



UNIVERSITY OF LEEDS

This is a repository copy of *Pan-cancer image-based detection of clinically actionable genetic alterations*.

White Rose Research Online URL for this paper:
<http://eprints.whiterose.ac.uk/159675/>

Version: Accepted Version

Article:

Kather, JN, Heij, LR, Grabsch, H orcid.org/0000-0001-9520-6228 et al. (25 more authors)
(2020) Pan-cancer image-based detection of clinically actionable genetic alterations.
Nature Cancer, 1 (8). pp. 789-799. ISSN 2662-1347

<https://doi.org/10.1038/s43018-020-0087-6>

© 2020, Springer Nature America, Inc.. This is an author produced version of an article published in Nature Cancer. Uploaded in accordance with the publisher's self-archiving policy.

Reuse

Items deposited in White Rose Research Online are protected by copyright, with all rights reserved unless indicated otherwise. They may be downloaded and/or printed for private study, or other acts as permitted by national copyright laws. The publisher or other rights holders may allow further reproduction and re-use of the full text version. This is indicated by the licence information on the White Rose Research Online record for the item.

Takedown

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.



eprints@whiterose.ac.uk
<https://eprints.whiterose.ac.uk/>

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 Loeffler¹, Amelie Echle¹, Hannah Sophie Muti¹, Jeremias Krause¹, Jan M. Niehues¹, Kai A. J. Sommer¹, Peter Bankhead⁹, Loes F. S. Kooreman⁷, Jefree J. Schulte¹⁰, Nicole A. Cipriani¹⁰, Roman D. Buelow⁶, Peter Boor⁶, Nadina Ortiz-Brüchle⁶, Andrew M. Hanby⁸, Valerie Speirs¹¹, Sara Kochanny¹², Akash Patnaik¹², Andrew Srisuwananukorn¹³, Hermann Brenner^{2,14,15}, Michael Hoffmeister¹⁴, Piet A. van den Brandt¹⁶, Dirk Jäger^{2,3}, Christian Trautwein¹, 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

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

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

24 ¹⁰ Department of Pathology, University of Chicago Medicine, Chicago, IL, USA

25 ¹¹ Institute of Medical Sciences, School of Medicine, Medical Sciences and Nutrition, University
26 of Aberdeen, Aberdeen, UK

27 ¹² Department of Medicine, University of Chicago Medicine, Chicago, IL, USA

28 ¹³ Department of Medicine, University of Illinois – Chicago, Chicago, IL, USA

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

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

33 ¹⁶ Department of Epidemiology, GROW School for Oncology and Developmental Biology, Maas-
34 tricht University Medical Center+, Maastricht, The Netherlands

35 ¹⁷ Division of Gastroenterology, Hepatology and GI Oncology, University Hospital RWTH Aachen,
36 Aachen, Germany

37 * these authors contributed equally to this work

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

40

41 **Abstract**

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

56 **Introduction**

57 Precision treatment of cancer relies on detection of genetic alterations which are diagnosed by
58 molecular biology assays.¹ These tests can be a bottleneck in oncology workflows because of high
59 turnaround time, tissue usage and costs.² Clinical guidelines recommend molecular testing of
60 tumor tissue for most patients with advanced solid tumors. However, in most tumor types, rou-
61 tine testing includes only a handful of alterations, such as KRAS, NRAS, BRAF mutations and mi-
62 crosatellite instability (MSI) in colorectal cancer.³ While new studies identify more and more mo-
63 lecular features of potential clinical relevance, current diagnostic workflows are not designed to
64 incorporate an exponentially rising load of tests. For example, in colorectal cancer, previous stud-
65 ies have identified consensus molecular subtypes (CMS)⁴ as a candidate biomarker, but sequenc-
66 ing costs and method complexity preclude widespread testing in clinical routine and clinical tri-
67 als.⁵ Therefore, there is a growing need to identify new, inexpensive and scalable biomarkers in
68 medical oncology.

69 While comprehensive molecular and genetic tests are hard to implement at scale, tissue sections
70 stained with hematoxylin and eosin (H&E) are ubiquitously available. We hypothesized that these
71 routine tissue sections contain information about established and candidate biomarkers and that
72 molecular biomarkers could be inferred directly from digitized whole slide images (WSI). The ra-
73 tionale for this hypothesis is that genetic changes in tumor cells cause functional changes, which
74 can influence tumor cell morphology.^{6,7} In addition to such first-order genotype-phenotype cor-
75 relations, genetic changes in tumor cells can influence the tumor microenvironment, resulting in
76 higher-order genotype-phenotype correlations. Specific examples for such correlations are
77 known for microsatellite instability (MSI)⁸, a clinically approved biomarker for cancer immuno-
78 therapy in colorectal cancer.⁹ In the case of MSI, the genotype-phenotype correlation is con-
79 sistent enough to robustly infer the genotype just by observing morphological features in a his-
80 tological image, as we have previously shown.¹⁰ Other previous studies have identified genotype-
81 phenotype links for selected genetic features in lung cancer^{11,12}, prostate cancer¹³, head and
82 neck¹⁴ and liver¹⁵ cancer, among others. Building on these previous studies, we systematically

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

95

96 **Methods**

97 Patient cohorts and ethics statement

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

111 Molecular labels

112 The aim of this study was to predict clinically relevant features, including genetic alterations, di-
113 rectly from routine histology slides. We systematically applied this screening approach to four
114 groups of molecular alterations: First, we used single-gene mutations, considering any genetic
115 variant. We used the most commonly mutated genes in the respective tumor types (derived from
116 the “cbioportal” database^{32,33} at <http://cbioportal.org>) and clinically targetable genes (level one
117 genes from OncoKB at <http://www.oncokb.org>, Pan Cancer Atlas Project³⁴). We required each
118 mutation to affect at least four patients in a given cohort. Second, we repeated the analysis on
119 putative and confirmed oncogenic driver mutations only, as defined in OncoKB. Third, we aimed
120 to predict gene expression subtypes, relevant gene expression signatures and immune-cell gene
121 expression signatures derived from systematic studies³⁵⁻³⁷. Fourth, we used “standard of care”
122 features derived from the TCGA database (data at <http://portal.gdc.cancer.gov/>), including hor-
123 mone receptor status in breast cancer. All labels (genetic variants, driver mutations, signatures

