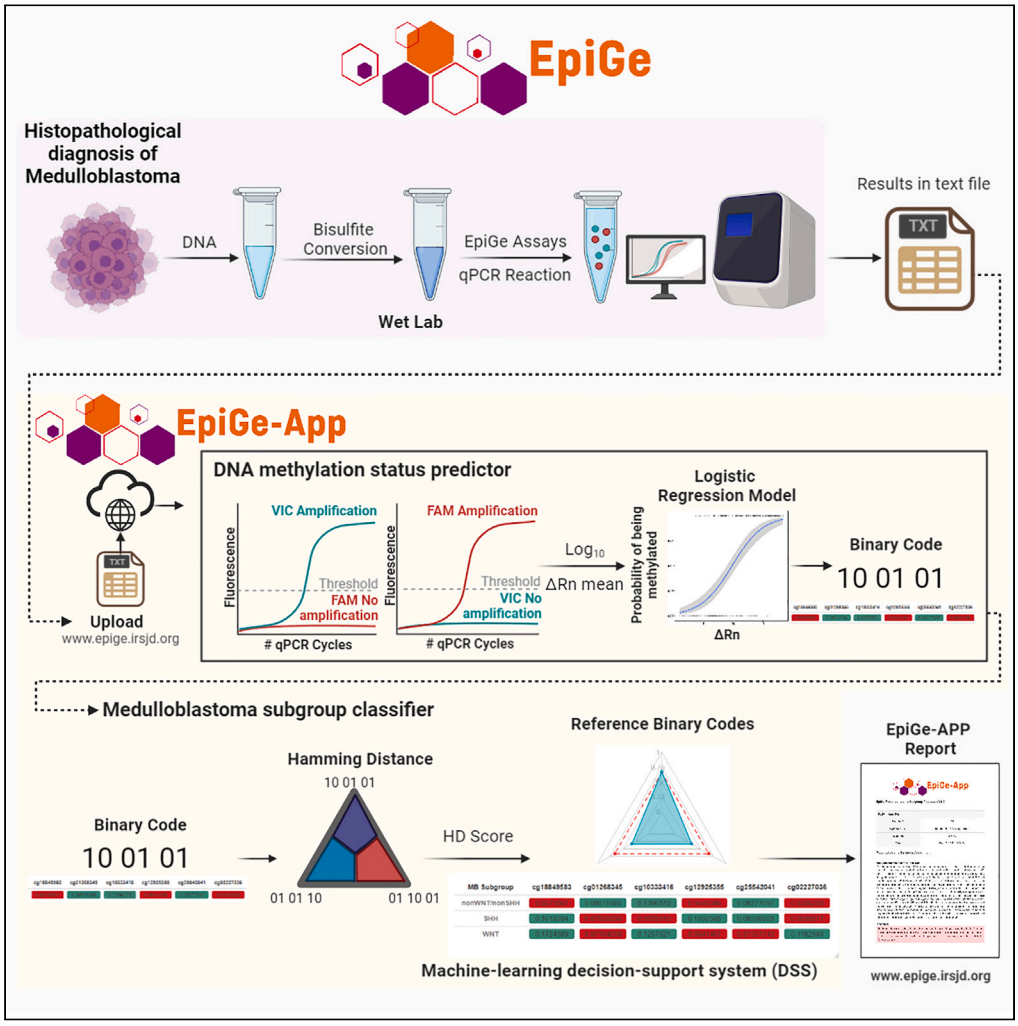
iScience

CellPress
OPEN ACCESS

Article

EpiGe: A machine-learning strategy for rapid classification of medulloblastoma using PCR-based methyl-genotyping



Soledad Gómez-González, Joshua Llano, Marta Garcia, ..., Andrés Morales La Madrid, Alexandre Perera, Cinzia Lavarino

cinzia.lavarino@sjd.es

Highlights

A PCR-based machine-learning strategy for medulloblastoma subgroup classification

MB subgroup-specific DNA methylation profiles analyzed using PCR methyl-genotyping

Automated website analysis of PCR methylation data predicts MB molecular subgroups

EpiGe, a comprehensive approach for rapid classification of MB principal subgroups

Gómez-González et al., iScience 26, 107598 September 15, 2023 © 2023 The Author(s).
https://doi.org/10.1016/j.isci.2023.107598

iScience

CellPress
OPEN ACCESS

Article

EpiGe: A machine-learning strategy for rapid classification of medulloblastoma using PCR-based methyl-genotyping

Soledad Gómez-González,[1,2,13,14] Joshua Llano,[2,3,4,13] Marta Garcia,[1,2] Alicia Garrido-Garcia,[1,2] Mariona Suñol,[5] Isadora Lemos,[6] Sara Perez-Jaume,[1,2] Noelia Salvador,[6] Nagore Gene-Olaciregui,[6] Raquel Arnau Galán,[5] Vicente Santa-María,[7] Marta Perez-Somarriba,[9] Alicia Castañeda,[10] José Hinojosa,[11] Ursula Winter,[12] Francisco Barbosa Moreira,[12] Fabiana Lubieniecki,[12] Valeria Vazquez,[12] Jaume Mora,[1,10] Ofelia Cruz,[7] Andrés Morales La Madrid,[7] Alexandre Perera,[2,3,4] and Cinzia Lavarino[1,2,6,*]

SUMMARY

Molecular classification of medulloblastoma is critical for the treatment of this brain tumor. Array-based DNA methylation profiling has emerged as a powerful approach for brain tumor classification. However, this technology is currently not widely available. We present a machine-learning decision support system (DSS) that enables the classification of the principal molecular groups—WNT, SHH, and non-WNT/non-SHH—directly from quantitative PCR (qPCR) data. We propose a framework where the developed DSS appears as a user-friendly web-application—EpiGe-App—that enables automated interpretation of qPCR methylation data and subsequent molecular group prediction. The basis of our classification strategy is a previously validated six-cytosine signature with subgroup-specific methylation profiles. This reduced set of markers enabled us to develop a methyl-genotyping assay capable of determining the methylation status of cytosines using qPCR instruments. This study provides a comprehensive approach for rapid classification of clinically relevant medulloblastoma groups, using readily accessible equipment and an easy-to-use web-application.

INTRODUCTION

DNA methylation–based machine learning algorithms represent an extremely useful diagnostic tool for brain tumor classification. The stability and specificity of DNA methylation tumor signatures, together with the high robustness of methylation on the DNA molecule, make methylation data suitable for the development of machine learning–based brain tumor classifiers.[1–3] A paradigmatic example is the molecular classification of medulloblastoma, the most common pediatric malignant brain tumor. Four principal subgroups of medulloblastoma have been described, which are characterised by distinct epigenetic and genetic profiles, and as well as differing clinical courses.[4–18] These groups are currently represented in the 2021 WHO classification of Central Nervous System tumors as four primary, clinically relevant groups: WNT, SHH-TP53 mutated, SHH-TP53-wildtype, and non-WNT/non-SHH tumors.[19] Patients with WNT medulloblastoma have an excellent prognosis with current therapy schemes (5-year event-free survival greater than 90%) and are currently considered for controlled reduction of treatment. The prognosis of SHH-activated medulloblastomas is largely dependent on patient's age and specific genetic features, where children with *TP53* mutated SHH tumors have poorer outcome. Subgroup driven clinical trials are currently being conducted aimed at the assessment of the efficacy of SHH pathway inhibitors (e.g., vismodegib) at diagnosis. Patients with non-WNT/non-SHH group have the most unfavourable prognosis, especially when associated with *MYC* amplification.[7,20–22]

[1]Laboratory of Developmental Tumor Biology, Institut de Recerca Sant Joan de Déu, Pediatric Cancer Center Barcelona, Hospital Sant Joan de Déu, Barcelona, Spain
[2]Institut de Recerca Sant Joan de Déu, Esplugues de Llobregat, Barcelona, Spain
[3]B2SLab, Departament d'Enginyeria de Sistemes, Automàtica i Informàtica Industrial, Universitat Politècnica de Catalunya, Barcelona, Spain
[4]Networking Biomedical Research Centre in the Subject Area of Bioengineering, Biomaterials and Nanomedicine (CIBER-BBN), Madrid, Spain
[5]Department of Pathology, Hospital Sant Joan de Déu, Barcelona, Spain
[6]Laboratory of Molecular Oncology, Pediatric Cancer Center Barcelona, Hospital Sant Joan de Déu, Barcelona, Spain
[7]Neuro Oncology Unit, Pediatric Cancer Center Barcelona, Hospital Sant Joan de Déu, Barcelona, Spain
[9]Children & Young People's Unit, The Royal Marsden NHS Foundation Trust, London, UK
[10]Pediatric Solid Tumor Unit, Pediatric Cancer Center Barcelona, Hospital Sant Joan de Déu, Barcelona, Spain
[11]Department of Neurosurgery, Hospital Sant Joan de Déu, Barcelona, Spain
[12]Department of Pathology, Pediatric Hospital S.A.M.I.C. Prof. Dr. Juan P. Garrahan, Buenos Aires, Argentina
[13]These authors contributed equally
[14]Lead contact
*Correspondence: cinzia.lavarino@sjd.es
https://doi.org/10.1016/j.isci.2023.107598

