Pan-cancer classifications of tumor histological images using deep learning

Javad Noorbakhsh[1], Saman Farahmand[2], Mohammad Soltanieh-ha[3], Sandeep Namburi[1], Kourosh Zarringhalam[2], Jeff Chuang[1]

[1]The Jackson Laboratory for Genomic Medicine, Farmington, CT; [2]University of Massachusetts Boston, Boston, MA; [3]Boston University, Boston, MA

Abstract

Histopathological images are essential for the diagnosis of cancer type and selection of optimal treatment. However, the current clinical process of manual inspection of images is time consuming and prone to intra- and inter-observer variability. Here we show that key aspects of cancer image analysis can be performed by deep convolutional neural networks (CNNs) across a wide spectrum of cancer types. In particular, we implement CNN architectures based on Google Inception v3 transfer learning to analyze 27815 H&E slides from 23 cohorts in The Cancer Genome Atlas in studies of tumor/normal status, cancer subtype, and mutation status. For 19 solid cancer types we are able to classify tumor/normal status of whole slide images with extremely high AUCs (0.995±0.008). We are also able to classify cancer subtypes within 10 tissue types with AUC values well above random expectations (micro-average 0.87±0.1). We then perform a cross-classification analysis of tumor/normal status across tumor types. We find that classifiers trained on one type are often effective in distinguishing tumor from normal in other cancer types, with the relationships among classifiers matching known cancer tissue relationships. For the more challenging problem of mutational status, we are able to classify TP53 mutations in three cancer types with AUCs from 0.65-0.80 using a fully-trained CNN, and with similar cross-classification accuracy across tissues. These studies demonstrate the power of CNNs for not only classifying histopathological images in diverse cancer types, but also for revealing shared biology between tumors. We have made software available at: https://github.com/javadnoorb/HistCNN

Introduction

Histology image analysis is the gold standard for diagnosis of cancer malignancy and subtypes and plays a major role in selecting specific treatment modalities (He et al. 2012). The current clinical process of image analysis involves manual examination of whole slide images (WSIs) for histopathological features by trained pathologists (Gurcan et al. 2009), a time-consuming process that could benefit from advances in computer-assisted technologies. Although pathologist agreement for severe cases is high, there is significant variability for borderline cases (Allison et al. 2014). For example, the inter-observer agreement for histological evaluation of adenocarcinoma and squamous cell carcinoma has been estimated as κ=0.48-0.64 (Stang et al. 2006), with additional confounding by tumor stage and the evaluator's pulmonary expertise

(Grilley-Olson et al. 2013). New computational approaches to standardize image analysis could be valuable, as large scale cancer image mining (Cooper et al. 2018) has great potential in analogy with pan-cancer genomic data mining (Bailey et al. 2018).

Over the past few years there have been major advances in machine learning models for supervised and unsupervised learning tasks including image analysis and classification (Russakovsky et al. 2015; Litjens et al. 2017). Tumor histopathology image analysis is well-suited to machine learning, as it involves assessments of visual features that are context-dependent and challenging to specify, such as cellular morphology, nuclear structure, and tissue architecture, but which can be learned computationally. Earlier work used support vector machines and random forests for tumor subtype classification and survival outcome analysis (Luo et al. 2017; Yu et al. 2016; Mousavi et al. 2015). However, these approaches use pre-specified image features that do not generalize well across tumor types.

Recent studies have focused on fully-automated approaches using convolutional neural networks (CNNs), bypassing the feature extraction step. For example, Schaumberg et. al., trained an ensemble of ResNet-50 CNNs to predict SPOP mutations using WSIs from 177 prostate cancer patients with 20 positive (Schaumberg, Rubin, and Fuchs 2018), achieving AUC = 0.74 in cross validation and AUC = 0.64 on an independent cohort. Yu et.al., utilized multiple CNN architectures, including AlexNet, GoogLeNet, VGGNet-16 (Simonyan and Zisserman 2014), and the ResNet-50 to identify transcriptomic subtypes of lung adenocarcinoma (LUAD) and squamous cell carcinoma (LUSC) (Yu et al. 2019). They were able to classify LUAD vs. adjacent dense benign tissues (AUC of 0.941-0.965), LUSC vs. adjacent dense benign tissues (AUCs 0.935-0.987), and LUAD vs. LUSC (AUCs 0.883-0.932). Moreover, they were able to predict the TCGA transcriptomic classical, basal, secretory, and primitive subtypes of LUAD (Wilkerson et al. 2010, 2012) with AUCs 0.771-0.892. Recently, Coudray et. al. (Coudray et al. 2018) proposed a CNN based on Google Inception v3 architecture to classify WSIs in LUAD and LUSC using TCGA samples. They achieved an AUC ~ 0.99 in tumor/normal classification using 1,176 tumor and 459 normal WSIs, a significant improvement over non-CNN based models. Using a three-way classifier, their model was able to distinguish LUAD, LUSC, and normal tissues with a macro average AUC ~ 0.97. Further, their models were able to predict mutations in 6 key genes (STK11, EGFR, FAT1, SETBP1, KRAS and TP53) in LUAD with AUC >= 0.74. In a subsequent work (Kim et al. 2019), the same group predicted mutations in BRAF (AUC ~ 0.75) or NRAS (AUC ~ 0.77) on 324 melanoma slides. Other groups have used CNNs to distinguish tumors with high or low mutation burden (Xu et al. 2019). These recent advances highlight the potential of CNNs in computer assisted diagnosis on WSIs.

Still, many open questions and challenges remain. Notably, because prior studies have focused on only a small number of tumor types, it is unclear how broadly such methods can be applied. This is a crucial question, as unique image pre-processing steps and neural network architectures might be needed to understand each cancer type. A systematic investigation across types would therefore be valuable, e.g. using The Cancer Genome Atlas -- a resource with centralized rules for image collection, sequencing and sample processing. Second, pan-cancer cross-tissue comparisons of the efficacy of neural network classifiers could provide important biological insights. This is in analogy with pan-cancer comparisons of sequence features, which have revealed many drivers shared or private to different types of tumors (Martincorena et al. 2017; Hoadley et al. 2018). Shared genomic drivers have motivated the cross-tissue translation of therapies e.g. the application of trastuzumab to HER2-amplified breast and gastric cancers (Bang et al. 2010; Piccart-Gebhart et al. 2005), and shared histopathological features have similar potential.

In this work we undertake such a pan-cancer study of tumor image data, systematically analyzing 27815 whole-slide hematoxylin and eosin (H&E) images from 23 TCGA cohorts. To do this, we have developed image processing and neural network training software applicable across tumor types within an Inception CNN architecture. Using these techniques, first we report on neural networks trained to classify 14,550 images from 19 TCGA cohorts according to their tumor/normal status, showing that almost all cancers can be distinguished from normal with high accuracy. Second, we report on the effectiveness of neural networks trained to classify slides by their cancer subtype. Third, we compare the ability of neural networks trained on one cancer type to classify images from another cancer type. We show that a network trained for one cancer can often classify other cancers as well, and that these relationships overlap with known tissue biology. Finally, we test the ability of neural networks to classify mutation status in selected genes. As part of this we investigate the effect of full training of the network on classification of TP53 mutation status in and across breast, lung, and gastric cancers. We also present software to enable the applications of these methods by other groups. We anticipate that these pan-cancer studies and tools will motivate new imaging-based approaches for biomarker development and treatment design.

Results

Convolutional neural networks for tumor/normal classification

We classified tumor/normal status of slides using a neural network that fed the last fully connected layer of an Inception v3-based CNN pre-trained on ImageNet into a fully connected layer with 1024 neurons (Figure 1a). These two final fully connected layers were trained on tiles of size 512x512 from WSIs. We trained this model separately on 19 TCGA cohorts having numbers of slides ranging from 205-1949 (Figure 2a). 70% of the slides were randomly assigned to the training set and the rest were assigned to the test set. To address the data imbalance problem, which has an appreciable impact on performance (Charte et al. 2015), the majority class was undersampled to match the minority class.

Figure 2b shows the classification results. We first considered classification of individual tiles using labels such that, after removal of background regions, all tiles in a normal image are assumed normal and all tiles in a tumor image are assumed tumor. The CNN accurately classifies the tiles for most tumor types (accuracy: 0.91±0.05, precision: 0.97±0.02, recall: 0.90±0.06, specificity: 0.86±0.07. Mean and std calculated across cohorts). We next examined the fraction of tiles classified as tumor or normal within each slide. The fractions of tiles matching the slide annotation are 0.88±0.14 and 0.90±0.13 for normal and tumor samples, respectively (Figure 2c) (mean and std calculated from all cohorts pooled together). These fractions are high in almost all slides, and the tumor-predicted fraction (TPF) is significantly different between tumors and normals ($p < 0.0001$ per-cohort comparison of tumor vs. normal distributions in Figure 2c; Welch's t-test). We also performed the classification on a per-slide basis. To do this, we used the TPF in each slide as a metric to classify it as tumor or normal. This approach yielded extremely accurate classification results for all cohorts (Figure 2d, mean AUC ROC = 0.995, mean PR AUC = 0.998).

We next investigated whether TPF is a meaningful measure of tumor purity. The distributions of TPF values were generally higher than the pathologist-reported tumor purities available on TCGA (Figure S1). However, we found significant positive correlations between TPF and this ground truth in the majority of cancer types (Figure 2d). These observations can be reconciled by the fact that TPF is based on neoplastic area while the pathologist annotation is based on cell counts, and tumor cells are larger than stromal cells. TPF is therefore a good predictor of tumor purity in most cohorts. Furthermore, we observe a stronger correlation for larger cohorts (e.g. BRCA: p=5e-17), indicating that training on larger datasets may improve development of classifiers for purity estimation. Overall these results suggest that our network can successfully classify WSIs as tumor or normal across many different cancer types. Per-tile classification results can potentially be improved by increasing data and developing intra-slide annotations for regions of interest.

