



OPEN

The Digital Brain Tumour Atlas, an open histopathology resource

DATA DESCRIPTOR

Thomas Roetzer-Pejrimovsky¹✉, Anna-Christina Moser¹, Baran Atli¹, Clemens Christian Vogel¹, Petra A. Mercea^{1,2}, Romana Prihoda^{2,3}, Ellen Gelpi¹, Christine Haberler¹, Romana Höftberger¹, Johannes A. Hainfellner¹, Bernhard Baumann¹ , Georg Langs¹  & Adelheid Woehrer¹

Currently, approximately 150 different brain tumour types are defined by the WHO. Recent endeavours to exploit machine learning and deep learning methods for supporting more precise diagnostics based on the histological tumour appearance have been hampered by the relative paucity of accessible digital histopathological datasets. While freely available datasets are relatively common in many medical specialties such as radiology and genomic medicine, there is still an unmet need regarding histopathological data. Thus, we digitized a significant portion of a large dedicated brain tumour bank based at the Division of Neuropathology and Neurochemistry of the Medical University of Vienna, covering brain tumour cases from 1995–2019. A total of 3,115 slides of 126 brain tumour types (including 47 control tissue slides) have been scanned. Additionally, complementary clinical annotations have been collected for each case. In the present manuscript, we thoroughly discuss this unique dataset and make it publicly available for potential use cases in machine learning and digital image analysis, teaching and as a reference for external validation.

Background & Summary

Brain tumours account for a large fraction of years of potential life lost as compared with tumours from other sites¹, and have a significant negative impact on patients' quality of life². Overall, they are relatively uncommon neoplasms with an incidence of approximately 24 per 100,000 person-years³. Current diagnostic guidelines published by the WHO define approximately 150 distinct brain tumour types and assign grades I to IV, based on malignancy and potential to malignant transformation or progression. They are mainly differentiated by their histopathological phenotypes and molecular alterations⁴. While the majority of tumours is diagnosed solely based on histopathology, an integrated approach is mandatory for 19 tumour types.

Still, more accurate diagnostic distinctions are needed in order to i) better assess individual patients' prognoses and ii) support more robust therapeutic decisions^{4,5}. Recently, diagnostic algorithms trained on DNA methylation data have been shown to significantly increase diagnostic accuracy⁶. Similar advances focusing on histopathological data have been hampered, so far, by the lack of freely available histopathology datasets⁷. Most available histopathology data such as those available through TCGA⁸, IvyGAP^{9,10} or TCIA¹¹ focus on only a few diagnostic entities. They mostly consist of digitized fresh frozen tissue sections, which feature relatively poor tissue morphology as compared to formalin-fixed and paraffin-embedded tissues. Still, even with these limited data, computational algorithms have been successfully trained - amongst others - for survival prediction¹², detection of tumour-infiltrating lymphocytes¹³, and assessments of tumour microvessels¹⁴. However, larger datasets encompassing an even wider range of brain tumours and featuring improved cellular and morphological characteristics are necessary to further develop these algorithms and extend their applicability to the entire spectrum of brain tumour types.

Thus, we set out to compile a comprehensive resource of digitized Haematoxylin-eosin(H&E)-stained brain tumour whole slide images (WSIs) with clinical annotations (Fig. 1). We aimed to capture the complete spectrum of brain tumours as encountered in day-to-day medical diagnostic practice. Importantly, we managed to specifically digitize slides of exceedingly rare pathologies, which are usually, if ever, seen only a few times in a

¹Division of Neuropathology and Neurochemistry, Department of Neurology, Medical University of Vienna, Vienna, Austria. ²Department of Neurosurgery, Medical University of Vienna, Vienna, Austria. ³Department of Neurosurgery, University Hospital St. Poelten, St. Poelten, Austria. ⁴Center for Medical Physics and Biomedical Engineering, Medical University of Vienna, Vienna, Austria. ⁵Department of Biomedical Imaging and Image-Guided Therapy, Computational Imaging Research Lab, Medical University of Vienna, Vienna, Austria. ✉e-mail: thomas.roetzer-pejrimovsky@meduniwien.ac.at

