Article:
Kather, JN, Pearson, AT, Halama, N et al. (14 more authors) (2019) Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. Nature Medicine, 25 (7). pp. 1054-1056. ISSN 1078-8956

https://doi.org/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[1,9], 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, Illinois, 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 Research Center (DKFZ), Heidelberg, Germany

[18] Cancer Epidemiology Group, University Cancer Center Hamburg, University Medical Center Hamburg-Eppendorf, Hamburg, Germany

*correspondence should be addressed to J.N.K. and T.L.: jkather@ukaachen.de and tluedde@ukaachen.de

Microsatellite instability (MSI) determines whether patients with gastrointestinal (GI) cancer respond exceptionally well to immunotherapy. In clinical practice however, 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 hematoxylin-eosin histology, which is ubiquitously available. This approach has the potential to provide immunotherapy to a much broader subset of GI cancer patients.

While immunotherapy now represents a cornerstone of cancer therapy, patients with gastrointestinal (GI) cancer usually do not benefit to an extent comparable to other solid malignancies such as melanoma or lung cancer[1] unless they belong 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 significant clinical benefit[4], resulting in recent approval by the Food and Drug Administration (FDA). MSI can be identified by immunohistochemistry or genetically[5], but not all patients are screened for MSI except in high-volume tertiary care centers[6]. Accordingly, a significant 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 survival and mutations from images in 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 hematoxylin-eosin (HE) histology slides. First, we compared five convolutional neural networks (CNN) on a three-class set of GI cancer tissues (N=94 slides, N=81 patients, Fig. 1a-c, Extended Data Fig. 1). Resnet18, a residual learning[12] CNN, was an efficient tumor detector with an out-of-sample area under the curve (AUC) of >0.99, which represented an improvement on the current state of the art[13,14]. Another resnet18 (Fig. 1d) was trained to classify microsatellite instability (MSI) versus stability (MSS, Fig. 1e) in large patient cohorts

from "The Cancer Genome Atlas" (TCGA): N=315 stomach adenocarcinoma[15] (formalin-fixed paraffin-embedded [FFPE], TCGA-STAD), N=360 CRC[16] (FFPE, TCGA-CRC-DX) and N=378 CRC patients (snap-frozen, TCGA-CRC-KR; Suppl. 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 Suppl. Table 2) had a median MSI score of 0.61 (95% confidence interval [CI] [0.12, 0.82], Fig. 2a) while true MSS tiles had an MSI score of 0.29 (CI [0.08, 0.57]; two-tailed t-test p-value = 1.1e-6, Fig. 2b). In the TCGA-CRC-KR test cohort, the MSI score for MSI tiles was 0.50 [0.17, 0.80] and 0.22 [0.06, 0.60] (p=7.3e-11) for MSS, indicating that our approach can robustly distinguish features predictive of MSI both in snap-frozen and FFPE samples. Patient-level AUC for MSI detection was 0.81 [0.69, 0.90] in TCGA-STAD, 0.84 [0.73, 0.91] in TCGA-CRC-KR and 0.77 [0.62, 0.87] in TCGA-CRC-DX (Extended Data Fig. 2a; MSI frequency is listed in Suppl. 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), patient-level AUC was 0.84 [0.72, 0.92] (Fig. 2d). "Train on FFPE, deploy on FFPE" was superior to "train on frozen, deploy on FFPE" and "train on CRC, deploy on CRC" was better than "train on STAD, deploy on CRC" (Extended Data Fig. 2a). To probe the limits of our proposed method, we validated the MSI detector on N=185 gastric cancer patients from Yokohama, Japan (KCCH cohort[19]). Asian gastric cancer has a very different histology and clinical course than non-Asian gastric cancer[20]. A classifier trained on TCGA-STAD (approximately 80% non-Asian) achieved an AUC of 0.69 [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 GI cancer, we additionally trained and tested our method in uterine cancer (UCEC, N=327 patients), which has a high prevalence of MSI[3], yielding an AUC for MSI detection in held-out patients of 0.75 [0.63, 0.83] (Extended Data Fig. 2a).

While our new 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: First, 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 (IHC) 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-gamma signature in colorectal cancer (Fig. 2e, Suppl. 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 a prognostic in addition to a predictive biomarker[21,22] and correspondingly, in MSS patients of the DACHS cohort, high MSIness defined a group with worse overall survival (univariable Cox hazard ratio [HR] 1.65 [1.00, 2.73], log rank p = 0.0207, multivariable models in Suppl. Table 5). Although this was not statistically significant in a four-variable model (HR 1.37 [0.88 – 2.14], Suppl. Table 5), future clinical trials could determine the response to cancer immunotherapy in these MSI-like patients.

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 (ASCO) 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-cancer-advances-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 wet lab tissue testing and can infer MSI status from ubiquitously existing data. After training on larger data sets and prospective validation, this could ultimately enable efficient identification of MSI patients, allowing to distribute the benefit of cancer immunotherapy to a broader target population.

