nature portfolio

Corresponding author(s): Anahita Fathi Kazerooni

Last updated by author(s): Jan 2, 2025

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	The descriptions of all codes used for generation of genomic data are provided in the Methods section, subsections "Study Cohort", "Next Generation Sequencing (NGS)", "Analysis of Copy Number Variation"
Data analysis	All data analysis methods for mutational signature, evolutionary trajectories, and image processing are detailed in the Methods section, subsections"Mutational Signature Analysis","Mutant-Allele Tumor Heterogeneity (MATH) Score; A Measure of Molecular Heterogeneity", "Prediction of Evolutionary Trajectories", "Radiogenomic Signatures"

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Data Availability

The MRI sequences, image acquisition parameters, and clinical variables used in this study can be found in the UPenn-GBM collection in the Cancer Imaging Archive (https://doi.org/10.7937/TCIA.709X-DN49) 73. The derived data (radiomic features and mutation status) that support the findings of this study are available on reasonable request from the corresponding author (C. D.). However, the raw data (images and genomic variables) are not distributable, as the patients were not consented for public release of their data. The numerical data used to create the figures in this manuscript are available in the Supplementary Information/Source Data file on Figshare, DOI: 10.6084/m9.figshare.28243670. All oligonucleotide sequences used in this study, including primers and probes are provided on Figshare, DOI: 10.6084/m9.figshare.28243673. Additional details are available upon reasonable request from the corresponding author.

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	The study included data from both sexes. The detailed demographic information is summarized in Supplemental Information. The default for CNVkit is to calculate log2 ratios and copy number with respect to a diploid chrX and haploid chrY, infer sample sex based on chrX and chrY coverage, and double chrX coverage for male samples. Because this panel targets only one gene on chrX and none on chrY, sex inference was not always accurate. Additionally, the 'genemetrics' and 'call' commands are not typically used in conjunction, and both attempt to determine sex, which can result in unexpected behavior. To address this all samples were treated as female at the 'genemetrics' step while sex was explicitly asserted in the 'call' command.
Population characteristics	All population characteristics are summarized in Supplementary Table 1.
Recruitment	This retrospective HIPAA-compliant study was approved by the IRB of the University of Pennsylvania (IRB protocol 706564). All patients provided their informed consent at the time of imaging. A retrospective cohort of n = 585 patients with GBM tumors who had undergone surgery at the Hospital of the University of Pennsylvania (HUP) between 2006 to 2019 was reviewed. The patients who had undergone maximal safe or partial resection, were histopathologically confirmed with de novo glioblastoma tumors, received radiotherapy, and concomitant and adjuvant chemotherapy with temozolomide, and at the very least had four conventional MRI scans, including pre- and post-gadolinium T1-weighted (T1, and T1-Gd), T2-weighted (T2), and T2 fluid attenuation inversion recovery (T2-FLAIR), acquired pre-operatively on a 3T MRI scanner (Siemens, Tim Trio, Erlangen, Germany) for a first-occurrence brain tumor were included. A cohort of 374 patients meeting the inclusion criteria was collected. Of these, 244 had additional advanced MRI sequences available, including diffusion tensor imaging (DTI) and dynamic susceptibility contrast (DSC) MRI with preload administration. All measurements described below were conducted on separate samples and patients.
Ethics oversight	The Institutional Review Board of the University of Pennsylvania approved this study.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	The sample size was determined based on the maximum number of patients with available molecular sequencing data.
Data exclusions	We excluded patients with mutations in IDH1 or IDH2 (n = 16).
Replication	We randomly split the data into discovery and replication sets, keeping a stratified split . We used an unseen replication set to test the predictive value of our models trained on the discovery set independently on the replication set.
Randomization	The samples were allocated to different groups of mutant vs wildtype based on their mutation status. For splitting the data into Discovery and Replication sets, we used a random splitting approach .
Blinding	There was no need to blind information as there were no qualitative assessments in this study.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☐	☒ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☐	☒ MRI-based neuroimaging

Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

Clinical trial registration	This wasn't a clinical trial study;
Study protocol	Not Applicable
Data collection	Retrospective as described before
Outcomes	The outcomes in this study were mutant/non-mutant tumors

Magnetic resonance imaging

Experimental design

Design type	Structural MRI
Design specifications	Not a functional MRI study
Behavioral performance measures	Not applicable

Acquisition

Imaging type(s)	structural, diffusion, perfusion
Field strength	3 Tesla
Sequence & imaging parameters	image acquisition parameters can be accessed at https://www.cancerimagingarchive.net/collection/upenn-gbm/
Area of acquisition	brain

Diffusion MRI ☒ Used ☐ Not used

Parameters: b-values of 0, 1000

Preprocessing

Preprocessing software	Pre-processing of multi-parametric MRI (mpMRI) scans was performed using the CaPTk software 31,32 (https://www.med.upenn.edu/cbica/captk/)
Normalization	co-registration and resampling to an isotropic resolution of 1mm3 based on a common anatomical SRI24 atlas 33, performed using the "Greedy" software 34; the image intensities were rescaled in the intensity range of 0-255 after removing outliers.
Normalization template	SRI24
Noise and artifact removal	Image intensity outliers outside the range of 99th percentile were removed
Volume censoring	not applicable

