Jenkinson *et al. BMC Bioinformatics* (2019) 20:175
https://doi.org/10.1186/s12859-019-2777-6

BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

Ranking genomic features using an information-theoretic measure of epigenetic discordance

Garrett Jenkinson[1,2,3], Jordi Abante[1], Michael A. Koldobskiy[2,4], Andrew P. Feinberg[2,5,6] and John Goutsias[1*]

Abstract

Background: Establishment and maintenance of DNA methylation throughout the genome is an important epigenetic mechanism that regulates gene expression whose disruption has been implicated in human diseases like cancer. It is therefore crucial to know which genes, or other genomic features of interest, exhibit significant discordance in DNA methylation between two phenotypes. We have previously proposed an approach for ranking genes based on methylation discordance within their promoter regions, determined by centering a window of fixed size at their transcription start sites. However, we cannot use this method to identify statistically significant genomic features and handle features of variable length and with missing data.

Results: We present a new approach for computing the statistical significance of methylation discordance within genomic features of interest in single and multiple test/reference studies. We base the proposed method on a well-articulated hypothesis testing problem that produces *p*- and *q*-values for each genomic feature, which we then use to identify and rank features based on the statistical significance of their epigenetic dysregulation. We employ the information-theoretic concept of mutual information to derive a novel test statistic, which we can evaluate by computing Jensen-Shannon distances between the probability distributions of methylation in a test and a reference sample. We design the proposed methodology to simultaneously handle biological, statistical, and technical variability in the data, as well as variable feature lengths and missing data, thus enabling its wide-spread use on any list of genomic features. This is accomplished by estimating, from reference data, the null distribution of the test statistic as a function of feature length using generalized additive regression models. Differential assessment, using normal/cancer data from healthy fetal tissue and pediatric high-grade glioma patients, illustrates the potential of our approach to greatly facilitate the exploratory phases of clinically and biologically relevant methylation studies.

Conclusions: The proposed approach provides the first computational tool for statistically testing and ranking genomic features of interest based on observed DNA methylation discordance in comparative studies that accounts, in a rigorous manner, for biological, statistical, and technical variability in methylation data, as well as for variability in feature length and for missing data.

Keywords: DNA methylation, Genomic feature analysis, Information theory, Mutual Information, Gene ranking, Methylation analysis, WGBS data analysis

*Correspondence: goutsias@jhu.edu
[1]Whitaker Biomedical Engineering Institute, Johns Hopkins University, Baltimore, MD, USA
Full list of author information is available at the end of the article

Background

Epigenetics encompasses all cellular processes that lead to heritable changes in gene expression without modifying the DNA sequence. Uncovering the role of epigenetics in regulating a cell's phenotype is a key problem in modern molecular biology and medicine with important implications for understanding and treating human diseases, such as cancer.

The most studied epigenetic mechanism in humans is DNA methylation. This process chemically marks the DNA by adding a methyl group (CH_3) to individual cytosines located immediately adjacent to guanines (CpG sites). In humans, a targeted enzymatic machinery allows for the heritable transmission of these epigenetic marks from parent to progeny cells during cell division. It turns out that this dynamic mechanism is partially responsible for the developmental establishment and ongoing maintenance of cellular identity, as well as for many deleterious biological processes, such as aging and carcinogenesis [1–4].

Whole genome bisulfite sequencing (WGBS) provides a genome-wide assessment of DNA methylation patterns along the genome at a single-base resolution. For this reason, many computational methods have been proposed in the literature for the extraction and analysis of methylation information from this type of data [5, 6]. One such technique, known as informME (information-theoretic analysis of methylation), provides the most advanced capabilities available to-date for the modeling and analysis of DNA methylation from WGBS data [7, 8]. The method employs principles from statistical physics and information theory to build statistical models for WGBS data that provide accurate and insightful assessment of methylation information well beyond the one performed by other approaches. This is accomplished through genome-wide methylation analysis by modeling the DNA methylation state within regions of the genome using joint probability distributions computed directly from WGBS data, by quantifying stochasticity using the information-theoretic notion of normalized methylation entropy (NME), and by detecting methylation discordances in test/reference comparisons using the Jensen-Shannon distance (JSD) between joint methylation probabilities in test and reference samples. These methylation metrics go beyond mean-based analysis and have shown great promise when studying development, aging, and cancer [7].

A fundamental problem when analyzing WGBS data is linking the genome-wide results back to specific genomic features of interest (e.g., genes), and ranking these features based on their significance. This problem is commonly addressed by a procedure that first labels CpG sites or regions of the genome as being differentially methylated between a reference and a test sample and then quantifies their degree of overlap with features of interest [9, 10]. Unfortunately, this method is confounded by the variable lengths of most genomic features of interest. For example, by recognizing the importance of intragenic DNA in gene expression [11, 12], we may consider gene bodies as the genomic features of interest, whose length varies appreciably throughout the genome. In this case, any differentially methylated region (DMR) that is randomly placed on the DNA will more likely overlap with long gene bodies than short ones, and this will seriously skew the analysis. In addition to the previous issue, we have argued in [8] that there is a fundamental loss of power when performing genome-wide statistical analysis for DMR detection followed by scoring features of interest using DMR overlaps, or other DMR metrics, as compared to a targeted approach that scores features by focusing the statistical analysis on the features themselves. For this reason, we developed in [8] an approach for ranking genes based on observed methylation discordances within their promoter regions, determined by a fixed window centered at their transcription start sites (TSSs), and showed that it outperforms DMR overlap-based analysis. However, this ranking method has two critical weaknesses. First, it cannot appropriately handle genomic features with variable lengths, since a genomic region is scored by a p-value whose calculation depends on the region's length. Although this is not a problem for promoter regions, which are often taken to be of fixed length, it excludes other genomic regions of interest, such as gene bodies, exons, introns, bivalent domains, and enhancers. In addition, our method calculates a p-value, which is used to score a genomic region, by computing multiple p-values within the region, which are then combined using Fisher's method [13]. This approach however generates a combined p-value or score that cannot be trusted when evaluating statistical significance, since the individual p-values are not necessarily statistically independent, as required by Fisher's method.

In this paper, we introduce a new statistical approach for ranking genomic features that addresses the previous shortcomings. This method allows the user to input a set of genomic features of interest and receive an annotated ranked list of these features and their corresponding statistical significance, quantified by appropriately computed p- and q-values [to control the false discovery rate (FDR)]. The proposed approach evaluates, within a genomic feature of interest, DNA methylation discordance between a test and a reference phenotype, by quantifying the amount of information that the methylation state of a genomic feature contains about the phenotype, and by using this information to score the genomic feature. This is accomplished by articulating an appropriate hypothesis testing problem whose test statistic is derived using the mutual information between the methylation state and the phenotype. Importantly, this test statistic can be directly evaluated

from WGBS data by computing JSD values within a given genomic feature using informME.