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

127 Image preprocessing

128 Scanned whole slide images of diagnostic tissue slides (formalin-fixed paraffin-embedded tissue)
129 stained with hematoxylin and eosin were acquired in SVS format. All images were downsampled
130 to 20x magnification, corresponding to 0.5 $\mu\text{m}/\text{pixel}$ (px). Each whole slide image was manually
131 reviewed and the tumor area with was annotated under direct supervision of a specialty
132 pathologist. During annotation, all observers were blinded with regard to any molecular or clinical
133 feature. Only those images containing at least 1 mm^2 contiguous tumor tissue were used for
134 downstream analysis. 6% of whole slide images, corresponding to 5% of patients were excluded
135 due to technical artifacts or lack of tumor (Suppl. Table 2). Tumor tissue on all other slides was
136 tessellated into square tiles of 512x512 px edge length, corresponding to 256x256 μm at a reso-
137 lution of 0.5 $\mu\text{m}/\text{px}$. Tiles with more than 50% background were discarded; background pixels
138 were defined by brightness over 0.86 (220/255). For the benchmark task (identification of an
139 optimal neural network model), these images were resized to 224x224 px (at 1.14 $\mu\text{m}/\text{px}$) to be
140 consistent with a previous study¹⁰. All steps in the data preprocessing pipeline (including prepro-
141 cessing of images and preprocessing of metadata) are documented in detail in our in-house man-
142 ual for data preparation, which is publicly available at [https://dx.doi.org/10.5281/ze-](https://dx.doi.org/10.5281/zenodo.3694994)
143 [nodo.3694994](https://dx.doi.org/10.5281/zenodo.3694994). All methods for whole slide image processing, including tessellation of images
144 and visualization of spatial activation maps, were implemented in QuPath v0.1.2 in Groovy
145 (<http://qupath.github.io>).

146 Patient-level cross-validation

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

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

156 Neural network training, model selection and hyperparameter optimization

157 Deep neural networks were trained on image tiles with the aim of predicting molecular labels.
158 All neural networks were pre-trained on the ImageNet database as described previously¹⁰ and
159 were specifically modified for the classification task at hand by replacing the three top layers with
160 a 1000-neuron fully connected layer, a softmax layer and a classification layer. For training, we
161 used on-the-fly data augmentation (random horizontal and vertical reflection) to achieve rota-
162 tional invariance of the classifiers. Hyperparameter selection was performed for five commonly
163 used deep neural networks: resnet18, alexnet, inceptionv3, densenet201 and shufflenet. The
164 sampled hyperparameter space was as follows: learning rate 5e-5 and 1e-4, maximum number
165 of tiles per whole slide image: 250, 500 and 750, number of trainable layers: 10, 20 and 30. We
166 trained for four epochs with a mini batch size of 512, similar to previous experiments.¹⁰ As a
167 benchmark task, we used MSI detection in colorectal cancer as described before.¹⁰

168 Inference of molecular status

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

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

186 External validation

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

194 Feature visualization

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

202 Alternative approaches

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

206 normalization of image tiles with the Macenko method³⁸ to mitigate differences in staining in-
207 tensity and hue (Suppl. Fig. 4). Some previous studies have used color normalization for deep
208 learning¹⁰, while other studies have shown that color normalization can bias histology image clas-
209 sification.³⁹ The second alternative approach we investigated was to use tiles from the whole
210 slide, as opposed to the tumor region only. In this “weakly supervised” approach, many tiles with-
211 out invasive cancer tissue were present in the training and inference sets (Suppl. Fig. 5). The third
212 alternative approach was to use frozen slides as opposed to FFPE slides in a weakly supervised
213 way (Suppl. Fig. 6).

214 Data availability

215 All data (including histological images) from the TCGA database are available at [https://por-](https://portal.gdc.cancer.gov/)
216 [tal.gdc.cancer.gov/](https://portal.gdc.cancer.gov/). All molecular data for patients in the TCGA cohorts are available at
217 <https://cbioportal.org>. Raw data for Figures and Suppl. Figures are shown in Suppl. Table 3.

218 Code availability

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

220

221 **Results**

222 Optimization of deep learning for inference of genotype from histology

223 We hypothesized that deep learning can infer molecular alterations directly from routine histol-
224 ogy images across multiple common solid tumor types. To test this, we developed, optimized and
225 extensively validated a new ‘one-stop-shop’ workflow to train and evaluate deep learning net-
226 works. To select an efficient network model and to optimize the deep learning hyperparameters,
227 prediction of microsatellite instability (MSI) in colorectal cancer was used as a clinically relevant
228 benchmark task¹⁰. In this benchmark, we sampled a large hyperparameter space with different
229 commonly used deep learning models^{10,11,14,40} which were modified specifically for this applica-
230 tion. Unexpectedly, ‘shufflenet’⁴¹, a lightweight neural network architecture performed similarly
231 to more complex networks including ‘densenet’⁴², ‘inception’⁴³ and ‘resnet’⁴⁴ networks, which
232 are used in many other studies⁴⁵ (Fig. 1c). Shufflenet demonstrated high accuracy at a low train-
233 ing time (raw data in Suppl. Table 1, N=426 patients in the TCGA cohort). Shufflenet is optimized
234 for mobile devices, making this deep neural network architecture attractive for decentralized
235 point-of-care image analyses or direct implementation in microscopes⁴⁶. We externally validated
236 the best shufflenet classifier by training on N=426 patients in the TCGA-CRC cohort¹⁰ and validat-
237 ing on N=379 patients with available MSI status in the DACHS cohort¹⁰, reaching an AUROC of
238 0.89 [0.88; 0.92]. This represents an improvement over the previous best performance of 0.84 in
239 that dataset¹⁰ and supports the notion that shufflenet is an efficient and powerful neural network
240 model which can infer clinically relevant molecular changes directly from histology images.

