Pan-cancer image-based detection of clinically actionable genetic alterations

Jakob Nikolas Kather[1,2,3], Lara R. Heij[4,5,6], Heike I. Grabsch[7,8], Chiara Loeffler[1], Amelie Echle[1], Hannah Sophie Muti[1], Jeremias Krause[1], Jan M. Niehues[1], Kai A. J. Sommer[1], Peter Bankhead[9], Loes F. S. Kooreman[7], Jefree J. Schulte[10], Nicole A. Cipriani[10], Roman D. Buelow[6], Peter Boor[6], Nadina Ortiz-Brüchle[6], Andrew M. Hanby[8], Valerie Speirs[11], Sara Kochanny[12], Akash Patnaik[12], Andrew Srisuwananukorn[13], Hermann Brenner[2,14,15], Michael Hoffmeister[14], Piet A. van den Brandt[16], Dirk Jäger[2,3], Christian Trautwein[1], Alexander T. Pearson[12,*], Tom Luedde[1,17,*]

[1] Department of Medicine III, University Hospital RWTH Aachen, Aachen, Germany

[2] German Cancer Consortium (DKTK), Heidelberg, Germany

[3] Applied Tumor Immunity, German Cancer Research Center (DKFZ), Heidelberg, Germany

[4] Department of Surgery and Transplantation, University Hospital RWTH Aachen, Aachen, Germany

[5] Department of Surgery, NUTRIM, School of Nutrition and Translational Research in Metabolism, Maastricht University, Maastricht, The Netherlands

[6] Institute of Pathology, University Hospital RWTH Aachen, Aachen, Germany

[7] Department of Pathology, GROW School for Oncology and Developmental Biology, Maastricht University Medical Center+, Maastricht, The Netherlands

[8] Pathology & Data Analytics, Leeds Institute of Medical Research at St James's, University of Leeds, Leeds, UK

[9] MRC Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, UK

[10] Department of Pathology, University of Chicago Medicine, Chicago, IL, USA

[11] Institute of Medical Sciences, School of Medicine, Medical Sciences and Nutrition, University of Aberdeen, Aberdeen, UK

[12] Department of Medicine, University of Chicago Medicine, Chicago, IL, USA

[13] Department of Medicine, University of Illinois – Chicago, Chicago, IL, USA

[14] Division of Clinical Epidemiology and Aging Research, German Cancer Research Center (DKFZ), Heidelberg, Germany

[15] Division of Preventive Oncology, German Cancer Research Center (DKFZ) and National Center for Tumor Diseases (NCT), Heidelberg, Germany

[16] Department of Epidemiology, GROW School for Oncology and Developmental Biology, Maastricht University Medical Center+, Maastricht, The Netherlands

[17] Division of Gastroenterology, Hepatology and GI Oncology, University Hospital RWTH Aachen, Aachen, Germany

* these authors contributed equally to this work

Correspondence should be addressed to jkather@ukaachen.de, apearson5@medicine.bsd.uchicago.edu and tluedde@ukaachen.de

Abstract

Molecular alterations in malignant tumors can cause phenotypic changes in tumor cells and their microenvironment. Routine histopathology tissue slides – which are ubiquitously available for patients with solid tumors – can reflect such morphological changes. Here, we show that deep learning can consistently infer a wide range of genetic mutations, molecular tumor subtypes, gene expression signatures and standard pathology biomarkers directly from routine histology images of cancer. We developed, systematically optimized, validated and publicly released a one-stop-shop workflow and applied it to routine tissue slides of more than 5000 patients across a broad spectrum of common solid tumors including lung, colorectal, breast and gastric cancer. Our findings show that a single deep learning algorithm can be trained to predict a wide range of molecular alterations from routine, paraffin-embedded histology slides stained with hematoxylin and eosin. These predictions generalize to other populations and yield spatially resolved predictions. Our method can be implemented on mobile hardware, potentially enabling point-of-care diagnostics for personalized cancer treatment. More generally, this approach can be used to elucidate and quantify genotype-phenotype links in cancer.

Introduction

Precision treatment of cancer relies on detection of genetic alterations which are diagnosed by molecular biology assays.[1] These tests can be a bottleneck in oncology workflows because of high turnaround time, tissue usage and costs.[2] Clinical guidelines recommend molecular testing of tumor tissue for most patients with advanced solid tumors. However, in most tumor types, routine testing includes only a handful of alterations, such as KRAS, NRAS, BRAF mutations and microsatellite instability (MSI) in colorectal cancer.[3] While new studies identify more and more molecular features of potential clinical relevance, current diagnostic workflows are not designed to incorporate an exponentially rising load of tests. For example, in colorectal cancer, previous studies have identified consensus molecular subtypes (CMS)[4] as a candidate biomarker, but sequencing costs and method complexity preclude widespread testing in clinical routine and clinical trials.[5] Therefore, there is a growing need to identify new, inexpensive and scalable biomarkers in medical oncology.

While comprehensive molecular and genetic tests are hard to implement at scale, tissue sections stained with hematoxylin and eosin (H&E) are ubiquitously available. We hypothesized that these routine tissue sections contain information about established and candidate biomarkers and that molecular biomarkers could be inferred directly from digitized whole slide images (WSI). The rationale for this hypothesis is that genetic changes in tumor cells cause functional changes, which can influence tumor cell morphology.[6,7] In addition to such first-order genotype-phenotype correlations, genetic changes in tumor cells can influence the tumor microenvironment, resulting in higher-order genotype-phenotype correlations. Specific examples for such correlations are known for microsatellite instability (MSI)[8], a clinically approved biomarker for cancer immunotherapy in colorectal cancer.[9] In the case of MSI, the genotype-phenotype correlation is consistent enough to robustly infer the genotype just by observing morphological features in a histological image, as we have previously shown.[10] Other previous studies have identified genotype-phenotype links for selected genetic features in lung cancer[11,12], prostate cancer[13], head and neck[14] and liver[15] cancer, among others. Building on these previous studies, we systematically

investigated the presence of genotype-phenotype links for a wide range of clinically relevant molecular features across all major solid tumor types. Specifically, we asked which molecular features leave a strong enough footprint in histomorphology so they can be inferred from histology images alone with deep learning. We aimed to use deep learning in a pan-molecular pan-cancer approach, with a focus on clinically relevant genetic molecular features. Such an approach could ultimately yield clinically useful biomarkers with favorable cost, time and material requirements. More specifically, this approach could guide a more narrow indication for molecular testing, increasing the pre-test probability of a given molecular feature. Independently of potential clinical application, inferring genetic changes from histology images could also elucidate biological mechanisms of downstream effects of molecular alterations in solid tumors. Therefore, we developed, optimized and externally validated a new deep learning pipeline to determine molecular features directly from histology images.

Methods

Patient cohorts and ethics statement

All experiments were conducted in accordance with the Declaration of Helsinki and the International Ethical Guidelines for Biomedical Research Involving Human Subjects. Anonymized scanned whole slide images were retrieved from The Cancer Genome Atlas (TCGA) project through the Genomics Data Commons Portal (https://portal.gdc.cancer.gov/). We applied our method to 14 of the most common solid tumor types: breast (BRCA)[16], cervical (CESC)[17], colorectal (COAD and READ)[18], gastric (STAD)[19], head and neck (HNSC)[20], hepatocellular (LIHC)[21], lung adeno (LUAD)[22], lung squamous (LUSC)[23], melanoma (SKCM)[24], pancreatic (PAAD)[25], prostate (PRAD)[26], renal chromophobe (KICH)[27], renal clear cell (KIRC)[28] and renal papillary cancer (KIRP)[29]. Melanoma (SKCM) tissue slides in the TCGA database comprised primary tumor samples as well as metastasis tissue. These groups were analyzed separately. For external validation, we acquired colorectal cancer tissue samples from the DACHS study[30,31], which were retrieved from the tissue bank of the National Center for Tumor Diseases (NCT, Heidelberg, Germany) as described before.[10]

Molecular labels

The aim of this study was to predict clinically relevant features, including genetic alterations, directly from routine histology slides. We systematically applied this screening approach to four groups of molecular alterations: First, we used single-gene mutations, considering any genetic variant. We used the most commonly mutated genes in the respective tumor types (derived from the "cbioportal" database[32,33] at http://cbioportal.org) and clinically targetable genes (level one genes from OncoKB at http://www.oncokb.org, Pan Cancer Atlas Project[34]). We required each mutation to affect at least four patients in a given cohort. Second, we repeated the analysis on putative and confirmed oncogenic driver mutations only, as defined in OncoKB. Third, we aimed to predict gene expression subtypes, relevant gene expression signatures and immune-cell gene expression signatures derived from systematic studies[35-37]. Fourth, we used "standard of care" features derived from the TCGA database (data at http://portal.gdc.cancer.gov), including hormone receptor status in breast cancer. All labels (genetic variants, driver mutations, signatures

and standard features) are listed in Suppl. Table 1. For each individual target label in each tumor type and each cross-validation run, we re-trained a single deep neural network, using identical hyperparameters. Features with continuous values were binarized at the mean.

Image preprocessing

Scanned whole slide images of diagnostic tissue slides (formalin-fixed paraffin-embedded tissue) stained with hematoxylin and eosin were acquired in SVS format. All images were downsampled to 20x magnification, corresponding to 0.5 µm/pixel (px). Each whole slide image was manually reviewed and the tumor area with was annotated under direct supervision of a specialty pathologist. During annotation, all observers were blinded with regard to any molecular or clinical feature. Only those images containing at least 1 mm^2 contiguous tumor tissue were used for downstream analysis. 6% of whole slide images, corresponding to 5% of patients were excluded due to technical artifacts or lack of tumor (Suppl. Table 2). Tumor tissue on all other slides was tessellated into square tiles of 512x512 px edge length, corresponding to 256x256 µm at a resolution of 0.5 µm/px. Tiles with more than 50% background were discarded; background pixels were defined by brightness over 0.86 (220/255). For the benchmark task (identification of an optimal neural network model), these images were resized to 224x224 px (at 1.14 µm/ px) to be consistent with a previous study[10]. All steps in the data preprocessing pipeline (including preprocessing of images and preprocessing of metadata) are documented in detail in our in-house manual for data preparation, which is publicly available at https://dx.doi.org/10.5281/zenodo.3694994. All methods for whole slide image processing, including tessellation of images and visualization of spatial activation maps, were implemented in QuPath v0.1.2 in Groovy (http://qupath.github.io).

Patient-level cross-validation

Aiming to develop a one-stop-shop method for systematic discovery of genotype-phenotype links in multiple cancer types, we developed a reusable pipeline of data processing steps. One or more whole slide images (WSI) per patient were collected tumor regions in these images were tessellated into tiles. All tiles inherited the molecular label of their parent patient. Before training, the patient cohort was randomly split in three partitions, keeping the target labels balanced between

partitions. Neural networks were trained on two partitions each and subsequently evaluated on the third partition. Thus, no tiles from a given patient were ever part of a training set and a test set for the same classifier. Before training, tile libraries were randomly undersampled in such a way that the number of tiles per label was identical for each label (Fig. 1a).

Neural network training, model selection and hyperparameter optimization

Deep neural networks were trained on image tiles with the aim of predicting molecular labels. All neural networks were pre-trained on the ImageNet database as described previously[10] and were specifically modified for the classification task at hand by replacing the three top layers with a 1000-neuron fully connected layer, a softmax layer and a classification layer. For training, we used on-the-fly data augmentation (random horizontal and vertical reflection) to achieve rotational invariance of the classifiers. Hyperparameter selection was performed for five commonly used deep neural networks: resnet18, alexnet, inceptionv3, densenet201 and shufflenet. The sampled hyperparameter space was as follows: learning rate 5e-5 and 1e-4, maximum number of tiles per whole slide image: 250, 500 and 750, number of trainable layers: 10, 20 and 30. We trained for four epochs with a mini batch size of 512, similar to previous experiments.[10] As a benchmark task, we used MSI detection in colorectal cancer as described before.[10]

Inference of molecular status

During inference, a categorical prediction was made for each tile by the neural network (Fig. 1b). The percentage of positive predicted tiles for each class was regarded as a "probability score" for each patient. This score was used as the free variable for a receiver operating characteristic (ROC) analysis with area under the ROC curve (AUROC) being the primary endpoint for each target feature. AUROC values are reported as mean with a confidence interval representing lower and upper range of a 10x bootstrapped experiment. To quantify if predictions for different classes of patients were statistically significant, the probability scores for patients in a given class were compared to probability scores of all other patients. Statistical significance of these differences was assessed with a two-tailed t-test with a pre-defined significance level of 0.05. To compensate for the large number of tested hypotheses in this study, we performed "false detection rate"

(FDR) correction for p-values with the Benjamini-Hochberg method on all p-values across all cancer types. All p-values smaller than 10^{-5} after FDR-correction are reported as 10^{-5}. Statistical methods are further described in Suppl. Fig. 1a-c. The number of tiles generated per whole slide image is shown in Suppl. Fig. 2. Training and inference were performed on our local computing cluster on 10 Nvidia RTX graphics processing units (GPUs), each with 24 GB of GPU RAM. Cumulative computing time for all experiments within this study was approximately 12,000 GPU-hours. All deep learning algorithms were implemented in Matlab R2019a (Mathworks, Natick, MA, USA).

External validation

To investigate if complex deep learning biomarkers generalize to external patient cohorts, we trained deep learning classifiers on all TCGA samples of a given tumor type and externally validated the predictions in patient cohorts from our respective institutions. External validation was performed for BRAF mutation status and CpG island methylator phenotype (CIMP) in colorectal cancer in N=408 patients, a subset of the multicenter DACHS study which was previously collected and described.[10] BRAF and CIMP were chosen as validation markers because of their biological relevance and availability of robust measurements of these markers in the DACHS cohort.

Feature visualization

To visualize the deep learning predictions and make them understandable to human observers, we used two approached: First, we rendered the tile-level soft predictions for each class as activation maps, visualizing prediction scores as a heatmap overlay on the original histology image. Second, we identified the highest-predicted tiles of the highest-predicted true positive patients for each class, allowing observers to identify histological patterns that are correlated with a molecular feature. These approaches were designed to allow human observers to identify which morphological features deep learning classifiers were most sensitive to.

Alternative approaches

In our baseline approach, image tiles from manually annotated tumor regions on formalin-fixed paraffin-embedded (FFPE) slides (diagnostic slides) were used. This approach was compared to several alternative approaches as shown in Suppl. Fig. 3. The first alternative approach used color

normalization of image tiles with the Macenko method[38] to mitigate differences in staining intensity and hue (Suppl. Fig. 4). Some previous studies have used color normalization for deep learning[10], while other studies have shown that color normalization can bias histology image classification.[39] The second alternative approach we investigated was to use tiles from the whole slide, as opposed to the tumor region only. In this "weakly supervised" approach, many tiles without invasive cancer tissue were present in the training and inference sets (Suppl. Fig. 5). The third alternative approach was to use frozen slides as opposed to FFPE slides in a weakly supervised way (Suppl. Fig. 6).