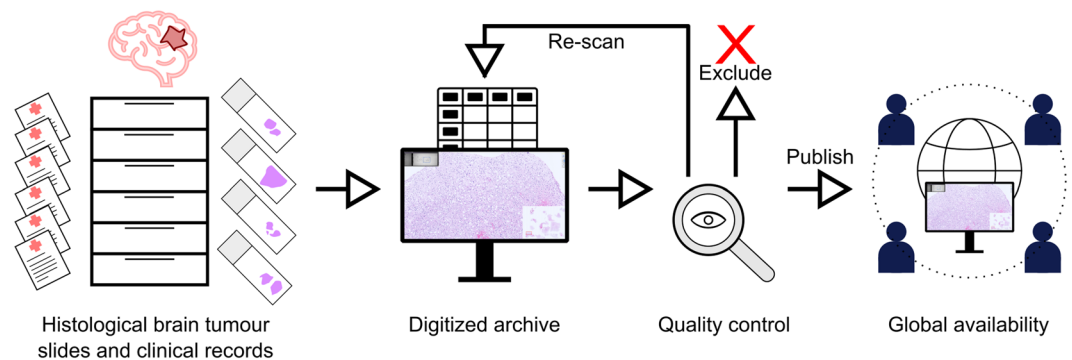


Fig. 1 Overview of the data acquisition and publication process. First, histological slides and clinical records of brain tumour patients were retrieved from the biobank of the Division of Neuropathology and Neurochemistry, Medical University of Vienna. Then, slides were digitized using a Hamamatsu slidescanner. Clinical data were translated into standardized annotations. At least two experienced neuropathologists checked each slide scan to ensure conformity of the diagnosis with the current revised 4th edition of the “WHO Classification of Tumours of the Central Nervous System” and sufficient scan quality. Ambiguous cases were excluded and WSIs of inferior quality were re-scanned. Finally, data were made available via EBRAINS to the international research community. (Brain illustration adapted from Meaghan Hendricks from the Noun Project).

Variable	Description
uuid	unique sample identifier
pat_id	unique patient identifier
diagnosis	primary diagnosis according to the WHO <i>Classification of Tumours of the Central Nervous System (2016)</i>
grade	WHO grade according to the WHO <i>Classification of Tumours of the Central Nervous System (2016)</i>
subtype	further specification of the histopathological subtype which is not a distinct entity as defined by the WHO, if applicable
secondary_diagnosis	secondary diagnosis in cases where two distinct diagnosis according to the WHO are applicable
control	1 if sample is a control sample without tumour tissue
age	patient age at the time of surgery
sex	biological patient sex
location	list (in square brackets) of all applicable tumour locations; empty if location is unknown
laterality	laterality of the tumour (left or right)
cellularity	estimated cell density of the tissue (given in $1/\text{mm}^2$)
tissue_area	estimated scanned tissue area (in mm^2)
recurrence	0 if the entry corresponds to a primary tumour resection; if the entry corresponds to a tumour recurrence, the number of the recurrence is given (e.g., 2 corresponds to the second recurrence)
comment	notable findings that do not fit in other columns (e.g., important mutations not yet integrated in the WHO classification; other non-tumour pathologies in the control samples)

Table 1. Recorded clinical variables and corresponding descriptions.

pathologist’s lifetime. By performing a manual review of each slide, we ensure high scan quality and actuality of provided diagnoses. We envisage this dataset to be used for advancing digital pathology-based machine learning and for teaching purposes. Importantly, this dataset can be used for (1) inter-tumour comparisons thanks to the wide inclusion of distinct brain tumour types as well as (2) within-tumour-type investigations thanks to the inclusion of a large number of samples for the common tumour types.

Methods

Sample acquisition. H&E stained tumour slides from FFPE tissues, which were collected for routine diagnostics in the time interval of 1995–2019 have been obtained from the biobank of the Division of Neuropathology & Neurochemistry, Medical University of Vienna. We digitized each slide in high magnification (40x objective, 228 nm/pixel) using a Hamamatsu NanoZoomer 2.0 HT slide-scanner. Each slide was manually reviewed to ensure high scan quality and sufficient diagnostic tumour tissue. Samples with equivocal diagnoses or missing molecular work-up otherwise needed to assign an integrated WHO 2016 diagnosis were excluded. A subset of glioblastoma scans ($n = 381$) has been published previously as part of the GBMatch study¹⁵.