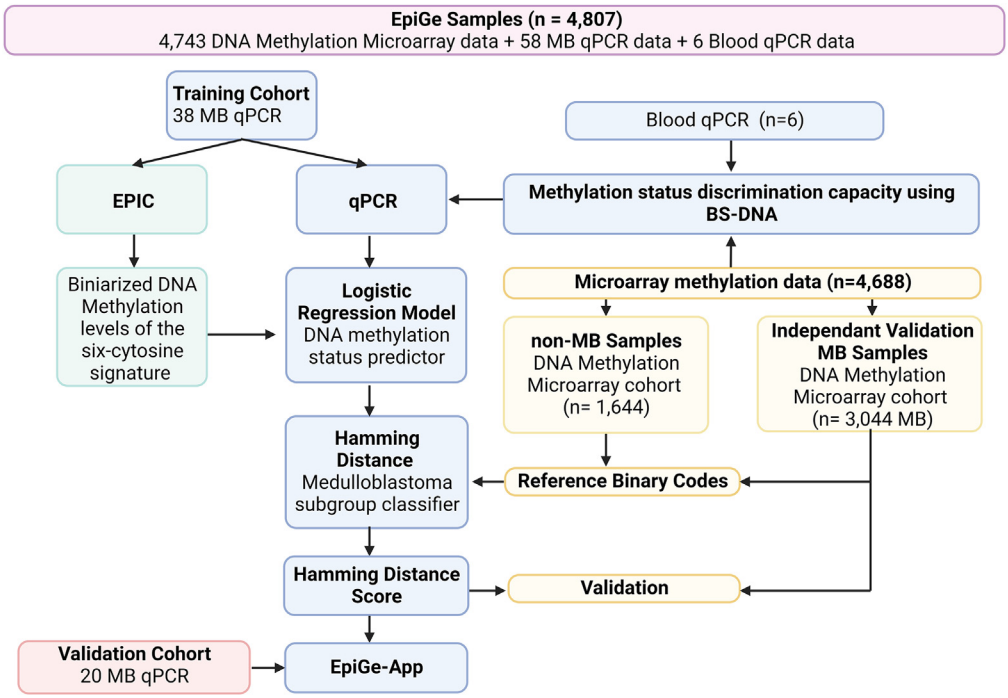


Figure 1. Patient flow diagram

CNS, central nervous system; DSS, decision support system; EPIC, Illumina methylation EPIC BeadChip array; HM450K, Illumina Infinium HumanMethylation 450 BeadChip; MB, medulloblastoma; qPCR, quantitative PCR. Images created with BioRender.

Molecular subgrouping of medulloblastoma tumors has become increasingly important in routine diagnosis, risk stratification and selection of patients eligible for subgroup-specific treatment. Genome-wide DNA methylation-based profiling is currently considered a gold standard for the classification of these molecular subgroups of medulloblastoma. However, the application of array-based technology in a routine diagnostic setting can be time consuming, costly, and sometimes inaccessible for many centers worldwide that treat patients with brain tumors. Consequently, a significant number of patients cannot benefit from the clinical advances associated with methylation-based medulloblastoma classification.

We recently developed an epigenetic classifier based on the methylation profile of a six-cytosine signature that allows for classification of medulloblastoma into the clinically relevant subgroups of WNT, SHH and non-WNT/non-SHH, with an accuracy (of 99% concordance) equivalent to genome-wide DNA methylation microarray and gene-signature profiling methods.[19,23] The six-cytosine classifier represents a simplified approach for accurate, rapid, and cost-effective classification of single medulloblastoma DNA samples.

On the basis of our six-cytosine signature, we have now developed a decision support system (DSS) to enable accurate classification of medulloblastoma tumors into the molecular subgroups WNT, SHH, or non-WNT/non-SHH, using a clinically applicable quantitative PCR (qPCR)-based approach. We also built an interactive, user-friendly web application that enables the automated interpretation of qPCR methylation data, defines the methylation status of cytosines, and predicts a molecular subgroup, and reports the methylation class of the medulloblastoma tumor. The overall design of the proposed workflow is shown in the graphical abstract.

RESULTS

The DSS was developed using DNA methylation data from 4,804 samples, comprising 3,157 primary medulloblastoma tumors, 1,613 non-medulloblastoma tumors, and 37 normal tissues (Figure 1). The purpose of the DSS was to enable the automated analysis and interpretation of qPCR methylation data to predict the methylation status of the six-cytosine signature. Two main components compose the DSS: an automated DNA methylation status predictor, and a medulloblastoma molecular subgroup classifier. The DSS was generated using 38 medulloblastoma cases (training cohort) with available methyl-genotyping qPCR data and methylation microarray data of the six-cytosine signature.

PCR-based allelic discrimination assay for the analysis of single CpG sites

Sodium bisulfite conversion of genomic DNA involves deamination of unmethylated cytosines to uracil, leaving methylated cytosines unchanged, enabling the identification of methylated CpG sites (Figure 2B). We tested the capacity of the PCR-based allelic discrimination assay (rhAmp SNP Genotyping System) to discriminate the methylation status of a single cytosine using bisulfite-converted DNA (BS-DNA). As BS controls, we selected two cytosines, cg13458561 (methylated) and cg22885965 (unmethylated), that displayed a stable methylation status

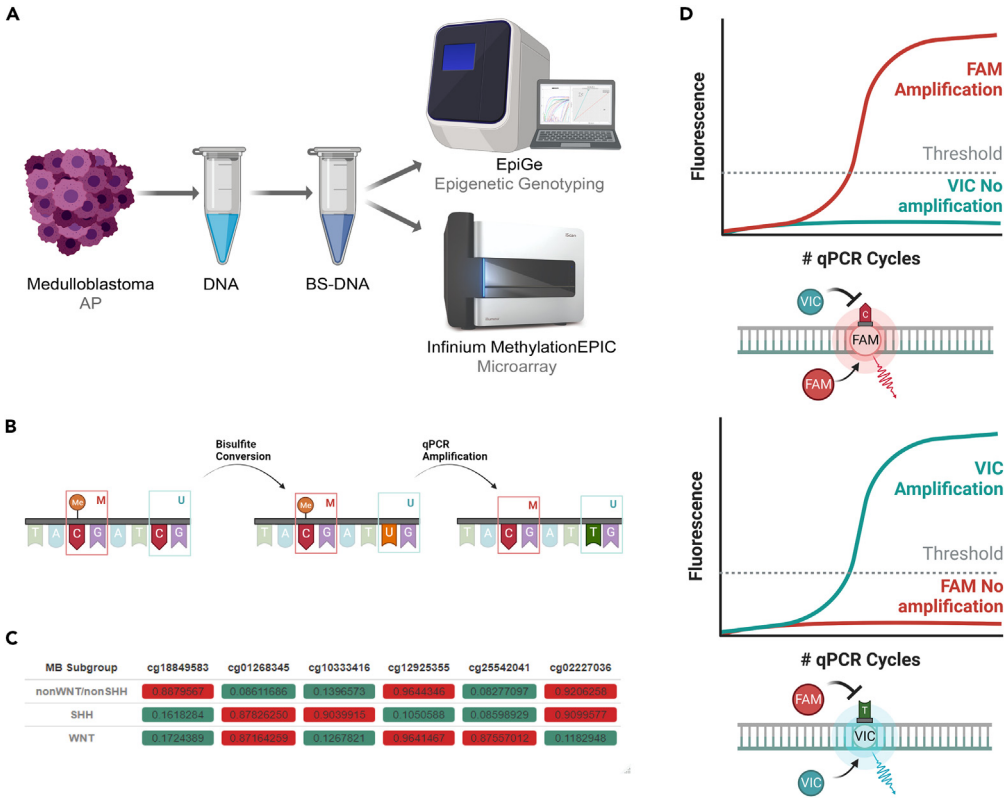


Figure 2. Development and testing of the methyl-genotyping assays

(A) The decision support system (DSS) was generated using 38 medulloblastoma cases (training cohort) with available methyl-genotyping qPCR data (EpiGe) and methylation microarray data (EPIC, Illumina methylation EPIC BeadChip array) of the six-cytosine signature.

(B) Bisulfite conversion (BS) scheme.

(C) The mean DNA methylation values of the six-cytosine signature that accurately discriminate the medulloblastoma subgroups WNT, SHH, and non-WNT/non-SHH (Gómez et al. 2018).