Data availability

All data (including histological images) from the TCGA database are available at https://portal.gdc.cancer.gov/. All molecular data for patients in the TCGA cohorts are available at https://cbioportal.org. Raw data for Figures and Suppl. Figures are shown in Suppl. Table 3.

Code availability

All source codes are available and documented at https://github.com/jnkather/DeepHistology.

Results

Optimization of deep learning for inference of genotype from histology

We hypothesized that deep learning can infer molecular alterations directly from routine histology images across multiple common solid tumor types. To test this, we developed, optimized and extensively validated a new 'one-stop-shop' workflow to train and evaluate deep learning networks. To select an efficient network model and to optimize the deep learning hyperparameters, prediction of microsatellite instability (MSI) in colorectal cancer was used as a clinically relevant benchmark task[10]. In this benchmark, we sampled a large hyperparameter space with different commonly used deep learning models[10,11,14,40] which were modified specifically for this application. Unexpectedly, 'shufflenet'[41], a lightweight neural network architecture performed similarly to more complex networks including 'densenet'[42], 'inception'[43] and 'resnet'[44] networks, which are used in many other studies[45] (Fig. 1c). Shufflenet demonstrated high accuracy at a low training time (raw data in Suppl. Table 1, N=426 patients in the TCGA cohort). Shufflenet is optimized for mobile devices, making this deep neural network architecture attractive for decentralized point-of-care image analyses or direct implementation in microscopes[46]. We externally validated the best shufflenet classifier by training on N=426 patients in the TCGA-CRC cohort[10] and validating on N=379 patients with available MSI status in the DACHS cohort[10], reaching an AUROC of 0.89 [0.88; 0.92].This represents an improvement over the previous best performance of 0.84 in that dataset[10] and supports the notion that shufflenet is an efficient and powerful neural network model which can infer clinically relevant molecular changes directly from histology images.

Pan-cancer prediction of genetic variants from histology

Having thus identified a deep neural network model and a set of suitable hyperparameters, we systematically applied this approach to hundreds of molecular alterations in 14 major tumor types, and trained and evaluated deep learning networks by three-fold cross-validation on each cohort. This yielded approximately 10^4 independently trained deep neural networks which were systematically evaluated and compared across molecular features across cancer types. The full list of candidate mutations (Suppl. Table 1) included all point mutations targetable by FDA-approved drugs (Level 1 evidence on www.oncokb.org, the 20 most common mutations shown in

Fig. 1d). First, we trained deep neural networks to detect any sequence variants in these target genes. We found that in 13 out of 14 tested tumor types, the mutation of one or more of such genes could be inferred from histology images alone, with statistical significance after correction for multiple testing (Fig. 2a-n, Suppl. Fig. 7). In particular, in major cancer types such as lung adenocarcinoma, colorectal cancer, breast cancer and gastric cancer, alterations of several genes of particular clinical and/or biological examples were detectable (Fig. 2a-d). Examples include mutations in TP53, which could be significantly detected in all four of these cancer types, as well as mutations of BRAF in colorectal cancer (TCGA-COAD and TCGA-READ[18], N=555, Fig. 2b), MTOR – a candidate for targeted treatment[47] – in gastric cancer (Fig. 2d) and FBXW7 mutation in lung adenocarcinoma (TCGA-LUAD[22], N=457, Fig. 2a) and gastric cancer (TCGA-STAD[19], N=321, Fig. 2d). Mutations of PIK3CA (which is directly targetable by a small molecule inhibitor[48]) was significantly detectable in breast cancer (TCGA-BRCA[16], N=995, Fig. 2c) and gastric cancer (Fig. 2d). In addition, in breast cancer, mutations of MAP2K4 (which is a potential biomarker for response to MEK inhibitors[49]) were significantly detectable (Fig. 2c). Among all tested tumor types, gastric cancer (Fig. 2d) and colorectal cancer (Fig. 2b) had the highest absolute number of detectable mutations. For all statistically significant features, the mean cross-validated area under the receiver operating curve (AUROC) for the top eight mutations ranged from 0.60 to 0.78 in lung adenocarcinoma (Suppl. Fig. 8); from 0.65 to 0.76 in colorectal cancer (Suppl. Fig. 9); from 0.62 to 0.78 in breast cancer (Suppl. Fig. 10) and from 0.66 to 0.78 in gastric cancer (Suppl. Fig. 11). Beyond these four tumor types, a range of notable mutations could be detected in other tumor types: While in melanoma (TCGA-SKCM[24]) primary tumors, few mutations were detectable (Suppl. Fig. 12a-h), in melanoma metastases, mutations in FBXW7 and PIK3CA were significantly detectable (Fig. 2e, Suppl. Fig. 12i-p). In prostate cancer (TCGA-PRAD[26], N=397 patients, Fig. 2f, Suppl. Fig. 13), our method detected TP53 and FOXA1 mutations from histology, among others. In pancreatic adenocarcinoma (TCGA-PAAD[25], N=171 patients, Fig. 2g, Suppl. Fig. 14), identifying KRAS wild type patients is of high clinical relevance because these patients are potential candidates for targeted treatment and our method significantly identified KRAS genotype in pancreatic cancer. Lung squamous cell carcinoma is known for its difficulty in molecular diagnosis and few molecularly or genetically targeted treatment options even in clinical trials. Thus, it is plausible

that in this cancer type, tumor histomorphology is not well correlated to mutations and correspondingly, few mutations were significantly detectable in this tumor type in our experiments (TCGA-LUSC, N=413, Fig. 2h, Suppl. Fig. 15). In hepatocellular carcinoma (TCGA-LIHC[21], N=358 patients, Fig. 2i), the β-catenin gene (CTNNB1) is a key driver gene with broad prognostic and predictive implications[50] and its mutational status was highly significantly detected from histology (Suppl. Fig. 16). In papillary (Fig. 2j, Suppl. Fig. 17) and clear cell renal cell carcinoma (Fig. 2k, Suppl. Fig. 18), alterations in multiple genes including KRAS and PBRM were highly significantly detectable while in and chromophobe renal cell carcinoma (Fig. 2l, Suppl. Fig. 19), no genetic variants were significantly detectable, possibly due to a low patient number in this cohort. In head and neck squamous cell carcinoma (TCGA-HNSC[20], N=435 patients), genotype of CASP8, which is linked to resistance to cell death[51], was significantly detected (Fig. 2m, Suppl. Fig. 20). In cervical cancer (TCGA-CESC[17], N=261 patients), mutations in TCERG1, STK11, AMER1, among others, were significantly detectable with high AUROC values (Fig. 2n, Suppl. Fig. 21). Raw data for prediction performance in any gene in any tumor type are available in Suppl. Table 3.

Pan-cancer prediction of oncogenic drivers from histology

Not all genetic variants are causative of malignant processes. Therefore, we repeated the screening experiment, limiting mutations to confirmed or putative oncogenic drivers (Fig. 3a-n). With this criterion, the absolute number of patients affected by a particular mutation was lower and thus, fewer genes met the threshold of at least four positive cases in a given tumor type. On the other hand, we hypothesized that oncogenic driver genes could leave a stronger pattern in histological morphology due to their higher biological relevance. Genetic variants in classical oncogenes such as TP53 and KRAS are almost always oncogenic drivers and correspondingly, mutations of these genes reached similar prediction accuracy valued in the “drivers only” experiment when compared to the “all variants” approach (Fig. 3a-n). For other genes, prediction accuracy increased when limited to oncogenic drivers: a notable example was EGFR in lung adenocarcinoma (Fig. 3a). In summary, these data show that deep learning can detect a wide range of targetable and potentially targetable point mutations directly from histology across multiple prevalent tumor types.

306 Inference of molecular subtypes and gene expression signatures

307 In the next step, we asked if established molecular subtypes and gene expression signatures of
308 cancer and immune cells could be detected by deep learning. Compared to single-gene muta-
309 tions, these changes occur at a higher functional level and we hypothesized that their morpho-
310 logical impact could be larger than that of single mutations. To address this hypothesis, we chose
311 features with known biological and potential clinical significance. A major group of such features
312 are immune-related gene expression signatures[37] of CD8-positive lymphocytes, macrophages,
313 cell proliferation, interferon-gamma (IFNg) signaling and transforming growth factor-beta (TGFb)
314 signaling (full list available in Suppl. Table 1). These biological processes are involved in response
315 to cancer treatment, including immunotherapy. Detecting their morphological correlates in his-
316 tology images could facilitate the development of more nuanced treatment strategies. Indeed,
317 across all investigated tumor types, we saw that these high-level biological features were much
318 better predictable than genetic variants or driver mutations (Fig. 4a-d and Suppl. Fig. 7) Again,
319 AUROC values for significantly ($p<0.05$ after FDR correction) predictable features were highest in
320 lung adenocarcinoma (Fig. 4e), colorectal cancer (Fig. 4f), breast cancer (Fig. 4g) and gastric can-
321 cer (Fig. 4h). In lung adenocarcinoma, signatures of proliferation, macrophage infiltration and T-
322 lymphocyte infiltration were significantly detectable from images with high AUROCs (Fig. 4e).
323 Similarly, significant AUROCs for these biomarkers were achieved in colorectal cancer (Fig. 4f)
324 breast cancer (Fig. 4g) and gastric cancer (Fig. 4h). In gastric cancer, we additionally found that a
325 signature of stem cell properties (stemness) was highly detectable directly from histology images
326 (Fig. 4h). Recent studies have clustered tumors into comprehensive 'molecular subtypes'[37]. We
327 found that our method could detect TCGA molecular subtypes[37] with up to AUROC 0.74 in lung
328 adenocarcinoma (Fig. 4e), pan-gastrointestinal subtypes[36] with up to AUROC 0.76 in colorectal
329 cancer (Fig. 4f) and PAM50 subtypes with up to AUROC 0.78 in breast cancer (Fig. 4g), among
330 other molecular subtypes. These findings could open up new options for clinical trials of cancer:
331 While accumulating evidence shows that such molecular clusters of tumors reflect biologically
332 distinct groups and are correlated to clinical outcome, deep molecular classification of these tu-
333 mors is usually not available in clinical routine or clinical trials. Detecting these subtypes merely
334 from histology would allow for these subtypes to be analyzed in clinical trials directly from

broadly available routine material, potentially helping to identify new biomarkers for treatment response or to guide specific molecular testing.

Prediction of standard histological biomarkers with deep learning

To comprehensively evaluate the potential clinical use of our new deep learning pipeline, we investigated classification accuracy for standard histopathological biomarkers. We found that deep learning could highly significantly predict most of these biomarkers for breast cancer (Fig. 4c and i), gastric cancer (Fig. 4d and j) and other tumor types (Suppl. Fig. 11-18). In particular, status of hormone receptors was predictable from routine histology in breast cancer, with an AUROC of 0.82 for estrogen receptor and 0.74 for progesterone receptor (Fig. 4i). Together, these results demonstrate that deep-learning-based inference of genetic alterations, high-level molecular alterations and established biomarkers from routine diagnostic histology slides is feasible.

Evaluation of alternative approaches

Deep learning-based inference of molecular features from histology is a relatively novel field of research and it can be anticipated that technical improvements can further improve prediction performance. We quantified the effect of alternative technical approaches in the colorectal cancer cohort (TCGA-COAD/READ). First, we investigated the role of color normalization of tiles. In a head-to-head comparison to the baseline approach, we found a tendency of Macenko's[38] color normalization to improve classifier performance for mutation prediction but not for predition of subtypes or gene expression signatures (Suppl. Fig. 4a-c). Second, we investigated a weakly supervised approach to our baseline of expert-annotated tumor regions and found that the weakly supervised approach was only slightly inferior to manual annotation (Suppl. Fig. 4d-f). Third, we analyzed prediction performance on frozen slides compared to diagnostic slides. While frozen slides are not generally available in a clinical setting, the TCGA database provides an opportunity to perform such a direct comparison. In a weakly supervised experiment, we found that prediction power for driver genes was on par, but prediction power for genetic variants and high-level subtypes/signatures was better in frozen slides than in diagnostic slides (Suppl. Fig. 4g-h). These data provide quantitative guidance for future large-scale validation studies.

External validation of the classification results

Deep learning approaches to a single dataset are prone to overfit and should be validated in external populations before clinical deployment. For external validation of our method, we used routine H&E slides of N=408 colorectal cancer patients from the DACHS study for which BRAF mutational status and CpG-island methylator phenotype (CIMP) was available. We trained deep learning classifiers for BRAF and CIMP on TCGA colorectal cancer samples and evaluated the patient-level accuracy on DACHS. Both features were statistically significantly detectable from DACHS H&E images alone: For BRAF mutants, AUROC was 0.77 (0.64 – 0.82, $p<10^{-5}$) and for CIMP-high, AUROC was 0.66 [0.56– 0.72, $p<10^{-5}$). These data show that deep-learning-based prediction of clinically relevant genetic features can generalize to external patient populations.

Discussion

Image-based genetic testing as a clinical and research tool

Our results demonstrate the feasibility of pan-cancer deep-learning-based inference of a broad range of molecular and genetic features directly from histological images. We show that a unified workflow yields reliably high performance across multiple clinically relevant scenarios without the need to tune technical parameters to a specific molecular target. Our systematic screening approach identifies candidate genetic variants, driver genes, gene expression signatures and standard of care features that can be significantly inferred from histology images, opening up perspectives for large-scale validation of these candidate markers. As a large-scale, systematic screening study, this work identifies a number of mutations which are significantly linked to a detectable phenotype in histological images, including those in key oncogenic pathways including TP53, FBXW7, KRAS, BRAF and CTNNB1. In addition to individually mutated genes, our data show that higher-level gene expression clusters or signatures can be inferred from histological images. Many of these clusters represent groups of patients with distinct and well-described cancer biology such as consensus molecular subtype (CMS) in colorectal cancer. By linking these molecularly defined groups to specific histological image features, our method constitutes a new tool to decipher downstream biological effects of molecular alterations in solid tumors. In an external validation cohort, we show that the models trained on images from the TCGA archive generalize to external patients, demonstrating the potential of applying these methods to routine material from real-world clinical cohorts. Of note, additional retrospective and prospective validation and regulatory approval is needed for histology-based deep learning methods to be implemented in clinical workflows. An example for clinical implementation would be the use as pre-screening tools to enrich patient populations for specific molecular testing. While it is expected that the first applications of deep learning technology in routine workflows will relate to the automatic identification of tumor tissues for the selection of specimens or regions of interest, our method could be easily added to such digital pathology workflows, providing a strong additional incentive for digitization of histopathology.

