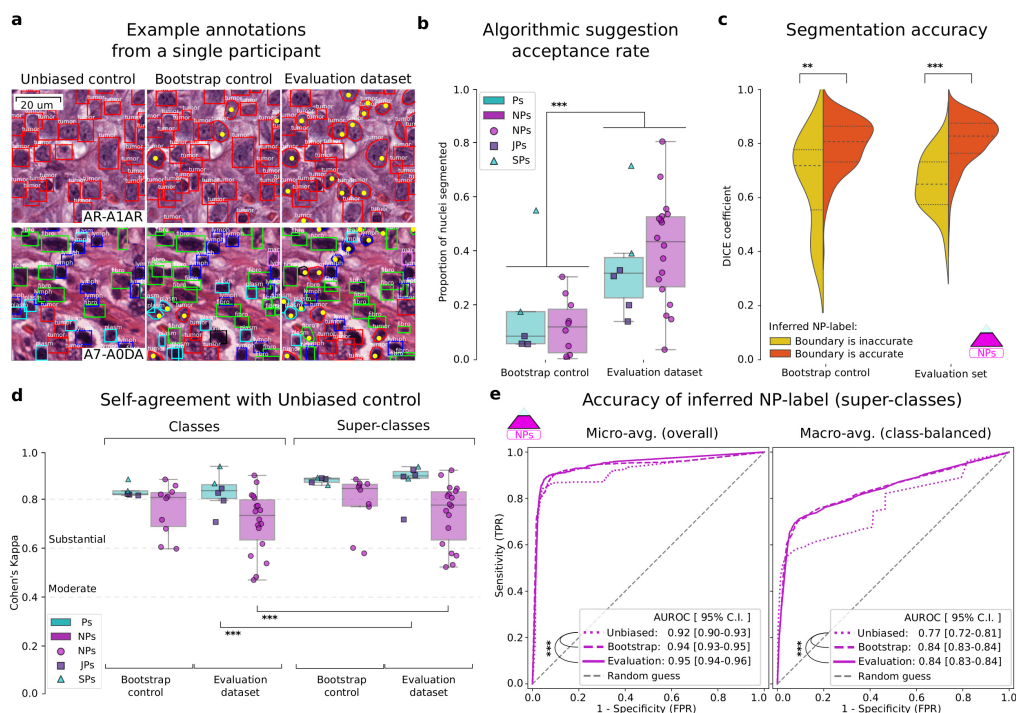


Figure 4. Effect of algorithmic suggestions on annotation abundance and accuracy. We compared annotations from the Evaluation dataset and controls to measure the impact of suggestions and Mask R-CNN refinement on the acquisition of nucleus segmentation data and the accuracy of annotations. a. Example annotations from a single participant. Algorithmic suggestions allow the collection of accurate nucleus segmentations without added effort. Yellow points indicate clicks to approve suggestions. b. The number of segmented nuclei clicked is significantly higher for the Evaluation dataset than for the Bootstrap control, indicating that refinement improves suggestion quality. c. Accuracy of algorithmic segmentation suggestions. The comparison is made against a limited set of manually traced segmentation boundaries obtained from one senior pathologist. Suggestions that were determined to be correct by the EM procedure had significantly more accurate segmentation boundaries. d. Self-agreement for annotations in the presence or absence of algorithmic suggestions. The agreement is substantial for NP and pathologist groups, indicating that algorithmic suggestions do not impact classification decisions adversely. Pathologists have higher self-agreement and are less impressionable than NPs. e. ROC curves for the classification accuracy of inferred NP-label, using inferred P-truth as our reference. Statistically-significant comparisons are indicated with a star (**, $p < 0.01$; ***, $p < 0.001$).

Heuristic nucleus segmentation. We used simple image processing heuristics to obtain noisy nucleus segmentations [31]. Images were analyzed at scan magnification (40x) with the following steps: 1. Hematoxylin stain unmixing using the Macenko method [51]. 2. Gaussian smoothing followed by global Otsu thresholding to identify foreground nuclei pixels [52]. This step was done for each region class separately to increase robustness. We used a variance of two pixels for lymphocyte-rich regions and five pixels for other regions. 3. Connected-component analysis split the nuclei pixel mask using 8-connectivity and a 3×3 structuring element [53]. 4. We computed the Euclidean distance from every nucleus pixel to the nearest background pixel and found the peak local maxima using a minimum distance of 10 [54]. 5. A watershed segmentation algorithm split the connected components from step 3 into individual nuclei using the local maxima from step 4 as markers [55, 56]. 6. Any object < 300 pixels in area was removed.

Bootstrapping noisy training data. Region annotations were used to assign a noisy class to each segmented nucleus. This decision was based on the observation that although tissue regions usually contain multiple cell types, there is often a single predominant cell type: tumor regions / tumor cells, stromal regions / fibroblasts, lymphocytic infiltrate / lymphocytes, plasmacytic infiltrate / plasma cells, other regions / other cells. One exception to this direct mapping is stromal regions, which contain a large number of sTILs in addition

to fibroblasts. Within stromal regions, a nucleus was considered a fibroblast if it had a spindle-like shape with an aspect ratio between 0.4 and 0.55 and circularity between 0.7 and 0.8.

Mask R-CNN refinement of bootstrapped suggestions. A Mask R-CNN model with a Resnet50 backbone was used as a function approximator to refine the bootstrapped nucleus suggestions. This model was trained using randomly cropped 128×128 tiles where the number of nuclei was limited to 30. Table S3 summarizes the hyperparameters used.

FOV sampling procedure. ROIs were tiled into non-overlapping potential FOVs. These were selected for inclusion in our study based on predefined stratified sampling criteria. 16.7% of FOVs were sampled such that the majority of refined suggestions were a single class, e.g., almost all suggestions are tumor. 16.7% were sampled to favor FOVs with two almost equally-represented classes, e.g., many tumor and fibroblast suggestions. Finally, 16.7% of FOVs were sampled to favor discordance between the bootstrapped suggestions and Mask R-CNN-refined suggestions, e.g., a stromal region with sTILs. The remaining 50% of FOVs were randomly sampled from the following pool, with the intent of favoring the annotation of difficult nuclei: a) the bottom 5% of FOVs containing high numbers of nuclei with low Mask R-CNN confidence; b) and the top 5% of FOVs containing extreme size detections, presumably clumped nuclei.

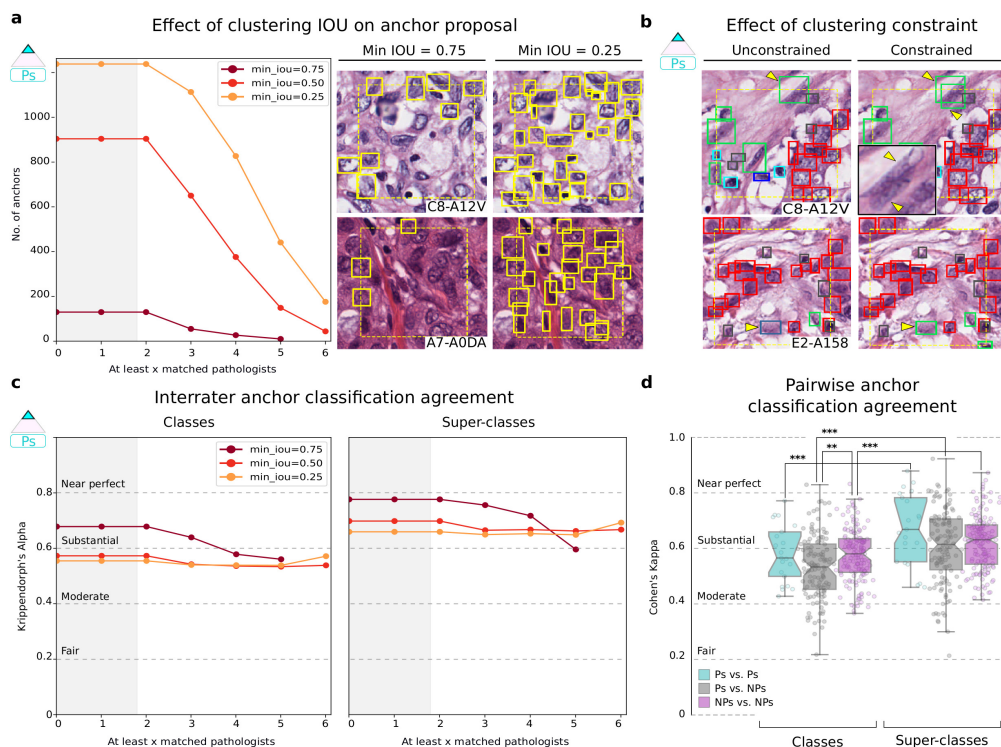


Figure 5. Effect of clustering on detection and interrater agreement. a. Stricter IOU thresholds reduce the number of anchor proposals generated by clustering but increase agreement. A threshold of 0.25 provides more anchor proposals with negligible difference in agreement from the 0.5 threshold. The shaded region indicates that by design, there are no anchor proposals with less than two clustered annotations. b. The clustering constraint prevents annotations from the same participant from being assigned to the same anchor, preserving participant intention when annotating overlapping nuclei. This results in better detection of overlapping nuclei during clustering (upper panel) and also impacts the inferred P-truth for anchors (bottom panel). c. Interrater classification agreement among pathologists for tested clustering thresholds. d. Pairwise interrater classification agreement (Cohen's Kappa) at 0.25 IOU threshold. Statistically-significant comparisons are indicated with a star (**, $p < 0.01$; ***, $p < 0.001$).

Annotation procedure and data management

The annotation protocol used is provided in the supplement. We asked the participants to annotate the single-rater dataset first because this also acted as their de-facto training. Participants were blinded to the multi-rater dataset name to avoid biasing them. The Unbiased control was annotated first for the same reason. A summary of the data management procedure is provided below.

HistoricsUI. We used the Digital Slide Archive, a web-based data management tool, to assign slides and annotation tasks ([digital-slidearchive.github.io](https://github.com/digital-slidearchive)) [45]. HistoricsUI, the associated annotation interface, was used for creating, correcting, and reviewing annotations. Using a centralized setup avoids participants installing software and simplifies the dissemination of images, control over view/edit permissions, monitoring progress, and collecting results. The annotation process is illustrated in [this video](#). The process of pathologist review of annotations is illustrated in Figure S1.

HistoricsTK API. The HistoricsTK Restful Application Programming Interface (API) was used to manage data, users, and annotations programmatically. This includes uploading algorithmic suggestions, downloading participant annotations, and scalable correction of systematic annotation errors where appropriate.

Obtaining labels from multi-rater datasets

Obtaining anchor proposals. We implemented a constrained agglomerative hierarchical clustering process to obtain anchor proposals (Figure 2a). The algorithm is summarized in Figure S10. In order to have a single frame of reference for comparison, annotations from all participants and for all multi-rater datasets were clustered. After clustering, we used two rules to decide which anchor proposals corresponded to actual nuclei (for each multi-rater dataset independently): 1. At least two pathologists must detect a nucleus. 2. The inferred P-truth must concur that the anchor is a nucleus.

Inference of NP-labels and P-truth. We used the Expectation-Maximization (EM) framework described by Dawid and Skene and implemented in Python by Zheng et al. [46, 47, 57]. Each participant was assigned an initial quality score of 0.7, and 70 EM iterations were performed. As illustrated in Figure 2b, undetected was considered a nucleus class for P-truth/NP-label inference. The same process was used to infer whether the boundary of an algorithmic suggestion was accurate. In effect, the segmentation accuracy was modeled as a binary variable (clicked vs. not clicked), and the EM procedure was applied to infer its value.

Class grouping

We defined two levels of grouping for nuclei classes as illustrated in Figure 2c. This was done for both the single-rater and multi-rater dataset annotations. Aggregate EM probability was calculated by summing probabilities across subsets.

Participant agreement

Overall interrater agreement was measured using Krippendorff's alpha statistic, implemented in Python by Santiago Castro and Thomas Grill [58, 59, 60]. This statistic was chosen because of its ability to handle missing values [61]. Pairwise interrater agreement was measured using Cohen's Kappa statistic [62]. Likewise, self-agreement was measured using Cohen's Kappa. All of these measures range from -1 (perfect disagreement) to +1 (perfect agreement). A kappa (or alpha) value of zero represents agreement that is expected by random chance. We used thresholds set by Fleiss for defining slight, fair, moderate, substantial, and near-perfect agreement [61].

Annotation redundancy simulations

We performed simulations to measure the impact of the number of NPs assigned to each FOV on the accuracy of NP-label inference (Figure 3e). We kept the total number of NPs constant at 18 and randomly removed annotations to obtain a desired number of NPs per FOV. No constraints were placed on how many FOVs any single NP had. This simulated the realistic scenario where participants can annotate as many FOVs as they want, and our decision-making focuses on FOV assignment. For each random realization, we calculated the inferred NP-labels using EM and measured accuracy against the static P-truth. This process was repeated for 1000 random realizations per configuration.

Software

Data management, machine learning models, and plotting were all implemented using Python 3+. Pytorch and Tensorflow libraries were used for various deep-learning experiments. Scikit-learn, Scikit-image, OpenCV, HistomicsTK, Scipy, Numpy, and Pandas libraries were used for matrix and image processing operations. Openslide library and HistomicsTK API were used for interaction with whole-slide images.

Statistical tests

The Mann-Whitney U test was used for unpaired comparisons. The Wilcoxon signed-rank test was used for paired comparisons. Confidence bounds for the AUROC values were obtained by bootstrap sampling with replacement using 1000 trials [63, 64]. AUROC values are presented in the format: value[5th percentile, 95th percentile].

Conclusion

In summary, we have described a scalable crowdsourcing approach that benefits from the participation of NPs to reduce pathologist effort and enables minimal-effort collection of segmentation boundaries. We systematically examined aspects related to the interrater agreement and truth inference. There are important limitations and opportunities to improve on our work. Our results suggest that the participation of NPs can help address the scarcity of pathologists' availability, especially for repetitive annotation tasks. This benefit, however, is restricted to annotating predominant and visually distinctive patterns. Naturally, pathologist input — and pos-

sibly full-scale annotation effort— would be needed to supplement uncommon and challenging classes that require greater expertise. Some nuclear classes may be challenging to annotate in H&E stained slides reliably and would be subject to considerable interrater variability even among practicing pathologists. In these settings, and where resources allow, IHC stains may be used as a more objective form of ground truth [65].

We chose to engage medical students and graduates with the presumption that familiarity with basic histology would help acquire higher-quality data. Whether this presumption was warranted or whether it was possible to engage a broader pool of participants was not investigated. On a related note, while we observed differences based on pathologist expertise, this was not our focus. We expect to address related questions such as the value of fellowship specialization in future work. Also, we did not measure the time it took participants to create annotations; we relied on the safe assumption that certain annotation types evidently take less time and effort than others.

Another limitation is that the initial bootstrapped nuclear boundaries were generated using classical image processing methods, which tend to underperform where nuclei are highly clumped or have very faint staining. This theoretically introduces some bias in our dataset, with an overrepresentation of simpler nuclear boundaries. We focused our annotation efforts on nucleus detection, as opposed to whole cells. Nuclei have distinct staining (hematoxylin) and boundaries, potentially reducing the interrater variability associated with the detection of cell boundaries. Finally, we would point out that dataset curation is context-dependent and likely differs depending on the problem. Nevertheless, we trust that most of our conclusions have broad implications for other histopathology annotation efforts.

Availability of supporting data and materials

The datasets used are available at the [NuCLS website](#).

Declarations

List of abbreviations

API: Application Programming Interface; BCSS: Breast Cancer Semantic Segmentation; CNN: Convolutional neural networks; EM: Expectation Maximization; FOV: Field of view; H&E: Hematoxylin and Eosin; IOU: Intersection over union; MDS: Multidimensional scaling; NPs: Non-pathologists; NP-label: Inferred label from multi-rater pathologist data; NuCLS: Nucleus classification, localization, and segmentation; P-truth: Inferred truth from multi-rater pathologist data; ROI: Region of Interest; TCGA: The Cancer Genome Atlas

Ethical Approval

Not applicable.

Consent for publication

Not applicable.

Competing Interests

The author(s) declare that they have no competing interests.

Funding

This work was supported by the U.S. National Institutes of Health National Cancer Institute grants U01CA220401 and U24CA19436201. Lee A.D. Cooper is the Principal Investigator for the grants. The funding body had no role in the design of the study, data collection, data analysis, or data interpretation, or writing the manuscript.

Author's Contributions

M.A. and L.A.D.C. conceived the hypothesis, designed the experiments, performed the analysis, and wrote the manuscript. D.M. and D.A.G. contributed support for the Digital Slide Archive software and database. B.D. and D.J. provided ideas for the interrater analysis. M.A. and M.A.T.E. were the study coordinators and corrected the single-rater dataset. H.E. provided feedback and approved the corrected single-rater dataset. E.H. provided manual nucleus segmentation data. H.E., H.H., and E.H. are senior pathologists and provided multi-rater annotations. L.A.A., K.H.M., P.A.P., and L.E.H. are junior pathologists and provided multi-rater annotations. M.A.T.E., A.M.A., M.A.A., A.M.E., R.A.S., A.R., A.M.S., A.M.A., I.A.R., A.A., N.M.E., A.A., A.F., A.E., A.G.E., Y.A., Y.A.A., A.M.R., M.K.N., M.A.T.E., A.A., A.G., and M.E. are non-pathologists and provided single- and multi-rater annotations. All experience designations are based on the time of annotation. All authors reviewed the manuscript draft.

Acknowledgements

We would like to acknowledge with gratitude the contributions made by the following participants: Eman Elsayed Sakr (El-Matariya Teaching Hospital, Egypt), Joumana Ahmed (Cairo University, Egypt); Mohamed Zalabia and Ahmed S. Badr (Menoufia University, Egypt); Ahmed M. Afifi (Ain Shams University, Egypt); Esraa B. Ghabban (Damascus University, Syria); Mahmoud A. Hashim (Baylor College of Medicine, USA). In addition, we are thankful to Uday Kurkure, Jim Martin, Raghavan Venugopal, Joachim Schmidt (Roche Tissue Diagnostics, USA), and Michael Barnes (Roche Diagnostic Information Solutions, USA) for support and discussions. We also thank Brian Finkelman for constructive feedback on the interrater analysis. Finally, we thank Jeff Goldstein and other members of the Cooper research group at Northwestern for constructive feedback and discussion.

Full list of author affiliations

¹Department of Pathology, Northwestern University, Chicago, IL, USA and ²Cairo Health Care Administration, Egyptian Ministry of Health, Cairo, Egypt and ³Department of Pathology, Nasser institute for research and treatment, Cairo, Egypt and ⁴Department of Pathology and Laboratory Medicine, University of Pennsylvania, PA, USA and ⁵Department of Clinical Laboratory Research, Theodor Bilharz Research Institute, Giza, Egypt and ⁶Department of Medicine, Cook County Hospital, Chicago, IL, USA and ⁷Department of Pathology, Baystate Medical Center, University of Massachusetts, Springfield, MA, USA and ⁸Faculty of Medicine, Menoufia University, Menoufia, Egypt and ⁹Faculty of Medicine, Al-Azhar University, Cairo, Egypt and ¹⁰Consultant for The Center for Applied Proteomics and Molecular Medicine (CAPMM), George Mason University, Manassas, VA, USA and ¹¹Department of Pathology, National Liver Institute, Menoufia University, Menoufia, Egypt and ¹²Faculty of Medicine, Ain Shams University, Cairo, Egypt and ¹³Cleveland Clinic Foundation, Cleveland, OH, USA and ¹⁴Department of Pathology, Indiana University, Indianapolis, IN, USA and ¹⁵Faculty of Medicine, Damascus University, Damascus, Syria and ¹⁶Faculty of Medicine, Mansoura University, Mansoura, Egypt and ¹⁷Faculty of Medicine, Cairo University, Cairo, Egypt and ¹⁸Department of Anaesthesia and Critical Care, Menoufia University Hospital, Menoufia, Egypt and ¹⁹Department of Clinical Pathology, Ain Shams University, Cairo, Egypt and ²⁰Research Department, Oncology Consultants, PA, Houston, TX, USA and ²¹Siparadigm Diagnostic Infor-

matics, Pine Brook, NJ, USA and ²²Department of Pathology and Laboratory Medicine, Emory University School of Medicine, Atlanta, GA, USA and ²³Kitware Inc., Clifton Park, NY, USA and ²⁴Department of Neurology, Emory University School of Medicine, Atlanta, GA, USA and ²⁵Department of Pathology, National Cancer Institute, Cairo, Egypt and ²⁶Department of Pathology, Children's Cancer Hospital Egypt (CCHE 57357), Cairo, Egypt and ²⁷Lurie Cancer Center, Northwestern University, Chicago, IL, USA and ²⁸Center for Computational Imaging and Signal Analytics, Northwestern University Feinberg School of Medicine, Chicago, IL, USA

References

- Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. *Med Image Anal* 2017 Dec;42:60–88.
- Abels E, Pantanowitz L, Aeffner F, Zarella MD, Laak J, Bui MM, et al., Computational pathology definitions, best practices, and recommendations for regulatory guidance: a white paper from the Digital Pathology Association; 2019.
- Hartman DJ, Van Der Laak JAWM, Gurcan MN, Pantanowitz L. Value of Public Challenges for the Development of Pathology Deep Learning Algorithms. *J Pathol Inform* 2020 Feb;11:7.
- Amgad M, International Immuno-Oncology Biomarker Working Group, Stovgaard ES, Balslev E, Thagaard J, Chen W, et al., Report on computational assessment of Tumor Infiltrating Lymphocytes from the International Immuno-Oncology Biomarker Working Group; 2020.
- Beck AH, Sangoi AR, Leung S, Marinelli RJ, Nielsen TO, van de Vijver MJ, et al. Systematic analysis of breast cancer morphology uncovers stromal features associated with survival. *Sci Transl Med* 2011 Nov;3(108):108ra113.
- Koh PW, Nguyen T, Tang YS, Musmann S, Pierson E, Kim B, et al. Concept bottleneck models. In: International Conference on Machine Learning PMLR; 2020. p. 5338–5348.
- Naik S, Doyle S, Agner S, Madabhushi A, Feldman M, Tomaszewski J, Automated gland and nuclei segmentation for grading of prostate and breast cancer histopathology; 2008.
- Cooper LAD, Kong J, Gutman DA, Wang F, Gao J, Appin C, et al. Integrated morphologic analysis for the identification and characterization of disease subtypes. *J Am Med Inform Assoc* 2012 Mar;19(2):317–323.
- Cooper LAD, Kong J, Gutman DA, Wang F, Cholleti SR, Pan TC, et al., An Integrative Approach for In Silico Glioma Research; 2010.
- Alexander J Lazar, Michael D McLellan, Matthew H Bailey, Christopher A Miller, Elizabeth L Appelbaum, Matthew G Cordes, Catrina C Fronick, The Cancer Genome Atlas Research Network. Comprehensive and Integrated Genomic Characterization of Adult Soft Tissue Sarcomas. *Cell* 2017 Nov;171(4):950–965.e28.
- Saltz J, Gupta R, Hou L, Kurc T, Singh P, Nguyen V, et al. Spatial Organization and Molecular Correlation of Tumor-Infiltrating Lymphocytes Using Deep Learning on Pathology Images. *Cell Rep* 2018 Apr;23(1):181–193.e7.
- Diao JA, Wang JK, Chui WF, Mountain V, Gullapally SC, Srinivasan R, et al. Human-interpretable image features derived from densely mapped cancer pathology slides predict diverse molecular phenotypes. *Nat Commun* 2021 Mar;12(1):1613.
- Lu W, Graham S, Bilal M, Rajpoot N, Minhas F. Capturing Cellular Topology in Multi-Gigapixel Pathology Images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops; 2020. p. 260–261.
- Alexander CB, Bruce Alexander C, Pathology graduate medical education (overview from 2006–2010); 2011.
- Kovashka A, Russakovsky O, Fei-Fei L, Grauman K, Crowdsourcing in Computer Vision; 2016.
- Ching T, Himmelstein DS, Beaulieu-Jones BK, Kalinin AA, Do BT, Way GP, et al. Opportunities and obstacles for deep learning in biology and medicine. *J R Soc Interface* 2018 Apr;15(141).
- Amgad M, Man Kin Tsui M, Liptrott SJ, Shash E. Medical Student Research: An Integrated Mixed-Methods Systematic Review and Meta-Analysis. *PLoS One* 2015 Jun;10(6):e0127470.
- Shaw S, Pajak M, Lisowska A, Tsafaris SA, O'Neil AQ. Teacher-student chain for efficient semi-supervised histology image classification. arXiv preprint arXiv:200308797 2020;

19. Hou L, Agarwal A, Samaras D, Kurc TM, Gupta RR, Saltz JH, Robust Histopathology Image Analysis: To Label or to Synthesize?; 2019.
20. Irshad H, Montaser-Kouhsari L, Waltz G, Bucur O, Nowak JA, Dong F, et al., Crowdsourcing image annotation for nucleus detection and segmentation in computational pathology: Evaluating experts, automated methods, and the crowd; 2014.
21. Campanella G, Hanna MG, Geneslaw L, Miraflor A, Werneck Krauss Silva V, Busam KJ, et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nat Med* 2019 Aug;25(8):1301–1309.
22. Alemi Koohbanani N, Jahanifar M, Zamani Tajadin N, Rajpoot N. NuClick: A deep learning framework for interactive segmentation of microscopic images. *Med Image Anal* 2020 Oct;65:101771.
23. Deshpande S, Minhas F, Graham S, Rajpoot N. SAFRON: Stitching Across the Frontier for Generating Colorectal Cancer Histology Images. arXiv preprint arXiv:200804526 2020;
24. Mahmood F, Borders D, Chen RJ, Mckay GN, Salimian KJ, Baras A, et al. Deep Adversarial Training for Multi-Organ Nuclei Segmentation in Histopathology Images. *IEEE Trans Med Imaging* 2020 Nov;39(11):3257–3267.
25. Koohbanani NA, Unnikrishnan B, Khurram SA, Krishnaswamy P, Rajpoot N. Self-Path: Self-supervision for Classification of Pathology Images with Limited Annotations. *IEEE Transactions on Medical Imaging* 2021;
26. Örteng S, Doyle A, van Hilten A, Hirth M, Inel O, Madan CR, et al. A survey of crowdsourcing in medical image analysis. arXiv preprint arXiv:190209159 2019;
27. Marzahl C, Aubreville M, Bertram CA, Gerlach S, Maier J, Voigt J, et al. Fooling the crowd with deep learning-based methods. arXiv preprint arXiv:191200142 2019;
28. Amgad M, Elfandy H, Hussein H, Atteya LA, Elsebaie MAT, Abo Elnasr LS, et al. Structured crowdsourcing enables convolutional segmentation of histology images. *Bioinformatics* 2019 Sep;35(18):3461–3467.
29. Graham S, Vu QD, Raza SEA, Azam A, Tsang YW, Kwak JT, et al. HoverNet: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. *Med Image Anal* 2019 Dec;58:101563.
30. Kumar N, Verma R, Anand D, Zhou Y, Onder OF, Tsougenis E, et al. A Multi-Organ Nucleus Segmentation Challenge. *IEEE Trans Med Imaging* 2020 May;39(5):1380–1391.
31. Xing F, Yang L. Robust Nucleus/Cell Detection and Segmentation in Digital Pathology and Microscopy Images: A Comprehensive Review. *IEEE Rev Biomed Eng* 2016 Jan;9:234–263.
32. Gamper J, Koohbanani NA, Benet K, Khuram A, Rajpoot N, PanNuke: An Open Pan-Cancer Histology Dataset for Nuclei Instance Segmentation and Classification; 2019.
33. Gamper J, Koohbanani NA, Benes K, Graham S, Jahanifar M, Khurram SA, et al. Pannuke dataset extension, insights and baselines. arXiv preprint arXiv:200310778 2020;
34. Veta M, Heng YJ, Stathonikos N, Bejnordi BE, Beca F, Wollmann T, et al. Predicting breast tumor proliferation from whole-slide images: The TUPAC16 challenge. *Med Image Anal* 2019 May;54:111–121.
35. Janowczyk A, Madabhushi A. Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. *J Pathol Inform* 2016 Jul;7:29.
36. Verma R, Kumar N, Patil A, Kurian NC, Rane S, Sethi A. Multi-organ nuclei segmentation and classification challenge 2020. *IEEE Trans Med Imaging* 2020;39:1380–1391.
37. Graham S, Jahanifar M, Azam A, Nimir M, Tsang YW, Dodd K, et al. Lizard: A Large-Scale Dataset for Colonic Nuclear Instance Segmentation and Classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2021. p. 684–693.
38. Verma R, Kumar N, Patil A, Kurian NC, Rane S, Graham S, et al. MoNuSAC2020: A Multi-organ Nuclei Segmentation and Classification Challenge. *IEEE Trans Med Imaging* 2021 Jun;PP.
39. Dudgeon SN, Wen S, Hanna MG, Gupta R, Amgad M, Sheth M, et al. A pathologist-annotated dataset for validating artificial intelligence: a project description and pilot study. arXiv preprint arXiv:201006995 2020;
40. Litjens G, Bandi P, Ehteshami Bejnordi B, Geessink O, Balkenhol M, Bult P, et al. 1399 H&E-stained sentinel lymph node sections of breast cancer patients: the CAMELYON dataset. *Gigascience* 2018 Jun;7(6).
41. Hou L, Gupta R, Van Arnam JS, Zhang Y, Sivalenka K, Samaras D, et al. Dataset of segmented nuclei in hematoxylin and eosin stained histopathology images of ten cancer types. *Sci Data* 2020 Jun;7(1):185.
42. Nalisnik M, Amgad M, Lee S, Halani SH, Velazquez Vega JE, Brat DJ, et al. Interactive phenotyping of large-scale histology imaging data with HistomicsML. *Sci Rep* 2017 Nov;7(1):14588.
43. Amgad M, Atteya L, Hussein H, Mohammed KH, Hafiz E, Elsebaie MAT, et al. Explainable nucleus classification using Decision Tree Approximation of Learned Embeddings. *Bioinformatics* 2021 Sep;
44. He K, Gkioxari G, Dollár P, Girshick R. Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision; 2017. p. 2961–2969.
45. Gutman DA, Khalilia M, Lee S, Nalisnik M, Mullen Z, Beezley J, et al. The Digital Slide Archive: A Software Platform for Management, Integration, and Analysis of Histology for Cancer Research. *Cancer Res* 2017 Nov;77(21):e75–e78.
46. Dawid AP, Skene AM, Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm; 1979.
47. Zheng Y, Li G, Li Y, Shan C, Cheng R, Truth inference in crowdsourcing; 2017.
48. Khoreva A, Benenson R, Hosang J, Hein M, Schiele B, Simple Does It: Weakly Supervised Instance and Semantic Segmentation; 2017.
49. Amgad M, Sarkar A, Srinivas C, Redman R, Ratra S, Bechert CJ, et al. Joint Region and Nucleus Segmentation for Characterization of Tumor Infiltrating Lymphocytes in Breast Cancer. *Proc SPIE Int Soc Opt Eng* 2019 Feb;10956.
50. Salgado R, Denkert C, Demaria S, Sirtaine N, Klauschen F, Pruneri G, et al., The evaluation of tumor-infiltrating lymphocytes (TILs) in breast cancer: recommendations by an International TILs Working Group 2014; 2015.
51. Macenko M, Niethammer M, Marron JS, Borland D, Woosley JT, Xiaojun Guan, et al. A method for normalizing histology slides for quantitative analysis. In: 2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro ieeexplore.ieee.org; 2009. p. 1107–1110.
52. Otsu N. A threshold selection method from gray-level histograms. *IEEE Trans Syst Man Cybern* 1979 Jan;9(1):62–66.
53. Gonzalez R, Woods R, Digital Image Processing, (March 1992). Addison-Wesley Publishing Company; 1992.
54. Maurer CR, Rensheng Qi, Raghavan V. A linear time algorithm for computing exact Euclidean distance transforms of binary images in arbitrary dimensions. *IEEE Trans Pattern Anal Mach Intell* 2003 Feb;25(2):265–270.
55. Beucher S. Use of watersheds in contour detection. In: Proceedings of the International Workshop on Image Processing; 1979. .
56. Soille PJ, Ansoult MM. Automated basin delineation from digital elevation models using mathematical morphology. *Signal Processing* 1990 Jun;20(2):171–182.
57. Zheng Y, Li G, Li Y, Shan C, Cheng R, Crowdsourcing truth inference (Github); Accessed: 2020-12-19. https://github.com/zhydhkws/crowd_truth_infer.
58. Krippendorff K. Krippendorff, Klaus, Content Analysis: An Introduction to its Methodology. Beverly Hills, CA: Sage, 1980 1980;
59. Castro S, Fast Krippendorff; Accessed: 2020-12-19. <https://github.com/pln-fing-udelar/fast-krippendorff>.
60. Grill T, Krippendorff alpha; Accessed: 2020-12-19. <https://github.com/grrrr/krippendorff-alpha>.
61. Fleiss JL. Measuring nominal scale agreement among many raters. *Psychol Bull* 1971 Nov;76(5):378–382.
62. Cohen J. A Coefficient of Agreement for Nominal Scales. *Educ Psychol Meas* 1960 Apr;20(1):37–46.
63. Mann HB, Whitney DR. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *Ann Math Stat* 1947;18(1):50–60.
64. Wilcoxon F. Individual Comparisons by Ranking Methods. In: Kotz S, Johnson NL, editors. Breakthroughs in Statistics: Methodology and Distribution New York, NY: Springer New York; 1992.p. 196–202.
65. Tellez D, Balkenhol M, Otte-Holler I, van de Loo R, Vogels R, Bult P, et al., Whole-Slide Mitosis Detection in H&E Breast Histology Using PHH3 as a Reference to Train Distilled Stain-Invariant Convolutional Networks; 2018.

Section 2.3

A pathologist-annotated dataset for validating artificial intelligence: a project description and pilot study

This section is an exact reproduction of the following open-access journal paper:

*Dudgeon, S.N., Wen, S., Hanna, M.G., Gupta, R., **Amgad, M.**, Sheth, M., Marble, H., Huang, R., Herrmann, M.D., Szu, C.H. and Tong, D., 2020. A pathologist-annotated dataset for validating artificial intelligence: a project description and pilot study. J Pathol Inform, 2021;12:45.*

Candidate's role: Developing the protocol in consultation with other co-authors, involvement in discussions on methodology and software interface, editing manuscript.

[Downloaded free from <http://www.jpathinformatics.org> on Wednesday, November 24, 2021, IP: 165.124.124.231]

Technical Note

A Pathologist-Annotated Dataset for Validating Artificial Intelligence: A Project Description and Pilot Study

Sarah N. Dudgeon¹, Si Wen¹, Matthew G. Hanna², Rajarsi Gupta³, Mohamed Amgad⁴, Manasi Sheth⁵, Hetal Marble⁶, Richard Huang⁶, Markus D. Herrmann⁶, Clifford H. Szu⁷, Darick Tong⁷, Bruce Werness⁷, Evan Szu⁷, Denis Larsimon⁸, Anant Madabhushi⁹, Evangelos Hytopoulos¹⁰, Weijie Chen¹, Rajendra Singh¹¹, Steven N. Hart⁶, Ashish Sharma¹², Joel Saltz², Roberto Salgado^{13,14}, Brandon D. Gallas¹

¹Division of Imaging Diagnostics and Software Reliability, Office of Science and Engineering Laboratories, Center for Devices and Radiologic Health, United States Food and Drug Administration, White Oak, MD, USA, ²Memorial Sloan Kettering Cancer Center, New York, NY, USA, ³Department of Biomedical Informatics, Stony Brook University, Stony Brook, NY, USA, ⁴Department of Pathology, Northwestern University, Chicago, IL, USA, ⁵Division of Biostatistics, Center for Devices and Radiologic Health, United States Food and Drug Administration, White Oak, MD, USA, ⁶Department of Pathology, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA, ⁷Arrive Bio, San Francisco, CA, USA, ⁸Department of Pathology, Institute Jules Bordet, Brussels, Belgium, ⁹Louis Stokes Cleveland Veterans Administration Medical Center, Cleveland, OH, USA, ¹⁰iRhythm Technologies Inc., San Francisco, CA, USA, ¹¹Northwell Health and Zucker School of Medicine, New York, NY, USA, ¹²Department of Biomedical Informatics, Emory University, Atlanta, GA, USA, ¹³Division of Research, Peter Mac Callum Cancer Centre, Melbourne, Australia, ¹⁴Department of Pathology, GZA-ZNA Hospitals, Antwerp, Belgium

Submitted: 28-Sep-2020

Revised: 23-Jan-2021

Accepted: 16-Mar-2021

Published: 15-Nov-2021

Abstract

Purpose: Validating artificial intelligence algorithms for clinical use in medical images is a challenging endeavor due to a lack of standard reference data (ground truth). This topic typically occupies a small portion of the discussion in research papers since most of the efforts are focused on developing novel algorithms. In this work, we present a collaboration to create a validation dataset of pathologist annotations for algorithms that process whole slide images. We focus on data collection and evaluation of algorithm performance in the context of estimating the density of stromal tumor-infiltrating lymphocytes (sTILs) in breast cancer. **Methods:** We digitized 64 glass slides of hematoxylin- and eosin-stained invasive ductal carcinoma core biopsies prepared at a single clinical site. A collaborating pathologist selected 10 regions of interest (ROIs) per slide for evaluation. We created training materials and workflows to crowdsource pathologist image annotations on two modes: an optical microscope and two digital platforms. The microscope platform allows the same ROIs to be evaluated in both modes. The workflows collect the ROI type, a decision on whether the ROI is appropriate for estimating the density of sTILs, and if appropriate, the sTIL density value for that ROI. **Results:** In total, 19 pathologists made 1645 ROI evaluations during a data collection event and the following 2 weeks. The pilot study yielded an abundant number of cases with nominal sTIL infiltration. Furthermore, we found that the sTIL densities are correlated within a case, and there is notable pathologist variability. Consequently, we outline plans to improve our ROI and case sampling methods. We also outline statistical methods to account for ROI correlations within a case and pathologist variability when validating an algorithm. **Conclusion:** We have built workflows for efficient data collection and tested them in a pilot study. As we prepare for pivotal studies, we will investigate methods to use the dataset as an external validation tool for algorithms. We will also consider what it will take for the dataset to be fit for a regulatory purpose: study size, patient population, and pathologist training and qualifications. To this end, we will elicit feedback from the Food and Drug Administration via the Medical Device Development Tool program and from the broader digital pathology and AI community. Ultimately, we intend to share the dataset, statistical methods, and lessons learned.

Keywords: Artificial intelligence validation, medical image analysis, pathology, reference standard, tumor-infiltrating lymphocytes

INTRODUCTION

Artificial intelligence (AI) is often used to describe machines or computers that mimic “cognitive” functions associated with the human mind, such as “learning” and “problem-solving.”^[1] Machine learning (ML) is an AI technique that can be used to design and train software algorithms to learn from and act on data. Although AI/ML has existed for some time, recent

Address for correspondence: Dr. Brandon D. Gallas, Division of Imaging, Diagnostics, and Software Reliability, Office of Science and Engineering Laboratories, Center for Devices and Radiological Health, US Food and Drug Administration, 10903 New Hampshire Ave, Silver Spring, Rm 4104, White Oke-62, MD 20993, USA.
E-mail: brandon.gallas@fda.hhs.gov

This is an open access journal, and articles are distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 License, which allows others to remix, tweak, and build upon the work non-commercially, as long as appropriate credit is given and the new creations are licensed under the identical terms.

For reprints contact: WKHLRPMedknow_reprints@wolterskluwer.com

How to cite this article: Dudgeon SN, Wen S, Hanna MG, Gupta R, Amgad M, Sheth M, *et al.* A pathologist-annotated dataset for validating artificial intelligence: A project description and pilot study. *J Pathol Inform* 2021;12:45.

Available FREE in open access from: <http://www.jpathinformatics.org/text.asp?2021/12/1/45/330486>

Access this article online	
Quick Response Code: 	Website: www.jpathinformatics.org
	DOI: 10.4103/jpi.jpi_83_20

[Downloaded free from <http://www.jpathinformatics.org> on Wednesday, November 24, 2021, IP: 165.124.124.231]

J Pathol Inform 2021, 1:45

<http://www.jpathinformatics.org/content/12/1/45>

advances in algorithm architecture, software tools, hardware infrastructure, and regulatory frameworks have enabled health-care stakeholders to harness AI/ML as a medical device. Such medical devices have the potential to offer enhanced patient care by streamlining operations, performing quality control, supporting diagnostics, and enabling novel discovery.

While AI/ML has already found utility in radiology, the role of AI/ML algorithms in pathology has been a matter of wide discussion.^[2-7] Recent technological advancements and market access of systems that scan glass slides to create digital whole slide images (WSIs) have opened the door to a myriad of opportunities for AI/ML applications in digital pathology.^[8,9] While pathology is new to digitization, the field is expected to extend algorithms to a broad range of clinical decision support tasks. This technology shift is reminiscent of the digitization of mammography in 2000^[10] and the first computer-aided detection (CADe) device in radiology in 1998, the R2 ImageChecker.^[11] The R2 CADe device marked regions of interest (ROIs) likely to contain microcalcifications or masses, initially evaluating digitized screen-film mammograms rather than digital acquisition of mammography images.

Fourteen years after the R2 ImageChecker was approved by the US Food and Drug Administration (FDA), regulatory guidance for CADe was finalized in two documents. While both guidance documents are specific to radiology, their principles are applicable to other specialties, including digital pathology. The first document generally delineates how to describe a CADe device and assess its “stand-alone” performance.^[12] In the pathology space, this might be referred to as analytical validation. The second guidance document covers clinical performance assessment, or clinical validation.^[13] The document was recently updated and discusses issues such as study design, study population, and the reference standard. Related issues are also discussed in a paper summarizing a meeting jointly hosted by the FDA and the Medical Imaging Perception Society.^[14]

Regardless of the technology providing the data or the algorithm architecture, Software as a Medical Device (SaMD) must be analytically and clinically validated to ensure safety and effectiveness before clinical deployment.^[15] One critical aspect of algorithm validation is to assess accuracy. Accuracy compares algorithm predictions to true labels using holdout validation data, data that are independent from data used in development. Validation data include patient data (images and metadata) on which the algorithm will make predictions as well as the corresponding reference standard (ground truth or label). The reference standard can be established using an independent “gold standard” modality, longitudinal patient outcomes, or when these are not available or appropriate, a reference standard established by human experts. What constitutes the “ground truth” and how to approach it is a topic of discussion even in more traditional diagnostic test paradigms, and certainly so in evolving areas such as SaMD.

In this work, we focus on the often challenging task of establishing a reference standard using pathologists. The “interpretation by a reviewing clinician” is listed as a reference standard in the radiology CADe guidance documents and acts as the reference standard (in full or in part) in many precedent-setting radiology applications.^[16-18] In pathology, the reference standard for evaluating performance in the Philips IntelliSite Pathology Solution regulatory submission, “was based on the original sign-out diagnosis rendered at the institution, using an optical (light) microscope.”^[8]

In this manuscript, we present a collaborative project to produce a validation dataset established by pathologist annotations. The project will additionally produce statistical analysis tools to evaluate algorithm performance. The context of this work is the validation of an algorithm that measures, or estimates, the density of tumor-infiltrating lymphocytes (TILs), a prognostic biomarker in breast cancer. Resulting tools and data may be used to facilitate the external validation of an algorithm within the applied context. Given the cross-disciplinary nature of the study, the volunteer effort comprises an international, multidisciplinary team working in the precompetitive space. Project participants include the FDA Center for Devices and Radiological Health’s Office of Science and Engineering Laboratories, clinician-scientists from international health systems, academia, professional societies, and medical device manufacturers. By incorporating diverse stakeholders, we aim to address multiple perspectives and emphasize interoperability across platforms.

We are pursuing qualification of the final validation dataset as an FDA Medical Device Development Tool (MDDT).^[19] In doing so, we have an opportunity to receive feedback from an FDA review team while building the dataset. If the dataset qualifies as an MDDT, it will be a high-value public resource that can be used in AI/ML algorithm submissions, and our work may guide others to develop their own validation datasets.

Definitions of terms in AI-based medical device development and regulation are evolving. For example, there has been inconsistent usage of “testing” versus “validation.” To avoid this confusion, we are referring to building, training, tuning, and validating algorithms, where tuning is for hyperparameter optimization, and validation is for assessing or testing the performance of AI/ML algorithms. There is also some confusion between the terms “algorithm” and “model.” In this work, we will use the term “algorithm” to refer to the SaMD, the device, the software that is or will be deployed. Some may refer to the SaMD as the “model,” but we shall use “model” to refer to the description of the algorithm (the architecture, image normalization, transfer learning, augmentation, loss function, training, hyperparameter selection, etc.).

Herein, we present our efforts to source a pathologist-driven reference standard and apply it to algorithm validation, with an eye toward generating a fit-for-regulatory-purpose dataset. Specifically, we review the clinical association between TILs and patient outcomes in the context of accepted guidelines for

[Downloaded free from <http://www.jpathinformatics.org> on Wednesday, November 24, 2021, IP: 165.124.124.231]

J Pathol Inform 2021, 1:45

<http://www.jpathinformatics.org/content/12/1/45>

estimating TIL density in tumor-associated stroma (stromal TIL [sTIL] density). We then imagine an algorithm that similarly estimates sTIL density and could use a sTIL density annotated dataset for validation. Next, we describe the breast cancer tissue samples used in our pilot study, the data collection methods and platforms, and the pathologists we recruited and trained to provide sTIL density estimates in ROIs using digital and microscope platforms. We also present some initial data and outline how we plan to account for pathologist variability when estimating algorithm performance.

TECHNICAL BACKGROUND

Tumor-infiltrating Lymphocytes

TILs are an inexpensively assessed, robust, prognostic biomarker that is a surrogate for antitumor, T-cell-mediated immunity. Clinical validity of TILs as a prognostic biomarker in early-stage, triple-negative breast cancer (TNBC), as well as in HER2+ breast cancer, has been well-established via Level 1b evidence.^[20-23] Two pooled analyses of TILs, in the adjuvant setting for TNBC^[21] and neoadjuvant setting across BC subtypes,^[22] included studies that have evaluated TILs in archived tissue samples based on published guidelines.^[24] Incorporating TILs into standard clinical practice for TNBC is endorsed by international clinical and pathology standards (St. Gallen 2019 recommendation, WHO 2019 recommendation, and ESMO2019 recommendation).^[25-28] It is expected that TILs will be assessed to monitor treatment response in the future.^[29,30] Further, evidence is emerging that TIL-assessment will be done in other tumor types as well, including melanoma, gastrointestinal tract carcinoma, non-small cell lung carcinoma and mesothelioma, and endometrial and ovarian carcinoma.^[31,32]

Visual and Computational Tumor-infiltrating Lymphocyte Assessment

Given the recent and evolving evidence of the prognostic value of TIL assessment, there have been several efforts to create algorithms to estimate TIL density in cancer tissue. Amgad *et al.* provide an excellent summary of this space, including a table of algorithms from the literature, an outline with visual aids for TIL assessment, as well as a discussion on validation and training issues.^[32,33] While some algorithms are leveraging details about the spatial distribution of individual TILs in different tissue compartments,^[34-36] the guidelines for pathologists are to calculate the sTIL density^[24] defined as the area of sTILs divided by the area of the corresponding tumor-associated stroma.

In this work, we imagine an algorithm that estimates the density of sTILs in pathologist-marked ROIs in WSIs of hematoxylin- and eosin-stained slides (H&E) containing breast cancer needle core biopsies. Amgad *et al.* refer to these quantitative values as computational TIL assessments and visual TIL assessments, respectively.^[32] Such an algorithm produces quantitative values^[37] that are equivalent to those proposed in the guidelines for pathologists. This provides the

opportunity for using pathologist evaluations as the reference standard for such an algorithm.

We propose the following clinical workflow: (1) patient imaging finds an abnormality suspected for breast cancer. Physicians order a needle core biopsy to assess the tissue. (2) TILs will be scored during histopathologic evaluation and diagnosis. Specifically, pathologists will score the TILs in each H&E-stained breast cancer core biopsy with assistance from an algorithm. Or, depending on the algorithm intended use, the sTIL score could be created automatically, without pathologist input. (3) The sTIL density will then be reported in the patient's pathology report.

Algorithm Validation

Before it can be marketed and applied in the clinical workflow, any algorithm/SaMD should be well validated. Validation of algorithms for clinical use comes after the building, training, and tuning phases of algorithm development. There are two main categories of algorithm validation: analytical and clinical. For both categories, a reference standard is needed. For algorithms that evaluate WSIs of H & E slides, there are generally three kinds of truth: patient outcomes, evaluation of the tissue with other diagnostic methods, and evaluation of the slide by pathologists. This work focuses on truth as determined by pathologists.

Analytical validation, or stand-alone performance assessment, focuses on the precision and accuracy of the algorithm, and compares algorithm outputs directly against the reference standard [Figure 1a]. In a clinical validation study, the algorithm end user, here a pathologist, evaluates cases without and with the algorithm outputs; Figure 1b shows an independent-crossover clinical validation study design. There is typically a washout period between the evaluations by the same pathologists evaluating the same cases without and with the algorithm outputs, where the order in which these viewing modes are executed is randomized and balanced across pathologists and batches of cases. Figure 1c shows a putative sequential clinical validation study design for an algorithm intended to be used as a decision support tool after the clinician makes their conventional evaluation. We have depicted two populations of pathologists in our proposed clinical validation studies: experts for establishing the reference standard and end users for evaluating performance without and with the algorithm outputs.

The current best practice for algorithm validation is to source slides from multiple independent sites different from the algorithm development site to ensure algorithm generalizability, also known as external validation.^[38-41] Developers should also be blinded to the validation data before a validation study, eliminating potential bias arising from developers' training to the test.^[41-44] These practices generally assume that the algorithm is locked; the architecture, parameters, weights, and thresholds should not be changed before the algorithm is released into the field. Validation of algorithms that are not locked – algorithms that rely on “active learning” and “online”

[Downloaded free from <http://www.jpathinformatics.org> on Wednesday, November 24, 2021, IP: 165.124.124.231]

J Pathol Inform 2021, 1:45

<http://www.jpathinformatics.org/content/12/1/45>

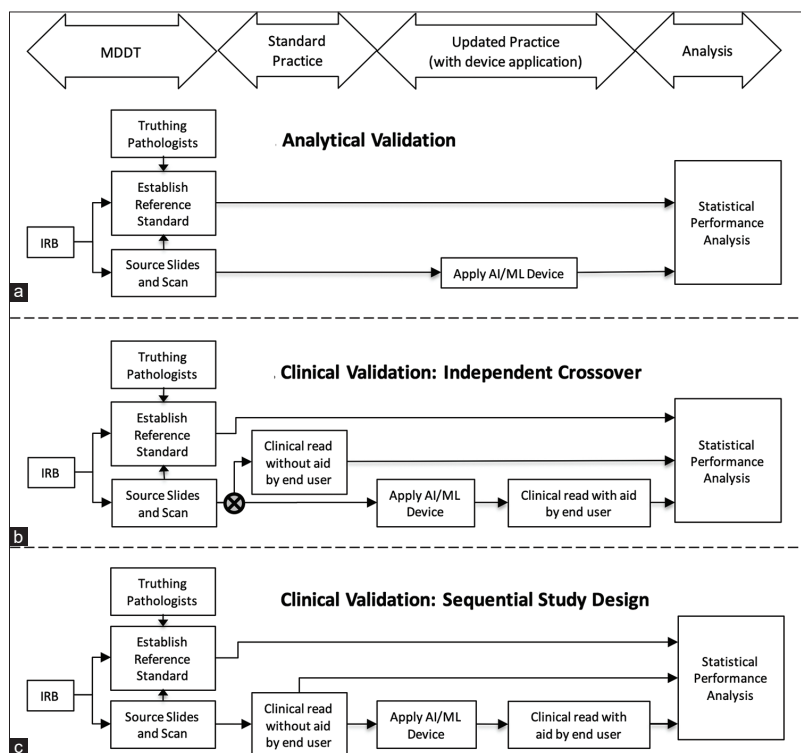


Figure 1: (a) Study design for analytical validation of an algorithm (stand-alone performance assessment). Algorithm outputs are compared to the reference standard. (b) Independent crossover study design for clinical validation has two arms corresponding to pathologist evaluations without and with the algorithm. We compare the performance of these two evaluation modes. (c) Sequential study design for clinical validation has one arm corresponding to end user evaluations first without and then with the algorithm as an aid. A comparison is made between the performance of these two evaluation modes

learning, or hard negative mining, where the training is done iteratively and continuously – is an area that is still evolving and not in the scope of this work.^[32,45-49]

APPROACH: PILOT STUDY

Data – Pathology Tissue and Images

We, through a partnership with the Institute Jules Bordet, Brussels, Belgium, sourced 77 matched core biopsies and surgical resections. Of these cases, 65 were classified as invasive ductal carcinoma and 12 were invasive lobular carcinoma. There was no patient information provided with these slides, no metadata such as age, race, cancer stage, or subtype (morphologic or molecular). This study was approved by the Ethics Commission of the Institute Jules Bordet.

The slides are 2019 recuts of formalin-fixed, paraffin-embedded tissue blocks from a single institution. Slide preparation was performed at the same institution by a single laboratory technician. Specifically, one 5 µm-thick section was mounted on a glass slide and stained with H & E.

The slides were scanned on a Hamamatsu Nanozoomer 2.0-RS C10730 series at $\times 40$ equivalent magnification (scale: 0.23 µm per pixel).

For our pilot study, we included eight batches of eight cases each; a case refers to the slide image pair. The remaining 13 slides were not used for the pilot study. All 64 cases were biopsies of invasive ductal carcinomas; no resection specimens were used. Batches split data collection into manageable chunks for pathologists. Each batch was expected to take about 30 min to annotate. Batches also allowed us to make assignments for pathologists that help distribute evaluations across all cases and ROIs. We targeted five pathologist evaluations per ROI for the pilot study.

Data Collection = Region of Interest Annotation

Data collection, or ROI annotation, is broken into ROI selection and ROI evaluation in this work. ROI selection is a data curation step preceding ROI evaluation. The purpose of selecting ROIs ahead of ROI evaluation is to allow multiple pathologists to evaluate the same ROIs quickly. For our pilot study, ROI selection was performed by a collaborating pathologist using the digital platforms. Subsequent ROI

[Downloaded free from <http://www.jpathinformatics.org> on Wednesday, November 24, 2021, IP: 165.124.124.231]

J Pathol Inform 2021, 1:45

<http://www.jpathinformatics.org/content/12/1/45>

evaluation was performed by recruited pathologists using digital and microscope platforms. The platforms, ROI selection and evaluation, and the pathologists that participated in the pilot study are described in more detail below.

Digital Platforms

For this work, we have two digital platforms for viewing and annotating WSIs: PathPresenter^[50] and caMicroscope.^[51] Screenshots of the user interfaces are shown in Figure 2a and b. Pathologists can log in from anywhere in the world, and annotate images using web-based viewers.

Both PathPresenter and caMicroscope leadership are collaborators in this project and supported development of controlled and standardized workflows to select ROIs and to evaluate ROIs. Both platforms can read and write annotations using the ImageScope XML format,^[52] and we have used that format to share ROIs and create an identical study on both platforms. Both platforms also record the pixel width and height and the zoom setting of the WSI area being viewed. We

have not yet imposed display requirements in the pilot study but that will be discussed for future phases of our project.

Using more than one platform, including the microscope platform described next, allows us to involve more partners that can provide different perspectives, build redundancy to mitigate against a collaborator leaving the team, and promote interoperability as we progress to future phases of the project. The validation dataset will be based on the microscope platform, and the digital platforms allow fast development and understanding of our study and also allow us to compare microscope mode to digital mode evaluations.

Microscope Platform

The microscope platform we use is a hardware and software system called Evaluation Environment for Digital and Analog Pathology (eeDAP).^[53] The system uses a computer-controlled motorized stage and digital camera mounted to a microscope. eeDAP software registers the location of what is seen in the physical tissue through the microscope to the corresponding location in a WSI. Registration is accomplished through an interactive process that links the coordinates of the motorized stage to the coordinates of a WSI image. Registration enables the evaluation of the same ROIs in both the digital and microscope domains.

Similar to the digital platforms, the eeDAP software includes a utility to read and write ImageScope XML files, and a graphical user interface (GUI) implementing the ROI evaluation workflow [Figure 2c].^[54] A research assistant supports the pathologist by entering data into the eeDAP GUI and monitoring registration accuracy. The square ROI is realized with a reticle in the eyepiece. As annotations are collected on the slide, they are scanner agnostic and may be mapped to any scanned version of the slide using the eeDAP registration feature.

Region of Interest Selection: Study Preparation

A board-certified collaborating pathologist marked 10 ROIs on each of the 64 cases using the digital platforms described above. The ROIs were 500 $\mu\text{m} \times 500 \mu\text{m}$ squares. The instructions were to target diverse morphology from various locations within the slide. More specific instructions were to target areas with and without tumor-associated stroma, areas where sTIL densities should and should not be evaluated. More details on selecting specific ROI types can be found in Table 1. An algorithm is expected to perform well in all these areas, so it is vital that the dataset include them.

Region of Interest Evaluation

In current project protocols, we crowdsource pathologists to participate in ROI evaluation, separate from the pathologist who completed ROI selection. These pathologists will first label the ROI by one of the four labels given in Table 1. Pathologists then mark if the ROI is appropriate for evaluating sTIL density. This question is designed to determine if the area has tumor-associated stroma or not. If there is no tumor-associated stroma, annotation is complete. If there is

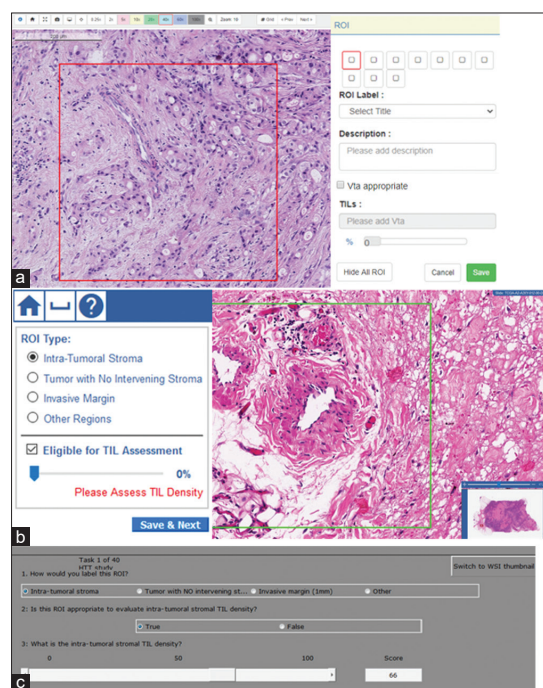


Figure 2: Screenshots from graphical user interfaces of three platforms used in data collection. All three collect a descriptive label of the regions of interest [Table 1], a binary evaluation of whether the regions of interest are appropriate for stromal tumor-infiltrating lymphocyte density estimation, and an estimate of stromal tumor-infiltrating lymphocyte density via slider bar or keyboard entry. (a) PathPresenter and (b) caMicroscope are digital platforms. (c) Evaluation environment for digital and analog pathology microscope platform. In data collection, the pathologist is at the microscope, while a study coordinator records evaluations through the graphical user interface

[Downloaded free from <http://www.jpathinformatics.org> on Wednesday, November 24, 2021, IP: 165.124.124.231]

J Pathol Inform 2021, 1:45

<http://www.jpathinformatics.org/content/12/1/45>

Table 1: Region of interest types

<p>Intra-tumoral stroma (aka tumor-associated stroma): Select ~3 ROIs</p> <ul style="list-style-type: none"> • Be sure to include regions with lymphocytes (TILs) • If there are lymphocytic aggregates, make sure to capture both lymphocyte-depleted and lymphocyte-rich areas within the same ROI if possible • Preferable to include some tumor in the same ROI — i.e. carcinoma cells as well as their associated stroma • If variable density within the slide, make sure to capture ROIs from different areas with different densities
<p>Invasive margin (Tumor-stroma transition): Select ~2 ROIs</p> <ul style="list-style-type: none"> • If heterogeneous tumor morphology, sample from different tumor-stroma transitions for each
<p>Tumor with no intervening stroma: Select ~2 ROIs, if possible</p> <ul style="list-style-type: none"> • If heterogeneous tumor morphology, sample from different morphologies • Be sure to sample from: vacuolated tumor cells, dying tumor cells, regions of different densities of tumor • Will be used to capture/assess intra-tumoral TILs and/or detect false positive TIL detections in purely cancerous regions.
<p>Other regions: Select ~3-4 ROIs</p> <ul style="list-style-type: none"> • ~1 from "empty"/distant/uneventful stroma • ~1 from hyalinized stroma, if any • ~2 other regions: <ul style="list-style-type: none"> ○ Necrosis transition (including comedo pattern) ○ Normal acini/ducts ○ Blood vessels ○ Others at pathologist discretion

tumor-associated stroma, the pathologist needs to estimate the density of TILs appearing in the tumor-associated stroma. The platforms allow integers 0–100, with no binning or thresholds. The motivation is to allow for thresholds to be determined later as the role of TILs becomes more clear and patient management guidelines are developed.

Pathologist Participants in Region of Interest Evaluation

Pathologist participants were recruited at a meeting of the Alliance for Digital Pathology immediately preceding the February 2020 USCAP [United States and Canadian Academy of Pathology] annual meeting.^[6] That meeting launched the in-person portion of pilot phase data collection. Board-certified anatomic pathologists and anatomic pathology residents were eligible to participate. To participate, they were asked to review the informed consent^[55] and the training materials: the guidelines on sTIL evaluation^[24] and a video tutorial and corresponding presentation about sTIL evaluation, the project, and using the platforms.^[56] Reviewing the sTIL evaluation training was required before participating and took about 30 min. Pathologists were asked to label the ROI according to the types given in Table 1, a true-false decision about whether sTIL densities should or should not be evaluated, and if true, an estimate of the sTIL density.

In total, 19 pathologists made 1645 ROI evaluations during the February event and the 2 weeks following. The primary platform at the event was the eDAP microscope system where 7 pathologists made 440 evaluations. Most of the evaluations made on the digital platforms were made by pathologists who could not attend in person. Data collection in digital mode took approximately 30–40 min per batch and twice that long in microscope mode. The increased time for microscope evaluation was due to the motorized stage movements.

Reference Standard (Truth) from Pathologists

The sTIL density measurements from pathologists are subject to bias and variance due to differences in pathologist expertise and training. In this work, we collected observations from multiple pathologists for each ROI, and then, we averaged over the pathologists. While the precision of these values can be estimated, averaging over pathologists ultimately ignores pathologist variability in the subsequent algorithm performance metric. As such, we also let the observations from each pathologist stand as noisy realizations of the truth. This approach is used in related research on inferring truth from the crowd for the purpose of training an algorithm.^[57] For our work, however, the purpose is to properly account for pathologist variability when estimating the uncertainty of algorithm performance.

Performance Metric for Stromal Tumor-Infiltrating Lymphocyte Density Values

The primary endpoint of an algorithm that produces quantitative values needs to measure how close the values from the algorithm (*Predicted_i*) are to the reference standard (*Truth_i*). To evaluate “closeness,” one appropriate performance metric that we are focusing on is the root mean squared error (RMSE):

$$RMSE = \sqrt{\frac{1}{N} \sum_i^N (Predicted_i - Truth_i)^2}, \quad (1)$$

where N is the number of ROIs. Smaller values of RMSE indicate that the predicted values are closer to the truth, and thus better algorithmic performance. Equation 1 shows the RMSE estimated from a finite population (e.g., a finite sample of ROIs). As we consider a statistical analysis for our work – estimating uncertainty, confidence intervals, and hypothesis tests – we look to the infinite population quantity without the square root.^[38,58-60]

$$MSE = E\left(\left[Predicted_i - Actual_i\right]^2\right) = bias^2 + variance, \quad (2)$$

Here, we see that mean squared error measures accuracy and precision, similar to Lin’s concordance correlation coefficient.^[61]

There are two main challenges to analyzing the differences between predictions and truth in our work. First, the sum in Equation 1 is really a sum over ROIs nested within cases. These values are not independent and identically distributed (iid), as is generally assumed for Equation 1. There should be a subscript for both case and ROI, and the statistical analysis needs to account for the correlation between values from ROIs within a case. In Figure 3, we show that sTIL densities are not iid across cases. The data are from one pathologist evaluating three cases that have different levels of sTIL infiltration. We see the sTIL densities are correlated within a case, and the variance is increasing with the mean. The distribution of sTIL densities is not the same for every case.

The second challenge in our work is to account for the variability from pathologist to pathologist. This variability

[Downloaded free from <http://www.jpathinformatics.org> on Wednesday, November 24, 2021, IP: 165.124.124.231]

J Pathol Inform 2021, 1:45

<http://www.jpathinformatics.org/content/12/1/45>

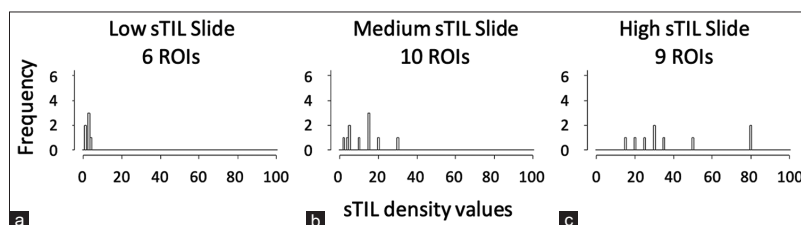


Figure 3: The distribution of stromal tumor-infiltrating lymphocyte densities in three slides with different levels of infiltration: (a) Low, (b) Medium, (c) High. The stromal tumor-infiltrating lymphocyte densities were from one pathologist. As not all region of interest labels are appropriate for stromal tumor-infiltrating lymphocyte density evaluation, not every case will contain tumor-infiltrating lymphocyte evaluations for all 10 regions of interests

is shown in Figure 4, which is a scatter plot showing the paired sTIL densities from two pathologists. Our strategy for addressing pathologist variability is to replace the single reference score in Equation 1 with pathologist-specific values.

To address these two challenges, we rewrite Equation 2 as

$$MSE = E \left([Y_{ki} - X_{jki}]^2 \right), \quad (3)$$

Where X_{jki} denotes the sTIL density from pathologist j evaluating the ROI i in case k and Y_{ki} denotes the sTIL density from the algorithm evaluating the ROI i in case k . Furthermore, the expected value averages over pathologists, cases, and ROIs. It is this quantity that we wish to estimate, and we are developing such methods to account for the correlation of ROIs within a case and pathologist variability. The estimate may take the form of a summation over readers, cases, and ROIs, or it may be the result of a model that needs to be solved by more sophisticated methods that do not permit an explicit closed-form expression. The methods build on previous work on so-called multi-reader multi-case methods^[62-65] and methods to evaluate intra- and inter-reader agreement.^[66]

DISCUSSION

The “high-throughput truthing” (HTT) moniker for this project reflects the data collection methods as well as the spirit of the effort. The project was inspired by perception studies that have been run at annual meetings of the Radiological Society of North America.^[67] Society meetings provide an opportunity to reach a high volume of pathologists away from the workload of their day job. A similar opportunity is available at organizations with many pathologists. We have explored both of these kinds of data collection opportunities via an event at the American Society of Clinical Pathology Annual Meeting 2018,^[68,69] and an event at the Memorial Sloan Kettering Cancer Center.^[70,71]

In addition to live events where we can use the eeDAP microscope system, our workflows on web-based platforms (PathPresenter and caMicroscope) can crowdsource pathologists from anywhere in the world. We have found these events to be low-cost, efficient opportunities to recruit pathologists and collect data. We plan to continue the project

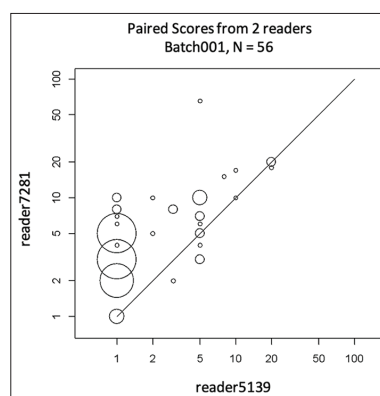


Figure 4: Scatter plot of stromal tumor-infiltrating lymphocyte densities from two pathologists on eight slides (one batch) that led to 56 paired observations. The plot is scaled by a log-base-10 transformation (with zero stromal tumor-infiltrating lymphocyte values changed to ones). The size of the circles is proportional to the number of observations at that point

by scaling our efforts to a pivotal phase and disseminating our final validation dataset.

Food and Drug Administration Medical Device Development Tool Program

A key aim of this project is to pursue the qualification of this dataset as a tool through the FDA MDDT program.^[19] Pursuing qualification offers an opportunity to receive feedback from an FDA review team about building the dataset to be fit for a regulatory purpose. As we disseminate our work, we believe that this feedback will be valuable for the project and more generally, for other public health stakeholders interested in the collection of validation datasets (industry, academia, health providers, patient advocates, professional societies, and government). A qualified tool has the potential to streamline the submission and review of validation data and allows the FDA to compare algorithms on the same prequalified data. In this way, the project may benefit the agency and medical device manufacturers, as well as the larger scientific community.

[Downloaded free from <http://www.jpathinformatics.org> on Wednesday, November 24, 2021, IP: 165.124.124.231]

J Pathol Inform 2021, 1:45

<http://www.jpathinformatics.org/content/12/1/45>

The MDDT program was created by the FDA as a mechanism by which any public health stakeholder may develop and submit a tool to the agency for formal review. Tools are not medical devices. Rather, tools facilitate and increase predictability in medical device development and evaluation. Each tool is qualified for a specific context of use and may be used in a manufacturer's submission without needing to reconfirm its suitability and utility.^[19] Qualified tools are expected to be made publicly available, which can include a licensing arrangement. In this way, qualified tools reduce burden to both the agency and the manufacturer and ultimately increase product quality and better patient outcomes. The proposed context of use for this work is given in Table 2.

The exact platform and mechanisms for sharing the dataset have yet to be determined. However, the dataset will be shared broadly at no cost with any entity, subject to applicable terms required by either the FDA or the MDDT program. Possible terms would protect against data being used to "train to the test" using strategies such as data access via containers or data governance by written agreements. We can look to public challenges^[72-75] to inform our data sharing plans and educational dissemination opportunities.

An MDDT dataset has the potential to significantly reduce the burden of manufacturers, especially small companies. Validation in the commercial space tends to be siloed, with each developer using distinct, licensed, and proprietary data. Our proposed MDDT may allow manufacturers and the FDA to avoid the time- and resource-consuming back-and-forth discussions to formulate a study design and protocol. Manufacturers may also be able to skip burdensome steps such as obtaining Investigational Review Board approvals, slide sourcing, reader recruitment, and collecting the data. Instead of planning statistical analyses from scratch, manufacturers may use the analyses developed from this project as an example to guide their work. These bypassed steps are represented in the column headings of Figure 1.

Data Representativeness/Generalizability

A random set of breast cancer biopsies are naturally expected to include the different immunophenotypic subtypes of TILs (CD4+, CD8+ T-cells, and natural killer cells) and a variety of shapes, locations, colors, and clustering of TILs.^[76-79] Our current strategy of selecting ROIs gathers areas for sTIL evaluation with and without tumor-associated stroma, areas where sTIL densities should and should not be

evaluated [Table 1]. Despite efforts to assemble a balanced and stratified sample of ROI types, our pilot study data yielded an abundant number of cases with nominal sTIL infiltration. While this may be the true clinical distribution, for our MDDT, we want to balance and stratify the sTIL density values across the expected range. For this, we intend to realize some data curation before ROI evaluation in our future pivotal study.

The MDDT dataset should also adequately represent the variability arising from preanalytic differences (slide preparation) and the intended population (clinical subgroups). As such, for our pivotal study, we intend to source slides from at least three sites and stratify the cases across important clinical subgroups. If possible, we will also create some cases that systematically explore the H & E staining protocol (incubation time, washing time, and stain strength).

There are several clinical subgroups that are appropriate to sample, such as patient age, breast cancer subtypes and stages,^[28,80-82] and treatment at various time intervals. Sampling from all possible subgroups is challenging if not impossible. While our inclusion and exclusion criteria limit the use of our MDDT to a selective population, we do not expect to sample all the subgroups that might be required in an algorithm submission, and we do not expect to have the same metadata for all cases. It is important to note that while TILs are known to have the most prognostic value in certain molecular (genomic) subtypes (e.g., TNBC and HER2+), a TIL algorithm is most likely to be confounded by histologic subtype and characteristics. While there is some correspondence between genomic and histologic classifications of breast tumors, the histological presentation (morphology) of, say, a ductal carcinoma does not necessarily correlate well with its genomic composition. Any data that is not part of the MDDT but is required for a regulatory submission of an algorithm will ultimately be the responsibility of the algorithm manufacturer. We do not intend to sample treatment methods or longitudinal data.

Pathologists and Pathologist Variability

In this work, our initial data shows notable variability in independent sTIL density estimates from multiple pathologists on each ROI [Figure 4], which is consistent with previous work in this area.^[32] These findings further reinforce the need to collect data from multiple pathologists and the need to better understand this variability. We intend to explore the difference between

Table 2: Proposed context of use for a stromal tumor-infiltrating lymphocyte density annotated dataset

The sTIL-density Annotated Dataset is a tool to be used to assess the accuracy of algorithms that quantify the density of stromal tumor infiltrating lymphocytes (sTILs). It is comprised of a dataset of slides, digital whole slide images and annotations in regions of interest (ROIs) compiled by pathologists using microscopes to evaluate glass slides of tissue samples from breast cancer needle core biopsies, where the tissue sections are stained with Hematoxylin and eosin (H&E).

[Downloaded free from <http://www.jpathinformatics.org> on Wednesday, November 24, 2021, IP: 165.124.124.231]

J Pathol Inform 2021, 1:45

<http://www.jpathinformatics.org/content/12/1/45>

averaging over pathologists and keeping them distinct when evaluating algorithm performance. In either case, we believe that a statistical analysis method should account for reader variability in addition to case variability. A final statistical analysis plan for our pivotal study, including sizing the number of pathologists and cases, will be developed based on the pilot data, simulation studies, and feedback from the FDA's MDDT review team.

As we are crowdsourcing pathologists, we have received questions regarding the expertise of the participating pathologists. Initially, we accepted any board-certified pathologist or anatomic pathology resident, but the reader variability observed in the pilot data has caused us to reconsider. As such, this is a limitation in the reliability of the pilot study data. Improving the expertise of annotating pathologists will reduce pathologist variability and allow us to reduce the number of pathologists. Therefore, for our pivotal study, we are expanding our current training materials to include testing with immediate feedback, providing the reference standard for each ROI. We are also creating a proficiency test. These training materials may be built from the pilot study dataset. A robust training program could additionally serve the community beyond our specific project need.

As relates to the RMSE performance metric, which summarizes the bias as well as the variance of an algorithm, it is not clear whether the bias comes from the algorithm or the pathologist. Amgad *et al.*^[83] found their algorithm to be biased low compared to the pathologists. They also found that the Spearman rank-based correlation was stronger for the algorithm-to-pathologist-consensus comparison compared to the pathologist-to-pathologist comparison ($R=0.73$ vs. $R=0.66$). The authors believe these results are related to pathologist bias and variability, and not the algorithm. While this may be true, it is difficult to know as only two pathologists provided sTIL density values. Furthermore, the comparison does not account for pathologist variability in either correlation result and is not an apple-to-apple comparison due to the consensus process. Still, we expect that our expanded pathologist training will improve pathologist correlation, and we will compare the correlation of our pivotal study data to that of Amgad *et al.* and to our pilot study results. In preparation for this comparison, we will explore Spearman's rank correlation and Kendall's tau on the pilot study data. These metrics treat pathologist sTIL density estimates as ordinal data rather than quantitative and calibrated data.^[66,84-86]

Relaunch and Future Pivotal Study

While the live event portion of pilot phase data collection was a burdensome process, we totaled 1645 evaluations in 10 h. The live event was set up with four evaluation stations: 2 digital platforms and 2 microscope platforms. We created training materials and hosted an online training seminar before the event. We assembled recruitment materials and sent invitations to pathologists. We trained study administrators to operate eeDAP and assist pathologists with data collection at the microscope. All equipment was shipped and assembled on site. Data aggregation was completed via APIs. Not surprisingly, the data are stored

quite differently on the two digital platforms, so we created scripts to clean and harmonize the raw data into common data frames. We began building a software package to analyze the clean data. In sum, the process took a lot of time and effort, but offered experiences to inform the next phase of our project.

To help pathologists improve their sTIL density estimates and collect more detailed data, we thought about what an algorithm generally would do: identify and segment the tumor, tumor-associated stroma, and sTILs. We thought that it would be worthwhile to parallel these steps. We were already asking pathologists to label ROIs by tumor, margin, and the presence of tumor-associated stroma. We decided that in our pivotal study, we would ask the pathologist to estimate the percent of the ROI area that contains tumor-associated stroma.

Data collection on the microscope system was put on hold because of the COVID-19 pandemic, but we relaunched data collection on the digital platforms in September 2020 to fill out observations across all batches of the pilot study. We invite board-certified pathologists to spend approximately 30 min on training and 30 min per batch on data collection.^[87] With newly established agreements for sharing materials, we are in the process of securing more slides and images to sample the patient subgroups mentioned from multiple sites in our future pivotal study, bolstering our single-site pilot data. We welcome parties that are able and willing to share such materials to contact us through the corresponding author. Similarly, we are looking for opportunities to set up HTT events or find collaborating sites interested in hosting data collection events on their own. There are opportunities to set up their own eeDAP microscope system or borrow an existing system from us. We are willing to supervise and assist remotely.

CONCLUSION

On the volunteer efforts of many and a nominal budget, we have created a team and a protocol, administrative materials, and infrastructure for our HTT project. We have sourced breast-cancer slides and crowdsourced pathologists in a pilot study, and we are actively planning a pivotal study with more data and better pathologist training. Our goal is to create a sTIL-density annotated dataset that is fit for a regulatory purpose. We hope that this project can be a roadmap and inspiration for other stakeholders (industry, academia, health providers, patient advocates, professional societies, and government) to work together in the precompetitive space to create similar high-value, fit-for-purpose, broadly accessible datasets to support the field in bringing algorithms to market and to monitor algorithms on the market.

ACKNOWLEDGMENTS

The HTT team acknowledges the work of collaborating pathology networks such as the International Immuno Oncology Biomarker Working Group and the Alliance for Digital Pathology in their participant pathologist recruitment (www.tilsinbreastcancer.org and <https://digitalpathologyalliance.org>).

[Downloaded free from <http://www.jpathinformatics.org> on Wednesday, November 24, 2021, IP: 165.124.124.231]

J Pathol Inform 2021, 1:45

<http://www.jpathinformatics.org/content/12/1/45>

org). Ioanna Laios and Ligia Craciun contributed technical support and expertise in the preparation of glass slides. Finally, the team thanks their developers: Krushnavadan Acharya (PathPresenter), Nan Li (caMicroscope), and Qi Gong (eeDAP).

FINANCIAL SUPPORT AND SPONSORSHIP

R. S. is supported by a grant from the Breast Cancer Research Foundation (grant No. 17 194), and J. S. is supported by grants from the NIH (UH3CA225021 and U24CA180924). A. M. is supported via grants from the NCI (1U24CA199374-01, R01CA249992-01A1, R01CA202752-01A1, R01CA208236-01A1, R01CA216579-01A1, R01CA220581-01A1, R01CA257612-01A1, 1U01CA239055-01, 1U01CA248226-01, 1U54CA254566-01), National Heart, Lung and Blood Institute (1R01HL15127701A1, R01HL15807101A1), National Institute of Biomedical Imaging and Bioengineering (1R43EB028736-01), National Center for Research Resources (1 C06 RR12463-01), VA Merit Review Award IBX004121A from the United States Department of Veterans Affairs Biomedical Laboratory Research and Development Service, the Office of the Assistant Secretary of Defense for Health Affairs, through the Breast Cancer Research Program (W81XWH-19-1-0668), the Prostate Cancer Research Program (W81XWH-15-1-0558, W81XWH-20-1-0851), the Lung Cancer Research Program (W81XWH-18-1-0440, W81XWH-20-1-0595), the Peer Reviewed Cancer Research Program (W81XWH-18-1-0404), the Kidney Precision Medicine Project (KPMP) Glue Grant, the Ohio Third Frontier Technology Validation Fund, the Clinical and Translational Science Collaborative of Cleveland (UL1TR0002548) from the National Center for Advancing Translational Sciences (NCATS) component of the National Institutes of Health and NIH roadmap for Medical Research, The Wallace H. Coulter Foundation Program in the Department of Biomedical Engineering at Case Western Reserve University. Sponsored research agreements from Bristol Myers-Squibb, Boehringer-Ingelheim, and AstraZeneca.

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.

CONFLICTS OF INTEREST

A. M. 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 and currently consults for Caris, Roche and Aiforia. He also has sponsored research agreements with Philips, AstraZeneca, Boehringer-Ingelheim 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 3 different R01 grants with Inspirata Inc.

REFERENCES

- Russell SJ, Norvig P. *Artificial Intelligence: A Modern Approach*. 3rd ed. Upper Saddle River, New Jersey: Prentice Hall; 2009.
- Fuchs TJ, Buhmann JM. Computational pathology: Challenges and promises for tissue analysis. *Comput Med Imaging Graph* 2011;35:515-30.
- Chang HY, Jung CK, Woo JI, Lee S, Cho J, Kim SW, *et al*. Artificial Intelligence in Pathology. *J Pathol Transl Med* 2019;53:1-2.
- Serag A, Ion-Margineanu A, Qureshi H, McMillan R, Saint Martin MJ, Diamond J, *et al*. Translational AI and deep learning in diagnostic pathology. *Front Med (Lausanne)* 2019;6:185.
- Bera K, Schalper KA, Rimm DL, Velcheti V, Madabhushi A. Artificial intelligence in digital pathology – New tools for diagnosis and precision oncology. *Nat Rev Clin Oncol* 2019;16:703-15.
- Marble HD, Huang R, Dudgeon SN, Lowe A, Herrmann MD, Blakely S, *et al*. A regulatory science initiative to harmonize and standardize digital pathology and machine learning processes to speed up clinical innovation to patients. *J Pathol Inform* 2020;11:22.
- Niazi MK, Parwani AV, Gurcan MN. Digital pathology and artificial intelligence. *Lancet Oncol* 2019;20:e253-61.
- FDA CDRH. “*De Novo* Request Evaluation of Automatic Class III Designation for Philips IntelliSite Pathology Solution (PIPS): Decision Summary;” 2017. Available from: https://www.accessdata.fda.gov/cdrh_docs/reviews/DEN160056.pdf. [Last accessed on 2018 May 17].
- FDA CDRH. “510(k) Summary Aperio AT2 DX System;” 2019. Available from: https://www.accessdata.fda.gov/cdrh_docs/pdf19/K190332.pdf. [Last accessed on 2020 Apr 21].
- FDA CDRH. “Summary of Safety and Effectiveness Data: GE FFDm (p990066);” 2001. Available from: http://www.accessdata.fda.gov/cdrh_docs/pdf/P990066b.pdf. [Last accessed on 2021 Sept 02].
- FDA CDRH. “Summary of Safety and Effectiveness Data: R2 Technology, Inc. ImageChecker M1000 (p990066);” 1998. Available from: https://www.accessdata.fda.gov/cdrh_docs/pdf/p970058.pdf. [Last accessed on 2020 Aug 06].
- FDA CDRH. “Guidance for Industry and FDA Staff – Computer-Assisted Detection Devices Applied to Radiology Images and Radiology Device Data – Premarket Notification [510(k)] Submissions.” FDA; 2012. Available from: <https://www.fda.gov/media/77635/download>. [Last accessed on 2020 Apr 21].
- FDA CDRH. “Guidance for Industry and FDA Staff – Clinical Performance Assessment: Considerations for Computer-Assisted Detection Devices Applied to Radiology Images and Radiology Device Data in Premarket Notification [510(k)] Submissions.” FDA; 2020. Available from: <https://www.fda.gov/media/77642/download>. [Last accessed on 2020 Aug 05].
- Gallas BD, Chan HP, D’Orsi CJ, Dodd LE, Giger ML, Gur D, *et al*. Evaluating imaging and computer-aided detection and diagnosis devices at the FDA. *Acad Radiol* 2012;19:463-77.
- FDA/CDRH. “Software as a Medical Device (SAMd): Clinical Evaluation.” FDA, Silver Spring, MD; 2017. Available from: <https://www.fda.gov/media/100714/download>. [Last accessed on 2018 Jan 05].
- FDA CDRH. “Evaluation of Automatic Class III Designation for QuantX: Decision Summary;” 2017. Available from: <https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfpmn/denovo.cfm?id=DEN170022>. [Last accessed on 2020 Sep 16].
- FDA CDRH. “Evaluation of Automatic Class III Designation for ContaCT: Decision Summary;” 2017. Available from: <https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfpmn/denovo.cfm?id=DEN170073>. [Last accessed on 2020 Sep 16].
- FDA CDRH. “Evaluation of Automatic Class III Designation for IDx-DR: Decision Summary;” 2017. Available from: <https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfpmn/denovo.cfm?id=DEN180001>. [Last accessed on 2020 Sep 16].
- FDA/CDRH. “Qualification of Medical Device Development Tools.” FDA; August 10, 2017. Available from: <https://www.fda.gov/media/87134/download>. [Last accessed on 2020 Aug 03].
- Simon RM, Paik S, Hayes DF. Use of archived specimens in evaluation of prognostic and predictive biomarkers. *J Natl Cancer Inst* 2009;101:1446-52.

