Cell-type-specific nuclear morphology predicts genomic instability and prognosis in multiple cancer types

John Abel[1], Suyog Jain[1], Deepta Rajan[1], Harshith Padigela[1], Kenneth Leidal[1], Aaditya Prakash[1], Jake Conway[1], Michael Nercessian[1], Christian Kirkup[1], Syed Ashar Javed[1], Robert Egger[1], Benjamin Trotter[1], Ylaine Gerardin[1], Jacqueline A. Brosnan-Cashman[1], Aditya Dhoot[1], Michael C. Montalto[1], Ilan Wapinski[1], Archit Khosla[1], Michael G. Drage[1], Limin Yu[1], Amaro Taylor-Weiner[1]*

[1]PathAI, Boston, Massachusetts

Correspondence to:
Amaro Taylor-Weiner
PathAI
1325 Boylston Street, Suite 10000
Boston, MA 02215
Phone: 617-500-8457
Email: amaro.taylor@pathai.com

Abstract

Background: Altered nuclei are ubiquitous in cancer, with changes in nuclear size, shape, and coloration accompanying cancer progression. However, comprehensive quantification of nuclear morphology across a whole-slide histologic image remains a challenge. Here, we describe the development of a pan-tissue, deep learning-based digital pathology pipeline for exhaustive nucleus detection, segmentation, and classification and the utility of this pipeline for nuclear morphologic biomarker discovery.

Methods: Manually-collected nucleus annotations (>29,000) were used to train an object detection and segmentation model for identifying nuclei, which was deployed to segment nuclei in H&E-stained slides from the BRCA, LUAD, and PRAD TCGA cohorts. Interpretable features describing the shape, size, color, and texture of each nucleus were extracted from model predictions. These quantitative nuclear features were compared to available metrics, including measurements of genomic instability, gene expression, and prognosis.

Results: The nuclear segmentation and classification model trained herein performed comparably to previously reported models. Features extracted from the model revealed differences sufficient to distinguish between BRCA, LUAD, and PRAD. Furthermore, cancer cell nuclear area was associated with increased genomic instability, as measured by aneuploidy score and homologous recombination deficiency. In BRCA, increased fibroblast nuclear area was indicative of poor progression-free and overall survival and was associated with gene expression signatures related to extracellular matrix remodeling and anti-tumor immunity.

Conclusions: We developed a powerful pan-tissue approach for nucleus segmentation and featurization, enabling the construction of predictive models and the identification of features linking nuclear morphology with clinically-relevant prognostic biomarkers across multiple cancer types.

Introduction

Histological assessment of tissue is central to the diagnosis and classification of malignancy, and critically informs patient management;pathologists routinely report visible alterations in nuclear morphology. Altered nuclear features are ubiquitous in cancer, and changes in nuclear size, shape, coloration, texture, nucleoli, and nuclear-cytoplasmic ratio, are important features of histologic grade, which has prognostic relevance independent of disease stage (1,2). The enumeration and morphologic features of mitoses also informs pathologist assessment of malignancy (3). Furthermore, nuclei can be well-established markers of specific cancers; for example, a clear nucleus ("Orphan Annie Eye") is a known feature of papillary thyroid carcinoma (4).

A complex interplay exists between nuclear morphology and the genetic, epigenetic, and transcriptomic milieu of cancer cells, reflecting the importance of the nucleus to the process of oncogenic transformation. Distorted nuclei can indicate dysregulated replication processes, aneuploidy, genomic instability, and genetic mutations that affect stability and function of the nuclear envelope (5). Indeed, many cancers have altered expression of nuclear envelope components, resulting in nuclear rupture and micronuclei formation, further increasing genomic instability (5,6). In addition, components of the nuclear envelope are known to bind to both chromatin and transcription factors, providing a spatial regulation to gene transcription and expression (5,7). Therefore, the visual appearance of a cancer cell nucleus has the potential to reveal key information about the biology of a tumor.

The quantitation of nuclear morphology has been a long sought-after goal (8). Early studies used semi-quantitative approaches to enumerate features such as nuclear size and shape; these works revealed relationships between increased nuclear area and altered nuclear shape with poor prognosis and advanced disease in breast cancer and prostate cancer, respectively (9–11). The use of computational approaches in pathology image analysis to identify and quantify nuclear changes has gained traction as modern computer vision methods have allowed for rapid, reproducible and cost-effective quantification of nuclear morphology. Using these methods, nuclear

morphometric features have been shown to correlate with relevant clinical and pathological metrics, such as oligodendroglioma component in glioblastoma (12), as well as stage (13), disease aggressiveness (14), recurrence (15–17), and outcome (18) in other cancer subtypes. In addition, increased nuclear size has been correlated with whole genome duplication (19,20), and nuclear morphometric features have allowed for the prediction of relevant molecular information, such as ER status (21) and Oncotype DX risk scores (22,23) in breast cancer. Most recently, Nimgaonkar et al. described an AI-derived histologic signature, the main component of which was variance in nuclear morphology in cancer cells, that predicted response to gemcitabine in patients with pancreatic adenocarcinoma (24).

Digitized whole slide images (WSIs) have enhanced the degree to which nuclear morphology can be studied in histological specimens (12,13). However, the large size of WSIs - up to billions of pixels and containing thousands of nuclei - makes exhaustive manual annotation infeasible; thus studies have relied on manually-selected subregions of interest rather than entire slides (20,25). Automated methods are, therefore, needed to fully quantify nuclear features in WSIs. We recently described a cell- and tissue-level computational pathology pipeline using WSIs for the automated computation of human interpretable features (HIFs), distinctive features with tangible methods for validation (26). This pipeline allows the use of HIFs to predict treatment-relevant molecular phenotypes and allows for integration with current pathological methods. Given that morphological analysis of histology features is central to pathology workflows, we sought to extend this work to identify nuclear human interpretable features (nuHIFs) in multiple cancer types.

In this paper, we present a multi-tissue model for the exhaustive detection, segmentation, and classification of nuclei from entire hematoxylin and eosin (H&E)-stained WSIs, allowing for the exhaustive analysis of slide-level descriptors of nuclear size, shape, texture and staining intensity. Furthermore, we demonstrate that these nuHIFs are predictive of clinically relevant information in multiple cancer types.

Methods

Brief Overview

Manually collected annotations were used to train and validate an object detection and segmentation model to detect and segment nuclei from H&E stained tissue slides. This model was deployed on whole-slide H&E images from The Cancer Genome Atlas (TCGA) to extract features from each nucleus in each slide, and the resulting features were used to analyze the relationship between nuclear morphology and underlying molecular markers of cancer, and patient outcomes.

Dataset Description and Annotation Collection

Over 29,000 manual annotations of cell nuclei were collected from H&E images from 21 tumor types at 40x and 20x magnification from TCGA (27), as well as a proprietary set of H&E-stained tissue biopsies of skin, liver non-alcoholic steatohepatitis, colon inflammatory bowel disease, and kidney lupus. Board-certified pathologists (MGD and LY) selected 1000 x 1000 pixel patches that were exemplary of varied tissue and nuclear morphology from the training slides and trained collaborators to perform exhaustive manual annotation of nuclei in the patches. Annotations were checked for quality, adjusted, and confirmed by MGD and LY. This process resulted in 67 WSI patches exhaustively annotated for nuclei. These patches were split into training, validation, and held-out test data sets to ensure distribution of tissue types (Table 1).

Nuclear Segmentation Model Architecture

A Mask-RCNN-style architecture was selected for nuclear segmentation. A ResNet50 backbone pretrained on the ImageNet dataset was used to produce the feature pyramid network. The first two of five modules that comprise ResNet50 were frozen during training to preserve the pretrained weights of early layers. Model development was performed using the PyTorch library (28) .

Nuclear Segmentation Model Training

The manually-collected annotations were used to train the model for detecting and segmenting cellular nuclei (Figure 1A). During training, the annotated patches were augmented by crops, flips, rotations, and affine deformations.

Cell Classification

Following nuclear segmentation, the cell class of each nucleus was assigned using previously developed cell classification models specific to breast cancer (BRCA) and lung adenocarcinoma (LUAD), and a prostate adenocarcinoma (PRAD) cell classification models was developed using similar methods (26); performance of the prostate cancer model is described in Figures S1 and S2. Example prediction results are shown in Figure 3.

Deployment Dataset and Feature Extraction

The nuclear segmentation model was deployed on publicly available images of H&E slides from the BRCA (N=892), PRAD (N=392), and LUAD (N=426) TCGA cohorts; a summary of clinicopathologic features of each cohort is shown in Table 2. Model performance was qualitatively assessed by board-certified pathologists and determined to be consistent with performance on the held-out test dataset. The features computed for each individual nucleus were: area, circularity, eccentricity, major and minor axis length, perimeter, solidity, and the mean and standard deviation of pixel grayscale intensity, pixel saturation, and pixel A and B channels in LAB colorspace. The mean and standard deviation of each feature from each nucleus class on the slides were used to summarize the nuclear morphology on each slide. This yielded 30 slide-level nuHIFs for each cell type, e.g. the mean area of cancer nuclei, the standard deviation of CAF nuclear eccentricity, or the mean pixel grayscale intensity of lymphocyte nuclei. Attributes and features described by nuHIFs are included in Figure 1B. Thus, the total number of features summarizing the morphology on each slide was 30 times the number of cell classes.

