Published in final edited form as:

Nat Med. 2019 July ; 25(7): 1054–1056. doi:10.1038/s41591-019-0462-y.

Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer

Jakob Nikolas Kather[1,2,3,4,5,*], **Alexander T. Pearson**[4], **Niels Halama**[2,5,6], **Dirk Jäger**[2,3,5], **Jeremias Krause**[1], **Sven H. Loosen**[1], **Alexander Marx**[7], **Peter Boor**[8], **Frank Tacke**[9], **Ulf Peter Neumann**[10], **Heike I. Grabsch**[11,12], **Takaki Yoshikawa**[13,14], **Hermann Brenner**[2,15,16], **Jenny Chang-Claude**[17,18], **Michael Hoffmeister**[15], **Christian Trautwein**[1], **Tom Luedde**[1,*]

[1]Department of Medicine III, University Hospital RWTH Aachen, Aachen, Germany [2]German Cancer Consortium (DKTK), German Cancer Research Center (DKFZ), Heidelberg, Germany [3]Applied Tumor Immunity, German Cancer Research Center (DKFZ), Heidelberg, Germany [4]Hematology/Oncology, Department of Medicine, University of Chicago, Chicago, IL, USA [5]Medical Oncology, National Center for Tumor Diseases (NCT), Heidelberg, Germany [6]Translational Immunotherapy, German Cancer Research Center (DKFZ), Heidelberg, Germany [7]Institute of Pathology, University Medical Center Mannheim, Heidelberg University, Mannheim, Germany [8]Institute of Pathology and Department of Nephrology, University Hospital RWTH Aachen, Aachen, Germany [9]Hepatology and Gastroenterology, Charité University Medicine, Berlin, Germany [10]Visceral and Transplant Surgery, University Hospital RWTH Aachen, Aachen, Germany [11]Pathology & Data Analytics, Leeds Institute of Medical Research at St James's, University of Leeds, Leeds, UK [12]Pathology and GROW School for Oncology and Developmental Biology, Maastricht University Medical Center+, Maastricht, the Netherlands [13]Department of Gastrointestinal Surgery, Kanagawa Cancer Center, Yokohama, Japan [14]Department of Gastric Surgery, National Cancer Center Hospital, Tokyo, Japan [15]Division of Clinical Epidemiology and Aging Research, German Cancer Research Center (DKFZ), Heidelberg, Germany [16]Division of Preventive Oncology, German Cancer Research Center (DKFZ) and National Center for Tumor Diseases (NCT), Heidelberg, Germany [17]Division of Cancer Epidemiology, German Cancer

[*]**Correspondence and requests for materials** should be addressed to J.N.K. or T.L. jkather@ukaachen.de; tluedde@ukaachen.de.

Author contributions

J.N.K., A.T.P. and T.L. designed the study; J.N.K. and J.K. performed the analysis; J.N.K., S.H.L. and T.L. performed the statistical analyses; N.H., D.J., A.M., H.I.G., T.Y., H.B., J.C.-C. and M.H. provided human tissue material; D.J., C.T., F.T., U.P.N. and T.L. supervised the study; A.M., P.B. and H.I.G. contributed histopathology expertise; all authors contributed to the interpretation of data and to the writing and revision of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41591-019-0462-y.

Supplementary information is available for this paper at https://doi.org/10.1038/s41591-019-0462-y.

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

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of code and data availability and associated accession codes are available at https://doi.org/10.1038/s41591-019-0462-y.

Research Center (DKFZ), Heidelberg, Germany [18]Cancer Epidemiology Group, University Cancer Center Hamburg, University Medical Center Hamburg-Eppendorf, Hamburg, Germany

Abstract

Microsatellite instability determines whether patients with gastrointestinal cancer respond exceptionally well to immunotherapy. However, in clinical practice, not every patient is tested for MSI, because this requires additional genetic or immunohistochemical tests. Here we show that deep residual learning can predict MSI directly from H&E histology, which is ubiquitously available. This approach has the potential to provide immunotherapy to a much broader subset of patients with gastrointestinal cancer.

Although immunotherapy now represents a cornerstone of cancer therapy, patients with gastrointestinal cancer usually do not benefit to the same extent as patients with other solid malignancies, such as melanoma or lung cancer[1], unless the tumor belongs to the group of microsatellite instable (MSI) tumors[2]. In this group, which accounts for approximately 15% of gastric (stomach) adenocarcinoma (STAD) and colorectal cancer (CRC)[3], immune checkpoint inhibitors demonstrated considerable clinical benefit[4], resulting in recent approval by the Food and Drug Administration (FDA). MSI can be identified by immunohistochemistry or genetic analyses[5], but not all patients are screened for MSI except in high-volume tertiary care centers[6]. Accordingly, a substantial group of potential responders to immunotherapy may not be offered timely treatment with immune checkpoint inhibitors, missing chances of disease control.

Deep learning has outperformed humans in some medical data analysis tasks[7] and can predict patient survival and mutations in tumors using images of lung[8], prostate[9] and brain[10,11] tumors. To facilitate universal MSI screening, we investigated whether deep learning can predict MSI status directly from H&E-stained histology slides. First, we compared five convolutional neural networks on a three-class set of gastrointestinal cancer tissues (n = 94 slides, n = 81 patients, Fig. 1a–c, Extended Data Fig. 1). Resnet18, a residual learning[12] convolutional neural network, was an efficient tumor detector with an out-of-sample area under the curve (AUC) > 0.99, which represented an improvement on the current state of the art[13,14]. Another resnet18 (Fig. 1d) was trained to classify MSI versus microsatellite stability (MSS, Fig. 1e) in large patient cohorts from The Cancer Genome Atlas (TCGA): n = 315 formalin-fixed paraffin-embedded (FFPE) samples of STAD[15] (TCGA-STAD), n = 360 FFPE samples of CRC[16] (TCGA-CRC-DX) and n = 378 snap-frozen samples of CRC (TCGA-CRC-KR; Supplementary Table 1).