Online content

Any methods, supplementary data, Nature Research Life Sciences Reporting Summary, source data and source codes and associated accession codes are available online.

Data availability

All whole slide images for data sets are available at https://portal.gdc.cancer.gov/. Training images for tumor detection are available at http://dx.doi.org/10.5281/zenodo.2530789. Training images for MSI detection are available at http://dx.doi.org/10.5281/zenodo.2530835 and http://dx.doi.org/10.5281/zenodo.2532612. Raw data for the figures are available in the online Supplementary Data. Source codes are available at https://github.com/jnkather/MSIfromHE.

References

1. Darvin, P., Toor, S.M., Sasidharan Nair, V. & Elkord, E. *Exp Mol Med* **50**, 165 (2018).
2. Le, D.T.*, et al. N Engl J Med* **372**, 2509-2520 (2015).
3. Bonneville, R.*, et al. JCO Precis Oncol* **2017**, 1-15 (2017).
4. Le, D.T.*, et al. Science* **357**, 409-413 (2017).
5. Kather, J.N., Halama, N. & Jaeger, D. *Semin Cancer Biol* **52**, 189-197 (2018).
6. Franke, A.J.*, et al. J Clin Oncol* **36**, 796-796 (2018).
7. Norgeot, B., Glicksberg, B.S. & Butte, A.J. *Nat Med* **25**, 14-15 (2019).
8. Coudray, N.*, et al. Nat Med* **24**, 1559-1567 (2018).
9. Schaumberg, A.J., Rubin, M.A. & Fuchs, T.J. *bioRxiv*, 064279 (2018).
10. Chang, P.*, et al. AJNR Am J Neuroradiol* **39**, 1201-1207 (2018).
11. Mobadersany, P.*, et al. Proc Natl Acad Sci USA*. **115**, E2970-E2979 (2018).
12. He, K., Zhang, X., Ren, S. & Sun, J. in *Proceedings of the IEEE conference on computer vision and pattern recognition* 770-778 (2016).

13. Kather, J.N.*, et al. PLoS Medicine* **16**, e1002730 (2019).
14. Kather, J.N.*, et al. Sci Rep* **6**, 27988 (2016).
15. The Cancer Genome Atlas Network*, et al. Nature* **513**, 202 (2014).
16. The Cancer Genome Atlas Network*, et al. Nature* **487**, 330 (2012).
17. Hoffmeister, M.*, et al. J Natl Cancer Inst* **107**, djv045 (2015).
18. Brenner, H., Chang-Claude, J., Seiler, C.M. & Hoffmeister, M. *J Clin Oncol* **29**, 3761-3767 (2011).
19. Aoyama, T.*, et al. Cancer Med* **7**, 4914-4923 (2018).
20. Rahman, R., Asombang, A.W. & Ibdah, J.A. *World J Gastroenterol* **20**, 4483-4490 (2014).

Acknowledgements

The results are in part based upon data generated by the TCGA Research Network: http://cancergenome.nih.gov/. Our funding sources are as follows. J.N.K.: RWTH University Aachen (START 2018-691906). A.T.P.: NIH/NIDCR (K08-DE026500). A.M.: German Federal Ministry of Education and Research (BMBF) (M2oBITE/13GW0091E). DACHS study: 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, HO 5117/2-1), BMBF (01ER0814, 01ER0815, 01ER1505A, 01ER1505B). P.B.: DFG (BO 3755/6-1, SFB-TRR57, SFB-TRR219). T.L.: 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, LU 1360/3-1), Ernst-Jung-Foundation Hamburg and IZKF (interdisciplinary center of clinical research) at RWTH Aachen.

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 statistics; 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 that no competing interests exist.

Figure legends

Fig. 1: Tumor detection and MSI prediction in hematoxylin-eosin histology. (a) A convolutional neural network was trained as a tumor detector for gastric and colorectal cancer. Scale bar 4 mm. (b) Tumor regions were cut into square tiles, which were (c) color-normalized and sorted into microsatellite instable (MSI) or stable (MSS). (d) Another network was trained to classify MSI versus MSS. (e) This automatic pipeline was applied to held-out patient sets. Scale bar: 256 µm.

Fig. 2: Classification performance in an external validation set. (a-b) Tissue slides of MSI and MSS patients in the TCGA-CRC-DX test set show spatial patterns of predicted MSI score (see also 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 (ROC) curve with bootstrapped 95% confidence interval in DACHS (N=378 patients), TPR = true positive rate (sensitivity), FPR = false positive rate (1 – specificity). (e) Pearson correlation of predicted MSIness to transcriptomic and immunohistochemical (IHC) data across test sets. Precise p-values are listed in Suppl. Table 5. Sample size per cohort are: STAD N=91, CRC-KR N=105, CRC-DX N=95, DACHS N=134 patients. No adjustments for multiple comparisons were made and all statistical tests were two-sided.

