Title: Sensitivity Analysis of the MGMT-STP27 Model and Impact of Genetic and Epigenetic Context to Predict the MGMT Methylation Status in Gliomas and Other Tumors.

Authors: Bady P, Delorenzi M, Hegi ME

Journal: The Journal of molecular diagnostics : JMD

Year: 2016 May

Volume: 18

Issue: 3

Pages: 350-61

DOI: 10.1016/j.jmoldx.2015.11.009

Sensitivity analysis of the MGMT-STP27 model and impact of genetic/epigenetic context to predict the *MGMT* methylation status in gliomas and other tumors

Pierre Bady[*,†,‡,§], Mauro Delorenzi[§,¶,‖], Monika E Hegi[*,†]

[*] Neurosurgery, Lausanne University Hospital, 1011 Lausanne, Switzerland

[†] Neuroscience Research Center, Lausanne University Hospital, 1011 Lausanne, Switzerland

[‡] Department of Education and Research, University of Lausanne, 1011 Lausanne, Switzerland

[§] Bioinformatics Core Facility, Swiss Institute for Bioinformatics, 1005 Lausanne, Switzerland;

[¶] Ludwig Center for Cancer Research, University of Lausanne, 1011 Lausanne, Switzerland,

[‖] Department of Oncology, University of Lausanne, 1011 Lausanne, Switzerland.

Manuscript 17 pages (including references, 25), 2 Tables, 6 Figures

Running head: Sensitivity analysis MGMT-STP27

Grant support: Swiss National Science Foundation (3100A-138116), the Swiss Bridge Award 2011, and the Swiss Cancer League (KFS-29-02-2012).

Corresponding Author: Monika E. Hegi, Laboratory of Brain Tumor Biology and Genetics, Department of Clinical Neurosciences, Lausanne University Hospital (CHUV-CLE C306), Ch des Boveresses 155, 1066 Epalinges, Switzerland

Phone: +41-21-314-2582, Email: monika.hegi@chuv.ch

Disclosures: No conflict of interest to report

Abstract

The methylation status of the O(6)-methylguanine-DNA methyltransferase (*MGMT*) gene is an important predictive biomarker for benefit from alkylating agent therapy in glioblastoma. Our model MGMT-STP27 allows prediction of the methylation status of the *MGMT* promoter using data from the HumanMethylationBeadChip (Illumina, HM-27K and HM-450K) that is publically available for many cancer datasets. Here we present investigations addressing the impact of the context of genetic and epigenetic alterations and tumor type on the classification, report on technical aspects, such as robustness of cut-off definition and preprocessing of the data. The association between gene copy number variation (CNV), predicted *MGMT* methylation and *MGMT* expression revealed a gene dosage effect on *MGMT* expression in lower grade glioma (WHO grade II/III) that in contrast to glioblastoma usually carry two copies of chromosome 10 on which *MGMT* resides (10q26.3). This implies some *MGMT* expression, potentially conferring residual repair function blunting the therapeutic effect of alkylating agents. A sensitivity analyses corroborated the performance of the original cut-off for various optimization criteria and for most data preprocessing methods. Finally, we propose a R package mgmtstp27 that allows prediction of the methylation status of the *MGMT* promoter and calculation of appropriate confidence and/or prediction intervals. Overall the MGMT-STP27 is a robust model for *MGMT* classification that is independent of tumor type, and is adapted for single sample prediction.

Introduction

Large scale analyses of the methylome of gliomas have provided relevant insights into tumor biology and cell of origin that has important implications for tumor classification and choice of therapy [1, 2]. The DNA methylation status of the promoter of the O(6)-methylguanine-DNA methyltransferase (*MGMT)* gene that encodes a DNA repair protein is the most important predictive factor for benefit from alkylating agents such as temozolomide in glioblastoma (GBM) [3-6]. However, in anaplastic and low grade glioma a prognostic versus a predictive value is more controversial [6-9]. A principle difference between GBM and lower grade glioma (WHO grade II and III) is the high frequency of mutations in the isocitrate dehydrogenase (IDH) genes 1 or 2 in lower grade glioma that is mechanistically linked with the development of a CpG island methylator phenotype (CIMP+) [10]. In glioma CIMP is almost invariably associated with *MGMT* promoter methylation regardless of tumor grade as we have reported previously [11]. This raises the question whether the mechanistic underpinnings of CIMP may lead to functionally relevant differences in the methylation pattern affecting epigenetic silencing of the *MGMT* gene. It has been shown that DNA hypermethylation in CIMP results from inhibition of α-ketoglutarate-dependent dioxygenases such as the epigenetic modifier TET2, by high concentrations of the oncometabolite 2-hydroxyglutarate produced by the neomorphic enzymatic function of the IDH1 and 2 mutants [10, 12, 13]. Furthermore, loss of 1 copy of chromosome 10, home of *MGMT* (10q26), is a hallmark of primary GBM (>80%), while it is a rare event in lower grade glioma. Hence in *MGMT* methylated lower grade gliomas *MGMT* could be transcribed from the second potentially intact strand.

Genome-wide DNA methylation data on human methylation 27K (HM-27K) or 450K (HM-450K) BeadChips have become publically available for large datasets of glioma. This data can be used to determine the *MGMT* methylation status using our previously developed logistic regression model, MGMT-STP27 [11]. The input into the model are measures of 2 key

CpG probes located in the *MGMT* promoter that we identified to be functionally highly relevant and which are available on both versions of the chip. The model was trained with a dataset of 63 GBM from homogenously treated patients, for which the *MGMT* methylation status was previously shown to be predictive for outcome, based on classification by methylation-specific PCR (MSP). The MGMT-STP27 model provided good classification properties and prognostic value (kappa=0.85; logrank p<0.001), and has been successfully validated in independent datasets including clinical trials, by us and other groups [2, 9, 11, 14, 15]. The original preprocessing procedure was based on the conversion of the Red/Green channel from the Illumina methylation array into the methylation signal, without using any normalization. However, the rising interest into epigenetics has stimulated development of methods to analyze DNA methylation data including numerous procedures for normalization and bias correction [16-19]. Triche et al. [17] listed no fewer than seven methods to correct background such as substraction of fifth percentile of negative control distribution (Illumina procedure) and normal-exponential deconvolution (Noob). The use of one of these new procedures may modify the estimation of signal intensities in ways that affect the suitability of the parameters in the current MGMT-STP27 model thereby impacting classification.

The aim of the present study was to determine the impact of methodological/computational procedures, sample type (frozen versus formalin fixed paraffin embedded, FFPE), and biological context [CIMP, gene copy number alterations (CNA), tumor type] on the evaluation of the *MGMT* status using the MGMT-STP27 method. The functional validity of the classification model, including the previously established cut-off, is tested across tumor grades, CIMP-status, and extended to non-brain tumor entities. This includes the investigation of the spatial correlations of CpG-methylation and *MGMT* expression that informs on the functionality of the methylation to actually impact *MGMT* expression and thereby indicating the potential of the tumor cells for DNA repair. The simultaneous effects of CIMP, promoter

methylation and gene dosage on *MGMT* expression are evaluated. To complete the sensitivity analysis for the model MGMT-STP27, we investigate how our classifier can be affected by different background and normalization procedures for data from the HM-27K and HM-450K platforms. Finally, we provide a R package called "mgmtstp27" (https://github.com/badozor/mgmtstp27) that allows easy computation of MGMT-STP27 classification for individual samples, and includes new features such as the calculation of the confidence intervals of the *MGMT* methylation scores (*MGMT* methylation probability), comparison of the score distribution of external datasets with the training set, and quality control.

Materials and methods

Datasets

Clinical information and DNA methylation data (HM-27K and 450K) from 7 publically available glioma data-sets (761 individuals, 119 WHO grade II, 258 WHO grade III and 384 GBM) were used for this study. The first, originally used as the training set, contained DNA methylation profiles and expression data for 63 GBM tissues from 59 patients treated within clinical trials and five non-tumoral brain tissues (epilepsy surgery) (M-GBM) [11, 20, 21]. The external datasets used are VB-Glioma-III, from patients treated within a clinical trial (n= 110 glioma grade III) [9]; T-Glioma-II/III (29 WHO grade II, 42 grade III) [10]; and the following datasets from The Cancer Genome Atlas (TCGA, http://cancergenome.nih.gov; https://tcga-data.nci.nih.gov/tcga/): TCGA-GBM-27, TCGA-GBM-450 (n= 321 GBM) and TCGA-Glioma-II/III (n=197; 90 WHO grade II, 106 WHO grade III, n=1, unspecified grade; website http://cancergenome.nih.gov/) [22-24]. Three additional TCGA datasets for non-brain tumors comprise colon adenocarcinoma (TCGA-COAD, n= 227), breast cancer (TCGA-BRCA, n=

305, randomly selected from a set of 642 samples), head and neck squamous cell carcinoma (TCGA-HNSC, n=442), and lung squamous cell carcinoma (TCGA-LUSC, n=328). The dbGaP accession number to the specific version of the TCGA data set is phs000178.v9.p8. The datasets and their accession numbers, including their corresponding expression datasets, are described in detail in the Supplemental Table S1. The clinical and molecular baseline description for the glioma datasets is summarized in Supplemental Table S2.

Procedures for preprocessing and *MGMT* promoter methylation prediction

The pipeline for computation of the *MGMT* classification is summarized in Supplemental Figure S1. The prediction of the DNA methylation status of *MGMT* promoter requires the conversion of the Red/Green channel information derived from the Illumina methylation array into signals for methylated and unmethylated, respectively, without normalization. .The M-values [25] (log2-ratio of methylated and unmethylated intensities corrected by an offset equal to 1,) for the methylation probes of interest located in the *MGMT* promoter, cg12434587 and cg12981137 (location see Figure 1) were used as input into the logistic regression model (MGMT-STP27) to predict the methylation status of the *MGMT* gene [11]. The calculation of the confidence intervals for the logistic regression model is described [26]. The *MGMT* score was obtained by logit-transformation of the probability that the *MGMT* promoter is methylated to obtain a quasi-normal score. The predicted values (probabilities and *MGMT* score), confidence intervals, and *MGMT* classification can be directly obtained by the function MGMTpredict from the R package mgmtstp27 (https://github.com/badozor/mgmtstp27).

The effect of normalization and preprocessing of the HM-450K data on the prediction of the *MGMT* status was tested for five additional procedures and compared to the original (raw)

preprocessing used for developing the method [11]: control normalization which requires the selection of a reference array (Genome Studio), preprocessing including only background correction, quantile normalization of the separated unmethylated and methylated signals, Subset-quantile within array normalization (SWAN) procedure [16] and Noob normalization, including background correction based on normal-exponential deconvolution with dye-bias correction [17].

Preprocessing for determination of gene copy number alterations from HM-450K and HM-27K

Gene copy number alterations (CNA) were calculated basically according to the procedure described by Feber et al [19] and adapted for the HM-27k platform and Genome Studio output. As proposed for Illumina Infinium Whole-genome SNP data [27], the quantile normalization was performed individually for each sample using intensity for unmethylated and methylated signals. The combined intensities for methylated and unmethylated (total intensity, T) was calculated from the normalized intensities. Because matched reference samples were not available, the value *log2(R)* was defined as the difference of intensity between samples and a synthetic reference corresponding to the median profile from a reference dataset containing eight non-tumor brain samples from the TCGA database and M-GBM [11].

$$log2(R) = log2\left(T_{observed} + 1\right) - log2(T_{reference} + 1)$$

An additional smoothing procedure was applied to remove the wave bias for more accurate breakpoint detection in profiles [28]. The unmethylated and methylated intensities from chemistry II (see Illumina technical sheet; http://www.illumina.com/content/dam/illumina-marketing/documents/products/datasheets/datasheet_humanmethylation450.pdf) were

corrected by a scaling factor method to reduce the chemistry type-bias before the computation of the total intensity. As indicated above, probes with non-significant p-values (typically >0.01) were excluded from our analysis when raw data served as input.

Determination of gene copy alteration state

For determination of CNA the R package CGHcall [29] was used that performs circular binary segmentation (CBS) [30] starting with normalized *log2(R)* values for each sample. Afterwards, each probe (CpG) was classified by a mixture model [29] into five classes: amplified, gained, normal, deleted and homozygously deleted. For genomic region (or gene), the CNA events were detected in using copy number probe means (CpGs) contained in the selected region (e.g. chromosomal arms 1p and 19q, region of 10q26.3).

Statistical Analysis

CIMP positive tumors were identified using unsupervised clustering methods (Ward's algorithm with Euclidean distance) as previously reported [22]. The relationships between categorical variables were assessed by Chi-squared tests with *p* values computed by Monte Carlo simulation, because cell counts were expected to be less than five [31].

The classical two-way ANOVA is replaced by Monte-Carlo version to test the effects of CNA and DNA methylation on expression of *MGMT* based on F-statistics (two-way ANOVA-like approach) [32, 33], this method is more robust for the unbalanced data and non-normal assumption for the distribution of the data.

Evaluation of cut-off robustness, including determination of optimal values and performances was tested for six criteria (cost functions) using the training dataset (M-GBM) for which classification by MSP is also available, which served as gold standard [11]: maximization of sensitivity and specificity, *MaxSpSE* [34]; maximization of the product of sensitivity and specificity, *MaxProdSpSe* [35]; equality (balance) of sensitivity and specificity, *SpEqualSe* [36]; maximization of the Youden's index [37];maximization of the accuracy, *MaxEfficiency* [38]; and maximization of the Kappa index, *MaxKappa* [39]. The optimal values and performances were provided by the R packages OptimalCutpoints [40] and epiR. The statistical tests, analyses and graphical representations were performed using R-3.2.0.

Results

Epigenetic context of *MGMT* promoter methylation and expression of *MGMT*

The fact that almost all CIMP+ glioma are predicted to have a methylated *MGMT* status using the MGMT-STP27 model [9, 11, 15] raised the question whether the functional correlation of the pattern of *MGMT* promoter methylation and *MGMT* expression is similar between CIMP+ and CIMP- glioma and thus the prediction model remains valid. The spatial pattern of the correlations between methylation of the 19 individual CpGs (7 for 27K) interrogated in the *MGMT* promoter region and *MGMT* expression is displayed separately for CIMP+ and CIMP- gliomas across tumor grades (WHO II, III, IV) (Figure 1). It was similar between CIMP+ and CIMP- gliomas, and across tumor grades. As previously observed, CpG methylation close to the initiation start site (ISS) displayed little correlation with expression. Methylation at the two CpGs (cg12434587 and cg12981137) comprised in the MGMT-STP27 model consistently exhibited substantial negative correlation with expression of *MGMT*, with maximal values close to -0.5, regardless of glioma subtype, CIMP-status, and tumor grade

(Figure 1). The pattern was also very similar in colon adenocarcinoma (TCGA-COAD), head and neck cancer (TCGA-HNSC), and lung squamous cell carcinoma (TCGA-LUSC), but not in breast cancer (TCGA-BRCA) (Supplemental Figure S2). In the latter, correlation between expression and methylation is very weak. However *MGMT* methylation is rare (see below).