Tumor tissue was automatically detected and subsequently tessellated into 100,570 (TCGA-STAD), 60,894 (TCGA-CRC-KR) and 93,408 (TCGA-CRC-DX) color-normalized tiles, in which the deep learning model scored MSI. In the TCGA-CRC-DX test cohort, true MSI image tiles (as defined in Supplementary Table 2) had a median MSI score of 0.61 (95% confidence interval (CI), 0.12–0.82; Fig. 2a), whereas true MSS tiles had an MSI score of 0.29 (95% CI, 0.08–0.57; two-tailed t-test $P = 1.1 \times 10^{-6}$; Fig. 2b). In the TCGA-CRC-KR test cohort, the MSI score was 0.50 (95% CI, 0.17–0.80) for MSI tiles and 0.22 (95% CI,

0.06–0.60; $P = 7.3 \times 10^{-11}$) for MSS tiles, indicating that our approach can robustly distinguish features that are predictive of MSI in both snap-frozen and FFPE samples. Patient-level AUCs for MSI detection were 0.81 (95% confidence interval, 0.69–0.90) in TCGA-STAD, 0.84 (95% CI, 0.73–0.91) in TCGA-CRC-KR and 0.77 (95% CI, 0.62–0.87) in TCGA-CRC-DX (Extended Data Fig. 2a; MSI frequency is listed in Supplementary Table 3).

The multi-center DACHS study[17,18] was used as an external validation set ($n = 378$ patients). Using the automatic tumor detector and the MSI detector trained on TCGA-CRC-DX (Fig. 2c), the patient-level AUC was 0.84 (95% CI, 0.72–0.92) (Fig. 2d). The model that was trained on FFPE samples and used on FFPE samples was superior to a model that was trained on frozen samples and used on FFPE samples. Similarly, a model that was trained on CRC samples and used on CRC samples performed better than a model that was trained on STAD samples and used on CRC samples (Extended Data Fig. 2a). To analyze the limits of our proposed method, we validated the MSI detector on $n = 185$ patients with gastric cancer from Yokohama, Japan (KCCH cohort)[19]. Gastric cancer in Asian individuals has a very different histology and clinical course than gastric cancer in non-Asian individuals[20]. A classifier trained on TCGA-STAD (approximately 80% non-Asian) achieved an AUC of 0.69 (95% CI, 0.52–0.82) in the KCCH cohort (0% non-Asian; Extended Data Fig. 2a). Because MSI is a pan-tumor biomarker with clinical usefulness beyond gastrointestinal cancer, we additionally trained and tested our method in samples of endometrial cancer (UCEC[21], $n = 327$ patients), which has a high prevalence of MSI[3], yielding an AUC for MSI detection in held-out patients of 0.75 (95% CI, 0.63–0.83; Extended Data Fig. 2a).

Although our method attained robust performance across a range of human tumors and exceeded the previously reported performance of predicting molecular features from histology[8,9], our experiments point to some limitations. The ability to classify does not necessarily extend beyond the cancer type and ethnicity present in the training set. Larger training cohorts are likely to boost classification performance because rare morphological variants can be learned by the network. Another limitation is the required tissue size. To define its lower limit, we generated ‘virtual biopsies’ and found that performance plateaued at approximately 100 tiles of 256 μm edge length, suggesting that biopsies are sufficient for MSI prediction (Extended Data Fig. 2b,c).

To reverse-engineer the black-box MSI detector, we correlated MSIness (the fraction of MSI-predicted tiles) to transcriptomic and immunohistochemical data across our test sets. MSIness was correlated to a lymphocyte gene expression signature in gastric cancer and to PD-L1 expression and an interferon-γ (IFNγ) signature in CRC (Fig. 2e, Supplementary Table 4). Spatially, predicted MSI overlapped with poorly differentiated and lymphocyte-rich tumor regions (Extended Data Fig. 3), which is consistent with histopathological knowledge. MSI is both a prognostic and predictive biomarker[22,23] and correspondingly, in patients with MSS tumors in the DACHS cohort, high MSIness defined a group with worse overall survival (univariable Cox hazard ratio, 1.65 (95% confidence interval, 1.00–2.73), log-rank test, $P = 0.0207$, multivariable models in Supplementary Table 5). Although this was not statistically significant in a four-variable model (hazard ratio, 1.37 (95% confidence

interval, 0.88–2.14); Supplementary Table 5), future clinical trials could determine the response to cancer immunotherapy in these patients with MSI-like tumors.

Cancer immunotherapy has changed the landscape of oncology but identifying patients who will benefit from immunotherapy has remained a key challenge. Recently, the American Society of Clinical Oncology has declared discovery of new biomarkers for immunotherapy as the top priority in cancer research in 2019 (https://www.asco.org/research-progress/reports-studies/clinical-canceradvances-2019/clinical-cancer-advances-2019-glance). However, even established biomarkers such as MSI are not universally tested today. Our method can be implemented at tertiary care centers at a low cost (Extended Data Fig. 4a,b). It does not require additional laboratory tissue testing and can infer MSI status from ubiquitously existing data. After training on larger datasets and prospective validation, this could ultimately enable efficient identification of patients with MSI tumors, enabling the distribution of the benefit of cancer immunotherapy to a broader target population.

Methods

Ethics statement

All experiments were conducted in accordance with the Declaration of Helsinki and the International Ethical Guidelines for Biomedical Research Involving Human Subjects. Anonymized archival tissue samples were retrieved from the tissue bank of the National Center for Tumor diseases (NCT; including samples from the DACHS trial[17,18]) and from the pathology archive at the University Medical Center Mannheim, Heidelberg University (UMM) after approval by the institutional ethics boards as described previously[13]. Clinical data for all cohorts are listed in Supplementary Table 1.