To address the confounding issue of feature length variability, we assume that the null probability density function (PDF) of the test statistic required for hypothesis testing depends on the length of the feature. We then approximate this PDF using the logit skewed Student's t distribution (logitSST) with length-dependent parameters, which we estimate from replicate reference WGBS data using heteroscedastic regression. We verify the appropriateness of the logitSST model for estimating the PDF of the test statistic under the null hypothesis by performing goodness-of-fit and model selection analyses.

To illustrate the utility of the proposed method and its integration with informME, we reanalyzed previously available WGBS data obtained from healthy individuals and patients diagnosed with two different types of pediatric high-grade gliomas (pHGGs), a highly malignant form of brain tumor in children. Using these data, we ranked genes in terms of observed methylation discordance within their promoter regions and gene bodies, as well as within bivalent domains, whose role in the epigenetic regulation of gene expression is increasingly acknowledged in the literature [14]. Our results provide a clear demonstration of the importance and credibility of the proposed approach for linking genome-wide WGBS analysis results back to specific genomic features of interest and for appropriately ranking these features using their statistical significance. Our analysis shows that our method is capable of identifying genes that have been previously reported in the literature to be important in pHGGs, illustrates the importance of using multiple features (e.g., promoter regions and gene bodies) for ranking genes, demonstrates its seamless integration with informME, and establishes its importance as an exploratory tool in a WGBS analysis framework that can effectively identify important genomic regions and features for subsequent in-depth analysis.

We have coded the proposed method using R and have integrated it with informME. A fully documented GPLv3 licensed software implementation can be downloaded from GitHub (https://github.com/GarrettJenkinson/informME).

Methods

Information-theoretic analysis of methylation

In a previous work [7, 8], we developed an information-theoretic approach to the modeling and analysis of WGBS data known as informME. This methodology performs methylation data analysis by partitioning the genome into non-overlapping regions and by estimating the probability mass function (PMF) of the methylation state within these regions genome-wide. Let us consider one of these genomic regions comprised of N CpG sites $1, 2, \ldots, N$, which are indexed by their order of appearance along the genome. informME associates with the n-th CpG site a binary random variable X_n that takes values 0 or 1 if the site is unmethylated or methylated, respectively. It then characterizes the methylation state $\mathbf{X} = (X_1, X_2, \ldots, X_N)$ by a (usually) high-dimensional joint probability distribution $\Pr[\mathbf{X} = \boldsymbol{x}]$, which is modeled using the 1D Ising model of statistical physics estimated from available WGBS data using statistical inference. To summarize this high-dimensional probability distribution, informME performs methylation analysis by partitioning the genome into small analysis regions of 150bp each, which we refer to as genomic units (GUs). The methylation state within a GU with L CpG sites $\ell = 1, 2, \ldots, L$ is then characterized by the methylation level $M = \frac{1}{L}\sum_{\ell=1}^{L} X_\ell$, whose probability distribution $\Pr[M = m]$, $m = 0, 1/L, \ldots, 1$, is computed from the estimated Ising distribution $\Pr[\mathbf{X} = \boldsymbol{x}]$. In turn, this produces two statistical summaries of interest that we use to describe the statistical behavior of methylation within a GU: the mean methylation level (MML), given by $\mathrm{E}[M] = \sum_m m \Pr[M = m]$, and the normalized methylation entropy (NME), given by $h = -\left\{\sum_m \Pr[M = m] \log_2 \Pr[M = m]\right\}/\log_2(L + 1)$.

Mutual information and test statistic

In this paper, we are interested in statistically detecting DNA methylation discordances between a test and a reference phenotype within genomic features of interest, specified by their start and end coordinates along the genome. In particular, we seek to score genomic features in terms of the potential of their methylation states to distinguish between the two phenotypes. Genomic features of interest might include gene promoters, gene bodies, exons, introns, enhancers, bivalent domains, or any other genomic regions deemed to be important in a specific application.

A powerful way to proceed is to evaluate the dependance of the methylation state on the phenotype using the concept of mutual information [15]. In the following, we model the phenotype by a random variable Q that takes values 1 or 0, indicating a test or a reference phenotype, respectively. Moreover, we specify the methylation state of a genomic region by the K-dimensional random vector $\boldsymbol{M} = (M_1, M_2, \ldots, M_K)$, where M_k is the methylation level of the k-th GU with data (i.e., with computed MML, NME and JSD values) that overlaps the region. We can then measure the dependance of $\boldsymbol{M}$ on Q using the average mutual information within the genomic region, given by

$$\bar{I}(\boldsymbol{M}; Q) = \frac{1}{K}\sum_{k=1}^{K} I(M_k; Q), \tag{1}$$

where

$$I(M_k;Q)=\sum_{q=0,1}\sum_{m_k}\Pr[M_k=m_k,Q=q]\log_2\frac{\Pr[M_k=m_k,Q=q]}{\Pr[M_k=m_k]\Pr[Q=q]}. \quad (2)$$

In Eqs. (1) & (2), $I(M_k;Q)$ is the mutual information between M_k and Q, which tells us how much information the methylation state within the k-th GU carries about the phenotype and accounts for higher order relationships between the variables than simple correlations.

In the absence of any prior information about the phenotype, we can set $\Pr[Q=1]=\Pr[Q=0]=1/2$. In this case, we can show (Additional file 1: Section 1) that $\bar{I}(M;Q)=(1/K)\sum_{k=1}^{K}[\mathrm{JSD}(k)]^2$, where $\mathrm{JSD}(k)$ is the Jensen Shannon distance (JSD) [16] between the conditional methylation PMFs $\Pr[M_k=m_k\,|\,Q=1]$ and $\Pr[M_k=m_k\,|\,Q=0]$ of the test and reference phenotypes within the k-th GU, respectively. This result motivates us to statistically score epigenetic discordance within a genomic region using $1/K\sum_{k=1}^{K}[\mathrm{JSD}(k)]^2$ as the test statistic, since large values of this quantity indicate that DNA methylation within the region carries, on the average, significant information about its phenotypic state. However, for reasons we explain in Additional file 1: Section 2, we would like our test statistic to satisfy the triangle inequality $T(q_1,q_2)+T(q_1,q_3)\geq T(q_2,q_3)$, where $T(p,q)$ is the test statistic used to distinguish between two phenotypes p and q. It turns out that this is not true for $1/K\sum_{k=1}^{K}[\mathrm{JSD}(k)]^2$. However, we can show that it is true for

$$T=\sqrt{\frac{1}{K}\sum_{k=1}^{K}[\mathrm{JSD}(k)]^2} \quad (3)$$

(see Additional file 1: Section 2), which is the test statistic we use in this paper. Notably, T is a normalized test statistic that takes its minimum value 0 when the methylation state within the genomic region carries no information about the phenotypic state and its maximum value 1 when the methylation state is maximally informative about the phenotypic state (Additional file 1: Section 2). Moreover, T can be readily computed from available WGBS data, since the JSD can be calculated using informME [8].

