

Clinical Epigenetics

DNA methylation signatures for WHO 2016 classification subtypes of diffuse gliomas --Manuscript Draft--

Manuscript Number:	CLEP-D-16-00217
Full Title:	DNA methylation signatures for WHO 2016 classification subtypes of diffuse gliomas
Article Type:	Research
Abstract:	<p>Background:</p> <p>Glioma is the most common of all primary brain tumors with poor prognosis and high mortality. The 2016 World Health Organization classification of the tumors of central nervous system uses molecular parameters in addition to histology to redefine many tumor entities. The new classification scheme divides diffuse gliomas into low grade glioma (LGG) and glioblastoma (GBM) as per histology. LGGs are further divided into IDH wild-type or mutant, which is further classified into either oligodendroglioma that harbours 1p/19q codeletion or diffuse astrocytoma that has an intact 1p/19q loci but enriched for ATRX loss and TP53 mutation. GBMs are divided into IDH wild-type that corresponds to primary or de novo GBMs and IDH mutant that corresponds to secondary or progressive GBMs. To make the WHO 2016 subtypes of diffuse gliomas more robust, we carried out Prediction Analysis of Microarrays (PAM) to develop DNA methylation signatures for these subtypes.</p> <p>Results</p> <p>In this study, we applied PAM on a training set of diffuse gliomas derived from TCGA and identified DNA methylation signatures to classify LGG IDH wild type from LGG IDH mutant, LGG IDH mutant with 1p/19q codeletion from LGG IDH mutant with intact 1p/19q loci and GBM IDH wild type from GBM IDH mutant with an accuracy of 99-100%. The signatures were validated using the test set of diffuse glioma samples derived from TCGA with an accuracy of 96 to 99%. Further, the methylation signature identified a fraction of samples as discordants, which were found to have molecular and clinical features typical of the class as identified by methylation signatures.</p> <p>Conclusions</p> <p>Thus, we identified methylation signatures that classified different subtypes of diffuse glioma accurately and propose that these signatures could complement WHO 2016 classification scheme of diffuse glioma.</p>

[Click here to view linked References](#)

DNA methylation signatures for WHO 2016 classification subtypes of diffuse gliomas

Yashna Paul^a, Baisakhi Mondal^a, Vikas Patil and Kumaravel Somasundaram^{*}

¹Department of Microbiology and Cell Biology, Indian Institute of Science, Bangalore 560012.

^a Equal contribution

^{*} Corresponding author

Tel: +91-80-23607171

Fax: +91-80-23602697

Email: ksomasundaram1@gmail.com, skumar@mcbl.iisc.ernet.in

Running title: DNA methylation signatures for diffuse glioma subtypes

Abstract

Background:

Glioma is the most common of all primary brain tumors with poor prognosis and high mortality. The 2016 World Health Organization classification of the tumors of central nervous system uses molecular parameters in addition to histology to redefine many tumor entities. The new classification scheme divides diffuse gliomas into low grade glioma (LGG) and glioblastoma (GBM) as per histology. LGGs are further divided into IDH wild-type or mutant, which is further classified into either oligodendroglioma that harbours 1p/19q codeletion or diffuse astrocytoma that has an intact 1p/19q loci but enriched for ATRX loss and TP53 mutation. GBMs are divided into IDH wild-type that corresponds to primary or *de novo* GBMs and IDH mutant that corresponds to secondary or progressive GBMs. To make the WHO 2016 subtypes of diffuse gliomas more robust, we carried out Prediction Analysis of Microarrays (PAM) to develop DNA methylation signatures for these subtypes.

Results

In this study, we applied PAM on a training set of diffuse gliomas derived from TCGA and identified DNA methylation signatures to classify LGG IDH wild type from LGG IDH mutant, LGG IDH mutant with 1p/19q codeletion from LGG IDH mutant with intact 1p/19q loci and GBM IDH wild type from GBM IDH mutant with an accuracy of 99-100%. The signatures were validated using the test set of diffuse glioma samples derived from TCGA with an accuracy of 96 to 99%. Further, the methylation signature identified a fraction of samples as discordants, which were found to have molecular and clinical features typical of the class as identified by methylation signatures.

Conclusions

Thus, we identified methylation signatures that classified different subtypes of diffuse glioma accurately and propose that these signatures could complement WHO 2016 classification scheme of diffuse glioma.

Keywords: Glioma, DNA methylation classification signature, IDH1/IDH2 mutation, WHO 2016, PAM, PCA.

Background

The neoplasia of non-neuronal glial cells in the brain is referred to as glioma and is the most common type of primary central nervous system (CNS) tumors [1]. The different histological subtypes of glioma are as follows: astrocytoma being the most common, accounting for 70% of all cases, while oligodendroglioma comprises 9% which include classic oligodendrogliomas as well as mixed oligoastrocytomas, and ependymoma comprises 6% [2].

Over the past decades, classification of brain tumors were based on the histopathological and microscopic features in hematoxylin and eosin stained sections, like cell type, level of differentiation, identifying necrotic lesions and presence of lineage-specific markers. According to the WHO 2007 based classification, grade II/diffused astrocytoma (DA) was described as low grade while high grade glioma comprised of grade III/anaplastic astrocytoma (AA) and grade IV/glioblastoma (GBM) [3]. The vast majority of GBM develops *de novo* in elderly patients with no prior clinical or histological evidence and are referred to as primary GBM. Secondary GBM progresses through low grade diffuse astrocytoma or anaplastic astrocytoma and are manifested in younger patients. Several studies have shown that glioma is highly heterogeneous which indicates that tumors of same grade have diverse genetic and epigenetic molecular aberrations [4-9]. With the invent of new technologies, many high throughput studies have reported different molecular signatures based on Glioma CpG-island methylator phenotype (GCIMP), expression-based studies for mRNA, miRNA and lncRNA in GBM [10-13]. One of the most exciting and clinically relevant observations was the discovery that a high percentage of grade II/III and grade IV secondary glioblastoma harbour mutations in the genes isocitrate dehydrogenase 1 and 2 [2]. Growing data indicate that these mutations play a causal role in gliomagenesis, have a major impact on tumor biology, and also have clinical and prognostic importance [2].

Nearly 12% of GBM patients have been identified to have point mutation in codon 132 (R132H) of the isocitrate dehydrogenase 1 (IDH1) gene located in the chromosome locus 2q33 [14]. IDH1 codes for a cytosolic protein that controls oxidative cellular damage [14, 15]. Several studies showed that the IDH1 mutation is inversely associated with grade in diffuse glial tumors, affecting 71% of grade II, 64% of grade III, and 6% of primary glioblastomas [14]. Interestingly, IDH mutation is found to be present in the secondary glioblastoma (76%) probably because these tumours have been derived from the lower grade gliomas [16]. IDH1 is an enzyme and it catalyses the oxidative decarboxylation of isocitrate to produce α -ketoglutarate (α -KG) [17].

IDH mutation has been shown to be associated with alterations in the methylome thus being sufficient to establish glioma hypermethylator phenotype [18]. At present, 2016 WHO CNS tumor classification has included both molecular markers along with histological features to identify and classify different subtypes of diffuse glioma which includes include the WHO grade II and grade III astrocytic tumors, the grade II and III oligodendrogliomas and the grade IV glioblastomas. The low grade gliomas (LGGs), which include the WHO grade II and grade III astrocytic tumors, the grade II and III oligodendrogliomas, are classified based on IDH mutation status. The LGG IDH mutant subtype is further classified based on the codeletion of 1p/19q where LGG IDH mutant patients harbouring 1p/19q codeletion is termed as oligodendrogliomas (ODG) while LGG IDH mutant patients having intact 1p/19q loci are termed as diffuse astrocytoma which may be enriched in TP53 mutation/ATRX loss. The other axis is the glioblastoma (GBM) which, similar to LGG, is further classified in to IDH WT and mutant.

In the present study, we investigated the altered methylation pattern amongst the different subtypes of diffuse gliomas as per WHO 2016 CNS tumour classification [19] and derived methylation-based classification signature for distinguishing different subtypes. Our

study sets up the premise of using methylation signature in combination to the WHO 2016 classification system with a higher precision of classification of the diffuse glioma patients, thereby helping better diagnosis and appropriate treatment therapy.

Result

The overall work flow of methylation-based signatures to distinguish diffuse glioma subtypes of WHO 2016 classification

To develop methylation-based signatures to distinguish diffuse glioma subtypes as per 2016 WHO CNS tumour classification (**Figure 1**), we subjected the 450K DNA methylation data of TCGA diffuse glioma samples (<https://tcga-data.nci.nih.gov/tcga/>) to various statistical tools and validation steps (**Figure 2**). The methylation signatures were developed to distinguish LGG IDH mutant from LGG IDH WT, LGG IDH mutant with 1p/19q codeletion (Oligodendroglioma) from LGG IDH mutant with intact 1p/19q loci (Diffuse astrocytoma) and GBM IDH mutant (Progressive GBM) from GBM IDH WT (de novo GBM). The TCGA samples were classified into these groups as per WHO 2016 classification scheme (**Figure 1**). For methylation signature development, to begin with, we performed a Wilcoxon-rank sum test between different diffuse glioma subtypes to identify a list of significantly differentially methylated CpG probes, which were further subjected to a differential β value ($\Delta\beta$) of 0.4 between groups. The TCGA samples were then divided randomly into two equal groups as training and test sets (**Additional File 2: Table S1**). The training set was subjected to Prediction Analysis of Microarrays (PAM) to identify the methylation signatures containing minimum number of CpGs with least error. The robustness of the identified signatures was internally cross validated within training set using Support Vector Machine (SVM) and subset validation. The signatures were further applied on the test set for the final validation. We also used Principle Component Analysis (PCA) to test the

ability of methylation signatures to separate the two compared groups into two distinct clusters. Additionally, 10-fold cross validation by PAM was carried out to identify the discordant samples, which were then subjected to further analysis to find out true nature of these samples.

14-CpG methylation signature to distinguish LGG IDH mutant from LGG IDH wild type (WT): Identification and validation

PAM analysis of differentially methylated CpGs (**Additional File 2: Table S2**; n=9554) in the training set (**Additional File 2: Table S1**; 217 LGG IDH mutant and 49 LGG IDH wild type) identified a set of 14 CpGs to distinguish IDH mutant from IDH WT in LGG at a threshold value of 18.9 with least error (**Figure 3A, Additional File 1: Figure S1A**). The robustness of this probe set was tested by internal cross validation using SVM, which gave a classification accuracy of 100% and subset validation with an accuracy of 100% (**Additional File 1: Figure S2A and B respectively; see materials and methods for more details**). The CpG probes of the signature were found to be hypermethylated in IDH mutant LGGs compared to IDH WT LGGs (**Figure 3B and Table 1**). Further, upon subjecting the 14 CpG probes to PCA, the two principle components were able to form two distinct clusters for IDH mutant and IDH WT LGGs (**Figure 3C**). Prediction accuracy estimation by 10-fold cross-validation by PAM, showed that out of 217 LGG IDH mutant samples, the 14 CpG probe methylation signature predicted all the samples accurately with no error (**Figure 3D**). Similarly, among 49 LGG IDH WT samples, all were rightly predicted to be LGG with WT IDH samples based on the 14 CpG probes methylation signature (**Figure 3D**). Thus, the 14 CpG DNA methylation signature was able to discriminate LGG IDH1 mutant from LGG IDH1 WT with an overall classification accuracy of 100%. The sensitivity and specificity of the signature for IDH1 mutant and WT in LGG is 100% (**Table 2**).

Next, we validated the strength of 14 CpG methylation signature using the test set (Additional File 2: Table S1; 216 LGG IDH mutant and 48 LGG IDH wild type). The 14 discriminatory probes were observed to be differentially methylated between LGG IDH mutant and LGG IDH WT in the test set also (Figure 4A and Table 1). The PCA demonstrated that the probes were able to distinguish IDH mutant from the WT groups as two distinct clusters (Figure 4B). Prediction accuracy estimation by 10-fold cross-validation using PAM, showed that out of 216 IDH mutant LGG samples, the 14 CpG probe methylation signature predicted 215 IDH mutant LGG samples accurately with an error rate of 0.004 (Figure 4C). Among 48 IDH WT LGG samples, all 48 samples were accurately predicted by the signature (Figure 4C). Thus, the 14 CpG methylation signature was able to discriminate between IDH mutant and WT LGG samples with an overall diagnostic accuracy of 99.62% in the test set. The sensitivity of the signature for IDH mutant LGG is 99.53% while for IDH WT LGG is 100% and the specificity for IDH mutant is 100% whereas for those of the IDH WT it is 99.53% (Table 2). The 14 CpG methylation signature, as identified in the training set and validated in the test set, was also used to classify the entire set of TCGA LGG (97 IDH WT samples and 433 IDH mutant samples). We found that the 14 discriminatory probes distinguished two groups (Additional File 1: Figure S3A, B and C) with an overall accuracy of 99.81% (Table 2).