Tumor detection, MSI detection and patient cohorts

To train an automatic tumor detector for histological images of gastrointestinal cancer, we used histological specimens of CRC and stomach cancer surgical specimens from the UMM and NCT tissue banks. This cohort was described previously and encompassed $n = 94$ whole-slide images from $n = 81$ patients[13]. Regions in these images were manually annotated and classified as tumor and two types of nontumor tissue (dense and loose tissue, representing muscle and/or stroma and fat and/or mucus, respectively), yielding 11,977 unique image tiles of 256 μm edge length. All of these images are freely available for download at https://doi.org/10.5281/zenodo.2530789. Image preprocessing was performed as previously described[13], including color normalization. For color normalization, we used the Macenko method, which converts all images to a reference color space as described previously[13,14,24].

We retrieved histology images of $n = 315$ patients with STAD (diagnostic slides, FFPE tissue), $n = 387$ patients with CRC (CRC-KR; cryosections, snap-frozen tissue), $n = 360$ patients with CRC (CRC-DX; diagnostic slides, FFPE tissue) and $n = 492$ patients with UCEC (diagnostic slides, FFPE tissue) from TCGA[16]. All slides contained tumor tissue (after manual review in a blinded manner) and had the resolution available as part of the metadata (μm per pixel). During training, 99 (STAD), 109 (CRC-KR), 100 (CRC-DX) and

110 (UCEC) randomly selected patients were held out and were used as a test set. In all cases, training and test sets were split on a patient level and no image tiles from test patients were present in any training set. A more extensive description of these datasets and all image files are freely available for download under an open source license at https://doi.org/10.5281/zenodo.2530835 and https://doi.org/10.5281/zenodo.2532612. All TCGA images can be downloaded from public repositories at the National Institutes of Health (https://portal.gdc.cancer.gov/).

For TCGA-CRC and TCGA-STAD, all patients who were previously defined as MSI-H[25] were included in the MSI group. All patients with unknown MSI status but with a mutation count of >1,000 (as defined previously[26]) were also included in the MSI group (this was the case for fewer than 10 patients in any cohort). Supplementary Table 2 lists the methods that were used to determine MSI in all cohorts. In the TCGA cohorts, patients with less than 10 image tiles per slide were not used for prediction. As an external validation cohort for CRC, we used $n = 378$ patients from the population-based DACHS study, a case–control study of CRC in the southwest of Germany with long-term follow-up of patients enrolled at more than 20 clinics in the study region. In addition, we analyzed data of $n = 185$ patients from KCCH as described previously[19]. Additional information about the cohorts is shown in Supplementary Tables 1–3.

Neural network models, tumor detection and MSI detection

For tumor detection in gastrointestinal cancer, we trained a convolutional neural network with deep residual learning (resnet18)[12] model to classify tumor versus normal tissue by transfer learning. In TCGA-STAD, TCGA-CRC-KR, TCGA-CRC-DX and DACHS, the automatic gastrointestinal tumor detector was used whereas in TCGA-UCEC and KCCH, tumor regions were delineated by a pathologist. For MSI detection, we trained another resnet18 model for each tumor type. We chose resnet18, because our initial experiments showed that among five popular neural network models[12,27–30] (Extended Data Fig. 1) that we compared on our tumor detection dataset, resnet18 had a short training time, excellent classification performance and fewer parameters than similarly performing models (alexnet, vgg19), reducing the risk of overfitting.

The number of image tiles per class was equalized by undersampling. Training was stopped if the validation accuracy in a held-out set of 12.5% of all training tiles did not increase for three successive validation checks (checked every 256 iterations). All convolutional neural networks were pretrained on the ImageNet (www.image-net.org) database as described previously[13]. Only the weights in the last 10 layers were trainable whereas all other weights were frozen. We used the Adam algorithm for training, counteracted overfitting by an L2-regularization of 1×10^{-4} and used a fixed learning rate of 1×10^{-6} for TCGA-STAD, TCGA-CRC-DX and TCGA-CRC-KR, and 1×10^{-4} for TCGA-UCEC. DACHS and KCCH were only used for prediction and not for training. All codes were implemented in MATLAB R2018a and run on desktop workstations with Nvidia graphics-processing units (GPUs; Titan Xp, Quadro P6000, Titan RTX). Performance was scored as AUC in a receiver operating characteristic analysis as in previous studies[8,9]. AUC values are given as median with 95% confidence intervals as calculated by 500-fold bootstrapping with the ‘bias

corrected and accelerated percentile method' unless otherwise noted[31]. Our source codes are freely available at https://github.com/jnkather/MSIfromHE and can be applied to any tumor type.

Statistics

Classifier performance was assessed by area under the receiver operating characteristic curve as calculated with 'perfcurve' in MATLAB R2018a. Correlations were calculated with R version 3.5.1 'cor.test' using the 'Pearson' method.