[Downloaded free from <http://www.jpathinformatics.org> on Wednesday, November 24, 2021, IP: 165.124.124.231]

J Pathol Inform 2021, 1:45

<http://www.jpathinformatics.org/content/12/1/45>

21. Loi S, Drubay D, Adams S, Pruneri G, Francis PA, Lacroix-Triki M, *et al*. Tumor-infiltrating lymphocytes and prognosis: A pooled individual patient analysis of early-stage triple-negative breast cancers. *J Clin Oncol* 2019;37:559-69.
22. Denkert C, von Minckwitz G, Darb-Esfahani S, Lederer B, Heppner BI, Weber KE, *et al*. Tumour-infiltrating lymphocytes and prognosis in different subtypes of breast cancer: A pooled analysis of 3771 patients treated with neoadjuvant therapy. *Lancet Oncol* 2018;19:40-50.
23. McShane LM, Altman DG, Sauerbrei W, Taube SE, Gion M, Clark GM, *et al*. Reporting recommendations for tumor marker prognostic studies. *J Clin Oncol* 2005;23:9067-72.
24. Salgado R, Denkert C, Demaria S, Sirtaine N, Klauschen F, Pruneri G, *et al*. The evaluation of tumor-infiltrating lymphocytes (TILs) in breast cancer: Recommendations by an International TILs Working Group 2014. *Ann Oncol* 2015;26:259-71.
25. Cardoso F, Kyriakides S, Ohno S, Penault-Llorca F, Poortmans P, Rubio IT, *et al*. Early breast cancer: ESMO clinical practice guidelines for diagnosis, treatment and follow-up†. *Ann Oncol* 2019;30:1194-220.
26. Morigi C. Highlights of the 16th St Gallen International Breast Cancer Conference, Vienna, Austria, 20-23 March 2019: Personalised treatments for patients with early breast cancer. *Ecancermedscience* 2019;13:924.
27. Balic M, Thomssen C, Würstlein R, Gnant M, Harbeck N. St. Gallen/Vienna 2019: A brief summary of the consensus discussion on the optimal primary breast cancer treatment. *Breast Care (Basel)* 2019;14:103-10.
28. International Agency for Research on Cancer. Breast Tumours. In: WHO Classification of Tumours Series. 5th ed., Vol. 2. Lyon (France): WHO Classification of Tumours Editorial Board; 2019: Available from: <https://tumourclassification.iarc.who.int/chapters/32>. [Last accessed on 2020 Jun 12].
29. Luen SJ, Salgado R, Dieci MV, Vingiani A, Curigliano G, Gould RE, *et al*. Prognostic implications of residual disease tumor-infiltrating lymphocytes and residual cancer burden in triple-negative breast cancer patients after neoadjuvant chemotherapy. *Ann Oncol* 2019;30:236-42.
30. Luen SJ, Griguolo G, Nuciforo P, Campbell C, Fasani R, Cortes J. On-treatment changes in tumor-infiltrating lymphocytes (TIL) during neoadjuvant HER2 therapy (NAT) and clinical outcome. *J Clin Oncol* 2019;37 Suppl 15:574.
31. Hendry S, Salgado R, Gevaert T, Russell PA, John T, Thapa B, *et al*. Assessing tumor-infiltrating lymphocytes in solid tumors: A practical review for pathologists and proposal for a standardized method from the international immunooncology biomarkers working group: Part 1: Assessing the host immune response, TILs in invasive breast carcinoma and ductal carcinoma *in situ*, metastatic tumor deposits and areas for further research. *Adv Anat Pathol* 2017;24:235-51.
32. Amgad M, Stovgaard ES, Balslev E, Thagaard J, Chen W, Dudgeon S, *et al*. Report on computational assessment of tumor infiltrating lymphocytes from the international immuno-oncology biomarker working group. *NPJ Breast Cancer* 2020;6:16.
33. Klauschen F, Müller KR, Binder A, Bockmayr M, Hägele M, Seegerer P, *et al*. Scoring of tumor-infiltrating lymphocytes: From visual estimation to machine learning. *Semin Cancer Biol* 2018;28:151-7.
34. Corredor G, Wang X, Zhou Y, Lu C, Fu P, Syrigos K, *et al*. Spatial architecture and arrangement of tumor-infiltrating lymphocytes for predicting likelihood of recurrence in early-stage non-small cell lung cancer. *Clin Cancer Res* 2019;25:1526-34.
35. Saltz J, Gupta R, Hou L, Kurc T, Singh P, Nguyen V, *et al*. Spatial organization and molecular correlation of tumor-infiltrating lymphocytes using deep learning on pathology images. *Cell Rep* 2018;23:181-93.e7.
36. AbdulJabbar K, Raza SE, Rosenthal R, Jamal-Hanjani M, Veeriah S, Akarca A, *et al*. Geospatial immune variability illuminates differential evolution of lung adenocarcinoma. *Nat Med* 2020;26:1054-62.
37. F. C. FDA/CDRH. "Technical Performance Assessment of Quantitative Imaging in Device Premarket Submissions: Draft Guidance for Industry and Food and Drug Administration Staff." FDA; 2019. Available from: <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/technical-performance-assessment-quantitative-imaging-device-premarket-submissions>. [Last accessed on 2020 Apr 28].
38. Altman DG, Royston P. What do we mean by validating a prognostic model? *Stat Med* 2000;19:453-73.
39. Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW, *et al*. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): Explanation and elaboration. *Ann Intern Med* 2015;162:W1-73.
40. Campanella G, Hanna MG, Geneslaw L, Mirafior A, Werneck Krauss Silva V, Busam KJ, *et al*. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nat Med* 2019;25:1301-9.
41. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): The TRIPOD statement. *Ann Intern Med* 2015;162:55-63.
42. De A, Meier K, Tang R, Li M, Gwise T, Gomatam S, *et al*. Evaluation of heart failure biomarker tests: A survey of statistical considerations. *J Cardiovasc Transl Res* 2013;6:449-57.
43. Pennello GA. Analytical and clinical evaluation of biomarkers assays: When are biomarkers ready for prime time? *Clin Trials* 2013;10:666-76.
44. Williams BJ, Treanor D. Practical guide to training and validation for primary diagnosis with digital pathology. *J Clin Pathol* 2020;73:418-22.
45. Doyle S, Monaco J, Feldman M, Tomaszewski J, Madabhushi A. An active learning based classification strategy for the minority class problem: Application to histopathology annotation. *BMC Bioinformatics* 2011;12:424.
46. FDA. "Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) – Discussion Paper and Request for Feedback." US Food and Drug Administration; April 02, 2019. Available from: <https://www.fda.gov/media/122535/download>. [Last accessed on 2020 Mar 15].
47. Feng J, Emerson S, Simon N. Approval policies for modifications to machine learning-based software as a medical device: A study of bio-creep. *Biometrics* 2020;77:31-44.
48. Pennello G, Sahiner B, Gossmann A, Petrick N. Discussion on 'Approval policies for modifications to machine learning-based software as a medical device: A study of bio-creep' by Jean Feng, Scott Emerson, and Noah Simon. *Biometrics* 2021;77:45-48.
49. McShane LM, Cavenagh MM, Lively TG, Eberhard DA, Bigbee WL, Williams PM, *et al*. Criteria for the use of omics-based predictors in clinical trials. *Nature* 2013;502:317-20.
50. "PathPresenter". Available from: <https://pathpresenter.net>. [Last accessed on 2020 Sep 16].
51. Saltz J, Sharma A, Iyer G, Bremer E, Wang F, Jasniewski A, *et al*. A containerized software system for generation, management, and exploration of features from whole slide tissue images. *Cancer Res* 2017;77:e79-82.
52. "Aperio ImageScope." Available from: <https://www.leicabiosystems.com/digital-pathology/manage/aperio-imagescope>. [Last accessed on 2020 Sep 16].
53. Gallas BD, Gavrielides MA, Conway CM, Ivansky A, Keay TC, Cheng WC, *et al*. Evaluation Environment for Digital and Analog Pathology (eeDAP): A platform for validation studies. *J Med Imaging* 2014;1: 037501.
54. Gallas BD. eeDAP: Evaluation environment for digital and analog histopathology. Available from: <https://github.com?DIDSR/eeDAP/releases>. [Last accessed on 2020 Aug 13]
55. "High-Throughput Truthing Project Informed Consent Form." Available from: <https://ncihub.org/groups/eedapstudies/wiki/HTTInformedConsent>. [Last accessed on 2020 Sep 17].
56. "High-Throughput Truthing Training Materials." Available from: <https://ncihub.org/groups/eedapstudies/wiki/HTTdataCollectionTraining>. [Last accessed on 2020 Sep 17].
57. Zheng Y, Li G, Li Y, Shan C, Cheng R. Truth inference in crowdsourcing: Is the problem solved? *Proc VLDB Endow* 2017;10:541-52.
58. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986;1:307-10.
59. Bland JM, Altman DG. Measuring agreement in method comparison studies. *Stat Methods Med Res* 1999;8:135-60.
60. Casella G, Berger RL. *Statistical Inference*. 2nd ed. Pacific Grove, CA: Duxbury; 2002.
61. Lin LL. A concordance correlation coefficient to evaluate reproducibility. *Biometrics* 1989;45:255-68.

[Downloaded free from <http://www.jpathinformatics.org> on Wednesday, November 24, 2021, IP: 165.124.124.231]

J Pathol Inform 2021, 1:45

<http://www.jpathinformatics.org/content/12/1/45>

62. Gallas BD, Bandos A, Samuelson F, Wagner RF. A framework for random-effects ROC analysis: Biases with the bootstrap and other variance estimators. *Commun Stat Theory* 2009;38:2586-603.
63. Gallas BD, Chen W, Cole E, Ochs R, Petrick N, Pisano ED, *et al*. Impact of prevalence and case distribution in lab-based diagnostic imaging studies. *J Med Imaging (Bellingham)* 2019;6: 015501.
64. Chen W, Gong Q, Gallas BD. Paired split-plot designs of multireader multicase studies. *J Med Imaging (Bellingham)* 2018;5: 031410.
65. Tabata K, Uraoka N, Benhamida J, Hanna MG, Sirintrapun SJ, Gallas BD, *et al*. Validation of mitotic cell quantification via microscopy and multiple whole-slide scanners. *Diagn Pathol* 2019;14:65.
66. Dawid AP, Skene AM. Maximum likelihood estimation of observer error-rates using the EM algorithm. *J R Stat Soc Ser C Appl Stat* 1979;28:20-8.
67. Toomey RJ, McEntee MF, Rainford LA. The pop-up research centre – Challenges and opportunities. *Radiography (Lond)* 2019;25 Suppl 1:S19-24.
68. Gallas BD, Amgad M, Chen W, Cooper LAD, Dudgeon S, Gilmore H. A collaborative project to produce regulatory-grade pathologist annotations to validate viewers and algorithms. In: Abstracts. *J Pathol Inform* 2019;10:28.
69. Gallas BD, Amgad M, Chen W, Cooper LAD, Dudgeon S, Gilmore H. A collaborative project to produce regulatory-grade pathologist annotations to validate viewers and algorithms. In: Abstracts, Supplementary Materials. Available from: <https://ncihub.org/groups/eedapstudies/wiki/HighthroughputTruthingYear2>. [Last accessed 2020 Aug 04].
70. Gallas BD. A reader study on a 14-head microscope In: Pathology Informatics Summit 2018. *J Pathol Inform* 2018;9:50.
71. Gallas BD. A Reader Study on a 14-Head Microscope, In: Pathology Informatics Summit 2018, Supplementary Materials; January 01, 2018. Available from: <https://ncihub.org/groups/eedapstudies/wiki/Presentation:ARReaderStudyona14headMicroscope>. [Last accessed on 2020 Aug 04].
72. van Ginneken B, Kerkstra S, Meakin J. “Grand Challenge.” Available from: <https://grand-challenge.org/>. [Last accessed on 2020 Aug 11].
73. Cha K. “Overview – Grand Challenge.” Available from: <https://breastpathq.grand-challenge.org/>. [Last accessed on 2020 Aug 11].
74. Ehteshami Bejnordi B, Veta M, Johannes van Diest P, van Ginneken B, Karssemeijer N, Litjens G, *et al*. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA* 2017;318:2199-210.
75. Litjens G, Bandi P, Ehteshami Bejnordi B, Geessink O, Balkenhol M, Bult P, *et al*. 1399 H&E-stained sentinel lymph node sections of breast cancer patients: The CAMELYON dataset. *Gigascience* 2018;7:giy065.
76. Zgura A, Galesa L, Bratila E, Anghel R. Relationship between tumor infiltrating lymphocytes and progression in breast cancer. *Maedica (Bucur)* 2018;13:317-20.
77. Romagnoli G, Wiedermann M, Hübner F, Wengers A, Mathiak M, Röcken C, *et al*. Morphological evaluation of Tumor-Infiltrating Lymphocytes (TILs) to investigate invasive breast cancer immunogenicity, reveal lymphocytic networks and help relapse prediction: A retrospective study. *Int J Mol Sci* 2017;18:1936.
78. Egeblad M, Ewald AJ, Askautrud HA, Truitt ML, Welm BE, Bainbridge E, *et al*. Visualizing stromal cell dynamics in different tumor microenvironments by spinning disk confocal microscopy. *Dis Model Mech* 2008;1:155-67.
79. Çelebi F, Agacayak F, Ozturk A, Ilgun S, Ucuncu M, Iyigun ZE, *et al*. Usefulness of imaging findings in predicting tumor-infiltrating lymphocytes in patients with breast cancer. *Eur Radiol* 2020;30:2049-57.
80. Tan PH, Ellis I, Allison K, Brogi E, Fox SB, Lakhani S, *et al*. The 2019 World Health Organization classification of tumours of the breast. *Histopathology* 2020;77:181-5.
81. Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature* 2012;490:61-70.
82. American Joint Committee on Cancer. Breast. In: *AJCC Cancer Staging Manual*. 8th ed. New York, NY: Springer; 2017.
83. Amgad M, Elfandy H, Hussein H, Atteya LA, Elsebaie MA, Abo Elnasr LS, *et al*. Structured crowdsourcing enables convolutional segmentation of histology images. *Bioinformatics* 2019;35:3461-7.
84. Stevens SS. On the theory of scales of measurement. *Science* 1946;103:677.
85. Kim JO. Predictive measures of ordinal association. *Am J Sociol* 1971;76:891-907.
86. Kendall MG. A new measure of rank correlation. *Biometrika* 1938;30:81-93.
87. “High-Throughput Truthing Project Data Collection Information.” Available from: <https://ncihub.org/groups/eedapstudies/wiki/HTTStartDataCollection>. [Last accessed on 2020 Sep 17].

Chapter 3

Deep-learning methods for automatic detection of histopathology structures

This chapter includes several convolutional neural network (CNN) modeling approaches for delineating tissue regions and nuclei in digitized scans of formalin-fixed paraffin-embedded H&E stained slides from invasive carcinomas of the breast. These CNN architectures were designed to be suitable for and adapted to histopathology applications. We also present an approach for improving the transparency of CNN models by providing explanations that are intuitive and sensible.

The first section is a consensus statement describing a set of recommendations that we published with the International Immuno-Oncology Working Group. The recommendations discuss key considerations for computational assessment of Tumor-Infiltrating Lymphocytes (TILs) in solid tumors in a manner that is consistent with clinical scoring guidelines:

- **Amgad, M., Stovgaard, E. S., Balslev, E., Thagaard, J., Chen, W., Dudgeon, S., Sharma, A., Kerner, J. K., Denkert, C., Yuan, Y., et al. (2020).** *Report on computational assessment of tumor-infiltrating lymphocytes from the International Immuno-oncology Biomarker Working Group. NPJ breast cancer, 6(1):1–13.*

Next, we present a set of original CNN modeling techniques for automatically detecting histopathologic regions and cells in breast carcinomas, including cancer regions and cells, stromal regions and cells, necrosis, TILs, and other structures. The following sections are presented:

- **Amgad, M., Atteya, L., Hussein, H., Mohammed, K. H., Hafiz, E., Elsebaie, M. A., Mobadersany, P., Manthey, D., Gutman, D. A., Elfandy, H., et al. (2021).** *Explainable nucleus classification using decision tree approximation of learned embeddings. Bioinformatics.*

- **Amgad, M.,** Sarkar, A., Srinivas, C., Redman, R., Ratra, S., Bechert, C. J., Calhoun, B. C., Mrazek, K., Kurkure, U., Cooper, L. A., et al. (2019). Joint region and nucleus segmentation for characterization of tumor-infiltrating lymphocytes in breast cancer. In *Medical Imaging 2019: Digital Pathology*, volume 10956, page 109560M. International Society for Optics and Photonics.
- *MuTILs: explainable, multiresolution computational scoring of Tumor-Infiltrating Lymphocytes in breast carcinomas using clinical guidelines*

Section 3.1

Report on computational assessment of tumor infiltrating lymphocytes from the International Immuno-Oncology Biomarker Working Group

This section is an exact reproduction of the following open-access journal paper:

Amgad, M., Stovgaard, E. S., Balslev, E., Thagaard, J., Chen, W., Dudgeon, S., Sharma, A., Kerner, J. K., Denkert, C., Yuan, Y., et al. (2020). Report on computational assessment of tumor infiltrating lymphocytes from the International Immuno-oncology Biomarker Working Group. NPJ breast cancer, 6(1):1–13.

REVIEW ARTICLE OPEN



Report on computational assessment of Tumor Infiltrating Lymphocytes from the International Immuno-Oncology Biomarker Working Group

Mohamed Amgad et al.[#]

Assessment of tumor-infiltrating lymphocytes (TILs) is increasingly recognized as an integral part of the prognostic workflow in triple-negative (TNBC) and HER2-positive breast cancer, as well as many other solid tumors. This recognition has come about thanks to standardized visual reporting guidelines, which helped to reduce inter-reader variability. Now, there are ripe opportunities to employ computational methods that extract spatio-morphologic predictive features, enabling computer-aided diagnostics. We detail the benefits of computational TILs assessment, the readiness of TILs scoring for computational assessment, and outline considerations for overcoming key barriers to clinical translation in this arena. Specifically, we discuss: 1. ensuring computational workflows closely capture visual guidelines and standards; 2. challenges and thoughts standards for assessment of algorithms including training, preanalytical, analytical, and clinical validation; 3. perspectives on how to realize the potential of machine learning models and to overcome the perceptual and practical limits of visual scoring.

npj Breast Cancer (2020)6:16; <https://doi.org/10.1038/s41523-020-0154-2>

INTRODUCTION

Very large adjuvant trials have illustrated how the current schemes fail to stratify patients with sufficient granularity to permit optimal selection for clinical trials, likely owing to application of an overly limited set of clinico-pathologic features^{1,2}. Histologic evaluation of tumor-infiltrating lymphocytes (TILs) is emerging as a promising biomarker in solid tumors and has reached level IB-evidence as a prognostic marker in triple-negative (TNBC) and HER2-positive breast cancer^{3–5}. Recently, the St Gallen Breast Cancer Expert Committee endorsed routine assessment of TILs for TNBC patients⁶. In the absence of adequate standardization and training, visual TILs assessment (VTA) is subject to a marked degree of ambiguity and interobserver variability^{7–9}. A series of published guidelines from this working group (also known as TIL Working group or TIL-WG) aimed to standardize VTA in solid tumors, to improve reproducibility and clinical adoption^{10–12}. TIL-WG is an international coalition of pathologists, oncologists, statisticians, and data scientists that standardize the assessment of Immuno-Oncology Biomarkers to aid pathologists, clinicians, and researchers in their research and daily practice. The value of these guidelines was highlighted in two studies systematically examining VTA reproducibility^{7,13}. Nevertheless, VTA continues to have inherent limitations that cannot be fully addressed through standardization and training, including: 1. visual assessment will always have some degree of inter-reader variability; 2. the time constraints of routine practice make comprehensive assessment of large tissue sections challenging^{7,13}; 3. perceptual limitations may introduce bias in VTA, for example, the same TILs density is perceived to be higher if there is limited stroma.

Research in using machine learning (ML) algorithms to analyze histology has recently produced encouraging results, fueled by improvements in both hardware and methodology. Algorithms that learn patterns from labeled data, based on “deep learning” neural networks, have obtained promising results in many challenging problems. Their success has translated well to digital

pathology, where they have demonstrated outstanding performance in tasks like mitosis detection, identification of metastases in lymph node sections, tissue segmentation, prognostication, and computational TILs assessment (CTA)^{14–17}. ‘Traditional’ computational analysis of histology focuses on complex image analysis routines, that typically require extraction of handcrafted features and that often do not generalize well across data sets^{18,19}. Although studies utilizing deep learning-based methods suggest impressive diagnostic performance, and better generalization across data sets, these methods remain experimental. Table 1 shows a sample of published CTA algorithms and discusses their strengths and limitations, in complementarity with a previous literature review by the TIL-WG^{16,20–31}.

This review and perspective provides a broad outline of key issues that impact the development and translation of computational tools for TILs assessment. The ideal intended outcome is that CTA is successfully integrated into the routine clinical workflow; there is significant potential for CTA to address inherent limitations in VTA, and partially to mitigate high clinical demands in remote and under-resourced settings. This is not too difficult to conceive, and there are documented success stories in the commercialization and clinical adoption of computational algorithms including pap smear cytology analyzers³², blood analyzers³³, and automated immunohistochemistry (IHC) workflows for ER, PR, Her2, and Ki67^{34–38}.

THE IMPACT OF STAINING APPROACH ON ALGORITHM DESIGN AND DEPLOYMENT

The type of stain and imaging modality will have a significant impact on algorithm design, validation, and capabilities. VTA guideline from the TIL-WG focus on assessment of stromal TILs (sTIL) using hematoxylin and eosin (H&E)-stained formalin-fixed paraffin-embedded sections, given their practicality and widespread availability, and the clear presentation of tissue

[#]A full list of authors and their affiliations appears at the end of the paper.

Table 1. Sample CTA algorithms from the published literature.

Stain Approach	Ref	Data set	Method	Ground truth	Notes
H&E Patch Classification	24	Multiple sites TCGA data set	CNN	Labeled patches (yes/no TILs) Annotations are open-access	Strengths: large-scale study with investigation of spatial TIL maps. AV includes molecular correlates. Limitations: does not distinguish sTIL and iTIL; does not classify individual TILs*. Other: we defined CTA TIL score as fraction of patches that contain TILs, and found this to be correlated with VTA ($R = 0.659, p = 2e-35$).
Semantic segmentation	16	Breast TCGA data set	FCN	Traced region boundaries (exhaustive) Annotations are open-access	Strengths: large sample size and regions; investigates inter-rater variability at different experience levels; delineation of tumor, stroma and necrosis regions. Limitations: only detects dense TIL infiltrates*; does not classify individual TILs*.
Semantic segmentation + Object detection	25	Breast TCGA data set	Seeding + FCN	Traced region boundaries (exhaustive)	Strengths: mostly follows TIL-WG VTA guidelines. AV includes correlation with consensus VTA scores and inter-pathologist variability.
Object detection	26	Private data set Breast METABRIC data set	SVM using morphology features	Labeled & segmented nuclei within labeled region Labeled nuclei	Limitations: heavy ground truth requirement*; underpowered CV; and limited manually annotated slides. Strengths: robust analysis and exploration of molecular TIL correlates.
Object detection + inferred TIL localization	27	Breast Private data set	RG and MRF	Qualitative density scores Labeled patches (low-medium-high density)	Limitations: individual labeled nuclei are limited; does not distinguish TILs in different histologic regions*. Strengths: explainable model and modular pipeline.
IHC Object detection + manual regions	28	NSCLC Private data sets Breast	Watershed + SVM classifier	Labeled nuclei	Limitations: does not distinguish sTIL and iTIL; does not classify individual TILs. Limited AV sample size. Strengths: explainable model; robust CV; captures spatial TIL clustering.
Object detection	29	Multiple Private data set	SVM classifier using morphology features	Labeled nuclei	Limitations: limited AV; does not distinguish sTIL and iTIL. Strengths: infers TIL localization using spatial localization. Robust CV. Investigation of spatial TIL patterns.
Object detection	30	Multiple Private data set	Complex pipeline (non-DL) Multiple DL pipelines	Qualitative density scores Overall density estimates Labeled nuclei within FOV (exhaustive)	Limitations: individual labeled nuclei are limited; not clear if spatial clustering has 1:1 correspondence with regions. Strengths: CTA within manual regions, including invasive margin. Limitations: unpublished AV. Strengths: large-scale, robust AV. Systematic benchmarking. Limitations: no CV; does not distinguish TILs in different regions*.

This non-exhaustive list has been restricted to H&E and chromogenic IHC, although excellent works exist showing CTA based on other approaches like multiplexed immunofluorescence²¹⁻²³. Published CTA algorithms vary markedly in their approach to TIL scoring, the robustness of their validation, their interpretability, and their consistency with published VTA guidelines. Strengths and limitations of each publication is highlighted, with general limitations (related to the broad approach used, not the specific paper) are marked with an asterisk (*). Going forward, nuanced approaches are needed, ideally incorporating workflows for robust quantification and validation as presented in this paper. Different approaches have different ground truth requirements (illustrated in Fig. 1, panel f), hence the need for large-scale ground truth data sets. We encourage all future CTA publications to open-access their data sets whenever possible. Of note are two major efforts: 1. A group of scientists, including the US FDA and the TIL-WG, is collaborating to crowdsourcing pathologists and collect images and pathologist annotations that can be qualified by the FDA medical device development tool program; 2. The TIL-WG is organizing a challenge to validate CTA algorithms against clinical trial outcome data (CV).
AV analytical validation, CNN convolutional neural network, DL deep learning, FCN fully convolutional network, FOV field of view, MRF markov random field, RG region growing, NSCLC non-small cell lung cancer, SVM support vector machine.

architecture this stain provides^{10–12,39}. Multiple studies have relied on in situ approaches like IHC, in situ hybridization (ISH), or genomic deconvolution in assessing TILs^{11,40,41}. These modalities, however, are not typically used in daily clinical TILs assessment, as they are either still experimental, rely on assays of variable

reliability, or involve stains not widely used in clinical practice, especially in low-income settings^{4,10,11}. It is also difficult to quantitate and establish consistent thresholds for IHC measurement of even well-defined epitopes, such as Ki67 and ER, between different labs^{42,43}. Moreover, there is no single IHC stain that

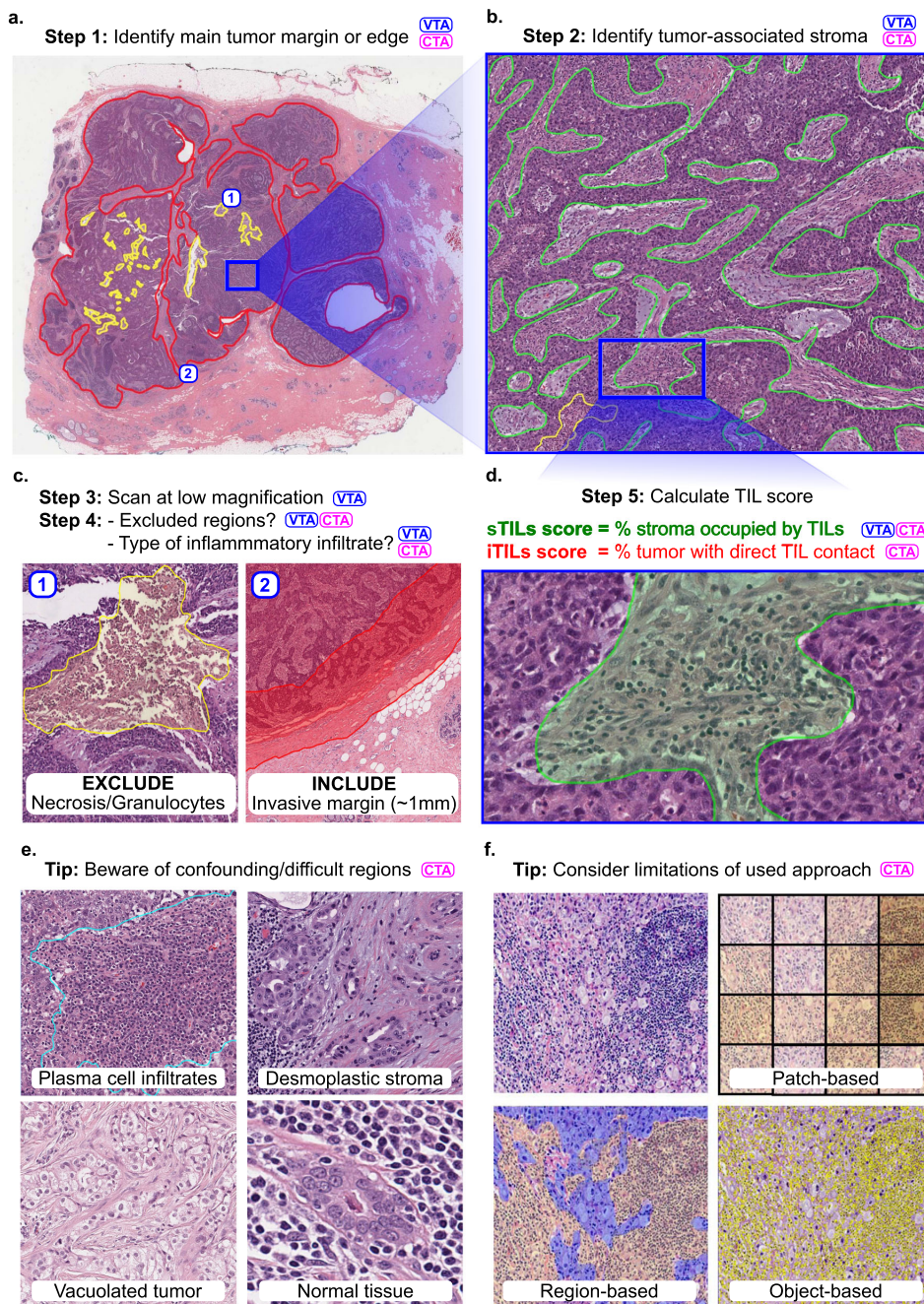


Fig. 1 Outline of the visual (VTA) and computational (CTA) procedure for scoring TILs in breast carcinomas. TIL scoring is a complex procedure, and breast carcinomas are used as an example. Specific guidelines for scoring different tumors are provided in the references. Steps involved in VTA and/or CTA are tagged with these abbreviations. CTA according to TIL-WG guidelines involves TIL scoring in different tissue compartments. **a** Invasive edge is determined (red) and key confounding regions like necrosis (yellow) are delineated. **b** Within the central tumor, tumor-associated stroma is determined (green). Other considerations and steps are involved depending on histologic subtype, slide quality, and clinical context. **c** Determination of regions for inclusion or exclusion in the analysis in accordance with published guidelines. **d** Final score is estimated (visually) or calculated (computationally). In breast carcinomas, stromal TIL score (sTIL) is used clinically. Intratumoral TIL score (iTIL) is subject to more VTA variability, which has hampered the generation of evidence demonstrating prognostic value; perhaps CTA of iTILs will prove less variable and, consequently, prognostic. **e** The necessity of diverse pathologist annotations for robust analytical validation of computational models. Desmoplastic stroma may be misclassified as tumor regions; Vacuolated tumor may be misclassified as stroma; intermixed normal acini or ducts, DCIS/LCIS, and blood vessels may be misclassified as tumor; plasma cells are sometimes misclassified as carcinoma cells. Note that while the term “TILs” includes lymphocytes, plasma cells and other small mononuclear infiltrates, lumping these categories may not be optimal from an algorithm design perspective; plasma cells tend to be morphologically different from lymphocytes in nuclear texture, size, and visible cytoplasm. **f** Various computational approaches may be used for computational scoring. The more granular the algorithm is, the more accurate/useful it is likely to be, but—as a trade-off—the more it relies on exhaustive manual annotations from pathologists. The least granular approach is patch classification, followed by region delineation (segmentation), then object detection (individual TILs). A robust computational scoring algorithm likely utilizes a combination of these (and related) approaches.

highlights all mononuclear cells with high sensitivity and specificity, so H&E remains the stain typically used in the routine clinical setting⁴⁴.

Despite these issues, there are significant potential advantages for using IHC with CTAs. By specifically staining TILs, IHC can make image analysis more reliable, and can also present new opportunities for granular TILs subclassification; different TIL subpopulations, including CD4+ T cells, CD8+ T cells, Tregs, NK cells, B cells, etc, convey pertinent information on immune activation and repression^{4,12}. IHC is already utilized in standardization efforts for TILs assessment in colorectal carcinomas^{45,46}. The specific highlighting of TILs by IHC can improve algorithm specificity^{47,48}, and enable characterization of TIL subpopulations that have potentially distinct prognostic or predictive roles^{49,50}. IHC can reduce misclassification of intratumoral TILs, which are difficult to reliably assess given their resemblance to tumor or supporting cells in many contexts like lobular breast carcinomas, small-blue-cell tumors like small cell lung cancer, and primary brain tumors^{4,12}.

CHARACTERISTICS OF CTA ALGORITHMS THAT CAPTURE CLINICAL GUIDELINES

TIL-WG guidelines for VTA are somewhat complex^{4,10–12}. There are VTA guidelines for many primary solid tumors and metastatic tumor deposits^{10,12}, for untreated infiltrating breast carcinomas¹¹, post-neoadjuvant residual carcinomas of the breast³⁹, and for carcinoma in situ of the breast³⁹. TILs score is defined as the fraction of a tissue compartment that is occupied by TILs (lymphoplasmacytic infiltrates). Different compartments have different prognostic relevance; tumor-associated sTILs is the most relevant in most solid tumors, whereas intratumoral TIL score (iTILs) has been reported to be prognostic, most notably in melanoma¹⁰. The spatial and visual context of TILs is strongly confounded by organ site, histologic subtype, and histomorphologic variables; therefore, it is important to provide situational context and instructions for clinical use of the CTA algorithms^{24,51,52}. For example, a CTA algorithm designed for general-purpose breast cancer TILs scoring should be validated on different subtypes (infiltrating ductal, infiltrating lobular, mucinous, etc) and on a wide array of slides that capture variabilities in tumor phenotype (e.g., vacuolated tumor, necrotic tumor, etc), stromal phenotype (e.g., desmoplastic stroma), TIL densities, and sources of variability like staining and artifacts. That being said, it is plausible to assume that the biology and significance of TILs may vary in different clinical and genomic subtypes of the same primary cancer site, and that a general-purpose TILs-scoring algorithm may not be applicable. Further research into the commonalities and differences in

the prognostic and biological value of TILs in different tissue sites and within different subtypes of the same cancer is warranted.

Clear inclusion criteria are helpful in deciding whether a slide is suitable for a particular CTA algorithm. For robust implementation, it is useful to: 1. detect when slides fail to meet its minimum quality; 2. provide some measure of confidence in its predictions; 3. be free of single points of failure (i.e., modular enough to tolerate failure of some sub-components); 4. be somewhat explainable, such that an expert pathologist can understand its limitations, common failure modes, and what the model seems to rely on in making decisions. Algorithms for measuring image quality and detecting artifacts will play an important role in the clinical implementation of CTA⁵³.

From a computer vision perspective, we can subdivide CTA in two separate tasks: 1. segmentation of the region of interest (e.g., intratumoral stroma in case of sTIL assessment) and 2. detection of individual TILs within that region. In practice, a set of complementary computer vision problems often need to be addressed to score TILs (Fig. 1). To segment the region in which TILs will be assessed, it is also often needed to explicitly segment regions for exclusion from the analysis. Although these can be manually annotated by pathologists, these judgements are a significant source of variability in VTA, and developing algorithms capable of performing these tasks could improve reproducibility and standardization^{7–9}.

Specifically, segmentation of the “central tumor” and the “invasive margin/edge” enable TILs quantitation to be focused in relevant areas, excluding “distant” stroma along with normal tissue and surrounding structures. A semi-precise segmentation of invasive margin also allows sTILs score to be broken down for the margin and central tumor regions (especially, in colorectal carcinomas) and to characterize peri-tumoral TILs independently¹⁰. Within the central tumor, segmenting carcinoma cell nests and intratumoral stroma enables separate measurements for sTIL and iTIL densities. Furthermore, segmentation helps exclude key confounder regions that need to be excluded from the analysis. This includes necrosis, tertiary lymphoid structures, intermixed normal tissue or DCIS/LCIS (in breast carcinoma), pre-existing lymphoid stroma (in lymph nodes and oropharyngeal tumors), perivascular regions, intra-alveolar regions (in lung), artifacts, etc. This step requires high-quality segmentation annotations, and may prove to be challenging. Indeed, for routine clinical practice, it may be necessary to have a pathologist perform a quick visual confirmation of algorithmic region segmentations, and/or create high-level region annotations that may be difficult to produce algorithmically.

When designing a TIL classifier, consideration of key confounding cells is important. Although lymphocytes are, compared with

tumor cells, relatively monomorphic, their small sizes offer little lymphocyte-specific texture information; small or perpendicularly cut stromal cells and even prominent nucleoli may result in misclassifications. Apoptotic bodies, necrotic debris, neutrophils, and some tumor cells (especially in lobular breast carcinomas and small-blue-round cell tumors) are other common confounders. Quantitation of systematic misclassification errors is warranted; some misclassifications will have contradictory consequences for clinical decision making. For example, neutrophils are evidently associated with adverse clinical outcomes, whereas TILs are typically associated with favorable outcomes⁵¹. Note that some of the TIL-WG clinical guidelines have been optimized for human scoring and are not very applicable in CTA algorithm design. For example, in breast carcinomas it is advised to “include but not focus on” tumor invasive edge TILs and TILs “hotspots”; CTA circumvents the need to address these cognitive biases¹¹. To fully adhere to clinical guidelines, segmentation of TILs is warranted, so that the fraction of intratumoral stroma occupied by TILs is calculated.

COMPUTER-AIDED VERSUS FULLY AUTOMATED TILS ASSESSMENT

The extent to which computational tools can be used to complement clinical decision making is highly context-dependent, and is strongly impacted by cancer type and clinical setting^{54–57}. In a computer-aided diagnosis paradigm, CTA is only used to provide guidance and increase efficiency in the workflow by any combination of the following: 1. calculating overall TILs score estimates to provide a frame-of-reference for the visual estimate; 2. directing the pathologist attention to regions of interest for TIL scoring, helping mitigate inconsistencies caused by heterogeneity in TILs density in different regions within the same slide; 3. providing a quantitative estimate for TILs density within regions of interest that the pathologist identifies, hence reducing ambiguity in visual estimation. Two models exist to assess this type of workflow during model development. In the traditional open assessment framework, the algorithm is trained on a set of manually annotated data points and evaluated on an independent held-out testing set. Alternatively, a closed-loop framework may be adopted, whereby pathologists can use the algorithmic output to re-evaluate their original decisions on the held-out set after exposure to the algorithmic results^{55,56}. Both frameworks have pros and cons, although the closed-loop framework enables assessment of the potential impact that CTA has on altering the clinical decision-making process⁵⁶.

The alternative paradigm is an entirely computational pipeline for CTA. This approach clearly provides efficiency gains, which could markedly reduce costs and accelerate development in a research setting. When the sample sizes are large enough, a few failures (i.e., “noise”) could be tolerated without altering the overall conclusions. This is contrary to clinical medicine, where CTA is expected to be highly dependable for each patient, especially when it is used to guide treatment decisions. Owing to the highly consequential nature of medical decision-making, a stand-alone CTA algorithm requires a higher bar for validation. It is also likely that even validated stand-alone CTA tools will need “sanity checks” by pathologists, guarding against unexpected failures. For example, a CTA report may be linked to a WSI display system to visualize the intermediate results (i.e., detected tissue boundaries and TILs locations) that were used by the algorithm to reach its decision (Fig. 2).

We do not envision computational models at their current level of performance replacing pathologist expertise. In fact, we would argue that quite the opposite is true; CTA enables objective quantitative assessment of an otherwise ambiguous metric, enabling the pathologist to focus more of his/her time on higher-order decision-making tasks⁵⁴. With that in mind, we argue

that the efficiency gains from CTA in under-resourced settings are likely to be derived from workflow efficiency, as opposed to reducing the domain expertise required to make diagnostic and therapeutic assessments. When used in a telepathology setting, i.e., off-site review of WSIs, CTA is still likely to require supervision by an experienced attending pathologist. Naturally, this depends on infrastructure, and one may argue that the cost-effectiveness of CTA is determined by the balance between infrastructure costs (WSI scanners, computing facilities, software, cloud support, etc) and expected long-term efficiency gains.

VALIDATION AND TRAINING ISSUES SURROUNDING COMPUTATIONAL TIL SCORING

CTA algorithms will need to be validated just like any prognostic or predictive biomarker to demonstrate preanalytical validation (Pre-AV), analytical validation (AV), clinical validation (CV), and clinical utility^{8,58,59}. In brief, Pre-AV is concerned with procedures that occur before CTA algorithms are applied, and include items like specimen preparation, slide quality, WSI scanner magnification and specifications, image format, etc; AV refers to accuracy and reproducibility; CV refers to stratification of patients into clinically meaningful subgroups; clinical utility refers to overall benefit in the clinical setting, considering existing methods and practices. Other considerations include cost-effectiveness, implementation feasibility, and ethical implications⁵⁹. VTA has been subject to extensive AV, CV, and clinical utility assessment, and it is critical that CTA algorithms are validated using the same high standards^{7,8}. The use-case of a CTA algorithm, specifically whether it is used for computer-aided assessment or for largely unsupervised assessment, is a key determinant of the extent of required validation. Key resources to consult include: 1. Recommendations by the Society for Immunotherapy of Cancer, for validation of diagnostic biomarkers; 2. Guidance documents by the US Food and Drug Administration (FDA); 3. Guidelines from the College of American Pathologists, for validation of diagnostic WSI systems^{60–64}. Granted, some of these require modifications in the CTA context, and we will highlight some of these differences here.

Pre-AV is of paramount importance, as CTA algorithm performance may vary in the presence of artifacts, variability in staining, tissue thickness, cutting angle, imaging, and storage^{65–68}. Trained pathologists, on the other hand, are more agile in adapting to variations in tissue processing, although these factors can still impact their visual assessment. Some studies have shown that the implementation of a DICOM standard for pathology images can improve standardization and improve interoperability if adopted by manufacturers^{67,69}. Techniques for making algorithms robust to variations, rather than eliminating the variations, have also been widely studied and are commonly employed^{69–72}. According to CAP guidelines, it is necessary to perform in-house validation of CTAs in all pathology laboratories, to validate the entire workflow (i.e., for each combination of tissue, stain, scanner, and CTA) using adequate sample size representing the entire diagnostic spectrum, and to re-validate whenever a significant component of the pre-analytic workflow changes⁶². Pre-AV and AV are most suitable in the in-house validation setting, as they can be performed with relatively fewer slides. It may be argued that proper in-house Pre-AV and AV suffice, provided large-scale prospective (or retrospective-prospective) AV, CV, and Clinical Utility studies were performed in a multi-center setting. Demonstrating local equivalency of Pre-AV and AV results can thus allow “linkage” to existing CV and Clinical Utility results assuming comparable patient populations.

AV typically involves quantitative assessment of CTA algorithm performance using ML metrics like segmentation or classification accuracy, prediction vs truth error, and area under receiver-operator characteristic curve or precision-recall curves. AV also includes validation against “non-classical” forms of ground truth like

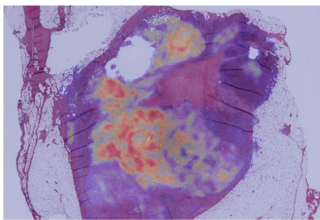
Patient Name / ID: DOE, Jane / AQH12CR3-DX-2		21/05/2020 03:22 PM		
Gender: Female Age: 46 Dx: Breast carcinoma, right, primary; Stage IB		Tx: Not initiated, No NACT		
Histology: Invasive ductal carcinoma / NST; Grade 3		Stain: H&E, FFPE	Other Markers: TN (ER-, PR-, Her2-); Ki67 < 25%	
	Global density: Whole-slide score	Local density: 50 μm x 50 μm fields	Local density: 100 μm x 100 μm fields	Local density: 200 μm x 200 μm fields
Stromal TILs	40.3 %	54.2 (\pm 20.1) %	52.1 (\pm 7.4) %	41.2 (\pm 5.1) %
Intra-tumoral TILs	5.6 %	0.1 (\pm 3.1) %	2.5 (\pm 2.1) %	4.9 (\pm 1.1) %
Invasive margin TILs	7.8 %	3.7 (\pm 4.1) %	6.2 (\pm 2.6) %	8.2 (\pm 0.8) %
Tissue delineation confidence: 0.95 TIL classification confidence: 0.86 TIL heatmap: See right; refer to WSI display for detailed tissue delineation, TIL classification, and zoomable heatmap. Distance from stromal TIL to nearest tumor: 62.1 (\pm 23.7) μ m Distance from tumor to nearest TIL: 726.9 (\pm 13.5) μ m Number of TIL clusters per unit area: 1.3 / mm ² TIL cluster morphology: Brisk, diffuse - moderate heterogeneity TIL cluster size: 320 (\pm 129) μ m Multivariable PFS prob.: 0.87 (1 yr) - 0.76 (3 yrs) - 0.67 (5 yrs) - 0.61 (10 yrs)				
On visual inspection, what is the quality of computational tissue delineation (tumor, stroma, etc) (circle one): <div style="display: flex; justify-content: center; gap: 10px;"> Very Poor Poor Acceptable Very good Excellent </div>				
On visual inspection, what is the quality of computational TIL localization (circle one): <div style="display: flex; justify-content: center; gap: 10px;"> Very Poor Poor Acceptable Very good Excellent </div>				
Pathologist Comments & Recommendations: <div style="border: 1px solid black; padding: 10px; min-height: 40px;"> None. Refer to pathology report for detailed histologic comment. </div>				
_____ Attending pathologist				

Fig. 2 Conceptual pathology report for computational TIL assessment (CTA). CTA reports might include global TIL estimates, broken down by key histologic regions, and estimates of classifier confidence. CTA reports are inseparably linked to WSI viewing systems, where algorithmic segmentations and localizations supporting the calculated scores are displayed for sanity check verification by the attending pathologist. Other elements, like local TIL estimates, TIL clustering results, and survival predictions may also be included.

co-registered IHC, in which case the registration process itself may also require validation. AV is a necessary prerequisite to CV as it answers the more fundamental question: "Do CTA algorithms detect TILs correctly?". AV should measure performance over the spectrum of variability induced by pre-analytic factors, and in cohorts that reflect the full range of intrinsic/biological variability. Naturally, this means that uncommon or rare subtypes of patterns are harder to validate owing to sample size limitations. AV of nucleus detection and classification algorithms has often neglected these issues, focusing on a large number of cells from a small number of cases.

Demonstrating the validity and generalization of prediction models is a complex process. Typically, the initial focus is on "internal" validation, using techniques like split-sample cross validation and bootstrapping. Later, the focus shifts to "external" validation, i.e., on an independent cohort from another institution. A hybrid technique called "internal-external" (cross-) validation may be appropriate when multi-institutional data sets (like the TCGA and METABRIC) are available, where training is performed on some hospitals/institutions and validation is performed on

others. This was recommended by Steyerberg and Harrell and used in some computational pathology studies^{16,73-75}.

Many of the events associated with cancer progression and subtyping are strongly correlated, so it may not be enough to show correspondence between global/slide-level CTA and VTA scores, as this shortcuts the AV process⁴⁹. AV therefore relies on the presence of quality "ground truth" annotations. Unfortunately, there is a lack of open-access, large-scale, multi-institutional histology segmentation and/or TIL classification data sets, with few exceptions^{16,24,76,77}. To help address this, a group of scientists, including the US FDA Center for Devices and Radiological Health (CDRH) and the TIL-WG, is collaborating to crowdsource pathologists and collect images and pathologist annotations that can be qualified by the FDA/CDRH medical device development tool program (MDDT). The MDDT qualified data would be available to any algorithm developer to be used for the analytic evaluation of their algorithm performance in a submission to the FDA/CDRH⁷⁸. The concept of "ground truth" in pathology can be vague and is often subjective, especially when dealing with H&E; it is therefore important to measure inter-rater variability by having multiple

experts annotate the same regions and objects^{7,8}. A key bottleneck in this process is the time commitment of pathologists, so collaborative, educational and/or crowdsourcing settings can help circumvent this limitation^{16,79}. It should be stressed, however, that although annotations from non-pathologists or residents may be adequate for CTA algorithm training; validation may require ground truth annotations created or reviewed by experienced practicing pathologists^{16,80}.

It is important to note that the ambiguity in ground truth (even if determined by consensus by multiple pathologists) typically warrants additional validation using objective criteria, most notably the ability to predict concrete clinical endpoints in validated data sets. One of the best ways to meet this validation bar is to use WSIs from large, multi-institutional randomized-controlled trials. To facilitate this effort, the TIL-WG is establishing strategic international partnerships to organize a machine learning challenge to validate CTA algorithms using clinical trials data. The training sets would be made available for investigators to train and fine tune their models, whereas separate blinded validation sets would only be provided once a locked-down algorithm has been established. Such resources are needed so that different algorithms and approaches can be directly compared on the same, high-quality data sets.

CTA FOR CLINICAL VERSUS ACADEMIC USE

Like VTA, CTA may be considered to fall under the umbrella of “imaging biomarkers,” and likely follows a similar validation roadmap to enable clinical translation and adoption^{38,81,82}. CTA may be used in the following academic settings, to name a few: 1. as a surrogate marker of response to experimental therapy in animal models; 2. as a diagnostic or predictive biomarker in retrospective clinical studies using archival WSI data; 3. as a diagnostic or predictive biomarker in prospective randomized-controlled trials. Incorporation of imaging biomarkers into prospective clinical trials requires some form of analytical and clinical validation (using retrospective data, for example), resulting in the establishment of Standards of Practice for trial use⁸¹. Establishment of clinical validity and utility in multicentric prospective trials is typically a prerequisite for use in day-to-day clinical practice. In a research environment, it is not unusual for computational algorithms to be frequently tweaked in a closed-loop fashion. This tweaking can be as simple as altering hyperparameters, but can include more drastic changes like modifications to the algorithm or (inter)active machine learning^{83,84}. From a standard regulatory perspective, this is problematic as validation requires a defined “lockdown” and version control; any change generally requires at least partial re-validation^{64,85}. It is therefore clear that the most pronounced difference between CTA use in basic/retrospective research, prospective trials, and routine clinical setting is the rigor of validation required^{38,81,82}.

In a basic/retrospective research environment, there is naturally a higher degree of flexibility in adopting CTA algorithms. For example, all slides may be scanned using the same scanner and using similar tissue processing protocols. In this setting, there is no immediate need for worrying about algorithm generalization performance under external processing or scanning conditions. Likewise, it may not be necessary to validate the model using ground truth from multiple pathologists, especially if some degree of noise can be tolerated. Operational issues and practicality also play a smaller role in basic/retrospective research settings; algorithm speed and user friendliness of a particular CTA algorithm may not be relevant when routine/repetitive TILs assessment is not needed. Even the nature of CTA algorithms may be different in a non-clinical setting. For instance, even though there is conflicting evidence on the prognostic value of iTILs in breast cancer, there are motivations to quantify them in a research environment. It should be noted, however, that this

flexibility is only applicable for CTA algorithms that are being used to support non-clinical research projects, not for those algorithms that are being validated for future clinical use.

THE FUTURE OF COMPUTATIONAL IMAGE-BASED IMMUNE BIOMARKERS

CTA algorithms can enable characterization of the tumor microenvironment beyond the limits of human observers, and will be an important tool in identifying latent prognostic and predictive patterns of immune response. For one, CTA enables calculation of local TIL densities at various scales, which may serve as a guide to “pockets” of differential immune activation (Fig. 2). This surpasses what is possible with VTA and such measurements are easy to calculate provided that CTA algorithms detect TILs with adequate sensitivity and specificity. Several studies have identified genomic features that in hindsight are associated with TILs, and CTA presents opportunities for systematic investigation of these associations^{24,26,74,86,87}. The emergence of assays and imaging platforms for multiplexed immunofluorescence and in situ hybridization will present new horizons for identifying predictive immunologic patterns and for understanding the molecular basis of tumor-immune interactions^{88,89}; these approaches are increasingly becoming commoditized.

Previous work examined how various spatial metrics from cancer-associated stroma relate to clinical outcomes, and similar concepts can be borrowed; for example, metrics capturing the complex relationships between TILs and other cells/structures in the tumor microenvironment⁹⁰. CTA may enable precise definitions of “intratumoral stroma”, for example using a quantitative threshold (i.e., “stroma within x microns from nearest tumor nest”). Similar concepts could be applied when differentiating tertiary lymphocytic aggregates, or other TIL hotspots, from infiltrating TILs that presumably have a direct role in anticancer response. It is also important to note that lymphocytic aggregation and other higher-order quantitative spatial metrics may play important prognostic roles yet to be discovered. A CTA study identified five broad categories of spatial organization of TILs infiltration, which are differentially associated with different cancer sites and subtypes²⁴. Alternatively, TILs can be placed on a continuum, such that sTILs that have a closer proximity to carcinoma nests get a higher weight. iTILs could be characterized using similar reasoning. Depending on available ground truth, numerous spatial metrics can be calculated. Nuanced assessment of immune response can be performed; for example, number of apoptotic bodies and their relation to nearby immune infiltrates. It is likely that there would be a considerable degree of redundancy in the prognostic value of CTA metrics; such redundancy is not uncommon in genomic biomarkers⁹¹. This should not be problematic as long as statistical models properly account for correlated predictors. In fact, the ability to calculate numerous metrics for a very large volume of cases enables large-scale, systematic discovery of histological biomarkers, bringing us a step closer to evidence-based pathology practice.

Learning-based algorithms can be utilized to learn prognostic features directly from images in a minimally biased manner (without explicit detection of TILs), and to integrate these with standard clinico-pathologic and genomic predictors. The approach of using deep learning algorithms to first detect and classify TILs and structures in histology, and then to calculate quantitative features of these objects, presents a way of closely modeling the clinical guidelines set forth by expert pathologists. Here, the power of learning algorithms is directed at providing highly accurate and robust detection and classification to enable reproducible and quantitative measurement. Although this approach is interpretable and provides a clear path for analytic validation, the limitation is that quantitative features are prescribed instead of learned. Recently, there have been

successful efforts to develop end-to-end prognostic deep learning models that learn to directly predict clinical outcomes from raw images without any intermediate classification of histologic objects like TILs^{17,92}. Although these end-to-end learning approaches have the potential to learn latent prognostic patterns (including those impossible to assess visually), they are less interpretable and thus the factors driving the predictions are currently unknown.

Finally, we would note that one of the key limitations of machine learning models, and deep learning models in particular, is their opaqueness. It is often the case that model accuracy comes at a cost to explainability, giving rise to the term “black box” often associated with deep learning. The problem with less explainable models is that key features driving output may not be readily identifiable to evaluate biologic plausibility, and hence the only safeguard against major flaws is extensive validation⁹³. Perhaps the most notorious consequence of this problem is “adversarial examples”, which are images that look natural to the human eye but that are specifically crafted (e.g., by malicious actors) to mislead deep learning models to make targeted misclassifications⁹⁴. Nevertheless, recent advances in deep learning research have substantially increased model interpretability, and have devised key model training strategies (e.g., generative adversarial neural networks) to increase performance robustness^{93,95–97}.

CONCLUSIONS

Advances in digital pathology and ML methodology have yielded expert-level performance in challenging diagnostic tasks. Evaluation of TILs in solid tumors is a highly suitable application for computational and computer-aided assessment, as it is both technically feasible and fills an unmet clinical need for objective and reproducible assessment. CTA algorithms need to account for the complexity involved in TIL-scoring procedures, and to closely follow guidelines for visual assessment where appropriate. TIL scoring needs to capture the concepts of stromal and intratumoral TILs and to account for confounding morphologies specific to different tumor sites, subtypes, and histologic patterns. Preanalytical factors related to imaging modality, staining procedure, and slide inclusion criteria are critical considerations, and robust analytical and clinical validation is key to adoption. In the clinical setting, CTA would ideally provide time- and cost-savings for pathologists, who face increasing demands for reporting biomarkers that are time-consuming to evaluate and subject to considerable inter- and intra- reader variability. In addition, CTA enables discovery of complex spatial patterns and genomic associations beyond the limits of visual scoring, and presents opportunities for precision medicine and scientific discovery.

Received: 15 July 2019; Accepted: 18 February 2020;
Published online: 12 May 2020

REFERENCES

- Piccart-Gebhart, M. et al. Adjuvant lapatinib and trastuzumab for early human epidermal growth factor receptor 2-positive breast cancer: results from the randomized phase III adjuvant lapatinib and/or Trastuzumab Treatment Optimization Trial. *J. Clin. Oncol.* **34**, 1034–1042 (2016).
- von Minckwitz, G. et al. Adjuvant pertuzumab and trastuzumab in early HER2-positive breast cancer. *N. Engl. J. Med.* **377**, 122–131 (2017).
- Denkert, C. et al. Tumor-associated lymphocytes as an independent predictor of response to neoadjuvant chemotherapy in breast cancer. *J. Clin. Oncol.* **28**, 105–113 (2010).
- Savas, P. et al. Clinical relevance of host immunity in breast cancer: from TILs to the clinic. *Nat. Rev. Clin. Oncol.* **13**, 228–241 (2016).
- Burns, P. B., Rohrich, R. J. & Chung, K. C. The levels of evidence and their role in evidence-based medicine. *Plast. Reconstr. Surg.* **128**, 305–310 (2011).
- Balic, M., Thomssen, C., Würstlein, R., Gnant, M. & Harbeck, N. St. Gallen/Vienna 2019: a brief summary of the consensus discussion on the optimal primary breast cancer treatment. *Breast Care* **14**, 103–110 (2019).
- Denkert, C. et al. Standardized evaluation of tumor-infiltrating lymphocytes in breast cancer: results of the ring studies of the international immuno-oncology biomarker working group. *Mod. Pathol.* **29**, 1155–1164 (2016).
- Wein, L. et al. Clinical validity and utility of tumor-infiltrating lymphocytes in routine clinical practice for breast cancer patients: current and future directions. *Front. Oncol.* **7**, 156 (2017).
- Brambilla, E. et al. Prognostic effect of tumor lymphocytic infiltration in resectable non-small-cell lung cancer. *J. Clin. Oncol.* **34**, 1223–1230 (2016).
- Hendry, S. et al. Assessing tumor-infiltrating lymphocytes in solid tumors: a practical review for pathologists and proposal for a standardized method from the international immuno-oncology biomarkers working group: part 1: assessing the host immune response, til in invasive breast carcinoma and ductal carcinoma in situ, metastatic tumor deposits and areas for further research. *Adv. Anat. Pathol.* **24**, 235–251 (2017).
- Salgado, R. et al. The evaluation of tumor-infiltrating lymphocytes (TILs) in breast cancer: recommendations by an International TILs Working Group 2014. *Ann. Oncol.* **26**, 259–271 (2015).
- Hendry, S. et al. Assessing tumor-infiltrating lymphocytes in solid tumors: a practical review for pathologists and proposal for a standardized method from the international immuno-oncology biomarkers working group: part 2: til in melanoma, gastrointestinal tract carcinomas, non-small cell lung carcinoma and mesothelioma, endometrial and ovarian carcinomas, squamous cell carcinoma of the head and neck, genitourinary carcinomas, and primary brain tumors. *Adv. Anat. Pathol.* **24**, 311–335 (2017).
- Brunyé, T. T., Mercan, E., Weaver, D. L. & Elmore, J. G. Accuracy is in the eyes of the pathologist: the visual interpretive process and diagnostic accuracy with digital whole slide images. *J. Biomed. Inform.* **66**, 171–179 (2017).
- Ehteshami Bejnordi, B. et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA* **318**, 2199–2210 (2017).
- Tellez, D. et al. Whole-slide mitosis detection in H&E breast histology using PHH3 as a reference to train distilled stain-invariant convolutional networks. *IEEE Trans. Med. Imaging* (2018). <https://doi.org/10.1109/TMI.2018.2820199>
- Amgad, M. et al. Structured crowdsourcing enables convolutional segmentation of histology images. *Bioinformatics* **35**, 3461–3467 (2019).
- Mobadersany, P. et al. Predicting cancer outcomes from histology and genomics using convolutional networks. *Proc. Natl Acad. Sci. USA* **115**, E2970–E2979 (2018).
- Litjens, G. et al. A survey on deep learning in medical image analysis. *Med. Image Anal.* **42**, 60–88 (2017).
- Janowczyk, A. & Madabhushi, A. Deep learning for digital pathology image analysis: a comprehensive tutorial with selected use cases. *J. Pathol. Inform.* **7**, 29 (2016).
- Klauschen, F. et al. Scoring of tumor-infiltrating lymphocytes: from visual estimation to machine learning. *Semin. Cancer Biol.* **52**, 151–157 (2018).
- Barua, S. et al. Spatial interaction of tumor cells and regulatory T cells correlates with survival in non-small cell lung cancer. *Lung Cancer* **117**, 73–79 (2018).
- Schalper, K. A. et al. Objective measurement and clinical significance of TILs in non-small cell lung cancer. *J. Natl. Cancer Inst.* **107**, dju435 (2015).
- Brown, J. R. et al. Multiplexed quantitative analysis of CD3, CD8, and CD20 predicts response to neoadjuvant chemotherapy in breast cancer. *Clin. Cancer Res.* **20**, 5995–6005 (2014).
- Saltz, J. et al. Spatial organization and molecular correlation of tumor-infiltrating lymphocytes using deep learning on pathology images. *Cell Rep.* **23**, 181–193.e7 (2018).
- Amgad, M. et al. Joint region and nucleus segmentation for characterization of tumor infiltrating lymphocytes in breast cancer. in *Medical Imaging 2019: Digital Pathology* (eds. Tomaszewski, J. E. & Ward, A. D.) 20 (SPIE, 2019).
- Yuan, Y. et al. Quantitative image analysis of cellular heterogeneity in breast tumors complements genomic profiling. *Sci. Transl. Med.* **4**, 157ra143 (2012).
- Basavanahally, A. N. et al. Computerized image-based detection and grading of lymphocytic infiltration in HER2+ breast cancer histopathology. *IEEE Trans. Biomed. Eng.* **57**, 642–653 (2010).
- Corredor, G. et al. Spatial architecture and arrangement of tumor-infiltrating lymphocytes for predicting likelihood of recurrence in early-stage non-small cell lung cancer. *Clin. Cancer Res.* **25**, 1526–1534 (2019).
- Yoon, H. H. et al. Intertumoral heterogeneity of CD3 and CD8 T-cell densities in the microenvironment of dna mismatch-repair-deficient colon cancers: implications for prognosis. *Clin. Cancer Res.* **25**, 125–133 (2019).
- Swiderska-Chadaj, Z. et al. Convolutional Neural Networks for Lymphocyte detection in Immunohistochemically Stained Whole-Slide Images. (2018).

31. Heindl, A. et al. Relevance of spatial heterogeneity of immune infiltration for predicting risk of recurrence after endocrine therapy of ER+ breast cancer. *J. Natl. Cancer Inst.* **110**, dxi137 (2018).
32. Stoler, M. H. Advances in cervical screening technology. *Mod. Pathol.* **13**, 275–284 (2000).
33. Vis, J. Y. & Huisman, A. Verification and quality control of routine hematology analyzers. *Int. J. Lab. Hematol.* **38**, 100–109 (2016).
34. Perkel, J. M. Immunohistochemistry for the 21st century. *Science* **351**, 1098–1100 (2016).
35. Lloyd, M. C. et al. Using image analysis as a tool for assessment of prognostic and predictive biomarkers for breast cancer: how reliable is it? *J. Pathol. Inform.* **1**, 29 (2010).
36. Holten-Rossing, H., Møller Talman, M.-L., Kristensson, M. & Vainer, B. Optimizing HER2 assessment in breast cancer: application of automated image analysis. *Breast Cancer Res. Treat.* **152**, 367–375 (2015).
37. Gavrielides, M. A., Gallas, B. D., Lenz, P., Badano, A. & Hewitt, S. M. Observer variability in the interpretation of HER2/neu immunohistochemical expression with unaided and computer-aided digital microscopy. *Arch. Pathol. Lab. Med.* **135**, 233–242 (2011).
38. Hamilton, P. W. et al. Digital pathology and image analysis in tissue biomarker research. *Methods* **70**, 59–73 (2014).
39. Dieci, M. V. et al. Update on tumor-infiltrating lymphocytes (TILs) in breast cancer, including recommendations to assess TILs in residual disease after neoadjuvant therapy and in carcinoma in situ: a report of the International Immuno-Oncology Biomarker Working Group on Breast Cancer. *Semin. Cancer Biol.* **52**, 16–25 (2018).
40. Finotello, F. & Trajanoski, Z. Quantifying tumor-infiltrating immune cells from transcriptomics data. *Cancer Immunol. Immunother.* **67**, 1031–1040 (2018).
41. Chakravarthy, A. et al. Pan-cancer deconvolution of tumour composition using DNA methylation. *Nat. Commun.* **9**, 3220 (2018).
42. Ács, B. et al. Ki-67 as a controversial predictive and prognostic marker in breast cancer patients treated with neoadjuvant chemotherapy. *Diagn. Pathol.* **12**, 20 (2017).
43. Yi, M. et al. Which threshold for ER positivity? a retrospective study based on 9639 patients. *Ann. Oncol.* **25**, 1004–1011 (2014).
44. Göransson, C. et al. Immunohistochemical characterization of lymphocytes in microscopic colitis. *J. Crohns. Colitis* **7**, e434–e442 (2013).
45. Pagès, F. et al. International validation of the consensus Immunoscore for the classification of colon cancer: a prognostic and accuracy study. *Lancet* **391**, 2128–2139 (2018).
46. Galon, J. et al. Towards the introduction of the 'Immunoscore' in the classification of malignant tumours. *J. Pathol.* **232**, 199–209 (2014).
47. Väyrynen, J. P. et al. An improved image analysis method for cell counting lends credibility to the prognostic significance of T cells in colorectal cancer. *Virchows Arch.* **460**, 455–465 (2012).
48. Singh, U. et al. Analytical validation of quantitative immunohistochemical assays of tumor infiltrating lymphocyte biomarkers. *Biotech. Histochem.* **93**, 411–423 (2018).
49. Buisseret, L. et al. Tumor-infiltrating lymphocyte composition, organization and PD-1/PD-L1 expression are linked in breast cancer. *Oncoimmunology* **6**, e1257452 (2017).
50. Blom, S. et al. Systems pathology by multiplexed immunohistochemistry and whole-slide digital image analysis. *Sci. Rep.* **7**, 15580 (2017).
51. Gentles, A. J. et al. The prognostic landscape of genes and infiltrating immune cells across human cancers. *Nat. Med.* **21**, 938–945 (2015).
52. Stanton, S. E., Adams, S. & Disis, M. L. Variation in the incidence and magnitude of tumor-infiltrating lymphocytes in breast cancer subtypes: a systematic review. *JAMA Oncol.* **2**, 1354–1360 (2016).
53. Janowczyk, A., Zuo, R., Gilmore, H., Feldman, M. & Madabhushi, A. HistoQC: an open-source quality control tool for digital pathology slides. *JCO Clin. Cancer Inf.* **3**, 1–7 (2019).
54. Hipp, J. et al. Computer aided diagnostic tools aim to empower rather than replace pathologists: Lessons learned from computational chess. *J. Pathol. Inform.* **2**, 25 (2011).
55. Gurcan, M. N. Histopathological image analysis: path to acceptance through evaluation. *Microsc. Microanal.* **22**, 1004–1005 (2016).
56. Fauzi, M. F. A. et al. Classification of follicular lymphoma: the effect of computer aid on pathologists grading. *BMC Med. Inform. Decis. Mak.* **15**, 115 (2015).
57. Madabhushi, A., Agner, S., Basavanahally, A., Doyle, S. & Lee, G. Computer-aided prognosis: predicting patient and disease outcome via quantitative fusion of multi-scale, multi-modal data. *Comput. Med. Imaging Graph.* **35**, 506–514 (2011).
58. Hayes, D. F. Precision medicine and testing for tumor biomarkers—are all tests born equal? *JAMA Oncol.* **4**, 773–774 (2018).
59. Selleck, M. J., Senthil, M. & Wall, N. R. Making meaningful clinical use of biomarkers. *Biomark. Insights* **12**, 11772719–17715236 (2017).
60. Masucci, G. V. et al. Validation of biomarkers to predict response to immunotherapy in cancer: Volume I - pre-analytical and analytical validation. *J. Immunother. Cancer* **4**, 76 (2016).
61. Dobbin, K. K. et al. Validation of biomarkers to predict response to immunotherapy in cancer: volume II - clinical validation and regulatory considerations. *J. Immunother. Cancer* **4**, 77 (2016).
62. Pantanowitz, L. et al. Validating whole slide imaging for diagnostic purposes in pathology: guideline from the College of American Pathologists Pathology and Laboratory Quality Center. *Arch. Pathol. Lab. Med.* **137**, 1710–1722 (2013).
63. Fda, M. Guidance for Industry and Food and Drug Administration Staff - Technical Performance Assessment of Digital Pathology Whole Slide Imaging Devices. (2016). Available at: <https://www.fda.gov/media/90791/download>.
64. US Food and Drug Administration. Device Advice for AI and Machine Learning Algorithms. NCIPhub - Food and Drug Administration. Available at: <https://nciphub.org/groups/eedapstudies/wiki/DeviceAdvice>. (Accessed: 4th July 2017).
65. Leo, P. et al. Evaluating stability of histomorphometric features across scanner and staining variations: prostate cancer diagnosis from whole slide images. *J. Med. Imaging (Bellingham)* **3**, 047502 (2016).
66. Pantanowitz, L., Liu, C., Huang, Y., Guo, H. & Rohde, G. K. Impact of altering various image parameters on human epidermal growth factor receptor 2 image analysis data quality. *J. Pathol. Inform.* **8**, 39 (2017).
67. Pantanowitz, L. et al. Twenty years of digital pathology: an overview of the road travelled, what is on the horizon, and the emergence of vendor-neutral archives. *J. Pathol. Inform.* **9**, 40 (2018).
68. Zarella, M. D., Yeoh, C., Breen, D. E. & Garcia, F. U. An alternative reference space for H&E color normalization. *PLoS ONE* **12**, e0174489 (2017).
69. Herrmann, M. D. et al. Implementing the DICOM standard for digital pathology. *J. Pathol. Inform.* **9**, 37 (2018).
70. Van Eycke, Y.-R., Allard, J., Salmon, I., Debeir, O. & Decaestecker, C. Image processing in digital pathology: an opportunity to solve inter-batch variability of immunohistochemical staining. *Sci. Rep.* **7**, 42964 (2017).
71. Tellez, D. et al. Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. <http://arxiv.org/abs/1902.06543> (2019).
72. Hou, L. et al. Unsupervised histopathology image synthesis. <http://arxiv.org/abs/1712.05021> (2017).
73. Steyerberg, E. W. & Harrell, F. E. Jr. Prediction models need appropriate internal, internal-external, and external validation. *J. Clin. Epidemiol.* **69**, 245–247 (2016).
74. Natrajan, R. et al. Microenvironmental heterogeneity parallels breast cancer progression: a histology-genomic integration analysis. *PLoS Med.* **13**, e1001961 (2016).
75. Loi, S. et al. Tumor-infiltrating lymphocytes and prognosis: a pooled individual patient analysis of early-stage triple-negative breast cancers. *J. Clin. Oncol.* **37**, 559–569 (2019).
76. Beck, A. H. Open access to large scale datasets is needed to translate knowledge of cancer heterogeneity into better patient outcomes. *PLoS Med.* **12**, e1001794 (2015).
77. Litjens, G. et al. 1399 H&E-stained sentinel lymph node sections of breast cancer patients: the CAMELYON dataset. *Gigascience* **7**, gij065 (2018).
78. US Food and Drug Administration. Year 2: High-throughput truthing of microscope slides to validate artificial intelligence algorithms analyzing digital scans of pathology slides: data (images + annotations) as an FDA-qualified medical device development tool (MDDT). Available at: <https://ncihub.org/groups/eedapstudies/wiki/HighthroughputTruthingYear2?version=2>.
79. Ørting, S. et al. A survey of crowdsourcing in medical image analysis. <http://arxiv.org/abs/1902.09159> (2019).
80. Hannun, A. Y. et al. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nat. Med.* **25**, 65–69 (2019).
81. O'Connor, J. P. B. et al. Imaging biomarker roadmap for cancer studies. *Nat. Rev. Clin. Oncol.* **14**, 169–186 (2017).
82. Abramson, R. G. et al. Methods and challenges in quantitative imaging biomarker development. *Acad. Radiol.* **22**, 25–32 (2015).
83. Nalnsnik, M. et al. Interactive phenotyping of large-scale histology imaging data with HistomicsML. *Sci. Rep.* **7**, 14588 (2017).
84. Kwolek, B. et al. Breast Cancer Classification on Histopathological Images Affected by Data Imbalance Using Active Learning and Deep Convolutional Neural Network: 28th International Conference on Artificial Neural Networks, Munich, Germany, September 17–19, 2019, Proceedings. in *Artificial Neural Networks and Machine Learning—ICANN 2019: Workshop and Special Sessions* (eds. Tetko, I. V., Kúrková, V., Karpov, P. & Theis, F.) 11731, 299–312 (Springer International Publishing, 2019).
85. U.S. Food and Drug Administration (FDA). Considerations for Design, Development, and Analytical Validation of Next Generation Sequencing-Based In Vitro Diagnostics Intended to Aim in the Diagnosis of Suspected Germline Diseases. (2018). Available

- at: <https://www.fda.gov/downloads/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/UCM509838.pdf>.
86. Mukherjee, A. et al. Associations between genomic stratification of breast cancer and centrally reviewed tumour pathology in the METABRIC cohort. *NPJ Breast Cancer* **4**, 5 (2018).
 87. Cooper, L. A. et al. PanCancer insights from The Cancer Genome Atlas: the pathologist's perspective. *J. Pathol.* **244**, 512–524 (2018).
 88. Anderson, R. Multiplex fluorescence in situ hybridization (M-FISH). *Methods Mol. Biol.* **659**, 83–97 (2010).
 89. Longuespée, R. et al. Tissue proteomics for the next decade? Towards a molecular dimension in histology. *OMICS* **18**, 539–552 (2014).
 90. Beck, A. H. et al. Systematic analysis of breast cancer morphology uncovers stromal features associated with survival. *Sci. Transl. Med.* **3**, 108ra113 (2011).
 91. Cantini, L. et al. Classification of gene signatures for their information value and functional redundancy. *NPJ Syst. Biol. Appl.* **4**, 2 (2018).
 92. Meier, A. et al. 77P-End-to-end learning to predict survival in patients with gastric cancer using convolutional neural networks. *Ann. Oncol.* **29** <https://doi.org/10.1093/annonc/mdy269.07510.1093/annonc/mdy269.075> (2018).
 93. Guidotti, R. et al. A survey of methods for explaining black box models. *ACM Comput. Surv.* **51**, 1–42 (2018).
 94. Yuan, X., He, P., Zhu, Q. & Li, X. Adversarial examples: attacks and defenses for deep learning. *IEEE Trans. Neural Netw. Learn. Syst.* **30**, 2805–2824 (2019).
 95. Zhang, Q.-S. & Zhu, S.-C. Visual interpretability for deep learning: a survey. *Front. Inf. Technol. Electron. Eng.* **19**, 27–39 (2018).
 96. Yousefi, S. et al. Predicting clinical outcomes from large scale cancer genomic profiles with deep survival models. *Sci. Rep.* **7**, 11707 (2017).
 97. Yi, X., Wallia, E. & Babyn, P. Generative adversarial network in medical imaging: a review. *Med. Image Anal.* **58**, 101552 (2019).

ACKNOWLEDGEMENTS

L.A.D.C. is supported in part by the National Institutes of Health National Cancer Institute (NCI) grants U01CA220401 and U24CA19436201. R.S. is supported by the Breast Cancer Research Foundation (BCRF), grant No. 17-194. J.S. is supported in part by NCI grants UG3CA225021 and U24CA215109. A.M. is supported in part by NCI grants 1U24CA199374-01, R01CA202752-01A1, R01CA208236-01A1, R01 CA216579-01A1, R01 CA220581-01A1, 1U01 CA239055-01, 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 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 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. S.G. is supported by Susan G Komen Foundation (CCR CCR18547966) and a Young Investigator Grant from the Breast Cancer Alliance. T.O.N. receives funding support from the Canadian Cancer Society. M.M.S. is supported by P30 CA16672 DHHS/NCI Cancer Center Support Grant (CCSG). A.S. is supported in part by NCI grants 1UG3CA225021, 1U24CA215109, and Leidos 14 × 138. This work includes contributions from, and was reviewed by, individuals at the F.D.A. This work has been approved for publication by the agency, but it does not necessarily reflect official agency policy. Certain commercial materials and equipment are identified in order to adequately specify experimental procedures. In no case does such identification imply recommendation or endorsement by the FDA, nor does it imply that the items identified are necessarily the best available for the purpose. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health, the US Department of Veterans Affairs, the Department of Defense, the United States Government, or other governments or entities. The following is a list of current members of the International Immuno-Oncology Working Group (TILs Working Group). Members contributed to the manuscript through discussions, including at the yearly TIL-WG meeting, and have reviewed and provided input on the manuscript. The authors alone are responsible for the views expressed in the work of the TILs Working Group and they do not necessarily represent the decisions, policy or views of their employer.

AUTHOR CONTRIBUTIONS

This report is produced as a result of discussion and consensus by members of the International Immuno-Oncology Biomarker Working Group (TILs Working Group). All

authors have contributed to: 1) the conception or design of the work, 2) drafting the work or revising it critically for important intellectual content, 3) final approval of the completed version, 4) accountability for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

COMPETING INTERESTS

J.T. is funded by Visiopharm A/S, Denmark. A.M. is an equity holder in Elucid Bioimaging and in Inspirata Inc. He is also a scientific advisory consultant for Inspirata Inc. In addition he has served as a scientific advisory board member for Inspirata Inc, AstraZeneca, Bristol Meyers-Squibb and Merck. He also has sponsored research agreements with Philips and Inspirata Inc. His technology has been licensed to Elucid Bioimaging and Inspirata Inc. He is also involved in an NIH U24 grant with PathCore Inc, and three different R01 grants with Inspirata Inc. S.R.L. received travel and educational funding from Roche/Ventana. A.J.L. serves as a consultant for BMS, Merck, AZ/Medimmune, and Genentech. He is also provides consulting and advisory work for many other companies not relevant to this work. FPL does consulting for AstraZeneca, BMS, Roche, MSD Pfizer, Novartis, Sanofi, and Lilly. S.L.d.H., A.K., M.K., U.K., and M.B. are employees of Roche. J.M.S.B. is consultant for Insight Genetics, BioNTech AG, Biothernostics, Pfizer, RNA Diagnostics, and OncoXchange. He received funding from Thermo Fisher Scientific, Genoptix, Agendia, NanoString technologies, Stratifyer GmbH, and Biothernostics. L.F.S.K. is a consultant for Roche and Novartis. J.K.K. and A.H.B. are employees of PathAI. D.L.R. is on the advisory board for Amgen, AstraZeneca, Cell Signaling Technology, Cepheid, Daiichi Sankyo, GSK, Konica/Minolta, Merck, Nanostring, Perkin Elmer, Roche/Ventana, and Ultivue. He has received research support from AstraZeneca, Cepheid, Navigate BioPharma, NextCure, Lilly, Ultivue, Roche/Ventana, Akoya/Perkin Elmer, and Nanostring. He also has financial conflicts of interest with BMS, Biocept, PixelGear, and Rarecyte. S.G. is a consultant for and/or receives funding from Eli Lilly, Novartis, and G1 Therapeutics. J.A.W.M.v.d.L. is a member of the scientific advisory boards of Philips, the Netherlands and ContextVision, Sweden, and receives research funding from Philips, the Netherlands and Sectra, Sweden. S.A. is a consultant for Merck, Genentech, and BMS, and receives funding from Merck, Genentech, BMS, Novartis, Celgene, and Amgen. T.O.N. has consulted for Nanostring, and has intellectual property rights and ownership interests from Bioclassifier LLC. S.L. receives research funding to her institution from Novartis, Bristol Meyers-Squibb, Merck, Roche-Genentech, Puma Biotechnology, Pfizer and Eli Lilly. She has acted as consultant (not compensated) to Seattle Genetics, Pfizer, Novartis, BMS, Merck, AstraZeneca and Roche-Genentech. She has acted as consultant (paid to her institution) to Aduro Biotech. J.H. is director and owner of Slide Score BV. M.M.S. is a medical advisory board member of OptraScan. R.S. has received research support from Merck, Roche, Puma; and travel/congress support from AstraZeneca, Roche and Merck; and he has served as an advisory board member of BMS and Roche and consults for BMS.

ADDITIONAL INFORMATION

Correspondence and requests for materials should be addressed to R.S. or L.A.D.C.

