A Clinically Annotated Transcriptomic Atlas of Nervous System Tumors

Chi H. Le, MD, MSc;[1] Ajai K. Nelson;[2] Janki R. Naidugari, BS;[3] Robert P. Naftel, MD;[4] Eyas M. Hattab, MD, MBA;[5] Brian J. Williams, MD;[6] Akshitkumar M. Mistry, MD[6*]

[1]*School of Medicine, Vanderbilt University, Nashville, TN USA*

[2]*Oberlin College, Oberlin, OH, USA*

[3]*School of Medicine, University of Louisville, Louisville, KY, USA*

[4]*Department of Neurological Surgery, Vanderbilt University Medical Center, Nashville, TN, USA*

[5]*Department of Department of Pathology and Laboratory Medicine, University of Louisville, Louisville, KY, USA*

[6]*Department of Neurological Surgery, University of Louisville, Louisville, KY, USA*

Running Title: Transcriptomic Atlas of Nervous System Tumors

Word counts:	Abstract: 234	Text: 3451 (Introduction to conclusion)
	References: 24	Figure legends: 4 (main); 3 (supplementary)

***Corresponding Author:**

Akshitkumar M. Mistry, MD

Department of Neurological Surgery

University of Louisville

220 Abraham Flexner Way, 15th Floor

Louisville, KY 40202

Tel: 502-588-2329; Fax: 502-407-3256

Email: axitamm@gmail.com

ABSTRACT

Background: While DNA methylation signatures are distinct across various nervous system neoplasms, it has not been comprehensively demonstrated whether transcriptomic signatures exhibit similar uniqueness. Additionally, no single, large-scale dataset is available for comparative gene expression analyses of these neoplasms. This study aims to address these knowledge and resource gaps.

Methods: Raw transcriptomic and any associated clinical data for nervous system neoplasms (5,402 samples) and non-neoplastic entities (1,973 samples) were obtained from publicly available sources. These data were generated using the Applied Biosystems™ (previously Affymetrix®) GeneChip™ Human Genome U133 Plus 2.0 Array and reprocessed simultaneously for a harmonized integration. Machine learning tools were used to visualize all the samples and evaluate cluster formation. Of them, 2,127 samples did not belong to a cluster or lacked a diagnosis according to current classifications. They were reclassified by training machine learning classifiers with 5,248 samples with a known diagnosis.

Results: We created a large-scale, clinically annotated transcriptomic dataset from public domain sources by reprocessing, integrating, and reclassifying samples with uncertain diagnoses. Visualization using machine learning tools revealed clustering primarily based on diagnosis.

Conclusions: We demonstrate that the diagnostic distinctiveness of bulk DNA methylation signatures also extends to gene expression across the diagnostic spectrum of nervous system neoplasms. Our dataset's broad coverage of diagnoses, including rarely studied entities, spans all ages and includes individuals from diverse geographical regions, enhancing its utility for comprehensive and robust comparative gene expression analyses.

KEYWORDS (five): Atlas; Transcriptome; Gene Expression; Brain Neoplasms; Nervous System Neoplasms

KEY POINTS:

- We present a transcriptomic atlas of nervous system tumors and non-tumor entities.
- Sample clustering in this atlas is primarily driven by diagnosis.
- This large-scale atlas allows for comparative gene expression analyses.

IMPORTANCE OF THE STUDY

In this study, we present an atlas that combines harmonized gene expression and manually curated clinical data from 5,402 neoplastic and 1,973 non-neoplastic nervous system samples from the public domain. Currently, there is a lack of comprehensive atlases covering a broad range of nervous system neoplasm diagnoses, including rarely studied entities, and spanning different geographic regions and age groups. They enable discoveries through comprehensive and robust comparative gene expression analyses across the diagnostic spectrum of nervous system neoplasms. Additionally, we demonstrate that the diagnostic distinctiveness of bulk DNA methylation signatures also extends to gene expression across the diagnostic spectrum of nervous system neoplasms and age groups. In the process, we identified specific entities within discrete gliomas that require further diagnostic refinement. Finally, the methods of this study can be applied to integrate and harmonize raw transcriptomic data from other rare conditions, enhancing their utility.

INTRODUCTION

The unique DNA methylation signature of neoplasms has enabled the development of various DNA methylation-based classifiers for diagnosing neoplasms.[1-3] DNA methylation, an epigenetic modification, regulates gene transcription. While the bulk transcriptomic signature of nervous system neoplasms is also believed to be distinctive, it has not been comprehensively demonstrated across all nervous system neoplasms and age groups. A large-scale dataset showing this woudld permit comparative gene expression analyses of nervous system neoplasms with statistical confidence. This study addresses these knowledge and resource gaps by harmonizing publicly available data to create a large-scale, clinically annotated transcriptomic dataset. This dataset includes 5,402 neoplastic samples affecting the nervous system and 1,973 non-neoplastic samples. This resource can be widely used for gene expression-based analyses to support clinical and biological research in neuro-oncology.