Exploring Cancer Type and Nuclear Morphology

To compare the nuHIFs quantifying cancer cell, cancer-associated fibroblast (CAF), and tumor-infiltrating lymphocyte (TIL) morphology, uniform manifold approximation and projection (UMAP) analysis was performed. Nuclear HIFs were z-scored across all cancer types for standardization. UMAP was parameterized with 100 neighbors, an embedding dimension of 2, and the Euclidean distance metric. Features characteristic of each cancer type were evaluated by averaging each feature across the samples of each cancer type and z-scoring for visualization; hierarchical clustering (using Euclidean distance with average linkage) identified features that varied across cancer types.

Classifying Cancer Type from Nuclear Morphology

Random forest (RF) binary classification models were trained and applied to each cell-type-specific nuHIF set to differentiate between pairs of cancer types. RF classification models were trained using 5-fold stratified cross-validation with balanced class weighting. The performance of each model was assessed using the area under the receiver operating characteristic curve (AUROC) on each held-out validation split. The mean AUROC on the held-out validation splits is reported. RF model training was performed in scikit-learn with default hyperparameters (100 trees) (29).

Association of Cancer nuHIFs with Metrics of Genome Instability

We used Spearman (rank-based) correlation to find the association between variation in cancer nuclear morphology and metrics of genomic instability. We captured variation in size by the nuHIF “standard deviation of cancer cell nuclear area” for each slide. For metrics of genomic instability, we selected previously published metrics: aneuploidy score (30) and homologous recombination deficiency (HRD) score (31).

Prediction of WGD from Cancer nuHIFs

RF binary classification models were trained and applied to the cancer nuHIF set from each cancer type to predict binarized whole-genome doubling (WGD; 1-2 doublings = 1; no doublings = 0). RF classification models were trained using 5-fold stratified cross-validation with balanced class weighting. The performance of each model was evaluated using AUROC on each held-out validation split. The mean AUROC on the

held-out validation splits is reported. RF model training was performed in scikit-learn with default hyperparameters. The mean RF Gini importance (also called the mean decrease in impurity) of the top five features for each cancer type across the five splits are reported.

Association of Cancer-Associated Fibroblast nuHIFs in BRCA with Clinical Outcomes

Cox proportional hazard models were utilized to explore the relationship between BRCA CAF nuHIFs and overall and progression-free survival (OS and PFS, respectively). Ordinal tumor stage (1-4) and patient age were included as clinical covariates; 17 subjects missing tumor stage and the one missing survival data were excluded. Robust z-scoring (i.e. using the median and scaled interquartile range) of each nuHIF before modeling was performed for simple interpretation of the hazard ratios (HRs). The p-values associated with each nuHIF were corrected for false discovery rate (FDR) by the Benjamini-Hochberg procedure. Survival analyses were performed using the Lifelines library (32).

Gene Expression Analysis

Gene expression data was acquired from the Genomic Data Commons (GDC)-processed TCGA BRCA cohort (release 18.0) from the UCSC Xena data portal (33). We mapped each probe to a gene and discarded non-unique mappings. Gene expression samples were paired to case-matched slides in our dataset, yielding 868 expression-nuHIF pairs. Spearman (rank-based) correlation was used to quantify the association between bulk RNAseq expression and the mean fibroblast nucleus area nuHIF for each gene, and corrected for FDR via Benjamini-Hochberg procedure. Genes with corrected $p < 0.05$ and Spearman correlation greater than 0.15 or less than -0.15 were selected to comprise the significant positively and negatively associated gene sets, respectively, for gene set enrichment analysis (GSEA). GSEA (34) was performed using the Molecular Signatures Database (MSigDB) (35) and the REACTOME pathway database (36), and the ten most significant pathway overlaps, with FDR-corrected $p < 0.05$, are reported.

Results

Model development, performance, and nuclear feature extraction

We collected annotations and trained a machine learning (ML) model to detect and segment nuclei in H&E-stained WSIs as described in the Methods and shown in Figure 1A. The model is not limited to sampling regions of interest from tissue samples, but rather can be utilized to exhaustively annotate WSIs. Application of the model to our held-out test data, including held-out tissue and disease types, demonstrated performance (mean Dice score=0.818, aggregated Jaccard index=0.619) comparable to models reported previously in the literature (37,38). Importantly, model speed was adequate to apply to multi-gigabyte WSIs at full resolution (approximately 0.25 um/pixel; roughly 30 minutes per slide). Examples of model performance in mesothelioma, head and neck squamous cell carcinoma, and stomach adenocarcinoma are shown in Figure 2.

Having validated our model's performance, we deployed the resulting model on primary diagnostic (DX1) H&E slides from the breast cancer (BRCA; N=892), prostate adenocarcinoma (PRAD; N=392), and lung adenocarcinoma (LUAD; N=426) TCGA cohorts (Figure 3); model performance was visually assessed to be consistent with test data. We extracted interpretable features describing the shape, size, staining intensity, and texture of every nucleus on each WSI (Table S2). We performed further analysis on nuHIFs specific to cancer cells, CAFs, and lymphocytes, as these three cell classes are common across all cancer types and have been implicated in clinical outcomes.

nuHIFs show within- and between-cancer type variation

To assess whether nuHIFs differ between cancer types, we performed UMAP to compare the nuHIFs from cancer cells, CAFs, and lymphocytes in BRCA, LUAD, and PRAD datasets. We observed notable inter- and intra-dataset variation in nuHIFs. For cancer cells, nuclear morphology was distinct between PRAD and LUAD datasets, while BRCA dataset cancer cells showed nuclear features similar to both PRAD and LUAD (Figure 4A). Unsupervised hierarchical clustering of z-scored features revealed specific nuHIFs enriched in these three cancer subtypes (Figure 4B). Features associated with

nuclear size were higher in PRAD cancer nuclei relative to LUAD, while features associated with nuclear staining intensity, color, and shape were higher in LUAD nuclei relative to PRAD. The clearest distinction between cancer subtypes was discerned through nuHIFs of CAFs in BRCA, LUAD, and PRAD (Figure 4B). Unsupervised hierarchical clustering revealed specific features enriched in CAFs from these cancer subtypes (Figure 4B). Interestingly, lymphocyte nuHIFs also differed between cancer types. Features associated with increased and decreased CAF nuclear size are shown in Tables S3 and S4, respectively.

Because of the apparent association between nuclear morphology and cancer type, we hypothesized that nuHIF-quantified nuclear morphology could be a distinguishing feature of cancer types. To test this, we constructed a simple random forest binary classification model for differentiating between each pair of cancer types (BRCA, PRAD, LUAD) using cancer, CAF, or lymphocyte nuclear HIFs. We performed five-fold cross validation to estimate the extent to which cancer types may be differentiated by nuclear morphology. We found consistently strong performance for differentiating between cancer types using nuclear morphology (Figure 4D). Although lymphocyte nuclear morphology was less distinct between cancer types when evaluated in an unsupervised fashion with UMAP, supervised analysis indicated that lymphocyte morphology differed between cancer types.

<u>Cancer nuclear morphology is associated with metrics of genomic instability in multiple cancer types</u>

Cancer nuclear atypia is used clinically as a marker of malignancy. We therefore hypothesized that underlying levels of genomic instability may partially explain the observed heterogeneity in cancer nuclear morphology within cancer subtypes, as well as between cancer types with known differences in malignancy. We tested this hypothesis by assessing the relationship between cancer nuclear morphology and genomic instability in LUAD, BRCA, and PRAD cohorts using aneuploidy score and homologous recombination deficiency (HRD) score as metrics of genomic instability. Indeed, using the standard deviation of cancer nuclear area as a metric of nuclear

atypia, we detected significant correlation between this nuHIF and both aneuploidy score (Figure 5A) and HRD score (Figure 5B). When assessed in a pan-cancer manner, the overall correlation increased, and the pattern observed in cancer nuHIF UMAP analysis persisted: PRAD displayed a lower level of genomic instability across both metrics compared to LUAD, while BRCA showed a wide range of genomic instability, with similarities to both PRAD and LUAD. These results confirm that cancer nuclear morphology, especially variability in nuclear size, correlates with the level of genomic instability.

Because aneuploidy score was correlated to variation in cancer nuclear area, we posited that cancer nuclear morphology was predictive of whole genome doubling. To address this hypothesis, we trained random forest models for predicting binarized whole-genome doubling using cancer nuHIFs from each of the BRCA, LUAD, and PRAD cancer types. We found that cancer nuclear morphology was predictive of WGD for each cancer type, with strongest predictive power in BRCA, and more variation in performance expected for PRAD, where WGD occurs less frequently (Figure 5C). The mean RF importance across the five splits is reported for the top five features for each cancer type in Table S5. Briefly, variation in cancer nuclear dimensions were most important for predicting WGD in BRCA, mean cancer nuclear dimensions were most important for predicting WGD in LUAD, and a mix of color and shape features were found to be most important for PRAD.

