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

Document status and date:
Published: 01/08/2020

DOI:
10.1038/s43018-020-0087-6

Document Version:
Publisher's PDF, also known as Version of record

Document license:
Taverne

Please check the document version of this publication:

• A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
• The final author version and the galley proof are versions of the publication after peer review.
• The final published version features the final layout of the paper including the volume, issue and page numbers.

Link to publication

Download date: 11 Oct. 2023

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 [12,19] ✉ and Tom Luedde [17,18,19] ✉

Molecular alterations in cancer can cause phenotypic changes in tumor cells and their microenvironment. Routine histopathology tissue slides, which are ubiquitously available, 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. We developed, optimized, validated and publicly released a one-stop-shop workflow and applied it to tissue slides of more than 5,000 patients across multiple solid tumors. 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 are spatially resolved. Our method can be implemented on mobile hardware, potentially enabling point-of-care diagnostics for personalized cancer treatment. More generally, this approach could elucidate and quantify genotype–phenotype links in cancer.

Precision treatment of cancer relies on the 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* and *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 (CMSs)[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 difficult to implement at scale, tissue sections stained with hematoxylin and eosin 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 (WSIs). 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 MSI[7]—a clinically approved biomarker for cancer immunotherapy in colorectal cancer[8]. 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 shown previously[9]. Other previous studies have identified

genotype–phenotype links for selected genetic features in lung cancer[10,11], prostate cancer[12], head and neck cancer[13] and liver cancer[14], 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 that 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 narrower indication for molecular testing, increasing the pre-test probability of a given molecular feature. Independent 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 deep learning pipeline to determine molecular features directly from histology images.

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 one-stop-shop workflow to train and evaluate deep learning networks (Fig. 1a,b). To select an efficient network model and to optimize the deep learning hyperparameters, prediction of MSI in colorectal cancer was used as a clinically relevant benchmark task[9]. In this benchmark, we sampled a large hyperparameter space with different commonly used deep learning models[9,10,13,15], which were modified specifically for this application. Unexpectedly, ShuffleNet[16]—a lightweight neural network architecture—performed similarly to more complex networks including the DenseNet[17], Inception[18] and ResNet[19] networks, which have been used in many other studies[20] (Fig. 1c). ShuffleNet showed high accuracy at a low training time (raw data in Supplementary Table 1; $n = 426$ patients in The Cancer Genome Atlas (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[21]. We externally validated the best ShuffleNet classifier by training on $n = 426$ patients in the TCGA colorectal cancer cohort[9] and validating on $n = 379$ patients with available MSI status in the DACHS ('Darmkrebs: Chancen der Verhütung durch Screening', which translates as 'colorectal cancer: chances for prevention through screening') cohort[9], reaching an area under the receiver operating curve (AUROC) of 0.89 (confidence interval, 0.88–0.92). This represents an improvement over the previous best performance of 0.84 in that dataset[9], and supports the notion that ShuffleNet is an efficient and powerful neural network model that 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 threefold 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 (Supplementary Table 1) included all point mutations targetable by Food and Drug Administration-approved drugs (level 1 evidence on www.oncokb.org; the 20 most common mutations are 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 such genes could be inferred from histology images alone, with statistical significance after correction for multiple testing (Fig. 2a–n and Extended Data Fig. 1). 