



Published in final edited form as:

Cancer Res. 2022 August 03; 82(15): 2792–2806. doi:10.1158/0008-5472.CAN-21-2318.

Intratumoral resolution of driver gene mutation heterogeneity in renal cancer using deep learning

Paul H. Acosta¹, Vandana Panwar², Vipul Jarmale¹, Alana Christie³, Jay Jasti¹, Vitaly Margulis^{3,4}, Dinesh Rakheja², John Cheville⁵, Bradley C. Leibovich⁶, Alexander Parker⁷, James Brugarolas^{3,8}, Payal Kapur^{2,3,4,#}, Satwik Rajaram^{1,2,3,#}

¹Lyda Hill Department of Bioinformatics, University of Texas Southwestern Medical Center, Dallas, TX, USA

²Department of Pathology, University of Texas Southwestern Medical Center, Dallas, TX, USA

³Kidney Cancer Program, Simmons Comprehensive Cancer Center, University of Texas Southwestern Medical Center, Dallas, TX, USA

⁴Department of Urology, University of Texas Southwestern Medical Center at Dallas, Dallas, TX, USA.

⁵Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, MN, USA.

⁶Department of Urology, Mayo Medical School and Mayo Clinic, Rochester, MN, USA.

⁷Office of Research Affairs, University of Florida, College of Medicine, Jacksonville, FL, USA.

⁸Department of Internal Medicine (Hematology-Oncology), University of Texas Southwestern Medical Center, Dallas, TX, USA

Abstract

Intratumoral heterogeneity arising from tumor evolution poses significant challenges biologically and clinically. Dissecting this complexity may benefit from deep learning (DL) algorithms, which can infer molecular features from ubiquitous H&E stained tissue sections. While DL algorithms

#Corresponding Authors: 1) Payal Kapur, MD, Department of Pathology, University of Texas Southwestern Medical Center, 5323 Harry Hines Boulevard, Dallas, TX 75390, Ph: 214.590.6592, payal.kapur@utsouthwestern.edu 2) Satwik Rajaram, PhD, Lyda Hill Department of Bioinformatics, University of Texas Southwestern Medical Center, 5323 Harry Hines Boulevard, Dallas, TX - 75390-9365, Ph: 214-645-1727, satwik.rajaram@utsouthwestern.edu.

Authors' Contributions P. Acosta: Conceptualization, data curation, software, formal analysis, validation, investigation, visualization, methodology, writing—original draft, writing—review and editing. V. Panwar: Data curation, validation, investigation, methodology, writing—review and editing. V. Jarmale: Software, formal analysis, validation, visualization, methodology. A. Christie: Data curation, formal analysis, methodology, writing—review and editing. J. Jasti: Data curation, software, validation, methodology. V. Margulis: Resources, data curation, writing—review and editing. D. Rakheja: Resources, data curation, investigation, writing—review and editing. J. Cheville: Resources, data curation, investigation, writing—review and editing. B.C. Leibovich: Resources, data curation, investigation. A. Parker: Resources, data curation, investigation, writing—review and editing. J. Brugarolas: Conceptualization, resources, data curation, funding acquisition, investigation, methodology, project administration, writing—review and editing. P. Kapur: Conceptualization, resources, data curation, formal analysis, supervision, funding acquisition, validation, investigation, methodology, writing—original draft, project administration, writing—review and editing. S. Rajaram: Conceptualization, resources, data curation, software, formal analysis, supervision, funding acquisition, validation, investigation, visualization, methodology, writing—original draft, project administration, writing—review and editing.

Authors' Disclosures

J. Brugarolas reports personal fees from Eisai, Johnson & Johnson, Exelixis, Calithera, and Arrowhead outside the submitted work. No disclosures were reported by the other authors.

Conflict of Interest Disclosure Statement: The authors declare no potential conflicts of interest.

have been developed to predict some driver mutations from H&E images, the ability of these DL algorithms to resolve intratumoral mutation heterogeneity at sub-clonal spatial resolution is unexplored. Here, we apply DL to a paradigm of intratumoral heterogeneity, clear-cell renal cell carcinoma (ccRCC), the most common type of kidney cancer. Matched immunohistochemistry and H&E images were leveraged to develop DL models for predicting intratumoral genetic heterogeneity of the three most frequently mutated ccRCC genes, *BAP1*, *PBRM1*, and *SETD2*. DL models were generated on a large cohort (N=1282) and tested on several independent cohorts, including a TCGA cohort (N=363 patients) and two tissue microarray (TMA) cohorts (N=118 and 365 patients). These models were also expanded to a patient-derived xenograft (PDX) tissue microarray, affording analysis of homotopic and heterotopic interactions of tumor and stroma. The status of all three genes could be inferred by DL, with BAP1 showing the highest sensitivity and performance within and across tissue samples (AUC=0.87-0.89 on holdout). BAP1 results were validated on independent human (AUC=0.77-0.84) and PDX (AUC=0.80) cohorts. Finally, BAP1 predictions correlated with clinical outputs such as disease-specific survival. Overall, these data show that DL models can resolve intratumoral heterogeneity in cancer with potential diagnostic, prognostic, and biological implications.

Introduction

Intra-tumor heterogeneity is a primary challenge in cancer. Indeed, metastasis and treatment resistance are often driven by mutations that are present in only a subpopulation of tumor cells(1-3) and may be missed by traditional bulk-sequencing methods. Multi-region sequencing studies spanning different tumor types(4-6) show that multiple areas of a tumor would need to be sequenced to reliably capture all “driver” events, likely rendering such an approach impractical in the clinic. In contrast, morphological tumor heterogeneity is far easier to observe in the ubiquitous hematoxylin and eosin (H&E) stained images that are a staple of routine pathologic diagnosis. Surprisingly, the relationship between heterogeneity observed in genetics and morphology – two pillars of modern oncology – is poorly understood, and the ability of H&E images to aid the deconvolution of genetic heterogeneity remains little studied. Moreover, recent studies suggest that morphology is not simply a rich downstream readout of signaling but may be contributing to it (7,8). Thus, there are undoubtedly mechanistic insights and associated vulnerabilities to be uncovered by unraveling the link between morphologic phenotypes and the underlying genetics.

Recent approaches using Deep Learning (DL) to predict tumor genetics from H&E images have shown great promise. For example, DL models, both from pan-cancer(9,10) and organ-specific cancer studies(11,12), predicted the likelihood of genetic mutations in various cancer types including lung, colorectal, breast and kidney cancer. Similarly, DL models categorized important biomarkers like estrogen receptor status in breast cancer(13), microsatellite instability in gastrointestinal cancer(14), and predicted associations with molecular transcriptomics in various cancer types(9,10). Although the accuracy of these predictions does not necessarily reach practical utility, they show an association (based on high AUCs) between mutational events and specific morphological phenotypes, which we define here as any histologic property evaluated from H&E-stained formalin-fixed paraffin-embedded (FFPE) tissue sections on glass slides. Other relevant studies have shown that DL models

can be used to investigate and quantify tumor heterogeneity in terms of cell types(15), tissue architectures(16), molecular subtypes and expression of specific (typically cell-type specific) genes(17).

Despite these advances, the ability of DL models to unravel genetic intra-tumor heterogeneity – a powerful, yet more complex application – is unexplored. Such investigation is hindered by the common use of modalities like bulk sequencing, which do not provide spatial assignment for genetic information. Additionally, if DL did discover any genotype-morphology connections, it remains unclear how one would probe the underlying mechanisms. Ideally, one would study these relationships in model systems like patient derived xenografts (PDX) which are more amenable to molecular manipulation. However, it is uncertain whether relationships identified in human patient samples survive the extensive perturbation (including tissue disassociation and replacement of microenvironment) needed to establish PDX models(18).

Motivated by these challenges, we investigate the link between genetic and morphological heterogeneity in the context of clear-cell renal cell carcinoma (ccRCC), the most common type of kidney cancer. We chose ccRCC as our model system for two reasons. First, ccRCC is a paradigm for intra-tumor heterogeneity, both in terms of genetics and morphology. With respect to genetics, the seminal multi-region sequencing experiments unequivocally connecting genetic heterogeneity to tumor evolution were performed in ccRCC(19). More recent studies have reinforced the widespread genetic heterogeneity in ccRCC: to capture 75% of variants(4), seven randomly sampled areas are needed on average. Similarly, concerning morphology, it has been well documented that individual ccRCC tumors frequently exhibit areas showing different tissue architectures, cytological states and microenvironmental features(20-24). Thus, even a single slide may contain multiple “clones” with different states of driver genes or morphologic phenotypes (Supplementary Fig. 1a-c). Recent DL studies analyzing data from The Cancer Genome Atlas (TCGA) have demonstrated the possibility of predicting specific driver mutations from H&E images(9-11), but the ability to use such analyses to predict intra-tumor heterogeneity is unexplored. Moreover, these studies failed to detect a relationship between morphology and mutations in the driver gene *BAP1*, whose impact on tumor grade (which in ccRCC is primarily based on nuclear morphology) has been previously established(25-27). The second reason for choosing ccRCC is that the main driver genes are two-hit tumor suppressors, and we can use loss of protein expression, which can be resolved at the cellular level using immunohistochemistry (IHC) assays, as a proxy for gene status. Specifically, we use validated IHC assays (with positive and negative predictive values >98%(26,28,29)) to ascertain genetic heterogeneity (both for training and validation) at much higher spatial resolution than is possible using bulk sequencing. Importantly, these IHC assays likely capture other mechanisms of gene inactivation besides intragenic mutation, such as gene silencing (*BAP1/PBRM1*), and in the case of SETD2 focus on loss of activity (i.e. histone methylation)(26). In this study, IHC assays will be used as surrogates for “gene loss.”

Leveraging these features of our ccRCC model system, we present the first investigation of H&E-based DL models to: a) test genetic state predictions at a regional intra-tumor scale; and b) recapitulate these predictions in PDX models. Specifically, we focus on the three

most common driver genes in ccRCC that we previously showed impact prognosis: BAP1, PBRM1, and SETD2(4,26,30-32). We attempt to predict the status of these genes from tissue morphology. For each gene, we trained models at two scales: a) whole-slide level - to parallel previous bulk-sequencing approaches, we predict if the gene is lost anywhere in a slide; and b) region level - to resolve intra-tumor heterogeneity, we predict gene loss status of local regions within slides. A defining aspect of this study is the diversity of cohorts used, which allowed us to test the robustness of our models to an unprecedented degree. For training and internal testing of both our region and slide level models, we use a large well annotated cohort(28,29,33,34) (WSI) collected at the Mayo Clinic. For external validation of the slide level models we use data from TCGA and for the region level models we employ two patient tissue microarray(25,26,35-39) (TMA) cohorts collected at UTSW (TMA1 and TMA2). Additionally, to interrogate whether the morphological phenotypes are cell-intrinsic or driven by interactions with stroma we test our predictions on a patient derived xenograft TMA cohort(26,40) (PDX1). As described below, we were able to predict gene loss status, specifically that of BAP1, across these cohorts. Our results suggest that DL based prediction of gene status from histopathological slides has the robustness and spatial resolution to serve as a low-cost in-silico method to resolve intratumoral heterogeneity in cancer.