14 CpG probe methylation signature to classify oligodendrogliomas (ODG) and diffuse astrocytoma (DA): Identification and validation

PAM analysis of differentially methylated CpGs (Additional File 2: Table S3; n=2817) on the training set (Additional File 2: Table S1; 86 oligodendrogliomas and 131 diffused astrocytoma) identified a set of 14 CpGs to distinguish IDH mutant with 1p/19q codeletion (designated as oligodendroglioma) from LGG IDH mutant with intact 1p/19q loci (designated as diffuse astrocytoma) at a threshold value of 9.491 with minimal error (Figure

5A, Additional File 1: Figure S1B). The robustness of this probe set was tested by internal cross validation using SVM, which gave a classification accuracy of 97.67 to 100 % and subset validation with an accuracy of 99 to 100% (**Additional File 1: Figure S2C and D respectively; see materials and methods for more detail**). The CpG probes that correspond to this signature were found to be hypermethylated in oligodendroglioma (LGG IDH mutant with codeletion) compared to diffuse astrocytoma (LGG IDH mutant with intact 1p/19q loci) (**Figure 5B and Table 3**). Further, upon subjecting the 14 CpG probes to PCA, the two principle components were able to separate these two groups into two distinct clusters (**Figure 5C**). Prediction accuracy estimation by 10-fold cross-validation using PAM, showed that out of 86 oligodendroglioma samples, the 14-CpG probe methylation signature predicted all the samples accurately with no error (**Figure 5D**). Similarly, among 131 diffuse astrocytoma samples, 129 sample were accurately predicted to be diffuse astrocytoma based on the 14 CpG probes methylation signature with an error rate of 0.0153 (**Figure 5D**). Thus, the 14 CpG DNA methylation signature was able to discriminate oligodendroglioma (LGG IDH mutant with codeletion) from diffuse astrocytoma (LGG IDH mutant with intact 1p/19q loci) with an overall diagnostic accuracy of 99.07 %. The sensitivity of the signature for oligodendroglioma is 100 % while for diffuse astrocytoma is 98.47 % and the specificity for oligodendroglioma is 98.47 % whereas for those of the diffuse astrocytomas is 100 % (**Table 2**).

Next, we validated the strength of 14 CpG methylation signature using the test set (**Additional File 2: Table S1; 86 oligodendrogliomas and 130 diffused astrocytoma**). The 14 discriminatory probes were observed to be differentially methylated between oligodendrogliomas and diffused astrocytoma similar to as seen in the training set (**Figure 6A and Table 3**). The PCA demonstrated that the probes were able to distinguish oligodendrogliomas from diffused astrocytoma as two distinct clusters (**Figure 6B**).

Prediction accuracy estimation by 10-fold cross-validation using PAM, showed that out of 86 oligodendrogliomas, the 14-CpG probe methylation signature predicted 85 oligodendrogliomas samples accurately with an error rate of 0.0117 (**Figure 6C**). Similarly, among 130 diffused astrocytoma samples, 123 samples were accurately predicted by the signature with an error rate of 0.0539 (**Figure 6C**). Thus, the 14 CpG methylation signature was able to discriminate between oligodendrogliomas and diffused astrocytoma samples with an overall diagnostic accuracy of 96.29 % in the test set. The sensitivity of the signature for oligodendrogliomas is 98.83% while for diffused astrocytoma, it is 94.61 % and the specificity for oligodendrogliomas is 94.61 % whereas for diffused astrocytoma, it is 98.83 % (**Table 2**). The 14 CpG methylation signature, as identified in the training set and validated in the test set, was also used to classify the entire TCGA LGG IDH mutant samples into oligodendroglioma and diffuse astrocytoma samples. We found that the 14 discriminatory probes behaved similar in the classification (**Additional File 1: Figure S4A, B and C**) with an overall accuracy of 97.69 % (**Table 2**).

13-CpG probe methylation based signature to classify IDH mutant from wild type (WT) in Glioblastoma (GBM): Identification and validation

PAM analysis of differentially methylated CpGs (**Additional File 2: Table S4**; n=259) in the training set (**Additional File 2: Table S1**; 4 GBM IDH mutant and 59 GBM IDH1 wild type) identified a set of 13 CpGs to distinguish IDH mutant from IDH WT in GBM at a threshold value of 2.694 with no error (**Figure 7A, Additional File 1: Figure S1C**). The robustness of this probe set was tested by internal cross validation using SVM, which gave a classification accuracy of 100% and subset validation with an accuracy of 100% (**Additional File 1: Figure S2E and F respectively; see materials and methods for more details**). The CpG probes of the signature were found to be hypermethylated in IDH mutant GBMs compared to IDH WT GBMs (**Figure 7B and Table 4**). Further, upon

subjecting the 13 CpG probes to PCA, the two principle components were able to form two distinct clusters for IDH mutant and IDH WT GBMs (**Figure 7C**). Prediction accuracy estimation by 10-fold cross-validation using PAM, showed that out of 4 GBM IDH mutant samples, the 13 CpG probe methylation signature predicted all the samples accurately with no error (**Figure 7D**). Similarly, among 59 GBM IDH wild type samples, all were rightly predicted to be GBM with WT IDH samples based on the 13 CpG probes methylation signature (**Figure 7D**). Thus, the 13 CpG DNA methylation signature was able to discriminate GBM IDH mutant from GBM IDH WT with an overall classification accuracy of 100%. The sensitivity and specificity of the signature for IDH1 mutant and WT in GBM is 100% (**Table 2**).

Next, we validated the strength of 13 CpG methylation signature using the test set (**Additional File 2: Table S1**; 3 GBM IDH mutant and 58 GBM IDH wild type). The 13 discriminatory probes were observed to be differentially methylated between GBM IDH mutant and GBM IDH WT in the test set also (**Figure 8A and Table 4**). The PCA demonstrated that the probes were able to distinguish IDH mutant from the WT group as two distinct clusters (**Figure 8B**). Prediction accuracy estimation by 10-fold cross-validation using PAM, showed that out of 3 GBM IDH mutant samples, the 13 CpG probe methylation signature predicted 3 IDH mutant GBM samples accurately with no error rate (**Figure 8C**). Among 58 IDH WT GBM samples, 57 samples were accurately predicted by the signature with an error rate of 0.0173 (**Figure 8C**). Thus, the 13 CpG methylation signature was able to discriminate between IDH mutant and WT GBM samples with an overall diagnostic accuracy of 98.36% in the test set. The sensitivity of the signature for IDH mutant GBM is 100 % while for IDH WT GBM is 98.27 % and the specificity for IDH mutant is 98.27 % whereas for those of the IDH WT it is 100 % (**Table 2**). The 13 CpG methylation signature, as identified in the training set and validated in the test set, was also used to classify the entire

set of TCGA GBM set (117 IDH WT samples and 7 IDH mutant samples). We found that the 13 discriminatory probes distinguished two groups (**Additional File 1: Figure S5A, B and C**) with an overall accuracy of 99.19% (**Table 2**).

Molecular analysis of discordant samples

While the DNA methylation signatures were able to distinguish different diffuse glioma subtypes, it also identified a fraction of samples as discordants. It is of our interest to find out the accurate molecular nature of these samples in order to assess the true nature of them. In the classification of LGG IDH mutant from IDH WT, the 14 CpG signature identified one IDH mutant LGG sample in the test set as a discordant. We carried out a careful assessment of the molecular markers of this sample using c-Bioportal (<http://www.cbioportal.org/>) from the TCGA dataset. For this purpose, we analysed TP53 mutation, ATRX loss and 1p/19q codeletion status of all the samples (**Additional File 2: Table S5, Table S6 and Table S7**). As per 2016 WHO CNS tumor classification, all LGG IDH mutant samples that have 1p/19q codeletion are designated as oligodendroglioma and those with intact 1p/19q loci and enriched for TP53 mutation/ATRX loss are designated as diffuse astrocytoma. The LGG IDH mutant discordant sample had intact 1p/19q, WT TP53 and ATRX genes indicating that this sample is not an oligodendroglioma. The presence of WT TP53 and ATRX genes raises the possibility of it not being a diffuse astrocytoma. Additional analysis revealed that the discordant sample is indeed carrying WT IDH1 as per DNA sequencing even though IDH1 antibody based scoring classified it as IDH1 mutant. Therefore it appears that IDH1 mutation scoring by IHC could be an error as evidenced by DNA sequencing and that the 14 CpG methylation signature is able to classify the LGGs more accurately.

In the classification of LGG oligodendroglioma from LGG diffuse astrocytoma, 14 CpG probe methylation signature identified ten samples as discordants which did not match the WHO 2016 tumour grading. In order to understand the true status of the discordant samples, we analysed the clinical information and molecular markers using c-Bioportal (<http://www.cbioportal.org/>) from the TCGA dataset. For this purpose, we analysed TP53 mutation, ATRX mutation and 1p/19q codeletion status in DA, ODG and discordant samples of LGG (**Additional File 2: Table S5, Table S6 and Table S7**). Based on the WHO 2016 CNS tumor classification, IDH mutant LGGs having intact 1p/19q with an enrichment of TP53 mutation and ATRX loss are classified as diffuse astrocytoma. IDH mutant LGG samples with 1p/19q codeletion are classified as oligodendroglioma. The analysis of discordant samples for the molecular markers and histological features revealed some interesting findings. While the single ODG discordant sample had 1p/19q codeletion and WT TP53/ATRX genes, this sample was identified as oligoastrocytoma as per histology. Among nine DA discordant samples, while all of them had intact 1p/19q loci, a majority of them were found to have WT TP53/ATRX genes.

In the classification of GBM IDH mutant from IDH WT, the 13 CpG probe methylation signature identified one GBM IDH WT sample as a discordant. In order to understand the true nature of the discordant sample, we analysed the clinical information and molecular markers using c-Bioportal (<http://www.cbioportal.org/>) from the TCGA dataset (**Additional File 2: Table S5, Table S7 and Table S8**). The discordant GBM IDH WT sample had WT IDH gene as per both immunohistochemical staining and DNA sequencing. However, this sample had no amplification of EGFR locus with an intact PTEN gene, unlike what is expected for a IDH1 WT GBM sample.

Discussion

Glioma is the most common and highly malignant primary brain tumour. The 2007 WHO classification of the glioma tumors were majorly based on microscopic appearance of cell type and histopathological markers largely segregating into three subtypes such as astrocytoma, oligodendroglioma and oligoastrocytoma (mixed) [3]. With the advent of the high-throughput technologies, comprehensive understanding of the heterogeneous genetic and epigenetic landscape of both glioblastoma and the low grades became vibrant [20, 21]. The histopathological grading of glioma tumours could be subjected to inter-observer variation which would lead to misclassification with a potential possibility of not providing right kind of treatment [22]. To combat this short-coming, several groups including work from our laboratory, carried out extensive studies and have identified several prognostic markers and molecular signatures based on mRNA, miRNA and DNA methylation that would aid in better classification and identifying best choice of therapy [10-13, 15, 23-26].

The meeting by the International Society of Neuropathology held in Haarlem, Netherland, established guidelines for how to incorporate molecular findings into brain tumor diagnosis thereby setting the platform for a major revision of the 2007 CNS WHO classification [27]. The current updated version is summarized in the 2016 CNS WHO classifications [19]. According to this new integrated classification, glioma is broadly sub-classified as low grade glioma (Astrocytoma, Oligodendroglioma, Oligoastrocytoma) and glioblastoma (GBM) based on histology. The low grade glioma is then further sub-divided into two groups based on mutation status of IDH- IDH mutant and IDH wild type. The low grade glioma samples having IDH1 mutation along with 1p/19q co-deletion were termed as oligodendroglioma (ODG) while those with intact 1p/19q loci were classified as diffuse astrocytoma (DA). The latter group is found to be enriched for TP53 mutation and ATRX2

loss, but these features are not essential for diagnosis. Similarly, GBM were sub-divided into IDH wild type that correspond to primary GBM (*de novo*) and IDH mutant that corresponds to secondary GBM (progressive) based on IDH mutation status. These two groups of GBM also differ from each other on various key characteristics like median age at diagnosis, mean length of clinical history, median overall survival, location of the tumor, extent of necrosis, TERT promoter mutation, TP53 mutation, ATRX mutation, PTEN mutation and EGFR amplification status. Further, it is well established that IDH1 mutation brings about a hypermethylation phenotype altering the methylome architecture of the cells [13, 28]. In this study, using TCGA 450K DNA methylation data, we developed methylation signatures that could distinguish different classes of diffuse glioma with a high accuracy.

Infinium HumanMethylation450K BeadChip array data for astrocytoma (grade II, III and IV/GBM), oligodendroglioma and oligoastrocytoma tumor samples from TCGA dataset was used in this study. Towards classifying LGG IDH mutant from LGG IDH wild type samples, PAM identified a 14 CpG probe methylation signature that could distinguish these two groups with a classification accuracy in the training set and test set 100% and 99.62% respectively. Interestingly, it was observed that the lone LGG IDH mutant discordant was found to carry no mutation in IDH1 and IDH2 genes as per DNA sequencing, while the sample was classified originally as IDH1 mutant based on immunohistochemistry using mutant IDH1-specific antibody. The 14 CpG probe methylation signature comprised of genes coding for KCNB1 (Potassium Voltage-Gated Channel Subfamily B Member 1), GNAO1 (G protein subunit alpha O1), FGFR1 (Fibroblast Growth Factor Receptor Like-1), TPPP3 (Tubulin Polymerization Promoting Protein Family Member 3), MMP23A (Matrix Metalloproteinase 23A Pseudogene), UCP2 (Uncoupling Protein 2), RHBDF2 (Rhomoid 5 Homolog 2 in Drosophila), GPR62 (G Protein-Coupled Receptor 62) and RAPGEFL1 (Rap Guanine Nucleotide Exchange Factor Like 1).