Cancer subtype classification

Next we applied our algorithm to classify tumor slides based on their cancer subtypes (Figure 1a). This analysis was performed on 10 tissues for which pathologist subtype annotation was available on TCGA: sarcoma (SARC), brain (LGG), breast (BRCA), cervix (CESC), esophagus (ESCA), kidney (KIRC/KIRP/KICH), lung (LUAD/LUSC), stomach (STAD), uterine (UCS/UCEC), and testis (TGCT). Cancer subtypes with at least 15 samples were considered, based on TCGA metadata (see Methods). Because comparable numbers of FFPE and flash frozen samples are present in TCGA cohorts (FFPE to frozen slide ratio: 1.0±0.5), both were included in our analysis (Figure 3a).

The following subtypes were considered for each tissue (Figure 3b): brain (oligocytoma, oligodendroglioma, astrocytoma), breast (mucinous, mixed, lobular, ductal), cervix (adenoma, squamous cell carcinoma), esophagus (adenocarcinoma, squamous cell carcinoma), kidney (chromophobe, clear cell, papillary), lung (adenocarcinoma, squamous cell carcinoma), sarcoma (MFS: myxofibrosarcoma, UPS: undifferentiated pleomorphic sarcoma, DDLS: dedifferentiated liposarcoma, LMS: leiomyosarcomas), stomach (diffuse, intestinal), testis (non-seminoma, seminoma), thyroid (tall, follicular, classical), uterine (carcinoma, carcinosarcoma). For the subtype classification task, we used the same CNN model as used for tumor/normal classification; however, for cohorts with more than two subtypes, a multi-class classification was used. Figure 3c and 3d show the per-tile and per-slide classification results (AUC ROCs alongside their micro- and macro-averages). As can be seen, the classifier can accurately identify the subtypes in most cases (AUC micro-average: 0.87±0.1; macro-average: 0.87±0.09). The highest AUC micro/macro-average was achieved in kidney (AUC 0.98), while the lowest is in brain with micro-average 0.60 and macro-average: 0.67. Taken together, these results demonstrate the ability of our classifiers to effectively detect cancer subtypes from WSIs in each tumor type.

Cross-classification relationships between tumor types

Next we used cross-classification to test the hypothesis that different tumor types share CNN-detectable morphological features distinct from those in normal tissues. For each cancer cohort, we re-trained the binary CNN classifier for tumor/normal status using all samples in the cohort. We then tested the ability of each classifier to correctly predict tumor/normal status in the samples from each other cohort. Figure 4 shows a heat map of per-slide ROC AUC for all cross-classifications, along with a hierarchical clustering on the rows and columns of the matrix of AUC values. Surprisingly, most neural networks trained on any single tissue were quite successful in classifying cancer vs. normal in most other tissues (average pairwise AUCs of off-diagonal elements: 0.88±0.11 across all 342 cross-classifications). There were exceptions to this, e.g. LIHC and PAAD performed poorly for multiple cross-classifications. Still, the prevalence of strong cross-classifications supports the common existence of morphological features shared across cancer cohorts but not normal tissues. In particular, classifiers trained on most cohorts successfully predicted tumor/normal status in BLCA (AUC=0.98±0.02), UCEC (AUC=0.97±0.03), and BRCA (AUC=0.97±0.04), suggesting that these cancers most clearly display features universal across types.

To test the biological significance of cross-classification relationships, we assessed associations between tissue of origin (Hoadley et al. 2018) and cross-classification clusters. Specifically, we labeled KIRC/KIRP/KICH as pan-kidney (Ricketts et al. 2018), UCEC/BRCA/OV as pan-gynecological (pan-gyn) (Berger et al. 2018), COAD/READ/STAD as pan-gastrointestinal (pan-GI) (Liu et al. 2018), and LUAD/LUSC as lung. The hierarchical clustering in Figure 4 provides a visualization demonstrating how cohorts of similar tissue of origin cluster closer together. We observed that the lung cohort clusters together on both axes, Pan-GI clusters on the test and partially the train axis, and Pan-Gyn also partially clusters on the test axis. Pan-Kidney partially clusters on both axes. To quantify this, we tested the associations between proximity of cohorts on each axis and similarity of their phenotype (i.e. tissue of origin/adeno-ness). Organ of origin was significantly associated with smaller distances in the hierarchical clustering (p-value=0.002 for test axis and p=0.009 for train axis; Gamma index permutation test. See Methods). We also grouped cohorts by adenocarcinoma/carcinoma status (Figure 4, second row from top), though SARC and ESCA do not fit either category. The cohort distances were significantly associated with adeno-ness on the test axis (p-value=0.008). This indicates that cross-prediction is influenced by commonality in tissue architecture.

We observed other intriguing relationships among cross-tissue classifications as well. Particularly, Pan-GI created a cluster with Pan-Gyn, sarcoma and COAD/READ, suggesting these tumor types have shared features related to malignancy. Likewise, Pan-Kidney and lung also cluster close to each other. On the other hand, the LIHC classifier is very poorly predictive of cancer in the Pan-Kidney cohort. LIHC has an intermediate sample size among the cohorts, indicating this is not due to lack of training data and likely reflects distinct biology.

Fully-trained CNNs for mutation classification

A more challenging task is to predict underlying genetic aberrations from WSIs, as this should be heavily dependent on the aberration. We investigated this question in the multi-cohort context by testing our ability to predict TP53 mutations in TCGA cohorts of BRCA, LUAD and STAD. For this instead of transfer learning we used a fully trained CNN including the Inception model based on an architecture described in (Coudray et al. 2018) (Figure 1b). The cohorts BRCA, LUAD and STAD were selected as they have the highest frequencies of TP53 mutations (Cancer Genome Atlas Research Network 2014a; Cancer Genome Atlas Network 2012; Cancer Genome Atlas Research Network 2014b). The same under-sampling technique as in the tumor/normal classifier was utilized to address data imbalance. The CNN model was trained on 70% of slides and testing was performed on the remaining 30%. Figures 5a and 5b show heatmaps of AUC for the classification results per-tile and per-slide, respectively. Self-cohort redictions (diagonal values) are generally good, with AUC values ranging from 0.65-0.80 for the per-slide and 0.63-0.78 for the per-tile evaluations. Stomach adenocarcinoma was notably more difficult to predict than lung adenocarcinoma (slide AUC=0.80), for which we found AUC values comparable to the AUC=0.76 LUAD results reported by (Coudray et al. 2018). It is noteworthy that the LUAD fully trained network outperformed a transfer learning simulation we implemented for the same data (slide AUC = 0.73), suggesting opportunities for improved training by optimizing early layers of the CNN. We also tested the ability of the model to cross-predict across cohorts in a similar fashion as for the tumor/normal classifiers. Cross-predictions yielded AUC values with a comparable range as the self-cohort analyses (cross-prediction AUCs 0.62-0.72 for slides; 0.60-0.70 for tiles), again supporting shared morphological features across tissues, though self-cohort analyses were slightly more accurate. The cross-classification results also supported the finding that stomach adenocarcinoma was the most difficult set on which to train and test.

Discussion

In this paper we have presented a versatile CNN-based classification framework for pan-cancer analysis of cancer histology images. We were able to train highly effective tumor/normal classifiers in nearly all cancer types (AUC: 0.995±0.008), based on cohorts as small as a few hundred whole slide images. Although we trained based on whole slide annotations, most tiles were classified consistently with their slide annotation, indicating that the classifiers learned global features of the WSIs. Moreover, the fraction of tumor-predicted tiles in each slide was strongly correlated with pathologist-annotated tumor purity, indicating that these CNNs have sub-slide predictive power. We also demonstrated that our algorithms are effective at classifying images based on histopathology subtypes. The accuracy of the model for subtype classification is dependent on the tissue type, but this is expected as genetic drivers of cancer subtypes vary and may manifest differently, and TCGA annotations may vary in quality as well. Notably, subtype classification in kidney is highly accurate, while lower grade glioma subtypes are more difficult to identify.

Remarkably, tumor/normal classifiers developed for most cancer cohorts were effective at distinguishing tumor/normal in most other cancers, indicating that cancers often share histopathological features that can be effectively learned by CNNs. Cancers known to have similar biology, e.g. various types of adenocarcinomas, carcinomas, or cancers with the same tissue of origin, showed commonalities in their trainability and classifiability. This was especially pronounced for UCEC, BLCA, and BRCA, for which classifiers trained on most cohorts could predict their cancer status in these 3 groups. Taken together, our cross classification analysis indicates that there are hierarchies of features that are common across all cancers or subtypes of histologically related cancers.

Our cross-classification analysis also revealed novel pairs of cancer types (Figure 4) likely to have shared predictive histopathological features (e.g. kidney and lung cancers), or that may have unique features (e.g. LIHC). Associating such features to molecular subtypes may provide new insights for shared molecular markers for drug discovery, and also aid in identifying cancers of unknown origin. In the future, using tissue relationships to train neural networks on multiple cancer types may lead to improved classification and provide a more realistic data augmentation procedure. For the TP53 mutation calling, we observed similar behaviors. Although tissues varied in their amenability to the CNN approach (LUAD best, STAD worst), all three tissues showed positive self- and cross-classifiability. This indicates that the mutation classification problem will benefit from the use of tissue relationships as well.

A number of key challenges need further exploration. For example, histopathology datasets are typically highly imbalanced in the samples of each class, which may result in biased predictions. This problem will be critical for analysis of regional variations within slides, e.g. in the analysis of intratumoral heterogeneity with multiple subclones. This is an example of multi-label, multi-instance supervised learning with highly imbalanced data, an active area of research in machine learning (Charte et al. 2015; Read et al. 2015; Li and Wang 2016). Another limitation of deep learning models is their challenge in interpretability, which may slow identification of targetable biological features. Fortunately, there have been rapid developments in interpretation in the last decade (Yosinski et al. 2015; Samek et al. 2017), and integrating the latest advancements into computational pathology applications will be crucial to increasing their utility.