The distribution of the *MGMT* score (logit-transformed probability of methylation) revealed bimodal distributions for all glioma subtypes clearly separating methylated from unmethylated (Figure 2, CIMP+ and CIMP- cases are visualized separately) and were almost superimposable onto the original GBM training set (M-GBM). Similar bimodal distributions were obtained for TCGA-COAD, TCGA-HNSC and TCGA-LUSC, while TCGA-BRCA basically only displays a peak for *MGMT* unmethylated tumors (Figure 3). The original cut-off, based on the maximized sum of sensitivity and specificity of the training cohort (M-GBM) was located at the nadir (lowest point between two populations) of the density plots in all glioma subpopulations, and including other tumor types, hence efficiently differentiating *MGMT* unmethylated and methylated (Figure 2 & 3). The majority of CIMP+ samples were *MGMT* methylated across all glioma datasets (Figure 2). Of note, samples with codeletion of 1p/19q were without exception *MGMT* methylated and displayed a high *MGMT* score confirmed in other datasets by other groups using MGMT-STP27 [14, 15]. The calculated proportions of *MGMT* methylation were 36.6% in TCGA-COAD, 31.2% for TCGA-HNSC, 16.2% in TCGA-LUSC, and 4.3 % in the TCGA-BRCA population (Figure 3) in line with the literature [41]. A meta-analysis based on 13 colon cancer studies using different technologies and comprising 2772 cases [42-53] revealed 37% (Supplemental Figure S3) that is in good agreement with the *MGMT* methylation proportion detected by MGMT-STP27 model in TCGA-COAD.

Robustness of the cut-off to varying optimization criteria

The assessment of cut-off robustness was conducted to determine how the definition of cut-off points would influence the dichotomization into unmethylated and methylated subgroups using the M-GBM dataset for which *MGMT* classification based on MSP is available. Six criteria (cost functions, see methods) were used to determine the optimal cut-off. Four yielded the same cut-off as obtained originally for the MGMT-STP27 model (0.358, Table 1). A different cut-off of 0.405 was obtained by two of the procedures (Table 1) that balance the errors among false positives (FP) and false negatives (FN) (as previously defined based on MSP) [11]. The use of this cut-off value reduced the sensitivity by 6%, but only slightly improved the specificity (<2%), while it had minor impact on the rate of good classification accuracy (Table 1). When testing the second cut-off (0.405) on the 788 glioma samples, we only identified five discrepancies, two for the training dataset (M-GBM), two for the TCGA-Glioma-II/III dataset and one for the T-Glioma-II/III dataset. No discrepancy was observed for TCGA-GBM-27, TCGA-GBM-450, and VB-Glioma-III datasets.

Association of CNA at the *MGMT* Locus and CIMP status on Expression of *MGMT*

Loss of the chromosomal region comprising the *MGMT* gene (10q26) is common in GBM (>80%) as opposed to lower grade glioma. We assessed, whether there is a statistical relation (an “effect”) between gene dosage, methylation, and expression of the *MGMT* gene using an additive model. Promoter methylation significantly affected *MGMT* expression in all glioma subtypes and grades (Table 2). Loss of 10q26 had a significant effect on expression in the lower grade glioma populations (p-value=0.003, T-Glioma-II/III; p-value=0.001, TCGA-Glioma-II/III; Table 2), while the effect was not significant in GBM (p-value=0.692, TCGA-GBM-450; p-value=0.848, TCGA-GBM-27; p-value=0.544, M-GBM; Table 2, Figure 4). In

the other cancer types, we observed that promoter methylation was significantly associated with *MGMT* expression (p-value=0.001, TCGA-COAD; p-value=0.001, TCGA-HNSC; p-value=0.001 TCGA-LUSC; Table 2, Supplemental Figure S4). No significant associations were detected between 10q26.3 deletion and *MGMT* expression, but such deletion events were rare in TCGA-LUSC (4%), TCGA-COAD (2%) and TCGA-HNSC (2%) datasets that can affect the robustness of the statistical tests (Table 2).

The interaction between deletion and methylation was not significant (p=0.196, Monte-Carlo ANOVA with 999 permutations) in the TCGA-Glioma-II/III dataset, suggesting an additive effect. The other datasets could not be analyzed because the distributions of patients in each cross-category were highly unbalanced, in particular due to the high frequency of loss of one copy of chromosome 10 in GBM that harbors *MGMT* (10q26) that can reduce the power of the statistical tests. Further, the CIMP status did not significantly affect the expression of the *MGMT* gene (Supplemental Table S3 and Supplemental Figure S5) in the LGG populations and it was not reasonably testable in the GBM populations considering the very low frequency of this event (7%, Supplemental Table S2).

Effect of tumor matrix (frozen versus FFPE)

The beadchip platform can be used for frozen and with the addition of a restoration step also for formalin fixed paraffin embedded (FFPE) samples. Here we tested whether datasets originating from different sample matrices can be combined. The VB-Glioma-III dataset, containing 51 frozen samples and 59 FFPE samples, was analyzed (Supplemental Table S1). The distributions of the *MGMT* scores calculated for FFPE and frozen samples, respectively, were not significantly different (p=0.253, Kolmogorov-Smirnov test, Supplemental Figure S6). Furthermore, the original cut-off of 0.3582 efficiently differentiated the unmethylated

and methylated *MGMT* promoters for FFPE tissues. Hence, the two datasets were combined for the present study.

Effect of data preprocessing

The datasets M-GBM and TCGA-GBM-450 were used to compare five normalization and preprocessing procedures for HM-450K with the original (raw) preprocessing used to build the model MGMT-STP27 (Figure 5, Supplemental Figure S7, Supplemental Table S4). The control normalization and preprocessing including only background correction lead to a slight underestimation of the methylation probabilities compared to the standard procedure. However, we only observed three (2.5%) differently reclassified samples for TCGA-GBM-450 (Figure S7) and four (5.9 %) for the training dataset, M-GBM (Figure 5). The background correction based on normal-exponential deconvolution (Noob) (Supplemental Table S4) similarly underestimated the methylation probabilities. Five and four samples were misclassified for TCGA-GBM-450 and M-GBM, respectively. In contrast, the SWAN normalization resulted in a slight overestimation of the methylation probabilities. Five (4.1%) and one (1.5%) reclassified samples were detected for TCGA-GBM-450 and M-GBM, respectively (Supplemental Table S4). In contrast, the concordance between the initial classification and outputs resulting from a procedure using quantile normalization separately on each signal was extremely low (Figure 5C and Supplemental Figure S7C), indicating incompatibility between this procedure and the current MGMT-STP27 default parameters.

For the HM-27K platform, we investigated the cohort of 241 TCGA GBM samples (TCGA-GBM-27) and compared the *MGMT* scores obtained with raw data (TCGA level 1) and already preprocessed data including Noob background correction (Level 2, preprocessed data) (Supplemental Figure S7F and G, Supplemental Table S4). The methylation probabilities

trended to be underestimated for data from Level 2 (Supplemental Figure S7G), with 9 (3.7%) misclassified samples in comparison with the original results [11]. The use of Level 1 (raw) data provided similar predictions as originally determined.

In spite of a moderate bias for probability estimation, the final *MGMT* classification was robust for both Infinium platforms, except for quantile normalization. The effect of data preprocessing on classification was limited. The strong bimodal distribution of the *MGMT* scores and the low proportion of samples contained in the intermediate probability range [0.3; 0.7] favor this robust behavior.

Discussion

In the present study we tested the robustness of the MGMT-STP27 model to predict the *MGMT* methylation status. Considerations included biological effects, such as the context of pathogenetic and epigenetic alterations of the tumors analyzed. On the other hand we investigated technical issues, ranging from impact of tissue matrix to preprocessing of the data and cut-off definitions.

First, we demonstrated that the functional relationship, corresponding to the pattern of the spatial correlation between methylation and expression was preserved across glioma subtypes, WHO grade and CIMP-status, and was also valid in other tumor types. The probes of the two CpGs used in the MGMT-STP27 model displayed a strong negative correlation between methylation and expression in all datasets. Clear bimodal distributions of the *MGMT* scores allowing classification into methylated and unmethylated samples was conserved across all datasets. The original cut-off used for dichotomization was located at the nadir of the distributions in all datasets analyzed including the non-glioma tumor cohorts. The robustness

of the original cut-off was further confirmed by comparing different procedures of cut-off optimization that had little effect on classification.

An essential issue for any model is the estimation of the uncertainty related to the prediction. The computation of the confidence intervals as proposed in the new R package mgmtstp27 permits evaluation of the pertinence and quality of the classification for a new sample as we have reported previously [11]. The implemented quality control procedures allow visualization of multiple or single sample predictions in comparison to the training set (Figure 6). The confidence intervals on the methylation status probability are important to assess the confidence in the classification, particularly useful when the prediction is close to the cut-off. This is clinically relevant in particular when deciding not to give TMZ, e.g in clinical trials where patients are selected according to their *MGMT* status [54], or to use TMZ as mono-therapy, as recommended for elderly patients whose GBM is *MGMT* methylated [4, 55]. In other tumor types, like metastatic colon cancer, alkylating agents may be a treatment option among others [56], and only patients with a higher *MGMT* score may be considered.

A significant effect of gene dosage on *MGMT* expression was observed in LGG that usually have two gene copies in contrast to GBM. This may indicate that not both copies are methylated, which cannot be distinguished by the assay, potentially yielding some expression conferring residual repair function in these tumors. In other words, residual *MGMT*-related resistance to TMZ may not be excluded in LGG, even when they are classified methylated. In GBM the effect of gene dosage was not statistically evaluable due to the characteristic high frequency of loss of one copy of chromosome 10, home of *MGMT*. In contrast, no effect on expression was observed for CIMP in LGG, while it was not testable in GBM. However, it is of note that the *MGMT* status in LGG is not independent of CIMP due to the nested relationship.

The effect of preprocessing on the classification was relatively moderate for the tested scenarios, except for quantile normalization that is clearly not suitable. For the other methods, the effect on classification was minor due to the strong bimodal distribution with few samples close to the cut-off. Additionally, the classification robustness can be explained by the limited difference of the probe specific bias in M-values among background correction methods for Infinium chemistry type I probes [17]. This corroborates our previous results [11] showing that the M-value distributions of the two selected probes from the training dataset (M-GBM) and TCGA-GBM-27 were not significantly different.

A major constraint for direct inter-study prediction are normalization procedures, such as quantile methods, as they can be affected by biological differences in the sample populations across studies and by study design (e.g. presence or absences of control or non-tumor samples, overrepresentation of subgroups). Testing of five preprocessing/normalizing procedures revealed that quantile normalization was clearly not compatible with MGMT-STP27, while for the other four only moderate differences were observed. Unless the compatibility is tested, we recommend to use the raw data (format IDAT), and convert the Red/Green channel from the Illumina methylation array into methylation signal, without using any normalization. This avoids potential dataset dependent biases associated with normalization procedures and allows for single sample prediction that is an essential requirement for clinical utility [57]. In practice, functions such as preprocessRaw or methylumIDAT from the R packages minfi [58] and methylumi [59] offer appropriate solutions to import and to preprocess the raw HM-450K and HM-27K data.

Overall the MGMT-STP27 is a robust model for classification of samples into *MGMT* methylated and unmethylated that is independent on glioma subtype, is adapted for single sample prediction, and is also valid in other tumor types.

Note Added in Proof

The new Infinium MethylationEPIC BeadChip (850K) proposed by Illumina contains both probes used in the model MGMT-STP27. The annotations (eg, chemistry type and probe location) suggest that our model can be extended to this new platform.

Acknowledgements

This work was supported by the Swiss National Science Foundation (3100A-138116), the Swiss Bridge Award 2011, and the Swiss Cancer League (KFS-29-02-2012). The results published here are in part based upon data generated by The Cancer TCGA Genome Atlas pilot project established by the NCI and NHGRI.

References

1. Sturm D, Witt H, Hovestadt V, Khuong-Quang DA, Jones DT, Konermann C, Pfaff E, Tonjes M, Sill M, Bender S, Kool M, Zapatka M, Becker N, Zucknick M, Hielscher T, Liu XY, Fontebasso AM, Ryzhova M, Albrecht S, Jacob K, Wolter M, Ebinger M, Schuhmann MU, van Meter T, Fruhwald MC, Hauch H, Pekrun A, Radlwimmer B, Niehues T, von Komorowski G, Durken M, Kulozik AE, Madden J, Donson A, Foreman NK, Drissi R, Fouladi M, Scheurlen W, von Deimling A, Monoranu C, Roggendorf W, Herold-Mende C, Unterberg A, Kramm CM, Felsberg J, Hartmann C, Wiestler B, Wick W, Milde T, Witt O, Lindroth AM, Schwartzentruber J, Faury D, Fleming A, Zakrzewska M, Liberski PP, Zakrzewski K, Hauser P, Garami M, Klekner A, Bognar L, Morrissy S, Cavalli F, Taylor MD, van Sluis P, Koster J, Versteeg R, Volckmann R, Mikkelsen T, Aldape K, Reifenberger G, Collins VP, Majewski J, Korshunov A, Lichter P, Plass C, Jabado N, Pfister SM: Hotspot mutations in H3F3A and IDH1 define distinct epigenetic and biological subgroups of glioblastoma. Cancer Cell 2012, 22:425-37.

2. Brennan CW, Verhaak RG, McKenna A, Campos B, Noushmehr H, Salama SR, Zheng S, Chakravarty D, Sanborn JZ, Berman SH, Beroukhim R, Bernard B, Wu CJ, Genovese G, Shmulevich I, Barnholtz-Sloan J, Zou L, Vegesna R, Shukla SA, Ciriello G, Yung WK, Zhang W, Sougnez C, Mikkelsen T, Aldape K, Bigner DD, Van Meir EG, Prados M, Sloan A, Black KL, Eschbacher J, Finocchiaro G, Friedman W, Andrews DW, Guha A, Iacocca M, O'Neill BP, Foltz G, Myers J, Weisenberger DJ, Penny R, Kucherlapati R, Perou CM, Hayes DN, Gibbs R, Marra M, Mills GB, Lander E, Spellman P, Wilson R, Sander C, Weinstein J, Meyerson M, Gabriel S, Laird PW, Haussler D, Getz G, Chin L, Network TR: The somatic genomic landscape of glioblastoma. Cell 2013, 155:462-77.

3. Hegi ME, Diserens AC, Gorlia T, Hamou MF, de Tribolet N, Weller M, Kros JM, Hainfellner JA, Mason W, Mariani L, Bromberg JE, Hau P, Mirimanoff RO, Cairncross JG, Janzer RC, Stupp R: MGMT gene silencing and benefit from temozolomide in glioblastoma. N Engl J Med 2005, 352:997-1003.

4. Malmstrom A, Gronberg BH, Marosi C, Stupp R, Frappaz D, Schultz H, Abacioglu U, Tavelin B, Lhermitte B, Hegi ME, Rosell J, Henriksson R, Nordic Clinical Brain Tumour Study G: Temozolomide versus standard 6-week radiotherapy versus hypofractionated radiotherapy in patients older than 60 years with glioblastoma: the Nordic randomised, phase 3 trial. Lancet Oncol 2012, 13:916-26.