Next, we identified a unique set of 14 CpG probe methylation signature that could classify LGG IDH mutant samples into diffuse astrocytoma (DA: samples with intact 1p/19q) and oligodendroglioma (ODG: samples with 1p/19q codeletion). The classification accuracy in the training set and test set were 99.07% and 96.29% respectively. Surprisingly, there were ten discordant samples with one ODG and nine DA discordants. While the lone ODG discordant carried 1p/19q codeletion, it was classified histologically as an oligoastrocytoma. Similarly, while all nine DA discordants carried intact 1p/19q loci, majority of them had wild type TP53/ATRX genes. These 14 CpG probe methylation signature comprised of genes coding for CD300LB (CD300 Molecule Like Family Member B), FLJ37543, FGFR2 (Fibroblast Growth Factor Receptor 2), TCF7L1 (Transcription Factor 7 Like 1), PLCG1 (Phospholipase C Gamma 1), PTPRN2 (Protein Tyrosine Phosphatase Receptor Type N2), PRKAG2 (Protein Kinase AMP-activated Non-catalytic Subunit Gamma 2), MAPKAP1 (Mitogen-Activated Protein Kinase Associated Protein 1) and GPR156 (G-Protein coupled Receptor 156).

Glioblastoma can arise *de novo* which is generally associated with IDH WT (primary GBM) or could progress through the low grades having IDH mutation (secondary GBM). IDH mutant GBM patients have been found to show GCIMP phenotype with an expression profile of proneural subtype with better survival than that of the IDH WT GBM patients [10]. In our analysis, GBM patients were categorically divided into IDH mutant and WT based on the clinical data available from the TCGA datasheet. As expected, IDH mutants comprised a minor per cent (5.6%; 7/124) showing significantly better survival than IDH WT GBM patients. In the present analysis, we identified 13 CpG probe methylation signature, using the TCGA dataset, which could distinguish IDH mutant from the IDH WT in GBM patient samples. The classification accuracy in the training set and test set were 100% and 98.36% respectively. The signature identified one GBM IDH wild type as a discordant sample. While

this GBM IDH WT discordant sample had WT IDH1 as per both immunohistochemical staining and DNA sequencing, this sample had no amplification of EGFR locus and intact PTEN gene. These 13 CpG probe methylation signature comprised of genes coding for PCDP1 (Primary Ciliary Dyskinesia Protein 1), LOC144571, PRR18 (Proline Rich Region 18), YPEL4 (Yippee Like 4), SRRM3 (Serine/Arginine Repetitive Matrix 3), GLUL (Glutamate Ammonia Ligase), OBFC2A (Oligonucleotide/Oligosaccharide-Binding Fold-Containing Protein 2A), ATP5G2 (ATP Synthase, H⁺ Transporting, Mitochondrial Fo Complex Subunit C2), TUBA4B (Tubulin Alpha 4B) and CHADL (Chondroadherin Like).

Conclusion

In conclusion, we were able to classify diffuse glioma subtypes with high accuracy. The discordant samples identified by the methylation signature were found to be either due to technical errors or mixed histological types. More importantly, we believe that the high levels of intra-tumoral heterogeneity reported in glioma could also be a reason for their misclassification [7, 22]. Collectively, our study indicates that the methylation based molecular profiles in combination with the revised 2016 WHO CNS tumour classification guidelines might be able to classify the samples more precisely.

Materials and Methods

Tumor samples and clinical details

Glioma TCGA dataset was used for this study. Methylation data for histologically defined WHO classification glioma types which included; astrocytoma (n=197), oligoastrocytoma (n=136), oligodendroglioma (n=197) and glioblastoma (n=124) samples, was used. Samples were then segregated according to the WHO 2016 CNS tumour IHC-based grading classification into three distinct groups namely 1. Lower Grade Glioma IDH

wild-type and mutant (LGG IDH WT and mutant), 2. Lower Grade Glioma IDH mutant with intact 1p/19q termed as diffuse astrocytoma and with 1p/19q codeletion termed as oligodendroglioma (DA and ODG), and 3. Glioblastoma IDH mutant and wild-type (GBM IDH WT and mutant). The clinical information for the same was also procured from TCGA.

With an aim to identify methylation differences between the diffuse glioma subtypes (based on IDH mutation and 1p/19q codeletion status) of each group a supervised machine learning approach through PAM (*Prediction Analysis of Microarrays*) was used. For this purpose the first step was to identify significantly differentially methylated CpG probes between Lower Grade Glioma IDH WT and mutant, DA and ODG, and between GBM IDH mutant and WT which is described in details below.

Identification of differentially methylated CpGs

In this study, three different comparisons were carried out – 1. LGG: IDH mutant versus WT, 2. LGG IDH mutant: 1p/19q codeletion (ODG) versus non-codeletion (DA), and 3. GBM: IDH mutant versus WT. For the first comparison between LGG IDH mutant versus WT, we have performed a Wilcoxon-rank sum test between IDH mutant and WT which yielded 2,69,442 CpG probes significantly differentially methylated in mutant versus WT. Next, a stringent cut-off of 0.4 absolute $\Delta\beta$ value was applied that showed 9,554 significantly differentially methylated (26 CpGs were hypomethylated and 9,528 CpGs were hypermethylated in IDH mutant LGG; **Additional File 2: Table S2**) CpG probes in mutant as compared to WT IDH LGG patients. Firstly, the TCGA 450K Human Methylation dataset for LGG patients with IDH mutation (n=433) and LGG patients with WT IDH (n=97) was randomized and 50% of each of the two classes formed the training set, and the remaining 50% was used as the test set. We randomized TCGA dataset 10 times to obtain 10 different training sets and their corresponding test sets. After performing PAM and internal cross

validation using SVM as well as subset analysis on each training set, the probe set that gave maximum SVM accuracy included a set of 14 discriminatory CpG probes. This 14 probe set signature that was prioritized by PAM was further applied on the test set (**Figure 2; Table 1**).

Similarly, analysis was carried out for LGG IDH mutant cohort with and without 1p/19q co-deletion (ODG and DA respectively) patients (**Figure 2**). For this comparison, between LGG IDH mutant 1p/19q codel (ODG) versus non-codel (DA), we have performed a Wilcoxon-rank sum test which yielded 1,60,288 CpG probes significantly differentially methylated in ODG versus DA. Next, a stringent cut-off of 0.2 absolute $\Delta\beta$ value was applied that showed 2,817 significantly differentially methylated (627 CpGs were hypomethylated and 2,190 CpGs were hypermethylated in ODG; **Additional File 2: Table S3**) CpG probes in mutant as compared to WT IDH LGG patients. The TCGA 450K Human Methylation dataset for LGG patients with 1p/19q codel (n=172) and non-codel (n=261) was randomized and 50% of each of the two classes formed the training set, and the remaining 50% was used as the test set. We randomized TCGA dataset 10 times. PAM and internal cross validation using SVM as well as subset analysis was performed on each training set, the probe set that gave maximum SVM accuracy included a set of 14 discriminatory CpG probes. This 14 probe set signature that was prioritized by PAM was further applied on the test set and these 14 CpGs are unique to this comparison (**Figure 2; Table 3**).

Likewise, the same work flow was followed to identify a methylation based signature that could distinguish the GBM IDH WT from mutant samples (**Figure 2**). In this comparison, between GBM IDH mutant versus WT patient samples, we have performed a Wilcoxon-rank sum test which yielded 69,669 CpG probes significantly differentially methylated in mutant versus WT. Next, a stringent cut-off of 0.2 absolute $\Delta\beta$ value was applied that showed 259 significantly differentially methylated (33 CpGs were hypomethylated and 226 CpGs were hypermethylated in mutant; **Additional File 2: Table**

S4) CpG probes in mutant as compared to WT IDH GBM patients. The TCGA 450K Human Methylation dataset for GBM patients with IDH mutation (n=7) and WT (n=117) was randomized and 50% of each of the two classes formed the training set, and the remaining 50% was used as the test set. We randomized TCGA dataset 10 times. PAM and internal cross validation using SVM as well as subset analysis was performed on each training set, the probe set that gave maximum SVM accuracy included a set of 13 discriminatory CpG probes. A 13 CpG probe set signature that was prioritized by PAM was further applied on the test set for further validation (**Figure 2; Table 4**).

Prediction Analysis of Microarray (PAM)

To identify a list of a minimal set of signatory probes from the significantly differentially methylated CpGs between each compared groups, Prediction analysis of microarrays (PAM) using the package pamr available in R software (version 3.1.0) was applied. PAM uses nearest shrunken centroid method for classifying samples. This method “shrinks” each of the class centroids towards the overall centroid by the threshold. In case of selecting a signature, it is ideal to choose a threshold value that would achieve a set of minimum number of genes with maximum accuracy thereby least error. For preparing input files for PAM analysis, the list of significantly methylated probes between each compared groups across all the tumor samples was randomized and 50% of each of the two classes formed the training set, and the remaining 50% was used as the test set. This randomization was performed 10 times which resulted into 10 different compositions of training set and their corresponding test set. Thereafter each of these 10 training sets was subjected to PAM analysis that uses 10-fold cross-validation to identify a predictive signature. Ten different training sets that were used to construct the PAM classifier resulted in ten non-identical predictive signatures; one for each iteration. The most promising signature which had the

maximum training and test set accuracies was chosen. We also performed an internal cross-validation on the training set of the most promising signature as predicted using PAM.

Internal cross-validation using Support Vector Machine (SVM) and Random Subset Sampling

For internal cross-validation, we have used Support Vector Machine (SVM).²⁰ Many prediction methods use SVM for classification of dataset into two or more classes. For a given set of binary classes training examples, SVM can map the input space into higher dimensional space and seek a hyperplane to separate the positive data examples from the negative ones with the largest margin. SVM based internal cross validation is used for the training sets of 1. LGG IDH mutant versus WT, 2. Diffuse astrocytoma versus oligodendroglioma, and 3. GBM IDH mutant versus WT. For each of the above mentioned cases, the samples were divided randomly into five sub-groups containing equal number of the respective samples. These five sub-groups of each cases, example LGG IDH mutant and WT, were made into five groups where each group contained one sub-group of LGG IDH mutant and one sub-group of LGG IDH WT samples. Consequently, one group of LGG IDH WT plus LGG IDH mutant was considered as a test set while, the rest four groups were considered as training set and this is referred to as a 'fold'. In this way, SVM models were built five times to give five folds, wherein every group was considered as a test set and the remaining groups as training set. The accuracy for each fold was checked by this method.

The predictive accuracy of the three signatures was also analyzed in a subset of the following cases: 1. LGG IDH mutant (217) versus WT (n=49), 2. Diffuse astrocytoma (n=131) versus oligodendroglioma (n=86), and 3. GBM IDH mutant (n=4) versus WT (n=59) by random subset sampling. PAM was used to predict the respective accuracies in the random subset sampling.

Principal Component analysis

Principal Component Analysis (PCA) uses orthogonal transformation to convert a set of variables into a set of values of linearly uncorrelated variables that are called principal components. The number of principal components can be less than or equal to the number of original variables. The first two principal components account for the largest possible variation in the dataset. PCA was performed using R package (version 3.1.0), on the training and test sets to know how well the identified methylation signature classifies LGG IDH mutant and WT.

This process was repeated for identifying a methylation signature between IDH mutant DA and ODG; and between GBM IDH mutant and WT (a cut off of 0.2 absolute $\Delta\beta$ was used here to identify significantly differently methylated probes between the two classes).

Additional Files contains two files: **Additional File 1** has five additional figures and their corresponding figure legends. **Additional File 2** has eight additional tables.

Abbreviation: GBM-Glioblastoma; AA-Anaplastic Astrocytoma; DA-Diffuse Astrocytoma; ODG-Oligodendroglioma; WHO-World Health Organization; TCGA-The Cancer Genome Atlas; IDH-Isocitrate dehydrogenase; PAM-Prediction analysis of Microarray; SVM-Support Vector Machine; PCA-Principle Component Analysis.

Ethics approval and consent to participate: Not Applicable

Consent for publication: Not Applicable

Availability of data and material: Information about TCGA methylation data are made publically available at <http://cancergenome.nih.gov>.

Competing interests: The authors declare that they have no competing interests.

Funding: Infrastructure support by funding from DST-FIST, DBT grant-in-aid and UGC (Centre for Advanced Studies in Molecular Microbiology) to MCB is acknowledged. KS thanks DBT, Government of India for financial support. KS is a JC Bose Fellow of the Department of Science and Technology.

Authors' contribution: KS coordinated the study; KS and BM conceived and wrote the paper; KS and BM designed while YP and VP performed analysis of the TCGA dataset for all the experiments.

Acknowledgement: The results published here are in whole or part based upon data generated by The Cancer Genome Atlas (TCGA) pilot project established by the NCI and NHGRI. Information about TCGA and the investigators and institutions which constitute the TCGA research network can be found at <http://cancergenome.nih.gov>. Infrastructure support by funding from DST-FIST, DBT grant-in-aid and UGC (Centre for Advanced Studies in Molecular Microbiology) to MCB is acknowledged. KS thanks DBT, Government of India for financial support. KS is a JC Bose Fellow of the Department of Science and Technology.