Statistical modeling & inference

Model type and settings	Described in the Methods Section Statistics and Reproducibility

In all DL and ML classification scenarios, the data was randomly split using a stratified approach to ensure the proportion of mutated to wildtype tumors was preserved in both the discovery and replication cohorts. The split assigned 75% of the data to the discovery cohort and 25% to an independent (unseen) replication cohort. For the DL and ML models utilizing conventional MRI, the discovery and replication cohorts included 268 and 89 patients, respectively. For the ML model incorporating both conventional and advanced MRI, the discovery and replication cohorts consisted of 171 and 57 patients, respectively. The performance of our predictive ML and DL models in stratifying mutation status was evaluated using the area under the receiver operating characteristic (ROC) curve (AUC) and balanced accuracy metrics. Pairwise correlation of the extracted radiomic features was assessed using Pearson's correlation coefficient. To analyze the separability of DL and ML signatures, we compared tumors with single alterations (i.e., exclusive tumors) to those with co-occurring multiple alterations using Cohen's D effect size. Additionally, cosine similarity metrics were calculated to compare molecular signatures across tumor groups stratified by spatiogenomic tumor location, thereby quantifying and validating distinct mutational profiles among these groups. Differences in MATH scores across brain regions were analyzed using the Wilcoxon rank-sum test.

Effect(s) tested: Cohen D's effect size

Specify type of analysis: ☐ Whole brain ☐ ROI-based ☒ Both

Anatomical location(s): brain tumors, and different labels of brain regions: frontal, temporal, occipital,

Statistic type for inference (See Eklund et al. 2016): N/A

Correction: N/A

Models & analysis

n/a	Involved in the study
☒	☐ Functional and/or effective connectivity
☒	☐ Graph analysis
☐	☒ Multivariate modeling or predictive analysis

Multivariate modeling and predictive analysis

Deep Learning-Based Radiogenomic Signatures

In the discovery cohort, the remaining radiomic features were ranked using recursive feature elimination with random forests and the optimal number of features was determined through nested cross-validation (10 folds for the outer and 5 folds for the inner folds). The selected feature subset was imported into the CNN classifier, as described below, for prediction of mutation status in the key driver genes, i.e., EGFR, NF1, PTEN, and TP53, and pathways, i.e., RTK and PI3K.
In the discovery cohort, we trained a CNN classifier based on a modified 34-layer ResNet architecture 40, with an initial 7×7 convolution and four layers (layers 1 to 4, composed of 3, 4, 6, and 3 residual blocks respectively; each residual block comprises two 3×3 convolution) 41. Slices of two structural images, i.e., T1-Gd and T2, along with slices of tumor segmentations (overall three image volumes) were passed as inputs for training the ResNet model. Specifically, the structural images and segmentation masks were cropped with margins of 8 pixels from the boundaries of brain tissue and resampled to 128×128×128 voxels. Intensity normalization of the brain tissue was performed using z-scoring for all structural MRI scans. Five axial slices were identified from the tumor segmentations, where the axial slice in the center included the maximum tumor mask area; 2 upper and 2 lower slices were also identified along with the center slice. After concatenating all three volumes together, the five slices were finally extracted from this concatenated volume. The discovery cohort was split into 70%-30% training-validation sets. The ResNet model was pretrained with input size of 5×3×128×128×128 for each subject in the training set (Supplementary Note 2, Supplementary Figure 1). After pretraining the ResNet, the weights from the initial 7×7 convolution to layer 3 were fed to the CNN classifier and were not used in any further training process. The radiomic features (including spatial location features) were imported as inputs to the fully connected layers that followed layer 4, and along with the weights from layer 4 were used to train the CNN classifier. Input images were further augmented using rotation. Optimization was performed using Adam optimizer (initial learning rate (x0) = 0.0001), with L2 penalty of 0.01, the cross-entropy cost function, and an exponential learning rate decay (xepoch = x0*0.99epoch). For both pretraining and training phases, the maximum training epoch was set to 75, and the model was saved when minimum validation loss was achieved.

Machine Learning-Based Radiogenomic Signatures

We also trained classical machine learning models, using support vector machines (SVM), with radiomic features for prediction of mutations in each of the key driver genes and pathways. For the patients in the discovery cohort, feature selection and classifier training were performed using Least Absolute Shrinkage and Selection Operator (LASSO) feature selection approach wrapped with SVM with a linear kernel. This method was applied in a nested cross-validation (nested-CV) schema with 5-fold CV in the inner loop for feature subset selection and with 10-fold CV in the outer loop for model optimization and hyperparameter selection to avoid data overfitting and ensure generalizability to the unseen replication cohort.

Classification Scenarios

Radiogenomic signatures of co-occurring mutations in key driver genes, i.e., EGFR, PTEN, TP53, NF1, and key pathways, i.e., RTK, PI3K, P53, and MAPK were generated based on (1) radiomic features extracted from multi-parametric conventional MRI scans using deep learning and SVM classification methods, and (2) the features calculated from multi-parametric conventional and advanced MRI scans using SVM classification.