5. Wick W, Platten M, Meisner C, Felsberg J, Tabatabai G, Simon M, Nikkhah G, Papsdorf K, Steinbach JP, Sabel M, Combs SE, Vesper J, Braun C, Meixensberger J, Ketter R, Mayer-Steinacker R, Reifenberger G, Weller M: Temozolomide chemotherapy alone versus radiotherapy alone for malignant astrocytoma in the elderly: the NOA-08 randomised, phase 3 trial. Lancet Oncol 2012, 13:707-15.

6. Weller M, Stupp R, Hegi ME, van den Bent M, Tonn JC, Sanson M, Wick W, Reifenberger G: Personalized care in neuro-oncology coming of age: why we need MGMT and 1p/19q testing for malignant glioma patients in clinical practice. Neuro Oncol 2012, 14 Suppl 4:iv100-iv8.

7. van den Bent MJ, Dubbink HJ, Sanson M, van der Lee-Haarloo CR, Hegi M, Jeuken JW, Ibdaih A, Brandes AA, Taphoorn MJ, Frenay M, Lacombe D, Gorlia T, Dinjens WN, Kros JM: MGMT promoter methylation is prognostic but not predictive for outcome to adjuvant PCV chemotherapy in anaplastic oligodendroglial tumors: A report from EORTC Brain Tumor Group Study 26951. J Clin Oncol 2009, 9:5881-6.

8. Wick W, Meisner C, Hentschel B, Platten M, Schilling A, Wiestler B, Sabel MC, Koeppen S, Ketter R, Weiler M, Tabatabai G, von Deimling A, Gramatzki D, Westphal M, Schackert G, Loeffler M, Simon M, Reifenberger G, Weller M: Prognostic or predictive value of MGMT promoter methylation in gliomas depends on IDH1 mutation. Neurology 2013, 81:1515-22.

9. van den Bent MJ, Erdem Eraslan L, Idbaih A, de Rooi JJ, Eilers PH, Spliet W, den Dunnen WF, Tijssen C, Wesseling P, Sillevis Smitt PA, Kros JM, Gorlia T, French PJ: MGMT-STP27 methylation status as predictive marker for response to PCV in anaplastic oligodendrogliomas and oligoastrocytomas. A report from EORTC study 26951. Clin Cancer Res 2013, 19:5513-22.

10. Turcan S, Rohle D, Goenka A, Walsh LA, Fang F, Yilmaz E, Campos C, Fabius AW, Lu C, Ward PS, Thompson CB, Kaufman A, Guryanova O, Levine R, Heguy A, Viale A, Morris LG, Huse JT, Mellinghoff IK, Chan TA: IDH1 mutation is sufficient to establish the glioma hypermethylator phenotype. Nature 2012, 483:479-83.

11. Bady P, Sciuscio D, Diserens AC, Bloch J, van den Bent MJ, Marosi C, Dietrich PY, Weller M, Mariani L, Heppner FL, McDonald DR, Lacombe D, Stupp R, Delorenzi M, Hegi ME: MGMT methylation analysis of glioblastoma on the Infinium methylation BeadChip identifies two distinct CpG regions associated with gene silencing and outcome, yielding a prediction model for comparisons across datasets, tumor grades, and CIMP-status. Acta Neuropathol 2012, 124:547-60.

12. Figueroa ME, Abdel-Wahab O, Lu C, Ward PS, Patel J, Shih A, Li Y, Bhagwat N, Vasanthakumar A, Fernandez HF, Tallman MS, Sun Z, Wolniak K, Peeters JK, Liu W, Choe SE, Fantin VR, Paietta E, Löwenberg B, Licht JD, Godley LA, Delwel R, Valk PJM, Thompson CB, Levine RL, Melnick A: Leukemic IDH1 and

IDH2 mutations result in a hypermethylation phenotype, disrupt TET2 function, and impair hematopoietic differentiation. Cancer Cell 2010, 18:553-67.
13. Xu W, Yang H, Liu Y, Yang Y, Wang P, Kim S-H, Ito S, Yang C, Wang P, Xiao M-T, Liu L-x, Jiang W-q, Liu J, Zhang J-y, Wang B, Frye S, Zhang Y, Xu Y-h, Lei Q-y, Guan K-L, Zhao S-m, Xiong Y: Oncometabolite 2-Hydroxyglutarate Is a Competitive Inhibitor of [alpha]-Ketoglutarate-Dependent Dioxygenases. Cancer Cell 2011, 19:17-30.
14. Wiestler B, Capper D, Sill M, Jones DT, Hovestadt V, Sturm D, Koelsche C, Bertoni A, Schweizer L, Korshunov A, Weiss EK, Schliesser MG, Radbruch A, Herold-Mende C, Roth P, Unterberg A, Hartmann C, Pietsch T, Reifenberger G, Lichter P, Radlwimmer B, Platten M, Pfister SM, von Deimling A, Weller M, Wick W: Integrated DNA methylation and copy-number profiling identify three clinically and biologically relevant groups of anaplastic glioma. Acta Neuropathol 2014, 128:561-71.
15. Eckel-Passow JE, Lachance DH, Molinaro AM, Walsh KM, Decker PA, Sicotte H, Pekmezci M, Rice T, Kosel ML, Smirnov IV, Sarkar G, Caron AA, Kollmeyer TM, Praska CE, Chada AR, Halder C, Hansen HM, McCoy LS, Bracci PM, Marshall R, Zheng S, Reis GF, Pico AR, O'Neill BP, Buckner JC, Giannini C, Huse JT, Perry A, Tihan T, Berger MS, Chang SM, Prados MD, Wiemels J, Wiencke JK, Wrensch MR, Jenkins RB: Glioma groups based on 1p/19q, IDH, and TERT promoter mutations in tumors. N Engl J Med 2015, 372:2499-508.
16. Maksimovic J, Gordon L, Oshlack A: SWAN: Subset-quantile within array normalization for illumina infinium HumanMethylation450 BeadChips. Genome Biol 2012, 13:R44.
17. Triche TJ, Jr., Weisenberger DJ, Van Den Berg D, Laird PW, Siegmund KD: Low-level processing of Illumina Infinium DNA Methylation BeadArrays. Nucleic Acids Res 2013, 41:e90.
18. Morris TJ, Butcher LM, Feber A, Teschendorff AE, Chakravarthy AR, Wojdacz TK, Beck S: ChAMP: 450k Chip Analysis Methylation Pipeline. Bioinformatics 2014, 30:428-30.
19. Feber A, Guilhamon P, Lechner M, Fenton T, Wilson GA, Thirlwell C, Morris TJ, Flanagan AM, Teschendorff AE, Kelly JD, Beck S: Using high-density DNA methylation arrays to profile copy number alterations. Genome Biol 2014, 15:R30.
20. Kurscheid S, Bady P, Sciuscio D, Samarzija I, Shay T, Vassallo I, van Criekinge W, Daniel RT, van den Bent MJ, Marosi C, Weller M, Mason WP, Domany E, Stupp R, Delorenzi M, Hegi ME: Chromosome 7 gain and DNA hypermethylation at the HOXA10 locus are associated with expression of a stem cell related HOX-signature in glioblastoma. Genome Biol 2015, 16.
21. Murat A, Migliavacca E, Gorlia T, Lambiv WL, Shay T, Hamou MF, de Tribolet N, Regli L, Wick W, Kouwenhoven MC, Hainfellner JA, Heppner FL, Dietrich PY, Zimmer Y, Cairncross JG, Janzer RC, Domany E, Delorenzi M, Stupp R, Hegi ME: Stem cell-related "self-renewal" signature and high epidermal growth factor receptor expression associated with resistance to concomitant chemoradiotherapy in glioblastoma. J Clin Oncol 2008, 26:3015-24.
22. Noushmehr H, Weisenberger DJ, Diefes K, Phillips HS, Pujara K, Berman BP, Pan F, Pelloski CE, Sulman EP, Bhat KP, Verhaak RG, Hoadley KA, Hayes DN, Perou CM, Schmidt HK, Ding L, Wilson RK, Van Den Berg D, Shen H, Bengtsson H, Neuvial P, Cope LM, Buckley J, Herman JG, Baylin SB, Laird PW, Aldape K, Cancer Genome Atlas Research Network: Identification of a CpG island methylator phenotype that defines a distinct subgroup of glioma. Cancer Cell 2010, 17:419-20.
23. The Cancer Genome Atlas Consortium: Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature 2008, 455:1061-8.
24. The Cancer Genome Atlas Network: Comprehensive, integrative genomic analysis of diffuse lower-grade gliomas. N Engl J Med 2015, 372:2481-98.
25. Du P, Zhang X, Huang CC, Jafari N, Kibbe WA, Hou L, Lin SM: Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. BMC Bioinformatics 2010, 11:587.
26. Faraway JJ: Extending linear models with R: Generalized linear, mixed effects and nonparametric regression models. Boca Raton, FL: Chapman & Hall/CRC, 2006.
27. Staaf J, Vallon-Christersson J, Lindgren D, Juliusson G, Rosenquist R, Hoglund M, Borg A, Ringner M: Normalization of Illumina Infinium whole-genome SNP data improves copy number estimates and allelic intensity ratios. BMC Bioinformatics 2008, 9:409.
28. van de Wiel MA, Brosens R, Eilers PH, Kumps C, Meijer GA, Menten B, Sistermans E, Speleman F, Timmerman ME, Ylstra B: Smoothing waves in array CGH tumor profiles. Bioinformatics 2009, 25:1099-104.
29. van de Wiel MA, Kim KI, Vosse SJ, van Wieringen WN, Wilting SM, Ylstra B: CGHcall: calling aberrations for array CGH tumor profiles. Bioinformatics 2007, 23:892-4.
30. Olshen AB, Venkatraman ES, Lucito R, Wigler M: Circular binary segmentation for the analysis of array-based DNA copy number data. Biostatistics 2004, 5:557-72.
31. Patefield WM: Algorithm AS159. An efficient method of generating r x c tables with given row and column totals. Applied Statistics 1981, 30:91-7.

32. Manly BFJ: Randomization, bootstrap and Monte-Carlo methods in biology. Third edition. London: Chapman & Hall/CRC, 2006.
33. Kherad-Pajouh S, Renaud O: An exact permutation method for testing any effect in balanced and unbalanced fixed effect ANOVA. Computational Statistics & Data Analysis 2010, 54:1881-93.
34. Riddle DL, Stratford PW: Interpreting validity indexes for diagnostic tests: an illustration using the Berg balance test. Phys Ther 1999, 79:939-48.
35. Lewis JD, Chuai S, Nessel L, Lichtenstein GR, Aberra FN, Ellenberg JH: Use of the noninvasive components of the Mayo score to assess clinical response in ulcerative colitis. Inflamm Bowel Dis 2008, 14:1660-6.
36. Hosmer DW, Lemeshow S: Applied logistic regression. Chichester, NY: Wiley Interscience, 2000.
37. Youden WJ: Index for rating diagnostic tests. Cancer 1950, 3:32-5.
38. Feinstein SH: The accuracy of diver sound localization by pointing. Undersea Biomed Res 1975, 2:173-84.
39. Cohen J: A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement 1960, 20:37-46.
40. López-Ratón M, Rodríguez-Álvarez MX, Cadarso-Suárez C, Gude-Sampedro F: OptimalCutpoints: An R Package for Selecting Optimal Cutpoints in Diagnostic Tests. J Stat Soft 2014, 61:1-36.
41. Esteller M, Herman JG: Generating mutations but providing chemosensitivity: the role of O6-methylguanine DNA methyltransferase in human cancer. Oncogene 2004, 23:1-8.
42. Alonso S, Dai Y, Yamashita K, Horiuchi S, Dai T, Matsunaga A, Sánchez-Muñoz R, Bilbao-Sieyro C, Díaz-Chico JC, Chernov AV, Strongin AY, Perucho M: Methylation of MGMT and ADAMTS14 in normal colon mucosa: biomarkers of a field defect for cancerization preferentially targeting elder African-Americans. Oncotarget 2015, 6:3420-31.
43. Azuara D, Rodriguez-Moranta F, de Oca J, Soriano-Izquierdo A, Mora J, Guardiola J, Biondo S, Blanco I, Peinado MA, Moreno V, Esteller M, Capellá G: Novel methylation panel for the early detection of colorectal tumors in stool DNA. Clinical Colorectal Cancer 2010, 9:168-76.
44. Farzanehfar M, Vossoughinia H, Jabini R, Tavassoli A, Saadatnia H, Khorashad AK, Ahadi M, Afzalaghaee M, Ghayoor Karimiani E, Mirzaei F, Ayatollahi H: Evaluation of methylation of MGMT (O(6)-methylguanine-DNA methyltransferase) gene promoter in sporadic colorectal cancer. DNA Cell Biol 2013, 32:371-7.
45. Shima K, Morikawa T, Baba Y, Nosho K, Suzuki M, Yamauchi M, Hayashi M, Giovannucci E, Fuchs CS, Ogino S: MGMT promoter methylation, loss of expression and prognosis in 855 colorectal cancers. Cancer Causes Control 2011, 22:301-9.
46. Esteller M, Toyota M, Sanchez-Cespedes M, Capella G, Peinado MA, Watkins DN, Issa J-PJ, Sidransky D, Baylin SB, Herman JG: Inactivation of the DNA repair Gene O6-Methylguanine-DNA Methyltransferase by promoter hypermethylation Is associated with G to A mutations in K-ras in colorectal tumorigenesis. Cancer Research 2000, 60:2368-71.
47. Lee K-H, Lee J-S, Nam J-H, Choi C, Lee M-C, Park C-S, Juhng S-W, Lee J-H: Promoter methylation status of hMLH1, hMSH2, and MGMT genes in colorectal cancer associated with adenoma–carcinoma sequence. Langenbecks Arch Surg 2011, 396:1017-26.
48. Coppedè F, Migheli F, Lopomo A, Failli A, Legitimo A, Consolini R, Fontanini G, Sensi E, Servadio A, Seccia M, Zocco G, Chiarugi M, Spisni R, Migliore L: Gene promoter methylation in colorectal cancer and healthy adjacent mucosa specimens: Correlation with physiological and pathological characteristics, and with biomarkers of one-carbon metabolism. Epigenetics 2014, 9:621-33.
49. Kim J, Choi J, Roh S, Cho D, Kim T, Kim Y: Promoter methylation of specific genes is associated with the phenotype and progression of colorectal adenocarcinomas. Ann Surg Oncol 2010, 17:1767-76.
50. Chen SP, Chiu SC, Wu CC, Lin SZ, Kang JC, Chen YL, Lin PC, Pang CY, Harn HJ: The association of methylation in the promoter of APC and MGMT and the prognosis of Taiwanese CRC patients. Genet Test Mol Biomarkers 2009, 13:67-71.
51. Krtolica K, Krajnovic M, Usaj-Knezevic S, Babic D, Jovanovic D, Dimitrijevic B: Comethylation of p16 and MGMT genes in colorectal carcinoma: Correlation with clinicopathological features and prognostic value. World Journal of Gastroenterology : WJG 2007, 13:1187-94.
52. Nagasaka T, Goel A, Notohara K, Takahata T, Sasamoto H, Uchida T, Nishida N, Tanaka N, Boland CR, Matsubara N: Methylation pattern of the O6-methylguanine-DNA methyltransferase gene in colon during progressive colorectal tumorigenesis. Int J Cancer 2008, 122:2429-36.
53. Nagasaka T, Sharp GB, Notohara K, Kambara T, Sasamoto H, Isozaki H, MacPhee DG, Jass JR, Tanaka N, Matsubara N: Hypermethylation of O6-methylguanine-DNA methyltransferase promoter may predict nonrecurrence after chemotherapy in colorectal cancer cases. Clin Cancer Res 2003, 9:5306-12.
54. Wick W, Gorlia T, Bady P, Platten M, van den Bent MJ, Taphoorn MJB, Steuve J, Brandes AA, Hamou MF, Wick A, Kosch MA, Weller M, Stupp R, Roth P, Golfinopoulos V, Frenel J-S, Campone M, Ricard D, Marosi C, Villa S, Weyerbrock A, Hopkins K, Homicsko K, Lhermitte B, Pesce GA, Hegi ME: Phase II study of