References:

1. Ostrom QT, Gittleman H, Fulop J, Liu M, Blanda R, Kromer C, Wolinsky Y, Kruchko C, Barnholtz-Sloan JS: **CBTRUS Statistical Report: Primary Brain and Central Nervous System Tumors Diagnosed in the United States in 2008-2012.** *Neuro-oncology* 2015, **17** Suppl 4:iv1-iv62.
2. Cohen AL, Holmen SL, Colman H: **IDH1 and IDH2 mutations in gliomas.** *Current neurology and neuroscience reports* 2013, **13**(5):345.
3. Louis DN, Ohgaki H, Wiestler OD, Cavenee WK, Burger PC, Jouvet A, Scheithauer BW, Kleihues P: **The 2007 WHO classification of tumours of the central nervous system.** *Acta neuropathologica* 2007, **114**(2):97-109.
4. Dunn GP, Rinne ML, Wykosky J, Genovese G, Quayle SN, Dunn IF, Agarwalla PK, Chheda MG, Campos B, Wang A *et al*: **Emerging insights into the molecular and cellular basis of glioblastoma.** *Genes & development* 2012, **26**(8):756-784.
5. Holland EC: **Glioblastoma multiforme: the terminator.** *Proceedings of the National Academy of Sciences of the United States of America* 2000, **97**(12):6242-6244.

6. Meyer M, Reimand J, Lan X, Head R, Zhu X, Kushida M, Bayani J, Pressey JC, Lionel AC, Clarke ID *et al*: **Single cell-derived clonal analysis of human glioblastoma links functional and genomic heterogeneity.** *Proceedings of the National Academy of Sciences of the United States of America* 2015, **112**(3):851-856.
7. Patel AP, Tirosh I, Trombetta JJ, Shalek AK, Gillespie SM, Wakimoto H, Cahill DP, Nahed BV, Curry WT, Martuza RL *et al*: **Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma.** *Science* 2014, **344**(6190):1396-1401.
8. Scherer HJ: **A Critical Review: The Pathology of Cerebral Gliomas.** *Journal of neurology and psychiatry* 1940, **3**(2):147-177.
9. Stupp R, Reni M, Gatta G, Mazza E, Vecht C: **Anaplastic astrocytoma in adults.** *Critical reviews in oncology/hematology* 2007, **63**(1):72-80.
10. Noushmehr H, Weisenberger DJ, Diefes K, Phillips HS, Pujara K, Berman BP, Pan F, Pelloski CE, Sulman EP, Bhat KP *et al*: **Identification of a CpG island methylator phenotype that defines a distinct subgroup of glioma.** *Cancer cell* 2010, **17**(5):510-522.
11. Shukla S, Pia Patric IR, Thinagararjan S, Srinivasan S, Mondal B, Hegde AS, Chandramouli BA, Santosh V, Arivazhagan A, Somasundaram K: **A DNA methylation prognostic signature of glioblastoma: identification of NPTX2-PTEN-NF-kappaB nexus.** *Cancer research* 2013, **73**(22):6563-6573.
12. Srinivasan S, Patric IR, Somasundaram K: **A ten-microRNA expression signature predicts survival in glioblastoma.** *PloS one* 2011, **6**(3):e17438.
13. Verhaak RG, Hoadley KA, Purdom E, Wang V, Qi Y, Wilkerson MD, Miller CR, Ding L, Golub T, Mesirov JP *et al*: **Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1.** *Cancer cell* 2010, **17**(1):98-110.
14. Cairns RA, Mak TW: **Oncogenic isocitrate dehydrogenase mutations: mechanisms, models, and clinical opportunities.** *Cancer discovery* 2013, **3**(7):730-741.
15. Parsons DW, Jones S, Zhang X, Lin JC, Leary RJ, Angenendt P, Mankoo P, Carter H, Siu IM, Gallia GL *et al*: **An integrated genomic analysis of human glioblastoma multiforme.** *Science* 2008, **321**(5897):1807-1812.
16. Yan H, Parsons DW, Jin G, McLendon R, Rasheed BA, Yuan W, Kos I, Batinic-Haberle I, Jones S, Riggins GJ *et al*: **IDH1 and IDH2 mutations in gliomas.** *The New England journal of medicine* 2009, **360**(8):765-773.
17. Zhao S, Lin Y, Xu W, Jiang W, Zha Z, Wang P, Yu W, Li Z, Gong L, Peng Y *et al*: **Glioma-derived mutations in IDH1 dominantly inhibit IDH1 catalytic activity and induce HIF-1alpha.** *Science* 2009, **324**(5924):261-265.
18. Turcan S, Rohle D, Goenka A, Walsh LA, Fang F, Yilmaz E, Campos C, Fabius AW, Lu C, Ward PS *et al*: **IDH1 mutation is sufficient to establish the glioma hypermethylator phenotype.** *Nature* 2012, **483**(7390):479-483.
19. Louis DN, Perry A, Reifenberger G, von Deimling A, Figarella-Branger D, Cavenee WK, Ohgaki H, Wiestler OD, Kleihues P, Ellison DW: **The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary.** *Acta neuropathologica* 2016, **131**(6):803-820.
20. Brennan CW, Verhaak RG, McKenna A, Campos B, Noushmehr H, Salama SR, Zheng S, Chakravarty D, Sanborn JZ, Berman SH *et al*: **The somatic genomic landscape of glioblastoma.** *Cell* 2013, **155**(2):462-477.
21. Frattini V, Trifonov V, Chan JM, Castano A, Lia M, Abate F, Keir ST, Ji AX, Zoppoli P, Niola F *et al*: **The integrated landscape of driver genomic alterations in glioblastoma.** *Nature genetics* 2013, **45**(10):1141-1149.
22. Coons SW, Johnson PC, Scheithauer BW, Yates AJ, Pearl DK: **Improving diagnostic accuracy and interobserver concordance in the classification and grading of primary gliomas.** *Cancer* 1997, **79**(7):1381-1393.

23. Nijaguna MB, Patil V, Hegde AS, Chandramouli BA, Arivazhagan A, Santosh V, Somasundaram K: **An Eighteen Serum Cytokine Signature for Discriminating Glioma from Normal Healthy Individuals.** *PloS one* 2015, **10**(9):e0137524.
24. Rao SA, Srinivasan S, Patric IR, Hegde AS, Chandramouli BA, Arimappamagan A, Santosh V, Kondaiah P, Rao MR, Somasundaram K: **A 16-gene signature distinguishes anaplastic astrocytoma from glioblastoma.** *PloS one* 2014, **9**(1):e85200.
25. Hegi ME, Diserens AC, Gorlia T, Hamou MF, de Tribolet N, Weller M, Kros JM, Hainfellner JA, Mason W, Mariani L *et al*: **MGMT gene silencing and benefit from temozolomide in glioblastoma.** *The New England journal of medicine* 2005, **352**(10):997-1003.
26. Colman H, Zhang L, Sulman EP, McDonald JM, Shooshtari NL, Rivera A, Popoff S, Nutt CL, Louis DN, Cairncross JG *et al*: **A multigene predictor of outcome in glioblastoma.** *Neuro-oncology* 2010, **12**(1):49-57.
27. Louis DN, Perry A, Burger P, Ellison DW, Reifenberger G, von Deimling A, Aldape K, Brat D, Collins VP, Eberhart C *et al*: **International Society Of Neuropathology--Haarlem consensus guidelines for nervous system tumor classification and grading.** *Brain pathology* 2014, **24**(5):429-435.
28. Issa JP: **CpG island methylator phenotype in cancer.** *Nature reviews Cancer* 2004, **4**(12):988-993.

Figure Legends:

Figure 1: Overview of the 2016 WHO CNS tumor classification based algorithm with the number of patients from TCGA dataset that is used in the present study.

Figure 2: The schematic representation of the work flow of statistical analysis.

To develop methylation-based signatures to distinguish diffuse glioma subtypes as per 2016 WHO CNS tumour classification, we subjected the 450K DNA methylation data of TCGA diffuse glioma samples (1. LGG IDH mutant versus WT; 2. LGG IDH mutant with 1p/19q intact versus LGG IDH mutant 1p/19q co-del and 3 GBM IDH mutant versus WT) to various statistical tools and validation steps (**Details in Materials and Methodology**).

Figure 3: Identification of 14-CpG probe methylation signature in training set (TCGA) for LGG IDH mutant and WT samples.

(A) Plot demonstrating classification error for 9,554 CpG probes from PAM analysis in training set. The threshold value 18.9 corresponded to 14 discriminatory CpG probes which classified IDH mutant (n=217) and WT (n=49) LGG samples with classification error of 0%.

(B) Heat map of the 14 CpG discriminatory probes identified from the PAM analysis between LGG IDH mutant and WT patient samples in the training set (TCGA). A dual color-code was used where yellow indicates more methylation (hypermethylation) and blue indicates less methylation (hypomethylation).

(C) PCA was performed using β (methylation) values of 14- PAM identified CpG probes between IDH mutant (n=217) and WT (n=49) LGG samples in training set. A scatter plot is generated using the first two principal components for each sample. The color code of the samples is as indicated.

(D) The detailed cross validation probabilities of 10-fold cross-validation for the samples of training set based on the β values of 14 CpG probes are shown. For each sample, its probability as LGG IDH mutant (red color) and WT (green color) are shown and it was predicted by the PAM program as either IDH mutant or WT in LGG samples based on which grade's probability is higher. The original histological grade of the samples is shown on the top.

Figure 4: PCA and cross validation probabilities of IDH mutant and WT in LGG samples of test set (TCGA).

(A) Heat map of the 14 CpG discriminatory probes identified in PAM analysis in IDH mutant (n=216) and WT (n=48) LGG patient samples in the test set (TCGA). A dual color-code was used where yellow indicates more methylation (hypermethylation) and blue indicates less methylation (hypomethylation).

(B) PCA was performed using β (methylation) values of 14- PAM identified CpG probes between IDH mutant (n=215) and WT (n=48) LGG samples in test set. A scatter plot is generated using the first two principal components for each sample. The color code of the samples is as indicated.

(C) The detailed probabilities of 10-fold cross-validation for the samples of training set based on the β values of 14 CpG probes are shown. For each sample, its probability as IDH mutant (red color) and WT (green color) of LGG samples are shown and it was predicted by the PAM program as either LGG IDH mutant or WT based on which grade's probability is higher. The original histological grade of the samples is shown on the top.

Figure 5: Identification of 14-CpG probe methylation signature in training set (TCGA) for Diffuse astrocytoma (DA) and Oligodendroglioma (ODG).

(A) Plot demonstrating classification error for 2,817 CpG probes from PAM analysis in training set. The threshold value 9.491 corresponded to 14 discriminatory CpG probes which classified DA (LGG IDH mutant with intact 1p/19q; n=131) and ODG (LGG IDH mutant with 1p/19q code; n=86) LGG samples with classification error of 0.93%.

(B) Heat map of the 14 CpG discriminatory probes identified from the PAM analysis between DA and ODG patient samples in the training set (TCGA). A dual color-code was used where yellow indicates more methylation (hypermethylation) and blue indicates less methylation (hypomethylation).

(C) PCA was performed using β (methylation) values of 14- PAM identified CpG probes between DA (n=131) and WT (n=86) LGG samples in training set. A scatter plot is generated using the first two principal components for each sample. The color code of the samples is as indicated.

(D) The detailed cross validation probabilities of 10-fold cross-validation for the samples of training set based on the β values of 14 CpG probes are shown. For each sample, its probability as ODG (red color) and DA (green color) are shown and it was predicted by the PAM program as either ODG or DA in LGG samples based on which grade's probability is higher. The original histological grade of the samples is shown on the top.

Figure 6: PCA and cross validation probabilities of Diffuse Astrocytoma (DA) and Oligodendroglioma (ODG) IDH mutant in LGG samples of test set (TCGA).

(A) Heat map of the 14 CpG discriminatory probes identified in PAM analysis in DA (n=130) and ODG (n=86) LGG patient samples in the test set (TCGA). A dual color-code was used where yellow indicates more methylation (hypermethylation) and blue indicates less methylation (hypomethylation).

(B) PCA was performed using β (methylation) values of 14- PAM identified CpG probes between DA (n=130) and ODG (n=86) LGG samples in test set. A scatter plot is generated using the first two principal components for each sample. The color code of the samples is as indicated.

(C) The detailed probabilities of 10-fold cross-validation for the samples of training set based on the β values of 14 CpG probes are shown. For each sample, its probability as ODG (red color) and DA (green color) of LGG samples are shown and it was predicted by the PAM program as either DA or ODG based on which grade's probability is higher. The original histological grade of the samples is shown on the top.

Figure 7: Identification of 13-CpG probe methylation signature in training set (TCGA) for IDH mutant and WT in GBM.

(A) Plot demonstrating classification error for 259 CpG probes from PAM analysis in training set. The threshold value 2.694 corresponded to 13 discriminatory CpG probes which classified IDH mutant (n=4) and WT (n=59) GBM samples with classification error of 0%.

(B) Heat map of the 13 CpG discriminatory probes identified from the PAM analysis between IDH mutant and WT GBM patient samples in the training set (TCGA). A dual color-code was used where yellow indicates more methylation (hypermethylation) and blue indicates less methylation (hypomethylation).

(C) PCA was performed using β (methylation) values of 13- PAM identified CpG probes between IDH mutant (n=4) and WT (n=59) GBM samples in training set. A scatter plot is generated using the first two principal components for each sample. The color code of the samples is as indicated.

(D) The detailed cross validation probabilities of 10-fold cross-validation for the samples of training set based on the β values of 14 CpG probes are shown. For each sample, its probability as IDH mutant (red color) and WT (green color) GBM samples are shown and it was predicted by the PAM program as either IDH mutant or WT in GBM samples based on which grade's probability is higher. The original histological grade of the samples is shown on the top.

Figure 8: PCA and cross validation probabilities of IDH mutant and WT in GBM samples of test set (TCGA).