Basic clinical annotations consisting of patient age and sex as well as tumour location and recurrence were acquired from local electronic records where available. Tumour locations have been assigned to the following 19 categories: frontal; parietal; insular; occipital; temporal; cerebellar; brain stem; spinal; lateral ventricle; diencephalon; third ventricle; fourth ventricle; sellar region; cranial nerves; basal ganglia; cerebral, NOS (not otherwise specified); posterior fossa, NOS; cranial, NOS; and other.

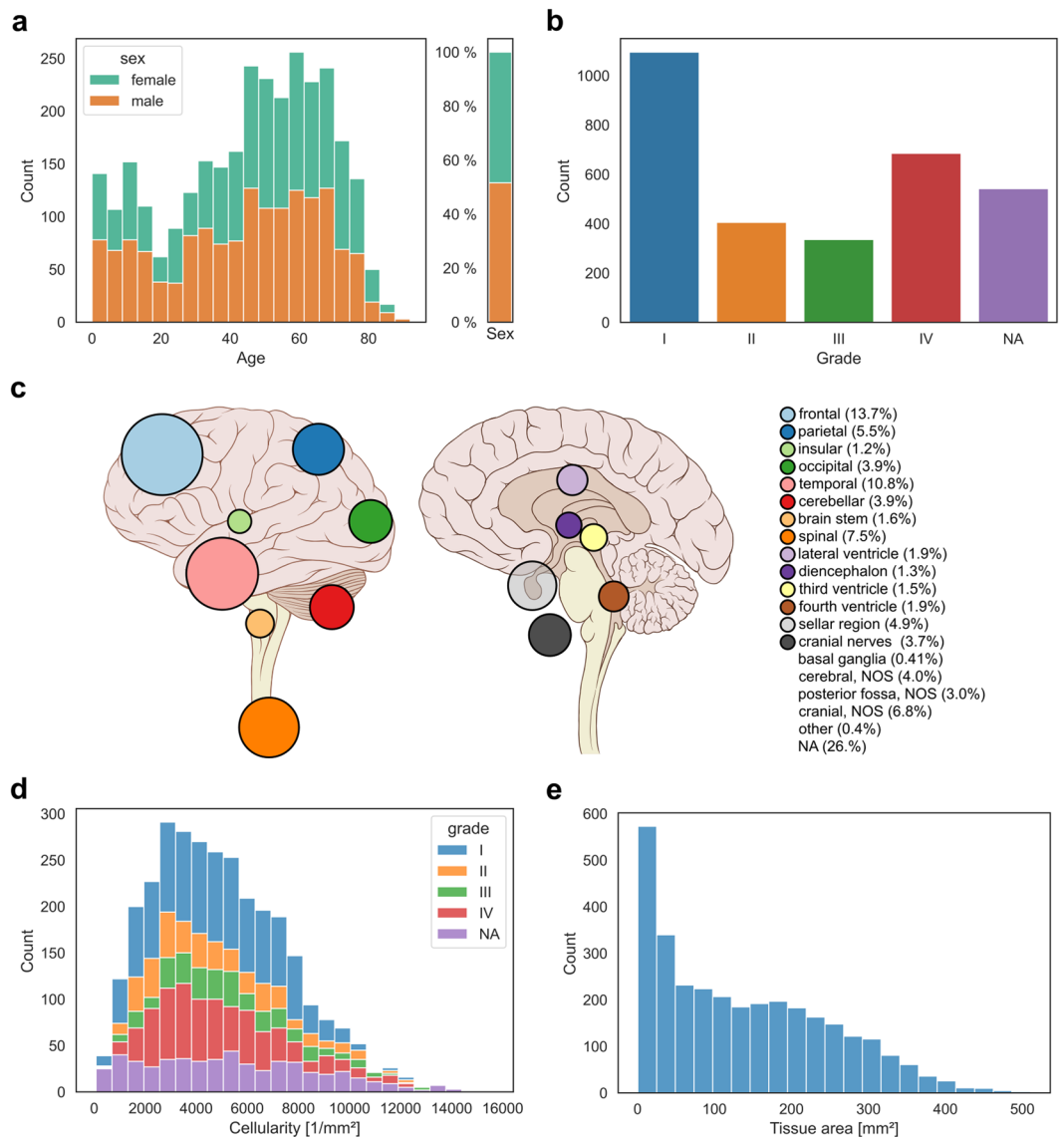


Fig. 2 Descriptive statistics of the ‘Digital Brain Tumour Atlas’ patient cohort (not including control patients). **(a)** The age distribution by sex shows a bimodal distribution with most patients belonging to the higher-age categories. Since some uncommon tumour types like medulloblastoma occur mainly in children and have been strategically over-sampled, there is also a peak in younger patients. **(b)** The distribution of the different WHO grades shows a slight predominance of grade I and grade IV tumours. Of note, some tumour entities are not assigned WHO grades (‘NA’) and very few tumour types are assigned intermediate grades II–III (a total of five cases, not shown in the figure). **(c)** Tumour distribution with colour-coded locations and ratio-specific circle sizes. (Brain illustration adapted from Patrick J. Lynch, wikimedia) **(d)** Distribution of the cell densities of all included tumour samples by tumour grade. Note that lower-grade tumours are not necessarily less cell dense (e.g., in the case of cellular schwannoma). **(e)** The distribution of the scanned tissue areas (per slide).

This study complies with the relevant ethical, legal and institutional regulations and the study protocol has been approved by the Ethics Committee of the Medical University of Vienna (EK1691–2017). Participant informed consent has been obtained as by institutional guidelines, necessitating restrictions on commercial use of the obtained data.

Estimation of cell density and scanned tissue area. Additionally, the total tissue area and the average cellularities were estimated for each scan using a custom MATLAB script (MATLAB R2017b, MathWorks) with a similar approach as previously published^{15,16}. In summary, H&E stained WSIs were first colour-deconvoluted into separate Haematoxylin and Eosin channels¹⁷. Then, global, Phansalkar and Otsu thresholding were applied to the Haematoxylin channel to identify nuclei^{18,19}. Watershedding was used to separate densely clustered cells²⁰. Only cells with a minimum size of 4 pixels were kept. The total tissue area was determined by averaging all colour channels, thresholding at a threshold of 220, followed by binary *close* and *open* operations.