Reporting Summary

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

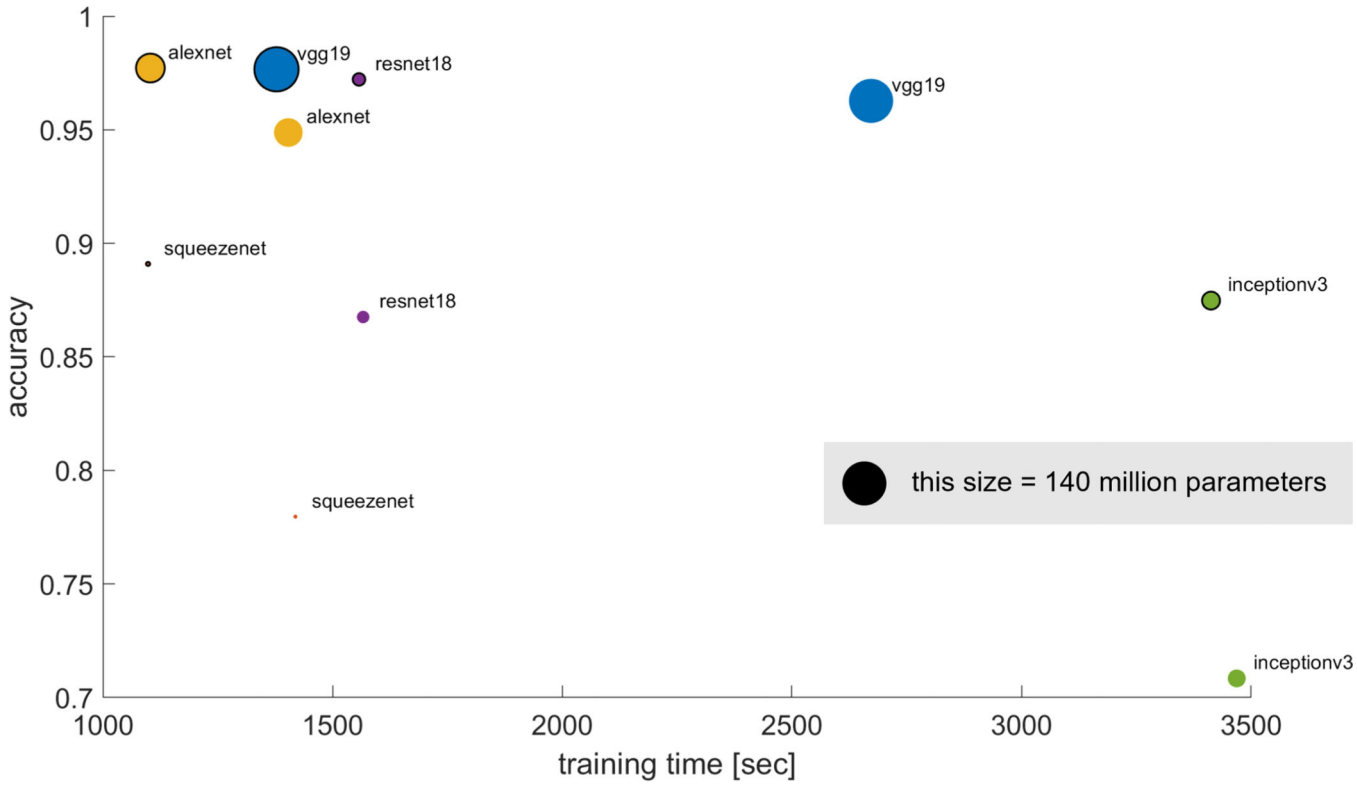
Data availability

All whole-slide images for datasets are available at https://portal.gdc.cancer.gov/. Training images for tumor detection are available at https://doi.org/10.5281/zenodo.2530789. Training images for MSI detection are available at https://doi.org/10.5281/zenodo.2530835 and https://doi.org/10.5281/zenodo.2532612. Source data for Fig. 1 are available in public repositories at https://doi.org/10.5281/zenodo.2530789, https://doi.org/10.5281/zenodo.2530835 and https://doi.org/10.5281/zenodo.2532612. Source Data for Figs. 1, 2 and Extended Data Figs. 1, 2 containing the raw data for these figures are available in the online version of the paper.

Code availability

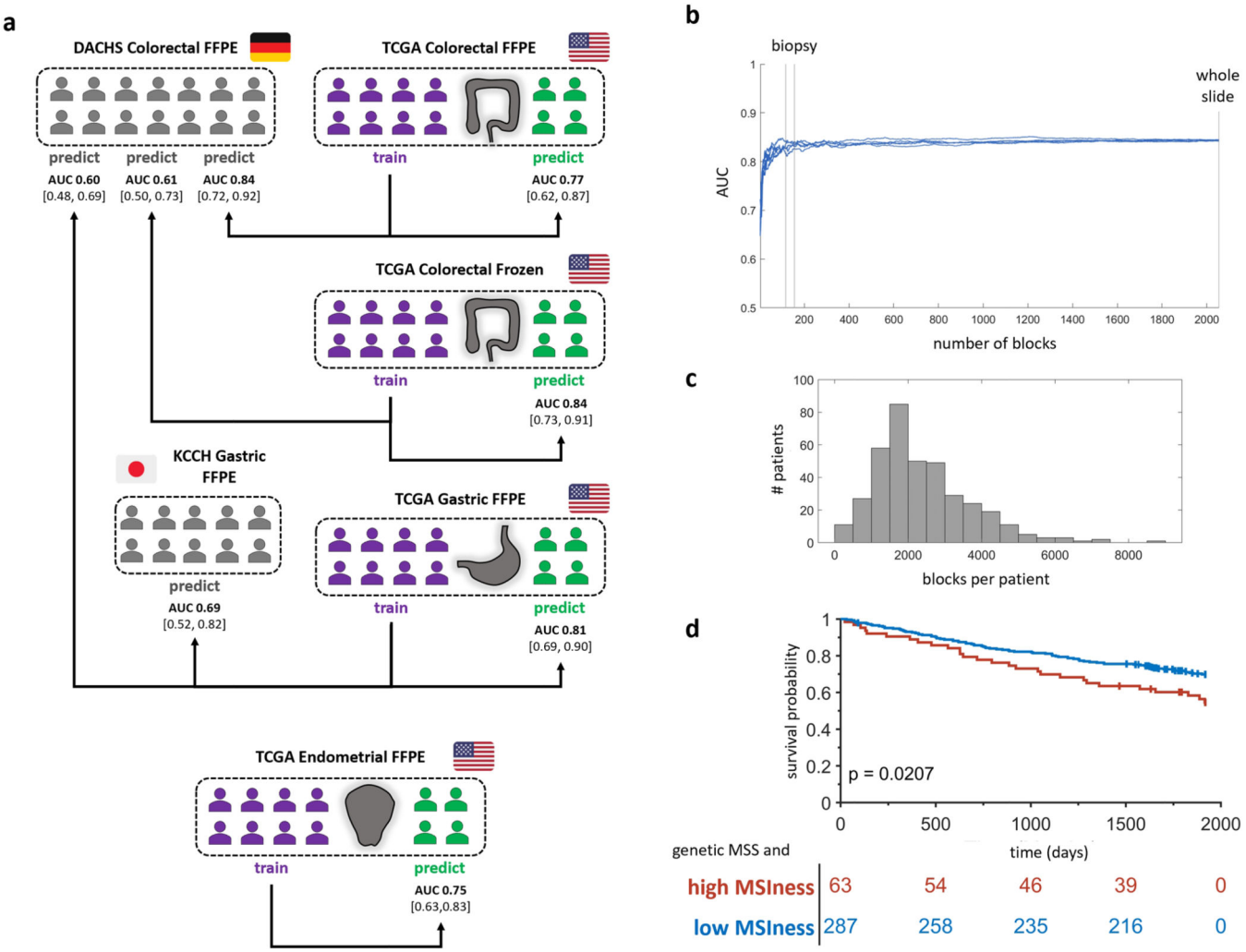
Source codes are available at https://github.com/jnkather/MSIfromHE.