METHODS

Transcriptomic Platform Selection

To ensure precision in the integrated data, we used only raw data generated on a single platform, allowing for uniform reprocessing. Our comprehensive search of major online genomic databases revealed that most raw transcriptomic data for neoplasms affecting the nervous system were generated using the Applied Biosystems™ (formerly Affymetrix®) GeneChip™ Human Genome U133 Plus 2.0 Array [National Center for Biotechnology Information (NCBI) ID: GPL570].[4] GPL570 provides extensive coverage of the human genome,[5] and its data are comparable to RNA sequencing data, offering similar potential for biological discoveries.[6]

Data Collection

Raw data files of neoplasms affecting the nervous system and non-neoplastic samples of the nervous system analyzed using GPL570 were retrieved from the Gene Expression Omnibus

(NCBI, National Institutes of Health, Bethesda, MD, USA), ArrayExpress (European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, Cambridgeshire, UK), and other less common websites (last searched on July 15, 2024). We also retrieved any clinical and data processing information associated with the samples (metadata; see Supplementary Table) from the repositories, publications, and, at times, the corresponding authors of the publications. Metadata were discarded if they were inconsistent between the respective publication and data repository, unless they had been reported in multiple publications or deposited more than once, in which case we used the latest metadata.

Raw Transcriptomic Data Reprocessing

We conducted all analyses using R version 4.0 on a computing instance with 40 CPU cores and 1.5 terabytes of RAM, running Windows 10.

We read and processed all raw data (.cel files) using the affyPara R package[7] (version 1.42) and a customized chip definition format with the updated probe set definitions (BrainArray[8] ENTREZG version 25). Background correction was performed using the RMA algorithm.[9] Quantile normalization of the data was performed. We used PM-only probes and summarized the data with the medianpolish algorithm. The probes were annotated with their respective gene Entrez identification numbers. Multiple probes for the same gene were summarized using the maxmean method and the collapseRows function in the WCGNA package.[10] Finally, we identified and removed duplicate samples, defined as those with identical values for every gene, resulting in a final dataset of 7,375 samples with 20,360 genes measured in each sample.

Dimensionality Reduction

We reduced the high-dimensional dataset to two dimensions to visualize samples on a two-dimensional plot using FFT-accelerated Interpolation-based t-distributed stochastic neighbor

embedding[11] (FIt-SNE version 1.2.1). The following recommended[12] parameters were used: perplexity = n/100, theta = 0.5, max_iter = 5000, stop_early_exag_iter = 250, exaggeration_factor = 12.0, learning_rate = n/12, initialization = 'pca', where n = the total number of samples, which is 7375.

Training Dataset Development

Clusters were visually identified on the two-dimension FIt-SNE plot primarily. However, the visually identified clusters were confirmed or adjusted using a combination of Ordering Points To Identify Clustering Structure (OPTICS) and the Density-Based Spatial Clustering of Applications with Noise (DBSCAN), which are unsupervised algorithms in the dbscan R package[13] (version 1.2.0). OPTICS' ability to detect clusters of varying density is an advantage over DBSCAN. The following parameters were used for OPTICS: eps = 10, minPts = 3, xi = 0.03. For DBSCAN, the parameters were: eps = 1.75, minPts = 3, k = 2. The identified clusters were visualized using convex hulls on the FIt-SNE plot (Supplementary Figure 1). Samples in the clusters were annotated with their respective diagnoses from the metadata. Each cluster was annotated with its majority diagnosis, and the samples of the majority diagnosis that clustered together on the FIt-SNE plot were incorporated into a subset dataset that we designated as the training dataset. We used it to train classifiers to identify or reclassify the remaining samples.

Sample Classification

Synthetic Balancing of Samples per Diagnosis

As expected, the proportions of samples per diagnosis in the generated training dataset were imbalanced. Therefore, to train machine learning classifiers to predict accurate results, we over-sampled samples within each diagnosis to match the maximum number of samples for a neoplastic diagnosis in the training dataset (n=966). We used the synthetic minority over-

sampling technique[14] in the R package smotefamily (version 1.3.1), setting the K nearest neighbor parameter at 10, or n-1 if the number of samples per diagnosis (n) was less than 10. Using the synthetically balanced core dataset (which contained 50,880 samples), we developed two different machine learning classifiers to predict the diagnoses of the remaining samples. The total number of classes to predict equaled 52 for this multiclass classification.

Machine Learning Classifiers

We generated one classifier using the random forest algorithm (ranger R package[15], version 0.13.1). The algorithm was trained for accuracy by tuning the parameters and performing 5-fold cross-validation with 3 repeats using the caret R package[16] (version 6.0-94). The following parameters and values were used in tuning: num.trees (1000, 2000, 3000, 4000), min.node.size (1, 2, 3, 4, 5, 6, 7, 8, 9), and mtry (142, 170, 175, 180, 200). After tuning, the classifier was developed using these final values of the parameters: num.trees=1000, min.node.size=5, mtry=175, splitrule = ‘gini’, and metric=‘accuracy’.

Next, we generated another classifier using a high-performance gradient boosting framework for decision tree-based learning algorithms called LightGBM[17] (lightgbm R package, version 4.3.0). The synthetically balanced core dataset was split into training and testing datasets (85% and 15%, respectively), balancing the diagnosis of the samples for training. The algorithm was trained to minimize the multiclass classification error rate in the test dataset by tuning the following parameters : boosting (‘gbdt’ or ‘dart’), number of iterations (100, 200, 300), number of leaves (30, 40, 50, 60), learning rate (0.1, 0.05, 0.01), and maximum bin (255, 375, 500). After tuning, the classifier was developed using these final parameters: boosting=’gbdt’, num_iterations=100 with early stopping if 10 sequential iterations did not reduce the error rate, learning_rate=0.1, num_leaves=30, and max_bin=255.