241 Pan-cancer prediction of genetic variants from histology

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

249 Fig. 1d). First, we trained deep neural networks to detect any sequence variants in these target
250 genes. We found that in 13 out of 14 tested tumor types, the mutation of one or more of such
251 genes could be inferred from histology images alone, with statistical significance after correction
252 for multiple testing (Fig. 2a-n, Suppl. Fig. 7). In particular, in major cancer types such as lung ad-
253 enocarcinoma, colorectal cancer, breast cancer and gastric cancer, alterations of several genes
254 of particular clinical and/or biological examples were detectable (Fig. 2a-d). Examples include
255 mutations in TP53, which could be significantly detected in all four of these cancer types, as well
256 as mutations of BRAF in colorectal cancer (TCGA-COAD and TCGA-READ¹⁸, N=555, Fig. 2b), MTOR
257 – a candidate for targeted treatment⁴⁷ – in gastric cancer (Fig. 2d) and FBXW7 mutation in lung
258 adenocarcinoma (TCGA-LUAD²², N=457, Fig. 2a) and gastric cancer (TCGA-STAD¹⁹, N=321, Fig.
259 2d). Mutations of PIK3CA (which is directly targetable by a small molecule inhibitor⁴⁸) was signif-
260 icantly detectable in breast cancer (TCGA-BRCA¹⁶, N=995, Fig. 2c) and gastric cancer (Fig. 2d). In
261 addition, in breast cancer, mutations of MAP2K4 (which is a potential biomarker for response to
262 MEK inhibitors⁴⁹) were significantly detectable (Fig. 2c). Among all tested tumor types, gastric
263 cancer (Fig. 2d) and colorectal cancer (Fig. 2b) had the highest absolute number of detectable
264 mutations. For all statistically significant features, the mean cross-validated area under the re-
265 ceiver operating curve (AUROC) for the top eight mutations ranged from 0.60 to 0.78 in lung
266 adenocarcinoma (Suppl. Fig. 8); from 0.65 to 0.76 in colorectal cancer (Suppl. Fig. 9); from 0.62
267 to 0.78 in breast cancer (Suppl. Fig. 10) and from 0.66 to 0.78 in gastric cancer (Suppl. Fig. 11).
268 Beyond these four tumor types, a range of notable mutations could be detected in other tumor
269 types: While in melanoma (TCGA-SKCM²⁴) primary tumors, few mutations were detectable
270 (Suppl. Fig. 12a-h), in melanoma metastases, mutations in FBXW7 and PIK3CA were significantly
271 detectable (Fig. 2e, Suppl. Fig. 12i-p). In prostate cancer (TCGA-PRAD²⁶, N=397 patients, Fig. 2f,
272 Suppl. Fig. 13), our method detected TP53 and FOXA1 mutations from histology, among others.
273 In pancreatic adenocarcinoma (TCGA-PAAD²⁵, N=171 patients, Fig. 2g, Suppl. Fig. 14), identifying
274 KRAS wild type patients is of high clinical relevance because these patients are potential candi-
275 dates for targeted treatment and our method significantly identified KRAS genotype in pancreatic
276 cancer. Lung squamous cell carcinoma is known for its difficulty in molecular diagnosis and few
277 molecularly or genetically targeted treatment options even in clinical trials. Thus, it is plausible

278 that in this cancer type, tumor histomorphology is not well correlated to mutations and corre-
279 spondingly, few mutations were significantly detectable in this tumor type in our experiments
280 (TCGA-LUSC, N=413, Fig. 2h, Suppl. Fig. 15). In hepatocellular carcinoma (TCGA-LIHC²¹, N=358
281 patients, Fig. 2i), the β -catenin gene (CTNNB1) is a key driver gene with broad prognostic and
282 predictive implications⁵⁰ and its mutational status was highly significantly detected from histol-
283 ogy (Suppl. Fig. 16). In papillary (Fig. 2j, Suppl. Fig. 17) and clear cell renal cell carcinoma (Fig. 2k,
284 Suppl. Fig. 18), alterations in multiple genes including KRAS and PBRM were highly significantly
285 detectable while in and chromophobe renal cell carcinoma (Fig. 2l, Suppl. Fig. 19), no genetic
286 variants were significantly detectable, possibly due to a low patient number in this cohort. In
287 head and neck squamous cell carcinoma (TCGA-HNSC²⁰, N=435 patients), genotype of CASP8,
288 which is linked to resistance to cell death⁵¹, was significantly detected (Fig. 2m, Suppl. Fig. 20). In
289 cervical cancer (TCGA-CESC¹⁷, N=261 patients), mutations in TCERG1, STK11, AMER1, among oth-
290 ers, were significantly detectable with high AUROC values (Fig. 2n, Suppl. Fig. 21). Raw data for
291 prediction performance in any gene in any tumor type are available in Suppl. Table 3.

292 Pan-cancer prediction of oncogenic drivers from histology

293 Not all genetic variants are causative of malignant processes. Therefore, we repeated the screen-
294 ing experiment, limiting mutations to confirmed or putative oncogenic drivers (Fig. 3a-n). With
295 this criterion, the absolute number of patients affected by a particular mutation was lower and
296 thus, fewer genes met the threshold of at least four positive cases in a given tumor type. On the
297 other hand, we hypothesized that oncogenic driver genes could leave a stronger pattern in his-
298 tological morphology due to their higher biological relevance. Genetic variants in classical onco-
299 genes such as TP53 and KRAS are almost always oncogenic drivers and correspondingly, muta-
300 tions of these genes reached similar prediction accuracy valued in the “drivers only” experiment
301 when compared to the “all variants” approach (Fig. 3a-n). For other genes, prediction accuracy
302 increased when limited to oncogenic drivers: a notable example was EGFR in lung adenocarci-
303 noma (Fig. 3a). In summary, these data show that deep learning can detect a wide range of tar-
304 getable and potentially targetable point mutations directly from histology across multiple preva-
305 lent 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 signatures³⁷ of CD8-positive lymphocytes, macrophages,
313 cell proliferation, interferon-gamma (IFN γ) signaling and transforming growth factor-beta (TGF β)
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 subtypes³⁷ with up to AUROC 0.74 in lung
328 adenocarcinoma (Fig. 4e), pan-gastrointestinal subtypes³⁶ 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

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

337 Prediction of standard histological biomarkers with deep learning

338 To comprehensively evaluate the potential clinical use of our new deep learning pipeline, we
339 investigated classification accuracy for standard histopathological biomarkers. We found that
340 deep learning could highly significantly predict most of these biomarkers for breast cancer (Fig.
341 4c and i), gastric cancer (Fig. 4d and j) and other tumor types (Suppl. Fig. 11-18). In particular,
342 status of hormone receptors was predictable from routine histology in breast cancer, with an
343 AUROC of 0.82 for estrogen receptor and 0.74 for progesterone receptor (Fig. 4i). Together, these
344 results demonstrate that deep-learning-based inference of genetic alterations, high-level molec-
345 ular alterations and established biomarkers from routine diagnostic histology slides is feasible.