Extended Data



Extended Data Fig. 1 |. Comparison of five deep neural network architectures.

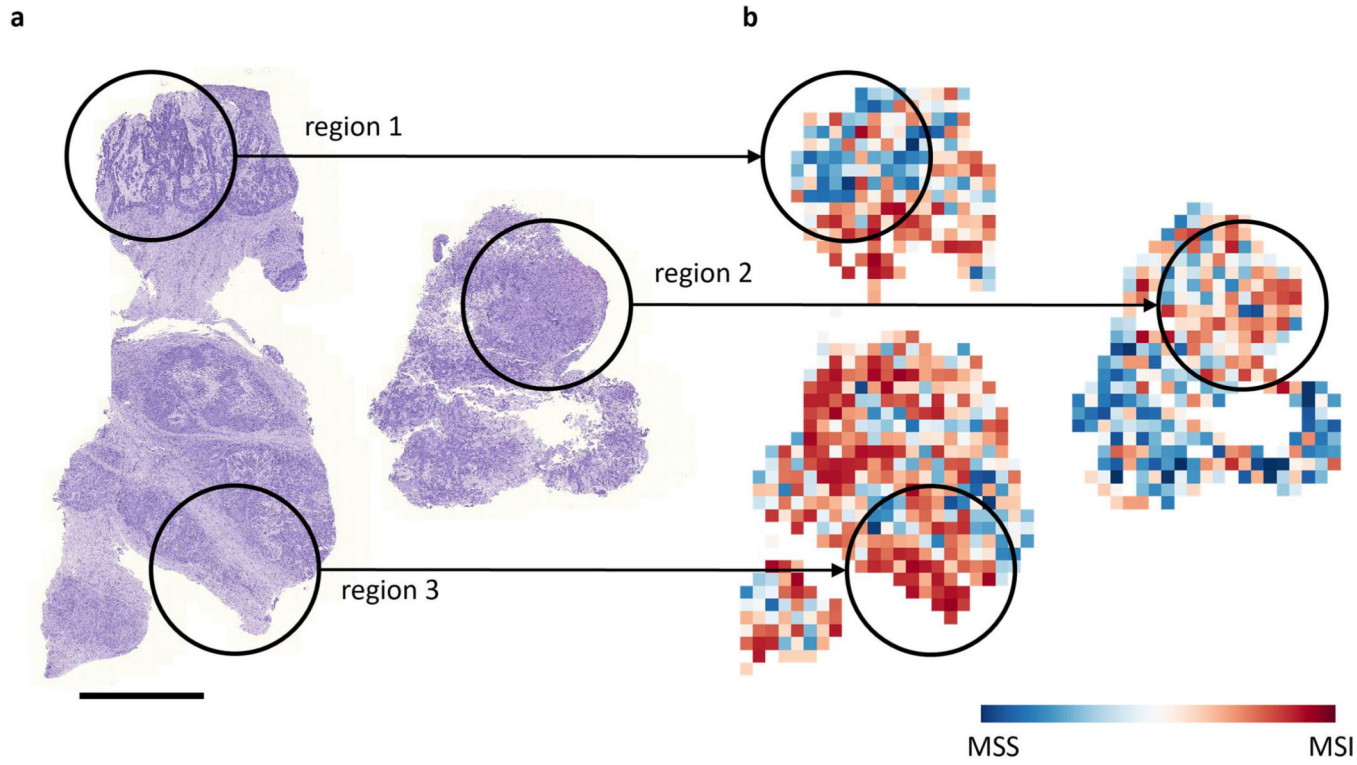
We compared accuracy and training time of five neural network architectures on the tumor detection dataset with three balanced classes. Alexnet[27], VGG19 (ref. [28])) and resnet18 (ref. [12]) achieved >95% accuracy in withheld images, whereas inceptionv3 (ref. [29]) and squeezenet[30] had a poor performance on this benchmark task. Among the well-performing models, resnet18 had the lowest number of parameters, making it potentially more portable and less prone to overfitting. In this comparison, we split the dataset into 70% training, 15% validation and 15% test images. Each network is shown twice in this graph: with a learning rate of 1×10^{-6} and 1×10^{-5} (outlined). Training was run for 25 epochs. Resnet18 was subsequently retrained on the dataset, attaining a median fivefold cross-validated out-of-sample AUC > 0.99 for tumor detection. The dataset was derived from $n = 94$ whole-slide images from $n = 81$ patients and is available at https://doi.org/10.5281/zenodo.2530789.