Limitations

Currently, limitations of our method are the low AUROC values for some molecular features (Fig. 2 and Fig. 3). A strategy to increase the diagnostic performance would be re-training on larger patient cohorts. Re-training can be expected to boost performance because previous studies have shown that performance of deep learning systems in histopathology scales with the number of patients in the training cohort.[40] In addition, the performance of deep learning systems could potentially be improved by technical modifications. Our systematic evaluation of alternative technical approaches provides a guidance for this on multiple levels: First, regarding the choice of neural network models, our results demonstrate that lightweight neural network models perform on par with more complex models, facilitating further evaluation of these methods on decentralized hardware, including desktop or ultimately mobile hardware. While this finding is based on a clinically relevant benchmark task and generalizes to an external population, we cannot exclude that other network models perform better in other histology applications. Second, regarding the type of input image data, other studies in digital pathology have used frozen histology sections[11]. In contrast, our baseline workflow was based on FFPE tissue slides (labeled as 'diagnostic slides' in the TCGA archive) due to their clinical relevance. In clinical settings, frozen specimens constitute only a small fraction of pathology samples and therefore, establishing methods on FFPE material is paramount for large-scale clinical validation. Our head-to head comparison shows that molecular inference generally works better on frozen slides, which is a limitation of the FFPE-based method. Further studies are needed to determine the reasons for this observation. Lastly, our baseline method relied on expert annotations of tumor tissue, constraining deep learning models to learn from invasive tumor tissue only. The rationale behind this design was that despite advances in computer vision, expert annotation of tumor tissue remains the gold standard in histopathology studies. Yet, in a head-to-head comparison, a weakly supervised approach without any manual annotation did not markedly reduce performance, demonstrating feasibility of even simpler data preprocessing pipelines. We publicly release all source codes of our method, enabling further optimization and validation on a larger scale.

Deciphering genotype-phenotype links

Beyond being a potentially useful tool for clinical applications, deep learning-based inference of molecular features from morphology could shed light on more fundamental properties of cancer biology. Our study systematically screens hundreds of molecular alterations and identifies candidates that are linked to detectable patterns in histology images. These patterns can be visualized through prediction maps (Fig. 5a-e). Such "spatialization" of genetic predictions is a key aspect lacking in conventional bulk genetic tests of tumor and could be useful to trace back molecular alterations to specific spatial regions. An alternative approach to understanding deep-learning-based predictions is through visualization of highly ranked image tiles (Fig. 5f-k). This approach can serve as a plausibility control and may help to discover new morphological features. Indeed, highly ranked tiles of CMS classes in colorectal cancer showed poorly differentiated tumor in CMS1 tiles (Fig. 5f), well-differentiated glands for CMS2-3 (Fig. 5g-h) and highly stromal tiles for CMS4 (Fig. 5i). These patterns correspond to known biological processes underlying CMS subclasses, corroborating the assumption that our deep learning system detects biologically meaningful features. Similarly, visualizing histomorphology in the highest predicted tiles in BRAF mutant patients in the validation cohort (Fig. 5j-k) demonstrated poorly differentiated areas and mucinous areas as recurring features in BRAF mutant image tiles, which is consistent with previous studies.[52] Visualizing highly predicted tiles in gastric cancer (Suppl. Fig. 22a-h) highlighted highly cellular areas as correlates of a "Proliferation" gene expression signature, but at the same time identified patterns for mutations (e.g. in AMER1 and MTOR) which could help to form new hypotheses on how these specific mutations influence cancer cell behavior and morphology. Interestingly, the prediction performance markedly varied between the 14 different types of cancer (Fig. 2, Suppl. Fig. 7). Variations in sample size between the cohorts could explain some of these differences, but additional biological effects could contribute to this. One hypothesis is that tumor types with few clinically targetable mutations (e.g. lung squamous cell cancer and pancreatic cancer) also display few detectable mutations. Further studies are warranted to investigate this.

Conclusion

Together, our results demonstrate that molecular changes in solid tumors can be inferred from routine histology alone with deep learning. This could be a useful tool to objectively elucidate

genotype-phenotype relationships in cancer and ultimately, could be used as a low-cost biomarker in clinical trials and routine clinical workflows.

Competing interests

JNK has an informal, unpaid advisory role at Pathomix, Heidelberg, Germany. No other competing interests exist.

Funding

The results are in part based upon data generated by the TCGA Research Network: http://cancergenome.nih.gov/. Our funding sources are as follows. J.N.K.: RWTH University Aachen (START 2018-691906). V.S.: Breast Cancer Now, P.Bo: DFG: (SFB/TRR57, SFB/TRR219, BO3755/3-1, and BO3755/6-1), the German Ministry of Education and Research (BMBF: STOP-FSGS-01GM1901A) and the German Ministry of Economic Affairs and Energy (BMWi: EMPAIA project). A.T.P.: NIH/NIDCR (#K08-DE026500), Institutional Research Grant (#IRG-16-222-56) from the American Cancer Society, Cancer Research Foundation Research Grant, and the University of Chicago Medicine Comprehensive Cancer Center Support Grant (#P30-CA14599). T.L.: Horizon 2020 through the European Research Council (ERC) Consolidator Grant PhaseControl (771083), a Mildred-Scheel-Endowed Professorship from the German Cancer Aid (Deutsche Krebshilfe), the German Research Foundation (DFG) (SFB CRC1382/P01, SFB-TRR57/P06, LU 1360/3-1), the Ernst-Jung-Foundation Hamburg and the IZKF (interdisciplinary center of clinical research) at RWTH Aachen.

Authors' contributions

JNK, ATP and TL designed the study. LH, HIG, NAC, JJS, PAVDB, LFSK, PBo and AP oversaw the tumor annotation. CL, AE, JK, HSM, JMN, RDB and KAJS manually annotated all tumors. JNK, JK, JMN and PBa designed and implemented the algorithm. JNK, CL, AS, SK, RDB and NOB curated the list of molecular alterations. HB, MH, ATP, AMH and VS provided external validation samples and gave statistical advice. CT, DJ, ATP, PBo, VS and TL provided infrastructure and supervised the study. All authors contributed to the data analysis and writing the manuscript.

Figure Legends

Fig. 1: Deep learning workflow for prediction of molecular features from histology images. We describe a comprehensive method pipeline for prediction of molecular features directly from histological images. (a) Training of the deep learning system comprised six steps. Step 1: Patient cohorts were randomly split into three partitions for cross-validation of deep classifiers. Step 2: The tumor region on each whole slide image (WSI) was tessellated into tiles. Step 3: Up to 500 randomly chosen tiles were collected. Step 4: Tiles from patients in the training partitions were collected, classes were equalized by random undersampling. Step 5: All training tiles were used to train a deep neural network (pre-trained on a non-medical task). Step 6: Classification performance was evaluated on patients from the test partition. (b) For patient-level inference of molecular labels in patients not seen during training, three successive steps were used. Step 1: Tiles were generated from the tumor region on WSI. Step 2: A prediction was made for each tile. Step 3: Tile-level class predictions were pooled on a patient level. (c) Hyperparameters of the deep learning system were optimized in a benchmark task (prediction of microsatellite instability status [MSI] in colorectal cancer). The opacity of each point corresponds to the number of trainable layers (Suppl. Table 3). Shufflenet, a lightweight neural network architecture was selected as a highly efficient network model. (d) This workflow was subsequently applied for prediction of four types of molecular features across 14 cancer types. In particular, this included genetic mutations. The distribution of the 20 most common among all analyzed mutations is shown for each tumor type.

Fig. 2: Inference of genetic mutations from histological images. A deep learning system was trained to predict mutational status (mutated or wild-type) of relevant genes in 14 cancer type and was evaluated by cross-validation. All mutations, including variants of unknown significance, were included in the 'mutated' class. For each gene, patient-level test set performance is shown as area under the receiver operating curve (AUROC) with p-value for prediction scores corrected for multiple testing (false detection rate, FDR). The significance level of 0.05 is marked with a line and the distribution of p-values in each panel is shown as a density plot. P values smaller than 10^{-5} are set to 10^{-5}. N denotes the number of patients per tumor type. (a-d) In lung adenocarcinoma,

colorectal cancer, breast cancer and gastric cancer, a number of relevant genes were significantly predictable from histology alone, including key oncogenic drivers such as TP53, BRAF and MTOR. (e-n) In all other tested tumor types, mutational status was predictable for some genes, with notable examples including KRAS in pancreatic cancer, CTNNB1 in hepatocellular carcinoma and TP53 and CASP8 in head and neck cancer.

Fig. 3: Inference of putative oncogenic drivers from histological images. A deep learning system was trained to predict oncogenic driver genes from histology. Only putative and confirmed drivers were included and variants of unknown significance were pooled with the "wild type" class. (a-n) This process uncovered significant predictability of multiple oncogenic drivers, including EGFR, BRAF and TP53.

Fig. 4: Inference of molecular subtypes, gene expression signatures and standard biomarkers directly from histology. In addition to prediction of single-gene mutations, the capability of deep learning to infer high-level molecular features was systematically assessed. (a-d) In lung, colorectal, breast and gastric cancer, gene expression signatures (such as TCGA molecular subtype in any tumor type) and standard of care features (such as hormone receptor status in breast cancer) were highly predictable from histology alone, as shown by the distribution of false-detection rate (FDR)-corrected p-values. (e-h) Gene expression signatures for Proliferation (Prolif), Wound Healing (WoundHeal), Macrophage infiltration (Mcrphg), Homologous Repair Deficiency (HRD), CD8-positive Lymphocyte (LymCD8), TCGA molecular subtypes (LUAD 1-6), pan-gastrointestinal (GI) molecular subtypes, consensus molecular subtypes (CMS), PAM50 subtypes and other key molecular features were highly predictable across multiple tumor types. Patient-level AUROC with bootstrapped confidence intervals, * denotes FDR-p-value < 0.05. (i-j) Standard of care biomarkers including estrogen and progesterone receptor (ER and PR) status in breast cancer, pathologic subtype and microsatellite instability (MSI) were highly predictable from routine histology alone by deep learning.

Fig. 5: Explainability of deep learning-based analysis of histological images. Deep learning-based predictions were visualized through genotype maps and comparison of highly ranked image tiles. (a-e) Prediction maps for consensus molecular subtype (CMS) in colorectal cancer show spatially

resolved prediction scores, unveiling intratumor heterogeneity of predicted genotype. As a generic tool, this visualization approach allows to identify spatial regions associated with a molecular feature. In this patient, the correct prediction of CMS4 correctly show that deep learning robustly predicts CMS from histology alone while highlighting potential intratumor heterogeneity (f-i) For each of the CMS classes, the most highly scored test set tiles are shown, enabling correlation of deep learning-predictions with histopathological features at high resolution. In this case, highly predicted CMS1 tiles contain numerous tumor-infiltrating lymphocytes while predicted CMS4 tiles contain abundant stroma, consistent with previous studies. (j-k) Highly scored tiles in the external test cohort DACHS for prediction of BRAF mutant and wild type (l-m) and CpG-island methylator phenotype (CIMP) high or non-CIMP.

Bibliography

1. Cheng, M.L., Berger, M.F., Hyman, D.M. & Solit, D.B. Clinical tumour sequencing for precision oncology: time for a universal strategy. *Nature Reviews Cancer* 18, 527-528 (2018).
2. Rusch, M*., et al.* Clinical cancer genomic profiling by three-platform sequencing of whole genome, whole exome and transcriptome. *Nature Communications* 9, 3962 (2018).
3. Kather, J.N., Halama, N. & Jaeger, D. Genomics and emerging biomarkers for immunotherapy of colorectal cancer. *Seminars in Cancer Biology* 52, 189-197 (2018).
4. Guinney, J*., et al.* The consensus molecular subtypes of colorectal cancer. *Nature Medicine* 21, 1350 (2015).
5. Fontana, E., Eason, K., Cervantes, A., Salazar, R. & Sadanandam, A. Context matters-consensus molecular subtypes of colorectal cancer as biomarkers for clinical trials. *Ann Oncol* 30, 520-527 (2019).
6. Shia, J*., et al.* Morphological characterization of colorectal cancers in The Cancer Genome Atlas reveals distinct morphology-molecular associations: clinical and biological implications. *Modern pathology : an official journal of the United States and Canadian Academy of Pathology, Inc* 30, 599-609 (2017).
7. Greenson, J.K*., et al.* Pathologic predictors of microsatellite instability in colorectal cancer. *The American journal of surgical pathology* 33, 126-133 (2009).
8. Greenson, J.K*., et al.* Pathologic predictors of microsatellite instability in colorectal cancer. *Am J Surg Pathol* 33, 126-133 (2009).
9. Le, D.T*., et al.* PD-1 Blockade in Tumors with Mismatch-Repair Deficiency. *New England Journal of Medicine* 372, 2509-2520 (2015).
10. Kather, J.N*., et al.* Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. *Nature Medicine* (2019).
11. Coudray, N*., et al.* Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. *Nature Medicine* 24, 1559-1567 (2018).
12. Sha, L*., et al.* Multi-Field-of-View Deep Learning Model Predicts Nonsmall Cell Lung Cancer Programmed Death-Ligand 1 Status from Whole-Slide Hematoxylin and Eosin Images. *J Pathol Inform* 10, 24 (2019).
13. Schaumberg, A.J., Rubin, M.A. & Fuchs, T.J. H&E-stained Whole Slide Image Deep Learning Predicts SPOP Mutation State in Prostate Cancer. *bioRxiv*, 064279 (2018).
14. Kather, J.N*., et al.* Deep learning detects virus presence in cancer histology. *bioRxiv*, 690206 (2019).
15. Zhang, H*., et al.* Predicting Tumor Mutational Burden from Liver Cancer Pathological Images Using Convolutional Neural Network. in *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* 920-925 (2019).
16. The Cancer Genome Atlas Network*, et al.* Comprehensive molecular portraits of human breast tumours. *Nature* 490, 61 (2012).
17. Burk, R.D*., et al.* Integrated genomic and molecular characterization of cervical cancer. *Nature* 543, 378-384 (2017).
18. Muzny, D.M*., et al.* Comprehensive molecular characterization of human colon and rectal cancer. *Nature* 487, 330-337 (2012).
19. The Cancer Genome Atlas Network*, et al.* Comprehensive molecular characterization of gastric adenocarcinoma. *Nature* 513, 202 (2014).
20. The Cancer Genome Atlas Network*, et al.* Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature* 517, 576 (2015).