563 radiotherapy and temsirolimus versus radiochemotherapy with temozolomide in patients with newly diagnosed
564 glioblastoma without MGMT promoter hypermethylation (EORTC 26082). Clin Cancer Res in press.
565 55. Weller M, Pfister SM, Wick W, Hegi ME, Reifenberger G, Stupp R: Molecular neuro-oncology in clinical
566 practice: a new horizon. Lancet Oncol 2013, 14:e370-9.
567 56. Amatu A, Sartore-Bianchi A, Moutinho C, Belotti A, Bencardino K, Chirico G, Cassingena A, Rusconi F,
568 Esposito A, Nichelatti M, Esteller M, Siena S: Promoter CpG island hypermethylation of the DNA repair
569 enzyme MGMT predicts clinical response to dacarbazine in a phase II study for metastatic colorectal cancer.
570 Clin Cancer Res 2013, 19:2265-72.
571 57. Cheng C, Shen K, Song C, Luo J, Tseng GC: Ratio adjustment and calibration scheme for gene-wise
572 normalization to enhance microarray inter-study prediction. Bioinformatics 2009, 25:1655-61.
573 58. Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, Irizarry RA: Minfi: a
574 flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays.
575 Bioinformatics 2014, 30:1363-9.
576 59. Davis S, Du P, Bilke S, Triche Tj, Bootwalla M: methylumi: Handle Illumina methylation data. R package
577 version 2.12.0. . 2014.

578

Table 1. Sensitivity analysis of the cut-offs associated with the model MGMT-STP27 compared to classification based on MSP (M-GBM dataset).

*Criterion	cutoff	FP	FN	opt criterion	prev meth	sens	spec	diag acc	Youden
†Youden [37]	0.3582	4	1	0.8576	0.5147	0.9688	0.8889	0.9265	0.8576
MaxEfficiency [38]	0.3582	4	1	0.9265	0.5147	0.9688	0.8889	0.9265	0.8576
MaxKappa [39]	0.3582	4	1	0.8532	0.5147	0.9688	0.8889	0.9265	0.8576
MaxProdSpSe[35]	0.3582	4	1	0.8611	0.5147	0.9688	0.8889	0.9265	0.8576
SpEqualSe [36]	0.4055	3	3	0.0104	0.4706	0.9063	0.9167	0.9118	0.8229
MaxSpSe [34]	0.4055	3	3	0.9063	0.4706	0.9063	0.9167	0.9118	0.8229

*See methods for explication of Criterion: *Youden*, maximization of Youden's index; *MaxEfficiency*, maximization of accuracy; *MaxKappa*, maximization of Kappa index; *MaxProdSpSe*, maximization of product of sensitivity and specificity; *SpEqualSe*, equality (balance) of sensitivity and specificity; *MaxSpSE*: maximization of sensitivity and specificty

†The maximization of the sum of specificity and sensitivity used for developing MGMT-STP27 [11] was identical to the maximization of Youden's index.

Abbreviations: FP, false positives; FN, false negatives; prev meth, prevalence of methylation; sens, sensitivity; spec, specificity; diag acc; diagnostic accuracy; Youden, Youden index

590 **Table 2** Effects of CNA and DNA methylation on expression of *MGMT* in Glioma and Non-
591 Glioma tumors.

Tumor	Dataset (N)	Type	Variables	% (N)	F-statistic	‡Pvalue
GLIOMA						
	M-GBM (59)	GBM	MGMTmeth	55.93 (33)	10.966	**0.003**
			*10q26.3 loss	93.22 (55)	0.402	0.544
	TCGA-GBM-27 (212)	GBM	MGMTmeth	50.94 (108)	139.656	**0.001**
			*10q26.3 loss	86.32 (183)	0.04	0.848
	TCGA-GBM-450 (67)	GBM	MGMTmeth	43.28 (29)	8.058	**0.007**
			*10q26.3 loss	73.13 (49)	0.175	0.692
	TCGA-Glioma-II/III (195)	LGG	MGMTmeth	84.62 (165)	20.63	**0.001**
			10q26.3 loss	21.54 (42)	15.232	**0.001**
	T-Glioma-II/III (48)	LGG	MGMTmeth	85.42 (41)	11.153	**0.005**
			10q26.3 loss	18.75 (9)	8.541	**0.003**
NON-GLIOMA						
	TCGA-COAD (212)	COAD	*MGMT* meth	37.26 (79)	91.4629	**0.001**
			†10q26.3 loss	1.89 (4)	0.0005	0.982
	TCGA-HNSC (393)	HNSC	*MGMT* meth	32.06 (126)	64.3487	**0.001**
			†10q26.3 loss	1.53 (6)	2.5321	0.089
	TCGA-LUSC (288)	LUSC	*MGMT* meth	16.32 (47)	53.5159	**0.001**
			†10q26.3 loss	4.17 (12)	3.6662	0.051

592 *CNA 10q26.3 very common event, unbalanced data!

593 †10q26.3 loss very rare event, unbalanced data!

594 ‡ simulated p-values estimated by Monte-Carlo procedures (999 permutations); significant p-
595 values are indicated in bold.

596

Figure Legends

Figure 1. Spatial correlation between *MGMT* expression and CpG methylation in the *MGMT* promoter. The correlation between the Infinium probes, in the *MGMT* promoter (genome assemble 37, hg19) present on the 450K and the 27K, respectively, and expression of *MGMT* is displayed for 5 glioma datasets (AFFYmetrix probe, ; RNA sequencing for TCGA-Glioma II/III). The black, green and red line correspond to the correlation for all samples, CIMP- and CIMP+ populations respectively. The CpG island located in the *MGMT* promoter region is illustrated with a green bar, and the location of the two Inifinium HM-450K/27K probes used in the model MGMT-STP27 are indicated with dark blue marks, and the transcription start site (TSS) with an arrow.

Figure 2. Distribution of the *MGMT* scores in glioma grade II-IV stratified by CIMP-status. The density plots of the *MGMT* scores, corresponding to the logit-transformed probabilities (*MGMT* score) that the *MGMT* promoter is methylated, are shown for the LGG (grade II and III) and GBM (grade IV) populations. The smoothened lines are provided by kernel density estimate, and indicate in green grade IV (GBM), in red grade III, and in blue for grade II glioma. The vertical dotted lines identify the position of the cut-off used to classify in into methylated and unmethylated *MGMT* promoter status.

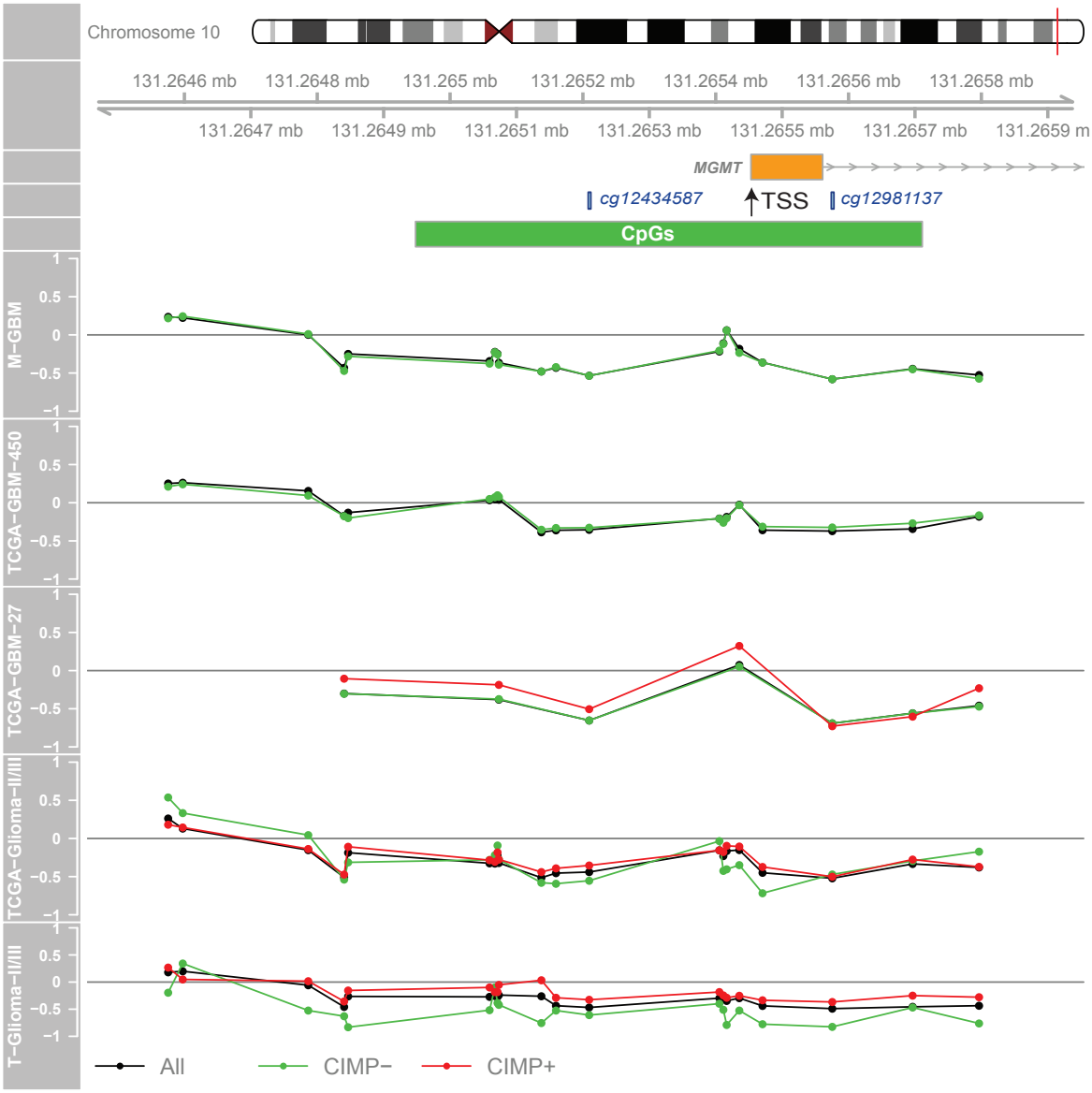
Figure 3. Distribution of *MGMT* score for non-Glioma datasets from TCGA. The score corresponds to the logit-transformed probabilities that *MGMT* promoter is methylated. The black smoothened line is provided by kernel density estimate. The vertical dotted line identifies the position of the cut-off used to determinate the *MGMT* promoter state [11]. The

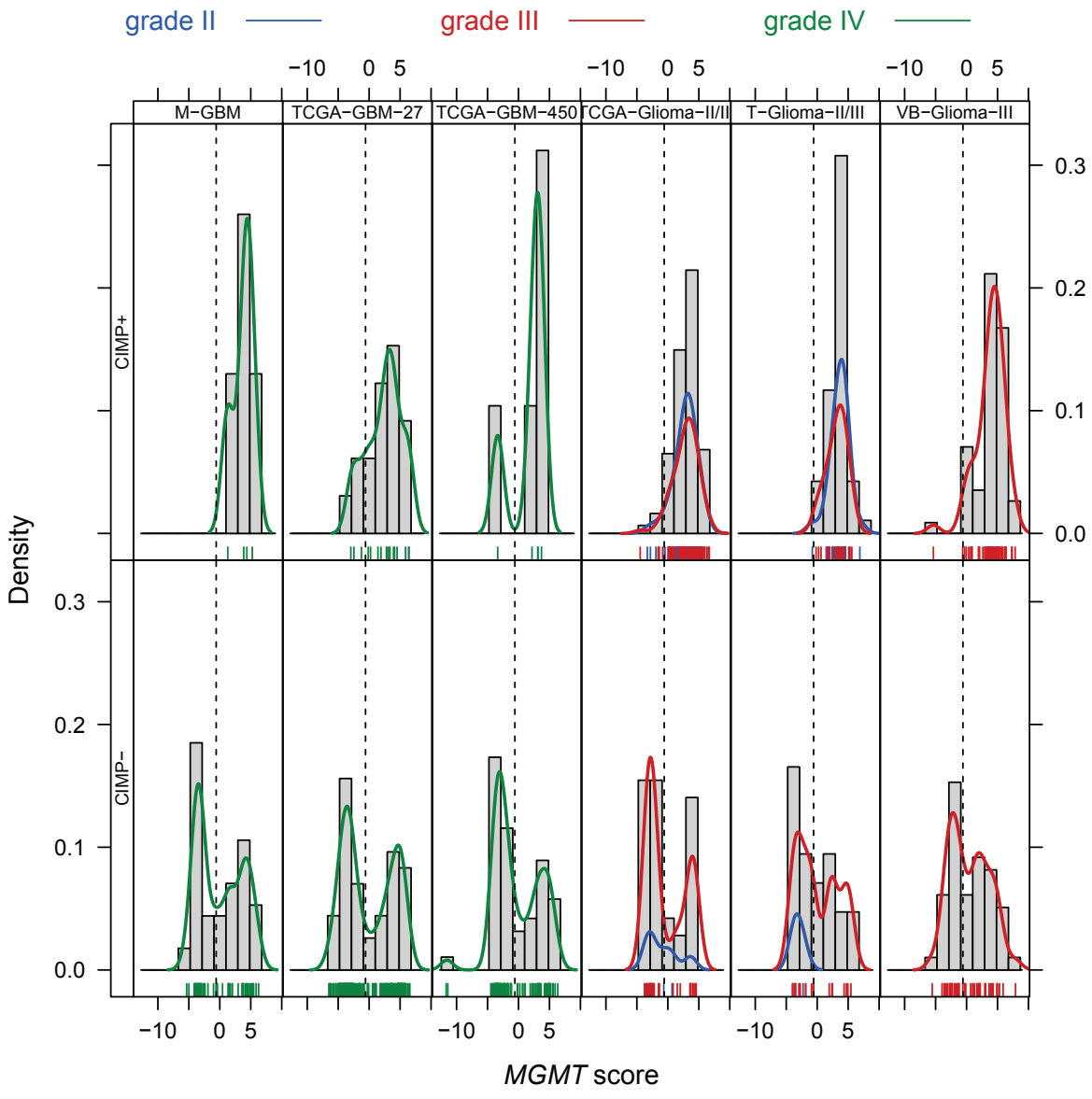
proportion of *MGMT* methylation for head and neck cancer (TCGA-HNSC) is 138/442 (31.2%, 95% confidence interval [CI, 26.9-35.8%]), 53/328 (16% [CI, 12.3-20.6%]) for lung squamous cell carcinoma (TCGA-LUSC), 13/305 (4.3% [CI, 2.3-7.2%]) for breast carcinoma (TCGA-BRCA), and 83/227 (36.6% [CI, 3.0-4.3]) for colon adenocarcinoma (TCGA-COAD).