Sample Proximity Classifiers

We developed two simple proximity-based classifiers. Mathematically calculated proximity is logically appealing, such as in a dendrogram of hierarchically arranged samples based on a measure of distance. However, in this case, clustering depends on the subjective selection of a cut point on a dendrogram. To address this, we calculated the shortest Euclidian distance between one sample and another in the training dataset using the expression values of all the genes measured (n=20,360). The maximum value for this set of shortest distances was used as a threshold for the proximity-based classifier, whereby the predicted diagnosis of a test sample was the diagnosis of the sample closest in distance that is below this proximity threshold.

Finally, using a similar approach to the proximity-based classifier, we leveraged FIt-SNE-based clustering to predict the diagnosis of samples. We calculated the linear distance between a test sample and all samples in the training dataset using the two FIt-SNE coordinates. The predicted diagnosis of a test sample was the diagnosis of the sample closest in distance, thus assigning the test sample to the cluster noted on the two-dimensional FIt-SNE plot of the training dataset.

RESULTS

We completed the transcriptomic atlas in two steps. First, we harmonized and integrated publicly available raw transcriptomic microarray data generated using the Applied Biosystems™ (previously Affymetrix®) GeneChip™ Human Genome U133 Plus 2.0 Array. Most of the data were generated using fresh frozen samples from living individuals (74.8%); 19.8% of the samples used were postmortem. We simultaneously processed and summarized the raw transcriptomic microarray data using the standard RMA algorithm.[9]

Accurate annotation of the samples with a diagnosis according to the current classification was the second step in creating the atlas. This was necessary because we integrated samples

processed from the year 2003 to 2018 (the raw data have a date and time stamp). Many samples had histopathological diagnoses that did not match the primary diagnosis of their cluster; their membership to a cluster was visually unclear on the FIt-SNE plot; or, they lacked a histopathological diagnosis altogether. During the time span of the samples, there have been changes in the diagnostic criteria for nervous system neoplasms. New diagnostic entities have been discovered, some have been deprecated, and the neuro-oncology community has shifted from histopathological to methylation-based diagnosis for its improved accuracy. The misdiagnosis rate of histopathological-based diagnosis can be as high as 12%.[3]

To complete this challenge, we first created a training dataset using a subset of the samples to train models for predicting the questionable diagnoses of the remaining samples. The training dataset included the majority of samples within a cluster identified on the FIt-SNE plot with a diagnosis according to the latest classification. In the absence of unequivocal 'truth', this method is acceptable and was used to validate the DNA methylation-based classifier for nervous system neoplasms.[18] Of the 7,375 samples, 5,248 samples were selected to be part of the training dataset (Figure 1) to predict or reclassify the diagnosis of the remaining 2,127 samples. The training dataset was used to develop four different classifiers to reclassify the remaining samples. Multiple classifiers using different approaches were chosen to ensure precision. These included the random forest, gradient boosting, FIt-SNE, and a Euclidean-based proximity classifier that classifies test samples based on the diagnosis of the closest sample in the training dataset according to the transcriptome. Cross-validation of these four classifiers using the training dataset showed high discriminating power with accuracies of 100% (95% confidence interval, CI, 99.93% to 100%) for the random forest classifier; 99.98% (95% CI, 99.89% to 100%) for the gradient boosting classifier; 99.92% (95% CI, 99.80% to 99.98%) for the FIt-SNE-based classifier; and, 99.58% (99.37% to 99.74%) for the Euclidean-based

proximity classifier (Supplementary Figure 2). Using these methods, the diagnoses of the 2,127 samples were predicted.

The four classifiers predicted the diagnoses of 1,893 of the 2,127 samples (89%) consistently, defined by prediction of the same diagnosis by 3 or all classifiers. The same diagnosis was predicted by all four classifiers in 1,495 samples and by three of the four classifiers in 398 samples. 155 samples (7.3%) were reclassified with the diagnoses predicted by two classifiers because their histopathological diagnosis (an expert's assessment) was also consistent with the predicted diagnoses. If the histopathological diagnosis was outdated, consistency of the predicted diagnoses with the general histopathological diagnosis (e.g., medulloblastoma or atypical teratoid rhabdoid tumor without further specification, such as SHH) or the diagnostic class of the histopathological diagnosis (e.g., 'Embryonal Tumors' or 'Diffuse Gliomas') was accepted. In 65 samples, the histological assessment supported two different diagnoses predicted by two different classifiers. In these rare cases, the diagnoses predicted by the t-SNE-based classifier was chosen to facilitate visual interpretation of the final results (Figure 2). For 79 samples (3.7%), a final diagnosis could not be confidently assigned because we could not resolve inconsistencies in the predicted diagnoses (Supplementary Figure 3).

Understanding how machine learning algorithms predict diagnoses can be highly beneficial. For instance, recent efforts to decipher how the random forest-based classifier,[3] trained on bulk genome-wide DNA methylation profiles, predicts brain tumor diagnoses have provided valuable biological insights.[19,20] Analyzing the different diagnoses frequently predicted for a given sample by the classifiers may reveal nuanced biological relationships between these diagnoses. Therefore, we examined the inconsistent predictions of diagnosis by four classifiers across 632 samples. Each sample received four predictions, resulting in six pairs of predictions per sample, totaling 3,792 pairs among the 632 samples. After excluding pairs with the same predicted