(D) The quantitative PCR (qPCR) amplification curve representation of bisulfite-converted cytosines for methylated cytosines (with FAM labeling) in red, and unmethylated cytosines (with VIC labeling), in blue.

across microarray methylation data obtained from 4,669 samples of different human tumors and normal tissues (Figure S2). rhAmp SNP genotyping assays for both cytosines (cg13458561 and cg22885965) and their corresponding synthetic controls were designed and tested using BS converted and non-converted genomic DNA (gDNA) obtained from peripheral blood (PB) samples of six healthy donors (Tables S1 and S4). Specific amplification was observed for BS-gDNA-PB and the synthetic controls, whereas no amplification was identified in the gDNA-PB samples, indicating that the BS control primers were specific for BS-DNA. This demonstrated that the hybridization of the rhAmp SNP assays to their targets was specific, with no significant non-specific signals (Figure S3; Table S5).

Next, we explored the rhAmp SNP Genotyping System for medulloblastoma classification. To this end, we designed and validated rhAmp SNP methyl-genotyping assays for single CpG site analysis using the six-cytosine signature (cg18849583, cg01268345, cg10333416, cg12925355, cg25542041, cg02227036) with medulloblastoma subgroup-specific differential methylation, which allows for accurate classification of medulloblastoma into the clinically relevant subgroups of WNT, SHH, and non-WNT/non-SHH[23] (Figure 2C). This six-cytosine signature was previously developed and validated by our group using DNA methylation data from 1,576 samples, including medulloblastoma, pediatric brain tumors, and normal tissue[23] (Figure 2C). The six cytosines of the signature are characterized by a bimodal subgroup-specific methylation profile, with a clear methylated or unmethylated status, which enables the analysis by single-nucleotide variant detection methods. The designed genotyping assays to discriminate between single-basepair changes represent opposed methylation states, cytosine (methylated) and thymine (unmethylated) (Figure S4).

A total of 26 EpiGe assays were tested by qPCR using synthetic double-stranded DNA sequences that recapitulated both bimodal methylation states. Regions enriched with CG dinucleotides were avoided for primer design, since conversion of cytosines to thymines after bisulfite treatment could interfere with annealing efficiency. Six EpiGe assays were selected for the robust and reproducible capacity to identify and distinguish between methylated (cytosine) and non-methylated (thymine) states of the cytosines (Table S5). The allelic discrimination plots corresponding to the EpiGe assays and synthetic controls for each of the six cytosines showed compact, non-overlapping, and well-differentiated clusters: methylated synthetic DNA, unmethylated synthetic DNA, and non-template controls (NTC) (Figure S5). These clusters

presented large separation angles between synthetic methylated and unmethylated DNA control clusters[24] (Figure S4B), and both synthetic control clusters clearly separated from the NTC cluster coordinates: the methylated control/NTC FAM signal ratio had a mean of 8.43 (range 8.00–8.72), and the unmethylated control/NTC VIC signal ratio had a mean of 5.08 (range 4.79–5.48) (Table S5).

Detection and discrimination of DNA methylation of medulloblastoma samples using EpiGe assays

To explored the performance of the methyl-genotyping assay using tumor DNA as the starting material, we used DNA obtained from 38 fresh-frozen medulloblastoma biopsies (training cohort), previously classified employing the Infinium MethylationEPIC BeadChip array data (EPIC) and the Molecular Neuropathology Platform brain tumor classifier[2] (https://www.molecularneuropathology.org/mnp) (Table S1; GEO ID: GSE210723). The methylation EPIC microarray data of well-characterized CNS tumors and normal pediatric tissues (210 medulloblastoma, 63 aypical teratoid rhabdoid tumor [ATRT], 37 embryonal tumor with multi-layered rosettes [ETMR], 30 ependymoma [EPN], and 6 normal cerebellum samples) were used for comparative and quality control analyses (Figure S6; Table S4).

Methylation levels of the six-cytosine signature were isolated from the DNA methylation data and used to classify the training cohort according to our previously described classification method[23] (Table S6). All samples could be assigned to a subgroup with an excellent degree of concordance (100% agreement; 95% CI [90.7%–100%]) with array-based DNA methylation profiling classification.

The DNA of the training cohort was BS-converted and analyzed using the genotyping primers (Figure S4A). The allelic discrimination plots of normalized fluorescence intensity of FAM and VIC at the Y- and X axis, respectively, showed four distinct clusters: methylated, unmethylated, hemimethylated, and NTC samples (Figure S4B). By manual calling, 92.98% (212/228) of the analyzed cytosines were assigned to a methylation state (Table S7). The qPCR methylation predictions enabled us to successfully classify 32 medulloblastoma samples (84.21%) with a 100% concordance with DNA methylation profiling classification (95% CI [89.11%–100%], Kappa Cohen (k) = 1). Samples were not classified if more than one cytosine of the panel was assigned to a hemimethylated methylation state (Table S7).

DNA methylation status predictor

We computed the normalized fluorescence qPCR endpoint values (ΔRn) mean and standard deviation[25] values between qPCR replicates, and observed low SD values between replicates (Figures 3A; S7). All cytosines presented a strong ΔRn correlation (R^2 of 0.98 and 0.95 for ΔRn Allele1 and 2, respectively, and p value < 0.01) (Figure 3A). Only 5 cytosines, corresponding to 5 different samples, presented SD > 0.5 in both alleles (Figure S7). The qPCR Allele1 and Allele2 ΔRn mean values in base 10 logarithmic (log_{10}) scale were used as independent values for training the logistic regression model (LRM). As a dependent variable, we used the binarized reference methylation microarray values (Figure 3B). The LRM was validated using LOPOCV, for which the methylation status of each single cytosine was predicted individually, but the set of six-cytosines were computed together to avoid biases or overfitted results. The LRM showed a significant cytosine methylation prediction capacity with an AUC of 0.98 (95% CI [0.96–0.99]). Based on the Youden's Index of 0.88 (95% CI [0.81–0.94]), sensitivity at 94% (95% CI [87.5%–97.3%]), and specificity at 93.8% (95% CI [87.1%–97.2%]), an optimal threshold value of 0.456 for the logistic regression output was obtained (Figure 3C). The performance metrics presented an excellent degree of accuracy (93.9%; 95% CI [89.9%–96.6%], k = 0.88 (95% CI [0.81–0.94])) (Tables S8 and S9).

Medulloblastoma subgroup classifier

To develop the subgroup classifier, we binarized and encoded the DNA methylation microarray data from 3,044 medulloblastoma and 1,644 non-medulloblastoma samples according to a fixed order of the cytosines: cg18849583, cg01268345, cg10333416, cg12925355, cg25542041, and cg02227036 (Figures 1 and S8). For the medulloblastoma tumors, we identified three primary binary codes that represented more than 96% of the cohort: 100101 (77.2%), 011001 (14.7%), and 010110 (4.4%). These codes were found to be subgroup-specific; 100101 categorised 97.1% of non-WNT/non-SHH; 011001, 94.7% of SHH; and 010110, 88.8% of WNT medulloblastomas. The binary codes 100101, 011001, and 010110 were defined as reference binary codes (RefBC) (Figures 3D and 3E; S8; Tables S10 and S11).

We investigated the training cohort based on the methylation status predicted by the LRM (binary code) in order to compute the minimum distance to the three RefBC using the Hamming distance (HD) (Figure S9A). The HD ranged from a 0, indicating that the two codes were identical, to 6, indicating that all the positions were different between the two codes. Samples were assigned to a subgroup if they had the minimum HD, but they were excluded if they had the same minimum HD for multiple RefBCs. Our training cohort was assigned to the nearest subgroup using HD with an excellent degree of accuracy (94.7%, 95% CI [82.3%–99.4%]; k = 0.87, 95% CI [0.71–1]), according to previously reported medulloblastoma classification data (Figures 3F; S6; Tables S9 and 12). By applying the HD score and the assignation criteria (HD score system) to the 64 possible combinations of the binary code, we obtained a total of 33 classifiable binary codes, none with an HD score of <0.67, whereby 1 = perfect match, 0.83 = high match, and 0.67 = low match (Figure 3G). After applying the HD score system to the training cohort, 36 of the 38 medulloblastoma samples (94.74%) were assigned to a molecular subgroup with 100% accuracy, according to previous reported classification subgrouping (95% CI [90.3%–100%], k = 1 95% CI [1-1]). Of these, all but one had an HD score >0.83 (Figure S10; Tables S9 and S12). The remaining two cases with low scores (0.67) could not be classified given that the minimum HD score was the same between two RefBCs (Table S12). An independent validation of the HD score system was performed using previously published DNA methylation microarray data from 3,044 medulloblastoma samples (2,834 from HM450K, and 210 from EPIC). Overall, 99.5% of the medulloblastoma samples presented classifiable binary codes and were classified with 99.8% accuracy (95% CI [99.6%–100%]; k = 1 (95% CI [0.9–1])) (Table S13).