Reprints and permission information is available at <http://www.nature.com/reprints>

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



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

© The Author(s) 2020

Mohamed Amgad¹, Elisabeth Specht Stovgaard², Eva Balslev², Jeppe Thagaard^{3,4}, Weijie Chen⁵, Sarah Dudgeon⁵, Ashish Sharma¹, Jennifer K. Kerner⁶, Carsten Denkert^{7,8,9}, Yinyin Yuan^{10,11}, Khalid AbdulJabbar^{10,11}, Stephan Wienert⁷, Peter Savas^{12,13}, Leonie Voorwerk¹⁴, Andrew H. Beck⁶, Anant Madabhushi^{15,16}, Johan Hartman¹⁷, Manu M. Sebastian¹⁸, Hugo M. Horlings¹⁹, Jan Hudeček²⁰, Francesco Ciompi²¹, David A. Moore²², Rajendra Singh²³, Elvire Roblin²⁴, Marcelo Luiz Balancin²⁵, Marie-Christine Mathieu²⁶, Jochen K. Lennerz²⁷, Pawan Kirtani²⁸, I-Chun Chen²⁹, Jeremy P. Braybrooke^{30,31}, Giancarlo Pruner³², Sandra Demaria³³, Sylvia Adams³⁴, Stuart J. Schnitt³⁵, Sunil R. Lakhani³⁶, Federico Rojo^{37,38}, Laura Comerma^{37,38}, Sunil S. Badve³⁹, Mehrnosh Khojasteh⁴⁰, W. Fraser Symmans⁴¹, Christos Sotiriou^{42,43}, Paula Gonzalez-Ericsson⁴⁴, Katherine L. Pogue-Geile⁴⁵, Rim S. Kim⁴⁵, David L. Rimm⁴⁶, Giuseppe Viale⁴⁷, Stephen M. Hewitt⁴⁸, John M. S. Bartlett^{49,50}, Frédérique Penault-Llorca^{51,52}, Shom Goel⁵³, Huang-Chun Lien⁵⁴, Sibylle Loibl⁵⁵, Zuzana Kos⁵⁶, Sherene Loi⁵⁷, Matthew G. Hanna⁵⁸, Stefan Michiels^{59,60}, Marleen Kok^{61,62}, Torsten O. Nielsen⁶³, Alexander J. Lazar^{64,65,66}, Zsuzsanna Bago-Horvath⁶⁷, Loes F. S. Kooreman^{68,69}, Jeroen A. W. M. van der Laak^{70,71}, Joel Saltz⁷¹, Brandon D. Gallas⁷², Uday Kurkure⁷³, Michael Barnes⁷², Roberto Salgado^{12,73}, Lee A. D. Cooper⁷⁴ and International Immuno-Oncology Biomarker Working Group

¹Department of Biomedical Informatics, Emory University School of Medicine, Atlanta, GA, USA. ²Department of Pathology, Herlev and Gentofte Hospital, University of Copenhagen, Herlev, Denmark. ³DTU Compute, Department of Applied Mathematics, Technical University of Denmark, Lyngby, Denmark. ⁴Visiopharm A/S, Hørsholm, Denmark. ⁵FDA/CDRH/OSEL/Division of Imaging, Diagnostics, and Software Reliability, Silver Spring, MD, USA. ⁶PathAI, Cambridge, MA, USA. ⁷Institut für Pathologie, Universitätsklinikum Gießen und Marburg GmbH, Standort Marburg, Philipps-Universität Marburg, Marburg, Germany. ⁸Institute of Pathology, Philipps-Universität Marburg, Marburg, Germany. ⁹German Cancer Consortium (DKTK), Partner Site Charité, Berlin, Germany. ¹⁰Centre for Evolution and Cancer, The Institute of Cancer Research, London, UK. ¹¹Division of Molecular Pathology, The Institute of Cancer Research, London, UK. ¹²Department of Research and Cancer Medicine, Peter MacCallum Cancer Centre, University of Melbourne, Victoria, Australia. ¹³Sir Peter MacCallum Department of Oncology, University of Melbourne, Parkville, Australia. ¹⁴Department of Tumor Biology & Immunology, The Netherlands Cancer Institute, Amsterdam, The Netherlands. ¹⁵Case Western Reserve University, Department of Biomedical Engineering, Cleveland, OH, USA. ¹⁶Louis Stokes Cleveland Veterans Administration Medical Center, Cleveland, OH, USA. ¹⁷Department of Oncology and Pathology, Karolinska Institutet and University Hospital, Solna, Sweden. ¹⁸Departments of Epigenetics and Molecular Carcinogenesis, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. ¹⁹Division of Molecular Pathology, The Netherlands Cancer Institute, Amsterdam, The Netherlands. ²⁰Department of Research IT, The Netherlands Cancer Institute, Amsterdam, The Netherlands. ²¹Department of Pathology, Radboud University Medical Center, Nijmegen, The Netherlands. ²²Department of Pathology, UCL Cancer Institute, London, UK. ²³Department of Pathology and Laboratory Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA. ²⁴Université Paris-Saclay, Univ. Paris-Sud, Villejuif, France. ²⁵Department of Pathology, Faculty of Medicine, University of São Paulo, São Paulo, Brazil. ²⁶Department of Medical Biology and Pathology, Gustave Roussy Cancer Campus, Villejuif, France. ²⁷Department of Pathology, Massachusetts General Hospital, Boston, MA, USA. ²⁸Department of Histopathology, Manipal Hospitals Dwarka, New Delhi, India. ²⁹Department of Oncology, National Taiwan University Cancer Center, Taipei, Taiwan. ³⁰Nuffield Department of Population Health, University of Oxford, Oxford, UK. ³¹Department of Medical Oncology, University Hospitals Bristol NHS Foundation Trust, Bristol, UK. ³²Pathology Department, Fondazione IRCCS Istituto Nazionale Tumori and University of Milan, School of Medicine, Milan, Italy. ³³Weill Cornell Medical College, New York, NY, USA. ³⁴Laura and Isaac Perlmutter Cancer Center, NYU Langone Medical Center, New York, NY, USA. ³⁵Department of Pathology, Brigham and Women's Hospital, Boston, MA, USA. ³⁶The University of Queensland Centre for Clinical Research and Pathology Queensland, Brisbane, Australia. ³⁷Pathology Department, CIBERONC-Instituto de Investigación Sanitaria Fundación Jiménez Díaz (IIS-FJD), Madrid, Spain. ³⁸GEICAM-Spanish Breast Cancer Research Group, Madrid, Spain. ³⁹Department of Pathology and Laboratory Medicine, Indiana University School of Medicine, Indianapolis, IN, USA. ⁴⁰Roche Tissue Diagnostics, Digital Pathology, Santa Clara, CA, USA. ⁴¹Department of Pathology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. ⁴²Breast Cancer Translational Research Laboratory, Institut Jules Bordet, Université Libre de Bruxelles (ULB), Brussels, Belgium. ⁴³ULB-Cancer Research Center (U-CRC) Université Libre de Bruxelles, Brussels, Belgium. ⁴⁴Breast Cancer Program, Vanderbilt-Ingram Cancer Center, Vanderbilt University Medical Center, Nashville, TN, USA. ⁴⁵NRG Oncology/NSABP, Pittsburgh, PA, USA. ⁴⁶Department of Pathology, Yale University School of Medicine, New Haven, CT, USA. ⁴⁷Department of Pathology, IEO, European Institute of Oncology IRCCS & State University of Milan, Milan, Italy. ⁴⁸Laboratory of Pathology, National Cancer Institute, National Institutes of Health, Bethesda, MD, USA. ⁴⁹Ontario Institute for Cancer Research, Toronto, ON, Canada. ⁵⁰Edinburgh Cancer Research Centre, Western General Hospital, Edinburgh, UK. ⁵¹Department of Pathology and Molecular Pathology, Centre Jean Perrin, Clermont-Ferrand, France. ⁵²UMR INSERM 1240, Université Clermont Auvergne, Clermont-Ferrand, France. ⁵³Victorian Comprehensive Cancer Centre building, Peter MacCallum Cancer Centre, Melbourne, Victoria, Australia. ⁵⁴Department of Pathology, National Taiwan University Hospital, Taipei, Taiwan. ⁵⁵German Breast Group, c/o GBG-Forschungs GmbH, Neu-Isenburg, Germany. ⁵⁶Department of Pathology, BC Cancer, Vancouver, British Columbia, Canada. ⁵⁷Peter MacCallum Cancer Centre, Melbourne, Australia. ⁵⁸Department of Pathology, Memorial Sloan Kettering Cancer Center, New York, NY, USA. ⁵⁹Gustave Roussy, Université Paris-Saclay, Villejuif, France. ⁶⁰Université Paris-Sud, Institut National de la Santé et de la Recherche Médicale, Villejuif, France. ⁶¹Division of Molecular Oncology & Immunology, The Netherlands Cancer Institute, Amsterdam, The Netherlands. ⁶²Department of Medical Oncology, The Netherlands Cancer Institute, Amsterdam, The Netherlands. ⁶³University of British Columbia, Vancouver, British Columbia, Canada. ⁶⁴Department of Genomic Medicine, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. ⁶⁵Department of Translational Molecular Pathology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. ⁶⁶Department of Dermatology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. ⁶⁷Department of Pathology, Medical University of Vienna, Vienna, Austria. ⁶⁸GROW - School for Oncology and Developmental Biology, Maastricht University Medical Centre, Maastricht, The Netherlands. ⁶⁹Department of Pathology, Maastricht University Medical Centre, Maastricht, The Netherlands. ⁷⁰Center for Medical Image Science and Visualization, Linköping University, Linköping, Sweden. ⁷¹Department of Biomedical Informatics, Stony Brook University, Stony Brook, NY, USA. ⁷²Roche Diagnostics Information Solutions, Belmont, CA, USA. ⁷³Department of Pathology, GZA-ZNA Ziekenhuizen, Antwerp, Belgium. ⁷⁴Department of Pathology, Northwestern University Feinberg School of Medicine, Chicago, IL, USA. [✉]email: roberto@salgado.be | lee.cooper@northwestern.edu

INTERNATIONAL IMMUNO-ONCOLOGY BIOMARKER WORKING GROUP

Aini Hyttiäinen⁷⁵, Akira I. Hida⁷⁶, Alastair Thompson⁷⁷, Alex Lefevre⁷⁸, Allen Gown⁷⁹, Amy Lo⁸⁰, Anna Sapino⁸¹, Andre Moreira⁸², Andrea Richardson⁸³, Andrea Vingiani⁸⁴, Andrew M. Bellizzi⁸⁵, Andrew Tutt⁸⁶, Angel Guerrero-Zotano⁸⁷, Anita Grigoriadis^{88,89}, Anna Ehinger⁹⁰, Anna C. Garrido-Castro⁹¹, Anne Vincent-Salomon⁹², Anne-Vibeke Laenkholm⁹³, Ashley Cimino-Mathews⁹⁴, Ashok Srinivasan⁹⁵, Balazs Acs⁹⁶, Baljit Singh⁹⁷, Benjamin Calhoun⁹⁸, Benjamin Haibe-Kans⁹⁹, Benjamin Solomon¹⁰⁰, Bibhusal Thapa¹⁰¹, Brad H. Nelson¹⁰², Carlos Castaneda^{103,104}, Carmen Ballesteros-Merino¹⁰⁵, Carmen Criscitiello¹⁰⁶, Carolien Boeckx⁷⁸, Cecile Colpaert¹⁰⁷, Cecily Quinn¹⁰⁸, Chakra S. Chennubhotla¹⁰⁹, Charles Swanton¹¹⁰, Cinzia Solinas¹¹¹, Crispin Hiley¹¹⁰, Damien Drubay^{59,60}, Daniel Bethmann¹¹², Deborah A. Dillon¹¹³, Denis Larsmont¹¹⁴, Dhanusha Sabanathan¹¹⁵, Dieter Peeters¹¹⁶, Dimitrios Zardavas¹¹⁷, Doris Höfmayr¹¹⁸, Douglas B. Johnson¹¹⁹, E. Aubrey Thompson¹²⁰, Edi Brogi⁵⁸, Edith Perez¹²¹, Ehab A. ElGabry¹²², Elizabeth F. Blackley¹⁰⁰, Emily Reisenbichler⁴⁶, Enrique Bellolio^{123,124}, Ewa Chmielik¹²⁵, Fabien Gaire¹²⁶, Fabrice Andre¹²⁷, Fang-I Lu¹²⁸, Farid Azmoudeh-Ardalan¹²⁹, Forbuis Tina Grusso¹³⁰, Franklin Peale¹³¹, Fred R. Hirsch¹³², Frederick Klauschen¹³³, Gabriela Acosta-Haab¹³⁴, Gelareh Farshid¹³⁵, Gert van den Eynden¹³⁶, Giuseppe Curigliano^{137,138}, Giuseppe Floris^{139,140}, Glenn Broeckx¹⁴¹, Harmut Joeppen⁸⁰, Harry R. Haynes¹⁴², Heather McArthur¹⁴³, Heikki Joensuu¹⁴⁴, Helena Olofsson¹⁴⁵, Ian Cree¹⁴⁶, Iris Nederlof¹⁴⁷, Isabel Frahm¹⁴⁸, Iva Brčić¹⁴⁹, Jack Chan¹⁵⁰, Jacqueline A. Hall¹⁵¹, James Ziai⁸⁰, Jane Brock¹⁵², Jelle Wesseling¹⁵³,



12

Jennifer Giltne⁸⁰, Jerome Lemonnier¹⁵⁴, Jiping Zha¹⁵⁵, Joana M. Ribeiro¹⁵⁶, Jodi M. Carter¹⁵⁷, Johannes Hainfellner¹⁵⁸, John Le Quesne¹⁵⁹, Jonathan W. Juco¹⁶⁰, Jorge Reis-Filho^{58,161}, Jose van den Berg¹⁶², Joselyn Sanchez¹⁰⁴, Joseph Sparano¹⁶³, Joël Cucherousset¹⁶⁴, Juan Carlos Araya¹²³, Julien Adam¹⁶⁵, Justin M. Balko¹⁶⁶, Kai Saeger¹⁶⁷, Kalliopi Siziopikou¹⁶⁸, Karen Willard-Gallo¹⁶⁹, Karolina Sikorska¹⁷⁰, Karsten Weber¹⁷¹, Keith E. Steele¹⁵⁵, Kenneth Emancipator¹⁶⁰, Khalid El Bairi¹⁷², Kim R. M. Blenman¹⁷³, Kimberly H. Allison¹⁷⁴, Koen K. van de Vijver¹⁷⁵, Konstany Korsi¹⁷⁶, Lajos Pusztai¹⁷⁷, Laurence Buisseret¹⁶⁹, Leming Shi¹⁷⁷, Liu Shi-wei¹⁷⁸, Luciana Molinero¹³¹, M. Valeria Estrada¹⁷⁹, Maartje van Seijen¹⁸⁰, Magali Lacroix-Triki¹⁸¹, Maggie C. U. Cheang¹⁸², Maise al Bakir¹¹⁰, Marc van de Vijver¹⁸³, Maria Vittoria Dieci¹⁸⁴, Marlon C. Rebelatto¹⁵⁵, Martine Piccart¹⁸⁵, Matthew P. Goetz¹²¹, Matthias Preusser¹⁵⁸, Melinda E. Sanders¹⁸⁶, Meredith M. Regan^{187,188}, Michael Christie¹⁸⁹, Michael Misialek¹⁹⁰, Michail Ignatiadis¹⁹¹, Michiel de Maaker¹⁸⁰, Mieke van Bockstal¹⁹², Miluska Castillo¹⁰⁴, Nadia Harbeck¹⁹³, Nadine Tung¹⁹⁴, Nele Laudus¹⁹⁵, Nicolas Sirtaine¹⁹⁶, Nicole Burchardi¹⁹⁷, Nils Ternes¹⁹⁸, Nina Radosevic-Robin¹⁹⁹, Oleg Gluz²⁰⁰, Oliver Grimm¹²⁶, Paolo Nuciforo²⁰¹, Paul Jank²⁰², Petar Jelinic¹⁶⁰, Peter H. Watson²⁰³, Prudence A. Francis^{13,57}, Prudence A. Russell²⁰⁴, Robert H. Pierce²⁰⁵, Robert Hills²⁰⁶, Roberto Leon-Ferre¹²¹, Roland de Wind¹⁹⁶, Ruohong Shui²⁰⁷, Sabine Declercq²⁰⁸, Sam Leung⁶³, Sami Tabbarah²⁰⁹, Sandra C. Souza²¹⁰, Sandra O'Toole²¹¹, Sandra Swain²¹², Scooter Willis²¹³, Scott Ely²¹⁴, Seong-Rim Kim²¹⁵, Shahinaz Bedri²¹⁶, Sheeba Irshad^{217,218}, Shi-Wei Liu²¹⁹, Shona Hendry²²⁰, Simonetta Bianchi²²¹, Sofia Braganca²²², Soonmyung Paik⁹⁵, Stephen B. Fox²²⁰, Stephen J. Luen¹², Stephen Naber²²³, Sua Luz²²⁴, Susan Fineberg²²⁵, Teresa Soler²²⁶, Thomas Gevaert²²⁷, Timothy d'Alfons⁵⁸, Tom John²²⁸, Tomohagu Sugie²²⁹, Veerle Bossuyt²³⁰, Venkata Manem⁹⁹, Vincente Peg Cámaea²³¹, Weida Tong²³², Wentao Yang²⁰⁷, William T. Tran²⁰⁹, Yihong Wang²³³, Yves Allory²³⁴ and Zahed Husain²³⁵

⁷⁵Department of Oral and Maxillofacial Diseases, Helsinki, Finland. ⁷⁶Department of Pathology, Matsuyama Shimin Hospital, Matsuyama, Japan. ⁷⁷Surgical Oncology, Baylor College of Medicine, Texas, USA. ⁷⁸Roche Diagnostics, Machelen, Belgium. ⁷⁹PhenoPath Laboratories, Seattle, USA. ⁸⁰Research Pathology, Genentech Inc., South San Francisco, USA. ⁸¹Department of Medical Sciences, University of Turin, Italy and Candiolo Cancer Institute - FPO, IRCCS, Candiolo, Italy. ⁸²Pulmonary Pathology, New York University Center for Biospecimen Research and Development, New York University, New York, NY, USA. ⁸³Department of Pathology, Johns Hopkins Hospital, Baltimore, USA. ⁸⁴Department of Pathology, Istituto Europeo di Oncologia, University of Milan, Milan, Italy. ⁸⁵Department of Pathology, University of Iowa Hospitals and Clinics, Iowa City, USA. ⁸⁶Breast Cancer Now Toby Robins Research Centre, The Institute of Cancer Research, London, UK. ⁸⁷Department of Oncology, Instituto Valenciano de Oncologia, Valencia, Spain. ⁸⁸Cancer Bioinformatics Lab, Cancer Centre at Guy's Hospital, London, UK. ⁸⁹School of Life Sciences and Medicine, King's College London, London, UK. ⁹⁰Department of Clinical Genetics and Pathology, Skåne University Hospital, Lund University, Lund, Sweden. ⁹¹Dana-Farber Cancer Institute, Boston, MA, USA. ⁹²Institut Curie, Paris Sciences Lettres Université, Inserm U934, Department of Pathology, Paris, France. ⁹³Department of Surgical Pathology Zealand University Hospital, Køge, Denmark. ⁹⁴Departments of Pathology and Oncology, The Johns Hopkins Hospital, Baltimore, USA. ⁹⁵National Surgical Adjuvant Breast and Bowel Project Operations Center/NRG Oncology, Pittsburgh, PA, USA. ⁹⁶Department of Pathology, Karolinska Institute, Solna, Sweden. ⁹⁷Department of Pathology, New York University Langone Medical Center, New York, USA. ⁹⁸Department of Pathology and Laboratory Medicine, UNC School of Medicine, Columbia, USA. ⁹⁹Québec Heart and Lung Institute Research Center, Laval University, Quebec city, Quebec, Canada. ¹⁰⁰Department of Medical Oncology, Peter MacCallum Cancer Centre, Melbourne, VIC, Australia. ¹⁰¹Department of Medicine, University of Melbourne, Parkville, Australia. ¹⁰²Trev & Joyce Deeley Research Centre, British Columbia Cancer Agency, Victoria, Canada. ¹⁰³Department of Medical Oncology, Instituto Nacional de Enfermedades Neoplásicas, Lima, Perú. ¹⁰⁴Department of Research, Instituto Nacional de Enfermedades Neoplásicas, Lima 15038, Peru. ¹⁰⁵Providence Cancer Research Center, Portland, Oregon, USA. ¹⁰⁶Department of Medical Oncology, Istituto Europeo di Oncologia, Milan, Italy. ¹⁰⁷Department of Pathology, AZ Turnhout, Turnhout, Belgium. ¹⁰⁸Department of Pathology, St Vincent's University Hospital and University College Dublin, Dublin, Ireland. ¹⁰⁹Department of Computational and Systems Biology, University of Pittsburgh, Pittsburgh, USA. ¹¹⁰Cancer Research UK Lung Cancer Centre of Excellence, University College London Cancer Institute, University College London, London, UK. ¹¹¹Azienda AUSL, Regional Hospital of Aosta, Aosta, Italy. ¹¹²University Hospital Halle (Saale), Institute of Pathology, Halle, Saale, Germany. ¹¹³Department of Pathology, Brigham and Women's Hospital, Boston, MA. ¹¹⁴Department of Pathology, Dana Farber Cancer Institute, Boston, MA, USA. ¹¹⁵Department of Pathology, Jules Bordet Institute, Brussels, Belgium. ¹¹⁶Department of Clinical Medicine, Macquarie University, Sydney, Australia. ¹¹⁷HistoGeneX NV, Antwerp, Belgium and AZ Sint-Maarten Hospital, Machelen, Belgium. ¹¹⁸Oncology Clinical Development, Bristol-Myers Squibb, Princeton, USA. ¹¹⁹Institut für Pathologie, UK Hamburg, Hamburg, Germany. ¹²⁰Department of Medicine, Vanderbilt University Medical Centre, Nashville, USA. ¹²¹Department of Cancer Biology, Mayo Clinic, Jacksonville, USA. ¹²²Department of Oncology, Mayo Clinic, Rochester, USA. ¹²³Roche, Tucson, USA. ¹²⁴Department of Pathology, Universidad de La Frontera, Temuco, Chile. ¹²⁵Tumor Pathology Department, Maria Skłodowska-Curie Memorial Cancer Center, Gliwice, Poland. ¹²⁶Pathology and Tissue Analytics, Roche, Machelen, Belgium. ¹²⁷Department of Medical Oncology, Gustave Roussy, Villejuif, France. ¹²⁸Sunnybrook Health Sciences Centre, Toronto, Canada. ¹²⁹Tehran University of Medical Sciences, Tehran, Iran. ¹³⁰Translational Research, Montreal, Canada. ¹³¹Oncology Biomarker Development, Genentech-Roche, Machelen, Belgium. ¹³²Division of Medical Oncology, Department of Medicine, University of Colorado Anschutz Medical Campus, Aurora, USA. ¹³³Institute of Pathology, Charité Universitätsmedizin Berlin, Berlin, Germany. ¹³⁴Department of Pathology, Hospital de Oncología María Curie, Buenos Aires, Argentina. ¹³⁵Directorate of Surgical Pathology, SA Pathology, Adelaide, Australia. ¹³⁶Department of Pathology, GZA-ZNA Hospitals, Wilrijk, Belgium. ¹³⁷University of Milano, Istituto Europeo di Oncologia, IRCCS, Milano, Italy. ¹³⁸Division of Early Drug Development for Innovative Therapy, IEO, European Institute of Oncology IRCCS, Milan, Italy. ¹³⁹Department of Imaging and Pathology, Laboratory of Translational Cell & Tissue Research, Leuven, Belgium. ¹⁴⁰KU Leuven- University Hospitals Leuven, Department of Pathology, Leuven, Belgium. ¹⁴¹Department of Pathology, University Hospital Antwerp, Antwerp, Belgium. ¹⁴²Translational Health Sciences, Department of Cellular Pathology, North Bristol NHS Trust, University of Bristol, Bristol, UK. ¹⁴³Medical Oncology, Department of Medicine, Cedars-Sinai Medical Center, Los Angeles, USA. ¹⁴⁴Helsinki University Central Hospital, Helsinki, Finland. ¹⁴⁵Department of Clinical Pathology, Akademiska University Hospital, Uppsala, Sweden. ¹⁴⁶International Agency for Research on Cancer (IARC), World Health Organization, Lyon, France. ¹⁴⁷Division of Tumor Biology & Immunology, The Netherlands Cancer Institute, Amsterdam, The Netherlands. ¹⁴⁸Department of Pathology, Sanatorio Mater Dei, Buenos Aires, Argentina. ¹⁴⁹Institute of Pathology, Medical University of Graz, Graz, Austria. ¹⁵⁰Department of Oncology, National Cancer Centre, Singapore, Singapore. ¹⁵¹Vivactiv Ltd, Bellingdon, Bucks, UK. ¹⁵²Department of Pathology, Brigham and Women's Hospital, Boston, USA. ¹⁵³Department of Pathology, Netherlands Cancer Institute, Amsterdam, The Netherlands. ¹⁵⁴R&D UNICANCER, Paris, France. ¹⁵⁵Translational Sciences, MedImmune, Gaithersburg, USA. ¹⁵⁶Breast Unit, Champalimaud Clinical Centre, Lisboa, Portugal. ¹⁵⁷Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, USA. ¹⁵⁸Department of Medicine, Clinical Division of Oncology, Comprehensive Cancer Centre Vienna, Medical University of Vienna, Vienna, Austria. ¹⁵⁹Leicester Cancer Research Centre, University of Leicester, Leicester, and MRC Toxicology Unit, University of Cambridge, Cambridge, UK. ¹⁶⁰Merck & Co., Inc, Kenilworth, NJ, USA. ¹⁶¹Human Oncology and Pathogenesis Program, Memorial Sloan Kettering Cancer Center, New York, NY, USA. ¹⁶²Department of Pathology, The Netherlands Cancer Institute, Amsterdam, The Netherlands. ¹⁶³Department of Medicine, Department of Obstetrics & Gynecology and Women's Health, Albert Einstein Medical Center, Bronx, USA. ¹⁶⁴GHI Le Raincy-Montfermeil, Chelles, Île-de-France, France. ¹⁶⁵Department of Pathology, Gustave Roussy, Grand Paris, France. ¹⁶⁶Departments of Medicine and Cancer Biology, Vanderbilt University Medical Centre, Nashville, USA. ¹⁶⁷Vm Scope, Berlin, Germany. ¹⁶⁸Department of Pathology, Breast Pathology Section, Northwestern University, Chicago, USA. ¹⁶⁹Molecular Immunology Unit, Institut Jules Bordet, Université Libre de Bruxelles, Brussels, Belgium. ¹⁷⁰Department of Biometrics, The Netherlands Cancer Institute, Amsterdam, The Netherlands. ¹⁷¹German Breast Group, Neu-Isenburg, Germany. ¹⁷²Cancer Biomarkers Working Group, Faculty of Medicine and Pharmacy, Université Mohamed Premier, Oujda, Morocco. ¹⁷³Yale Cancer Center Genetics, Genomics and Epigenetics Program, Yale School of Medicine, New Haven, CT, USA. ¹⁷⁴Pathology Department, Stanford University Medical Centre, Stanford, USA. ¹⁷⁵Department of Pathology, University Hospital Ghent, Ghent, Belgium. ¹⁷⁶Pathology and Tissue Analytics, Roche Innovation Centre Munich, Penzberg, Germany. ¹⁷⁷Center for Pharmacogenomics and Fudan-Zhangjiang, Center for Clinical Genomics School of Life Sciences and Shanghai Cancer Center, Fudan University, Fudan, China. ¹⁷⁸Sichuan Cancer Hospital, Chengdu, China. ¹⁷⁹Biorepository and Tissue Technology Shared Resources, University of California San Diego, San Diego, USA. ¹⁸⁰Division of Molecular Pathology, The Netherlands Cancer Institute, Amsterdam, The Netherlands. ¹⁸¹Department of Pathology, Gustave Roussy, Villejuif, France. ¹⁸²Institute of Cancer Research Clinical Trials and Statistics Unit, The Institute of Cancer Research, Surrey, UK. ¹⁸³Department of Pathology, Academic Medical Center, Amsterdam, The Netherlands. ¹⁸⁴Department of Surgery, Oncology and Gastroenterology, University of

Padova, Padua, Italy. ¹⁸⁵Institut Jules Bordet, Université Libre de Bruxelles, Brussels, Belgium. ¹⁸⁶Department of Pathology, Microbiology and Immunology, Vanderbilt University Medical Centre, Nashville, USA. ¹⁸⁷Harvard Medical School, Boston, USA. ¹⁸⁸Division of Biostatistics, Dana-Farber Cancer Institute, Boston, USA. ¹⁸⁹Department of Anatomical Pathology, Royal Melbourne Hospital, Parkville, Australia. ¹⁹⁰Vernon Cancer Center, Newton-Wellesley Hospital, Newton, USA. ¹⁹¹Department of Medical Oncology, Institut Jules Bordet, Université Libre de Bruxelles, Brussels, Belgium. ¹⁹²Department of Pathology, Cliniques universitaires Saint-Luc, Brussels, Belgium. ¹⁹³Breast Center, Dept. OB&GYN and CCC (LMU), University of Munich, Munich, Germany. ¹⁹⁴Division of Hematology-Oncology, Beth Israel Deaconess Medical Center, Boston, USA. ¹⁹⁵University of Leuven, Leuven, Belgium. ¹⁹⁶Department of Pathology, Institut Jules Bordet, Université Libre de Bruxelles, Brussels, Belgium. ¹⁹⁷German Breast Group GmbH, Neu-Isenburg, Germany. ¹⁹⁸Service de Biostatistique et d'Epidémiologie, Gustave Roussy, CESP, Université-Paris Sud, Université Paris-Saclay, Villejuif, France. ¹⁹⁹Department of Surgical Pathology and Biopathology, Jean Perrin Comprehensive Cancer Centre, Clermont-Ferrand, France. ²⁰⁰Johanniter GmbH - Evangelisches Krankenhaus Bethesda Mönchengladbach, West German Study Group, Mönchengladbach, Germany. ²⁰¹Molecular Oncology Group, Vall d'Hebron Institute of Oncology, Barcelona, Spain. ²⁰²Department of Pathology, University of Marburg, Marburg, Germany. ²⁰³Department of Pathology and Laboratory Medicine, University of British Columbia, Columbia, USA. ²⁰⁴Department of Anatomical Pathology, St Vincent's Hospital Melbourne, Fitzroy, Australia. ²⁰⁵Cancer Immunotherapy Trials Network, Central Laboratory and Program in Immunology, Fred Hutchinson Cancer Research Center, Seattle, USA. ²⁰⁶Clinical Trial Service Unit & Epidemiological Studies Unit, University of Oxford, Oxford, UK. ²⁰⁷Department of Pathology, Fudan University Shanghai Cancer Center, Shanghai, China. ²⁰⁸Department of Pathology, GZA-ZNA Hospitals, Antwerp, Belgium. ²⁰⁹Department of Radiation Oncology, Odette Cancer Centre, Sunnybrook Research Institute, Toronto, Canada. ²¹⁰Oncology Merck & Co, New Jersey, USA. ²¹¹The Cancer Research Program, Garvan Institute of Medical Research, Darlinghurst, Australian Clinical Labs, Darlinghurst, Australia. ²¹²Georgetown University Medical Center, Washington DC, USA. ²¹³Department of Molecular and Experimental Medicine, Avera Cancer Institute, Sioux Falls, SD, USA. ²¹⁴Translational Medicine, Bristol-Myers Squibb, Princeton, USA. ²¹⁵National Surgical Adjuvant Breast and Bowel Project Operations Center/NRG Oncology, Pittsburgh, USA. ²¹⁶Anatomic Pathology, Boston, MA, USA. ²¹⁷King's College London, London, UK. ²¹⁸Guy's Hospital, London, UK. ²¹⁹Peking University First Hospital Breast Disease Center, Beijing, China. ²²⁰Department of Pathology, Peter MacCallum Cancer Centre, Melbourne, Australia. ²²¹Dipartimento di Scienze della Salute (DSS), Firenze, Italy. ²²²Department of Oncology, Champalimaud Clinical Centre, Lisbon, Portugal. ²²³Department of Pathology and Laboratory Medicine, Tufts Medical Center, Boston, USA. ²²⁴Department of Pathology, Fundación Valle del Lili, Cali, Colombia. ²²⁵Department of Pathology, Montefiore Medical Center and the Albert Einstein College of Medicine, Bronx, NY, USA. ²²⁶Department of Pathology, University Hospital of Bellvitge, Oncobell, IDIBELL, L'Hospitalet del Llobregat, Barcelona 08908 Catalonia, Spain. ²²⁷Department of Development and Regeneration, Laboratory of Experimental Urology, KU Leuven, Leuven, Belgium. ²²⁸Department of Medical Oncology, Austin Health, Heidelberg, Australia. ²²⁹Department of Surgery, Kansai Medical University to Tomoharu Sugie, Breast Surgery, Kansai Medical University Hospital, Hirakata, Japan. ²³⁰Department of Pathology, Massachusetts General Hospital, Boston, USA. ²³¹Pathology Department, H.U. Vall d'Hebron, Barcelona, Spain. ²³²Division of Bioinformatics and Biostatistics, U.S. Food and Drug Administration, Wuhan, USA. ²³³Department of Pathology and Laboratory Medicine, Rhode Island Hospital and Lifespan Medical Center, Providence, USA. ²³⁴Université Paris-Est, Créteil, France. ²³⁵Praava Health, Dhaka, Bangladesh.

Section 3.2

Explainable nucleus classification using Decision Tree Approximation of Learned Embeddings

This section is an exact authorised reproduction of the following journal paper:

Amgad, M., Atteya, L., Hussein, H., Mohammed, K. H., Hafiz, E., Elsebaie, M. A., Mobadersany, P., Manthey, D., Gutman, D. A., Elfandy, H., et al. (2021). Explainable nucleus classification using decision tree approximation of learned embeddings. Bioinformatics.

The open-access code base is available at this [Github repository](#).

An abstract version was also presented at the 2021 Pathology Visions Conference (Las Vegas, NV).

Bioinformatics, 2021, 1–7

doi: 10.1093/bioinformatics/btab670

Advance Access Publication Date: 29 September 2021

Original Paper

OXFORD

Bioimage informatics

Explainable nucleus classification using Decision Tree Approximation of Learned Embeddings

Mohamed Amgad ¹, Lamees A. Atteya^{2,†}, Hagar Hussein^{3,†},
Kareem Hosny Mohammed^{4,†}, Ehab Hafiz ^{5,†}, Maha A. T. Elsebaie^{6,†},
Pooya Mobadersany¹, David Manthey⁷, David A. Gutman⁸, Habiba Elfandy⁹ and
Lee A. D. Cooper^{1,*}

¹Department of Pathology, Northwestern University, Chicago, IL, USA, ²Egyptian Ministry of Health, Cairo, Egypt, ³Department of Pathology, Nasser Institute for Research and Treatment, Cairo, Egypt, ⁴Department of Pathology and Laboratory Medicine, University of Pennsylvania, Philadelphia, PA, USA, ⁵Department of Clinical Laboratory Research, Theodor Bilharz Research Institute, Giza, Egypt, ⁶Department of Medicine, Cook County Hospital, Chicago, IL, USA, ⁷Kitware Inc., Clifton Park, NY, USA, ⁸Department of Neurology, Emory University, Atlanta, GA, USA and ⁹Department of Pathology, National Cancer Institute, Cairo, Egypt

*To whom correspondence should be addressed.

†The authors wish it to be known that these authors contributed equally.

Associate Editor: Jinbo Xu

Received on May 8, 2021; revised on August 5, 2021; editorial decision on September 19, 2021; accepted on September 23, 2021

Abstract

Motivation: Nucleus detection, segmentation and classification are fundamental to high-resolution mapping of the tumor microenvironment using whole-slide histopathology images. The growing interest in leveraging the power of deep learning to achieve state-of-the-art performance often comes at the cost of explainability, yet there is general consensus that explainability is critical for trustworthiness and widespread clinical adoption. Unfortunately, current explainability paradigms that rely on pixel saliency heatmaps or superpixel importance scores are not well-suited for nucleus classification. Techniques like Grad-CAM or LIME provide explanations that are indirect, qualitative and/or nonintuitive to pathologists.

Results: In this article, we present techniques to enable scalable nuclear detection, segmentation and explainable classification. First, we show how modifications to the widely used Mask R-CNN architecture, including decoupling the detection and classification tasks, improves accuracy and enables learning from hybrid annotation datasets like NuCLS, which contain mixtures of bounding boxes and segmentation boundaries. Second, we introduce an explainability method called Decision Tree Approximation of Learned Embeddings (DTALE), which provides explanations for classification model behavior globally, as well as for individual nuclear predictions. DTALE explanations are simple, quantitative, and can flexibly use any measurable morphological features that make sense to practicing pathologists, without sacrificing model accuracy. Together, these techniques present a step toward realizing the promise of computational pathology in computer-aided diagnosis and discovery of morphologic biomarkers.

Availability and implementation: Relevant code can be found at github.com/CancerDataScience/NuCLS

Contact: lee.cooper@northwestern.edu

Supplementary information: [Supplementary data](#) are available at *Bioinformatics* online.

1 Introduction

Nucleus classification, localization and segmentation (NuCLS) are fundamental pattern recognition tasks commonly performed in computational pathology workflows (Xing and Yang, 2016). Nuclear identification and morphologic assessment are integral to most histopathology diagnostic and clinical grading schemes, and are

used for determining how aggressive certain malignancies are, and whether the patient is likely to respond to certain therapeutics. By extension, computational assessment of nuclei is important for computer-aided diagnosis and patient prognostication (Abels *et al.*, 2019). Moreover, nuclear segmentation and/or extraction of nuclear morphometric and spatial features is the first and most important step in exploratory research to discover genomic and clinical

correlates of quantitative morphologic features (Cooper et al., 2012; Diao et al., 2021; Saltz et al., 2018). Computational pathology most commonly make use of digitized histopathology slides known as whole-slide images (WSIs). A number of unique challenges contribute to the difficulty of translating traditional image processing and machine learning algorithms to histopathology contexts, including the extreme sizes of WSIs, typically $80k \times 80k$ pixels, and high variability in image quality and appearance due to differences in tissue processing, staining and slide scanning equipment and protocols (Amgad et al., 2020; Pantanowitz et al., 2013). In situations where the data variability is high, machine learning algorithms typically need a large number of examples to capture the full spectrum of cases that would be seen after deployment. This variability stems in part from preanalytical factors such as specimen preparation and staining protocols, slide scanner specifications, image formats and compression, etc. (Masucci et al., 2016; Pantanowitz et al., 2013). Unfortunately, the lack of publicly available datasets limits the development and benchmarking of deep-learning models, due to: (i) logistical and legal difficulties of health data sharing and (ii) time constraints of practicing pathologists whose expertise is required to produce ground truth data (Abels et al., 2019; Amgad et al., 2020; Hartman et al., 2020).

In previous work, we developed a crowdsourcing approach that scales the acquisition of nucleus segmentation and classification data and produces hybrid datasets containing both bounding boxes and segmentation data (Fig. 1) (Amgad et al., 2021). This assisted labeling protocol relied on a decentralized web-based annotation platform, HistomicsUI, and involved asking the users to click on accurate annotation suggestions generated by weak segmentation and classification algorithms, and to place bounding boxes around missing or inaccurately segmented nuclei (Gutman et al., 2017). The weak algorithm used to produce the annotation suggestions uses simple image processing operations and therefore has no reliance on training data. This procedure was used to generate the 220,000 annotations that comprise the NuCLS datasets, and motivates the development and/or adaptation of deep-learning approaches to handle hybrid ground truth data. More generally, as we discuss later, strategies are needed for mitigating systematic differences between typical object detection in natural images and nucleus detection and classification (Fig. 2). It should be noted that machine learning using hybrid datasets is a combination of object detection and segmentation. Hence, for consistency, we use the term ‘detection’ throughout this article whenever segmentations are not necessarily needed for the task being discussed.

Besides achieving high accuracy, deep-learning models for clinical applications are most useful when they are explainable (Fig. 3). Not only does explainability increase confidence in model decisions and hence the likelihood of clinical adoption but it also helps guard against catastrophic failures and spurious correlations (Amgad et al., 2020; D’Amour et al., 2020). This emphasis on explainability



Fig. 1. Example hybrid bounding box and segmentation data. Hybrid annotation datasets combine bounding boxes generated by humans with segmentations and classifications generated by a weak algorithm. They can be generated more scalably and require less effort from annotators, but require new algorithms that can learn from a mixture of boxes and segmentation boundaries. Segmentations enable the computation of morphologic features to discover biological associations and can provide valuable explanations of model inference

is being increasingly recognized by the deep-learning community, and a number of algorithms have been devised to explain model decisions in image classification and natural language processing contexts (Samek et al., 2021).

Unfortunately, the literature is sparse on explainability techniques for object detection and classification, especially in the context of nucleus classification. For consistency, the lexicon we are using here is derived from Rudin (2018) and Marcinkevics and Vogt (2020). These authors distinguish between interpretable models and explanation methods. Interpretable models are transparent—think of the weights of a linear model or criteria from a decision tree, and are commonly provided by modeling approaches that are inherently simple. This of course comes at a price, since model simplicity can result in underfitting and lack of generalizability (Hastie et al., 2017). Transparency is more difficult for complex models like neural networks, although some research attempts to tackle this challenge. In contrast, explanation methods are surrogate post-hoc techniques that demystify the black-box model decisions. Explanations could be global, explaining the full range of decisions a model can produce, or sample-specific, explaining the inference performed for a single sample (Marcinkevics and Vogt, 2020).

A very popular set of explanation techniques, including variants of Grad-CAM, rely on using gradient backpropagation to estimate pixel saliency (Selvaraju et al., 2017). These methods produce a heatmap that, when overlaid over the input image, can give an idea about where the model is ‘looking’ during inference. This approach, while an important advance, has two problems. First, the heatmaps produced tend to be quite blurry and do not follow natural boundaries. In fact, heatmaps only tell us whether certain pixels are important for classification, not how they are used to distinguish between alternative classification decisions. Second, there are concerns over the misuse of this explainability approach, particularly its qualitative nature and lack of falsifiability (Leavitt and Morcos, 2020; Rudin, 2018). Falsifiability is the ability of a hypothesis to be disproven, and is a fundamental guardrail against confirmation bias (Popper, 1959). When using saliency heatmaps for, say, a dog versus wolf classifier what could a wrong answer possibly be? Not clear. More recently, a technique called Local Interpretable Model-agnostic Explanations (LIME) has gained popularity for its simplicity and general-purpose nature (Ribeiro et al., 2016). LIME relies on decomposition of the input into interpretable components, superpixels in the imaging context, which are repeatedly perturbed. The predicted classification probability then is used to identify the most important superpixel, and hence provides clear boundaries that cannot be obtained using Grad-CAM. While more quantitative than Grad-CAM, LIME is not directly applicable in our context because superpixels cannot account for discrete object morphometric measurements like size, shape and texture.

In this article, we make two contributions toward nucleus segmentation and explainable classification using hybrid box and segmentation annotation data. First, we systematically examine modifications to Mask R-CNN, the state-of-the-art object detection model, to optimize for the specific task of nucleus detection and to learn how to segment from hybrid annotation datasets (He et al., 2017). Second, we describe an explainability technique we call Decision Tree Approximation of Learned Embeddings (DTALE) that provides falsifiable, quantitative and intuitive explanations of decisions by nucleus detection and classification models. We believe these contributions will enable the development of scalable systems for mapping the tumor microenvironment, with implications in computer-aided diagnostics and biomarker discovery.

2 Materials and methods

2.1 Training and validation data

The NuCLS datasets were used for training and validating the NuCLS model, our Mask R-CNN variant (Amgad et al., 2021). The scans come from hematoxylin and eosin stain, formalin-fixed paraffin embedded slides from 144 breast cancer patients from The Cancer Genome Atlas. These NuCLS datasets contain 220 000

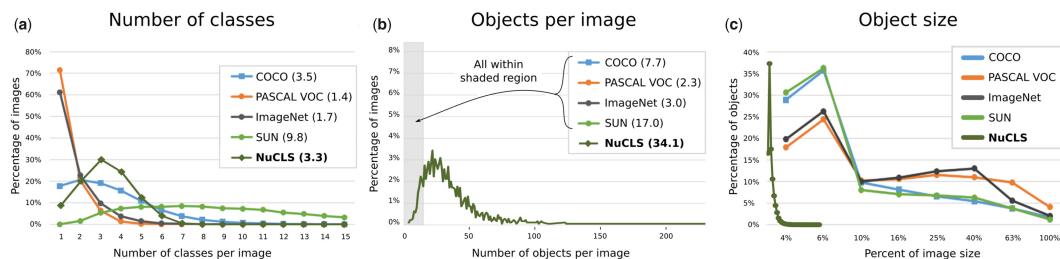


Fig. 2. Comparison of the NuCLS dataset with canonical ‘natural’ object detection datasets. Nucleus detection datasets typically contain objects that are much smaller and more densely packed than imaging datasets of natural or day-to-day scenery. NuCLS images are $\sim 380 \times 380$ pixel patches at 0.2 microns-per-pixel resolution, and contain on average 34 nuclei, each of which filling only $\sim 1\%$ of the image area. These systemic differences motivate the adaptation of existing methods like Mask R-CNN to accommodate numerous small objects and to revisit some of the assumptions about object morphology that do not apply in the context of nucleus detection. Modified with permission from Lin *et al.* (2014)

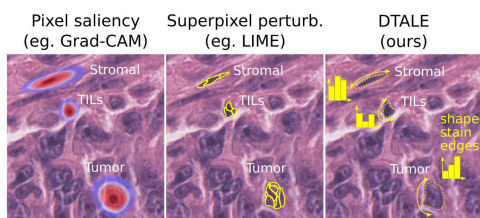


Fig. 3. DTALE provides falsifiable, meaningful and quantitative explanations of nucleus detection model decisions. Unlike other approaches, DTALE can provide explanations that reference object-level morphological measurements such as nuclear size, shape, staining intensity, chromatin texture, perinuclear cytoplasmic staining, etc. In fact, DTALE can use any set of measurable features that make sense to a pathologist to provide quantitative decision tree approximations for black-box classification model decisions. These explanations include global decision criteria, e.g. ‘tumor nuclei are large and have irregular shapes’, as well as decision criteria for individual nuclei of interest

annotations of nucleus segmentation and classification. For this article, we use the following dataset subsets: corrected single-rater datasets, which were annotated by nonpathologists and corrected and approved by pathologists, and multirater evaluation dataset annotated by multiple pathologists. The NuCLS datasets contain three nucleus superclasses (tumor, stroma and TILs), each of which is subdivided into two granular subclasses. The annotation data were found to be reliable for superclasses, but less so for the granular subclasses.

2.2 NuCLS model

Our NuCLS model modifies the Pytorch implementation of the Mask R-CNN architecture (He *et al.*, 2017), as illustrated in Figure 4. Further details can be found in the Supplementary Methods (He *et al.*, 2016, 2017; Kuhn, 1955; Macenko *et al.*, 2009; Tellez *et al.*, 2018, 2019).

2.3 DTALE

DTALE relies on the fact that Mask R-CNN (and by extension, our NuCLS model) learns to predict object segmentation boundaries as well as their classifications (He *et al.*, 2017). The DTALE procedure involves four steps (Fig. 6): (i) learning embeddings, (ii) generating interpretable features, (iii) fitting the decision tree and (iv) calculating node statistics.

2.3.1 Learned embeddings

Starting with a trained NuCLS model, we extracted the terminal, per-nucleus, 1024-dimensional classification feature vectors (just before the logits). Hyperbolic UMAP was applied to these features to generate a two-dimensional (2D) embedding (McInnes *et al.*, 2018).

2.3.2 Interpretable features

The same FOVs that were input into the NuCLS model were processed to enable extraction of interpretable features. Macenko stain unmixing was used to separate the hematoxylin channel (Macenko *et al.*, 2009). Both the hematoxylin intensity channel and the segmentation mask predictions from the NuCLS model were input into the HistomicsTK function `compute_nuclei_features`, which uses image processing operations to extract feature vectors encoding 62 morphologic features describing shape, intensity, edges and texture (Supplementary Table S5) (Haralick *et al.*, 1973; Kokoska and Zwillinger, 2000; Zhang *et al.*, 2001).

2.3.3 Regression decision tree

A regression decision tree was fitted to produce predictions in the embedding space using the interpretable features as inputs (Hastie *et al.*, 2017). This maps the interpretable features directly into the 2D embedding space to connect morphology with NuCLS model behavior. The rationale for using a regression tree, as opposed to a classification tree, is twofold. First, any accurate classification model will produce similar classification decisions. In contrast, the 2D embedding is a compressed version of a 1024-feature space that is highly specific to our trained NuCLS model. Second, using a regression tree allows us to produce fine-grained *within-class* explanations for individual nuclei (see Fig. 6). This technique is broadly similar to some existing works that use soft decision trees to approximate deep-learning model behavior (Dahlin *et al.*, 2020). We constrained the tree to a maximum depth of 7 and a minimum of 250 nuclei per leaf.

2.3.4 Node fit statistics

Once the DTALE tree was fitted, we traversed nodes to find paths that best represented NuCLS class predictions. For each classification class C_i , and for each tree node N_j , we calculate precision, recall and F1 scores for the downstream subtree as if all nuclei were classified as C_i and using actual NuCLS model classifications as ground truth. This generates an F1 and precision score for each node/class pair. For each class, we identify the node with the highest F1 score as the most representative of NuCLS model predictions for that class, whereas the highest precision node corresponds to interpretable features that are the most discriminative.

3 Results and discussion

3.1 NuCLS: a Mask R-CNN variant using hybrid datasets

Nucleus detection differs from natural object detection tasks in several important respects. Nuclei have lower variability in size and coarse morphology than objects in natural scenes, and different nucleus classes are mostly distinguished by fine detail and spatial context. Models designed for detection in natural images, including Mask R-CNN, produce inferences that integrate the concepts of

4

M.Amgad et al.

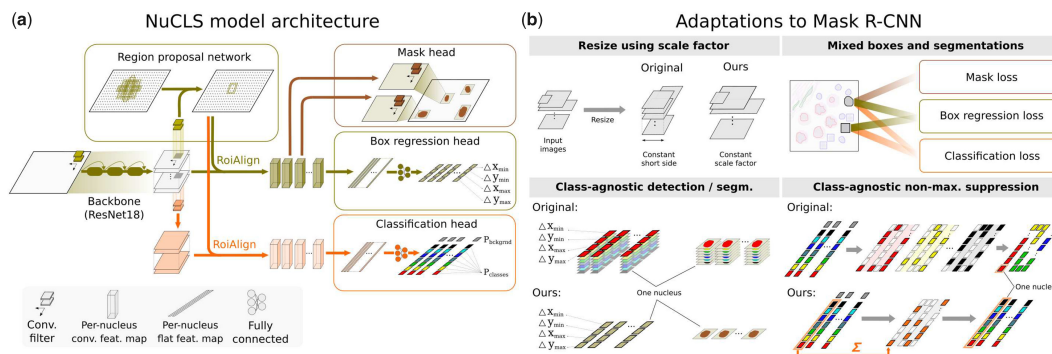


Fig. 4. NuCLS model architecture. (a) The Mask R-CNN architecture was adapted for nucleus detection and classification, allowing some independence of the classification and detection tasks, which improves performance. (b) Other adaptations we made include: (i) supporting variable-size images at inference while preserving scale and aspect ratio; (ii) supporting hybrid training data that mixes bounding boxes and segmentations; (iii) simplifying object detection and (iv) generating full class probability vectors for each nucleus at inference

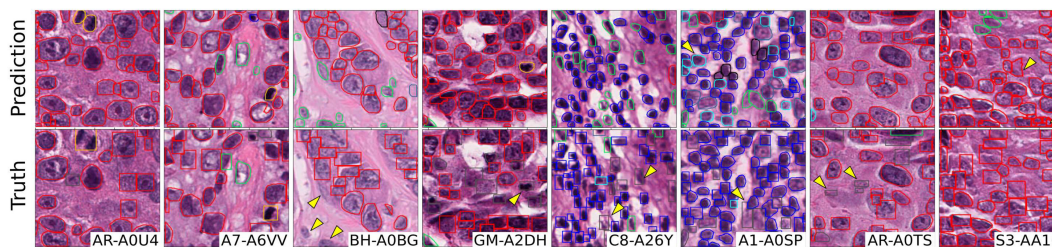


Fig. 5. Qualitative performance of NuCLS model on testing sets. The displayed ground truth comes from the pathologist-corrected single-rater dataset. The images are representative of a number of different hospitals in each of the testing sets. Detection and classification performance closely matches the ground truth, and discrepancies are marked by arrows. Not all discrepancies are algorithmic errors, including: (i) adjacent nuclei that could conceivably be viewed as a single nucleus; (ii) missing annotations and (iii) morphologically ambiguous nuclei

detection and classification (e.g. person, 82% probability) (He *et al.*, 2017). In contrast, for the purpose of detection, nuclei belong to a single metaclass with an ovoid morphology. Treating nuclei as a single metaclass allows calculation of a full classification probability vector for each nucleus, which would be useful where nuclear morphology is ambiguous, especially in computer-assisted diagnostic settings. Nuclei are also typically much smaller and more numerous than natural objects, even at high magnification, which makes accurate detection more challenging (Fig. 2) (Lin *et al.*, 2014). Moreover, scalable deployment of trained nucleus detection models requires the flexibility to perform inference for very large images without resizing and distorting nuclei (Chandrasevan *et al.*, 2020; Yousefi and Nie, 2019).

We modified Mask R-CNN for the specific task of nucleus detection and to handle the hybrid annotations generated by our assisted annotation method, as well as pure segmentation data (Fig. 4). We call our architecture the NuCLS model, for consistency with the NuCLS annotation datasets used for training and validation (Amgad *et al.*, 2021). The pathologist-corrected *single-rater dataset* was used for training and validation. The *multi-rater evaluation dataset* was used for additional validation, although it should be noted that the single-rater dataset contains many more unique fields of view (FOVs) compared to the multirater dataset (1744 versus 52 FOVs). Our key modifications included increased independence of the jointly trained detector and classifier, and enabled: (i) training with hybrid box/segmentation annotations; (ii) generating class probability vectors for all detections; (iii) inference with variable input image sizes without distortion of scale or aspect ratios. To account for the scale and density of nuclei, we also made the following changes to

improve detection performance: (i) increasing the density of region proposals relative to natural image datasets and (ii) digitally increasing magnification beyond 40 \times objective (Supplementary Table S1). Since detection and classification have disparate clinical utility, we report their accuracies separately. We also trained a baseline Mask R-CNN model (with discounting of segmentations from mask loss), and show that that achieves a lower performance (Supplementary Table S2).

We used an internal-external cross-validation scheme to assess the generalization performance of our trained models (Supplementary Fig. S1). This separates training and testing data by hospital rather than image to better reflect the challenge of external generalization (Amgad *et al.*, 2020; Steyerberg and Harrell, 2016). NuCLS models were trained on the single-rater dataset, and reached convergence within 40 epochs (Supplementary Fig. S2). They converged smoothly despite being trained using a mixture of box and segmentation annotations. Trained NuCLS models had high generalization accuracy for detection (AP = 74.8 \pm 0.5), segmentation (DICE = 88.5 \pm 0.8) and superclass classification (AUROC = 93.5 \pm 2.7) (Table 1 and Supplementary Table S3). For classification of sTILs (stromal tumor-infiltrating lymphocytes), a clinically salient problem, NuCLS models had a testing AUROC of 94.7 \pm 2.1 (Supplementary Table S4) (Amgad *et al.*, 2020). This was also reflected on qualitative examination of predictions (Fig. 5 and Supplementary Fig. S3).

The performance of NuCLS models was consistent with limitations of the training data. Accuracy was lower for classes with higher interrater variability (e.g. plasma cells) or for classes where nonpathologists were not reliable annotators (mitotic figures and macrophages) (Supplementary Fig. S4 and Fig. 6b and g).

Table 1. Generalization accuracy of NuCLS models trained and evaluated on the corrected single-rater dataset using internal-external cross-validation

Fold	Detection					Segmentation					Classification							
	N	AP@.5	mAP@.5-.95	N	Median IOU	Median DICE	N	Superclasses?	Accuracy	MCC	AUROC (micro)	AUROC (macro)	N	Superclasses?	Accuracy	MCC	AUROC (micro)	AUROC (macro)
1 (Val.)	6102	75.3	34.4	1389	78.5	87.9	5351	No	71.0	58.1	93.3	84.6	5351	No	71.0	58.1	93.3	84.6
2	15442	74.9	33.2	3474	78.0	87.6	13597	Yes	77.5	65.2	93.7	89.0	13597	Yes	77.5	65.2	93.7	89.0
3	12672	74.0	33.8	1681	80.2	89.0	11176	No	70.1	56.9	93.8	83.6	11176	No	70.1	56.9	93.8	83.6
4	8260	75.3	33.5	1948	80.9	89.5	7288	Yes	79.4	68.2	94.6	86.5	7288	Yes	79.4	68.2	94.6	86.5
5	7295	74.9	31.5	1306	78.1	87.7	6294	No	79.0	68.6	93.5	87.1	6294	No	79.0	68.6	93.5	87.1
Mean (Std)	—	74.8 (0.5)	33.0 (0.9)	—	79.3 (1.3)	88.5 (0.8)	—	Yes	73.1	61.8	94.4	89.4	—	Yes	73.1	61.8	94.4	89.4
								No	68.4 (4.2)	55.7 (5.4)	92.8 (2.0)	83.7 (2.9)		No	68.4 (4.2)	55.7 (5.4)	92.8 (2.0)	83.7 (2.9)
								Yes	77.7 (5.7)	65.6 (7.9)	93.5 (2.7)	86.0 (3.2)		Yes	77.7 (5.7)	65.6 (7.9)	93.5 (2.7)	86.0 (3.2)

Note: All accuracy values are percentages. Fold 1 acted as the validation set for hyperparameter tuning, so the bottom row shows mean and standard deviation of four values (folds 2-5). Note that the number of testing set nuclei varied by fold because the split happens at the level of hospitals and not nuclei. Note that the classification accuracy is consistently higher when the assessment was done at the level of superclasses. Abbreviations: AP@.5, average precision when a threshold of 0.5 is used for considering a detection to be true; mAP@.5-.95, mean average precision at detection thresholds between 0.5 and 0.95.

Interestingly, we found that superclass accuracy was higher when trained on granular classes than on superclasses (config 2 versus 6 in [Supplementary Table S2](#)). This indicates that uncommon classes, while noisy, provide signal to improve the function approximation by placing nuclei that look morphologically different (e.g. inactive lymphocytes versus plasma cells) into different ‘buckets’. We also found that NuCLS models outperform approaches that decouple detection and classification into independent, sequential stages (config 2 versus 4 in [Supplementary Table S2](#)) ([Chandradevan et al., 2020](#)).

3.2 DTALE

From a clinical perspective, nucleus detection and classification are arguably more relevant than precise segmentation of nuclei. Segmentation, however, enables the extraction of quantitative and interpretable morphologic nuclear features, which may contain latent prognostic information and help to discover novel biological associations ([Beck et al., 2011](#); [Cooper et al., 2012, 2010](#); [Lazar et al., 2017](#)). Here, we show how segmentation can also be used to enhance the explainability of nucleus classification models, thereby improving confidence in model decisions, a key requirement for clinical adoption ([Amgad et al., 2020](#)).

We developed DTALE, an intuitive quantitative method to explain models like NuCLS. DTALE uses segmentation boundaries predicted by NuCLS to extract interpretable features of nuclear morphometry (shape, staining, edges, etc.), that are used to create a decision tree approximation of our black-box model ([Fig. 6](#)). The outputs of the DTALE tree and the black-box model can be quantitatively compared to evaluate the fidelity of the approximation. We made a distinction between representative explanations of model decisions (e.g. *what features describe most nuclei predicted as tumor?*) and discriminative explanations (e.g. *what features are most specific to tumor predictions?*). The former optimizes for the F1 score, while the latter optimizes for precision ([Supplementary Fig. S5](#)).

DTALE has an important advantage over existing methods like Grad-CAM or LIME in that it provides both an overall explanation of the model decision-making process, as well as explanations for individual nuclei ([Fig. 7](#)) ([Ribeiro et al., 2016](#); [Selvaraju et al., 2017](#)). DTALE fitting accurately explained NuCLS decisions for the most common classes [precision = 0.99 (tumor), 0.89 (stroma), 0.98 (STILs)]. The DTALE tree suggests that tumor nuclei are identified by their large size, globular shape and sharp chromatin edges (i.e. nucleoli or chromatin clumping), that stromal nuclei are identified by their slender shape and rough texture, and that lymphocytes are identified by their small size, circular shape and hyperchromatic staining. Approximations for uncommon classes were not reliable, likely due to: (i) the noisy nature of the ground truth for these classes and (ii) NuCLS model relying on visual characteristics that are not reliably captured by our interpretable features ([D’Amour et al., 2020](#)).

4 Conclusions

This article presented computational techniques that enable the development of scalable and explainable models for nuclear segmentation and classification, with implications in computer-aided diagnostic pathology and discovery of novel quantitative morphologic biomarkers and correlations. We showed how existing general-purpose object segmentation models can be adapted for improved performance in the context of nuclear segmentation and classification. The adaptations also enable learning from hybrid bounding box and segmentation datasets that can be crowdsourced scalably. We also presented DTALE, a novel technique for explaining nucleus classification models using morphologic features obtained by segmentation. DTALE provides global explanations that approximate model behavior as a whole, as well as explanations for individual nuclear predictions, paving the way for trustworthiness and clinical adoption. Contrary to existing approaches, DTALE explanations better capture how pathologists assess histological specimens, and are falsifiable and quantitative by design.

6

M.Amgad et al.

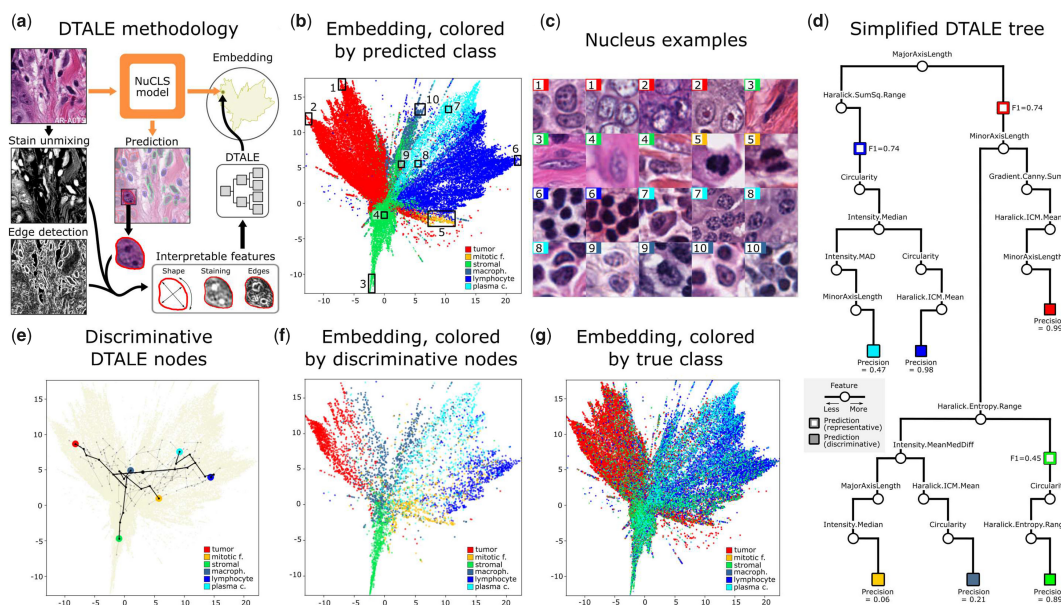


Fig. 6. Explaining NuCLS model decisions using DTALE. (a) Illustrative explanation of the DTALE method. Two-dimensional UMAP embeddings were obtained from the flattened nucleus classification feature maps. A regression decision tree was then fitted to produce predictions in the embedding space using interpretable nucleus features as inputs. (b) Classification embeddings, colored by the prediction that the NuCLS model eventually assigns to the nuclei. (c) Sample nuclei from the embeddings in b. Peripheral regions (1–3, 5–7, 10) contain textbook example nuclei, while nuclei closer to the class boundaries have a more ambiguous morphology. (d) A simplified version of the DTALE tree, showing representative nodes for the three common classes and discriminative nodes for all classes. To reach a discriminative node, DTALE naturally incorporates more features downstream of the representative nodes. (e) An overlay of the fitted DTALE tree (light gray) on top of the NuCLS classification embeddings. In black, we show paths to the nodes that allow discriminative, high-precision, approximation of NuCLS decisions. (f) Nuclei within the embedding, belonging to and colored by, discriminative DTALE nodes. (g) Embeddings are colored by the true class. The three superclasses are well-separated

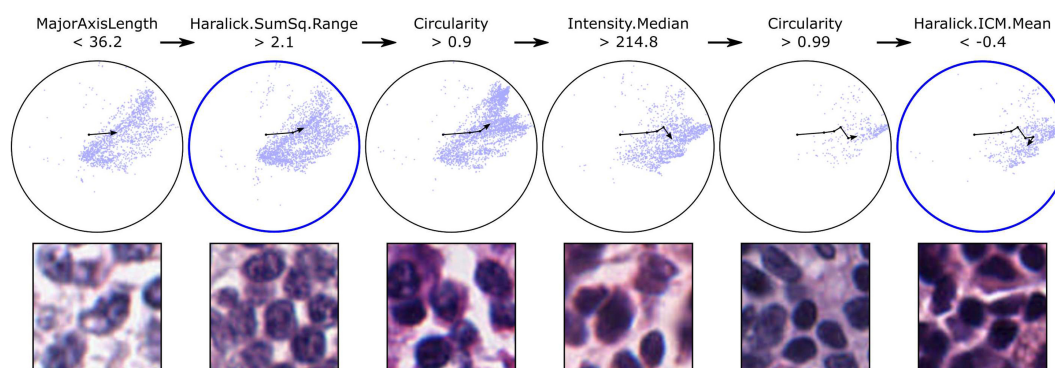


Fig. 7. DTALE enables fine-grained approximation of NuCLS model decisions. Here, we approximate the process by which NuCLS classifies nuclei as lymphocytes. The UMAP embedding is shown, along with an overlay of the DTALE path for lymphocyte classification. An intermediate node in the DTALE path corresponds to the most representative global explanation of NuCLS lymphocyte decisions (left blue circle). The initial set of decision criteria (MajorAxisLength < 36.2 and Haralick.SumSq.Range > 2.1) are our best global explanation for arriving at a lymphocyte classification (F1 = 0.74). When four extra decision criteria are met, we arrive at the most discriminative explanations (second blue circle). These criteria are highly specific to lymphocyte classifications (precision = 0.98). In addition to providing global per-class explanations, DTALE also provides fine-grained, within-class, approximations of NuCLS decision-making. Because DTALE relies on regression trees, we can provide six explanations for different lymphocytes, ranging from ambiguous to highly typical morphology

We would like to note some of the limitations of the work presented. We showed that NuCLS models can handle hybrid data with relatively few segmentation boundaries; only ~37% of the nuclei in the NuCLS hybrid dataset have segmentations. Nonetheless, we did not systematically examine how low this fraction can be before segmentation performance degrades. The NuCLS dataset contains

segmentations for nuclei as opposed to whole cells. This meant that while data collection were more standardized, modeling was more difficult for some classes. Plasma cells, for instance, are distinguishable not only by their (often nonspecific) cartwheel nuclear morphology, but also their perinuclear halo and abundant cytoplasm. Additionally, our NuCLS modeling did not incorporate low-magnification, region-

level patterns. We proposed potential region-cell integration strategies in the past, and we expect this would improve nuclear classification performance (Amgad et al., 2019). Finally, we would note that DTALE explanations are only as rich as the underlying morphologic features used, and the decision tree may not adequately approximate model behavior in all contexts.

Acknowledgements

We would like to acknowledge all participants who annotated the NuCLS datasets, including Ahmed M. Alhusseiny, Mohamed Atef AlMoslemany, Abdelmagid M. Elmatboly, Philip A. Pappalardo, Rokia Adel Sakr, Ahmad Rachid, Anas M. Saad, Ahmad M. Alkashash, Inas A. Ruhban, Anas Alrefai, Nada M. Elgazar, Ali Abdulkarim, Abo-Alela Farag, Amira Etman, Ahmed G. Elsaed, Yahya Alagha, Yonna A. Amer, Ahmed M. Raslan, Menatalla K. Nadim, Mai A.T. Elsebaie, Ahmed Ayad, Liza E. Hanna, Ahmed Gadallah and Mohamed Elkady.

Funding

This work was supported by the U.S. National Institutes of Health National Cancer Institute [grants U01CA220401 and U24CA19436201].

Data availability

The data underlying this article are available in <https://sites.google.com/view/nucls/>.

Conflict of Interest: none declared.

References

- Abels, E. et al. (2019) Computational pathology definitions, best practices, and recommendations for regulatory guidance: a white paper from the digital pathology association. *J. Pathol.*, **249**, 286–294.
- Amgad, M. et al. (2019) Joint region and nucleus segmentation for characterization of tumor infiltrating lymphocytes in breast cancer. *Proc. SPIE Int. Soc. Opt. Eng.*, 10956, 109560M.
- Amgad, M. et al. (2020) Report on computational assessment of tumor infiltrating lymphocytes from the International Immunology Biomarker Working Group. *npj Breast Cancer*, **6**, 16.
- Amgad, M. et al. (2021) NuCLS: a scalable crowdsourcing, deep learning approach and dataset for nucleus classification, localization and segmentation. *Comput. Vis. Pattern Recognit.* arXiv:2102.09099.
- Beck, A.H. et al. (2011) Systematic analysis of breast cancer morphology uncovers stromal features associated with survival. *Sci. Transl. Med.*, **3**, 108ra113.
- Chandradevan, R. et al. (2020) Machine-based detection and classification for bone marrow aspirate differential counts: initial development focusing on nonneoplastic cells. *Lab Invest.*, **100**, 98–109.
- Cooper, L.A.D. et al. (2010) An integrative approach for in silico glioma research. *IEEE Trans. Biomed. Eng.*, **57**, 2617–2621.
- Cooper, L.A.D. et al. (2012) Integrated morphologic analysis for the identification and characterization of disease subtypes. *J. Am. Med. Inform. Assoc.*, **19**, 317–323.
- Dahlin, N. et al. (2020) Designing interpretable approximations to deep reinforcement learning with soft decision trees. *arXiv preprint*, arXiv:2010.14785.
- D'Amour, A. et al. (2020) Underspecification presents challenges for credibility in modern machine learning. *Mach. Learn.* arXiv:2011.03395.
- Diao, J.A. et al. (2021) Human-interpretable image features derived from densely mapped cancer pathology slides predict diverse molecular phenotypes. *Nat. Commun.*, **12**, 1613.
- Gutman, D.A. et al. (2017) The digital slide archive: a software platform for management, integration, and analysis of histology for cancer research. *Cancer Res.*, **77**, e75–e78.
- Haralick, R.M. et al. (1973) Textural features for image classification. *IEEE Trans. Syst. Man Cybern.*, **SMC-3**, 610–621.

- Hartman, D.J. et al. (2020) Value of public challenges for the development of pathology deep learning algorithms. *J. Pathol. Inform.*, **11**, 7.
- Hastie, T. et al. (2017) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd edn. Springer, Berlin, Germany.
- He, K. et al. (2016) Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, Nevada, pp. 770–778.
- He, K. et al. (2017) Mask R-CNN. *Comput. Vis. Pattern Recognit.* arXiv:1703.06870.
- Kokoska, S. and Zwillinger, D. (2000) *CRC Standard Probability and Statistics Tables and Formulae*. Student edn. CRC Press, London, England.
- Kuhn, H.W. (1955) The Hungarian method for the assignment problem. *Nav. Res. Logist. Q.*, **2**, 83–97.
- Lazar, A.J. et al. (2017) Comprehensive and integrated genomic characterization of adult soft tissue sarcomas. *Cell*, **171**, 950–965.
- Leavitt, M.L. and Morcos, A. (2020) Towards falsifiable interpretability research. *Comput. Soc. arXiv:2010.12016*.
- Lin, T.-Y. et al. (2014) Microsoft COCO: common objects in context. *Comput. Vis. Pattern Recognit.* arXiv:1405.0312.
- Macenko, M. et al. (2009) A method for normalizing histology slides for quantitative analysis. In: *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, Boston, MA, pp. 1107–1110.
- Marcinkevics, R. and Vogt, J.E. (2020) Interpretability and explainability: a machine learning zoo mini-tour. *Mach. Learn.* arXiv:2012.01805.
- Masucci, G.V. et al. (2016) Validation of biomarkers to predict response to immunotherapy in cancer: volume I—pre-analytical and analytical validation. *J. Immunother. Cancer*, **4**, 76.
- Mclnnes, L. et al. (2018) UMAP: uniform manifold approximation and projection for dimension reduction. *Mach. Learn.* arXiv:1802.03426.
- Pantanowitz, L. et al. (2013) Validating whole slide imaging for diagnostic purposes in pathology: guideline from the College of American Pathologists Pathology and Laboratory Quality Center. *Arch. Pathol. Lab. Med.*, **137**, 1710–1722.
- Popper, K.R. (1959) *Logic of Scientific Discovery: Basic Books*. Routledge, Oxfordshire, England, UK.
- Ribeiro, M.T. et al. (2016). Why should I trust you? In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, New York, NY.
- Rudin, C. (2018) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.*, **1**, 206–215.
- Saltz, J. et al. (2018) Spatial organization and molecular correlation of tumor-infiltrating lymphocytes using deep learning on pathology images. *Cell Rep.*, **23**, 181–193.e7.
- Samek, W. et al. (2021) Explaining deep neural networks and beyond: a review of methods and applications. *Proc. IEEE*, **109**, 247–278.
- Selvaraju, R.R. et al. (2017) Grad-CAM: visual explanations from deep networks via gradient-based localization. *Comput. Vis. Pattern Recognit.* arXiv:1610.02391.
- Steyerberg, E.W. and Harrell, F.E. (2016) Prediction models need appropriate internal, internal–external, and external validation. *J. Clin. Epidemiol.*, **69**, 245–247.
- Tellez, D. et al. (2018) Whole-slide mitosis detection in H&E breast histology using PHH3 as a reference to train distilled stain-invariant convolutional networks. *IEEE Trans. Med. Imaging*, **37**, 2126–2136.
- Tellez, D. et al. (2019) Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. *Med. Image Anal.*, **58**, 101544.
- Xing, F. and Yang, L. (2016) Robust nucleus/cell detection and segmentation in digital pathology and microscopy images: a comprehensive review. *IEEE Rev. Biomed. Eng.*, **9**, 234–263.
- Yousefi, S. and Nie, Y. (2019) Transfer learning from nucleus detection to classification in histopathology images. In: *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, Venice, Italy, pp. 957–960.
- Zhang, D. et al. (2001) A comparative study on shape retrieval using Fourier descriptors with different shape signatures. In: *Proceedings of the International Conference on Intelligent Multimedia and Distance Education (ICIMADE01)*. John Wiley & Sons Inc., Hoboken, New Jersey, pp. 1–9.

Section 3.3

Joint region and nucleus segmentation for characterization of tumor infiltrating lymphocytes in breast cancer

This section is an authorized exact reproduction of the following paper from the SPIE 2019 conference (San Diego, CA):

Amgad, M., Sarkar, A., Srinivas, C., Redman, R., Ratra, S., Bechert, C. J., Calhoun, B. C., Mrazek, K., Kurkure, U., Cooper, L. A., et al. (2019). Joint region and nucleus segmentation for characterization of tumor infiltrating lymphocytes in breast cancer. In Medical Imaging 2019: Digital Pathology, volume 10956, page 109560M. International Society for Optics and Photonics.

Joint Region and Nucleus Segmentation for Characterization of Tumor Infiltrating Lymphocytes in Breast Cancer

Mohamed Amgad^{†a,b}, Anindya Sarkar^{†b}, Chukka Srinivas^{†b}, Rachel Redman^c, Simrath Ratra^b, Charles J Bechert^c, Benjamin C Calhoun^d, Karen Mrazek^d, Uday Kurkure^b, Lee AD Cooper^{*a,e,f}, Michael Barnes^{*c}

[†] These authors contributed equally to this work

^aDepartment of Biomedical Informatics, Emory University School of Medicine, Atlanta, GA, USA;

^bRoche Tissue Diagnostics, Digital Pathology, Mountain View, CA, USA; ^cRoche Diagnostics Information Solutions, Belmont, CA, USA; ^dDepartment of Pathology, Robert J. Tomsich Pathology and Laboratory Medicine Institute, Cleveland Clinic, Cleveland, OH USA; ^eWinship Cancer Institute, Emory University, Atlanta, GA, USA; ^fDepartment of Biomedical Engineering, Emory University and Georgia Institute of Technology, Atlanta, GA, USA

ABSTRACT

Histologic assessment of stromal tumor infiltrating lymphocytes (sTIL) as a surrogate of the host immune response has been shown to be prognostic and potentially chemo-predictive in triple-negative and HER2-positive breast cancers. The current practice of manual assessment is prone to intra- and inter-observer variability. Furthermore, the inter-play of sTILs, tumor cells, other microenvironment mediators, their spatial relationships, quantity, and other image-based features have yet to be determined exhaustively and systemically. Towards analysis of these aspects, we developed a deep learning based method for joint region-level and nucleus-level segmentation and classification of breast cancer H&E tissue whole slide images. Our proposed method simultaneously identifies tumor, fibroblast, and lymphocyte nuclei, along with key histologic region compartments including tumor and stroma. We also show how the resultant segmentation masks can be combined with seeding approaches to yield accurate nucleus classifications. Furthermore, we outline a simple workflow for calibrating computational scores to human scores for consistency. The pipeline identifies key compartments with high accuracy (Dice= overall: 0.78, tumor: 0.83, and fibroblasts: 0.77). ROC AUC for nucleus classification is high at 0.89 (micro-average), 0.89 (lymphocytes), 0.90 (tumor), and 0.78 (fibroblasts). Spearman correlation between computational sTIL and pathologist consensus is high ($R=0.73$, $p<0.001$) and is higher than inter-pathologist correlation ($R=0.66$, $p<0.001$). Both manual and computational sTIL scores successfully stratify patients by clinical progression outcomes.

Keywords: Tumor infiltrating lymphocytes, convolutional networks, deep learning, computational pathology

1. INTRODUCTION

Tumor Infiltrating Lymphocytes (TiL's) have seen increasing interest in recent years as important surrogate markers of immune response and cancer prognosis in multiple tumor types¹. In breast cancer, and specifically in the Her2+ and triple-negative subtypes (lacking markers for estrogen, progesterone, and Her2), they are known to have strong prognostic value, and have recently been incorporated into clinical guidelines². The most important metric used in clinical practice is sTIL, stromal TILs, which is defined as the fraction of intra-tumoral stroma occupied by lymphocytes. Unlike many histopathology workflows, however, the manual quantification of sTIL (m-sTIL) is particularly well-known for its subjectivity and inter-observer variability, given the difficulty of accurate gauging of which regions to include, and how to accurately estimate the area occupied by lymphocytes². To address this challenge, we developed a streamlined pipeline for integrated, joint region- and cell-level semantic segmentation of Whole-Slide histopathology Images (WSI's). Specifically, we quantified lymphocytic infiltration of the tumor microenvironment in triple-negative breast cancer (TNBC) (Figure 1).

*Address correspondence to: lee.cooper@emory.edu; michael.barnes.mbl@roche.com.

Medical Imaging 2019: Digital Pathology, edited by John E. Tomaszewski, Aaron D. Ward,
Proc. of SPIE Vol. 10956, 109560M · © 2019 SPIE · CCC code:
1605-7422/19/\$18 · doi: 10.1117/12.2512892

Proc. of SPIE Vol. 10956 109560M-1

Whereas traditional approaches rely on feature engineering, we exploited fully-convolutional neural networks (FCN-8), using ImageNet-pretrained VGG16 architecture, for an unbiased approach that outputs pixel-wise class probability maps³. Moreover, we avoided step-wise “classical” computational pathology approaches, where nuclei are segmented, post-processed, and then used to infer region types⁴. Instead, we encoded both the cell- and region- level information in the ground truth itself, hence ensuring that biologically-infeasible (or irrelevant) region-cell combinations are excluded during training; for example, fibroblasts cannot be found in tumor regions. This combined approach helps focus the training process in an integrated fashion, with reduced or minimized expert review or post-processing. Combining this deep-learning workflow with traditional seeding methods results in accurate segmentation and cell classification results, which are used to obtain computational sTIL scores (c-sTIL) that correlate well with pathologist consensus.

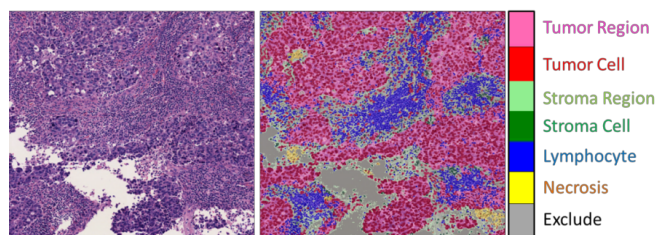


Figure 1. Problem setting. Quantification of tumor infiltrating lymphocytes is a complex task involving segmentation of diverse histological structures. (Left) A representative tile from a testing set slide containing dense sTIL infiltration. (Right) Segmentation output from our model, trained to jointly segment region and cell-level information. Necrotic regions and excluded areas (white spaces, artifacts, etc) are important to segment so that they do not skew the sTIL score calculations.

2. METHODS

The overall workflow used to obtain segmentation and classification result is illustrated in Figure 2.

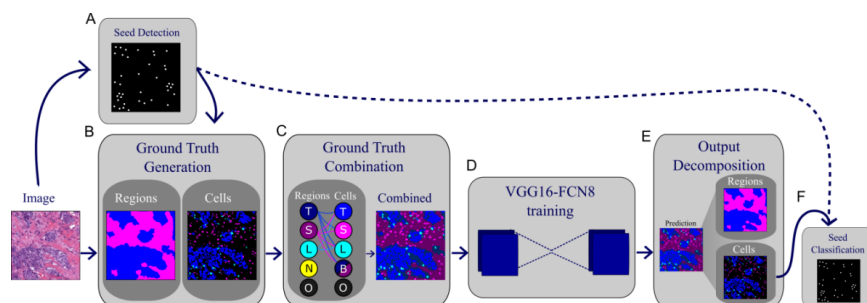


Figure 2. Overall workflow used to obtain region and nucleus classification, as well as seed classification from tiled slides. A. Seeds are extracted from RGB images after deconvolution. B. Region-level ground truth is annotated and semi-automated nucleus segmentation ground truth is vetted by a pathologist. C. Region and nucleus-level ground truth is combined into one common mask to be used for training. This process ensures consistency and excludes biologically-infeasible combinations. D. A fully-convolutional network is trained to output a combined mask. E. Output is decomposed into region and nucleus segmentation masks. F. Seed classifications are obtained from the cell segmentation mask.

2.1 Dataset used and ground truth generation

The cohort used in this study consists of 120 anonymized H&E stained slides, which were obtained from the Cleveland Clinic Foundation, and scanned using a single Aperio scanner at 20x magnification. The slides had m-sTIL scores from two practicing pathologists, who resolved inconsistent scores via consensus. 14 slides had available ground truth, 5 of which were held-out to measure segmentation and classification generalization accuracy. Two models were trained: 1. A model to calculate segmentation and classification testing set accuracies (trained on 9 slides); and 2. A model to calculate sTIL scores over the entire dataset (trained on all 14 slides with available ground truth). The 14 slides with available annotations were divided into overlapping tiles of size 1024x1024. The annotated slides were chosen to represent as many of the histological structures as possible within the dataset. Note that the ground truthing process is extremely labor-

intensive, and the regions chosen for annotation are fairly large (on the order of $\sim 5K$ pixels squared) to ensure adequate training and trustable accuracy metrics. Two types of segmentation ground truth were obtained:

1. Region level ground truth: this was manually annotated by drawing polygon boundaries at tissue interfaces.
2. Cell-level segmentation ground truth: this was obtained in a semi-automatic manner. Traditional methods based on radial symmetry were used to extract seeds and segment nuclei, whose class was then determined using size and shape heuristics to provide a first-cut approximation of nucleus classification⁵.

The results were overlooked and corrected by a senior pathologist.

2.2 Combining region and nucleus ground truths

Region and nucleus segmentation masks were combined such that pixel value encodes both region and nucleus membership information; essentially the fully-convolutional model was trained to classify pixels into 12 different classes (Table 1). Combining different classification problems in the same framework had two advantages. First, it reduces reliance on post-processing heuristics, such as ensemble learning or parameter tuning approaches, to combine the two sets of results. Second, it utilizes a priori biological knowledge to generate consistent results. For example this framework disallows any cell classifications within necrotic regions (nuclear debris and dead cells are counted as part of the necrotic region and not delineated individually). It also disallows stromal cells (fibroblasts) within non-stromal regions. To incorporate non-nuclear components of a histologic region, including cell cytoplasm, extracellular matrix and other structural elements, we also included region categories that correspond to non-nuclear elements.

Table 1. Combined mask code and corresponding region and cell information encoding.

Code	Region	Nucleus	Code	Region	Nucleus
1	Background	N/A	7	Stroma	Lymphocyte
2	Tumor	Tumor	8	Lymphocyte	Lymphocyte
3	Stroma	Tumor	9	Tumor	N/A
4	Lymphocyte	Tumor	10	Stroma	N/A
5	Stroma	Stroma	11	Necrosis	N/A
6	Tumor	Lymphocyte	12	Lymphocyte	N/A

2.3 Fully Convolutional model training and inference

We tried two architectures for training: VGG16-FCN8 and FC-DensNet103^{3,6} (Figure 3). VGG16-FCN8 showed better training and convergence properties and was hence chosen.

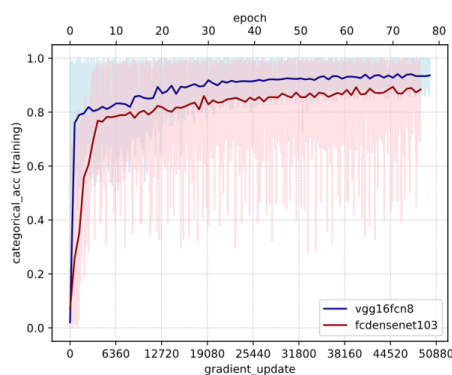


Figure 3. Effect of architecture on training categorical accuracy (model fitting). Light colors represent the batch-level accuracy, while darker colors represent epoch-level accuracies. Pre-trained VGG16-FCN8 has better model fitting properties than the

deeper and more complex fully-connected DenseNet-103 for this problem setting. Also note the remarkably higher batch-to-batch variability in FC-DenseNet compared to VGG. Both networks were trained on the same set of slides with exactly the same set of hyperparameters, including batch size, optimizer type and learning rate, for comparability.

Our model that has been pre-trained on ImageNet in tensorflow, but we only used pre-training as a weight initialization strategy, and allowed the full 16 layer weights to be optimized during the training process. We used Adam optimizer and learning rate of $1e-5$ ^{3,7}. Single machine, 4-GPU data parallelism with gradient averaging was used. The main model was instantiated on each of GPUs with weight sharing. A batch of 4 images is sent to each of the GPUs for gradient calculation. These are sent back to the CPU and averaged to get the overall gradient update⁸. Weighted categorical cross entropy loss was used to train the model, with the class-specific weights being determined as:

$$W_c = 1 - \frac{N_c}{\sum_{c=1}^{12} N_c} \quad (1)$$

Where N_c is the number of pixels belonging to class c in the training dataset. This helps handle class imbalance during training by assigning higher weight to less abundant classes. Categorical accuracies reported are defined as the argmax of soft class prediction probabilities. Two methods of data augmentation were used to improve robustness of the training process: 1. Tiles were generated with an a shift overlap of 250 pixels; 2. A random FOV of size 768x768 was cropped on the fly and is what is actually input to train the model. After the network has been trained, the tile size used for inference ranged from 1024x1024 to 2048x2048. Note that the trained convolutional weights in a fully-convolutional network can be applied to any tile size as long as it fits in GPU memory. The combined prediction mask from the trained network is then decomposed back into region and cell-specific masks by reverse-mapping the coding scheme in Table 1. Note, however, that the “background in lymphocyte region” (i.e. pixel does not belong to a cell, but region was annotated as lymphocyte-rich) was mapped to stromal regions. It is important to note that, technically-speaking there is no “lymphocyte region”, and that while lymphocytes may tend to aggregate there is no clear threshold for the density at which lymphocytic aggregates are considered to be a region. The lymphocyte region class, therefore, was just used to facilitate ground-truthing (create a single boundary rather than click on hundreds of cells) and to train the model to detect these aggregates. For all slides (training/testing), a hard threshold of 220 (for all R, G, B channels) was used to map all white regions to the exclude class. Segmentation accuracy was quantified using the DICE coefficient (2x intersection over bag union).

2.4 Seed classification by pixel class majority

The soft scores for seed class membership are obtained by counting the proportion of a circle of radius r pixels that belongs to the class of interest. The final classification of a seed is therefore determined by the equation:

$$\text{Seed Classification} = \operatorname{argmax}_{c \in \{\text{tumor}, \text{fibroblast}, \text{lymphocyte}, \text{other}\}} (n_c) \quad (2)$$

Where n_c is the number of pixels belonging to class c within a radius r of the nucleus seed. This helps de-noise some of the segmentation inaccuracies. A radius of 5 pixels was used in our experiments.

2.5 c-sTIL scoring and progression outcomes analysis

We focused on sTILs, as opposed to intra-tumoral TILs (tTILs) to faithfully adhere to the clinical scoring guidelines. The guidelines mention a set of rules that determine which regions are suitable for calculating sTIL scores, most notably²:

1. Do not focus on “hot spots” too much: This rule was set to facilitate manual scoring, but is not very relevant to computational quantification, since we calculate the statistics globally across all included tiles without the inherent biases of manual scoring.
2. Do not include cells in necrotic regions: we segment necrotic regions and exclude them in the calculation.
3. Focus on stromal areas proximal to the tumor: we addressed this by discarding the bottom x percentile of tiles by tumor fraction, where x is determined by supervised hyperparameter tuning. The slides used to train the segmentation algorithm were combined with 16 others (randomly chosen) to construct a new training set. The process was repeated 30 times (Monte-Carlo cross validation). During each “trial”, the training set was used to:
 - a. Find the optimal exclusion threshold which maximizes correlation with pathologist sTIL scores on the training set; and
 - b. Learn the linear calibration bias to map absolute algorithmic scores to manual scores.

A threshold value of 10% (based on clinical guidelines), was used to dichotomize the sTIL scores for outcome correlative analysis². A “progression” event was defined as the earliest occurrence of local, regional, or distant metastasis events.

3. RESULTS AND DISCUSSION

Model predictions on the held-out testing set are accurate and correspond well to ground truth and underlying tissue boundaries (Figure 4). Discrepancies between the model and ground truth arise from inaccuracies in either the prediction or the ground truth itself. Ground truth may be inaccurate for a number of reasons:

1. Human limitations. Human annotators tend to prefer smooth contours and connected regions, even when the true underlying tissue structures have jagged edges and are composed of multiple scattered regions. Moreover, there is considerable difficulty in noticing and segmenting every lymphocyte in the dataset, whether manually or through vetting of moderately accurate H&E radial symmetry segmentation algorithms. Essentially, the model is learning the latent representation within a noisy ground truth, resulting in predictions that oftentimes surpass the limitations of ground truth⁹.
2. Limitations related to the accuracy of deconvolution, seeding and segmentation used to generate the cell-level ground truth.