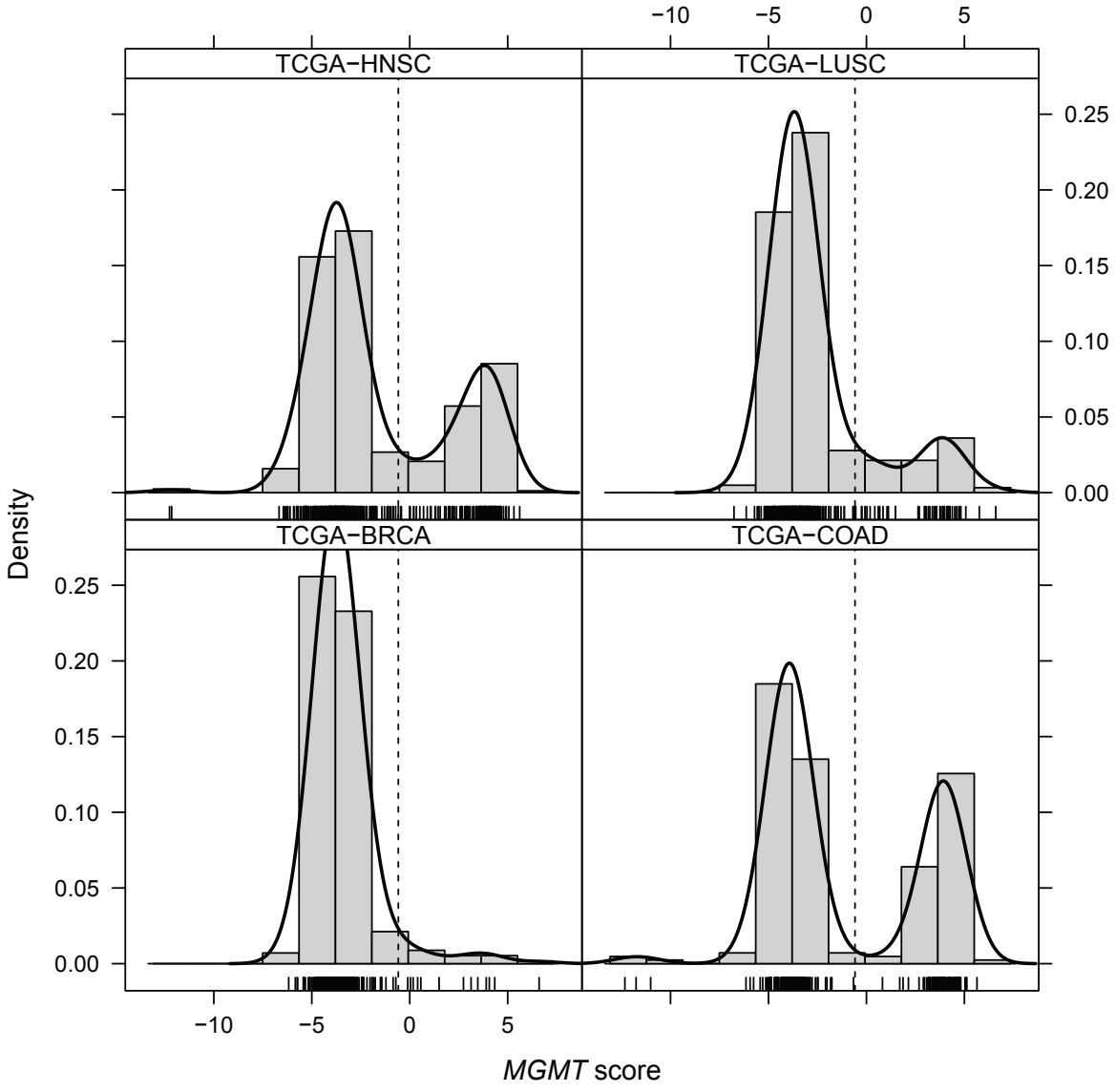
Figure 4. Boxplot representation of *MGMT* expression in function of CNA and *MGMT* methylation status in glioma grade II to IV. For each dataset the number of samples for each subpopulation is provided next to the box. Subpopulations with deletions at 10q26.3 (del) are indicated in white, the ones with normal copy number (no-del) in black. *MGMT* methylated, M; *MGMT* unmethylated, U.

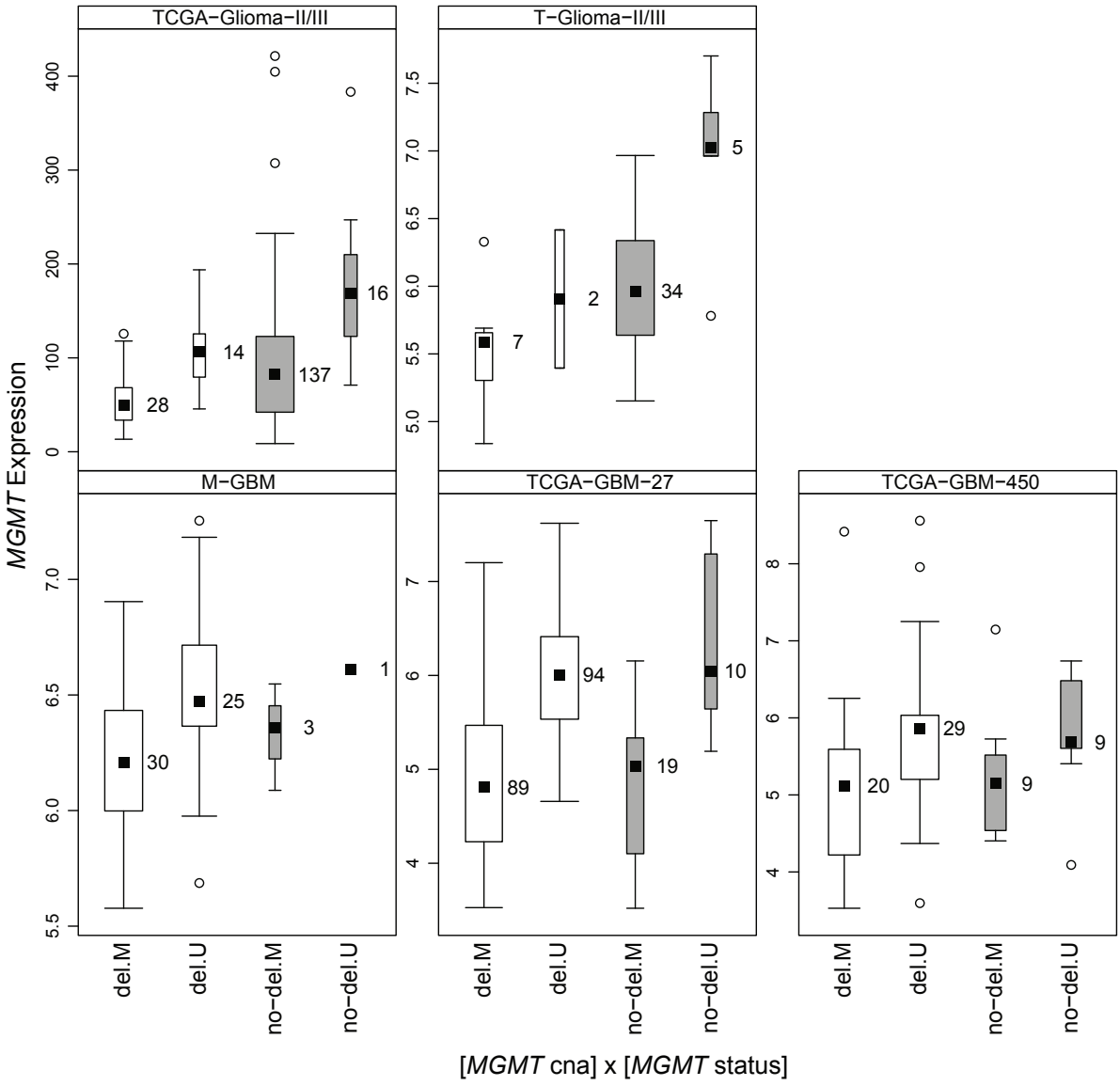
Figure 5. Effect of data preprocessing procedures on *MGMT* classification. Paired comparisons of the probabilities of *MGMT* promoter methylation (MGMT-STP27) between preprocessing procedures for the M-GBM dataset. Five preprocessing procedures for the HM-450K platform were compared with the initial procedure used to build the model MGMT-STP27. The outputs from recommended preprocessing were compared with (A) outputs from the Illumina-like procedure based on control normalization (a reference sample was used during the normalization step), (B) preprocessing with Illumina-like background correction only, (C) quantile normalization, (D) SWAN normalization, and (E) Noob normalization. Each dataset contained exactly the same samples. The grey dashed lines identify the original cut-off of 0.3582. The straight, dashed black line corresponds to the equation y=x and the grey line to the loess regression, respectively. The proportions of good classification (diagnostic accuracy, DA) are provided for the original cut-off on each panel.

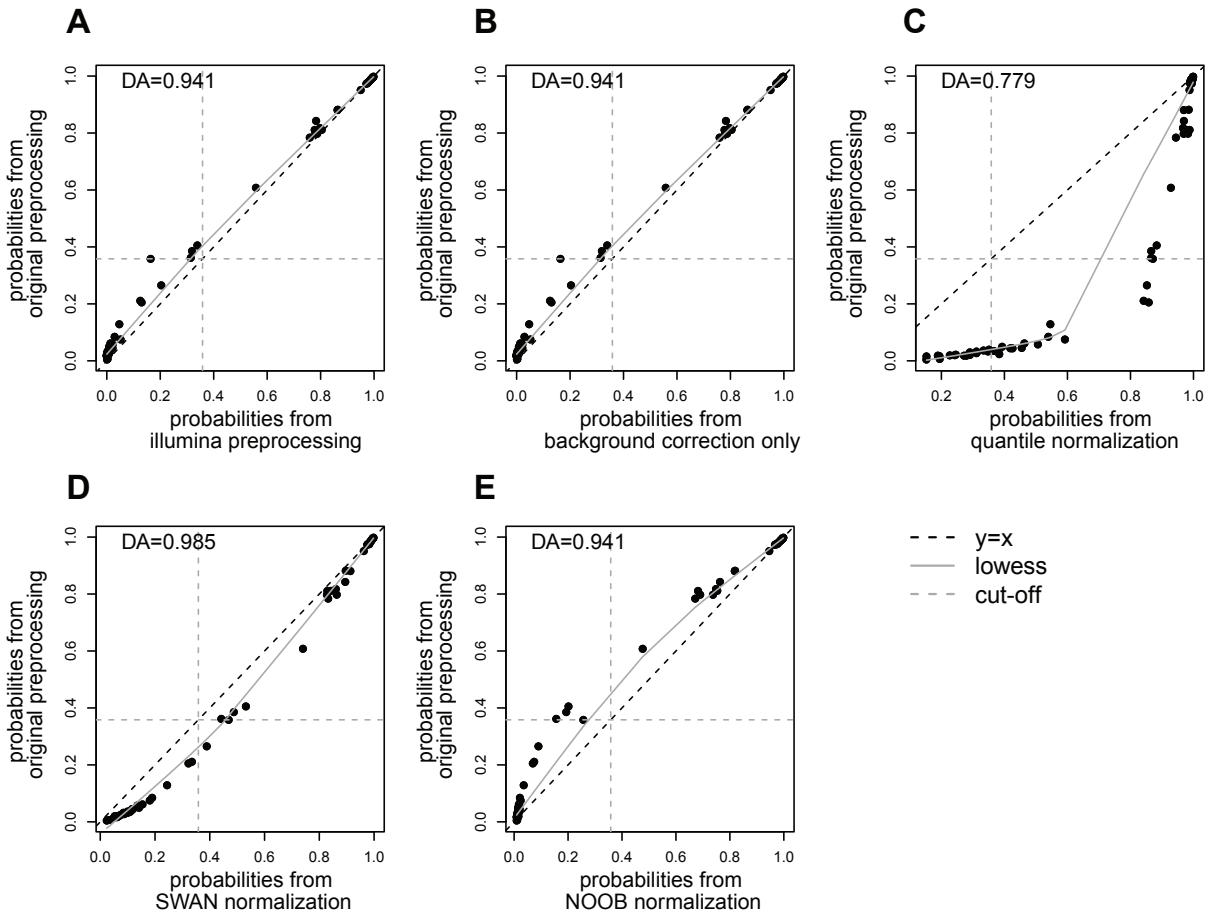
Figure 6. Quality control visualization for multi-sample and single sample predictions from R package mgmtstp27. The M-values of the two probes *cg12434587* and *cg12981137* are illustrated in (A) for multi-sample predictions and (D) for single sample prediction. The inertia ellipses identify the training dataset and the dots correspond to the location of the new sample prediction. The red and blue colors visualize methylated and unmethylated status, respectively. (B) illustrates the comparison of the *MGMT* score distribution of a new multi-sample dataset (black curve) with the training dataset (M-GBM, green curve, histogram). For single sample prediction, the new sample is indicated by the black vertical line (E). The multi-sample predictions (*MGMT* score and Probabilities) for the dataset TCGA-GBM-27 (black points and lines) associated with their prediction intervals (grey polygons) are shown in (C). The prediction for the sample TCGA-02-0057 from the dataset TCGA-GBM-27 is indicated in (F) associated with the prediction interval. As reference, the green curve and grey polygons correspond to the prediction and confidence intervals for the training dataset (M-GBM).