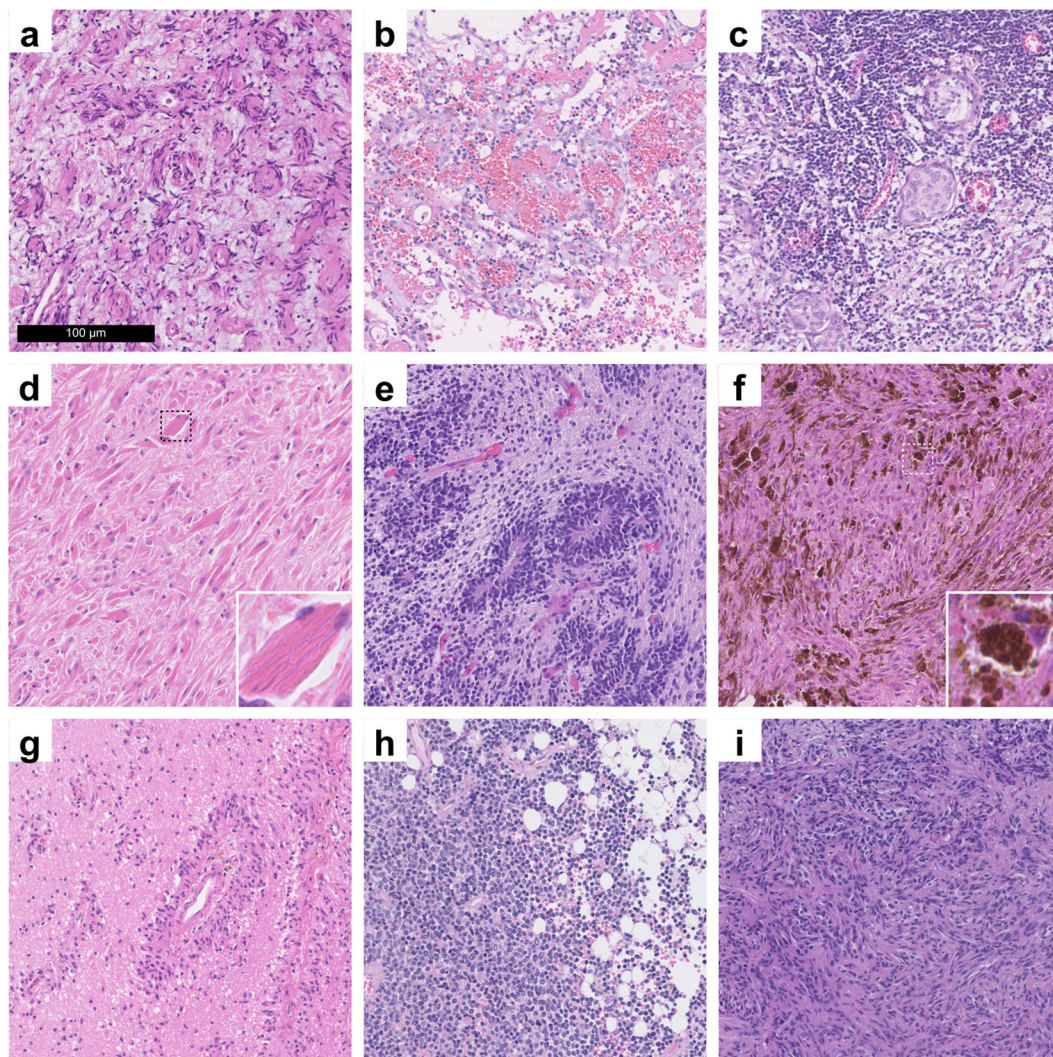


Fig. 3 Exemplary images from exceedingly rare brain tumours, which are included in the DBTA. (a) Perineurioma component of a hybrid nerve sheath tumour. (b) Angiosarcoma. (c) Lymphoplasmacyte-rich meningioma. (d) Crystal-storing histiocytosis. (e) Embryonal tumour with multilayered rosettes. (f) Melanotic schwannoma. (g) Angiocentric glioma. (h) Cerebellar liponeurocytoma. (i) Pituicytoma.

Data Records

Data are provided via EBRAINS²¹ as one ndpi-file per sample, sorted by diagnostic tumour type (in alphabetical order) for easier access. It is possible to download single files directly or all files of a specific tumour type or the whole dataset using a download manager (such as the Chrono Download Manager for the Google Chrome browser). Furthermore, supplementary clinical information, estimated cell densities and scanned tissue area is provided in a csv-spreadsheet with one row per tumour sample. An overview of all spreadsheet variables and descriptions is given in Table 1.

A total of 3,115 histological slides of 2,880 patients have been scanned. A total of 126 distinct diagnostic tumour types could be included. There are 1,395 female and 1,462 male patients in the dataset. The mean patient age at brain tumour surgery was 45 years, ranging from 9 days to 92 years. 2,530 of the scanned slides originated from primary operations and 538 from re-operations. See online-only Table 1 for descriptive properties broken down by tumour type. Descriptive visualizations of patient age, sex, tumour location, cellularity, and scanned tissue area are given in Fig. 2. Of note, we also scanned exceptionally rare tumour types such as melanotic schwannomas or liponeurocytomas (Fig. 3). A total of 47 non-tumour slides from different non-tumour CNS regions and with different pathologies were included as controls.

Technical Validation

All cases were initially selected based on the given diagnosis in the diagnostic electronic records. To ensure conformity with the WHO 2016 diagnosis, all slides have been independently reviewed by two neuropathologists experienced in neuro-oncology. In disputed cases, a third senior neuropathologist was consulted. Older cases with missing necessary molecular analyses were not included in the dataset.