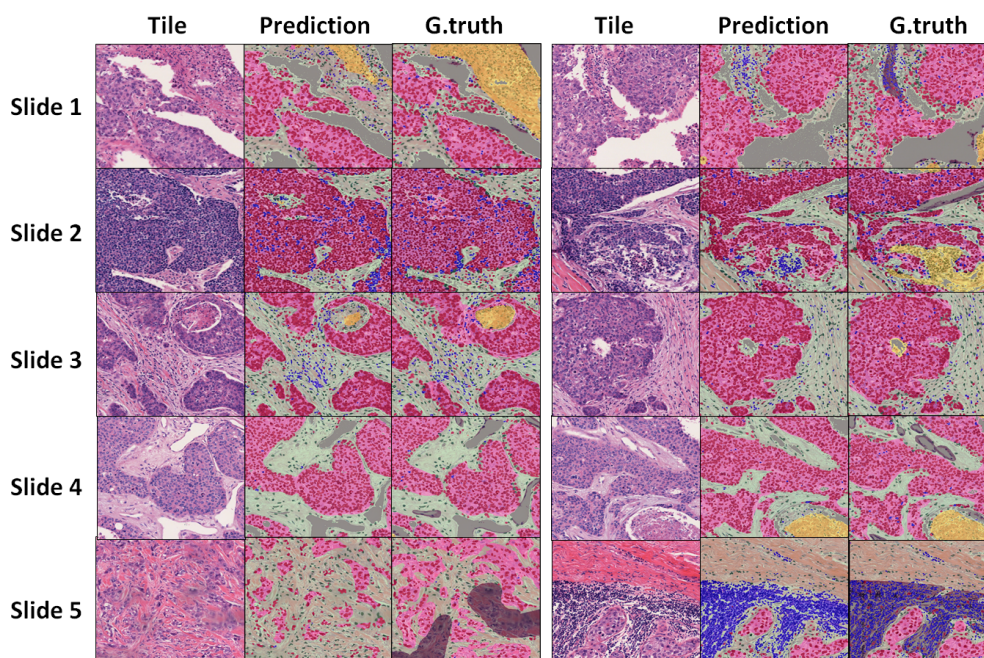


Figure 4. Qualitative examination of segmentation results on the testing set. Representative tiles from each the testing set. Slide 1 (right) and Slide 5 (right): Non-cellular components of “lymphocyte regions” (grey) were present in ground truth (for training) but were mapped to stroma in output. Slide 2 (left): enclosed stromal region within a tumor nest is missed in ground truth but is picked up by trained model. Slide 2 (right) and Slide 3 (right) algorithm misclassified small necrotic region as stroma. Slide 5 (left): Ground truth connects small, scattered tumor nests under one tumor “region”, whereas the model learns to more accurately delineate region boundaries.

The accuracy of region segmentation, as measured using the Dice coefficient, was: 0.78 (overall), 0.83 (tumor) and 0.77 (stroma). Seed classification area under receiver-operator characteristics curve was 0.89 (micro-average), 0.89 (lymphocytes), 0.90 (tumor), and 0.78 (fibroblasts) (Figure 5).

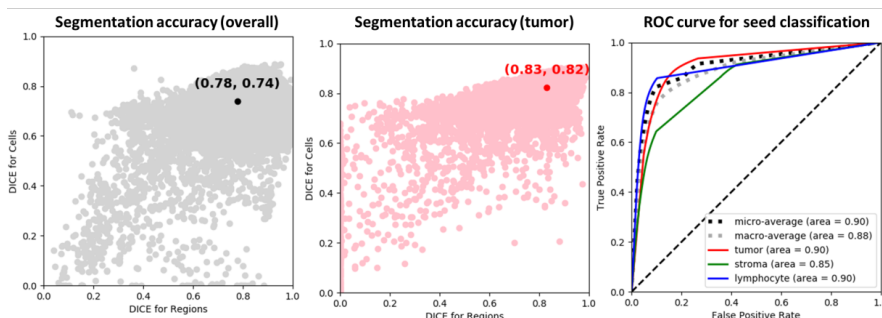


Figure 5. Accuracy of segmentation and classification. (Left) Overall semantic segmentation accuracy, measured by the DICE coefficient. The accuracy was calculated after decomposition of the model output into separate region and nucleus segmentation masks. Every point represents the accuracy over one tile in the testing set. Note the general correspondence between segmentation accuracy for regions and for nuclei. (Middle) Segmentation accuracy for tumor classification. (Right) Receiver-Operator Characteristics curve for final seed classification by pixel class majority.

Computational sTIL scores were strongly and significantly correlated with consensus pathologist scores (Figure 6). Spearman Correlation between computational sTIL and pathologist consensus is high ($R=0.73$, $p<0.001$) and is higher than inter-pathologist correlation ($R=0.66$, $p<0.001$), though smaller in magnitude (hence the rationale for learning the linear calibration). We believe this magnitude difference is related to the inherent biases and ambiguity in estimating area by human observers.

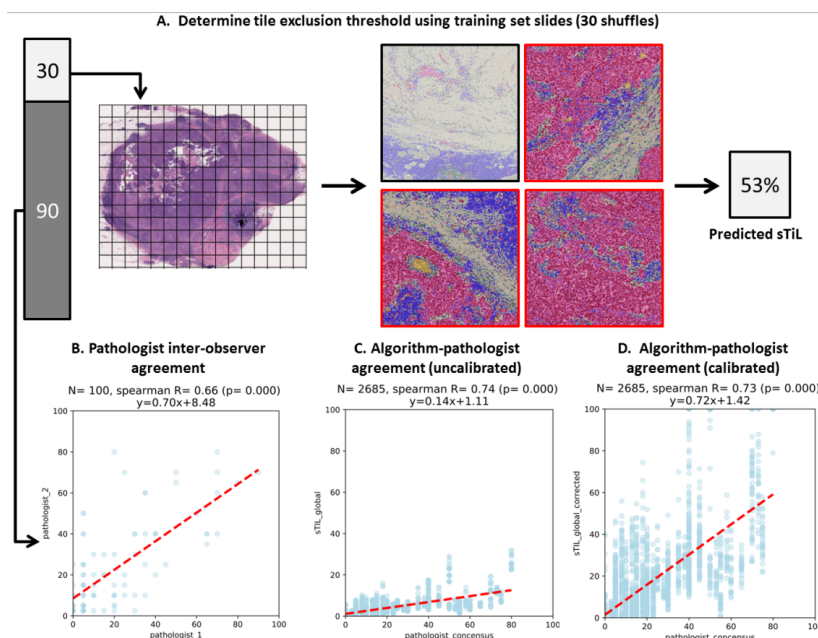


Figure 6. Calculating sTIL scores and correlating with consensus manual pathologist scores. A. Supervised tile selection process using the training set. The 30 training slides are used to learn a threshold for excluding irrelevant tiles by tumor fraction and to learn a linear calibration to map true c-sTIL fraction to what would be perceived as the sTIL fraction by a practicing pathologist. B. Agreement between the two pathologists. C. Agreement between algorithmic TiL scores and pathologist consensus. Each point represents one testing set slide from one of the 30 shuffles. D. Same as C, but after linear calibration.

The dichotomized sTIL score is 85% accurate in identifying low sTIL slides. This low sTIL group is characterized by poor survival outcomes, both using m-sTIL and c-sTIL scoring, consistent with existing literature and providing an additional layer of validation to the computational pipeline described here (Figure 7).

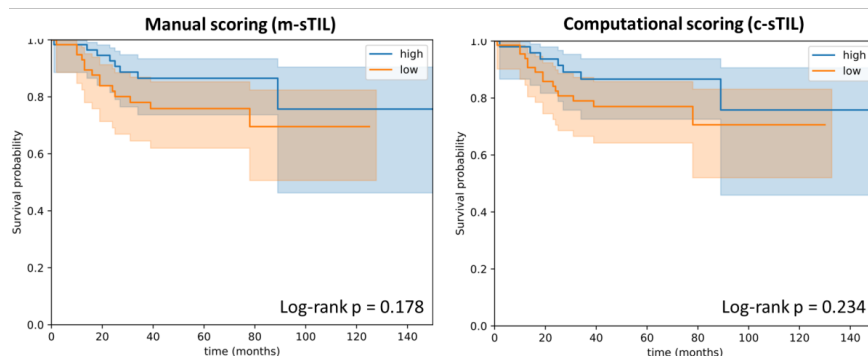


Figure 7. Kaplan-Meier curves for dichotomized human and computational sTIL scores. A threshold of 10% was used to distinguish low- from moderate or high infiltrates, consistent with the published guidelines.

4. LIMITATIONS AND CONCLUSIONS

This work, like most others in the current computational pathology space, is limited by the lack of large-scale validated ground truth for segmentation of salient tissue components in breast cancer H&E images. Nonetheless, the limited dataset we have illustrates the validity of methods presented, both analytically (segmentation and classification accuracy) and clinically (TIL score and disease outcomes). The rarity of TNBC resulted in a relative scarcity of progression events, causing the Kaplan-Meier analysis to be slightly under-powered. Nevertheless, the trends are in the right direction and are almost indistinguishable for m-sTIL and c-sTIL scoring. Future work will investigate generalization of this algorithm to independent datasets derived from institutions not included in the model training and optimization process.

Our results illustrate how an end-to-end framework enables accurate and consistent estimation of tumor infiltrating lymphocytes in breast cancer. The results are highly concordant with consensus scores from pathologists and successfully stratify patients by clinical progression outcomes. In the future we intend to extract various spatial sTIL metrics for correlation with clinical and genomic variables.

REFERENCES

- [1] Hendry, S., Salgado, R., Gevaert, T., Russell, P. A., John, T., Thapa, B., Christie, M., van de Vijver, K., Estrada, M. V., Gonzalez-Ericsson, P. I., Sanders, M., Solomon, B., Solinas, C., Van den Eynden, G. G. G. M., Allory, Y., Preusser, M., Hainfellner, J., Pruneri, G., Vingiani, A., et al., “Assessing Tumor-Infiltrating Lymphocytes in Solid Tumors: A Practical Review for Pathologists and Proposal for a Standardized Method from the International Immuno-Oncology Biomarkers Working Group: Part 2: TILs in Melanoma, Gastrointestinal Tract Carcinom,” *Adv. Anat. Pathol.* **24**(6), 311–335 (2017).
- [2] Salgado, R., Denkert, C., Demaria, S., Sirtaine, N., Klauschen, F., Pruneri, G., Wienert, S., Van den Eynden, G., Baehner, F. L., Penault-Llorca, F., Perez, E. A., Thompson, E. A., Symmans, W. F., Richardson, A. L., Brock, J., Criscitiello, C., Bailey, H., Ignatiadis, M., Floris, G., et al., “The evaluation of tumor-infiltrating lymphocytes (TILs) in breast cancer: recommendations by an International TILs Working Group 2014,” *Ann. Oncol. Off. J. Eur. Soc. Med. Oncol.* **26**(2), 259–271 (2015).
- [3] Long, J., Shelhamer, E. and Darrell, T., “Fully Convolutional Networks for Semantic Segmentation” (2014).
- [4] AP, R., Khan, S. S., Anubhav, K. and Paul, A., “Gland Segmentation in Histopathology Images Using Random Forest Guided Boundary Construction” (2017).

- [5] Xing, F. and Yang, L., “Robust Nucleus/Cell Detection and Segmentation in Digital Pathology and Microscopy Images: A Comprehensive Review.,” *IEEE Rev. Biomed. Eng.* **9**, 234–263 (2016).
- [6] Jégou, S., Drozdal, M., Vazquez, D., Romero, A. and Bengio, Y., “The One Hundred Layers Tiramisu: Fully Convolutional DenseNets for Semantic Segmentation” (2016).
- [7] Kingma, D. P. and Ba, J., “Adam: A Method for Stochastic Optimization” (2014).
- [8] Tensorflow., “Cifar 10 multi-GPU training tutorial” (2018).
- [9] Khoreva, A., Benenson, R., Hosang, J., Hein, M. and Schiele, B., “Simple Does It: Weakly Supervised Instance and Semantic Segmentation” (2016).

Section 3.4

MuTILs: explainable, multiresolution computational scoring of Tumor-Infiltrating Lymphocytes in breast carcinomas using clinical guidelines

3.4.1 Introduction

Tumor-Infiltrating Lymphocytes (TILs) are an important prognostic and predictive biomarker in basal and Her2+ breast carcinomas [158]. The stromal TILs score is the fraction of stroma within the tumor bed occupied by lymphoplasmacytic infiltrates (Figure 3.4.1).

TILs are assessed visually by pathologists through examination of formalin-fixed paraffin-embedded, hematoxylin and eosin (FFPE H&E) stained slides from tumor biopsies or resections. They are subject to considerable inter- and intraobserver variability, and hence a set of standardized recommendations was developed by the international Immuno-Oncology Working Group [155, 94]. Nevertheless, observer variability remains a critical limiting factor in the widespread clinical adoption of TILs in research and clinical settings. Therefore, a set of recommendations was published for developing computational tools for TILs assessment [11]. This brief report describes the development and validation of *MuTILs*, an explainable deep-learning model for the evaluation of TILs.

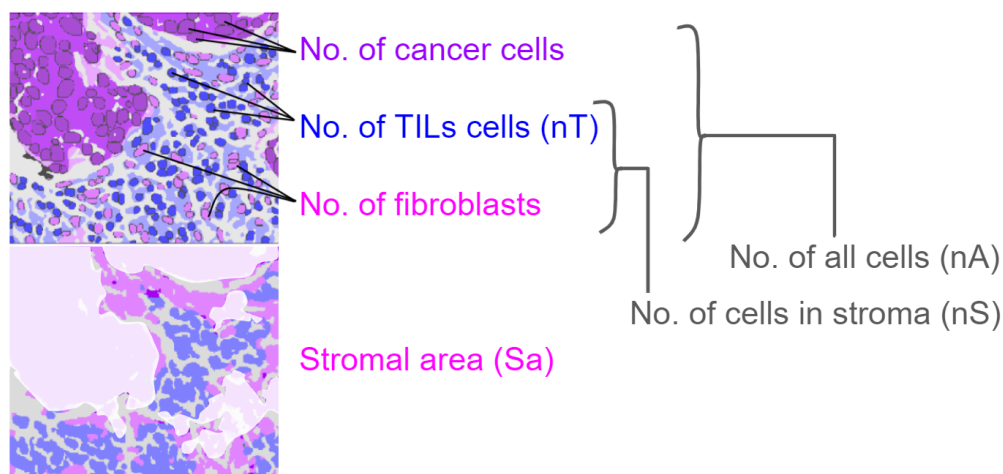


Figure 3.4.1: Components of various variants of the computational TILs score.

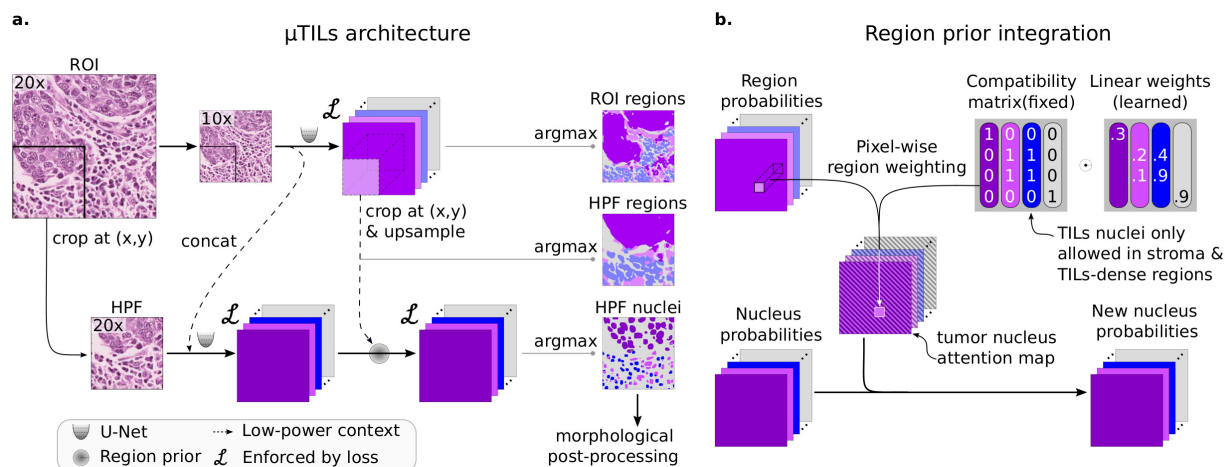


Figure 3.4.2: **MuTILs model architecture.** a. The MuTILs architecture utilizes two parallel U-Net models to segment regions at 1 micron-per-pixel (MPP) and nuclei at a 0.5 MPP resolution. Inspired by HookNet, we passed information down from the low-resolution branch to the high-resolution branch by concatenation. Additionally, region predictions from the low-resolution branch are upsampled and used to constrain the nucleus predictions in the high-resolution branch. The model was trained using a multi-task loss that gives equal weight to ROI and high-power-field (HPF) region predictions, unconstrained HPF nuclear predictions, and region-constrained nuclear predictions. b. Region predictions are used to constrain nucleus predictions to enforce compatible cell predictions through class-specific attention maps. Attention maps are derived by modeling the nucleus class prior probability as a linear combination of the corresponding region probability vector. User-defined manual compatibility kernels mask out incompatible predictions.

3.4.2 Methods

MuTILs jointly segments tissue regions and cell nuclei and extends our earlier work on this topic (Figure 3.4.2) [10]. It comprises two parallel U-Nets (each with a depth of 5) for segmenting regions and nuclei at 1 and 0.5 microns-per-pixel (MPP), respectively [146]. Inspired by the HookNet architecture, information is passed from the region branch down to the nucleus branch to provide low-power context [186]. Additionally, we employed a series of constraints to promote compatible, biologically sensible predictions.

We relied on images from 125 infiltrating ductal breast carcinoma patients from the BCSS and NuCLS datasets [7, 8]. Additionally, we supplemented the training set with annotations from 85 slides from a private cohort. The slides were separated into training and testing sets using 5-fold internal-external cross-validation, using the same folds as the NuCLS modeling paper [7, 171]. For training, we extrapolated the nuclear labels from the small 256x256 pixel high-power fields to large 1024x1024 pixel regions of interest (ROIs) by using NuCLS models to perform inference on

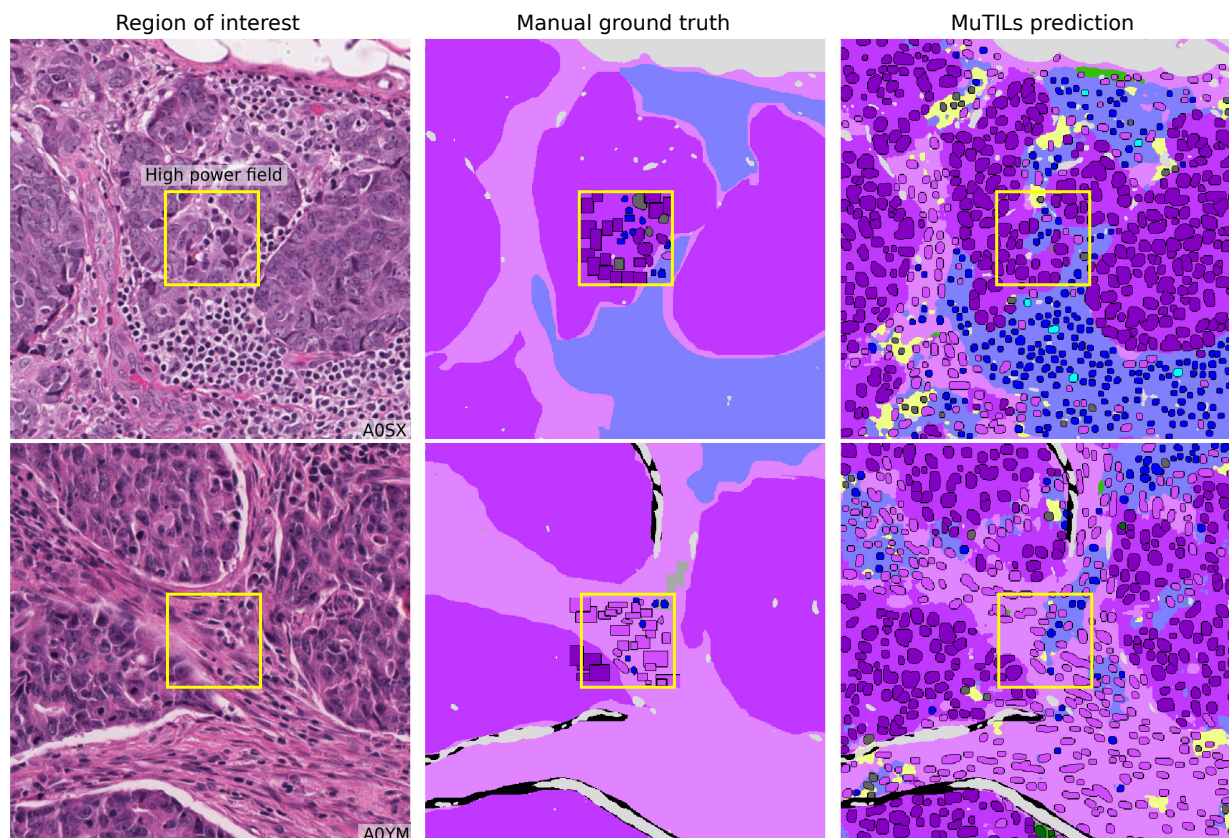


Figure 3.4.3: **Reconciliation of manual region and nucleus ground truth for model validation.** Each high power field from the pathologist-corrected single-rater NuCLS dataset was padded to 1024x1024 at 0.5 MPP resolution (20x objective). As a result, each ROI had region segmentation for the entire field (from the BCSS dataset) and nucleus segmentation and classification for the central portion (from the NuCLS dataset). Note that the nucleus ground truth contains a mixture of bounding boxes and segmentation. The fields shown here are from the testing sets.

the same slides they were trained on to obtain bootstrapped “weak” labels. Generalization results presented here use manual labels (Figure 3.4.3).

For whole-slide image (WSI) inference, we relied on data from 305 breast carcinoma patients for validation, 269 of whom were infiltrating ductal carcinomas, and 156 were Her2+. Visual scores were assessed by one pathologist (Dr. Roberto Salgado, GZA-ZNA Ziekenhuizen, Antwerp, Belgium) and used as the baseline. The WSI accession and tiling workflow used the *histolab* and *large_image* packages and included: 1. Tissue detection; 2. Detection and exclusion of empty space and markers/inking; 3. Tiling the slide and scoring tiles at a very low resolution (2 MPP); 4. Analyzing the top 300 tiles [53, 112]. Fixing the number of analyzed ROIs ensured a near-constant run time of less than two hours per slide. Low-resolution tiles with a high composition of cellular

(hematoxylin-rich) and acellular (eosin-rich) regions received a higher informativeness score. This favored tiles with more peritumoral stroma. Color deconvolution was performed using the Macenko method from the HistomicsTK package [58, 110]. Each of the top informative tiles was assigned one of the trained MuTILs models in a grid-like fashion. This scheme acted as a form of ensembling without increasing the overall inference time.

Trained MuTILs models were then used to segment tissue and nuclear components. A euclidean distance transform was applied to detect stroma within 32 microns from the tumor boundary. The fraction of image pixels occupied by this peritumoral stroma was considered a saliency score. We assessed the following variant of the TILs score (Figure 3.4.1):

- Number of TILs / Stromal area (nT_{Sa}).
- Number of TILs / Number of cells in stroma (nT_{nS}).
- Number of TILs / Total Number of cells (nT_{nA}).

We obtained these score variants both globally (aggregating region and nuclear counts from all ROIs) and through saliency-weighted averaging of scores obtained for each ROI independently. A simple linear calibration was then used to ensure the scores occupied a similar range as the visual scores.

3.4.3 Results

Table 3.4.1 shows the region segmentation and nucleus classification accuracy on the testing sets. MuTILs achieves high accuracy for stromal region segmentation (DICE=80.8±0.4), as well as the classification of fibroblasts (AUROC=91.0±3.6), lymphocytes (AUROC=93.0±1.1), and plasma cells (AUROC=81.6±6.6) — all contributors to the computational TILs score. This accuracy is also supported by qualitative examination of model predictions on both the ROIs from BCSS and NuCLS datasets (Figure 3.4.3) and the full WSI (Figure 3.4.4). Computational TILs score variants had a modest-to-high correlation with the visual scores (Spearman R ranges between 0.55 - 0.58) (Figure 3.4.5). Some slides were outliers with discrepant visual and computational scores; the causes for this discrepancy are discussed below. Both global and ROI saliency-weighted scores were significantly correlated with the visual scores ($p < 0.001$).

Table 3.4.1: **Generalization accuracy for region segmentation and nucleus classification using manual ground truth.** Results are on testing sets from the internal-external 5-fold cross-validation scheme (separation by hospital). Fold 1 contributed to hyperparameter tuning, so it is not included in the mean and standard deviation calculation. MuTILs achieves a high classification performance for components of the computational TILs score. Region segmentation performance is variable and class-dependent, with the predominant classes (cancer, stroma, and empty) being the most accurate. The region constraint improves nuclear classification accuracy by 2-3% overall, mainly by reducing misclassification of immature fibroblasts and large TILs/plasma cells as cancer (see qualitative examination figure).

* Classes that contribute to the computational TILs score.

† Performance for Necrosis/Debris and TILs-dense regions is modest, primarily because of the inherent subjectivity of the task and variability in the ground truth. For example, how dense should the infiltrate be to be considered “dense”? Necrotic regions also often have TILs infiltrates at the margin or adjacent areas of fibrosis, which are inconsistently labeled as necrosis, stroma, or TILs-dense in the ground truth. Nonetheless, the classification of cells/material that comprise necrotic regions (neutrophils, apoptotic bodies, debris, etc.) is reasonable at higher magnification.

‡ From the table, it is clear that the model essentially fails to segment normal breast acini at 10x magnification. This failure is likely caused by: 1. The low representation of normal breast tissue in the validation data from NuCLS and BCSS datasets; 2. Inconsistency in defining “normal,” which is sometimes used in the sense of “non-cancer” (including benign proliferation), and sometimes only refers to terminal ductal and lobular units (TDLUs). At high resolution, the distinction between cancer versus normal/benign epithelial nuclei is reasonable.

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean	Std
Regions at 10x objective (DICE)							
Cancer	84.4	82.1	83	82.8	82.8	82.7	0.4
Normal ‡	1.6	2.3	2.1	2.3	2.3	2.3	0.1
Stroma *	81.3	80.2	81	80.8	81	80.8	0.4
TILs-dense †	64.8	64	65.3	65.6	65.6	65.1	0.8
Necrosis/Debris †	64.1	55.6	56.7	57.3	57.1	56.7	0.8
Empty	83.5	83.5	84	84.2	84.3	84.0	0.4
Nuclei at 20x objective (AUROC)							
Cancer	96.5	97.2	98	97.4	91.1	95.9	3.2
Normal ‡		84.6	89.3	80	74.7	82.2	6.3
Fibroblast *	90.4	93	91.8	93.5	85.8	91.0	3.6
Lymphocyte *	93.3	92.3	93.6	91.9	94.2	93.0	1.1
Plasma Cell *	80.9	73.5	88	78.9	85.8	81.6	6.6
Debris †	82.8	84.9	80.1	93.9	57.1	79.0	15.7
Micro-avg.	91.9	92.2	95.6	93.5	88.9	92.6	2.8
Macro-avg.	85.4	83.9	86.3	85.2	75.3	82.7	5.0
Nuclei without region constraint (AUROC)							
Micro-avg.	90.5	91.1	95.4	91.9	86.2	91.2	3.8
Macro-avg.	84.5	78.1	86.9	81.5	73.1	79.9	5.8

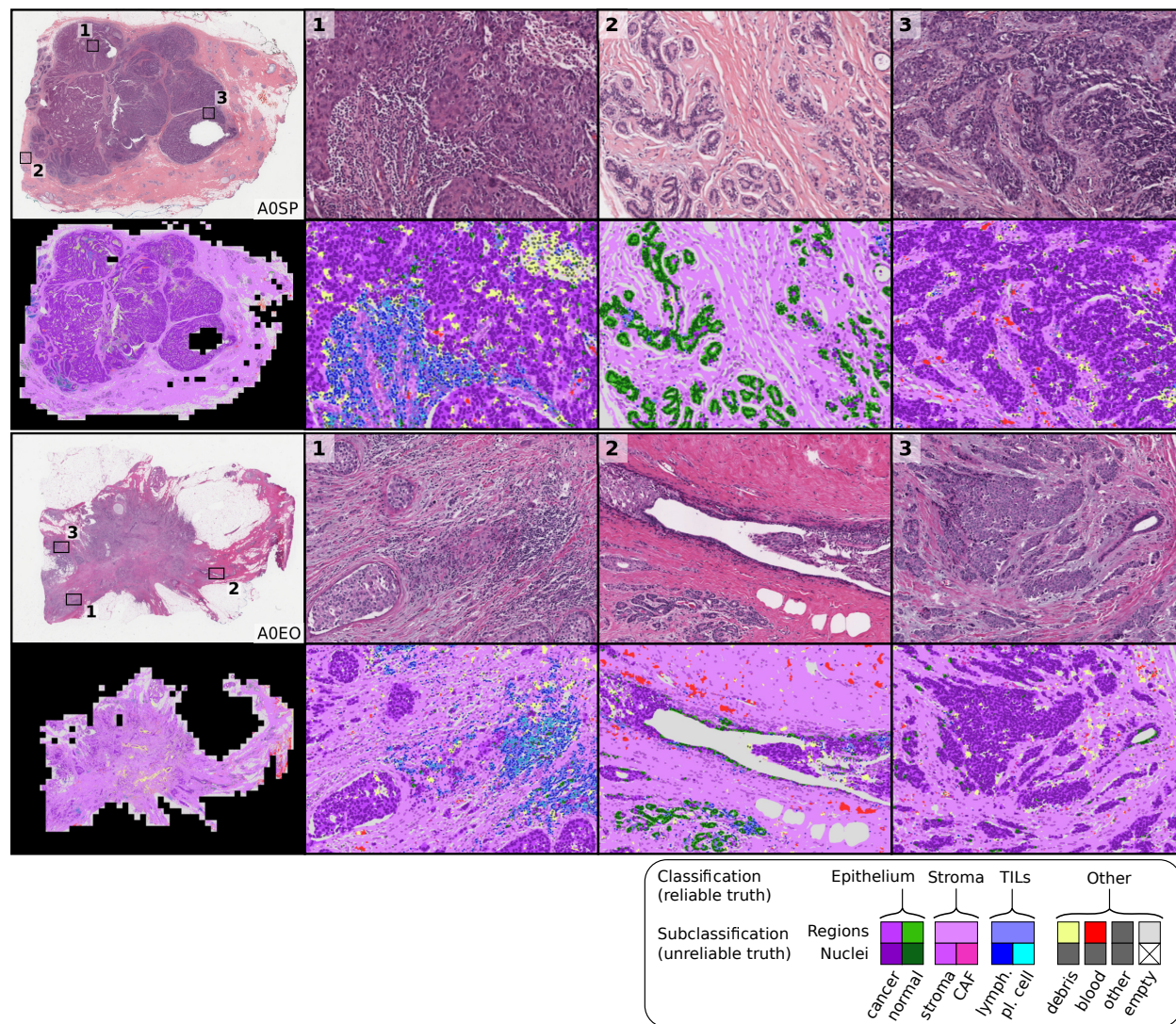


Figure 3.4.4: **Sample whole-slide predictions from trained MuTILs models.** The predictions show full WSI inference for illustration. Our analysis, however, only admitted the 300 most informative ROIs to the MuTILs model to ensure a constant run time of less than two hours per slide for practical applicability. ROI “informativeness” was measured at a very low resolution (2 MPP) during WSI tiling and favored ROIs with more peritumoral stroma.

We examined the prognostic value of MuTILs on infiltrating ductal carcinomas and Her2+ carcinomas. While we had access to visual scores from the basal cohort, the number of outcomes was limited, and neither visual nor computational scores had prognostic value. Progression-free interval (PFI) is the endpoint used per recommendations from Liu et al. for TCGA, with progression events including local and distant spread, recurrence, or death [105]. First, we examined the Kaplan-Meier curves for patient subgroups using a TILs-score threshold of 10% for stromal TILs score and the median value for the nTnA computational score variant (Figure 3.4.6). Both visual and

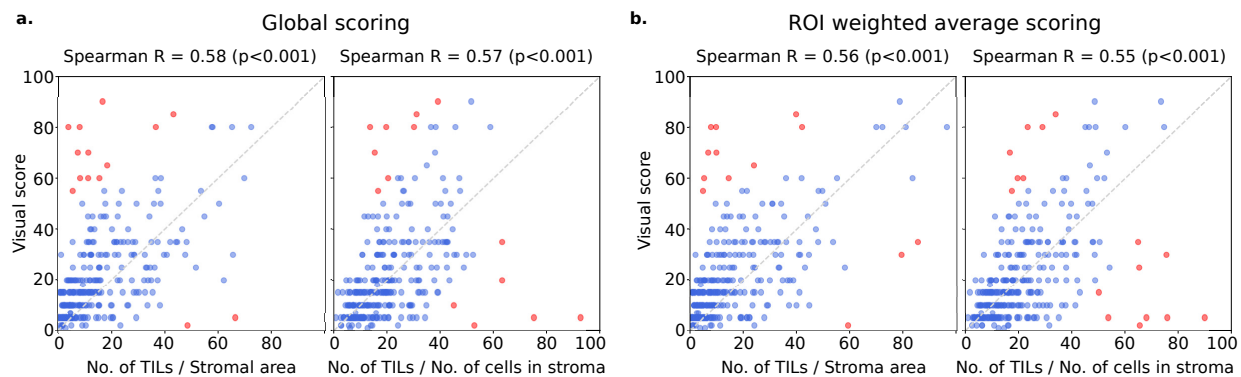


Figure 3.4.5: Correlation between visual and computational TILs assessment scores. Visual scores were obtained from one pathologist using clinical scoring recommendations from the TILs Working Group. MuTILs is a concept bottleneck model with a strong emphasis on explainability; it segments individual regions and nuclei, which are then used to calculate the computational scores. Two variants of computational scores were obtained: either the number of stromal TILs was divided by the stromal region area, or the number of TILs was divided by the total number of cells within the stromal region. We then calibrated these numbers to the visual scores for easy comparison. While this scatter plot shows the calibrated scores, the correlation coefficients were obtained using the raw scores to avoid optimistic results. Each point represents a single patient. Points in red are outliers that contributed to the correlation metric but not to the calibration. a. Computational scores are computed globally by aggregating data from all ROIs. b. Computational scores are computed independently for each ROI, and the slide-level score is calculated by weighted averaging.

computational scores had good separation within the infiltrating ductal cohort, although only the nTnS and nTnA computational scores had significant log-rank p-values ($p=0.009$ and $p=0.006$, respectively). Within the Her2+ cohort, all metrics had good separation on the Kaplan-Meier, although the visual score had a borderline p-value. All computational scores were significant within this cohort ($p=0.018$ for nTnS, $p=0.002$ for nTnS, and $p=0.006$ for nTnA).

We also examined the prognostic value of the continuous (untresholded) TILs scores using Cox proportional hazards regression, with and without controlling for clinically-salient covariates including patient age, AJCC pathologic stage, histologic subtype, and basal status (Table 3.4.2). Within the infiltrating ductal cohort, the only metric with significant independent prognostic value on multivariable analysis was the nTnS computational score. Within the Her2+ cohort, the visual score was not independently prognostic ($p=0.158$), while the computational scores all had independent prognostic value, with the most prognostic being the nTnS variant ($p=0.003$, $HR<0.001$). Saliency-weighted ROI scores almost always had better prognostic value than global computational

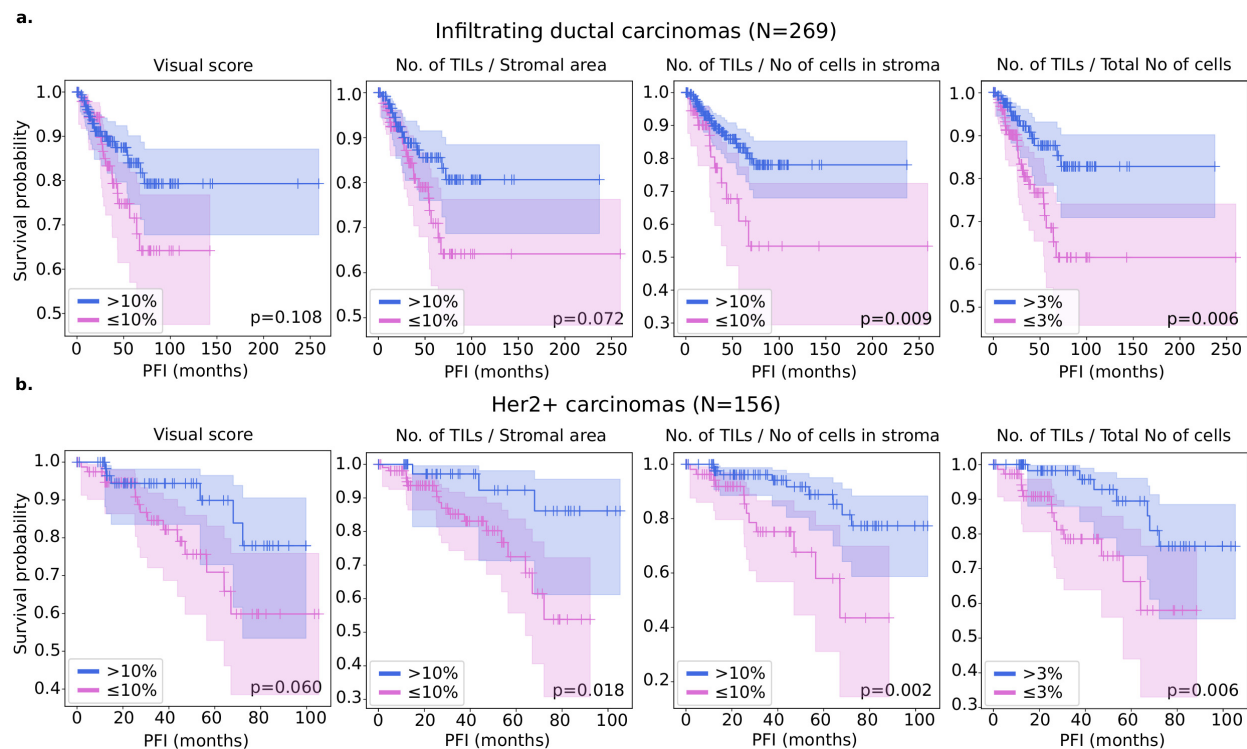


Figure 3.4.6: **Kaplan-Meier analysis of visual and computational TILs assessment in predicting breast cancer progression.** A threshold of 10% was used for visual and calibrated computational scores consistent with some of the research literature. Note that there is no recommended threshold for stromal TILs scoring, and so these results should be considered along with continuous results used in Cox regression modeling. For comparison, we also included a metric that looks into the predictive value of TILs when the denominator includes all cells, not just those in the stromal compartment. All metrics were obtained by weighted averaging of computational scores from 300 ROIs.

scores.

3.4.4 Discussion

MuTILs is a concept bottleneck model; it learns to predict the individual components that contribute to the TILs score (i.e., peritumoral stroma and TILs cells) and uses those to make the final predictions [92]. This setup makes its predictions explainable and helps identify sources of error.

The region constraint helped provide context for the nuclear predictions at high resolution, which helped reduce misclassification of immature fibroblasts and plasma cells as cancer (Figure 3.4.7). A qualitative examination of slides with discrepant visual and computational TILs scores shows there are three major contributors to discrepancies:

Table 3.4.2: **Cox regression survival analysis of the predictive value of visual and computational TILs scores for breast cancer progression.** The analysis was restricted to slides where visual TILs scores were available for a fair comparison. In the multivariable setting, each metric was part of an independent model along with clinically-salient covariates. We controlled all multivariable models for patient age and AJCC pathologic stage I and II status. Additionally, we controlled models using the infiltrating ductal carcinoma subset for basal genomic subtype status, and we controlled models using the Her2+ subset for infiltrating ductal histologic subtype status. Significant p-values are outlined in bold, using a significance threshold of 0.05. The * symbol indicates values < 0.001. Abbreviations used: HR, Hazard Ratio; 95%CI, upper and lower bounds of the 95% confidence interval; C-index, concordance index; No., number; Avg, weighted average.

Metric	Type	Univariable				Multivariable					
		HR	95% CI	P-value	C-index	HR	95% CI	P-value	C-index		
Infiltrating ductal carcinoma (N=269)											
Visual score		0.466	0.074	2.951	0.418	0.520	0.334	0.039	2.881	0.318	0.681
No of TILs / Stromal area	Global	*	*		0.287	0.548	*	*	*	0.321	0.667
No of TILs / No of cells in stroma	Global	0.098	0.004	2.711	0.170	0.546	0.081	0.002	3.428	0.188	0.670
No of TILs / Total No of cells	Global	0.078	*	16.98	0.353	0.526	0.073	*	29.87	0.393	0.667
No of TILs / Stromal area	ROI avg.	*	*		0.159	0.577	*	*	*	0.192	0.668
No of TILs / No of cells in stroma	ROI avg.	0.005	*	0.832	0.042	0.600	0.002	*	0.722	0.038	0.675
No of TILs / Total No of cells	ROI avg.	0.001	*	11.56	0.151	0.579	0.001	*	18.33	0.164	0.679
Her2+ carcinoma (N=156)											
Visual score		0.073	0.001	3.919	0.198	0.581	0.029	*	3.952	0.158	0.725
No of TILs / Stromal area	Global	*	*		0.039	0.644	*	*	*	0.011	0.816
No of TILs / No of cells in stroma	Global	*	*	0.201	0.015	0.673	*	*	0.057	0.007	0.813
No of TILs / Total No of cells	Global	*	*	0.719	0.045	0.621	*	*	0.001	0.007	0.800
No of TILs / Stromal area	ROI avg.	*	*		0.020	0.679	*	*	*	0.010	0.837
No of TILs / No of cells in stroma	ROI avg.	*	*	0.010	0.005	0.704	*	*	0.002	0.003	0.837
No of TILs / Total No of cells	ROI avg.	*	*	0.014	0.021	0.660	*	*	*	0.006	0.833

- Misclassifications of some benign or low-grade tumor nuclei as TILs.
- Variations in TILs density in different areas within the slide, which causes inconsistencies in visual scoring. This phenomenon is also a well-known contributor to inter-observer variability in visual TILs scoring [94].
- Variable influence of tertiary lymphoid structures on the WSI-level score.

Our results show that the most prognostic TILs score variant (nTnS) is derived from dividing the number of TILs cells by the total number of cells within the stromal region. The visual scoring guidelines rely on the nTnSa, which is reflected in the slightly higher correlation of the nTnSa variant with the visual scores compared to nTnS [155]. So why is nTnS more prognostic than nTnSa? There are two potential explanations. First, it may be that nTnS is better controlled for stromal cellularity since it would be the same in low- vs. high-cellularity stromal regions as long as the proportion of stromal cells that are TILs is the same. Second, nTnS may be less noisy since it relies entirely on nuclear assessment at 20x objective, while stromal regions are segmented at half that resolution.

Finally, we note that this validation was done only using the TCGA cohort, and future work will include validation on more breast cancer cohorts. In addition, we note that MuTILs has limited ability to distinguish cancer from normal breast tissue at low resolution, which may necessitate manual curation of the analysis region, especially for low-grade cases.

3.4.5 Conclusion

MuTILs is a lightweight deep learning model for reliable computational assessment of TILs scores in breast carcinomas. It jointly classifies tissue regions and cell nuclei at different resolutions and uses these predictions to derive patient-level TILs scores. We show that MuTILs can produce predictions that have good generalization for the predominant tissue and cell classes relevant for TILs scoring. Furthermore, computational scores are significantly correlated with visual assessment and have strong independent prognostic value in infiltrating ductal carcinoma and Her2+ breast cancer.

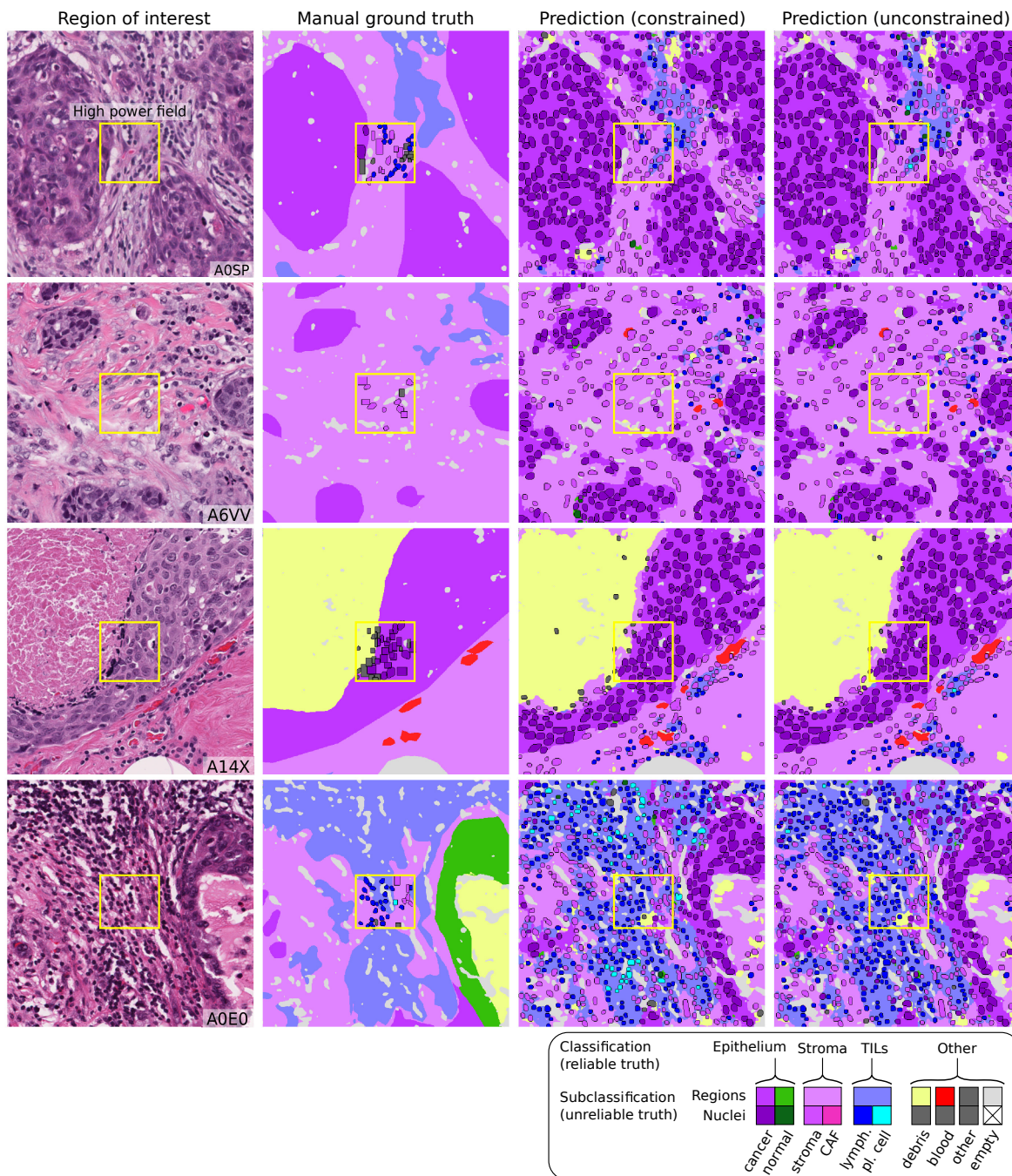


Figure 3.4.7: **Qualitative examination of sample testing set predictions and sources of misclassification.** The training dataset contained several subclassifications for region and nuclear data with unreliable or variable ground truth. Hence, we assessed performance at the level of grouped classes with reliable ground truth (tumor, stroma, TILs) at evaluation. The low representativeness of normal breast acini in training makes raw MuTILs predictions unreliable for differentiating normal and cancerous epithelial tissue (bottom row). This issue can be mitigated by expanding the training set or downstream modeling of architectural patterns, which is beyond the scope of this work. Note how the region constraint improves nuclear classifications (third vs fourth column). This improvement is most notable for large TILs (first row) and immature fibroblasts (second row), which are misclassified as cancer without the region constraint.

Chapter 4

Histopathologic correlates of clinical and genomic phenotypes

This chapter contains two explorative studies of morphologic histopathologic correlates of clinical and genomic phenotypes. First, we build on our CNN modeling work from Chapter 3 by representing each whole-slide image (WSI) scan by a small set of numerical features summarizing the morphologic appearance of computationally-delineated tissue regions and nuclei. These features are then used for downstream modeling of patient survival outcomes and genomic phenotypes of interest. The following sections are presented:

- *Histomic Prognostic Score: a computational morphologic signature with independent prognostic value in invasive carcinomas of the breast*
- Yang, X., **Amgad, M.**, Cooper, L. A., Du, Y., Fu, H., and Ivanov, A. A. (2020). High expression of *MKK3* is associated with worse clinical outcomes in African American breast cancer patients. *Journal of translational medicine*, 18(1):1–19.

Section 4.1

Histomic Prognostic Score: a computational morphologic signature with independent prognostic value in invasive carcinomas of the breast

4.1.1 Introduction

So far, the work presented in this dissertation has focused on the accurate delineation of tissue regions and nuclei from digitized WSI scans of H&E stained slides of invasive carcinomas of the breast. This section is a continuation of the work described in Section 3.4. After we had trained and validated the MuTILs model, we were able to automatically delineate all tissue regions and nuclei within WSIs. For each patient, we now had 1000+ tissue region boundaries and 100,000+ nuclei locations, boundaries, and classifications. These regions and nuclei were then used to extract a number of morphological and contextual features that we used for downstream prognostic modeling. This *concept bottlenecking* approach was introduced in Section 1.2, and is also summarized in Figure 4.1.1.

The main question we attempt to answer is: can we extract a morphological signature from WSI scans of H&E stained slides that are objective, accurate, and have equivalent prognostic value to manual Nottingham grading [12, 101]? There are two aligned but somewhat distinct objectives here. First, we wanted to extract morphologic features that correspond to the criteria used by pathologists in the day-to-day grading of invasive carcinomas. The main gains here stem from the objective and repeatable nature of computational algorithms. Second, we wanted to capture morphologic features that have prognostic value but are not captured by existing grading criteria. For example, Nottingham grading only relies on epithelial criteria without representing stromal or immune elements [12, 101]. Here, the main gains would be derived from gaps in the current assessment criteria. We rely on two datasets for this analysis. The first is called the Cancer Prevention Study II and is a long-term prospective cohort study organized by the American Cancer Society [27]. While all CPS-II participants were cancer-free at the time of recruitment, our analysis focuses on the survival outcomes of those participants who developed invasive breast cancer during the study period. CPS-II was used as our discovery cohort for model fitting. We used The Cancer Genome Atlas (TCGA) as our validation cohort.

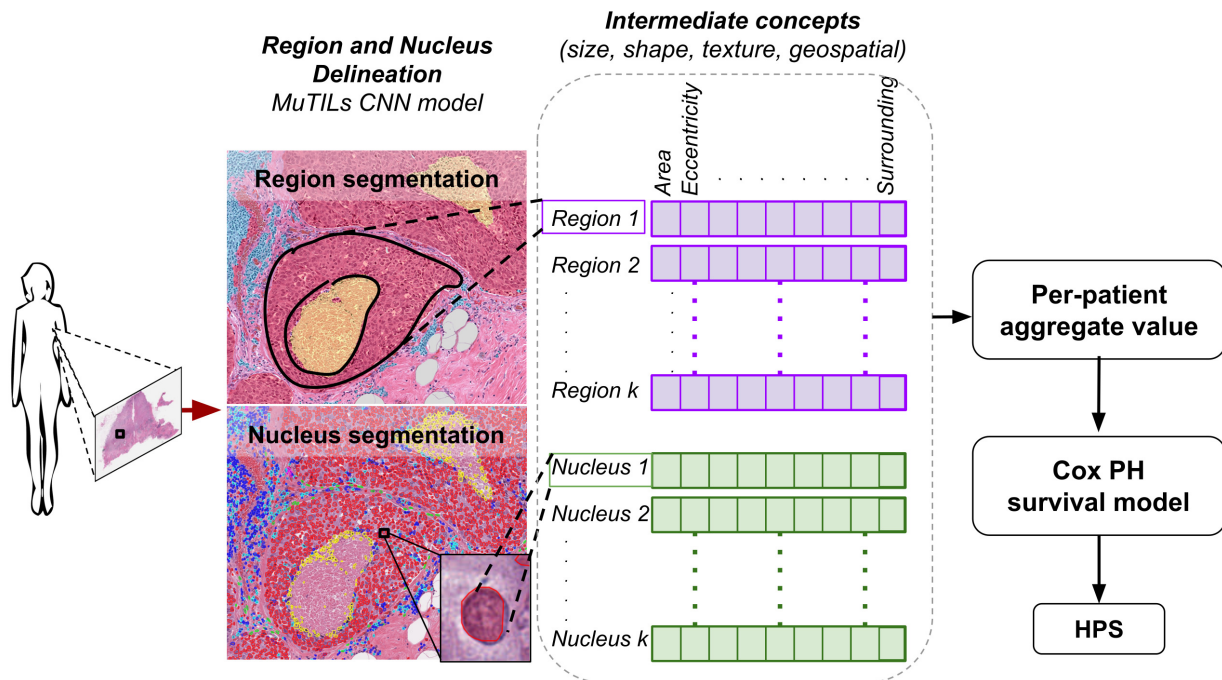


Figure 4.1.1: **Concept bottleneck modeling in invasive carcinomas of the breast.** The trained MuTILs model is used to delineate tissue regions and cells, which are then used to extract a number of morphological and spatial histomic features. These per-region and per-cell histomic features are aggregated using weighted mean and standard deviation to obtain patient-level histomic features that are used for downstream Cox Proportional Hazards (PH) survival modeling. The Histomic Prognostic Score (HPS) is a continuous score mapped to the 0-10 range using log partial likelihood predictions from the fitted survival model.

4.1.2 Methods

Detected regions and nuclei were summarised using a set of compact features that were used for later prognostic modeling. Figure 4.1.2 shows two sample histomic features we extracted: epithelial chromatin clumping and peri-fibroblast stromal matrix intensity heterogeneity (shorthand: *PeriFibroblMatrixHeteroIn512uMROI*).

Our extracted features can be broadly categorized into:

- *Global features*: Such as overall cancer cell density, overall TILs density, the global amount of necrosis within the WSI, and so on. All of these values were normalized to the amount of tissue analyzed.
- *Region morphology*: Standard size and shape measurements, extracted per region, and averaged for tissue regions that belong to the same class. These were extracted using the sklearn

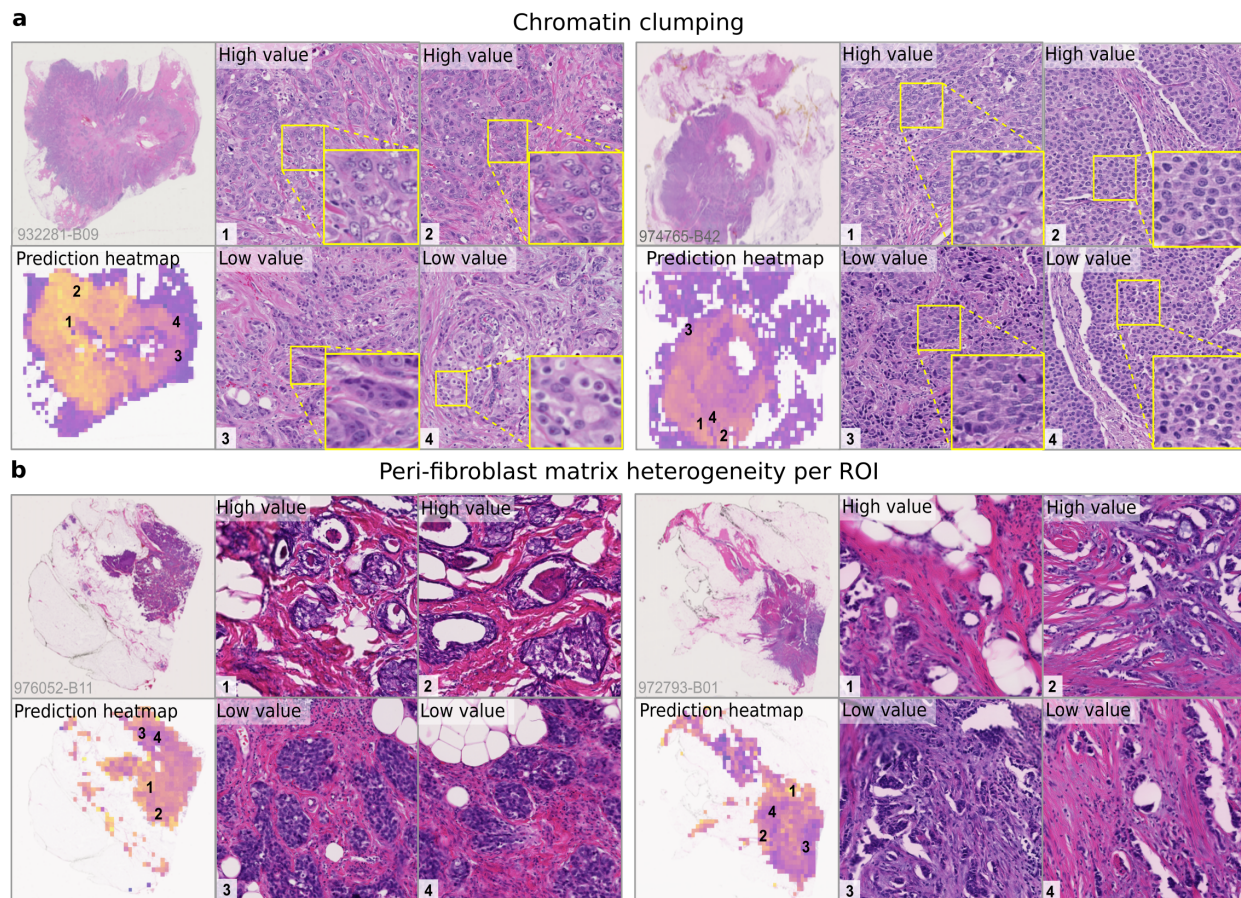


Figure 4.1.2: **Visualization of two prognostically important histomic features.** **a.** Epithelial chromatin clumping is one of the most prognostic features we found and is strongly and significantly correlated with manual nuclear grade. Our metric capturing *variation* of chromatin clumping is also prognostically important. Our computational measure of clumping relies on a Canny edge detector, with higher values indicating more edges inside the nucleus. Cancer cells are transcriptionally-active and haphazardly dividing, hence the clearing of some areas within the genome to transcribe specific gene subsets. **b.** Peri-fibroblast stromal matrix intensity heterogeneity per ROI, a highly prognostic histomic feature.

library’s *regionprops* method. Boundary complexity was measured using fractal (box counting) dimension, as implemented by Nicholas P. Rougier in this [Github repository](#) . We note that the MuTILs model detects contiguous TILs aggregates at a one micron-per-pixel (MPP) resolution. Hence, we did not need to use global clustering or graph-based methods to identify TILs cluster boundaries as they were determined in a data-driven manner using semantic segmentation.

- *Simple nuclear morphology*: Nuclear size, shape, staining intensity, boundary complexity,

edges (chromatin clumping) and texture were all extracted using the HistomicsTK standard library's *compute_nuclei_features* method [58]. The HistomicsTK implementation is largely based on sklearn library's *regionprops* method for the size and shape features, Canny edge detectors and Haralick features for texture, and fractal dimension for boundary complexity [64, 54, 58].

- *Deep nuclear morphology*: We observed, based on results presented in Sections 3.2 and 3.4, that nuclei do not always have typical morphology and that they are often ambiguous and difficult to classify in H&E images without IHC markers. Deep-learning models produce a classification probability vector, and we used those probabilities to capture the degree of conformity of nuclei to various classifications of interest.

TILs activation is the probability that a certain nucleus is classified as a plasma cell, divided by the overall probability that the nucleus belongs to the TILs superclass. We note that in the ground truth, the plasma cell class was not determined using IHC, nor was it limited to the most typical morphology. Hence, it refers to large TILs, including plasma cells and others.

Nuclear atypia is the probability that a certain nucleus is classified as a cancer cell divided by the total probability that it belongs to the epithelial class.

CAF epithelialization is the probability that a certain nucleus is classified as a fibroblast, divided by the sum of its fibroblast and cancer cell classification probabilities.

- *Cytoplasmic texture and staining*: We note that the delineation of cytoplasmic boundaries in H&E is not reliable, but we used the standard method of calculating texture statistics within 4 microns of nuclear boundaries. This search area is determined by dilating nuclear boundaries. This HistomicsTK library was used for extracting these features.
- *Local cell density*: These were defined as the average number of cells of a certain class within a predefined radius from the "central" cell. The central nucleus can have the same class as the surrounding nuclei; for example, the *LocalTILsDensity32uM* metric measures how many TILs are within 32 microns of the typical TILs cell? Alternatively, the central and surrounding can have different classifications; for example, the *TILsDensityWithin32uMOfEpithCell* metric

measures how many TILs are within 32 microns of the typical epithelial cell. This statistic was calculated using a fast K-D trees implementation, based loosely on the implementation by Sam P. Ingram in this [Github repository](#) [20].

- *Local cell clustering*: These were based on Ripley’s K-function at a single distance, which is a measure of clustering beyond that expected from random chance [145, 99]. We obtained this metric by normalizing local cell density estimates to ”complete spatial randomness.” For example, there is a higher chance that more lymphocytes will surround another lymphocyte by random chance, just because there are so many of them. Hence, high density does not necessarily indicate clustering beyond random chance. On the other hand, just a few fibroblasts surrounding each other may result in a high clustering value since they are (globally) less dense, so there is a lower chance of this dense local aggregation occurring by random chance.
- *Region composition*: Cellular composition of various histopathologic tissue compartments. For example, *NoOfLowGradeNucleiPerEpithNest* measures the number of epithelial nuclei that were considered low-grade by the MuTILs model, per epithelial nest. Region composition metrics also enabled us to estimate nuclei-to-cytoplasmic (N/C) ratio. As we mentioned earlier, cytoplasmic boundaries cannot be precisely determined in H&E stained slides (even visually), but we relied on a simple heuristic to calculate the N/C ratio: divide the total nuclear area within an epithelial nest by the overall area of the epithelial nest.
- *Region neighborhood composition*: Region masks were morphologically dilated to identify the tissue and cellular composition within a prespecified distance of the margin. For example, *CAFDensityAtEpithNestMargin* measures the density of CAFs within a 128-micron margin around epithelial nests.
- *Stromal matrix and collagen features*: Figure 4.1.3 illustrates some features that capture stromal matrix, including abstract texture and intensity-based analysis, as well as a more sophisticated analysis of the separation, length, and disorder of collagen fiber orientations. We captured collagen disorder by three separate approaches. First, we hypothesized that collagen separation and stromal matrix discoloration (e.g., desmoplasia) would reflect on

abstract intensity and texture measurements from the collagen stroma (upper-right panel). Peri-fibroblast stromal matrix intensity heterogeneity per ROI is a complex feature that captures the variation in stromal matrix caused by collagen fiber separation as well as the interface between desmoplastic and quiescent stroma. The metric is calculated by measuring the average intensity within a very thin rim around each fibroblast and calculating the variance in that intensity across a 512×512 μM ROI.

Second, we took a direct approach whereby we detected the collagen fibers themselves largely following the methodology described by Li et al. [102]. In brief, a Canny edge detection algorithm is used to detect the interface where collagen fibers separate from each other. Then we used connected component analysis to isolate individual edges and extracted standard morphological descriptors from each of these edges. Fibers that have a very small minor-to-major axis ratio were considered straight fibers (thickened in Figure 4.1.3), and these were further admitted to calculate the CFOD metric described by Li et al. [102]. This is a measure of the degree of disorder in collagen orientation, calculated from a length-weighted orientation co-occurrence matrix (bottom-right panels in Figure 4.1.3). Finally, we also measured collagen entropy indirectly by calculating the entropy of orientations of fibroblast nuclei within a certain radius of each other. We hypothesized that in some settings, fibroblast nuclei might be more reliably detected than collagen fibers.

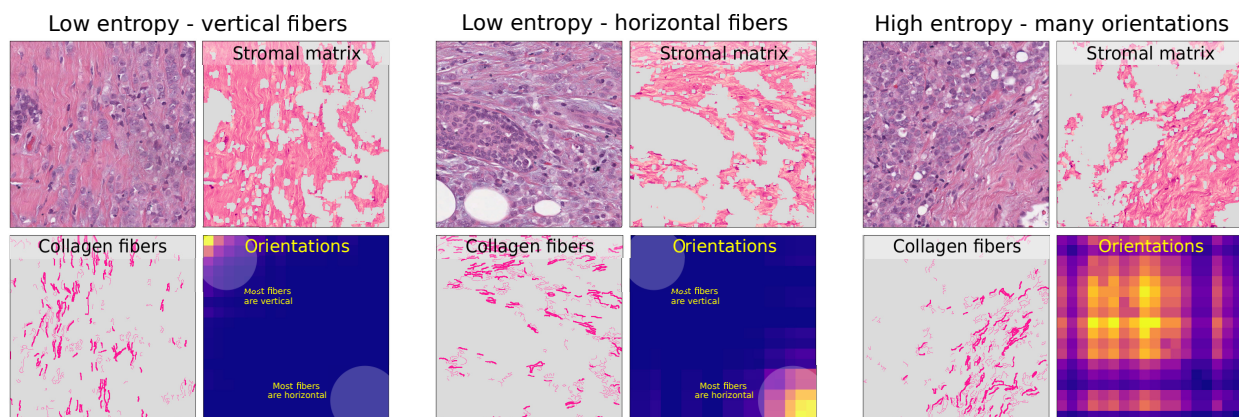


Figure 4.1.3: **Features capturing stromal matrix and collagen entropy.**

Thematic extraction of histomic features

Histomic features fall under five major themes, which are further divided into 26 sub-themes. Themes and sub-themes were engineered in a hypothesis-driven manner and are meant to capture distinct biological phenomena and processes. The themes are: epithelial features, stromal features, TILs features, necrosis abundance, and features capturing various interactions between different cells. Sub-themes encompass specific phenomena. For example, the TILs theme is subdivided into: TILs abundance, TILs clustering, TILs cluster morphology, TILs cluster pleomorphism, and so on.

We expect histomic features within the same themes and sub-themes to be correlated with each other and less correlated with other themes. The exception, of course, is the Interactions theme, which may or may not be correlated with others. On examination of the absolute correlation matrix, we can see that the correlation is stronger towards the diagonal, within the squares representing themes and sub-themes (Figure 4.1.4). The off-diagonal correlations are visibly stronger within the TCGA cohort, likely because it is mostly composed of high-grade, advanced cases. Hence the variation and distinctiveness of biological phenomena are less in TCGA. This higher variability of cases is the reason we chose CPS-II as our training cohort for the prognostic modeling. As we will discuss later, feature selection is made within each biological sub-theme; the most prognostic feature within each sub-theme on univariable analysis is used for further prognostic modeling.

One of the themes we focused on is characterizing the cancer-associated stroma since standard Nottingham grading does not capture it. Stromal features include morphological descriptors of fibroblast nuclei, characterization of TILs, and detailed analysis of the stromal matrix, which is primarily composed of type I Collagen fibers (Figure 4.1.3).

All of these measurements were assessed as potential prognostic indicators, and the most prognostic ones were admitted into the final prognostic score.

Histomic prognostic model fitting

The most prognostic feature from each biological sub-theme at the univariable level was admitted into this model, along with the standard IHC marker panel: Estrogen Receptor (ER) expression, Progesterone Receptor (PR) expression, and Her2+ overexpression (Figure 4.1.5). Additionally, we created a composite metric for triple-negative status (TNBC), defined as the absence of all standard

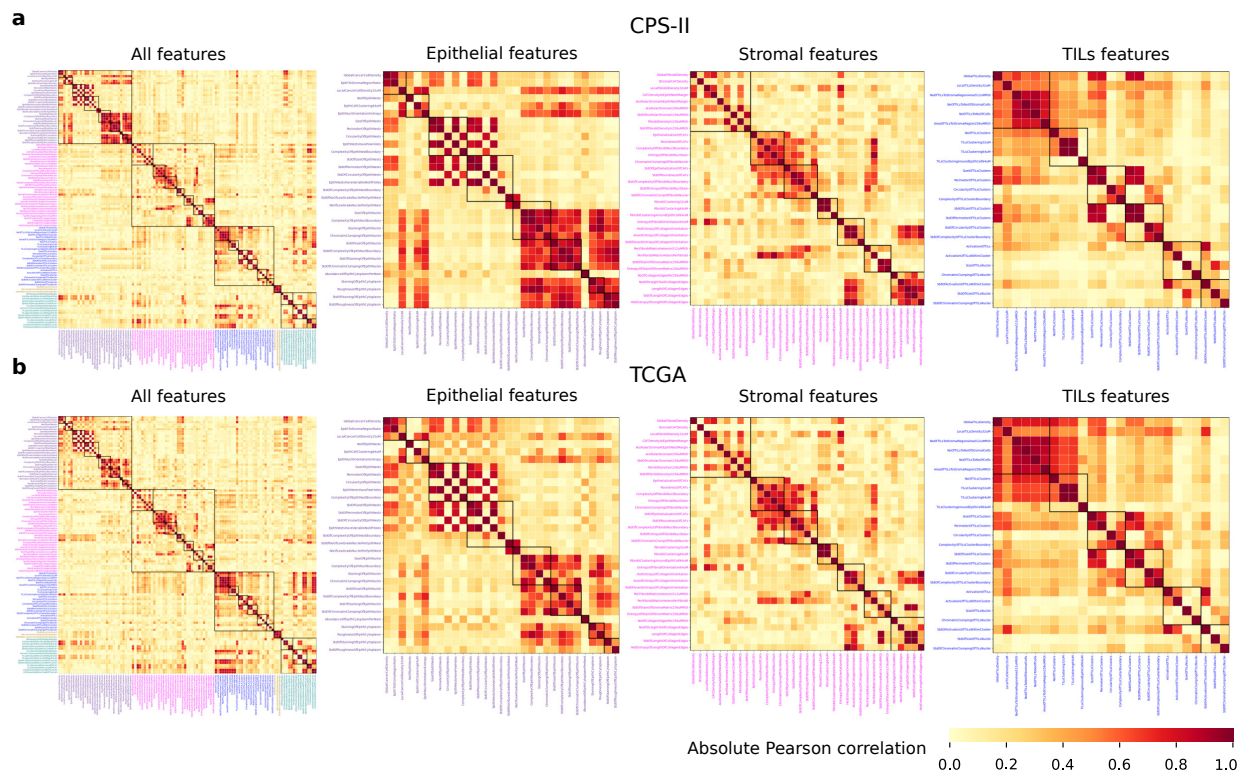


Figure 4.1.4: **Correlation between histomic features, by biological theme and sub-theme.**

markers. The rationale for incorporating IHC markers is to attempt to find histomic features that provide excess prognostic values beyond that already defined by the expression of hormone and Her2 receptors. We did not want to learn histological features that were very highly correlated with, say, TNBC status since they would not be clinically helpful. Virtually all breast cancer patients undergo standard breast panel testing, so in a clinical setting, the practicing pathologist and oncologist always have access to at least the histological slide and standard IHC markers [101, 12]. A total of 30 features (26 sub-themes and 4 IHC panel markers) were entered into an elastic-net regularized Cox proportional hazards survival model, predicting breast cancer-specific survival with the CPS-II patient cohort. The optimal hyperparameters (alpha and beta) for model regularization were obtained by cross-validation. The trained model was then used to predict the log partial hazard for the entire training population, and the predictions were binned into ten equal intervals. The end result is a continuous score that ranges from 0 to 10, where 10 is the highest risk of death. Additionally, we modeled the predicted risk scores as a mixture of three Gaussian curves, representing the low-, intermediate- and high-risk populations. The points where curves

cross were then considered to be data-driven cut-off for dividing patients into three risk groups.

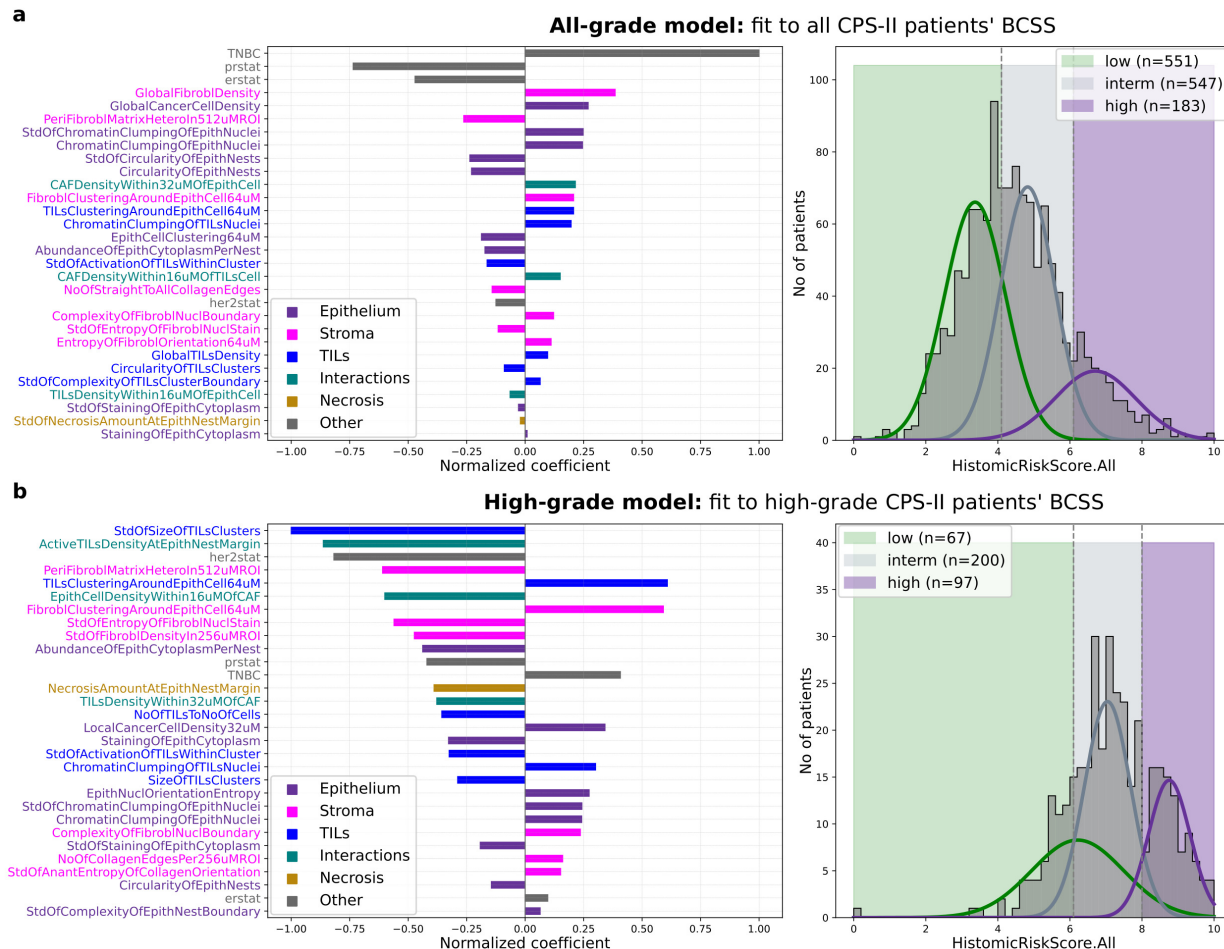


Figure 4.1.5: **Model fitting to obtain the Histomic Prognostic Score and groups.** **A.** The All-grade model is fit to the entire discovery cohort from the Cancer Prevention Study II. **B.** The *High-grade model* fit only to the high-grade CPS-II cases, based on Nottingham grades from clinical records.

The manual Nottingham grade is a semi-quantitative metric that is assessed by pathologists through visual examination of three epithelial features: tubule formation, nuclear atypia and pleomorphism, and mitotic figure count in hotspot regions [12, 101]. Standard of care also includes assessment of the standard IHC panel (ER, PR, and Her2+ expression status). To ensure a fair head-to-head comparison of our Histomic Prognostic Score and Histomic Prognostic Group, we used the exact same methodology to develop a baseline model fit only to the standard panel available in day-to-day practice (Figure 4.1.6). This baseline model also yields a risk score in the range 0-10 and three risk groups. Since Nottingham grading is discrete, unlike our histomic features,

which are continuous, the resulting histogram of predicted risks contains discrete bins, so it could not be faithfully represented by a mixture of Gaussian. Instead, we divided the score range into three equal proportions. Nottingham grade 1 and triple-negative status were the most important contributors to this model.

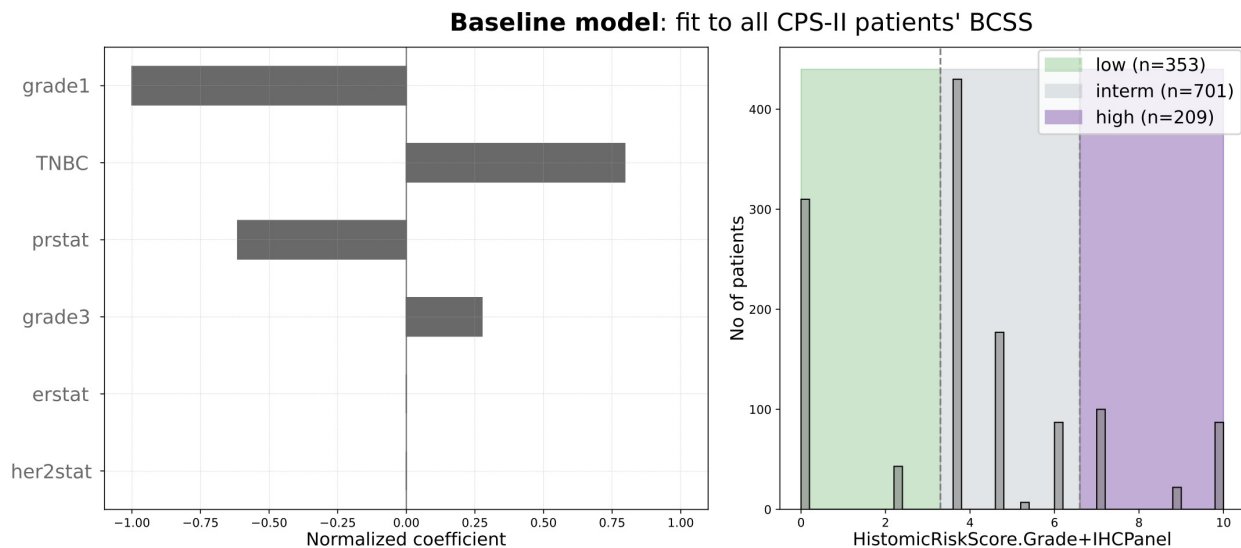


Figure 4.1.6: **Model fitting to obtain the baseline model based on manual Nottingham grading and the standard IHC panel.**

Hence, we had three sets of weights that we learned using the CPS-II cohort:

- *All-grade model*: This is a learned weighted combination of 26 histomic features and 4 IHC features based on the entire CPS-II patient cohort.
- *High-grade model*: This is a learned weighted combination of 26 histomic features and 4 IHC features based on high-grade CPS-II patient cohort. The high-grade designation is based on Nottingham grading obtained from pathology reports.
- *Baseline model*: This is a learned weighted combination of pathologist-determined Nottingham grade and 4 IHC features based on the entire CPS-II patient cohort.

Multi-variable Cox Proportional Hazards regression

After we learned the optimum combination of histomic features comprising the All-grade, High-grade, and Baseline models, and the optimum thresholds to learn discrete risk groups, we produced the following features for each patient:

- *Histomic Prognostic Score (HPS)*: This is a number ranging from 0-10, where 10 is the highest risk. If the patient's clinical record indicated that they had a high-grade (i.e., advanced) cancer, they had two HPS scores, one from the All-grade model and one from the High-grade model.
- *Histomic Prognostic Group (HPG)*: One of three risk groups, low, intermediate, and high-risk. If the patient's clinical record indicated that they had high-grade cancer, they had two HPG assignments corresponding to the two HPS values.
- *Baseline Risk Score*: A risk score in the range 0-10 using the baseline model.
- *Baseline Risk Group*: One of three risk group assignments based on the baseline risk score.

In the CPS-II cohort, we fit multivariable models to predict Breast Cancer-Specific Survival (BCSS), while in the TCGA, we fit the models to predict Overall Survival (OS). There was missing clinical data in both cohorts, so we explored two multivariable models. The first only controls for pathologic stage and tumor size and is a robust model with maximal sample size. We fit another model using a smaller set of patients with complete clinical information on pathologic stage, tumor size, whether the cancer was detected using proactive screening (in CPS-II), menopausal status at diagnosis, race, smoking history (in CPS-II), age at diagnosis, body mass index (for CPS-II), expression of basal markers CK5/6 or EGFR (for CPS-II) and genomic subtype (for TCGA).

4.1.3 Results

Epithelial histomic features are associated with Nottingham grades

We extracted a number of morphological and contextual descriptors that capture the tubule formation component of Nottingham grading criteria for breast cancer. We restricted this analysis to invasive ductal carcinomas to avoid confounding by special histological subtypes with distinct architectures. We obtained the detailed histological grades for TCGA patients from Ping et al. [139], which included the tubule formation of the Nottingham grade, parsed from the pathology report by three independent pathologists. The CPS-II cohort only had the overall grade, which was used here. Many measurements have a significant correlation with manual grading, and the strength of association is notably variable by measurement and cohort. We noted that the association was

weaker within the CPS-II cohort, likely because the overall grade is impacted by nuclear morphology and mitotic figures, not captured by the architecture-focused metrics we examine here. Notable associations include: 1. Epithelial cell clustering is negatively associated with grade, as high-grade cancers have cells that are spreading out and less likely to form well-defined acini. 2. Size and number of holes within the epithelial nests are positively associated with grade. Low-grade cancers are smaller and only have one hole (the central gland or duct lumen), whereas high-grade cancers do not have a well-defined structure and therefore have many holes. 3. The *variation/pleomorphism* in the size of acini and number of holes within them is also positively associated with grade.

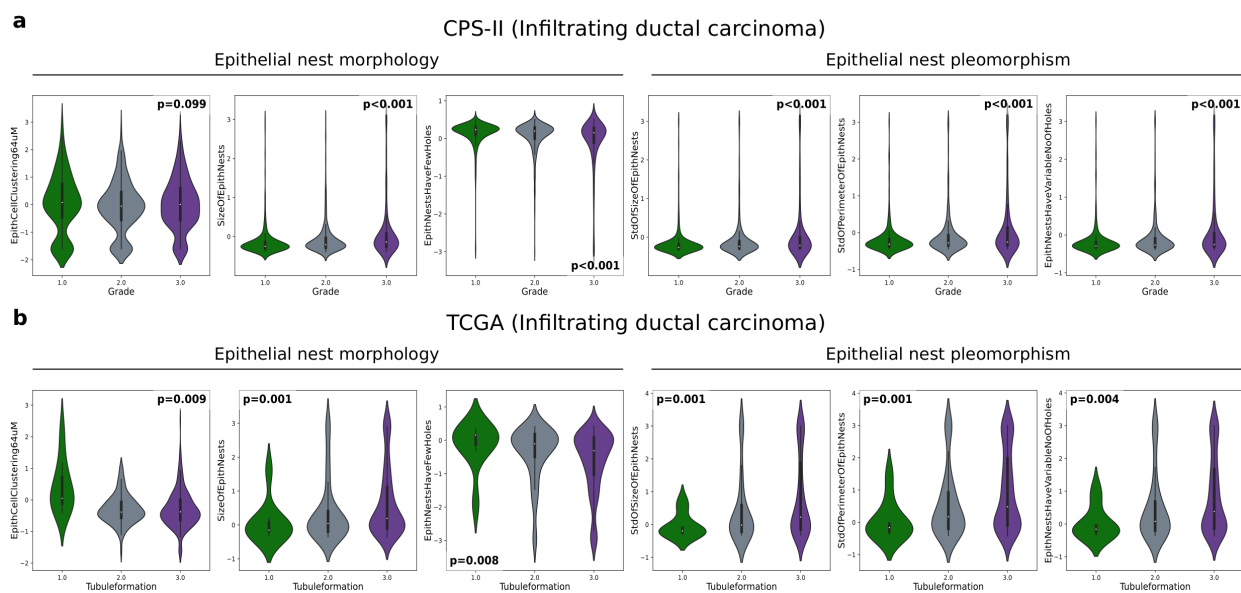


Figure 4.1.7: Histomic epithelial architecture measurements are associated with Nottingham grades. Note that the distributions are not Gaussian, and tend to be bi-modal since many cancers have a minority of normal or low-grade acini. Therefore, we used the non-parametric Wilcoxon rank-sum test for reporting p-values.

We extracted a number of morphological and contextual descriptors that capture the nuclear grade component of Nottingham criteria for breast cancer. We obtained the detailed histological grades for TCGA patients from Ping et al. [139], which included the nuclear component of the Nottingham grade, parsed from the pathology report by three independent pathologists. The CPS-II cohort only had the overall grade, which was used here. There are many histomic features with a significant association with the nuclear grade, and the strength of association is stronger than architectural features within the CPS-II cohort. As with the architectural features, many of the distributions are bi-modal due to normal or low-grade epithelial acini/nests. Since MuTILs can al-

ready distinguish between normal and cancerous cells, we assigned each epithelial nucleus an *atypia score* as the probability it is a normal epithelial cell, divided by the probability it is an epithelial cell. The number of nuclei with a low atypia score per epithelial nest was strongly and significantly associated with nuclear grade in both the CPS-II and TCGA cohorts. Other histomic measures of atypia, including nuclear size and chromatin clumping, were significantly positively associated with nuclear grade. Additionally, we found that histomic measures of nuclear pleomorphism were significantly associated with nuclear grade, including variation in nuclear size, staining, and chromatin clumping.