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 ($P < 0.05$ after FDR correction) in all four of these cancer types, as well as mutations of *BRAF* in colorectal cancer (colon adenocarcinoma (COAD) and rectum adenocarcinoma (READ) TCGA cohorts[22]; $n = 555$; Fig. 2b), *MTOR*—a candidate for targeted treatment[23]—in gastric cancer (Fig. 2d), and *FBXW7* mutation in lung adenocarcinoma (LUAD TCGA cohort[24]; $n = 457$; Fig. 2a) and gastric cancer (stomach adenocarcinoma (STAD) TCGA cohort[25]; $n = 321$; Fig. 2d). Mutations of *PIK3CA* (which are directly targetable by a small molecule inhibitor[26]) were significantly detectable ($P = 7 \times 10^{-9}$) in breast cancer (BRCA TCGA cohort[27]; $n = 995$; 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[28]) were significantly detectable ($P = 0.0008$) (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 AUROC for the top eight mutations ranged from 0.60–0.78 in lung adenocarcinoma (Extended Data Fig. 2a–h), from 0.65–0.76 in colorectal cancer (Extended Data Fig. 2i–p), from 0.62–0.78 in breast cancer (Extended Data Fig. 2q–x) and from 0.66–0.78 in gastric cancer (Extended Data Fig. 3a–h). Beyond these four tumor types, a range of notable mutations could be detected in other tumor types. While in melanoma (skin cutaneous melanoma (SKCM) TCGA cohort[29]) primary tumors few mutations were detectable (Extended Data Fig. 3i–p), in melanoma metastases mutations in *FBXW7* ($P = 0.0129$) and *PIK3CA* ($P = 0.0052$) were significantly detectable (Fig. 2e and Extended Data Fig. 3q–x). In prostate cancer (prostate adenocarcinoma (PRAD) TCGA cohort[30]; $n = 397$ patients; Fig. 2f and Extended Data Fig. 4a–h), our method detected *TP53* and *FOXA1* mutations from histology, among others. In pancreatic adenocarcinoma (PAAD TCGA cohort[31]; $n = 171$ patients; Fig. 2g and Extended Data Fig. 4i–p), identifying *KRAS* wild-type patients is of high clinical relevance because these patients are potential candidates for targeted treatment and our method significantly identified mutations in the *KRAS* gene in pancreatic cancer ($P = 0.0016$). 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 (lung squamous cell carcinoma (LUSC) TCGA cohort[32]; $n = 413$; Fig. 2h and Extended Data Fig. 4q–x) In hepatocellular carcinoma (liver hepatocellular carcinoma (LIHC) TCGA cohort[33]; $n = 358$ patients; Fig. 2i), the product of the β-catenin gene (*CTNNB1*) is a key driver gene with broad prognostic and predictive implications[34], and its mutational status was highly significantly detected ($P = 2 \times 10^{-7}$) from histology (Extended Data Fig. 5a–h). In papillary[35] (Fig. 2j and Extended Data Fig. 5i–p) and clear cell[36] renal cell carcinoma (Fig. 2k and Extended Data Fig. 5q–x), alterations in multiple genes including *KRAS* and *PBRM* were highly detectable, while in chromophobe[37] renal cell carcinoma (Fig. 2l and Extended Data Fig. 6a–h) no genetic variants were significantly detectable, possibly due to a low patient number in this cohort. In head and neck squamous cell carcinoma (HNSC TCGA cohort[38]; $n = 435$ patients), the *CASP8* gene, which is linked to resistance to cell death[39], was significantly detected ($P = 3 \times 10^{-6}$) (Fig. 2m and
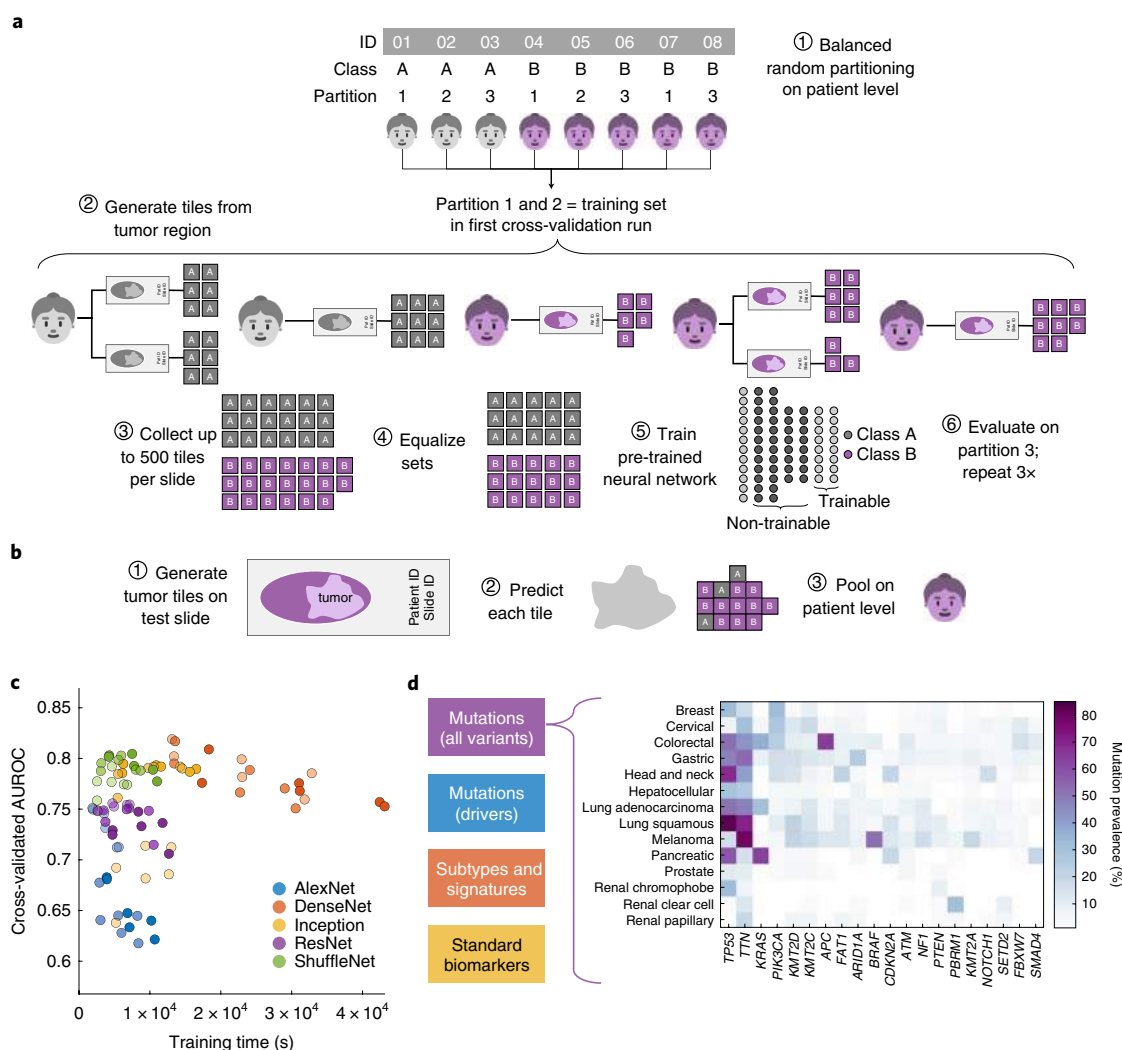
Fig. 1 | Deep learning workflow for the prediction of molecular features from histology images. a, Training of the deep learning system comprised six steps: (1) patient cohorts were randomly split into three partitions for cross-validation of deep classifiers; (2) the tumor region on each WSI was tessellated into tiles; (3) up to 500 randomly chosen tiles were collected; (4) tiles from patients in the training partitions were collected and classes were equalized by random undersampling; (5) all training tiles were used to train a deep neural network (pre-trained on a non-medical task); and (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: (1) tiles were generated from the tumor region on WSIs; (2) a prediction was made for each tile; and (3) tile-level class predictions were pooled on the patient level. **c**, Hyperparameters of the deep learning system were optimized in a benchmark task (the prediction of MSI in colorectal cancer). The opacity of each point corresponds to the number of trainable layers. ShuffleNet—a lightweight neural network architecture—was selected as a highly efficient network model. **d**, Pan-cancer application. This workflow was subsequently applied for the prediction of four types of molecular features across 14 cancer types. In particular, this included genetic mutations. The distribution of the 20 most common mutations among all analyzed mutations is shown for each tumor type. Icons are from Twitter Twemoji (CC-BY 4.0 license).