21. The Cancer Genome Atlas Consortium. Comprehensive and Integrative Genomic Characterization of Hepatocellular Carcinoma. *Cell* 169, 1327-1341.e1323 (2017).
22. The Cancer Genome Atlas Network*, et al.* Comprehensive molecular profiling of lung adenocarcinoma. *Nature* 511, 543 (2014).
23. Hammerman, P.S.*, et al.* Comprehensive genomic characterization of squamous cell lung cancers. *Nature* 489, 519-525 (2012).
24. Cancer Genome Atlas, N. Genomic Classification of Cutaneous Melanoma. *Cell* 161, 1681-1696 (2015).
25. The Cancer Genome Atlas Network. Integrated Genomic Characterization of Pancreatic Ductal Adenocarcinoma. *Cancer Cell* 32, 185-203.e113 (2017).
26. The Cancer Genome Atlas Network. The Molecular Taxonomy of Primary Prostate Cancer. *Cell* 163, 1011-1025 (2015).
27. Davis, C.F.*, et al.* The somatic genomic landscape of chromophobe renal cell carcinoma. *Cancer Cell* 26, 319-330 (2014).
28. Creighton, C.J.*, et al.* Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature* 499, 43-49 (2013).
29. Linehan, W.M.*, et al.* Comprehensive Molecular Characterization of Papillary Renal-Cell Carcinoma. *N Engl J Med* 374, 135-145 (2016).
30. Hoffmeister, M.*, et al.* Statin use and survival after colorectal cancer: the importance of comprehensive confounder adjustment. *J Natl Cancer Inst* 107, djv045 (2015).
31. Brenner, H., Chang-Claude, J., Seiler, C.M. & Hoffmeister, M. Long-term risk of colorectal cancer after negative colonoscopy. *J Clin Oncol* 29, 3761-3767 (2011).
32. Cerami, E.*, et al.* The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov* 2, 401-404 (2012).
33. Gao, J.*, et al.* Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci Signal* 6, pl1 (2013).
34. Bailey, M.H.*, et al.* Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell* 173, 371-385 e318 (2018).
35. Berger, A.C.*, et al.* A Comprehensive Pan-Cancer Molecular Study of Gynecologic and Breast Cancers. *Cancer Cell* 33, 690-705.e699 (2018).
36. Liu, Y.*, et al.* Comparative Molecular Analysis of Gastrointestinal Adenocarcinomas. *Cancer Cell* 33, 721-735.e728 (2018).
37. Thorsson, V.*, et al.* The Immune Landscape of Cancer. *Immunity* 48, 812-830.e814 (2018).
38. Macenko, M.*, et al.* A method for normalizing histology slides for quantitative analysis. in *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro* 1107-1110 (2009).
39. Bianconi, F., Kather, J.N. & Reyes-Aldasoro, C.C. Evaluation of Colour Pre-processing on Patch-Based Classification of H&E-Stained Images. in *European Congress on Digital Pathology* 56-64 (Springer, 2019).
40. Campanella, G.*, et al.* Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature Medicine* (2019).
41. Zhang, X., Zhou, X., Lin, M. & Sun, J. Shufflenet: An extremely efficient convolutional neural network for mobile devices. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 6848-6856 (2018).
42. Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K.Q. Densely connected convolutional networks. in *Proceedings of the IEEE conference on computer vision and pattern recognition* 4700-4708 (2017).

43. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. & Wojna, Z. Rethinking the inception architecture for computer vision. in *Proceedings of the IEEE conference on computer vision and pattern recognition* 2818-2826 (2016).
44. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. in *Proceedings of the IEEE conference on computer vision and pattern recognition* 770-778 (2016).
45. Srinidhi, C.L., Ciga, O. & Martel, A.L. Deep neural network models for computational histopathology: A survey. *arXiv preprint arXiv:1912.12378* (2019).
46. Chen, P.C*., et al.* An augmented reality microscope with real-time artificial intelligence integration for cancer diagnosis. *Nature Medicine* (2019).
47. Fukamachi, H*., et al.* A subset of diffuse-type gastric cancer is susceptible to mTOR inhibitors and checkpoint inhibitors. *Journal of Experimental & Clinical Cancer Research* 38, 127 (2019).
48. André, F*., et al.* Alpelisib for PIK3CA-Mutated, Hormone Receptor–Positive Advanced Breast Cancer. *New England Journal of Medicine* 380, 1929-1940 (2019).
49. Xue, Z*., et al.* MAP3K1 and MAP2K4 mutations are associated with sensitivity to MEK inhibitors in multiple cancer models. *Cell Research* 28, 719-729 (2018).
50. Khalaf, A.M*., et al.* Role of Wnt/beta-catenin signaling in hepatocellular carcinoma, pathogenesis, and clinical significance. *J Hepatocell Carcinoma* 5, 61-73 (2018).
51. Li, C., Egloff, A.M., Sen, M., Grandis, J.R. & Johnson, D.E. Caspase-8 mutations in head and neck cancer confer resistance to death receptor-mediated apoptosis and enhance migration, invasion, and tumor growth. *Molecular oncology* 8, 1220-1230 (2014).
52. Barresi, V., Bonetti, L.R. & Bettelli, S. KRAS, NRAS, BRAF mutations and high counts of poorly differentiated clusters of neoplastic cells in colorectal cancer: observational analysis of 175 cases. *Pathology* 47, 551-556 (2015).

Supplementary Figure Legends

Suppl. Fig. 1: Additional details on the statistical procedures. (a) For patient-level three-fold cross-validation, the patient cohort was split into three random partitions. Each partition had approximately the same proportion of patients within each class. Three classifiers were trained and their patient-level predictions on the respective test set were concatenated. Thus, a prediction was gained for each patient in the cohort, but no patient was ever part of a training set and a test set of the same classifier at the same time. (b) The percentage of predicted tiles for each class was used for a receiver operating characteristic (ROC) analysis with 10x bootstrapped pointwise confidence bounds. (c) In addition to the ROC analysis, the prediction scores (percent predicted tiles) for patients in each class was compared to prediction scores for patients in all other classes. The resulting false-detection-rate (FDR)-corrected p-value in a two-tailed t-test for this comparison was reported for each feature of interest. Icons are from Twitter Twemoji (CC-BY 4.0 license).

Suppl. Fig. 2: Distribution of tumor content across slides in all tumor types. Central mark = median, bottom and top edge of the box = 25^{th} and 75^{th} percentile, line extends to the most extreme data points, circles = outliers. Outliers larger than 2000 mm^2 are not plotted. Median tumor content on slide is 139 mm^2 of tumor tissue per slide for colorectal cancer (CRC).

Suppl. Fig. 3: Design of additional technical optimization experiments. The baseline approach in this study was to perform image analysis of tiles based on manual tumor annotations in every single tissue slide, without performing any color normalization. This approach was compared to three alternative approaches as shown here.

Suppl. Fig. 4: Results of additional technical optimization experiments: Normalization. (a) Comparison of cross-validated absolute differences in AUROC to the baseline model (no normalization), genetic variants. (b) Comparison of AUROC differences for genetic driver mutations. (c) Comparison of AUROC differences for expression signatures and subtypes.

Suppl. Fig. 5: Results of additional technical optimization experiments: Weakly supervised. (a) Comparison of cross-validated absolute differences in AUROC to the baseline model (no normalization), genetic variants. (b) Comparison of AUROC differences for genetic driver mutations. (c) Comparison of AUROC differences for expression signatures and subtypes.

Suppl. Fig. 6: Results of additional technical optimization experiments: Frozen tissue. (a) Comparison of cross-validated absolute differences in AUROC to the baseline model (no normalization), genetic variants. (b) Comparison of AUROC differences for genetic driver mutations. (c) Comparison of AUROC differences for expression signatures and subtypes.

Suppl. Fig. 7: Distribution of predictability scores for feature classes in all cancer types. Target features were assigned to one of four categories as shown in Suppl. Table 1: Genetic variants, oncogenic drivers, high-level signatures and standard-of-care features. For each of these classes, predictability by deep learning was assessed and the distribution of false-detection-rate (FDR)-corrected p-values is shown, with low p-values capped at 10^{-5}. High-level signatures were highly predictable in most tumor types.

Suppl. Fig. 8: Detailed prediction statistics for lung adenocarcinoma (LUAD). (a-c) Area under the receiver operating curve (AUROC) with corresponding p-values, for each feature. (e-h) Detailed view of the features with highest AUROC values. Low p-values capped at 10^{-5}.

Suppl. Fig. 9: Detailed prediction statistics for colorectal cancer (COAD, READ). (a-c) Area under the receiver operating curve (AUROC) with corresponding p-values, for each feature. (e-h) Detailed view of the features with highest AUROC values. Low p-values capped at 10^{-5}.

Suppl. Fig. 10: Detailed prediction statistics for breast cancer (BRCA). (a-c) Area under the receiver operating curve (AUROC) with corresponding p-values, for each feature. (e-h) Detailed view of the features with highest AUROC values. Low p-values capped at 10^{-5}.

Suppl. Fig. 11: Detailed prediction statistics for gastric cancer (STAD). (a-c) Area under the receiver operating curve (AUROC) with corresponding p-values, for each feature. (e-h) Detailed view of the features with highest AUROC values. Low p-values capped at 10^{-5}.

Suppl. Fig. 12: Detailed prediction statistics for melanoma (SKCM) primary tumors and metastases. (a-c) Area under the receiver operating curve (AUROC) with corresponding p-values, for each feature, for primary tumors. (e-h) Detailed view of the features with highest AUROC values. Low p-values capped at 10^{-5}, for primary tumors. (i-l)

Suppl. Fig. 13: Detailed prediction statistics for prostate cancer (PRAD). (a-c) Area under the receiver operating curve (AUROC) with corresponding p-values, for each feature. (e-h) Detailed view of the features with highest AUROC values. Low p-values capped at 10^{-5}.

Suppl. Fig. 14: Detailed prediction statistics for pancreatic cancer (PAAD). (a-c) Area under the receiver operating curve (AUROC) with corresponding p-values, for each feature. (e-h) Detailed view of the features with highest AUROC values. Low p-values capped at 10^{-5}.

Suppl. Fig. 15: Detailed prediction statistics for lung squamous cell carcinoma (LUSC). (a-c) Area under the receiver operating curve (AUROC) with corresponding p-values, for each feature. (e-h) Detailed view of the features with highest AUROC values. Low p-values capped at 10^{-5}.

Suppl. Fig. 16: Detailed prediction statistics for hepatocellular carcinoma (LIHC). (a-c) Area under the receiver operating curve (AUROC) with corresponding p-values, for each feature. (e-h) Detailed view of the features with highest AUROC values. Low p-values capped at 10^{-5}.

Suppl. Fig. 17: Detailed prediction statistics for renal papillary cancer (KIRP). (a-c) Area under the receiver operating curve (AUROC) with corresponding p-values, for each feature. (e-h) Detailed view of the features with highest AUROC values. Low p-values capped at 10^{-5}.

Suppl. Fig. 18: Detailed prediction statistics for renal clear cell cancer (KIRC). (a-c) Area under the receiver operating curve (AUROC) with corresponding p-values, for each feature. (e-h) Detailed view of the features with highest AUROC values. Low p-values capped at 10^{-5}.

Suppl. Fig. 19: Detailed prediction statistics for renal chromophobe cancer (KICH). (a-c) Area under the receiver operating curve (AUROC) with corresponding p-values, for each feature. (e-h) Detailed view of the features with highest AUROC values. Low p-values capped at 10^{-5}.

Suppl. Fig. 20: Detailed prediction statistics for head and neck cancer (HNSC). (a-c) Area under the receiver operating curve (AUROC) with corresponding p-values, for each feature. (e-h) Detailed view of the features with highest AUROC values. Low p-values capped at 10^{-5}.

Suppl. Fig. 21: Detailed prediction statistics for cervical cancer (CESC). (a-c) Area under the receiver operating curve (AUROC) with corresponding p-values, for each feature. (e-h) Detailed view of the features with highest AUROC values. Low p-values capped at 10^{-5}.

Suppl. Fig. 22: Highest scoring tiles for molecular features in gastric cancer (STAD). (a-b) Top tiles corresponding to AMER1 mutational status. (c-d) Top tiles corresponding to MTOR mutational status. (e-f) Top tiles corresponding to high or low values of a proliferation signature. (a-b) Top tiles corresponding to hypermutated samples.

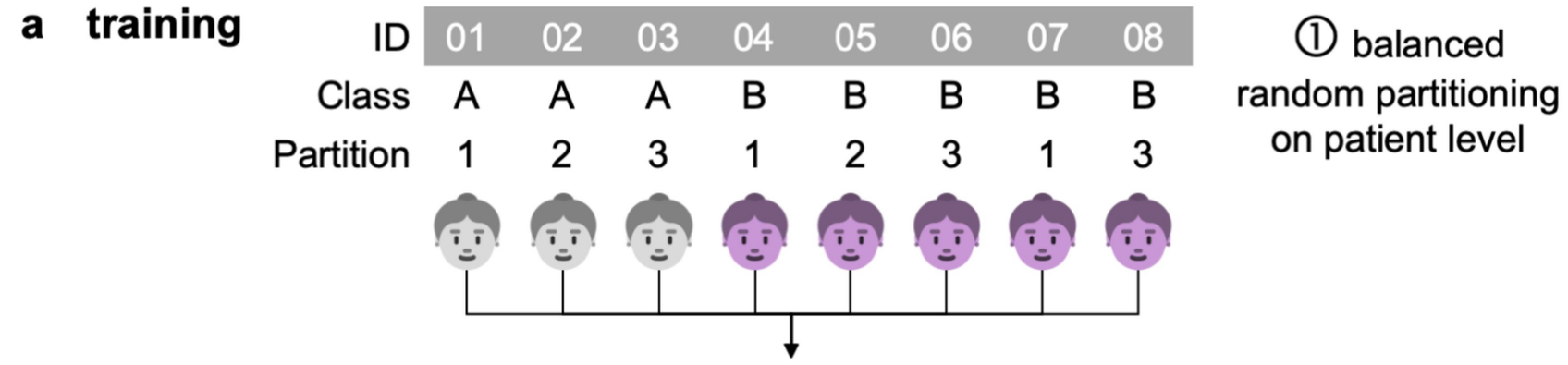
Supplementary Table Legends

Suppl. Table 1: All investigated molecular labels.

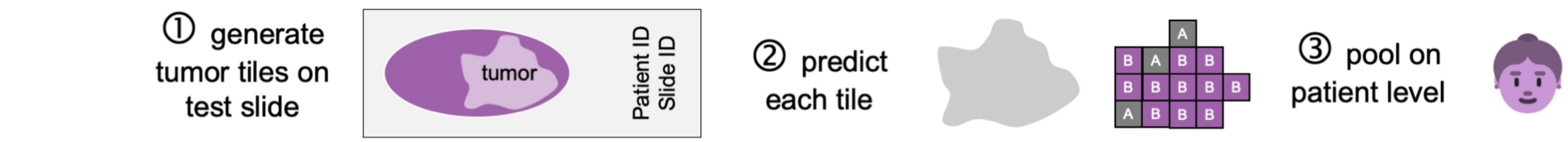
Suppl. Table 2: Slide numbers and case numbers for each cohort (diagnostic slides, TCGA). For melanoma (TCGA-SKCM), the total number of patients included in the analysis was N=430, of which N=290 had a tissue slide of the primary tumor available and N=141 had a tissue slide of metastatic tissue available.

Suppl. Table 3: All raw values for prediction experiments, alternative methods and hyperparameter optimization experiments.