Methods

The study was conducted with approval from the UT Southwestern Medical Center (UTSW) Institutional Review Board (STU 022013-052) and in concordance with the Health Insurance Portability and Accountability Act (HIPAA) guidelines.

Assays:

This study makes use of tissue slides stained using: 1) H&E as a readout of morphology and 2) IHC as a ground truth readout of genetic status. Our IHC staining made use of previously developed(26) antibodies against BAP1(26,28), PBRM1(26) and H3K36me3(29) which have been validated and accurately capture the functional status of BAP1, PBRM1 and SETD2 respectively. Note that although, for simplicity, we refer to SETD2 loss, this is in fact based on an assessment of H3K36me3, a histone modification catalyzed by SETD2.

IHC-based Definition of Genetics Ground Truth:

IHC (for all three proteins) and H&E staining were performed on proximal tissue sections in the WSI, TMA1, TMA2 and PDX1 cohorts. Determination of loss status made by pathologists on tumor regions of IHC slides were manually mapped to the corresponding H&E slides. Specifically, for each gene, whole slides were classified as either a) wildtype (WT) if the entire tumor area displayed wildtype staining, b) Universal Loss (UL) if the entire tumor area displayed loss-like staining or c) Localized Loss (LL) if the tumor region exhibited a mix of WT and loss-like staining (Supplementary Fig. 2a). Tissue micro-array punches in contrast were simply classified as WT or Loss as little heterogeneity was observed in this limited area. For patient level scoring in the TMAs, if any of the punches for a patient displayed gene loss, the patient was considered to exhibit gene loss.

Datasets:

In this study, we made use of four of our previously published(25,26,28,29,33-40) datasets with matched H&E and IHC slides and the TCGA KIRC dataset(41) with matching H&E slides and slide-level mutation calls from cBioPortal(42,43). Unless explicitly noted, we only analyzed the H&E images from those studies, while using the previously generated pathologist (P.K.) scoring of the IHC slides (WSI and TMA cohorts) and cBioPortal mutation calls (TCGA) as ground truth. All five cohorts consisted solely of FFPE samples, which are standard in clinical practice and better preserve tissue morphology than frozen sections. Our four cohorts (WSI, TMA1, TMA2, PDX1) with matching IHCs were imaged using Aperio scanners. WSI and TMA1 were imaged at 20X magnification, while PDX1, TMA2 and TCGA, were imaged at 40X (but were computationally down-sampled to 20X for analysis).

1. WSI: This cohort was used primarily as training data for our models. It consists of 1362 H&E-stained whole slide images of ccRCC tumors, each representing a single patient treated at the Mayo Clinic(34). As described above, based on previous IHC staining, slides were classified as WT, UL or LL. Additionally, in the case of the 19 BAP1 Localized loss cases, a pathologist (V.P.) manually compared matching IHC and H&E staining and manually annotated the BAP1 Loss areas within the tumor regions of the slides. For quality control purposes, we excluded samples where H&E images were corrupted, our tumor model did not detect enough tumor area for patch generation (see below) or IHC status for any of the three genes was unreliable. Of the 1362 whole slide images, 1282 were used in this study.
2. TCGA: This cohort was used primarily as validation for our slide-level model. It consists of 383 H&E stained whole slide images of ccRCC tumors, with each representing a single patient with genetic sequencing data from the Pan Cancer Atlas study(41). The mutational status of *BAP1*, *PBRM1*, and *SETD2* in these samples was acquired from the *KIRC TCGA Pan Cancer Atlas* project on cBioPortal (RRID:SCR_014555). Samples with any mutation calls (missense/splice/truncating) for our genes of interest were considered to exhibit gene loss, while those lacking any mutation calls were treated as universal WT. The corresponding FFPE whole slide images were downloaded directly from *NIH GDC Data Portal* under the *TCGA-KIRC* project by selecting only the samples which listed *Diagnostic Slide* as their experimental strategy. Of the 402 samples with mutation data, we only considered the 383 samples with matching diagnostic slides. For quality control purposes, we further excluded 20 samples where H&E images did not contain enough tumor tissue for slide level predictions (see Slide Level Inference). Of the available 383 FFPE whole slide images with mutation data, 363 were used in this study.
3. TMA1: This cohort was used primarily for validation of our region model and evaluation of clinical readouts. It was acquired from a study conducted at UTSW(25). For quality control purposes, the H&E image of each individual core was evaluated by a pathologist and cores containing less than 10% tumor cells or

exhibiting tears, staining issues, or poor focus were discarded. In all, 257 TMA cores corresponding to 118 patients passed our QC and were used in downstream analysis. We only evaluate the BAP1 prediction in this cohort. Nuclear grade was determined at a punch level by pathologist examination. Except when explicitly noted (e.g., patient level predictions), each core in the H&E-stained slide image was treated as an independent sample with corresponding cores in the matching BAP1 IHC slide used for defining BAP1 genotype.

4. TMA2: This cohort was used primarily for validation of our BAP1 region model and evaluation of clinical variables. Similar to TMA1, it was acquired from a study conducted at UTSW(36, 39). As with our other TMA cohorts, we only evaluated the BAP1 prediction. Key differences from the TMA1 cohort are that BAP1/grade status was available only at a per patient level, and imaging was performed at 40X (and synthetically down-sampled to 20X). Quality control was performed to preserve only high quality ccRCC punches with both grade and survival information. In all, 679 TMA cores corresponding to 365 patients passed our QC and were used in downstream analysis. We note that there were 23 patients whose tissue (albeit from different areas) was included both in TMA1 and TMA2.
5. PDX1: This cohort was used primarily for testing the performance in PDX tissue of our BAP1 region model trained on human patient tissue. It consists of tissue cores extracted from PDX models established at UTSW(40). Like TMA2, imaging was performed at 40X, which was then computationally down-sampled to 20X. As with TMA1, each core in an H&E stained slide was treated as an independent sample whose BAP1 status was determined previously based on a matching IHC slide(40). Based on the same eligibility criterion as TMA1, we were able to use 147 tissue cores corresponding to 46 patients.

Tumor Region Identification.

To identify the parts of our H&E slide images that contained tumor cells, we developed a DL model that segmented images into six region classes (tumor, background, normal, fat, immune, and stroma). The ground truth for this model consisted of pathologist-annotated regions, belonging to one of these six classes, in H&E stained ccRCC whole slide images. We then sampled 256x256 pixels image patches, at 5X magnification, from these annotated regions and trained a custom fully-convolutional DL model to predict the region-class corresponding to the center pixel of each patch. This region classifier was then applied to the images from the WSI cohort and only areas classified as tumor were used for training our genetic classification models (Supplementary Fig. 2a-c). Similarly, inference on TCGA was restricted to areas classified as tumor. For tissue microarray cohorts (TMA1, TMA2 and PDX1), we analyzed all tissue (i.e., not just tumor, identified by ignoring all “background” pixels) as the cores already contain a limited amount of tissue and were curated to focus on tumor regions.

Patch Generation:

Our genetics models were trained exclusively on patches from the WSI cohort. We aim to randomly sample 1200, 224x224 pixel (at 20X magnification) patches lying within the tumor area of WSI H&E images. For slides having limited tumor area, to avoid resampling the same tissue, we reduce the number of patches extracted such that the area covered by patches doesn't exceed (roughly) a third of the tumor area. For slide level classification, the patches inherit the ground truth label (Universal Loss/Localized Loss as Loss, Wild Type as WT) of whole slide genotype for the corresponding gene as determined from IHC slides as described above.

Data Splits in WSI Cohort:

To allow us to evaluate the performance of the models in the WSI cohort, we split the slides (and correspondingly the patches extracted from them) into two groups: ~60% of the slides (n=775) were used in training and the remaining ~40% (n=507) for testing, with no slide and thus patient present in both groups. The training data was further divided using a three-fold stratified cross validation scheme. The same splitting of data was used to train and test all three gene (BAP1, PBRM1, SETD2) models. In making these different splits of the data we sought to ensure that each split contained a roughly equal proportion of slides assigned as WT, UL or LL, for all three genes (Supplementary Table 1-2). Each of the folds was used to train a slide-level classification model and a region-level classification model for each driver gene, resulting in three slide models and three region models for each gene. For both slide and region level models, the results of the models from the three folds were averaged (see description of Inference) before further analysis, essentially constituting a consensus model prediction.

Slide Classification Model:

The goal of this model is to predict whether a slide contains *any* areas that exhibit loss (Supplementary Fig. 2, 3). It aims to classify WT slides as WT, but LL and UL slides as Loss. As the majority of patches in a Localized loss case could be WT, we make use of a Multiple Instance Learning (MIL) framework(44) that makes predictions over a collection of patches rather than on individual patches. Because it is impractical to store a whole slide in memory, we instead use batches of data from a single slide and train a model to predict the genetic status of the batch in MIL fashion. A unique aspect of our MIL implementation is the use of an attention mechanism: rather than predict a fixed number of the most mutant patches(44), our network learns to pay more attention to informative patches. 224 x 224-pixel patches feed into a VGG backbone which then splits into two arms, a feature extractor and an attention arm (Supplementary Fig. 4). We then combine information across all the patches by merging these two branches to calculate a single weighted (by attention outputs) mean output feature for the batch, which is ultimately used to predict whether the batch should be considered Loss or not (Supplementary Fig. 3).

Training Color Augmentation:

To increase the robustness of our models to staining variation, we subject each patch used in training to "HED" color augmentation(45) that randomly modifies the color of individually

patches during training to prevent overfitting. Our HED augmentation splits RGB image into channels corresponding to hematoxylin(H), eosin(E), and the residual DAB(D). For each channel it rescales the staining intensity by a random factor (between 0.93 and 1.08 chosen separately for each channel) across all pixels, and then finally merges back to produce an augmented RGB image.

Slide Classification Model Training:

We train independent networks, based on the shared MIL architecture, for each of 3 cross-validation folds for our 3 genes (BAP1, PBRM1 and SETD2). Each batch of training data contains randomly selected patches from a single (randomly selected) sample in the training cohort (augmented as described above). If the sample is UL or LL the ground truth target is Loss, while if it is homogeneous wildtype the target is WT. We initialized a VGG backbone with weights pretrained on ImageNet(46), froze the first four layers after the input layer and fine-tuned the remainder of the network on the training data. We used a class-weighted binary cross-entropy loss function and stochastic gradient decent (SGD) optimization, with a batch-size of 48 patches, learning rate of 0.00001, and momentum of 0.9 for training the weights. All models were trained for 15 epochs.

Slide Level Inference:

Prediction for a slide was performed for both cohorts (WSI and TCGA) by a) dividing the extracted tissue-region patches from the slide into batches of 128 patches b) predicting the batch level scores based on the pre-trained model and c) averaging the result across all batches and models from all 3 folds for the gene, i.e., a consensus model prediction. We note that 128 patches are needed to perform inference on a slide in this way. Thus, samples with fewer patches, which based on our patch generation strategy corresponds to tumor area less than about 5mm², were not analyzed.

