H&E-stained Whole Slide Deep Learning Predicts SPOP Mutation State in Prostate Cancer

Andrew J. Schaumberg[a,b,*], **Mark A. Rubin**[c,d,e], **and Thomas J. Fuchs**[f,g,*]

[a]Memorial Sloan Kettering Cancer Center and the Tri-Institutional Training Program in Computational Biology and Medicine; [b]Weill Cornell Graduate School of Medical Sciences; [c]Caryl and Israel Englander Institute for Precision Medicine, New York Presbyterian Hospital–Weill Cornell Medicine; [d]Sandra and Edward Meyer Cancer Center at Weill Cornell Medicine; [e]Department of Pathology and Laboratory Medicine, Weill Cornell Medicine; [f]Department of Medical Physics, Memorial Sloan Kettering Cancer Center; [g]Department of Pathology, Memorial Sloan Kettering Cancer Center

This manuscript was compiled on July 16, 2016

The genetic basis of histological phenotype is well known in molecular pathology, such as CDH1 loss leading to a lobular rather than ductal phenotype in breast cancer. Unfortunately, these phenotypes are qualitative. Moreover, a single genetic alteration may evidence multiple and non-unique histologic features, such as TMPRSS2-ERG fusion giving rise to macronuclei, blue-tinged mucin, and cribriform pattern. A quantitative model to genetically interpret the histology is desirable to guide downstream immunohistochemistry, genomics, and precision medicine. We constructed a statistical model that predicts whether or not SPOP is mutated in prostate cancer, given only the digital whole slide after standard hematoxylin and eosin [H&E] staining. Using a cohort of 177 prostate cancer patients where 20 had mutant SPOP, we trained multiple ensembles of residual networks, accurately distinguishing SPOP mutant from SPOP wild type patients. To our knowledge, this is the first statistical model to predict a genetic mutation in cancer directly from the patient's digitized H&E-stained whole microscope slide.

cancer | molecular pathology | deep learning | whole slide image

Genetic drivers of cancer morphology, such as CDH1 loss promoting lobular rather than ductal phenotypes in breast, are well known. SPOP mutation in prostate cancer has a number of known morphological traits, including blue-tinged mucin, cribriform pattern, and macronuclei [1]. Computational pathology methods [2] typically predict clinical or genetic features as a function of histological imagery, e.g. whole slide images. Our central hypothesis is that the morphology shown in these whole slide images, having nothing more than standard hematoxylin and eosin staining, is a function of the underlying genetic drivers. To test this hypothesis, we gathered a cohort of 499 prostate adenocarcinoma patients from The Cancer Genome Atlas [TCGA][1], 177 of which were suitable for analysis, with 20 of those having mutant SPOP (Fig 1). We then used ensembles of deep convolutional neural networks to accurately predict whether or not SPOP was mutated in the patient, given only the patient's whole slide image. Our classifier's generalization error bounds and receiver operating characteristic support our hypothesis, in agreement with earlier work suggesting SPOP mutants are a distinct subtype of prostate cancer [3].

Prior analyses have manually derived morphological features from slide images with the help of trained pathologists, then correlated these features to genetic aberrations, such as mutations and copy number alterations in a variety of genes [4, 5]. In contrast, our deep learning approach learns features automatically and analyzes the images directly, without a pathologist, using a single mutation from whole exome sequencing as a class label for an image, in a completely automated pipeline that predicts a 95% confidence interval of the mutation probability, representing both the mutation prediction and its uncertainty.

Others used support vector machine techniques to predict molecular subtypes in a bag-of-features approach over Gabor filters [6]. The authors note the similarity to convolutional features in deep learning methods, such as ours, but did not pursue deep learning due to limited data available. Gabor filters are similar to the first layer in a deep convolutional neural network, with more complex features learned in deeper layers of the network, and each of our networks is at least 50 layers deep. We address limited data through pre-training, Monte Carlo cross validation, and ensemble methods to support deep learning. Our method also predicts the state of a single mutation, which we believe to be more clinically actionable for precision medicine, rather than predicting broad gene expression subtypes.

Support vectors machines, Gaussian mixture modeling, and principal component analyses have predicted PTEN deletion and copy number variation in cancer, but relied on fluorescence *in situ* hybridization [FISH], a very specific stain [7]. Our approach uses standard hematoxylin and eosin, a non-specific stain that we believe could be utilized to predict more genetic aberrations than only the SPOP mutation that is our focus here. However, our method does not quantify tumor

DRAFT

Significance Statement

The authors present the first automatic pipeline predicting gene mutation probability in cancer from digitized light microscopy slides having standard hematoxylin and eosin staining. To predict whether or not the speckly-type POZ protein [SPOP] gene is mutated in prostate cancer, the pipeline (i) identifies diagnostically salient regions in the whole slide at low magnification, (ii) identifies the salient region having the dominant tumor, (iii) within this region finds the high magnification subregion most enriched for abnormal cells, and (iv) trains an ensemble of deep learning binary classifiers that together predict a 95% confidence interval of mutation probability. This work enables fully-automated histologic diagnoses based on probabilities of underlying molecular aberrations. Such probabilities may directly guide immunohistochemistry choices, genetic tests, and precision medicine.