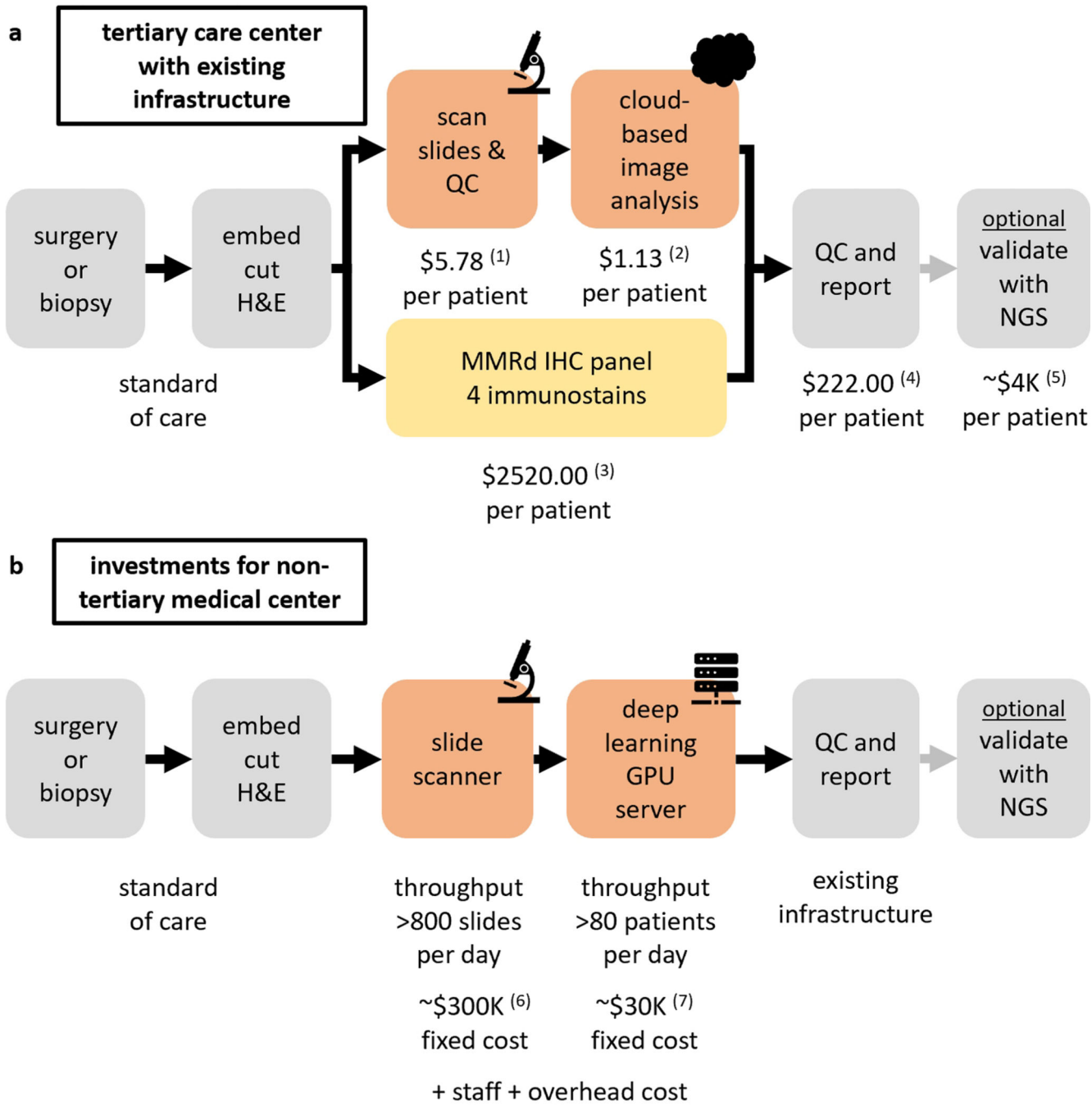
Extended Data Fig. 2 |. Additional data for classifier performance.

a, Flowchart of all experiments. The area under the receiver operating characteristic curve gives an overall measure of patient-level classifier accuracy as measured in held-out test sets. Flag symbols are from https://twemoji.twitter.com/ (licensed under a CC-BY 4.0 license). **b**, Classification performance in virtual biopsies. We predicted MSI status in all patients in the DACHS cohort, varying the number of blocks (tiles) from 3 to 2,054, which was the median number of blocks per whole-slide image This experiment was repeated five times with different randomly picked blocks being used. As one block has an edge length of 256 μm, a 1-cm tissue cylinder with 100% tumor tissue from a standard 18G biopsy needle corresponds to 117 blocks and a 16G needle corresponds to 156 blocks. In clinical routine, usually only a part of each biopsy core contains tumor, but multiple biopsy cores are collected. With increasing tissue size, performance stabilizes at AUC = 0.84. This shows that a typical biopsy would be sufficient for MSI prediction. CI, confidence interval. **c**, Distribution of the numbers of blocks for all patients in DACHS (n = 378 patients). **d**, Overall survival of patients with genetic MSS tumors stratified by high or low predicted MSIness. In this group, patients with high MSIness had a shorter survival than patients with

low MSIness. The table shows the number of patients at risk. The P value was calculated by two-sided log-rank test (n = 350 patients).



Extended Data Fig. 3 |. Morphological correlates of intratumor heterogeneity of MSI. **a**, Histological image of a test set patient who was genetically determined as MSI. **b**, Corresponding predicted MSI map for the image shown in **a**. Three regions are highlighted. Region 1 is a glandular region with necrosis and extracellular mucus; this region was predominantly predicted to be MSS. Region 2 is a solid, dedifferentiated region, which was predicted to be MSI. Region 3 contained mostly budding tumor cells mixed with immune cells, this region was strongly predicted to be MSI. Together, these representative examples show that different morphologies elicit different predictions and that these predictions can be traced back to patterns that are understandable for humans. Scale bar, 2.5 mm. This figure is representative of $n = 378$ patients in the DACHS cohort.