ROC Curve Calculation:

For each gene, we used the IHC/sequencing-based status (Loss or WT, with LL being treated as Loss) as ground truth and compared it to the average consensus activation score for the sample. We obtain 95% confidence estimates on the AUC, following past work(9), by performing bootstrapping (n=10,000).(9)

Region Classification Model Data:

We trained this model using patches from WT and UL cases for all genes as their genotypes could be assigned unambiguously. For BAP1, we additionally included patches from LL cases but assigned them labels based on the manual annotations made by a pathologist (by comparison to the corresponding IHC slide). For PBRM1 and SETD2, we did not possess such manual annotations and thus patches from LL cases were not used in training (this is unlikely to have had an impact on performance as Loss/WT patches from LL cases would anyway have shown up independently for classification, the LL cases represent a minority of the total cases and were not used for evaluation of PBRM1/SETD2). All regional models were trained on the same three-fold data splits as above.

Region Classification Model Training:

All regional classification models used the same network architecture: a VGG19 backbone with a custom final layer to produce a fully convolutional network predicting two classes, WT or Loss (Supplementary Table 3, Supplementary Fig. 2c). We initialized our model with weights pretrained on ImageNet(46), froze the first four layers after the input layer and fine-tuned the remainder of the network with the same augmentation as the slide classification model. We used a class-weighted binary cross-entropy loss function and stochastic gradient descent (SGD) optimization, with learning rate of 0.0001, and momentum of 0.9 for training the weights. All models were trained for 10 epochs. We trained a different model on each cross-validation fold of each gene (BAP1, PBRM1, SETD2), giving us 9 models total.

Region Level Predictions in Whole-Slide-Images:

Inference was performed using a tessellation strategy(11,14), where patches (224x224px) were staggered in steps of 100 px and a prediction was made on each patch. To generate whole-slide region level predictions for a gene we: 1) Profiled the H&E whole slide images (by tessellation) using each of the three “fold” models for the gene to generate three activation maps. 2) averaged these activation maps to form a single final “consensus” activation map for each gene 3) partitioned the final activation map into a 1mmx1mm (TMA core size) grid, 4) averaged the activation for all grids containing at least 70% tumor tissue (the other grids are treated as background) to produce a region level prediction map. The tumor locations in this map were treated as independent predictions on which we measured our performance. The ground truth for each region was acquired by using the tumor masks created by the tumor classifier mentioned above to isolate the tumor regions and assign labels to each region based on corresponding IHC. For WT and UL cases, which were used to train all three genes, the label of all tumor regions was inherited from the assigned slide label. For Localized Loss cases, which were only used in BAP1 training/testing, the loss regions were manually annotated by a pathologist and the wildtype regions were determined by the non-loss tumor regions of the tumor mask. ROC-AUC analysis used the regional scores and ground truth for all the samples in the testing cohort.

Stain Normalization:

We performed slide level stain-normalization on our independent TMA cohorts to reduce the impact of staining variation, and to bring the staining intensities in the TMA cohort within the range present in our training data. This was achieved using an established stain normalization method(47) implemented in the StainTools package (<https://github.com/Peter554/StainTools>) which takes as input RGB values from pixels in source and target images and determines a transform such that the color distributions of the transformed source pixels match that of target pixels. This transformation is then used to normalize all pixels in the source image. We used a single target distribution for all slides, consisting of randomly sampled pixels from tumor patches in the WSI training cohort. For each TMA slide, our source distribution consisted of pixels randomly sampled from areas determined to be tissue (i.e., not in the background class) in that slide. A different transform was determined per slide and applied to all pixels in that slide.

Tests on Independent Cohorts.

We tested the BAP1 region classification model on the TMA1, PDX1, and TMA2 cohorts described in the datasets section. The inference was performed similarly to that for the regional model section above with modifications to best use the limited TMA tissue. Specifically, for each model, instead of profiling on a uniform grid we used QuPath (<https://qupath.github.io/>) to identify boxes around each TMA core, and calculated the average activation across all non-background (identified as described above) pixels in the box. The final “consensus” activation score for each punch was found by averaging scores for each fold-model. We note that for the TMA1 and PDX1 cohorts, we had punch level BAP1 assignments, and thus the above punch level activations and BAP1 status were compared. However, for the TMA2 cohort, BAP1 status was only available at a patient level (with each patient being represented by ~2 punches on average). As a patient was considered to have BAP1 loss if any of the corresponding punches displayed loss, we calculated patient level activation as the minimum (i.e., more BAP1-loss like) activation across all punches for the patient.

Nuclear Feature Based Classification.

We developed a U-Net based model to identify the location and shape of individual nuclei in our H&E images as shown previously(48). Ground truth for this model, an assignment of pixels as nucleus/edge/background, was based on annotations of the boundaries of all nuclei in 186 – ~1000x1000 pixel areas from 31 H&E slides at 20X magnification. The trained U-Net classifier was then used to classify pixels in whole slides, and individual nuclei were identified as connected components of pixels belonging to the “nucleus” class. As quality control, we dropped nuclei smaller than 30 pixels or larger than 5000 pixels. For each image in the WSI cohort, for each nucleus in the tumor region, we extracted a library of 36 features quantifying aspects of nuclear size, shape, color (in different color spaces), and texture (Supplementary Table 4). We then calculated the median value of each feature across all the (tumor region) nuclei to produce a 36-dimensional profile vector for the sample. Using the WSI Training Cohort, we trained random forest classifiers to predict the mutational status of the slides based on their nuclear features (which we then evaluated on the WSI Testing cohort).

Relationship to Tumor Grade.

Two pathologists (V.P and P.K) independently reviewed each core in the human tissue microarray, and then together assigned a consensus grade score (1-4) for each core in the TMA1 cohort, while grade was obtained at the patient level for the TMA2 cohort.

Survival Prediction.

Using the Lifelines python package, we performed a Cox-proportional hazards model analysis relating our BAP1 model’s continuous score (range 0-1) to disease-specific survival (based on time between diagnosis and ccRCC-related-death or censored at last follow up/death related to other causes), and quantified performance in terms of p-value, hazard-ratio, and c-index. We also performed a similar analysis after stratifying patients based on the IHC assessment of their True BAP1 status. Finally, for visualization purposes we used a Kaplan

Meier Curve to compare the IHC stratification to one based on predicted BAP1 status: patient predictions were considered Loss if at least one activation score across all cores for that patient was below 0.5, and WT if all activation scores for those cores were above 0.5.

Data/Code Availability.

Our H&E slide images for the WSI, TMA1, TMA2 and PDX1 cohorts are available on FigShare+ at <https://doi.org/10.25452/figshare.plus.c.5983795>. The WSI cohort can be found at <https://doi.org/10.25452/figshare.plus.19310870> and the TMA cohorts can be found at <https://doi.org/10.25452/figshare.plus.19324118>. TCGA KIRC can be downloaded from the TCGA GDC portal, while the corresponding genetic assignments are available from cBioPortal. The deep learning models for tumor identification, nuclear segmentation and gene status prediction (both at local and region level) are available at <https://doi.org/10.25452/figshare.plus.19353830>. The source code and metadata for UTSW cohorts are available on GitHub at <https://github.com/Rajaram-Lab/cancres-2022-intratumoral-heterogeneity-dl-paper>. A singularity container image to run the code/models is available on DockerHub at https://hub.docker.com/layers/srajaram/lab_img/v2.14/images/sha256-f307fdf489b509740758813abef603f931993ce9ea1979eb9d980853a4b52415.

Results

In this study, we attempt to predict the status of three commonly mutated genes in ccRCC, *BAP1*, *PBRM1*, and *SETD2*(4,26,30-32), from H&E images (Fig. 1a). For each gene, we trained models at two scales: a) whole-slide level - to allow comparison to previous bulk-sequencing approaches, we predict if the gene is lost anywhere in a slide; and b) region level - to resolve intra-tumor heterogeneity, we predict gene loss status of local regions within slides (Fig. 1b, Supplementary Fig. 2). The characteristics of different cohorts used are summarized in Table 1 and detailed below.

Our primary training cohort, the WSI cohort (Table 1), consists of 1282 whole slide H&E-stained images (each from a different ccRCC patient)(28,29,33,34) with spatially resolved gene loss status determined by a trained pathologist's inspection (Methods; Supplementary Fig. 2a) of serial sections stained with IHC antibodies against BAP1/PBRM1/H3K36me3 (SETD2). Consistent with expected frequencies of gene mutation, BAP1, PBRM1, and SETD2 inactivation was observed in approximately 12.40%, 52.18% and 27.22% of slides respectively. We split our samples into two separate cohorts (Fig. 2a), a training cohort that contains 60% of the slides (775 samples) and a testing cohort (WSI Testing) that contains the remaining 40% of the data (507 samples) (Fig. 2a). All analysis of whole slide images (which often contain areas of normal/stromal cells) was performed on tumor areas only (Supplementary Fig. 2a), which we identify in H&E WSIs using an independent model trained using pathologist annotation (Methods). From these tumor regions, we then extract 224px (~111 microns) square patches at 20X magnification and train models to assign gene loss status based on the collection of these patches. Crucially, because of characteristic ccRCC intra-tumor heterogeneity, for a significant proportion of slides (~8-20% of cases exhibiting loss, depending on gene; Supplementary Table 1), the protein loss was localized to a focal region of the tumor area visualized on the whole slide (Supplementary Fig. 1).

Thus, in generating our training data we distinguish between slides that exhibit Wild-Type (WT, wildtype-like staining across whole slide), Universal Loss (UL, entire tumor area exhibiting gene loss), or Localized Loss (LL, subset of tumor area exhibiting gene loss) for each gene, which we ensure are spread approximately equally across our data splits (Supplementary Table 2).

Deep Learning Can Predict Gene Loss at Whole Slide Level.

To enable comparison to published results, we first built a model that called (for example) a slide as BAP1 Loss if any part of the slide (even if in a minority of cells) exhibited this loss (Fig. 1b). Thus, all WT slides should be predicted as WT, but UL and LL slides should be predicted as Loss. To overcome the intra-slide heterogeneity in gene loss status, we adopt a custom multiple instance learning (MIL) approach(44) (Fig. 2b, Supplementary Fig. 2b, 3, 4, Methods). This approach learns to pay more attention to loss-like phenotypes in a collection of patches and is thereby able to identify gene loss at the slide level, even if exhibited only in a minority of patches. We trained “consensus” predictors (i.e., average of multiple models based on the training data; Methods) to infer, from H&E images, the occurrence of BAP1, PBRM1, and SETD2 loss on whole slides. The output of each predictor is an “activation” score of gene function, spanning a range of 0 for a confident Loss call to 1 for a confident WT call. In cases where a hard Loss/WT call is needed, an activation below 0.5 is treated as a Loss (and above 0.5 as WT).