Conception, Writing: AJS, TJF. Algorithms, Learning: AJS. Analysis: AJS, MAR, TJF. Supervision: MAR, TJF.

Authors declare no conflicts of interest.

[1]TCGA data courtesy the TCGA Research Network http://cancergenome.nih.gov/

*To whom correspondence should be addressed. E-mail: ajs625@cornell.edu fuchst@mskcc.org

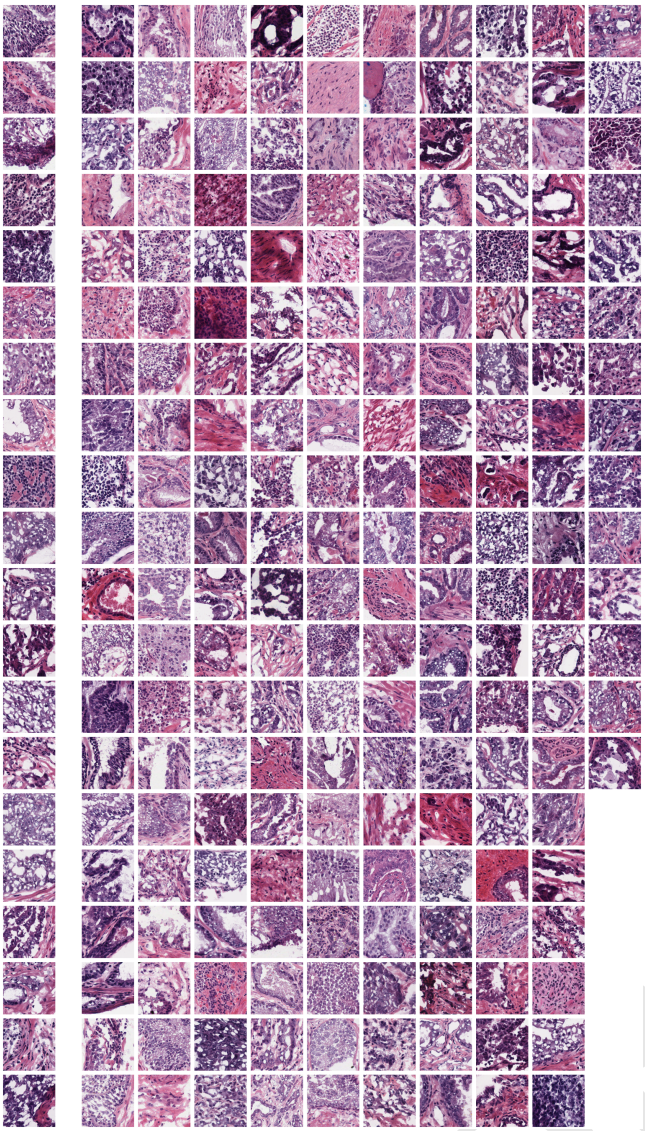

Fig. 1. Twenty SPOP mutants shown in left column. Remaining right are 157 SPOP wild type patients, where 25 patients had 2 and 6 patients had 3 acceptable slides available. Unacceptable slides discarded by Algorithm 1.

heterogeneity.

Tumor heterogeneity has been analyzed statistically by other groups [8], as a spatial analysis of cell type entropy cluster counts in H&E-stained whole slides. A high count, when combined with genetic mutations such as TP53, improves patient survival prediction. This count is shown to be independent of underlying genetic characteristics, whereas our method predicts a genetic characteristic (SPOP mutation) from convolutional features of the imaging.

Histological review to identify patients having a specific genetic mutation for the development of targeted inhibitors has been previously proposed in the context of fibroblast growth factor receptor 3 [FGFR3] mutations in bladder cancer [9]. FGFR3 mutants revealed unique histological features, this represents a potential therapeutic target, and follow-up confirmatory genotyping is suggested. Our approach with SPOP mutants in prostate cancer is similar, with deep learning ensembles in place of pathologists to analyze the slide.

Results

Molecular information as labels of pathology images opens a new field of molecular pathology. Rather than correlating or combining genetic and histologic data, we predict a gene mutation directly from a whole slide image having unbiased H&E stains. This not only opens the door to systematic investigation of other genotype and phenotype relationships on the basis of deep learning models such as ours, but serves as a new supervised learning paradigm for molecular targets, independent of clinician-supplied labels of the histology. Epigenetic, copy number alteration, gene expression, and post-translational modification data may all be brought to bear as labels for histology. Deep learning ensemble approaches such as ours may then learn to predict these labels from images. Future work may refine these predictions to single-cell resolution, combined with single-cell sequencing, immunohistochemistry, fluorescence *in situ* hybridization, or other technologies to label corresponding H&E images. We suggest focusing on labels that are clinically actionable, such as gains or losses of function. Soon we hope to interpret our deep network's predictions through DeepVis or similar tool [10].