Prior studies have distinguished tumor and normal samples and their subtypes when coupled with detailed spatial annotations from expert pathologists (Wei et al. 2019); however because of the labor-intensive nature of pathological annotation and the engineering effort needed for image data processing, such studies have typically been limited to a single cancer type. Our approach indicates that even with only slide-level annotations, high classification accuracies are possible by leveraging pan-cancer datasets, systematic data processing and job submission, and multi-tissue cross-classification findings. We expect these results will be improved if fine-grained spatial pathological annotations can be generated at scale by the community, e.g. through further curation of TCGA and other cohorts. There is precedent for such community-based effort in computer vision and image analysis (e.g. ImageNet, CIFAR-10). The development of such resources with new computational approaches has the potential to not only accelerate pathological annotation, but also to join histopathology to genomics as pillars of cancer data science.

Materials and Methods

Transfer learning

Sample selection for tumor/normal classification: Since there are very few normal FFPE WSIs on TCGA, we only considered flash frozen samples (with barcodes ending with BS, MS, or TS). We selected 19 TCGA cohorts that had at least 25 normal samples. The samples were randomly divided into 70% training and 30% testing. Stratified sampling was used to balance the ratio of positives and negatives into train and test sets.

Sample selection for subtype classification: WSI images from 10 tissue types were used for subtype classification. FFPE and flash frozen samples are approximately balanced among the tumor WSIs; hence we used both for subtype classification. The samples were randomly divided into 70% training and 30% testing. Some cancer tissues had subtypes that were available as individual cohorts within TCGA. These 3 tissues were LUAD/LUSC (lung); KICH/KIRC/KIRP (kidney); and UCS/UCEC (uterine). For all other tissues, TCGA provided single cohorts that spanned multiple subtypes designated by pathologist annotations. Only clinical subtype annotations with at least 15 samples were considered. Samples with ambiguous or uninformative annotations were not included.

CNN architecture and training:
We used a Google Inception v3-based architecture for pan-cancer tumor/normal classification of TCGA H&E slides. Our CNN architecture uses transfer learning on the Inception module with a modified last layer to perform the classification task. For predicting mutational status we utilize the same architecture as in Coudray et.al. (Coudray et al. 2018) and fully trained the model on TCGA WSIs.

The output of the last fully connected layer of Inception v3 (with 2048 neurons) was fed into a fully connected layer with 1024 neurons. The output was encoded as a one-hot-encoded vector. A softmax function was utilized to generate class probabilities. Each training simulation was run for 2000 steps in batches of 512 samples, with 20% dropout. Mini-batch gradient descent was performed using Adam optimizer (Kingma and Ba 2014). To mitigate the effects of label imbalance in tumor/normal classification, undersampling was performed during training by rejecting inputs from the larger class according to class imbalances, such that, on average, the CNN receives equal number of tumor and normal tiles as input. Per-tile classification ROCs were calculated based on thresholding softmax probabilities and per-slide classification ROCs were based on voting on maximum softmax probability.

Preprocessing and transfer learning steps:

1. Aperio SVS files from primary solid tumors or solid tissue normal samples with 20X or 40X magnification were selected.
2. Each SVS file was randomly assigned to train or test set.
3. 40X images were resized to 20X.
4. Background was removed as in (Coudray et al. 2018).
5. Images were tiled into non-overlapping patches of 512x512 pixels.
6. Tiles were used as inputs of the Inception v3 network (pretrained on ImageNet; downloaded from http://download.tensorflow.org/models/image/imagenet/inception-2015-12-05.tgz), in a forward pass and the values of last fully connected layer ('pool_3/_reshape:0') were stored as 'caches' (vectors of 2048 floating point values).
7. Caches from similar holdout group were shuffled and assigned to a TFRecords in groups of 10,000.
8. TFRecords were used as input to the transfer learning layers.

Programming details: All analysis was performed in Python. Neural network codes were written in TensorFlow (Martín et al. 2016). Images were analyzed using OpenSlide (Goode et al. 2013). Classification metrics were calculated using Scikit-learn (Pedregosa et al. 2011). All transfer learning analysis including preprocessing was performed on the Google Cloud Platform (GCP). The following GCP services were used in our analysis: Kubernetes, Datastore, Cloud Storage, and Pub/Sub. During the preprocessing steps we used up to 1,000 compute instances (each 8 vCPUs and 52GB memory) and up to 4,000 Kubernetes pods. Cloud Storage was used as shared storage system, while Pub/Sub asynchronous messaging service in conjunction with Datastore were used for task distribution and job monitoring of the Kubernetes cluster. This architecture ensures scalability and a fault tolerant process. We leveraged a similar architecture for the pan-cancer training/testing process.

Mutational classification

Sample selection for mutational classification: We selected flash frozen WSIs of BRCA, LUAD and STAD cohorts. Impactful TP53 mutations were determined using masked somatic mutations maf files called by MuTect2 (Cibulskis et al. 2013). Mutations which were categorized as MODERATE/HIGH (by VEP software (McLaren et al. 2016)) in the IMPACT column were considered as impactful mutations. If the gene had at least one impactful mutation in the sample, it was counted as mutated and was considered as wild-type otherwise. Table 1 shows the number of wild type and mutated slides in each cancer type. For cross classification, the model was trained on the entire training cohort and predictions were made on the entire test cohort.

Table 1. Numbers of wild type and mutated slides in each cohort.

Cancer type	No. Wild type slides	No. Mutated slides

BRCA	647	338
LUAD	295	270
STAD	237	200

CNN architecture and training: We utilized the Inception v3 architecture (Coudray et al. 2018) to predict TP53-associated mutations in BRCA, LUAD and STAD cohorts. Unlike the tumor/normal analysis, transfer learning was not used for mutational classifiers. Instead models were fully trained on input slides. As a pre-processing step, we used a fully trained normal/tumor classifier to identify and exclude normal tiles within each tumor slide. This filtering step ensures that tiles with positive mutation class label are also labeled as tumor. To predict mutations in the TP53 gene, we trained 2-way classifiers, assigning 70% of the images in each tissue to training and the remaining 30% to the test set. Cross-tissue mutational classification was performed by training the model on the entire train set of a cohort and performing prediction on other cohorts. The model outputs for tiles were used to produce slide level prediction by averaging probabilities. A similar downsampling as in the tumor/normal classifier was performed to handle data imbalance issues.

Computational configuration: All of the computational tasks for mutation prediction were performed on linux High performance computing clusters with following specification: 8 CPUs, RAM: 64 GB, and Tesla V100 GPUs, 256 GB RAM. Furthermore, The GPU-supported TensorFlow needed CUDA 8.0 Toolkit and cuDNN v5.1. All GPU settings and details were obtained from TensorFlow and TF-slim documentations and NVIDIA GPUs support.

Cross-classification statistics: Hierarchical clustering was applied to cross-classification per-slide AUC values using UPGMA with Euclidean distance. To determine the association between clustering and independent phenotypic labels (i.e. organ and adeno-ness), we used Gamma index of spatial autocorrelation from the Python package PySal (Rey and Anselin 2007). Gamma index is defined as (Hubert, Golledge, and Costanzo 1981):

$$\Gamma = \sum_{i,j} \quad A_{ij}W_{ij},$$

where A is the feature matrix and $W$is the weight matrix, and indices range over cancer cohorts. For each axis and each phenotype group (i.e. organ or adeno-ness), we calculate a separate Gamma index. We define $A_{ij} = 1$ if cohorts i and j have the same phenotype (e.g both are adenocarcinoma) and $A_{ij} = 0$ otherwise. For weights we set $W_{ij} = 1$ if cohorts i and $j$are immediately clustered next to each other and $W_{ij} = 0$ otherwise. P-values are then calculated by permutation test using the PySal package. We dropped any cohort with 'Other' phenotype from this analysis.

Data/Code Availability

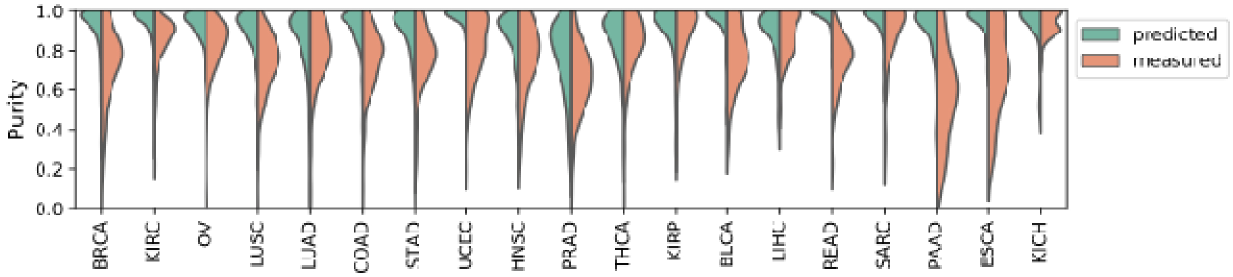
Codes used in this analysis can be found on the following GitHub page: https://github.com/javadnoorb/HistCNN

Acknowledgment

This research benefited from the use of credits from the National Institutes of Health (NIH) Cloud Credits Model Pilot, a component of the NIH Big Data to Knowledge (BD2K) program. This material is based upon work supported by Google Cloud. JHC acknowledges support from NCI grant R01CA230031. JN and JHC would like to thank Yun-Suhk Suh for his helpful comments. JN would also like to thank Ali Foroughipour for his helpful comments regarding statistical analyses.