Scoring genomic features

Two important issues arise when using the test statistic T in Eq. (3) to score a genomic feature: scoring is subject to biological, statistical, and technical variability, whereas the test statistic depends on the number of GUs with data that overlap the genomic region associated with the feature, which is affected by its length as well as by the number of overlapping GUs with missing data. Although we can sufficiently address the first issue by collecting replicate reference data, by performing all possible reference/reference comparisons, and by properly including the results of such analysis in the test/reference hypothesis testing problem, the second issue is more complex. In this section, we propose a method that takes into account biological, statistical, and technical variability, as well as the "sizes" of the genomic features under consideration. We quantify the size of a genomic feature by $s=\log_2 K$, where we introduce the logarithm to handle the large dynamic range of the number K of GUs with data that overlap the corresponding genomic region. In a given application, we may also wish to consider genomic features with sizes no smaller than a minimum size $s_{\min}=\log_2(K_{\min})$ to ensure that only features with sufficient length and/or data enter into the analysis.

We can simultaneously address the previous issues by computing the *null* PDF $f_0(t;s)$ of the test statistic T associated with a genomic region of size s, under the hypothesis that methylation discordance observed within the region is only due to biological, statistical, and technical variability. We can then compute the p-value, associated with an observation t_* of T in a test/reference comparison for a genomic feature of size s, by $p(s)=\int_{t_*}^{1}f_0(t;s)dt$, and use this p-value to rank the genomic region based on evidence against the null hypothesis that the observed value t can be explained by normal biological, statistical, and technical variability.

In theory, we could use replicate reference WGBS samples to empirically estimate $f_0(t;s)$. However, for this estimation to be sufficiently accurate, it is required that a large number of reference replicate data must be available and a prohibitively large number of reference/reference comparisons must be performed, which is not feasible in practice. We address this problem by employing a recently developed method for heteroscedastic regression, which we discuss next.

For a size s, we assume that the null PDF $f_0(t;s)$ can be sufficiently approximated by a logit skewed Student's t distribution (logitSST) $\phi(t;\theta_\mu(s),\sigma(s),\nu(s),\tau(s))$ with parameters μ, σ, ν, and τ that depend on s, in which case we set

$$f_0(t;s)\simeq\widehat{f_0}(t;s)=\phi(t;\mu(s),\sigma(s),\nu(s),\tau(s)). \quad (4)$$

The logitSST distribution is the PDF of the random variable $Y=1/\left(1+e^{-X}\right)$, $-\infty<X<\infty$, where X follows the skewed Student's t distribution [17].

To compute $\widehat{f_0}(t;s)$ in Eq. (4), we need to estimate the four parameters $\mu(s)$, $\sigma(s)$, $\nu(s)$ and $\tau(s)$ for any size s using observations of the test statistic T obtained from the reference samples. We perform this task by using the `gamlss` package of R [18], which assumes that $\mu(s)$, $\ln\sigma(s)$, $\ln\nu(s)$, and $\ln\tau(s)-2$ are smooth functions of s [19] and estimates these functions from data using generalized additive regression models based on penalized

splines [18, 20]. Smoothness ensures that the approximation $\widehat{f}_0(t;s)$ of the null PDF will be changing smoothly as the size s varies. In the "Results" section, we perform goodness-of-fit and model selection analyses to verify that the logitSST model provides an acceptable approximation to $f_0(t;s)$.

Scoring genes in a single test/reference comparison

As an example of the method we propose for scoring genomic features, we rank genes based on their significance in exhibiting differential DNA methylation discordance in a test/reference sample. One way to do this is to rank genes based on methylation discordance within their promoters. Towards this goal, we perform hypothesis testing by employing the test statistic T_p calculated within promoter regions using Eq. (3), and score the significance of each gene using the computed p-value for that gene, where higher significance is associated with a lower p-value. We can also use a similar approach to rank genes based on methylation discordance within their bodies by employing the test statistic T_b calculated within gene bodies.

In addition to the previous rankings, we may obtain a more informative gene ranking if we could simultaneously test for methylation discordance within their promoters and bodies. In this case, and for each gene, we can test the null hypothesis that methylation discordance between a test and a reference WGBS sample observed within its promoter region and gene body is only associated with biological, statistical, or technical variability in the reference sample, against the alternative hypothesis that this discordance is due to other factors within at least one of the two genomic regions (promoter or gene body). We can perform this hypothesis testing using Fisher's summary test statistic [13]

$$T_{pb} = -2\ln P_p - 2\ln P_b, \tag{5}$$

where P_p and P_b are the p-values obtained by separately testing, using the test statistics T_p and T_b, respectively methylation discordance within promoter regions or gene bodies.

If the two hypothesis testing problems were statistically independent, then T_{pb} would follow, under the combined null hypothesis, a χ^2_4 distribution with 4 degrees of freedom [13] from which a p-value for rejecting the combined null hypothesis could be readily obtained. However, due to the correlative nature of DNA methylation, the individual hypothesis tests may in general depend on each other, in which case, T_{pb} will not follow a χ^2_4 distribution.

To address this problem, we characterize the test statistic T_{pb} using its cumulative distribution function (CDF) $F_{pb}(t)$, which we empirically estimate by

$$\widehat{F}_{pb}(t) = \frac{1}{N_r}\sum_{n=1}^{N_r} I\left[t_n \le t\right], \tag{6}$$

where N_r is the number of observations t_n, $n = 1, 2, \ldots, N_r$, of the test statistic T_{pb} in the reference samples, and $I[\cdot]$ is the Iverson bracket, taking value 1 when its argument is true and 0 otherwise. In contrast to the theoretical χ^2_4 distribution, using $\widehat{F}_{pb}(t)$ will incorporate existing correlations into the problem and result in a more conservative and accurate statistical analysis than using the χ^2_4 distribution. In the "Results" section, we perform goodness-of-fit analysis and show that the empirical CDF $\widehat{F}_{pb}(t)$ provides a more appropriate characterization of the probability distribution of the test statistic T_{pb} than the theoretically derived χ^2_4 distribution.

As a consequence of the above, to rank genes based on observed methylation discordance within their promoters and bodies, we perform hypothesis testing using the T_p and T_b statistics to calculate (genome-wide) the p-values P_p and P_b for a given test/reference comparison, and compute the values of the test statistic T_{pb} using Eq. (5). For a given gene with observed test statistic value t_* between the test and the reference samples, we use the estimated null CDF $\widehat{F}_{pb}(t)$ in Eq. (6) and approximately calculate the probability (p-value) $P_{pb} \simeq 1 - \widehat{F}_{pb}(t_*)$ that, under the null hypothesis (of methylation discordance between a test and a reference sample observed within the gene's promoter region and body being only associated with biological, statistical, and technical variability), the test statistic T_{pb} is at least as large as the observed value t_*. We then employ this p-value to score the gene and use these scores to rank genes in terms of the significance of their methylation discordance in the test sample, with higher significance being associated with a lower score.