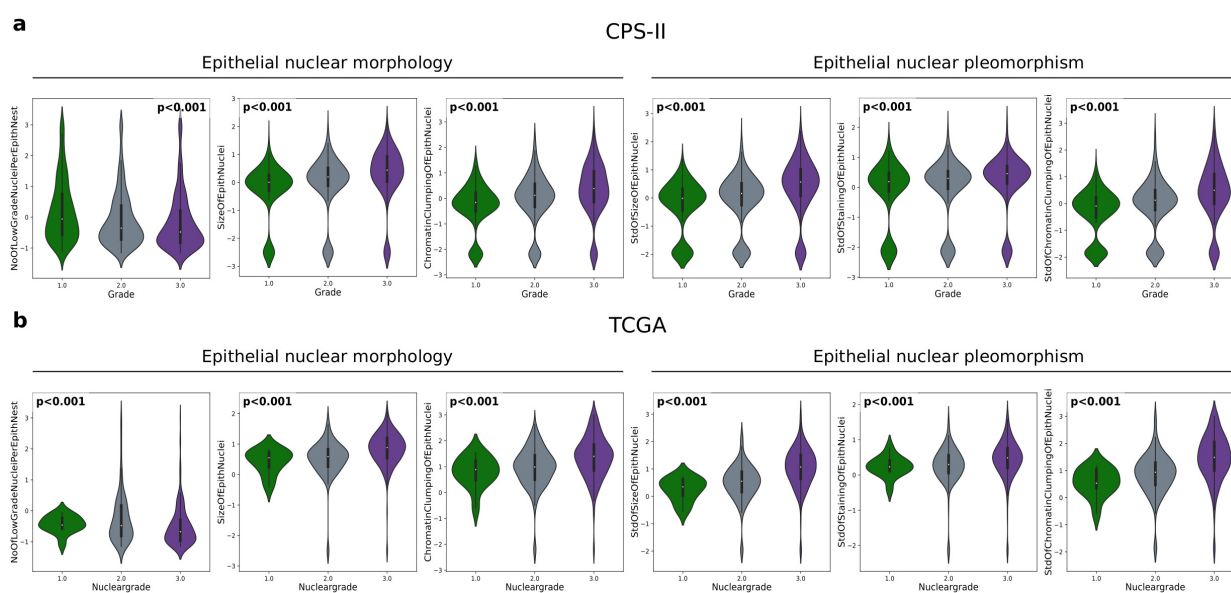


Figure 4.1.8: **Histomic epithelial nuclear measurements are associated with Nottingham grades.**

Computational scores are correlated with visual scores from pathologists

Unlike the results shown in Figure 3.4.5, where only the top 300 ROIs were assessed per slide for efficiency, this analysis looked at the entire slide. One practicing pathologist provided the TILs scores for the TCGA cohort, and another provided the scores for the CPS-II cohort. We extracted many histomic features capturing TILs; we selected three that capture TILs abundance in various ways. Specifically, we measured the number of TILs, divided by the stromal region area (nTSA), the number of TILs to the total number of stromal cells (nTnS), and the local TILs density. The first two measurements were explored and discussed in Section 3.4. Local TILs density is defined as

the number of TILs within 32 microns of the typical TILs cell. Hence, this measure is of the density within areas where TILs already exist. All of these measurements had a significant but modest correlation with the manual scores. Within the TCGA, nTSA had a stronger correlation than the other two measurements; it is the metric most closely capturing the clinical recommendations. In CPS-II, however, local TILs density had a stronger correlation with visual scores. Note that all visual and histomic metric distributions were skewed since more patients had low TILs infiltration, hence our usage of the non-parametric Spearman's correlation metric.

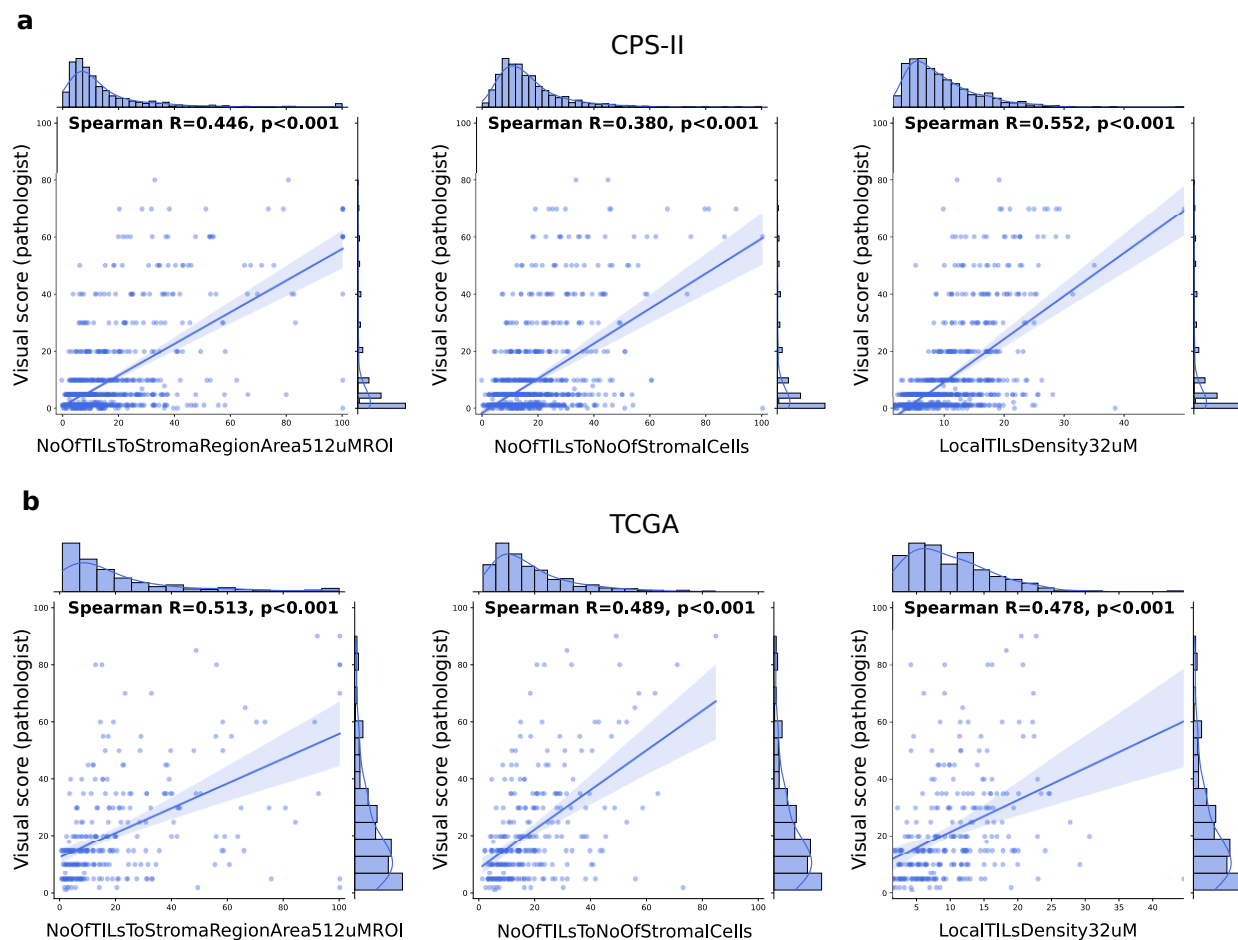


Figure 4.1.9: Association between histomic TILs measurements and pathologist TILs scores.

Most prognostic histomic features

Figure 4.1.10 shows the univariable Cox proportional hazards model coefficients of histomic features on the discovery cohort (CPS-II). The four biggest themes, tumor, stroma TILs, and interac-

tions are shown here. The coefficients shown here are from univariable Cox proportional-hazards models predicting breast cancer-specific survival within the CPS-II cohort. Since these models do not control for critical factors like hormone receptor status and Her2 over-expression, they should be interpreted with caution. Nevertheless, we used these univariable coefficients to pick the top histomic feature per biological sub-theme to be admitted into the histomic prognostic model described later. The top three features within the epithelial theme were: circularity of epithelial nests, chromatin clumping of epithelial nuclei, and variation of circularity of epithelial nests. The top three stromal features were: global fibroblast density, peri-fibroblast stromal matrix intensity heterogeneity within the ROI, and density of cancer-associated fibroblasts (CAF) at the epithelial nest margin. The top 3 prognostic TILs features were: Chromatin clumping of TILs nuclei, TILs clustering within 64 microns of epithelial cells, and variation in the activation of TILs within TILs aggregates. *Activation* of TILs is a term we use for a deep-learning score we obtain for each TILs cell, capturing how closely it resembles large TILs with plasma cell-like morphology. The top 3 interaction features were all related to the density of Cancer-Associated Fibroblasts (CAF) within various radii from epithelial and TILs cells.

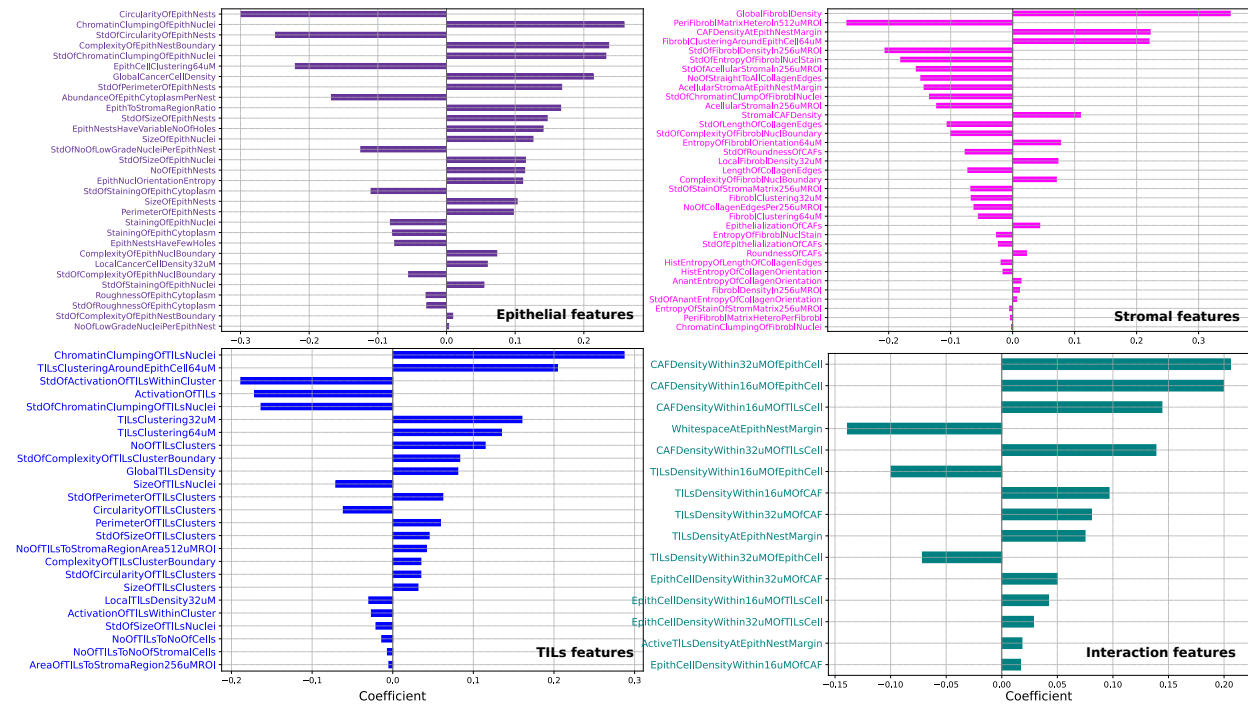


Figure 4.1.10: Univariable coefficients for histomic features for BCSS on CPS-II.

Figure 4.1.5 shows the multivariable histomic model coefficients, where the top feature from

each of the 26 sub-themes are controlled for each other, along with the expression of standard IHC markers ER, PR, and Her2. The most prognostic features were different for the general patient population compared to the high-grade cohort.

Within the general population, the standard IHC panel had the highest prognostic value, with triple-negative (TNBC) status as the most adverse prognostic features, followed by ER and PR expression. The top adversely prognostic histomic features were global fibroblast density, global cancer cell density, and chromatin clumping of epithelial nuclei (and its variation). The most protective prognostic indicators were peri-fibroblast stromal matrix heterogeneity and circularity of epithelial nests (and its variation).

Epithelial features have a less important prognostic role within the advanced, high-grade cases. Instead, the top features within this cohort were variation in the size of TILs clusters, the density of large/”active” TILs at the epithelial nest margin, peri-fibroblast stromal matrix heterogeneity, and TILs clustering within 64 μM of epithelial cells.

Distribution shift between the CPS-II and TCGA datasets

We wanted to quantitatively investigate the degree of dissimilarity between the two cohorts in terms of the histomic features comprising the histomic prognostic score. To do so, we Z-score normalized both cohorts relative to CPS-II and examined the feature histograms. Then we calculated the Kullback–Leibler (KL) divergence between CPS-II and TCGA histograms. The larger this value, the most dissimilar are the two distributions. Three epithelial features that are influential on the prognostic score also exhibit a high degree of dissimilarity between the CPS-II and TCGA patient cohorts. Figure 4.1.11 shows differences in distributions between the CPS-II and TCGA cohorts for histomic features comprising the all-grade model. For global cancer cell density, many CPS-II cases are concentrated just below the mean, whereas most TCGA cases are to the right of the CPS-II mean. The same is true for chromatin clumping and variance of chromatin clumping. In contrast, Figure 4.1.12 shows differences in feature distributions between the CPS-II and TCGA cohorts for histomic features comprising the high-grade model. Since the high-grade model was trained on high-grade CPS-II cases, it learns features that are can differentiate advanced cases like those comprising most of the TCGA cohort. Notice that epithelial features are much less influential in this model. TILs clustering and interactions emerge as top contributors to the score.

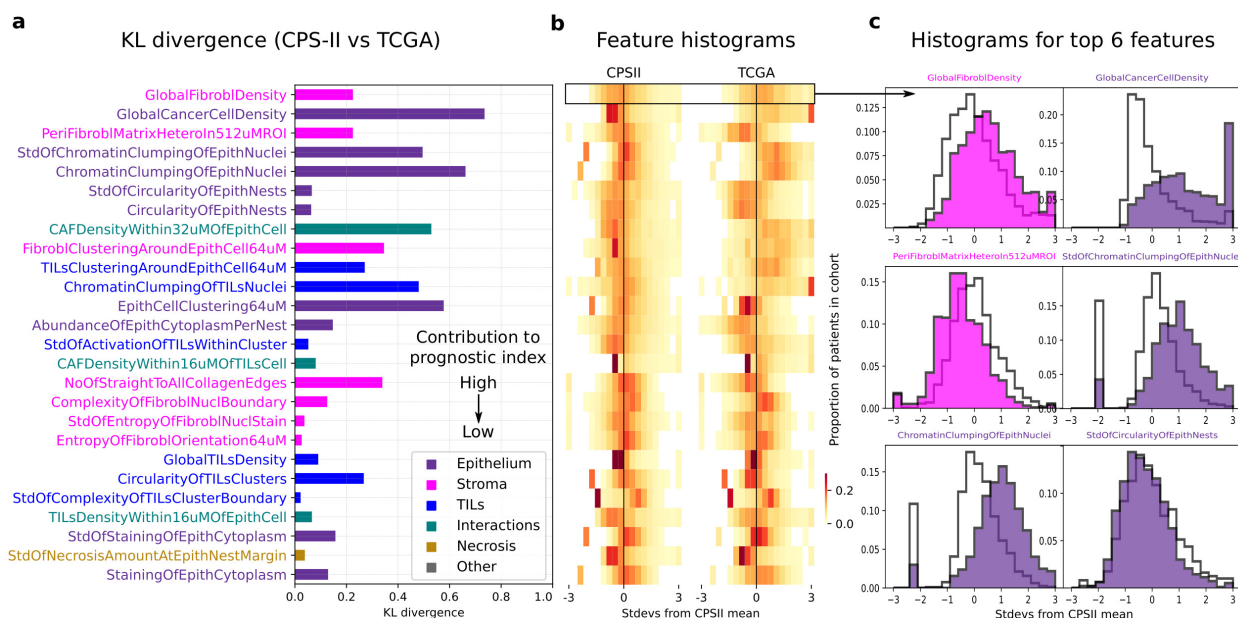


Figure 4.1.11: Differences in feature distributions between the CPS-II and TCGA cohorts for histomic features comprising the All-grade model. **a.** Kullback–Leibler (KL) divergence between CPS-II and TCGA histograms. **b.** A heatmap encoding the feature histograms for both patient cohorts. Note the difference in density in features where KL divergence is high. **c.** Histograms for the top six histomic features within the Histomic Prognostic Score.

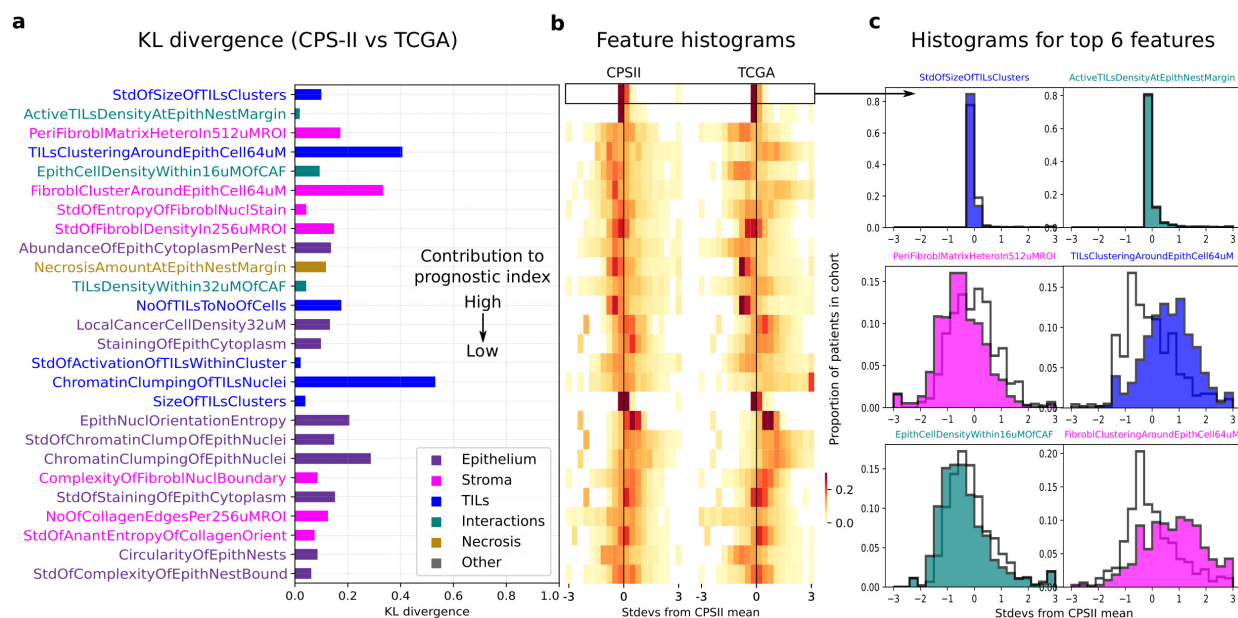


Figure 4.1.12: Differences in feature distributions between the CPS-II and TCGA cohorts for histomic features comprising the High-grade model. This figure is organized the same way as Figure 4.1.11, but measures dissimilarity using histomic features from the high-grade model instead.

Histomic Prognostic Groups stratify patients better than the baseline model

Within the CPS-II cohort, Histomic Prognostic Groups from the All-grade model can stratify patients into three subgroups better than both grading alone and the baseline model (grading + IHC) (Figure 4.1.13). This is evidenced by the better separation of KM curves and the larger log-rank test statistic. This stratification improvement is driven in large part by re-assignment of patients identified as intermediate risk by the baseline model into either the low- or high-risk groups (Figure 4.1.14).

Moreover, The histomic model is able to identify low-risk patients that were classified as grade 3 by the pathologist. Again, the stratification is better than the baseline model using manual grading and standard IHC markers.

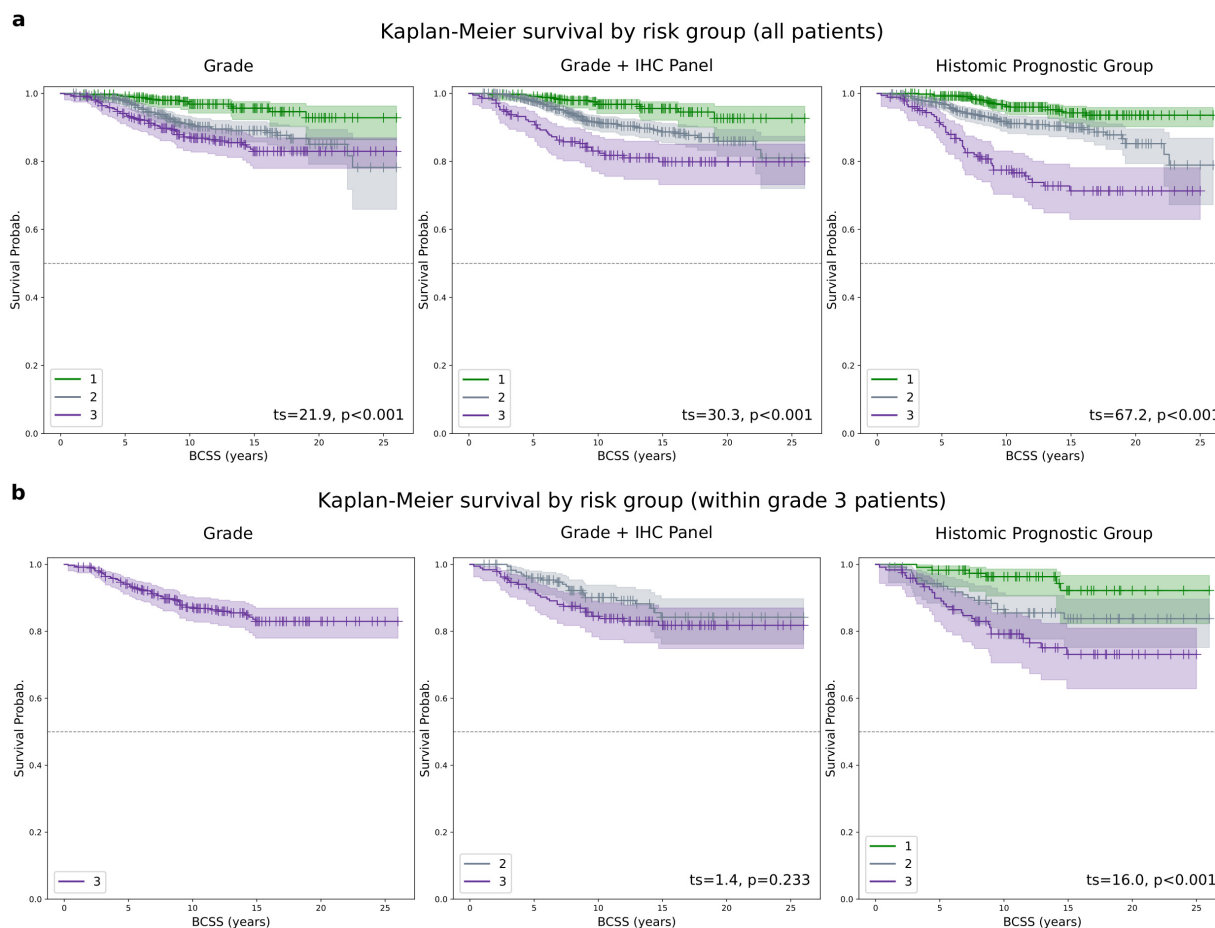


Figure 4.1.13: Kaplan-Meier survival curves for the CPS-II cohort, using Histomic Prognostic Groups from the All-grade model.

TCGA patients were harder to subdivide into different cohorts through manual grading or the



Figure 4.1.14: **Sankey plot showing risk group changes if Histomic Prognostic Grades are adopted.** These are the changes in risk grouping of patients responsible for prognostic gains shown in the Kaplan-Meier plots above, and later in the multivariable survival models.

baseline model. The intuitive explanation is that most TCGA breast cancer cases are advanced, so further subdivisions would be difficult. That said, the Histomic Prognostic Grouping successfully stratifies TCGA patients into three prognostically-distinct groups, although the p-value was not significant (Figure 4.1.15). It should be noted that TCGA has a higher censorship rate than the CPS-II cohort.

Histomic Prognostic Score has independent prognostic value

The Histomic Prognostic Score is prognostic in both the CPS-II and TCGA cohorts and consistently beats the baseline model. Figure 4.1.16 shows the minimal multivariable model, controlled only for stage and tumor size to maximize robustness and sample size. In CPS-II, the Histomic Prognostic Score has a stronger independent prognostic value than the baseline model. Since the score ranges

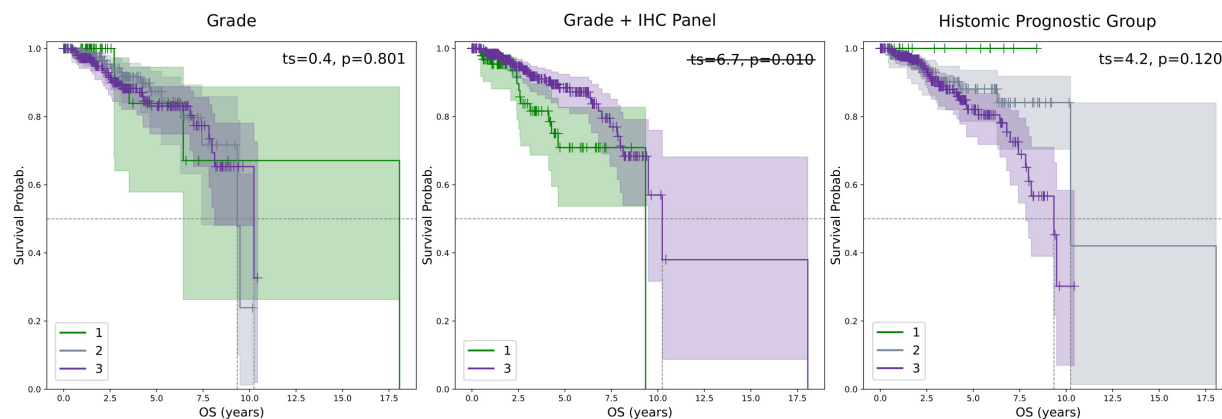


Figure 4.1.15: Kaplan-Meier survival curves for the TCGA cohort, using Histomic Prognostic Groups from the high-grade model.

from 0-10, the hazard ratio can be interpreted as follows: every unit increase on the risk score is associated with a 15%-45% increase in the risk of death from breast cancer. Discretized Histomic prognostic groups have independent prognostic value, but the baseline model is stronger. The Histomic Prognostic Score also has an independent prognostic value in TCGA; every unit increase is associated with a 10%-67% increase in the risk of death. None of the other measures had prognostic value in TCGA, including the Histomic Prognostic Groups and the baseline model scores and groups.

Figure 4.1.16 shows a more comprehensive multivariable model, controlled for a wide range of covariates. In CPS-II, the Histomic Risk Score has a stronger independent prognostic value than the baseline model. The risk groups have independent prognostic value but are roughly equivalent to the control model. The Histomic Prognostic Score also has an independent prognostic value in TCGA; every unit increase is associated with a 3%-71% increase in the risk of death. None of the other measures had prognostic value in TCGA, including the Histomic Prognostic Groups and the baseline model scores and groups.

The issue of confounding by treatment is complex and is further discussed below. Neither the CPS-II nor TCGA datasets have granular treatment information to allow for robust control for this issue. Where available, the treatment information is coarse and only available for a small subset of patients, substantially reducing the sample size.

Nonetheless, we did examine the effect of coarse treatment variables on the CPS-II cohort in a separate multivariable model. The variables admitted into that model included all variables

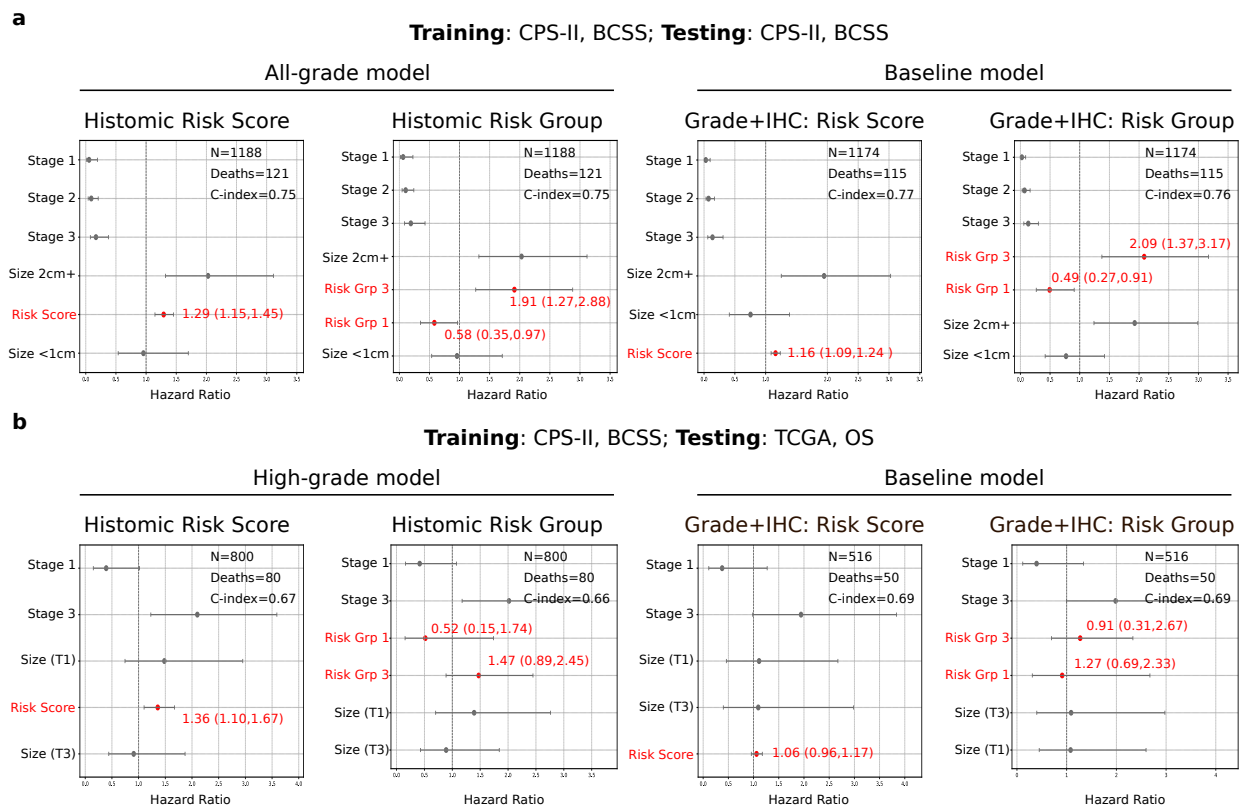


Figure 4.1.16: **Multivariable Cox PH of Histomic Prognostic Scores and Groups (v1).** This version only controls for cancer stage and tumor size to maximize sample size and model fit. CPS-II was always used as the training cohort for learning the All-grade, High-grade and Baseline model parameters.

from Figure 4.1.17, along with three indicator variables for targeted therapy, chemotherapy, and radiotherapy. Inclusion of the treatment variables results in a sample size reduction to 727 patients (72 events). Controlling for all these factors, HPS was independently prognostic (HR=1.38; 95%CI: 1.14,1.66; $p<0.001$). HPG group 1 was not independently prognostic (HR=0.55; 95%CI: 0.27,1.11; $p=0.093$), unlike HPG group 3 which was independently prognostic (HR=1.78; 95%CI: 1.01,3.12; $p=0.045$).

For comparison, the baseline risk score was also independently prognostic but had a lower hazard ratio than HPS (HR=1.23; 95%CI:1.10,1.39; $p<0.001$). Like the histomic groups, baseline group 1 was not independently prognostic (HR=0.63; 95%CI: 0.30,1.33; $p=0.227$), unlike baseline group 3, which has a stronger independent prognostic value than the histomic model (HR=2.64; 95%CI: 1.32,5.27; $p=0.006$).

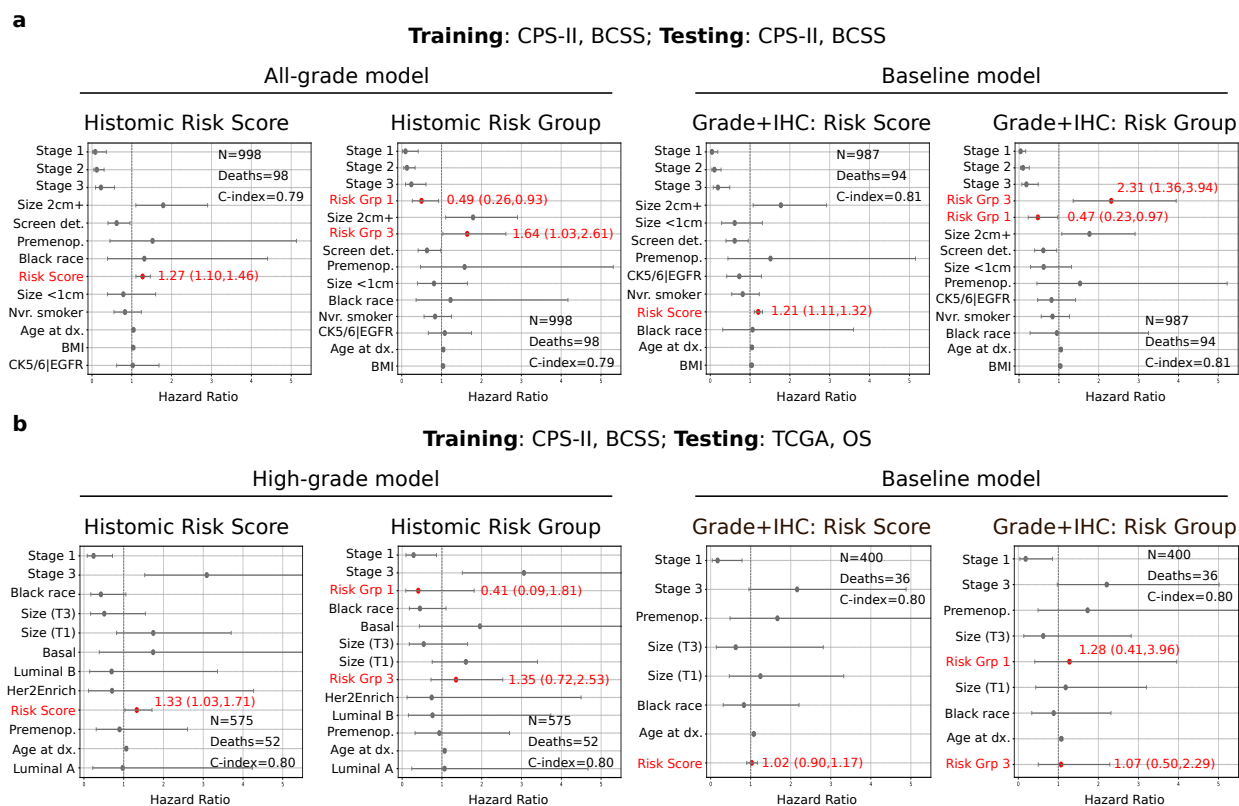


Figure 4.1.17: **Multivariable Cox PH of Histomic Prognostic Scores and Groups (v2).** This version was controlled for a wide range of covariates including stage, tumor size, whether the cancer was detected using screen (in CPS-II), menopausal status at diagnosis, race, smoking history (in CPS-II), age at diagnosis, body mass index (for CPS-II), expression of basal markers CK5/6 or EGFR (for CPS-II) and genomic subtype (for TCGA).

4.1.4 Discussion

We found that the All-grade model trained on all CPS-II cases was prognostic in TCGA but not as robustly and reliably as the high-grade model trained only on high-grade CPS-II cases. This makes intuitive sense since the TCGA cases are more advanced (mostly grade 2 and 3) and are almost entirely derived from tertiary care centers where advanced cases tend to get treated. CPS-II slides, on the other hand, are from patients in the general population who enrolled in a prospective cohort study and have a much more balanced representation of low and high-grade cases.

Future work may help elucidate the mechanistic biological basis for the prognostic value of our top histomic features. Some prognostic features have a well-known mechanism. For example, histomic measurements of epithelial architectural disruption, nuclear atypia, and nuclear pleomorphism are a more objective and quantitative measure of the standard visual clues that form the

basis of Nottingham grading [12, 139, 101]. Similarly, histomic features that measure TILs abundance provide a quantitative computational equivalent of the visually-assessed stromal TILs score, which has been extensively studied as an intuitive measure of the physical reach of immune cells to their site of action [155, 94, 11].

Some histomic features have plausible but somewhat speculative biological explanations. Within the high-grade patient cohort, we found a stronger favorable prognostic value of the variation in the area of TILs aggregates than of simple abundance measurements. This is consistent with recent work from other groups [125, 1]. TILs clustering may capture the interaction between various immune cells as they relay signals to coordinate the immune response [125, 1]. By design, the Histomic Prognostic Score relies on H&E stained images to ensure broad applicability to day-to-day clinical practice. Capturing different immune cell subsets such as CD4+ T-cells, CD8+ T-cells, B-cells, and macrophages requires systematic studies using immunohistochemistry (IHC) or immunofluorescence. Likewise, TILs' "activation" and morphological differences in TILs' nuclear appearance may indicate phenotypic changes within the same cell type to activate transcriptional programs, or they may reflect true differentiation into specialized cell types, such as B-cells differentiating into plasma cells [84, 35]. Several prognostic features also capture TILs density, activation, and clustering within a prespecified neighborhood of cancer cells, cancer-associated fibroblasts (CAFs), and epithelial nest margins. These immune features capture morphological changes within the local tumor microenvironment and have a more substantial influence within the high-grade cancer cases in our analysis [84, 35].

We found a strong adverse prognostic value of global fibroblast density within the general breast cancer population and CAF clustering within the high-grade cancer cases. These measurements are proxies for inflammation and wound healing. Consistent with Beck et al. and Abubakar et al., our results show a strong prognostic significance of stromal features, comparable to epithelial ones [19, 3]. Compared to these two previous important works, our analysis has the advantage of explicitly detecting TILs and excluding them from this "stromal cell" calculation, ensuring that the metric truly captures fibroblast/CAF density. Consistent with Yuan et al., we found that CAF clustering has an adverse prognostic value, especially in the high-grade cohort [200]. Some of our measurements also focused on phenotypic differences in the appearance of fibroblasts. There is a well-established body of literature documenting morphological changes in CAFs in response to var-

ious physical and biochemical signals within the tumor microenvironment [59, 106, 153]. We found that the complexity of fibroblast nuclear boundary to be adversely prognostic, especially within high-grade cases. Some of these morphologic changes may be capturing epithelial-to-mesenchymal transition (EMT) of leading metastatic cancer cells as they acquire a CAF-like morphology and transcriptional profile [106, 153, 63, 51, 201].

We note that the directionality of the prognostic value of the average measurement of histomic phenomena may or may not be consistent with the variation of measurement across the slide. For instance, both the size and variation in size of TILs clusters are protective in high-grade cases. In contrast, the complexity of fibroblast nuclear boundary is adversely prognostic, while variation in complexity is protective within the general patient cohort.

A few of the feature sub-themes focused on the characterization of the non-cellular component of the cancer-associated stroma, including measures of stromal matrix staining, collagen fiber separation, length, and orientation disorder (entropy). Some of these features measure very precise phenomena; for instance, we replicated the measure of collagen fiber orientation disorder from Li et al. [102]. One difference between our implementation and theirs is that we blocked out nuclei before applying the edge detection algorithm, which we believed was necessary to minimize confounding by nuclear material. While we found measures of collagen length and orientation to be somewhat prognostic, the most prognostic non-cellular stromal feature was the heterogeneity of peri-fibroblast stromal matrix intensity within a 512 uM region. This is an abstract measure that captures variation in intensity in neighboring stromal areas and is increased at the interface between quiescent and desmoplastic stromal regions, and when collagen fibers are separated from each other. Li et al. speculated that collagen fiber disorder may act as a physical barrier to slow the spread of cancer cells [102]. We found this metric had robust positive prognostic value in both the general cancer population and the high-grade population. This finding is consistent with seminal findings by Beck et al., who found that a related metric was a top prognostic stromal feature in their cohort [19]. Beck et al. relied on absolute differences in overall intensity between neighboring contiguous stromal regions, which may have been liable to some confounding by segmentation errors or non-stromal matrix elements like small vessels and vacuoles. To minimize this confounding, we relied on the peri-fibroblast stromal matrix within 4 uM. All images were color normalized using the Macenko method to maximize robustness to staining and scanner differences.

4.1.5 Limitations and future work

It is important to note that the interpretation of the prognostic value of histomic features is a function of more than just fundamental biology. Detection accuracy, and the robustness of algorithms in capturing the same phenomenon consistently, are also key considerations. Let's consider the prognostic value of variation in peri-fibroblast stromal matrix intensity. We consistently found this metric to be more prognostic than collagen fiber orientation, but is this improvement because we measure a distinct biological phenomenon that is more important or because pixel intensity is a robust abstract measurement with fewer moving parts? Further systematic exploration of this question can help disentangle these issues.

As Figure 4.1.1 illustrates, there are thousands of regions and hundreds of thousands of nuclei per patient. Each tissue region and nucleus is described by a set of morphological and spatial features, which are then aggregated using weighted mean and standard deviation per patient. This aggregation simplifies downstream modeling but results in the loss of potentially useful information. One of the questions we intend to address in future work is heterogeneity in histomic feature values within the same patient. A good aggregate learning model captures the complexity in individual nuclei and can learn high-order logic operations that may characterize how histology contributes to observed outcomes. For example, a combination of inflammatory infiltration and small tumor nest sizes may capture immune success, while inflammatory infiltrates alone without tumor size reduction may not. Likewise, small nests may encode a shrinking tumor if other favorable histologic or clinical characteristics are present, or may indicate an invasive tumor phenotype instead (budding). Sometimes, a very small subset of patterns can dramatically affect the overall observed phenotype; basement membrane invasion or angioinvasion, which are very subtle phenomena, indicate local and distant tumor invasion and worse outcomes [12, 101]. Different strategies have been proposed in the literature with variable success, including Multiple Instance Learning, Recurrent neural networks, recurrent attention models, and attention-based transformers, among others [196, 95, 28, 75, 26, 5, 120, 117, 187].

Another future avenue is the systematic exploration of genomic correlates of various histomic features, especially the most prognostic ones. One of the most important and distinctive aspects of the TCGA dataset is the availability of genomic, transcriptomic, epigenomic, clinical, and other

data for the same patients where WSI scans are available. This presents a unique opportunity to correlate the histomic feature data we extracted with the genomic records from the TCGA. These correlations will not only open new avenues for discovery but also help validate and understand the histomic features themselves. For instance, we may want to correlate the genomic measurements of wound healing with various stromal histomic features. We may correlate the expression of pathways related to EMT with CAF morphology and test the hypothesis that our measurements of CAF cellular and nuclear morphology are indeed related to EMT. It should be emphasized, however, that simple correlations do not necessarily imply or prove a causal biological chain of events. Genomic measurements are obtained from crushed tissue samples containing a heterogeneous mixture of cells, and the tissue sections used for genomic analysis are not the same ones used for diagnostic purposes [42, 40]. The ideal way to study these correlates is through hypothesis-driven experimentation, although simple correlative analyses can help generate hypotheses and point us in the right direction.

Finally, we would like to acknowledge the limitations associated with the retrospective nature of this analysis. After diagnostic slides were obtained, patients underwent treatment and various events until either the survival outcomes were observed or the patients were lost to follow-up. This period after diagnostic assessment, of course, has an impact on the outcomes and is not accounted for in our modeling. This limitation is not specific to our work, and virtually all research works in this niche suffer from this limitation. Nonetheless, retrospective exploratory analyses like ours are critical to making initial discoveries and observations and provide initial validation for identifying promising biomarkers for prospective randomized controlled trials, the golden standard [12, 25, 121, 180]. Prospective randomized controlled trials are expensive and logistically complex, and there is an ethical obligation not to enroll patients into a prospective trial unless there has been rigorous validation in pre-clinical studies, including retrospective data-driven analyses like the one we presented [121, 180].

Section 4.2

High expression of MKK3 is associated with worse clinical outcomes in African American breast cancer patients

This section is an exact reproduction of the following open-access paper:

*Yang, X., **Amgad, M.**, Cooper, L. A., Du, Y., Fu, H., and Ivanov, A. A. (2020). High expression of MKK3 is associated with worse clinical outcomes in african american breast cancer patients. Journal of translational medicine, 18(1):1–19.*

Candidate's role: Computational pathology component, including semantic segmentation of histologic regions, extraction of morphologic features, correlations with gene expression data (figure 2), and drafting corresponding methods and results.

RESEARCH

Open Access



High expression of MKK3 is associated with worse clinical outcomes in African American breast cancer patients

Xuan Yang^{1,2}, Mohamed Amgad³, Lee A. D. Cooper⁴, Yuhong Du^{1,2,5}, Haian Fu^{1,2,5,6*} and Andrey A. Ivanov^{1,2,5*} 

Abstract

Background: African American women experience a twofold higher incidence of triple-negative breast cancer (TNBC) and are 40% more likely to die from breast cancer than women of other ethnicities. However, the molecular bases for the survival disparity in breast cancer remain unclear, and no race-specific therapeutic targets have been proposed. To address this knowledge gap, we performed a systematic analysis of the relationship between gene mRNA expression and clinical outcomes determined for The Cancer Genome Atlas (TCGA) breast cancer patient cohort.

Methods: The systematic differential analysis of mRNA expression integrated with the analysis of clinical outcomes was performed for 1055 samples from the breast invasive carcinoma TCGA PanCancer cohorts. A deep learning fully-convolutional model was used to determine the association between gene expression and tumor features based on breast cancer patient histopathological images.

Results: We found that more than 30% of all protein-coding genes are differentially expressed in White and African American breast cancer patients. We have determined a set of 32 genes whose overexpression in African American patients strongly correlates with decreased survival of African American but not White breast cancer patients. Among those genes, the overexpression of mitogen-activated protein kinase kinase 3 (MKK3) has one of the most dramatic and race-specific negative impacts on the survival of African American patients, specifically with triple-negative breast cancer. We found that MKK3 can promote the TNBC tumorigenesis in African American patients in part by activating of the epithelial-to-mesenchymal transition induced by master regulator MYC.

Conclusions: The poor clinical outcomes in African American women with breast cancer can be associated with the abnormal elevation of individual gene expression. Such genes, including those identified and prioritized in this study, could represent new targets for therapeutic intervention. A strong correlation between MKK3 overexpression, activation of its binding partner and major oncogene MYC, and worsened clinical outcomes suggests the MKK3-MYC protein-protein interaction as a new promising target to reduce racial disparity in breast cancer survival.

Keywords: Triple-negative breast cancer, Racial disparity, Differential expression, MKK3

Background

Breast cancer is the most common cancer and the leading cause of cancer-related death in women [1]. Recent studies have shown up to a twofold higher incidence of triple-negative breast cancer (TNBC) among African American women as compared to White women [2–4]. Moreover, African Americans die from breast cancer at

*Correspondence: hf@emory.edu; andrey.ivanov@emory.edu
¹ Department of Pharmacology and Chemical Biology, Emory University School of Medicine, Emory University, 1510 Clifton Road, Atlanta, GA 30322, USA
Full list of author information is available at the end of the article



© The Author(s) 2020. This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

up to 40% higher rate than White and Hispanic women [5–7]. The American College of Radiology (ACR) has assigned a special status for African American women at higher-than-average risk for breast cancer [8].

Previous studies have revealed significant differences in the mutation rates of several cancer driver genes in African American and White breast cancer patients (Table 1) [9–13].

For example, it was shown that African American women with at least 50% African ancestry have a higher rate of mutations in the major tumor suppressor gene TP53 (43%) as compared to White women with at least 90% European ancestry (28%) [11, 12]. Huo et al. [12] also demonstrated that the mutation frequency in the ubiquitin ligase FBXW7 is almost four times higher in African American breast cancer patients (4.2%) than in White patients (1.2%). Furthermore, African American patients show a higher mutation frequency of BRCA1 (10.2%) and BRCA2 (5.7%) tumor suppressor genes comparing to European non-Ashkenazi Jews White patients (BRCA1: 6.9%, BRCA2: 5.2%) [9, 10]. In contrast, mutations in the catalytic subunit of the Alpha isoform of the Phosphatidylinositol 4,5-Bisphosphate 3-Kinase (PIK3CA) were rarer in African American patients than in White breast cancer patients (20% vs 34%). This difference was even more significant between European White patients (36%) and a cohort of Nigerian breast cancer patients (17%) [13]. In the same study [13], Pitt et al. also determined a significantly lower mutation rate of Cadherin 1 (CDH1) in Nigerian patients (0.8%) and TCGA African American patients (6.4%) as compared to White patients (16.2%).

Besides the mutation rates, the frequency of the DNA copy number alterations has been recently analyzed [12]. It was shown that retinoblastoma protein 1 (RB1), a cell cycle suppressor and the CUB And Sushi Multiple Domains 1 (CSMD1), a tumor suppressor that control cell proliferation, invasion, and migration, are more frequently deleted in Black/African American breast cancer patients (14.5% and 8.6%, respectively) as compared to White patients (8.7% and 4.1%, respectively). Conversely, MYC and Cyclin E1, critical activators of the cell cycle, are more frequently amplified in Black/African American breast cancer patients (30.9% and 9.2%, respectively) than in White patients (20.4% and 3.6%). Together, accumulating clinical and genomics data reveal unique molecular features that may contribute to survival disparity in breast cancer. As summarized in Table 1, the majority of genes that are differentially altered in White and African American breast cancer patients play critical functions in cell proliferation and survival. Meanwhile, most of those genes, including TP53, BRCA1/2, FBXW7, RB1, CDH1, and CSMD1, are tumor suppressors lost due to the inactivating mutations or deletions. The discovery of race-specific and therapeutically actionable targets to decrease the mortality in African American breast cancer patients remains a challenge.

To address this unmet medical need, we performed a systematic analysis of clinical outcomes and gene expression determined for the TCGA PanCancer cohorts of White and African American breast cancer patients. We have identified 32 genes as potential targets to decrease the mortality of African American breast cancer patients. The mitogen-activated protein kinase 3

Table 1 Frequency of tumor driver gene alterations in Black/African American and White breast cancer patients

	Black or African American	White	Oncogenic function	Regulated pathways	References
Mutation, %					
TP53	43	28	TSG	Apoptosis, senescence, DNA repair	[11]
BRCA1	10.2	6.9	TSG	DNA repair Checkpoint control	[9]
BRCA2	5.7	5.2	TSG	DNA repair, checkpoint control	[9]
PIK3CA	20	34	OG	Cell survival, proliferation	[11]
FBXW7	4.2	1.2	TSG	Cell cycle, apoptosis, differentiation	[12]
CDH1	6.4	16.2	TSG	Proliferation, adhesion, polarity, EMT	[13]
Deletion, %					
CSMD1	14.5	8.7	TSG	Proliferation, migration and invasion	[12]
RB1	8.6	4.1	TSG	Cell cycle, apoptosis	[12]
Amplification, %					
MYC	30.9	20.4	OG	Cell growth, survival, immune response, other	[12]
CCNE1	9.2	3.6	OG	Cell cycle	[12]

TSG tumor suppressor gene, OG oncogene, EMT epithelial-to-mesenchymal transition

(MKK3) appeared among the proteins with the most dramatic impact on the survival of African American TNBC patients. We determined that MKK3 promotes TNBC tumorigenesis in African American but not White or Asian patients, and its overexpression leads to the activation of the transcriptional program of major tumor driver MYC.

Together, our data revealed multiple proteins as new promising targets for therapeutic intervention in breast cancer African American patients. As one example, we showed that MKK3 has critical oncogenic functions and promotes TNBC tumorigenesis in African Americans through the activation of the MYC program. The discovery of small-molecule inhibitors to control MKK3 signaling may provide a new therapeutic strategy to decrease mortality in African American TNBC patients.

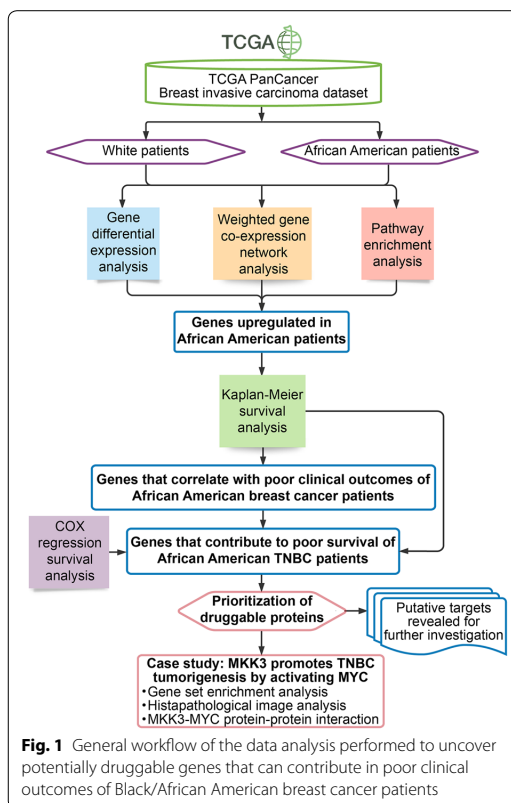
Methods

Breast cancer patient cohort

In this study, the clinical and genomics data from the Breast Invasive Carcinoma TCGA PanCancer cohorts [14] that consists of a total of 1055 female patients with determined DNA copy-number and mRNA expression were analyzed (Fig. 1). The gene RNA expression, DNA copy number, and breast cancer patient survival data were obtained from the NCI Genomics Data Commons (GDC) [15]. The dataset included samples from 729 White patients (69% of all samples) and 178 samples from Black or African American (BAA) patients (17%), as well as 60 Asian patient samples (6%), and 88 samples (8%) from patients with unspecified race. The breast cancer subtype annotations were added based on the original publication [14].

Differential expression

The subset of 17,211 protein-coding genes was identified based on the HUGO Gene Nomenclature Committee (HGNC) annotations [16]. The DNA amplifications or deletions were determined based on the GISTIC 2.0 scores (2-amplification, -2-homozygous deletion) [17]. For the differential expression analysis, the TCGA RNA-seqV2 expression data (EBPlusPlusAdjustPANCAN_IlluminaHiSeq_RNASeqV2.geneExp.tsv; <http://api.gdc.cancer.gov/data/3586c0da-64d0-4b74-a449-5ff4d9136611>) were used. For each gene, the log₂ fold change was calculated as $\log_2 \text{fold change} = \mu_{BAA} - \mu_{WT}$, where μ_{BAA} and μ_{WT} are the mean values of the log₂ (x+1)-transformed gene expression obtained for the Black/African American and White patient cohorts, respectively. The p-values were calculated with the Wilcoxon test. The false discovery rate adjusted q-values were calculated with the Benjamini–Hochberg procedure [18].



The mRNA overexpression was determined based on the z-scores. First, the average (μ) and standard deviation (σ) values were calculated for the samples in which gene is diploid. Then, the z-score was calculated as $(\tau - \mu)/\sigma$, where τ is the gene mRNA expression in the sample. Z-score > 2 and z-score < -2 indicates gene overexpression or underexpression, respectively.

The signed weighted gene co-expression network was constructed for 5256 genes differentially expressed in Black/African American and White breast cancer patients using the WGCNA R package [19]. Pearson correlation coefficients between the expression of all gene pairs were also calculated and used to construct the adjacency correlation matrix and the topological overlap matrix (TOM). The optimal value of the soft threshold power $\beta = 11$ was selected using the pickSoftThreshold function to maintain the scale-free topology and sufficient node connectivity [20]. The hierarchical clustering of genes was performed based on the TOM matrix using the average agglomeration method implemented in the flashClust function [21]. The gene modules were

identified using the dynamic tree cut method [22]. Specifically, the `cutreeDynamic` R function was used with the `minModuleSize = 100` and `method = "tree"` options.

Survival analysis

The Kaplan–Meier survival curves and the logrank p-values have been calculated using the Lifelines python package. The mean survival time (MST) values were calculated with the Lifelines package based on the area under the survival curve. The COX regression analysis has been performed using the fit proportional hazards regression model function `coxph` from the Survival R package.

Enrichment analysis

The disease association and pathway enrichment analysis were performed using the DisGENet [23], KEGG [24], and Reactome [25] datasets. The p-values were calculated with the Fisher Exact test using 17,211 protein-coding genes as the reference set. The false discovery rate adjusted q-values were calculated with the Benjamini–Hochberg procedure [18]. The gene set enrichment analysis (GSEA) was performed using the GSEA program [26]. The High and Low phenotypes were defined as the 10% of samples with the highest and the lowest gene expression, respectively. The GSEA curves were rebuilt using the GseaPy python package.

The breast cancer histological image analysis

Fully-convolutional model training

To extract tumor features we used our established standard 16-layer VGG fully-convolutional neural network (VGG16-FCN8) constructed using ImageNet [27] pre-trained weights as described previously [28]. We have previously shown that for this particular dataset, the VGG-16 FCN-8 architecture shows more favorable model convergence and fitting properties than the deeper and more complex DenseNet architecture [29]. Using this particular architecture and number of layers enabled us to leverage the publicly available pre-trained weights, hence improving accuracy [28, 29].

The model is trained to classify pixels into one of five classes: tumor (including DCIS), stroma, tumor-infiltrating lymphocytes (including plasma cells and mixed inflammatory infiltrates), necrosis or debris, and others. Regions of interest were divided into 800×800 pixel tiles that are overlapping, where the amount of overlap increased for smaller regions of interest to create a balanced training dataset. Random cropping of 768×768 pixel regions was used as a data augmentation strategy to improve robustness during training. The model was trained on 4 GPUs with a per-GPU batch size of 4 tiles (16 tiles per batch) using data parallelization and gradient

averaging. Adam optimizer was used with a starting learning rate of $1e-5$. The loss function used is weighted categorical cross-entropy, where the weight associated with each region class, W_c , is calculated using the equation:

$$W_c = \begin{cases} 0 & \text{if } c = 0 \\ 1 - \frac{N_c}{N} & \text{if } c > 0 \end{cases}$$

where N is the total number of pixels and N_c is the total number of pixels belonging to region class c .

Fully-convolutional model inference

We used whole-slide images (WSI) formalin-fixed paraffin-embedded hematoxylin and eosin-stained slides from the TCGA cohort. The analysis was focused on WSIs from African-American patients with triple-negative breast cancer, and limited to infiltrating ductal histologic subtype (determined using TCGA clinical records). The focus on infiltrating ductal subtype is for pragmatic reasons since the fully-convolutional model has been trained and optimized on this histologic subset. Only one diagnostic slide was used per patient (“-DX” designation in TCGA) and only WSIs scanned at $40\times$ were used in the analysis. The analysis was performed at scan magnification.

Analysis regions were chosen semi-automatically and constituted the main tumor bulk within a WSI. A low-resolution RGB image of the slide (at $0.3-0.5\times$) was loaded and converted to the Hue-Saturation-Intensity (HSI) space. Default thresholds for each of the HSI channels were manually adjusted for each slide to capture the majority or entirety of the tumor within the slide. This region of interest was divided into non-overlapping 1024×1024 pixel tiles and fed into the trained FCN-8 model after color normalization using the Reinhard method [30]. The Reinhard normalization used target statistics derived from the RGB image corresponding to the mask called “TCGA-A2-A3XS-DX1_xmin21421_ymin37486.png” [28].

Feature extraction of tumor nests

A total of nine features (four global and five local) were derived from the slides. “Local” features are those features derived from each individual tumor nest (defined as a coherent collection of carcinoma cells) and are averaged to get slide-level features. The global features were: tumor-to-stroma ratio, stromal tumor-infiltrating lymphocyte score, necrosis-to-tumor ratio, and the number of tumor nests, normalized for the area of the region of interest (i.e. “per pixel”). Local features included area and shape descriptors for each tumor nest.

Histologic-genomic correlation

Histological descriptors were compared against gene expression data derived from the same patients in the TCGA cohort. Spearman correlation coefficient was used and the Benjamini–Hochberg adjustment was used for multiple hypothesis testing.

Results

Differential gene expression in African American and White breast cancer patients

To determine the differences between gene expression in White and Black/African American breast cancers, we have performed the differential expression (DE) analysis for a total of 17,211 protein-coding genes. We found that 7195 genes showed statistically significant differences in expression between White and Black/African American cohorts, as determined with the Wilcoxon test *p*-values adjusted for the false discovery rate (*q*-value < 0.001, Additional file 1: Table S1). To increase the stringency of the analysis, we further prioritized 5268 genes with *q*-values < 0.001 and at least 20% difference in the mRNA expression in White and Black/African American patients (Fig. 2a). Among those genes, expression of 2501 genes was decreased in Black/African American patients, as compared to White breast cancer patients (BAA_{low} gene set). In contrast, the expression of 2767 genes was significantly higher in the Black/African Americans cohort than in White patients (BAA_{high} gene set). These data indicate that White and Black/African American breast cancer patients have very different genomic backgrounds with over 30% of the protein-coding genes expressed differently in these two patient cohorts.

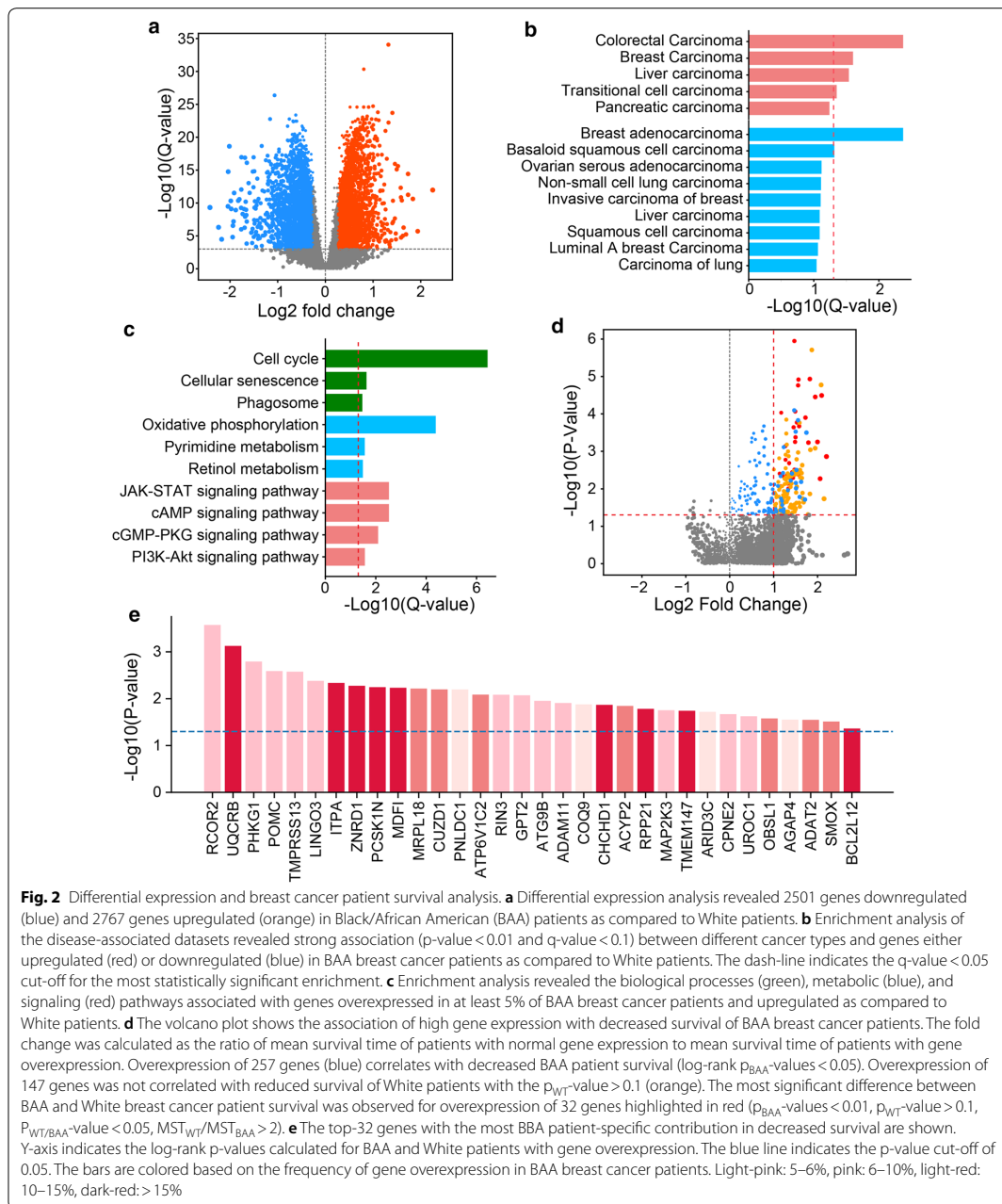
Cancer association and pathway enrichment analysis

To determine whether the differentially expressed genes are associated with the regulation of oncogenic processes, we performed the enrichment analysis. First, using the carcinoma-associated gene sets defined in DisGeNET database [23], we found that both, BAA_{high} and BAA_{low} gene sets are significantly enriched (*p*-value < 0.01, *q*-value < 0.1) in genes associated with different cancer types, including colon (*q*-value = 0.006), liver (*q*-value = 0.036 for BAA_{high} and *q*-value = 0.087 for BAA_{low}), pancreatic (*q*-value = 0.068), and lung cancers (*q*-value = 0.091) (Fig. 2b). Moreover, genes associated with breast carcinoma were among the most significantly overrepresented genes in both, BAA_{high} (*p*-value < 0.001, *q*-value < 0.031) and BAA_{low} sets (*p*-value < 0.001, *q*-value = 0.004, Fig. 2b). 2567 out of 2767 BAA_{high} genes (92%) are overexpressed in at least 5% of Black/African

American breast cancer patients, supporting their potential roles in breast carcinogenesis.

Then, we sought to determine specific biological programs associated with identified differentially expressed genes. Through the enrichment analysis of signaling and metabolic pathways defined in the KEGG database [24], we found that overexpressed BAA_{high} genes (BAA_{OV}R genes) showed the enrichment in genes associated with several major oncogenic pathways. The most significant enrichment (*p*-value 0.001, *q*-value 0.001) was observed for the cell cycle-associated genes. Furthermore, genes involved in senescence, phagosome maturation, JAK-STAT, cAMP, cGMP-PKG, and PIK3-AKT signaling pathways, retinol and pyrimidine metabolism, and oxidative phosphorylation (Fig. 2c) were significantly overrepresented in the BAA_{OV}R genes (*p*-value < 0.001, *q*-value < 0.05). Thus, overexpression of genes upregulated in Black/African American patients may promote breast cancer development and progression through the dysregulation of multiple oncogenic processes.

To identify functional modules of co-regulated genes we applied the weighted gene co-expression network analysis (WGCNA) [19, 20]. The WGCNA performed for all 5268 BAA_{high} and BAA_{low} genes revealed 12 distinct modules of significantly co-expressed genes (Additional file 1: Table S2, Additional file 2: Figure S1). The “pink” (332 genes), “black” (355 genes), “cyan” (846 genes), “red” (368 genes), “green” (431 genes), and “magenta” (292 genes) modules were comprised almost completely by BAA_{high} genes. The “yellow” (521 genes), “blue” (798 genes), “brown” (648 genes), “purple” (174 genes), “greenyellow” (164 genes), and “tan” (143 genes) modules included mostly BAA_{low} genes. To uncover the biological pathways associated with individual modules, we performed the enrichment analysis using the gene sets defined in the KEGG database [24] (Additional file 1: Table S3; Additional file 2: Figure S2). We found that five modules were more than tenfold overrepresented by genes involved in pathways defined in the KEGG database as compared to the reference human genome. Among all modules, the most significant enrichment was determined for the “pink” module, which appeared to be overrepresented (overrepresentation fold, OVF = 21.08, *q*-value = 8.78×10^{-26}) in the cell cycle regulating genes (Additional file 1: Table S3). An equally high overrepresentation (OVF = 20.83, *q*-value = 2.38×10^{-6}) was determined for the “magenta” module that was enriched in genes that control primary immunodeficiency, including ADA, CD19, CD79A, IGLL1, IGLL1, TAP1, TAP2, TNFRSF13C, and ZAP70. The “green” module appeared to be enriched in genes involved in oxidative phosphorylation (OVF = 12.5, *q*-value = 8.50×10^{-14}), ribosome (OVF = 12.43, *q*-value = 4.83×10^{-12}), and genes involved



in neurodegenerative disorders, such as the Parkinson’s disease (PD) ($OVF = 11.31$, q -value $= 4.83 \times 10^{-12}$). Notably, multiple studies have suggested that the development

of PD and cancer, including breast cancer, can progress through the same genes and molecular mechanisms [31–33]. In contrast to “pink”, “magenta”, and “green”

modules, the “tan” and “greenyellow” modules are comprised of genes with higher expression in White breast cancer patients as compared to Black/African American patients. We found that the “tan” module is enriched in genes involved in extracellular matrix receptor interactions (OVF=15.31, $q\text{-value}=2.39 \times 10^{-6}$). The “greenyellow” module appeared to be overrepresented in the ATP-binding cassette (ABC) transporters (OVF=19.77, $q\text{-value}=4.85 \times 10^{-5}$), and genes that control tyrosine metabolism (OVF=10.15, $q\text{-value}=2.98 \times 10^{-2}$) and the complement and coagulation cascades (OVF=14.37, $q\text{-value}=4.85 \times 10^{-5}$). The role of the ATP-binding cassette (ABC) transporters in tumorigenesis of White breast cancer patients is further supported by the enrichment of the “brown” module in basal transcription factors, including the ATP-binding cassette subfamily members ABCA9, ABCC9, ABCG2, ABCB1, ABCA6, and ABCA8. Previous studies have demonstrated the association of ATP-binding cassette transporters with breast cancer aggressiveness and reduced survival of breast cancer patients [34, 35]. We also noticed that the “blue” module is enriched (OVF=6.21, $q\text{-value}=0.036$) in genes that control sphingolipid metabolism that play critical functions in cancer growth and progression [36]. Furthermore, the “purple” module appeared to be enriched in genes that are associated with the dilated cardiomyopathy, a known side effect of breast cancer radiotherapy and chemotherapy [37, 38].

Differential expression and survival disparity

The genes with abnormally high expression may represent putative targets for therapeutic intervention. We used the set of 2567 BAA_{OV}R genes to determine the impact of their overexpression on breast cancer patient survival. We found that overexpression of 257 BAA_{OV}R genes (Additional file 1: Table S4, Group I) correlates with decreased survival of Black/African American patients ($p_{\text{BAA}}\text{-value}<0.05$) (Fig. 2d). Furthermore, the overexpression of 174 out of 257 genes (Additional file 1: Table S4, Group II) correlated with more than twofold decreased survival. Among the 174 genes, overexpression of 147 genes (Additional file 1: Table S4, Group III) was associated with the reduced survival of Black/African American patients, but not White patients ($p_{\text{WT}}\text{-value}>0.1$, Fig. 2d). This group of genes includes several genes previously linked with breast cancer

development and progression. For example, overexpression of protein arginine methyltransferase 1 (PRMT1) has been associated with the methylation of the transcription factor C/EBP α and inhibition of its tumor suppressor function in breast cancer [39]. Interestingly, PRMT1 knockdown was also correlated with decreased EGFR activity and suppressed proliferation of in MDA-MB-468 breast cancer cells that are derived from an African American breast cancer patient [40]. Kinesins KIF1C and KIFC3 promotes breast cancer cell growth and survival and mediate taxane resistance [41, 42]. Syndecan-1 (SDC1) has been linked with the accelerated metastasis of breast cancer to the brain [43]. Meanwhile, our data revealed genes previously not associated with the increased breast cancer progression, providing new opportunities for therapeutic interventions in breast cancer.

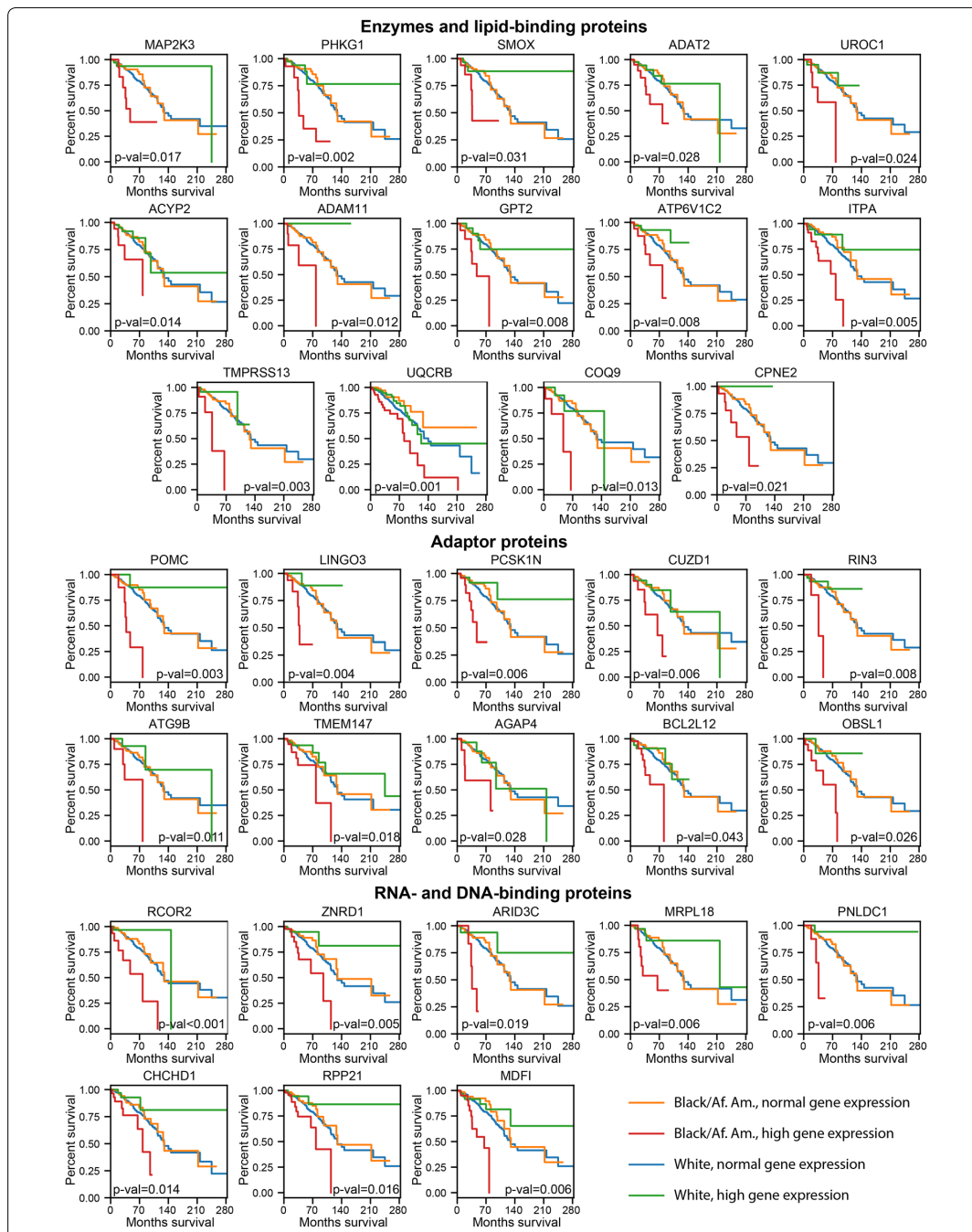
To further prioritize genes with the most significant contribution to the survival disparity between African American and White patients, we applied more stringent statistical cut-offs: $p_{\text{BAA}}\text{-value}<0.01$, $p_{\text{WT}}\text{-value}>0.1$, a significant difference between survival time of White and Black/African American patients with the overexpressed gene ($p_{\text{WT/BAA}}\text{-value}<0.05$), and at least twofold decreased mean survival time of Black/African American patients (MST_{BAA}) comparing to the MST of White patients (MST_{WT}) determined for the samples with the overexpressed gene. Using these parameters, a total of 32 genes with a most significant and race-specific impact on the breast cancer tumorigenesis in Black/African American patients have been prioritized (Additional file 1: Table S4, Group IV, Fig. 2e, and Fig. 3).

Evaluation of protein druggability for therapeutic discovery

To assess the potential druggability of the top-32 prioritized proteins, we have classified them into three groups based on the protein types (Fig. 3). RCOR2, ZNRD1, ARID3C, MRPL18, PNLDC1, CHCHD1, RPP21, MDFI are either DNA- or RNA-binding proteins. These proteins may represent the most challenging targets for direct interrogation with small molecules or specific antibodies due to their nuclear localization the lack of a defined pocket for a small-molecule binding. POMC, LINGO3, PCSK1N, CUZD1, RIN3, ATG9B, TMEM147, AGAP4, BCL2L12, OBSL1 also lack an enzymatic activity

(See figure on next page.)

Fig. 3 Survival curves for the top-32 genes that contribute to survival disparity between Black/African American and White patients. Orange and red lines indicate the survival of Black/African American breast cancer patients with normal and overexpressed gene levels, respectively. Blue and green lines indicate the survival of White breast cancer patients with normal and overexpressed gene levels, respectively. The log-rank p-values calculated for the survival rates of White and Black/African American patients with gene overexpression are indicated



and contribute in breast cancer tumorigenesis acting as adaptors for other proteins. A large area, hydrophobicity, and relatively flat configuration of the protein–protein interaction (PPI) interface surfaces are among the limiting factors for the design and discovery of low molecular weight PPI inhibitors [44]. On the other hand, the growing number of potent cell-permeable inhibitors for PPI discovered over the past decades, including the FDA-approved BCL2 inhibitor venetoclax [45], indicates the PPI druggability for therapeutic discovery [46]. Meanwhile, enzymes and receptors represent the largest class of therapeutic targets [47]. We found that 14 out of 32 proteins belong to protein families known to be druggable by low molecular weight compounds. Specifically, COQ9 and CPNE2 are the lipid-binding proteins with a defined binding site for a lipid molecule that can be targeted by small molecules [48]. Furthermore, 12 proteins belong to different types of enzymes, including a subunit of the ubiquinol-cytochrome c oxidoreductase UQCRB, serine protease TMPRSS13, inosine triphosphate pyrophosphatase ITPA, proton ATPase ATP6V1C2, Alanine aminotransferase GPT2, metalloproteinase ADAM11, acylphosphatase ACYP2, urocanate hydratase UROC1, tRNA-specific adenosine deaminase ADAT2, spermine oxidase SMOX, and two kinases: PHKG1 and MKK3 also known as MAP2K3. The discovery of potent inhibitors for these enzymes may lead to new therapeutic strategies for African American breast cancer patients.

COX regression survival analysis for TNBC Black/African American patients

The COX regression analysis is a widely used approach to identify predictive biomarkers of poor clinical outcomes [49, 50]. We applied the COX regression analysis to determine the overall impact of the prioritized genes on clinical outcomes of Black/African American patients specifically with the triple-negative breast cancer subtype. First, we built the univariate COX regression models to determine the hazard ratios and significance for each of the 32 prioritized genes. We found that for each gene the Hazard ratio values (HR) were higher than 1 indicating a positive correlation between gene expression and decreased patient survival (Additional file 1: Table S5). This result is consistent with the Kaplan–Meier analysis performed for all breast cancer subtypes (Fig. 3). Eight out of 32 genes demonstrated highly significant correlation with poor clinical outcomes with the Hazard ratio (HR) > 2 and the p-values ≤ 0.05, including ACYP2, ADAT2, AGAP4, CHCHD1, MKK3, MRPL18, RPP21, and ZNRD1 (Additional file 1: Table S5; Additional file 2: Figure S3).

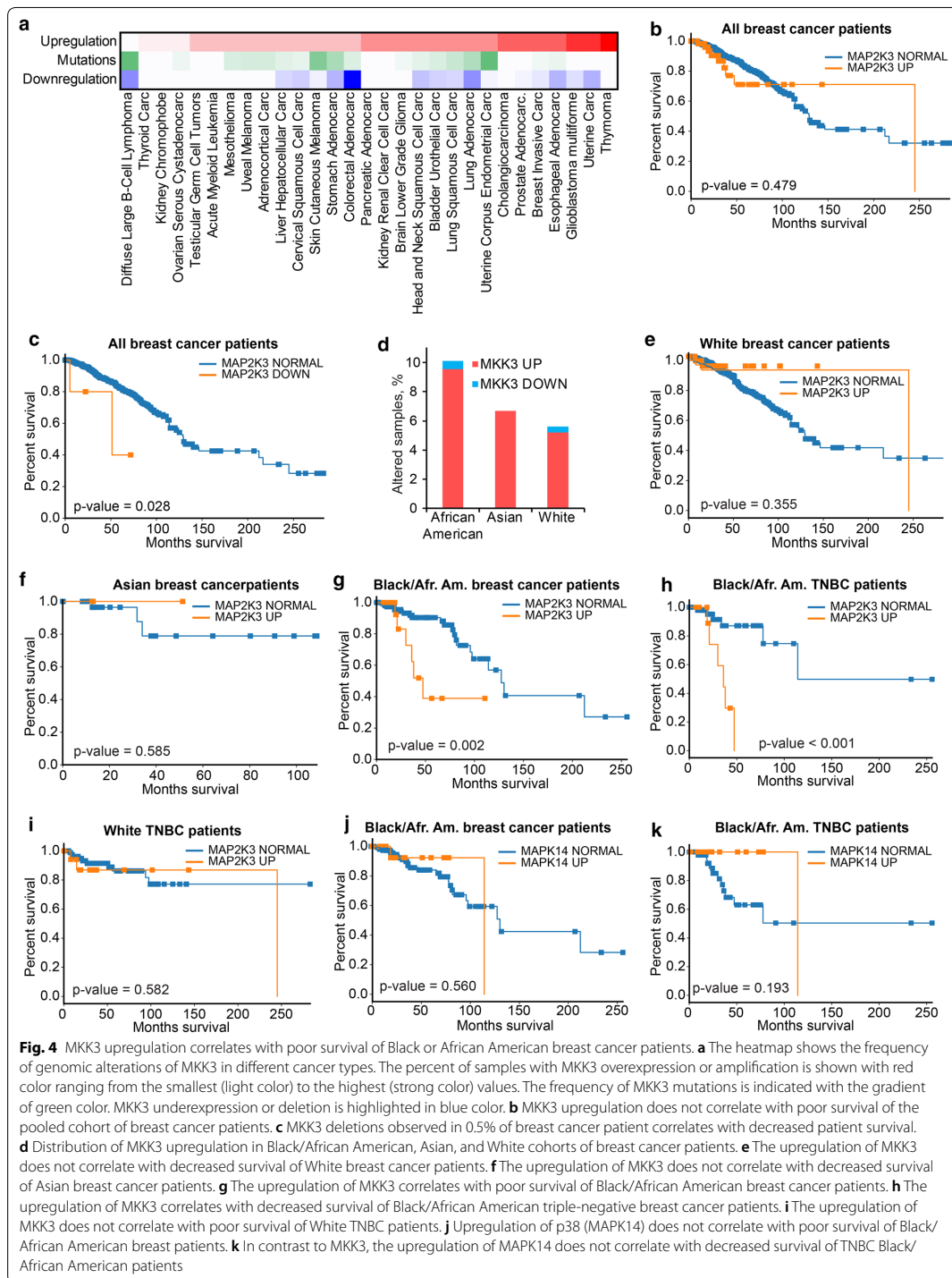
To evaluate the combined effect of these 8 genes on the clinical outcomes of TNBC Black/African American

patients, we built a multivariate COX regression model. The resulting Model 1 demonstrated a high concordance index (c-index=0.93) and statistical significance (p-value = 1×10^{-4}), indicating the satisfactory prognostic ability of the model (Additional file 1: Table S6). The detailed evaluation of the model revealed that expression of MKK3 (HR=27.98, p-value=0.002), AGAP4 (HR=1.73, p-value=0.017), and ACYP2 (HR=1.30, p-value=0.04) made the most significant contribution to the model. To determine if MKK3, AGAP4, and ACYP2 can be used as markers for poor clinical outcomes, we built another model based on these three genes only. The resulting Model 2 (Additional file 1: Table S6) was characterized by an equally high c-index of 0.91 and improved statistical significance (p-value = 2×10^{-5}) as compared to the 8-parameter model. We noticed that in both Model 1 and Model 2 the highest HR value was obtained for MKK3, suggesting its significance for clinical outcomes of Black/African American TNBC patients.

MKK3 overexpression promotes triple-negative breast cancer in African American patients

MKK3 is frequently altered in different cancers and recent studies have suggested that MKK3 may contribute in tumorigenesis in multiple cancer types [51–55]. Analysis of the TCGA PanCancer datasets indicates that MKK3 is mutated in 5% of uterine carcinoma, 5% of B-cell lymphoma, and 4% of skin melanoma patients. MKK3 is homozygously deleted in 6% of colon cancer patients. On the other hand, MKK3 is either overexpressed or amplified in 3 to 8% of patients in the vast majority of cancers, including thymoma (8%), glioblastoma multiform (7%), and breast invasive carcinoma (6%) (Fig. 4a, Additional file 1: Table S7). Furthermore, MKK3 overexpression can be triggered by TP53 mutations [56], that can link MKK3 to TP53-dependent cancers, such as breast cancer, particularly in African American patients.