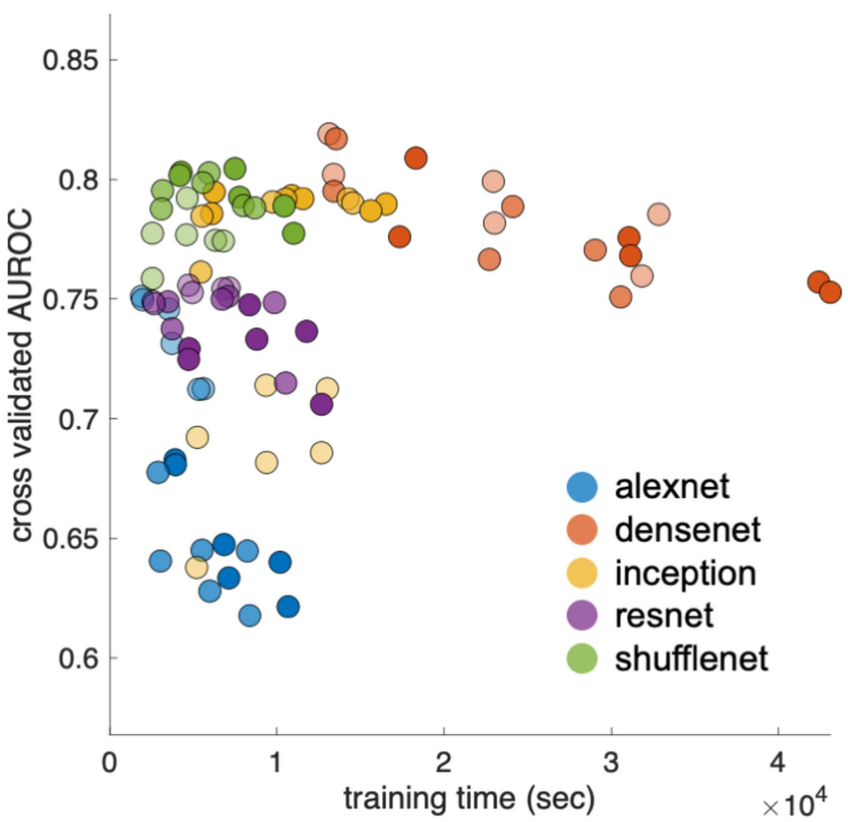
a training



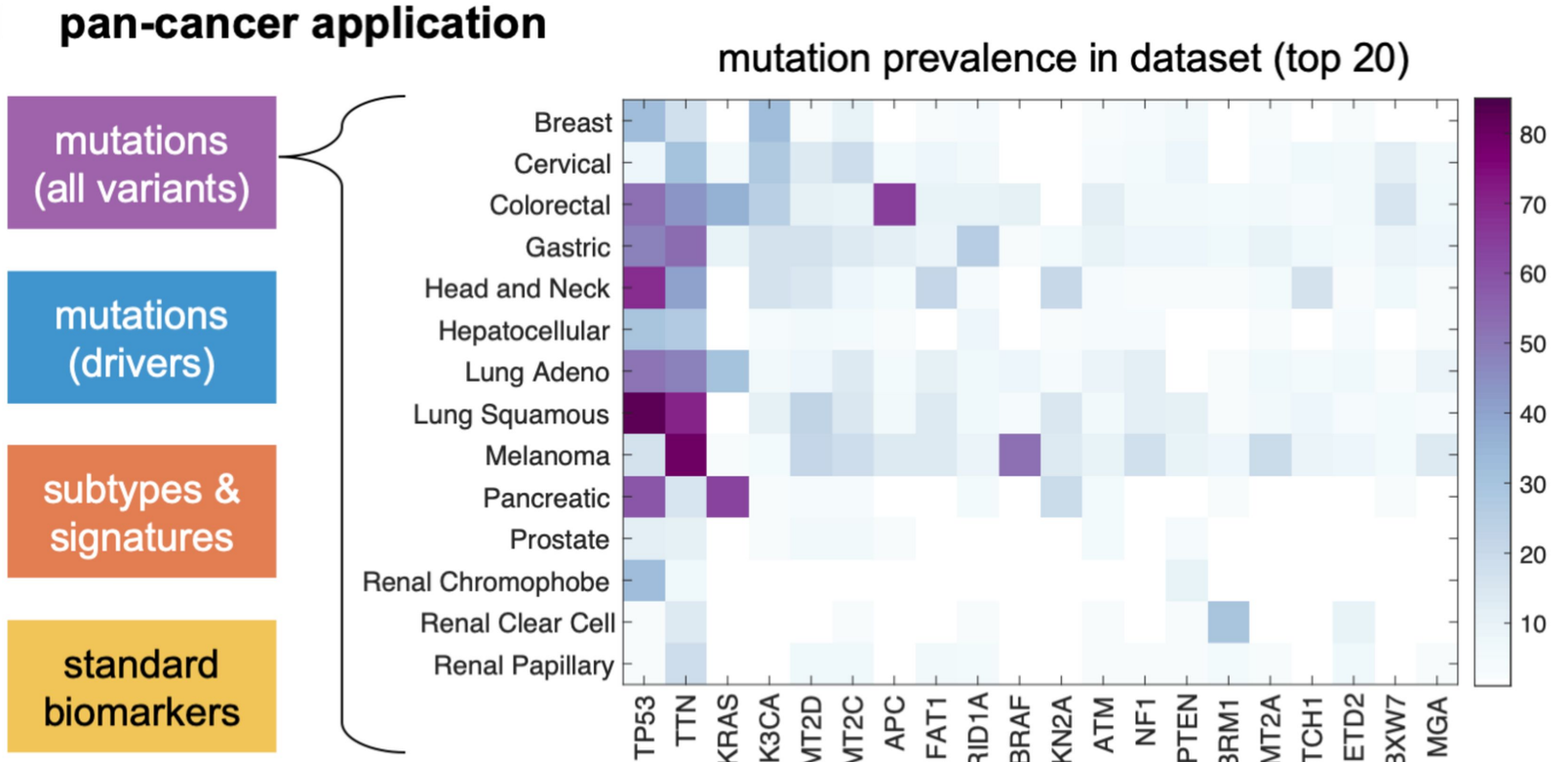
b inference



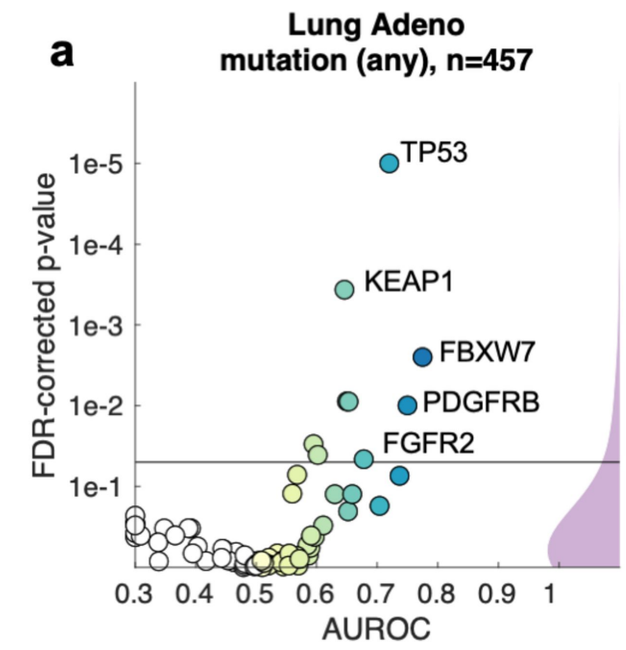
c model selection

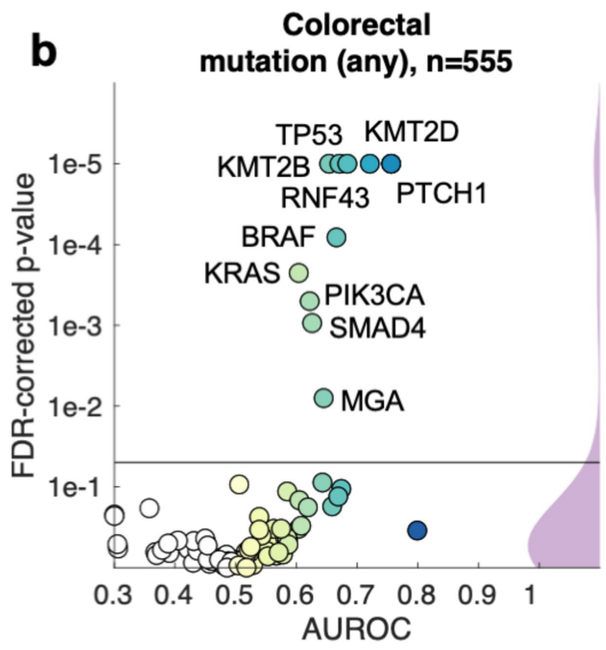


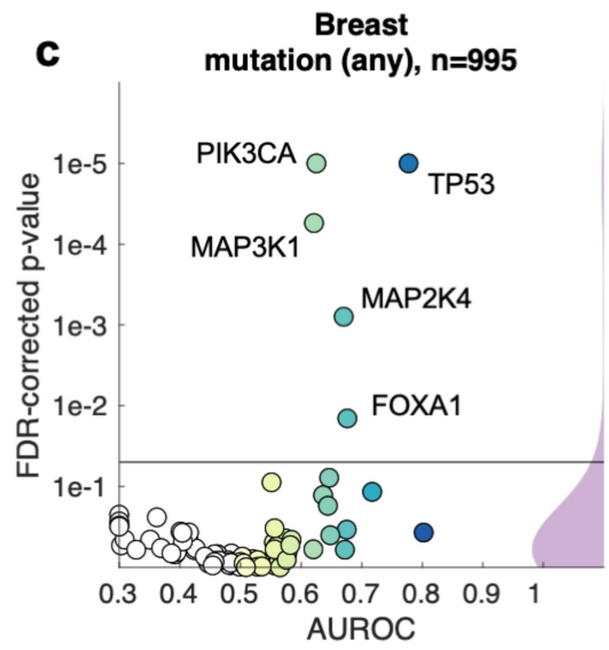
d pan-cancer application

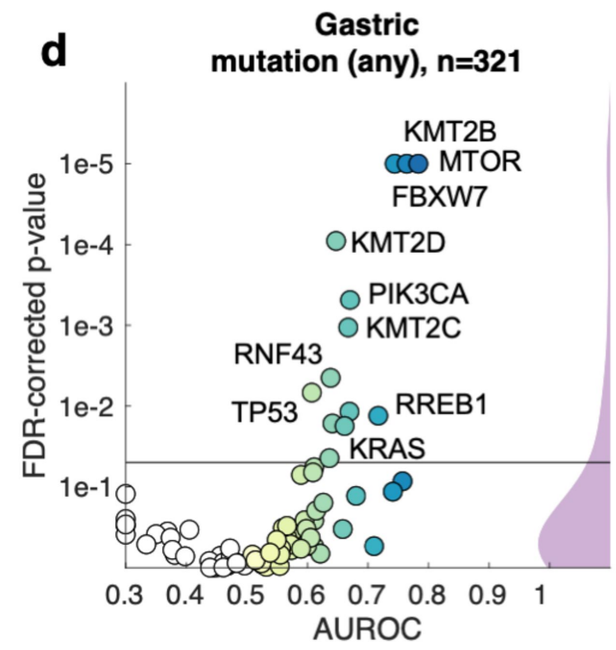


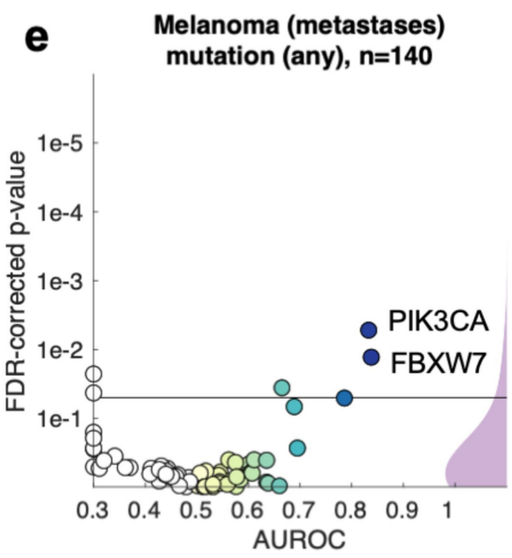
mutations (all variants)