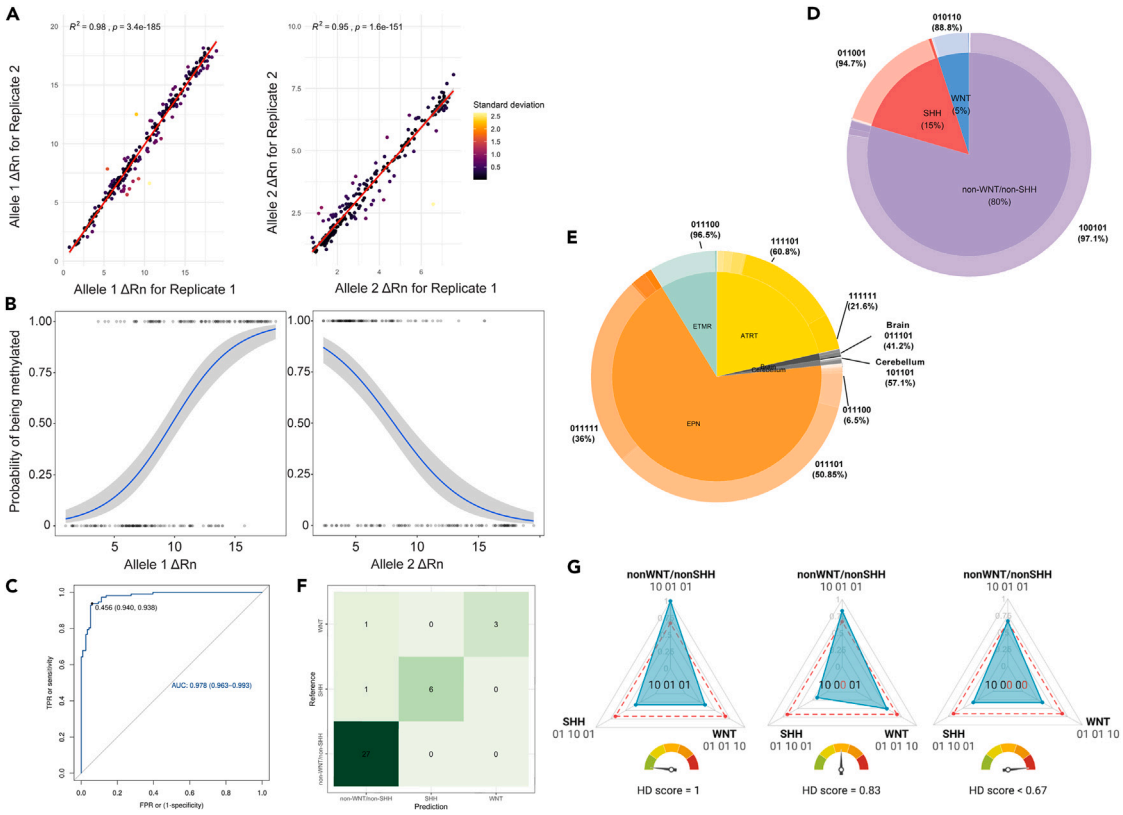


Figure 3. DNA methylation status predictor

(A) Allele 1 and 2 ΔRn correlation between replicates. Pearson correlation coefficient (R^2) is 0.98 and 0.95 for ΔRn Allele 1 and 2, respectively; p value <0.01 for both correlations.

(B) Left, Allele 1 ΔRn mean value database in a base 10 logarithmic scale. Right, the Allele 2 ΔRn mean value database in a base 10 logarithmic scale.

(C) Receiver operating characteristic (ROC) curve of the logistic regression method (LRM). TPR, true positive rate; FPR, false positive rate.

(D) Donut pie of previously published genome-wide DNA methylation array data of medulloblastoma (MB) tumors (n = 3,044) The outer pie represents the distribution of the binarized methylation status.

(E) Donut pie of non-MB tumors analyzed by genome-wide DNA methylation array data of (n = 1,644).

(F). HD score radar plots of three prototypical binary codes, of perfect match (HD = 1) with no mismatches, 1 mismatch (HD = 0.83), or 2 mismatches (the maximum number accepted) (HD = 0.67).

(G) Confusion matrix of the Hamming distance (HD) prediction of the training cohort analyzed by EpiGe. All sample replicates were assigned to the nearest MB subgroup reference binary code (94.74% [95% CI, 82.25%–99.36%]; k = 0.87).

EpiGe web application

The entire DSS was encapsulated in an easy-to-use web application, named EpiGe-App (https://www.epige.irsjd.org/), for automatic classification of medulloblastoma tumors using the six-cytosine methylation signature. We tested an independent set of qPCR data from 20 medulloblastoma samples, including 1 WNT, 2 SHH, and 17 non-WNT/non-SHH medulloblastoma tumors from pediatric patients (henceforth, validation cohort), all of which had been previously classified by microarray DNA profiling[16,17,23,26] (Table S1). qPCR data of the validation cohort was directly uploaded to the EpiGe-App. A total of 19 of the 20 (95%) binary codes obtained from the samples were classified with a 100% agreement (95% CI [82.4%–100%]; k = 1, 95% CI [1–1]) with the previously reported medulloblastoma molecular subgrouping data (Figures S9B and S11; Tables S9 and S14).

The EpiGe-App incorporates the analysis of the raw qPCR data into the result, which is delivered to the user as an easy-to-understand report. No programming knowledge is required for the use of the EpiGe-App. Supplementary material is provided on the EpiGe-App to guide the user through the entire procedure, from protocols that outline the basic principles of the qPCR experiment to how to upload data onto the web application. Information uploaded by the user to the platform is automatically anonymized and stored in the database for a period of 30 days, after which it is erased. Currently, EpiGe-App can analyze data generated by the following qPCR systems: Applied Biosystem 7500 Fast Real-time PCR System and the QuantStudio 3, QuantStudio 5, and QuantStudio 6 Flex System instruments. Users can access EpiGe-App through most web browsers, including Google Chrome, Mozilla Firefox, Apple Safari, and Microsoft Edge. The mean turnaround time for an analysis is 10.95 ± 0.65 s, if the task queue is empty. Afterward, the user has access to the analysis results and can download the report in a pdf format (Supplementary Appendix: EpiGe-App report). To use the application, the user must first read and accept the terms

and conditions agreement of use drafted by the legal team of the Hospital Sant Joan de Déu. EpiGe-App is a research tool that is intended only for scientific purposes: the EpiGe-App has not been validated clinically and was not designed to be used for diagnostic purpose nor to replace the services of a licensed, trained physician or health professional or to be a substitute for medical advice.

DISCUSSION

In this study, we developed, tested, and validated a machine-learning DSS that enables classification of medulloblastoma tumors directly from qPCR-based DNA methylation data. Our machine-learning DSS showed accurate performance (96% accuracy) for differentiating the principal, clinically relevant molecular subgroups of WNT, SHH, and non-WNT/non-SHH of medulloblastoma. Specifically, the performance of our model was similar to previously reported microarray-based molecular classification studies.[6,16] Molecular classification of medulloblastoma is critical for the correct treatment of patients with this malignant pediatric brain tumor. Array-based genome-wide DNA methylation profiling has proven to be a powerful analytical tool and is currently considered a gold standard for the molecular classification of medulloblastoma. However, using this genomic technology in routine clinical practice can be time-consuming, costly, and/or inaccessible for many centers around the world that treat patients with central nervous system tumors.

Our previous study showed that the clinically relevant molecular subgroups of medulloblastoma can be accurately classified using a reduced set of six cytosines with distinctive, subgroup-specific methylation profiles.[23] Our epigenetic classifier classified the WNT, SHH, and non-WNT/non-SHH subgroups with an accuracy (99% concordance) equivalent to DNA methylation microarray profiling.[23] In this study, we developed and trained a multistep DSS based on the methylation profiles of our panel of six CpG markers analyzed by qPCR. To increase the applicability, we developed an interactive, user-friendly web application—EpiGe-App—that has the potential for automated interpretation of qPCR methylation data and subsequent molecular subgroup prediction, and for reporting the methylation class of the medulloblastoma tumor. By using automated analysis of qPCR data, our classification approach should not only be easier to use but also reduce analyser variability and interpretative errors between users, thus improving efficiency of the algorithm.

Our classification method has several strengths. It is a qPCR-based, methyl-genotyping approach, which is overall more accessible and cost-effective than array-based profiling. To our knowledge, our study is the first to perform methylation-based classification of medulloblastoma using qPCR. Our approach is based on the analysis of the methylation status of a reduced panel of six-cytosine, providing a simple and low labour-dependent approach for accurate and rapid classification of medulloblastoma. Recent studies have demonstrated the feasibility of using small sets of tumour-defining epigenetic alterations as a tool for the molecular classification of gliomas and brain metastases.[27,28] Similar to our work, these studies also used PCR-based approaches, methylation-specific qPCR, or methylation-sensitive high-resolution.[27,28] In contrast to these, our approach enables the methylation status of single CpGs to be determined after bisulfite conversion without requiring pre-configuration or generation of serial dilutions of universal methylated control standard curves. No further calculations or acquisition of specialized software is needed, making it easy to use in a clinical setting. Moreover, our quantitative methyl-genotyping method performed consistently across different PCR platforms, suggesting that the proposed approach can be implemented using existing PCR instrumentation found in a vast majority of centers worldwide. Additionally, the web application EpiGe-App provides a simple, free-cost approach for automated interpretation of qPCR methylation data and subsequent molecular subgroup prediction, and reporting of the methylation class of a medulloblastoma tumor.