Scoring genes in multiple test/reference comparisons

We can also score a gene when multiple pairs of test/reference samples are available. To do so, we test the null hypothesis that epigenetic discordance observed within its promoter region and gene body in the test/reference comparisons is only associated with biological, statistical, or technical variability in the reference samples, against the alternative hypothesis that this discordance is due to other factors within at least one of the two genomic regions (promoter or gene body) in at least one of the test/reference samples. To address this problem, we use Fisher's summary test statistic

$$T_{\text{mult}} = -2\sum_{n=1}^{N_t} \ln P^{(n)}, \tag{7}$$

where N_t is the number of available test/reference comparisons and $P^{(n)}$ is the p-value obtained from the previous single comparison hypothesis testing problem (i.e., P_p, P_b, or P_{pb}) applied on the n-th test/reference pair.

The individual hypothesis testing problems can be considered to be statistically independent, in which case, the test statistic T_{mult} follows a $\chi^2_{2N_t}$ distribution with $2N_t$ degrees of freedom under the null hypothesis. From this distribution, we can compute a p-value for rejecting the null hypothesis, which we can use to score a gene's significance in terms of its methylation discordance in the test samples and produce a list of ranked genes. Note however that there might be ties in the resulting list, due for example to numerical issues that do not allow calculation of the p-values with arbitrary precision. We break such ties to the extent possible by using another list of ranked genes produced for this purpose, which we generate by combining the rankings obtained from each single test/reference comparison using the method of rank products [21]. Finally, to evaluate the statistical significance of each ranking while controlling for the FDR, we compute q-values using the Benjamini-Hochberg procedure [22].

Results

WGBS data samples

We now demonstrate the applicability of the proposed ranking method by reanalyzing, using informME, previously available WGBS data (Additional file 2: Table S1) obtained from normal fetal brain tissue or primary patient-derived pHGG tumor samples. For test samples, we use pHGG WGBS data from [23], which includes 7 primary pHGG samples harboring the H3.3 K27M mutation [a mutation within the histone H3.3 gene *H3F3A* that results in substitution of lysine 27 on the amino-terminal tail of H3.3 with methionine (K27M)], and 6 primary pHGG samples without K27M mutations (H3.3-WT). For reference samples, we use data (four samples) from normal fetal cerebellum tissue [24]. Boxplots of genome-wide distributions of JSD values obtained from our WGBS data confirm the appropriateness of the reference samples for providing a quantitative assessment of normal biological, statistical, and technical variability in our analyses (Additional file 1: Figure S1).

Goodness of fit

Estimation of null PDF of the T statistic

To demonstrate the appropriateness of the logitSST distribution for approximating the null PDF $f_0(t;s)$ of the test statistic T in Eq. (3), we considered a number of statistical models available in the `gamlss` R package and fitted each model to promoter region and gene body null T statistics obtained from six reference/reference comparisons. We then compared the results by employing two different model selection criteria, namely Akaike's information criterion (AIC) and the Bayesian information criterion (BIC). We took the promoter region of a gene to be the genomic region covered by a 4-kb window centered at the gene's TSS, and its gene body to be the genomic region between the gene's TSS and its termination site that does not overlap with its promoter region. We obtained this information by using the R package `TxDb.Hsapiens.UCSC.hg19.knownGene`. Moreover, we considered non-inflated and inflated distribution models based on the beta, generalized beta type 1, logit normal, logit t-family, and logit skewed Student's t distributions. The inflated models considered here include extra parameters to account for discrete probabilities of the test statistic taking its minimum and maximum values of 0 and 1 [18].

Ours results, summarized in Table 1, clearly show that the non-inflated logitSST model is superior under both criteria, in the sense that it produces the lowest AIC and BIC values, and this is true regardless of the type of genomic features considered (promoters or gene bodies). This model uses 25 degrees of freedom when fitted to promoter regions and 24 degrees of freedom when fitted to gene bodies. Note that a lower AIC means that the logitSST model is considered to be closer to the true model, whereas a lower BIC means that the logitSST model is considered more likely to be the true model.

In addition to the previous important result, we can also demonstrate that logitSST produces a high-quality fit to the null distribution of the T statistic data computed from the reference samples. Note that the size of gene bodies varies substantially more than the size of promoter regions. This is due to the fact that the size of a genomic feature is influenced by its actual length in bp's, which we choose to be fixed at 4-kb for the case of promoter regions. For this reason, we focus our discussion here on gene bodies. A similar approach applies for the case of promoter regions (see Additional file 1 for results pertaining to promoters).

To compute the approximation $\widehat{f_0}(t;s)$ of the true null PDF $f_0(t;s)$ of the T statistic within gene bodies, we applied informME on the four normal fetal cerebellum tissue samples, computed the JSD values in all six comparisons genome-wide, and calculated the T statistic values within all gene bodies via Eq. (3). To ensure that only features with sufficient length and/or data enter into our analysis, we considered only gene bodies with sizes $s \geq s_{\min} = \log_2(10)$ (i.e., we considered only bodies that overlap at least 10 GUs with data). This resulted in 104,694 paired observations (t_k, s_k) of null T statistic values t_k and gene body sizes s_k, which we passed to `gamlss` to produce $\widehat{f_0}(t;s)$.

In Fig. 1, we depict a scatter plot of pairs of (t_k, s_k) values for the case of gene bodies, together with α-centile curves computed from the estimated logitSST-based null PDF

Table 1 AIC and BIC values, computed using `gamlss`, for a number of non-inflated and inflated distribution models of the null probability density function of the *T* statistic, along with their (rounded) effective degrees of freedom (DF)

	Non-inflated Model			Inflated Model		
Model	AIC	BIC	DF	AIC	BIC	DF
Promoter						
BE	-486,026	-485,855	17	-486,229	-486,041	19
GB1	-488,175	-487,961	22	-488,408	-488,181	23
logitNO	-487,985	-487,895	9	-492,544	-492,384	16
logitTF	-493,963	-493,792	18	-493,944	-493,762	19
logitSST	**-494,260**	**-494,019**	**25**	-494,242	-493,992	26
Gene body						
BE	-482,633	-482,474	17	-482,608	-482,440	18
GB1	-484,822	-484,615	27	-484,798	-484,581	23
logitNO	-490,890	-490,730	17	-490,872	-490,702	18
logitTF	-495,777	-495,594	19	-495,751	-495,560	20
logitSST	**-498,362**	**-498,133**	**24**	-498,337	-498,099	25

The results were obtained by fitting each model to *T* statistic values within promoter regions and gene bodies obtained from all six reference/reference comparisons. Entries in bold highligth the proposed model, which is shown here to produce the best AIC and BIC scores
BE: beta distribution; GB1: generalized beta type 1 distribution; logitNO: logit normal distribution; logitTF: logit t-family distribution; logitSST: logit skewed Student's *t* distribution

$\widehat{f}_0(t;s)$ using different values of α (see Additional file 1: Figure S2 for results pertaining to promoter regions). An α-centile curve indicates that α% of the data points are below the curve. The results demonstrate that the estimated centile curves match the data very well. This is due to the fact that the percentage $\widehat{\alpha}$ of the empirically observed data points below a given estimated centile curve is close to its centile value α, indicating that the estimated null PDF $\widehat{f}_0(t;s)$ is consistent with the data. Note also that the data depicted in Fig. 1 demonstrate clear