The evaluation of the overall survival data for the pooled dataset of 1055 samples revealed no correlation between the MKK3 overexpression and patient survival (p=0.479, Fig. 4b, Additional file 1: Table S8). Instead, five patients with lost MKK3 demonstrated decreased survival comparing to patients with normal MKK3 (p=0.028, Fig. 4c). This observation is consistent with a previous report that MKK3 may play a tumor-suppressive role in breast cancer [57]. Meanwhile, we found that MKK3 is the most frequently overexpressed in the Black/African American cohort (9.6%) (Fig. 4d). In the Asian and White breast cancer patients, MKK3 is overexpressed in 6.7% and 5.2%, respectively. Conversely, MKK3 downregulation is not frequent in breast cancer patients. MKK3 is not underexpressed or deleted in the Asian cohort, and it



was deleted in 3 White patients (0.4%), 1 Black patient (0.6%), and 1 patient with the unspecified race (1.1%). In agreement with the genomic status of MKK3, its upregulation does not correlate with poor survival of White ($p=0.355$, Fig. 4e) nor Asian ($p=0.585$, Fig. 4f) patients. In contrast, a strong decrease in patient survival ($p=0.002$) was observed for the Black/African American cohort (Fig. 4g).

The analysis of the histological subtypes of breast patients indicates, that the majority of Black/African American patients in the breast cancer TCGA PanCancer cohort ($N=178$) have either basal-like/triple-negative (TNBC) (63 patients) or Luminal A (61 patients) breast cancer. The number of patients with Luminal B, HER2, and Normal breast cancers was 28, 16, and 10, respectively. Surprisingly, MKK3 was upregulated in only one patient with the Luminal A breast cancer. In contrast, MKK3 was overexpressed in 19% of Black/African American TNBC patients.

Similar to the combined set of breast cancer samples of all subtypes, the MKK3 overexpression correlates with poor survival of TNBC Black/African American patients ($p<0.001$, Fig. 4h), but not White patients ($p=0.582$, Fig. 4i). Moreover, through a systematic analysis of all breast cancer subtypes in all racial groups of patients, we have determined that MKK3 upregulation correlates uniquely with the poor survival of Black/African American patients specifically with the TNBC, and not with any other race or other breast cancer subtypes (Additional file 1: Table S9).

MKK3 promotes TNBC through a p38-distinct mechanism

MKK3 is the main activator of its only known substrate p38 which plays a key role in the induction of apoptosis and regulation of inflammation in response to extracellular stress [58, 59]. It can be expected that the poor survival of Black/African American patients is also associated with p38 activation. p38 (encoded by the MAPK14 gene) is amplified or overexpressed in 9.5% of the Black/African American breast cancer patients. However, in contrast to MKK3, p38 upregulation does not correlate with decreases survival of Black/African American patients neither for all breast cancer subtypes ($p=0.986$) (Fig. 4j) nor specifically for the TNBC ($p=0.193$, Fig. 4k). These data suggest a p38-distinct role for MKK3 in TNBC tumorigenesis. These results are further supported by the recent discovery of MKK3 as a hub protein in the PPI network determined for cancer-associated proteins [60, 61]. It was shown that besides p38, MKK3 can bind to multiple other proteins, including several drivers of breast cancer, such as CDK4, AURKA, FGFR4, EPHA2, and MYC [60].

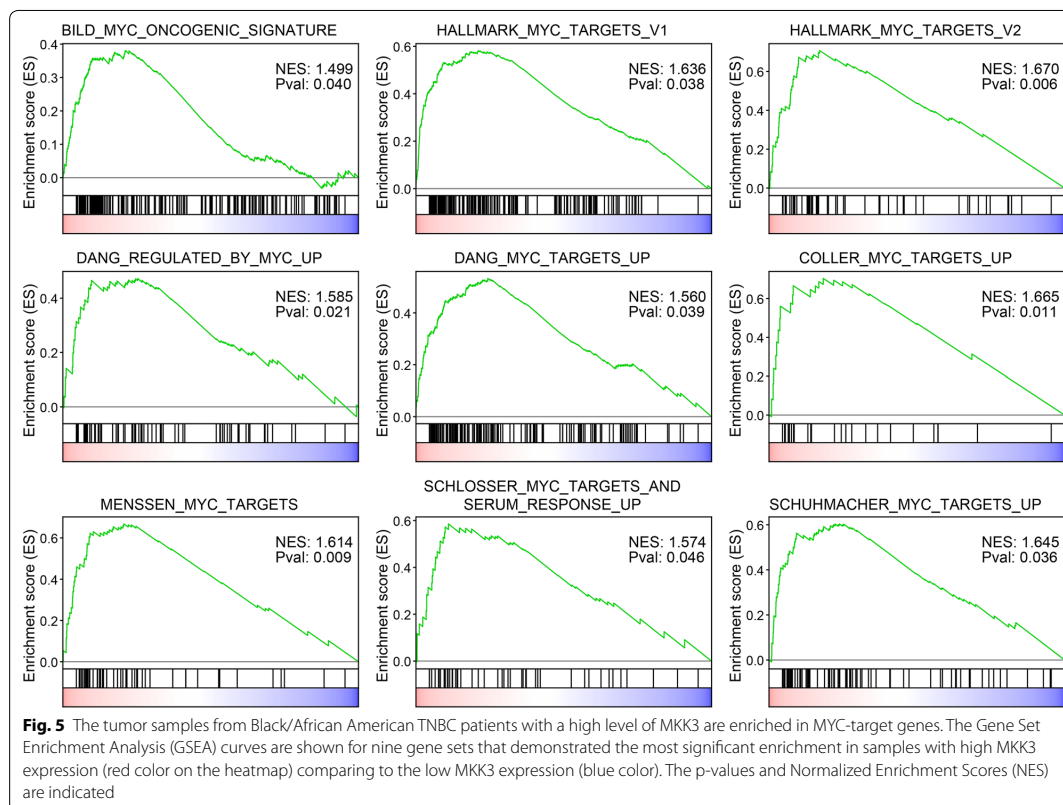
MKK3 activates MYC transcriptional program in TNBC African American patients

To uncover the molecular bases underlying the decreased survival of TNBC Black/African American patients, we performed the Gene Set Enrichment Analysis (GSEA) [62] against 50 hallmark sets of genes that define signatures of specific biological state or process [26, 63, 64] (Additional file 1: Table S10). Only five gene sets showed significant enrichment in samples with upregulated MKK3 expression ($p\text{-value}<0.05$ and $\text{FDR}<25\%$), including genes involved in unfolded protein response, mTORC1 signaling, response to the UV irradiation, and two sets of MYC target genes [64, 65] (Additional file 1: Table S10).

To further increase the confidence in the MKK3-MYC functional association, we have expanded the GSEA analysis using 16 more sets of MYC-upregulated genes independently defined in different studies (Additional file 1: Table S11). We found that 17 out of 18 tested MYC-target gene sets demonstrate the enrichment in Black/African American TNBC samples with a high level of MKK3. Furthermore, 9 out of 18 sets showed a statistically significant enrichment with the $p\text{-value}<0.05$, including the MYC oncogenic signature genes derived from the DNA microarray analysis of the breast cancer cells ($p=0.040$, $\text{FDR}=5.6\%$, normalized enrichment score, $\text{NES}=1.5$) (Fig. 5). Meanwhile, no enrichment in MYC-target genes was found for the samples with upregulated p38, further supporting p38-distinct functions of MKK3 in Black/African American TNBC patients.

As a master regulator, MYC controls multiple oncogenic programs. We sought to determine biological pathways that could be dysregulated specifically in response to MKK3-mediated MYC activation. Based on the GSEA analysis for each MYC-dependent gene set we determined a total of 323 core genes that contribute the most to the enrichment. The pathway overrepresentation analysis revealed a strong association of 222 of MYC-regulated genes enriched in patients with overexpressed MKK3 with 117 signaling and metabolomic pathways defined in the Reactome database ($q\text{-value}<0.01$, at least twofold overrepresentation as compared to the reference human genome). We found that the cell cycle and RNA metabolism and processing appeared among the pathway with the most significant overrepresentation in MKK3-MYC core enrichment genes (Additional file 1: Table S12). Interestingly, the REACTOME_DEASES gene set also appeared within the top-10 the most overrepresented pathways suggesting the pathological functions for the genes upregulated through the MKK3-MYC interaction.

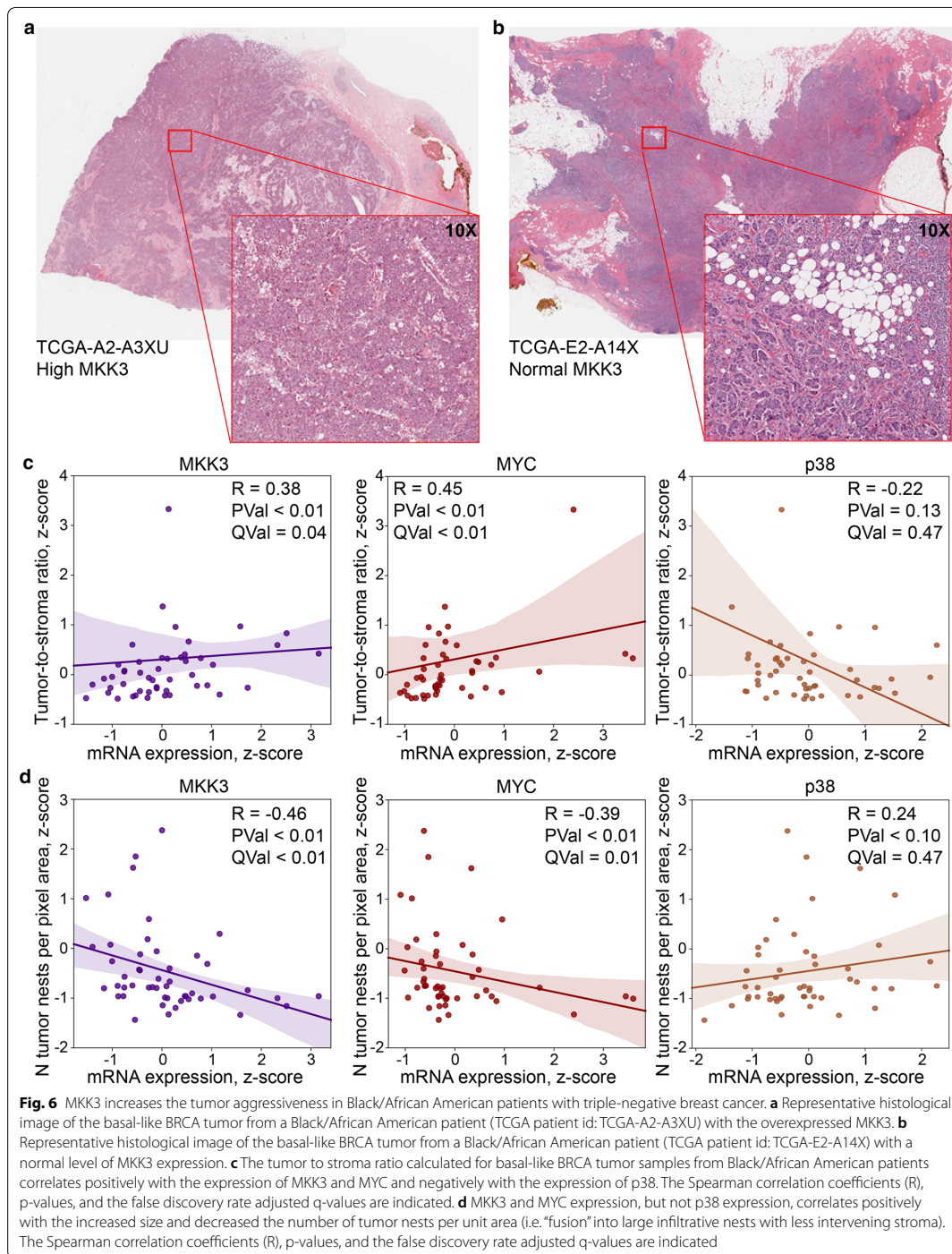
To support the clinical significance of MKK3 as a mediator of TNBC pathology, we performed a quantitative



analysis of the histopathological images from Black/African American TNBC patients (Fig. 6a, b). We found that the high level of MKK3 expression is associated with the increased overall tumor to stroma ratio (Spearman $R=0.38$, $p\text{-value}<0.01$, $q\text{-value}=0.04$, Fig. 6c), and fewer discrete tumor “nests” (Spearman $R=-0.46$, $p\text{-value}<0.01$, $q\text{-value}<0.01$, Fig. 6d). Note that the smaller number of discrete nests is, in this phenotype, a consequence of their larger size, causing less intervening stroma and apparent “fusion” into large invasive tumors (Fig. 6a versus b). Similar trends have been observed for MYC. The elevation of MYC expression leads to increased overall tumor-to-stroma ratio (Spearman $R=0.45$, $p\text{-value}<0.01$, $q\text{-value}<0.01$, Fig. 6c), as well as fewer discrete tumor nests (Spearman $R=-0.39$, $p\text{-value}<0.01$, $q\text{-value}=0.01$, Fig. 6c), which are, individually, significantly larger in size (Spearman $R=0.42$, $p\text{-value}<0.01$, $q\text{-value}<0.01$, not shown). Unlike MKK3 and MYC, p38 upregulation does not correlate with either the overall tumor-to-stroma ratio (Spearman $R=-0.22$, $p\text{-value}=0.13$, $q\text{-value}=0.47$, Fig. 6b) or the

number of discrete tumor nests (Spearman $R=0.24$, $p\text{-value}=0.10$, $q\text{-value}=0.47$, Fig. 6c). Moreover, the observed trends, although not statistically significant, were opposite compared to trends determined for MKK3 and MYC. Together, these data suggest a critical role of MKK3 in promoting the TNBC tumorigenesis in African American patients and its strong association with the activation of the MYC program.

To identify which of MKK3-activated MYC-regulated genes can contribute most in poor clinical outcomes of African American patients, we evaluated the correlations between overexpression of MKK3-MYC core enrichment genes and TNBC patient survival. We prioritized 8 MKK3-MYC core enrichment genes whose overexpression correlates with decreased survival of TNBC African American patients, including EIF5AL1 (log-rank test $p\text{-value}=0.029$), EIF5A ($p\text{-value}=0.015$), SNAI1 ($p\text{-value}=0.050$), TAF12 ($p\text{-value}=0.004$) as well APEX1 ($p\text{-value}=0.001$), FASN ($p\text{-value}=0.033$), HNRNPA2B1 ($p\text{-value}=0.036$), and GRSF1 ($p\text{-value}<0.001$). Notably, overexpression of these genes does not worsen clinical



outcomes in Caucasian TNBC patients (p -values >0.1), suggesting their unique functions in African American patients. We found that EIF5A, EIF5AL1, and SNAI1 are the most frequently overexpressed genes ($>20\%$) in African American TNBC patients. These genes also demonstrate the highest correlation with both MKK3 and MYC expression (Pearson correlation $p < 0.01$, Fig. 7a) and decreased survival of African American patients (Fig. 7b). Importantly, both Snail Family Transcriptional Repressor 1 (SNAI1) and Eukaryotic Translation Initiation Factor 5A (EIF5A) have been associated with the induction of the epithelial-to-mesenchymal transition (EMT) in breast cancer, promotion of breast cancer metastasis, and chemoresistance [66–68]. Together, these findings suggest a new function for MKK3 as an inducer of MYC-dependent epithelial-to-mesenchymal transition in African American TNBC patients (Fig. 7c).

Discussion

Breast invasive carcinoma is the most common cancer type in women. It is especially aggressive in African American patients. The discovery of new therapeutic targets is urgently needed to decrease breast cancer mortality and reduce the racial disparity in breast cancer outcomes. In contrast to the tumor suppressor genes, such as TP53 or BRCA1/2, that are lost due to deletions or mutations, the mRNA overexpression represents an actionable alteration that can be reached therapeutically. The identification of therapeutically actionable upregulated genes that contribute in poor clinical outcomes may facilitate the development of new clinical strategies in breast cancer. Toward this goal, we have performed a systematic analysis of clinical outcomes and differential gene expression in White and African American breast cancer patients.

We found that more than 2500 genes overexpressed in African American patients are also significantly upregulated in African Americans as compared to the White breast cancer patients. Our analysis has also confirmed 117 out of 142 (82%) genes previously reported as differentially expressed in African American and White/European cohorts of breast cancer patients [12]. The enrichment analysis revealed a strong functional association of these genes with breast cancer as well as several other cancer types, and multiple key oncogenic pathways including cell cycle, PI3K-AKT, and JAK-STAT pathways. Through the gene co-expression analysis integrated with the analysis of pathway overrepresentation, we determined specific modules of co-regulated genes. Notably, three distinct gene modules of genes with higher expression in African American patients as compared to White patients were significantly enriched in genes that control cell cycle progression, immunodeficiency, and oxidative

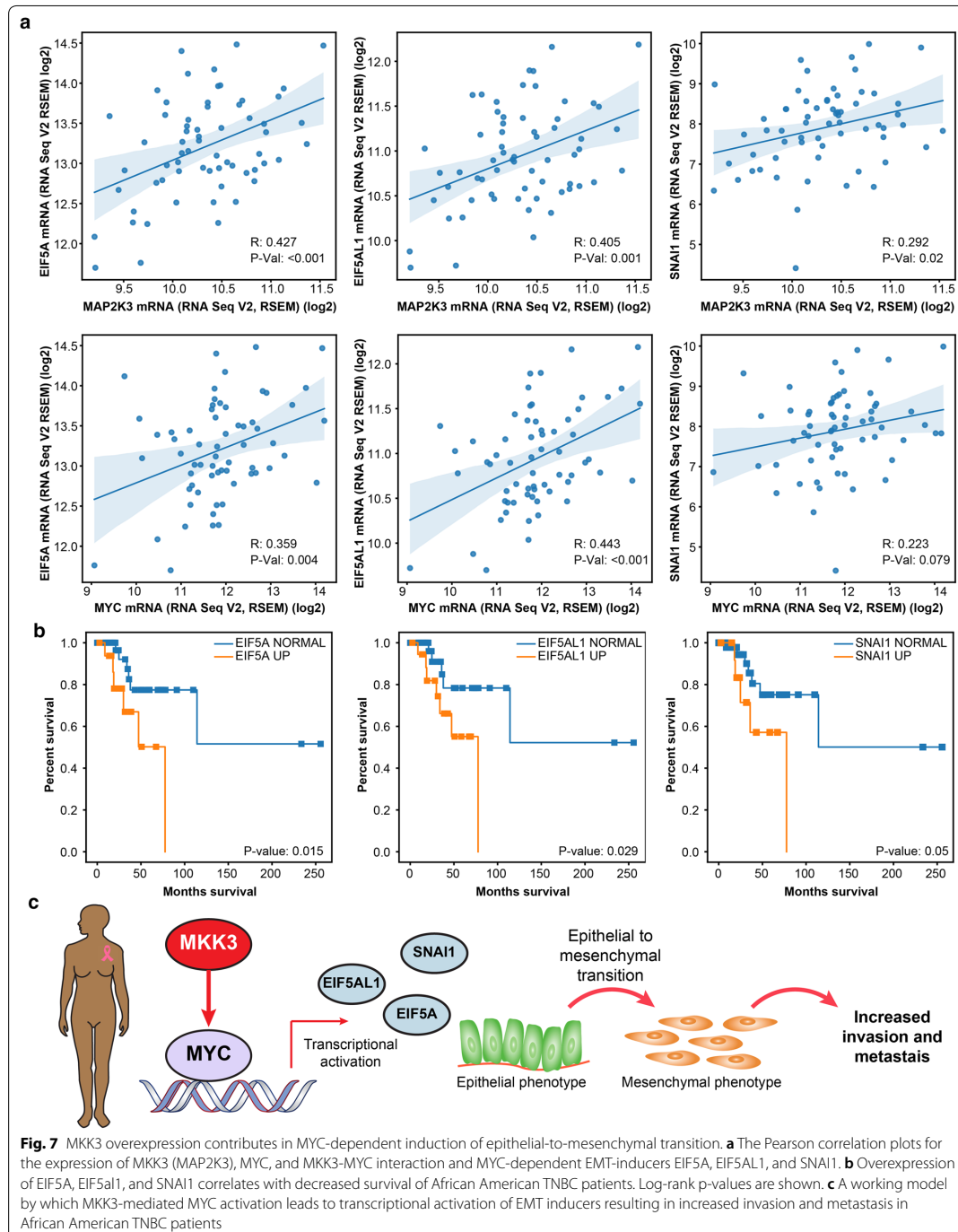
phosphorylation. The identification and targeting of the key druggable regulators of these fundamental oncogenic processes may facilitate the development of new clinical strategies to reduce survival disparity in breast cancer. Meanwhile, the diversity of differentially expressed genes and dysregulated pathways (summarized in Additional file 1: Table S3) indicate the heterogeneity and complexity of the molecular mechanisms underlying survival disparity in breast cancer. Thus, the discovery and prioritization of the most biologically clinically essential genes is critical to facilitate the translation of breast cancer patient genomics data into the clinic.

Through the rigorous statistical analysis, we prioritized 32 proteins that demonstrate the most prominent and race-specific association with decreased survival of African American women. We found that 14 of these prioritized proteins belong to proteins classes known to be druggable and thus represent promising targets for therapeutic discovery. Indeed, at least two proteins, ITPA and GPT2, are well-established therapeutic targets for rheumatoid arthritis and anxiety disorders with the FDA-approved inhibitors, azathioprine and phenelzine, respectively. Currently, phenelzine is also in clinical trials in patients with different cancer types, including patients with advanced or metastatic breast cancer. Our data may open new opportunities for the repurposing of these approved drugs and other reported inhibitors as the anticancer agents for African American breast cancer patients.

Through the COX regression analysis, we have identified several proteins as new promising targets for the therapeutic discovery in TNBC. The multivariate COX regression model suggests that expression of MKK3, AGAP2, and ACYP2 has a significant negative impact on clinical outcomes of TNBC Black/African American patients, and these genes may serve as putative biomarkers for decreased patient survival. Among them, MKK3 showed the most dramatic impact on the survival of African American patients, specifically with triple-negative breast cancer.

The integration of the survival data analysis, gene set enrichment analysis, and the analysis of the breast cancer histopathological images revealed that MKK3 can promote TNBC tumorigenesis through the activation of the MYC transcriptional program.

MKK3 is a well-established activator of the p38 pro-inflammatory and pro-apoptotic pathway [58], and MKK3 functions have been associated primarily with the regulation of p38 signaling [69–74]. The loss of p38-activation may promote tumor growth in cancers with a decreased level of MKK3 [57, 75], suggesting its tumor-suppressive function. On the other hand, the oncogenic role for MKK3 has been reported in



multiple tumor types, including melanoma, colorectal, liver, esophageal, cervical, and breast cancers [51, 52, 76–80].

Analysis of genomics data shows that MKK3 is either up- or downregulated in different cancer types and different groups of cancer patients. Thus, MKK3 can play a dual role in cancer: one as a lost tumor suppressor acting through the p38-pathway [81, 82], and another as an oncogene through upregulation of different oncogenic programs, such as the MYC transcription. MYC is a major tumor driver, and the master regulator of multiple key cellular processes, including cell growth and proliferation, immune response, and metabolism. Over the past decades, MYC became a well-established and highly-appealing therapeutic target in breast cancer [83]. Therapeutic regulation of MYC activation may provide new clinical strategies to suppress different oncogenic mechanisms in African American breast cancer patients [84–86].

Recent studies have established strong functional connectivity between TP53 mutations in breast cancer patients and MYC activation [87]. The frequency of TP53 mutations is more than 40% higher in African American patients than in White patients [11, 12]. Furthermore, MKK3 overexpression was linked to TP53 mutations in colon and breast cancer cells [56]. These data suggest that MKK3 may cooperate with TP53 to activate MYC and promote TNBC progression in the Black/African American cohort. This model is further supported by a synthetic lethal relationship [88] and a physical protein–protein interaction observed between MKK3 and MYC in cancer cells [60, 61, 89, 90]. Thus, the MKK3-MYC oncogenic axis may represent a new promising target for therapeutic discovery for African American TNBC patients.

Through a systematic gene set enrichment analysis and clinical outcome profiling, we have discovered a new oncogenic function for MKK3 in African American TNBC patients as an activator of MYC-dependent epithelial-to-mesenchymal transition, specifically through EIF5A, EIF5AL1, and SNAI1 genes. Overexpression of these MKK3-MYC signature genes has been linked with the induction of epithelial-to-mesenchymal transition in breast cancer and strongly correlates with worsened clinical outcomes in African American patients. These findings suggest that the inhibition of MKK3-MYC interaction itself and its downstream-activated genes EIF5A, EIF5AL1, and SNAI1 may provide new therapeutic options for African American patients with triple-negative breast cancer.

Conclusions

In this study, the relationship between gene expression and survival disparity in breast cancer patients has been investigated. Through the integrative statistical analyses of clinical and genomics data, we identified 32 genes as putative targets for therapeutic intervention in Black or African American breast cancer patients. The success of the translation of these findings into the clinic would certainly rely on the further rigorous experimental validation and can be complicated by diverse molecular mechanisms underlying survival disparity. To facilitate this process, the identification and prioritization of the most biologically relevant, clinically significant, and druggable targets is crucial. Toward this goal, we performed a systematic analysis of the genomics and clinical data available for MKK3 gene that demonstrated one of the most significant negative impacts on the survival of African Americans with triple-negative breast cancer. Through a comprehensive systems biology approach, we have linked MKK3-mediated worsened clinical outcomes in African American TNBC patients with the activation of MYC transcriptional program. We have determined that besides its well-defined function in the p38-inflammatory pathway, MKK3 can induce MYC-dependent epithelial-to-mesenchymal transition in breast cancer patients in part through upregulation of EIF5A, EIF5AL1, and SNAI1 genes. These findings suggest new oncogenic functions for MKK3 in breast cancer and define MKK3-MYC interaction as a promising target to reduce survival disparity in African American TNBC patients.

Supplementary information

Supplementary information accompanies this paper at <https://doi.org/10.1186/s12967-020-02502-w>.

Additional file 1: Table S1. Differential gene expression analysis. **Table S2.** Modules of co-expressed genes identified with the Weighted Gene Co-expression Network Analysis (WGCNA). **Table S3.** Gene module pathway overrepresentation analysis. **Table S4.** Breast cancer patient survival analysis for genes upregulated in African American patients as compared to White patients. **Table S5.** Univariate COX regression survival analysis for TNBC Black/African American patients. **Table S6.** Multivariate COX regression survival analysis for TNBC Black/African American patients. **Table S7.** Analysis of genomic alterations of MKK3. **Table S8.** Genomic status of MKK3 (MAP2K3 gene) and p38 (MAPK14 gene) in BRCA patients and associated clinical data. **Table S9.** Correlation between MKK3 overexpression and breast cancer patient survival. **Table S10.** MKK3 gene set enrichment analysis using the MSigDB cancer Hallmark sets. **Table S11.** MKK3 gene set enrichment analysis for MYC-regulated gene sets. **Table S12.** Pathway overrepresentation analysis for MYC-dependent genes enriched in TNBC patients with overexpressed MKK3.

Additional file 2. Additional figures.

Abbreviations

ACR: American College of Radiology; ACYP2: Acylphosphatase 2; ADAT2: Adenosine deaminase TRNA specific 2; AGAP2: ArfGAP with GTPase domain, Ankyrin repeat and PH domain 2; BAA: Black or African American; CHCHD1: Coiled-coil-helix-coiled-coil-helix domain containing 1; CDH1: Cadherin 1;

CCNE1: Cyclin E1; EIF5A: Eukaryotic translation initiation factor 5A; EIF5AL1: Eukaryotic translation initiation factor 5A like 1; EMT: Epithelial-to-mesenchymal transition; FDA: The United States Food and Drug Administration; FBXW7: F-Box and WD repeat domain containing 7; GDC: Genomics Data Commons; GSEA: Gene set enrichment analysis; HGNC: HUGO Gene Nomenclature Committee; HSI: Hue-Saturation-Intensity; HUGO: Human Genome Organization; MKK3: Mitogen-activated protein kinase kinase 3; MAP2K3: Mitogen-activated protein kinase kinase 3; MRPL18: Mitochondrial ribosomal protein L18; MST: Mean survival time; MYC: MYC proto-oncogene, basic helix-loop-helix transcription factor; NCI: National Cancer Institute; PIK3CA: Phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit alpha; RPP21: Ribonuclease P/MRP subunit P21; RB1: Retinoblastoma protein 1; SNAI1: Snail family transcriptional repressor 1; TCGA: The Cancer Genome Atlas; TNBC: Triple-negative breast cancer; TP53: Tumor protein 53; WGCNA: Weighted gene co-expression network analysis; WSI: Whole-slide images; ZNRD1: RNA Polymerase I Subunit H, also known as POLR1H.

Acknowledgements

Not applicable.

Authors' contributions

XY and AAI designed the studies. XY, MA, and AAI conducted expression, clinical, and histological image analysis and calculations. XY, MA, LADC, YD, HF, AAI participated in data analysis, discussion, manuscript preparation and editing. XY, HF, and AAI wrote the paper. All authors read and approved the final manuscript.

Funding

This work was supported in part by the National Cancer Institute of the NIH (Cancer Target Discovery and Development Network grants U01CA168449 and U01CA217875, H.F.), Woodruff Health Sciences Center Synergy Award (H.F.), Winship Cancer Institute #IRG-17-181-06 from the American Cancer Society (A.A.I.), NCI Emory Lung Cancer SPORE (NIH P50CA217691) Career Enhancement Program awardee (A.A.I.), and Winship Cancer Institute (NIH 5P30CA138292).

Availability of data and materials

Training images and annotations are publicly available from TCGA sources.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

¹ Department of Pharmacology and Chemical Biology, Emory University School of Medicine, Emory University, 1510 Clifton Road, Atlanta, GA 30322, USA. ² Emory Chemical Biology Discovery Center, Emory University School of Medicine, Emory University, Atlanta, GA, USA. ³ Department of Biomedical Informatics, Emory University School of Medicine, Emory University, Atlanta, GA, USA. ⁴ Department of Pathology, Northwestern University Feinberg School of Medicine, Chicago, IL, USA. ⁵ Winship Cancer Institute, Emory University, Atlanta, GA, USA. ⁶ Department of Hematology & Medical Oncology, Emory University, Atlanta, GA, USA.

Received: 22 May 2020 Accepted: 25 August 2020

Published online: 01 September 2020

References

- Torre LA, Islami F, Siegel RL, Ward EM, Jemal A. Global cancer in women: burden and trends. *Cancer Epidemiol Biomarkers Prev*. 2017;26(4):444–57.
- Newman LA, Kaljee LM. Health disparities and triple-negative breast cancer in African American women: a review. *JAMA Surg*. 2017;152(5):485–93.
- Yedjou CG, Sims JN, Miele L, Noubissi F, Lowe L, Fonseca DD, et al. Health and racial disparity in breast cancer. *Adv Exp Med Biol*. 2019;1152:31–49.
- Amirikia KC, Mills P, Bush J, Newman LA. Higher population-based incidence rates of triple-negative breast cancer among young African-American women: implications for breast cancer screening recommendations. *Cancer*. 2011;117(12):2747–53.
- Nolan TS, Ivankova N, Carson TL, Spaulding AM, Dunovan S, Davies S, et al. Life after breast cancer: 'Being' a young African American survivor. *Ethn Health*. 2019. <https://doi.org/10.1080/13557858.2019.1682524>.
- Yedjou CG, Tchounwou PB, Payton M, Miele L, Fonseca DD, Lowe L, et al. Assessing the racial and ethnic disparities in breast cancer mortality in the United States. *Int J Environ Res Public Health*. 2017;14(5):486.
- DeSantis CE, Ma J, Gaudet MM, Newman LA, Miller KD, Goding Sauer A, et al. Breast cancer statistics, 2019. *CA Cancer J Clin*. 2019;69(6):438–51.
- Monticciolo DL, Newell MS, Moy L, Niell B, Monsees B, Sickles EA. Breast cancer screening in women at higher-than-average risk: recommendations from the ACR. *J Am Coll Radiol*. 2018;15(3 Pt A):408–14.
- Friebel TM, Andrusis IL, Balmana J, Blanco AM, Couch FJ, Daly MB, et al. BRCA1 and BRCA2 pathogenic sequence variants in women of African origin or ancestry. *Hum Mutat*. 2019;40(10):1781–96.
- Hall MJ, Reid JE, Burbidge LA, Pruss D, Deffenbaugh AM, Frye C, et al. BRCA1 and BRCA2 mutations in women of different ethnicities undergoing testing for hereditary breast-ovarian cancer. *Cancer*. 2009;115(10):2222–33.
- Keenan T, Moy B, Mroz EA, Ross K, Niemierko A, Rocco JW, et al. Comparison of the genomic landscape between primary breast cancer in African American versus White women and the Association of racial differences with tumor recurrence. *J Clin Oncol*. 2015;33(31):3621–7.
- Huo D, Hu H, Rhie SK, Gamazon ER, Cherniack AD, Liu J, et al. Comparison of breast cancer molecular features and survival by African and European Ancestry in The Cancer Genome Atlas. *JAMA Oncol*. 2017;3(12):1654–62.
- Pitt JJ, Riestler M, Zheng Y, Yoshimatsu TF, Sanni A, Oluwasola O, et al. Characterization of Nigerian breast cancer reveals prevalent homologous recombination deficiency and aggressive molecular features. *Nat Commun*. 2018;9(1):4181.
- Berger AC, Korkut A, Kanchi RS, Hegde AM, Lenoir W, Liu W, et al. A comprehensive pan-cancer molecular study of gynecologic and breast cancers. *Cancer Cell*. 2018;33(4):690–705.
- Grossman RL, Heath AP, Ferretti V, Varmus HE, Lowy DR, Kibbe WA, et al. Toward a shared vision for cancer genomic data. *N Engl J Med*. 2016;375(12):1109–12.
- Yates B, Braschi B, Gray KA, Seal RL, Tweedie S, Bruford EA. Genenames.org: the HGNC and VGNC resources in 2017. *Nucleic Acids Res*. 2017;45(D1):D619–25.
- Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhim R, Getz G. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol*. 2011;12(4):R41.
- Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Roy Stat Soc: Ser B*. 1995;57(1):289–300.
- Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinform*. 2008;9:559.
- Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. *Stat Appl Genet Mol Biol*. 2005. <https://doi.org/10.2202/1544-6115.1128>.
- Langfelder P, Horvath S. Fast R functions for robust correlations and hierarchical clustering. *J Stat Softw*. 2012;46(11):11.
- Langfelder P, Zhang B, Horvath S. Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics*. 2008;24(5):719–20.
- Pinero J, Ramirez-Anguita JM, Sauch-Pitarch J, Ronzano F, Centeno E, Sanz F, et al. The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Res*. 2020;48(D1):D845–55.
- Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*. 2000;28(1):27–30.
- Fabregat A, Jupe S, Matthews L, Sidiropoulos K, Gillespie M, Garapati P, et al. The Reactome pathway knowledgebase. *Nucleic Acids Res*. 2018;46(D1):D649–55.
- Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA*. 2005;102(43):15545–50.

27. Deng J, Dong W, Socher R, Li L, Kai L, Li F-F, editors. ImageNet: A large-scale hierarchical image database. 2009 IEEE Conference on Computer Vision and Pattern Recognition; 2009 20–25 June 2009.
28. Amgad M, Elfandy H, Hussein H, Atteya LA, Elsebaie MAT, Abo Elnasr LS, et al. Structured crowdsourcing enables convolutional segmentation of histology images. *Bioinformatics*. 2019;35(18):3461–7.
29. Amgad M, Sarkar A, Srinivas C, Redman R, Ratra S, Bechert CJ, et al. Joint region and nucleus segmentation for characterization of tumor infiltrating lymphocytes in breast cancer. *Proc SPIE Int Soc Opt Eng*. 2019. <https://doi.org/10.1117/12.2512892>.
30. Reinhard E, Adhikhmin M, Gooch B, Shirley P. Color transfer between images. *IEEE Comput Graphics Appl*. 2001;21(5):34–41.
31. Li Z, Zheng Z, Ruan J, Li Z, Zeng CM. Chronic inflammation links cancer and Parkinson's disease. *Front Aging Neurosci*. 2016;8:126.
32. Li L. Secondary Parkinson disease caused by breast cancer during pregnancy: a case report. *World J Clin Cases*. 2019;7(23):4052–6.
33. Feng DD, Cai W, Chen X. The associations between Parkinson's disease and cancer: the plot thickens. *Transl Neurodegener*. 2015;4:20.
34. Park S, Shimizu C, Shimoyama T, Takeda M, Ando M, Kohno T, et al. Gene expression profiling of ATP-binding cassette (ABC) transporters as a predictor of the pathologic response to neoadjuvant chemotherapy in breast cancer patients. *Breast Cancer Res Treat*. 2006;99(1):9–17.
35. Yamada A, Ishikawa T, Ota I, Kimura M, Shimizu D, Tanabe M, et al. High expression of ATP-binding cassette transporter ABC11 in breast tumors is associated with aggressive subtypes and low disease-free survival. *Breast Cancer Res Treat*. 2013;137(3):773–82.
36. Ogretmen B. Sphingolipid metabolism in cancer signalling and therapy. *Nat Rev Cancer*. 2018;18(1):33–50.
37. Serie DJ, Crook JE, Necela BM, Axenfeld BC, Dockter TJ, Colon-Otero G, et al. Breast cancer clinical trial of chemotherapy and Trastuzumab: potential tool to identify cardiac modifying variants of dilated cardiomyopathy. *J Cardiovasc Dev Dis*. 2017;4(2):6.
38. Bellmann B, Alushi B, Bigalke B, Landmesser U, Morguet AJ. Restrictive cardiomyopathy: delayed occurrence after radiotherapy of breast cancer. *Wien Klin Wochenschr*. 2017;129(7–8):278–83.
39. Liu LM, Sun WZ, Fan XZ, Xu YL, Cheng MB, Zhang Y. Methylation of C/EBPalpha by PRMT1 inhibits its tumor-suppressive function in breast cancer. *Cancer Res*. 2019;79(11):2865–77.
40. Nakai K, Xia W, Liao HW, Saito M, Hung MC, Yamaguchi H. The role of PRMT1 in EGFR methylation and signaling in MDA-MB-468 triple-negative breast cancer cells. *Breast Cancer*. 2018;25(1):74–80.
41. Zou JX, Duan Z, Wang J, Sokolov A, Xu J, Chen CZ, et al. Kinesin family deregulation coordinated by bromodomain protein ANCCA and histone methyltransferase MLL for breast cancer cell growth, survival, and tamoxifen resistance. *Mol Cancer Res*. 2014;12(4):539–49.
42. Tan MH, De S, Bebek G, Orloff MS, Wesolowski R, Downs-Kelly E, et al. Specific kinesin expression profiles associated with taxane resistance in basal-like breast cancer. *Breast Cancer Res Treat*. 2012;131(3):849–58.
43. Sayyad MR, Puchalapalli M, Vergara NG, Wangenstein SM, Moore M, Mu L, et al. Syndecan-1 facilitates breast cancer metastasis to the brain. *Breast Cancer Res Treat*. 2019;178(1):35–49.
44. Ivanov AA, Khuri FR, Fu H. Targeting protein-protein interactions as an anticancer strategy. *Trends Pharmacol Sci*. 2013;34(7):393–400.
45. Walensky LD. Targeting BAX to drug death directly. *Nat Chem Biol*. 2019;15(7):657–65.
46. Mabonga L, Kappo AP. Protein-protein interaction modulators: advances, successes and remaining challenges. *Biophys Rev*. 2019;11(4):559–81.
47. Santos R, Ursu O, Gaulton A, Bento AP, Donadi RS, Bologa CG, et al. A comprehensive map of molecular drug targets. *Nat Rev Drug Discov*. 2017;16(1):19–34.
48. Niphakis MJ, Lum KM, Cognetta AB 3rd, Correia BE, Ichu TA, Olucha J, et al. A global map of lipid-binding proteins and their ligandability in cells. *Cell*. 2015;161(7):1668–80.
49. Cui L, Li H, Hui W, Chen S, Yang L, Kang Y, et al. A deep learning-based framework for lung cancer survival analysis with biomarker interpretation. *BMC Bioinform*. 2020;21(1):12.
50. Tang Z, Lei S, Zhang X, Yi Z, Guo B, Chen JY, et al. Gsslasso Cox: a Bayesian hierarchical model for predicting survival and detecting associated genes by incorporating pathway information. *BMC Bioinform*. 2019;20(1):94.
51. Baldari S, Ubertini V, Garufi A, D'Orazi G, Bossi G. Targeting MKK3 as a novel anticancer strategy: molecular mechanisms and therapeutical implications. *Cell Death Dis*. 2015;6:e1621.
52. Bossi G. MKK3 as oncotarget. *Aging*. 2016;8(1):1–2.
53. Boyle DL, Hammaker D, Edgar M, Zaiss MM, Teufel S, David JP, et al. Differential roles of MAPK kinases MKK3 and MKK6 in osteoclastogenesis and bone loss. *PLoS ONE*. 2014;9(1):e84818.
54. Samulin Erdem J, Skaug V, Haugen A, Zienolddiny S. Loss of MKK3 and MK2 copy numbers in non-small cell lung cancer. *J Cancer*. 2016;7(5):512–5.
55. Stramucci L, Pranteda A, Bossi G. Insights of crosstalk between p53 protein and the MKK3/MKK6/p38 MAPK signaling pathway in cancer. *Cancers*. 2018;10(5):131.
56. Gurtner A, Starace G, Norelli G, Piaggio G, Sacchi A, Bossi G. Mutant p53-induced up-regulation of mitogen-activated protein kinase kinase 3 contributes to gain of function. *J Biol Chem*. 2010;285(19):14160–9.
57. MacNeil AJ, Jiao SC, McEachern LA, Yang YJ, Dennis A, Yu H, et al. MAPK kinase 3 is a tumor suppressor with reduced copy number in breast cancer. *Cancer Res*. 2014;74(1):162–72.
58. Schieven GL. The biology of p38 kinase: a central role in inflammation. *Curr Top Med Chem*. 2005;5(10):921–8.
59. Cuadrado A, Nebreda AR. Mechanisms and functions of p38 MAPK signaling. *Biochem J*. 2010;429(3):403–17.
60. Ivanov AA, Gonzalez-Pecchi V, Khuri L, Niu T, Wang Y, Xu R, et al. OncoPPI-informed discovery of mitogen-activated protein kinase kinase 3 as a novel binding partner of c-Myc. *Oncogene*. 2017;36(42):5852–60.
61. Li Z, Ivanov AA, Su R, Gonzalez-Pecchi V, Qi Q, Liu S, et al. The OncoPPI network of cancer-focused protein-protein interactions to inform biological insights and therapeutic strategies. *Nat Commun*. 2017;8:14356. <https://doi.org/10.1038/ncomms>.
62. Bild AH, Yao G, Chang JT, Wang Q, Potti A, Chasse D, et al. Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature*. 2006;439(7074):353–7.
63. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdottir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0. *Bioinformatics*. 2011;27(12):1739–40.
64. Liberzon A, Birger C, Thorvaldsdottir H, Ghandi M, Mesirov JP, Tamayo P. The molecular signatures database (MSigDB) hallmark gene set collection. *Cell systems*. 2015;1(6):417–25.
65. Bouvard C, Lim SM, Ludka J, Yazdani N, Woods AK, Chatterjee AK, et al. Small molecule selectively suppresses MYC transcription in cancer cells. *Proc Natl Acad Sci USA*. 2017;114(13):3497–502.
66. Wang Y, Shi J, Chai K, Ying X, Zhou BP. The Role of Snail in EMT and Tumorigenesis. *Curr Cancer Drug Targets*. 2013;13(9):963–72.
67. Liu Y, Du F, Chen W, Yao M, Lv K, Fu P. EIF5A2 is a novel chemoresistance gene in breast cancer. *Breast Cancer*. 2015;22(6):602–7.
68. Ma SY, Park JH, Jung H, Ha SM, Kim Y, Park DH, et al. Snail maintains metastatic potential, cancer stem-like properties, and chemoresistance in mesenchymal mouse breast cancer TUBOP2J cells. *Oncol Rep*. 2017;38(3):1867–76.
69. Holand T, Riffo-Vasquez Y, Spina D, O'Connor B, Woisin F, Sand C, et al. A role for mitogen kinase kinase 3 in pulmonary inflammation validated from a proteomic approach. *Pulm Pharmacol Ther*. 2014;27(2):156–63.
70. Inoue T, Boyle DL, Corr M, Hammaker D, Davis RJ, Flavell RA, et al. Mitogen-activated protein kinase kinase 3 is a pivotal pathway regulating p38 activation in inflammatory arthritis. *Proc Natl Acad Sci USA*. 2006;103(14):5484–9.
71. Kang Y, Wang F, Lu Z, Ying H, Zhang H, Ding W, et al. MAPK kinase 3 potentiates Chlamydia HSP60-induced inflammatory response through distinct activation of NF-kappaB. *J Immunol*. 2013;191(1):386–94.
72. Kim EK, Choi EJ. Pathological roles of MAPK signaling pathways in human diseases. *Biochem Biophys Acta*. 2010;1802(4):396–405.
73. Dong C, Davis RJ, Flavell RA. MAP kinases in the immune response. *Annu Rev Immunol*. 2002;20:55–72.
74. Zhang W, Liu HT. MAPK signal pathways in the regulation of cell proliferation in mammalian cells. *Cell Res*. 2002;12(1):9–18.
75. Wang L, Chen C, Feng S, Lei P, Tian J. Mitogen-activated protein kinase kinase 3 induces cell cycle arrest via p38 activation mediated Bmi-1 downregulation in hepatocellular carcinoma. *Mol Med Rep*. 2016;13(1):243–8.

76. Peng R, Cheng X, Zhang Y, Lu X, Hu Z. miR-214 down-regulates MKK3 and suppresses malignant phenotypes of cervical cancer cells. *Gene*. 2020;724:144146.
77. Zhou M, Yu X, Jing Z, Wu W, Lu C. Overexpression of microRNA21 inhibits the growth and metastasis of melanoma cells by targeting MKK3. *Mol Med Rep*. 2019;20(2):1797–807.
78. Stramucci L, Pranteda A, Stravato A, Amoreo CA, Pennetti A, Diodoro MG, et al. MKK3 sustains cell proliferation and survival through p38DELTA MAPK activation in colorectal cancer. *Cell Death Dis*. 2019;10(11):842.
79. Luo S, Ren B, Zou G, Liu J, Chen W, Huang Y, et al. SPAG9/MKK3/p38 axis is a novel therapeutic target for liver cancer. *Oncol Rep*. 2019;41(4):2329–36.
80. Xie X, Liu K, Liu F, Chen H, Wang X, Zu X, et al. Gossypetin is a novel MKK3 and MKK6 inhibitor that suppresses esophageal cancer growth in vitro and in vivo. *Cancer Lett*. 2019;442:126–36.
81. Gupta J, del Barco Barrantes I, Igea A, Sakellariou S, Pateras IS, Gorgoulis VG, et al. Dual function of p38alpha MAPK in colon cancer: suppression of colitis-associated tumor initiation but requirement for cancer cell survival. *Cancer Cell*. 2014;25(4):484–500.
82. Wakeman D, Schneider JE, Liu J, Wandu WS, Erwin CR, Guo J, et al. Deletion of p38-alpha mitogen-activated protein kinase within the intestinal epithelium promotes colon tumorigenesis. *Surgery*. 2012;152(2):286–93.
83. Xu J, Chen Y, Olopade OI. MYC and breast cancer. *Genes Cancer*. 2010;1(6):629–40.
84. Siddharth S, Sharma D. Racial disparity and triple-negative breast cancer in African-American Women: a multifaceted affair between obesity, biology, and socioeconomic determinants. *Cancers*. 2018;10(12):514.
85. Naab TJ, Gautam A, Ricks-Santi L, Esnakula AK, Kanaan YM, DeWitty RL, et al. MYC amplification in subtypes of breast cancers in African American women. *BMC cancer*. 2018;18(1):274.
86. Khan F, Ricks-Santi LJ, Zafar R, Kanaan Y, Naab T. Expression of p27 and c-Myc by immunohistochemistry in breast ductal cancers in African American women. *Ann Diag Pathol*. 2018;34:170–4.
87. Santoro A, Vlachou T, Luzzi L, Melloni G, Mazzarella L, D'Elia E, et al. p53 loss in breast cancer leads to Myc activation, increased cell plasticity, and expression of a mitotic signature with prognostic value. *Cell Rep*. 2019;26(3):624–38.
88. Toyoshima M, Howie HL, Imakura M, Walsh RM, Annis JE, Chang AN, et al. Functional genomics identifies therapeutic targets for MYC-driven cancer. *Proc Natl Acad Sci USA*. 2012;109(24):9545–50.
89. Mo XL, Qi Q, Ivanov AA, Niu Q, Luo Y, Havel J, et al. AKT1, LKB1, and YAP1 revealed as MYC interactors with NanoLuc-based protein-fragment complementation assay. *Mol Pharmacol*. 2017;91(4):339–47.
90. Heidelberger JB, Voigt A, Borisova ME, Petrosino G, Ruf S, Wagner SA, et al. Proteomic profiling of VCP substrates links VCP to K6-linked ubiquitylation and c-Myc function. *EMBO Rep*. 2018;19(4):e44754.

Publisher's Note

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

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions



Chapter 5

HistomicsTK: an open-source software for computational pathology

HistomicsTK is a python toolkit for organizing, annotating, and analyzing WSI data. It is built as a complement to the *Digital Slide Archive* (DSA) and can be used either as a pure python package for the application of image analysis algorithms or as a server-side plugin for web-based analytics. HistomicsTK is built in collaboration with the company Kitware and is an open-source project at the HistomicsTK [Github repository](#). The graphical user interface associated that enables viewing HistomicsTK annotations within the DSA environment is called HistomicsUI (see this [video](#)).

Candidate's role: Development of workflows to handle annotations and segmentation masks; extending color normalization and augmentation; detection of tissue boundary; efficient detection of highly-cellular regions in WSIs.

Section 5.1

Creation, parsing and expert review of WSI annotations

As we have explored in chapter 2, structured crowdsourcing approaches enable the scalable collection of tissue and cell type annotations, which are in turn critical for developing accurate prognostic models in histopathology. *Structured* crowdsourcing is a hierarchical approach in which non-experts create the majority of data, and pathologists only need to review and approve the data and supplement minor annotation classes. This setup, therefore, necessitates the development of intuitive tools that enable efficient creation, parsing, and expert review of WSI annotations.

5.1.1 Annotation review

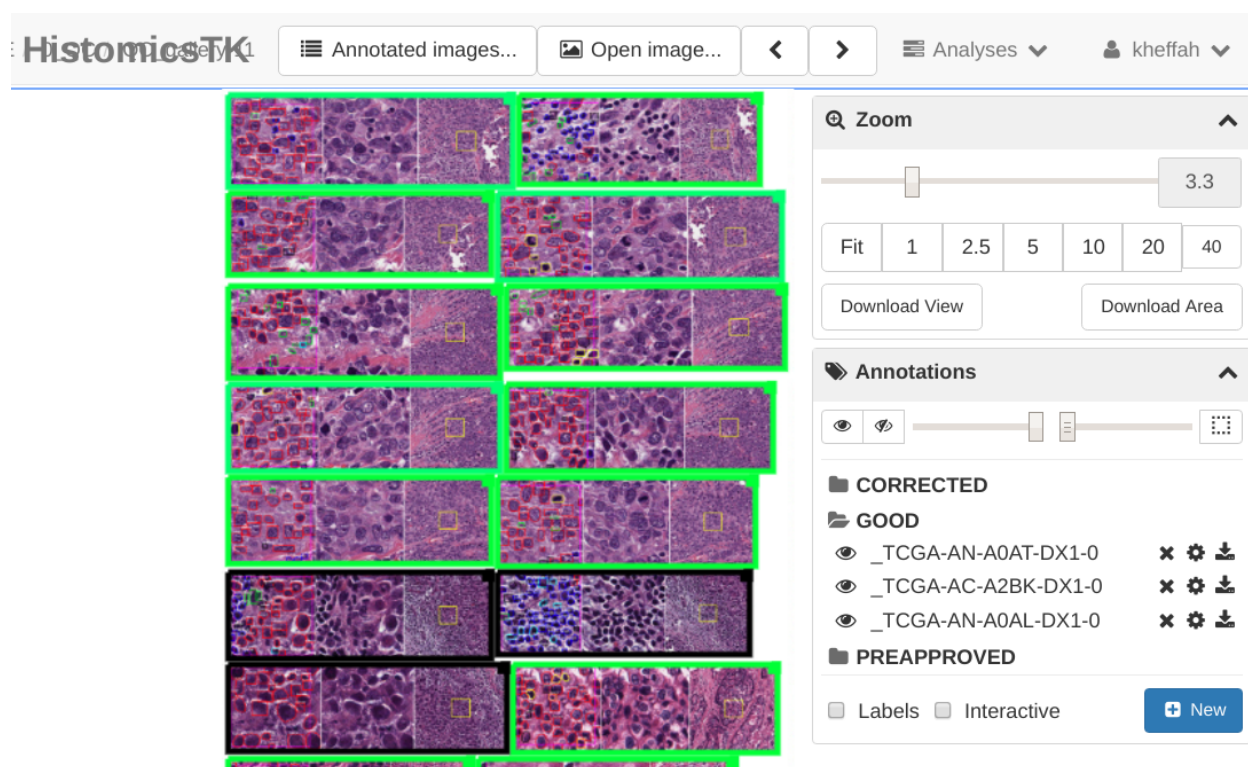


Figure 5.1.1: Use of review galleries for rapid expert review of annotations. A demonstration of this feature can be viewed [here](#).

Annotation studies often focus on small ROIs that are orders of magnitude smaller than whole slide images and sparsely distributed over many slides. Reviewing these annotations involves significant time spent navigating from one ROI to another within and across slides. To aid in review,

we developed a simple tool to generate mosaic gallery images that condense these ROIs into dense multiresolution images that can be viewed in HistomicsUI (Figure 5.1.1). These gallery images speed up the review process by minimizing navigation and the need for toggling annotations (see this [video](#)). Moreover, each ROI within the gallery is hyperlinked to its location within the full WSI, allowing the pathologist to navigate at pan around at various resolutions for wider context. This [documentation page](#) contains further details on the creation of these galleries, geared towards developers and project coordinators.

5.1.2 Annotation storage



Figure 5.1.2: **Back and forth conversion of the annotation database between SQL to MongoDB formats.**

Annotations represent a significant time investment for the users who generate them, and they should be backed up frequently. The simplest way to backup the annotations in a DSA database is to perform a [mongodump](#) operation, which relies on the intrinsic MongoDB database. While frequent mongodump operations are always important to guard against failures, they have the following disadvantages:

- You need to have access to the server where the annotations are hosted.
- The entire MongoDB database is backed up, not just the folder you care about.
- You cannot query the database using SQL queries.

We have created functionality that enables cross-conversion between the unstructured MongoDB-type format and structured tabular formats that are compatible with SQL database ingestion and querying. This functionality allows the recursive backup of a girder database locally as a combination of *.json* files (most similar to the raw MongoDB format), tabular files (*.csv*), and/or an SQLite database. This [documentation page](#) contains further details, geared towards developers and project coordinators.

5.1.3 Supporting annotation workflows for segmentation tasks

The DSA database stores annotations in an (x,y) coordinate list format. For many tasks that process annotation data like training machine learning algorithms or measuring interrater agreement, a mask image representation where pixel values encode ground truth information is more useful. Different segmentation tasks require distinct encoding for the ground truth masks used for training and validating image analysis models. Common encoding schemes include:

- **Semantic segmentation encoding:** The value of each pixel encodes the classification of the corresponding pixel in the image. For example, a value of 1 encodes cancer, 2 encodes stroma, 3 encodes TILs, and so on. If there are multiple objects that have the same class, they are not differentiated. For example, this format would not distinguish between touching TILs cells as they would all be encoded by the same value.
- **Object segmentation encoding:** The value of each pixel encodes the object to which it belongs. For example, 1 encodes a single TIL, 2 encodes another TIL, 3 encodes yet another TIL, 4 encodes a tumor cell, etc. This encoding is usually accompanied by an extra file that maps the objects to their classes, i.e., objects 1, 2, and 3 are TILs, while object 4 is a cancer cell. These masks are typically stored in a binary format (not standard image formats like *.png*) to allow encoding more than 255 objects per image.
- **Panoptic segmentation encoding:** combines semantic and object segmentation encodings in the same mask. One channel in this mask encodes the semantic classification encoding, while the other encodes object ID [91].

Our contributions include the development of workflows for handling both pure semantic segmentation (Figure 5.1.4, as well as panoptic segmentation tasks (Figure 5.1.3). Specifically, we

developed tools for back-and-forth conversion between unstructured records containing coordinates, suitable for viewing on HistomicsUI, and mask formats suitable for training and validating image analysis models. This [documentation page](#) contains details and usage examples, geared towards developers and project coordinators.

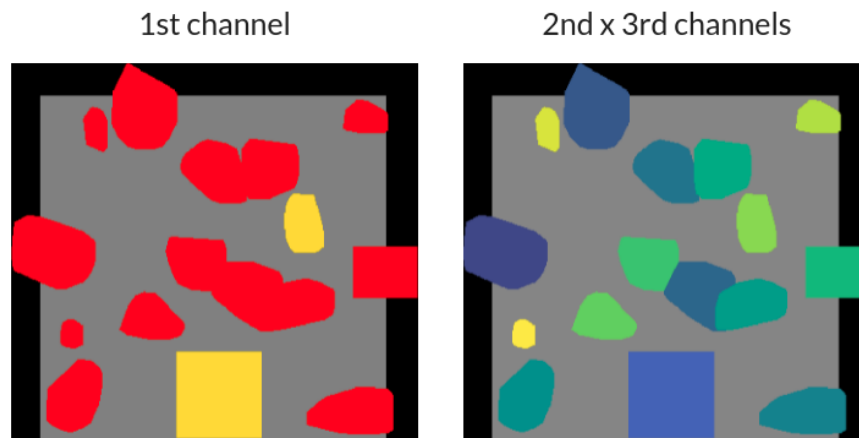


Figure 5.1.3: **Compact format for encoding object and panoptic segmentation masks.** Unlike other applications, there are numerous objects (e.g., nuclei) per histopathology image. This emphasizes the importance of efficient data representation. HistomicsTK saves object segmentation masks as an $(m, n, 3)$ unsigned 8-bit integer array that can be saved as a convenient png image. The first channel encodes semantic labels, for example, whether the nucleus is cancerous. Multiplication of the second and third channels gives the object ID. Hence, there could be a maximum of 32,385 unique objects per image. This arrangement is more compact than traditional object segmentation mask formats, which are very sparse (one object per channel). Note that since the first channel encodes segmentation class, this format can also be used for *panoptic* segmentation tasks. In that case, the segmentation channel can encode "thing" entities (like stromal collagen) with a void object encoding in the second and third channels [91].

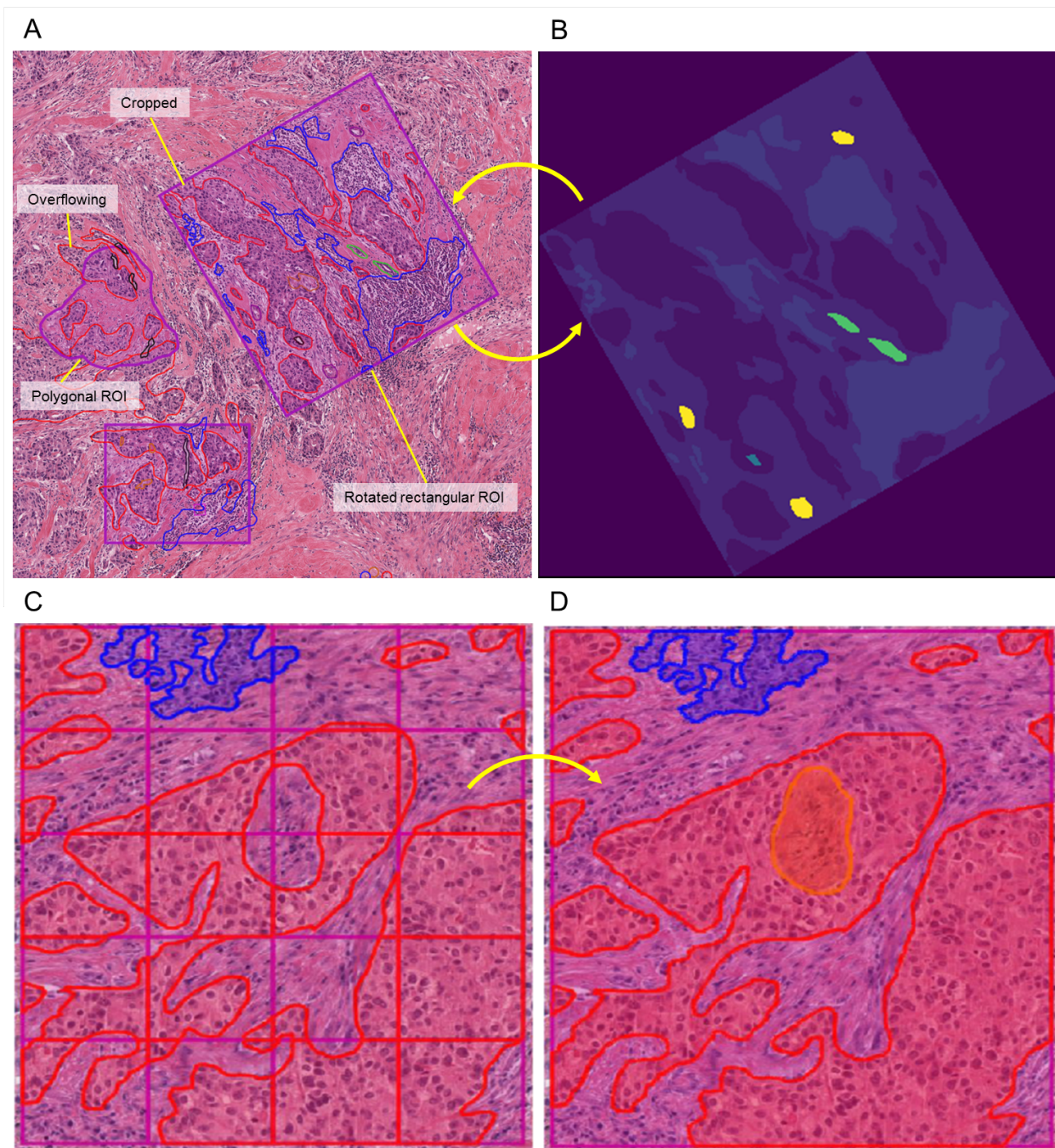


Figure 5.1.4: **Workflows for handling annotations and masks for semantic segmentation tasks.** A. The annotations are drawn by pathologists using the HistomicsTK user interface and are stored in the MongoDB database in the HistomicsTK server. These are parsed, using tools we developed, into labeled masks (where pixel values encode class labels) for use in image segmentation model training and validation. B. Alternatively, labeled masks may be generated by an algorithm (e.g., a semantic segmentation CNN), and HistomicsTK tools we developed are used to extract contours and parse them for viewing and editing in the HistomicsTK user interface. C. and D. Tile-wise outputs from semantic segmentation algorithms are converted to polygons using standard image processing libraries. Then, the spatial relationships between these adjacent polygons are analyzed using HistomicsTK software to enable efficient fusion into WSI-level polygons. This enables the extraction of meaningful morphological features like the perimeter of discrete tumor nests.

Section 5.2

Image processing operations for computational pathology

Empty space and non-tissue elements like pen markings cannot be modeled as a mixture of hematoxylin and eosin stains, and therefore need to be excluded from color normalization and augmentation routines for optimal results. To address this, we updated existing color normalization and augmentation routines in HistomicsTK to allow masking out of non-tissue elements. This results in more natural images with less color artifacts (Figure 5.2.1).

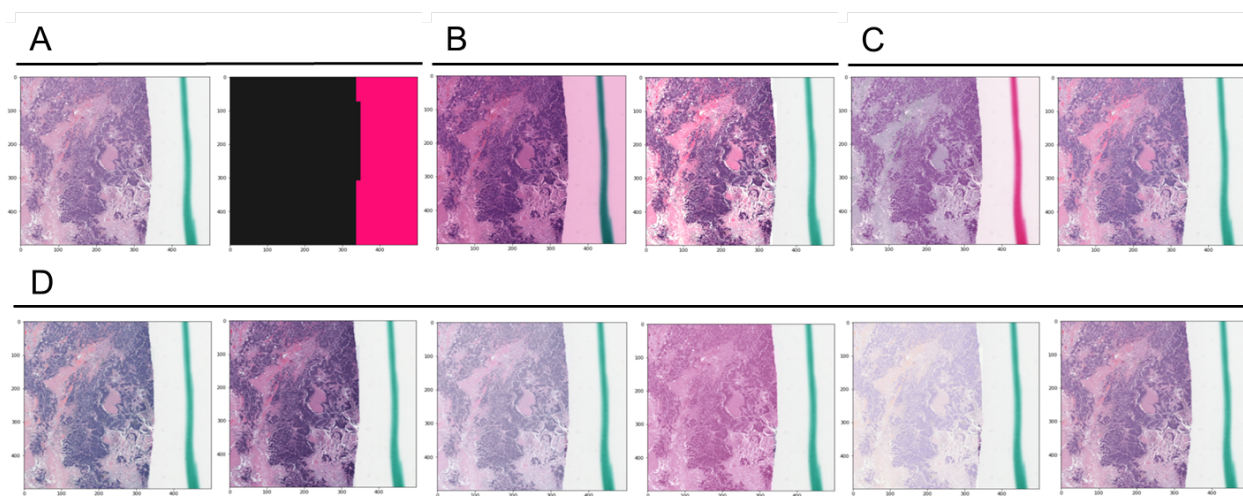


Figure 5.2.1: **Masked color normalization and augmentation** A. Sample tile from the TCGA dataset that includes a portion of tissue, some empty space, and a green pen marking. Simple thresholding routines are used to delineate non-tissue elements, shown in pink. B. Reinhard color normalization without (left) and with (right) masking of non-tissue elements. Masking results in a more natural normalized image. C. Macenko color normalization without (left) and with (right) masking of non-tissue elements. Masking results in a more natural normalized image. D. Masked color augmentation using the stain perturbation method described by Tellez et al.. Only the tissue stain is perturbed, and normalization mimics the expected variation in stain concentration between various slides and labs.

Section 5.3

Simple workflows for detection of salient tissue

Before any image processing routines are applied to a whole-slide image, a set of critical steps need to be performed to ensure that the analysis is focused on relevant and diagnostically-salient tissue areas. These steps include:

- Exclusion of glass and isolation of tissue regions.
- Exclusion of pen markings and inking.
- Exclusion of irrelevant regions of geographic hemorrhage (e.g., bleeding from surrounding vessels at the time of biopsy or resection).
- In the case of invasive carcinomas like breast cancer, isolation of cellular/cancerous tissue regions and exclusion of distant stroma, geographic necrosis, and large fibrotic scar tissue.

We developed tools to support each of the above tasks, as summarized in Figures 5.3.1, 5.3.2, 5.3.3, and 5.3.4. This [documentation page](#) contains details and usage examples. While machine learning methods can technically be used for these tasks, it usually makes more sense to use simple, lightweight methods that rely on the color properties of the image and simple image processing operations for efficient analysis. Besides efficiency, simple workflows like the ones described here do not require any training data and are hence easily portable from one context to another.

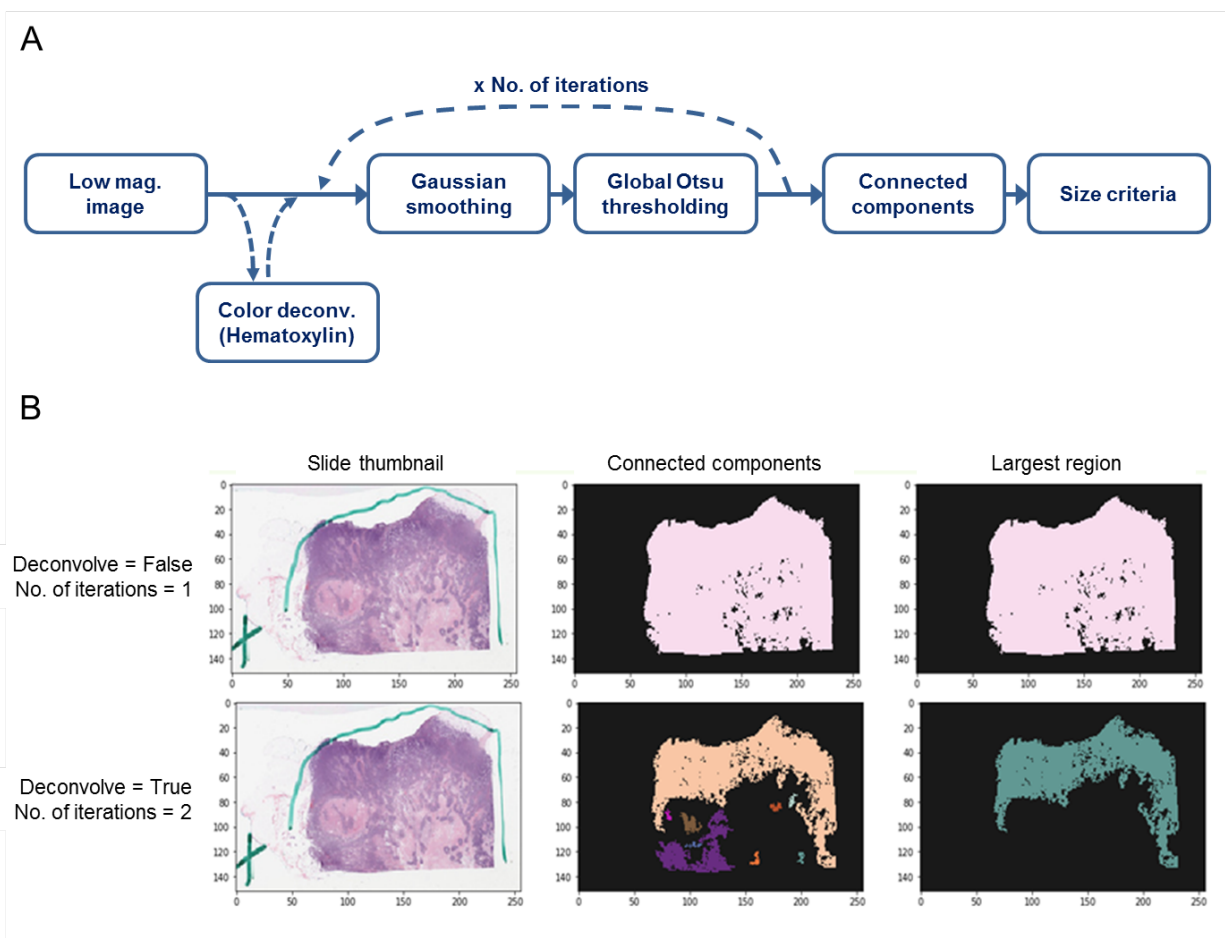


Figure 5.3.1: **Simple thresholding-based tissue detection workflow.** A. A number of image processing steps are used to efficiently segment tissue from non-tissue elements at very low resolution. Dashed arrows indicate optional steps. B. Illustrating the impact of color deconvolution and the number of thresholding steps on tissue detection results.

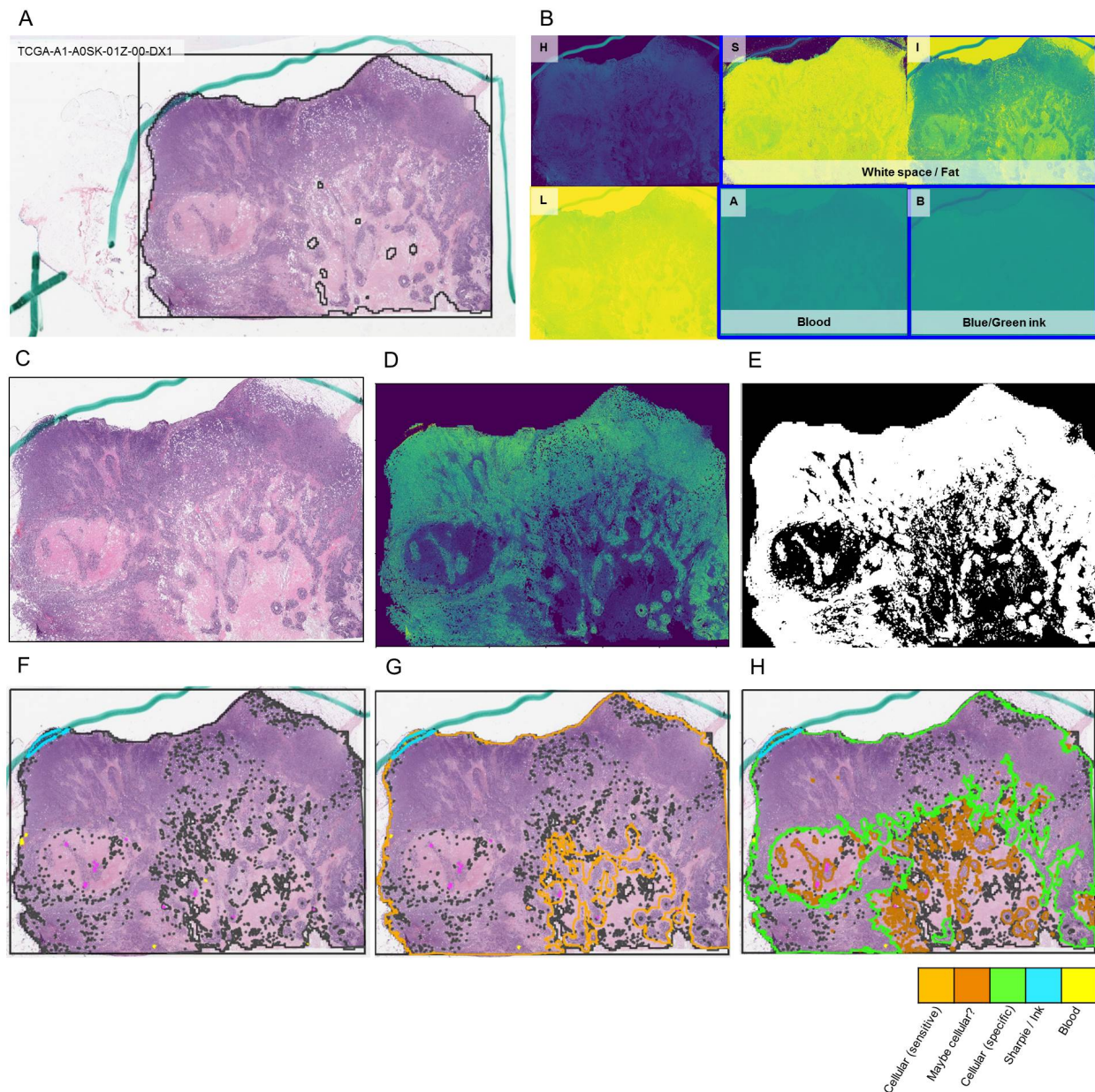


Figure 5.3.2: **Color thresholding-based semantic segmentation at low resolution (methodology)**. A. The first step involves segmentation of tissue elements using, for example, serial Gaussian smoothing and Otsu thresholding (described in Figure 5.3.1) [132]. B. The image is converted to the HSI and LAB color spaces. Different channels preferentially highlight specific image components, which are segmented by setting lower and upper thresholds. This step reliably segments empty space, adipose tissue, blood, blue pen/inking, and green pen/inking. C. The tissue image is color-deconvolved using the Macenko method to isolate the hematoxylin channel. Note that the components from step B cannot be modeled as a mixture of hematoxylin and eosin stains, so they are masked out of the deconvolution. D-F. The hematoxylin intensity channel is thresholded using Otsu's method to isolate cellular regions. G. A number of Gaussian smoothing, Otsu thresholding, and connected component analysis steps are applied to the output from F, similar to Figure 5.3.1. H. Using a smaller Gaussian smoothing threshold results in more fragmented distinct cellular regions. This is a parameter set by the user depending on their specific use case. Size criteria may be adjusted to keep only the largest regions to increase specificity.

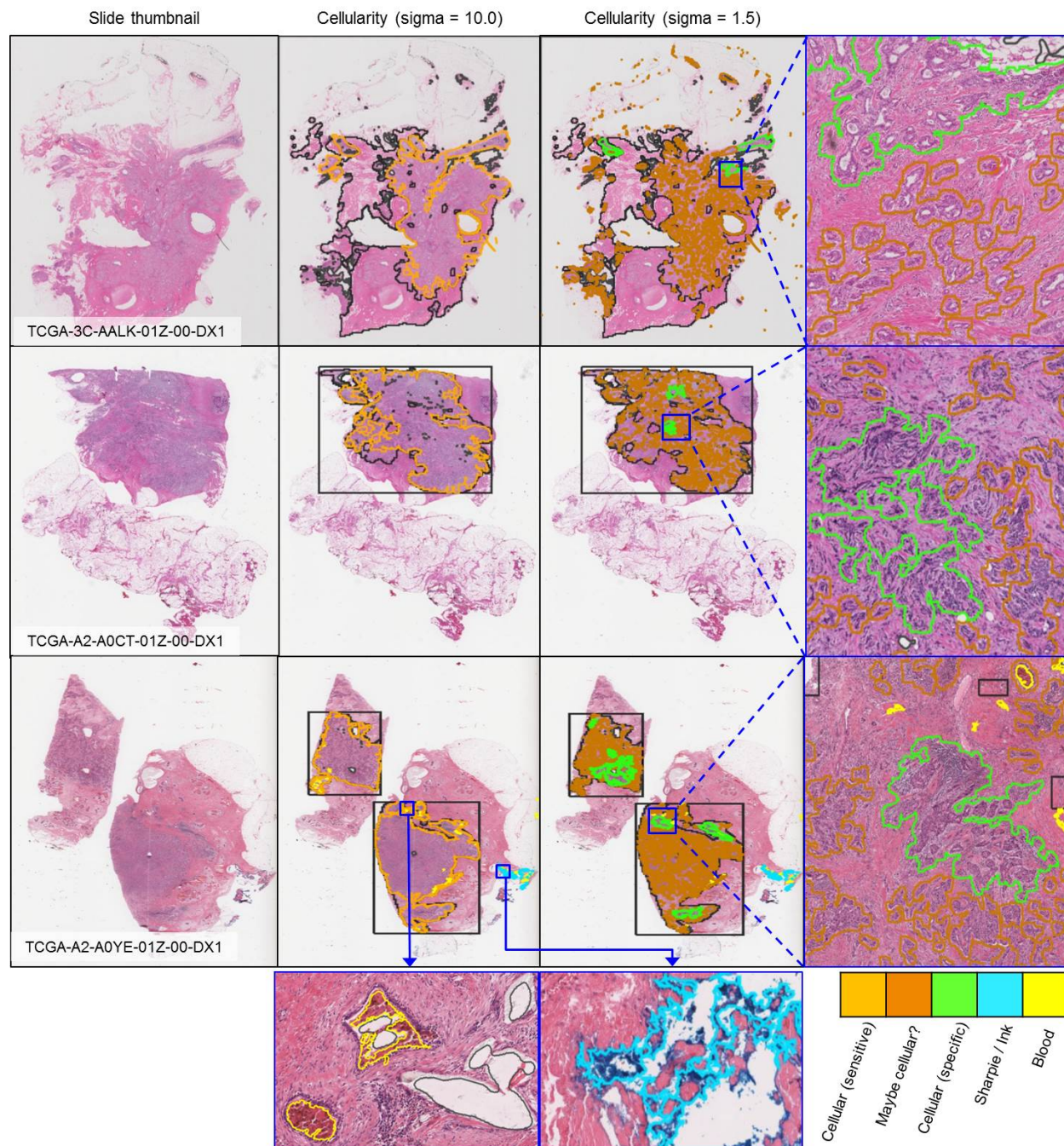


Figure 5.3.3: **Color thresholding-based semantic segmentation (results).** Cellular tissue regions are segmented, and irrelevant material is excluded, based on the color properties of the image. The second column shows using a large Gaussian smoothing filter size, which favors large contiguous regions. A smaller filter size is illustrated in the third column, favoring small discrete regions.

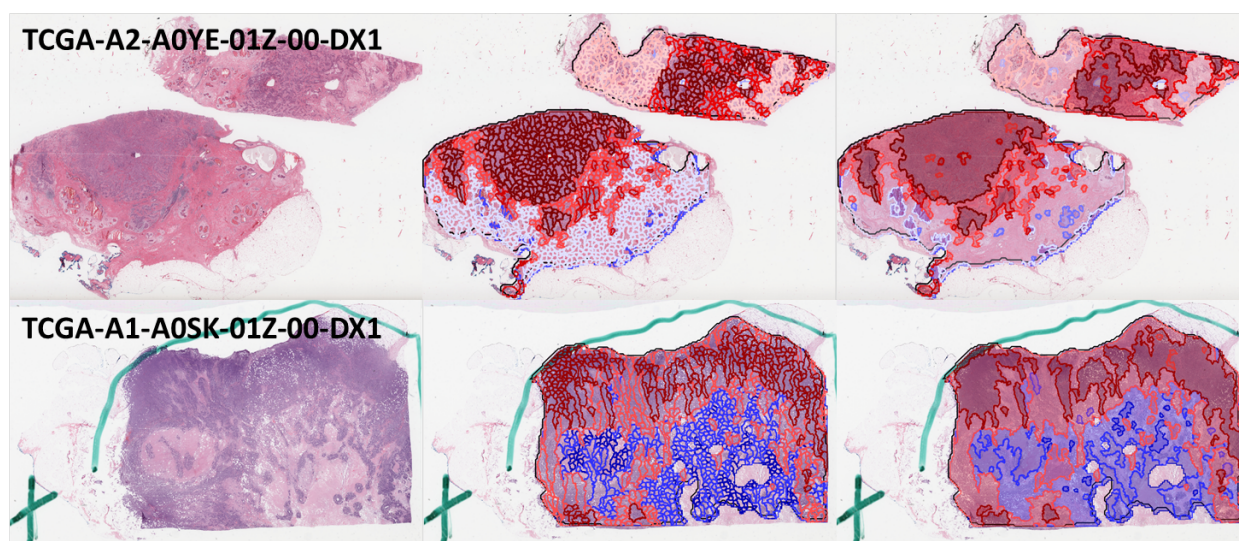


Figure 5.3.4: **Semantic segmentation of cellular regions using superpixel decomposition.** Tissue regions are subdivided into superpixels using Simple Linear Iterative Clustering (SLIC) [4]. The hematoxylin channel is then isolated using Macenko color deconvolution, and various intensity and texture features are extracted for each superpixel independently. Then, a Gaussian mixture model is fit to the features, yielding multiple superpixel clusters. These clusters are ranked from most cellular (red) to least cellular (blue) based on median hematoxylin intensity.

Chapter 6

Summary of conclusions and future directions

In this dissertation, we have presented a body of work that describes the methodology and empirical validation of a framework for the systematic computational discovery of histomic prognostic biomarkers in invasive carcinomas of the breast. *Histomic biomarkers* are a set of prognostically-relevant descriptors that summarize the visual appearance of an H&E stained whole-slide image (WSI) scan using objective and quantitative computational tools. We showed that a weighted combination of these descriptors produces the *Histomic Prognostic Score*, a continuous risk assessment tool that has independent prognostic value in two cohorts of patients with breast carcinomas.

We relied on a modeling paradigm called *concept bottlenecking*, whereby a set of intermediate concepts are extracted first, then used to extract features for prognostic modeling. Concept bottlenecking improves interpretability and trustworthiness in model decisions and acts as a guard against learning spurious correlations. Specifically, we used supervised convolutional neural networks (CNN) to automatically delineate histopathologic regions and nuclei in WSI scans, which were then summarized by a set of morphological and contextual features that were later used for predicting patient survival outcomes.

6.0.1 Structured Crowdsourcing is a viable data collection strategy

One key hurdle we overcame was the lack of large-scale open-access annotation data for training supervised CNN models to delineate tissue regions and nuclei in breast carcinomas. This prompted us to pursue a series of studies investigating a novel data collection approach called *structured crowdsourcing*. We relied on the fact that medical students and fresh graduates have some exposure to pathology as part of their training and have more time and career incentives than practicing pathologists to participate in research projects focused on distributed data collection. Using a hi-

erarchical supervision and quality control scheme, we were able to collect and publish two of the largest open-access datasets for this task in invasive carcinomas of the breast, comprised of 20,000 region annotations and 200,000 annotations of nuclei. Our crowdsourcing approach relies on two facets: 1. a large single-rater dataset produced by non-pathologists, after some initial training, and corrected or approved by practicing pathologists; 2. a much smaller multi-rater dataset to calculate interrater statistics and determine participant reliability. Optionally, the same participant may be assigned the same image under different experimental conditions to assess the impact of experimental design choices. Our key findings were that medical students are reliable annotators for predominant, visually-distinctive patterns but are not reliable for annotating uncommon or ambiguous tissue regions or nuclei. We also show that for some nuclear classes, the burden of pathologist supervision may be reduced by asking multiple non-pathologists to annotate the same image; the exact number of non-pathologists needed to obtain reliable data depends on the nuclear class of interest. Moreover, we showed that the scalability and accuracy of non-pathologist annotations could be improved by offering participants a set of algorithmic suggestions. These suggestions may be *bootstrapped* from heuristic nucleus segmentation algorithms and low-power region classification data and have little or no reliance on annotation data.

6.0.2 Open-source visualization and data management tools are indispensable

Our crowdsourcing projects relied heavily on the open-source whole-slide image (WSI) data and metadata management tool the *Digital Slide Archive*, along with its associated viewer *HistomicsUI* and image processing library *HistomicsTK*. In fact, as we outlined in chapter 5, the crowdsourcing projects prompted us to make systematic improvements to the software in close collaboration with Kitware. We relied on the web-based nature of this software to enable collaborative annotation from tens of participants in multiple countries, with no special installation requirements. This ease of use was critical to our success, and so was the tight coupling of the *HistomicsUI* viewer with an Applications Programming Interface (API) for back-and-forth interaction with the database and for pushing algorithmic outputs.