346 Evaluation of alternative approaches

347 Deep learning-based inference of molecular features from histology is a relatively novel field of
348 research and it can be anticipated that technical improvements can further improve prediction
349 performance. We quantified the effect of alternative technical approaches in the colorectal can-
350 cer cohort (TCGA-COAD/READ). First, we investigated the role of color normalization of tiles. In a
351 head-to-head comparison to the baseline approach, we found a tendency of Macenko's³⁸ color
352 normalization to improve classifier performance for mutation prediction but not for prediction of
353 subtypes or gene expression signatures (Suppl. Fig. 4a-c). Second, we investigated a weakly su-
354 pervised approach to our baseline of expert-annotated tumor regions and found that the weakly
355 supervised approach was only slightly inferior to manual annotation (Suppl. Fig. 4d-f). Third, we
356 analyzed prediction performance on frozen slides compared to diagnostic slides. While frozen
357 slides are not generally available in a clinical setting, the TCGA database provides an opportunity
358 to perform such a direct comparison. In a weakly supervised experiment, we found that predic-
359 tion power for driver genes was on par, but prediction power for genetic variants and high-level
360 subtypes/signatures was better in frozen slides than in diagnostic slides (Suppl. Fig. 4g-h). These
361 data provide quantitative guidance for future large-scale validation studies.

362 External validation of the classification results

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

372

373 **Discussion**

374 Image-based genetic testing as a clinical and research tool

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

400 Limitations

401 Currently, limitations of our method are the low AUROC values for some molecular features (Fig.
402 2 and Fig. 3). A strategy to increase the diagnostic performance would be re-training on larger
403 patient cohorts. Re-training can be expected to boost performance because previous studies
404 have shown that performance of deep learning systems in histopathology scales with the number
405 of patients in the training cohort.⁴⁰ In addition, the performance of deep learning systems could
406 potentially be improved by technical modifications. Our systematic evaluation of alternative
407 technical approaches provides a guidance for this on multiple levels: First, regarding the choice
408 of neural network models, our results demonstrate that lightweight neural network models per-
409 form on par with more complex models, facilitating further evaluation of these methods on de-
410 centralized hardware, including desktop or ultimately mobile hardware. While this finding is
411 based on a clinically relevant benchmark task and generalizes to an external population, we can-
412 not exclude that other network models perform better in other histology applications. Second,
413 regarding the type of input image data, other studies in digital pathology have used frozen his-
414 tology sections¹¹. In contrast, our baseline workflow was based on FFPE tissue slides (labeled as
415 'diagnostic slides' in the TCGA archive) due to their clinical relevance. In clinical settings, frozen
416 specimens constitute only a small fraction of pathology samples and therefore, establishing
417 methods on FFPE material is paramount for large-scale clinical validation. Our head-to-head com-
418 parison shows that molecular inference generally works better on frozen slides, which is a limi-
419 tation of the FFPE-based method. Further studies are needed to determine the reasons for this
420 observation. Lastly, our baseline method relied on expert annotations of tumor tissue, constrain-
421 ing deep learning models to learn from invasive tumor tissue only. The rationale behind this de-
422 sign was that despite advances in computer vision, expert annotation of tumor tissue remains
423 the gold standard in histopathology studies. Yet, in a head-to-head comparison, a weakly super-
424 vised approach without any manual annotation did not markedly reduce performance, demon-
425 strating feasibility of even simpler data preprocessing pipelines. We publicly release all source
426 codes of our method, enabling further optimization and validation on a larger scale.

427 Deciphering genotype-phenotype links

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

453 Conclusion

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

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

458

459 **Competing interests**

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

462 **Funding**

463 The results are in part based upon data generated by the TCGA Research Network: [http://can-](http://cancergenome.nih.gov/)
464 cancergenome.nih.gov/. Our funding sources are as follows. J.N.K.: RWTH University Aachen (START
465 2018-691906). V.S.: Breast Cancer Now, P.Bo: DFG: (SFB/TRR57, SFB/TRR219, BO3755/3-1, and
466 BO3755/6-1), the German Ministry of Education and Research (BMBF: STOP-FSGS-01GM1901A)
467 and the German Ministry of Economic Affairs and Energy (BMWi: EMPAIA project). A.T.P.:
468 NIH/NIDCR (#K08-DE026500), Institutional Research Grant (#IRG-16-222-56) from the American
469 Cancer Society, Cancer Research Foundation Research Grant, and the University of Chicago Med-
470 icine Comprehensive Cancer Center Support Grant (#P30-CA14599). T.L.: Horizon 2020 through
471 the European Research Council (ERC) Consolidator Grant PhaseControl (771083), a Mildred-
472 Scheel-Endowed Professorship from the German Cancer Aid (Deutsche Krebshilfe), the German
473 Research Foundation (DFG) (SFB CRC1382/P01, SFB-TRR57/P06, LU 1360/3-1), the Ernst-Jung-
474 Foundation Hamburg and the IZKF (interdisciplinary center of clinical research) at RWTH Aachen.