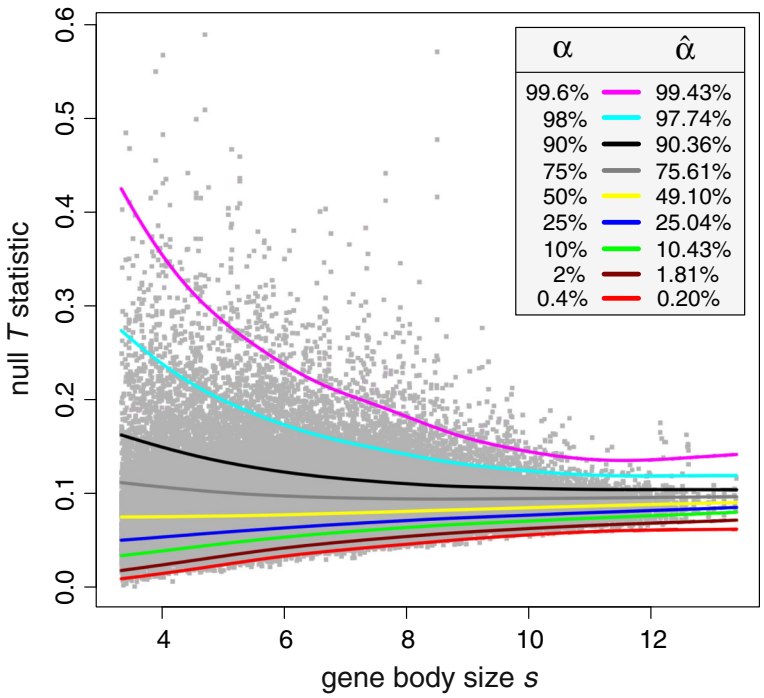

Fig. 1 α-centile curves, calculated for different values of α from the estimated logitSST-based null PDF $\widehat{f}_0(t;s)$ within gene bodies, drawn over a scatter plot of 104,694 observed pairs (t_k, s_k) of null *T* statistic values t_k and gene body sizes s_k. The percentage $\widehat{\alpha}$ of empirically observed data points that fall below a centile curve agrees well with the corresponding α value, indicating that $\widehat{f}_0(t;s)$ is consistent with the data

heteroscedasticity, indicating that the null PDF $f_0(t;s)$ depends on the gene body size s, as expected.

We further assessed the goodness of fit of the estimated PDF $\widehat{f_0}(t;s)$ using quantile residuals [25], as we explain next. Let $F_0(t;s)$ and $\widehat{F_0}(t;s)$ be the CDFs respectively associated with the true (but unknown) null PDF $f_0(t;s)$ and the logitSST-based estimated null PDF $\widehat{f_0}(t;s)$. Note that, for any size s, $U = F_0(T;s)$ is a random variable that is uniformly distributed over the unit interval $[0,1]$ regardless of the particular PDF of the T statistic. This implies that $G = \Phi^{-1}(U) = \Phi^{-1}(F_0(T;s))$, where Φ is the CDF of the standard normal distribution, is a zero-mean Gaussian random variable with unit standard deviation. We therefore expect that, when the logitSST-based estimated null PDF $\widehat{f_0}(t;s)$ is a good approximation of the true null PDF $f_0(t;s)$, the estimated quantile residuals, computed by $g_k = \Phi^{-1}(\widehat{F_0}(t_k;s_k))$ will be samples drawn from the standard normal PDF, where $t_1 \leq t_2 \leq \ldots$. If this turns-out to be true, then we can claim that the estimated null PDF provides a good fit of the true null PDF.

In Fig. 2a, we depict a kernel density approximation of the logitSST-estimated quantile residuals (shown in red at the bottom of the figure), obtained for the case of gene bodies using the reference samples [see Additional file 1: Figure S3a for the case of promoter regions]. Moreover, we provide values of four measures of location and shape of the approximated PDF, with the values in parentheses corresponding to the standard normal distribution. This result shows that the logitSST-estimated quantile residuals follow a probability distribution that is very close to being standard normal. We also depict in Fig. 2b a quantile-quantile (Q-Q) plot (green marks) of the logitSST-estimated quantile residuals for the case of gene bodies against the corresponding true quantile residuals [see Additional file 1: Figure S3b for the case of promoter regions]. The fact that the Q-Q plot is very close to the diagonal (red) line, is another indication of standard normality of the logitSST-estimated quantile residuals obtained from the reference samples.

Estimation of null CDF of the T_{pb} statistic

As we discussed earlier, scoring genes in a single test/reference comparison requires evaluation of the null probability distribution of the test statistic T_{pb} in Eq. (5). Due to the correlative nature of methylation, T_{pb} may not follow a χ^2_4 distribution, as theoretically expected by Fisher's method. For this reason, we chose to approximate the true null CDF of the test statistic T_{pb} using the empirical estimate $\widehat{F}_{pb}(t)$, given by Eq. (6).

To show the superiority of empirically approximating the null CDF of T_{pb}, we again performed quantile residual analysis. In Fig. 3a, we depict kernel density approximations of the quantile residuals obtained by using the null CDF associated with the χ^2_4 distribution (left), as well as the empirical null CDF (right), computed by using Eq. (6). The values of the location and shape parameters associated with these densities demonstrate that the empirical CDF provides a better approximation to the true distribution, since the computed values in this case are closer to the ones that correspond to the standard normal distribution (values in parentheses). This is also corroborated by the Q-Q plots depicted in Fig. 3b. In the Q-Q plot that corresponds to the χ^2_4 distribution (left), large quantile residuals exhibit noticeable deviation from the true residuals. However, the Q-Q plot corresponding to the empirical distribution (right) shows excellent match, indicating that this distribution is very close to the true distribution.

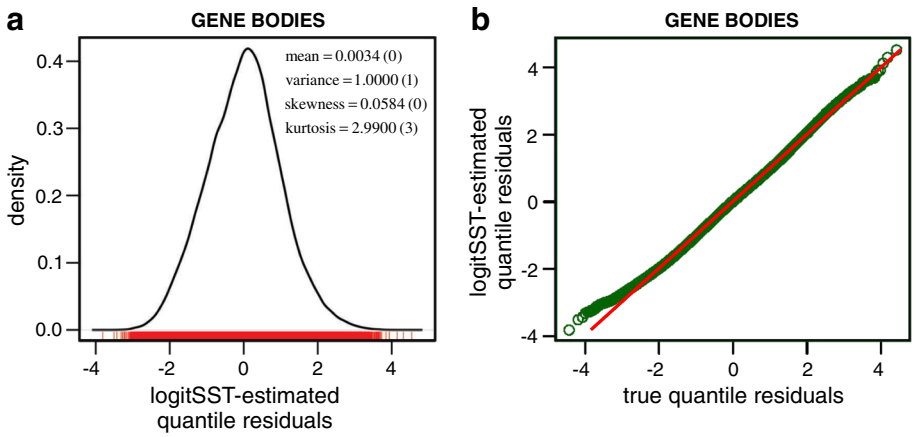

Fig. 2 Quantile residual analysis of the logitSST-estimated null PDF of the T statistic in the case of gene bodies. **a** The kernel density approximation of the distribution of the logitSST-estimated quantile residuals (bottom red marks) demonstrates close agreement with standard normality. **b** The Q-Q plot (green marks) of the logitSST-estimated quantile residuals against the corresponding true quantile residuals is very close to the diagonal (red) line, suggesting a close agreement of the logitSST-estimated null PDF of the T statistic to its true distribution