SPOP mutation state prediction is learnable from whole slides stained with hematoxylin and eosin. Our pipeline learns to predict SPOP mutations, separating true positives from false positives at an efficiency well above chance (AUROC=0.7128, p=0.000122, Fig 2). To our knowledge this is the first time a mutation in cancer has been predicted through deep learning on whole slides with standard H&E staining.

SPOP mutation state prediction is accurate and the prediction uncertainty is bounded within a confidence interval. Our pipeline's accuracy in SPOP mutation state prediction is statistically significant (mean=0.7, 95% CI 0.5785-0.8241, Fig 3). The confidence interval calculated using five resnet ensembles allows every prediction to be evaluated statistically: is there significant evidence for SPOP mutation in the patient, is there significant evidence for SPOP wild type in the patient, or is the patient's SPOP mutation state inconclusive.

SPOP mutation state prediction is fully automated does not rely on human interpretation. Unlike Gleason score, which relies on a clinician's interpretation of the histology, our pipeline is fully automated (Fig 4). The pipeline's input is simply the whole digital slide and the output is the probability that SPOP is mutated, with the probability bounded within a confidence interval.

Molecular pathology, such as characterizing histology in terms of SPOP mutation state, leads directly to precision medicine. Predictions of the underlying genetic state proceeds directly to precision medicine. For instance, wild type SPOP ubiquitinylates androgen receptor [AR], to mark AR for degradation, but mutant SPOP does not. Antiandrogen drugs, such as flutamide, promote degradation of excess AR to treat the cancer, though mutant AR confers resistance [15, 16].

SPOP mutation state prediction provides information regarding other mutational states. SPOP mutation is mutually exclusive with TMPRSS2-ERG gene fusion [3], so our SPOP mutation predictor provides indirect information regarding the TMPRSS2-ERG state.

DRAFT

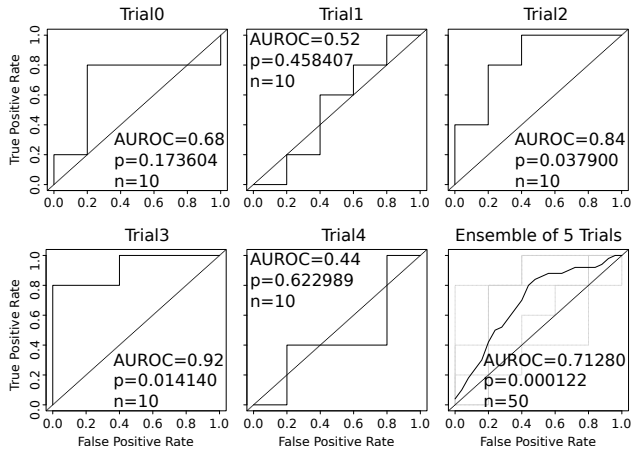

Fig. 2. Receiver operating characteristic curves, AUROC converted to p-values via normal approximation of Mann-Whitney U [11]. Each trial, an ensemble of eleven residual networks is learned to distinguish SPOP mutant from SPOP wild type, with low area under the receiver operating characteristic [AUROC] usually indicating this ensemble is a weak learner, in addition to p-values > 0.05. Each residual network is trained only on 10 SPOP mutants and 10 SPOP wild types. However, the ensemble of all five trials, a metaensemble, has an AUROC that is on average better that the individual trials, showing the metaensemble is a strong learner [12, 13]. Average ROC curve calculated in ROCR [14].

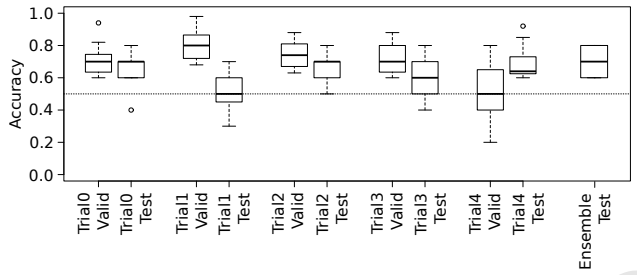

Fig. 3. For each Monte Carlo cross validation run within each trial, the distribution of validation and test set accuracies is plotted. At farthest right, for each ensemble composed of all eleven classifiers learned through all cross validation runs produces the ensemble for a trial, the distribution of all five ensemble test accuracies is plotted. The 95% CI for ensemble test accuracy is 0.5785-0.8241, Student's T, df=4, chance=0.5.