Inter- and intraobserver variability is one factor that contributes to misdiagnoses or discrepant diagnoses. We mitigated the risk by including only cases that had already undergone thorough routine diagnostic work-up and were additionally reviewed independently by at least two neuropathologists as described above. In this way, we also ensured excellent image quality and the presence of sufficient diagnostic tumour tissue on each WSI. Scans with suboptimal image quality were either re-scanned (if possible) or excluded.

Usage Notes

Data access. The data can be accessed via EBRAINS²¹. In order to download the data set, users have to register with EBRAINS and agree to the general terms of use, access policy as well as the data use agreement for pseudonymised human data (<https://ebrains.eu/terms>). The data are distributed under the conditions that users cite the respective DOI, adhere to EBRAINS' Data Use Agreement and do not use the data for commercial purposes.

WSI processing. The `ndp.view2` (© Hamamatsu) software can be freely used to view and annotate slide scans saved in the `ndpi` format²². Alternatively, most other WSI programs such as the open-source OMERO software platform²³ and the open-source QuPath software²⁴ can work directly on `ndpi`-files. However, most programming languages and non-specialized image processing software cannot handle `ndpi`-files out of the box. Thus, we also provide a toolbox of MATLAB scripts that depend on the `openslide` library²⁵ and can be used to

1. Automatically tile large slide scans and export multiple smaller image patches in a given magnification.
2. Convert annotation-files (`.ndpa`) to overlays, which can be used to extract specific regions of interest.
3. Estimate the total tissue area on a WSI.
4. Estimate the cell density on a WSI.

Of note, slide thickness and staining intensity vary to some degree, resulting in a slightly different histological appearance of each slide. Thus, for machine learning applications, we recommend astain normalization step such as WSICS²⁶, more recent methods employing generative adversarial networks²⁷ or style transfer learning²⁸. Moreover, heavy stain colour augmentation should be performed²⁹. Of note, the stain normalization step can be omitted with only a negligible drop in performance as has been shown by Tellez *et al.*²⁹.

Code availability

The custom-made MATLAB toolbox for loading, viewing and processing of `ndpi` & `ndpa` files and for estimating the total tissue area and average cell density of a WSI can be accessed at: https://github.com/tovaroe/WSI_histology.

Received: 13 April 2021; Accepted: 22 December 2021;

Published online: 15 February 2022

References

1. Rouse, C., Gittleman, H., Ostrom, Q. T., Kruchko, C. & Barnholtz-Sloan, J. S. Years of potential life lost for brain and CNS tumors relative to other cancers in adults in the United States, 2010. *Neuro. Oncol.* **18**, 70–77 (2015).
2. Liu, R., Page, M., Solheim, K., Fox, S. & Chang, S. M. Quality of life in adults with brain tumors: Current knowledge and future directions. *Neuro. Oncol.* **11**, 330–339 (2009).
3. Ostrom, Q. T. *et al.* CBTRUS Statistical Report: Primary Brain and Other Central Nervous System Tumors Diagnosed in the United States in 2013–2017. *Neuro. Oncol.* **22**, iv1–iv96 (2020).
4. International Agency for Research on Cancer & Wiestler, O. D. *WHO Classification of Tumours of the Central Nervous System*. (International Agency for Research on Cancer, 2016).
5. van den Bent, M. J. *et al.* A clinical perspective on the 2016 WHO brain tumor classification and routine molecular diagnostics. *Neuro. Oncol.* **19**, 614–624 (2017).
6. Capper, D. *et al.* DNA methylation-based classification of central nervous system tumours. *Nature* **555**, 469–474 (2018).
7. Djuric, U., Zadeh, G., Aldape, K. & Diamandis, P. Precision histology: how deep learning is poised to revitalize histomorphology for personalized cancer care. *npj Precision Oncology* **1**, 1–5 (2017).
8. The Cancer Genome Atlas Program. <https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga> (2018).
9. Puchalski, R. B. *et al.* An anatomic transcriptional atlas of human glioblastoma. *Science* **360**, 660–663 (2018).
10. Ivy Glioblastoma Atlas Project. <https://glioblastoma.alleninstitute.org/> (2018).
11. National Cancer Institute Clinical Proteomic Tumor Analysis Consortium (CPTAC). Radiology Data from the Clinical Proteomic Tumor Analysis Consortium Glioblastoma Multiforme [CPTAC-GBM] collection. <https://doi.org/10.7937/K9/TCIA.2018.3RJE41Q1> (2018).
12. Mobadersany, P. *et al.* Predicting cancer outcomes from histology and genomics using convolutional networks. *Proc. Natl. Acad. Sci. USA* **115**, E2970–E2979 (2018).
13. Spatial Organization and Molecular Correlation of Tumor-Infiltrating Lymphocytes Using Deep Learning on Pathology Images. *Cell Rep.* **23**, 181–193.e7 (2018).
14. Li, X., Tang, Q., Yu, J., Wang, Y. & Shi, Z. Microvasculature detection and quantification in glioma: a novel deep-learning-based framework. *Lab. Invest.* **99**, 1515–1526 (2019).
15. Klughammer, J. *et al.* The DNA methylation landscape of glioblastoma disease progression shows extensive heterogeneity in time and space. *Nat. Med.* **24**, 1611–1624 (2018).
16. Roetzer, T. *et al.* Evaluating cellularity and structural connectivity on whole brain slides using a custom-made digital pathology pipeline. *J. Neurosci. Methods* **311**, 215–221 (2019).
17. Ruifrok, A. C. & Johnston, D. A. Quantification of histochemical staining by color deconvolution. *Anal. Quant. Cytol. Histol.* **23**, 291–299 (2001).
18. A Threshold Selection Method from Gray-Level Histograms. <https://ieeexplore.ieee.org/document/4310076>.
19. Adaptive local thresholding for detection of nuclei in diversity stained cytology images. <https://ieeexplore.ieee.org/abstract/document/5739305>.
20. Topographic distance and watershed lines. *Signal Processing* **38**, 113–125 (1994).