Our approach has also several limitations. First, our machine-learning model is a classification tool that has been developed and optimized exclusively for qPCR methylation values obtained from tumors with histopathological diagnosis of medulloblastoma. Unlike array-based, genome-wide DNA methylation profiling tools, our qPCR approach, which is based on the methylation profile of only six cytosines, does not support a diagnostic analysis of medulloblastoma or other brain tumor entities. qPCR data generated from non-medulloblastoma samples that are submitted to the EpiGe-App will be detected by the DSS only as "non-matching" with any of the medulloblastoma subgroup–specific methylation profiles. Second, our approach does not distinguish between group 3 and group 4 tumors. The WHO Classification of Tumors of the Central Nervous System, fifth edition, includes four principal molecular groups: WNT-activated, SHH-activated divided on the basis of *TP53* status (mutated or wildtype tumors), and the non-WNT/non-SHH, including both group 3 and group 4 medulloblastomas.[19] Our method thus needs to be implemented alongside with p53 immunohistochemistry assays and/or *TP53* sequencing strategies. Third, although the analysis can use fresh, frozen, or FFPE embedded samples, the usefulness of FFPE material is often limited by the quality of the FFPE tissue and resulting DNA. This can hamper the use of the analysis in centers that only have access to FFPE samples. Fourth, the sample size of our validation cohort was small. As with all machine learning methods, our approach relies on the data available for training and validation. A large cohort study is needed to support the value of this classifier. Future work will include prospective validation with further samples of medulloblastoma tumors. Despite these limitations, the EpiGe-App classifier has the potential to assist classification of medulloblastoma, especially in centers with limited access to genome technologies for methylation and transcriptome profiling.

In conclusion, this study provides a comprehensive approach for rapid classification of clinically relevant medulloblastoma entities, using readily accessible equipment, even in poorly equipped laboratories, and an easy-to-use, free-cost web application. The proposed strategy will be broadly applicable to medulloblastoma research and has shown potential to support clinical application. Prospective validation in large, representative cohort of tumors will be crucial to support the potential clinical application. Finally, a similar classification strategy may prove to be useful also in the context of other pediatric tumors.

Limitations of the study

Our approach has also several limitations. First, our machine-learning model is a classification tool that has been developed and optimized exclusively for qPCR methylation values obtained from tumors with histopathological diagnosis of medulloblastoma. Unlike array-based,

genome-wide DNA methylation profiling tools, our qPCR approach, which is based on the methylation profile of only six cytosines, does not support a diagnostic analysis of medulloblastoma or other brain tumor entities. qPCR data generated from non-medulloblastoma samples that are submitted to the EpiGe-App will be detected by the DSS only as "non-matching" with any of the medulloblastoma subgroup–specific methylation profiles. Second, our approach does not distinguish between group 3 and group 4 tumors. The WHO Classification of Tumors of the Central Nervous System, fifth edition, includes four principal molecular groups: WNT-activated, SHH-activated divided on the basis of *TP53* status (mutated or wildtype tumors), and the non-WNT/non-SHH, including both group 3 and group 4 medulloblastomas.[19] Our method thus needs to be implemented alongside p53 immunohistochemistry assays and/or *TP53* sequencing strategies. Third, although the analysis can use fresh, frozen, or FFPE embedded samples, the usefulness of FFPE material is often limited by the quality of the FFPE tissue and resulting DNA. This can hamper the use of the analysis in centers that only have access to FFPE samples. Fourth, the sample size of our validation cohort was small. As with all machine learning methods, our approach relies on the data available for training and validation. A large cohort study is needed to support the value of this classifier. Future work will include prospective validation with further samples of medulloblastoma tumors. Despite these limitations, the EpiGe-App classifier has the potential to assist classification of medulloblastoma, especially in centers with limited access to genome technologies for methylation and transcriptome profiling.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
- METHOD DETAILS
  - Genomic DNA extraction and bisulfite conversion
  - Sequencing
  - Quantitative PCR-genotyping analysis
  - Synthetic controls
  - Binarization
  - Encoding
  - DNA methylation data processing
  - DNA methylation status prediction
  - Medulloblastoma subgroup classifier
- QUANTIFICATION AND STATISTICAL ANALYSIS
- ADDITIONAL RESOURCES

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2023.107598.

ACKNOWLEDGMENTS

The study was supported by Associations of Parents and Families of Children with Cancer and by funding of the Spanish Ministry of for Science, Innovation and University (grant PI20/00519; PI CL) and the Foundation La Marató TV3 (grant 201921-30; PI CL). We acknowledge the multidisciplinary team who helped in the molecular analyses and care of patients, and the BioBank Hospital Sant Joan de Déu of the Spanish BioBank Network for sample procurement. We also acknowledge Marta Fortuny for communication strategy advice and Eduard Puig for legal assistance and data protection regulations. Authors acknowledge the SJD Fundraising Team.

AUTHOR CONTRIBUTIONS

S.G.G. and C.L. conceived the study; S.G.G., J.L., A.P., and C.L. participated in designing the study; A.P. and C.L. supervised the study; S.G.G., J.L., M.G., A.G.G., A.P., and C.L. contributed to model design, development, training, and data analysis; S.P.J., S.G.G., and J.L. contributed to statistical analysis; A.M., O.C., V.S.M., M.S., M.P.S., A.C., and J.M. contributed to expert review, data interpretation, and literature review; S.G.G., J.L., A.P., and C.L. designed figures, data visualization, and the user interface; M.G., A.G.G., I.L., R.A.G., U.W., F.B.M., F.L., and V.V. contributed to molecular validations; and C.L., J.L., A.P., and S.G.G. drafted the manuscript; N.S., N.G., M.S., J.H., U.W., F.B.M., F.L., and V.V. contributed to providing data and tumor material used in this study. All authors had access to the resulting data presented in the final manuscript, approved the final manuscript, and agreed with the decision to submit the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

INCLUSION AND DIVERSITY

We support inclusive, diverse, and equitable conduct of research.

Received: March 31, 2023
Revised: June 26, 2023
Accepted: August 8, 2023
Published: August 12, 2023