<u>CAF nuclear morphology is associated with survival and gene expression patterns in breast cancer</u>

The interplay between CAFs and cancer cells is complex and prognostically relevant, as associations between CAFs and cancer progression have been recently described (39–42). Notably, in breast cancer, CAFs have been shown to contribute to prognosis (43), while CAF subset heterogeneity correlates with metastasis (44). We therefore hypothesized that CAF nuHIFs in BRCA would be clinically prognostic. We sought to identify CAF nuHIFs that are associated with progression-free (PFS) and/or overall survival (OS). We performed regression between each CAF nuHIF and PFS and OS

using Cox proportional hazards models. After FDR correction, multiple CAF nuHIFs were significantly prognostic of survival (Tables S6 and S7). Features quantifying the same general attribute, *e.g.* nuclear area and nuclear axis length as measures of size, were indeed found to be correlated with one another (mean pairwise Pearson r = 0.82 for fibroblast nuclear area, major axis length, minor axis length, and perimeter). We selected the mean CAF nucleus area ("MEAN[FIBROBLAST_NUCLEUS_AREA]_H & E") for further evaluation, and show Kaplan-Meier survival curves for PFS and OS for the population binarized by this feature median value. High nuclear area was prognostic of worse outcomes (Figure 6, PFS HR = 1.63, 95% CI [1.19-2.23], p = 0.002; OS HR = 1.53, 95% CI [1.06, 2.20], p = 0.022).

Having identified this relationship between CAF nuclear size and prognosis, we sought to assess whether mean CAF nuclear area was associated with differences in bulk gene expression in breast cancer. We computed the rank-based (Spearman) correlation between CAF mean nuclear area and each gene in TCGA bulk gene expression to identify genes associated with this nuHIF (see Methods for details). CAF nuclear area was weakly (absolute r > 0.15) but significantly associated with expression of numerous individual genes (Supplementary Table S8). Gene set enrichment analysis revealed significant relationships between CAF nuclear size and levels of several previously identified expression pathways. Notably, larger CAF nuclear size showed positive association with gene expression in pathways associated with degradation and remodeling of the extracellular matrix (Table S3) indicating higher CAF activity. Meanwhile, larger CAF nuclear size showed negative association with the expression of genes in pathways relating to immune response to the tumor, such as B cell receptor signaling and lymphoid cell interactions with non-lymphoid cells (Table S4). Taken together, these results indicate that CAF nuclear morphology is indicative of underlying patterns of gene expression with potential prognostic utility.

Discussion

In this study, we have presented a pan-tissue approach for nucleus segmentation, classification, and featurization on entire whole-slide pathology images. This method enabled the construction of predictive models and the identification of features linking nuclear morphology with quantitative biomarkers across BRCA, PRAD, and LUAD. These results highlight the potential of ML-enabled quantification of nuclear morphometry as a prognostic feature of many cancer types and a potential biomarker to be used by pathologists. Furthermore, this approach enables the quantitative testing of hypotheses and numerical quantification of histological relationships proposed by pathologists (*e.g.*, by establishing a numerical relationship between nuclear atypia and disease metrics). In addition, our approach enables the data-driven identification of sub-visual changes that may be clinically meaningful.

One particular strength of our approach is the ability to not only measure morphologic features associated with nuclei in a cancer specimen, but to assign a cell class to each nucleus, as well. To our knowledge, this work provides the first characterization of nuclear morphologies of specific cell types in different cancers at scale. As such, we were not only able to assess the associations of cancer cell nuclear morphology with clinically relevant metrics, but we were also able to examine these relationships using nuclear features of CAFs and lymphocytes. For example, CAF nuHIFs provided a clear separation of cancer types in both unsupervised and supervised analyses, indicating that the nuclear morphologies of CAFs differ in breast, lung, and prostate cancers. Given recent observations that CAFs can be classified into multiple functional subtypes based on gene expression (45), the distinctive nuclear morphologies seen in CAFs of breast, lung, and prostate cancers suggests that CAFs may contribute to cancer progression differently in these three cancer types.

Herein, we observed that nuclear morphology differed between cancers as assessed using nucleus segmentation models. This result was observed not only for cancer epithelial cells and fibroblasts, but also, surprisingly, for lymphocytes. However, caution is warranted in interpretation – it is plausible that batch effects between slides from

different tumor groups could drive variation in nuclear presentation, especially due to differences in pre-analytic variables such as slide preparation and staining. However, it is also plausible that this finding reflects the differences in genetic and epigenetic landscapes between tumor types, levels of genomic instability, and overall differences in cancer evolution between these cancer types that may manifest as disparate nuclear morphologies.

The observed relationship between greater variation in cancer nuclear area and genomic instability was consistent across cancer types, indicating a quantitative link between nuclear pleomorphism and genomic instability pertinent to numerous cancer histologies. Prior analyses have noted an association between increased variation in nuclear size and whole genome doubling, suggesting a direct link between variation in nuclear size and genomic instability (19)(20). Given that nuclear size reflects DNA content, variation in nuclear size features between cells may be linked to underlying genomic instability. Similarly, recent work identified a histologic signature based on variability in nuclear morphology in pancreatic cancer cells that was associated with improved response to gemcitabine but was not associated with a previously defined gene expression-based disease subtype (24). Pancreatic cancer patients with BRCA1/2 mutations, associated with increased genomic instability, are known to respond more favorably to therapy regimens involving gemcitabine (46); thus, our result that nuclear variation is associated with genomic instability may explain this recent finding. To this end, our observation that variability in nuclear size (measured here by standard deviation of cancer cell nuclear area) is consistent with these prior hypotheses and allows for them to be tested on a larger scale for each case (all cells for each cell type in the WSI).

One mitigation to potential batch effects is to analyze nuclear morphology within a single cancer type, and additionally to focus on size and shape features that are more likely to be robust to tissue preparation variabilities. For example, in breast cancer, we observed a clear relationship between CAF nuclear size, prognosis, and gene expression patterns. In breast cancer, increased CAF nuclear area was positively

correlated with gene expression in extracellular matrix remodeling pathways and negatively correlated with genes in anti-tumor immune response pathways. The CAF subtypes present in a breast cancer sample may impact the tumor immune microenvironment (45). While it would be interesting to posit that CAF nuclear morphology could reflect these CAF subtypes, the ability to explore this is precluded by the use of bulk RNAseq data, since fibroblast nuclear features and the bulk expression profiling reflect a summarization of a whole slide. However, because nuclear morphology is quantified at single-cell resolution, this approach could be tied directly to single-cell expression analysis. Further work is necessary to delineate the functional relevance of nuclear changes in CAFs.

The approach that we undertook for nucleus segmentation and morphometry analysis in this paper has several key strengths. First, the ability to compute human-interpretable nuclear features at scale enables testing quantitative biological hypotheses, rather than relying on by-eye estimation of parameters such as variation in nuclear morphology. The ability to perform these analyses on WSIs of H&E-stained cancer tissue additionally obviates the need to hand-select regions of interest, which may contribute to biased analyses. In addition, we were able to train and deploy our model on tissues from diverse cancer types, suggesting that the model can be readily deployed in a cancer-agnostic fashion (47).

A particular strength of this approach is the interpretability of the predictions made. While HIF-based predictive clinical models are inherently less flexible than end-to-end black-box approaches (and, thus, can yield lower performance), they benefit from the lower dimensionality of features as a method of regularization, as the HIFs used herein directly map to low-dimensional representations of the tissue image. Furthermore, HIF-based models allow researchers and clinicians to learn from the features and generate novel hypotheses without discarding the wealth of known biology.

Although our results point to the potential of nuclear segmentation, classification, and feature analysis as a clinical screening tool, our study is limited in that our biomarker

analysis was focused on academically curated datasets. These datasets were selected due to their size, completeness, and rich genomic and transcriptomic profiling data. Construction and validation of generalizable predictive machine-learning models requires the inclusion of a broad range of training and validation data, and future efforts should focus on validating these hypotheses in additional cohorts. The technical approaches we describe here have been validated by their application to other clinical data sets, showing their generalizability of this methodology and robustness of these models (data not shown).

In sum, this work highlights the power of ML-driven quantitative nuclear morphometry in multiple cancer types. The models and resulting features described herein have the capability not only to aid pathologists and research teams in discerning novel biomarkers but to provide meaningful prognostic information for cancer patients. The ability to measure these features robustly and consistently at scale may enable the development of improved clinical tools for advancing precision medicine.