Discussion

In this work, we summarize a whole histology slide with a maximally abnormal subregion within the dominant tumor, such that the vast majority of the slide is not used for deep learning. Tissue microarrays take a similar although manual approach, by using a small representative circle of tissue to characterize a patient's cancer. In a sense, our patch extraction, saliency prediction, and TMARKER-based cell counting pipeline stages together model how a representative patch may be selected to describe the slide overall. By using a high magnification abnormal subregion, we aim to capture cellular-level image features driven by the underlying molecular change, such as SPOP mutation. This subregion has many desirable properties:

1. it is diagnostically relevant at low power to a pathologist at the microscope making a diagnosis (saliency),

2. it is taken from the dominant tumor (maximum number of cancer cells at low power),

3. it has the most nuanced cellular appearance (maximum

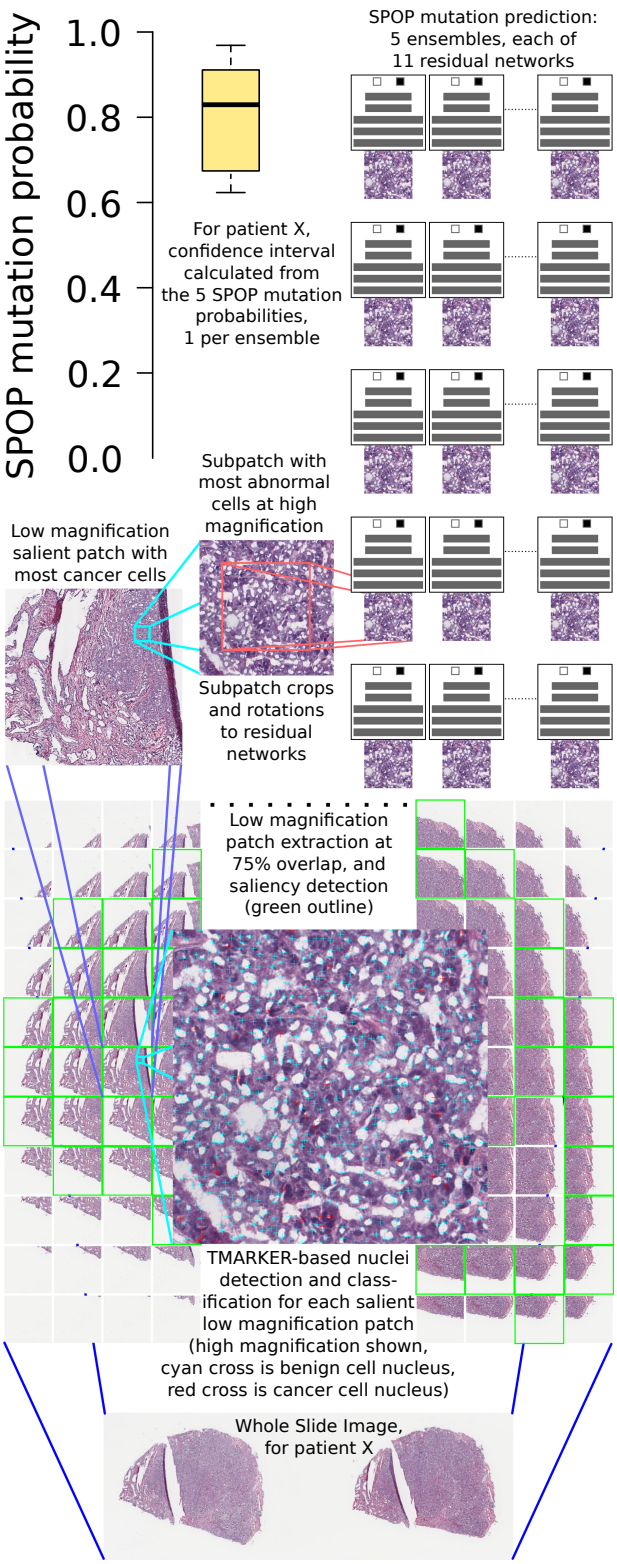

Fig. 4. Pipeline: a whole slide image is split into patches (bottom) at low magnification. Salient patches are identified. The salient patch with the most cancer cells is deemed the "dominant tumor" patch and further analyzed. At high magnification, a sliding window within the dominant tumor patch finds the region with maximum abnormal cells. Deep neural networks then predict SPOP mutation and a confidence interval is calculated over these predictions.

number of abnormal cells at high power), and

4. it is large enough at high power to capture histology, such as gland structure (800x800 pixel [px] patch).

A holistic approach that considers for deep learning more patches than this one subregion could improve performance and provide further insight. Gleason grading involves such a holistic approach, identifying first and second most prevalent cancer morphologies in a whole slide. We leave for future work the complexities of multiple representatives per slide, varying numbers of representatives per slide, and considering simultaneously multiple slides per patient.

Rather than use a single classifier, we use multiple classifier ensembles to predict the SPOP mutation state. Though each classifier, a residual network [17], tends to predict SPOP mutation probability in a bimodal distribution – being either close to 0 or close to 1, averaging these classifiers within an ensemble provides a more uniform distribution (Table S1). Each of five ensembles then emits the uniform-distributed probability SPOP is mutated, and from these five probabilities a 95% confidence interval is calculated, following a Student's T distribution. We believe statistical confidence in predictions is especially important in medical decision-making. Thus our pipeline, in addition to predicting SPOP mutation probability, indicates one of three outcomes: (i) confident SPOP mutation in patient, (ii) confident SPOP wild type in patient, (iii) inconclusive SPOP mutation state in patient.