REFERENCES

1. Capper, D., Stichel, D., Sahm, F., Jones, D.T.W., Schrimpf, D., Sill, M., Schmid, S., Hovestadt, V., Reuss, D.E., Koelsche, C., et al. (2018). Practical implementation of DNA methylation and copy-number-based CNS tumor diagnostics: the Heidelberg experience. Acta Neuropathol. *136*, 181–210. https://doi.org/10.1007/s00401-018-1879-y.
2. Capper, D., Jones, D.T.W., Sill, M., Hovestadt, V., Schrimpf, D., Sturm, D., Koelsche, C., Sahm, F., Chavez, L., Reuss, D.E., et al. (2018). DNA methylation-based classification of central nervous system tumours. Nature *555*, 469–474. https://doi.org/10.1038/nature26000.
3. Perez, E., and Capper, D. (2020). Invited Review: DNA methylation-based classification of paediatric brain tumours. Neuropathol. Appl. Neurobiol. *46*, 28–47. https://doi.org/10.1111/nan.12598.
4. Northcott, P.A., Korshunov, A., Witt, H., Hielscher, T., Eberhart, C.G., Mack, S., Bouffet, E., Clifford, S.C., Hawkins, C.E., French, P., et al. (2011). Medulloblastoma comprises four distinct molecular variants. J. Clin. Oncol. *29*, 1408–1414. https://doi.org/10.1200/JCO.2009.27.4324.
5. Northcott, P.A., Shih, D.J.H., Peacock, J., Garzia, L., Morrissy, A.S., Zichner, T., Stütz, A.M., Korshunov, A., Reimand, J., Schumacher, S.E., et al. (2012). Subgroup-specific structural variation across 1,000 medulloblastoma genomes. Nature *488*, 49–56. https://doi.org/10.1038/nature11327.
6. Hovestadt, V., Remke, M., Kool, M., Pietsch, T., Northcott, P.A., Fischer, R., Cavalli, F.M.G., Ramaswamy, V., Zapatka, M., Reifenberger, G., et al. (2013). Robust molecular subgrouping and copy-number profiling of medulloblastoma from small amounts of archival tumour material using high-density DNA methylation arrays. Acta Neuropathol. *125*, 913–916. https://doi.org/10.1007/s00401-013-1126-5.
7. Hovestadt, V., Ayrault, O., Swartling, F.J., Robinson, G.W., Pfister, S.M., and Northcott, P.A. (2020). Medulloblastomics revisited: biological and clinical insights from thousands of patients. Nat. Rev. Cancer *20*, 42–56. https://doi.org/10.1038/s41568-019-0223-8.
8. Gilbertson, R.J., and Ellison, D.W. (2008). The origins of medulloblastoma subtypes. Annu. Rev. Pathol. *3*, 341–365. https://doi.org/10.1146/annurev.pathmechdis.3.121806.151518.
9. Gajjar, A., Bowers, D.C., Karajannis, M.A., Leary, S., Witt, H., and Gottardo, N.G. (2015). Pediatric Brain Tumors: Innovative Genomic Information Is Transforming the Diagnostic and Clinical Landscape. J. Clin. Oncol. *33*, 2986–2998. https://doi.org/10.1200/JCO.2014.59.9217.
10. Cho, Y.J., Tsherniak, A., Tamayo, P., Santagata, S., Ligon, A., Greulich, H., Berhoukim, R., Amani, V., Goumnerova, L., Eberhart, C.G., et al. (2011). Integrative genomic analysis of medulloblastoma identifies a molecular subgroup that drives poor clinical outcome. J. Clin. Oncol. *29*, 1424–1430. https://doi.org/10.1200/jco.2010.28.5148.
11. Pugh, T.J., Weeraratne, S.D., Archer, T.C., Pomeranz Krummel, D.A., Auclair, D., Bochicchio, J., Carneiro, M.O., Carter, S.L., Cibulskis, K., Erlich, R.L., et al. (2012). Medulloblastoma exome sequencing uncovers subtype-specific somatic mutations. Nature *488*, 106–110. http://www.nature.com/nature/journal/v488/n7409/abs/nature11329.html#supplementary-information.
12. Robinson, G., Parker, M., Kranenburg, T.A., Lu, C., Chen, X., Ding, L., Phoenix, T.N., Hedlund, E., Wei, L., Zhu, X., et al. (2012). Novel mutations target distinct subgroups of medulloblastoma. Nature *488*, 43–48. http://www.nature.com/nature/journal/v488/n7409/abs/nature11213.html#supplementary-information.
13. Koelsche, C., and von Deimling, A. (2022). Methylation classifiers: Brain tumors, sarcomas, and what's next. Genes Chromosomes Cancer *61*, 346–355. https://doi.org/10.1002/gcc.23041.
14. Kool, M., Korshunov, A., Remke, M., Jones, D.T.W., Schlanstein, M., Northcott, P.A., Cho, Y.J., Koster, J., Schouten-van Meeteren, A., van Vuurden, D., et al. (2012). Molecular subgroups of medulloblastoma: an international meta-analysis of transcriptome, genetic aberrations, and clinical data of WNT, SHH, Group 3, and Group 4 medulloblastomas. Acta Neuropathol. *123*, 473–484. https://doi.org/10.1007/s00401-012-0958-8.
15. Taylor, M.D., Northcott, P.A., Korshunov, A., Remke, M., Cho, Y.J., Clifford, S.C., Eberhart, C.G., Parsons, D.W., Rutkowski, S., Gajjar, A., et al. (2012). Molecular subgroups of medulloblastoma: the current consensus. Acta Neuropathol. *123*, 465–472. https://doi.org/10.1007/s00401-011-0922-z.
16. Northcott, P.A., Shih, D.J.H., Remke, M., Cho, Y.J., Kool, M., Hawkins, C., Eberhart, C.G., Dubuc, A., Guettouche, T., Cardentey, Y., et al. (2012). Rapid, reliable, and reproducible molecular sub-grouping of clinical medulloblastoma samples. Acta Neuropathol. *123*, 615–626. https://doi.org/10.1007/s00401-011-0899-7.
17. Ramaswamy, V., Remke, M., Bouffet, E., Bailey, S., Clifford, S.C., Doz, F., Kool, M., Dufour, C., Vassal, G., Milde, T., et al. (2016). Risk stratification of childhood medulloblastoma in the molecular era: the current consensus. Acta Neuropathol. *131*, 821–831. https://doi.org/10.1007/s00401-016-1569-6.
18. Holgado, B.L., Guerreiro Stucklin, A., Garzia, L., Daniels, C., and Taylor, M.D. (2017). Tailoring Medulloblastoma Treatment Through Genomics: Making a Change, One Subgroup at a Time. Annu. Rev. Genomics Hum. Genet. *18*, 143–166. https://doi.org/10.1146/annurev-genom-091416-035434.
19. Louis, D.N., Perry, A., Wesseling, P., Brat, D.J., Cree, I.A., Figarella-Branger, D., Hawkins, C., Ng, H.K., Pfister, S.M., Reifenberger, G., et al. (2021). The 2021 WHO Classification of Tumors of the Central Nervous System: a summary. Neuro Oncol. *23*, 1231–1251. https://doi.org/10.1093/neuonc/noab106.
20. Li, Y., Song, Q., and Day, B.W. (2019). Phase I and phase II sonidegib and vismodegib clinical trials for the treatment of paediatric and adult MB patients: a systemic review and meta-analysis. Acta Neuropathol. Commun. *7*, 123. https://doi.org/10.1186/s40478-019-0773-8.
21. Fang, F.Y., Rosenblum, J.S., Ho, W.S., and Heiss, J.D. (2022). New Developments in the Pathogenesis, Therapeutic Targeting, and Treatment of Pediatric Medulloblastoma. Cancers *14*, 2285. https://doi.org/10.3390/cancers14092285.
22. Lazow, M.A., Palmer, J.D., Fouladi, M., and Salloum, R. (2022). Medulloblastoma in the Modern Era: Review of Contemporary Trials. Neurotherapeutics *19*, 1733–1751. https://doi.org/10.1007/s13311-022-01273-0.
23. Gómez, S., Garrido-Garcia, A., Garcia-Gerique, L., Lemos, I., Suñol, M., de Torres, C., Kulis, M., Pérez-Jaume, S., Carcaboso, Á.M., Luu, B., et al. (2018). A Novel Method for Rapid Molecular Subgrouping of Medulloblastoma. Clin. Cancer Res. *24*, 1355–1363. https://doi.org/10.1158/1078-0432.CCR-17-2243.
24. Huijsmans, C.J.J., Poodt, J., Damen, J., van der Linden, J.C., Savelkoul, P.H.M., Pruijt, J.F.M., Hilbink, M., and Hermans, M.H.A. (2012). Single nucleotide polymorphism (SNP)-based loss of heterozygosity (LOH) testing by real time PCR in patients suspect of myeloproliferative disease. PLoS One *7*, e38362. https://doi.org/10.1371/journal.pone.0038362.