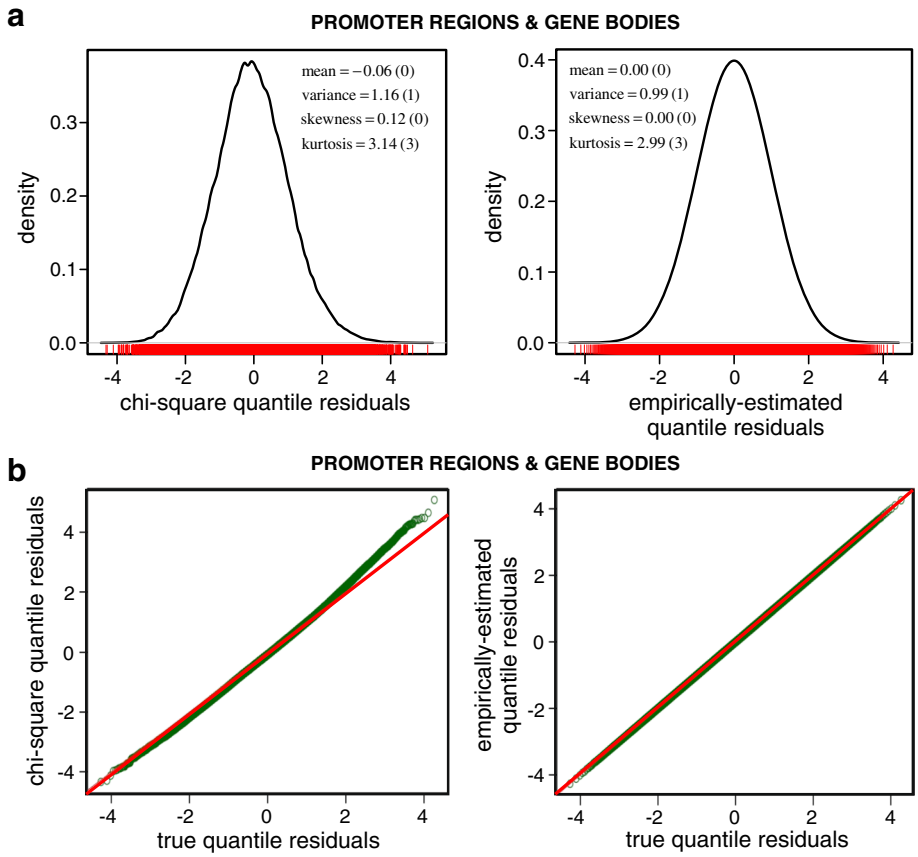

Fig. 3 Quantile residual analysis of the empirically-estimated null CDF of the T_{pb} statistic versus the null CDF associated with the χ^2_4 distribution, theoretically suggested by Fisher's method. **a** The kernel density approximations of the distributions of the χ^2_4 (left) and the empirically-estimated (right) quantile residuals demonstrate closer agreement of the latter with standard normality than the former. **b** The Q-Q plot (green marks) of the χ^2_4 (left) and the empirically-estimated (right) quantile residuals against the corresponding true quantile residuals corroborate the result in (**a**)

Gene ranking in pHGGs

In a previous paper [8], we introduced a method for identifying genes that exhibit significant epigenetic discordance within their promoters in a test/reference comparison. However, this method will miss genes that show significant epigenetic discordance only within their gene bodies. To find such genes, we may attempt to use a commonly applied heuristic that identifies genes overlapping differentially methylated regions (DMRs) of the genome determined by an effective DMR finder, such as the JSD-based DMR (jsDMR) detector of informME [8]. Nevertheless, results obtained by this method are confounded by gene length, which is problematic from a statistical perspective. More importantly, there is a fundamental issue with this strategy when analyzing cancer data characterized by profound changes in DNA methylation, such as the K27M mutant samples we consider in this paper, since detected DMRs may cover most of the genome. In this case, we will not be able to differentiate between most genes, since DMR overlap will be exceedingly common. Moreover, genes will tend to be completely covered by large DMRs, implying that more detailed heuristics, such as scoring genes based on the percentage of their overlap with DMRs, will fail in this case.

To demonstrate these issues, we processed a K27M mutant sample using informME and applied the jsDMR finder. In Additional file 2: Table S2, we list all genes whose promoters or bodies overlap a jsDMR, as well as the location of the corresponding jsDMR. Nearly all genes overlap a jsDMR and, therefore, this "target" list of genes is of little use when attempting to identify important genes. Not surprisingly, using this list for quantitative assessment by means of gene ontology (GO) enrichment analysis [26] or gene set enrichment analysis (GSEA) [27], produces no significant results, since there is no opportunity for enrichment when nearly every gene is included in the target list.

To properly analyze the pHGG data, we applied the gene ranking method proposed in this paper to find genes exhibiting significant DNA methylation discordance within their promoter regions or gene bodies. We obtained results by comparing the H3.3-WT and K27M

mutant pHGG samples to the normal fetal brain samples, which we summarize in Additional file 2: Tables S3 and S4, respectively.

As expected, a large number of genes demonstrate statistically significant methylation discordance (q-value $\leq$ 0.05) in the combined promoter/body lists (PB lists). Crucially, however, our method retains the ability to rank the genes and find those that are most differentially methylated. Many of the top genes in the PB lists, such as *KHDRBS2*, *PCDHGA2*, *PCDHGA3*, *RALYL*, *SLC25A21*, and *THSD7B*, for the case of H3.3-WT, as well as *EPHA3*, *EPHA6*, *ESR1*, *HBE1*, *HLX*, and *MTUS2*, for the case of K27M mutant, are not highly ranked in terms of their promoter region alone. As a consequence, their importance could be missed if only the promoter region list (PR list) is used. These genes exhibit profound methylation discordance within their bodies and, for this reason, most are placed at the top of the gene body list (GB list). Notably, *KHDRBS2*, *RALYL*, *SLC25A21*, *THSD7B*, *EPHA3*, *ESR1*, *EPHA6*, *HBE1*, and *HLX* have been implicated in brain tumors [28–39], whereas *PCDHGA2*, *PCDHGA3*, and *MTUS2* have known relationships to other types of cancer [40, 41].