Tables

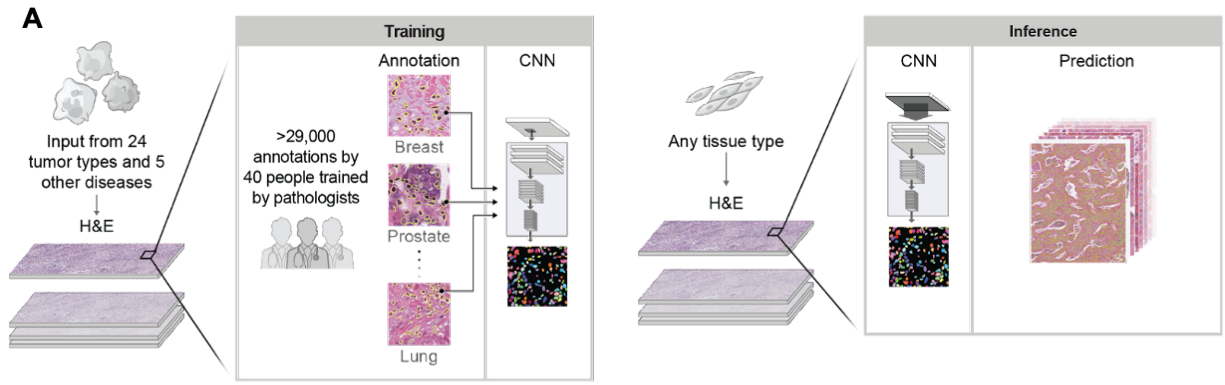
Table 1. Samples used for training and evaluating segmentation model.				
Source	**# Frames**	**Magnification**	**Indication**	**Dataset**
TCGA-A2-A3XS-01Z-00-DX1	3	40X	Breast, Infiltrating Ductal Carcinoma	Train
TCGA-XF-AAMJ-01Z-00-DX1	2	40X	Bladder, Urothelial Carcinoma	Train
TCGA-UW-A72Q-01Z-00-DX1	3	40X	Liver, Hepatocellular Carcinoma	Train
TCGA-G3-A7M6-01Z-00-DX1	2	40X	Liver, Hepatocellular Carcinoma	Train
TCGA-KM-A7QL-01Z-00-DX1	1	40X	Kidney, Renal Cell Carcinoma	Train
TCGA-CV-A6JD-01Z-00-DX1	1	40X	Head and Neck Squamous Cell Carcinoma	Train
TCGA-02-0009-01Z-00-DX1	1	20X	Brain, Glioblastoma	Train
TCGA-LB-A8F3-01Z-00-DX1	3	40X	Pancreas, Pancreatic Adenocarcinoma	Train
TCGA-36-2547-01A-01-TS1	1	40X	Ovary, Ovarian Serous Cystadenocarcinoma	Train
TCGA-W5-AA30-01Z-00-DX	3	40X	Bile Duct, Cholangiocarcinoma	Train

1				
TCGA-86-7701-01Z-00-DX1	2	40X	Lung, Lung Adenocarcinoma	Train
TCGA-B0-4823-01Z-00-DX1	3	40X	Kidney, Renal Clear Cell Carcinoma	Train
TCGA-02-0010-01Z-00-DX1	1	40X	Liver, Hepatocellular Carcinoma	Train
TCGA-2G-AAKO-05Z-00-DX1	1	40X	Testicle, Acute Myeloid Leukemia	Train
PathAI	3	40X	Liver, Hepatitis B	Train
PathAI	3	40X	Liver, Non-alcoholic steatohepatitis	Train
PathAI	3	40X	Kidney, Lupus	Train
PathAI	1	40X	Colon, Inflammatory Bowel Disease	Train
PathAI	2	20X	Small Intestine, Carcinoid	Train
PathAI	1	40X	Prostate, Benign Prostate Tissue	Train
TCGA-EJ-5495-01Z-00-DX1	2	40X	Prostate, Prostate Adenocarcinoma	Validation
TCGA-IG-A3YA-01Z-00-DX1	2	40X	Esophagus, Esophageal Carcinoma	Validation
TCGA-GS-A9U3-01Z-00-DX1	2	40X	Lymph, Diffuse Large B Cell Lymphoma	Validation
TCGA-LL-A5YM-01Z-00-DX1	2	40X	Breast, Invasive Ductal Carcinoma	Validation

TCGA-P4-AAVO-01Z-00-DX1	3	40X	Kidney, Renal Papillary Cell Carcinoma	Validation
PathAI	2	40X	Liver, Non-alcoholic steatohepatitis	Validation
PathAI	1	20X	Colon, Human Papillomavirus	Validation
PathAI	2	40X	Brain, Glioma	Validation
PathAI	2	40X	Skin, Normal	Validation
TCGA-VQ-A8DV-01Z-00-DX1	1	40X	Stomach, Stomach Adenocarcinoma	Test
TCGA-CV-6934-01Z-00-DX1	2	40X	Head and Neck Squamous Cell Carcinoma	Test
TCGA-MQ-A6BR-01Z-00-DX1	1	40X	Lung, Mesothelioma	Test
TCGA-OR-A5J8-01Z-00-DX1	1	40X	Adrenal, Adrenal Cortical Carcinoma	Test
PathAI	1	40X	Colon, Inflammatory Bowel Disease	Test
PathAI	1	20X	Breast, Invasive Ductal Carcinoma	Test
PathAI	2	40X	Prostate, Benign Prostate Tissue	Test

Table 2. Patient characteristics			
	TCGA BRCA N=891	TCGA LUAD N=423	TCGA PRAD N=392
Age at initial pathologic diagnosis (median, range)	58 (26-90; N=890)	66 (33-88)	61 (41-77)
Sex (male), n (%)	10 (1.1%)	190 (44.9%)	392 (100%)
American Joint Committee on Cancer (AJCC) Tumor Stage (n)	Stage I: 78 Stage IA: 76 Stage IB: 7 Stage II: 6 Stage IIA: 286 Stage IIB: 207 Stage III: 2 Stage IIIA: 130 Stage IIIB: 18 Stage IIIC: 51 Stage IV: 13 Unknown: 17	Stage I: 5 Stage IA: 122 Stage IB: 107 Stage II: 1 Stage IIA: 47 Stage IIB: 58 Stage IIIA: 49 Stage IIIB: 5 Stage IV: 21 Unknown: 8	Unknown: 392 (100%)
Whole genome doublings (WGD)	0: 440, 56.8% 1: 298, 38.5% 2: 37, 4.8% Unknown: 116	0: 168, 42.2% 1: 191, 48.0% 2: 39, 9.8% Unknown: 25	0: 337, 92.1% 1: 29, 7.9% Unknown: 26
Progression-free survival (PFS, median)	122.3 mo (N=889)	29.3 mo	116.7 mo
Overall survival (OS, median)	131.4 mo (N=889)	49.7 mo	Not Reached

Figures



B

Attribute	Feature
Size	Area Major Axis Length Minor Axis Length Perimeter Length
Shape	Circularity Eccentricity
Texture	Solidity Standard deviation of the grayscale intensity Standard deviation of the saturation value in HSV color space Standard deviation of the A value in LAB color space Standard deviation of the B value in LAB color space
Stain Intensity	Mean greyscale intensity Mean saturation value in HSV color space Mean A value in LAB color space Mean B value in LAB color space

Figure 1. Machine learning model annotation collection, training, and application. A) Model workflow. Briefly, pathologists trained expert annotators to perform exhaustive annotations of nuclei on H&E slide patches from diverse tissue sources. These were used to train a pan-H&E nucleus detection and segmentation model, which was subsequently evaluated on held-out patches and applied to exhaustively segment nuclei in three WSI datasets. B) Features extracted from the model. Mean and standard deviation values were calculated for these features at the whole-slide level for cancer cells, lymphocytes, and fibroblasts.