475 **Authors' contributions**

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

482

483 **Figure Legends**

484 **Fig. 1:** Deep learning workflow for prediction of molecular features from histology images. We
485 describe a comprehensive method pipeline for prediction of molecular features directly from his-
486 tological images. (a) Training of the deep learning system comprised six steps. Step 1: Patient
487 cohorts were randomly split into three partitions for cross-validation of deep classifiers. Step 2:
488 The tumor region on each whole slide image (WSI) was tessellated into tiles. Step 3: Up to 500
489 randomly chosen tiles were collected. Step 4: Tiles from patients in the training partitions were
490 collected, classes were equalized by random undersampling. Step 5: All training tiles were used
491 to train a deep neural network (pre-trained on a non-medical task). Step 6: Classification perfor-
492 mance was evaluated on patients from the test partition. (b) For patient-level inference of mo-
493 lecular labels in patients not seen during training, three successive steps were used. Step 1: Tiles
494 were generated from the tumor region on WSI. Step 2: A prediction was made for each tile. Step
495 3: Tile-level class predictions were pooled on a patient level. (c) Hyperparameters of the deep
496 learning system were optimized in a benchmark task (prediction of microsatellite instability sta-
497 tus [MSI] in colorectal cancer). The opacity of each point corresponds to the number of trainable
498 layers (Suppl. Table 3). Shufflenet, a lightweight neural network architecture was selected as a
499 highly efficient network model. (d) This workflow was subsequently applied for prediction of four
500 types of molecular features across 14 cancer types. In particular, this included genetic mutations.
501 The distribution of the 20 most common among all analyzed mutations is shown for each tumor
502 type.

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

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

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

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

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

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

549

550 Bibliography

- 551 1. Cheng, M.L., Berger, M.F., Hyman, D.M. & Solit, D.B. Clinical tumour sequencing for precision
552 oncology: time for a universal strategy. *Nature Reviews Cancer* 18, 527-528 (2018).
- 553 2. Rusch, M., *et al.* Clinical cancer genomic profiling by three-platform sequencing of whole genome,
554 whole exome and transcriptome. *Nature Communications* 9, 3962 (2018).
- 555 3. Kather, J.N., Halama, N. & Jaeger, D. Genomics and emerging biomarkers for immunotherapy of
556 colorectal cancer. *Seminars in Cancer Biology* 52, 189-197 (2018).
- 557 4. Guinney, J., *et al.* The consensus molecular subtypes of colorectal cancer. *Nature Medicine* 21,
558 1350 (2015).
- 559 5. Fontana, E., Eason, K., Cervantes, A., Salazar, R. & Sadanandam, A. Context matters-consensus
560 molecular subtypes of colorectal cancer as biomarkers for clinical trials. *Ann Oncol* 30, 520-527
561 (2019).
- 562 6. Shia, J., *et al.* Morphological characterization of colorectal cancers in The Cancer Genome Atlas
563 reveals distinct morphology-molecular associations: clinical and biological implications. *Modern
564 pathology : an official journal of the United States and Canadian Academy of Pathology, Inc* 30,
565 599-609 (2017).
- 566 7. Greenson, J.K., *et al.* Pathologic predictors of microsatellite instability in colorectal cancer. *The
567 American journal of surgical pathology* 33, 126-133 (2009).
- 568 8. Greenson, J.K., *et al.* Pathologic predictors of microsatellite instability in colorectal cancer. *Am J
569 Surg Pathol* 33, 126-133 (2009).
- 570 9. Le, D.T., *et al.* PD-1 Blockade in Tumors with Mismatch-Repair Deficiency. *New England Journal of
571 Medicine* 372, 2509-2520 (2015).
- 572 10. Kather, J.N., *et al.* Deep learning can predict microsatellite instability directly from histology in
573 gastrointestinal cancer. *Nature Medicine* (2019).
- 574 11. Coudray, N., *et al.* Classification and mutation prediction from non-small cell lung cancer
575 histopathology images using deep learning. *Nature Medicine* 24, 1559-1567 (2018).
- 576 12. Sha, L., *et al.* Multi-Field-of-View Deep Learning Model Predicts Nonsmall Cell Lung Cancer
577 Programmed Death-Ligand 1 Status from Whole-Slide Hematoxylin and Eosin Images. *J Pathol
578 Inform* 10, 24 (2019).
- 579 13. Schaumberg, A.J., Rubin, M.A. & Fuchs, T.J. H&E-stained Whole Slide Image Deep Learning
580 Predicts SPOP Mutation State in Prostate Cancer. *bioRxiv*, 064279 (2018).
- 581 14. Kather, J.N., *et al.* Deep learning detects virus presence in cancer histology. *bioRxiv*, 690206
582 (2019).
- 583 15. Zhang, H., *et al.* Predicting Tumor Mutational Burden from Liver Cancer Pathological Images Using
584 Convolutional Neural Network. in *2019 IEEE International Conference on Bioinformatics and
585 Biomedicine (BIBM)* 920-925 (2019).
- 586 16. The Cancer Genome Atlas Network, *et al.* Comprehensive molecular portraits of human breast
587 tumours. *Nature* 490, 61 (2012).
- 588 17. Burk, R.D., *et al.* Integrated genomic and molecular characterization of cervical cancer. *Nature*
589 543, 378-384 (2017).
- 590 18. Muzny, D.M., *et al.* Comprehensive molecular characterization of human colon and rectal cancer.
591 *Nature* 487, 330-337 (2012).
- 592 19. The Cancer Genome Atlas Network, *et al.* Comprehensive molecular characterization of gastric
593 adenocarcinoma. *Nature* 513, 202 (2014).
- 594 20. The Cancer Genome Atlas Network, *et al.* Comprehensive genomic characterization of head and
595 neck squamous cell carcinomas. *Nature* 517, 576 (2015).