We quantified the performance of our classifier on the WSI testing cohort using receiver operating characteristics (ROC) of the output activation for slides relative to their ground truth gene loss status. We restricted our analysis to 489 (out of 507) samples that had adequate tissue for reliable slide level prediction (Methods). Following past work(9) (Methods), we used a bootstrap-based estimation of 95% AUC confidence intervals that are indicated in parenthesis below. All three genes exhibited high AUCs, BAP1=0.87 (0.82-0.91), PBRM1= 0.77 (0.73-0.81), and SETD2 0.71 (0.66-0.76) (Supplementary Table 5), suggesting that in the hold-out portion of the training cohort (Fig. 2c), it is possible to predict expression of a whole slide purely based on tissue morphology.

Models trained on one cohort often perform poorly in cohorts that show significant differences in preparation (e.g., staining) and imaging (e.g., different microscopes) from the original training cohort(49). To validate our results, we used the independent KIRC TCGA Pan Cancer Atlas dataset(41) (referred to simply as TCGA), which contains H&E-stained whole slide images as well as mutational status for our genes of interest (Supplementary Table 1), albeit at the whole sample, rather than regional, level. To serve as ground truth, we used mutational status for *BAP1/PBRM1/SETD2* from cBioPortal(42,43) and all called mutations (missense/splice/truncating) for these genes were treated as representing gene loss for our analysis (Methods; excluding missense mutations did not have a significant impact on results - Figure 2d vs Supplementary Fig. 5a). We found H&E based features offered value towards gene loss prediction for all three genes with AUCs of 0.77(0.70-0.84)/0.69(0.63-0.76)/0.60(0.51-0.68) (Supplementary Table 5) for BAP1, PBRM1 and SETD2 respectively. Interestingly, we were able to further improve the performance of the models, by restricting predictions to just low-grade tumors (Supplementary Fig. 5b-d; based on

nuclear grade information available for TCGA). This is consistent with low-grade tumors being less evolved (with fewer mutations making it easier to identify the gene specific phenotypes) and more genetically homogenous.

Our model makes few false BAP1 Loss predictions (negative predictive value (NPV) of 0.94 on WSI Testing and 0.93 on TCGA), but as suggested by the flat top of the AUC curve, the high-confidence BAP1 WT predictions are especially reliable (Fig 2c,d). For both cohorts, considering ~60% of the samples deemed most likely to have BAP1 loss (Fig 2e, top) captures 96% of samples that truly exhibit BAP1 loss (i.e., sensitivity=0.96). Moreover, among the 33% of samples deemed most likely to be WT (Fig. 2e bottom), none exhibited BAP1 loss in the TCGA and the two apparent mistakes in the top 33% (n=167) in the WSI Testing cohort both exhibited BAP1 loss heterogeneity within the slide (suggesting a more conservative model may be able to do better). Taken together, our results suggest that our models for predicting gene loss status at the whole slide level, especially for ruling out BAP1 loss, are particularly robust and generalizable.

Deep Learning Can Predict Gene Loss at Regional Level.

We sought to build a framework to profile intra-tumor heterogeneity in H&E-stained images by localizing the gene loss calls to small intra-tumor regions (Fig 1b; as opposed to predicting if loss was exhibited anywhere on the slide as above). Here, we used a fixed region size of 1 mm diameter as our unit of prediction. We reasoned that this size a) strikes a good balance between capturing enough cells for a stable prediction, while still being small enough for low probability of genetic heterogeneity within a region and b) simplifies our validation process since it corresponds to the size of punches of our TMA cores.

For regional classification of gene loss, we built a more traditional convolutional neural network (Supplementary Fig. 2c) which sought to classify individual patches (rather than collections of patches for the MIL whole slide model) to the loss status they inherited from matching IHC regions. Similar to the whole-slide approach, we trained separate consensus predictors (by averaging multiple DL(46) models, Methods) for BAP1, PBRM1 and SETD2. To achieve TMA scale predictions, we first used the models to generate activation maps for the tumor portions of each slide (Fig. 3a-b, Supplementary Fig. 3), which we then divided into TMA sized grid regions (Fig. 3c) and finally calculated region scores by averaging activation values within a grid square (Fig. 3d-e).

In this way, we were able to generate regional predictions within all our slides in the WSI Testing cohort (n=507). To quantify performance, we compared the region level activation scores to the ground truth for each punch size area. For all three genes, we achieved similarly high levels of performance (Fig. 3f) as their slide level counterparts with AUCs of 0.89(0.88-0.89)/0.76(0.75-0.77)/0.68(0.68-0.69) for BAP1, PBRM1 and SETD2 respectively, suggesting that morphology is predictive of gene loss status even at the regional length scale (Fig. 3f). Notably, even at this higher spatial resolution, BAP1 Loss continues to have the strongest relationship with tissue morphology and as discussed below, BAP1 model predictions are highly sensitive (Supplementary Fig. 6a). Given how much stronger the results are for BAP1 than PBRM1 or SETD2 (Fig. 3F), all subsequent analysis and validations were performed in the context of BAP1 alone.

We next considered the ability to deconvolute heterogeneity in slides that contain instances of both Loss and WT status (BAP1 LL cases). Fig 3 shows an example of how our prediction of BAP1 loss areas in a BAP1 LL cases compares to the IHC ground truth (other examples in Supplementary Fig. 7a-e). Beyond a hard WT/Loss call, our models output can also be used to guide sequencing, for example to preferentially select areas likely to exhibit BAP1 loss. Among the LL cases not used in training, we sorted the different 1mmx1mm areas based on likelihood of BAP1 loss and found that by picking either the highest or second highest ranked areas we would capture BAP1 loss better than 98.7% of random selections, while choosing the third most confident area would still be better than better than 86.6% of random selections (Supplementary Fig. 8a-b).

Regional Model can predict BAP1 Status in multiple independent Tissue Microarrays.

We tested the robustness of our BAP1 region model on independent cohorts that a) consisted of samples not used in training b) used tissue microarrays (TMA) rather than whole tissue slides. An additional challenge was the stain color distribution differed from WSI and TCGA (Supplementary Fig. 9a-c), which we chose to address using slide level color normalization (although the overall results are largely unaffected by this; Supplementary Fig. 9d). We found that the TMA punches were largely homogenous and largely made up of tumor regions, obviating the need for filtering using a tumor classifier. We then used a region level approach akin to Fig. 3 (as our regions there were the same 1mm size as the TMA punches) to produce a single gene loss activation score for each punch which was then compared to known gene loss status (Methods).

Our first independent tissue microarray cohort (TMA1) contained matched IHC and H&E data for 118 human ccRCC patients assayed as 257 punches(25,26). This is an ideal data set for testing the robustness of our predictions because in addition to differing from our training cohort in preparation (tissue microarray vs training on whole slides and being generated and assayed at a different institution), the gene loss assignments are of high quality, having been validated by sequencing(26). While a tissue microarray does not exhibit the spatial heterogeneity of a whole slide, reliable predictions at a punch level will capture some of the genetic variation in a patient as each patient is represented by multiple TMA cores. Applying the BAP1 region classification models to the punches from TMA1 (Fig. 4a, b) we achieved an AUC of 0.77 (0.69-0.85) which, although expectedly lower than our training cohort with an AUC of 0.89(0.88-0.89), still represents a well performing classification, especially considering the differences to the training cohort.

On both the WSI (Negative predictive value (NPV) = 0.971) cohort and TMA1(NPV=0.975), our BAP1 region model is highly sensitive. In both cohorts, the top 25% of samples deemed most likely to be BAP1 WT, contain no BAP1 loss cases, and it is possible to capture 96% of BAP1 loss cases by examining just the top 63% of samples deemed most likely to exhibit BAP1 loss (Supplementary Fig. 6a,b; see Supplementary Fig. 10a-d for a more detailed view of tradeoffs across cohorts). Comparable performance (AUC=0.84) and similar sensitivity can also be observed in the second human cohort: TMA2 (Supplementary Fig. 6c, 10,11). Furthermore, like in our WSI cohort, in TMA1 we find that for patients exhibiting heterogeneity in BAP1 loss, representing the patient by the single

punch deemed most likely to exhibit BAP1 Loss by our model does better than a similar random selection 98.9% of the time (Supplementary Fig. 8c). Taken together, these results demonstrate the robustness of our region level BAP1 classification models and the high sensitivity of its predictions.

Next, we tested whether deep learning phenotypes learnt on patient tissue are also applicable to tissue from patient derived xenograft (PDX) models. Given the dramatic experimental differences between patient tissue and PDXs, this represents a severe test of the robustness of the gene models. Intriguingly, as the generation of a PDX tumor involves replacement of the stroma by mouse tissue(40), the extent to which BAP1 status can be successfully predicted in PDXs highlights the tumor-intrinsic (as opposed to microenvironmentally driven) nature of these tissue phenotypes. We tested our models by applying the same approach we used for TMA1 to 147 punches in the PDX1 cohort from 46 PDX models established at UTSW. Despite the significant differences from the training cohort, which consists of human patient tissue, we were able to achieve performance comparable to TMA1 with AUC of 0.80 (0.69-0.90) in PDX1 (Fig. 4b, c, Supplementary Fig. 6d). These data shows that the DL model does not rely on homotypic interactions between tumor and stromal cells, and that the features are preserved after tumor transplantation into a rodent. The ability to achieve such high performance using human trained models suggests that analysis of tissue phenotypes in PDX models could potentially inform future human studies.

Different Features Predict BAP1, PBRM1 and SETD2 Loss.

To better understand our models, we first sought to test whether morphology changes were gene-specific or were shared across driver mutations (e.g. reflecting overall tumor progression which can occur through a variety of mutations). If the latter were true, one might expect that when one gene model (e.g., BAP1) makes a Loss prediction that would increase the likelihood of a different gene model (e.g., PBRM1) also making a Loss prediction on that sample. However, our results indicate that in fact, the likelihoods for BAP1 and PBRM1 loss perceived by our models are negatively correlated (Spearman Correlation Coefficient= -0.326, p-value=6.73e-13) (Fig 5A-Bottom). This suggests that features used to predict BAP1 and PBRM1 are distinct and aligns with previous studies that indicate that mutations in BAP1 and PBRM1 are typically mutually exclusive in ccRCC(26). Moreover, PBRM1 status does not impact our model's perception of BAP1 phenotypes (Fig 5a-Top; gray vs blue curves): both PBRM1 WT and Loss samples show similar distributions of BAP1 mutation likelihood. Similar results show that BAP1 is only weakly correlated with SETD2 classification status (Supplementary Fig. 12a). In contrast, the correlation between PBRM1 and SETD2 is consistent with previously reported co-occurrence of their loss(26,32) (Supplementary Fig. 12b). Taken together, predicted co-occurrence appears to mirror true co-occurrence of gene loss, suggesting that our models can identify gene specific tissue phenotypes.

We next sought to identify the morphological features that were most impacted by BAP1 and PBRM1 mutations. We focused primarily on nuclear phenotypes based on 36 nuclear features (including quantification of nuclear shape, size and hematoxylin texture) from all tumor nuclei in the WSI training cohort samples (See Methods, Supplementary Fig.