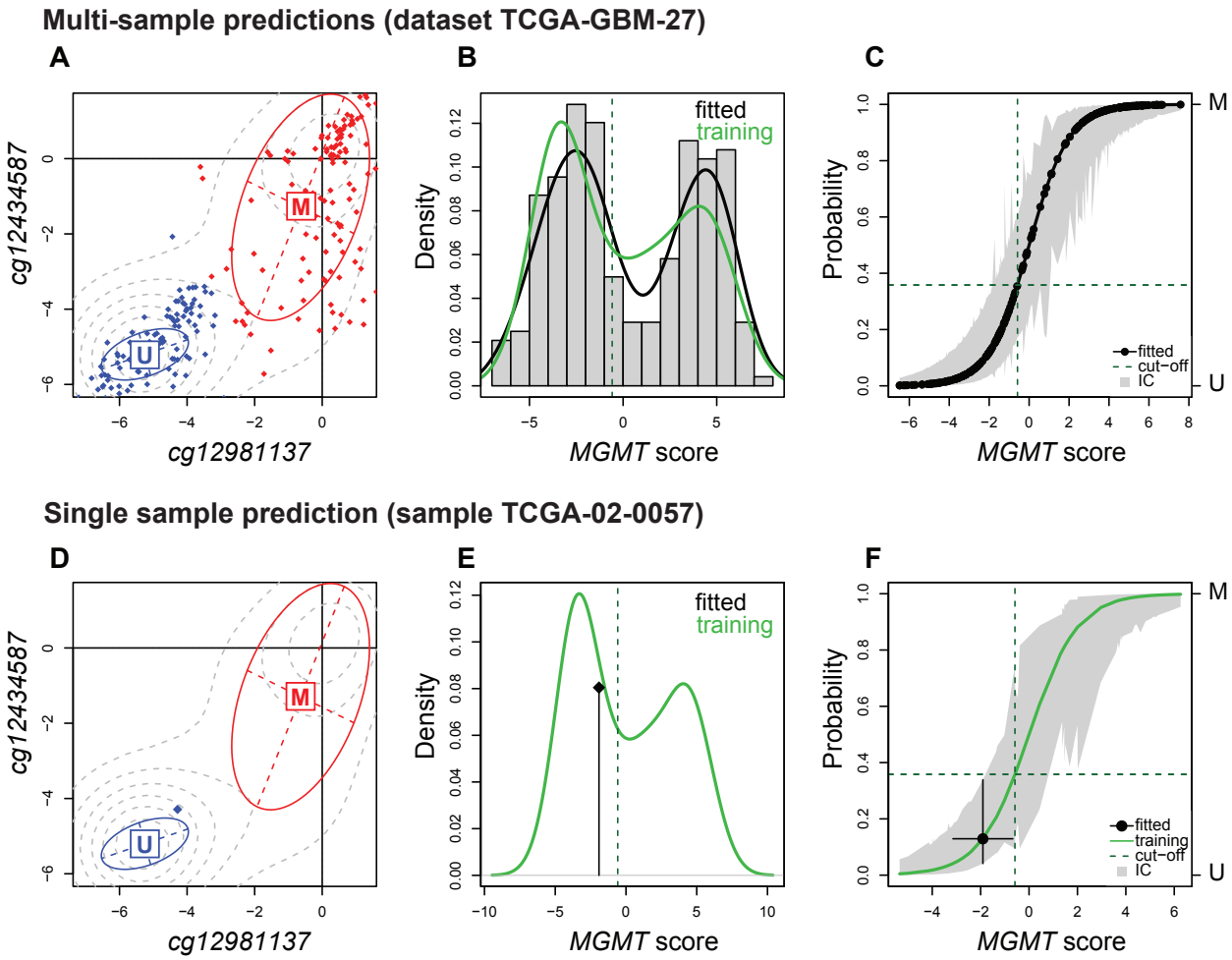

Chromosome 10
131.2646 mb
131.2647 mb
131.2648 mb
131.2649 mb
131.265 mb
131.2651 mb
131.2652 mb
131.2653 mb
131.2654 mb
131.2655 mb
131.2656 mb
131.2657 mb
131.2658 mb
131.2659 m
MGMT
cg12434587
TSS
cg12981137
CpGs
M−GBM
TCGA−GBM−450
TCGA−GBM−27
TCGA−Glioma−II/III
T−Glioma−II/III
1
0.5
0
−0.5
−1
All
CIMP−
CIMP+


grade II
grade III
grade IV
M−GBM
TCGA−GBM−27
TCGA−GBM−450
TCGA−Glioma−II/II
T−Glioma−II/III
VB−Glioma−III
CIMP+
CIMP−
Density
0.0
0.1
0.2
0.3
−10
0
5
MGMT score


TCGA−HNSC
TCGA−LUSC
TCGA−BRCA
TCGA−COAD
Density
0.25
0.20
0.15
0.10
0.05
0.00
−10
−5
0
5
MGMT score


TCGA−Glioma−II/III
T−Glioma−II/III
M−GBM
TCGA−GBM−27
TCGA−GBM−450
MGMT Expression
del.M
del.U
no−del.M
no−del.U
[MGMT cna] x [MGMT status]
28
14
137
16
7
2
34
5
30
25
3
1
89
94
19
10
20
29
9
9


A
DA=0.941
probabilities from original preprocessing
probabilities from illumina preprocessing
B
DA=0.941
probabilities from original preprocessing
probabilities from background correction only
C
DA=0.779
probabilities from original preprocessing
probabilities from quantile normalization
D
DA=0.985
probabilities from original preprocessing
probabilities from SWAN normalization
E
DA=0.941
probabilities from original preprocessing
probabilities from NOOB normalization
y=x
lowess
cut-off


Multi-sample predictions (dataset TCGA-GBM-27)
A
B
C
cg12434587
cg12981137
M
U
Density
fitted
training
MGMT score
Probability
fitted
cut−off
IC
Single sample prediction (sample TCGA-02-0057)
D
E
F

Supplementary Figures ¬ Tables

Sensitivity analysis of the MGMT-STP27 model and impact of genetic/epigenetic context to predict the *MGMT* methylation status in gliomas and other tumors

Pierre Bady[*,†,‡,§], Mauro Delorenzi[§,¶,ǀ], Monika E Hegi[*,†]

[*] Neurosurgery, Lausanne University Hospital, 1011 Lausanne, Switzerland

[†] Neuroscience Research Center, Lausanne University Hospital, 1011 Lausanne, Switzerland

[‡] Department of Education and Research, University of Lausanne, 1011 Lausanne, Switzerland

[§] SIB Bioinformatics Core Facility, Swiss Institute for Bioinformatics, 1005 Lausanne, Switzerland;

[¶] Ludwig Center for Cancer Research, University of Lausanne, 1011 Lausanne, Switzerland,

[ǀ] Department of Oncology, University of Lausanne, 1011 Lausanne, Switzerland;

Running head: Sensitivity analysis MGMT-STP27

Grant support: Swiss National Science Foundation (3100A-138116), the Swiss Bridge Award 2011, and the Swiss Cancer League (KFS-29-02-2012).

Corresponding Author: Monika E. Hegi, Laboratory of Brain Tumor Biology and Genetics, Department of Clinical Neurosciences, Lausanne University Hospital (CHUV-CLE C306), Ch des Boveresses 155, 1066 Epalinges, Switzerland

Phone: +41-21-314-2582, Email: monika.hegi@chuv.ch

Legends Supplementary Figures

Figure S1. Pipeline for computation of *MGMT* classification using the R package mgmtstp27. The R package minfi and methylumi can be used to import and to preprocess raw data. The prediction of the DNA methylations status of *MGMT* promoter requires preprocessed intensities for the signals for unmethylated and methylated as initially proposed for HM-27k in Illumina Genome Studio software in 2009-2011 and originally used in TCGA database. For raw HM-450K data, this operation was performed by the function preprocessRaw from R package minfi. When the raw IDAT format was not available, we assumed an adequate normalization procedure.

Figure S2. Spatial correlation between *MGMT* expression and CpG methylation in the *MGMT* promoter for Non-Glioma Tumors from TCGA. The correlation between expression and DNA methylation for the Infinium HM-450K probes in *MGMT* promoter (genome assemble 37, hg19) is given for TCGA-COAD, TCGA-BRCA, TCGA-HNSC and TCGA-LUSC datasets. The green rectangle corresponds to the CpG island located in the *MGMT* promoter region and the two dark blue rectangles identify the location of the two Inifinium HM-450K/27K probes used in the model MGMT-STP27.

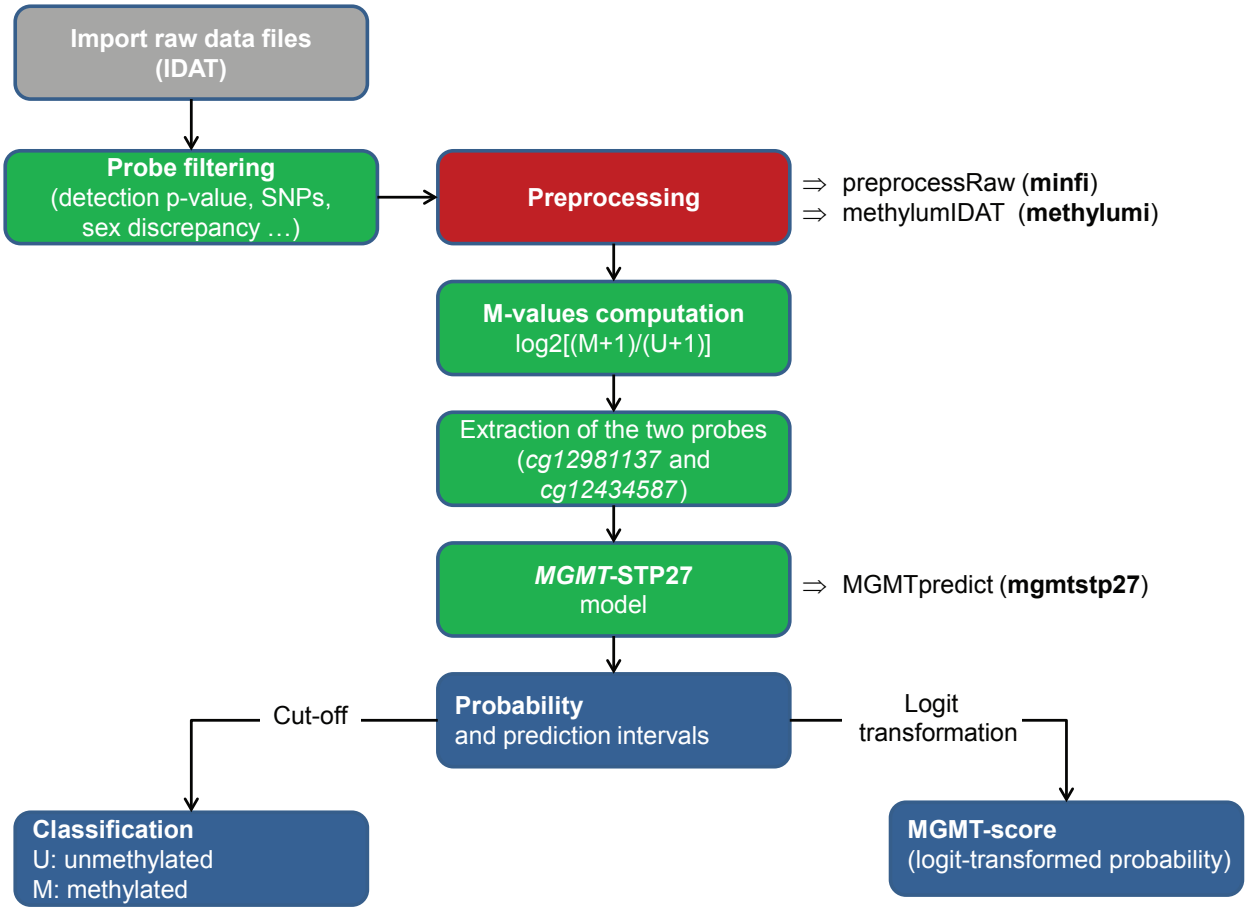
Figure S3. Forest plot of the meta-analysis for the proportion of *MGMT* methylation in colon cancer. The calculation of an overall proportion of *MGMT* methylation from 13 studies (2779 patients). This analysis used logit transformation and inverse variance method. DerSimonian-Laird estimate was used in the random effects model and Clopper-Pearson intervals were given for *MGMT* proportion in each study ('exact' binomial interval).

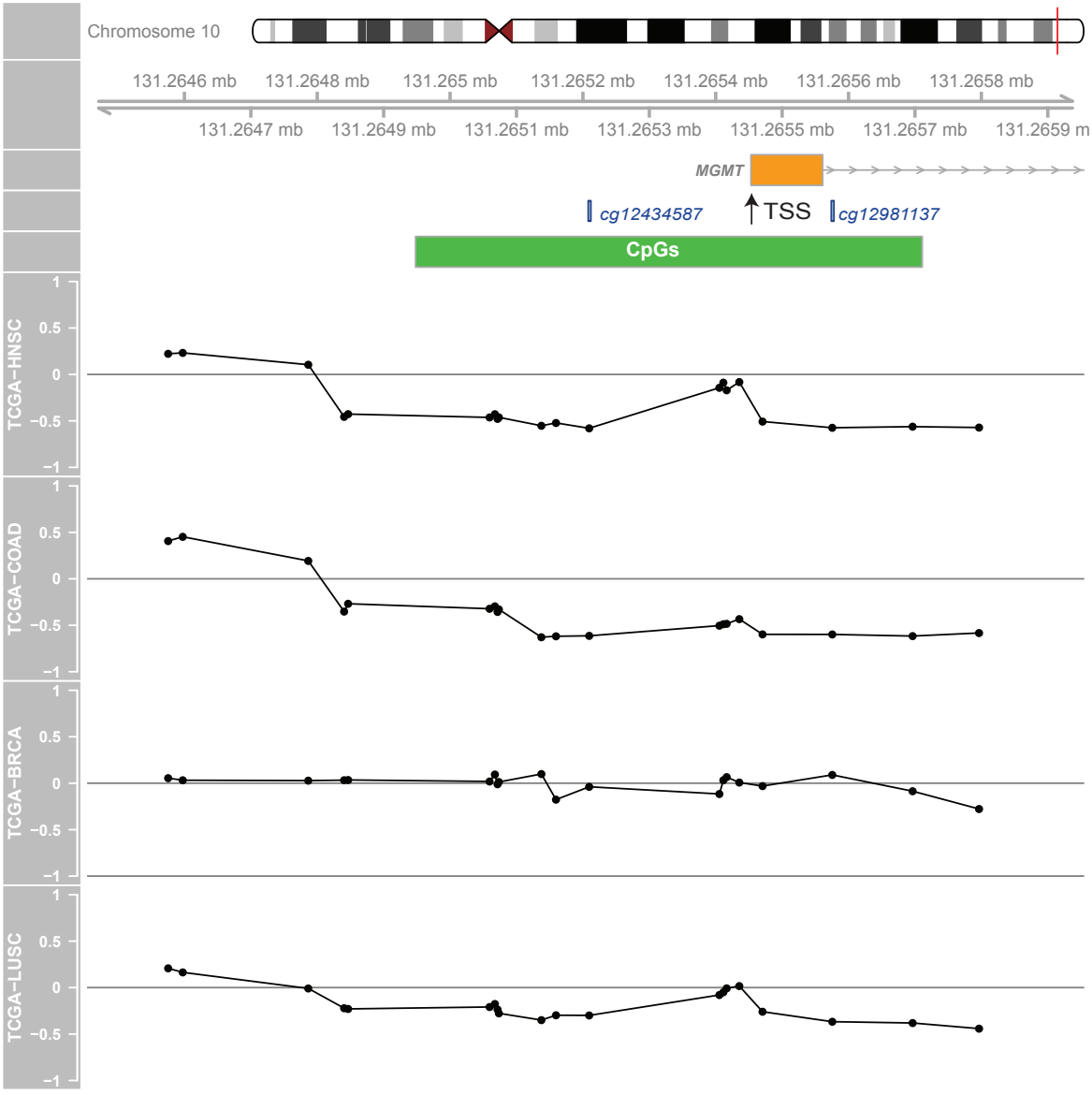
Figure S4. Boxplot representation of *MGMT* expression in function of CNA and *MGMT* methylation status in non-Glioma datasets from TCGA (TCGA-COAD, TCGA-BRCA,TCGA-HNSC and TCGA-LUSC). For each dataset the number of samples for each subpopulation is provided next to the box. Subpopulations with deletions at 10q26.3, del; subpopulations with normal copy number, no-del; *MGMT* methylated, M; *MGMT* unmethylated, U.