Extended Data Fig. 4 |. Estimated cost for MSI screening with deep learning.

a, Workflow for MSI screening with deep learning versus immunohistochemistry in tertiary care centers with existing digital pathology core facilities such as the University of Chicago Medical Center. Costs differ by country and are usually cheaper in Europe than in the United States. Here, we list the costs that apply in the United States. **b**, Set-up cost (fixed cost) for a digital pathology and deep learning infrastructure. H&E, hematoxylin and eosin; MMRd, mismatch repair deficiency; NGS, next-generation sequencing; QC, quality control. Sources and assumptions were as follows. (1) Prices were obtained from https://htrc.uchicago.edu/

fees.php?fee=2&fee=2, retrieved on 11 March 2019. We assume ×20 magnification on a high-volume whole-slide scanner. (2) Prices were obtained from https://techcrunch.com/2019/03/07/scaleway-releases-cloud-gpu-instances-for-e1-per-hour/ and https://www.scaleway.com/, retrieved on 11 March 2019. We assume that 1 h of GPU computing on a Nvidia Tesla P100 GPU is required to process whole-slide images for one patient to prediction. (3) US Current Procedural Terminology (CPT) code 88342, four-antibody panel at US$852.00 per staining. (4) Personal communication by the Pathology Department, University of Chicago Medicine, March 2019. (5) Personal communication, Medical Oncology, National Center for Tumor Diseases, Germany. (6) Personal experience of cost for a high-throughput slide scanner plus a limited storing capacity, based on offers by multiple digital pathology vendors. (7) Assuming a tower server with one NVidia Tesla V100 GPU or similar GPU, based on multiple offers by providers for professional hardware, March 2019. Staff cost and infrastructure cost are not accounted for in this schematic.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

The results are in part based on data generated by the TCGA Research Network (http://cancergenome.nih.gov/). J.N.K. was funded by RWTH University Aachen (START 2018-691906). A.T.P. was funded by NIH/NIDCR (K08-DE026500). A.M. was funded by the German Federal Ministry of Education and Research (BMBF) (M2oBITE/13GW0091E). The DACHS study was funded by the Interdisciplinary Research Program of the National Center for Tumor Diseases (NCT), Germany and German Research Council DFG (BR 1704/6-1, BR 1704/6-3, BR 1704/6-4, BR 1704/17-1, CH 117/1-1 and HO 5117/2-1), BMBF (01ER0814, 01ER0815, 01ER1505A and 01ER1505B). P.B. was funded by the DFG (BO 3755/6-1, SFB-TRR57 and SFB-TRR219). T.L. was funded by Horizon 2020 through the European Research Council (ERC) Consolidator Grant PhaseControl (771083), Mildred-Scheel-Endowed Professorship from the German Cancer Aid, DFG (SFB-TRR57/P06 and LU 1360/3-1), Ernst-Jung-Foundation Hamburg and IZKF (Interdisciplinary Center of Clinical Research) at RWTH Aachen.

References

1. Darvin P, Toor SM, Sasidharan Nair V & Elkord E Exp. Mol. Med. 50, 165 (2018).
2. Le DT et al. N. Engl. J. Med. 372, 2509–2520 (2015). [PubMed: 26028255]
3. Bonneville R et al. JCO Precis. Oncol. 2017, 1–15 (2017).
4. Le DT et al. Science 357, 409–413 (2017). [PubMed: 28596308]
5. Kather JN, Halama N & Jaeger D Semin. Cancer Biol. 52, 189–197 (2018). [PubMed: 29501787]
6. Franke AJ et al. J. Clin. Oncol. 36, 796 (2018).
7. Norgeot B, Glicksberg BS & Butte AJ Nat. Med. 25, 14–15 (2019). [PubMed: 30617337]
8. Coudray N et al. Nat. Med. 24, 1559–1567 (2018). [PubMed: 30224757]
9. Schaumberg AJ, Rubin MA & Fuchs TJ Preprint at https://www.biorxiv.org/content/10.1101/064279v9 (2018).
10. Chang P et al. AJNR Am. J. Neuroradiol. 39, 1201–1207 (2018). [PubMed: 29748206]
11. Mobadersany P et al. Proc. Natl Acad. Sci. USA 115, E2970–E2979 (2018). [PubMed: 29531073]
12. He K, Zhang X, Ren S & Sun J In Proc. IEEE Conference on Computer Vision and Pattern Recognition 770–778 (2016).
13. Kather JN et al. PLoS Med. 16, e1002730 (2019). [PubMed: 30677016]
14. Kather JN et al. Sci. Rep. 6, 27988 (2016). [PubMed: 27306927]
15. The Cancer Genome Atlas Network Nature 513, 202–209 (2014).
16. The Cancer Genome Atlas Network Nature 487, 330–337 (2012).

17. Hoffmeister M et al. J. Natl Cancer Inst. 107, djv045 (2015). [PubMed: 25770147]
18. Brenner H, Chang-Claude J, Seiler CM & Hoffmeister MJ Clin. Oncol. 29, 3761–3767 (2011).
19. Aoyama T et al. Cancer Med. 7, 4914–4923 (2018). [PubMed: 30160049]
20. Rahman R, Asombang AW & Ibdah JA World J. Gastroenterol. 20, 4483–4490 (2014). [PubMed: 24782601]
21. Levine DA & The Cancer Genome Atlas Research Network. Nature 497, 67–73 (2013). [PubMed: 23636398]
22. Kawakami H, Zaanan A & Sinicrope FA Curr. Treat. Options Oncol. 16, 30 (2015). [PubMed: 26031544]
23. Zhu L et al. Mol. Clin. Oncol. 3, 699–705 (2015). [PubMed: 26137290]

References

24. Macenko M et al. In Proc. IEEE International Symposium on Biomedical Imaging 1107–1110 (2009).
25. Liu Y et al. Cancer Cell 33, 721–735 (2018). [PubMed: 29622466]
26. Bailey MH et al. Cell 173, 371–385 (2018). [PubMed: 29625053]
27. Krizhevsky A, Sutskever I & Hinton GE in Proc. Advances in Neural Information Processing Systems 1097–1105 (2012).
28. Simonyan K & Zisserman A Preprint at https://arxiv.org/abs/1409.1556 (2014).
29. Szegedy C, Vanhoucke V, Ioffe S, Shlens J & Wojna Z in Proc. IEEE Conference on Computer Vision and Pattern Recognition 2818–2826 (2016).
30. Iandola FNet al. Preprint at https://arxiv.org/abs/1602.07360 (2016).
31. DiCiccio TJ & Efron B Stat. Sci. 11, 189–228 (1996).