25. Forsmo, H.M., Erichsen, C., Rasdal, A., Körner, H., and Pfeffer, F. (2017). Enhanced recovery after colorectal surgery (ERAS) in elderly patients is feasible and achieves similar results as in younger patients. Gerontol. Geriatr. Med. *3*, 2333721417706299.
26. Cavalli, F.M.G., Remke, M., Rampasek, L., Peacock, J., Shih, D.J.H., Luu, B., Garzia, L., Torchia, J., Nor, C., Morrissy, A.S., et al. (2017). Intertumoral Heterogeneity within Medulloblastoma Subgroups. Cancer Cell *31*, 737–754.e6. https://doi.org/10.1016/j.ccell.2017.05.005.
27. Majchrzak-Celińska, A., Dybska, E., and Barciszewska, A.M. (2020). DNA methylation analysis with methylation-sensitive high-resolution melting (MS-HRM) reveals gene panel for glioma characteristics. CNS Neurosci. Ther. *26*, 1303–1314. https://doi.org/10.1111/cns.13443.
28. Orozco, J.I.J., Knijnenburg, T.A., Manughian-Peter, A.O., Salomon, M.P., Barkhoudarian, G., Jalas, J.R., Wilmott, J.S., Hothi, P., Wang, X., Takasumi, Y., et al. (2018). Epigenetic profiling for the molecular classification of metastatic brain tumors. Nat. Commun. *9*, 4627. https://doi.org/10.1038/s41467-018-06715-y.
29. Aryee, M.J., Jaffe, A.E., Corrada-Bravo, H., Ladd-Acosta, C., Feinberg, A.P., Hansen, K.D., and Irizarry, R.A. (2014). Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. Bioinformatics *30*, 1363–1369. https://doi.org/10.1093/bioinformatics/btu049.
30. Khoo, C.K., Vickery, C.J., Forsyth, N., Vinall, N.S., and Eyre-Brook, I.A. (2007). A prospective randomized controlled trial of multimodal perioperative management protocol in patients undergoing elective colorectal resection for cancer. Ann. Surg. *245*, 867–872.
31. Hosmer, D.W., Lemeshow, S., and Sturdivant, R.X. (2013). Applied Logistic Regression (Wiley).
32. Bierbrauer, J. (2016). Introduction to Coding Theory (CRC Press).
33. Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. Educ. Psychol. Meas. *20*, 37–46. https://doi.org/10.1177/001316446002000104.
34. Migaly, J., Bafford, A.C., Francone, T.D., Gaertner, W.B., Eskicioglu, C., Bordeianou, L., Feingold, D.L., and Steele, S.R.; Clinical Practice Guidelines Committee of the American Society of Colon and Rectal Surgeons (2019). The American Society of Colon and Rectal Surgeons Clinical Practice Guidelines for the use of bowel preparation in elective colon and rectal surgery. Dis. Colon Rectum *62*, 3–8.
35. Denost, Q., Rouanet, P., Faucheron, J.-L., Panis, Y., Meunier, B., Cotte, E., Meurette, G., Kirzin, S., Sabbagh, C., Loriau, J., et al. (2017). To drain or not to drain infraperitoneal anastomosis after rectal excision for cancer. Ann. Surg. *265*, 474–480.
36. Youden, W.J. (1950). Index for rating diagnostic tests. Cancer *3*, 32–35. https://doi.org/10.1002/1097-0142(1950)3:1<32::AID-CNCR2820030106>3.0.CO;2-3.
37. Skaltsa, K., Jover, L., and Carrasco, J.L. (2010). Estimation of the diagnostic threshold accounting for decision costs and sampling uncertainty. Biom. J. *52*, 676–697. https://doi.org/10.1002/bimj.200900294.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Chemicals, peptides, and recombinant proteins		
Gentra Puregene Tissue Kit	QIAGEN	Cat. No. 158689
QIAamp DNA FFPE kit	QIAGEN	Cat. No. 56404
EpiTect Plus Bisulfite Conversion	QIAGEN	Cat. No. 59124
rhAmp® Reporter Mix w/Reference	Integrated DNA Technologies	Cat. No.1076020
rhAmp Genotyping Master Mix	Integrated DNA Technologies	Cat. No.1076014
Oligonucleotides		
gBlocks® Gene Fragments 125-500 bp	Integrated DNA Technologies	https://eu.idtdna.com/site/order/gblockentry
rhAmp SNP Genotyping Assays	Integrated DNA Technologies	https://eu.idtdna.com/site/order/designtool/index/GENOTYPING_PREDESIGN
		SNP Assay, Design ID: CD.GT.SPDR8897.10
		SNP Assay, Design ID: CD.GT.GYJV9231.1
		SNP Assay, Design ID: CD.GT.JBHB3172.1
		SNP Assay, Design ID: CD.GT.JYVK6101.1
		SNP Assay, Design ID: CD.GT.WFVP0601.1
		SNP Assay, Design ID: CD.GT.PVWC1084.1
		SNP Assay, Design ID: CD.GT.BFJY0393.1
		SNP Assay, Design ID: CD.GT.BMJJ6372.1
Deposited data		
Raw and analyzed data	This paper	NCBI under GEO Accession ID GSE210723 https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE210723
Raw data	Clinical and genetic diversity and recurrent CXorf67 mutations across distinct molecular subgroups of posterior fossa type A (PFA) ependymoma.	GSE104210
Raw data	Proteogenomic Analysis of Medulloblastoma	GSE109402
Raw data	Heterogeneity within the PF-EPN-B subgroup	GSE117130
Raw data	The Molecular Landscape of ETMR at Diagnosis and Relapse	GSE122038
Raw data	Second-generation molecular subgrouping of medulloblastoma: an international meta-analysis of Group 3 and Group 4 subtypes	GSE130051
Raw data	DNA methylation data from 153 ATRT tumor samples	GSE141039
Raw data	Methylation Profiling of Medulloblastoma in a Clinical Setting Permits Sub-Classification and Reveals New Outcome Predictions	GSE142627
Raw data	Metabolic Regulation of the Epigenome Drives Lethal Infantile Ependymoma	GSE146426
Raw data	Epigenetic methylation chip analysis of childhood medulloblastoma patient samples	GSE156012

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Raw data	Atypical teratoid/rhabdoid tumors (ATRTs) with SMARCA4 mutation are molecularly distinct from SMARCB1 deficient cases	GSE161692
Raw data	Validation of the MethylationEPIC BeadChip for fresh-frozen and formalin-fixed paraffin-embedded tumours	GSE92580
Raw data	Clinical and genetic diversity and recurrent CXorf67 mutations across distinct molecular subgroups of posterior fossa type A (PFA) ependymoma	GSE104210
Raw data	Proteogenomic landscape of medulloblastoma subgroups	GSE104728
Raw data	DNA methylation-based classification of human central nervous system tumors	GSE109381
Raw data	The Molecular Landscape of ETMR at Diagnosis and Relapse	GSE122038
Raw data	HM450K-based DNA methylation analysis of normal brain and glioma samples	GSE123678
Raw data	Second-generation molecular subgrouping of medulloblastoma: an international meta-analysis of Group 3 and Group 4 subtypes	GSE130051
Raw data	DNA methylation data from 153 ATRT tumor samples	GSE141039
Raw data	Recurrent Variations in DNA Methylation in Human Pluripotent Stem Cells and their Differentiated Derivatives	GSE30654
Raw data	Epigenomic Alterations Define Lethal CIMP-positive Ependymomas of Infancy	GSE45353
Raw data	Illumina Infinium 450K array data for Diffuse Intrinsic Pontine Glioma	GSE50022
Raw data	DNA methylation changes at CpG and non-CpG sites are associated with development and clinical behavior in neuroblastoma	GSE54719
Raw data	Microarray-based DNA methylation profiling of medulloblastoma and normal cerebellum samples	GSE54880
Raw data	The genomic and epigenomic landscape of atypical teratoid rhabdoid tumors	GSE70460
Raw data	DNA methylation profiling of primary medulloblastoma samples	GSE85212
Raw data	A biobank of 30 molecularly characterized patient-derived xeongraft models of pediatric brain tumors	GSE99994
Software and algorithms		
QuantStudio™ Design & Analysis Software	Thermo Fisher Scientific	https://www.thermofisher.com/es/es/home/global/forms/life-science/quantstudio-3-5-software.html
R Version 4.3.0	R Foundation	https://cran.r-project.org/

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Python Version 3.9	Python Software Foundation	https://www.python.org
minfi	Bioconductor software	https://bioconductor.org/packages/release/bioc/html/minfi.html
stats	Rcore statistical functions	https://rdocumentation.org/packages/stats/versions/3.6.2
stringdist	CRAN package	https://cran.r-project.org/web/packages/stringdist/index.html
vcd	CRAN package	https://cran.r-project.org/web/packages/vcd/index.html
pROC	CRAN package	https://cran.r-project.org/web/packages/pROC/index.html
ThresholdROC	CRAN package	https://cran.r-project.org/web/packages/ThresholdROC/index.html
Django	Python package	https://www.djangoproject.com
Celery Version 5.1.2	Python package	https://docs.celeryq.dev/en/v5.1.2/changelog.html
PostgreSQL	Framework for front-end web development	https://www.postgresql.org
Bootstrap	Open source object-relational database system	https://getbootstrap.com
cloudUPC	Private cloud from the Universitat Politècnica de Catalunya (UPC)	https://serveistic.upc.edu/ca/cloud-upc
Other		
EpiGe-APP	This paper	https://www.epige.irsjd.org/

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Soledad Gómez-González (soledad.gomezg@sjd.es).

Materials availability

rhAmp SNP assays were designed, optimized, and synthesized with IDT technical support. Primer sequences are available at Table S5 and the reference numbers are listed in the key resources table.

Data and code availability

- DNA methylation microarray data have been deposited at GEO and are publicly available as of the date of publication. Accession number is listed in the key resources table.
- This paper does not report original code.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

The data used to develop our decision-support system consisted of DNA methylation data from 4,807 samples, comprising 3,157 primary medulloblastomas, 1,613 non-medulloblastoma tumours, and 37 normal tissues. The dataset included DNA methylation microarray data from 4,743 samples obtained from databases that are publicly available or generated by our research group, together with methylation data from DNA of medulloblastoma biopsies (53 fresh-frozen and 5 fixed formalin paraffin embedded [FFPE]) and of six normal peripheral blood samples generated using a PCR-based methyl-genotyping approach (Table S1). Tumour biopsies were obtained from primary medulloblastoma tumours diagnosed and treated at Hospital Sant Joan de Déu (HSJD), Barcelona and at collaborative centres. All samples included in the study were obtained from patients 18 years old or younger. Clinical data and molecular subgroup affiliation of medulloblastoma samples were available for all the cases. Sample flow diagram is detailed in Figure 1.

The study was approved by the Institutional Research Ethics Committee of the HSJD (CEIC PIC-116-19). Written informed consent was obtained from patients/guardians before sample collection.

METHOD DETAILS

Genomic DNA extraction and bisulfite conversion

Genomic DNA was isolated using Gentra Pure gene Tissue kit for fresh-frozen samples or a QIAamp DNA FFPE kit for FFPE samples (Qiagen Technologies). Bisulfite conversion of genomic DNA was performed using EpiTect Plus Bisulfite Conversion kit (Qiagen Technologies), following the manufacturer's protocol.

Bisulfite conversion optimization: The bisulfite conversion reaction was optimised using 5µl at a concentration of 100ng/µL DNA to obtain 500ng of total DNA in a high-concentration range (1ng-2µL) after testing the low-concentration protocol (1-500ng). The bisulfite reaction was composed of 5µl of sample DNA volume (100ng/µl), 15µL of RNase-free water, 85µL of Bisulfite Mix, 35µL DNA Protect Buffer. The quality and concentration of the bisulfite converted DNA improved when we eluted the bisulfite converted DNA in 30µL of Endonuclease-free EB buffer.