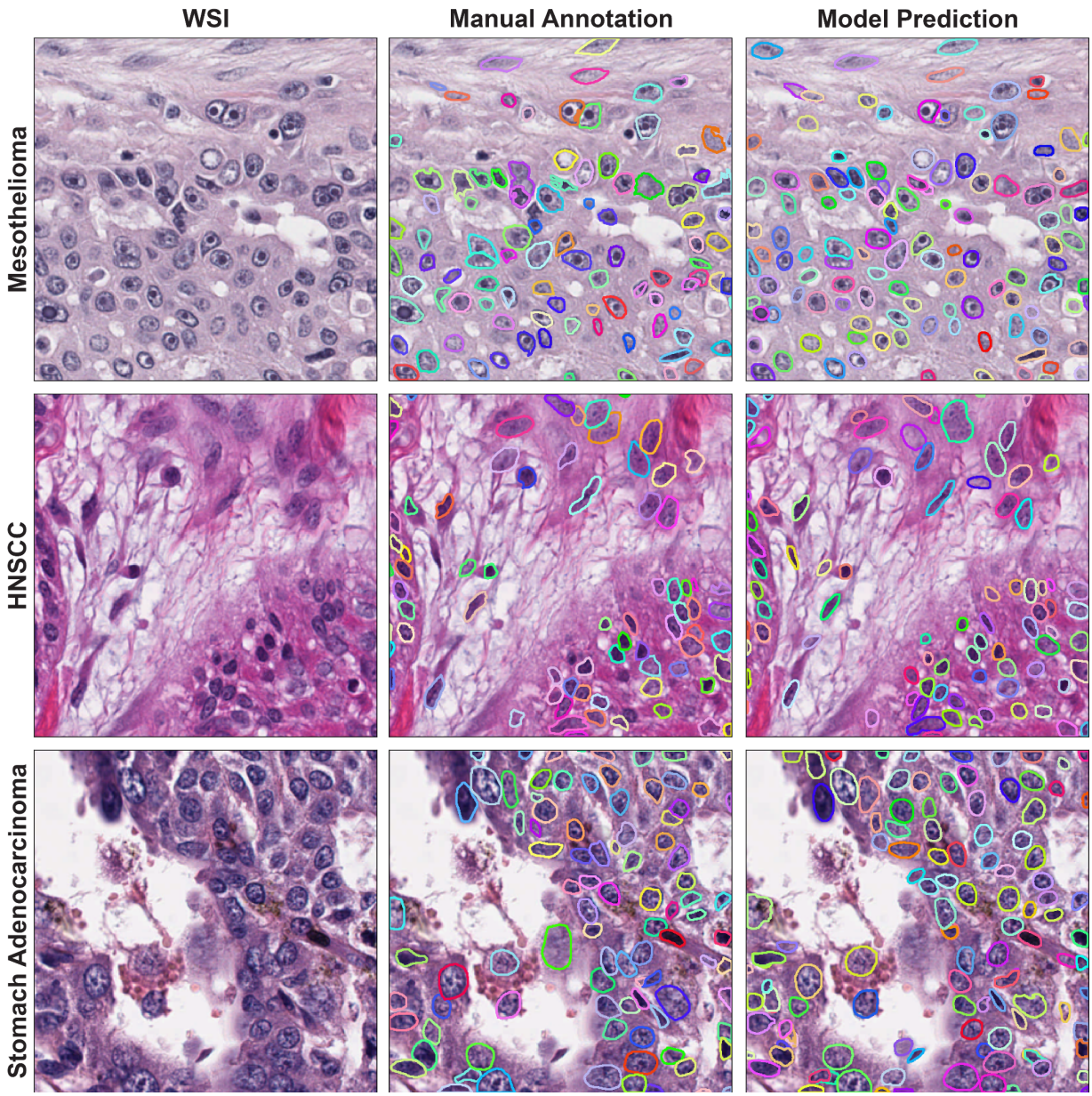


Figure 2. Example of model performance. A) Representative whole slide image patches from mesothelioma, head and neck squamous cell carcinoma (HNSCC), and stomach adenocarcinoma stained with H&E. B) Ground truth nuclei identified manually. C) Nuclei predicted by model. Each color represents a nucleus instance.

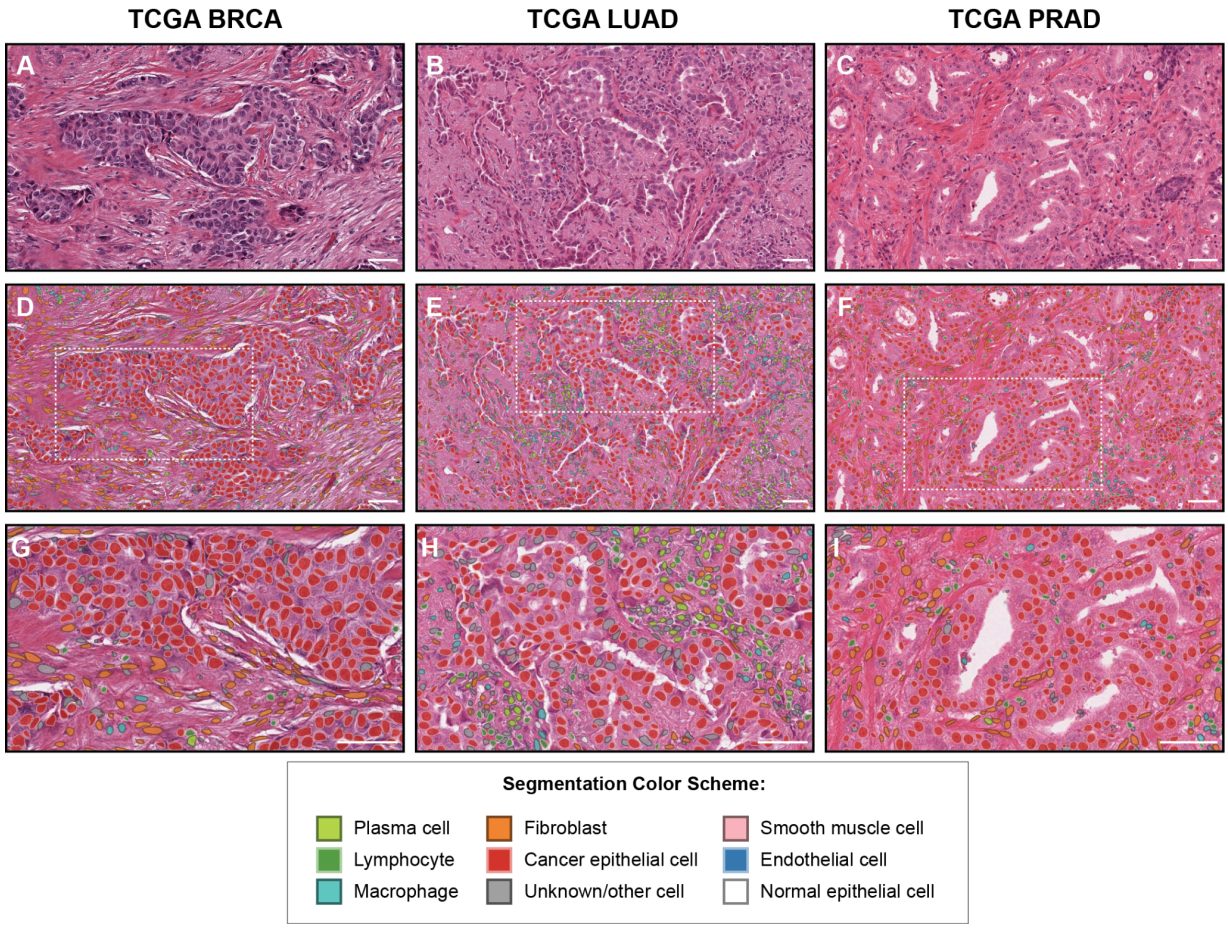


Figure 3. Nuclear segmentation and cell type identification in multiple cancer types. Representative H&E images of A) breast cancer (TCGA BRCA), B) lung adenocarcinoma (TCGA LUAD), and C) prostate adenocarcinoma (TCGA PRAD) are shown at 40X magnification. D-F) Nuclear segmentation and cell type identification masks are overlaid onto H&E images shown in panels A-C. H-I) High-magnification images of BRCA (G), LUAD (H), and PRAD (I). Magnified regions are indicated by dashed boxes in panels D-F. Scale bars indicate a distance of 50 μm.