Figure S5. Boxplot representation of *MGMT* expression in function of CIMP status and *MGMT* methylation status in glioma grade II to IV. The number of samples for each subpopulation is provided next to the box for each dataset. The combined effect of the two variables CIMP status and *MGMT* methylation status on the expression of *MGMT* was not efficiently testable because the data was strongly unbalanced. Presence of CIMP, CIMP+; absence of CIMP, CIMP-; *MGMT* methylated, M; *MGMT* unmethylated, U.

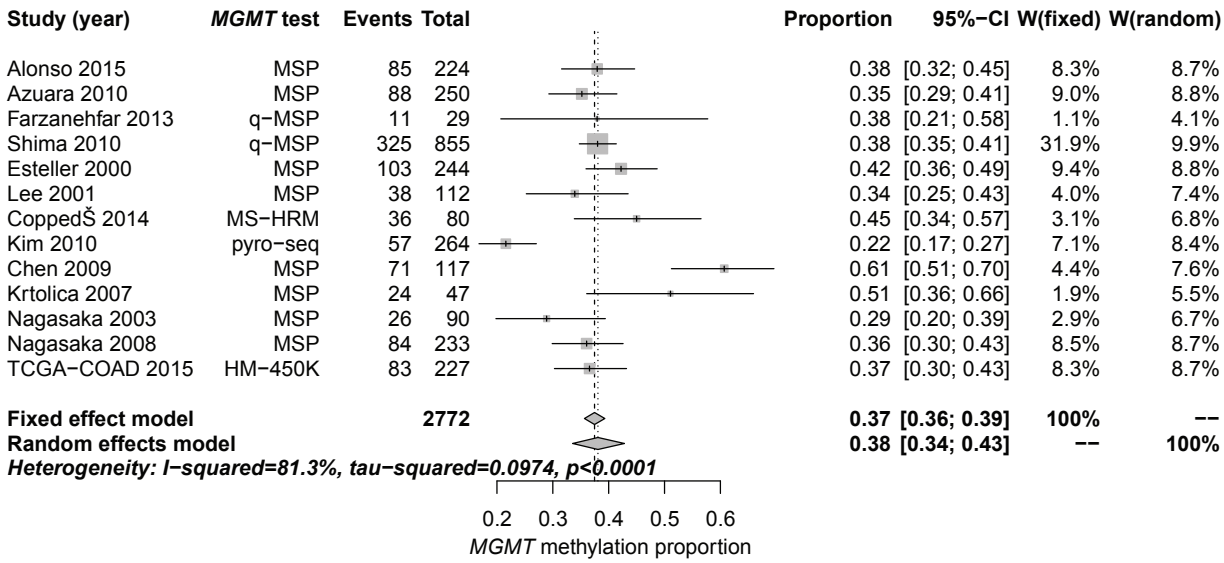
Figure S6. Comparison of *MGMT* score distributions (logit-transformed probability) among FFPE and Frozen Tissues from VB-Glioma-III dataset. The *MGMT* score distributions were represented by histogram for frozen tissue (A, n=51), for FFPE tissue (B, n=59) and for aggregated data (C, n=110). The dotted, dashed and solid red curves correspond to kernel density estimates for frozen tissues, FFPE tissues and all samples. The vertical dashed black line identifies the position of the cut-off used to determinate the *MGMT* promoter state (0.3582). The QQ-plot representation (D) compares the *MGMT* score distributions from Frozen and FFPE data (VB-Glioma-II/III). The distributions were compared by Smirnov-Kolmogorov tests (D=0.187, p-value=0.253). The solid red line corresponds to line of equation y=x.

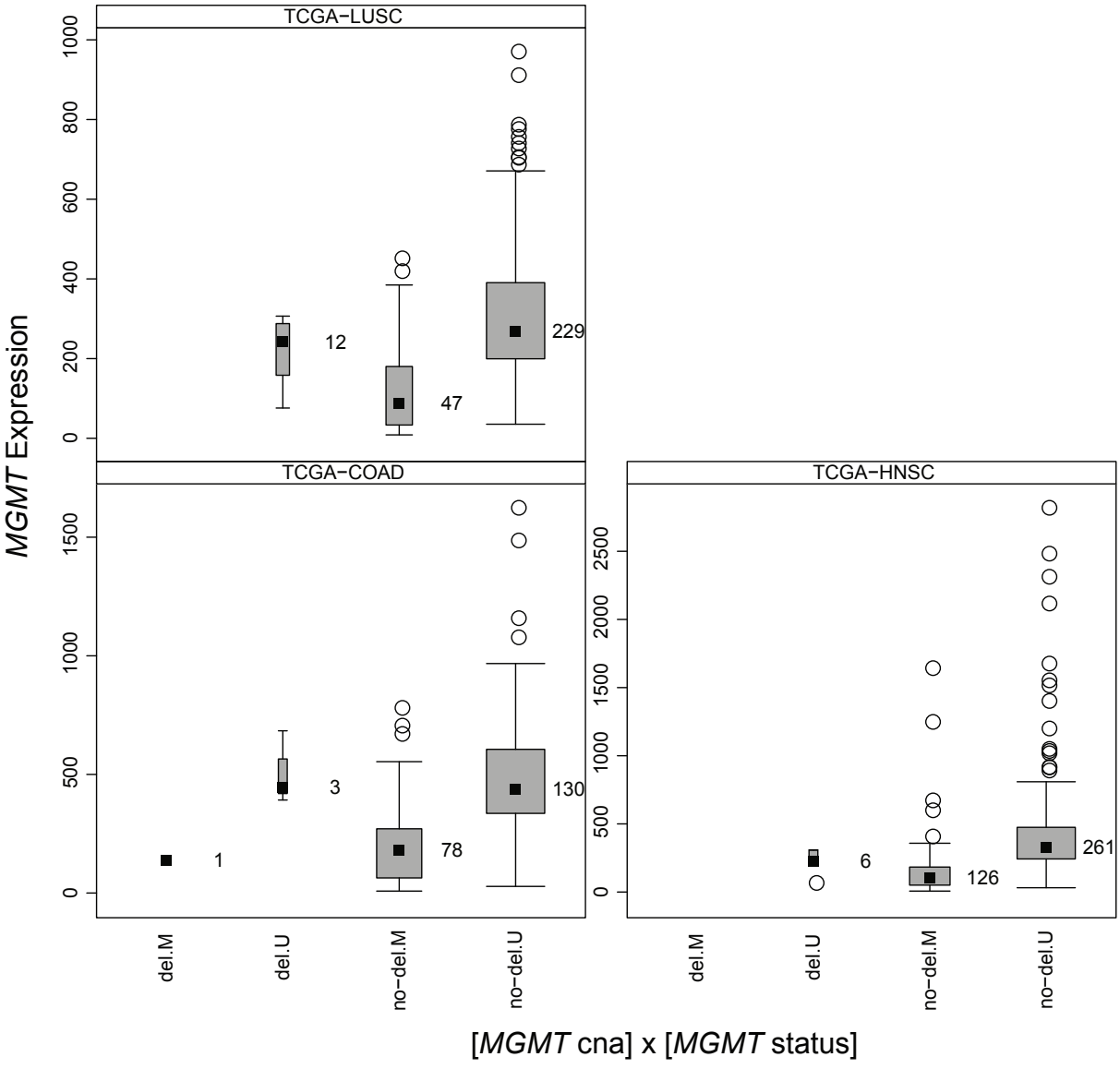
Figure S7. Effect of preprocessing procedures on *MGMT* classification. Paired comparison of the probabilities that *MGMT* promoter was methylated to evaluate the effect of preprocessing procedure for TCGA datasets (TCGA-GBM-450, TCGA-Glioma-II/III). Five preprocessing procedures for the HM-450K platform were compared with the initial procedure used to build the model MGMT-STP27. The outputs from recommended preprocessing were compared with outputs from (A) Illumina-like procedure based on control normalization (a reference sample was used during the normalization step), (B) preprocessing with Illumina-like background correction only, (C) quantile normalization, (D) SWAN normalization and (E) Noob normalization. Each dataset contained exactly the same samples. The predictions from the level 1 (F) and level 2 (G) for HM-27k data from TCGA GBM database were compared with outputs of the originally calculated probabilities [11]. The grey dashed lines identify the original cut-off of 0.3582. The straight, dashed black line corresponds to the equation y=x and the grey line to the loess regression, respectively. The proportions of good classification (diagnostic accuracy, DA) are provided for the original cut-off on each figure.


Import raw data files (IDAT)
Probe filtering (detection p-value, SNPs, sex discrepancy …)
Preprocessing
⇒ preprocessRaw (minfi)
⇒ methylumIDAT (methylumi)
M-values computation log2[(M+1)/(U+1)]
Extraction of the two probes (cg12981137 and cg12434587)
MGMT-STP27 model
⇒ MGMTpredict (mgmtstp27)
Probability and prediction intervals
Cut-off
Logit transformation
Classification U: unmethylated M: methylated
MGMT-score (logit-transformed probability)

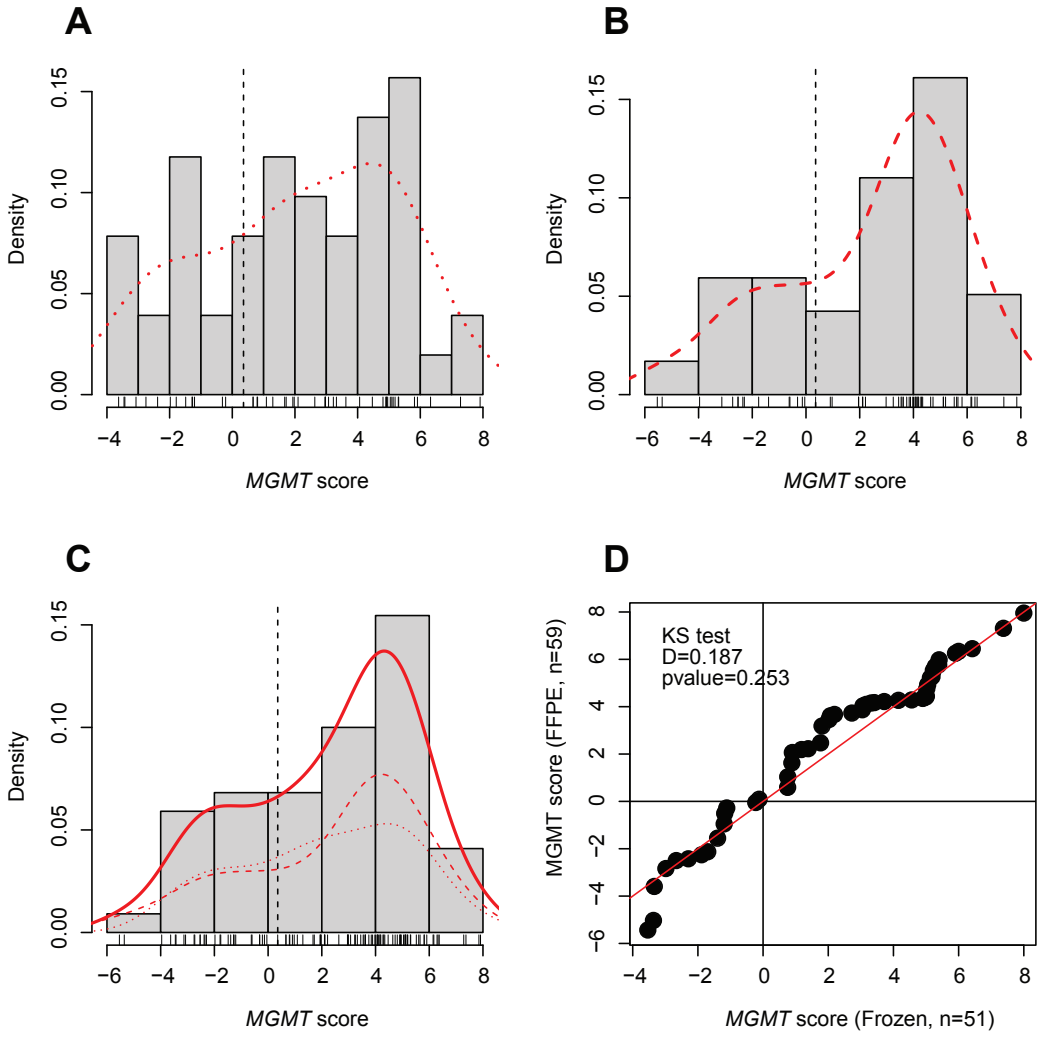

Chromosome 10
131.2646 mb
131.2648 mb
131.265 mb
131.2652 mb
131.2654 mb
131.2656 mb
131.2658 mb
131.2647 mb
131.2649 mb
131.2651 mb
131.2653 mb
131.2655 mb
131.2657 mb
131.2659 m
MGMT
cg12434587
TSS
cg12981137
CpGs
TCGA−HNSC
TCGA−COAD
TCGA−BRCA
TCGA−LUSC
1
0.5
0
−0.5
−1

Meta-analysis for *MGMT* methylation proportion in colon cancer




TCGA-LUSC
TCGA-COAD
TCGA-HNSC
MGMT Expression
0
200
400
600
800
1000
500
1000
1500
2000
2500
12
47
229
1
3
78
130
6
126
261
del.M
del.U
no-del.M
no-del.U
[MGMT cna] x [MGMT status]

TCGA−Glioma−II/III
T−Glioma−III
M−GBM
TCGA−GBM−27
TCGA−GBM−450
MGMT expression
[CIMP status] x [MGMT status]
CIMP−.M
CIMP−.U
CIMP+.M
CIMP+.U
14
22
151
8
5
7
36
29
26
4
94
101
14
3
26
37
3
1


A
B
C
D
Density
0.00
0.05
0.10
0.15
−6
−4
−2
0
2
4
6
8
MGMT score
MGMT score (FFPE, n=59)
MGMT score (Frozen, n=51)
KS test
D=0.187
pvalue=0.253