Sequencing

The presence and correct position of methylation-dependent single-nucleotide variant C/T SNPs in the synthetic DNA sequences was verified by targeted Sanger sequencing. Briefly, sequencing analyses (DNA and bisulfite sanger sequencing) were performed using BigDye® Terminator Cycle Sequencing kit (Applied Biosystems) on an ABI Prism 3130XL sequencer (Applied Biosystems), following standard procedures.[23]

Quantitative PCR-genotyping analysis

Genotyping analysis was performed using the rhAmp® SNP technology (Integrated DNA Technologies, USA (IDT)) according to manufacturer's instructions. Briefly, rhAmp genotyping assay was setup using 1 µl of bisulfite-converted genomic DNA with rhAmp Genotyping Mix, composed of rhAmp Genotyping Master Mix (IDT), rhAmp Reporter Mix with a reference dye (IDT), and custom rhAmp SNP assays (IDT). The rhAMP SNP assays contain an allele-specific primer 1 (methylated allele), an allele-specific primer 2 (unmethylated allele), and the locus-specific reverse primer. Allelic specificity of the rhAMP SNP assays was provided by two target-specific fluorescent probes, whereby the methylated allele probe was labelled with FAM dye and the unmethylated allele probe, with Yakima Yellow dye; both were detected using the VIC channel. Normalization across samples was performed using the ROX Passive Reference dye. A total of 10 ng synthetic control sequence was used for each qPCR assay. EpiGe primers were used at a 20× concentration, following the manufacture's protocol. Each experiment was run in duplicate to validate the reproducibility, including two non-template controls (NTC). Samples were analysed using a 7500 Real-Time PCR System (Thermo Fisher Scientific) or a QuantStudio™ Flex Real-Time PCR System, version 3, 5, or 6 (Thermo Fisher Scientific). Based on the thermal cycling parameters for rhAmp SNP genotyping specified in the IDT protocol, we tested different times for the extension step ranging from 20 to 45 seconds, the latter being the extension time with best performance. The optimized thermal cycle program is shown in Table S3. qPCR results (sample setup, raw data, amplification data, multicomponent data, results, and reagent information) were exported in a single text file (*.txt format) from the real-time PCR system.

Synthetic controls

Double-stranded, sequence-verified gBlocks® Gene Fragments (IDT) of 300 to 500 base pairs were designed and used as positive or negative control sequences. Two gBlocks were designed specifically for each cytosine of interest, one for each possible methylation state: methylated or unmethylated. The total DNA input for gBlock control sequences in all experiments was 10 ng. Synthetic controls were tested using a 1:10 dilution series (from 10 ng to 0.01 ng) (Figure S1).

Binarization

The DNA methylation status of each cytosine was binarizated into binary numbers, applying a cut-off of 0.5. Cytosines whose methylation ranged from 50% to 100% were assigned a 1 (methylated), and cytosines whose methylation was <50% were assigned a 0 (unmethylated).

Encoding

The encoding is a six-digit representation of the two possible values (1 or 0) obtained from the binarization of the six cytosines (binary code) according to a fixed order: cg18849583, cg01268345, cg10333416, cg12925355, cg25542041, and cg02227036.

DNA methylation data processing

DNA methylation microarray (Illumina Infinium HumanMethylation 450 BeadChip (HM450K) and Illumina methylation EPIC BeadChip array (EPIC)) RAW data (two colour iDAT files) were normalized using SWAN from minfi package available through Bioconductor.[29] In order to exclude technical biases, we used an optimized pipeline with stringent selection filters: probes detecting SNPs, probes with poor detection P-values (P>0.01) and those with sex-specific DNA methylation were removed from the initial dataset.[23] Single cytosine methylation values (β-values) were calculated as the ratio of the methylated signal intensity to the sum of methylated and unmethylated signals (Table S2; Figure 1).

DNA methylation status prediction

The DNA methylation predictor was developed to enable automated interpretation of qPCR data and prediction of the methylation status of cytosines. Normalised fluorescence qPCR endpoint values (ΔRn) were used as a data source for prediction of DNA methylation status. Methylated Allele1 ΔRn (FAM) and Unmethylated Allele2 ΔRn[30] values were obtained from a text file directly from the qPCR thermocycler.

First, we computed the ΔRn mean and standard deviation (SD)[25] values between qPCR replicates. The qPCR ΔRn mean values were transformed into a base 10 logarithmic (log10) scale. To train the logistic regression model (*glm* function from R v4.2.0),[31] the independent variables used were log10-transformed qPCR ΔRn mean values from Allele1 and Allele2, and the dependent variable used was the DNA methylation microarray data (EPIC data) of the six-cytosine signature that had been binarized (whereby 1 indicates methylated, and 0, unmethylated) applying a cut-off of 0.5. The performance of the model was assessed using the leave-one-out cross-validation (LOOCV) method of cytosines, grouped by sample (leave-one-patient-out cross-validation [LOPOCV]) to avoid having a potential biased performance induced by cytosines of the same patient assigned to both training and test sets.[31] N-folds were performed, with *N* equal to the number of patients.

Medulloblastoma subgroup classifier

To define the Hamming distance (HD) reference codes, we binarized the six-cytosine microarray methylation data obtained from 3,044 medulloblastoma and 1,644 non-medulloblastoma samples (Table S1). The binarized methylation data was encoded according to a fixed order of the cytosines (cg18849583, cg01268345, cg10333416, cg12925355, cg25542041, and cg02227036). Medulloblastoma subgroup data was represented by a specific, unique binary code: WNT (010110), SHH (011001), and non-WNT/non-SHH (100101) (termed reference binary code, RefBC).

To develop the classifier, we computed the minimum HD[32] between the binary encoding that described the methylation status of the six-cytosine signature (binary code) and the RefBC. The binary codes (obtained from the DNA methylation status prediction) and the RefBC were compared using HD. Samples were assigned to a subgroup by including those with the minimum HD, but excluding those with equal minimum HD to multiple RefBC. The HD was calculated using the *stringdist* function (*stringdist R* package v0.9.8). The HD distance to the three RefBC ranged from 0 (when the two codes were identical) to 6 (when the two codes differed at each position).

We generated a scoring system for an input sample *i*, based on the distance range to the three RefBCs by applying the following formula: $S_{ir} = 1 - \frac{d_{ir}}{d_{max}}$, where d_{ir} is the distance of *i* to the reference *r*, and d_{max} is the maximum possible distance. The obtained HD score ranged from 0, maximum distance, to 1, when the binary code was identical to the RefBC. We computed the HD score of the training cohort to the three RefBC.

QUANTIFICATION AND STATISTICAL ANALYSIS

The diagnostic ability of our method was assessed by computing sensitivity and specificity. We also computed other metrics to evaluate the agreement between our method and the gold standard: the accuracy (percentage of samples correctly classified) and Cohen's kappa coefficient (k).[33] The *accuracy* = $\frac{TN+TP}{TN+TP+FN+FP}$ (true positive [TP], false positive [FP], true negative [TN],[34] and false negative [FN]), and the 95% confidence intervals (CI) were computed using the *confusionMatrix* function (caret R package v6.0). The Cohen's kappa coefficient and the 95% CI were computed using *kappa* function (vcd R package v1.4-10).

The area under the receiver operating characteristic (ROC) curve[35] and its 95% CI were calculated using the *roc* function (*pROC* R package v1.18.0). The optimal threshold for the logistic regression output was obtained computing Youden's index (J),[36,37] J = sensitivity + specificity -1. Youden's index, sensitivity, and specificity, and the respective 95% CI were computed using the *diagnostic* function (ThresholdROC R package v2.9.0).

ADDITIONAL RESOURCES

The entire methodology developed was encapsulated in a web application named EpiGe-App: https://www.epige.irsjd.org/. This web server was developed in the Django (https://www.djangoproject.com) framework based on Python3.9 (https://www.python.org) that provides programming resources for the development of server web services and data management. EpiGe-App uses Celery as a task distributor (v5.1.2) to analyse samples as asynchronous tasks and used PostgreSQL (https://www.postgresql.org) as a database to store the information of each sample analysis performed. The database is anonymized, each sample has an internal code for traceability of the analysis. No personal information, such as email address or registration, are required to use the web server. The client-side has been designed with HTML5 styled with Bootstrap (https://getbootstrap.com). Users can communicate with the web server through an HTTPS protocol implemented in a Nginx HTTP server. Finally, the web server is hosted on a dedicated virtual machine in the private cloud from the *Universitat Politècnica de Catalunya* (UPC) (cloudUPC, https://serveistic.upc.edu/ca/cloud-upc). The hardware specifications are two virtual CPUs, with two processing cores each of virtual CPUs at 2.2GHz, 4GB of RAM memory, and a data storage capacity of 35GB. The virtual machine uses a Linux operating system (Ubuntu OS).