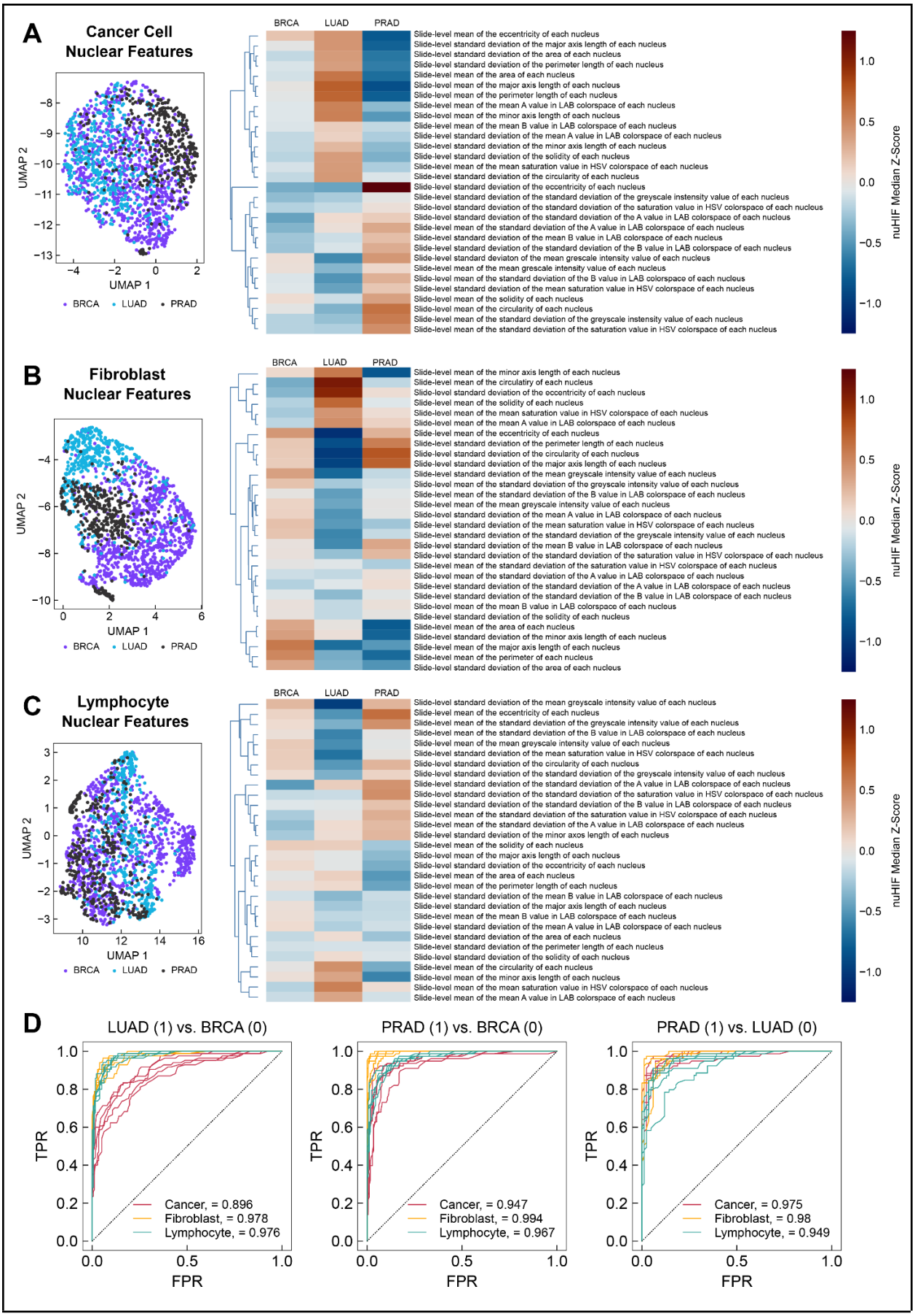

A
Cancer Cell Nuclear Features
UMAP 2
UMAP 1
BRCA
LUAD
PRAD
Slide-level mean of the eccentricity of each nucleus
Slide-level standard deviation of the major axis length of each nucleus
Slide-level standard deviation of the area of each nucleus
Slide-level standard deviation of the perimeter length of each nucleus
Slide-level mean of the area of each nucleus
Slide-level mean of the major axis length of each nucleus
Slide-level mean of the perimeter length of each nucleus
Slide-level mean of the mean A value in LAB colorspace of each nucleus
Slide-level mean of the minor axis length of each nucleus
Slide-level mean of the mean B value in LAB colorspace of each nucleus
Slide-level standard deviation of the mean A value in LAB colorspace of each nucleus
Slide-level standard deviation of the minor axis length of each nucleus
Slide-level standard deviation of the solidity of each nucleus
Slide-level mean of the mean saturation value in HSV colorspace of each nucleus
Slide-level standard deviation of the circularity of each nucleus
Slide-level standard deviation of the eccentricity of each nucleus
Slide-level standard deviation of the standard deviation of the greyscale instensity value of each nucleus
Slide-level standard deviation of the standard deviation of the saturation value in HSV colorspace of each nucleus
Slide-level standard deviation of the standard deviation of the A value in LAB colorspace of each nucleus
Slide-level mean of the standard deviation of the A value in LAB colorspace of each nucleus
Slide-level standard deviation of the mean B value in LAB colorspace of each nucleus
Slide-level standard deviation of the standard deviation of the B value in LAB colorspace of each nucleus
Slide-level standard deviaton of the mean grescale intensity value of each nucleus
Slide-level mean of the mean grescale intensity value of each nucleus
Slide-level mean of the standard deviation of the B value in LAB colorspace of each nucleus
Slide-level standard deviation of the mean saturation value in HSV colorspace of each nucleus
Slide-level mean of the solidity of each nucleus
Slide-level mean of the circularity of each nucleus
Slide-level mean of the standard deviation of the greyscale instensity value of each nucleus
Slide-level mean of the standard deviation of the saturation value in HSV colorspace of each nucleus
nuHIF Median Z-Score
B
Fibroblast Nuclear Features
Slide-level mean of the minor axis length of each nucleus
Slide-level mean of the circulatiry of each nucleus
Slide-level standard deviation of the eccentricity of each nucleus
Slide-level mean of the solidity of each nucleus
Slide-level mean of the mean saturation value in HSV colorspace of each nucleus
Slide-level mean of the mean A value in LAB colorspace of each nucleus
Slide-level mean of the eccentricity of each nucleus
Slide-level standard deviation of the perimeter length of each nucleus
Slide-level standard deviation of the circularity of each nucleus
Slide-level standard deviation of the major axis length of each nucleus
Slide-level standard deviation of the mean greyscale intensity value of each nucleus
Slide-level mean of the standard deviation of the greyscale intensity value of each nucleus
Slide-level mean of the standard deviation of the B value in LAB colorspace of each nucleus
Slide-level mean of the mean greyscale intensity value of each nucleus
Slide-level standard deviation of the mean A value in LAB colorspace of each nucleus
Slide-level standard deviation of the mean saturation value in HSV colorspace of each nucleus
Slide-level standard deviation of the standard deviation of the greyscale intensity value of each nucleus
Slide-level standard deviation of the mean B value in LAB colorspace of each nucleus
Slide-level standard deviation of the standard deviation of the saturation value in HSV colorspace of each nucleus
Slide-level mean of the standard deviation of the saturation value in HSV colorspace of each nucleus
Slide-level mean of the standard deviation of the A value in LAB colorspace of each nucleus
Slide-level standard deviation of the standard deviation of the A value in LAB colorspace of each nucleus
Slide-level standard deviation of the standard deviation of the B value in LAB colorspace of each nucleus
Slide-level mean of the mean B value in LAB colorspace of each nucleus
Slide-level standard deviation of the solidity of each nucleus
Slide-level mean of the area of each nucleus
Slide-level standard deviation of the minor axis length of each nucleus
Slide-level mean of the major axis length of each nucleus
Slide-level mean of the perimeter of each nucleus
Slide-level standard deviation of the area of each nucleus
C
Lymphocyte Nuclear Features
Slide-level standard deviation of the mean greyscale intensity value of each nucleus
Slide-level mean of the eccentricity of each nucleus
Slide-level mean of the standard deviation of the greyscale intensity value of each nucleus
Slide-level mean of the standard deviation of the B value in LAB colorspace of each nucleus
Slide-level mean of the mean greyscale intensity value of each nucleus
Slide-level standard deviation of the mean saturation value in HSV colorspace of each nucleus
Slide-level standard deviation of the circularity of each nucleus
Slide-level standard deviation of the standard deviation of the greyscale intensity value of each nucleus
Slide-level standard deviation of the standard deviation of the A value in LAB colorspace of each nucleus
Slide-level standard deviation of the standard deviation of the saturation value in HSV colorspace of each nucleus
Slide-level standard deviation of the standard deviation of the B value in LAB colorspace of each nucleus
Slide-level mean of the standard deviation of the saturation value in HSV colorspace of each nucleus
Slide-level mean of the standard deviation of the A value in LAB colorspace of each nucleus
Slide-level standard deviation of the minor axos length of each nucleus
Slide-level mean of the solidity of each nucleus
Slide-level mean of the major axis length of each nucleus
Slide-level standard deviation of the eccentricity of each nucleus
Slide-level mean of the area of each nucleus
Slide-level mean of the perimeter length of each nucleus
Slide-level standard deviation of the mean B value in LAB colorspace of each nucleus
Slide-level standard deviation of the major axis length of each nucleus
Slide-level mean of the mean B value in LAB colorspace of each nucleus
Slide-level standard deviation of the mean A value in LAB colorspace of each nucleus
Slide-level standard deviation of the area of each nucleus
Slide-level standard deviation of the perimeter length of each nucleus
Slide-level standard deviation of the solidity of each nucleus
Slide-level mean of the circularity of each nucleus
Slide-level mean of the minor axis length of each nucleus
Slide-level mean of the mean saturation value in HSV colorspace of each nucleus
Slide-level mean of the mean A value in LAB colorspace of each nucleus
D
LUAD (1) vs. BRCA (0)
Cancer, = 0.896
Fibroblast, = 0.978
Lymphocyte, = 0.976
PRAD (1) vs. BRCA (0)
Cancer, = 0.947
Fibroblast, = 0.994
Lymphocyte, = 0.967
PRAD (1) vs. LUAD (0)
Cancer, = 0.975
Fibroblast, = 0.98
Lymphocyte, = 0.949
TPR
FPR

Figure 4. nuHIFs show variation within and between cancer types. Uniform manifold approximation and projection (UMAP) visualization of BRCA, LUAD, and PRAD defined by nuclear human interpretable feature (HIF) for A) cancer cells, B) CAFs, and C) tumor-infiltrating lymphocytes. Clustered heatmaps of median Z-scores for all 30 nuHIFs are shown for each cell type. D) Receiver operating characteristic (ROC) curves for binary classification between paired cancer types using nuHIFs from each of cancer, CAF, or lymphocyte nuclei. ROCs are shown for the five held-out validation splits and mean area under ROC (AUROC) is shown for each classification problem. In particular, CAF and lymphocyte nuclear features are highly able to differentiate between cancer types. Mean AUROC is shown for each class of nuclear HIF.