13, Supplementary Table 4). Random forest classifiers trained using these features were able to classify BAP1 mutational status (Methods), with an AUC of 0.72, suggesting that nuclear features can contribute to identification of BAP1 loss (Supplementary Fig. 14a). By analyzing the differences in feature values between Loss and WT for BAP1 (Fig. 5b, Supplementary Fig. 15a-i), we were able to determine that BAP1 mutant tumors typically had larger nuclei and lighter nuclear hematoxylin staining than their WT counterparts (also see Supplementary Fig. 15a,d), consistent with previous reports(26,27). These features were significantly less affected by PBRM1 and SETD2 Loss (Fig. 5b, Supplementary Figures 14b-c, 16a-i, 17a-i).

BAP1 predictions correlate with clinical prognostic factors.

Next, we sought to investigate the relationship between our BAP1 model predictions and known clinical prognostic parameters. For these experiments, we used TMA2, which while overlapping with TMA1 (23 of the 118 patients in TMA1 are shared) contains a larger number of patients (N=365) with relevant clinical annotations. Next, as BAP1 mutations are known to be found largely in high-grade tumors(25), we chose to investigate how much value our classifier offers beyond nuclear tumor grade. As expected, tissue with higher grade is more likely to truly exhibit BAP1 loss and this is mirrored in the confidence of the BAP1 loss prediction. However, even *within* low- and high-grade tumors, samples with a BAP1 loss have higher confidence of being predicted as BAP1 loss than WT ones. These results are also supported by the TMA1 cohort where grade assessments were made at a punch (rather than patient) level, and within individual grades rather than just low/high (Supplementary Fig. 18a-b). Similarly, using just the grade to predict BAP1 would give poorer performance than our model with an AUC of 0.71 (0.63-0.79) on TMA1. Taken together, these results indicate that, while the phenotypes identified by our BAP1 classifier likely share some commonalities with grade, they capture BAP1 signature information that goes beyond grade.

Finally, because previous studies have shown that BAP1 mutations may influence patient outcome(25), we wanted to investigate the impact of our BAP1 predictions on survival. We performed a Cox-proportional hazards regression in TMA2 using our predicted BAP1 likelihood as a continuous variable and found that it significantly contributes to prediction of disease specific survival ($p=0.0087$, $HR=5.43$, $c\text{-index}=0.65$). Interestingly, we found our prediction outperformed stratification based on IHC based BAP1 status ($c\text{-index}=0.55$). For visualization purposes, we additionally stratified patients in TMA2 by whether their activation was greater than 0.5 (more likely to be WT) or less (more likely BAP1 loss) and compared the disease specific survival to stratification based on IHC based determination of BAP1 status (Fig. 5d). Based on this 0.5 cutoff, the BAP1 classifier based stratified survival curves showed an improved hazard ratio ($HR=2.53$ vs 1.8) and p-value (0.0027 vs 0.11) compared to stratification based on BAP1 status determined by IHC.

Discussion

In this study, we investigated the ability of DL to analyze H&E images to predict functional status of driver genes in ccRCC along two largely unexplored axes. First, we explored the

effect of spatial length scale (slide vs region) and whether genetic predictions can be made with high confidence at whole-slide level, as well as at sub-clonal regions, within tumors. The use of IHCs as ground truth is a unique aspect of our study enabling spatial resolution far surpassing previous sequencing-based approaches. Second, we test the robustness of our models across cohorts that are increasingly divergent from the initial training set, including the first test of human-tissue-trained models on PDX samples. Our results indicate that the status of all 3 genes evaluated, BAP1, PBRM1 and SETD2, can be gleaned based on morphology. A particularly robust relationship was observed with BAP1.

At the whole-slide level, we report AUCs of 0.87/0.77/0.71 for BAP1/PBRM1 and SETD2, respectively, on the held-out portion of the WSI cohort. These differences likely reflect the fact that BAP1 loss has a large impact on tissue morphology (evidenced by its greater impact on nuclear grade(25-27)), and while differences exist among the three IHC assays they all had very high positive and negative predictive values (26,29). Additionally, in ccRCCs, PBRM1/SETD2 loss is often associated with increased ITH (e.g., in the “PBRM1→SETD2 evolutionary subtype”) in contrast to “BAP1 driven” tumors that have low ITH with clonal expansion(4), potentially making it harder to identify their specific morphologic signatures. Our performance for all three genes was reduced when the classifier was applied to independent slides from TCGA (AUC=0.77/0.69/0.60). However, this is not unexpected for several reasons. First, whereas our models were trained on IHC calls easily co-registered with H&E sections and providing superb accuracy, gene loss calls for TCGA were performed on spatially disjointed samples from those stained for H&E analyses. Second, factors other than mutations may result in lack of protein expression. We have previously reported that samples lacking a mutation where IHC showed BAP1 loss, did not express BAP1 protein on Western blot(26). Finally, the bulk-sequencing based calling of mutations performed by TCGA is prone to missing sub-clonal events, represented by a small number of cells and therefore reads, that could be picked up by our assays. Consistent with these ideas, for the *BAP1* gene, high confidence WT predictions seem to be always correct (NPV: WSI=0.94, TCGA=0.93), and it is the fewer sequencing-based detections of high confidence BAP1 loss predictions (PPV: WSI=0.44, TCGA=0.25) in TCGA that seem to drive the drop in AUC.

Two previous pan-cancer studies based on FFPE(9) and frozen(10) tumor sections as well as a ccRCC specific study(12) attempted to predict mutation in our genes of interest from H&E images. These studies trained and tested their models on the TCGA KIRC cohort and may be expected to outperform our DL model on this dataset. However, despite training on a different dataset our DL model rivaled their performance with reported AUCs for PBRM1 of 0.59(10)-0.63(9) and SETD2 0.39(10)-0.65(9). The difference is even more striking for BAP1, where our model shows an AUC of 0.77 vs AUC=0.65(10). We suspect that our higher scores are probably due to the larger number of examples of BAP1 loss in our cohort (n=159) compared to the TCGA (n≈33), although our IHC assays which can capture other mechanisms of protein loss with high reliability (26,29) may also have contributed. It is observed in the field that predictions made on the held-out part of the training cohort are often significantly better than those on an independent cohort. Among mutation predictions from H&E images that have been validated on an independent cohort, our BAP1 model with an AUC of 0.87/0.77 (holdout/independent) is among the strongest. It compares favorably

to scores of 0.83/0.75 for EGFR in lung cancer(11), 0.83/0.63 for TP53 in breast cancer(10) and 0.72/0.67 for BRAF in colorectal cancer(9). Similarly, our region level BAP1 model showed consistent performance across multiple cohorts (AUCs=0.89 in WSI holdout vs 0.77 in TMA1 and 0.84 in TMA2). This speaks to the robustness of the BAP1-morphology relationship.

The strong and robust morphology-genetics connection we detect for BAP1 has both practical and conceptual implications. Practically, our slide-level BAP1 score distribution may provide high enough sensitivity to be confidently actionable, although validation on additional cohorts is needed. Studies in other cancers have reported similarly high AUC scores but our ability to confidently rule out BAP1 loss on a large number of samples has not been rivaled except in much smaller cohorts(50). Moreover, the continued performance of the model at shorter length scales (regions of 1 mm diameter) suggest that these predictions can help guide studies to dissect intra-tumor heterogeneity. To capture 75% of all driver mutations in ccRCC, multi-regional sequencing of 7 randomly-selected regions is needed(4). As our models can rule out BAP1 loss with high sensitivity, they may help identify areas for sequencing, thereby increasing efficiency. From a more conceptual perspective, our results emphasize the scientific opportunities in studying the strong connection between genetics and morphology. Despite the presence of many other mutations, and no known mechanistic link between BAP1 and morphology, the effect of BAP1 stood out. The fact that this link is recapitulated in PDXs models, where the stroma is replaced by mouse tissue(40), shows that the effect is likely independent of homotypic tumor-microenvironment interactions. Moreover, the strength of the association is such that it may be possible to design experiments to understand how a deubiquitinase like BAP1 ends up constructing such reproducible morphological phenotypes.

Our results highlight the association of BAP1 loss with larger nuclei and opened chromatin (lighter nuclear hematoxylin staining than their WT counterparts), an association not observed in the PBRM1 and SETD2 Loss tumors. These findings are consistent with our prior work demonstrating that BAP1 and PBRM1 drive ccRCC histologic grade(25,26). We have shown(30) that targeting *Bap1* and *Pbrm1* in the mouse induces tumors of different nuclear grade where *Bap1*-deficient tumors are of high grade, a finding consistently observed in human ccRCCs. Our deep learning model, which was not privy to such information, seemed to have rediscovered the relationship between BAP1 and aggressiveness (as measured by grade), but also seems to be able to make predictions within a tumor grade. Surprisingly, our BAP1 classifier was predictive of disease-specific survival, apparently more so than BAP1 status as measured by IHC. While we cannot rule out that grade-associated phenotypes picked up by the current BAP1 model could be influencing these predictions, future models which explicitly correct for grade effects could help better understand the interaction between, BAP1 loss, BAP1 associated morphological phenotypes and prognosis.

These findings also raise questions and offer avenues for improvement in future studies. Despite the best-in-class AUC for BAP1 prediction, our model is not ready to replace IHC. Our high AUC scores (0.87, 0.89 on slide and region level respectively) reflect an enrichment in the relatively rare BAP1 mutations (~10-15% of cases). However, this only

translates to a precision of approximately 44%. There are multiple potential explanations for these false positives. This may be a limitation of our training process and an even larger training cohort with a broader range of BAP1 mutant examples (our current model is trained on ~98 BAP1 loss cases), would perform better, as has been demonstrated for metastasis detection(49). A second possibility is that as morphology is determined by several other factors including other mutations, (epi)genomics and environmental factors(7,8), we have hit an information threshold on how much morphology can tell us about the functional status of BAP1. Perhaps the most intriguing possibility is that the morphological signature we identify corresponds to a deeper cellular state where BAP1 mutations represent simply one means to reach this state. In other words, there may be other mechanisms to deregulate the BAP1 pathway besides BAP1 loss. Representing an endpoint of multiple cellular events, morphology could thus be a better indicator of survival than BAP1 status.

This study highlights the convergence of the two pillars of cancer diagnosis, histopathology and molecular genetics. Until recently it was unclear how morphology correlated with genomics. Multiple studies demonstrate a connection between morphology and particular gene mutations(9-11,13,14). These links are detectable using DL and our results show that the approach is valid to also profile tumor subclones. Intra-tumor heterogeneity has been a bane of biomarker studies. It is challenging to capture intra-tumor heterogeneity with sequencing-based approaches, which are also expensive, but heterogeneity is readily appreciable by visual inspection. Provocatively, it has been proposed that tissue morphology may not simply reflect a downstream readout of cellular state but may actively modulate signaling(7,8). This emerging field has primarily focused on phenotypes of disassociated single cells, but our results suggest the links between morphology and molecular state in tissues are robust enough to be studied in more tractable model systems. Taken together, it is apparent that machine learning will have a key role to play in integrating tumor morphology and molecular biology, a key area of future research. Furthermore, as our data show, machine learning has the potential to provide important diagnostic and prognostic cues. Finally, inasmuch as mutations, such as in *BAP1*, may also predict therapeutic response, DL models may also influence therapy selection.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank all of the patients who provided tissues that enabled this research project. J.B., A.C., D.R. and P.K. are supported by NIH (Specialized Program in Research Excellence in Kidney Cancer P50 CA196516), J.B., A.C., and P.K. by the Cancer Research & Prevention Institute of Texas (CPRIT; RP180192). J.B. and A.C. are also supported by CPRIT (RP180191). P.K. is supported by NIH (R01CA244579, R01CA154475, and R01DK115986), DOD (W81XWH1910710), and CPRIT (RP200233). S.R. is supported by startup funds provided through the Lyda Hill Department of Bioinformatics.