Supplementary Material

Figure S1: Distribution of tumor purity as predicted by our CNN model compared to the pathologist reports



Figures

Figure 1) Classification pipelines. a) Transfer learning pipeline for tumor/normal and subtype classification b) Full training pipeline for mutation classification

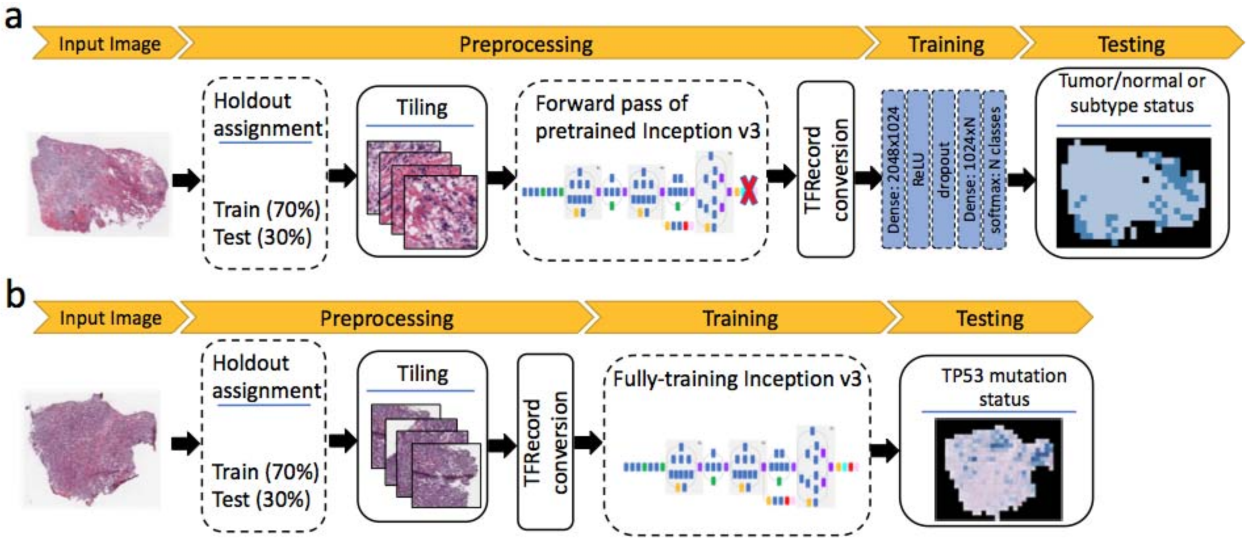


Figure 2) Tumor/normal classification using CNNs. (a) Numbers of tumor and normal slides in test and training sets. (b) Per-tile classification metrics (c) Fraction of tiles within each slide predicted as tumor (d) per slide AUC values for tumor/normal classification for ROC and precision-recall curve (PR) (e) Pearson correlation coefficients between predicted and pathologist evaluation of tumor purity. P-values are based on permutation test of the dependent variable after Bonferroni correction across all cohorts.

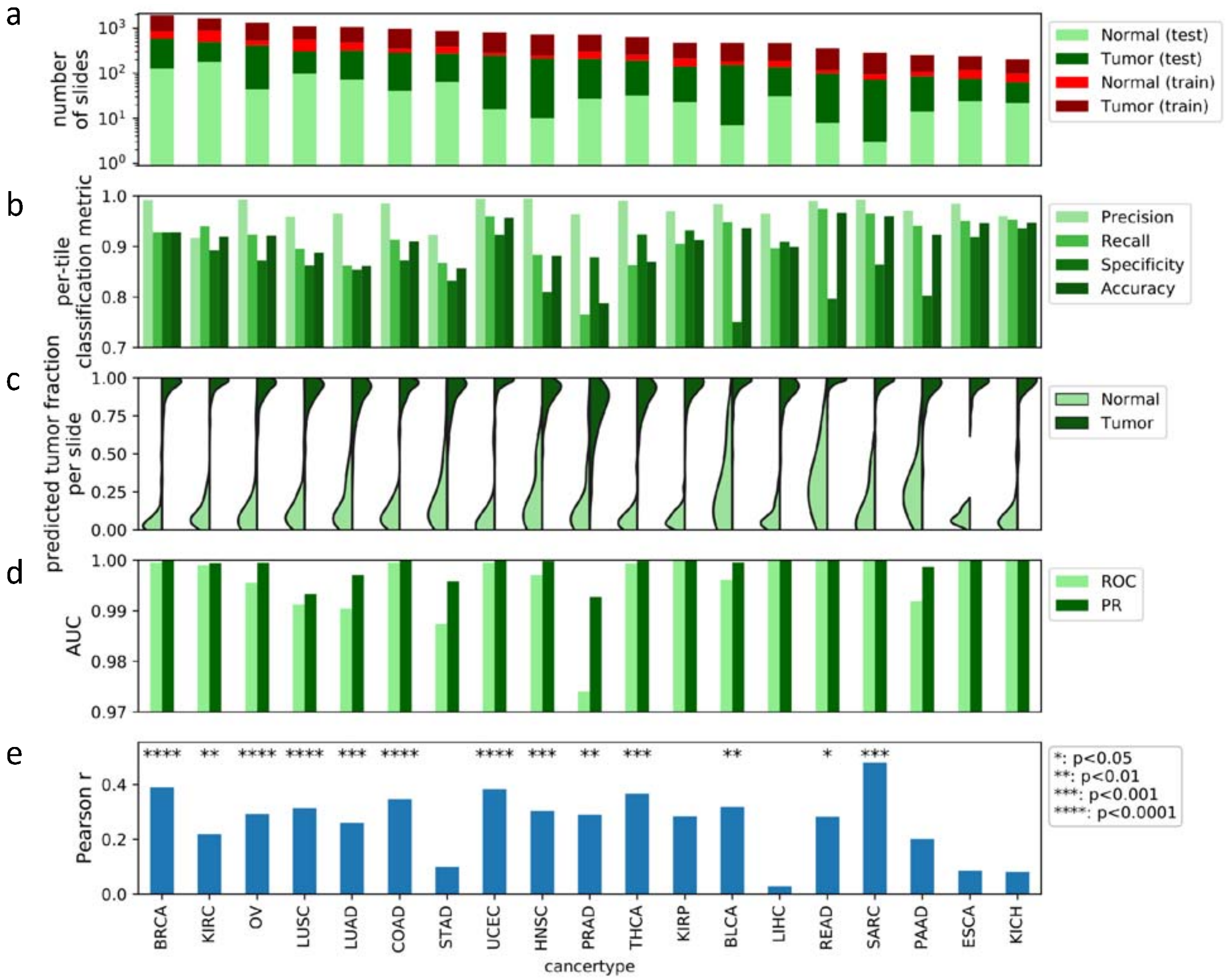


Figure 3) Subtype classification using CNNs. (a) Number of samples used for training. (b) number of samples for each subtype (c) AUC ROC for subtype classification per tile (d) and per slide.

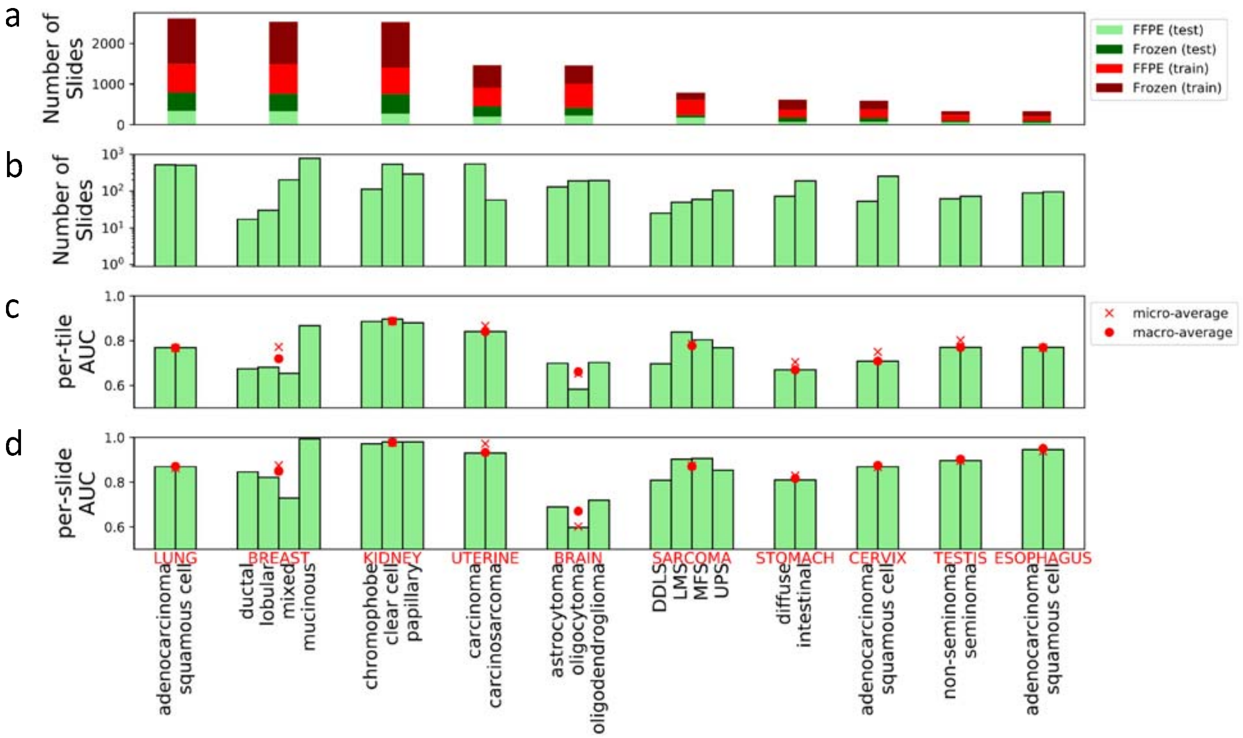


Figure 4) Per-slide AUC values for cross classification of tumor/normal status. The hierarchically clustered heatmap shows pairwise AUC values of CNNs trained on the tumor/normal status of one cohort (train axis) tested on the tumor/normal status of another cohort (test axis). Adeno-ness (adenocarcinoma vs. non-adeno carcinoma) and organ of origin (lung, kidney, gastrointestinal, gynecological) for each cohort are marked with colors on the margins. Cohorts with ambiguous or mixed phenotype are marked as 'Other'.