Extended Data Fig. 6i–p). In cervical cancer (cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC) TCGA cohort[40]; $n = 261$ patients), mutations in *TCERG1*, *STK11* and *AMER1*, among others, were detectable with high AUROC values (Fig. 2n and Extended Data Fig. 6x–q).

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; thus, fewer genes met the threshold of at least four positive cases in a given tumor type. In contrast, 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 values in the 'drivers only' experiment compared with the 'all variants' approach (Fig. 3a–n). For mutations in 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 targetable and potentially targetable point mutations in a wide range of genes directly from histology across multiple prevalent tumor types.

Inference of molecular subtypes and gene expression signatures. In the next step, we asked whether established molecular subtypes and gene expression signatures of cancer and immune cells could be detected by deep learning. Compared with single-gene mutations,
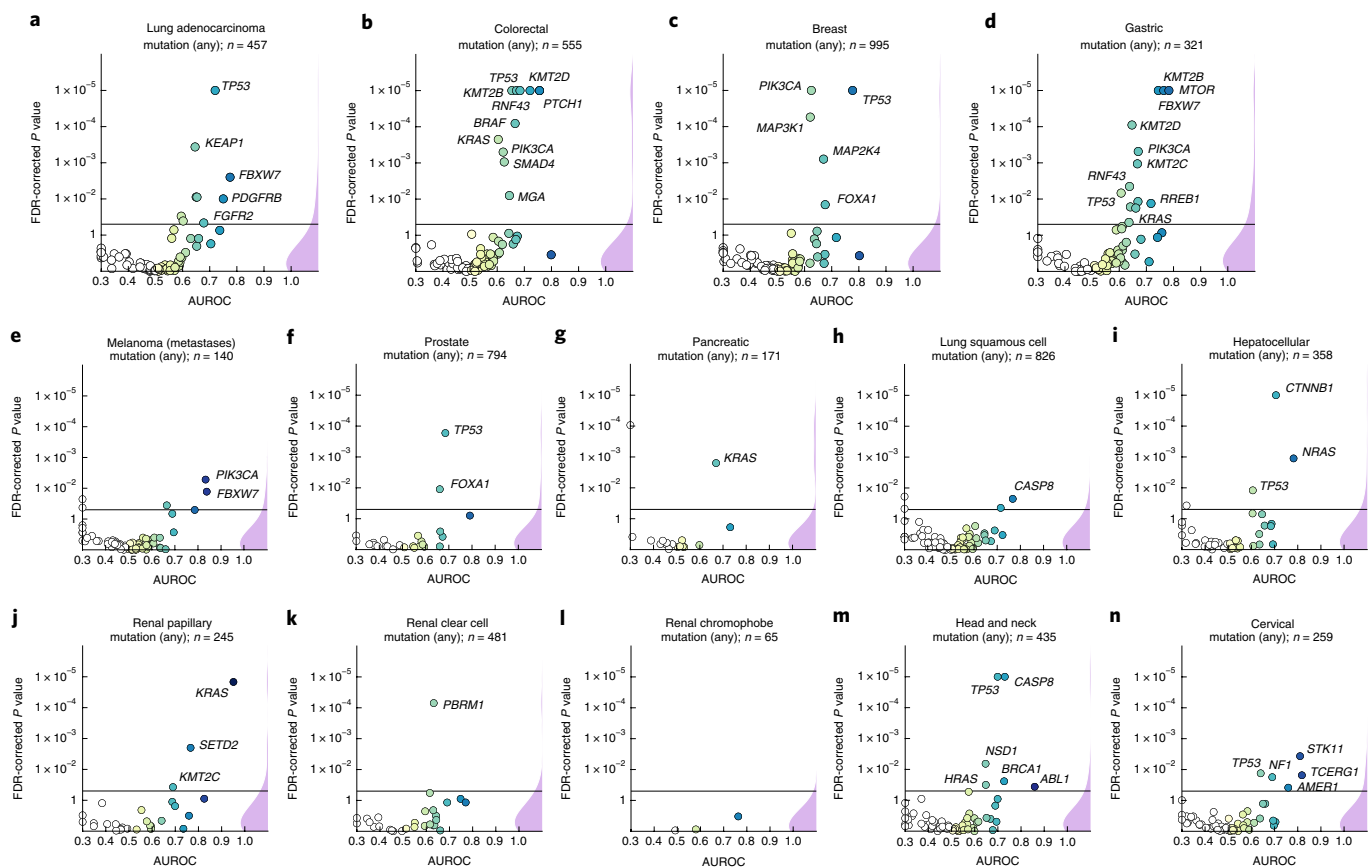
Fig. 2 | Inference of genetic mutations from histological images. A deep learning system was trained to predict the mutational status (mutated or wild type) of relevant genes in 14 cancer types and was evaluated by cross-validation. All mutations, including variants of unknown significance, were included in the mutated class. For each gene, the patient-level test set performance is shown as the AUROC, with two-sided t-test P value for prediction scores corrected for multiple testing (FDR). The significance level of 0.05 is marked with a solid horizontal line. P values smaller than 10^{-5} were set to 10^{-5}. On the right-hand side of each panel, a kernel density estimate shows the distribution of all plotted data points. n denotes the number of patients with available genetic information and matched histology images in each tumor type. **a–d,** In lung adenocarcinoma (**a**), colorectal cancer (**b**), breast cancer (**c**) and gastric cancer (**d**), a number of relevant genes were significantly predictable ($P < 0.05$) from histology alone, including key oncogenic drivers such as *TP53*, *BRAF* and *MTOR*. **e–n,** In all other tested tumor types (that is, melanoma (**e**), prostate (f), pancreatic (**g**), lung squamous cell (h), hepatocellular (**i**), renal papillary (j), renal clear cell (**k**), renal chromophobe (**l**), head and neck (m) and cervical (**n**)), 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.

these changes occur at a higher functional level and we hypothesized that their morphological impact could be larger than that of single mutations. To address this hypothesis, we chose features with known biological and potential clinical significance. A major group of such features are immune-related gene expression signatures[41] of CD8-positive lymphocytes and macrophages, cell proliferation, interferon-γ signaling and transforming growth factor-β signaling (a full list is available in Supplementary Table 1). These biological processes are involved in response to cancer treatment, including immunotherapy. Detecting their morphological correlates in histology images could facilitate the development of more nuanced treatment strategies. Indeed, across all investigated tumor types, we saw that these high-level biological features had much greater predictability than genetic variants or driver mutations (Fig. 4a–d and Extended Data Fig. 1). Again, AUROC values for significantly ($P < 0.05$ after false discovery rate (FDR) correction) predictable features were highest in lung adenocarcinoma (Fig. 4e), colorectal cancer (Fig. 4f), breast cancer (Fig. 4g) and gastric cancer (Fig. 4h). In lung adenocarcinoma, signatures of proliferation, macrophage infiltration and T-lymphocyte infiltration were significantly detectable from images with high AUROCs (Fig. 4e). Similarly, significant AUROCs for these biomarkers were achieved in colorectal cancer (Fig. 4f), breast