References

1. Dagogo-Jack I, Shaw AT. Tumour heterogeneity and resistance to cancer therapies. Nat Rev Clin Oncol 2018;15:81–94 [PubMed: 29115304]

2. Lawson DA, Kessenbrock K, Davis RT, Pervolarakis N, Werb Z. Tumour heterogeneity and metastasis at single-cell resolution. *Nat Cell Biol* 2018;20:1349–60 [PubMed: 30482943]
3. Massague J, Obenauf AC. Metastatic colonization by circulating tumour cells. *Nature* 2016;529:298–306 [PubMed: 26791720]
4. Turajlic S, Xu H, Litchfield K, Rowan A, Horswell S, Chambers T, et al. Deterministic Evolutionary Trajectories Influence Primary Tumor Growth: TRACERx Renal. *Cell* 2018;173:595–610 e11 [PubMed: 29656894]
5. Hao JJ, Lin DC, Dinh HQ, Mayakonda A, Jiang YY, Chang C, et al. Spatial intratumoral heterogeneity and temporal clonal evolution in esophageal squamous cell carcinoma. *Nat Genet* 2016;48:1500–7 [PubMed: 27749841]
6. Yates LR, Gerstung M, Knappskog S, Desmedt C, Gundem G, Van Loo P, et al. Subclonal diversification of primary breast cancer revealed by multiregion sequencing. *Nat Med* 2015;21:751–9 [PubMed: 26099045]
7. Mohan AS, Dean KM, Isogai T, Kasitinon SY, Murali VS, Roudot P, et al. Enhanced Dendritic Actin Network Formation in Extended Lamellipodia Drives Proliferation in Growth-Challenged Rac1(P29S) Melanoma Cells. *Dev Cell* 2019;49:444–60 e9 [PubMed: 31063759]
8. Park JS, Burckhardt CJ, Lazcano R, Solis LM, Isogai T, Li L, et al. Mechanical regulation of glycolysis via cytoskeleton architecture. *Nature* 2020;578:621–6 [PubMed: 32051585]
9. Kather JN, Heij LR, Grabsch HI, Loeffler C, Echle A, Mützi HS, et al. Pan-cancer image-based detection of clinically actionable genetic alterations. *Nature Cancer* 2020;1:789–99 [PubMed: 33763651]
10. 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* 2020;1:800–10 [PubMed: 35122049]
11. Coudray N, Ocampo PS, Sakellaropoulos T, Narula N, Snuderl M, Fenyo D, et al. Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. *Nat Med* 2018;24:1559–67 [PubMed: 30224757]
12. Marostica E, Barber R, Denize T, Kohane IS, Signoretti S, Golden JA, et al. Development of a histopathology informatics pipeline for classification and prediction of clinical outcome in subtypes of renal cell carcinoma. *Clin Cancer Res* 2021
13. Naik N, Madani A, Esteva A, Keskar NS, Press MF, Ruderman D, et al. Deep learning-enabled breast cancer hormonal receptor status determination from base-level H&E stains. *Nature Communications* 2020;11:5727
14. Kather JN, Pearson AT, Halama N, Jäger D, Krause J, Loosen SH, et al. Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. *Nat Med* 2019;25:1054–6 [PubMed: 31160815]
15. Tan X, Su AT, Hajiabadi H, Tran M, Nguyen Q. Applying Machine Learning for Integration of Multi-Modal Genomics Data and Imaging Data to Quantify Heterogeneity in Tumour Tissues. In: Cartwright H, editor. *Artificial Neural Networks*. New York, NY: Springer US; 2021. p 209–28.
16. Faust K, Roohi A, Leon AJ, Leroux E, Dent A, Evans AJ, et al. Unsupervised Resolution of Histomorphologic Heterogeneity in Renal Cell Carcinoma Using a Brain Tumor-Educated Neural Network. *JCO Clinical Cancer Informatics* 2020;8:11–21 [PubMed: 32946287]
17. He B, Bergenstråhle L, Stenbeck L, Abid A, Andersson A, Borg Å, et al. Integrating spatial gene expression and breast tumour morphology via deep learning. *Nature Biomedical Engineering* 2020;4:827–34
18. Hidalgo M, Amant F, Biankin AV, Budinská E, Byrne AT, Caldas C, et al. Patient-Derived Xenograft Models: An Emerging Platform for Translational Cancer Research. *Cancer Discovery* 2014;4:998–1013 [PubMed: 25185190]
19. Gerlinger M, Rowan AJ, Horswell S, Larkin J, Endesfelder D, Gronroos E, et al. Intratumor Heterogeneity and Branched Evolution Revealed by Multiregion Sequencing. *New England Journal of Medicine* 2012;366:883–92 [PubMed: 22397650]
20. Cai Q, Christie A, Rajaram S, Zhou Q, Araj E, Chintalapati S, et al. Ontological analyses reveal clinically-significant clear cell renal cell carcinoma subtypes with convergent evolutionary trajectories into an aggressive type. *EBioMedicine* 2020;51:102526 [PubMed: 31859241]

21. Payal Kapur M, Alana Christie, MS, Satwik Rajaram, PhD, and James Brugarolas, MD, PhD. What morphology can teach us about renal cell carcinoma clonal evolution. *Kidney Cancer Journal* 2020;18:68–75 [PubMed: 34178206]
22. Grignon DJ, Che M. Clear cell renal cell carcinoma. *Clin Lab Med* 2005;25:305–16 [PubMed: 15848738]
23. Reuter VE, Tickoo SK. Differential diagnosis of renal tumours with clear cell histology. *Pathology* 2010;42:374–83 [PubMed: 20438412]
24. Verine J, Colin D, Nheb M, Prapotnich D, Ploussard G, Cathelineau X, et al. Architectural Patterns are a Relevant Morphologic Grading System for Clear Cell Renal Cell Carcinoma Prognosis Assessment: Comparisons With WHO/ISUP Grade and Integrated Staging Systems. *Am J Surg Pathol* 2018;42:423–41 [PubMed: 29356723]
25. Kapur P, Peña-Llopis S, Christie A, Zhrebker L, Pavia-Jiménez A, Rathmell WK, et al. Effects on survival of BAP1 and PBRM1 mutations in sporadic clear-cell renal-cell carcinoma: a retrospective analysis with independent validation. *The Lancet Oncology* 2013;14:159–67 [PubMed: 23333114]
26. Pena-Llopis S, Vega-Rubin-de-Celis S, Liao A, Leng N, Pavia-Jimenez A, Wang S, et al. BAP1 loss defines a new class of renal cell carcinoma. *Nat Genet* 2012;44:751–9 [PubMed: 22683710]
27. Gallan AJ, Parilla M, Segal J, Ritterhouse L, Antic T. BAP1-Mutated Clear Cell Renal Cell Carcinoma: A Clinicopathologic Characterization of 14 Molecularly Confirmed Tumors. *American Journal of Clinical Pathology* 2020
28. Joseph RW, Kapur P, Serie DJ, Eckel-Passow JE, Parasramka M, Ho T, et al. Loss of BAP1 protein expression is an independent marker of poor prognosis in patients with low-risk clear cell renal cell carcinoma. *Cancer* 2014;120:1059–67 [PubMed: 24382589]
29. Ho TH, Kapur P, Joseph RW, Serie DJ, Eckel-Passow JE, Tong P, et al. Loss of histone H3 lysine 36 trimethylation is associated with an increased risk of renal cell carcinoma-specific death. *Modern Pathology* 2016;29:34–42 [PubMed: 26516698]
30. Gu YF, Cohn S, Christie A, McKenzie T, Wolff N, Do QN, et al. Modeling Renal Cell Carcinoma in Mice: Bap1 and Pbrm1 Inactivation Drive Tumor Grade. *Cancer Discov* 2017;7:900–17 [PubMed: 28473526]
31. Creighton CJ, Morgan M, Gunaratne PH, Wheeler DA, Gibbs RA, Gordon Robertson A, et al. Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature* 2013;499:43–9 [PubMed: 23792563]
32. Brugarolas J Molecular Genetics of Clear-Cell Renal Cell Carcinoma. *Journal of Clinical Oncology* 2014;32:1968–76 [PubMed: 24821879]
33. Ho TH, Kapur P, Eckel-Passow JE, Christie A, Joseph RW, Serie DJ, et al. Multicenter Validation of Enhancer of Zeste Homolog 2 Expression as an Independent Prognostic Marker in Localized Clear Cell Renal Cell Carcinoma. *Journal of Clinical Oncology* 2017;35:3706–13 [PubMed: 28976794]
34. Joseph RW, Kapur P, Serie DJ, Parasramka M, Ho TH, Cheville JC, et al. Clear Cell Renal Cell Carcinoma Subtypes Identified by BAP1 and PBRM1 Expression. *J Urol* 2016;195:180–7 [PubMed: 26300218]
35. Gayed BA, Youssef RF, Bagrodia A, Darwish OM, Kapur P, Sagalowsky A, et al. Ki67 is an independent predictor of oncological outcomes in patients with localized clear-cell renal cell carcinoma. *BJU Int* 2014;113:668–73 [PubMed: 23937277]
36. Gayed BA, Youssef RF, Bagrodia A, Kapur P, Darwish OM, Krabbe LM, et al. Prognostic role of cell cycle and proliferative biomarkers in patients with clear cell renal cell carcinoma. *J Urol* 2013;190:1662–7 [PubMed: 23792148]
37. Haddad AQ, Luo JH, Krabbe LM, Darwish O, Gayed B, Youssef R, et al. Prognostic value of tissue-based biomarker signature in clear cell renal cell carcinoma. *BJU Int* 2017;119:741–7 [PubMed: 28075543]
38. Krabbe LM, Westerman ME, Bagrodia A, Gayed BA, Darwish OM, Haddad AQ, et al. Dysregulation of beta-catenin is an independent predictor of oncologic outcomes in patients with clear cell renal cell carcinoma. *J Urol* 2014;191:1671–7 [PubMed: 24291548]