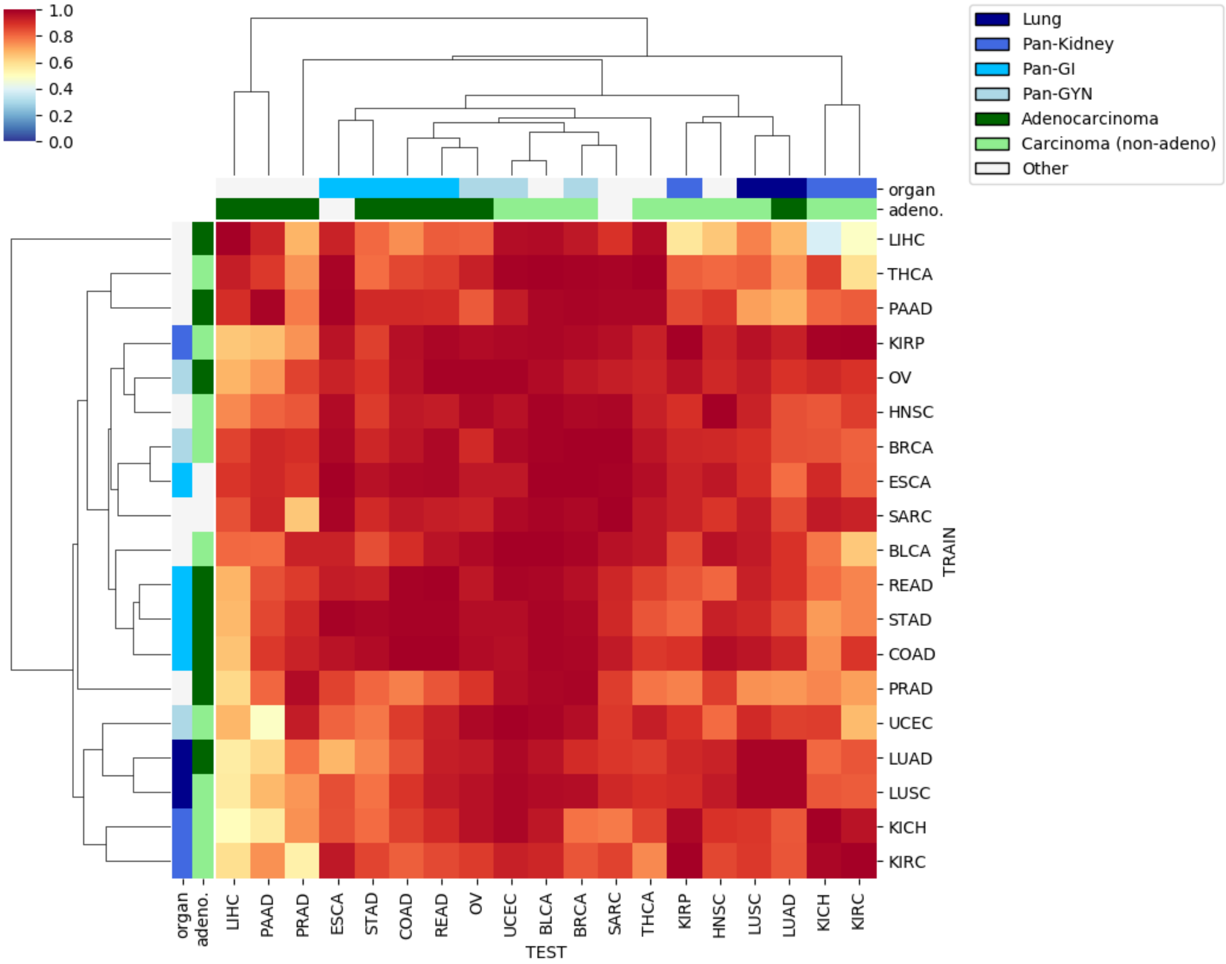
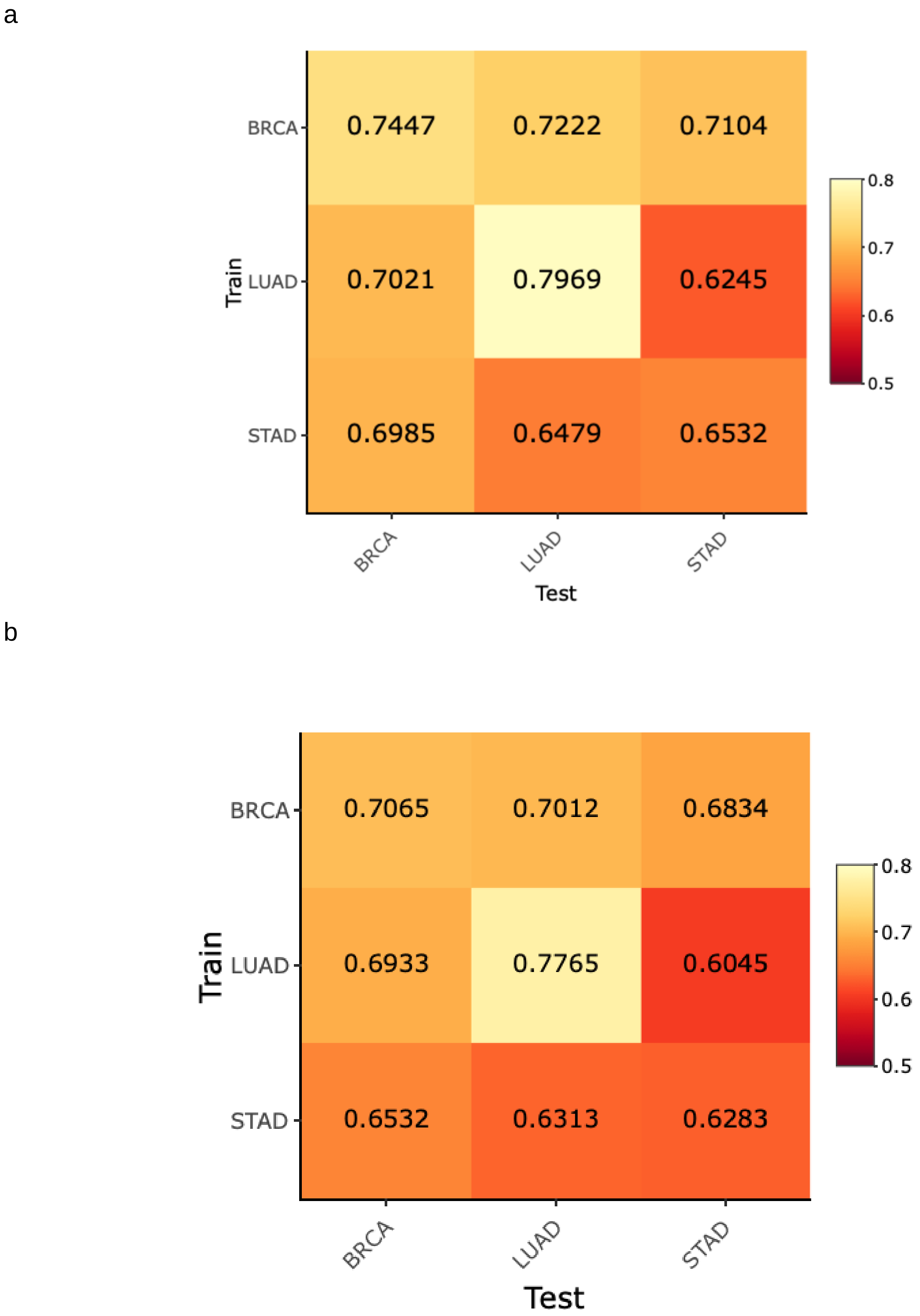


Figure 5) Classification of TP53 mutation status for breast cancer, lung adenocarcinoma, and stomach adenocarcinoma. Cross- and self- classification AUC values from balanced deep learning models (with 95% CIs) are given (a) per-slide and (b) and per-tile.



References

Allison, Kimberly H., Lisa M. Reisch, Patricia A. Carney, Donald L. Weaver, Stuart J. Schnitt, Frances P. O'Malley, Berta M. Geller, and Joann G. Elmore. 2014. "Understanding Diagnostic Variability in Breast Pathology: Lessons Learned from an Expert Consensus Review Panel." *Histopathology* 65 (2): 240–51.

Bailey, Matthew H., Collin Tokheim, Eduard Porta-Pardo, Sohini Sengupta, Denis Bertrand, Amila Weerasinghe, Antonio Colaprico, et al. 2018. "Comprehensive Characterization of Cancer Driver Genes and Mutations." *Cell* 174 (4): 1034–35.

Bang, Yung-Jue, Eric Van Cutsem, Andrea Feyereislova, Hyun C. Chung, Lin Shen, Akira Sawaki, Florian Lordick, et al. 2010. "Trastuzumab in Combination with Chemotherapy versus Chemotherapy Alone for Treatment of HER2-Positive Advanced Gastric or Gastro-Oesophageal Junction Cancer (ToGA): A Phase 3, Open-Label, Randomised Controlled Trial." *The Lancet*. https://doi.org/10.1016/s0140-6736(10)61121-x.

Berger, Ashton C., Anil Korkut, Rupa S. Kanchi, Apurva M. Hegde, Walter Lenoir, Wenbin Liu, Yuexin Liu, et al. 2018. "A Comprehensive Pan-Cancer Molecular Study of Gynecologic and Breast Cancers." *Cancer Cell* 33 (4): 690–705.e9.

Cancer Genome Atlas Network. 2012. "Comprehensive Molecular Portraits of Human Breast Tumours." *Nature* 490 (7418): 61–70.

Cancer Genome Atlas Research Network. 2014a. "Comprehensive Molecular Profiling of Lung Adenocarcinoma." *Nature* 511 (7511): 543–50.

———. 2014b. "Comprehensive Molecular Characterization of Gastric Adenocarcinoma." *Nature* 513 (7517): 202–9.

Charte, Francisco, Antonio J. Rivera, María J. del Jesus, and Francisco Herrera. 2015. "Addressing Imbalance in Multilabel Classification: Measures and Random Resampling Algorithms." *Neurocomputing*. https://doi.org/10.1016/j.neucom.2014.08.091.

Cibulskis, Kristian, Michael S. Lawrence, Scott L. Carter, Andrey Sivachenko, David Jaffe, Carrie Sougnez, Stacey Gabriel, Matthew Meyerson, Eric S. Lander, and Gad Getz. 2013. "Sensitive Detection of Somatic Point Mutations in Impure and Heterogeneous Cancer Samples." *Nature Biotechnology* 31 (3): 213–19.