cancer (Fig. 4g) and gastric cancer (Fig. 4h). In gastric cancer, we additionally found that a signature of stem cell properties (stemness) was highly detectable directly from histology images (Fig. 4h). Recent studies have clustered tumors into comprehensive molecular subtypes[41]. We found that our method could detect TCGA molecular subtypes[41] with up to AUROC 0.74 in lung adenocarcinoma (Fig. 4e), pan-gastrointestinal subtypes[42] with up to AUROC 0.76 in colorectal cancer (Fig. 4f), and PAM50 subtypes with up to AUROC 0.78 in breast cancer (Fig. 4g), among other molecular subtypes. These findings could open up new options for clinical trials of cancer. While accumulating evidence shows that such molecular clusters of tumors reflect biologically distinct groups and are correlated with clinical outcome, deep molecular classification of these tumors is usually not available in clinical routine or clinical trials. Detecting these subtypes merely 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 deep learning pipeline, we investigated classification accuracy
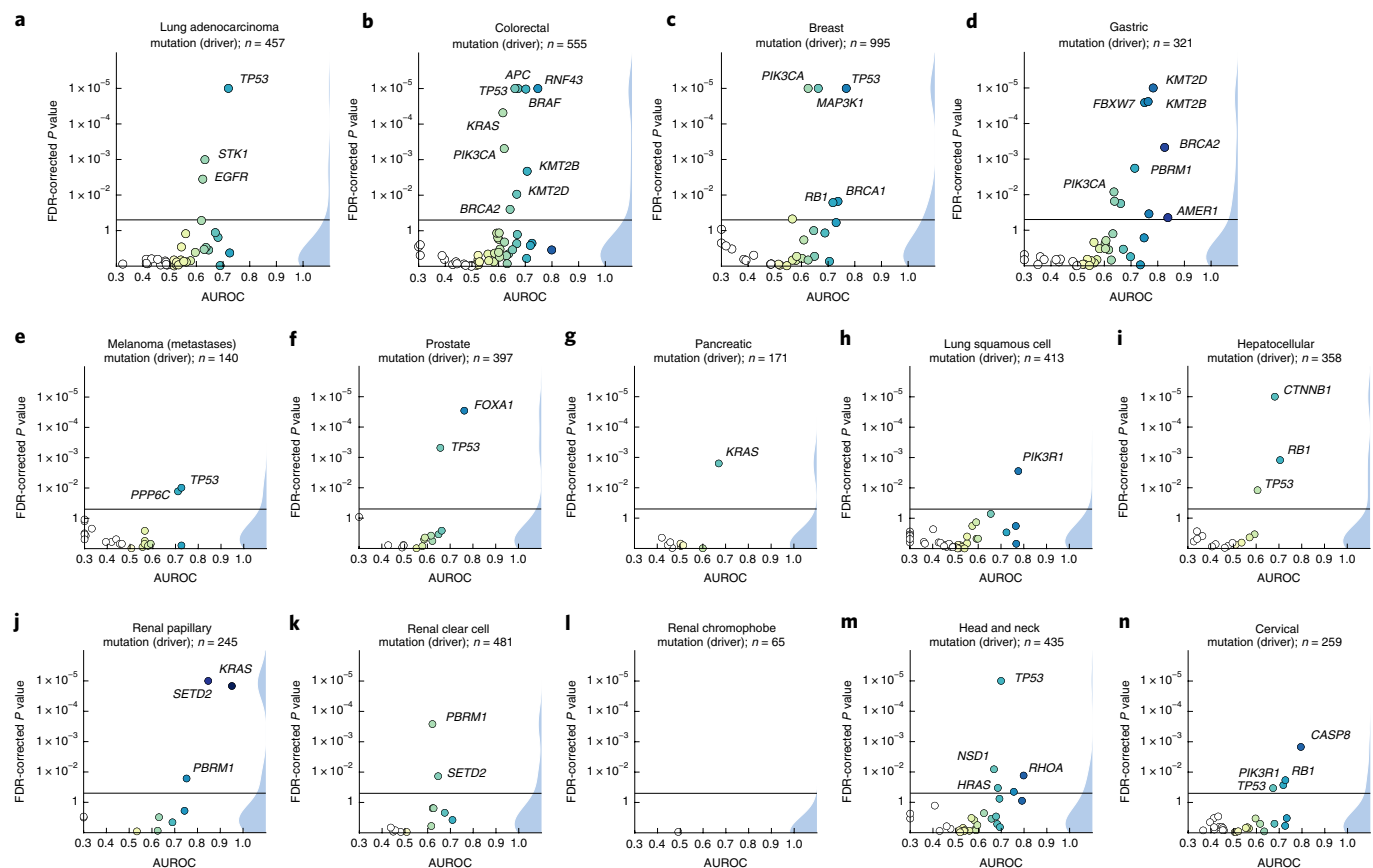
Fig. 3 | Inference of putative oncogenic drivers from histological images. A deep learning system was trained to predict oncogenic driver genes from histology. Only putative and confirmed drivers were included, and variants of unknown significance were pooled with the wild-type class. On the right-hand side of each panel, a kernel density estimate shows the distribution of all plotted data points. *n* denotes the number of patients with available genetic information and matched histology images in each tumor type. The layout of this figure corresponds to that of Fig. 2. **a–n**, This process (in lung adenocarcinoma (**a**), colorectal cancer (**b**), breast cancer (**c**), gastric cancer (**d**), melanoma (**e**), prostate cancer (f), pancreatic cancer (**g**), lung squamous cell carcinoma (h), hepatocellular carcinoma (i), renal papillary carcinoma (j), renal clear cell carcinoma (**k**), renal chromophobe carcinoma (**l**), head and neck cancer (m) and cervical cancer (**n**)) uncovered the significant predictability of multiple oncogenic drivers, including *EGFR*, *BRAF* and *TP53*.

for standard histopathological biomarkers. We found that deep learning could predict most of these biomarkers for breast cancer (Fig. 4c,i), gastric cancer (Fig. 4d,j) and other tumor types. In particular, the status of hormone receptors was predictable from routine histology in breast cancer, with an AUROC of 0.82 for estrogen receptors and 0.74 for progesterone receptors (Fig. 4i). Together, these results show 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 (COAD and READ TCGA cohorts). First, we investigated the role of color normalization of tiles. In a head-to-head comparison with the baseline approach, we found a tendency of Macenko's[43] color normalization to improve classifier performance for mutation prediction but not for the prediction of subtypes or gene expression signatures (Extended Data Fig. 7a–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 (Extended Data Fig. 8a–c). Third, we analyzed prediction