6.0.3 Crowdsourced annotations can train accurate convolutional models

We showed that crowdsourced annotation data could train accurate supervised CNN models to delineate and classify tissue regions and nuclei in breast cancer. For tissue region delineation, we showed that a standard fully-convolutional architecture (VGG-FCN) could train to automatically detect tissue region patterns with high accuracy, including cancer, stroma, tumor-infiltrating lymphocyte (TILs) aggregates, and necrosis. Nuclear detection and classification, on the other hand, required some modification to existing state-of-the-art. Specifically, the nucleus annotation dataset we obtained differed from standard "natural" object detection datasets in that: 1. objects (nuclei) were much smaller and more numerous per image, and 2. it was a *hybrid dataset*: there was a mixture of bounding boxes (manually placed) and segmentation boundaries (approved algorithmic suggestions). Building on top of the state-of-the-art object detection model, Mask R-CNN, we showed that decoupling the detection and classification tasks improves classification performance. Additionally, we showed that the model learned to produce highly accurate segmentation predictions even though less than half of the training data contained segmentation boundaries.

6.0.4 Customized modeling approaches better suit histopathology applications

Existing CNN explainability approaches fall short when trying to explain the decisions of nucleus classification models. Techniques based on saliency heatmaps, such as GRAD-CAM, are qualitative in nature, suffer from confirmation bias, and provide explanations that do not align with the criteria that practicing pathologists use in their own practice. We introduced a technique called *Decision Tree Approximation of Learned Embeddings* (DTALE), which provides explanations that are highly intuitive, referencing nuclear size, shape, staining, and chromatin clumping. DTALE provides these explanations without compromising accuracy.

We built *MuTILs*, a custom multi-resolution modeling approach that is well-suited for efficient, joint delineation of tissue regions and cells in WSIs. MuTILs has two customizations that make it well-suited for histopathology applications: 1. it has two branches operating at low- and high-resolution; 2. it incorporates a predefined *region prior*, which encourages compatibility between the tissue region and cell type predictions. The region constraint improves predictive accuracy for certain nuclear classes like round fibroblasts and large TILs.

6.0.5 Computational histomic features enable hypothesis-driven biomarker discovery

Our concept bottlenecking approach allowed us to extract a set of highly interpretable features that we used for downstream prognostic modeling. These features included descriptors of cellular abundance, region morphology, nuclear morphology, nuclear and cytoplasmic staining, stromal matrix and collagen disorder, cell-cell interactions, cell clustering, and cell-region interactions. We showed that histomic features measuring architectural disruption and nuclear atypia and pleomorphism represent a computational equivalent of the criteria used in Nottingham grading of breast cancer and that measures of TILs abundance are computational equivalents of the visually-assessed stromal TILs score. This helped provide some extra validation of the meaningfulness of the histomic features we extracted beyond raw accuracy values. We then went on to show how a weighted combination of 26 histomic features, along with the standard IHC panel measuring ER, PR, and Her2+ expression, produces the *Histomic Prognostic Score*, which is highly prognostic in invasive carcinomas of the breast. We showed that most prognostic histomic features differ between the general and high-grade patient populations. Specifically, epithelial features are highly successful in distinguishing low and high-risk patients but are less important in detecting subtle differences in patient survival outcomes within the high-grade sub-population. Within the high-grade patient sub-population, stromal and TILs features have a stronger prognostic value. The Histomic Prognostic Score is prognostic independently of pathologic stage, various risk factors, expression of basal markers, and (in the case of CPS-II) treatment. This prognostic value is stronger than a baseline model based on manual grading and the standard IHC panel.

6.0.6 Future directions

In this dissertation, we have touched on a wide variety of issues ranging from data collection, management, deep-learning image analysis, and prognostic modeling. A set of future research directions and recommendations was mentioned at the end of each section, and we offer a very broad overview here.

Data collection needs to be prioritized in the computational pathology domain. There is a pressing need for computational techniques and methodologies to scale up the data acquisition process. Active and online data collection techniques should be considered for this. Additionally,

there should be some systematic exploration of the possibility of engaging the general public in histopathology data collection, including gamification and other innovations to align the motives and incentives of participants. As we have shown, some histopathologic annotation tasks are almost never going to be relegated to non-pathologists, especially those involving uncommon, visually ambiguous, or high-stakes histology patterns. To collect this data, it may be necessary to explore passive data collection techniques. For example, screen and audio capture may be run in the background while attending physicians teach residents at large academic institutions.

The CNN modeling approach we used relied on MuTILs, a lightweight semantic segmentation-based approach. We showed that this method produces accurate results, but we believe there is room for improvement. There are two avenues for improvement. First, our CNN models did not perform well for detecting normal breast ducts and acini, likely due to limitations in the training data. Expansion of the training dataset is warranted to address this issue. Second, a future version of this approach should combine semantic segmentation of region detection and an object detection-type branch that better accommodates nuclear overlap.

Our prognostic modeling relied on per-patient aggregate data that were obtained using weighted averaging and standard deviation. As we discussed in detail in Section 4.1, this simplification may result in loss of prognostic information and does not model the full heterogeneity within the tumor microenvironment. A future expansion of this work will use more sophisticated modeling to handle this limitation. Finally, our prognostic modeling is limited by the retrospective nature of our data, and future work could address this issue by testing the prognostic value of our histomic signature in a prospective randomized controlled trial setting.

List of recurrent abbreviations

CNN	Convolutional Neural Network
DL	Deep Learning
DSA	Digital Slide Archive
FDA	Food and Drug Administration
FOV	Field of View
GSEA	Gene Set Enrichment Analysis
H&E	Hematoxylin & Eosin stain
HPF	High Power Field
IHC	Immunohistochemistry
ML	Machine Learning
MPP	Microns per pixel
OS	Overall Survival
PCA	Principal Component Analysis
PFI	Progression-free interval
RGB	Red-Green-Blue
ROI	Region of Interest
sTIL	Stromal TIL
TCGA	The Cancer Genome Atlas
TIL	Tumor-Infiltrating Lymphocyte
TNBC	Triple-Negative Breast Cancer
WSI	Whole-Slide Image

Supplementary methods and results

Supplement for Section 2.1: Amgad et al., 2019a

SUPPLEMENTARY METHODS

Triple-Negative status assessment

Evidence of HER2/Neu status was obtained from IHC or FISH results in the clinical file, with positive HER2/Neu status being assigned to cases where there is disparity between the IHC and FISH studies.

Participant training process

Google suite tools, including Docs, Sheets, Slides, and Drive, were utilized to distribute training materials and slide assignments. A preliminary review of slides was first performed to describe their histologic subtypes and patterns. This information was captured in a spreadsheet to facilitate the assignment of slides to participants.

Annotation processing workflow

The DSA server stores polygonal annotations (including corrections) in a Mongo database in a coordinate list format. These coordinates are queried using the DSA REST API and are converted to a mask image format offline, where pixel values encode region class (Figure S1). This mask image conversion greatly simplifies the integration of corrections and calculation of inter-participant agreement statistics. Mask images are then processed to extract the updated polygonal coordinates that are pushed back to DSA for secondary review by the SPs and study coordinator.

Phases of review and correction

Two phases of review and corrections were used (Figure S3). During primary review, only the annotation polygon boundaries are displayed to maximize visibility of underlying tissue structures. This was meant to facilitate detection of major errors, such as polygon misclassifications or missing annotations, and is mainly done by the SPs. During secondary review, polygons are displayed in solid (filled) form, to give a better impression of what the final annotation masks will look like and to maximize visibility of minor artifacts and gaps. Minor corrections to polygon boundaries are made during this phase, mostly by the study coordinator.

Annotation discordance calculation and visualization

We summarize discordance using a statistic we describe as median slide-wise discordance:

$$\widetilde{\Delta} = \text{median}_{s1..10}(\text{median}_{ij1..k}(\Delta_{i,j}))$$

Where $\Delta_{(i,j,s)}$ represents the discordance between participants i and j for evaluation slide s , and k represents number of participant pairs. This statistic first calculates the median discordance between participant pairs for each of the 10 evaluation ROIs, and then calculates the median across these discordances. The expected discordance between two participants on any given slide is represented by Δ . We visualized inter-participant discordance directly on the images to determine where discordance between SPs and NPs occurs within evaluation set regions (Figure S4). To do this, we generated pixel-wise discordance maps using the following procedure for each region class: (1) SP masks were averaged (pixel-wise) to obtain soft ground truth masks. The averaged masks have values in the range [0,1], where 0.5 represents maximum discordance, and 0/1 represents maximum concordance. (2) Discordance masks were generated for each NP by taking the absolute difference between NP masks and the ground truth masks from step 1. (3) Perform pixel-wise averaging of the discordance masks from step 2 over all NPs to visualize the localization of SP-NP discordance in each region.

Fully-convolutional model training

The 16-layer, FCN-8 variant of VGG fully-convolutional network was used in our experiments. The network was trained to map pixels into five region classes: tumor, stroma, inflammatory infiltration, necrosis and other. Regions that belong to rare classes were grouped with predominant were classes where appropriate, as follows: Grouped with "tumor": angioinvasion, DCIS; Grouped with "inflammatory infiltrates": lymphocytes, plasma cells, other immune infiltrates. Each ROI was divided into overlapping 800x800 pixel tiles, and tiles where over 90% of the area was composed of the "don't care" class were ignored. The tiles were stored into *.trecords* files to be used in training. Slides from these institutes were not used in training the final segmentation model on the core set (i.e. were used as an unseen testing set to report accuracy): OL, LL, C8, BH, AR, A7 and A1. The amount of overlap used (and therefore, the amount of shift-augmentation) was inversely proportional to the size of the region of interest; this was done to ensure balanced representation of various histologic patterns regardless of ROI size. Crop augmentation to further increase robustness of training; a randomly-located 768x768 pixel image was cropped on the fly from each tile (after loading in memory) and was used in model training. The models were trained on 3 GPUs with a per-GPU batch size of 4 (total batch size of 12) using data parallelization. Adam optimizer was used with a

learning rate of $1e-5$. Weighted categorical cross-entropy loss was used to mitigate class imbalance, with the weight associated with each class determined by:

$$W_c = \begin{cases} 0 & : \text{if } c = 0 \\ 1 - \frac{N_c}{N} & : \text{if } c > 0 \end{cases}$$

Where N is the total number of pixels in training dataset and N_c is total number of pixels belonging to class c in training dataset.

Patch classification model training

We trained a VGG-16 network to classify 224×224 pixel patches from the three predominant classes: tumor, stroma and inflammatory infiltration, using the same train/test assignment used in the semantic segmentation model. Each ROI was divided into non-overlapping patches at $20 \times$ objective magnification, and patches where the majority class falls below the 50% area were discarded as ambiguous. The convolutional layers and first fully-connected layer were derived from the pre-trained ImageNet VGG-16 network and fixed as non-trainable. Two fully-connected layers were added to the fixed network and trained using cross-entropy loss in TensorFlow. A static testing set was derived from 43 ROIs (13,888 patches), while a variable number of the remaining ROIs were randomly selected for training.

Statistical tests

Mann-Whitney U test was used for unpaired comparisons, and Wilcoxon signed-rank test was used for paired comparisons. A threshold level of 0.05 was used to determine statistical significance.

SUPPLEMENTARY TABLES

Table S1: Guiding instructions given and tips learned from our experience.

Instruction	Rationale
General instructions	
<i>Explain importance of annotation quality for algorithm training. Emphasize importance of accurate conformation to region boundaries.</i>	Some participants may have a misconception that their annotations are only meant to indicate "where the tumor generally is", rather than accurately delineating boundaries.
<i>Explain importance of comfortable workstation, including correct posture, frequent breaks, and mouse usage.</i>	The health and comfort of the participants is important in its own right, and is critical to ensure compliance and high quality annotations.
<i>Provide rules about general diagnostic workflow - zoom in and out and pan the slide for general orientation. Provide criteria on minimum and maximum magnification for annotating classes.</i>	Some region classes are easier to recognize at lower magnification, such as infiltrating lymphocytes which have a "salt and pepper" appearance. Very low magnifications can result in inaccurate boundaries, while very high magnifications significantly increase workload, potentially degrading quality or reducing compliance..
<i>Provide an unambiguous template of histological classes to annotate, including definitions of those classes and how they can be effectively recognized.</i>	This helps standardize the annotation process, reduces variability, and guards against common confusions made by novices.
Instructions to prevent annotation mask artifacts	
<i>When a region extends beyond the ROI, extend the annotation to slightly overlap the ROI boundary.</i>	This prevents gaps between region polygons and the ROI boundary in the annotation mask (See Figure 3B and Supplementary Figure 1).
<i>When a region is too large to enclose in a single polygon, use multiple overlapping polygons.</i>	This avoids inaccuracies caused by participant fatigue. Overlapping annotations of the same region class are fused offline.
<i>Minimize gaps between annotations of different region classes.</i>	This increases accuracy of resultant masks and avoids misclassification of pixels at the interface between non-background region classes as background.
<i>When two regions are enclosed within one another (eg lymphocytic infiltrate within a tumor region), make sure the polygon boundaries are also completely enclosed and non-crossing.</i>	This facilitates conversion of polygonal coordinates into masks by detecting the hierarchy of polygon enclosure (and hence, overlay order) using common image analysis libraries (Figure 3B).
<i>Clear definition of "baseline" or background class</i>	Having a default background class (in our case, stroma) significantly reduces the annotation workload and reduces chances of error.
<i>Clear rules about what constitutes a legal polygon: 1- Moderate-sized closed polygons preferred, even if this means having multiple overlapping polygons to enclose a single anatomical structure; 2- Keep mental image of where the interior of a polygon is; 3- Avoid self-crossing polygon boundaries.</i>	Illegal or malformed polygons are difficult to handle and correct offline, and may significantly degrade the quality of the resultant masks.

Table S2: Number of annotations in final dataset, broken down by region class. * stromal polygon counts only reflect stroma enclosed within non-stromal regions. Other stromal areas are considered the default background class so no polygons were extracted for them.

Broad category	Region class	Annotation count (%)
<i>Predominant classes</i>	Tumor	6536 (32.1%)
	Stroma *	2531 (12.4%)
	Lymphocyte-rich	5066 (24.9%)
	Necrosis or debris	506 (2.5%)
<i>Non-predominant classes</i>	Exclude (artifacts, tears, empty lumina, etc)	1943 (9.6%)
	Adipose tissue (fat)	1108 (5.4%)
	Blood vessel	633 (3.1%)
	Blood (intravascular or extravasated red blood cells)	611 (3.0%)
	Glandular secretions	93 (0.5%)
	Extracellular mucoid material	63 (0.3%)
<i>Challenging classes</i>	Plasma cells	806 (4.0%)
	Mixed inflammatory infiltrates	122 (0.6%)
	Metaplastic changes (osteoid, cartilaginous matrix etc)	4 (0.0%)
	Lymph vessel	11 (0.1%)
	Skin adnexa	2 (0.0%)
	Angioinvasion	12 (0.1%)
	Nerves	1 (0.0%)
	DCIS	9 (0.0%)
	Normal acinus or duct	138 (0.7%)
	Undetermined (eg cannot be determined without IHC)	145 (0.7%)

Table S3: Patch classification AUC and accuracy improves with larger training datasets. Summary of area under receiver-operator characteristics curve (ROC AUC) and accuracy of a convolutional neural network trained to classify patches into tumor, stroma, and inflammatory classes. Each row represents a set of 10-20 experiments where a fixed number of randomly chosen ROIs/slides were assigned to the training set. Number of patches is variable because the size of the ROIs varies between different slides. Numbers presented represent the mean and standard deviation (in brackets) of classification accuracy and AUC on a static testing set composed of 43 ROIs (13,888 patches).

No. of training ROIs	No. of training patches	AUC (Macro-average)	AUC (tumor)	AUC (stroma)	AUC (inflammatory)	Accuracy (overall)	Accuracy (tumor)	Accuracy (stroma)	Accuracy (inflammatory)
2	778.40 (389.94)	0.88 (0.04)	0.88 (0.04)	0.88 (0.03)	0.89 (0.08)	0.73 (0.04)	0.79 (0.04)	0.79 (0.03)	0.88 (0.04)
4	1526.10 (468.80)	0.92 (0.01)	0.92 (0.02)	0.90 (0.02)	0.94 (0.01)	0.78 (0.03)	0.83 (0.03)	0.82 (0.02)	0.90 (0.03)
5	1661.20 (437.09)	0.92 (0.01)	0.92 (0.02)	0.92 (0.01)	0.94 (0.01)	0.78 (0.03)	0.84 (0.03)	0.84 (0.01)	0.90 (0.02)
8	2691.00 (312.69)	0.94 (0.01)	0.94 (0.01)	0.92 (0.01)	0.95 (0.01)	0.81 (0.01)	0.86 (0.01)	0.84 (0.01)	0.92 (0.01)
9	2982.90 (498.45)	0.93 (0.01)	0.93 (0.01)	0.92 (0.01)	0.94 (0.01)	0.80 (0.01)	0.84 (0.01)	0.85 (0.01)	0.92 (0.00)
12	4496.70 (1338.68)	0.94 (0.01)	0.94 (0.01)	0.92 (0.02)	0.95 (0.00)	0.80 (0.05)	0.85 (0.03)	0.85 (0.02)	0.90 (0.04)
16	5160.60 (663.00)	0.94 (0.00)	0.94 (0.01)	0.93 (0.01)	0.95 (0.01)	0.81 (0.01)	0.86 (0.01)	0.84 (0.02)	0.92 (0.01)
32	10854.00 (1016.51)	0.95 (0.01)	0.95 (0.00)	0.93 (0.02)	0.96 (0.00)	0.81 (0.03)	0.87 (0.01)	0.85 (0.02)	0.90 (0.04)
41	14493.10 (1116.62)	0.95 (0.00)	0.95 (0.00)	0.94 (0.00)	0.96 (0.00)	0.83 (0.01)	0.87 (0.01)	0.86 (0.01)	0.92 (0.00)
49	17091.40 (755.45)	0.95 (0.00)	0.95 (0.00)	0.94 (0.00)	0.96 (0.00)	0.83 (0.02)	0.88 (0.01)	0.86 (0.01)	0.92 (0.01)
65	23284.10 (465.63)	0.95 (0.00)	0.96 (0.00)	0.94 (0.00)	0.96 (0.00)	0.82 (0.02)	0.87 (0.02)	0.85 (0.03)	0.93 (0.00)
82	28541.00 (0.00)	0.95 (0.00)	0.96 (0.00)	0.94 (0.00)	0.96 (0.00)	0.83 (0.01)	0.88 (0.01)	0.86 (0.02)	0.92 (0.00)

SUPPLEMENTARY FIGURES

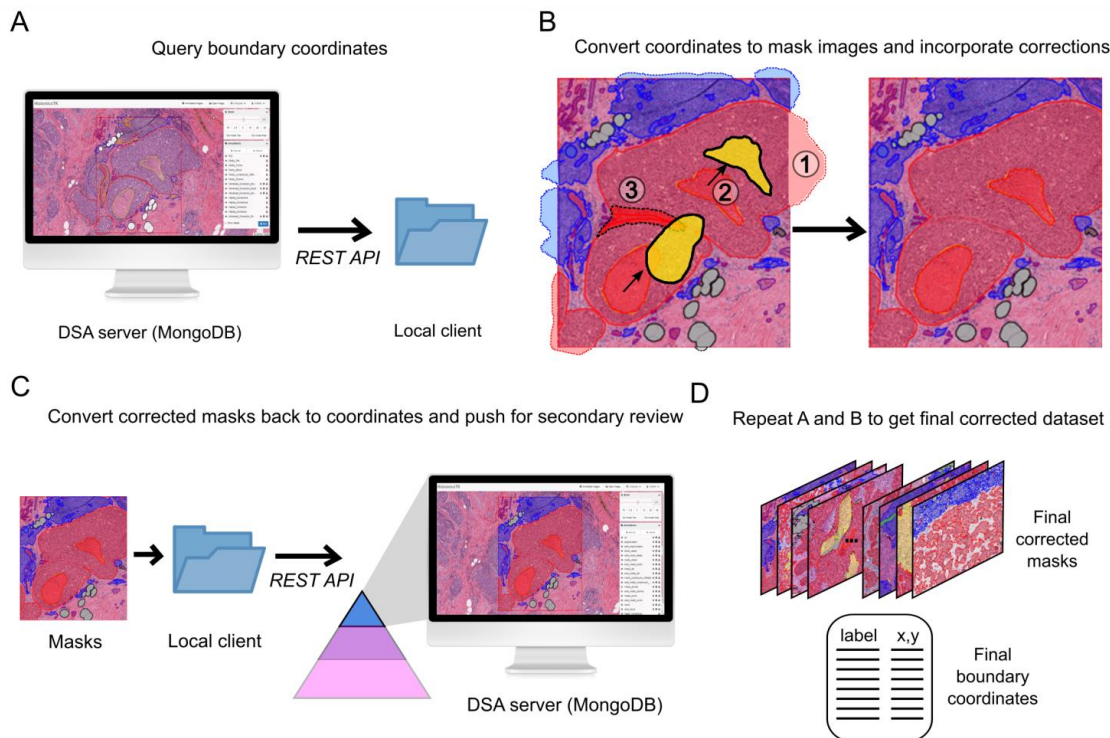


Figure S1: Processing annotations and integrating corrections. (A) The DSA server stores annotations and corrections as polygonal coordinates. The process of integrating corrections and generating masks begins by downloading these polygonal descriptions from the server using the REST API provided by DSA. Making corrections and calculating inter-participant agreement is more easily performed with mask images than with raw polygonal coordinates, while polygons are easier to store in the DSA database. (B) Before integrating corrections, annotations are converted to mask label images in a process that includes 1. Cropping polygons to the ROI boundary 2. Determining order of enclosure (polygons enclosed within other polygons are overlaid on top) and 3. Fusing the correction mask image with uncorrected mask image. (C) The fused mask images are then converted to polygons and uploaded back to DSA using the REST API. (D) Steps A and B are repeated to obtain final corrected dataset.

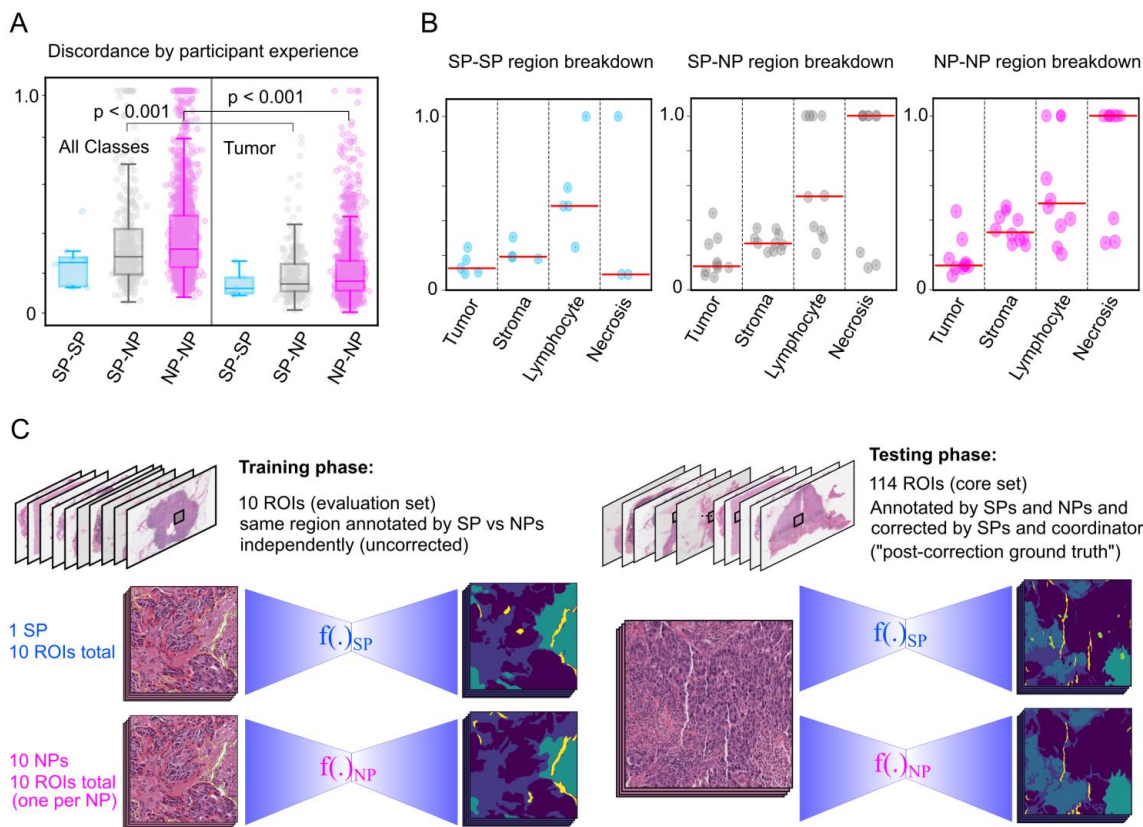


Figure S2: Evaluation slide set discordance and model training process. (A) Discordance by participant experience. Each point represents the discordance between one participant pair for a single slide. (B) Slidewise discordance breakdown by region class. The position of each circle represents the median inter-participant discordance for one slide and one region class. The diameter of the circle is proportional to the number of masks aggregated, eg. some necrotic regions are annotated by some participant pairs, but not others. The red lines indicate the $\bar{\Delta}$ value. (C) Investigating the impact of variability and lack of experience on model accuracy. **Left** - Two training sets were used, one based on a single SP evaluation set annotations, and another based on 10 NP annotations using the same evaluation set ROIs. While some vetting was done by the study coordinator (choosing which mask from which NPs to include), these annotations were used as is. **Right** - The two trained models, one based on a single SP and another based on 10 NPs were evaluated against 114 infiltrating ductal post-correction ground truth annotations from the core slide set.

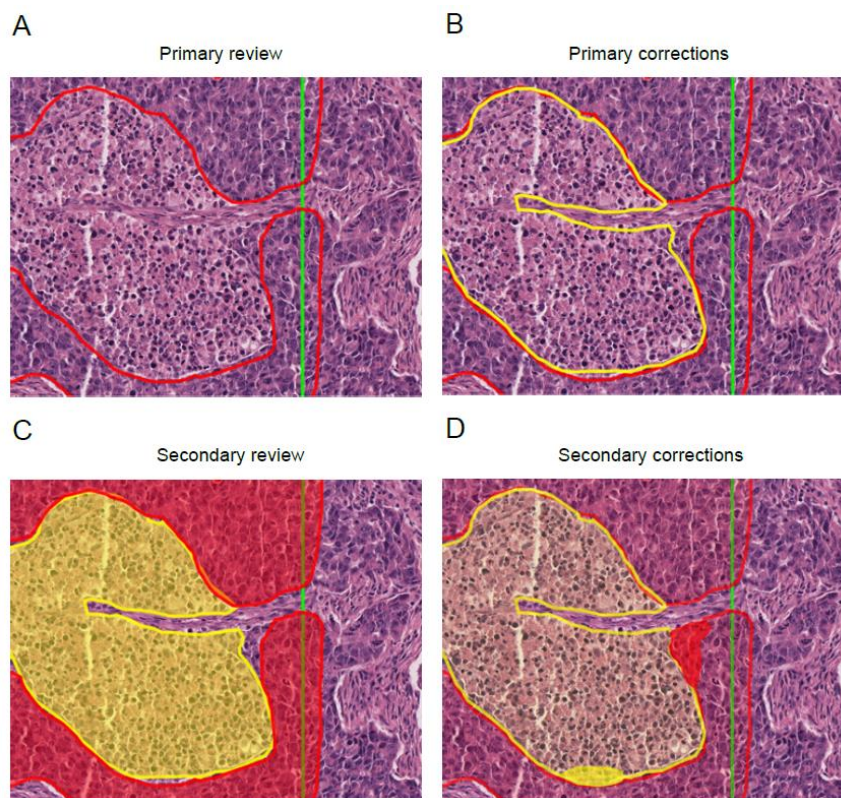


Figure S3: Two-stage review and correction process. The vertical green line represents the region-of-interest boundary. Notice how the participant extended his/her tumor annotation slightly beyond the green region-of-interest boundary in accordance with the annotation instructions in Supplementary Table 1. **(A)** Primary review process involves visualizing the annotation polygon boundaries without fill. **(B)** Major corrections, in this case a missing necrosis/debris region, are made during primary review. **(C)** Secondary review process involves visualization of solid polygons after incorporation of primary corrections. **(D)** Any gaps or artifacts are corrected during secondary review.

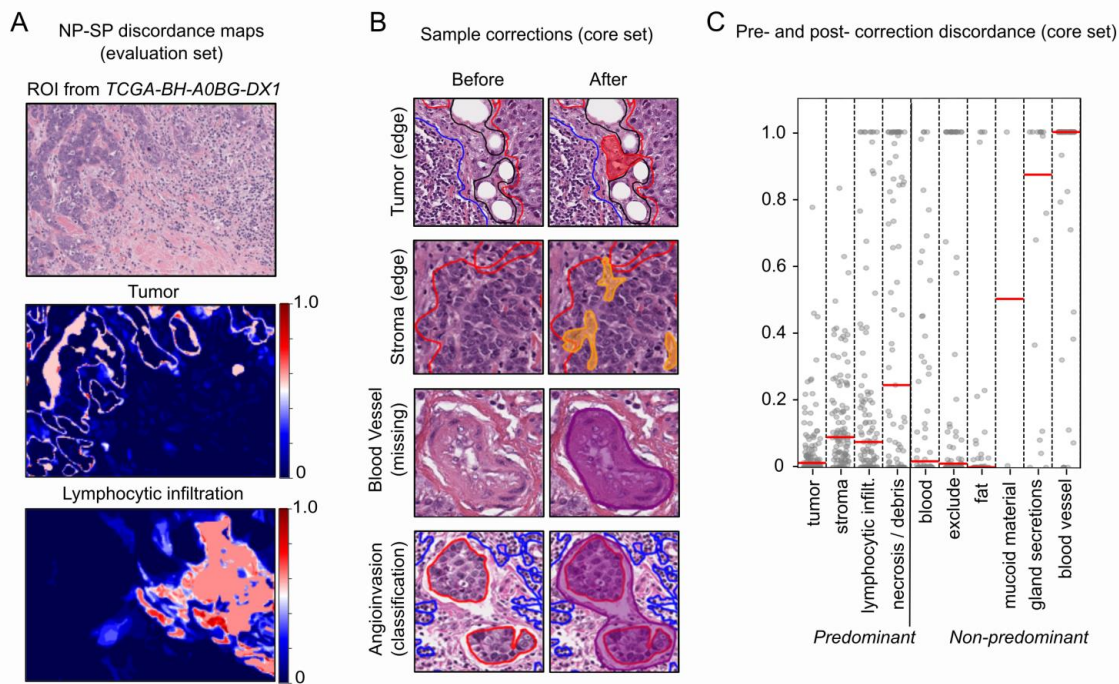


Figure S4: Sources of errors and the annotation correction process. (A) Pixel-wise discordance between senior pathologists and non-pathologists on two evaluation set slides (see supplementary methods). (B) Sample corrections on the core slide set. Original annotations appear as polygon boundaries, while corrections appear as solid polygons. (C) Discordance between pre-correction and post-correction masks for non-pathologists in the core set. The position of each dot represents the median inter-participant discordance for one slide and one region class. The red lines indicate the overall $\bar{\Delta}$ value per region class.

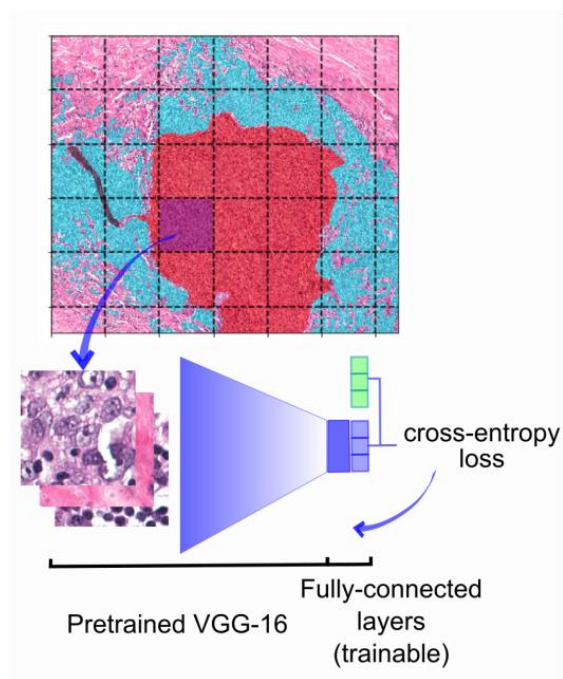


Figure S5: Investigating effect of training dataset size on model generalization. We trained a convolutional neural network to classify patches by pixel class majority into three predominant classes (tumor, stroma, and inflammatory infiltrate). An Imagenet-pretrained VGG-16 architecture was used, with non-trainable weights up to and including the first fully-connected layer, and two trainable fully-connected layers.

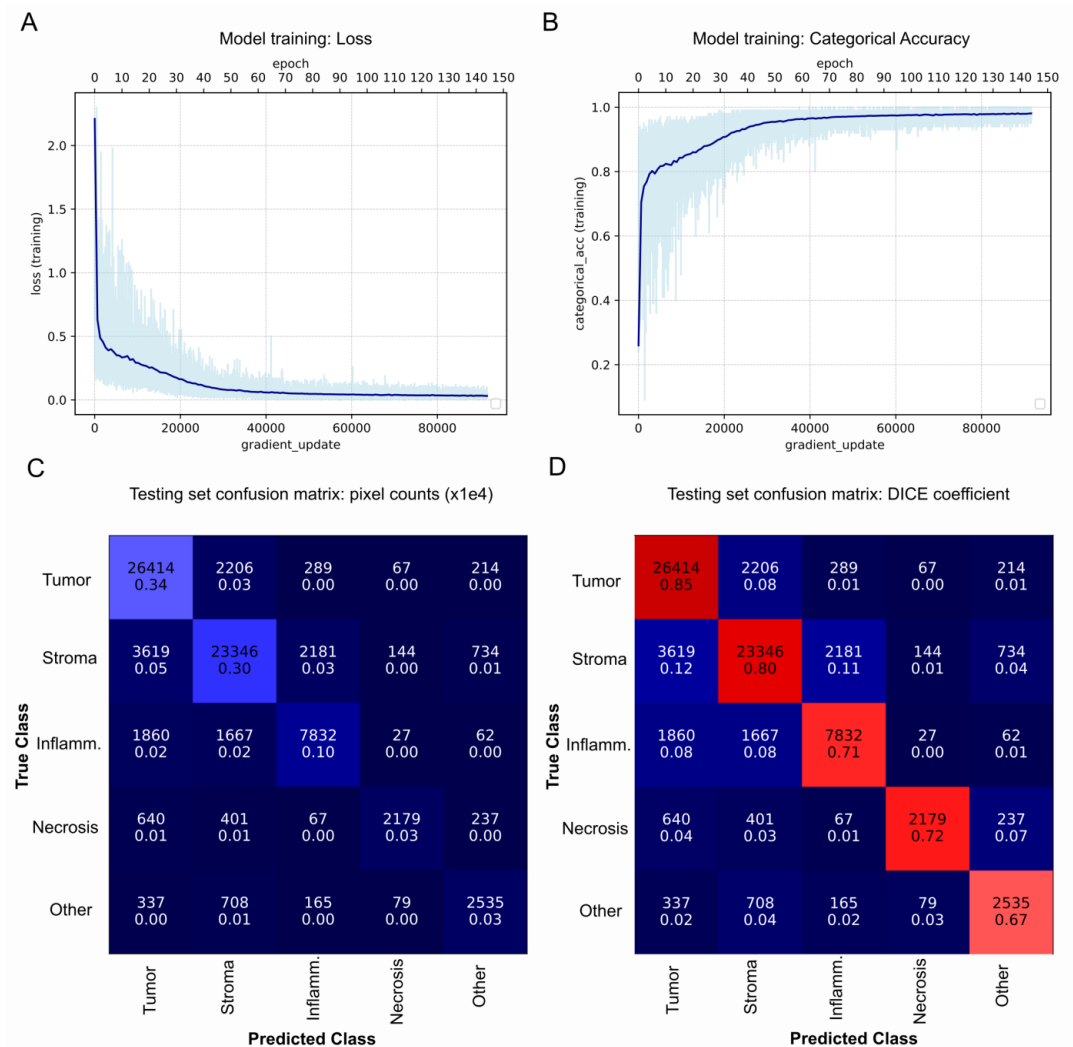


Figure S6: Semantic segmentation accuracy. (A) Loss over the training set. (B) Categorical accuracy over the training set. (C) Confusion matrix over testing set, normalized to pixel counts. Top numbers are pixel counts (x1e4) and bottom numbers represent fraction of total pixel counts. (D) Confusion matrix over testing set using DICE statistic. Top numbers are pixel counts (x1e4) and bottom numbers represent DICE coefficient.

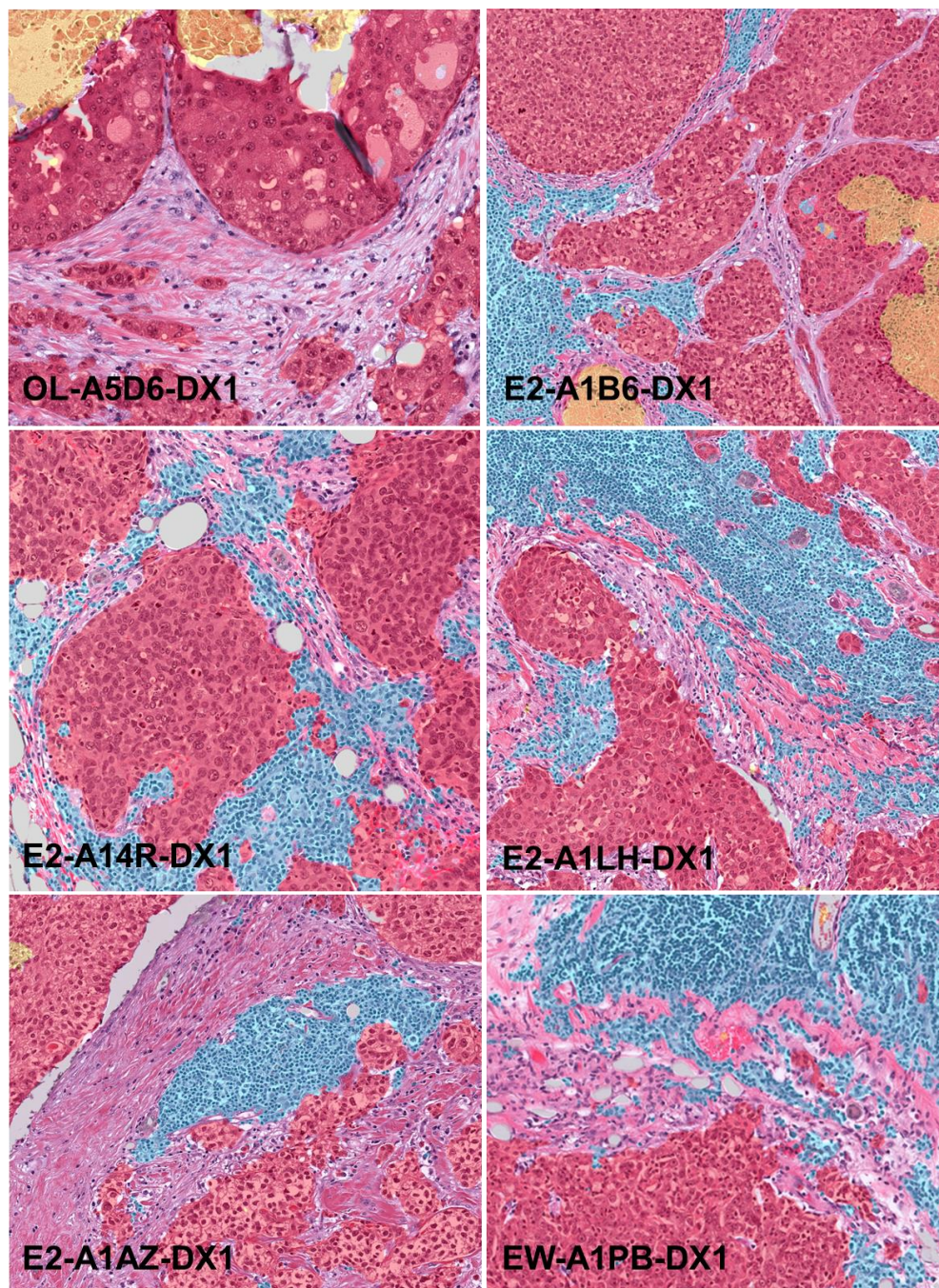


Figure S7: Semantic segmentation visualization over selected testing set ROI's (1). Color codes used: red (tumor); transparent (stroma); cyan (inflammatory infiltrates); yellow (necrosis).

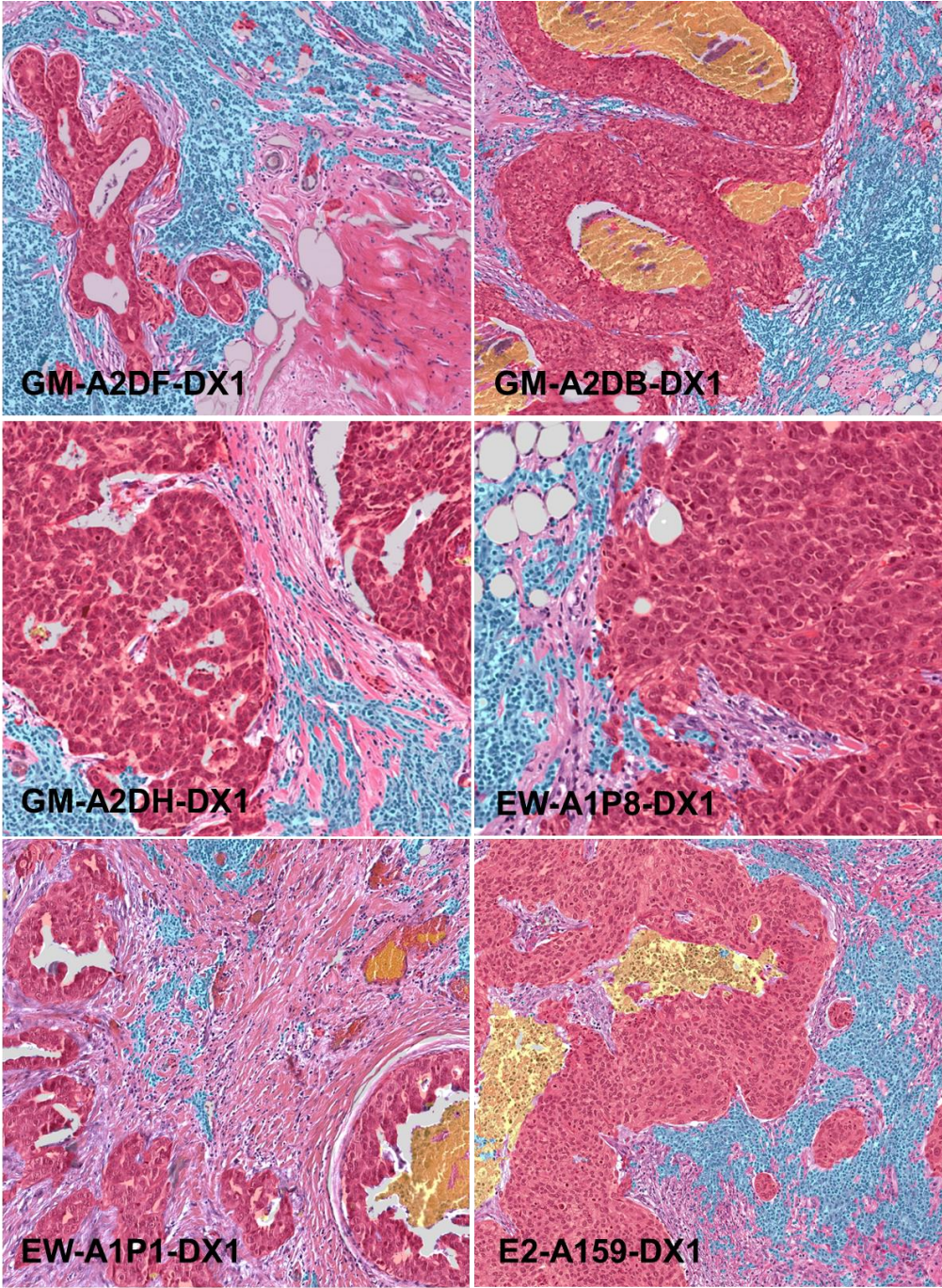


Figure S8: Semantic segmentation visualization over selected testing set ROI's (2). Color codes used: red (tumor); transparent (stroma); cyan (inflammatory infiltrates); yellow (necrosis).

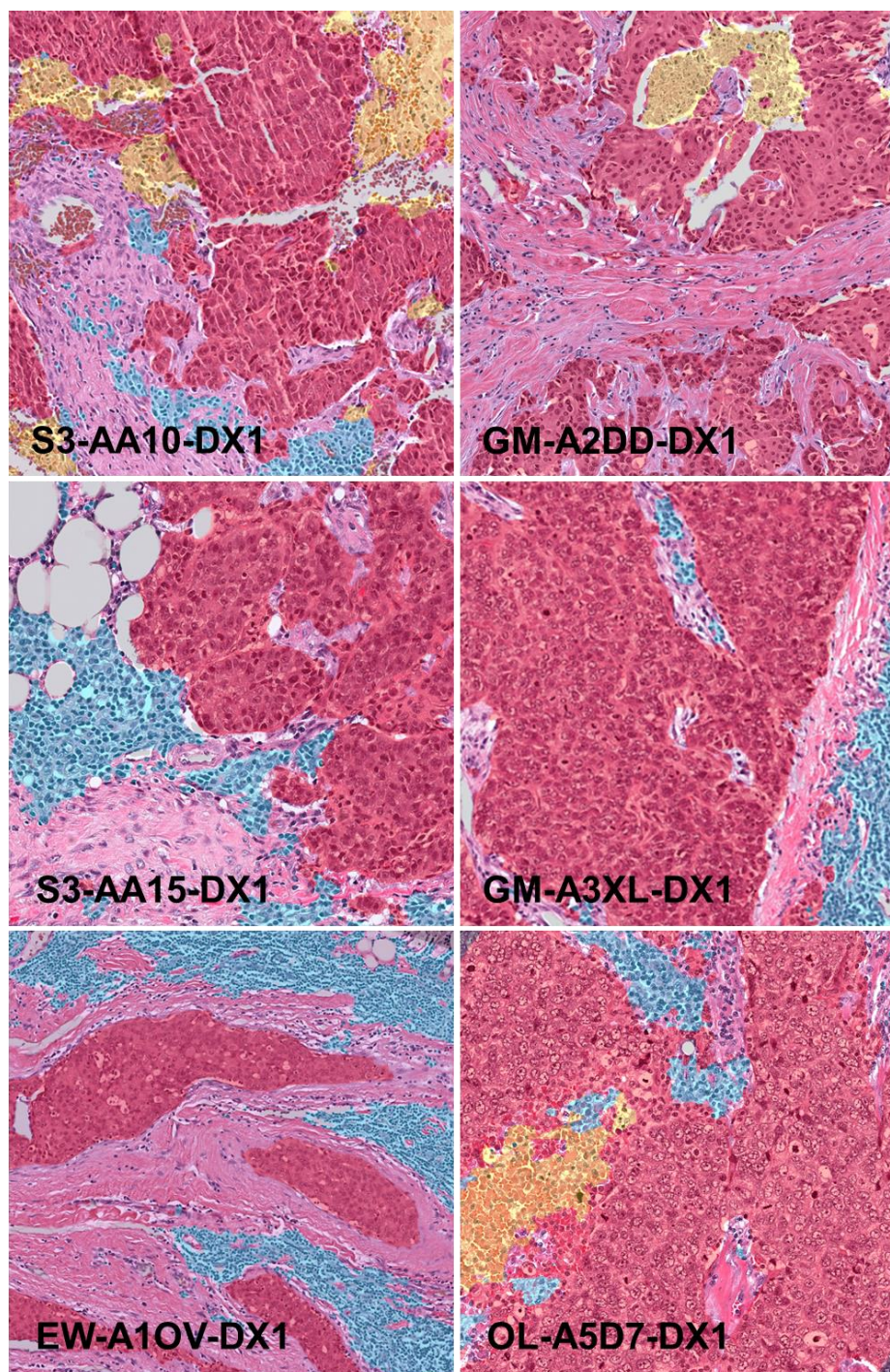


Figure S9: Semantic segmentation visualization over selected testing set ROI's (3). Color codes used: red (tumor); transparent (stroma); cyan (inflammatory infiltrates); yellow (necrosis).

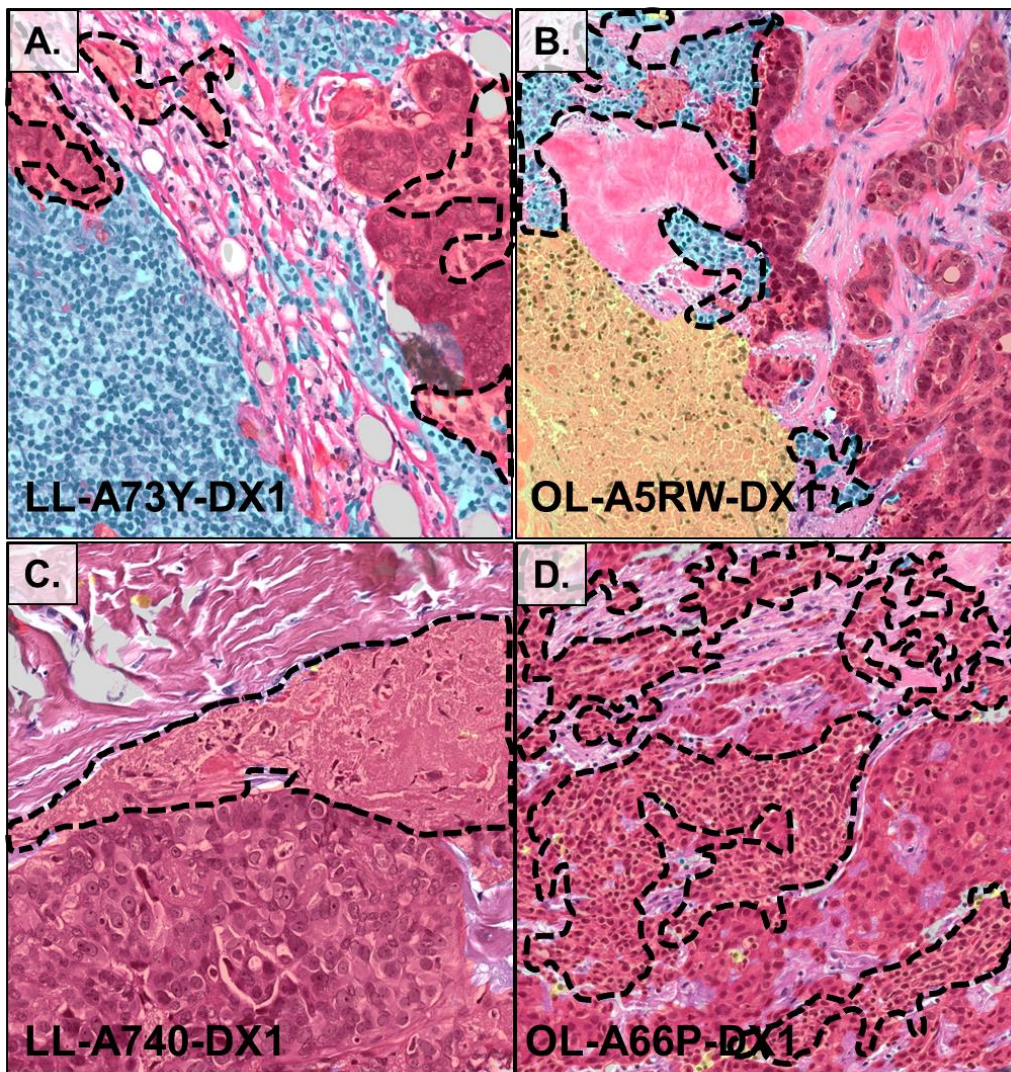
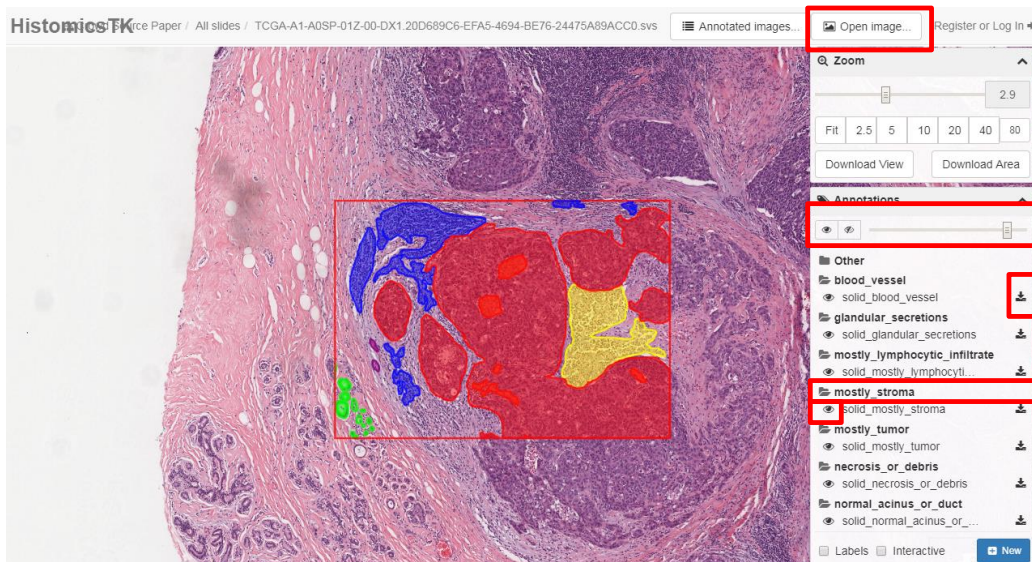


Figure S10: Patterns where semantic segmentation algorithm departs from ground truth annotations (testing set). (A) Stroma intervening small tumor nests misclassified as tumor. (B) Necrotic regions infiltrated by inflammatory cells classified as inflammatory infiltrates. The dataset was constructed such that these patterns are identified as necrosis instead. From a pathology standpoint, this is not a misclassification, though it does differ from ground truth and contributes to accuracy metrics. (C) Hyalinized, acellular stroma misclassified as tumor. This pattern is uncommon and was not represented during model training. (D) Dense plasma cell infiltrates misclassified as tumor. Plasma cells were most commonly present in admixtures with other inflammatory cells, especially lymphocytes, during model training. This pattern is also uncommon and was not represented during model training.

SUPPLEMENTARY DATASET

Dataset visualization (Kitware): The curated dataset can be visualized at the following public instance of the Digital Slide Archive from Kitware Inc.:

<http://demo.kitware.com/histomicstk/histomicstk?image=5bbdee62e629140048d01b0d>



To view a specific slide, click “Open image” in the upper right corner and choose the “Crowdsourced paper” collection. Use the eye icon under the “Annotations” tab to view all annotations, and use the slide bar to control annotation transparency. To view any particular group of annotations, click on the corresponding folder icon, then click the eye icon to toggle visibility. You may download any individual annotation by clicking the downward arrow symbol, although it is probably easier to access the ground truth masks directly using the link and instructions below. Note: for visualization purposes, stromal regions were only explicitly annotated if enclosed within another region class.

Ground truth masks: The ground truth masks for this dataset (to be used for model training and validation) can be found at the following link: <https://figshare.com/s/eae85d914fcb2920da23>. Each mask is a .png image, where pixel values encode region class membership. The meaning of ground truth encoded can be found at the file *gtruth_codes.tsv* found in the same directory. The name of each mask encodes all necessary information to extract the corresponding RGB images from TCGA slides, as follows:

TCGA-¹A1-²A0SK-³DX1_xmin⁴45749_ymin⁵25055_

1. Center/hospital where the patient was treated; 2. TCGA unique patient I.D.; 3. Slide I.D.; 4. Minimum x-coordinates in pixels relative to slide .svs file (at native magnification); 5. Minimum y-coordinates in pixels relative to slide .svs file (at native magnification).

Please be aware that some of the regions of interest are rotated, and that zero pixels represent regions outside the region of interest (“don’t care” class) and should be assigned zero-weight during model training; they do not represent an “other” class. This rotation was done in the interest of capturing adjacent, yet diverse histologic patterns with minimal annotator fatigue.

Supplementary_Tables.xlsx: Raw data and tables used for concordance analysis and convolutional network accuracy reporting. This excel file contains various sheets, described below:

- **Concordance_evaluation_set:** Concordance statistics for participant pairs over the evaluation set. The columns have the following meaning:
 - *Slide_name*: name of the slide from which the evaluation ROI was taken.

- *Participant1 / Participant2*: identifier for study participant who performed the annotation.
 - *Label*: Region class for which the concordance was calculated.
 - *Intersect*: Intersection of the two binary masks. This is the number of pixels commonly classified by both participants as belonging to the label of interest.
 - *Sums*: Bag union of the two binary masks. This is the sum of annotated pixels by each of the participant pairs.
 - *Dice*: Dice coefficient.
- **Concordance_core_set**: Concordance statistics of pre- and post- correction masks. The columns have the same meaning as the evaluation set concordance set.
 - **Patch_CNN_testing_accuracy**: CNN patch classification testing accuracy and AUC for each experiment. The columns have the following meaning:
 - *N_slides_train*: Number of ROI's (each from a unique slide) in the training set.
 - *N_patches_train*: Number of patches in the training set.
 - *Accuracy*: Overall accuracy.
 - *Accuracy_tumor / Accuracy_stroma / Accuracy_inflammatory*: Accuracy breakdown by patch class.
 - *ROCAUC*: Macro-averaged area under receiver operator characteristics curve over testing set.
 - *ROCAUC_tumor / ROCAUC_stroma / ROCAUC_inflammatory*: ROCAUC breakdown by patch class.
 - **FCN_AUC**: Area under ROC curve for softmax pixel values for each slide in the testing set, broken down by region class.
 - **FCN_confusion**: Overall confusion matrix over the testing set, numbers represent pixel counts.

Supplement for Section 2.2: Amgad et al., 2021b

Supplementary tables

Table S1. Definitions and abbreviations used. The following paper from the Digital Pathology Association can be consulted for an expanded list of relevant concepts: *Abels, E. et al., The Journal of Pathology. 2019. 249: 286–294.*

Term	Abbr.	Definition
Basic definitions		
Whole slide image	WSI	High-resolution scanned image of a histopathology slide. Most WSIs of solid tumors are scanned at a 20-40x magnification and are extremely large (~80k pixels side)
Annotation	-	Manual markup of the image to indicate the location, boundary, or class of an anatomical structure. Examples include a point at the centroid of a nucleus, a bounding box indicating the extent of a nucleus, or tracing the nucleus boundary.
Segmentation	-	A boundary delineating the edge of a structure like a histologic region or a nucleus.
Ground truth	-	The true location/boundary/class of a particular nucleus. This term is used loosely in this paper to refer to the truth against which the deep-learning models are evaluated. This truth will be different under different circumstances, depending on the experiment being discussed.
Region of interest	ROI	A ~1 mm ² region of a WSI from which FOVs are selected. Each ROI is accompanied by low-power annotations of tissue regions used for generating suggestions.
Field of view	FOV	A ~65 x 65 μm field selected from within an ROI. FOVs were annotated at high power to indicate the location and class of all nuclei contained in the FOV.
Application Programming Interface	API	A set of functions that allow developers to interact with a database or other software programmatically.
Participant groups		
Non-pathologists	NPs	Medical students/graduates who did not receive pathology residency training.
Junior pathologists	JPs	Pathology residents with < 2 years of anatomical pathology training.
Senior pathologists	SPs	Attendings or pathology residents with > 2 years of anatomical pathology training.
Pathologists	Ps	Junior or senior pathologists.
Datasets		
Hybrid dataset	-	A dataset where participants click accurate segmentation boundary suggestions and draw bounding boxes around all other nuclei. The resultant dataset contains a mixture of segmentation boundaries and bounding boxes.
Single-rater dataset	-	A collection of FOVs that NPs annotated in a single-rater manner. NPs received pathologist feedback during the annotation process. NPs were shown both region (low-power) and nucleus (high-power) suggestions while annotating.
Corrected single-rater dataset	-	A subset of single-rater dataset FOVs (approximately half) whose annotations have been manually corrected by study coordinators based on feedback from a senior pathologist. A senior pathologist approved all corrected single-rater dataset annotations.
Uncorrected single-rater dataset	-	Single-rater dataset FOVs whose annotations were not manually corrected. The quality of these annotations is participant-dependent.
Multi-rater datasets	-	A collection of FOVs that were annotated by multiple participants under different experimental conditions. NPs were not given feedback on these FOVs. These are used for interrater comparisons.

Evaluation dataset	-	A multi-rater dataset where Mask R-CNN refined algorithmic suggestions were shown to the participants. Refinement was applied to bootstrap suggestions to improve quality. These suggestions were the same type used in single-rater dataset annotation.
Bootstrap control	-	A multi-rater dataset where noisy bootstrapped algorithmic suggestions were shown to the participants. These suggestions were generated using a heuristic segmentation algorithm and processing of low power region annotations and shape data from segmentation.
Unbiased control	-	A multi-rater dataset where no annotation suggestions were shown to the participants. This was the first multi-rater dataset annotated to obtain annotations not biased by algorithmic suggestions.
Nucleus suggestions and labels		
Bootstrapped suggestions	-	A set of noisy nuclear boundary suggestions using simple image processing heuristics. Each boundary also had an associated classification suggestion, inherited from the histologic region where the presumed nucleus resides. Thus, for example, a suggested boundary in a tumor region would be associated with a tumor classification suggestion. These were an intermediate step in producing refined suggestions (see below) and were only shown to participants for the Bootstrap control dataset.
Mask R-CNN refined suggestions	-	The result of fitting a Mask R-CNN model to the bootstrap suggestion. Mask R-CNN acts as a function approximator to smooth out noise. These were shown to participants for the single-rater and Evaluation datasets.
Label	-	This term is used in the broad sense, as in <i>labeled data</i> used for supervised machine learning. A label is a tag associated with a potential nucleus location (anchor proposal, defined below). Labels include assessing whether an anchor proposal corresponds to a nucleus (i.e., detection), what class to assign (e.g., tumor) and whether or not the suggested segmentation boundary is correct.
Anchor proposal	-	A <i>potential</i> bounding box location of a nucleus. Anchor proposals are generated by clustering annotations from multi-rater datasets.
Class	-	A type of label that assigns a nucleus to a set of predefined biological categories (e.g., tumor, fibroblast, and TILs).
Raw nucleus classes	-	The set of 12 nucleus classes that were directly obtained from the participants, without class grouping.
Nucleus classes	-	A set of 7 nucleus classes, obtained by grouping related raw classes together.
Nucleus super-classes	-	Three clinically salient nucleus classes (tumor, stroma, sTILs), obtained by grouping nucleus classes.
Uncommon nucleus classes	-	Any raw nucleus classes other than tumor, fibroblasts, and lymphocytes.
Inferred pathologist truth	P-truth	A single label is generated from the analysis of multi-rater datasets using pathologist annotations. For each anchor proposal from clustering, we use EM to infer whether the proposal is an actual nucleus, the class, and the correctness of the suggested boundary. This was used to measure the accuracy of NP annotations and NP-label.
Inferred non-pathologist label	NP-label	A single label is generated from the analysis of multi-rater datasets using NP annotations (see inferred P-truth for comparison).
Machine learning and image processing		
Convolutional neural network	CNN	A deep-learning model that operates on image data.
Mask R-CNN	-	A CNN model that learns to jointly predict nucleus bounding-box localization, segmentation, and class.

Agglomerative hierarchical clustering	-	A bottom-up clustering approach that builds a hierarchy of clusters starting with each data point as its own cluster and grouping data points and clusters by similarity.
Expectation-Maximization	EM	An iterative method for estimating the parameters of a statistical model by maximizing a <i>likelihood</i> measure. It was used to simultaneously estimate participant reliability and nucleus locations, class, and correctness of boundaries.
Heuristic nucleus segmentation	-	Delineation of nuclear boundaries using simple image processing operations that have no dependence on annotation data (unlike machine learning models). This was used to generate bootstrapped algorithmic suggestions or segmentation boundaries.
Measures of accuracy and agreement		
Intersection over union	IOU	A quantitative measure of overlap of prediction and truth.
DICE coefficient	-	Similar to IOU, it is a measure of overlap of prediction and truth.
Area under Receiver-Operator Characteristic (ROC) curve	AUROC	It is a measure of accuracy, where a value of 0.5 corresponds to random chance, and a value of 1.0 is the maximum. There are two ways of obtaining this value: - <i>Micro-average</i> : This is the overall accuracy, where different nucleus classes contribute to the result in proportion to their abundance in the dataset - <i>Macro-average</i> : Is the class-balanced accuracy, where different nucleus classes are equally weighted, such that an uncommon class like macrophages will have the same contribution as a common class like sTILs.
Average precision	AP	The area under the precision-recall curve is used to measure detection performance. AP@.5 refers to the area measured with a minimum IOU of 0.5 for defining correct detections. mAP@.5:.95 is a more stringent measure that averages areas for a range of IOU thresholds from 0.5 to 0.95.
F1 score	-	The harmonic mean of precision and recall values.
Matthew's Correlation Coefficient	MCC	A balanced measure of classification accuracy considers all components of the confusion matrix, including true negatives (unlike the F1 score).
Cohen's Kappa statistic	-	A measure of agreement between two participants, ranging from -1 (perfect disagreement) to +1 (perfect agreement).
Krippendorff's Alpha statistic	-	A multi-rater generalization of Cohen's Kappa, which handles missing values.

Table S2. Accuracy of algorithmic suggestions. The accuracy is measured against the corrected single-rater dataset. Mask R-CNN refinement of the bootstrapped algorithmic suggestions results in better detection suggestions. Low-power region-based classification was more accurate than Mask R-CNN-derived classes. Note, however, that this was FOV-dependent, and there were some FOVs in which the Mask R-CNN prediction was better than relying on low-power regions for classification.

Stage	Class	N	Accuracy	MCC	F1	Precision	Sensitivity	Specificity	
Bootstrap suggestions	Detection	58598	18.8	-	31.7	40.4	26.1	-	
	Classification (region-inherited)	Overall	11029	86.8	77.9	-	-	-	-
		Tumor		95.6	91.2	95.2	93.6	96.8	94.7
		Stromal		90.3	21.5	12.4	80.0	6.7	99.8
		sTILs		89.6	80.3	89.2	82.6	97.0	83.7
		Other		98.2	16.8	17.7	16.9	18.6	99.0
Suggestions after Mask R-CNN refinement	Detection	75908	31.5	-	47.9	47.6	48.1	-	
	Classification (region-inherited)	Overall	23874	78.9	67.6	-	-	-	-
		Tumor		93.5	85.9	91.1	90.1	92.1	94.2
		Stromal		82.0	46.1	57.2	53.4	61.6	87.0
		sTILs		83.6	66.4	80.2	84.2	76.6	88.9
		Other		99.3	30.5	25.9	56.0	16.9	99.9
	Classification (Mask R-CNN prediction)	Overall		69.1	52.7	-	-	-	-
		Tumor		83.2	63.0	74.9	82.1	68.8	91.4
		Stromal		82.2	26.3	17.5	91.3	9.6	99.8
		sTILs		75.5	58.1	77.4	64.5	96.8	59.0
		Other		97.9	12.0	11.9	8.5	19.9	98.5

Table S3. Hyperparameters used for Mask R-CNN model training.

Backbone	Resnet50
Pretraining	Imagenet
Input (cropped) image size	128 x 128
Max. ground truth nuclei per image	30
Max. detections per image (inference)	200
Batch size	8
Optimizer	SGD
Learning rate	1.00E-04
Momentum	9.00E-01
Length of anchor sides in pixels	8,16,32,64,128
ROIs after NMS (training)	500
ROIs after NMS (inference)	1000
NMS threshold for RPN proposals	0.7

Supplementary figures

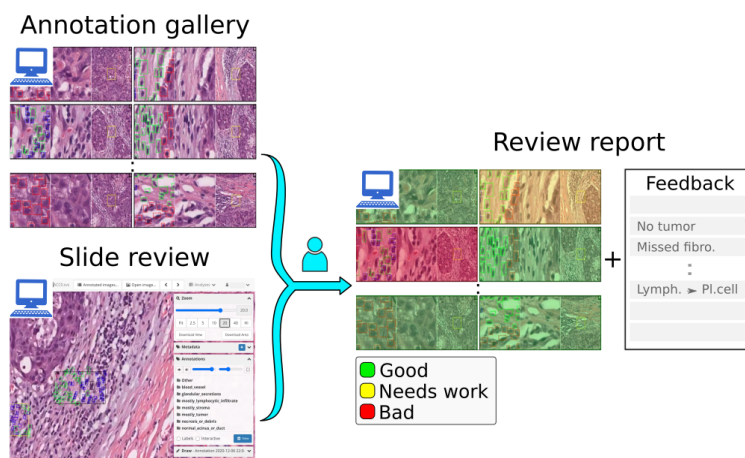


Figure S1. Use of review galleries for scalable review of single-rater annotations. Single-rater annotations were corrected by two study coordinators, in consultation with a senior pathologist. The pathologist was provided with a mosaic review gallery showing a bird's eye view of each FOV, with and without annotations, and at high and low power. The pathologist was asked to assign a per-FOV quality assessment. If the pathologist wanted further context, they were able to click on the FOV and pan around the full whole-slide image. They were also able to provide brief comments to be addressed by the coordinators, for eg. "change all to tumor". A demo is provided at the following video: https://youtu.be/Plh39obBq_0.

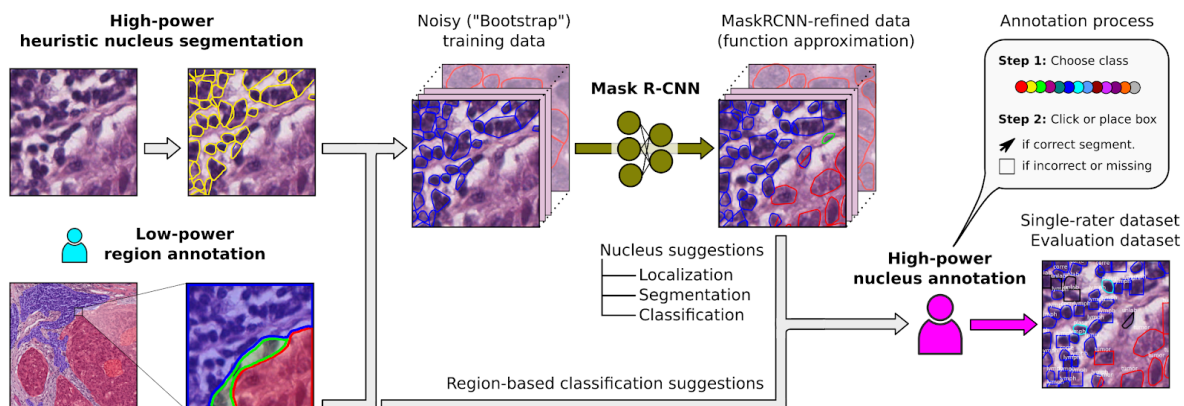


Figure S2. Process for obtaining algorithmic suggestions for scalable assisted annotation. Nucleus segmentation boundaries were derived using image processing heuristics at a high magnification. Low-power region annotations from the BCSS dataset, approved by a practicing pathologist, were then used to assign an initial class to nuclei. This combination of noisy nuclear segmentation boundaries and region-derived classifications are the *bootstrapped* suggestions. These noisy algorithmic suggestions were the basis for annotating the Bootstrap control multi-rater dataset. A Mask R-CNN model was then used as a function approximator to smooth out some of the noise in the bootstrapped suggestions. Participants were able to view these refined suggestions, along with low-power region annotations, when annotating the single-rater and Evaluation datasets.

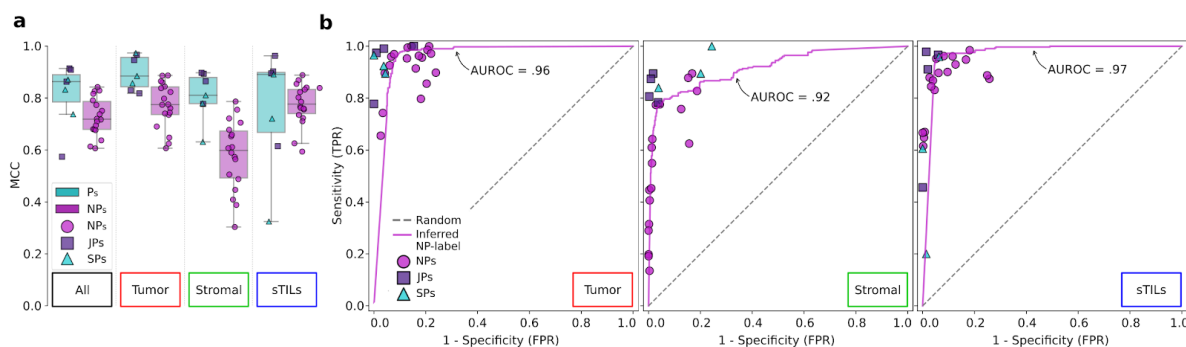


Figure S3. Super-class accuracy of participant annotations and inferred NP-labels (Evaluation dataset). The accuracy is measured against the inferred P-truth.

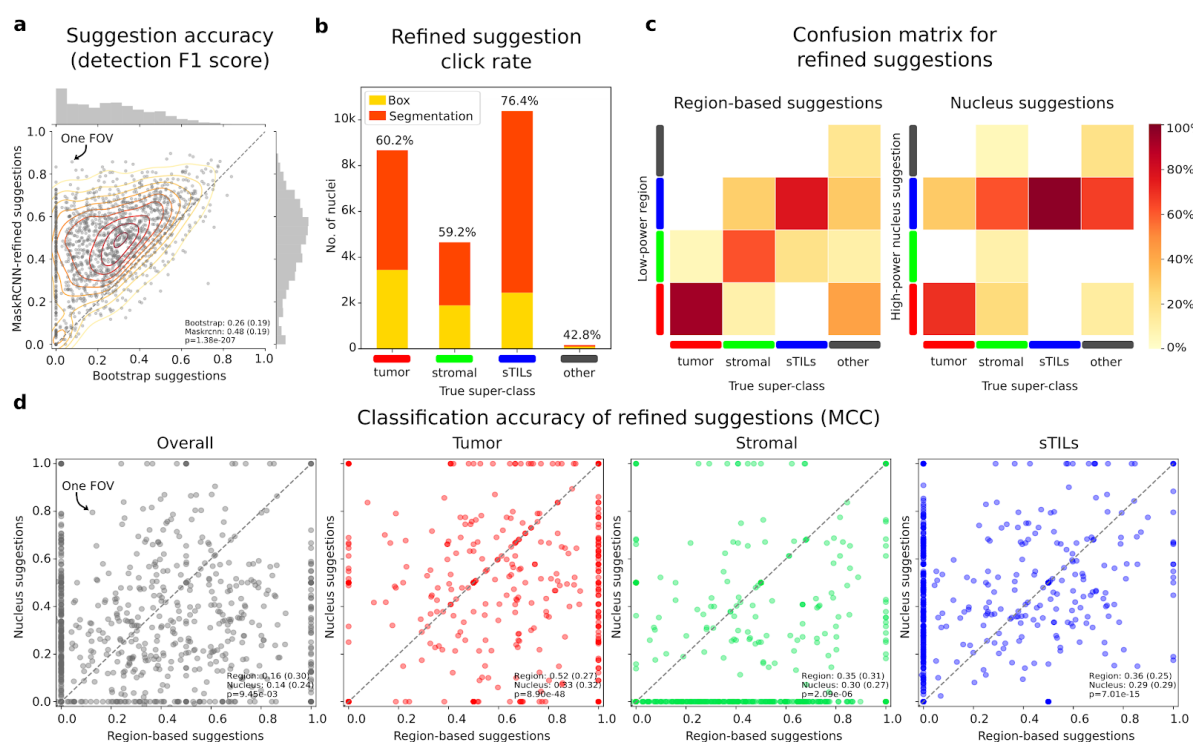


Figure S4. Accuracy of algorithmic suggestions (single-rater dataset). The accuracy is measured against the corrected single-rater dataset. **a.** Per-FOV detection accuracy of algorithmic data at the two stages of obtaining algorithmic suggestions; i.e. how well do the suggestions correspond to real nuclei? Mask R-CNN refinement significantly improves suggestion accuracy. **b.** Number of Mask R-CNN-refined suggestions that correspond to a segmentation (i.e. were clicked) or a bounding box. **c.** Concordance between suggested classes and classes assigned by participants. Region-based suggestions were, broadly-speaking, more concordant with the true classes, but nucleus suggestions had a higher recall for sTILs. **d.** Comparison of the classification accuracy (MCC) of low-power region class and high-power Mask R-CNN-derived nucleus class. Numbers are normalized column-wise, i.e. represent percentages of true nuclei of a particular class. Note how region-based and nucleus-based suggestions have disparate accuracies for different FOVs and classes. Hence, there was value in providing the participants with both forms of suggestion.

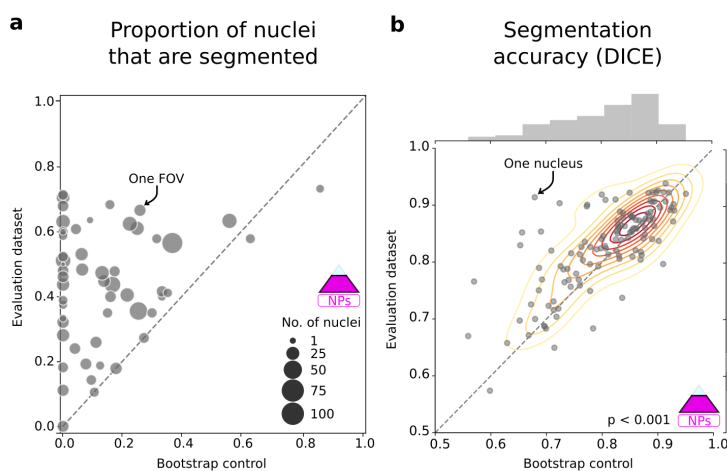


Figure S5. Abundance and segmentation accuracy of clicked algorithmic suggestions (multi-rater datasets). **a.** Proportion of nuclei in the FOV that were inferred to have good segmentation. Circle size represents the number of nuclei in that FOV. The proportion is notably higher for the Evaluation dataset than the Bootstrap control. **b.** Accuracy of algorithmic segmentation boundaries for nuclei that were inferred to have accurate segmentation boundaries in both the Evaluation dataset and Bootstrap control. The comparison is made against manual segmentations obtained for the same nuclei from one senior pathologist. Most clicked algorithmic segmentations were very accurate, and have a DICE coefficient above 0.8. The accuracy was slightly higher for Mask R-CNN-refined suggestions.

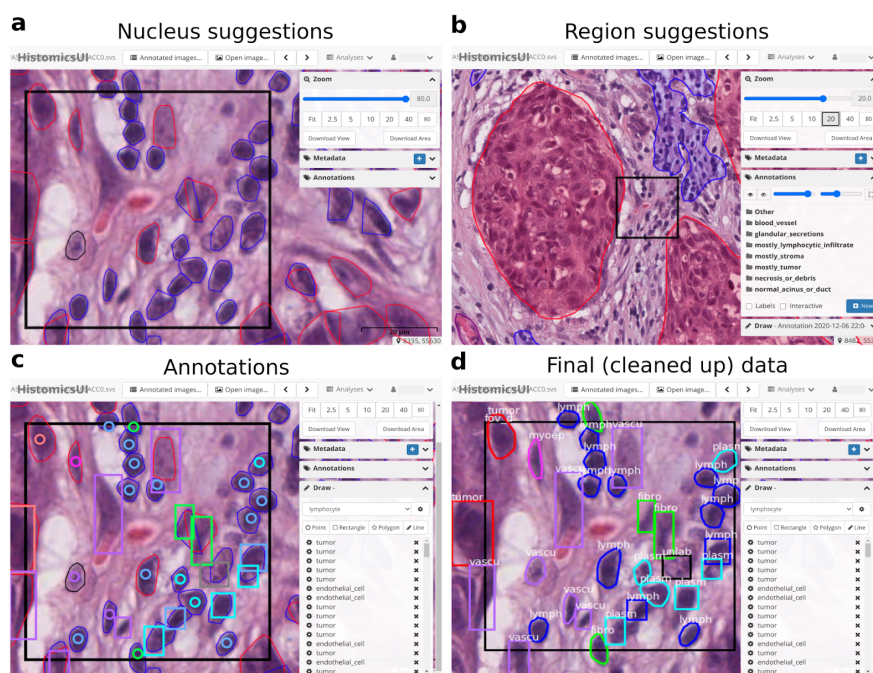


Figure S6. Annotation procedure on HistomicsUI. **a-b.** Participants were shown suggestions for nucleus segmentation boundaries, as well as two types of classification suggestions: low-power region suggestions and high-power nucleus classification suggestions. The FOV shown here is almost entirely present in a stromal region, but contains multiple scattered sTILs that were not dense enough to be captured as a sTILs "region". **c.** Participants' annotations were either points/clicks, for accurate segmentations, or bounding boxes. They picked the color/class of their annotations beforehand, and were told to simply ignore any inaccurate suggestions. Participants were able to turn the suggestions off for a clear view of the underlying tissue. **d.** Participant annotations and algorithmic suggestions were ingested into a database and processed to provide cleaned up data, which was then pushed for viewing on HistomicsUI for correction and review.

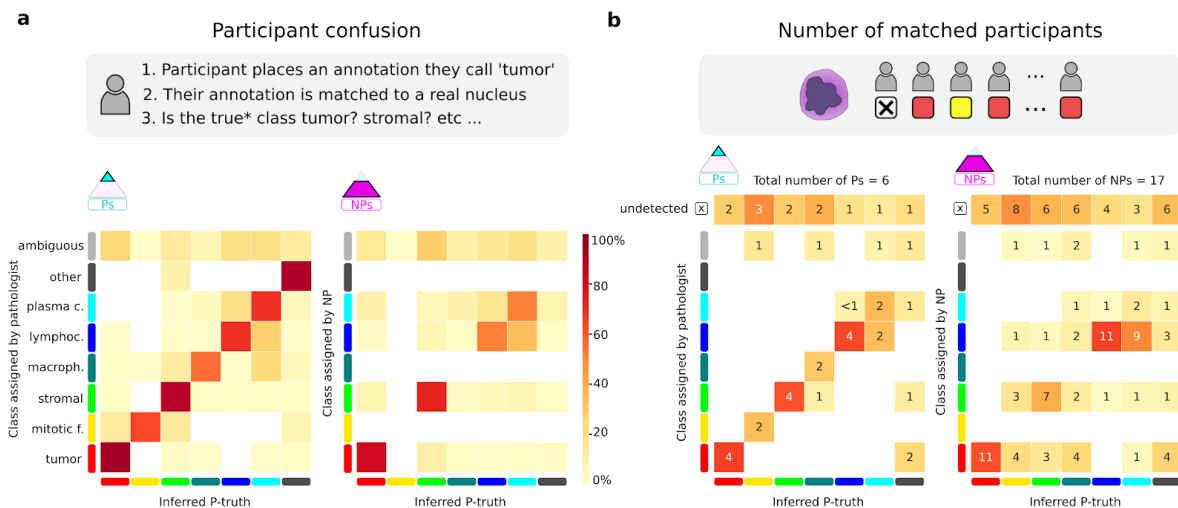


Figure S7. Confusion matrix of participant annotations (Evaluation dataset). **a.** Confusion of annotations placed by the participants, putting aside detection. Here, we ask the question, if a participant places an annotation they call tumor, and it matches a true nucleus, what is the class of that nucleus? By definition, there are no “ambiguous” true nuclei. **b.** For each true nucleus, how many of the participants detected it, and if so, what class did they assign? Note that since truth inference takes participant reliability into account, the inferred P-truth does not have to correspond to the most commonly assigned class. Empty entries are values <1.

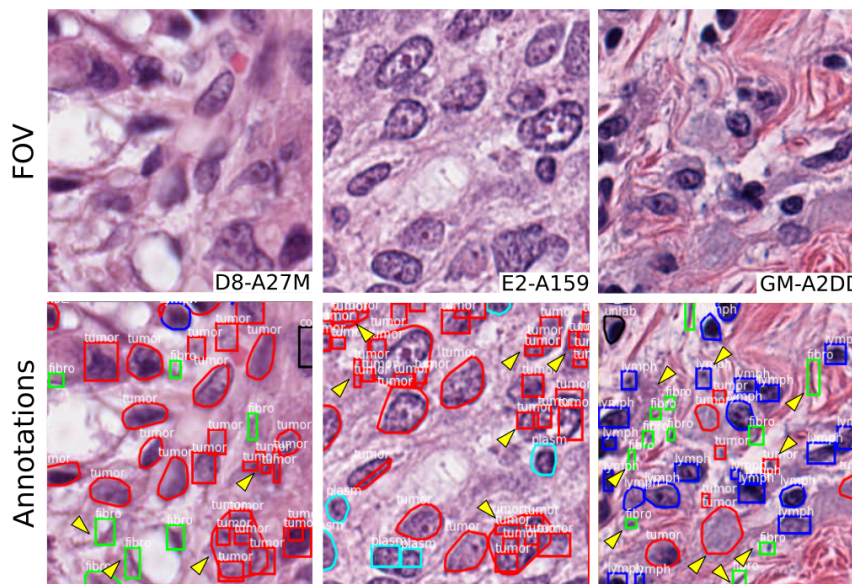


Figure S8. Sample poor annotation data excluded during the single-rater dataset correction process. Despite having received initial training and feedback, the NP who generated these annotations was confused about what is a nucleus, and frequently considered chromatin clumps or artifacts as nuclei (arrows). This underlines the need for quality control.