- 596 21. The Cancer Genome Atlas Consortium. Comprehensive and Integrative Genomic Characterization
597 of Hepatocellular Carcinoma. *Cell* 169, 1327-1341.e1323 (2017).
- 598 22. The Cancer Genome Atlas Network, *et al.* Comprehensive molecular profiling of lung
599 adenocarcinoma. *Nature* 511, 543 (2014).
- 600 23. Hammerman, P.S., *et al.* Comprehensive genomic characterization of squamous cell lung cancers.
601 *Nature* 489, 519-525 (2012).
- 602 24. Cancer Genome Atlas, N. Genomic Classification of Cutaneous Melanoma. *Cell* 161, 1681-1696
603 (2015).
- 604 25. The Cancer Genome Atlas Network. Integrated Genomic Characterization of Pancreatic Ductal
605 Adenocarcinoma. *Cancer Cell* 32, 185-203.e113 (2017).
- 606 26. The Cancer Genome Atlas Network. The Molecular Taxonomy of Primary Prostate Cancer. *Cell*
607 163, 1011-1025 (2015).
- 608 27. Davis, C.F., *et al.* The somatic genomic landscape of chromophobe renal cell carcinoma. *Cancer*
609 *Cell* 26, 319-330 (2014).
- 610 28. Creighton, C.J., *et al.* Comprehensive molecular characterization of clear cell renal cell carcinoma.
611 *Nature* 499, 43-49 (2013).
- 612 29. Linehan, W.M., *et al.* Comprehensive Molecular Characterization of Papillary Renal-Cell
613 Carcinoma. *N Engl J Med* 374, 135-145 (2016).
- 614 30. Hoffmeister, M., *et al.* Statin use and survival after colorectal cancer: the importance of
615 comprehensive confounder adjustment. *J Natl Cancer Inst* 107, djv045 (2015).
- 616 31. Brenner, H., Chang-Claude, J., Seiler, C.M. & Hoffmeister, M. Long-term risk of colorectal cancer
617 after negative colonoscopy. *J Clin Oncol* 29, 3761-3767 (2011).
- 618 32. Cerami, E., *et al.* The cBio cancer genomics portal: an open platform for exploring
619 multidimensional cancer genomics data. *Cancer Discov* 2, 401-404 (2012).
- 620 33. Gao, J., *et al.* Integrative analysis of complex cancer genomics and clinical profiles using the
621 cBioPortal. *Sci Signal* 6, pl1 (2013).
- 622 34. Bailey, M.H., *et al.* Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell*
623 173, 371-385 e318 (2018).
- 624 35. Berger, A.C., *et al.* A Comprehensive Pan-Cancer Molecular Study of Gynecologic and Breast
625 Cancers. *Cancer Cell* 33, 690-705.e699 (2018).
- 626 36. Liu, Y., *et al.* Comparative Molecular Analysis of Gastrointestinal Adenocarcinomas. *Cancer Cell*
627 33, 721-735.e728 (2018).
- 628 37. Thorsson, V., *et al.* The Immune Landscape of Cancer. *Immunity* 48, 812-830.e814 (2018).
- 629 38. Macenko, M., *et al.* A method for normalizing histology slides for quantitative analysis. in *2009*
630 *IEEE International Symposium on Biomedical Imaging: From Nano to Macro* 1107-1110 (2009).
- 631 39. Bianconi, F., Kather, J.N. & Reyes-Aldasoro, C.C. Evaluation of Colour Pre-processing on Patch-
632 Based Classification of H&E-Stained Images. in *European Congress on Digital Pathology* 56-64
633 (Springer, 2019).
- 634 40. Campanella, G., *et al.* Clinical-grade computational pathology using weakly supervised deep
635 learning on whole slide images. *Nature Medicine* (2019).
- 636 41. Zhang, X., Zhou, X., Lin, M. & Sun, J. Shufflenet: An extremely efficient convolutional neural
637 network for mobile devices. in *Proceedings of the IEEE Conference on Computer Vision and Pattern*
638 *Recognition* 6848-6856 (2018).
- 639 42. Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K.Q. Densely connected convolutional
640 networks. in *Proceedings of the IEEE conference on computer vision and pattern recognition* 4700-
641 4708 (2017).

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

665

666

667 **Supplementary Figure Legends**

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

680 **Suppl. Fig. 2:** Distribution of tumor content across slides in all tumor types. Central mark = me-
681 dian, bottom and top edge of the box = 25th and 75th percentile, line extends to the most extreme
682 data points, circles = outliers. Outliers larger than 2000 mm² are not plotted. Median tumor con-
683 tent on slide is 139 mm² of tumor tissue per slide for colorectal cancer (CRC).