Cooper, Lee Ad, Elizabeth G. Demicco, Joel H. Saltz, Reid T. Powell, Arvind Rao, and Alexander J. Lazar. 2018. "PanCancer Insights from The Cancer Genome Atlas: The Pathologist's Perspective." *The Journal of Pathology* 244 (5): 512–24.

Coudray, Nicolas, Paolo Santiago Ocampo, Theodore Sakellaropoulos, Navneet Narula, Matija Snuderl, David Fenyö, Andre L. Moreira, Narges Razavian, and Aristotelis Tsirigos. 2018. "Classification and Mutation Prediction from Non–small Cell Lung Cancer Histopathology Images Using Deep Learning." *Nature Medicine* 24 (10): 1559–67.

Grilley-Olson, Juneko E., D. Neil Hayes, Dominic T. Moore, Kevin O. Leslie, Matthew D. Wilkerson, Bahjat F. Qaqish, Michele C. Hayward, et al. 2013. "Validation of Interobserver Agreement in Lung Cancer Assessment: Hematoxylin-Eosin Diagnostic Reproducibility for Non-Small Cell Lung Cancer: The 2004 World Health Organization Classification and Therapeutically Relevant Subsets." *Archives of Pathology & Laboratory Medicine* 137 (1): 32–40.

Gurcan, M. N., L. E. Boucheron, A. Can, A. Madabhushi, N. M. Rajpoot, and B. Yener. 2009.

"Histopathological Image Analysis: A Review." *IEEE Reviews in Biomedical Engineering*. https://doi.org/10.1109/rbme.2009.2034865.

He, Lei, L. Rodney Long, Sameer Antani, and George R. Thoma. 2012. "Histology Image Analysis for Carcinoma Detection and Grading." *Computer Methods and Programs in Biomedicine* 107 (3): 538–56.

Hoadley, Katherine A., Christina Yau, Toshinori Hinoue, Denise M. Wolf, Alexander J. Lazar, Esther Drill, Ronglai Shen, et al. 2018. "Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer." *Cell* 173 (2): 291–304.e6.

Hubert, Lawrence James, Reg G. Golledge, and Carmen M. Costanzo. 1981. "Generalized Procedures for Evaluating Spatial Autocorrelation." *Geographical Analysis* 13 (3): 224–33.

Kim, Randie H., Sofia Nomikou, Zarmeena Dawood, George Jour, Douglas Donnelly, Una Moran, Jeffrey S. Weber, et al. 2019. "A Deep Learning Approach for Rapid Mutational Screening in Melanoma." *bioRxiv*. https://doi.org/10.1101/610311.

Li, Li, and Houfeng Wang. 2016. "Towards Label Imbalance in Multi-Label Classification with Many Labels." http://arxiv.org/abs/1604.01304.

Litjens, Geert, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A. W. M. van der Laak, Bram van Ginneken, and Clara I. Sánchez. 2017. "A Survey on Deep Learning in Medical Image Analysis." *Medical Image Analysis* 42 (December): 60–88.

Liu, Yang, Nilay S. Sethi, Toshinori Hinoue, Barbara G. Schneider, Andrew D. Cherniack, Francisco Sanchez-Vega, Jose A. Seoane, et al. 2018. "Comparative Molecular Analysis of Gastrointestinal Adenocarcinomas." *Cancer Cell* 33 (4): 721–35.e8.

Luo, Xin, Xiao Zang, Lin Yang, Junzhou Huang, Faming Liang, Jaime Rodriguez-Canales, Ignacio I. Wistuba, Adi Gazdar, Yang Xie, and Guanghua Xiao. 2017. "Comprehensive Computational Pathological Image Analysis Predicts Lung Cancer Prognosis." *Journal of Thoracic Oncology: Official Publication of the International Association for the Study of Lung Cancer* 12 (3): 501–9.

Martincorena, Iñigo, Keiran M. Raine, Moritz Gerstung, Kevin J. Dawson, Kerstin Haase, Peter Van Loo, Helen Davies, Michael R. Stratton, and Peter J. Campbell. 2017. "Universal Patterns of Selection in Cancer and Somatic Tissues." *Cell* 171 (5): 1029–41.e21.

McLaren, William, Laurent Gil, Sarah E. Hunt, Harpreet Singh Riat, Graham R. S. Ritchie, Anja Thormann, Paul Flicek, and Fiona Cunningham. 2016. "The Ensembl Variant Effect Predictor." *Genome Biology* 17 (1): 122.

Mousavi, Hojjat Seyed, Vishal Monga, Ganesh Rao, and Arvind U. K. Rao. 2015. "Automated Discrimination of Lower and Higher Grade Gliomas Based on Histopathological Image Analysis." *Journal of Pathology Informatics* 6 (March): 15.

Piccart-Gebhart, Martine J., Marion Procter, Brian Leyland-Jones, Aron Goldhirsch, Michael Untch, Ian Smith, Luca Gianni, et al. 2005. "Trastuzumab after Adjuvant Chemotherapy in HER2-Positive Breast Cancer." *The New England Journal of Medicine* 353 (16): 1659–72.

Read, Jesse, Luca Martino, Pablo M. Olmos, and David Luengo. 2015. "Scalable Multi-Output Label Prediction: From Classifier Chains to Classifier Trellises." *Pattern Recognition*. https://doi.org/10.1016/j.patcog.2015.01.004.

Rey, Sergio J., and Luc Anselin. 2007. "PySAL: A Python Library of Spatial Analytical Methods." *The Review of Regional Studies* 37 (1): 5–27.

Ricketts, Christopher J., Aguirre A. De Cubas, Huihui Fan, Christof C. Smith, Martin Lang, Ed Reznik, Reanne Bowlby, et al. 2018. "The Cancer Genome Atlas Comprehensive Molecular Characterization of Renal Cell Carcinoma." *Cell Reports* 23 (12): 3698.

Russakovsky, Olga, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, et al. 2015. "ImageNet Large Scale Visual Recognition Challenge." *International Journal of Computer Vision* 115 (3): 211–52.

Samek, Wojciech, Alexander Binder, Gregoire Montavon, Sebastian Lapuschkin, and Klaus-

Robert Muller. 2017. “Evaluating the Visualization of What a Deep Neural Network Has Learned.” *IEEE Transactions on Neural Networks and Learning Systems* 28 (11): 2660–73.

Schaumberg, Andrew J., Mark A. Rubin, and Thomas J. Fuchs. 2018. “H&E-Stained Whole Slide Image Deep Learning Predicts SPOP Mutation State in Prostate Cancer.” *bioRxiv*. https://doi.org/10.1101/064279.

Simonyan, Karen, and Andrew Zisserman. 2014. “Very Deep Convolutional Networks for Large-Scale Image Recognition.” *arXiv [cs.CV]*. arXiv. http://arxiv.org/abs/1409.1556.

Stang, Andreas, Hermann Pohlabeln, Klaus M. Müller, Ingeborg Jahn, Klaus Giersiepen, and Karl-Heinz Jöckel. 2006. “Diagnostic Agreement in the Histopathological Evaluation of Lung Cancer Tissue in a Population-Based Case-Control Study.” *Lung Cancer* 52 (1): 29–36.

Wei, Jason W., Laura J. Tafe, Yevgeniy A. Linnik, Louis J. Vaickus, Naofumi Tomita, and Saeed Hassanpour. 2019. “Pathologist-Level Classification of Histologic Patterns on Resected Lung Adenocarcinoma Slides with Deep Neural Networks.” *Scientific Reports* 9 (1): 3358.

Wilkerson, Matthew D., Xiaoying Yin, Katherine A. Hoadley, Yufeng Liu, Michele C. Hayward, Christopher R. Cabanski, Kenneth Muldrew, et al. 2010. “Lung Squamous Cell Carcinoma mRNA Expression Subtypes Are Reproducible, Clinically Important, and Correspond to Normal Cell Types.” *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research* 16 (19): 4864–75.

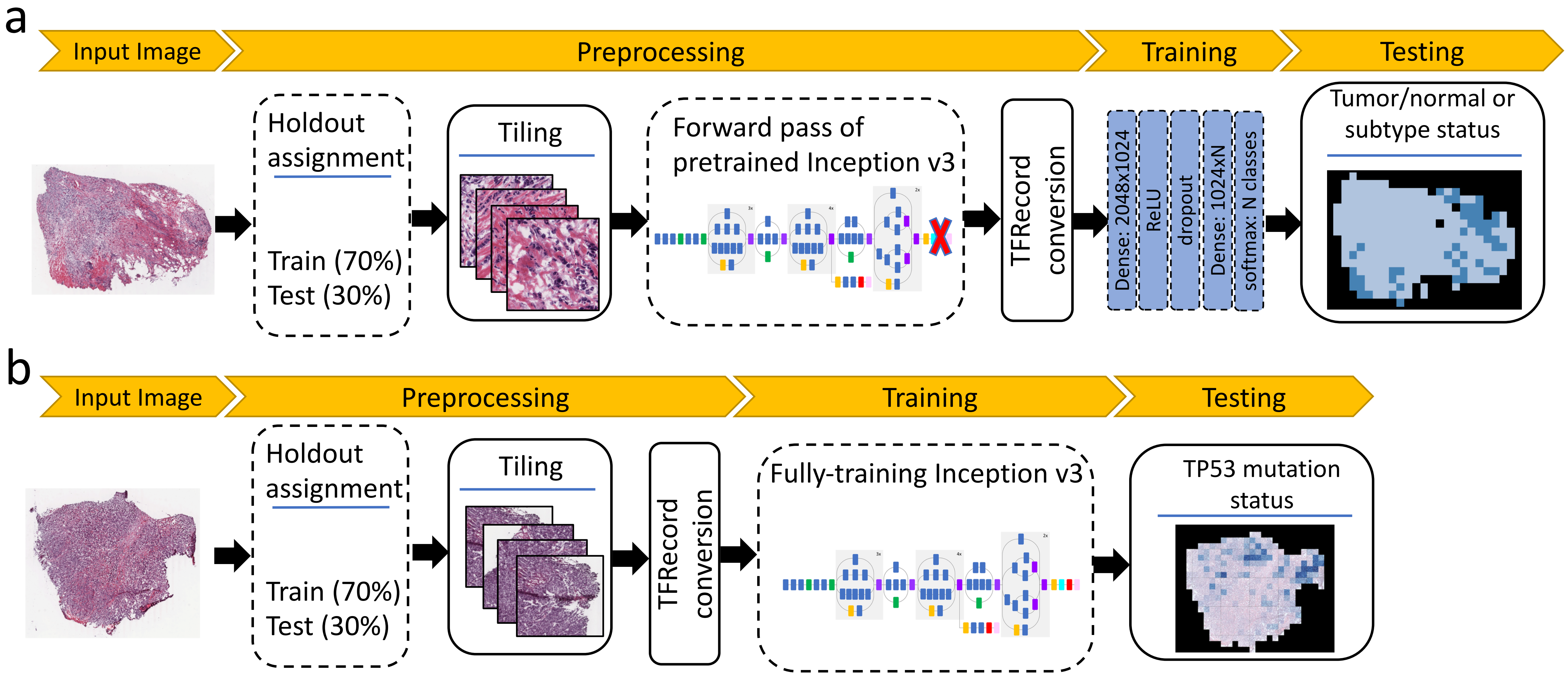
Wilkerson, Matthew D., Xiaoying Yin, Vonn Walter, Ni Zhao, Christopher R. Cabanski, Michele C. Hayward, C. Ryan Miller, et al. 2012. “Differential Pathogenesis of Lung Adenocarcinoma Subtypes Involving Sequence Mutations, Copy Number, Chromosomal Instability, and Methylation.” *PloS One* 7 (5): e36530.