(A) Heat map of the 13 CpG discriminatory probes identified in PAM analysis in IDH mutant (n=3) and WT (n=58) GBM patient samples in the test set (TCGA). A dual color-code was used where yellow indicates more methylation (hypermethylation) and blue indicates less methylation (hypomethylation).

(B) PCA was performed using β (methylation) values of 13- PAM identified CpG probes between IDH mutant (n=3) and WT (n=58) GBM patient samples in test set. A scatter plot is generated using the first two principal components for each sample. The color code of the samples is as indicated.

(C) The detailed probabilities of 10-fold cross-validation for the samples of training set based on the β values of 13 CpG probes are shown. For each sample, its probability as IDH1 mutant (red color) and WT (green color) of GBM patient samples are shown and it was predicted by the PAM program as either GBM IDH mutant or WT based on which grade's probability is higher. The original histological grade of the samples is shown on the top.

Table 1: List of the 14-CpG methylation signature for LGG IDH1 Mutant vs WT												
No.	CpG ID	Gene Name	Training Set					Test Set				
			Average β in Mutant	Average β in WT	$\Delta\beta$ =(Avg β in Mutant -Avg β in WT)	p value	FDR	Average β in Mutant	Average β in WT	$\Delta\beta$ =(Avg β in Mutant-Avg β in WT)	p value	FDR
1	cg00976453	KCNB1	0.792	0.039	0.753	0	0	0.789	0.041	0.748	0	0
2	cg02423318	#NA	0.861	0.085	0.775	0	0	0.861	0.075	0.786	0	0
3	cg03300177	GNAO1	0.834	0.083	0.751	0	0	0.828	0.103	0.724	0	0
4	cg05866411	FGFRL1	0.782	0.119	0.662	0	0	0.779	0.137	0.642	0	0
5	cg07355841	TPPP3	0.832	0.055	0.777	0	0	0.828	0.055	0.772	0	0
6	cg08231710	MMP23A	0.873	0.147	0.726	0	0	0.873	0.167	0.705	0	0
7	cg08442798	#NA	0.765	0.023	0.743	0	0	0.759	0.022	0.736	0	0
8	cg10064339	UCP2	0.775	0.053	0.722	0	0	0.771	0.064	0.707	0	0
9	cg10504751	GNAO1	0.834	0.088	0.746	0	0	0.821	0.109	0.712	0	0
10	cg11302533	#NA	0.777	0.035	0.742	0	0	0.770	0.034	0.736	0	0
11	cg12565681	RHBDF2	0.834	0.061	0.773	0	0	0.834	0.069	0.764	0	0
12	cg20564913	FGFRL1	0.833	0.124	0.708	0	0	0.828	0.141	0.687	0	0
13	cg25499397	GPR62	0.817	0.087	0.730	0	0	0.811	0.098	0.713	0	0
14	cg25813864	RAPGEFL1	0.846	0.060	0.786	0	0	0.840	0.056	0.785	0	0

Table 2: For the methylation-based signatures: Overall diagnostic accuracy, sensitivity and specificity 1. Low grade glioma IDH1 WT vs Mutant 2. Diffuse astrocytoma (DA) vs Oligodendroglioma (ODG) 3. GBM IDH1 WT vs Mutant

1. Low grade glioma IDH1 WT vs Mutant: for 14 CpG methylation-based signature

Cohort	Dataset	Overall accuracy (%) ^a	Sensitivity (%) ^b		Specificity (%) ^c		Overall Error (%)	IDH1 Mutant error (%)	IDH1 WT error (%)
			IDH1 Mutant	IDH1 WT	IDH1 Mutant	IDH1 WT			
TCGA	Training set	100 (266/266)	100 (217/217)	100 (49/49)	100 (49/49)	100 (217/217)	0	0	0
TCGA	Test set	99.62 (263/264)	99.53 (215/216)	100 (48/48)	100 (48/48)	99.53 (215/216)	0.38	0.47	0
TCGA	Combined set	99.81 (529/530)	99.76 (432/433)	100 (97/97)	100 (97/97)	99.76 (432/433)	0.19	0.24	0

2. Diffuse astrocytoma (IDH1 Mutant and non-codeletion of 1p/19q; DA) vs Oligodendroglioma (IDH1 Mutant and 1p/19q codeletion; ODG): for 14 CpG methylation-based signature

Cohort	Dataset	Overall accuracy (%) ^a	Sensitivity (%) ^b		Specificity (%) ^c		Overall Error (%)	DA error (%)	ODG error (%)
			DA	ODG	DA	ODG			
TCGA	Training set	99.07 (215/217)	98.47 (129/131)	100 (86/86)	100 (86/86)	98.47 (129/131)	0.93	1.53	0
TCGA	Test set	96.29 (208/216)	94.61 (123/130)	98.83 (85/86)	98.83 (85/86)	94.61 (123/130)	3.71	5.39	1.17
TCGA	Combined set	97.69 (423/433)	96.55 (252/261)	99.41 (171/172)	99.41 (171/172)	96.55 (252/261)	2.31	3.45	0.59

3. For GBM IDH1 WT vs Mutant: for 13 CpG methylation-based signature

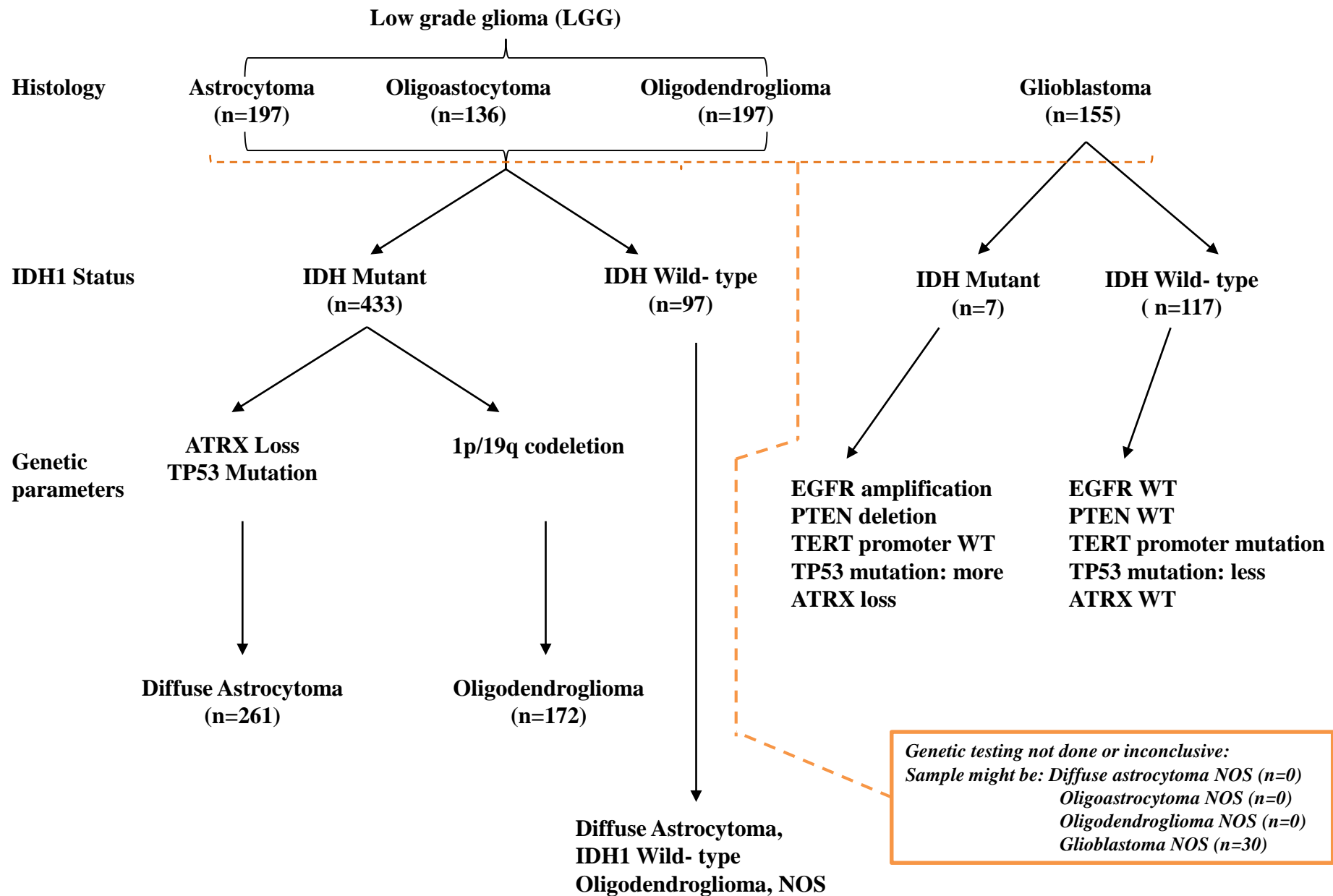
Cohort	Dataset	Overall accuracy (%) ^a	Sensitivity (%) ^b		Specificity (%) ^c		Overall Error (%)	GBM IDH1 Mutant error (%)	GBM IDH1 WT error (%)
			GBM IDH1 Mutant	GBM IDH1 WT	GBM IDH1 Mutant	GBM IDH1 WT			
TCGA	Training set	100 (63/63)	100 (4/4)	100 (59/59)	100 (59/59)	100 (4/4)	0	0	0

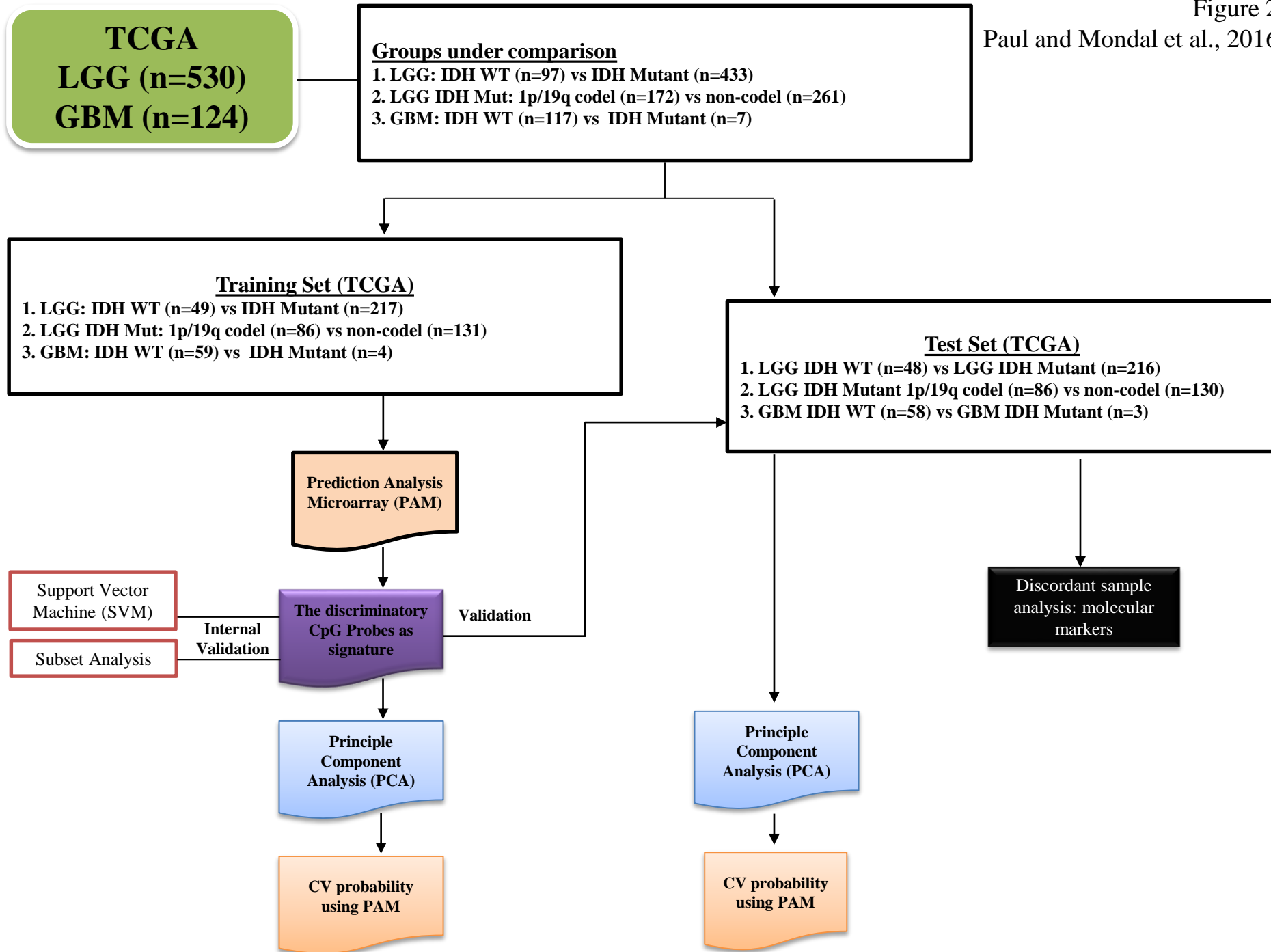
Table 3: List of the 14-CpG methylation signature for Diffuse astrocytoma (DA) versus Oligodendroglioma (ODG)