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

753

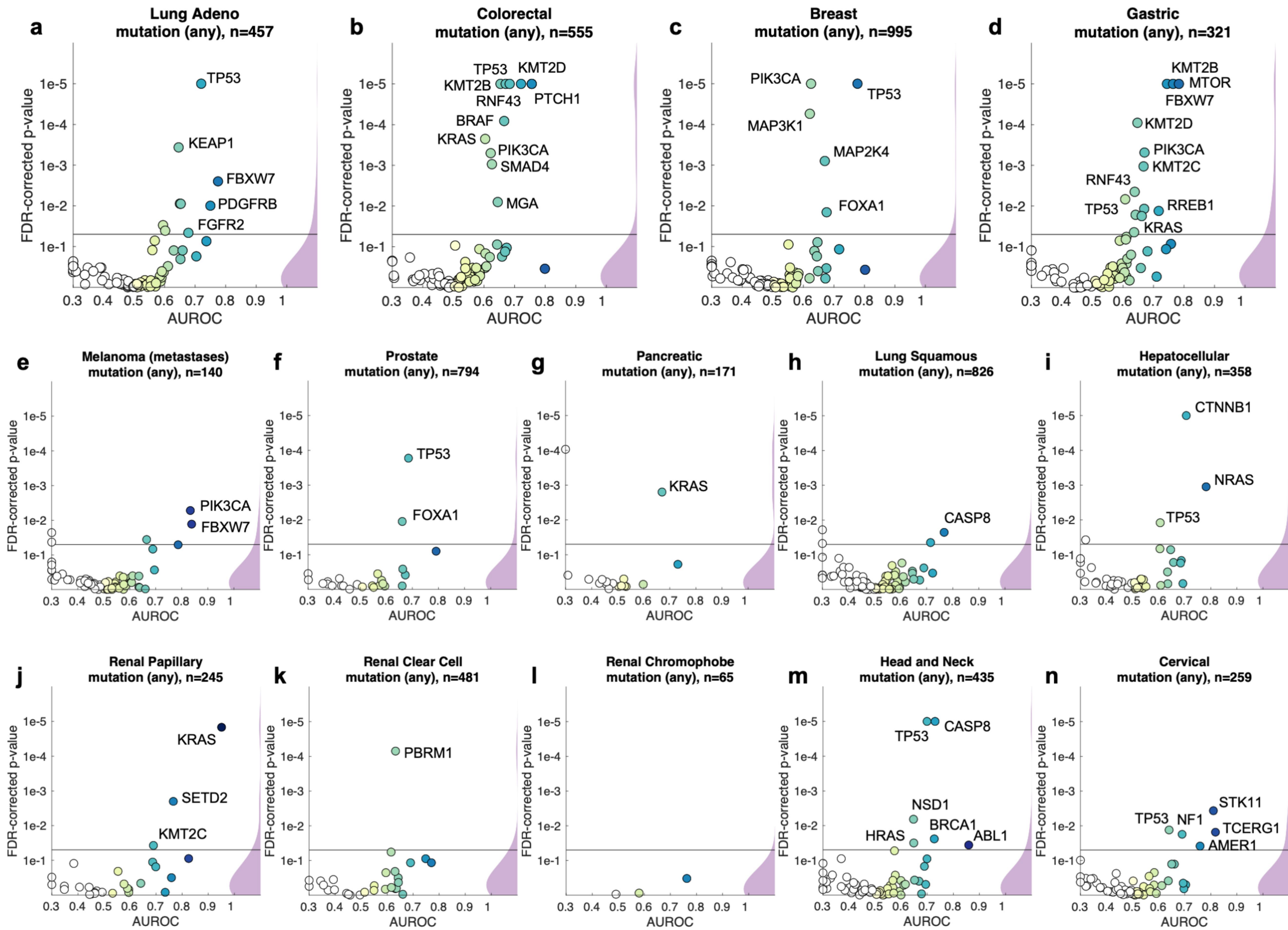
754 **Supplementary Table Legends**

755 **Suppl. Table 1:** All investigated molecular labels.

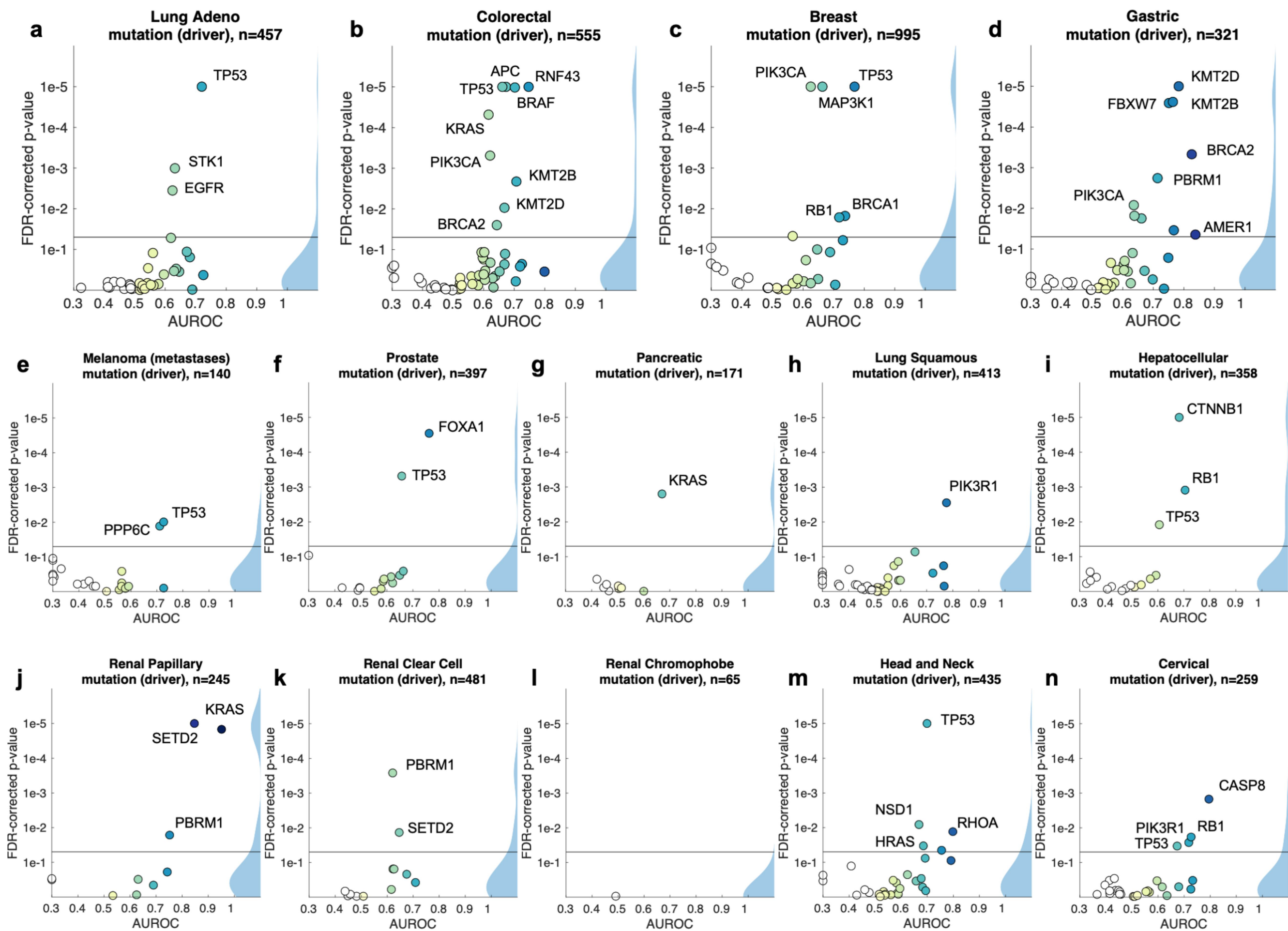
756 **Suppl. Table 2:** Slide numbers and case numbers for each cohort (diagnostic slides, TCGA). For
757 melanoma (TCGA-SKCM), the total number of patients included in the analysis was N=430, of
758 which N=290 had a tissue slide of the primary tumor available and N=141 had a tissue slide of
759 metastatic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