performance on frozen slides compared with 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 the prediction power for driver genes was on par, but the prediction power for genetic variants and high-level subtypes and signatures was better in frozen slides than in diagnostic slides (Extended Data Fig. 9a–c). 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 hematoxylin and eosin slides of $n = 408$ patients with colorectal cancer from the DACHS study for whom *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 hematoxylin and eosin images alone. For BRAF mutants, the AUROC was 0.77 (0.64–0.82; $P < 10^{-5}$), whereas for CIMP-high, the 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.
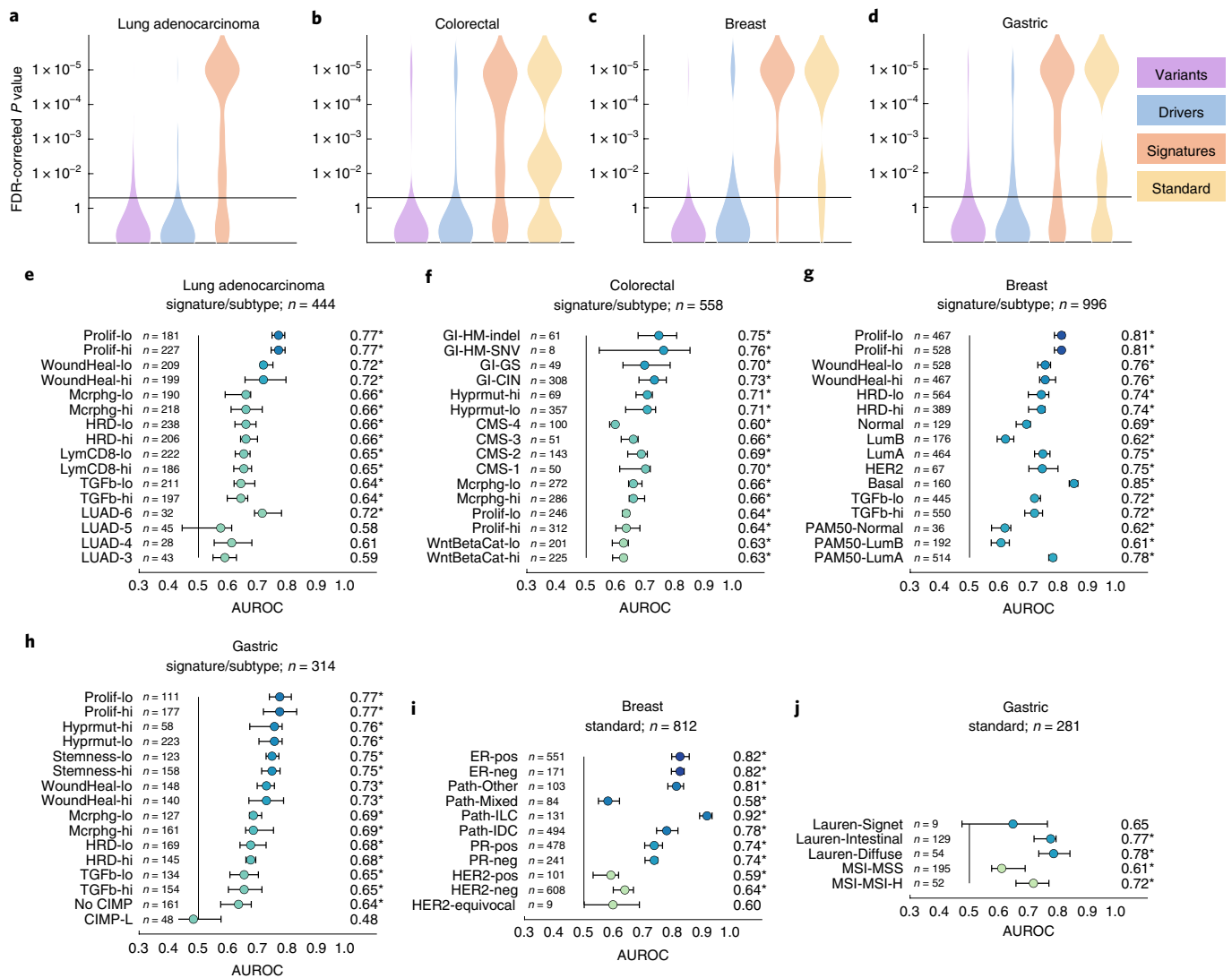
Fig. 4 | Inference of molecular subtypes, gene expression signatures and standard biomarkers directly from histology. In addition to the prediction of single-gene mutations, the capability of deep learning to infer high-level molecular features was systematically assessed. **a–d**, In lung (**a**), colorectal (**b**), breast (**c**) and gastric cancer (**d**), 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 two-sided t-test FDR-corrected P values as visualized by a kernel density estimate. Individual data points are shown in Extended Data Figs. 2–6. **e–h**, Gene expression signatures (in lung (**e**), colorectal (**f**), breast (**g**) and gastric cancer (**h**)) 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, CMSs, PAM50 subtypes and other key molecular features were highly predictable across multiple tumor types. Mean patient-level AUROC values are shown with bootstrapped confidence intervals (error bars). Asterisks denote a two-sided t-test FDR-corrected P value of <0.05. n refers to the number of patients. **i,j**, Standard-of-care biomarkers, including estrogen receptor (ER) and progesterone receptor (PR) status, in breast cancer (**i**) and gastric cancer (**j**). In gastric cancer (**j**), pathologic subtype and MSI were highly predictable from routine histology alone by deep learning. Solid vertical line (**e–j**) indicates AUROC = 0.5. hi, high; lo, low; neg, negative; pos, positive; Basal, breast cancer expression subtype basal; CIMP-L, hypermethylation category low CpG island methylator phenotype; CIN, chromosomally instable subtype; CMS, consensus molecular subtype; ER, estrogen receptor; GI, gastrointestinal; GS, genome stable subtype; HER2, human epidermal growth factor receptor 2; HM-indel, hypermutated indel subtype; HM-SNV, hypermutated single-nucleotide variant subtype; HRD, homologous recombination repair deficiency; Hyprmut, hypermutated; IDC, invasive ductal carcinoma; ILC, invasive lobular carcinoma; Lauren, Lauren histological subtype; LUAD, lung adenocarcinoma subtype; LumA, luminal A; LumB, luminal B; LymCD8, CD8⁺ lymphocyte expression signature; Mcrphg, macrophage signature; MSI-MSS, microsatellite instability status: microsatellite stable; MSI-MSI-H, microsatellite instability status: microsatellite instability high; neg, negative; Path, pathological subtype; PR, progesterone receptor; Prolif, proliferation sigature; Signet, signet ring cell differentiation; TGFb, transforming growth factor-β response signature; WntBetaCat, Wnt–β-catenin signaling signature; WoundHeal, wound healing signature.

Discussion

Image-based genetic testing as a clinical and research tool. Our results show 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
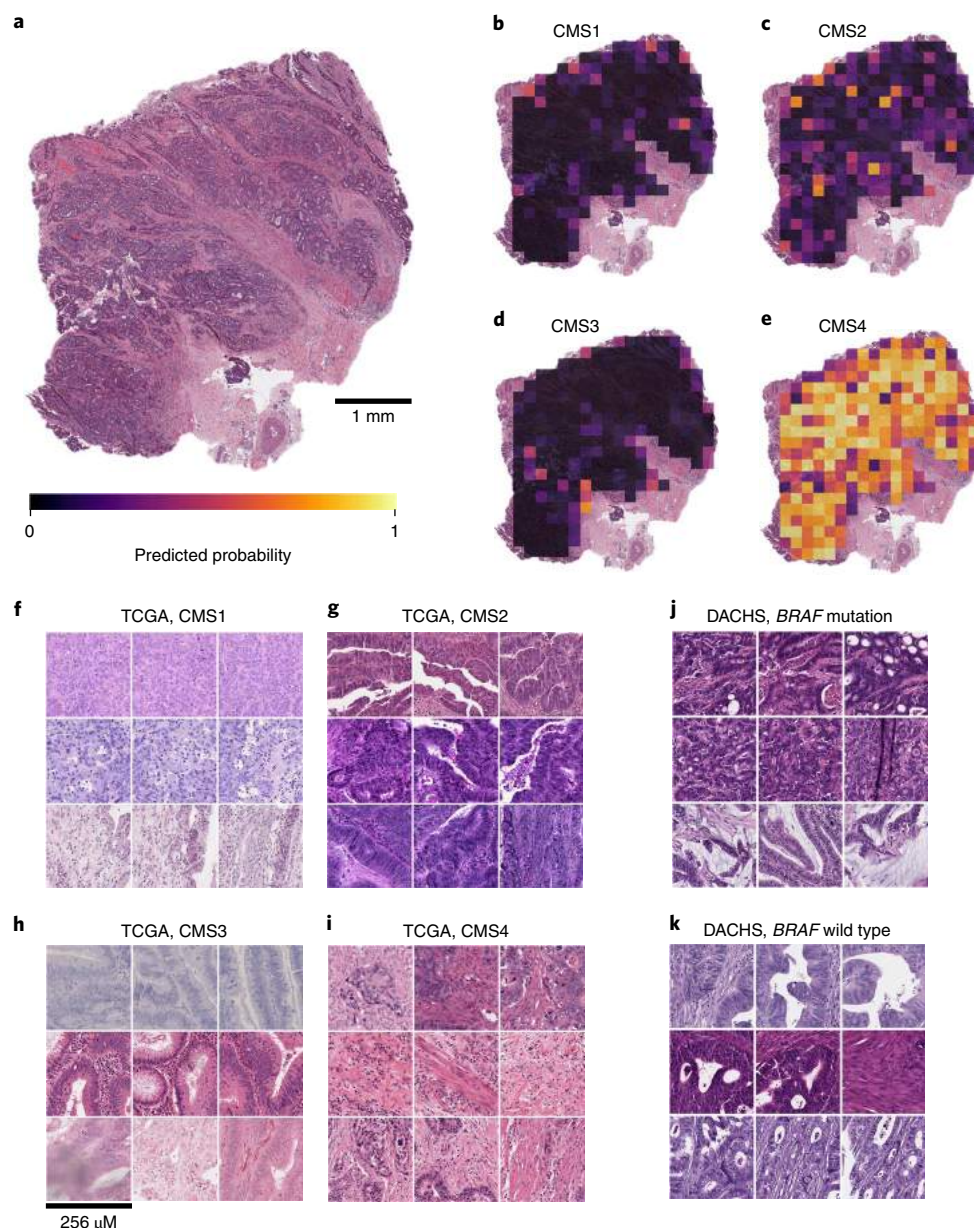
Fig. 5 | Explainability of deep learning-based analysis of histological images. Deep learning-based predictions were visualized through genotype maps and comparison of highly ranked image tiles. **a–e**, Prediction maps for CMS1 (**b**), CMS2 (**c**), CMS3 (**d**) and CMS4 (**e**) from a colorectal cancer image (**a**) show spatially resolved prediction scores, unveiling the intratumor heterogeneity of predicted genotype. As a generic tool, this visualization approach allows us to identify spatial regions associated with a molecular feature. In this patient, the correct prediction of CMS4 showed that deep learning robustly predicts CMSs from histology alone while highlighting potential intratumor heterogeneity. **f–i**, For each of the CMS classes (CMS1 (**f**), CMS2 (**g**), CMS3 (**h**) and CMS4 (**i**)), 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, whereas predicted CMS4 tiles contain abundant stroma, consistent with previous studies. **j,k**, Highly scored tiles in the external test cohort DACHS for the prediction of the *BRAF* mutation versus the wild type.