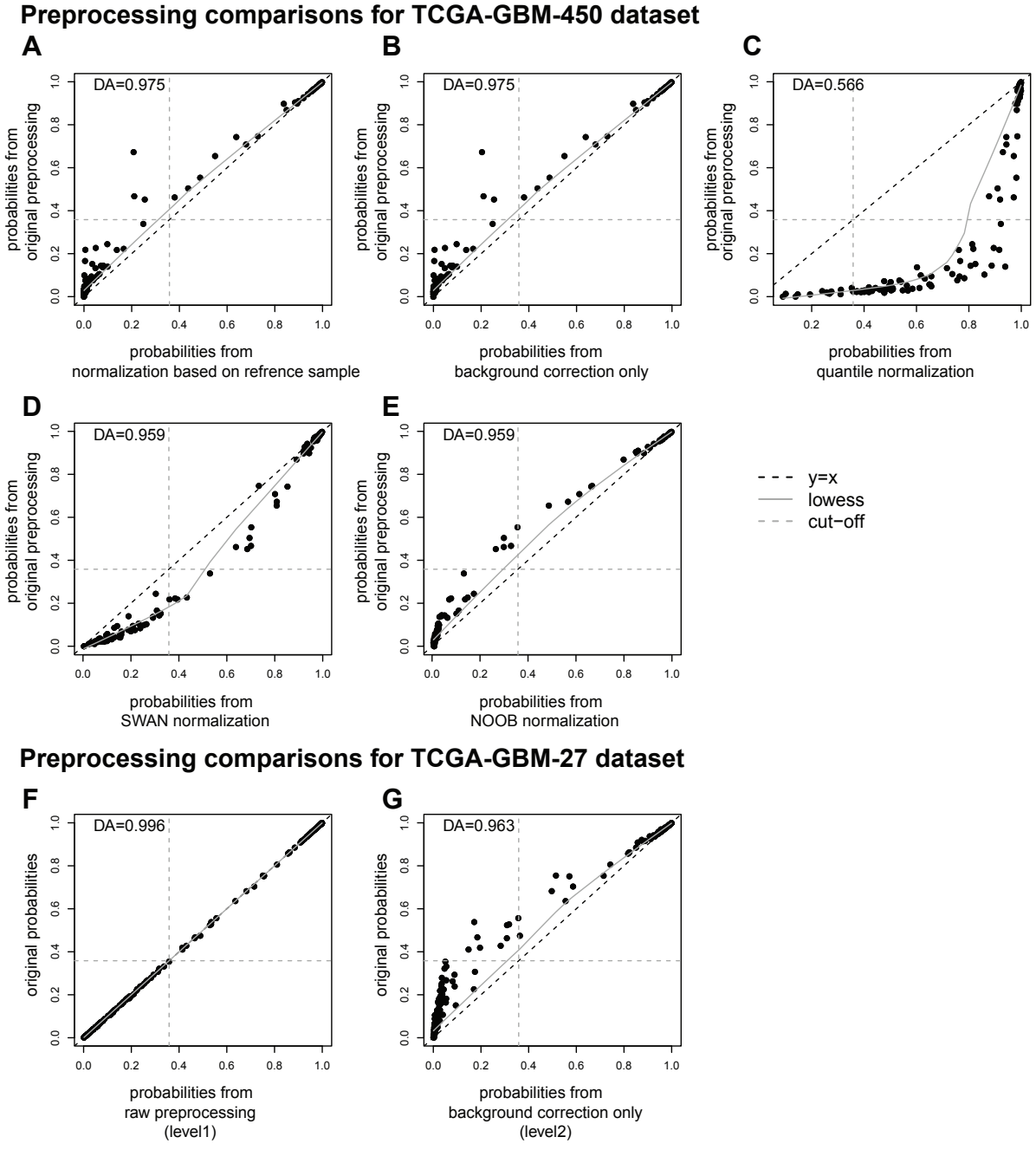

Preprocessing comparisons for TCGA-GBM-450 dataset
A
DA=0.975
probabilities from original preprocessing
probabilities from normalization based on refrence sample
B
DA=0.975
probabilities from original preprocessing
probabilities from background correction only
C
DA=0.566
probabilities from original preprocessing
probabilities from quantile normalization
D
DA=0.959
probabilities from original preprocessing
probabilities from SWAN normalization
E
DA=0.959
probabilities from original preprocessing
probabilities from NOOB normalization
y=x
lowess
cut−off
Preprocessing comparisons for TCGA-GBM-27 dataset
F
DA=0.996
original probabilities
probabilities from raw preprocessing (level1)
G
DA=0.963
original probabilities
probabilities from background correction only (level2)

Table S1. Description of datasets

Dataset	No samples	Trial	DNA methylation platform	†Acc No	Expression platform	†Acc No	Tissue type	References
GLIOMA datasets								
M-GBM	63	yes	HM-450K	GSE60274	Affy U133plus2	GSE7696	Frozen	[21, 20]
TCGA-GBM-27	217	no	HM-27K	TCGA	Affy U133A	TCGA	Frozen	[22, 23, 2]
TCGA-GBM-450	104	no	HM-450K	TCGA	Affy U133A	TCGA	Frozen	[22, 23, 2]
VB-Glioma-III	51	yes	HM-27K	GSE48460			Frozen	[7]
	59	yes	HM-450K	GSE48461			FFPE	[9]
Turcan-Glioma-II/III	71	no	HM-450K	GSE30338	Affy U133plus2	GSE30336	Frozen	[10]
TCGA-Glioma-II/III	197	no	HM-450K	TCGA	RNA-seq (level 3)	TCGA	Frozen	[24]
NON-Glioma datasets								
TCGA-COAD	227	no	HM-450K	TCGA	RNA-seq (level 3)	TCGA	Frozen	TCGA Consortium
TCGA-HNSC	442	no	HM-450K	TCGA	RNA-seq (level 3)	TCGA	Frozen	TCGA Consortium
***TCGA-BRCA**	305	no	HM-450K	TCGA	RNA-seq (level 3)	TCGA	Frozen	TCGA Consortium
TCGA-LUSC	328	no	HM-450K	TCGA	RNA-seq (level 3)	TCGA	Frozen	TCGA Consortium

* Randomly selected

†Accession number: Gene Expression Omnibus, www.ncbi.nlm.nih.gov/**geo**/ ;The Cancer Genome Atlas (TCGA), https://tcga-data.nci.nih.gov/tcga/

Table S2. Description of the main clinical and molecular variables of the Glioma datasets (WHO grade II, III and IV).

Study	Variable	Modality	n	Proportion	*Lower	*Upper
M-GBM (63)	Gender	F	15	0.2381	0.1398	0.3621
		M	48	0.7619	0.6379	0.8602
	MGMT meth	U	28	0.4444	0.3192	0.5751
		M	35	0.5556	0.4249	0.6808
	Grade	II	0	0.0000	0.0000	0.0569
		III	0	0.0000	0.0000	0.0569
		IV	63	1.0000	0.9431	1.0000
	hCIMP	CIMP-	59	0.9365	0.8453	0.9824
		CIMP+	4	0.0635	0.0176	0.1547
	CD-CIMP	none	59	0.9365	0.8453	0.9824
		cimp	3	0.0476	0.0099	0.1329
		cdcimp	1	0.0159	0.0004	0.0853
	MGMT CNA	none	6	0.0952	0.0358	0.1959
		del	57	0.9048	0.8041	0.9642
	Codel 1p19q	cd	1	0.0159	0.0004	0.0853
		n	62	0.9841	0.9147	0.9996
	†Age	middle	42	0.6667	0.5366	0.7805
		old	11	0.1746	0.0905	0.2910
		young	9	0.1429	0.0675	0.2539
TCGA-GBM450 (104)	Gender	F	47	0.4519	0.3541	0.5526
		M	57	0.5481	0.4474	0.6459
	MGMT meth	U	58	0.5577	0.4570	0.6550
		M	46	0.4423	0.3450	0.5430
	Grade	II	0	0.0000	0.0000	0.0348
		III	0	0.0000	0.0000	0.0348
		IV	104	1.0000	0.9652	1.0000
	hCIMP	CIMP-	99	0.9519	0.8914	0.9842
		CIMP+	5	0.0481	0.0158	0.1086
	CD-CIMP	none	99	0.9519	0.8914	0.9842
		cimp	5	0.0481	0.0158	0.1086
		cdcimp	0	0.0000	0.0000	0.0348
	MGMT CNA	none	22	0.2115	0.1376	0.3026
		del	82	0.7885	0.6974	0.8624
	Codel 1p19q	cd	0	0.0000	0.0000	0.0348
		n	104	1.0000	0.9652	1.0000
	†Age	middle	49	0.4712	0.3725	0.5715
		old	52	0.5000	0.4003	0.5997
		young	3	0.0288	0.0060	0.0820
TCGA-GBM27 (217)	Gender	F	83	0.3825	0.3175	0.4507
		M	134	0.6175	0.5493	0.6825
	MGMT meth	U	109	0.5023	0.4338	0.5707
		M	108	0.4977	0.4293	0.5662
	Grade	II	0	0.0000	0.0000	0.0169
		III	0	0.0000	0.0000	0.0169
		IV	217	1.0000	0.9831	1.0000
	hCIMP	CIMP-	200	0.9217	0.8775	0.9537
		CIMP+	17	0.0783	0.0463	0.1225
	CD-CIMP	none	191	0.8802	0.8294	0.9202
		cimp	16	0.0737	0.0427	0.1170
		cdcimp	1	0.0046	0.0001	0.0254
	MGMT CNA	none	30	0.1382	0.0953	0.1914
		del	187	0.8618	0.8086	0.9047
	Codel 1p19q	cd	10	0.0461	0.0223	0.0831
		n	207	0.9539	0.9169	0.9777
	†Age	middle	83	0.3825	0.3175	0.4507
		old	106	0.4885	0.4202	0.5571
		young	28	0.1290	0.0875	0.1811

TCGA-Glioma-II/III (197)	Gender	F	86	0.4365	0.3662	0.5089
		M	111	0.5635	0.4911	0.6338
	MGMT meth	U	31	0.1574	0.1095	0.2159
		M	166	0.8426	0.7841	0.8905
	[‡] Grade	II	90	0.4569	0.3859	0.5291
		III	106	0.5381	0.4658	0.6092
		IV	0	0.0000	0.0000	0.0186
	hCIMP	CIMP-	37	0.1878	0.1358	0.2495
		CIMP+	160	0.8122	0.7505	0.8642
	CD-CIMP	none	37	0.1878	0.1358	0.2495
		cimp	110	0.5584	0.4861	0.6289
		cdcimp	50	0.2538	0.1946	0.3206
	MGMT CNA	none	154	0.7817	0.7175	0.8373
		del	43	0.2183	0.1627	0.2825
	Codel 1p19q	cd	50	0.2538	0.1946	0.3206
		n	147	0.7462	0.6794	0.8054
	[†] Age	middle	76	0.3858	0.3175	0.4576
		old	22	0.1117	0.0713	0.1642
		young	99	0.5025	0.4306	0.5744
VB-Glioma-III (110)	Gender	F	40	0.3636	0.2740	0.4608
		M	70	0.6364	0.5392	0.7260
	MGMT meth	U	25	0.2273	0.1528	0.3170
		M	85	0.7727	0.6830	0.8472
	Grade	II	0	0.0000	0.0000	0.0330
		III	110	1.0000	0.9670	1.0000
		IV	0	0.0000	0.0000	0.0330
	hCIMP	CIMP-	51	0.4636	0.3680	0.5612
		CIMP+	59	0.5364	0.4388	0.6320
	CD-CIMP	none	48	0.4364	0.3420	0.5342
		cimp	26	0.2364	0.1606	0.3268
		cdcimp	33	0.3000	0.2163	0.3948
	MGMT CNA	none	65	0.5909	0.4931	0.6837
		del	45	0.4091	0.3163	0.5069
	Codel 1p19q	cd	36	0.3273	0.2408	0.4233
		n	74	0.6727	0.5767	0.7592
	[†] Age	middle	67	0.6091	0.5114	0.7007
		old	10	0.0909	0.0445	0.1608
		young	33	0.3000	0.2163	0.3948
Turcan-Glioma-II/III (71)	Gender	F	26	0.3662	0.2550	0.4890
		M	45	0.6338	0.5110	0.7450
	MGMT meth	U	14	0.1972	0.1122	0.3086
		M	57	0.8028	0.6914	0.8878
	Grade	II	29	0.4085	0.2932	0.5316
		III	42	0.5915	0.4684	0.7068
		IV	0	0.0000	0.0000	0.0506
	hCIMP	CIMP-	22	0.3099	0.2054	0.4308
		CIMP+	49	0.6901	0.5692	0.7946
	CD-CIMP	none	22	0.3099	0.2054	0.4308
		cimp	24	0.3380	0.2300	0.4601
		cdcimp	25	0.3521	0.2424	0.4746
	MGMT CNA	none	60	0.8451	0.7397	0.9200
		del	11	0.1549	0.0800	0.2603
	Codel 1p19q	cd	25	0.3521	0.2424	0.4746
		n	46	0.6479	0.5254	0.7576
	[†] Age	middle	36	0.5070	0.3856	0.6278
		old	13	0.1831	0.1013	0.2927
		young	22	0.3099	0.2054	0.4308

[*] The proportions were associated with their exact binomial confidence intervals at 95%.
[†] The age was encoded in three categories: young for age ≤ 40 , middle for age > 40 and ≤ 60 and for age > 60.
[‡] one missing value

Table S3. Effects of CIMP and DNA methylation status on expression of *MGMT*.

Dataset (N)	Type	Variables	% (N)	F-statistic	†Pvalue
M-GBM (59)	GBM	*MGMT* meth	55.93 (33)	10.933	0.003
		*CIMP+	6.78 (4)	0.232	0.627
TCGA-GBM-27 (212)	GBM	*MGMT* meth	50.94 (108)	141.068	0.001
		*CIMP+	8.02 (17)	2.154	0.145
TCGA-GBM-450 (67)	GBM	*MGMT* meth	43.28 (29)	8.103	0.008
		*CIMP+	5.97 (4)	0.529	0.46
TCGA-Glioma-II/III (195)	LGG	*MGMT* meth	84.62 (165)	19.114	0.001
		CIMP+	81.54 (159)	0.002	0.97
T-Glioma-II/III (48)	LGG	*MGMT* meth	85.42 (41)	9.374	0.005
		CIMP+	75 (36)	0.002	0.97

* CIMP+ very rare event, unbalanced data!
† simulated p-values estimated by Monte-Carlo procedures (999 permutations)

Table S4. Description of preprocessing and normalization procedures for HM-27K and HM-450K.

Platform	Preprocessing	Descrition	TCGA-GBM Missclassified (%)	M-GBM Missclassified (%)	R Function	R Packages	Reference
HM-27K	Raw	Preprocessing used initially to preprocess HM-27K	1 (0.4)		methylumIDAT	methylumi	[58]
	Noob	backgournd correction based on normal-exponential deconvolution (TCGA level2 in 2014)	9 (3.7)		methylumi.bgcorr	methylumi	[58]
HM-450K	Raw	Preprocessing initially designed for HM-27K	-	-	methylumIDAT preprocessRaw	methylumi minfi	[58, 25]
	Illumina	Control normalization and background correction (subtraction of the fifth percentile from background intensity distribution)	3 (2.5)	4 (5.9)	preprocessIllumina	minfi	[25]
	Background only	background correction based on the subtraction of the fifth percentile from background intensity distribution	3 (2.5)	4 (5.9)	preprocessIllumina	minfi	[25]
	Noob	background correction based on normal-exponential deconvolution with dye-bias correction	5 (4.1)	4 (5.9)	preprocessNoob,	minfi	[25,17]
	Quantile	separate quantile normalization of unmethylated and methylated signals	53 (43.4)	18 (26.7)	preprocessQuantile	minfi	[25]
	SWAN	Subset-quantile Within Array Normalisation for Illumina Infinium HumanMethylation450 BeadChips	5 (4.1)	1 (1.5)	preprocessSWAN	minfi	[16]