Deep learning typically requires large sets of training data, yet we have only 20 patients with SPOP mutation, nearly an order of magnitude fewer than the 157 patients with SPOP wild type. Deep learning in small data is a challenging setting. Our metaensemble approach confronts this challenge by training many residual networks on small draws of the data, in equal proportions of SPOP mutants and SPOP wild type, then combining the residual networks as weak learners, to produce strong learners [12, 13]. In practice, the great depth of residual networks appears to be important because CaffeNet [18] – a shallower 8-layer neural network based on AlexNet [19] – rarely achieved validation accuracy of 0.6 or more predicting SPOP mutation state.

Pretraining the deep networks is essential in small data regimes such as ours, and we use the ResNet-50 model pretrained on ImageNet [17]. We used both the published ResNet-50 and our customized ResNet-50 that included an additional 50% dropout [20] layer and a 1024-neuron fully-connected layer. In practice, for difficult training sets, at least one of these architectures would often converge to validation accuracy of 0.6 or more. For data augmentation, the 800x800px images at high power are trimmed to the centermost 512x512px in 6 degree rotations, scaled to 256x256px, flipped and unflipped, then randomly cropped to 224x224 within Caffe [18] for training. Learning rates varied from 0.001 to 0.007, and momentums from 0.3 to 0.99. Test error tended to be worse with momentum less than 0.8, though low momentum allowed convergence for difficult training sets. Learning rates of 0.001, 0.0015, or 0.002 with momentum of 0.9 was typical. We used nVidia Titan-X GPUs with 12GB RAM for training. At 12GB, one GPU could train either resnet architecture using a minibatch size of 32 images.

Though most genetic aberrations effect a single Cul3 binding pocket in SPOP (Figs 5 and 6), our method's performance appears robust to two of 20 patients having genetic changes outside this pocket. In future work, for loss of function mutations like those that appear in SPOP, we may also consider copy number loss. Gene silencing through epigenetic changes may also be used to increase the number of positive training examples. The prediction target is then whether or not the function of the gene, such as SPOP, is lost. We believe this remains clinically actionable. Gain of function may similarly leverage epigenetics, and perhaps copy number amplification. By restricting ourselves to only consider mutations, our predictions are focused to a specific function loss mechanism (mutation) and our methods must overcome "deep learning in small data" issues, potentially increasing the applicability of our method to study other genetic drivers of histology with little available data.

Algorithm 1 Preprocessing: Each patient's SPOP mutation state is paired with the whole slide image patch having the maximum number of abnormal cells, where abnormal is either cancer or unknown cell type, rather than healthy cell type. This patch is taken from the dominant tumor patch. The dominant tumor patch is both salient and has the maximum number of cancer cells. The dominant tumor patch is at low magnification, while the abnormal patch is at high magnification within the primary.

```
for all prostate adenocarcinoma patients do
    spop ← SPOP normal/mutated state as 0/1
    slide ← whole slide image of cancer biopsy
    patches ← 75% overlap 800x800px images ← slide
    salient_patches ← predict_saliency(patches)
    pri_tumor_patch ← max_cancer(salient_patches)
    abn_patch ← max_abnormal(pri_tumor_patch)
    if abn_patch < 50% blurred then
        if abn_patch < 50% background then
            append (abn_patch, spop) to data_set
return data_set
```

Materials and Methods

We studied a TCGA cohort of 499 prostate adenocarcinoma patients, in particular the 177 patients that had both acceptable pathology images and SPOP mutation state available (Fig 1, Alg 1). SPOP mutation state was downloaded from cBioportal [23, 24]. In future work, we plan to extend our analysis to data sets, for example MSK IMPACT or Cornell EXaCT-1 [25, 26].

Microscope slides were scanned at 0.25 ± 0.003 microns per pixel [mpp], using an Aperio AT2 scanner. The resulting SVS data file consists of multiple levels, where level 0 is not downsampled, level 1 is downsampled by a factor of 4, level 2 by a factor of 16, and level 3 by a factor of 32. From each level, 800x800px patches were extracted via the OpenSlide software library [27]. We refer to level 2 as low magnification and level 0 as high magnification. Level 2 approximately corresponds to a 10x eyepiece lens and 10x objective lens at the microscope when the scan is 0.5mpp. Our saliency predictor assumed 0.5mpp scans, though the scans here were 0.25mpp, but appeared robust.