21. Roetzer-Pejrimovsky, T. *et al.* The Digital Brain Tumour Atlas, an open histopathology resource. *EBRAINS* <https://doi.org/10.25493/WQ48-ZGX> (2021).
22. U12388-01. <https://www.hamamatsu.com/us/en/product/type/U12388-01/index.html>.
23. Allan, C. *et al.* OMERO: flexible, model-driven data management for experimental biology. *Nat. Methods* **9**, 245–253 (2012).
24. Bankhead, P. *et al.* QuPath: Open source software for digital pathology image analysis. *Sci. Rep.* **7**, 1–7 (2017).
25. Goode, A., Gilbert, B., Harkes, J., Jukic, D. & Satyanarayanan, M. OpenSlide: A vendor-neutral software foundation for digital pathology. *J. Pathol. Inform.* **4**, 27 (2013).
26. Stain Specific Standardization of Whole-Slide Histopathological Images. <https://ieeexplore.ieee.org/abstract/document/7243333>.
27. Zanjani, F. G., Zinger, S., Bejnordi, B. E., van der Laak, J. A. W. M. & de With, P. H. N. Stain normalization of histopathology images using generative adversarial networks. *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)* <https://doi.org/10.1109/isbi.2018.8363641> (2018).
28. Bug, D. *et al.* Context-Based Normalization of Histological Stains Using Deep Convolutional Features. in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support* 135–142 (Springer, Cham, 2017).
29. Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. *Med. Image Anal.* **58**, 101544 (2019).

Acknowledgements

T.R. is a recipient of a DOC Fellowship (25262) of the Austrian Academy of Sciences at the Division of Neuropathology and Neurochemistry, Department of Neurology, Medical University of Vienna. The present work has been further supported by the Austrian Science Fund 1000 ideas project TAI98-B to A.W.

Author contributions

T.R. and A.W. conceived and designed the project. T.R., A.C.M., C.C.V., P.M., B.A. and R.P. collected the data. T.R., E.G., R.H., C.H., J.A.H. and A.W. reviewed the data. T.R., B.B. and G.L. performed the image analysis. T.R. and A.W. wrote the paper with contributions from all authors.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to T.R.-P.

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

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



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

The Creative Commons Public Domain Dedication waiver <http://creativecommons.org/publicdomain/zero/1.0/> applies to the metadata files associated with this article.

© The Author(s) 2022