candidate genetic variants, driver genes, gene expression signatures and standard-of-care features that can be 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 that are significantly linked to a detectable phenotype in histological images, including those in key oncogenic pathways including the products of *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 CMSs in colorectal cancer. By linking these molecularly defined groups to specific histological image features, our method constitutes a 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
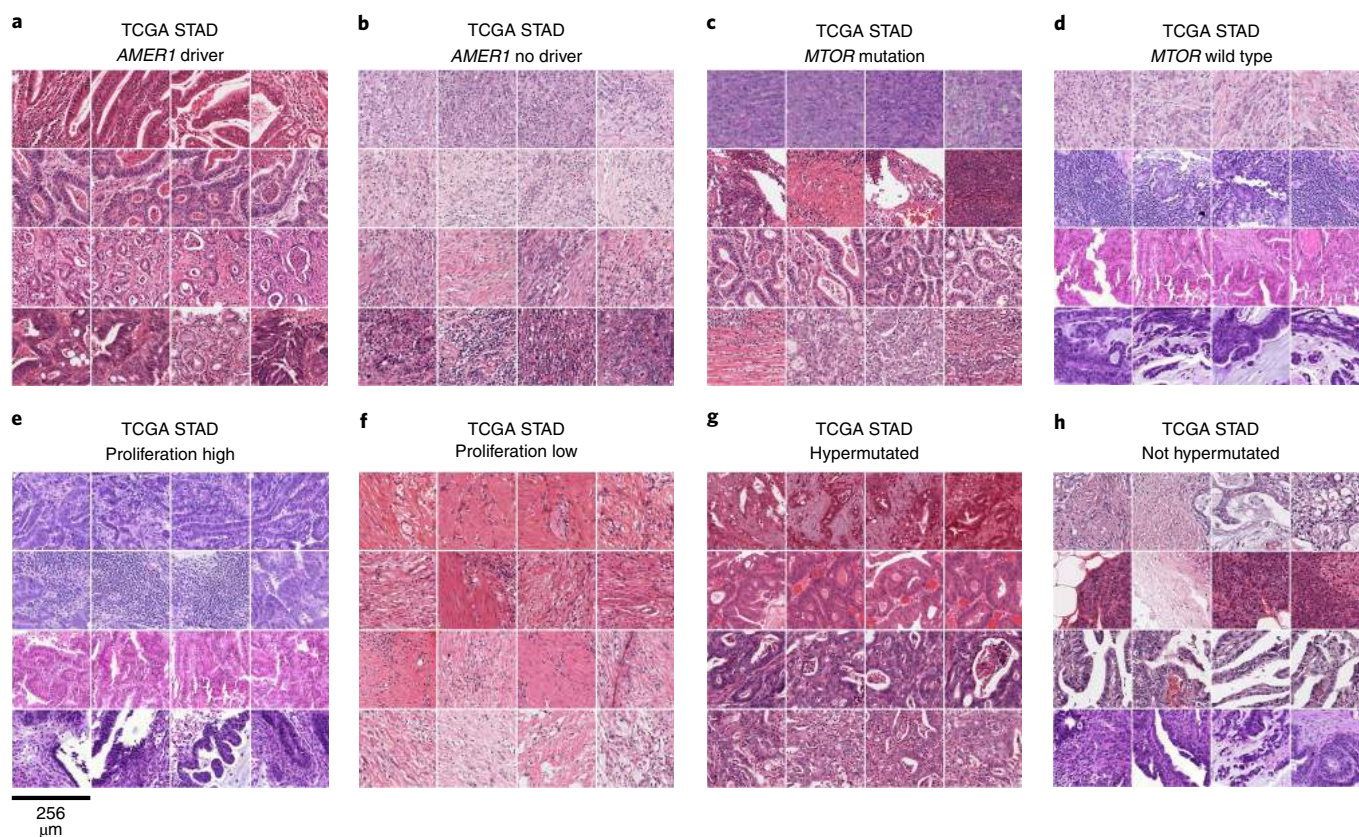
Fig. 6 | Highest-scoring image tiles for molecular features in gastric cancer. a,b, Highest-scoring tiles in the highest-scoring patients corresponding to *AMER1* mutational status (driver (**a**) or no driver (**b**)) in the STAD TCGA dataset. **c,d**, Tiles corresponding to *MTOR* mutational status (mutant (**c**) and wild type (**d**)). **e,f**, Tiles corresponding to high (**e**) or low values of a proliferation signature (**f**). **g,h**, Tiles corresponding to hypermutation status (hypermutated (**g**) versus not hypermutated (**h**)).

histology-based deep learning methods to be implemented in clinical workflows. An example for clinical implementation would be their 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, a limitation of our method is the low AUROC values for some molecular features (Figs. 2 and 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 the performance of deep learning systems in histopathology scales with the number of patients in the training cohort[15]. In addition, the performance of deep learning systems could potentially be improved by technical modifications. Our systematic evaluation of alternative technical approaches provides guidance for this on multiple levels. First, regarding the choice of neural network models, our results show 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[10]. In contrast, our baseline

workflow was based on formalin-fixed, paraffin-embedded (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; therefore, establishing methods on FFPE material is paramount for large-scale clinical validation. Our head-to-head comparison showed 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 the feasibility of even simpler data preprocessing pipelines. Ultimately, fully automatic workflows can be expected to be superior to manual workflows in terms of scalability and reproducibility. We have publicly released all of the source codes of our method, enabling further optimization and validation on a larger scale (see 'Code availability').