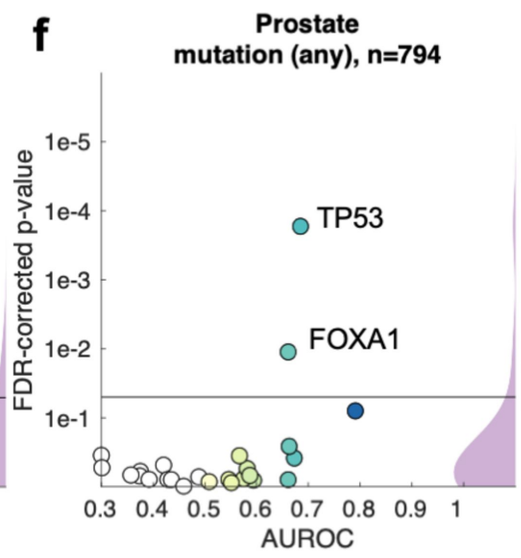


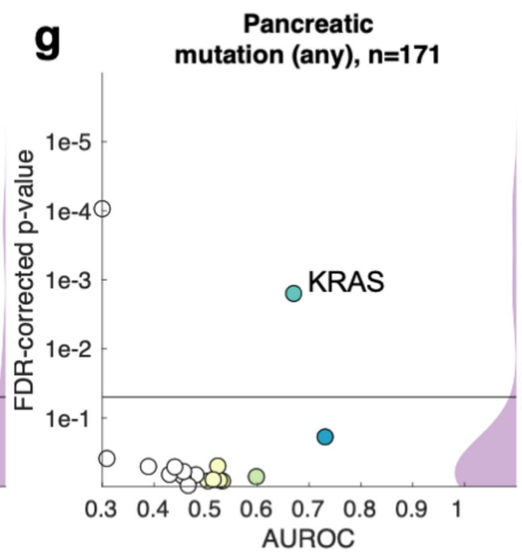


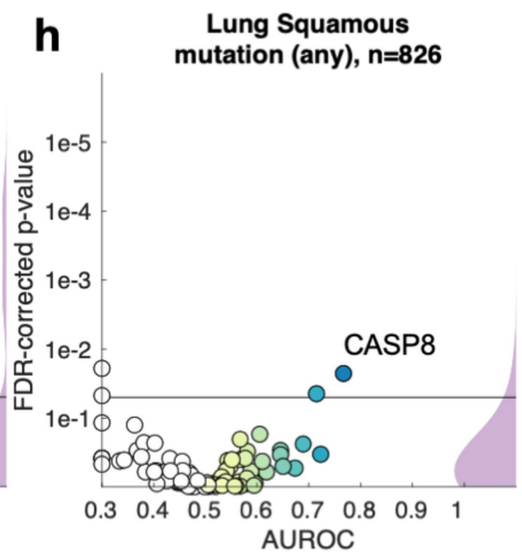


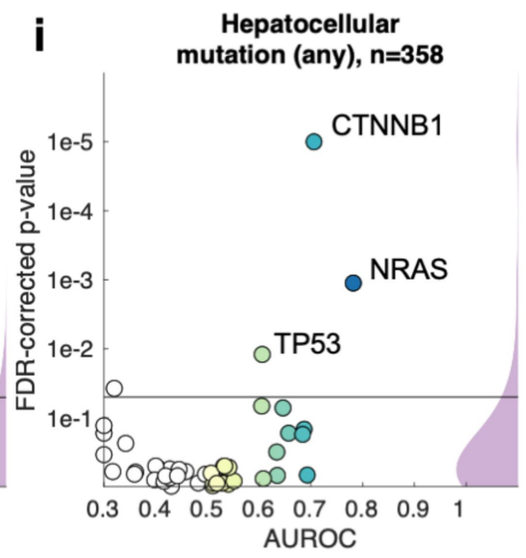


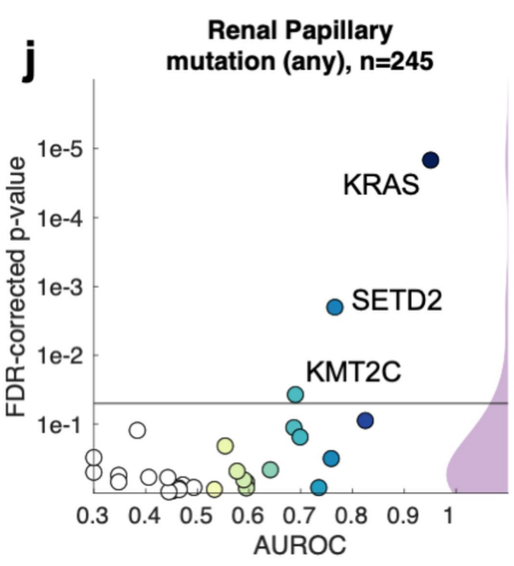


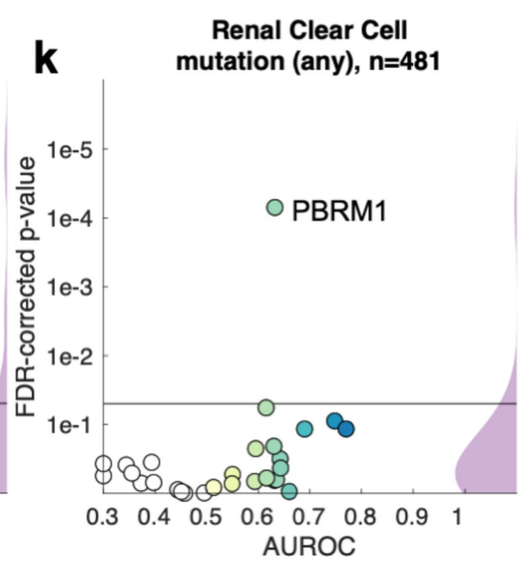


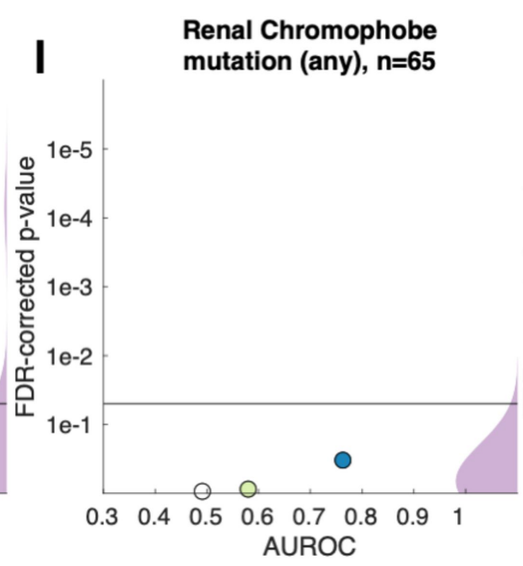


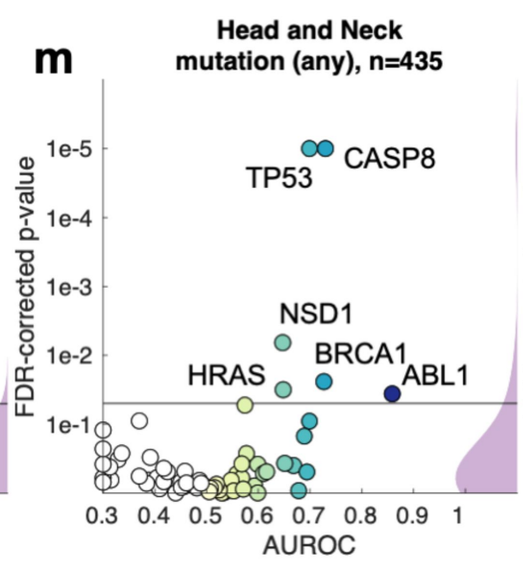


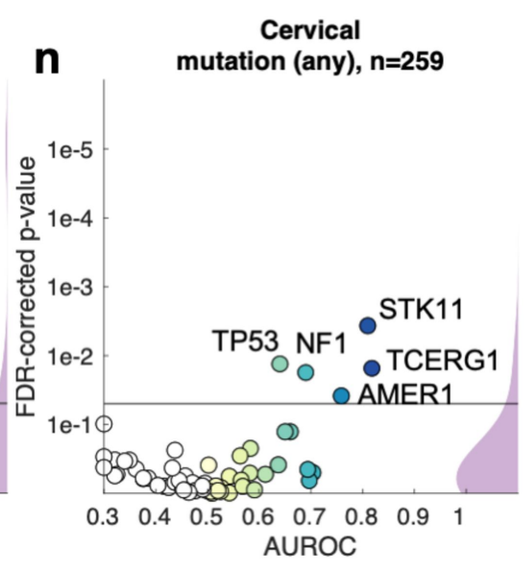




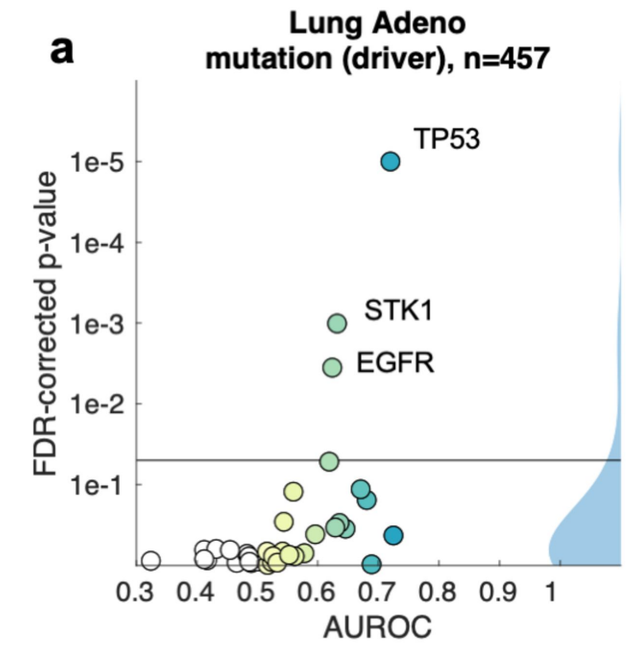


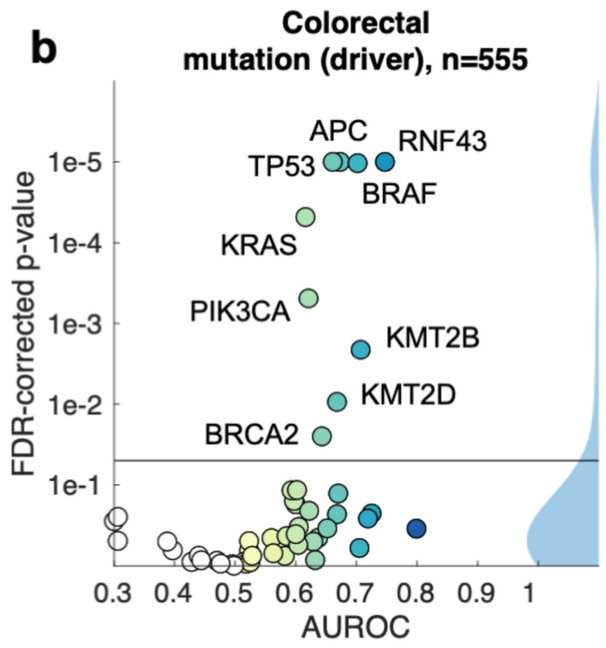


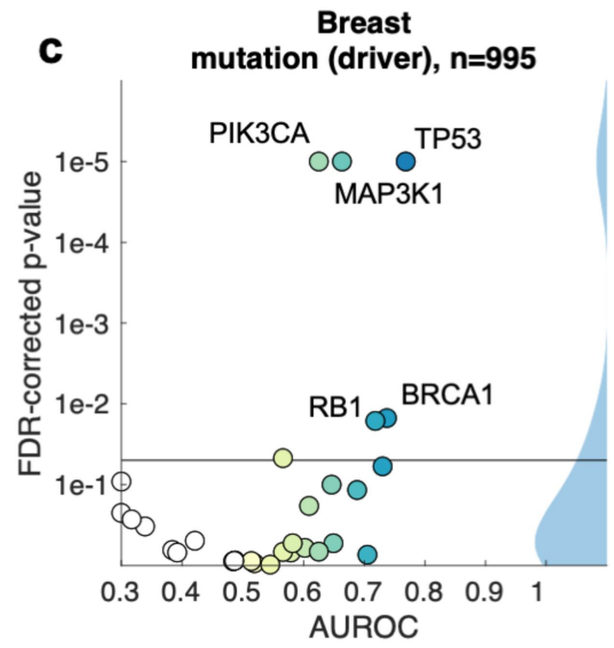


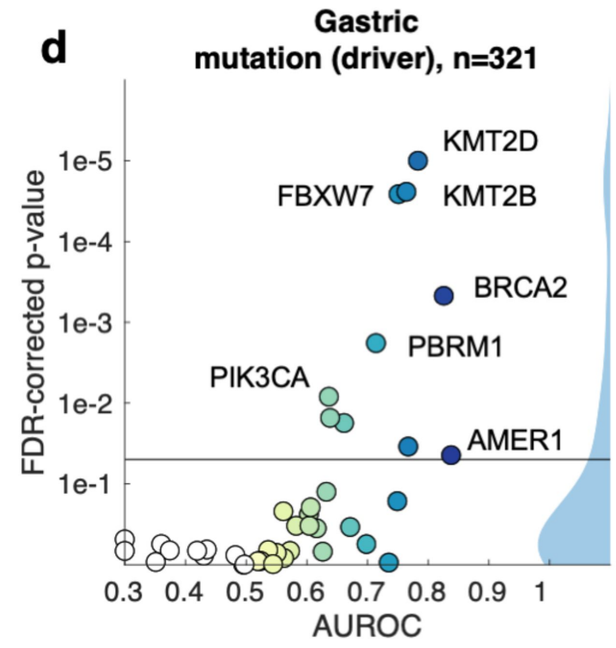


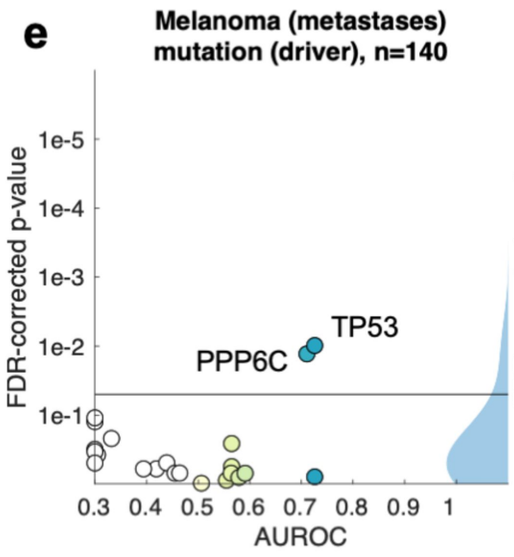
mutations (putative drivers)