39. Darwish OM, Kapur P, Youssef RF, Bagrodia A, Belsante M, Alhalabi F, et al. Cumulative number of altered biomarkers in mammalian target of rapamycin pathway is an independent predictor of outcome in patients with clear cell renal cell carcinoma. *Urology* 2013;81:581–6 [PubMed: 23290145]
40. Sivanand S, Peña-Llopis S, Zhao H, Kucejova B, Spence P, Pavia-Jimenez A, et al. A Validated Tumorgraft Model Reveals Activity of Dovitinib Against Renal Cell Carcinoma. *Science Translational Medicine* 2012;4:137ra75–ra75
41. Ricketts CJ, De Cubas AA, Fan H, Smith CC, Lang M, Reznik E, et al. The Cancer Genome Atlas Comprehensive Molecular Characterization of Renal Cell Carcinoma. *Cell Reports* 2018;23:313–26.e5 [PubMed: 29617669]
42. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, et al. The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data. *Cancer Discovery* 2012;2:401–4 [PubMed: 22588877]
43. Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci Signal* 2013;6:p11 [PubMed: 23550210]
44. Ilse M, Tomczak JM, Welling M. Attention-Based Deep Multiple Instance Learning. *arXiv* 2018
45. Tellez D, Litjens G, Bándi P, Bulten W, Bokhorst J-M, Ciompi F, et al. Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. *Medical Image Analysis* 2019;58:101544 [PubMed: 31466046]
46. Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* 2014
47. Vahadane A, Peng T, Sethi A, Albarqouni S, Wang L, Baust M, et al. Structure-Preserving Color Normalization and Sparse Stain Separation for Histological Images. *IEEE Trans Med Imaging* 2016;35:1962–71 [PubMed: 27164577]
48. Caicedo JC, Roth J, Goodman A, Becker T, Karhohs KW, Broisin M, et al. Evaluation of Deep Learning Strategies for Nucleus Segmentation in Fluorescence Images. *Cytometry Part A* 2019;95:952–65
49. 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;25:1301–9 [PubMed: 31308507]
50. Swiderska-Chadaj Z, Hebeda KM, van den Brand M, Litjens G. Artificial intelligence to detect MYC translocation in slides of diffuse large B-cell lymphoma. *Virchows Archiv* 2020

Statement of Significance

This work demonstrates the potential for deep learning analysis of histopathologic images to serve as a fast, low-cost method to assess genetic intratumoral heterogeneity.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

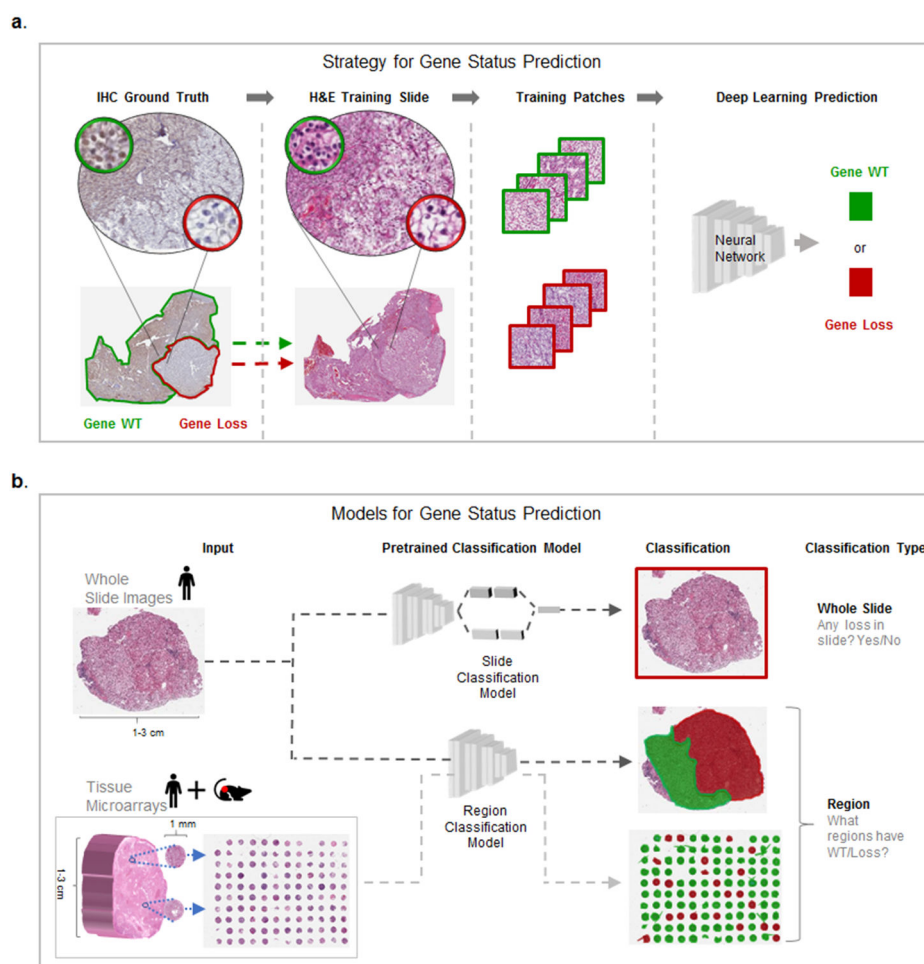


Figure 1. Overview

a. Strategy for training deep learning networks to predict gene mutation status (Loss or WT) from H&E images. Nuclear signal of Immunohistochemistry (IHC) antibodies previously developed to identify the presence/absence of target, serves as a ground truth for gene loss status across different areas of a slide. 224x224-pixel at 20x (111x111 microns) H&E-stained patches assigned ground truth labels based on matching IHC slides are used to train the deep learning networks. **b.** Predictions are performed at two scales: 1) whole slide: where we detect whether any part of the slide (even a minority of tissue), exhibits loss or not, and 2) region: where we predict the areas within a slide that exhibit gene loss (applicable to whole slide images and tumor microarrays).

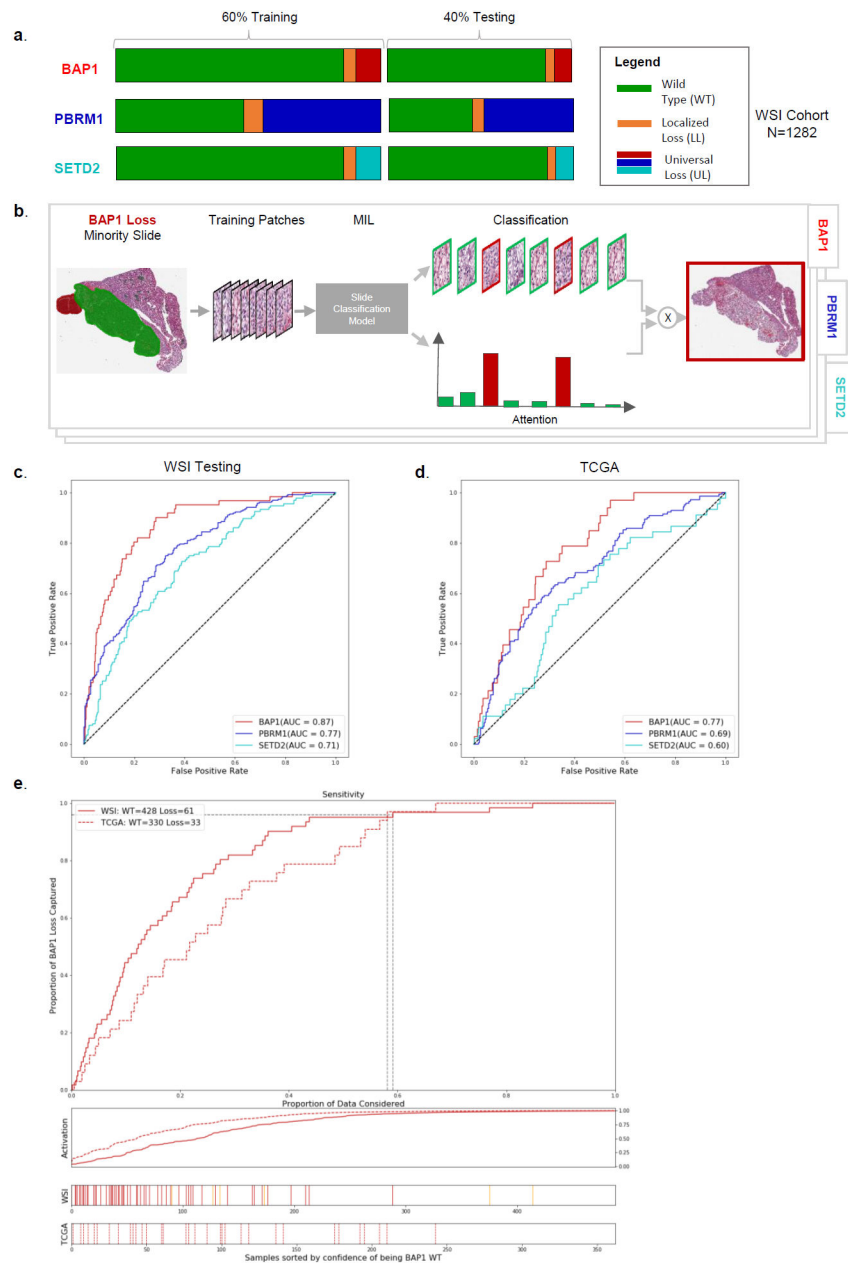


Figure 2. Slide Level Workflow and Performance

a. The WSI cohort was divided into a 60-40 (Training/Testing) split, with similar proportions of Wild Type (WT), Universal Loss (UL), and Localized Loss (LL) samples for all three driver genes in both splits. **b.** A Multiple Instance Learning (MIL) approach was used to train the slide classification model to detect the presence of gene loss even if the loss is only present in a minority of tissue. Patches from the same sample are analyzed individually for their mutation status and their importance to a prediction (based on attention mechanism, a sub-network which learns how much weight to give to each patch, see Methods, Supplementary Fig 4). Predictions from the patches in a slide are aggregated, while accounting for the importance of each patch, to provide a mutational status prediction for the whole slide. Independent models are constructed for each gene. **c.** Receiver operating

characteristic (ROC) curves for three driver mutations- *BAP1*, *PBRM1*, *SETD2*- for whole slide prediction in WSI Testing (i.e., the held-out portion of our WSI cohort). **d**, Same as **c** but using the TCGA cohort. **e**, Proportion of BAP1 Loss samples captured (i.e., sensitivity; Y-axis, top plot) by preferential selection of samples based on the model's assessment of loss likelihood. Shared x-axis represents samples ranked by BAP1 activation, with samples on left mostly likely to exhibit loss for each cohort (WSI-solid, TCGA-dashed). Dashed horizontal and vertical lines indicate 96% sensitivity and the fraction of samples needed to achieve this sensitivity in the two cohorts respectively. The middle plot shows the corresponding activations (Y-axis; middle) with higher values indicating higher model certainty of being WT. Bottom two bars indicate the BAP1 status (white-WT, red-UL, orange-LL only for WSI cohort) of the samples for the WSI and TCGA cohorts respectively.