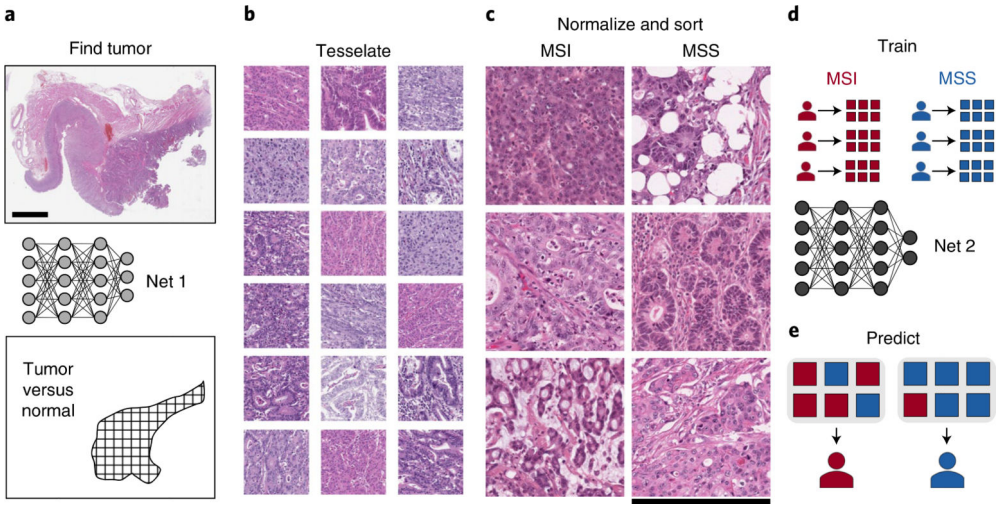


Fig. 1 |. Tumor detection and MSI prediction in H&E histology.

a, A convolutional neural network was trained as a tumor detector for STAD and CRC. Scale bar, 4 mm. **b**,**c**, Tumor regions were cut into square tiles (**b**), which were color-normalized and sorted into MSI and MSS (**c**). Scale bar, 256 μm. **d**, Another network was trained to classify MSI versus MSS. **e**, This automatic pipeline was applied to held-out patient sets.

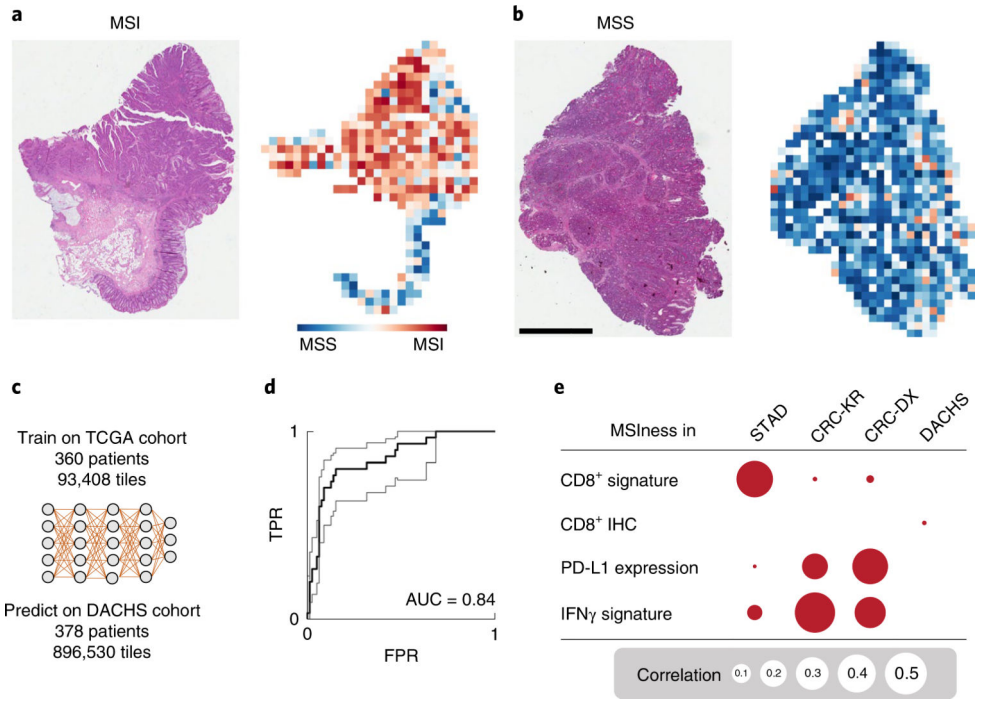


Fig. 2 |. Classification performance in an external validation set.

a,**b**, Tissue slides of patients with MSI and MSS tumors in the TCGA-CRC-DX test set show the spatial patterns of predicted MSI score (Extended Data Fig. 4). These images are representative of $n = 378$ patients. **c**, A network was trained on the TCGA-CRC-DX training cohort ($n = 260$ patients) and deployed on the DACHS cohort ($n = 378$ patients). **d**, Patient-level receiver operating characteristic curve with bootstrapped 95% CI in DACHS ($n = 378$ patients). FPR, false-positive rate (1 − specificity); TPR, true-positive rate (sensitivity). **e**, Pearson correlation of predicted MSIness to transcriptomic and immunohistochemical (IHC) data across test sets. *P* values are listed in Supplementary Table 4. Sample sizes per cohort are: TCGA-STAD $n = 91$, TCGA-CRC-KR $n = 105$, TCGA-CRC-DX $n = 95$, DACHS $n = 134$ patients. No adjustments for multiple comparisons were made, and all statistical tests were two-sided.