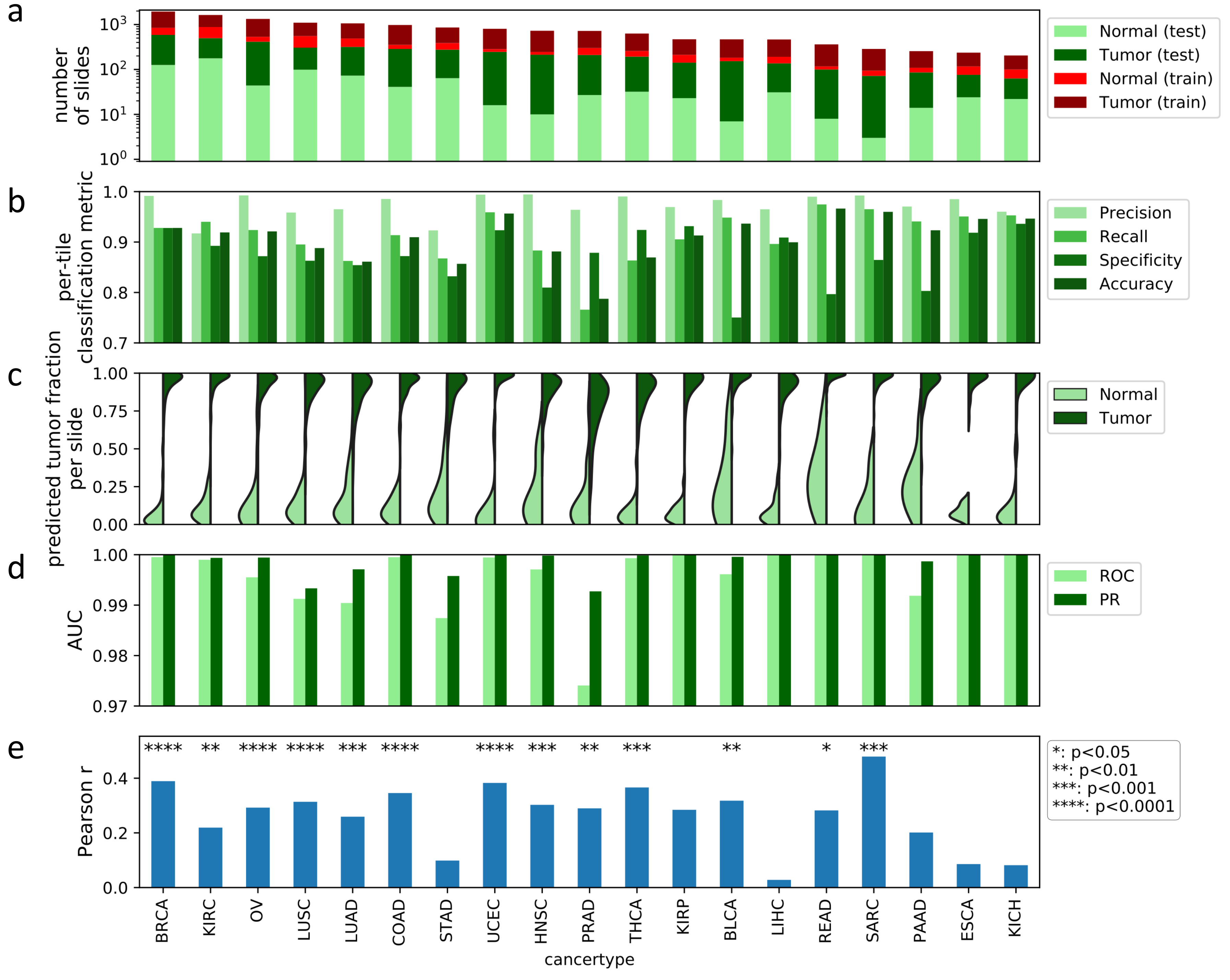
Xu, Hongming, Sunho Park, Sung Hak Lee, and Tae Hyun Hwang. 2019. “Using Transfer Learning on Whole Slide Images to Predict Tumor Mutational Burden in Bladder Cancer Patients.” *bioRxiv*. https://doi.org/10.1101/554527.

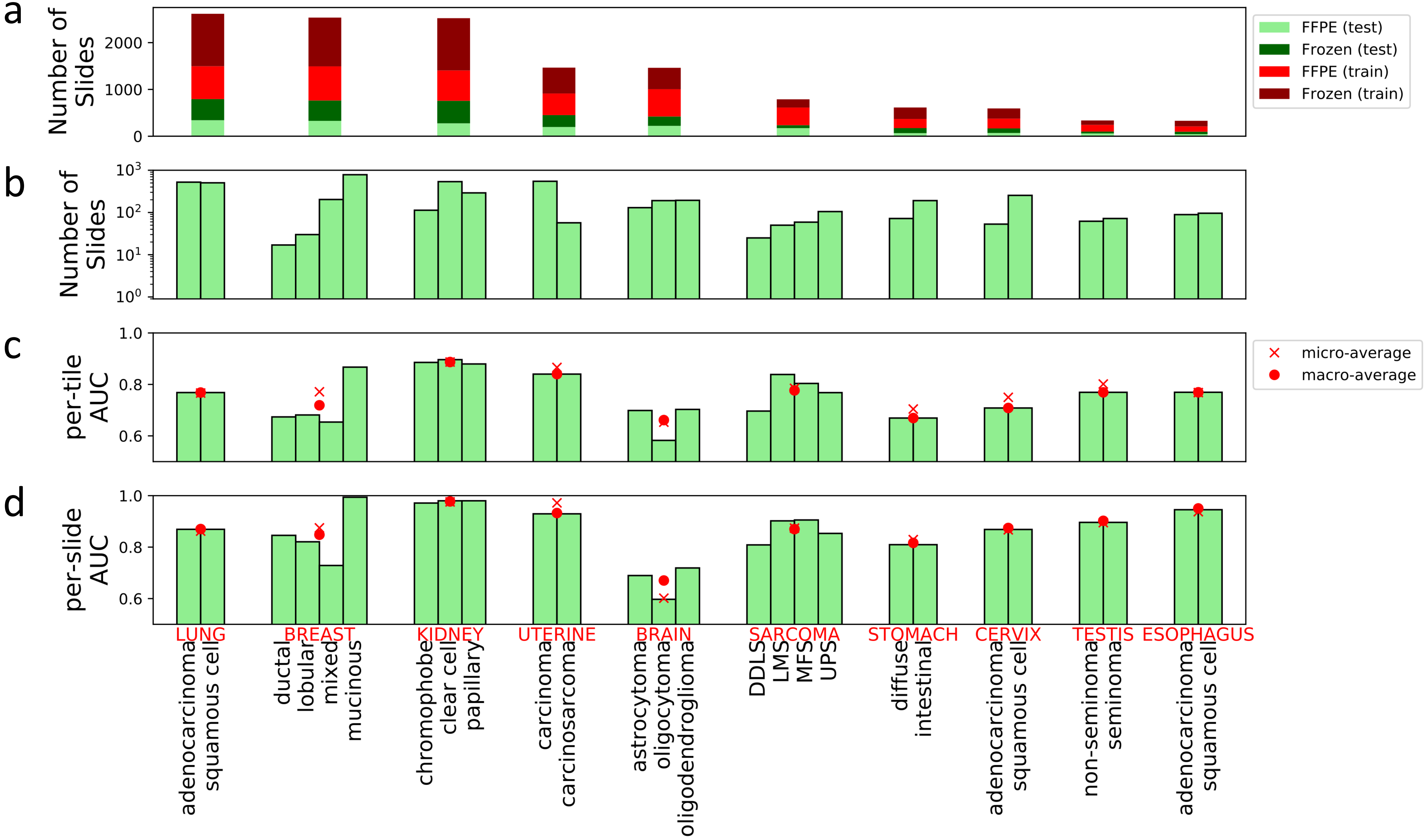
Yosinski, Jason, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson. 2015. “Understanding Neural Networks Through Deep Visualization.” *arXiv [cs.CV]*. arXiv. http://arxiv.org/abs/1506.06579.