Algorithm 1 describes data preprocessing for training. In prior work, we developed a patch saliency predictor [28]. A TMARKER classifier was trained to determine cell types [29]. We define the dominant tumor patch as having the maximum number of cancer cells amount all 800x800px salient patches at low magnification. Within the 800x800px dominant tumor patch, we select an 800x800px patch at high mgnification having the maximum number of abnormal cells. Cancer cells count 1 towards the maximum abnormal cell count,

DRAFT

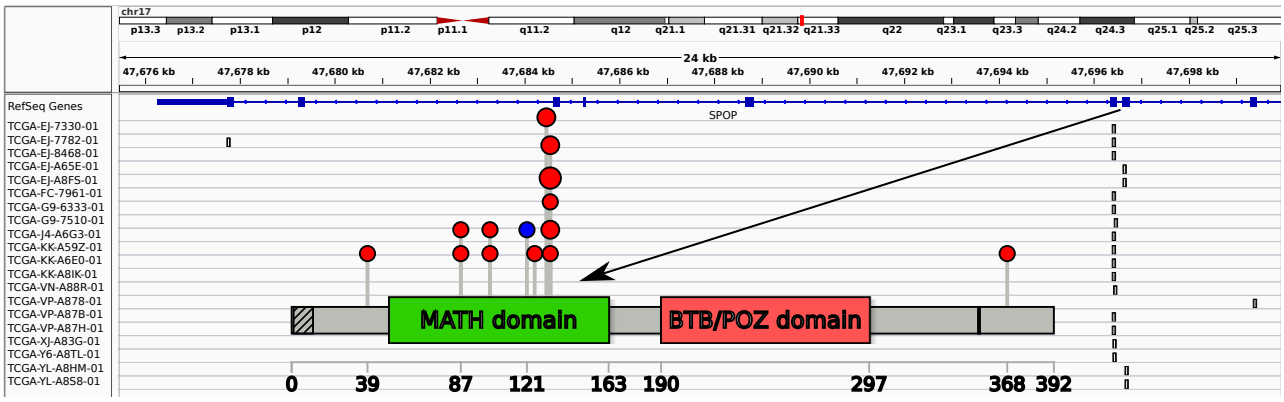

Fig. 5. SPOP mutations in the Integrated Genomics Viewer [21, 22], with lollipop plot showing mutations in most frequently mutated domain.

Algorithm 2 Training: Train a residual network on the abnormal patch representing each patient, labeled with SPOP mutation state. The final predictor is an ensemble of eleven of such resnets. After drawing a test set, training and validation sets are drawn from the remaining patients in the preprocessed data set (1). Within a Monte Carlo cross validation run, training and validation sets do not overlap. All draws are without replacement. Thirteen Monte Carlo cross validation runs are attempted in parallel, where training stops after eleven have validation accuracy of 0.6 or more. Some runs do not achieve 0.6 after many attempts. See Fig 7 for architecture details of $r50$ as "ResNet-50" and $drp5$ as "Drop50".

```
for trial in 0, 1, 2, 3, 4 do
    test_set ← draw(5 SPOP mutants, 5 SPOP normal)
    for monte in 1, 2, ..., 13 in parallel do
        train_set ← draw(10 mutants, 10 normal)
        valid_set ← draw(5 mutants, 5 normal)
        testable_model ← null
        repeat
            r50_model ← train_r50(train_set, valid_set)
            drp5_model ← train_drp5(train_set, valid_set)
            r50_acc ← accuracy(valid_set, r50_model)
            drp5_acc ← accuracy(valid_set, drp5_model)
            if r50_acc >= 0.6 then
                testable_model ← r50_model
            else if drp5_acc >= 0.6 then
                testable_model ← drp5_model
        until 11 testable models or testable_model ≠ null
        append testable_model to testable_models
    ensemble ← testable_models with ensemble averaging
    ensemble_acc ← accuracy(test_set, ensemble)
    append ensemble_acc to ensemble_accs
    append ensemble to ensembles
return (ensembles, ensemble_accs)
```

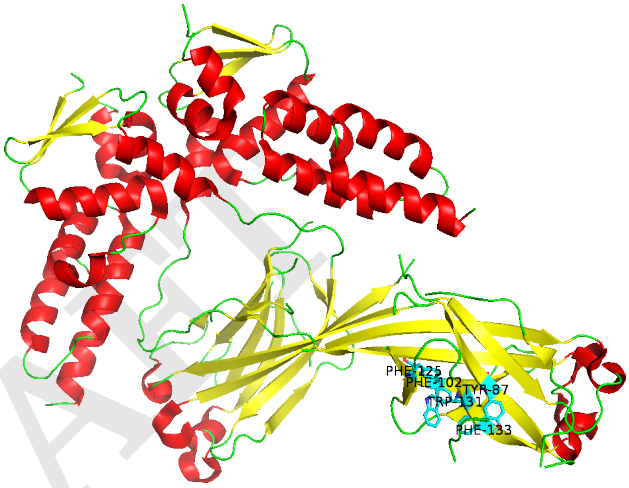

Fig. 6. SPOP tertiary structure with mutated residues labeled, common to one site. Mutation chr17:47677762 R368H in patient TCGA-EJ-7782-01 not shown and not in site. Deletion chr17:47699392 in patient TCGA-VP-A878 not shown and not in site.