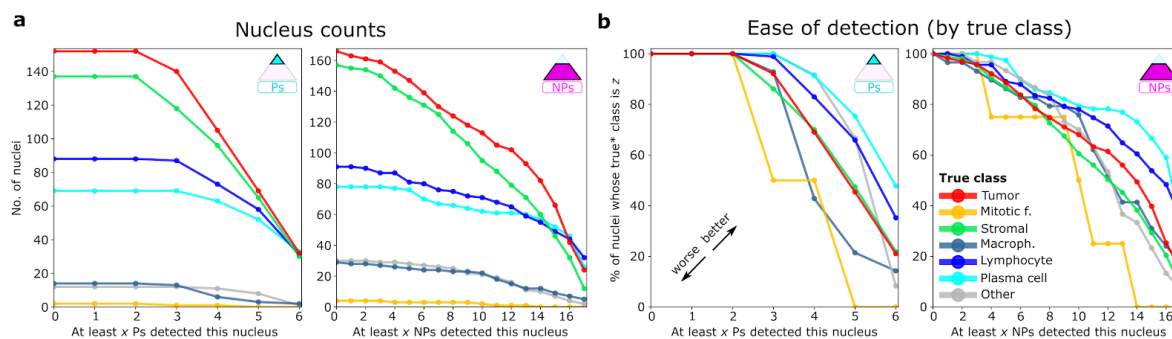


Figure S9. Ease of detection of various nucleus classes (Evaluation dataset). If we know for a fact this is, say, a lymphocyte, how many participants detected it, even if they called it something else?. True class is the inferred P-truth. The color coding used is explained in panel b. **a.** Nuclei counts, broken down by class and the number of matched participants. **b.** Ease of detection of nuclei by true class. Interpreting, say, the blue curve goes like this: 100% of lymphocytes were detected by at least 3 pathologists, ~80% were detected by 4 pathologists, and so on.

- 1 Do unconstrained agglomerative clustering with maximum linkage using bounding boxes from participants $\{P_1, P_2, \dots\}$
 - 2 Cut at linkage threshold $1 - t^*$ (where t^* is the threshold IOU)
 - 3 For each cluster C_i (corresponding to top-level node N_i)
 - 3.1 For each "don't-link" set S_j
 - 3.1.1 Check if more than one member in S_j is present in C_i
 - 3.1.2 For each extra member S_{jk} in C_i
 - 3.1.2.1 Check the next low-level node N_{i-1} :
 - > If there are no members from S_j in N_{i-1} :
 - If C_{i-1} does not exist:
 - Set N_{i-1} as a new cluster C_{i-1}
 - Assign S_{jk} to C_{i-1}
 - > Else:
 - Check the next low-level node N_{i-2} (repeat 3.1.2.1)
 - 3.1.2.2 If no nodes without members from S_j found:
 - > Assign S_{jk} as a separate one-leaf cluster
- 4 For each cluster C_i
 - 4.1 Find IOU of bounding boxes of members $\{C_{i1}, C_{i2}, \dots, C_{iN}\}$
 - 4.2 Assign member $C_{im} = \text{argmax}(\text{mean IOU})$ as the medoid

Figure S10. Algorithm for obtaining anchor proposals through constrained agglomerative clustering. We cluster bounding boxes from participants to get the *anchor proposals* corresponding to potential nucleus locations. Note that the threshold we use for maximum linkage, t^* , is influential in determining how many anchors we get. We make sure that annotations from the same participant do not end up in the same cluster by creating sets of "don't-link" bounding boxes. The final anchor proposals are the anchor medoids; using medoids ensures that the box anchor proposals correspond to real nucleus boundaries.

Supplementary file: Annotation protocol

Welcome to the breast cancer nucleus annotation project! The purpose of this project is to investigate a scalable data collection and refinement procedure, and to create a large-scale dataset for training and validation of machine learning algorithms.

> **Please view the introductory video before diving into this document.**

> **Please read this document in its entirety before making annotations.**

There are three categories of participants:

- **NP (Non-pathologist)** - Did not receive anatomical pathology residency training.
- **JP (Junior pathologist)** - Pathology residents with < 2 years of training.
- **SP (Senior pathologist)** - Attendings or pathology residents with > 2 years of training.

There are two required annotation assignments for each NP:

- **Single-rater dataset:** You can ask questions and receive feedback from pathologists.
- **Multi-rater datasets:** No feedback will be provided. Annotate to the best of your ability.

> **General remarks:**

- Use a **comfortable mouse, table and monitor**. This greatly impacts comfort and quality.
- When in doubt, take a screenshot and post a question on **Slack** for review & feedback.
- Remember, the algorithm is **learning** what we teach it (Garbage In → Garbage Out).

> **Annotation workflow:**

- **Step 1:** View the **region-level annotations**. These are the low-power classification suggestions.
- **Step 2:** Go to **medium power** (20x) and **reduce transparency**. Examine the underlying tissue.
- **Step 3:** Zoom on the FOV at **maximum power** (40x or 80x, depending on slide).
- **Step 4: Start annotating.** The process is illustrated in the introductory video. Briefly, the steps are:
 - > Pick an annotation class/color (feel free to rely on or ignore algorithmic suggestions)
 - > If an algorithmic boundary is correct, place a dot.
 - > Otherwise, place a bounding box around the nucleus.

> Specific annotation rules:

- Only annotate the Fields-of-View (FOVs) that were picked for you.
- If a nucleus extends beyond the FOV boundary, make sure your bounding box covers its full extent (i.e. extend your rectangle outside the FOV as well).
- Make sure each FOV is complete before moving to the next. Missing annotations may confuse our algorithms and make validation difficult!
- Make sure to annotate **in this order: Single-rater dataset → Multi-rater dataset 1 → Multi-rater dataset 2 and/or Multi-rater dataset 3**. SPs and JPs do not have a single-rater dataset (but they *do* have multi-rater datasets). Pathologists are kindly asked to respond to questions on Slack.

***Explanatory note:** We asked the participants to annotate the single-rater dataset first because this also acted as their de-facto training, and they received feedback and could ask questions. The multi-rater datasets were blinded to avoid biasing the participants. Multi-rater dataset 1 is the unbiased control dataset (no algorithmic suggestions), and was annotated first for the same reason.*

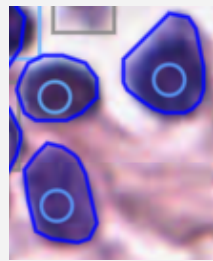

- After you annotate your **first FOV**, take a screenshot and share it on the **Slack group** to get **approval & feedback before continuing**. This acts as a test of your understanding.
- After every slide in the single-rater dataset, please ask for feedback from the SPs and/or study coordinator. Do not post a screenshot of every single FOV, simply post the slide ID on the group and SPs/coordinator will go to the slide and make suggestions/corrections where necessary.
- Share a screenshot of anything that you are unsure of, making sure to also share the slide name so that the SPs and study coordinators can take a closer look at various magnifications. Nuclei are often vague. If you are unsure about the class of a nucleus, either:
 - Ask what it is on the group and receive feedback from SP (preferred).
 - Assign is the class *unlabeled*.Make as much effort to classify nuclei as possible; only use the *unlabeled* class in a minority of cases.
- Make sure the **bounding box is tight** around the nucleus **NOT** the entire cell.
- Do not trust the computer suggestions too much. If the algorithmic boundaries are just slightly off then it's OK, otherwise use a bounding box instead.
- Never rotate the slide before annotating. All boxes should have the same orientation as the FOV.



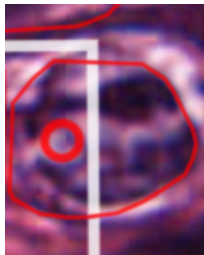

> Notes about specific annotation classes:

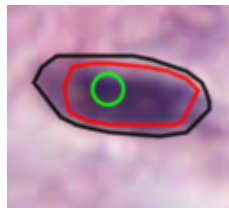
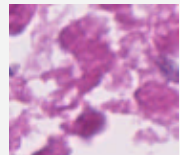
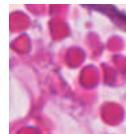


Notes that address some frequently asked questions on the Slack group.


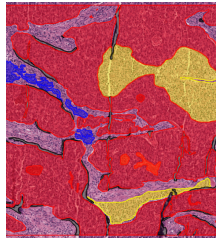
- **Tumor:** Malignant cells are very heterogeneous in shape. They tend to have hyperchromatic, eccentric nuclei, and tend to be crowded and irregular. See any standard pathology textbook.
- **Fibroblasts:** Stromal nuclei tend to be elongated and shaped like a cigar. May also have a rounder shape. The tell-tale sign is their presence in stroma in alignment with the collagen fibres. Some fibroblasts close to the tumor may be **activated** (i.e. have tumor-like morphology).
- **Lymphocytes:** Small, round, condensed, central nucleus. Tend to be grouped together.
- **Plasma cells:** May confuse with lymphocytes. Plasma cells are less common than lymphocytes; when in doubt, ask on Slack. They tend to have an eccentric, large, textured nucleus (described as *cart-wheel*, but rarely seen as such) with a pale perinuclear halo. Also tend to have eosinophilic cytoplasm
- **Macrophages:** Usually difficult to ascertain. They tend to be larger than lymphocytes, sometimes have vacuolated or frothy cytoplasm, have thin round-to-uniform (bean shaped) nuclei with variable nucleoli.

> Troubleshooting:

#	Situation	How to handle	Examples
1	Annotations take a long time to load.	<ul style="list-style-type: none"> - Close all programs running in the background. - Close all other Google Chrome tabs, especially videos (except this document, which you always have to refer to) - Switch from tablet to computer - Switch to a faster internet connection 	N/A
2	Algorithm correctly predicts both nucleus boundary and class	Place a dot inside the nucleus with the correct class	
3a	Algorithm correctly predicts nucleus boundary but assigns incorrect class, and you know the correct class	- If you know correct class: Place a dot with correct class inside the nucleus	

3b	Same as 3a, but you do not know the correct class	<ul style="list-style-type: none"> - If you are new to this or are in doubt, take a snapshot and ask for pathologist feedback. - If you are confident in your ability (eg you have been annotating many FOVs or are a pathologist), i.e. the nucleus is vague and cannot be classified using just H&E: place a dot with the class <i>unlabeled</i>. 	
4	Algorithm incorrectly predicts nucleus boundary or completely misses the nucleus	Place a rectangle with the correct class and color around the nucleus. The rectangle must be tight (i.e. it should be precise, not too large or too small).	
5	The algorithm clumps multiple nuclei together	Place a rectangle around each nucleus and ignore the algorithmic suggestion.	
6a	A nucleus extends beyond the edge of the FOV and I need to place a <u>dot</u>	Place the dot inside the FOV.	
6b	A nucleus extends beyond the edge of the FOV and I need to place a <u>rectangle</u>	Extend your rectangle to encompass the full extend of the nucleus	
7	I cannot see the underlying tissue	Reduce the annotation transparency	
8	I know the nucleus class but it is not in the	<ul style="list-style-type: none"> - If you are an NP: ask on Slack; a pathologist may recommend a class. - If you are a pathologist, create your own 	

	standard classes	class, and notify the study coordinator.	
9	The algorithm predict two overlapping boundaries for the same nucleus; only one is correct	If it is possible to place the dot inside the inside the correct boundary, but outside the incorrect one, do so.	
10	There is necrotic debris or collagen	Ignore it. Do not annotate debris or non-nuclear material.	
11	There are red blood cells	Ignore. Do not annotate RBCs	
12	There is a multinucleated giant cell or a cell-eat-cell phenomenon (cannibalism)	Classify each nucleus independently. We operate at the level of nuclei, not cells, in this project.	
13	There are overlapping nuclei and the bounding boxes will have to overlap to capture full extent.	No problem; use overlapping bounding boxes in this case.	
14	The nuclei are very textured and have prominent nucleoli	Don't be fooled!! Malignant nuclei can have a very textured appearance and prominent nucleoli so you may think they are multiple nuclei but are one nucleus!! By the way, in the image below, there are many vacuoles that were mistaken as being nuclei. This is a vacuolated phenotype.	

		<p>INCORRECT: over-segmented nuclei CORRECT: one bounding box per nucleus</p> 	
<p>15</p>	<p>The slide is quite difficult; stroma is difficult to distinguish from tumor.</p>	<p>Make sure you follow step 1 in the <i>Annotation workflow</i> section. Anything outside tumor regions may still be a tumor nucleus, but is more likely to be a fibroblast, lymphocyte, plasma cell etc.</p>	
<p>16</p>	<p>After finishing many FOVs, I discovered (or was told) that I have a systematic error in classifying nuclei (eg. all plasma cells mistakenly called tumor)</p>	<p>Notify one of the study coordinators and we will run a program (python script) to do this automatically for you.</p>	

Supplement for Section 3.2: Amgad et al., 2021a

Supplementary Methods

1 Internal-external cross-validation

Training and testing data were separated at the level of hospitals/institutions (Fig S1). To balance the size of various folds, we made sure each fold contained at least one "large" institution. Large institutions were defined as those having a minimum of 9 unique patients.

2 NuCLS model

Our NuCLS model modifies the Pytorch implementation of the Mask R-CNN architecture (He *et al.*, 2017).

2.1 Hyperparameters

We used a ResNet18 backbone that was pretrained on ImageNet. Single-GPU training was done using a batch size of 4, using a stochastic gradient descent optimiser with a learning rate of $2e-3$ and a momentum of $9e-1$. The learning rate and momentum were identified using grid search on the validation dataset during prototyping. All ground truth nuclei were kept per image at training, while detections were limited to a maximum of 300 nuclei at inference. 3,000 anchors were kept from the region proposal network after non-maximum suppression (NMS), using an NMS threshold of 0.7. The length of anchor sides used in pixels (relative to upsampled images, see below) is 12, 24 and 48.

2.2 Resize using scale factor

Mask R-CNN resizes input images to have a constant short side. While this may work for datasets where the variability in image size is modest, or where the camera distance is variable, it is not suitable in computational pathology applications where large tile sizes are favorable for efficient and scalable inference. Resizing to a constant short side would shrink nuclei during inference. To remedy this NuCLS resizes using a scale factor, instead, thus preserving the nuclear size and aspect ratio at inference for any tile size. We used a scale factor of 4.0, meaning that images were digitally zoomed to a 0.05 micron-per-pixel resolution before being analyzed. This corresponded to a sTILs diameter of 4.4 "pixels" in the feature map generated by the ResNet18 backbone. As a form of scale augmentation, we jittered this scale factor by up to 10% during training.

2.3 Training with hybrid datasets

Our annotation protocol generates a mixture of manually placed bounding boxes and approved suggestions of segmented nuclei. We train from this data by ignoring bounding boxes when calculating the mask loss.

2.4 Specialized classification convolutions

Four extra convolutional filters were applied to the feature map output from the ResNet18 backbone (He *et al.*, 2016). The filters had a kernel size of 3, a stride of 1, and a dilation and padding of 1 to preserve feature map size (Fig 4a). The resultant feature map was only used for classification and only contributed to the classification loss. The same procedure used for box regression was used for classification: 1. ROIAlign to obtain per-object convolutional feature maps; 2. flattening of the feature map; 3. passage through a single fully-connected layer.

2.5 Class-agnostic detection & segmentation

Both the box regression output and nucleus masks were simplified and made classification-agnostic. We relied on the fact that nucleus shapes and sizes are fairly homogeneous to simplify the learning problem and preserve classification probability vectors at inference. Specifically, we relied on a global NMS process (Fig 4b). We summed the classification probabilities for all classes (i.e. everything except background), and concatenated all these "objectness" scores for each FOV. An NMS process was then carried out as usual. That is, boxes were sorted by objectness score, and if a box overlapped with a higher-scoring box by more than a particular IOU threshold (0.2 in our case), it was removed.

2.6 Data augmentation

Previous research has shown that the combined use of color normalization and augmentation improves performance of deep learning models in histopathology applications (Tellez *et al.*, 2019). All FOVs were color normalized using the Macenko method before training began (Macenko *et al.*, 2009). During training, FOVs also underwent a stain augmentation routine (Tellez *et al.*, 2018). This augmentation routine randomly perturbed the hematoxylin and eosin channels each time the image was loaded, using a sigma of 0.5 for the random uniform distribution. The HistomicsTK package was used for both the color normalization and augmentation operations ([digitalSlideArchive.github.io](https://github.com/digitalSlideArchive/digitalSlideArchive)). Additionally, each training image was cropped at a random location after loading to memory (300×300 pixel region) to increase robustness.

2.7 Handling class imbalance

Nucleus class imbalance was mitigated by weighted random sampling with replacement. With the exception of ambiguous nuclei, which received zero weight, class weights were inversely proportional to the frequency of occurrence in the training set. Since we load data on a per-FOV basis, each FOV f was assigned a sampling weight W_f that favors FOVs with a high density of uncommon nuclear classes, as follows:

$$W_f = U_f \div \sum_{i=1}^F U_i \quad (1)$$

$$U_f = \sum_{c=1}^C (W_c N_{cf}) \div A_f \quad (2)$$

Where, C is the number of classes, F is the number of FOVs in the training set, N_{cf} is the number of nuclei of class c in FOV f , and A_f is the area of FOV f . W_c is the weight assigned to class c and is determined as follows:

$$W_c = V_c \div \sum_{i=1}^C V_i \quad (3)$$

$$V_c = 1 \div \sum_{f=1}^F N_{cf} \quad (4)$$

2.8 Matching detections

Algorithmic detections were matched to ground truth using linear sum assignment from the Scipy library (Kuhn, 1955).

Supplementary Tables

Table S1. NuCLS model tuning for the nucleus detection task on the validation set (fold 1). All accuracy values are percentages. After passage through the model backbone, the feature map is markedly smaller than original images due to the max pooling operations. This means that without digital zooming, the diameter of a 'typical' small nucleus, say TILs, is very small in the feature map. As a consequence, when the object-specific part of the feature map is pooled using ROIAlign, there is very little information to use for box regression or classification. Abbreviations: MPP, microns-per-pixel; AP@0.5, average precision when a threshold of 0.5 is used for validating a detection.

Scale factor	Equivalent MPP	Backbone	TILs diameter (image, pixels)	TILs diameter (featmap, 'pixels')	AP @ 0.5
1	0.2	Resnet18	30	1.1	61.7
1	0.2	Resnet34	30	1.1	63
1	0.2	Resnet50	30	1.1	62
2.67	0.075	Resnet18	80	3	76.4
2.67	0.075	Resnet34	80	3	74.3
2.67	0.075	Resnet50	80	3	Mem.Err.
4	0.05	Resnet18	120	4.4	75
4	0.05	Resnet34	120	4.4	72.9
4	0.05	Resnet50	120	4.4	Mem.Err.

Table S2. NuCLS model tuning for the nucleus classification task on the validation set (fold 1). All accuracy values are percentages. Empty entries correspond to metrics which were not applicable for the configuration (config) being studied. Classification AUROC statistics were not possible for configs where each nucleus had a single classification as opposed to a classification probability vector, as in the baseline Mask R-CNN model. The baseline model achieves a lower performance. We show that this is due in large part to the coupling of detection and classification, which may not be ideal for datasets with many small and clustered objects. After decoupling, the performance dramatically improves. Configs where the model was trained on super-classes do not have accuracy statistics for the main classes. On the other hand, when models were trained on the main classes, super-class predictions were easily obtained by aggregating the predicted class probabilities.

Config	Detection AP @ .5	Overall classification accuracy						Classification accuracy breakdown (AUROC)								
		MCC		Micro		Macro		Tumor			Stromal			sTILs		
		Supercl.?		Supercl.?		Supercl.?		Subclasses		Superclass	Subclasses		Superclass	Subclasses		Superclass
		No	Yes	No	Yes	No	Yes	Non-mitotic	Mitotic		Stromal	Macrophage		Lymphocyte	Plasma cell	
1	70	1.8	-3	-	-	-	-	-	-	-	-	-	-	-	-	-
2	74.5	57	65	93.4	94.3	85.2	88.2	93.1	91.5	93.2	88.8	71	83.6	95	78.6	95
3	75.4	59.6	66	93.5	93.7	84.7	85.2	94.2	90.6	94.5	89.1	73.5	82	95.2	84.2	95.7
4	72.2	52.6	60.9	91	92.3	82.4	83.6	92.5	90.8	92.1	86.7	61.7	78.9	94.7	82.9	93.4
4+	72.2	54.5	62.5	90.3	91.9	84.1	85.8	92.2	88.5	92	88.1	68.4	81.5	93.7	84.4	93.4
5	72.6	-	-5	-	-	-	-	-	-	-	-	-	-	-	-	-
6	74.8	-	63.6	-	93.5	-	85.9	-	-	92.8	-	-	81.3	-	-	95
7	72.2	-	63.1	-	93.1	-	82.8	-	-	91.9	-	-	81	-	-	94.9
7+	72.2	-	64.8	-	92.7	-	83.7	-	-	93.1	-	-	83.1	-	-	94.8

Config 1: Baseline Mask R-CNN implementation. We discounted bounding boxes from the mask loss to enable training on our hybrid data.

Config 2: Config 1, but with class-agnostic detection and non-maximum suppression.

Config 3: Config 2, but with 4 extra convolutions that specialize in classification.

Config 4: Config 1 for nucleus detection, then an independent nucleus classification model using thumbnails of detected nuclei.

Config 4+: Same model from config 4, but with test-time augmentation (random shift) at the classification stage.

Config 5: Config 1 but trained using supercategories.

Config 6: Config 2 but trained using supercategories.

Config 7: Config 4 but trained using supercategories.

Config 7+: Same model from config 7, but with test-time augmentation (random shift) at the classification stage.

Table S3. Generalization accuracy of the NuCLS models trained on the corrected single-rater dataset, and evaluated on the multi-rater dataset using internal-external cross-validation. All accuracy values are percentages. Fold 1 acted as the validation set for hyperparameter tuning, so the bottom row shows mean and standard deviation of three values (folds 3-5). Note that the number of testing set nuclei varied by fold because the split happens at the level of hospitals and not nuclei. There were no testing set slides with available multi-rater truth to assess the performance on fold 2. Notice that the classification accuracy is consistently higher when the assessment was done at the level of super-classes. Abbreviations: AP@.5, average precision when a threshold of 0.5 is used for considering a detection to be true; mAP@.5:.95, mean average precision at detection thresholds between 0.5 and 0.95.

Fold	Detection			Segmentation			Classification					
	N	AP @.5	mAP @.5:.95	N	Median IOU	Median DICE	N	Super-classes?	Accuracy	MCC	AUROC (micro)	AUROC (macro)
1 (Val.)	209	62.9	21.0	42	67.6	80.7	173	No	70.5	63.6	94.2	85.6
								Yes	86.1	79.0	95.7	95.6
3	66	65.2	29.0	7	76.9	86.9	52	No	63.5	42.4	80.7	85.5
								Yes	61.5	42.5	75.1	84.7
4	317	71.5	32.6	82	76.2	86.5	278	No	68.0	54.3	94.3	89.3
								Yes	84.9	75.5	96.9	92.0
5	213	58.3	22.9	49	71.8	83.6	174	No	67.8	55.8	92.2	90.4
								Yes	75.3	65.6	91.4	95.2
Mean (Std)	-	65.0 (5.4)	28.2 (4.0)	-	74.9 (2.3)	85.7 (1.5)	-	No	66.4 (2.1)	50.8 (6.0)	89.1 (6.0)	88.4 (2.1)
								Yes	73.9 (9.6)	61.2 (13.8)	87.8 (9.2)	90.6 (4.4)

Table S4. Generalization accuracy of the trained NuCLS models - broken down by superclass. All accuracy values are percentages. Note that the corrected single-rater dataset is likely more reflective of the generalization accuracy, since it contains 1,744 unique FOVs. The multi-rater dataset only has 52 unique FOVs, hence the large variation in performance.

Fold	N	MCC				AUROC				
		Overall	Tumor	Stromal	sTILs	Micro-avg.	Macro-avg.	Tumor	Stromal	sTILs
Training: Single-rater dataset; Testing: Single-rater dataset										
1 (Val.)	5351	65.2	72.9	47.1	73.7	93.7	89.0	94.2	83.2	95.3
2	13597	68.2	73.7	53.0	76.6	94.6	86.5	94.5	87.4	96.2
3	11176	68.1	74.9	46.9	77.9	94.4	89.4	96.1	84.3	95.7
4	7288	73.5	80.6	56.9	79.6	96.1	87.4	97.2	89.1	95.9
5	6294	52.4	57.4	40.7	60.1	89.0	80.8	88.8	80.7	91.0
Mean (Std)	-	65.6 (7.9)	71.7 (8.6)	49.4 (6.1)	73.5 (7.8)	93.5 (2.7)	86.0 (3.2)	94.2 (3.2)	85.4 (3.2)	94.7 (2.1)
Training: Single-rater dataset; Testing: Multi-rater dataset										
1 (Val.)	173	79.0	88.0	73.0	78.6	95.7	95.6	97.7	94.4	95.5
3	52	42.5	38.5	26.3	73.9	75.1	84.7	87.1	83.0	90.9
4	278	75.5	77.8	53.1	90.2	96.9	92.0	96.4	91.9	99.2
5	174	65.6	60.0	67.1	72.1	91.4	95.2	96.6	92.2	97.9
Mean (Std)	-	61.2 (13.8)	58.8 (16.1)	48.8 (16.9)	78.8 (8.2)	87.8 (9.2)	90.6 (4.4)	93.4 (4.4)	89.0 (4.3)	96.0 (3.6)

Table S5. List of interpretable features used as input for DTALE, which were extracted using the HistomicsTK package.

Category	N	Description	Feature	Category	N	Description	Feature		
Size	4	Pixels occupied by the nucleus	Area	Edges	8	Gradients and canny edge filters (hematoxylin channel)	Mag.Mean		
		Length of major/minor axes of the ellipse with the same 2nd central moments	MajorAxis				Mag.Std		
		Pixelated perimeter using 4-connectivity	MinorAxis				Mag.Skew		
Shape	6	Similarity to the shape of a circle	Circularity				2	Angular 2nd moment (ASM): A measure of homogeneity	Mag.Kurt.
		Eccentricity of fitted ellipse (a measure of aspect ratio)	Eccentricity						His.Entropy
		Diameter of a circle with the same area	Equiv.Diam.				His.Energy		
	Ratio of nucleus area to its bounding box	Extent	2				Contrast: Intensity variation for neighbouring pixels	Canny.Sum	
	Aspect ratio of a fitted ellipse	Min.Maj.Axis						Canny.Mean	
	A measure of convexity	Solidity	6	Fourier simplifications of object shape.	2	Correlation: Intensity correlation for neighboring pixels	Mean		
Intensity	12	Nucleus hematoxylin intensity features.					FSD1	2	Sum of squares: A measure of variance
					FSD2	Mean			
					FSD3	2	Inverse difference moment: A measure of homogeneity	Range	
					FSD4			Mean	
FSD5					4	Sum average & Sum variance for all features	Mean		
FSD6			Range						
Haralick texture features			12	Nucleus hematoxylin intensity features.	Min	2	Sum entropy features	Mean	
					Max			Range	
					Mean	2	Entropy	Mean	
					Median			Range	
					MeanMed.Diff	4	Difference variance & Difference entropy	Mean	
					Std			Range	
	IQR	4			Information Measure of Correlation (IMC) (2 types)	Mean			
	MAD					Range			
Skewness	4	Information Measure of Correlation (IMC) (2 types)	Mean						
Kurtosis			Range						
HistEnergy	4	Information Measure of Correlation (IMC) (2 types)	Mean						
HistEntropy			Range						

Supplementary Figures

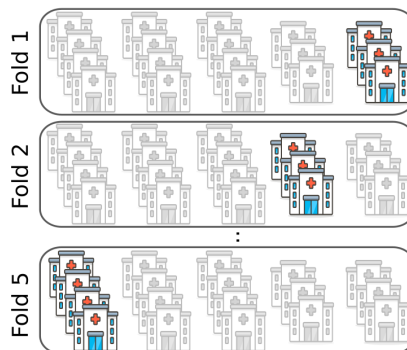


Fig. S1. Internal-external cross-validation procedure. The TCGA dataset originates from multiple institutions, and we used this fact to obtain an estimate of the external analytic validity of our models. Fold 1 was used for tuning hyper parameters, while folds 4-5 were used as external testing sets.

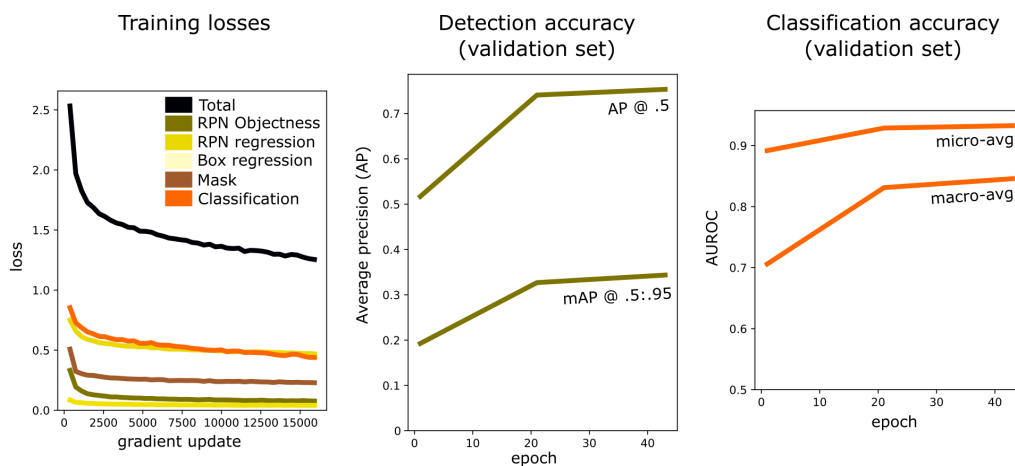


Fig. S2. Progression of NuCLS model training and convergence on fold 1. Our prototyping experiments on fold 1 (not shown) showed that the detection model started overfitting after 15k detection updates, so we froze detection weights after 15k iterations and allowed 1k extra iterations for fine-tuning of the classification layers. Abbreviations: RPN, region proposal network; AP@.5, average precision when a threshold of 0.5 is used for considering a detection to be true, mAP@.5:.95, mean average precision at a range of detection thresholds between 0.5 and 0.95; AUROC, area under receiver-operator characteristics curve.

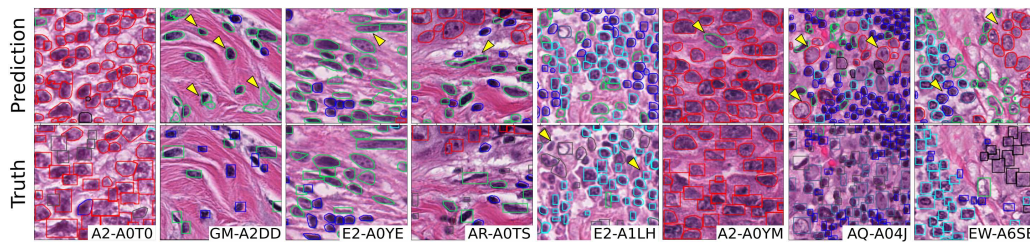


Fig. S3. Additional examples showing qualitative performance of NuCLS model on testing sets. The displayed ground truth comes from the pathologist-corrected single-rater dataset. The images are representative of a number of different hospitals in each of the testing sets from the cross-validation scheme. Detection and classification performance closely matches the ground truth, and discrepancies are marked by arrows. Not all discrepancies are algorithmic errors, including: *i.* adjacent nuclei that could conceivably be viewed as a single nucleus; *ii.* missing annotations; *iii.* morphologically ambiguous nuclei. Some errors arise from the lack of incorporation of contextual information in our models. Without low power context, macrophages and normal ductal/acinar cells may look morphologically similar to tumor cells.

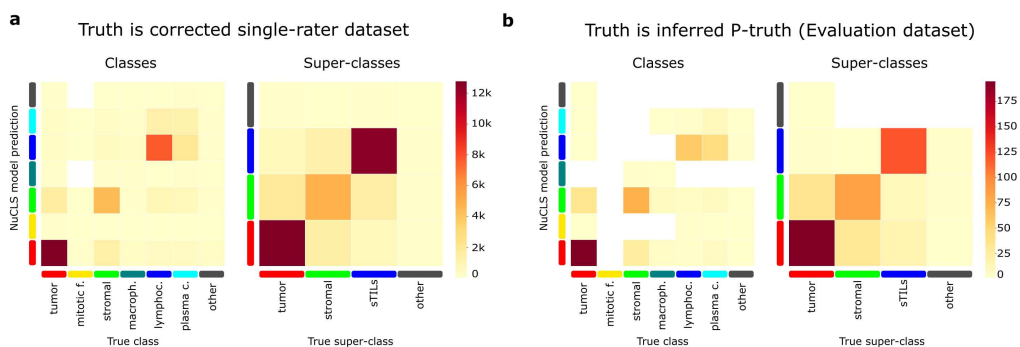


Fig. S4. Confusion matrix of NuCLS model predictions on the testing sets. For each of folds 2-5, the NuCLS model trained on the single-rater dataset training slides was used to predict FOVs from the corresponding testing set slides. The counts shown are aggregated over all testing sets. a. The single-rater dataset is considered to be the truth. b. Inferred truth from pathologists (inferred P-truth) on the multi-rater Evaluation dataset is considered to be the truth.

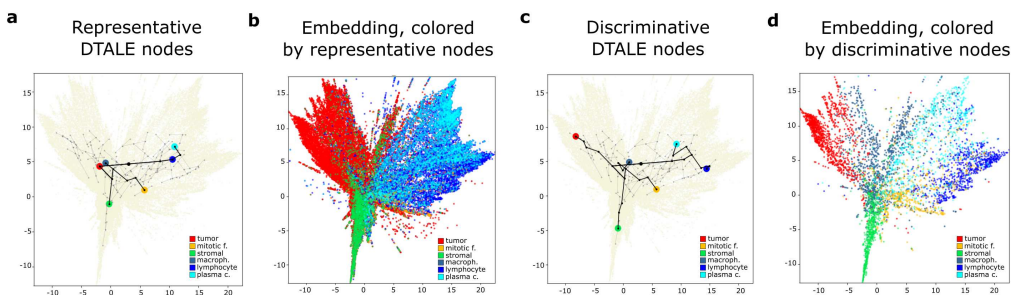


Fig. S5. Representative vs discriminative approximation of NuCLS model decisions using DTALE. a. Overlay of the full DTALE tree (light gray) on top of the embedding to which it was fitted. In black, we show paths to the nodes that allow representative approximation of NuCLS decisions, i.e. highest F-1 score. b. Nuclei that correspond to representative DTALE nodes. c. DTALE nodes that correspond to the most discriminative approximation of the NuCLS decisions, i.e. highest precision. d. Nuclei that correspond to discriminative DTALE nodes.

Bibliography

- [1] AbdulJabbar, K., Raza, S. E. A., Rosenthal, R., Jamal-Hanjani, M., Veeriah, S., Akarca, A., Lund, T., Moore, D. A., Salgado, R., Al Bakir, M., Zapata, L., Hiley, C. T., Officer, L., Sereno, M., Smith, C. R., Loi, S., Hackshaw, A., Marafioti, T., Quezada, S. A., McGranahan, N., Le Quesne, J., TRACERx Consortium, Swanton, C., and Yuan, Y. (2020). Geospatial immune variability illuminates differential evolution of lung adenocarcinoma. *Nat. Med.*, 26(7):1054–1062.
- [2] Abels, E., Pantanowitz, L., Aeffner, F., Zarella, M. D., van der Laak, J., Bui, M. M., Vemuri, V. N., Parwani, A. V., Gibbs, J., Agosto-Arroyo, E., Beck, A. H., and Kozlowski, C. (2019). Computational pathology definitions, best practices, and recommendations for regulatory guidance: a white paper from the digital pathology association. *J. Pathol.*, 249(3):286–294.
- [3] Abubakar, M., Zhang, J., Ahearn, T. U., Koka, H., Guo, C., Lawrence, S. M., Mutreja, K., Figueroa, J. D., Ying, J., Lissowska, J., et al. (2021). Tumor-associated stromal cellular density as a predictor of recurrence and mortality in breast cancer: Results from ethnically-diverse study populations. *Cancer Epidemiology and Prevention Biomarkers*.
- [4] Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., and Süsstrunk, S. (2012). Slic superpixels compared to state-of-the-art superpixel methods. *IEEE transactions on pattern analysis and machine intelligence*, 34(11):2274–2282.
- [5] Agarwalla, A., Shaban, M., and Rajpoot, N. M. (2017). Representation-aggregation networks for segmentation of multi-gigapixel histology images. *arXiv preprint arXiv:1707.08814*.
- [6] Amgad, M., Atteya, L., Hussein, H., Mohammed, K. H., Hafiz, E., Elsebaie, M. A., Mobadersany, P., Manthey, D., Gutman, D. A., Elfandy, H., et al. (2021a). Explainable nucleus classification using decision tree approximation of learned embeddings. *Bioinformatics*.

- [7] Amgad, M., Atteya, L. A., Hussein, H., Mohammed, K. H., Hafiz, E., Elsebaie, M. A., Alhusseiny, A. M., AlMoslemany, M. A., Elmatboly, A. M., Pappalardo, P. A., et al. (2021b). Nucls: A scalable crowdsourcing, deep learning approach and dataset for nucleus classification, localization and segmentation. *arXiv preprint arXiv:2102.09099*.
- [8] Amgad, M., Elfandy, H., Hussein, H., Atteya, L. A., Elsebaie, M. A., Abo Elnasr, L. S., Sakr, R. A., Salem, H. S., Ismail, A. F., Saad, A. M., et al. (2019a). Structured crowdsourcing enables convolutional segmentation of histology images. *Bioinformatics*, 35(18):3461–3467.
- [9] Amgad, M., Itoh, A., and Tsui, M. M. K. (2015). Extending ripley’s K-Function to quantify aggregation in 2-D grayscale images. *PLoS One*, 10(12):e0144404.
- [10] Amgad, M., Sarkar, A., Srinivas, C., Redman, R., Ratra, S., Bechert, C. J., Calhoun, B. C., Mrazek, K., Kurkure, U., Cooper, L. A., et al. (2019b). Joint region and nucleus segmentation for characterization of tumor infiltrating lymphocytes in breast cancer. In *Medical Imaging 2019: Digital Pathology*, volume 10956, page 109560M. International Society for Optics and Photonics.
- [11] Amgad, M., Stovgaard, E. S., Balslev, E., Thagaard, J., Chen, W., Dudgeon, S., Sharma, A., Kerner, J. K., Denkert, C., Yuan, Y., et al. (2020). Report on computational assessment of tumor infiltrating lymphocytes from the international immuno-oncology biomarker working group. *NPJ breast cancer*, 6(1):1–13.
- [12] Amin, M. B., Edge, S. B., Greene, F. L., Byrd, D. R., Brookland, R. K., Washington, M. K., Gershenwald, J. E., Compton, C. C., Hess, K. R., Sullivan, D. C., Milburn Jessup, J., Brierley, J. D., Gaspar, L. E., Schilsky, R. L., Balch, C. M., Winchester, D. P., Asare, E. A., Madera, M., Gress, D. M., and Meyer, L. R. (2018). *AJCC Cancer Staging Manual*. Springer International Publishing.
- [13] Aran, D., Sirota, M., and Butte, A. J. (2015). Systematic pan-cancer analysis of tumour purity. *Nat. Commun.*, 6:8971.
- [14] Aresta, G., Araújo, T., Kwok, S., Chennamsetty, S. S., Safwan, M., Alex, V., Marami, B., Prastawa, M., Chan, M., Donovan, M., Fernandez, G., Zeineh, J., Kohl, M., Walz, C., Ludwig,

- F., Braunewell, S., Baust, M., Vu, Q. D., To, M. N. N., Kim, E., Kwak, J. T., Galal, S., Sanchez-Freire, V., Brancati, N., Frucci, M., Riccio, D., Wang, Y., Sun, L., Ma, K., Fang, J., Kone, I., Boulmane, L., Campilho, A., Eloy, C., Polónia, A., and Aguiar, P. (2019). BACH: Grand challenge on breast cancer histology images. *Med. Image Anal.*, 56:122–139.
- [15] Ashley, E. A. (2016). Towards precision medicine. *Nat. Rev. Genet.*, 17(9):507–522.
- [16] Balkenhol, M. C. A., Tellez, D., Vreuls, W., Clahsen, P. C., Pinckaers, H., Ciompi, F., Bult, P., and van der Laak, J. A. W. M. (2019). Deep learning assisted mitotic counting for breast cancer. *Lab. Invest.*, 99(11):1596–1606.
- [17] Ballman, K. V. (2015). Biomarker: Predictive or prognostic? *J. Clin. Oncol.*, 33(33):3968–3971.
- [18] Barbie, D. A., Tamayo, P., Boehm, J. S., Kim, S. Y., Moody, S. E., Dunn, I. F., Schinzel, A. C., Sandy, P., Meylan, E., Scholl, C., Fröhling, S., Chan, E. M., Sos, M. L., Michel, K., Mermel, C., Silver, S. J., Weir, B. A., Reiling, J. H., Sheng, Q., Gupta, P. B., Wadlow, R. C., Le, H., Hoersch, S., Wittner, B. S., Ramaswamy, S., Livingston, D. M., Sabatini, D. M., Meyerson, M., Thomas, R. K., Lander, E. S., Mesirov, J. P., Root, D. E., Gilliland, D. G., Jacks, T., and Hahn, W. C. (2009). Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature*, 462(7269):108–112.
- [19] Beck, A. H., Sangoi, A. R., Leung, S., Marinelli, R. J., Nielsen, T. O., van de Vijver, M. J., West, R. B., van de Rijn, M., and Koller, D. (2011). Systematic analysis of breast cancer morphology uncovers stromal features associated with survival. *Sci. Transl. Med.*, 3(108):108ra113.
- [20] Bentley, J. L. (1975). Multidimensional binary search trees used for associative searching. *Communications of the ACM*, 18(9):509–517.
- [21] Berger, A. C., Korkut, A., Kanchi, R. S., Hegde, A. M., Lenoir, W., Liu, W., Liu, Y., Fan, H., Shen, H., Ravikumar, V., Rao, A., Schultz, A., Li, X., Sumazin, P., Williams, C., Mestdagh, P., Gunaratne, P. H., Yau, C., Bowlby, R., Robertson, A. G., Tiezzi, D. G., Wang, C., Cherniack, A. D., Godwin, A. K., Kuderer, N. M., Rader, J. S., Zuna, R. E., Sood, A. K., Lazar, A. J., Ojesina, A. I., Adebamowo, C., Adebamowo, S. N., Baggerly, K. A., Chen, T.-W., Chiu, H.-S.,

- Lefever, S., Liu, L., MacKenzie, K., Orsulic, S., Roszik, J., Shelley, C. S., Song, Q., Vellano, C. P., Wentzensen, N., Cancer Genome Atlas Research Network, Weinstein, J. N., Mills, G. B., Levine, D. A., and Akbani, R. (2018). A comprehensive Pan-Cancer molecular study of gynecologic and breast cancers. *Cancer Cell*, 33(4):690–705.e9.
- [22] Blazeby, J. M., Avery, K., Sprangers, M., Pikhart, H., Fayers, P., and Donovan, J. (2006). Health-related quality of life measurement in randomized clinical trials in surgical oncology. *J. Clin. Oncol.*, 24(19):3178–3186.
- [23] Boellner, S. and Becker, K.-F. (2015). Reverse phase protein Arrays-Quantitative assessment of multiple biomarkers in biopsies for clinical use. *Microarrays (Basel)*, 4(2):98–114.
- [24] Bradburn, M. J., Clark, T. G., Love, S. B., and Altman, D. G. (2003). Survival analysis part II: multivariate data analysis—an introduction to concepts and methods. *Br. J. Cancer*, 89(3):431–436.
- [25] Burns, P. B., Rohrich, R. J., and Chung, K. C. (2011). The levels of evidence and their role in evidence-based medicine. *Plastic and reconstructive surgery*, 128(1):305.
- [26] Bychkov, D., Linder, N., Turkki, R., Nordling, S., Kovanen, P. E., Verrill, C., Walliander, M., Lundin, M., Haglund, C., and Lundin, J. (2018). Deep learning based tissue analysis predicts outcome in colorectal cancer. *Scientific reports*, 8(1):1–11.
- [27] Calle, E. E., Rodriguez, C., Jacobs, E. J., Almon, M. L., Chao, A., McCullough, M. L., Feigelson, H. S., and Thun, M. J. (2002). The american cancer society cancer prevention study ii nutrition cohort: rationale, study design, and baseline characteristics. *Cancer: Interdisciplinary International Journal of the American Cancer Society*, 94(9):2490–2501.
- [28] Campanella, G., Hanna, M. G., Geneslaw, L., Mirafior, A., Werneck Krauss Silva, V., Busam, K. J., Brogi, E., Reuter, V. E., Klimstra, D. S., and Fuchs, T. J. (2019). Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nat. Med.*, 25(8):1301–1309.
- [29] Cancer Genome Atlas Network (2012). Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70.

- [30] Cancer Genome Atlas Network (2015). Genomic classification of cutaneous melanoma. *Cell*, 161(7):1681–1696.
- [31] Cancer Genome Atlas Research Network (2015). The molecular taxonomy of primary prostate cancer. *Cell*, 163(4):1011–1025.
- [32] Cancer Genome Atlas Research Network (2017). Comprehensive and integrated genomic characterization of adult soft tissue sarcomas. *Cell*, 171(4):950–965.e28.
- [33] Cancer Genome Atlas Research Network, Brat, D. J., Verhaak, R. G. W., and et al (2015). Comprehensive, integrative genomic analysis of diffuse Lower-Grade gliomas. *N. Engl. J. Med.*, 372(26):2481–2498.
- [34] Canziani, A., Paszke, A., and Culurciello, E. (2016). An analysis of deep neural network models for practical applications. *arXiv*.
- [35] Cardenas, M. A., Prokhnevskaya, N., and Kissick, H. T. (2021). Organized immune cell interactions within tumors sustain a productive t-cell response. *International Immunology*, 33(1):27–37.
- [36] Chandradevan, R., Aljudi, A. A., Drumheller, B. R., Kunananthaseelan, N., Amgad, M., Gutman, D. A., Cooper, L. A. D., and Jaye, D. L. (2019). Machine-based detection and classification for bone marrow aspirate differential counts: initial development focusing on nonneoplastic cells. *Lab. Invest.*
- [37] Chen, R. J., Lu, M. Y., Wang, J., Williamson, D. F. K., Rodig, S. J., Lindeman, N. I., and Mahmood, F. (2020a). Pathomic fusion: An integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis. *IEEE Trans. Med. Imaging*, PP.
- [38] Chen, Y., Janowczyk, A., and Madabhushi, A. (2020b). Quantitative assessment of the effects of compression on deep learning in digital pathology image analysis. *JCO Clin Cancer Inform*, 4:221–233.
- [39] Clark, T. G., Bradburn, M. J., Love, S. B., and Altman, D. G. (2003). Survival analysis part i: basic concepts and first analyses. *Br. J. Cancer*, 89(2):232–238.

- [40] Cooper, L. A., Demicco, E. G., Saltz, J. H., Powell, R. T., Rao, A., and Lazar, A. J. (2018). PanCancer insights from the cancer genome atlas: the pathologist’s perspective. *J. Pathol.*, 244(5):512–524.
- [41] Cooper, L. A. D., Kong, J., Gutman, D. A., Dunn, W. D., Nalisnik, M., and Brat, D. J. (2015). Novel genotype-phenotype associations in human cancers enabled by advanced molecular platforms and computational analysis of whole slide images. *Lab. Invest.*, 95(4):366–376.
- [42] Cooper, L. A. D., Kong, J., Gutman, D. A., Wang, F., Gao, J., Appin, C., Cholleti, S., Pan, T., Sharma, A., Scarpace, L., Mikkelsen, T., Kurc, T., Moreno, C. S., Brat, D. J., and Saltz, J. H. (2012). Integrated morphologic analysis for the identification and characterization of disease subtypes. *J. Am. Med. Inform. Assoc.*, 19(2):317–323.
- [43] Cox, D. R. (1972). Regression models and life-tables. *J. R. Stat. Soc.*, 34(2):187–202.
- [44] Diao, J. A., Wang, J. K., Chui, W. F., Mountain, V., Gullapally, S. C., Srinivasan, R., Mitchell, R. N., Glass, B., Hoffman, S., Rao, S. K., Maheshwari, C., Lahiri, A., Prakash, A., McLoughlin, R., Kerner, J. K., Resnick, M. B., Montalto, M. C., Khosla, A., Wapinski, I. N., Beck, A. H., Elliott, H. L., and Taylor-Weiner, A. (2021). Human-interpretable image features derived from densely mapped cancer pathology slides predict diverse molecular phenotypes. *Nat. Commun.*, 12(1):1613.
- [45] Dudgeon, S. N., Wen, S., Hanna, M. G., Gupta, R., Amgad, M., Sheth, M., Marble, H., Huang, R., Herrmann, M. D., Szu, C. H., et al. (2021). A pathologist-annotated dataset for validating artificial intelligence: a project description and pilot study. *J Pathol Inform*, 12:45.
- [46] Ehteshami Bejnordi, B., Mullooly, M., Pfeiffer, R. M., Fan, S., Vacek, P. M., Weaver, D. L., Herschorn, S., Brinton, L. A., van Ginneken, B., Karssemeijer, N., Beck, A. H., Gierach, G. L., van der Laak, J. A. W. M., and Sherman, M. E. (2018). Using deep convolutional neural networks to identify and classify tumor-associated stroma in diagnostic breast biopsies. *Mod. Pathol.*, 31(10):1502–1512.
- [47] Ehteshami Bejnordi, B., Veta, M., Johannes van Diest, P., van Ginneken, B., Karssemeijer, N., Litjens, G., van der Laak, J. A. W. M., the CAMELYON16 Consortium, Hermsen, M.,

- Manson, Q. F., Balkenhol, M., Geessink, O., Stathonikos, N., van Dijk, M. C., Bult, P., Beca, F., Beck, A. H., Wang, D., Khosla, A., Gargeya, R., Irshad, H., Zhong, A., Dou, Q., Li, Q., Chen, H., Lin, H.-J., Heng, P.-A., Haß, C., Bruni, E., Wong, Q., Halici, U., Öner, M. Ü., Cetin-Atalay, R., Berseth, M., Khvatkov, V., Vylegzhanin, A., Kraus, O., Shaban, M., Rajpoot, N., Awan, R., Sirinukunwattana, K., Qaiser, T., Tsang, Y.-W., Tellez, D., Annuschein, J., Hufnagl, P., Valkonen, M., Kartasalo, K., Latonen, L., Ruusu vuori, P., Liimatainen, K., Albarqouni, S., Mungal, B., George, A., Demirci, S., Navab, N., Watanabe, S., Seno, S., Takenaka, Y., Matsuda, H., Ahmady Phoulady, H., Kovalev, V., Kalinovsky, A., Liauchuk, V., Bueno, G., Fernandez-Carrobles, M. M., Serrano, I., Deniz, O., Racoceanu, D., and Venâncio, R. (2017). Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA*, 318(22):2199–2210.
- [48] Evans, A. J., Bauer, T. W., Bui, M. M., Cornish, T. C., Duncan, H., Glassy, E. F., Hipp, J., McGee, R. S., Murphy, D., Myers, C., O’Neill, D. G., Parwani, A. V., Rampy, B. A., Salama, M. E., and Pantanowitz, L. (2018). US food and drug administration approval of whole slide imaging for primary diagnosis: A key milestone is reached and new questions are raised. *Arch. Pathol. Lab. Med.*, 142(11):1383–1387.
- [49] Fallahpour, S., Navaneelan, T., De, P., and Borgo, A. (2017). Breast cancer survival by molecular subtype: a population-based analysis of cancer registry data. *CMAJ Open*, 5(3):E734–E739.
- [50] Fend, F. and Raffeld, M. (2000). Laser capture microdissection in pathology. *J. Clin. Pathol.*, 53(9):666–672.
- [51] Friedl, P. and Gilmour, D. (2009). Collective cell migration in morphogenesis, regeneration and cancer. *Nature reviews Molecular cell biology*, 10(7):445–457.
- [52] Ghassemi, M., Oakden-Rayner, L., and Beam, A. L. (2021). The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health*, 3(11):e745–e750.
- [53] Github (2019). histolab: Library for digital pathology image processing.

- [54] Gonzalez, R. and Woods, R. (1992). Digital image processing, (march 1992).
- [55] Goode, A., Gilbert, B., Harkes, J., Jukic, D., and Satyanarayanan, M. (2013). OpenSlide: A vendor-neutral software foundation for digital pathology. *J. Pathol. Inform.*, 4(1):27.
- [56] Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press.
- [57] Goodwin, S., McPherson, J. D., and McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.*, 17(6):333–351.
- [58] Gutman, D. A., Khalilia, M., Lee, S., Nalisnik, M., Mullen, Z., Beezley, J., Chittajallu, D. R., Manthey, D., and Cooper, L. A. D. (2017). The digital slide archive: A software platform for management, integration, and analysis of histology for cancer research. *Cancer Res.*, 77(21):e75–e78.
- [59] Ha, S. Y., Yeo, S.-Y., Xuan, Y.-h., and Kim, S.-H. (2014). The prognostic significance of cancer-associated fibroblasts in esophageal squamous cell carcinoma. *PloS one*, 9(6):e99955.
- [60] Haas, M., Seshan, S. V., Barisoni, L., Amann, K., Bajema, I. M., Becker, J. U., Joh, K., Ljubanovic, D., Roberts, I. S. D., Roelofs, J. J., Sethi, S., Zeng, C., and Jennette, J. C. (2020). Consensus definitions for glomerular lesions by light and electron microscopy: recommendations from a working group of the renal pathology society. *Kidney Int.*, 98(5):1120–1134.
- [61] Hamilton, N. (2009). Quantification and its applications in fluorescent microscopy imaging. *Traffic*, 10(8):951–961.
- [62] Hanahan, D. and Weinberg, R. A. (2000). The hallmarks of cancer. *Cell*, 100(1):57–70.
- [63] Hanahan, D. and Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. *Cell*, 144(5):646–674.
- [64] Haralick, R. M., Shanmugam, K., and Dinstein, I. (1973). Textural features for image classification. *IEEE Trans. Syst. Man Cybern.*, SMC-3(6):610–621.
- [65] Hartman, D. J., Van Der Laak, J. A. W. M., Gurcan, M. N., and Pantanowitz, L. (2020). Value of public challenges for the development of pathology deep learning algorithms. *J. Pathol. Inform.*, 11:7.

- [66] Hastie, T., Tibshirani, R., and Friedman, J. (2017). The elements of statistical learning: Data mining, inference, and prediction (springer series in statistics).
- [67] He, B., Bergenstr hle, L., Stenbeck, L., Abid, A., Andersson, A., Borg,  ., Maaskola, J., Lundberg, J., and Zou, J. (2020). Integrating spatial gene expression and breast tumour morphology via deep learning. *Nat Biomed Eng*, 4(8):827–834.
- [68] He, K., Gkioxari, G., Doll r, P., and Girshick, R. (2017a). Mask R-CNN. *arXiv*.
- [69] He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778. openaccess.thecvf.com.
- [70] He, L., Ren, X., Gao, Q., Zhao, X., Yao, B., and Chao, Y. (2017b). The connected-component labeling problem: A review of state-of-the-art algorithms. *Pattern Recognit.*, 70:25–43.
- [71] Heller, M. J. (2002). DNA microarray technology: devices, systems, and applications. *Annu. Rev. Biomed. Eng.*, 4:129–153.
- [72] Herrmann, M. D., Clunie, D. A., Fedorov, A., Doyle, S. W., Pieper, S., Klepeis, V., Le, L. P., Mutter, G. L., Milstone, D. S., Schultz, T. J., Kikinis, R., Kotecha, G. K., Hwang, D. H., Andriole, K. P., Iafrate, A. J., Brink, J. A., Boland, G. W., Dreyer, K. J., Michalski, M., Golden, J. A., Louis, D. N., and Lennerz, J. K. (2018). Implementing the DICOM standard for digital pathology. *J. Pathol. Inform.*, 9:37.
- [73] Hinton, G. E. and Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507.
- [74] Hoadley, K. A., Yau, C., Hinoue, T., Wolf, D. M., Lazar, A. J., Drill, E., Shen, R., Taylor, A. M., Cherniack, A. D., Thorsson, V., Akbani, R., Bowlby, R., Wong, C. K., Wiznerowicz, M., Sanchez-Vega, F., Robertson, A. G., Schneider, B. G., Lawrence, M. S., Noushmehr, H., Malta, T. M., Cancer Genome Atlas Network, Stuart, J. M., Benz, C. C., and Laird, P. W. (2018). Cell-of-Origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. *Cell*, 173(2):291–304.e6.

- [75] Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.
- [76] Hou, L., Agarwal, A., Samaras, D., Kurc, T. M., Gupta, R. R., and Saltz, J. H. (2019). Robust histopathology image analysis: To label or to synthesize?
- [77] Hou, L., Samaras, D., Kurc, T. M., Gao, Y., Davis, J. E., and Saltz, J. H. (2016). Patch-based convolutional neural network for whole slide tissue image classification. *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2016:2424–2433.
- [78] Ioffe, S. and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In Bach, F. and Blei, D., editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 448–456, Lille, France. PMLR.
- [79] Irshad, H., Veillard, A., Roux, L., and Racoceanu, D. (2014). Methods for nuclei detection, segmentation, and classification in digital histopathology: A Review—Current status and future potential. *IEEE Rev. Biomed. Eng.*, 7:97–114.
- [80] Ishwaran, H., Kogalur, U. B., Blackstone, E. H., and Lauer, M. S. (2008). Random survival forests. *arXiv*.
- [81] Islami, F., Ward, E. M., Sung, H., Cronin, K. A., Tangka, F. K. L., Sherman, R. L., Zhao, J., Anderson, R. N., Henley, S. J., Yabroff, K. R., Jemal, A., and Benard, V. B. (2021). Annual report to the nation on the status of cancer, part 1: National cancer statistics. *J. Natl. Cancer Inst.*
- [82] Janowczyk, A. and Madabhushi, A. (2016). Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. *J. Pathol. Inform.*, 7:29.
- [83] Janowczyk, A., Zuo, R., Gilmore, H., Feldman, M., and Madabhushi, A. (2019). HistoQC: An Open-Source quality control tool for digital pathology slides. *JCO Clin Cancer Inform*, 3:1–7.
- [84] Jansen, C. S., Prokhnevskaya, N., and Kissick, H. T. (2019). The requirement for immune infiltration and organization in the tumor microenvironment for successful immunotherapy in

- prostate cancer. In *Urologic Oncology: Seminars and Original Investigations*, volume 37, pages 543–555. Elsevier.
- [85] Jaume, G., Pati, P., Bozorgtabar, B., Foncubierta, A., Anniciello, A. M., Feroce, F., Rau, T., Thiran, J.-P., Gabrani, M., and Goksel, O. (2021). Quantifying explainers of graph neural networks in computational pathology. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8106–8116.
- [86] Kalra, S., Tizhoosh, H. R., Choi, C., Shah, S., Diamandis, P., Campbell, C. J. V., and Pantanowitz, L. (2020a). Yottixel - an image search engine for large archives of histopathology whole slide images. *Med. Image Anal.*, 65(101757):101757.
- [87] Kalra, S., Tizhoosh, H. R., Shah, S., Choi, C., Damaskinos, S., Safarpour, A., Shafiei, S., Babaie, M., Diamandis, P., Campbell, C. J. V., and Pantanowitz, L. (2020b). Pan-cancer diagnostic consensus through searching archival histopathology images using artificial intelligence. *NPJ Digit Med*, 3:31.
- [88] Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *J. Am. Stat. Assoc.*, 53(282):457–481.
- [89] Khoreva, A., Benenson, R., Hosang, J., Hein, M., and Schiele, B. (2017). Simple does it: Weakly supervised instance and semantic segmentation.
- [90] Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv*.
- [91] Kirillov, A., He, K., Girshick, R., Rother, C., and Dollár, P. (2019). Panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9404–9413.
- [92] Koh, P. W., Nguyen, T., Tang, Y. S., Mussmann, S., Pierson, E., Kim, B., and Liang, P. (2020). Concept bottleneck models. *arXiv*.
- [93] Kohler, B. A., Sherman, R. L., Howlader, N., Jemal, A., Ryerson, A. B., Henry, K. A., Boscoe, F. P., Cronin, K. A., Lake, A., Noone, A.-M., Henley, S. J., Ehemann, C. R., Anderson, R. N., and Penberthy, L. (2015). Annual report to the nation on the status of cancer, 1975-2011, featuring

incidence of breast cancer subtypes by Race/Ethnicity, poverty, and state. *J. Natl. Cancer Inst.*, 107(6):djv048.

- [94] Kos, Z., Roblin, E., Kim, R. S., Michiels, S., Gallas, B. D., Chen, W., van de Vijver, K. K., Goel, S., Adams, S., Demaria, S., Viale, G., Nielsen, T. O., Badve, S. S., Symmans, W. F., Sotiriou, C., Rimm, D. L., Hewitt, S., Denkert, C., Loibl, S., Luen, S. J., Bartlett, J. M. S., Savas, P., Pruneri, G., Dillon, D. A., Cheang, M. C. U., Tutt, A., Hall, J. A., Kok, M., Horlings, H. M., Madabhushi, A., van der Laak, J., Ciompi, F., Laenkholm, A.-V., Bellolio, E., Grusso, T., Fox, S. B., Araya, J. C., Floris, G., Hudeček, J., Voorwerk, L., Beck, A. H., Kerner, J., Larsimont, D., Declercq, S., Van den Eynden, G., Pusztai, L., Ehinger, A., Yang, W., AbdulJabbar, K., Yuan, Y., Singh, R., Hiley, C., Bakir, M. A., Lazar, A. J., Naber, S., Wienert, S., Castillo, M., Curigliano, G., Dieci, M.-V., André, F., Swanton, C., Reis-Filho, J., Sparano, J., Balslev, E., Chen, I.-C., Stovgaard, E. I. S., Pogue-Geile, K., Blenman, K. R. M., Penault-Llorca, F., Schnitt, S., Lakhani, S. R., Vincent-Salomon, A., Rojo, F., Braybrooke, J. P., Hanna, M. G., Soler-Monsó, M. T., Bethmann, D., Castaneda, C. A., Willard-Gallo, K., Sharma, A., Lien, H.-C., Fineberg, S., Thagaard, J., Comerma, L., Gonzalez-Ericsson, P., Brogi, E., Loi, S., Saltz, J., Klausen, F., Cooper, L., Amgad, M., Moore, D. A., Salgado, R., and International Immuno-Oncology Biomarker Working Group (2020). Pitfalls in assessing stromal tumor infiltrating lymphocytes (sTILs) in breast cancer. *NPJ Breast Cancer*, 6:17.
- [95] Kraus, O. Z., Ba, J. L., and Frey, B. J. (2016). Classifying and segmenting microscopy images with deep multiple instance learning. *Bioinformatics*, 32(12):i52–i59.
- [96] Krizhevsky, A. (2021). Learning multiple layers of features from tiny images. <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.222.9220&rep=rep1&type=pdf>. Accessed: 2021-10-30.
- [97] Kukačka, J., Golkov, V., and Cremers, D. (2017). Regularization for deep learning: A taxonomy. *arXiv*.
- [98] Kundu, S. (2021). AI in medicine must be explainable. *Nat. Med.*, 27(8):1328.
- [99] Lagache, T., Lang, G., Sauvonnnet, N., and Olivo-Marin, J.-C. (2013). Analysis of the spatial organization of molecules with robust statistics. *PLoS One*, 8(12):e80914.

- [100] Lee, S., Amgad, M., Mobadersany, P., McCormick, M., Pollack, B. P., Elfandy, H., Hussein, H., Gutman, D. A., and Cooper, L. A. D. (2021). Interactive classification of Whole-Slide imaging data for cancer researchers. *Cancer Res.*, 81(4):1171–1177.
- [101] Lester, S. C., Bose, S., Chen, Y.-Y., Connolly, J. L., de Baca, M. E., Fitzgibbons, P. L., Hayes, D. F., Kleer, C., O’Malley, F. P., Page, D. L., et al. (2009). Protocol for the examination of specimens from patients with invasive carcinoma of the breast. *Archives of pathology & laboratory medicine*, 133(10):1515–1538.
- [102] Li, H., Bera, K., Toro, P., Fu, P., Zhang, Z., Lu, C., Feldman, M., Ganesan, S., Goldstein, L. J., Davidson, N. E., Glasgow, A., Harbhajanka, A., Gilmore, H., and Madabhushi, A. (2021). Collagen fiber orientation disorder from H&E images is prognostic for early stage breast cancer: clinical trial validation. *NPJ Breast Cancer*, 7(1):104.
- [103] Lin, W.-C. and Tsai, C.-F. (2020). Missing value imputation: a review and analysis of the literature (2006–2017). *Artificial Intelligence Review*, 53(2):1487–1509.
- [104] Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A. W. M., van Ginneken, B., and Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Med. Image Anal.*, 42:60–88.
- [105] Liu, J., Lichtenberg, T., Hoadley, K. A., Poisson, L. M., Lazar, A. J., Cherniack, A. D., Kovatich, A. J., Benz, C. C., Levine, D. A., Lee, A. V., Omberg, L., Wolf, D. M., Shriver, C. D., Thorsson, V., Cancer Genome Atlas Research Network, and Hu, H. (2018). An integrated TCGA Pan-Cancer clinical data resource to drive High-Quality survival outcome analytics. *Cell*, 173(2):400–416.e11.
- [106] Liu, T., Zhou, L., Li, D., Andl, T., and Zhang, Y. (2019). Cancer-associated fibroblasts build and secure the tumor microenvironment. *Frontiers in cell and developmental biology*, 7:60.
- [107] Long, J., Shelhamer, E., and Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440.

- [108] Louis, D. N., Perry, A., Wesseling, P., Brat, D. J., Cree, I. A., Figarella-Branger, D., Hawkins, C., Ng, H. K., Pfister, S. M., Reifenberger, G., Soffiatti, R., von Deimling, A., and Ellison, D. W. (2021). The 2021 WHO classification of tumors of the central nervous system: a summary. *Neuro. Oncol.*, 23(8):1231–1251.
- [109] Lu, M. Y., Chen, T. Y., Williamson, D. F. K., Zhao, M., Shady, M., Lipkova, J., and Mahmood, F. (2021). AI-based pathology predicts origins for cancers of unknown primary. *Nature*, 594(7861):106–110.
- [110] Macenko, M., Niethammer, M., Marron, J. S., Borland, D., Woosley, J. T., Xiaojun Guan, Schmitt, C., and Thomas, N. E. (2009). A method for normalizing histology slides for quantitative analysis. In *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 1107–1110. ieeexplore.ieee.org.
- [111] Maleki, F., Ovens, K., Hogan, D. J., and Kusalik, A. J. (2020). Gene set analysis: Challenges, opportunities, and future research. *Front. Genet.*, 11:654.
- [112] Manthey, D. (2021). large image: Python modules to work with large multiresolution images.
- [113] Marcinkevičs, R. and Vogt, J. E. (2020). Interpretability and explainability: A machine learning zoo mini-tour. *arXiv*.
- [114] Mardis, E. R. (2011). A decade’s perspective on DNA sequencing technology. *Nature*, 470(7333):198–203.
- [115] MATLAB (2021). Types of morphological operations - MATLAB & simulink. <https://www.mathworks.com/help/images/morphological-dilation-and-erosion.html>. Accessed: 2021-10-30.
- [116] Metzker, M. L. (2010). Sequencing technologies - the next generation. *Nat. Rev. Genet.*, 11(1):31–46.
- [117] Mnih, V., Heess, N., Graves, A., et al. (2014). Recurrent models of visual attention. In *Advances in neural information processing systems*, pages 2204–2212.

- [118] Mobadersany, P., Yousefi, S., Amgad, M., Gutman, D. A., Barnholtz-Sloan, J. S., Velázquez Vega, J. E., Brat, D. J., and Cooper, L. A. D. (2018). Predicting cancer outcomes from histology and genomics using convolutional networks. *Proc. Natl. Acad. Sci. U. S. A.*, 115(13):E2970–E2979.
- [119] Molavi, D. W. (2017). *The Practice of Surgical Pathology: A Beginner’s Guide to the Diagnostic Process*. Springer.
- [120] Momeni, A., Thibault, M., and Gevaert, O. (2018). Deep recurrent attention models for histopathological image analysis. *bioRxiv*, page 438341.
- [121] Murad, M. H., Asi, N., Alsawas, M., and Alahdab, F. (2016). New evidence pyramid. *BMJ Evidence-Based Medicine*, 21(4):125–127.
- [122] Nair, V. and Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. *arXiv*.
- [123] Nalisnik, M., Amgad, M., Lee, S., Halani, S. H., Velazquez Vega, J. E., Brat, D. J., Gutman, D. A., and Cooper, L. A. D. (2017). Interactive phenotyping of large-scale histology imaging data with HistomicsML. *Sci. Rep.*, 7(1):14588.
- [124] National Cancer Institute (NIH) (2019). Cancer statistics. 2017. <https://www.cancer.gov/about-cancer/understanding/statistics>. Accessed: 2019-12-28.
- [125] Nawaz, S., Heindl, A., Koelble, K., and Yuan, Y. (2015). Beyond immune density: critical role of spatial heterogeneity in estrogen receptor-negative breast cancer. *Modern Pathology*, 28(6):766–777.
- [126] Network, C. G. A. et al. (2012a). Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, 487(7407):330.
- [127] Network, C. G. A. R. et al. (2011). Integrated genomic analyses of ovarian carcinoma. *Nature*, 474(7353):609.
- [128] Network, C. G. A. R. et al. (2012b). Comprehensive genomic characterization of squamous cell lung cancers. *Nature*, 489(7417):519.

- [129] Network, C. G. A. T. R. et al. (2008). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, 455(7216):1061.
- [130] of Health, U. N. I. (2021). The cost of sequencing a human genome. <https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost>. Accessed: 2021-10-31.
- [131] Oskoei, M. A. and Hu, H. (2010). A survey on edge detection methods. *University of Essex, UK*, 33.
- [132] Otsu, N. (1979). A threshold selection method from gray-level histograms. *IEEE Trans. Syst. Man Cybern.*, 9(1):62–66.
- [133] Pantanowitz, L., Sinard, J. H., Henricks, W. H., Fatheree, L. A., Carter, A. B., Contis, L., Beckwith, B. A., Evans, A. J., Lal, A., Parwani, A. V., and College of American Pathologists Pathology and Laboratory Quality Center (2013). Validating whole slide imaging for diagnostic purposes in pathology: guideline from the college of american pathologists pathology and laboratory quality center. *Arch. Pathol. Lab. Med.*, 137(12):1710–1722.
- [134] Pathology, L. (2021a). Basics. <https://librepathology.org/w/index.php?title=Basics&oldid=45439>. Accessed: 2021-10-29.
- [135] Pathology, L. (2021b). Stains. <https://librepathology.org/wiki/Stains>. Accessed: 2021-10-29.
- [136] Pedersen, A. B., Mikkelsen, E. M., Cronin-Fenton, D., Kristensen, N. R., Pham, T. M., Pedersen, L., and Petersen, I. (2017). Missing data and multiple imputation in clinical epidemiological research. *Clin. Epidemiol.*, 9:157–166.
- [137] Pernick), P. O. N. (2021). IHC procedure. <https://www.pathologyoutlines.com/topic/stainsihcprocedure.html>. Accessed: 2021-10-29.
- [138] Piedbois, P. and Buyse, M. (2008). Endpoints and surrogate endpoints in colorectal cancer: a review of recent developments. *Curr. Opin. Oncol.*, 20(4):466–471.
- [139] Ping, Z., Xia, Y., Shen, T., Parekh, V., Siegal, G. P., Eltoum, I.-E., He, J., Chen, D., Deng,

- M., Xi, R., et al. (2016). A microscopic landscape of the invasive breast cancer genome. *Scientific reports*, 6(1):1–10.
- [140] Popper, K. (2005). *The logic of scientific discovery*. Routledge.
- [141] Qu, H., Wu, P., Huang, Q., Yi, J., Yan, Z., Li, K., Riedlinger, G. M., De, S., Zhang, S., and Metaxas, D. N. (2020). Weakly supervised deep nuclei segmentation using partial points annotation in histopathology images.
- [142] Rabbani, B., Tekin, M., and Mahdieh, N. (2014). The promise of whole-exome sequencing in medical genetics. *J. Hum. Genet.*, 59(1):5–15.
- [143] Reinhard, E., Adhikhmin, M., Gooch, B., and Shirley, P. (2001). Color transfer between images. *IEEE Comput. Graph. Appl.*, 21(5):34–41.
- [144] Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.*, 28:91–99.
- [145] Ripley, B. D. (1976). The second-order analysis of stationary point processes. *Journal of applied probability*, 13(2):255–266.
- [146] Ronneberger, O., Fischer, P., and Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241. Springer International Publishing.
- [147] Rüdiger, T., Höfler, H., Kreipe, H.-H., Nizze, H., Pfeifer, U., Stein, H., Dallenbach, F. E., Fischer, H.-P., Mengel, M., von Wasielewski, R., and Müller-Hermelink, H. K. (2002). Quality assurance in immunohistochemistry: results of an interlaboratory trial involving 172 pathologists. *Am. J. Surg. Pathol.*, 26(7):873–882.
- [148] Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215.
- [149] Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, 323(6088):533–536.

- [150] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. (2015). ImageNet large scale visual recognition challenge. *Int. J. Comput. Vis.*, 115(3):211–252.
- [151] Saha, M., Chakraborty, C., Arun, I., Ahmed, R., and Chatterjee, S. (2017). An advanced deep learning approach for ki-67 stained hotspot detection and proliferation rate scoring for prognostic evaluation of breast cancer. *Sci. Rep.*, 7(1):3213.
- [152] Saha, M., Chakraborty, C., and Racoceanu, D. (2018). Efficient deep learning model for mitosis detection using breast histopathology images. *Comput. Med. Imaging Graph.*, 64:29–40.
- [153] Sahai, E., Astsaturov, I., Cukierman, E., DeNardo, D. G., Egeblad, M., Evans, R. M., Fearon, D., Greten, F. R., Hingorani, S. R., Hunter, T., et al. (2020). A framework for advancing our understanding of cancer-associated fibroblasts. *Nature Reviews Cancer*, 20(3):174–186.
- [154] Sakamoto, T., Furukawa, T., Lami, K., Pham, H. H. N., Uegami, W., Kuroda, K., Kawai, M., Sakanashi, H., Cooper, L. A. D., Bychkov, A., and Fukuoka, J. (2020). A narrative review of digital pathology and artificial intelligence: focusing on lung cancer. *Transl. Lung Cancer Res.*, 9(5):2255–2276.
- [155] Salgado, R., Denkert, C., Demaria, S., Sirtaine, N., Klauschen, F., Pruneri, G., Wienert, S., Van den Eynden, G., Baehner, F. L., Penault-Llorca, F., Perez, E. A., Thompson, E. A., Symmans, W. F., Richardson, A. L., Brock, J., Criscitiello, C., Bailey, H., Ignatiadis, M., Floris, G., Sparano, J., Kos, Z., Nielsen, T., Rimm, D. L., Allison, K. H., Reis-Filho, J. S., Loibl, S., Sotiriou, C., Viale, G., Badve, S., Adams, S., Willard-Gallo, K., and Loi, S. (2015). The evaluation of tumor-infiltrating lymphocytes (TILs) in breast cancer: recommendations by an international TILs working group 2014. *Annals of Oncology*, 26(2):259–271.
- [156] Saltz, J., Gupta, R., Hou, L., Kurc, T., Singh, P., Nguyen, V., Samaras, D., Shroyer, K. R., Zhao, T., Batiste, R., Van Arnem, J., Cancer Genome Atlas Research Network, Shmulevich, I., Rao, A. U. K., Lazar, A. J., Sharma, A., and Thorsson, V. (2018). Spatial organization and molecular correlation of Tumor-Infiltrating lymphocytes using deep learning on pathology images. *Cell Rep.*, 23(1):181–193.e7.

- [157] Sanger, F. and Coulson, A. R. (1975). A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J. Mol. Biol.*, 94(3):441–448.
- [158] Savas, P., Salgado, R., Denkert, C., Sotiriou, C., Darcy, P. K., Smyth, M. J., and Loi, S. (2016). Clinical relevance of host immunity in breast cancer: from TILs to the clinic. *Nat. Rev. Clin. Oncol.*, 13(4):228–241.
- [159] Schmauch, B., Romagnoni, A., Pronier, E., Saillard, C., Maillé, P., Calderaro, J., Kamoun, A., Sefta, M., Toldo, S., Zaslavskiy, M., Clozel, T., Moarii, M., Courtiol, P., and Wainrib, G. (2020). A deep learning model to predict RNA-Seq expression of tumours from whole slide images. *Nat. Commun.*, 11(1):3877.
- [160] Schömig-Markiefka, B., Pryalukhin, A., Hulla, W., Bychkov, A., Fukuoka, J., Madabhushi, A., Achter, V., Nieroda, L., Büttner, R., Quaas, A., and Tolkach, Y. (2021). Quality control stress test for deep learning-based diagnostic model in digital pathology. *Mod. Pathol.*
- [161] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE.
- [162] Sezgin, M. and Sankur, B. (2004). Survey over image thresholding techniques and quantitative performance evaluation. *JEI*, 13(1):146–165.
- [163] Shendure, J., Balasubramanian, S., Church, G. M., Gilbert, W., Rogers, J., Schloss, J. A., and Waterston, R. H. (2017). DNA sequencing at 40: past, present and future. *Nature*, 550(7676):345–353.
- [164] Shorten, C. and Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):1–48.
- [165] Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for Large-Scale image recognition. *arXiv*.
- [166] Sims, D., Sudbery, I., Ilott, N. E., Heger, A., and Ponting, C. P. (2014). Sequencing depth and coverage: key considerations in genomic analyses. *Nat. Rev. Genet.*, 15(2):121–132.

- [167] Skrede, O.-J., De Raedt, S., Kleppe, A., Hveem, T. S., Liestøl, K., Maddison, J., Askautrud, H. A., Pradhan, M., Nesheim, J. A., Albrechtsen, F., Farstad, I. N., Domingo, E., Church, D. N., Nesbakken, A., Shepherd, N. A., Tomlinson, I., Kerr, R., Novelli, M., Kerr, D. J., and Danielsen, H. E. (2020). Deep learning for prediction of colorectal cancer outcome: a discovery and validation study. *Lancet*, 395(10221):350–360.
- [168] Springenberg, J. T., Dosovitskiy, A., Brox, T., and Riedmiller, M. (2014). Striving for simplicity: The all convolutional net. *arXiv*.
- [169] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1):1929–1958.
- [170] Stanton, S. E. and Disis, M. L. (2016). Clinical significance of tumor-infiltrating lymphocytes in breast cancer. *Journal for immunotherapy of cancer*, 4(1):1–7.
- [171] Steyerberg, E. W. and Harrell, Jr, F. E. (2016). Prediction models need appropriate internal, internal-external, and external validation. *J. Clin. Epidemiol.*, 69:245–247.
- [172] Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., and Mesirov, J. P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.*, 102(43):15545–15550.
- [173] Sun, P., He, J., Chao, X., Chen, K., Xu, Y., Huang, Q., Yun, J., Li, M., Luo, R., Kuang, J., Wang, H., Li, H., Hui, H., and Xu, S. (2021). A computational Tumor-Infiltrating lymphocyte assessment method comparable with visual reporting guidelines for Triple-Negative breast cancer. *EBioMedicine*, 70:103492.
- [174] Surveillance, Epidemiology, And End (2019). Cancer of the breast (female) - cancer stat facts. <https://seer.cancer.gov/statfacts/html/breast.html>. Accessed: 2019-12-28.
- [175] Talo, M. (2019). Automated classification of histopathology images using transfer learning. *Artif. Intell. Med.*, 101:101743.

- [176] Tan, W. C. C., Nerurkar, S. N., Cai, H. Y., Ng, H. H. M., Wu, D., Wee, Y. T. F., Lim, J. C. T., Yeong, J., and Lim, T. K. H. (2020). Overview of multiplex immunohistochemistry/immunofluorescence techniques in the era of cancer immunotherapy. *Cancer Commun.*, 40(4):135–153.
- [177] Tellez, D., Balkenhol, M., Otte-Höller, I., van de Loo, R., Vogels, R., Bult, P., Wauters, C., Vreuls, W., Mol, S., Karssemeijer, N., et al. (2018). Whole-slide mitosis detection in h&e breast histology using phh3 as a reference to train distilled stain-invariant convolutional networks. *IEEE transactions on medical imaging*, 37(9):2126–2136.
- [178] Tellez, D., Litjens, G., Bándi, P., Bulten, W., Bokhorst, J.-M., Ciompi, F., and van der Laak, J. (2019). Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. *Med. Image Anal.*, 58:101544.
- [179] Tellez, D., Litjens, G., van der Laak, J., and Ciompi, F. (2021). Neural image compression for gigapixel histopathology image analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(2):567–578.
- [180] Tenny, S. and Varacallo, M. (2018). Evidence based medicine (ebm). *Treasure Island: Stat-Pearls Publishing*.
- [181] Tizhoosh, H. R. and Pantanowitz, L. (2018). Artificial intelligence and digital pathology: Challenges and opportunities. *J. Pathol. Inform.*, 9:38.
- [182] Tomczak, K., Czerwińska, P., and Wiznerowicz, M. (2015). The cancer genome atlas (TCGA): an immeasurable source of knowledge. *Contemp. Oncol.*, 19(1A):A68–77.
- [183] Vahadane, A., Peng, T., Albarqouni, S., Baust, M., Steiger, K., Schlitter, A. M., Sethi, A., Esposito, I., and Navab, N. (2015). Structure-preserved color normalization for histological images. In *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*, pages 1012–1015.
- [184] van der Laak, J., Litjens, G., and Ciompi, F. (2021). Deep learning in histopathology: the path to the clinic. *Nat. Med.*, 27(5):775–784.

- [185] Van Eycke, Y.-R., Allard, J., Salmon, I., Debeir, O., and Decaestecker, C. (2017). Image processing in digital pathology: an opportunity to solve inter-batch variability of immunohistochemical staining. *Sci. Rep.*, 7:42964.
- [186] van Rijthoven, M., Balkenhol, M., Siliņa, K., van der Laak, J., and Ciompi, F. (2021). HookNet: Multi-resolution convolutional neural networks for semantic segmentation in histopathology whole-slide images. *Med. Image Anal.*, 68:101890.
- [187] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.
- [188] Verhaak, R. G. W., Hoadley, K. A., Purdom, E., Wang, V., Qi, Y., Wilkerson, M. D., Miller, C. R., Ding, L., Golub, T., Mesirov, J. P., Alexe, G., Lawrence, M., O’Kelly, M., Tamayo, P., Weir, B. A., Gabriel, S., Winckler, W., Gupta, S., Jakkula, L., Feiler, H. S., Hodgson, J. G., James, C. D., Sarkaria, J. N., Brennan, C., Kahn, A., Spellman, P. T., Wilson, R. K., Speed, T. P., Gray, J. W., Meyerson, M., Getz, G., Perou, C. M., Hayes, D. N., and Cancer Genome Atlas Research Network (2010). Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell*, 17(1):98–110.
- [189] Vijayalakshmi A and Rajesh Kanna B (2020). Deep learning approach to detect malaria from microscopic images. *Multimed. Tools Appl.*, 79(21):15297–15317.
- [190] Wallace, S. E. and Bean, L. J. H. (2020). *Educational Materials - Genetic Testing: Current Approaches*. University of Washington, Seattle.
- [191] Wang, C.-W., Huang, S.-C., Lee, Y.-C., Shen, Y.-J., Meng, S.-I., and Gaol, J. L. (2021). Deep learning for bone marrow cell detection and classification on whole-slide images. *Med. Image Anal.*, 75:102270.
- [192] Wang, J., Chen, R. J., Lu, M. Y., Baras, A., and Mahmood, F. (2020). Weakly supervised prostate tma classification via graph convolutional networks. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 239–243.

- [193] Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, 10(1):57–63.
- [194] Xing, F. and Yang, L. (2016). Robust Nucleus/Cell detection and segmentation in digital pathology and microscopy images: A comprehensive review. *IEEE Rev. Biomed. Eng.*, 9:234–263.
- [195] Xu, J., Xiang, L., Wang, G., Ganesan, S., Feldman, M., Shih, N. N., Gilmore, H., and Madabhushi, A. (2015). Sparse non-negative matrix factorization (SNMF) based color unmixing for breast histopathological image analysis. *Comput. Med. Imaging Graph.*, 46 Pt 1:20–29.
- [196] Xu, Y., Mo, T., Feng, Q., Zhong, P., Lai, M., Eric, I., and Chang, C. (2014). Deep learning of feature representation with multiple instance learning for medical image analysis. In *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 1626–1630. IEEE.
- [197] Yadav, V. K. and De, S. (2015). An assessment of computational methods for estimating purity and clonality using genomic data derived from heterogeneous tumor tissue samples. *Brief. Bioinform.*, 16(2):232–241.
- [198] Yang, X., Amgad, M., Cooper, L. A., Du, Y., Fu, H., and Ivanov, A. A. (2020). High expression of *mkk3* is associated with worse clinical outcomes in african american breast cancer patients. *Journal of translational medicine*, 18(1):1–19.
- [199] Yousefi, S., Amrollahi, F., Amgad, M., Dong, C., Lewis, J. E., Song, C., Gutman, D. A., Halani, S. H., Velazquez Vega, J. E., Brat, D. J., and Cooper, L. A. D. (2017). Predicting clinical outcomes from large scale cancer genomic profiles with deep survival models. *Sci. Rep.*, 7(1):11707.
- [200] Yuan, Y., Failmezger, H., Rueda, O. M., Ali, H. R., Gräf, S., Chin, S.-F., Schwarz, R. F., Curtis, C., Dunning, M. J., Bardwell, H., Johnson, N., Doyle, S., Turashvili, G., Provenzano, E., Aparicio, S., Caldas, C., and Markowitz, F. (2012). Quantitative image analysis of cellular heterogeneity in breast tumors complements genomic profiling. *Sci. Transl. Med.*, 4(157):157ra143.
- [201] Zoeller, E. L., Pedro, B., Konen, J., Dwivedi, B., Rupji, M., Sundararaman, N., Wang, L.,

Horton, J. R., Zhong, C., Barwick, B. G., et al. (2019). Genetic heterogeneity within collective invasion packs drives leader and follower cell phenotypes. *Journal of cell science*, 132(19):jcs231514.