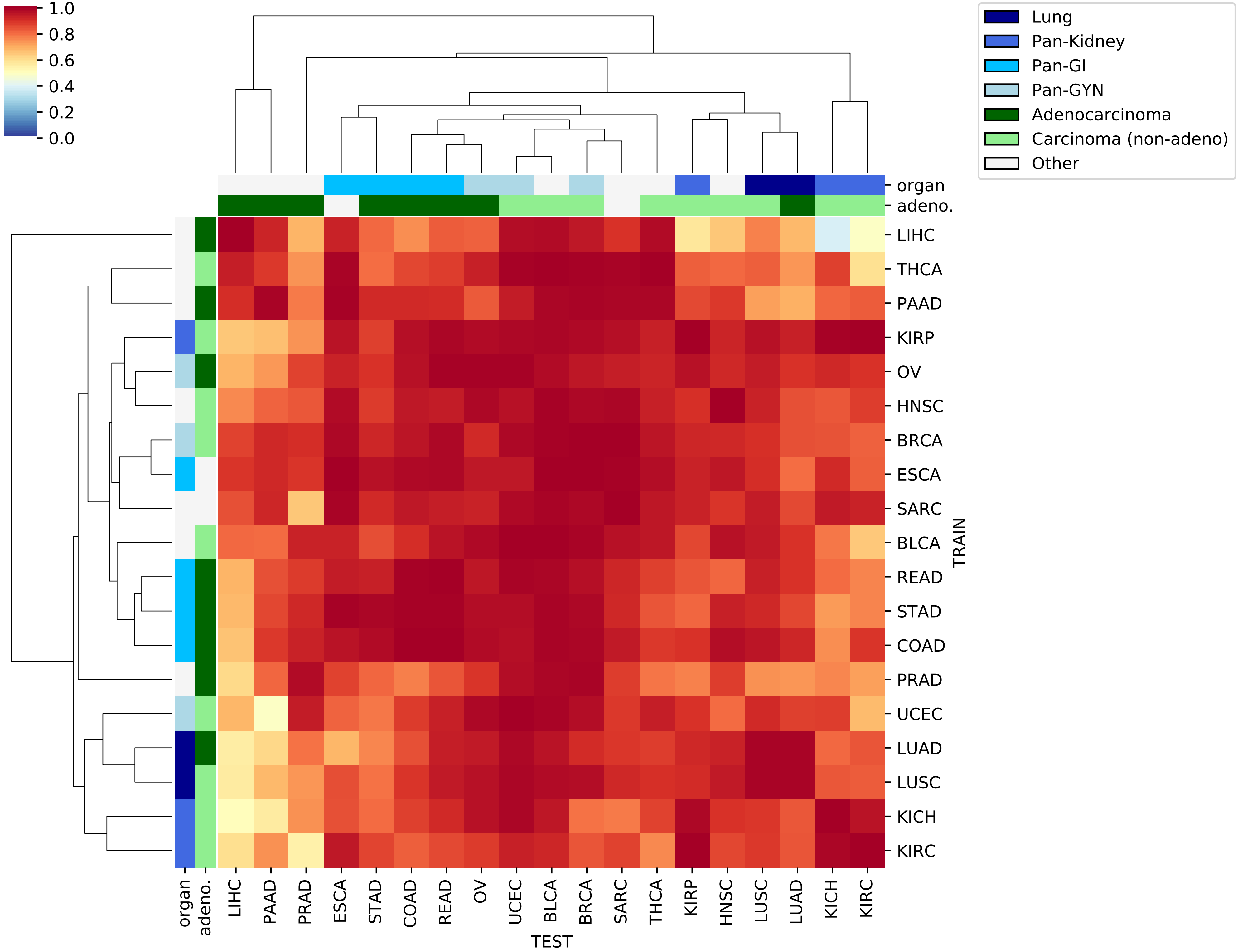
Yu, Kun-Hsing, Feiran Wang, Gerald J. Berry, Christopher Re, Russ B. Altman, Michael Snyder, and Isaac S. Kohane. 2019. “Classifying Non-Small Cell Lung Cancer Histopathology Types and Transcriptomic Subtypes Using Convolutional Neural Networks.” *bioRxiv*. https://doi.org/10.1101/530360.

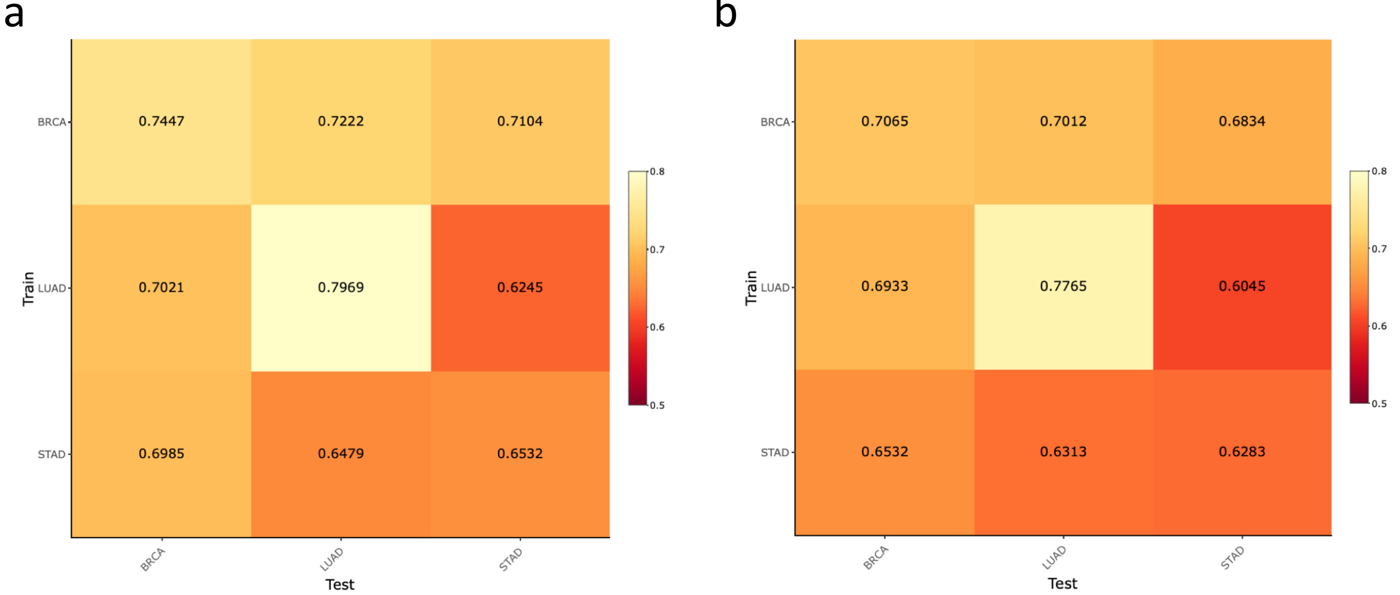
Yu, Kun-Hsing, Ce Zhang, Gerald J. Berry, Russ B. Altman, Christopher Ré, Daniel L. Rubin, and Michael Snyder. 2016. “Predicting Non-Small Cell Lung Cancer Prognosis by Fully Automated Microscopic Pathology Image Features.” *Nature Communications* 7 (August): 12474.


a
Input Image
Preprocessing
Training
Testing
Holdout assignment
Train (70%)
Test (30%)
Tiling
Forward pass of pretrained Inception v3
3x
4x
2x
TFRecord conversion
Dense: 2048x1024
ReLU
dropout
Dense: 1024xN
softmax: N classes
Tumor/normal or subtype status
b
Input Image
Preprocessing
Training
Testing
Holdout assignment
Train (70%)
Test (30%)
Tiling
TFRecord conversion
Fully-training Inception v3
3x
4x
2x
TP53 mutation status


a
number of slides
10^3
10^2
10^1
10^0
Normal (test)
Tumor (test)
Normal (train)
Tumor (train)
b
per-tile classification metric
1.0
0.9
0.8
0.7
Precision
Recall
Specificity
Accuracy
c
predicted tumor fraction per slide
1.00
0.75
0.50
0.25
0.00
Normal
Tumor
d
AUC
1.00
0.99
0.98
0.97
ROC
PR
e
Pearson r
0.4
0.2
0.0
**** ** **** **** *** **** **** *** ** *** ** * ***
*: p<0.05
**: p<0.01
***: p<0.001
****: p<0.0001
BRCA
KIRC
OV
LUSC
LUAD
COAD
STAD
UCEC
HNSC
PRAD
THCA
KIRP
BLCA
LIHC
READ
SARC
PAAD
ESCA
KICH
cancertype


a
Number of Slides
0
1000
2000
FFPE (test)
Frozen (test)
FFPE (train)
Frozen (train)
b
Number of Slides
10^0
10^1
10^2
10^3
c
per-tile AUC
0.6
0.8
1.0
micro-average
macro-average
d
per-slide AUC
0.6
0.8
1.0
LUNG
adenocarcinoma
squamous cell
BREAST
ductal
lobular
mixed
mucinous
KIDNEY
chromophobe
clear cell
papillary
UTERINE
carcinoma
carcinosarcoma
BRAIN
astrocytoma
oligocytoma
oligodendroglioma
SARCOMA
DDLS
LMS
MFS
UPS
STOMACH
diffuse
intestinal
CERVIX
adenocarcinoma
squamous cell
TESTIS
non-seminoma
seminoma
ESOPHAGUS
adenocarcinoma
squamous cell


1.0
0.8
0.6
0.4
0.2
0.0
Lung
Pan-Kidney
Pan-GI
Pan-GYN
Adenocarcinoma
Carcinoma (non-adeno)
Other
organ
adeno.
LIHC
THCA
PAAD
KIRP
OV
HNSC
BRCA
ESCA
SARC
BLCA
READ
STAD
COAD
PRAD
UCEC
LUAD
LUSC
KICH
KIRC
TRAIN
organ
adeno.
LIHC
PAAD
PRAD
ESCA
STAD
COAD
READ
OV
UCEC
BLCA
BRCA
SARC
THCA
KIRP
HNSC
LUSC
LUAD
KICH
KIRC
TEST


a
BRCA
LUAD
STAD
Train
0.7447
0.7222
0.7104
0.7021
0.7969
0.6245
0.6985
0.6479
0.6532
BRCA
LUAD
STAD
Test
0.8
0.7
0.6
0.5
b
BRCA
LUAD
STAD
Train
0.7065
0.7012
0.6834
0.6933
0.7765
0.6045
0.6532
0.6313
0.6283
BRCA
LUAD
STAD
Test
0.8
0.7
0.6
0.5