In addition to the above, by further investigating the PB list of ranked genes associated with H3.3-WT (Additional file 2: Table S3), we observed that homeobox (HOX) genes, which are important in determining cell fate and identity during embryonic development and have been implicated in many cancers, rank particularly high in that list. Notably, it was recently suggested that *HOXB3*, a HOX gene ranked at the top of the PB list (although it is ranked 72 in the PR list and 45 in the GB list), promotes tumor cell proliferation and invasion in glioblastoma [42]. Moreover, *HOXA9* has been implicated as an oncogene in glioblastoma and its aberrant expression seems to be independently predictive of shorter survival rates [43]. In addition, an expression signature dominated by HOX genes has also emerged as a predictor for poor survival in patients treated with concomitant chemo-radiotherapy [44].

To further assess the biological significance of our gene rankings, we performed GO analysis using the PB list of genes ranked based on their discordance within their promoter regions and gene bodies, which we obtained from Additional file 2: Table S3 (for H3.3-WT) and Additional file 2: Table S4 (for the K27M mutant). We summarize the results in Additional file 2: Tables S5 and S6. Notably, GO analysis respectively identified 1293 and 232 significantly enriched GO process categories in the case of H3.3-WT and K27M mutant, indicating that our rankings are highly enriched near the top with genes of known biological significance. The GO results include several biological processes that play important roles in cancer initiation and progression, such as signaling, transcription regulation, cell communication, differentiation, commitment, morphogenesis, migration, and motility, as well as processes involved specifically in neuron development, differentiation, commitment, proliferation, and migration.

To bolster the biological relevance of the top genes in our lists, we also performed GSEA by using the top 500 significant genes from each ranked PB list and by computing overlaps with gene sets in the Molecular Signatures Database (MSigDB). We summarize these results in Additional file 2: Tables S7 and S8 for the case of H3.3-WT and K27M mutant, respectively. GSEA produced enrichments with gene sets related to stemness, as well as with genes known to exhibit DNA methylation and expression discordance in various cancers. In addition, GSEA showed a striking enrichment of genes whose promoters are bound by two functional enzymatic components (EED and SUZ12) of the Polycomb repressive complex 2 (PRC2), which promotes chromatin compaction by establishing, through another enzymatic component (EZH2), dimethylated H3K27 (H3K27me2) and trimethylated H3K27 (H3K27me3) marks [45]. This implies that, in pHGG, PRC2 targets genes that exhibit significant DNA methylation discordance.

The previous observation is not surprising, considering evidence that H3K27me3 and DNA methylation are compatible throughout most of the genome [46]. Notably, it has been suggested in [23] that the K27M mutation acts as a dominant negative inhibitor of H3K27 di- and trimethylation, due to an aberrant recruitment of the PRC2 complex to K27M mutant H3.3 and reduction of PRC2 activity through enzymatic inhibition of EZH2, and that this behavior is accompanied by hypomethylation, which is clearly seen in our K27M mutant samples (see Additional file 1: Figure S1). As a consequence, it has been suggested in [23] that inhibition of PRC2 activity and DNA hypomethylation may provide a mechanism of an altered gene expression program that drives tumor progression in K27M mutant cells. However, our finding that PRC2 also targets genes that exhibit significant DNA methylation discordance in H3.3-WT pHGG and the fact that this type of tumor exhibits hypomethylation as well (see Additional file 1: Figure S1), raises the possibility that a similar mechanism involving DNA hypomethylation at PRC2-regulated sites may also explain aberrant gene expression in H3.3-WT cells that leads to tumorigenesis.

In addition to the above, our observation that PRC2-regulated genes play a central role in both H3.3-WT and K27M mutant tumors is particularly intriguing, considering our previous suggestion that the PRC2 complex may be an important regulator of epigenetic stochasticity [7]. In the present framework, this is supported by the fact that both H3.3-WT and K27M mutant samples are globally characterized by gains in methylation entropy in most of our pHGG data (see Additional file 1: Figure S1).

After identifying a gene of interest using our ranked lists, we can employ informME to further investigate properties of the methylation state within its promoter region and gene body in a test/reference comparison. In Fig. 4, we depict the JSDs, MMLs, and NMEs, computed by informME when comparing an H3.3-WT and a K27M mutant sample to a normal fetal brain sample, within a genomic region that contains the Ephrin type-A receptor 3 (*EPHA3*) gene. This gene ranks at the top of the PB list in the K27M sample (Additional file 2: Table S4), despite the fact that it is ranked 1874 in the PR list. This is due to the fact that *EPHA3* exhibits profound methylation discordance within its gene body (ranked at the top of the corresponding GB list). Interestingly, *EPHA3* does not rank highly when using the H3.3-WT samples (2965 in the PB list, 8976 in the PR list, and 1326 in the GB list; see Additional file 2: Table S3). Indeed, the JSD tracks depicted in Fig. 4 indicate small differences in DNA methylation when comparing the H3.3-WT sample to the fetal brain sample, but large differences within the gene body of *EPHA3* in the case of the K27M sample. In agreement with our previous discussion, this discordance is associated with appreciable hypomethylation, as well as with gain in entropy that approaches its maximum value, indicating a highly disordered methylation state over *EPHA3*. In Additional file 1: Figure S4, we demonstrate that this behavior is consistent across all samples.

The previous finding about *EPHA3* is of potential clinical significance, since it indicates that its methylation state is distinctly regulated in pHGG tumors harboring the K27M mutation. Importantly, *EPHA3* has been identified as a therapeutic target in glioblastomas using small-molecule inhibitors or targeted antibodies [33, 47]. However, our results provide a specific hypothesis regarding *EPHA3* targeting in pHGG: tumors harboring the