diagnosis, 2,349 inconsistent pairs remained, representing 188 unique pairs where one classifier predicted one diagnosis and another classifier predicted a different diagnosis. The frequency with which a diagnosis is predicted by a classifier can depend on the number of unique samples with that diagnosis used to train the classifier. Synthetically balancing the number of samples per diagnosis may not completely overcome this limitation. Therefore, we normalized the rate of a diagnosis predicted in the inconsistent pairs with the proportion of samples with that diagnosis in the training dataset. Figure 3A displays the normalized values, indicating how frequently a given diagnosis was predicted by a classifier beyond what would be expected based on its incidence in the training dataset, particularly when there is uncertainty in the predicted diagnosis. Diagnoses such as gangliоglioma, desmoplastic infantile gangliоglioma, angiocentric glioma, and pleomorphic xanthoastrocytoma were predicted more frequently than expected under uncertain conditions. When one classifier predicted these diagnoses, Figure 3B shows the other diagnoses predicted by another classifier (i.e., the inconsistent pairs). It reveals that samples classified as gangliоglioma, desmoplastic infantile gangliоglioma, and angiocentric glioma by one classifier were most likely to be classified as pilocytic astrocytoma or diffuse glioma by another. Similarly, samples classified as pleomorphic xanthoastrocytoma by one classifier were most likely to be classified as diffuse glioma by another.

Last, to increase the utility of the dataset, we manually curated and harmonized the publicly available clinical metadata associated with the samples. Our dataset is broadly representative across geographical regions (Figure 4A), spans a wide age spectrum (from fetus to a maximum age of 106 years; Figure 4B), and covers various anatomical locations (Figure 4C). When available, we also extracted relevant genetic information associated with the samples (Figure 4D) and overall survival (Figure 4E-N) of patients from whom the samples were collected. These and other data are tabulated in Supplementary Table.

DISCUSSION

This study harmonized publicly available data to create a large-scale, clinically annotated transcriptomic dataset comprising 5,402 samples of nervous system neoplasms and 1,973 non-neoplastic nervous system samples. Not needing to adjust for surrogate variable effects (similar to “batch” effects) greatly facilitated this work. Initially, we anticipated significant surrogate variable effects with sample clustering driven primarily by surrogate variables like time, place of origin, and sample collection/preparation (fresh frozen or formalin-fixed paraffin-embedded) rather than diagnosis reflecting the underlying biology. However, we found that standard simultaneous processing of the raw transcriptomic microarray data (using the RMA algorithm[9]) resulted in sample clustering primarily based on diagnosis rather than surrogate variables. Specifically, as we included more samples, the clustering of the samples became increasingly diagnosis-based. This motivated us to conduct a thorough search to include as many eligible samples as possible and to delay our work until the frequency of new publicly available and eligible raw data significantly decreased. To confirm that standard processing of a large amount of raw transcriptomic microarray data led to primarily biologically driven data, we applied established methods (such as ComBat[21] and SVA[22]) to reduce the effect of any surrogate variables (year, place, or preparation/collection of the samples). Applying these methods to either raw intensity values or summarized gene expression values resulted in a drastic loss of diagnosis-based clustering (data not shown). This rarely studied behavior of genomic microarray data was leveraged to integrate methylation profiles of nervous system tumors and to develop a clinically used classifier that does not need to adjust for any surrogate variable effect to allow the emergence of underlying biology for accurate diagnosis of neoplasms.[3,18] We noticed this behavior for the first time in transcriptomic data generated from microarray platform, which greatly facilitated our integration.

The dataset we generated is high-dimensional and nearly comprehensive in its coverage of various nervous system neoplasms. Machine learning tools used for visualization reveal that the transcriptomic signatures of neoplastic and non-neoplastic samples are diagnostically unique. We leveraged this uniqueness to reclassify samples with unknown, obsolete, or questionable diagnoses. The study demonstrates that the diagnostic distinctiveness of bulk DNA methylation signatures also extends to gene expression across a broad range of nervous system neoplasms and age groups. Additionally, the neuro-oncology community can use this dataset for comparative gene expression analysis among nervous system neoplasms, such as elucidating differences between rare and common neoplasms.

The strengths of this large-scale, harmonized, high-dimensional dataset are best illustrated using a dimensionally reduced plot with the unsupervised FIt-SNE algorithm. It is crucial to note that the linear distance between distant clusters on the FIt-SNE plot does not necessarily correlate with biological differences. For instance, it is incorrect to conclude that retinoblastoma is more biologically different from neuroblastoma than medulloblastoma based solely on the linear distance between these clusters on the FIt-SNE plot. However, the linear distance is meaningful over closer distances or within a cluster, i.e., the strength of its correlation with biological difference is greater. With this understanding, the proximity of clusters revealing known biological relationships makes the dataset reliable for comparative gene expression analyses. Several relationships between non-neoplastic and neoplastic entities are highlighted: non-neoplastic supratentorial white matter and gliomas, choroid plexus and choroid plexus papillomas, pituitary and pituitary neuroendocrine tumors, peripheral nerve and neurofibromas, retina and retinoblastomas. Non-neoplastic fetal tissues are closely associated with their primitive neuroectodermal neoplasms, such as fetal retina and retinoblastoma and fetal cerebellum and medulloblastoma (the SHH subgroup). The biological similarity of the following pairs of neoplasms is also reflected by the proximity of their clusters: meningiomas and solitary

fibrous tumors, subgroups of medulloblastomas, pituitary neuroendocrine tumors and peripheral neuroendocrine tumors like pheochromocytomas/paragangliomas, neuroblastoma and ganglioneuroma, and embryonal tumors with multilayered rosettes and atypical teratoid/rhabdoid tumors. Recently, a close biological relationship between high-grade neuroepithelial tumors with MN1 alteration and ependymal layer tumors was demonstrated,[23] and this is also evident in the FIt-SNE plot. The maintenance and revelation of these expected biological relationships after harmonized integration of samples, despite their diverse geographic origins, acquisition times, and preparation/collection methods, is a notable strength of our methods and the generated dataset.