Algorithm 3 Prediction: Use each of 5 ensembles to predict SPOP mutation state, then compute a 95% confidence interval of SPOP mutation state given these predictions. Each prediction is a probability of SPOP mutation in the patient. If the 95% CI lower bound is > 0.5, there is significant confidence that the patient has SPOP mutation. If the 95% CI upper bound is < 0.5, there is significant confidence that the patient does not have SPOP mutation. Otherwise, the patient's SPOP mutation state cannot be confidently predicted.

```
slide ← whole slide image of cancer biopsy
patches ← 75% overlap 800x800px images ← slide
salient_patches ← predict_saliency(patches)
pri_tumor_patch ← max_cancer(salient_patches)
abn_patch ← max_abnormal(pri_tumor_patch)
for ensemble in ensembles do
    spop_prediction ← predict(abn_patch, ensemble)
    append spop_prediction to spop_predictions
return confidence_interval(spop_predictions)
```

DRAFT

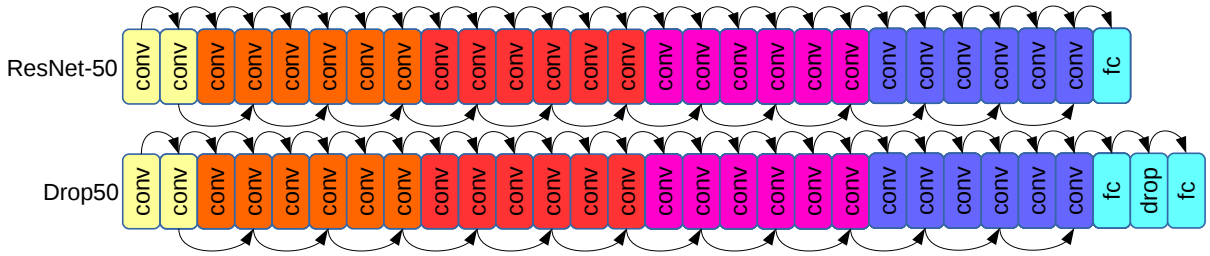

Fig. 7. ResNet-50 architecture [17], at right. Customized Drop50 networks supports ResNet-50 pretraining, but has additional dropout [20] and fully connected neuron layers. In practice at least one of these architectures converges to a validation accuracy of 0.6 or more. Convolutional layers "conv", fully-connected layers "fc", 50% dropout layer "drop".

unknown cells count 0.5, and healthy cells count 0. The dominant tumor is explored in increments of 10 pixels, until the bounding box with the maximum number of abnormal cells is found. Whereas our saliency predictor operated on low-power tissue-level details of a patch, our SPOP predictor operates on high-power cell-level details of a patch. A patient is discarded from the study if the abnormal patch is $> 50\%$ blurred or $> 50\%$ background, as determined by visual inspection.

Algorithm 2 describes neural network ensemble for training. This procedure allows learning to occur despite a mere 20 patients having an SPOP mutation, compared to 157 not having an SPOP mutation. Additionally, having 5 ensembles – each trial yielding one ensemble of 11 resnets – allows a 95% confidence interval in predicted SPOP mutation state to be calculated for both the generalization error estimate during training and any future patient (Algorithm 3). Moreover, such a large number of resnets can fully sample the SPOP wild type patients, while each resnet is still trained with an equal proportion of SPOP mutants and wild types. Two different neural network architectures are used, both the published ResNet-50 architecture ($r50$ in Alg 2) and our custom ResNet-50 with a 50% dropout [20] layer with an additional 1024 fully-connected layer as the top layer ($drp5$ in 2, and shown as Drop50 in Fig 7). In practice, at least one architecture tended to have validation accuracy > 0.6. A variety of architectures may increase the variance among the neural networks within the ensemble, and the variance in errors should average out in the ensemble. Trials 0, 2, and 3 used two Drop50 learners and nine ResNet-50 classifiers (Fig 7, Table S1). Trial 1 had six Drop50 and five ResNet-50 classifiers. Trial 4 had three Drop50 and eight ResNet-50 classifiers. Trials 1 and 4 had the worst performance by AUROC (Fig 2), both had at least one ResNet-50 predictor with 0.3 or worse test set accuracy (Table S1), and both had more than two Drop50 learners in the final ensemble for the trial. For challenging draws of training, validation, and test sets, Drop50 learners may slightly outperform ResNet-50 learners, though generalization accuracy may remain low due to the clustering of the data in the draws and limited sample sizes.

There is remarkable variability in test accuracy among the testable models from each Monte Carlo cross validation run (Table S1). If the test set is drawn from approximately the same distribution as the validation set, where resnet validation accuracy is 0.6+ and resnets are uncorrelated, then we can expect 6 of the 11 resnets ($6/11 < 0.6$) in an ensemble to correctly predict SPOP mutation state on average. In this way the ensemble is a strong learner based on resnet weak learners [12, 13]. Through ensemble averaging, the mean SPOP mutation probability is computed over all 11 constituent resnets, to provide the final probability from the ensemble. Finally, learning rate and momentum are tuned between attempts within a Monte Carlo cross validation run until acceptable validation accuracy is achieved.

ACKNOWLEDGMENTS. AJS was supported by the Tri-Institutional Training Program in Computational Biology and Medicine (via NIH training grant T32GM083937). Nvidia GPUs were provided in part by the GPU Center of Excellence Award to TJF. AJS and TJF were supported by MSK Functional Genomics Initiative grant to TJF.