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 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 tumors, 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 tumors 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 patients with a *BRAF* mutation in the validation cohort (Fig. 5j,k) showed poorly differentiated areas and mucinous areas as recurring features in *BRAF* mutant image tiles, which is consistent with previous studies[44]. Visualizing highly predicted tiles in gastric cancer (Fig. 6a–h) highlighted highly cellular areas as correlates of a proliferation gene expression signature, but at the same time identified patterns for mutations (for example, in *AMER1* and *MTOR*) that 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 and Extended Data Fig. 1). 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 (for example, lung squamous cell cancer and pancreatic cancer) also display few detectable mutations. Further studies are warranted to investigate this.

Conclusion

Together, our results show that molecular changes in solid tumors can be inferred from routine histology alone with deep learning. This could be a useful tool for objectively elucidating genotype–phenotype relationships in cancer, and ultimately could be used as a low-cost biomarker in clinical trials and routine clinical workflows.

Methods

Patient cohorts and ethics. All experiments were conducted in accordance with the Declaration of Helsinki and the International Ethical Guidelines for Biomedical Research Involving Human Subjects. Anonymized scanned WSIs were retrieved from the TCGA project through the Genomic Data Commons Portal (https://portal.gdc.cancer.gov/). We applied our method to 14 of the most common solid tumor types: breast (BRCA)[27], cervical (CESC)[40], colorectal (COAD and READ)[22], gastric (STAD)[25], head and neck (HNSC)[38], hepatocellular (LIHC)[33], lung adenocarcinoma (LUAD)[24], lung squamous (LUSC)[32], melanoma (SKCM)[29], pancreatic (PAAD)[31], prostate (PRAD)[30], renal cell chromophobe (KICH)[37], renal cell clear cell (KIRC)[36] and renal cell papillary cancer (KIRP)[35]. Melanoma 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[45,46], which were retrieved from the tissue bank of the National Center for Tumor Diseases (NCT; Heidelberg, Germany), as described before[9]. Ethics oversight of the TCGA study is described at https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga/history/policies and ethical approval of the DACHS study was given by the ethics committee of the Medical Faculty of the University of Heidelberg. Informed consent was obtained by all participants in the TCGA and DACHS studies.

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[47,48] at http://cbioportal.org) and clinically targetable genes (level 1 genes from OncoKB at http://www.oncokb.org; Pan-Cancer Atlas Project[49]). 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[41,42,50]. 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 Supplementary 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 WSIs of diagnostic tissue slides (FFPE tissue) stained with hematoxylin and eosin were acquired in SVS format. All images were downsampled to 20× magnification, corresponding to 0.5 μm px⁻¹. Each WSI was manually reviewed and the tumor area 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 mm² contiguous tumor tissue were used for downstream analysis. In total, 6% of WSIs, corresponding to 5% of patients, were excluded due to technical artefacts or a lack of tumor (Supplementary Table 2). Tumor tissue on all other slides was tessellated into square tiles of 512 px × 512 px edge length, corresponding to 256 μm × 256 μm at a resolution of 0.5 μm px⁻¹. Tiles with more than 50% background were discarded. Background pixels were defined by a brightness of >0.86 (220/255). For the benchmark task (identification of an optimum neural network model), these images were resized to 224 px × 224 px (at 1.14 μm px⁻¹) to be consistent with a previous study[9]. 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://doi.org/10.5281/zenodo.3694994. All methods for WSI processing, including tessellation of images and visualization of spatial activation maps, were implemented in QuPath[51] version 0.1.2 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 WSIs per patient were collected and 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 into 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[9], and were specifically modified for the classification task at hand by replacing the three top layers with a 1,000-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: ResNet-18, AlexNet, Inception-V3, DenseNet-201 and ShuffleNet. The sampled hyperparameter space was as follows: learning rate: 5×10^{-5} and 1×10^{-4}; maximum number of tiles per WSI: 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[9]. As a benchmark task, we used MSI detection in colorectal cancer as described before[9].

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 analysis, with AUROC being the primary endpoint for each target feature.

Alternative approaches. In our baseline approach, image tiles from manually annotated tumor regions on FFPE slides (diagnostic slides) were used. This approach was compared with several alternative approaches, as shown in Extended Data Figs. 7–9. The first alternative approach used color normalization of image tiles with the Macenko method[43] to mitigate differences in staining intensity and hue (Extended Data Fig. 7). Some previous studies have used color normalization for deep learning[9], while other studies have shown that color normalization can bias histology image classification[52]. 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 (Extended Data Fig. 8). The third alternative approach was to use frozen slides as opposed to FFPE slides in a weakly supervised way (Extended Data Fig. 9).

Statistics and reproducibility. AUROC values are reported as means with a confidence interval representing the lower and upper range of a 10× bootstrapped experiment. To quantify whether predictions for different classes of patients

were statistically significant, the probability scores for patients in a given class were compared with the probability scores of all other patients. The statistical significance of these differences was assessed with a two-sided 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 FDR correction, using 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 Extended Data Fig. 10a–c. The number of tiles generated per WSI is shown in Extended Data Fig. 10d. No statistical method was used to predetermine sample size. The investigators were blinded to the molecular status of samples during manual annotation, image processing procedures and outcome assessment. Source codes are publicly available, allowing replication of our findings. The investigators re-ran the computer codes three times, receiving identical results.

Implementation and hardware. Training and inference were performed on our local computing cluster on ten Nvidia RTX graphics-processing units (GPUs), each with 24 GB of GPU random-access memory. The 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).

External validation. To investigate whether 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 CIMP in colorectal cancer in $n = 408$ patients—a subset of the multicenter DACHS study from whom data were previously collected, as described[9]. *BRAF* and CIMP were chosen as validation markers because of their biological relevance and the 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 approaches. 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 were 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.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

All data, including histological images and information about the age and sex of the participants from the TCGA database are available at https://portal.gdc.cancer.gov/. Genetic data for patients in the TCGA cohorts are available at https://portal.gdc.cancer.gov/ and https://cbioportal.org. Raw data for the DACHS cohort are stored and administered by the DACHS consortium (more information is available from http://dachs.dkfz.org/dachs/). The corresponding authors of this study are not involved in data sharing decisions of the DACHS consortium. All other data supporting the findings of this study are available from the corresponding author upon reasonable request. Source data are provided with this paper.

Code availability

All source codes are available under an open-source license at https://github.com/jnkather/DeepHistology/releases/tag/v0.2.