The dataset has a few notable limitations. First, although diagnosis-based clustering is primarily observed, the effects of surrogate variables, especially the place of origin of the data, are notable within some clusters. This is particularly evident in large clusters like non-neoplastic central nervous system and diffuse gliomas, but also in smaller clusters like retinoblastoma. Therefore, this dataset should not be used to identify new subgroups within known diagnostic entities or their well-established subgroups, such as group 3 and group 4 medulloblastomas or the different subgroups of ependymomas and atypical teratoid/rhabdoid tumors. This also applies to subgroup analyses of the diffuse glioma cluster according to histopathological grade. However, this limitation can also be a strength. The presence of any surrogate variable effect, despite our harmonized integration, helps decrease false positive results in comparative gene expression analyses using this dataset. This effect represents real-world variability in gene expression and broadly captures the transcriptomic heterogeneity within diagnostic entities. Second, the high accuracies of our classifiers may reflect an over-fitting phenomenon. Although, we jointly used them to increase the precision of predictions, the predicted diagnoses were not consistent among the classifiers in 3.7% of the samples. The classifiers inconsistently predicted many discrete gliomas, especially ganglioglioma and desmoplastic infantile ganglioglioma,

which were also likely to be predicted as pilocytic astrocytoma. In the original work that developed the DNA methylation-based classifier of central nervous system tumors, there were supratentorial samples designated as pilocytic astrocytoma/ganglioglioma (labeled PA/GG ST), which were acknowledged as a “combined”, non-equivalent diagnosis with respect to the World Health Organization classification scheme.[3] In light of this data from the prior methylation-focused work,[3] our gene expression-based work supports further investigations into the biological differences between pilocytic astrocytomas and gangliogliomas to refine their classification. Third, despite this work, we believe that a machine learning classifier using transcriptomic data generated with a microarray transcriptomic platform (like GPL570) to diagnose nervous system neoplasms is far from being ready for clinical use. Currently, the practical ease of retrieving and handling DNA makes a DNA methylation-based classifier using a microarray DNA methylation platform more suitable for clinical use.

Conclusion

We created a transcriptomic dataset through harmonized integration of publicly available 5,402 neoplastic and 1,973 non-neoplastic samples of the nervous system. This dataset covers a wide range of diagnoses, including rarely studied entities, and includes clinical data such as age, sex, tumor location, genetic information, and survival rates for many samples, enhancing its utility. Reclassifying many samples according to the latest classification increases the value of older samples and boosts the statistical power of comparative gene expression analyses. By using publicly available data, we incorporated samples from around the world, broadening the dataset’s ethnic representation. Finally, our methodological workflow can be used to integrate and harmonize existing raw data of other rare diagnoses and conditions generated on GPL570, increasing their utility and informing future research.

FUNDING: Kentucky Pediatric Cancer Research Trust Fund (to AMM); Vanderbilt Institute for Clinical and Translational Research (VR52534 to AMM).

CONFLICT OF INTEREST: None declared.

AUTHORSHIP

Conception: AMM

Design of the work: AMM, CL

Acquisition of data: AMM, JRN

Data analysis: AMM, CL, AKN

Data interpretation: AMM, CL, AKN, EMH, RPN, BJW

Drafting the work: AMM

Reviewing it critically for important intellectual content: AMM, CL, JRN, AKN, EMH, RPN, BJW

Final approval of the version to be published: AMM, CL, JRN, AKN, EMH, RPN, BJW

Agreement to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved: AMM

DATA AVAILABILITY: Raw data (.cel files) are available from the primary source listed in Supplementary Table. Final processed transcriptomic data in the format of an .rds file to be loaded directly into R can be downloaded from the link in the github page https://github.com/axitamm/BrainTumorAtlas. The codes that reproduce the dataset from the raw data, generate the classifiers using the 'core' dataset, use the classifiers to diagnose samples in the 'test' dataset, and make the figures in this manuscript are also available from the github page.

ACKNOWLEDGEMENTS: None

REFERENCES

1. Dragomir MP, Calina TG, Perez E, et al. DNA methylation-based classifier differentiates intrahepatic pancreato-biliary tumours. *EBioMedicine.* 2023; 93: 104657.
2. Jurmeister P, Gloss S, Roller R, et al. DNA methylation-based classification of sinonasal tumors. *Nat Commun.* 2022; 13(1): 7148.
3. Capper D, Jones DTW, Sill M, et al. DNA methylation-based classification of central nervous system tumours. *Nature.* 2018; 555(7697): 469-474.
4. Technical Note: Design and Performance of the GeneChip® Human Genome U133 Plus 2.0 and Human Genome U133A 2.0 Arrays. 2003; https://assets.thermofisher.com/TFS-Assets%2FLSG%2Fbrochures%2Fhgu133_p2_technote.pdf. Accessed July 31, 2024.
5. Lakiotaki K, Vorniotakis N, Tsagris M, Georgakopoulos G, Tsamardinos I. BioDataome: a collection of uniformly preprocessed and automatically annotated datasets for data-driven biology. *Database (Oxford).* 2018; 2018.
6. Guo Y, Sheng Q, Li J, Ye F, Samuels DC, Shyr Y. Large scale comparison of gene expression levels by microarrays and RNAseq using TCGA data. *PLoS One.* 2013; 8(8): e71462.
7. Schmidberger M, Vicedo E, Mansmann U. affyPara-a Bioconductor Package for Parallelized Preprocessing Algorithms of Affymetrix Microarray Data. *Bioinform Biol Insights.* 2009; 3: 83-87.
8. Dai M, Wang P, Boyd AD, et al. Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Res.* 2005; 33(20): e175.
9. Irizarry RA, Hobbs B, Collin F, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics.* 2003; 4(2): 249-264.
10. Miller JA, Cai C, Langfelder P, et al. Strategies for aggregating gene expression data: the collapseRows R function. *BMC Bioinformatics.* 2011; 12: 322.