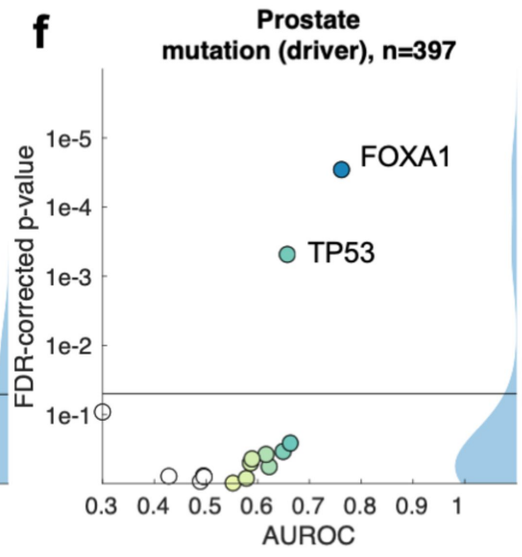


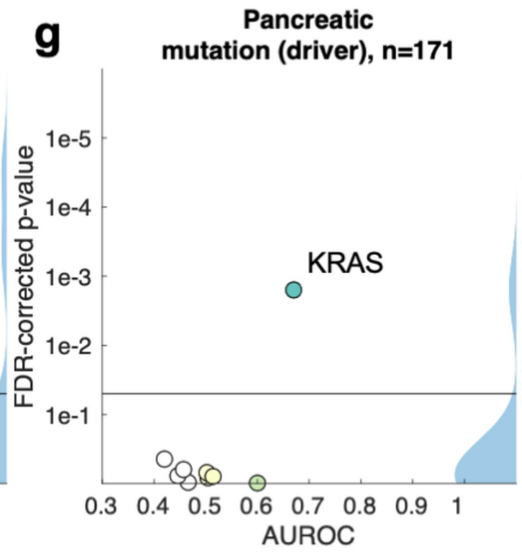


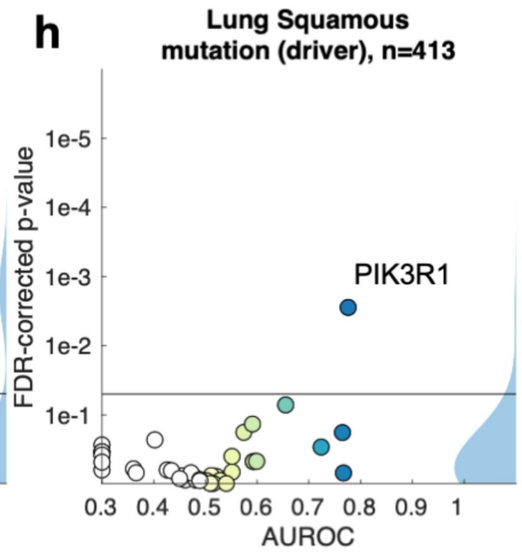


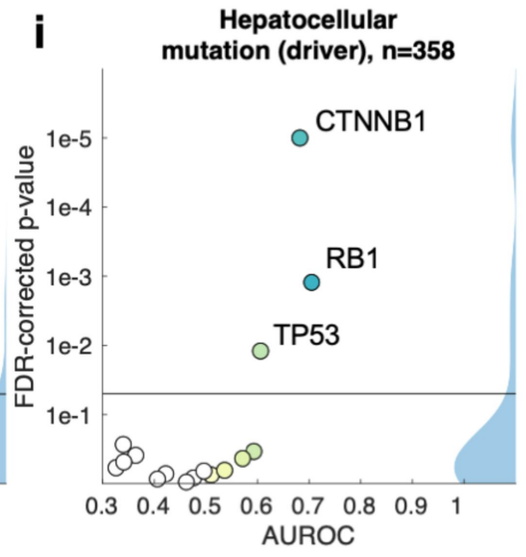


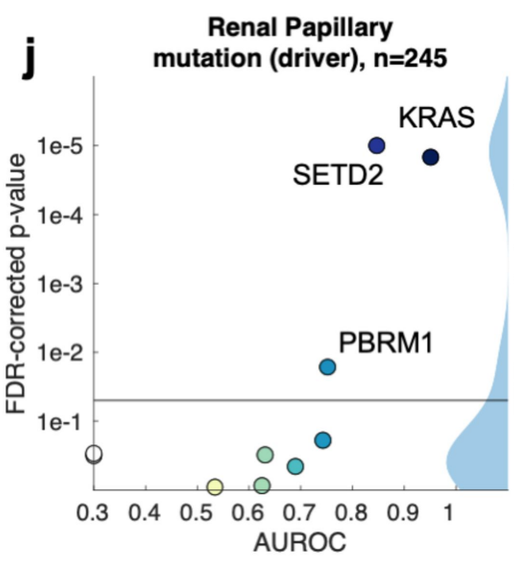


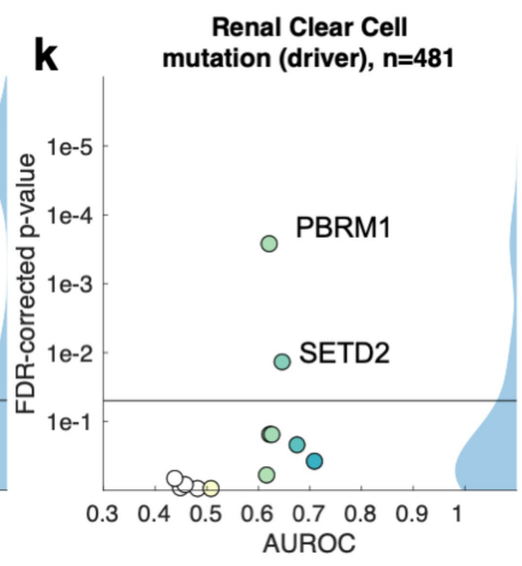


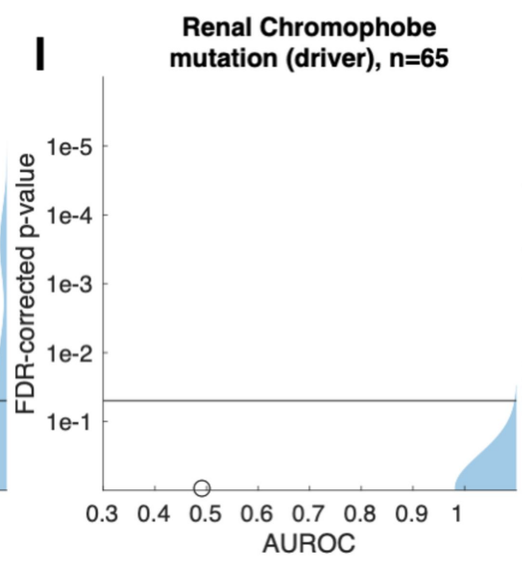


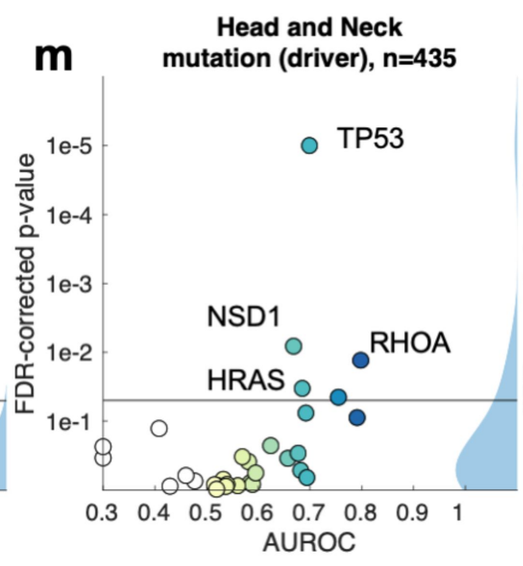


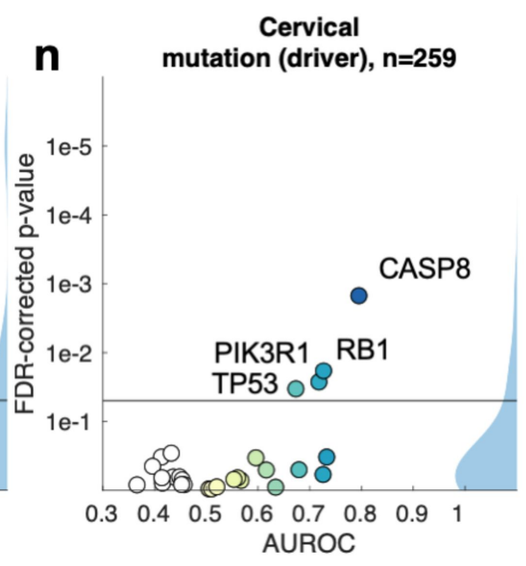


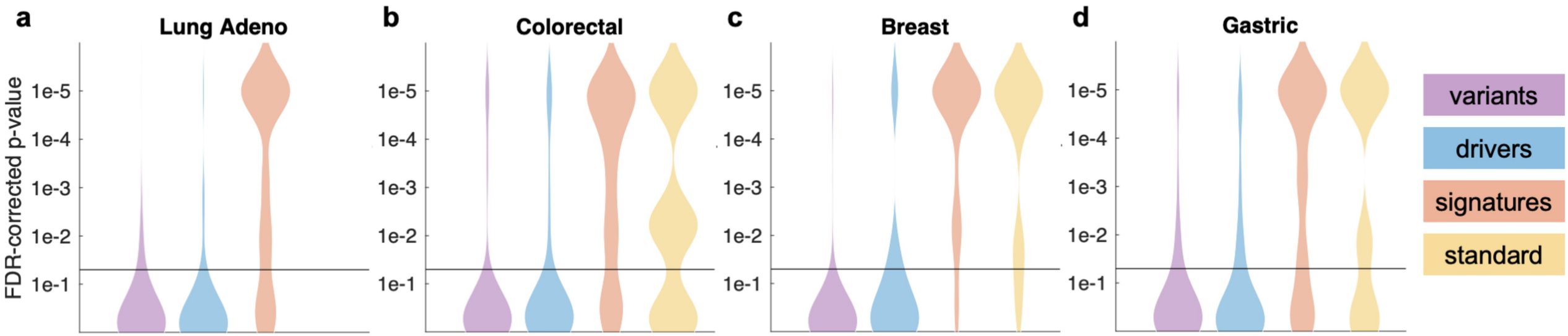




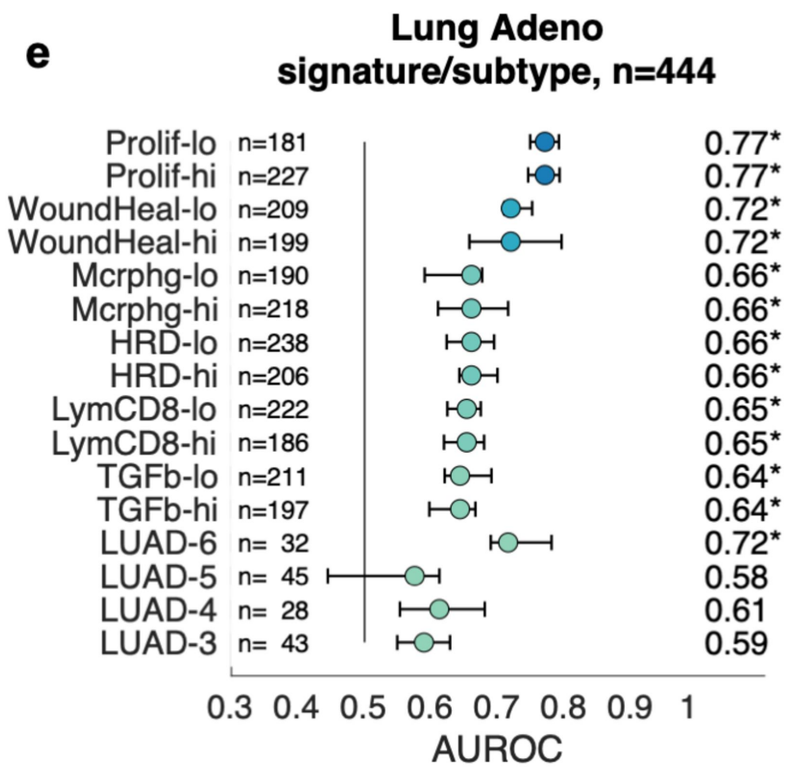


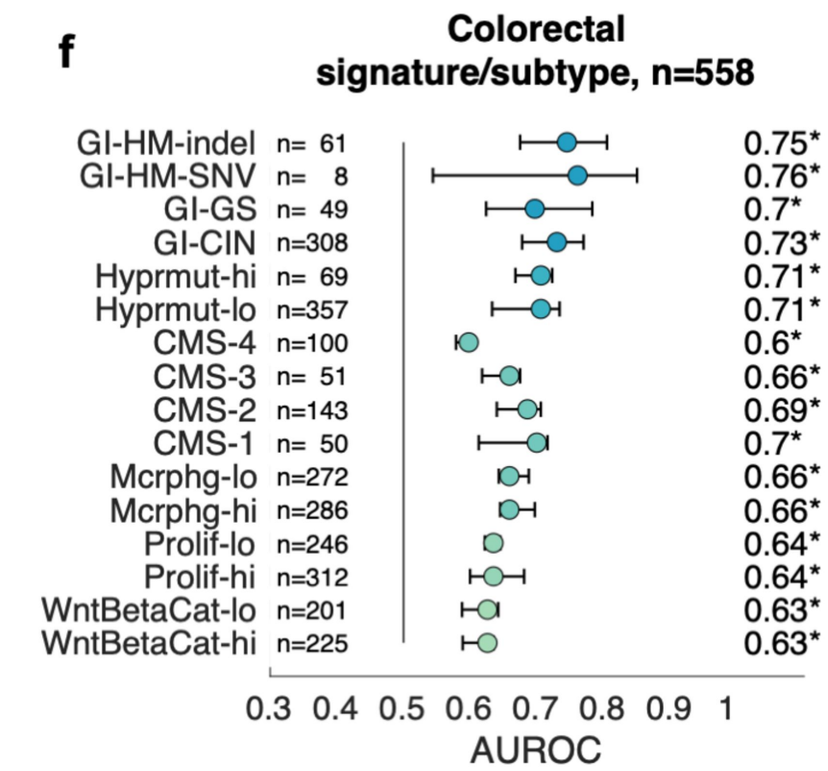