References

1. Cheng, M. L., Berger, M. F., Hyman, D. M. & Solit, D. B. Clinical tumour sequencing for precision oncology: time for a universal strategy. *Nat. Rev. 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. *Nat. Commun.* **9**, 3962 (2018).
3. Kather, J. N., Halama, N. & Jaeger, D. Genomics and emerging biomarkers for immunotherapy of colorectal cancer. *Semin. Cancer Biol.* **52**, 189–197 (2018).
4. Guinney, J. et al. The consensus molecular subtypes of colorectal cancer. *Nat. Med.* **21**, 1350–1356 (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 Pathol.* **30**, 599–609 (2017).
7. Greenson, J. K. et al. Pathologic predictors of microsatellite instability in colorectal cancer. *Am. J. Surg. Pathol.* **33**, 126–133 (2009).
8. Le, D. T. et al. PD-1 blockade in tumors with mismatch-repair deficiency. *N. Engl. J. Med.* **372**, 2509–2520 (2015).
9. Kather, J. N. et al. Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. *Nat. Med.* **25**, 1054–1056 (2019).
10. Coudray, N. et al. Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. *Nat. Med.* **24**, 1559–1567 (2018).
11. 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).
12. Schaumberg, A. J., Rubin, M. A. & Fuchs, T. J. H&E-stained whole slide image deep learning predicts SPOP mutation state in prostate cancer. Preprint at *bioRxiv* https://doi.org/10.1101/064279 (2018).
13. Kather, J. N. et al. Deep learning detects virus presence in cancer histology. Preprint at *bioRxiv* https://doi.org/10.1101/690206 (2019).
14. 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 (Institute of Electrical and Electronics Engineers, 2019); https://doi.org/10.1109/BIBM47256.2019.8983139
15. Campanella, G. et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nat. Med.* **25**, 1301–1309 (2019).
16. 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 (Institute of Electrical and Electronics Engineers, 2018); https://doi.org/10.1109/CVPR.2018.00716
17. 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 (Institute of Electrical and Electronics Engineers, 2017); https://doi.org/10.1109/CVPR.2017.243
18. 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 (Institute of Electrical and Electronics Engineers, 2016); https://doi.org/10.1109/CVPR.2016.30
19. 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 (Institute of Electrical and Electronics Engineers, 2016); https://doi.org/10.1109/CVPR.2016.90
20. Srinidhi, C. L., Ciga, O. & Martel, A. L. Deep neural network models for computational histopathology: a survey. Preprint at https://arxiv.org/abs/1912.12378 (2019).
21. Chen, P. C. et al. An augmented reality microscope with real-time artificial intelligence integration for cancer diagnosis. *Nat. Med.* **25**, 1453–1457 (2019).
22. Muzny, D. M. et al. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**, 330–337 (2012).
23. Fukamachi, H. et al. A subset of diffuse-type gastric cancer is susceptible to mTOR inhibitors and checkpoint inhibitors. *J. Exp. Clin. Cancer Res.* **38**, 127 (2019).
24. The Cancer Genome Atlas Network et al. Comprehensive molecular profiling of lung adenocarcinoma. *Nature* **511**, 543–550 (2014).
25. The Cancer Genome Atlas Networket al. Comprehensive molecular characterization of gastric adenocarcinoma. *Nature* **513**, 202–209 (2014).
26. André, F. et al. Alpelisib for *PIK3CA*-mutated, hormone receptor-positive advanced breast cancer. *N. Engl. J. Med.* **380**, 1929–1940 (2019).
27. The Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61–70 (2012).
28. Xue, Z. et al. *MAP3K1* and *MAP2K4* mutations are associated with sensitivity to MEK inhibitors in multiple cancer models. *Cell Res.* **28**, 719–729 (2018).
29. The Cancer Genome Atlas Network. Genomic classification of cutaneous melanoma. *Cell* **161**, 1681–1696 (2015).
30. The Cancer Genome Atlas Research Network. The molecular taxonomy of primary prostate cancer. *Cell* **163**, 1011–1025 (2015).
31. Cancer Genome Atlas Research Network. Integrated genomic characterization of pancreatic ductal adenocarcinoma. *Cancer Cell* **32**, 185–203.e13 (2017).
32. Hammerman, P. S. et al. Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **489**, 519–525 (2012).
33. The Cancer Genome Atlas Research Network. Comprehensive and integrative genomic characterization of hepatocellular carcinoma. *Cell* **169**, 1327–1341.e23 (2017).
34. Khalaf, A. M. et al. Role of Wnt/β-catenin signaling in hepatocellular carcinoma, pathogenesis, and clinical significance. *J. Hepatocell. Carcinoma* **5**, 61–73 (2018).
35. Linehan, W. M. et al. Comprehensive molecular characterization of papillary renal-cell carcinoma. *N. Engl. J. Med.* **374**, 135–145 (2016).

36. Creighton, C. J. et al. Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature* **499**, 43–49 (2013).
37. Davis, C. F. et al. The somatic genomic landscape of chromophobe renal cell carcinoma. *Cancer Cell* **26**, 319–330 (2014).
38. The Cancer Genome Atlas Network. Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature* **517**, 576–582 (2015).
39. 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. *Mol. Oncol.* **8**, 1220–1230 (2014).
40. Burk, R. D. et al. Integrated genomic and molecular characterization of cervical cancer. *Nature* **543**, 378–384 (2017).
41. Thorsson, V. et al. The immune landscape of cancer. *Immunity* **48**, 812–830.e14 (2018).
42. Liu, Y. et al. Comparative molecular analysis of gastrointestinal adenocarcinomas. *Cancer Cell* **33**, 721–735.e8 (2018).
43. 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 (Institute of Electrical and Electronics Engineers, 2009); https://doi.org/10.1109/ISBI.2009.5193250
44. 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).
45. Hoffmeister, M. et al. Statin use and survival after colorectal cancer: the importance of comprehensive confounder adjustment. *J. Natl Cancer Inst.* **107**, djv045 (2015).
46. 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).
47. Cerami, E. et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* **2**, 401–404 (2012).
48. Gao, J. et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal* **6**, pl1 (2013).
49. Bailey, M. H. et al. Comprehensive characterization of cancer driver genes and mutations. *Cell* **173**, 371–385.e18 (2018).
50. Berger, A. C. et al. A comprehensive pan-cancer molecular study of gynecologic and breast cancers. *Cancer Cell* **33**, 690–705.e9 (2018).
51. Bankhead, P. et al. QuPath: open source software for digital pathology image analysis. *Sci. Rep.* **7**, 16878 (2017).
52. 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* (eds. Reyes-Aldasoro, C. et al.) 56–64 (Lecture Notes in Computer Science Volume 11435, Springer, 2019).

Author contributions

J.N.K., A.T.P. and T.L. designed the study. L.R.H., H.I.G., N.A.C., J.J.S., P.A.v.D.B., L.F.S.K., P. Boor and A.P. oversaw the tumor annotation. C.L., A.E., J.K., H.S.M., J.M.N., R.D.B. and K.A.J.S. manually annotated all of the tumors. J.N.K., J.K., J.M.N. and P. Bankhead designed and implemented the algorithm. J.N.K., C.L., A.S., S.K., R.D.B. and N.O.-B. curated the list of molecular alterations. H.B., M.H., A.T.P., A.M.H. and V.S. provided external validation samples and gave statistical advice. C.T., D.J., A.T.P., P. Boor, V.S. and T.L. provided infrastructure and supervised the study. All authors contributed to the data analysis and writing of the manuscript.

Competing interests

J.N.K. has an informal, unpaid advisory role at Pathomix (Heidelberg, Germany) that does not relate to this research. J.N.K. declares no other relationships or competing interests. All other authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s43018-020-0087-6.

Supplementary information is available for this paper at https://doi.org/10.1038/s43018-020-0087-6.

Correspondence and requests for materials should be addressed to J.N.K., A.T.P. or T.L.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.