11. Linderman GC, Rachh M, Hoskins JG, Steinerberger S, Kluger Y. Fast interpolation-based t-SNE for improved visualization of single-cell RNA-seq data. *Nat Methods.* 2019; 16(3): 243-245.

12. Kobak D, Berens P. The art of using t-SNE for single-cell transcriptomics. *Nat Commun.* 2019; 10(1): 5416.

13. Hahsler M, Piekenbrock M, Doran D. dbscan: Fast Density-Based Clustering with R. *Journal of Statistical Software.* 2019; 91(1): 1-30.

14. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research.* 2002; 16: 321-357.

15. Wright MN, Ziegler A. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software.* 2017; 77(1): 1-17.

16. Kuhn M. Building predictive models in R using the caret package. *Journal of Statistical Software.* 2008; 28: 1-26.

17. Ke G, Meng Q, Finley T, et al. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems.* 2017; 30.

18. Maros ME, Capper D, Jones DTW, et al. Machine learning workflows to estimate class probabilities for precision cancer diagnostics on DNA methylation microarray data. *Nat Protoc.* 2020; 15(2): 479-512.

19. Benfatto S, Hovestadt V. shinyMNP. 2023; https://hovestadtlab.shinyapps.io/shinyMNP/ or https://github.com/hovestadt/shinyMNP. Accessed Aug 1, 2024.

20. Benfatto S, Hovestadt V. METB-10. Explainable Artificial Intelligence Reveals Dna Methylation Patterns Underlying Brain Tumor Classification. *Neuro Oncol (Society of Neuro-Oncology).* 2023; 25(Suppl 1): i32.

21. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics.* 2007; 8(1): 118-127.

22. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics.* 2012; 28(6): 882-883.

23. Lehman NL, Spassky N, Sak M, et al. Astroblastomas exhibit radial glia stem cell lineages and differential expression of imprinted and X-inactivation escape genes. *Nat Commun.* 2022; 13(1): 2083.

24. Tauziede-Espariat A, Siegfried A, Nicaise Y, et al. Supratentorial non-RELA, ZFTA-fused ependymomas: a comprehensive phenotype genotype correlation highlighting the number of zinc fingers in ZFTA-NCOA1/2 fusions. *Acta Neuropathol Commun.* 2021; 9(1): 135.