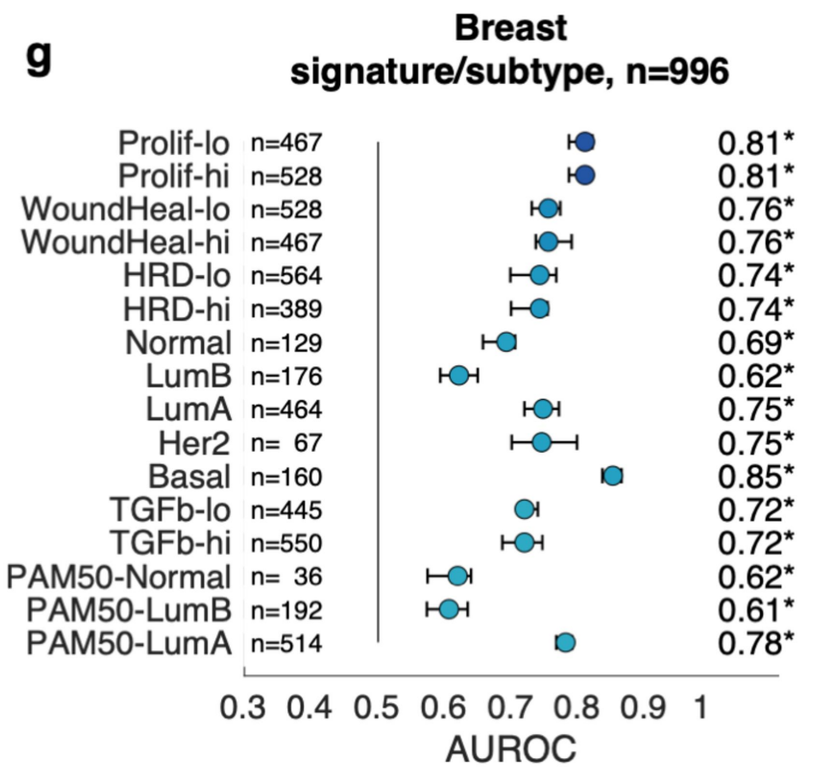


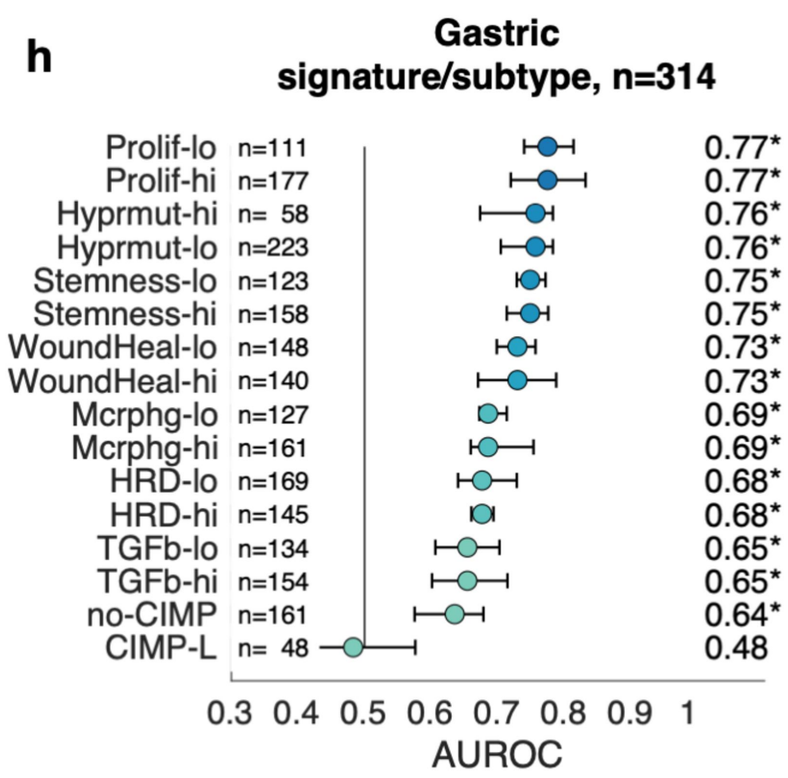


molecular subtypes and gene expression signatures

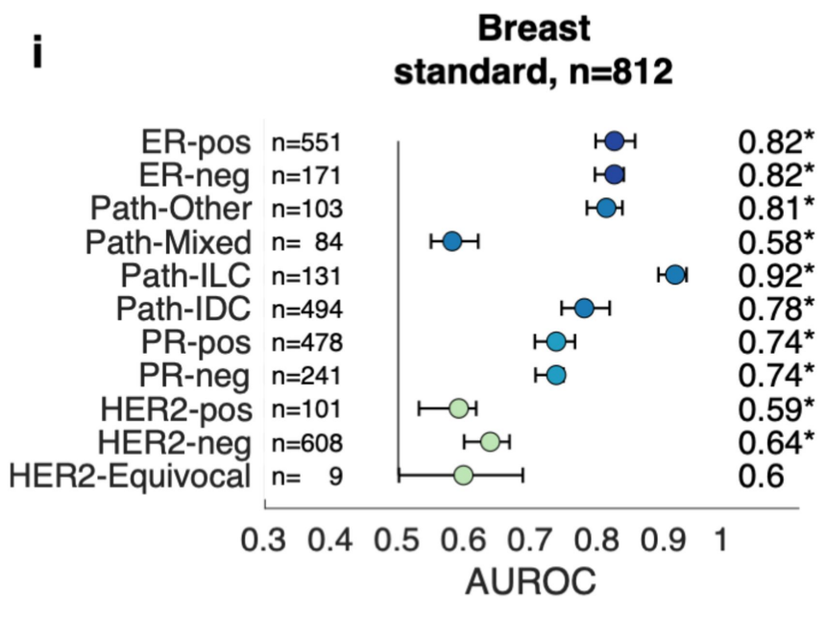


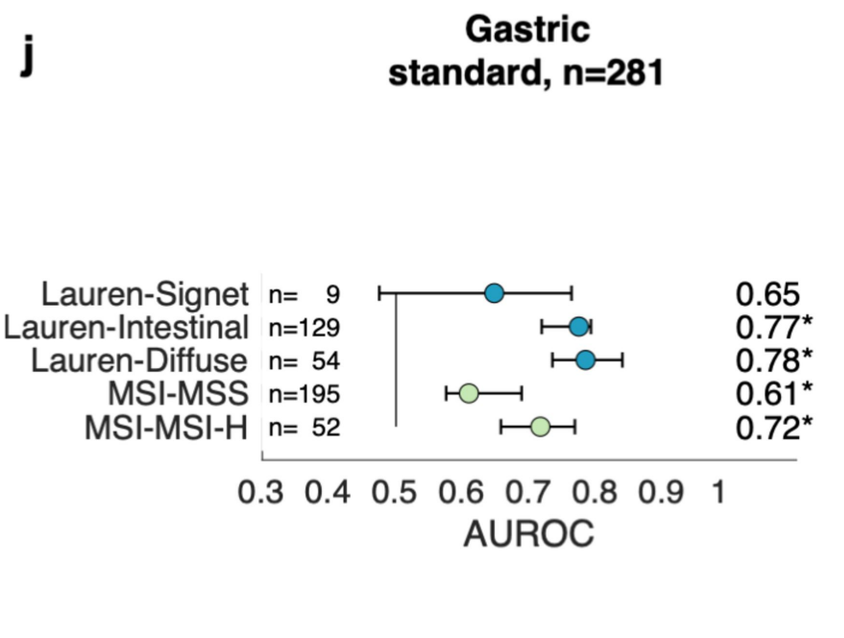


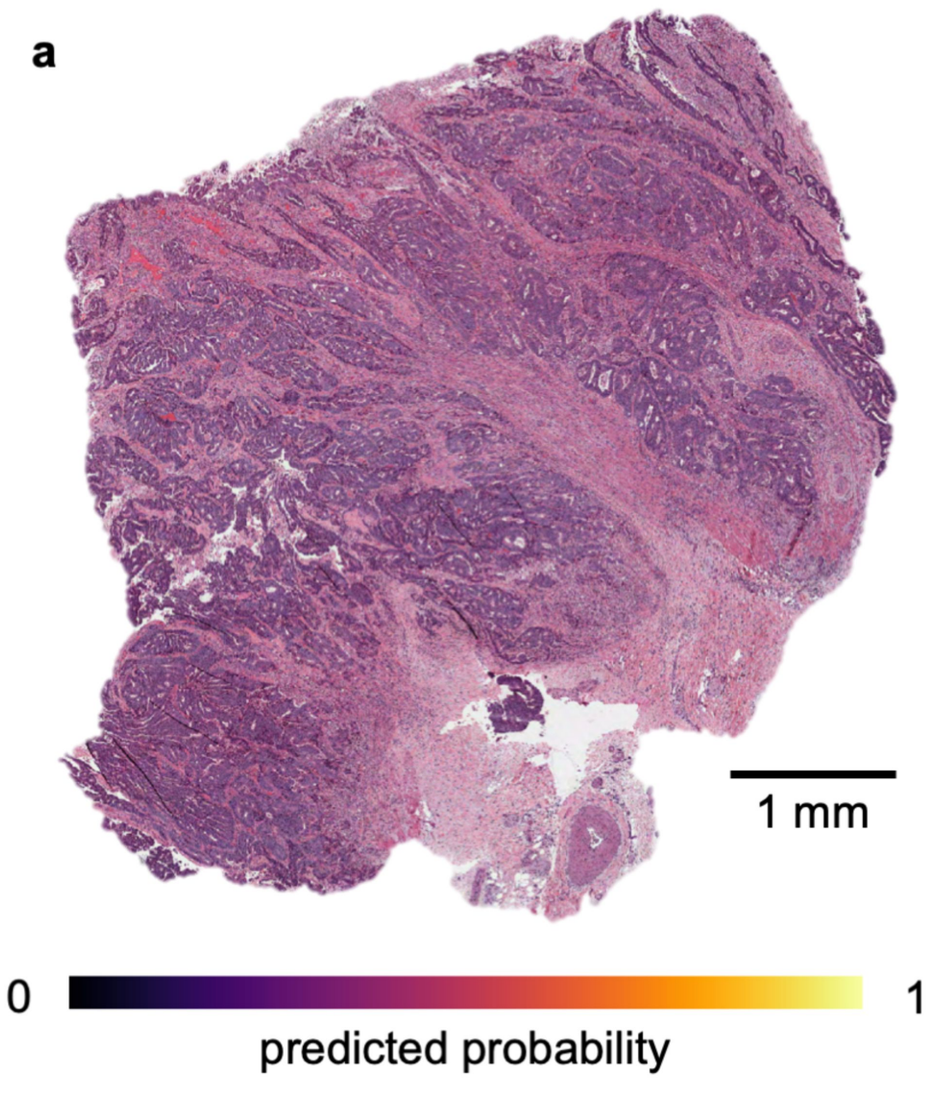


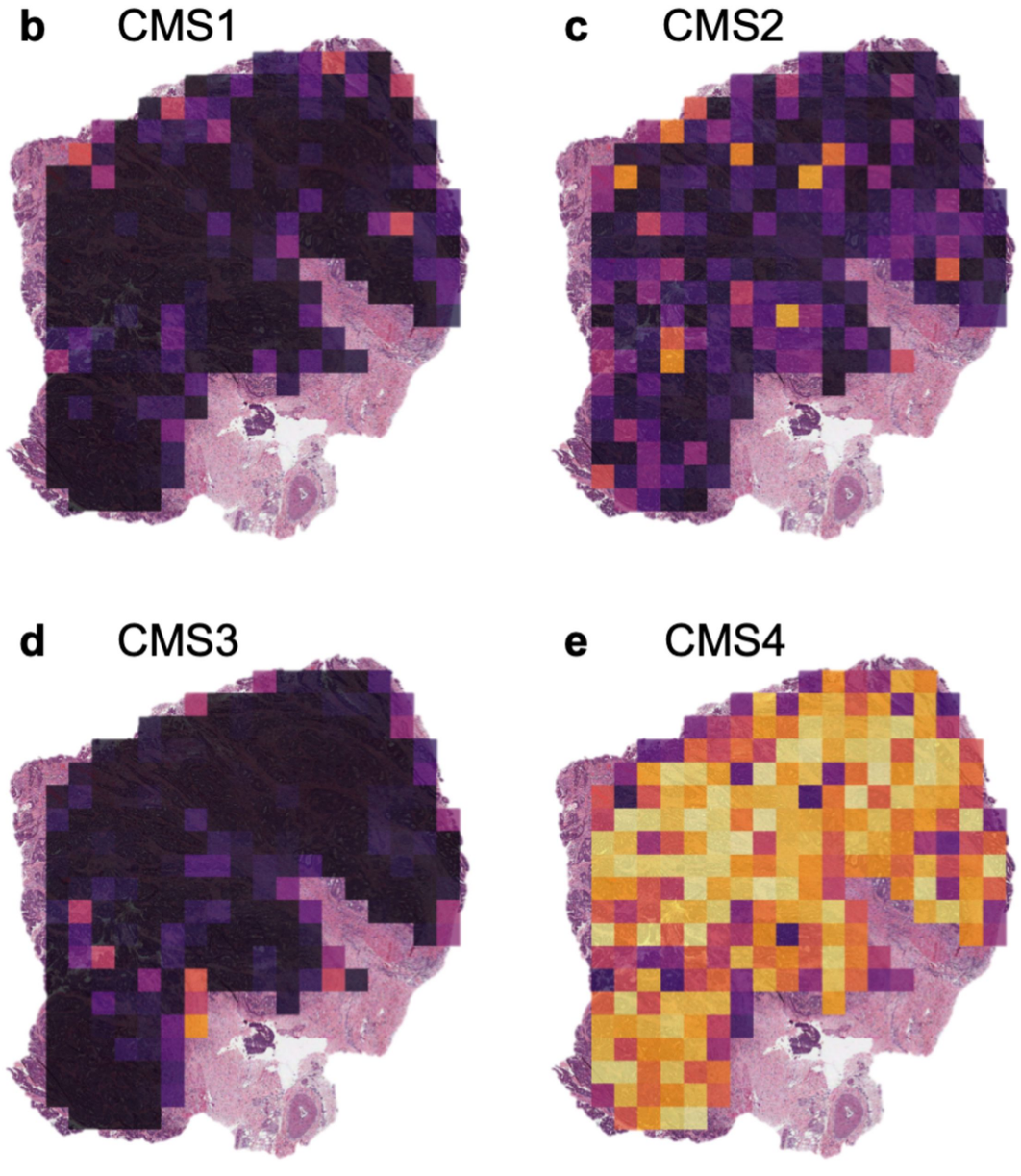


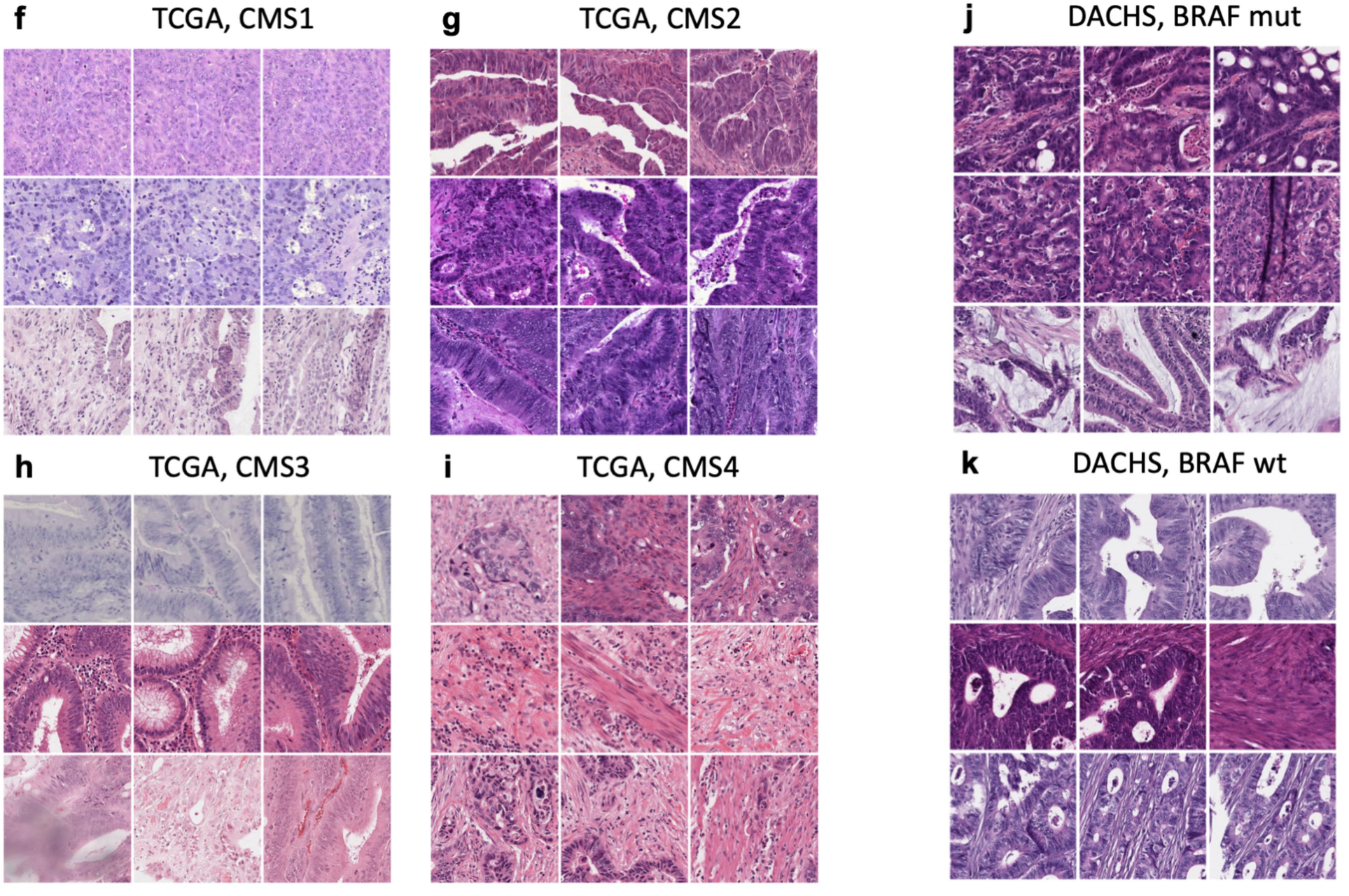
standard biomarkers






a
1 mm
0
1
predicted probability


b CMS1
c CMS2
d CMS3
e CMS4


f TCGA, CMS1
g TCGA, CMS2
j DACHS, BRAF mut
h TCGA, CMS3
i TCGA, CMS4
k DACHS, BRAF wt