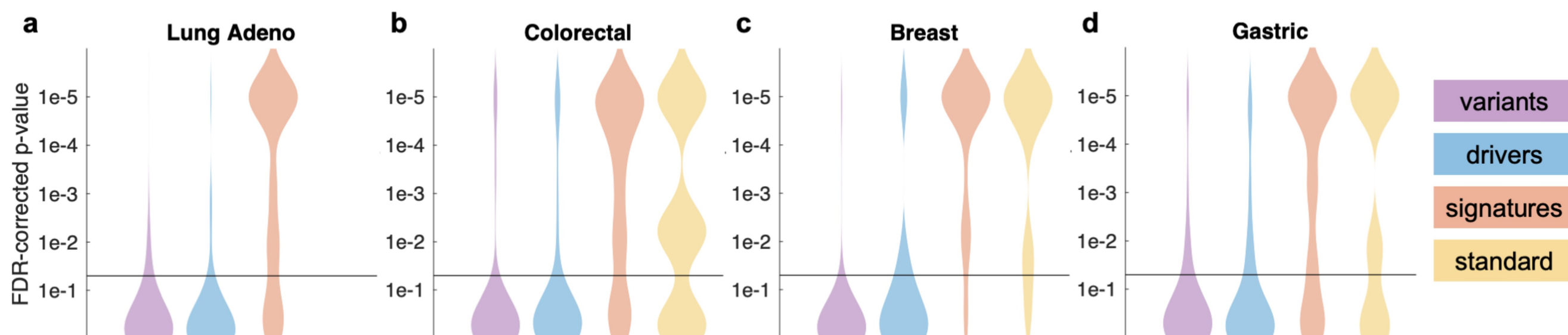
760 **Suppl. Table 3:** All raw values for prediction experiments, alternative methods and hyperparam-
761 eter optimization experiments.

mutations (all variants)

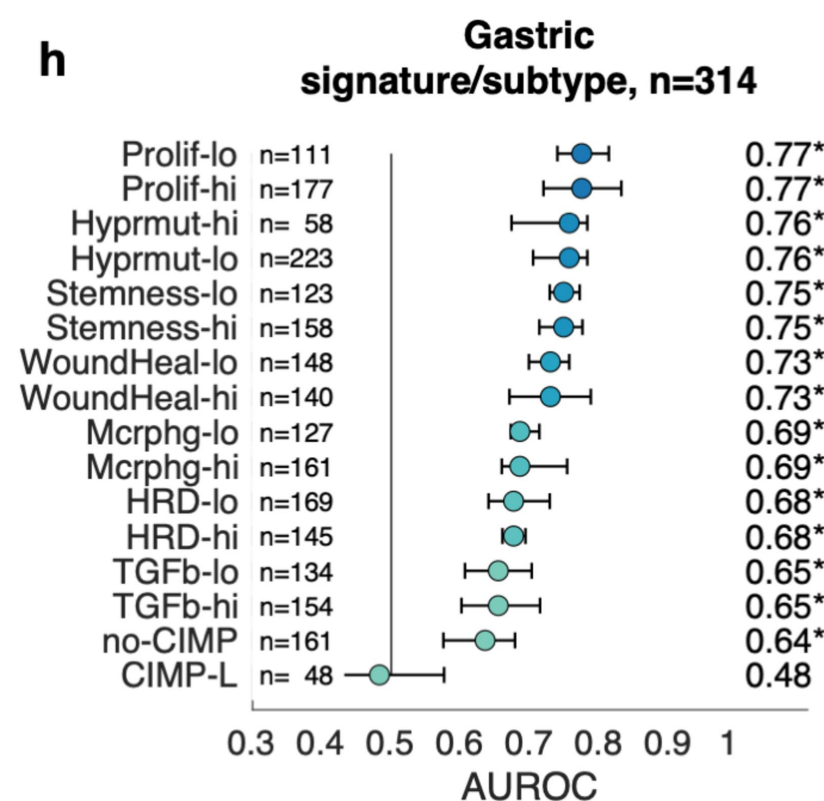
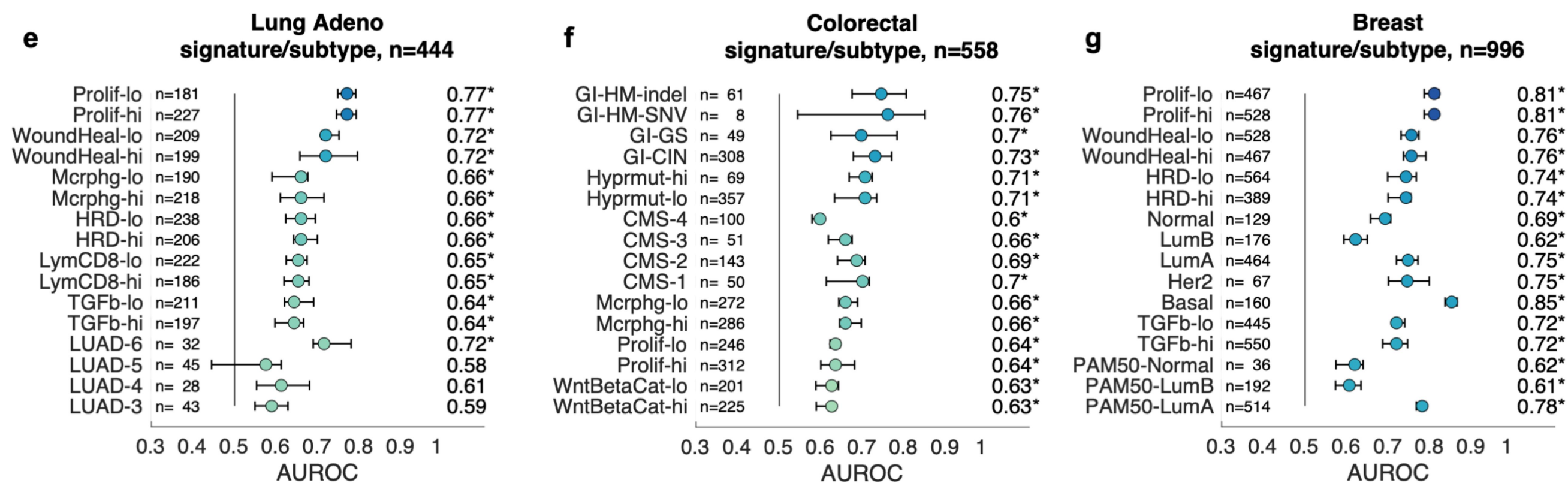


mutations (putative drivers)





molecular subtypes and gene expression signatures



standard biomarkers