FIGURES and FIGURE CAPTIONS

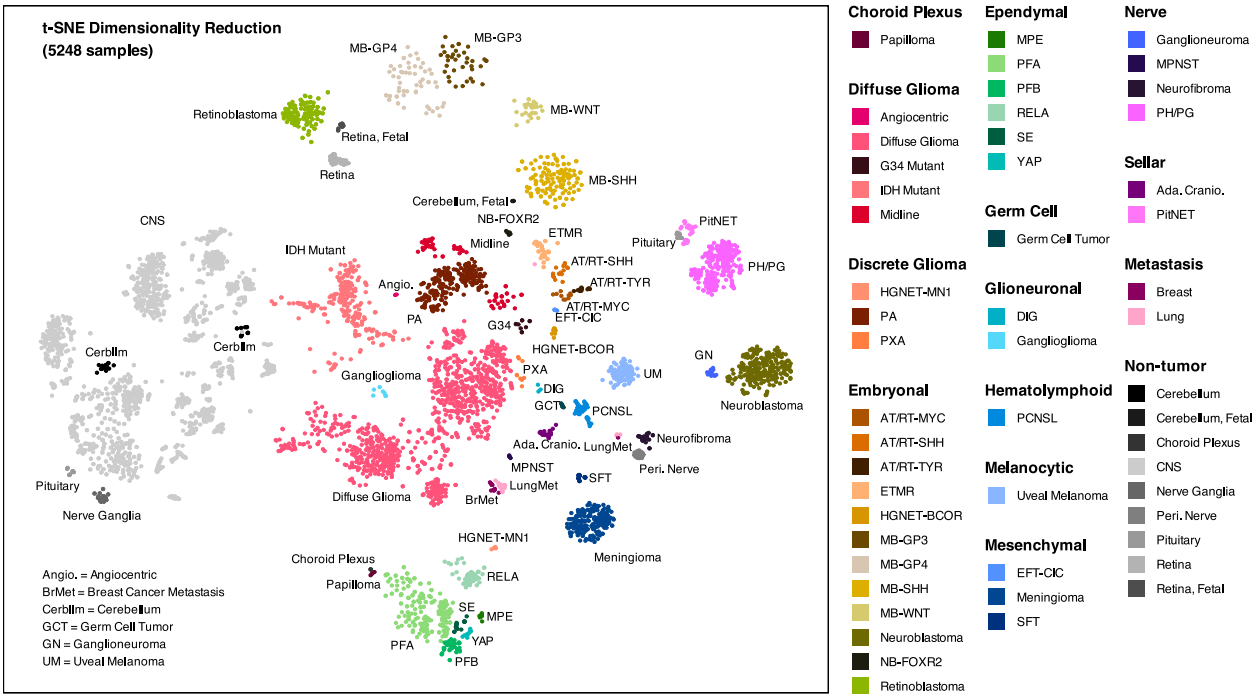

Figure 1. Representation of the training dataset (5,248 samples) in the t-SNE dimensionality reduction performed on the full dataset. Individual samples (dots) are color-coded and labelled according to the diagnosis listed in the side legend. Full names of the 52 diagnostic entities are provided in Supplementary Table. Of note, we chose to label a specific type of supratentorial ependymomas as “RELA” instead of the most recent nomenclature “ZFTA fusion-positive” because there are non-RELA, ZFTA-fused ependymomas[24] and these samples were identified as RELA-fusion positive.

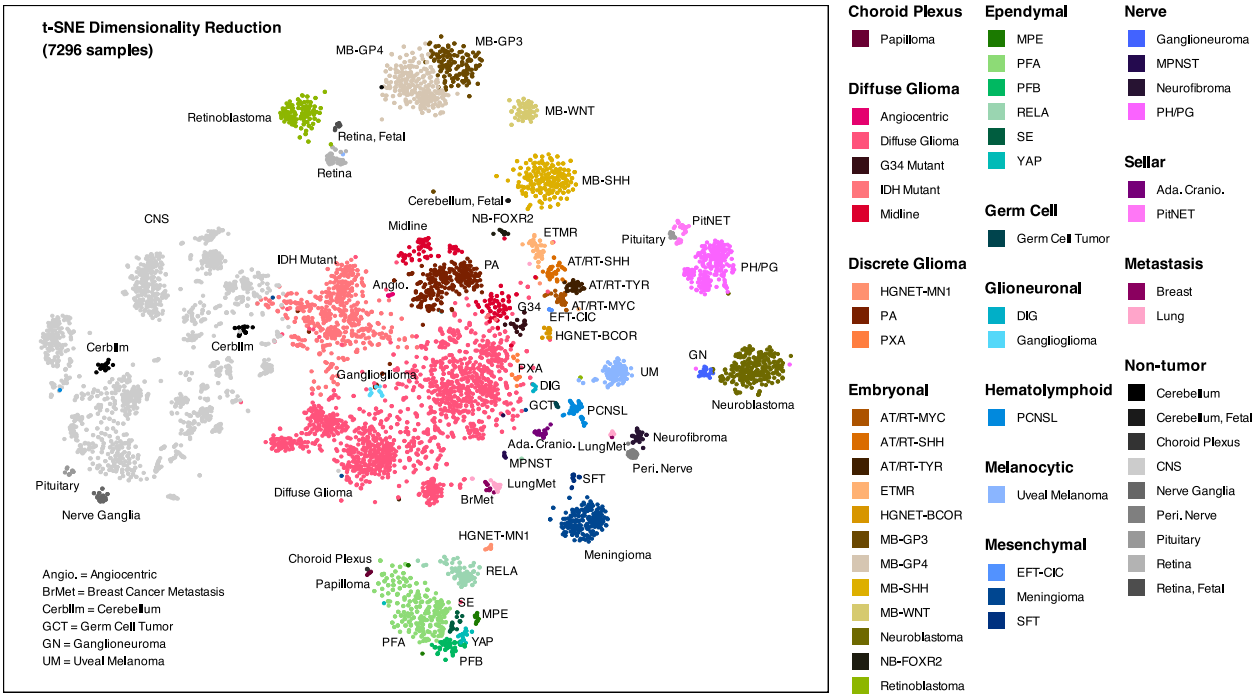

Figure 2. Representation of all samples (5,248 samples in the training dataset and 2,048 samples with an uncertain diagnosis that are reclassified) in the t-SNE dimensionality reduction performed on the full dataset. Individual samples (dots) are color-coded and labelled according to the diagnosis listed in the side legend. Full names of the 52 diagnostic entities are provided in Supplementary Table. Of note, we chose to label a specific type of supratentorial ependymomas as “RELA” instead of the most recent nomenclature “ZFTA fusion-positive” because there are non-RELA, ZFTA-fused ependymomas[24] and samples used in the training dataset were identified as RELA-fusion positive.