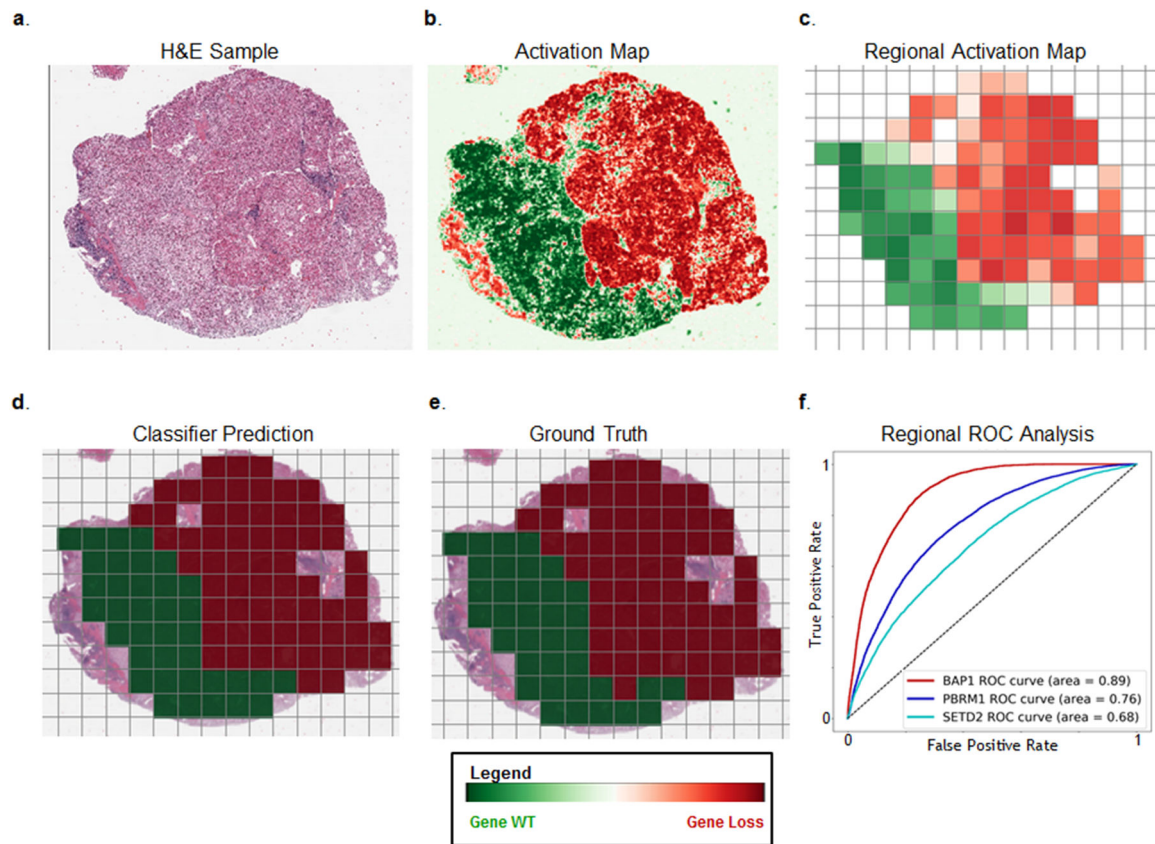


Figure 3. Predicting at high spatial resolution allows us to probe tissue heterogeneity (region level performance)

a-e, Process of BAP1 region classification model demonstrated on a sample slide with both BAP1 WT (green) and Loss (red) regions. **a,** H&E sample input for region classification model. **b,** Activation map from the consensus (i.e., average of multiple models) region classification prediction for the entire slide. **c,** Regional activation map derived from raw activation map. Ignoring any non-tumor tissue, the region scores are calculated by creating a tumor grid of the activation map and averaging the prediction scores within each grid square. **d,** Model predictions: discrete calls (Loss or WT) based on binarizing regional activations. **e,** Ground truth map for BAP1 mutation status on H&E image based on IHC sectioned into a regional grid. **f,** ROC curves from region score analysis for BAP1, PBRM1, and SETD2.

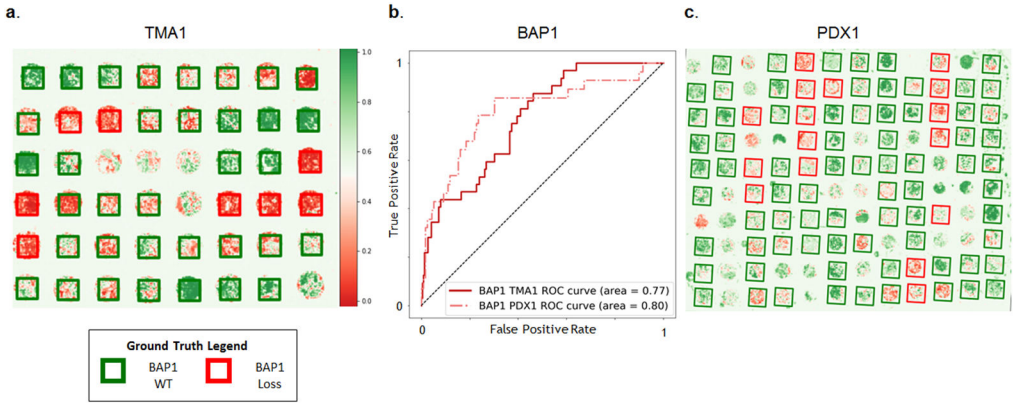


Figure 4. Region classification validation on external cohorts

BAP1 region classification model was applied to two additional cohorts: a human cohort (TMA1) and a PDX cohort (PDX1). **a, c**, Example areas from BAP1 activation maps for region level predictions in TMA1 and PDX1 TMA slides showing consensus prediction scores for each TMA core. The colors of the bounding boxes for each core indicate the gene-loss ground truth acquired from the IHC (cores lacking a box are those that failed quality control). **b**, BAP1 ROC curves for the human TMA1 and the PDX1 samples. Each core contributes a single point to the AUC curve based on the average activation across its all tissue (excludes any background activation). Cores that did not pass our quality control criteria (Methods) (lacking outer box) were not used in the ROC analysis.

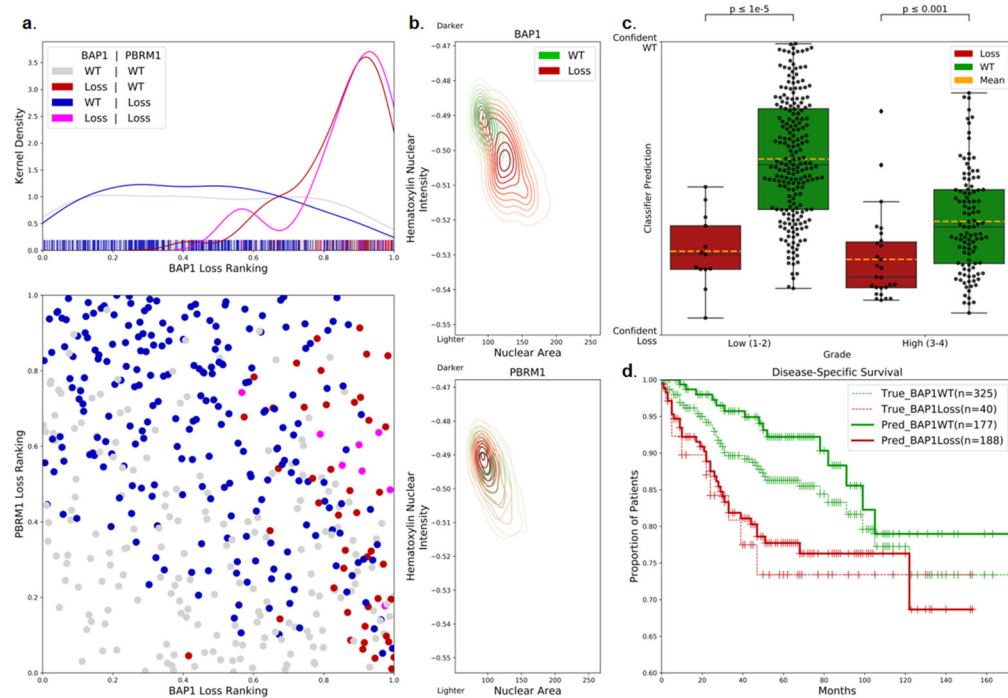


Figure 5. Interpretation of Results

a. Measuring classification independence for BAP1 loss and PBRM1 loss calls in WSI cohort. (BOTTOM) Relationship between BAP1 Loss ranking (X-axis, points on right are highest confidence BAP1 loss) and PBRM1 Loss ranking (Y-axis, points on top are highest confidence PBRM1 loss) based on slide model activations. Each dot represents a single sample color-coded based on its true loss status for BAP1 and PBRM1 as indicated in the legend. (TOP) Distribution of BAP1 confidence rankings for different genotypes. Each line in the rug plot indicates the genotype of a single sample, and the curves represent kernel density estimates of the ranking distribution for a single genotype. **b.** Effect of BAP1 and PBRM1 mutations on selected nuclear features. For each H&E slide in the WSI cohort, nuclei were segmented, nuclear image features were extracted from each nucleus and averaged across all nuclei in tumor region. Density plots show the distribution of average nuclear bounding box area (x-axis, Supp. Fig. 15a) and hematoxylin intensity (y-axis; see Supp Fig. 15d, Supp. Table 4 for details of feature calculation) across Loss and WT forms of BAP1 (top) and PBRM1 (bottom). **c.** BAP1 classification goes beyond tumor grade. Box-whisker plot where each point represents a single patient from the TMA2 cohort, with background bar color reflecting true BAP1 status, and y-axis indicating model assessment (activation) of BAP1 status (higher indicates more likely WT). Points are grouped by grade (low grades 1-2 and high grades 3-4) and BAP1 status (loss/WT). p-values calculated using Mann-Whitney test with Bonferroni correction. **d.** Disease Specific Survival curves stratified by predicted (solid lines) and true (dashed lines) BAP1 mutation status for patients from the TMA2 cohort. True BAP1 Loss (red) or True BAP1 WT (green) was assigned based on patient level IHC assessment. The output of the BAP1 model is an activation score between 0 and 1, rather than a loss status. Patients were classified as Predicted BAP1 Loss

if any punch from that patient exhibited an activation below 0.5 and as Predicted BAP1 WT otherwise.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 1.**Datasets**

Summary of the datasets used in this study. The model is trained on whole-slide-images from the WSI cohort and its performance is benchmarked on data from TCGA and three TMA cohorts, two from human tumors and another from patient-derived xenografts.

Cohort Name	System	Assay	Intra-Sample Gene Loss	Institution	Samples	Notes
WSI (34)	Human	Slide	From IHC Assays	Mayo Clinic	1282 WSIs (775 training, 507 holdout)	Subset of this cohort was used to train all gene models, remainder used as holdout to test performance of all gene models
TCGA (41)	Human	Slide	Unavailable	TCGA	363 WSIs	Slides with whole exome sequencing for all three genes. Used for validation
TMA1 (25)	Human	TMA	From IHC Assays	UTSW	257 Cores	Used to validate region-level BAP1 predictions on tissue microarray cores
PDX1 (40)	PDX	TMA	From IHC Assays	UTSW	147 Cores	Used to test BAP1 human-tissue-trained model performance in PDX tissue microarrays
TMA2 (36)	Human	TMA	From IHC Assays	UTSW	365 Patients, (679 Cores)	Used to relate patient-level BAP1 prediction to clinical readouts