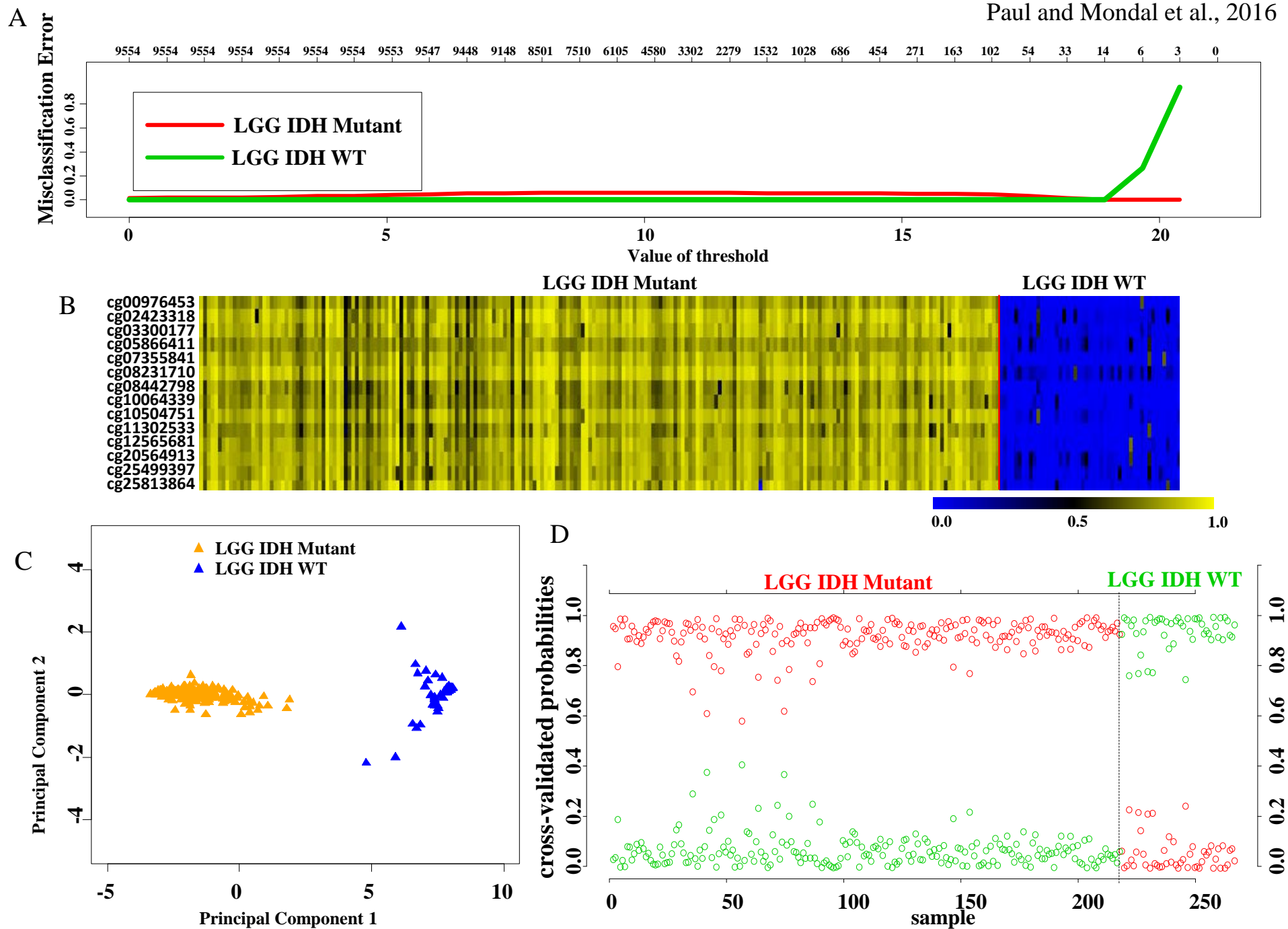
No.	CpG IDs	Gene Name	Training Set					Test Set				
			Average β in ODG	Average β in DA	$\Delta\beta$ =(Avg β in ODG-Avg β in DA)	p value	FDR	Average β in ODG	Average β in DA	$\Delta\beta$ =(Avg β in ODG-Avg β in DA)	p value	FDR
1	cg00873351	CD300LB	0.753	0.245	0.508	0	0	0.750	0.262	0.488	0	0
2	cg03492827	#NA	0.631	0.202	0.429	0	0	0.615	0.213	0.402	0	0
3	cg04437966	FLJ37543	0.481	0.096	0.385	0	0	0.463	0.105	0.358	0	0
4	cg07250222	FGFR2	0.737	0.189	0.548	0	0	0.720	0.191	0.529	0	0
5	cg07847030	TCF7L1	0.715	0.176	0.538	0	0	0.699	0.190	0.509	0	0
6	cg07893801	PLCG1	0.812	0.326	0.485	0	0	0.816	0.341	0.476	0	0
7	cg08935418	PTPRN2	0.721	0.226	0.495	0	0	0.707	0.221	0.486	0	0
8	cg09772154	FGFR2	0.715	0.193	0.522	0	0	0.696	0.191	0.505	0	0
9	cg10363569	PRKAG2	0.667	0.172	0.495	0	0	0.660	0.183	0.477	0	0
10	cg12210255	#NA	0.646	0.157	0.489	0	0	0.629	0.174	0.455	0	0
11	cg13412754	MAPKAP1	0.774	0.330	0.445	0	0	0.766	0.349	0.418	0	0
12	cg13598010	#NA	0.766	0.197	0.568	0	0	0.753	0.220	0.533	0	0
13	cg19093820	GPR156	0.206	0.712	-0.506	0	0	0.202	0.708	-0.507	0	0
14	cg23759393	PTPRN2	0.715	0.204	0.511	0	0	0.708	0.204	0.504	0	0

Table 4: List of the 13-CpG methylation signature for GBM IDH1 Mutant vs WT

No.	CpG IDs	Gene Name	Training Set					Test Set				
			Average β in Mutant	Average β in WT	$\Delta\beta$ =(Avg β in Mutant-Avg β in WT)	p value	FDR	Average β in Mutant	Average β in WT	$\Delta\beta$ =(Avg β in Mutant-Avg β in WT)	p value	FDR
1	cg02629106	PCDP1	0.372	0.171	0.201	0.001	0.002	0.880	0.141	0.739	0.004	0.008
2	cg03539765	LOC144571	0.382	0.157	0.225	0.001	0.002	0.721	0.125	0.596	0.005	0.008
3	cg08173692	PRR18	0.357	0.151	0.205	0.001	0.002	0.803	0.124	0.679	0.006	0.008
4	cg10366093	YPEL4	0.553	0.249	0.304	0.001	0.002	0.632	0.268	0.364	0.023	0.023
5	cg12662576	#NA	0.487	0.268	0.219	0.001	0.002	0.813	0.306	0.507	0.011	0.012
6	cg15198101	SRRM3	0.392	0.167	0.226	0.001	0.002	0.802	0.156	0.647	0.006	0.008
7	cg15389472	GLUL	0.445	0.219	0.225	0.001	0.002	0.772	0.213	0.558	0.008	0.010
8	cg15454486	OBFC2A	0.492	0.291	0.201	0.001	0.002	0.909	0.300	0.609	0.005	0.008
9	cg16264705	ATP5G2	0.392	0.159	0.233	0.002	0.002	0.815	0.140	0.675	0.005	0.008
10	cg16725050	TUBA4B	0.493	0.266	0.227	0.001	0.002	0.838	0.263	0.575	0.005	0.008
11	cg16917193	#NA	0.427	0.194	0.234	0.001	0.002	0.784	0.150	0.634	0.006	0.008
12	cg21000447	CHADL	0.376	0.126	0.250	0.001	0.002	0.656	0.105	0.551	0.006	0.008
13	cg25664381	#NA	0.418	0.200	0.218	0.001	0.002	0.926	0.164	0.763	0.005	0.008







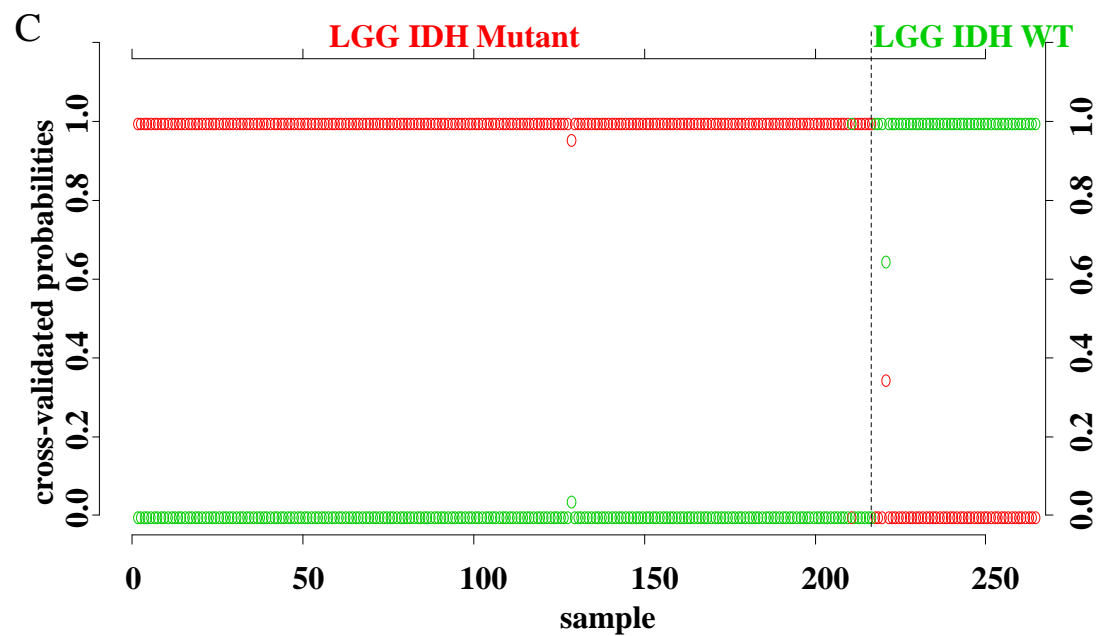
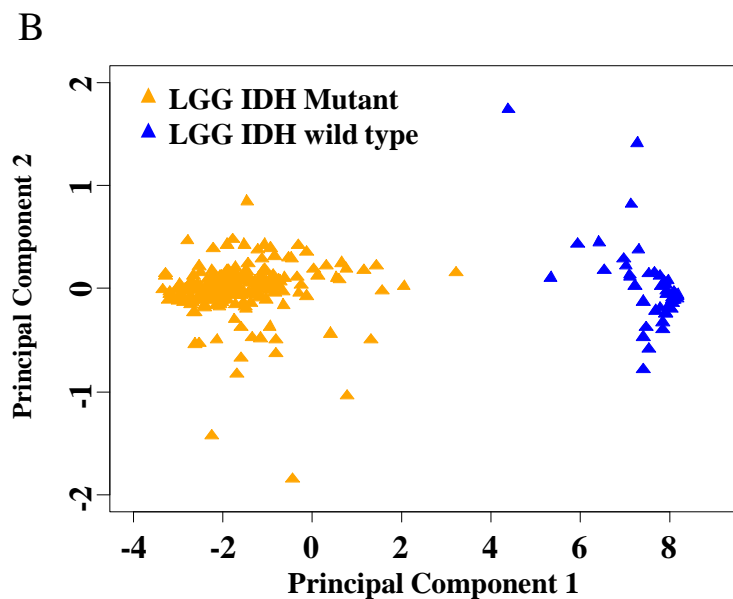
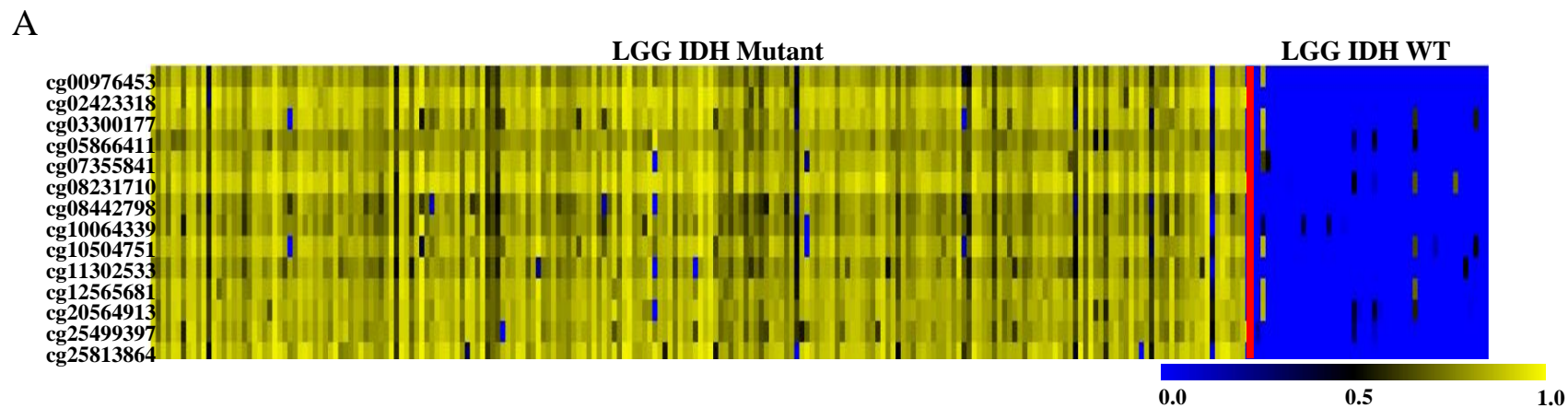
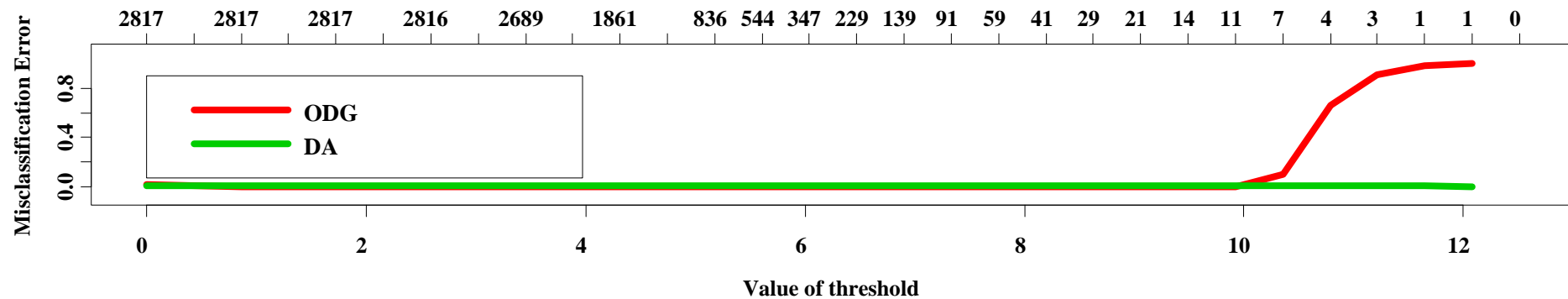