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 (CIOMS). Anonymized archival tissue samples were retrieved from the tissue bank of the National Center for Tumor diseases (NCT, Heidelberg, Germany; including samples from the DACHS trial[17,18]) and from the pathology archive at UMM (University Medical Center Mannheim, Heidelberg University, Mannheim, Germany) after approval by the institutional ethics boards as described before[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 GI cancer, we used histological specimens of colorectal and stomach cancer surgical specimen from UMM and NCT tissue bank. This cohort was described before 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 non-tumor tissue (dense and loose tissue, representing muscle/stroma and fat/mucus, respectively), yielding 11,977 unique image tiles of 256 µm edge length. All of these images are freely available for download at http://dx.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 by Macenko et al.[13,14,23]

We retrieved histology images of N=315 STAD patients (diagnostic slides, FFPE tissue), N=387 CRC-KR patients (kryosections, snap-frozen tissue), N=360 CRC-DX patients (diagnostic slides, FFPE tissue) and N=492

UCEC patients (diagnostic slides, FFPE tissue) from "The Cancer Genome Atlas" (TCGA)[24]. All slides contained tumor tissue (after manual review in a blinded way) and had resolution available as part of the metadata (microns per pixel, MPP). 99 (STAD), 109 (CRC-KR), 100 (CRC-DX) and 110 (UCEC) randomly selected patients were held out during training and were used as a test set. In all cases, training and test set 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 http://dx.doi.org/10.5281/zenodo.2530835 and http://dx.doi.org/10.5281/zenodo.2532612. All TCGA images can be downloaded from public repositories at the National Institutes of Health (NIH, USA) at https://portal.gdc.cancer.gov/.

For TCGA-CRC and TCGA-STAD, all patients who were previously defined as MSI-H (by Liu et al.[25]) were included in the MSI group. All patients with unknown MSI status but with a mutation count of >1000 (as defined by Bailey et al.[26]) were also included in the MSI group (this was the case for less than 10 patients in any cohort). Suppl. 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 on CRC in the southwest of Germany with long-term followed-up patients enrolled in more than 20 clinics of the study region. Also, we analyzed data of N=185 patients from Kanagawa Cancer Center, Yokohama, Japan (KCCH) as described previously[19]. More information about the cohorts is shown in Suppl. Tables 1-3.

Neural network models, tumor detection and MSI detection

For tumor detection in GI cancer, we trained a convolutional neural network (CNN) with deep residual learning ("resnet18")[27] model to classify tumor tissue vs. normal tissue by transfer learning. In TCGA-STAD, TCGA-CRC-KR, TCGA-CRC-DX and DACHS, the automatic GI tumor detector was used while 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 (Extended Data Fig. 1) which 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 CNNs were pre-trained on the ImageNet (www.image-net.org) database as described before[13]. Only the weights in the last 10 layers were trainable while all other weights were frozen. We used the Adam algorithm for training, counteracted overfitting by an L2-regularization of 1e-4 and used a fixed learning rate of 1e-6 for TCGA-STAD, TCGA-CRC-DX and TCGA-CRC-KR and 1e-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 GPUs (Titan Xp, Quadro P6000, Titan RTX). Performance was scored as area under the curve (AUC) in a receiver operating characteristic (ROC) 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[28]. 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 curve (AUC under ROC) as calculated with "perfcurve" in MATLAB R2018a. Correlations were calculated with R version 3.5.1 "cor.test" using the "Pearson" method.

Extended data figure legends

*See "inventory of supporting information"***.**

Additional references

21. Kawakami, H., Zaanan, A. & Sinicrope, F.A. *Curr Treat Options Oncol* **16**, 30 (2015).
22. Zhu, L.*, et al. Mol Clin Oncol* **3**, 699-705 (2015).
23. Macenko, M.*, et al.* in *IEEE International Symposium on Biomedical Imaging,* 1107-1110 (2009).
24. The Cancer Genome Atlas Network, *Nature* **487**, 330-337 (2012).
25. Liu, Y.*, et al. Cancer Cell* **33**, 721-735 e728 (2018).
26. Bailey, M.H.*, et al. Cell* **173**, 371-385 e318 (2018).
27. He, K., Zhang, X., Ren, S. & Sun, J in *Proceedings of the IEEE conference on computer vision and pattern recognition,* 770-778 (2016).
28. DiCiccio, T.J. & Efron, B. *Statist. Sci.* **11**, 189-228 (1996).
29. Krizhevsky, A., Sutskever, I. & Hinton, G.E. in *Advances in neural information processing systems,* 1097-1105 (2012).
30. Simonyan, K. & Zisserman, A. *arXiv* 1409.1556 (2014).
31. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. & Wojna, Z in *Proceedings of the IEEE conference on computer vision and pattern recognition,* 2818-2826 (2016).
32. Iandola, F.N.*, et al. arXiv*, 1602.07360 (2016).