DRAFT

References

1. Mosquera JM et al. (2007) Morphological features of TMPRSS2-ERG gene fusion prostate cancer. *The Journal of pathology* 212(1):91–101.
2. Fuchs T, Buhmann J (2011) Computational pathology: challenges and promises for tissue analysis. *Computerized medical imaging and graphics : the official journal of the Computerized Medical Imaging Society* 35(7-8):515–530.
3. Barbieri C et al. (2012) Exome sequencing identifies recurrent SPOP, FOXA1 and MED12 mutations in prostate cancer. *Nature Genetics* 44(6):685–689.
4. Hakimi A et al. (2015) TCEB1-mutated renal cell carcinoma: a distinct genomic and morphological subtype. *Modern pathology : an official journal of the United States and Canadian Academy of Pathology, Inc* 28(6):845–853.
5. Weisman P et al. (2016) Genetic alterations of triple negative breast cancer by targeted next-generation sequencing and correlation with tumor morphology. *Mod Pathol* 29(5):476–488.
6. Budinska E, Bosman F, Popovici V (2016) Experiments in molecular subtype recognition based on histopathology images. (IEEE), pp. 1168–1172.
7. Zhong Q et al. (2016) Image-based computational quantification and visualization of genetic alterations and tumour heterogeneity. *Scientific Reports* 6:24146.
8. Natrajan R et al. (2016) Microenvironmental Heterogeneity Parallels Breast Cancer Progression: A Histology–Genomic Integration Analysis. *PLoS Med* 13(2):e1001961.
9. Al-Ahmadie H et al. (2011) Somatic mutation of fibroblast growth factor receptor-3 (FGFR3) defines a distinct morphological subtype of high-grade urothelial carcinoma. *The Journal of pathology* 224(2):270–279.
10. Yosinski J, Clune J, Nguyen A, Fuchs T, Lipson H (2015) Understanding Neural Networks Through Deep Visualization.
11. Mason S, Graham N (2002) Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: Statistical significance and interpretation. *Quarterly Journal of the Royal Meteorological Society* 128(584):2145–2166.
12. Krogh A, Vedelsby J (1995) Neural network ensembles, cross validation, and active learning. *Advances in neural information processing systems* 7:231–238.
13. Sollich P, Krogh A (1996) Learning with ensembles: How over-fitting can be useful. Vol. 8, pp. 190–196.
14. Sing T, Sander O, Beerenwinkel N, Lengauer T (2005) ROCR: visualizing classifier performance in R. *Bioinformatics (Oxford, England)* 21(20):3940–3941.
15. An J, Wang C, Deng Y, Yu L, Huang H (2014) Destruction of full-length androgen receptor by wild-type SPOP, but not prostate-cancer-associated mutants. *Cell reports* 6(4):657–669.
16. Urbinati G et al. (2016) Knocking Down TMPRSS2-ERG Fusion Oncogene by siRNA Could be an Alternative Treatment to Flutamide. *Molecular therapy. Nucleic acids* 5.
17. He K, Zhang X, Ren S, Sun J (2015) Deep Residual Learning for Image Recognition.
18. Jia Y et al. (2014) Caffe: Convolutional Architecture for Fast Feature Embedding.
19. Krizhevsky A, Sutskever I, Hinton G (2012) Imagenet classification with deep convolutional neural networks.
20. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R (2014) Dropout: A Simple Way to Prevent Neural Networks from Overfitting. Vol. 15, pp. 1929–1958.
21. Thorvaldsdóttir H, Robinson J, Mesirov J (2013) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics* 14(2):178–192.
22. Robinson J et al. (2011) Integrative genomics viewer. *Nature Biotechnology* 29(1):24–26.
23. Gao J et al. (2013) Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Science signaling* 6(269):pl1–pl1.
24. Cerami E et al. (2012) The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data. *Cancer Discovery* 2(5):401–404.
25. Cheng D et al. (2015) Memorial Sloan Kettering-Integrated Mutation Profiling of Actionable Cancer Targets (MSK-IMPACT): A Hybridization Capture-Based Next-Generation Sequencing Clinical Assay for Solid Tumor Molecular Oncology. *The Journal of molecular diagnostics : JMD* 17(3):251–264.
26. Beltran H et al. (2015) Whole-Exome Sequencing of Metastatic Cancer and Biomarkers of Treatment Response. *JAMA oncology* 1(4):466–474.
27. Goode A, Gilbert B, Harkes J, Jukic D, Satyanarayanan M (2013) OpenSlide: A vendor-neutral software foundation for digital pathology. *Journal of pathology informatics* 4.
28. Schaumberg A, Sirintrapun J, Al-Ahmadie H, Schueffler P, Fuchs T (2016) DeepScope: Non-intrusive Whole Slide Saliency Annotation and Prediction from Pathologists at the Microscope.
29. Schüffler P et al. (2013) TMARKER: A free software toolkit for histopathological cell counting and staining estimation. *Journal of pathology informatics* 4(Suppl).