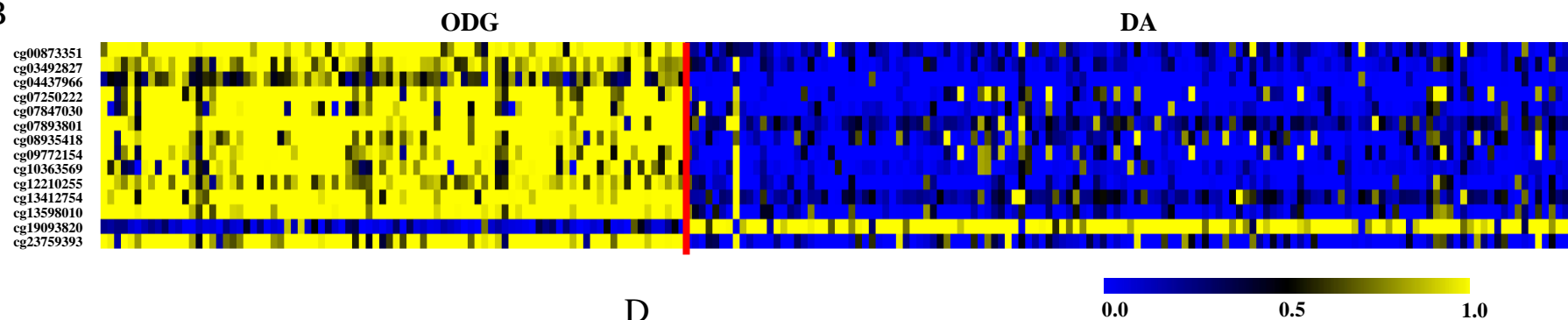
Figure 5

Paul and Mondal et al., 2016

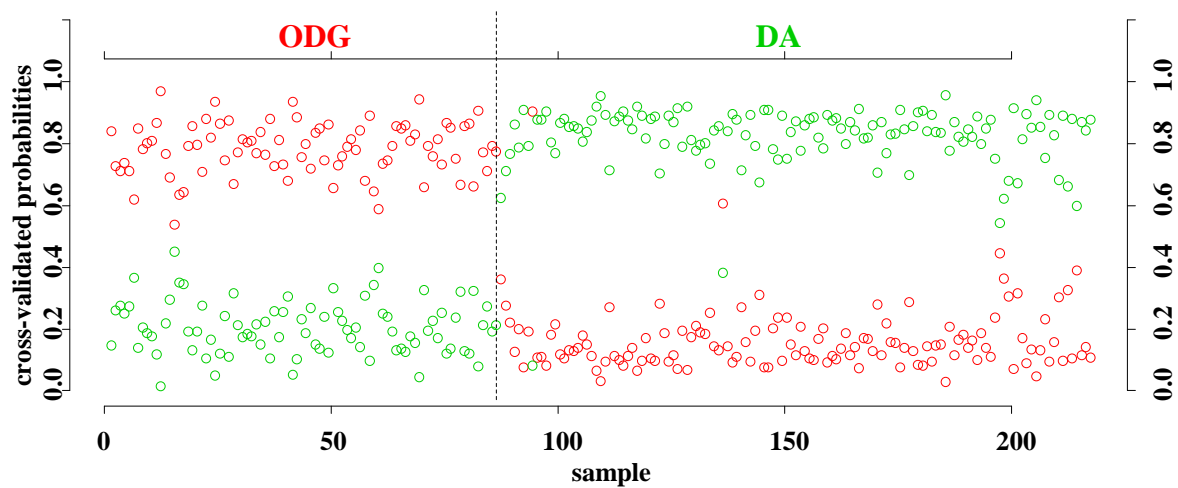
A



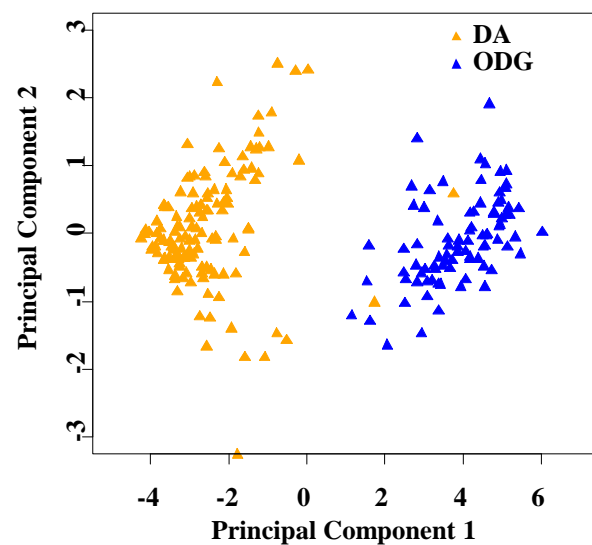
B



D



C



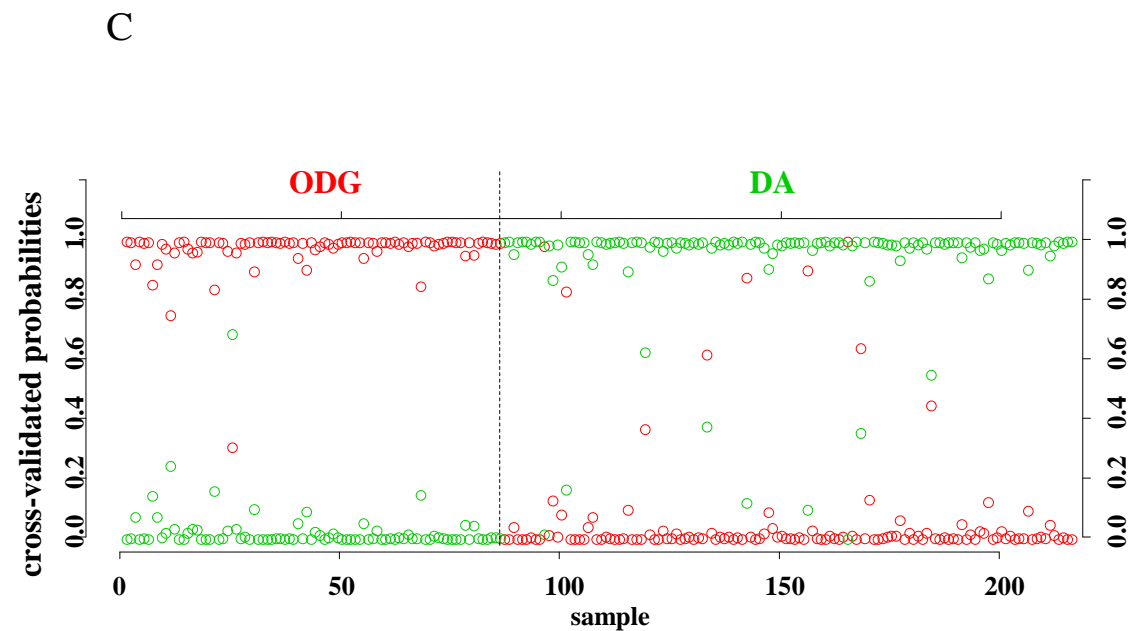
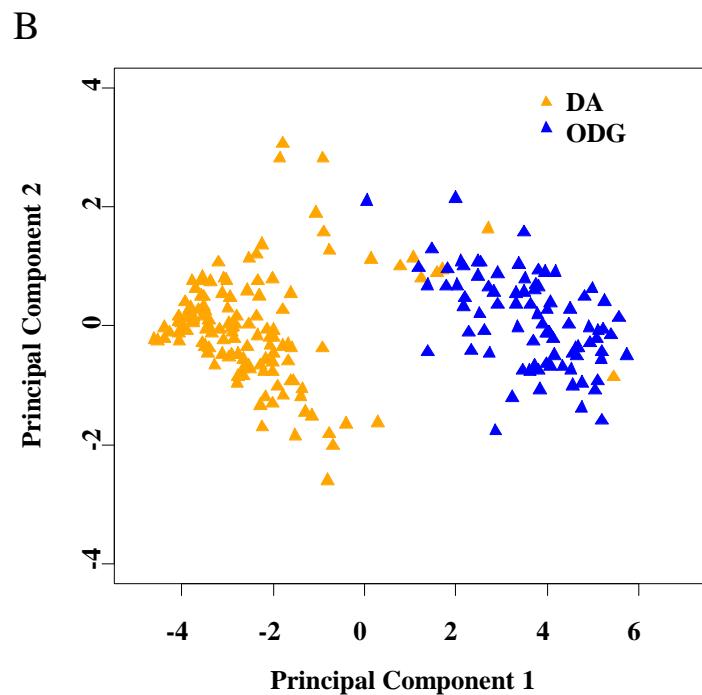
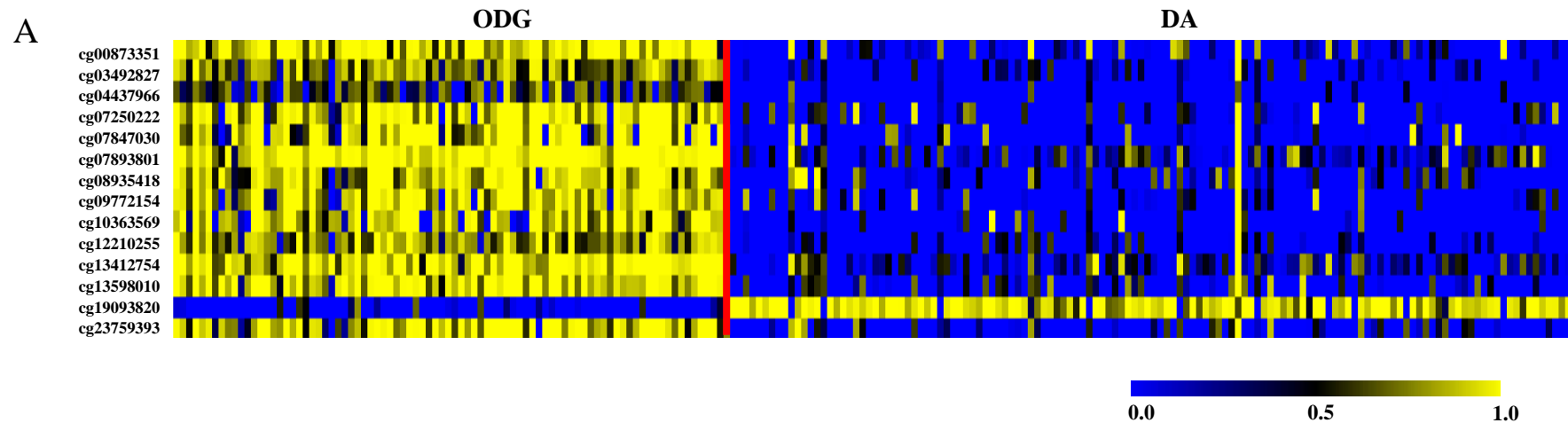
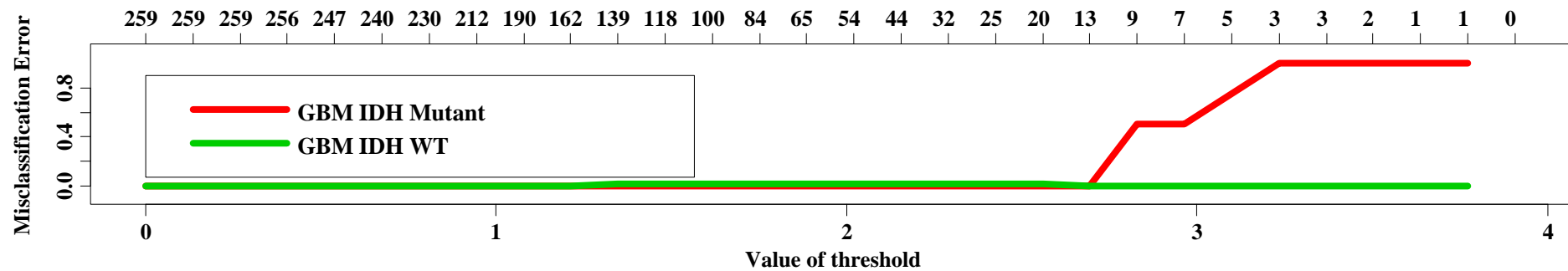