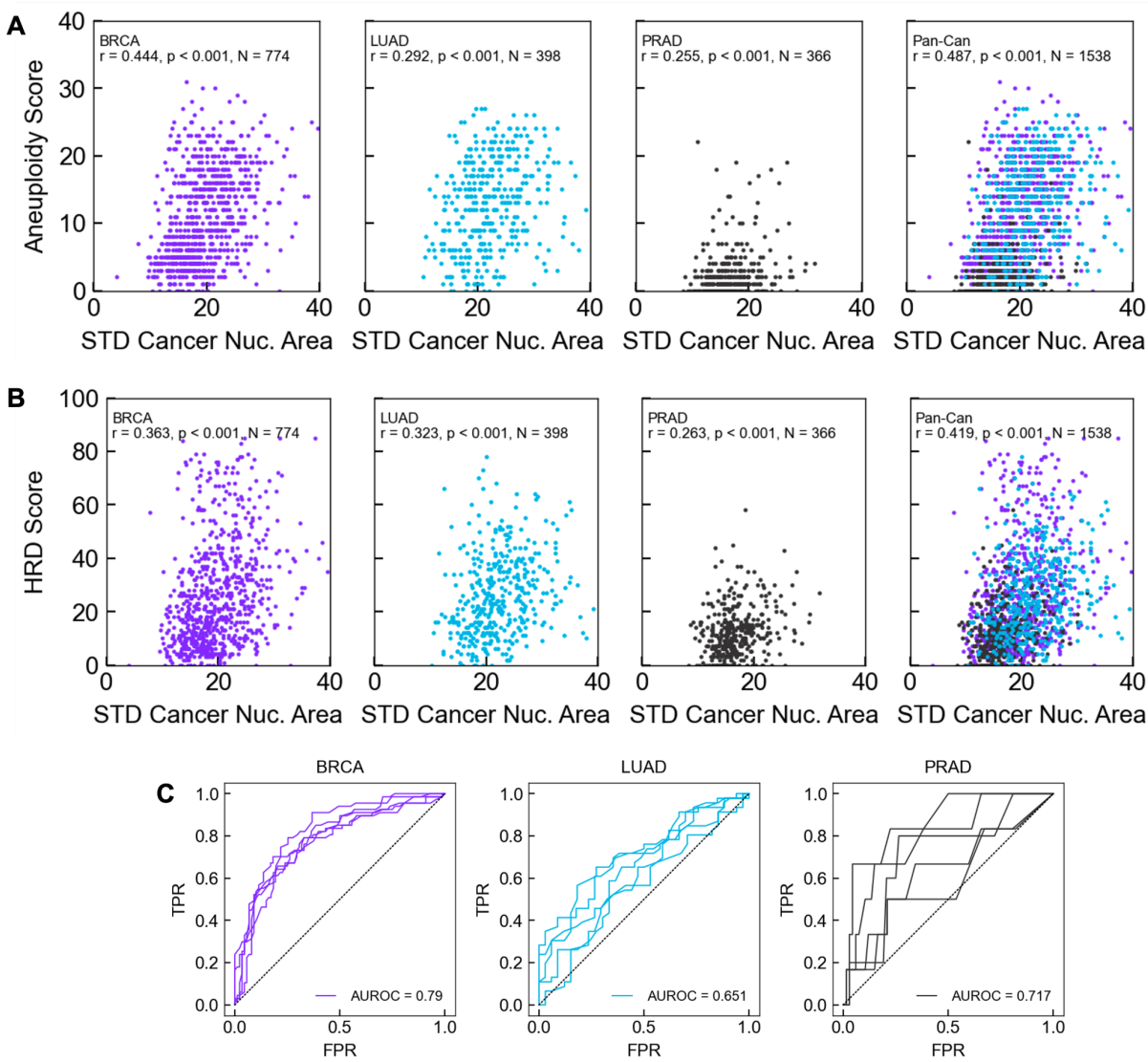


Figure 5. Variation in cancer nuclear size correlates with metrics of genomic instability. Standard deviation of cancer cell nuclear area was compared to A) aneuploidy score and B) homologous recombination deficiency (HRD) score for BRCA, LUAD, and PRAD. C) Receiver operating characteristic (ROC) curves for prediction of whole-genome doublings in BRCA, LUAD, and PRAD. ROCs are shown for the five held-out validation splits; mean AUROC is shown for each cancer type.

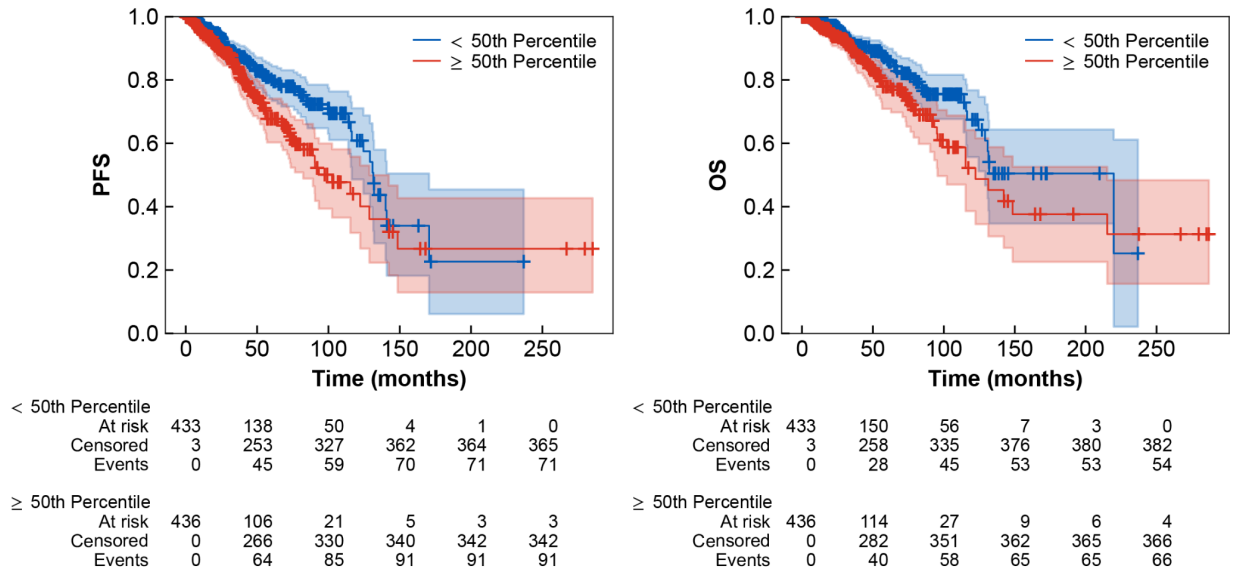


Figure 6. Association between CAF nuclear area and survival in breast cancer. Increased CAF nuclear area (≥ 50th percentile) corresponds to poor PFS (HR = 1.63, 95% CI [1.19-2.23], p = 0.002) and OS (HR = 1.53, 95% CI [1.06, 2.20], p = 0.022).

References

1. Zink D, Fischer AH, Nickerson JA. Nuclear structure in cancer cells. Nat Rev Cancer. 2004;4:677–87.

2. Fischer EG. Nuclear Morphology and the Biology of Cancer Cells. Acta Cytol. 2020;64:511–9.

3. Cree IA, Tan PH, Travis WD, Wesseling P, Yagi Y, White VA, et al. Counting mitoses: SI(ze) matters! Mod Pathol. 2021;34:1651–7.

4. Hapke MR, Dehner LP. The optically clear nucleus. A reliable sign of papillary carcinoma of the thyroid? Am J Surg Pathol. 1979;3:31–8.

5. Chow K-H, Factor RE, Ullman KS. The nuclear envelope environment and its cancer connections. Nat Rev Cancer. 2012;12:196–209.

6. Shah P, Wolf K, Lammerding J. Bursting the bubble--nuclear envelope rupture as a path to genomic instability? Trends Cell Biol. Elsevier; 2017;27:546–55.

7. Zuleger N, Robson MI, Schirmer EC. The nuclear envelope as a chromatin organizer. Nucleus. 2011;2:339–49.

8. Stenkvist BO, Westman-Naeser 2. Sighild, Holmquist J, Nordin B, Bengtsson E, Vegelius J. Computerized Nuclear Morphometry as an Objective Method for Characterizing Human Cancer Cell Populations1.