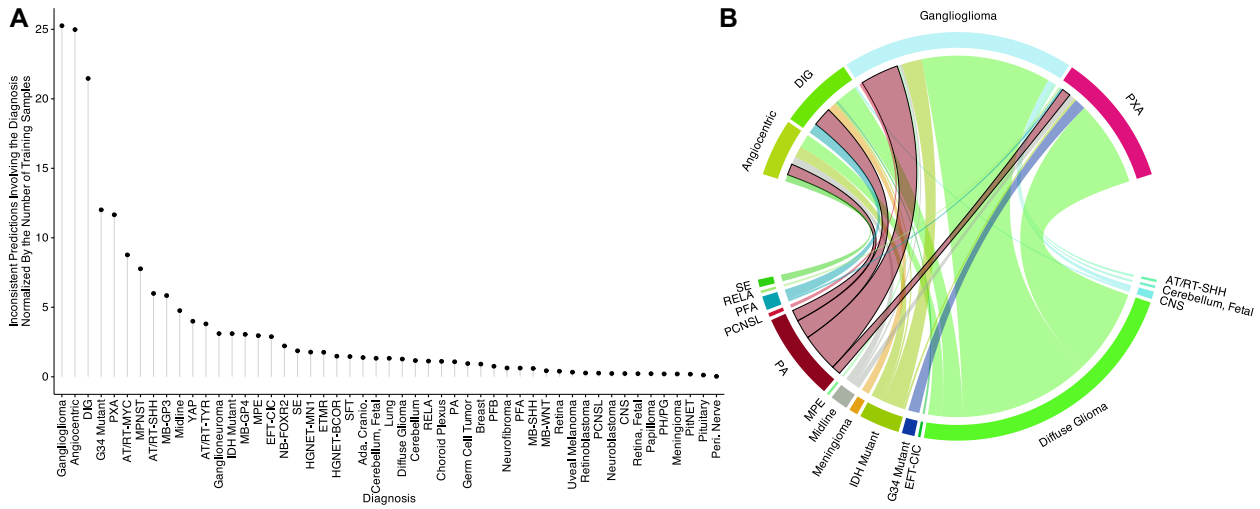

Figure 3. In the 632 samples with non-unanimous predictions among the classifiers, (**A**) the frequency of each diagnosis prediction is normalized to its proportion in the 'core' dataset used for training the classifiers. (**B**) When one classifier predicted a sample as ganglioglioma, desmoplastic infantile ganglioglioma, angiocentric glioma, or pleomorphic xanthoastrocytoma, this chord diagram illustrates the other diagnoses predicted by another classifier, highlighting pilocytic astrocytoma with a black border. Full names of the 52 diagnostic entities are provided in Supplementary Table.

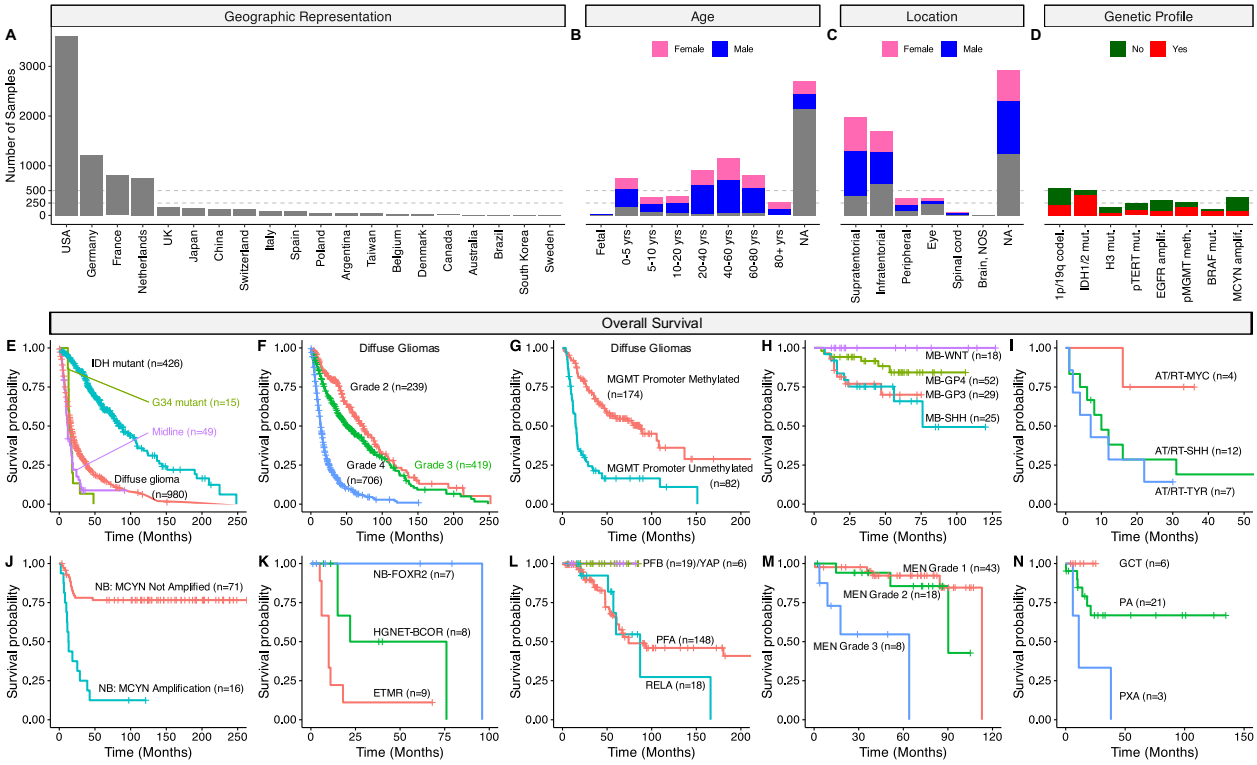

Figure 4. Manually curated metadata and clinical data associated with the 7,375 samples in the full dataset. The number of samples per (**A**) country listed in the contact information of the deposited raw data, (**B**) age group, and (**C**) anatomic location. The sex of the samples is color-coded (pink/blue or grey if not available) in the latter two plots. (**D**) The number of samples with associated genetic information. (**E-N**) Overall survival, visualized using Kaplan-Meier curves, based on the final diagnosis. **E-G** survival curves are of diffuse gliomas according to different subgroups. Tick marks on the curves represent censored values. Full names of the diagnostic entities are provided in Supplementary Table. Abbreviations: amplif = amplification; codel = co-deletion; meth = methylation; mut = mutation; NA = not available; NOS = not otherwise specified; pTERT/pMGMT = promoter of the respective gene; yrs = years.