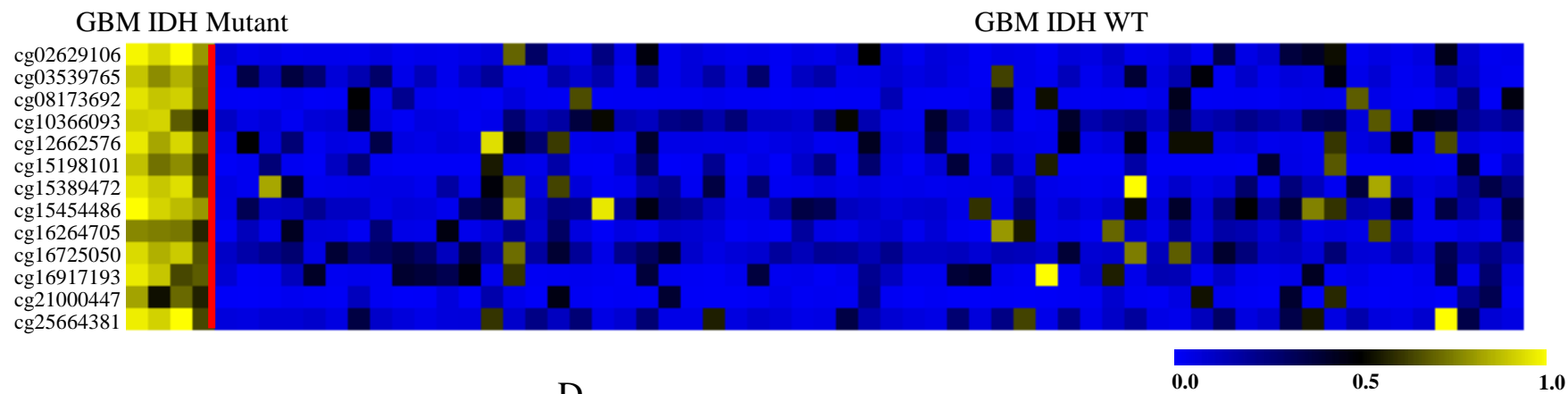
Figure 7

Paul and Mondal et al., 2016

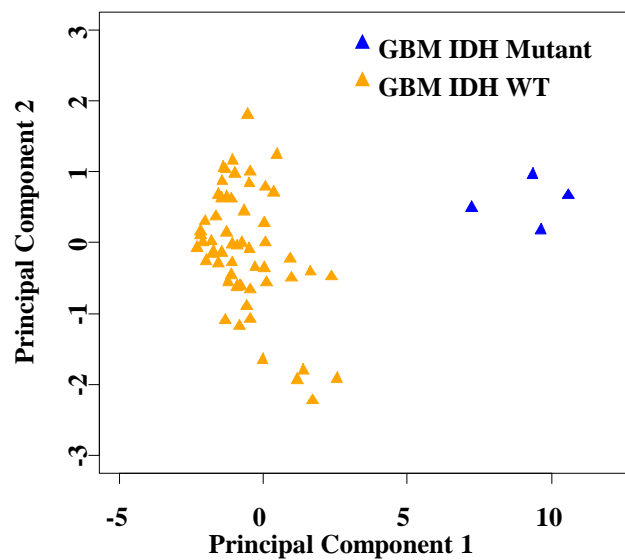
A



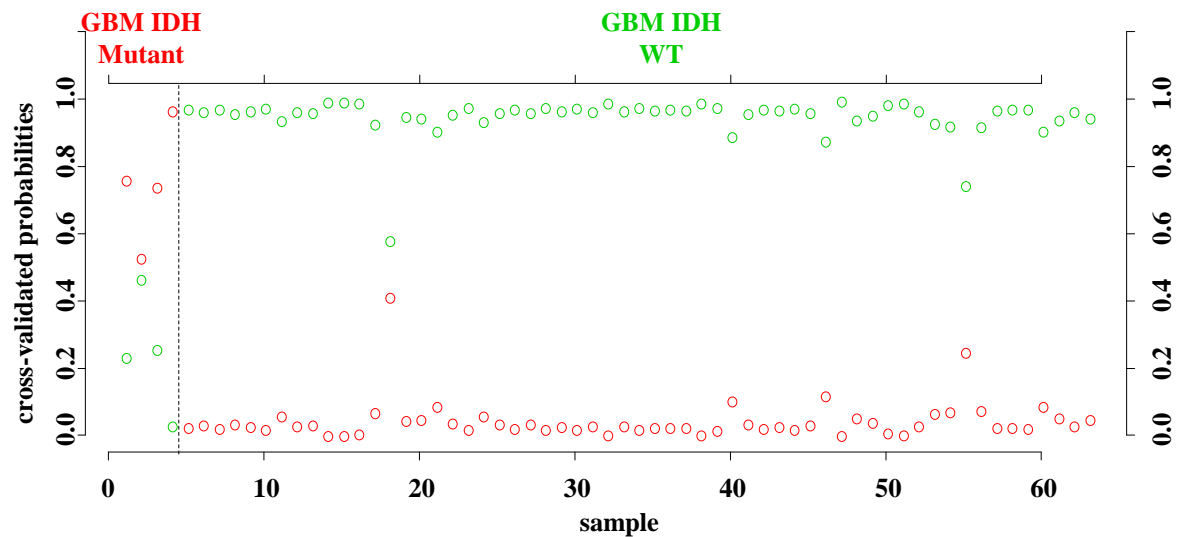
B



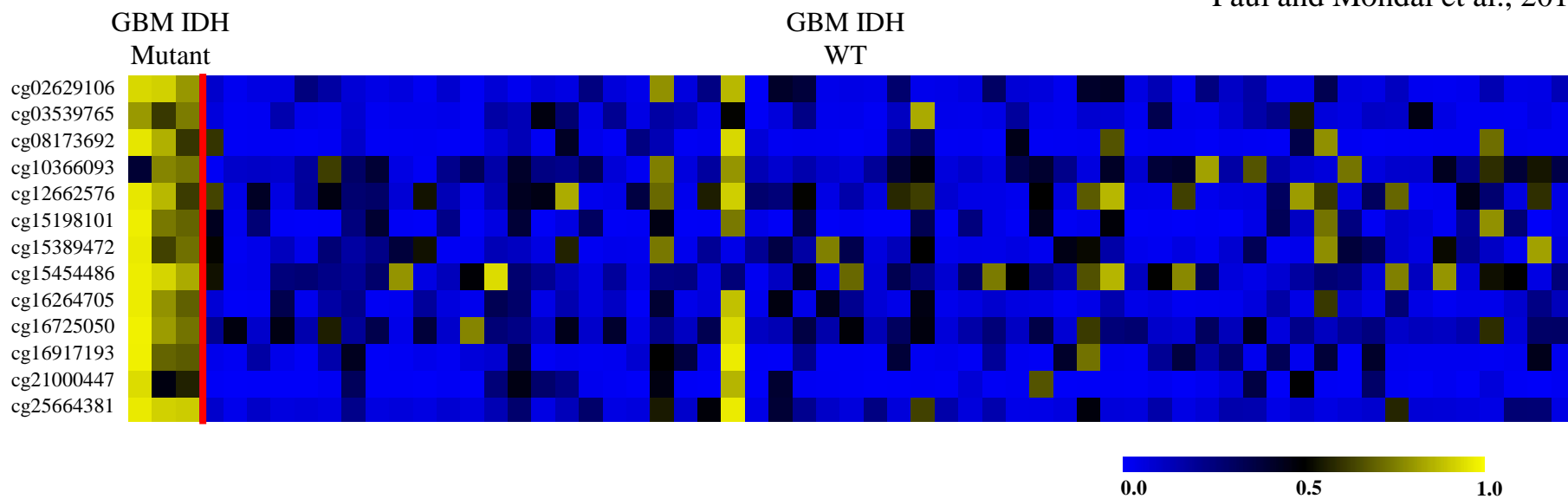
C



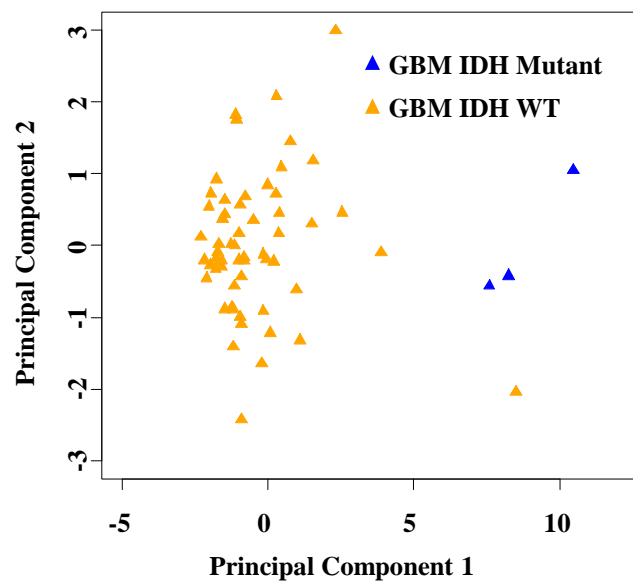
D



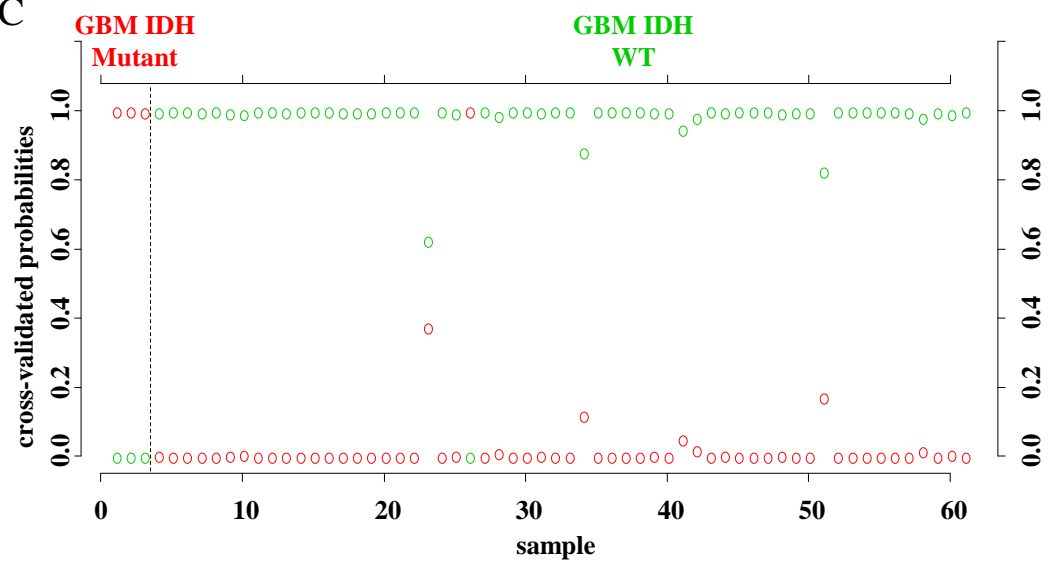
A



B



C





[Click here to access/download](#)

Supplementary Material

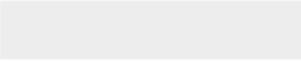
Paul and Mondal et al 2016 Additional File 1 Legends of
Figure S1-5.docx




[Click here to access/download](#)

Supplementary Material

Paul and Mondal et al 2016 Additional File 1 Figures S1-5.pptx






[Click here to access/download](#)

Supplementary Material

Paul and Mondal et al 2016 Additional File 2 Table
S1.xlsx






[Click here to access/download](#)

Supplementary Material

Paul and Mondal et al 2016 Additional File 2 Table
S2.xlsx

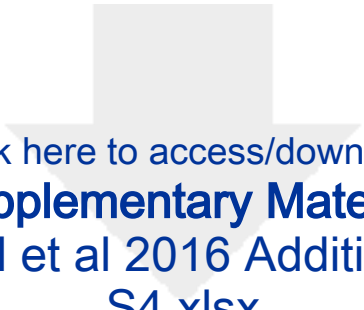




[Click here to access/download](#)

Supplementary Material

Paul and Mondal et al 2016 Additional File 2 Table
S3.xlsx




[Click here to access/download](#)

Supplementary Material

Paul and Mondal et al 2016 Additional File 2 Table
S4.xlsx






[Click here to access/download](#)

Supplementary Material

Paul and Mondal et al 2016 Additional File 2 Table
S5.xlsx





[Click here to access/download](#)

Supplementary Material

Paul and Mondal et al 2016 Additional File 2 Table
S6.xlsx



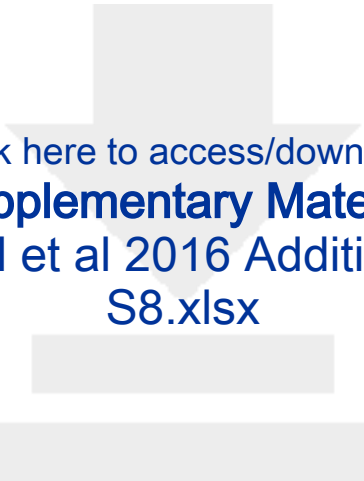


[Click here to access/download](#)

Supplementary Material

Paul and Mondal et al 2016 Additional File 2 Table
S7.xlsx





[Click here to access/download](#)

Supplementary Material

Paul and Mondal et al 2016 Additional File 2 Table
S8.xlsx