9. Baak JPA, Van Dop PH, Kurver t. PETH, Hermans JO. The Value of Morphometry to Classic Prognosticators in Breast Cancer. Available from: http://dx.doi.org/10.1002/1097-0142(19850715)56:2

10. Pienta KJ, Coffey DS. Correlation of Nuclear Morphometry With Progression of Breast Cancer. Available from: http://dx.doi.org/10.1002/1097-0142(19911101)68:9

11. Partin AW, Steinberg GD, Pitcock RV, Wu L, Piantadosi t. S, Cofey t. DS, et al. Use of Nuclear Morphometry, Gleason Histologic Scoring, Clinical Stage, and Age to Predict Disease-Free Survival Among Patients With Prostate Cancer. Available from: http://dx.doi.org/10.1002/1097-0142(19920701)70:1

12. Kong J, Cooper LAD, Wang F, Gao J, Teodoro G, Scarpace L, et al. Machine-based morphologic analysis of glioblastoma using whole-slide pathology images uncovers clinically relevant molecular correlates. PLoS One. 2013;8:e81049.

13. Kumar N, Verma R, Chen C, Lu C, Fu P, Willis J, et al. Computer-extracted features of nuclear morphology in hematoxylin and eosin images distinguish stage II and IV colon tumors. J Pathol. 2022;257:17–28.

14. Lewis JS Jr, Ali S, Luo J, Thorstad WL, Madabhushi A. A quantitative

histomorphometric classifier (QuHbIC) identifies aggressive versus indolent p16-positive oropharyngeal squamous cell carcinoma. Am J Surg Pathol. 2014;38:128–37.

15. Tokuyama N, Saito A, Muraoka R, Matsubara S, Hashimoto T, Satake N, et al. Prediction of non-muscle invasive bladder cancer recurrence using machine learning of quantitative nuclear features. Mod Pathol [Internet]. 2021; Available from: http://dx.doi.org/10.1038/s41379-021-00955-y

16. Ji M-Y, Yuan L, Jiang X-D, Zeng Z, Zhan N, Huang P-X, et al. Nuclear shape, architecture and orientation features from H&E images are able to predict recurrence in node-negative gastric adenocarcinoma. J Transl Med. 2019;17:92.

17. Lee G, Veltri RW, Zhu G, Ali S, Epstein JI, Madabhushi A. Nuclear Shape and Architecture in Benign Fields Predict Biochemical Recurrence in Prostate Cancer Patients Following Radical Prostatectomy: Preliminary Findings. Eur Urol Focus. 2017;3:457–66.

18. Yu K-H, Zhang C, Berry GJ, Altman RB, Ré C, Rubin DL, et al. Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. Nat Commun. 2016;7:12474.

19. Cancer Genome Atlas Research Network. Comprehensive and Integrated Genomic Characterization of Adult Soft Tissue Sarcomas. Cell. 2017;171:950–65.e28.

20. Fu Y, Jung AW, Torne RV, Gonzalez S, Vöhringer H, Shmatko A, et al. Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis. Nature Cancer. Nature Publishing Group; 2020;1:800–10.

21. Rawat RR, Ruderman D, Macklin P, Rimm DL, Agus DB. Correlating nuclear morphometric patterns with estrogen receptor status in breast cancer pathologic specimens. NPJ Breast Cancer. 2018;4:32.

22. Whitney J, Corredor G, Janowczyk A, Ganesan S, Doyle S, Tomaszewski J, et al. Quantitative nuclear histomorphometry predicts oncotype DX risk categories for early stage ER+ breast cancer. BMC Cancer. 2018;18:610.

23. Li H, Whitney J, Bera K, Gilmore H, Thorat MA, Badve S, et al. Quantitative nuclear histomorphometric features are predictive of Oncotype DX risk categories in ductal carcinoma in situ: preliminary findings. Breast Cancer Res. 2019;21:114.

24. Nimgaonkar V, Krishna V, Krishna V, Tiu E, Joshi A, Vrabac D, et al. Development of an artificial intelligence-derived histologic signature associated with adjuvant gemcitabine treatment outcomes in pancreatic cancer. Cell Rep Med [Internet]. Elsevier; 2023 [cited 2023 Apr 12];0. Available from: https://www.cell.com/cell-reports-medicine/fulltext/S2666-3791(23)00123-4?_returnURL=https%3A%2F%2Flinkinghub.elsevier.com%2Fretrieve%2Fpii%2FS2666379123001234%3Fshowall%3Dtrue

25. Joshi RP, Osinski BL, Beig N, Sha L, Ingale K, Stumpe MC. Imaging-based histological features are predictive of MET alterations in Non-Small Cell Lung Cancer [Internet]. arXiv [cs.CV]. 2022. Available from: http://arxiv.org/abs/2203.10062

26. 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;12:1613.

27. Gutman DA, Cobb J, Somanna D, Park Y, Wang F, Kurc T, et al. Cancer Digital Slide Archive: an informatics resource to support integrated in silico analysis of TCGA pathology data. J Am Med Inform Assoc. 2013;20:1091–8.

28. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: An imperative style, high-performance deep learning library [Internet]. arXiv [cs.LG]. 2019 [cited 2023 Jan 6]. Available from: https://proceedings.neurips.cc/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html

29. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python [Internet]. arXiv [cs.LG]. 2012 [cited 2023 Jan 6]. page 2825–30. Available from: https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf?ref=https://githubhelp.com

30. Taylor AM, Shih J, Ha G, Gao GF, Zhang X, Berger AC, et al. Genomic and Functional Approaches to Understanding Cancer Aneuploidy. Cancer Cell. 2018;33:676–89.e3.

31. Marquard AM, Eklund AC, Joshi T, Krzystanek M, Favero F, Wang ZC, et al. Pan-cancer analysis of genomic scar signatures associated with homologous recombination deficiency suggests novel indications for existing cancer drugs. Biomark Res. 2015;3:9.

32. Davidson-Pilon C. lifelines: survival analysis in Python. J Open Source Softw. The Open Journal; 2019;4:1317.

33. Goldman MJ, Craft B, Hastie M, Repečka K, McDade F, Kamath A, et al. Visualizing and interpreting cancer genomics data via the Xena platform. Nat Biotechnol. 2020;38:675–8.

34. 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 U S A. 2005;102:15545–50.

35. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0. Bioinformatics. 2011;27:1739–40.

36. Gillespie M, Jassal B, Stephan R, Milacic M, Rothfels K, Senff-Ribeiro A, et al. The reactome pathway knowledgebase 2022. Nucleic Acids Res. 2022;50:D687–92.

37. 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;39:1380–91.

38. Graham S, Vu QD, Raza SEA, Azam A, Tsang YW, Kwak JT, et al. Hover-Net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. Med Image Anal. 2019;58:101563.

39. Friedman G, Levi-Galibov O, David E, Bornstein C, Giladi A, Dadiani M, et al. Cancer-associated fibroblast compositions change with breast cancer progression linking the ratio of S100A4+ and PDPN+ CAFs to clinical outcome. Nat Cancer. 2020;1:692–708.

40. Becker LM, O'Connell JT, Vo AP, Cain MP, Tampe D, Bizarro L, et al. Epigenetic Reprogramming of Cancer-Associated Fibroblasts Deregulates Glucose Metabolism and Facilitates Progression of Breast Cancer [Internet]. Cell Reports. 2020. page 107701. Available from: http://dx.doi.org/10.1016/j.celrep.2020.107701

41. Suh J, Kim D, Lee Y, Jang J, Surh Y. Fibroblast growth factor‐2, derived from cancer‐associated fibroblasts, stimulates growth and progression of human breast cancer cells via FGFR1 signaling [Internet]. Molecular Carcinogenesis. 2020. page 1028–40. Available from: http://dx.doi.org/10.1002/mc.23233

42. Houthuijzen JM, Jonkers J. Cancer-associated fibroblasts as key regulators of the breast cancer tumor microenvironment. Cancer Metastasis Rev. 2018;37:577–97.

43. Ren J, Smid M, Iaria J, Salvatori DCF, van Dam H, Zhu HJ, et al. Cancer-associated fibroblast-derived Gremlin 1 promotes breast cancer progression. Breast Cancer Res. 2019;21:109.

44. Pelon F, Bourachot B, Kieffer Y, Magagna I, Mermet-Meillon F, Bonnet I, et al. Cancer-associated fibroblast heterogeneity in axillary lymph nodes drives metastases in breast cancer through complementary mechanisms [Internet]. Nature Communications. 2020. Available from: http://dx.doi.org/10.1038/s41467-019-14134-w

45. Costa A, Kieffer Y, Scholer-Dahirel A, Pelon F, Bourachot B, Cardon M, et al. Fibroblast Heterogeneity and Immunosuppressive Environment in Human Breast Cancer. Cancer Cell. 2018;33:463–79.e10.

46. Rosen MN, Goodwin RA, Vickers MM. BRCA mutated pancreatic cancer: A change is coming. World J Gastroenterol. 2021;27:1943–58.

47. Michener CM, Kirkup C, Rahsepar B, Iyer JS, Abel J, Leidal K, et al. 593P AI-powered analysis of nuclear morphology associated with prognosis in high-grade serous carcinoma. Ann Oncol. Elsevier; 2022;33:S818.