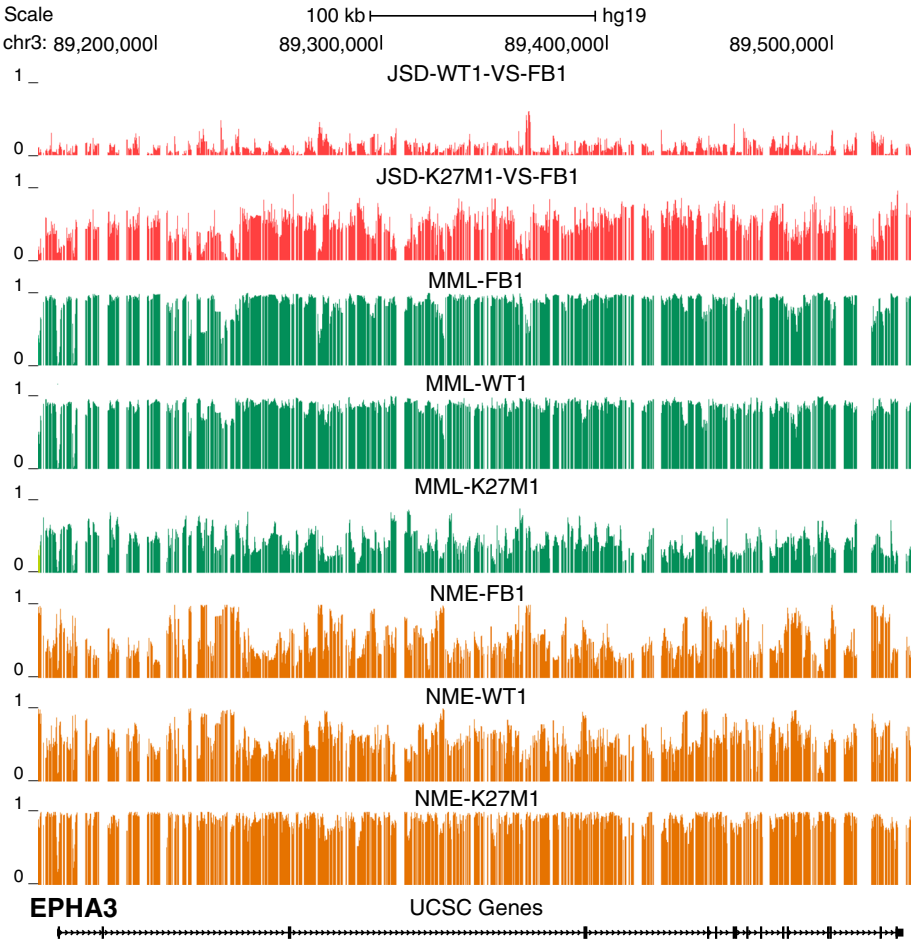

Fig. 4 UCSC genome browser images of JSD, MML, and NME tracks within the genomic region [chr3: 89,145,180–89,536,200] that contains *EPHA3*, obtained by informME in the WT1 vs. FB1 and K27M1 vs. FB1 comparisons. The JSD tracks indicate small differences in DNA methylation when comparing WT1 to FB1, but large differences when comparing K27M1 to FB1. This is associated with widespread hypomethylation in K27M1 (MML tracks) and a gain in methylation entropy (NME tracks) close to its maximum value

H3.3 K27M mutation have a markedly dysregulated *EPHA3* epigenetic signature while H3.3-WT pHGGs do not, suggesting that *EPHA3* therapeutic targeting may be more effective in patients with the H3.3 K27M histone mutation. In agreement with this hypothesis, a recent paper employed super-enhancer profiling of patient-derived diffuse intrinsic pontine glioma (DIPG) cell lines to demonstrate that Ephrin signaling is crucial for DIPG invasiveness, which mostly exhibits the H3.3 K27M mutation [48]. Therefore, *EPHA3* warrants further investigation as a therapeutic target in pHGG, while taking into account the K27M mutational status of the underlying tumor, which might enable new precision medicine efforts for this disease cohort.

Ranking bivalent domains in pHGGs

The results obtained by GSEA applied in H3.3-WT and K27M mutant pHGG (Additional file 1: Tables S7 and S8) revealed enrichment for genes in neural progenitor cells (NPCs) whose promoters bear the repressing H3K4 trimethylation mark (H3K4me3), as well as the activating histone H3K27 trimethylation mark (H3K27me3), which are distinctive histone modifications within bivalent domains [14]. Therefore, and as an additional example of the utility of our method for ranking genomic features, we sought to rank bivalent domains. The presence of bivalent domains within gene promoters and enhancers keeps the expression level of these "bivalent" genes at low levels, with the genes poised for rapid activation upon availability of suitable cues.

To determine bivalent domains, we used the genomic regions labeled "TssBiv" or "EnhBiv" under the ENCODE accession ENCSR567BIT. To ensure that only bivalent domains with sufficient length and/or data enter into the analysis, we considered only those with sizes $s \geq s_{\min} = \log_2(5)$ (i.e., we considered only bivalent domains overlapping at least 5 GUs with data). This resulted in 7446 paired observations (t_k, s_k) of null T statistic values t_k and bivalent domain sizes s_k, which were passed to `gamlss` to produce $\widehat{f}_0(t;s)$. The goodness-of-fit results depicted in Additional file 1: Figures S5 and S6 are similar to the ones obtained for promoter regions and gene bodies and show that the logitSST-estimated null PDF $\widehat{f}_0(t;s)$ provides a good approximation of the true null PDF of the T statistic within bivalent domains as well.

In Additional file 2: Table S9, we provide a ranking of bivalent domains produced by our method when comparing the H3.3-WT pHGG samples to a fetal brain sample, as well as the genes located nearest to these domains. From this ranking, we identified many bivalent domains with significant discordance in DNA methylation, with the most highly ranked domains being associated, for example, with *ZNF467*, *PCDH8*, *ISLR2*, *HLX*, *NR2E1*, *WT1*, *AGAP2*, *TGFB1I1*, *LRFN1*, *KLF4*, and *SOX10*, which have been previously implicated in brain tumors [39, 49–58]. We obtained similar results when comparing the K27M mutant pHGG samples to the same fetal brain sample (Additional file 2: Table S10). In this case, however, some genes that were located lower in the previous list, such as *ATXN10*, *CLDN5*, *EBF4*, and *SOX9*, were found in the top of the list. These genes have also been implicated in brain tumors [59–62].

The previous results indicate that using our methodology to rank bivalent domains in test/reference studies can be quite useful for identifying important bivalent genes in a test/reference comparison for further experimental analysis and validation. In conjunction with informME, ranking bivalent domains and genes can also help to increase our understanding regarding their role in a particular disease, such as pHGGs. For example, when comparing the H3.3-WT samples to the normal fetal brain samples, a bivalent domain located at [chr22: 38,379,200–38,389,200] and associated with *SOX10* is ranked much higher than a bivalent domain located at [chr17: 70,112,800–70,114,000] and associated with *SOX9*, while the opposite is true when comparing the K27M mutant and normal fetal brain samples (see Additional file 2: Tables S9 and S10). Since these genes have been implicated in gliomas [58, 62–65], we sought to further investigate these differences using informME.

In Fig. 5a, we depict the JSDs, MMLs, and NMEs computed by informME when comparing an H3.3-WT sample to a normal fetal brain sample within regions that contain *SOX10* (left) and *SOX9* (right). The bivalent domain associated with *SOX10*, which partially overlaps its promoter region, exhibits strong DNA methylation discordance, as indicated by the large JSD values over this domain. This is associated with hypermethylation in the H3.3-WT sample accompanied by substantial loss of methylation entropy. Interestingly, hypermethylation of the SOX10 promoter was found to be associated with shorter survival rates in glioblastoma [58]. On the other hand, the bivalent domain associated with *SOX9* shows smaller DNA methylation discordance than the bivalent domain in *SOX10*, which can be associated with moderate hypermethylation in the H3.3-WT sample accompanied by an appreciable gain in methylation entropy.

Interestingly, *SOX9* and *SOX10* exhibit a different behavior when comparing K27M mutant pHGG to normal fetal brain. Figure 5b shows that the bivalent domain associated with *SOX10* (left) exhibits again strong DNA methylation discordance. However, this discordance is now associated with appreciable hypomethylation in the K27M mutant sample (as opposed to hypermethylation in the H3.3-WT case), accompanied by a moderate loss of methylation entropy. On the other hand, the bivalent domain associated with *SOX9* shows strong methylation discordance (as opposed to smaller discordance in the

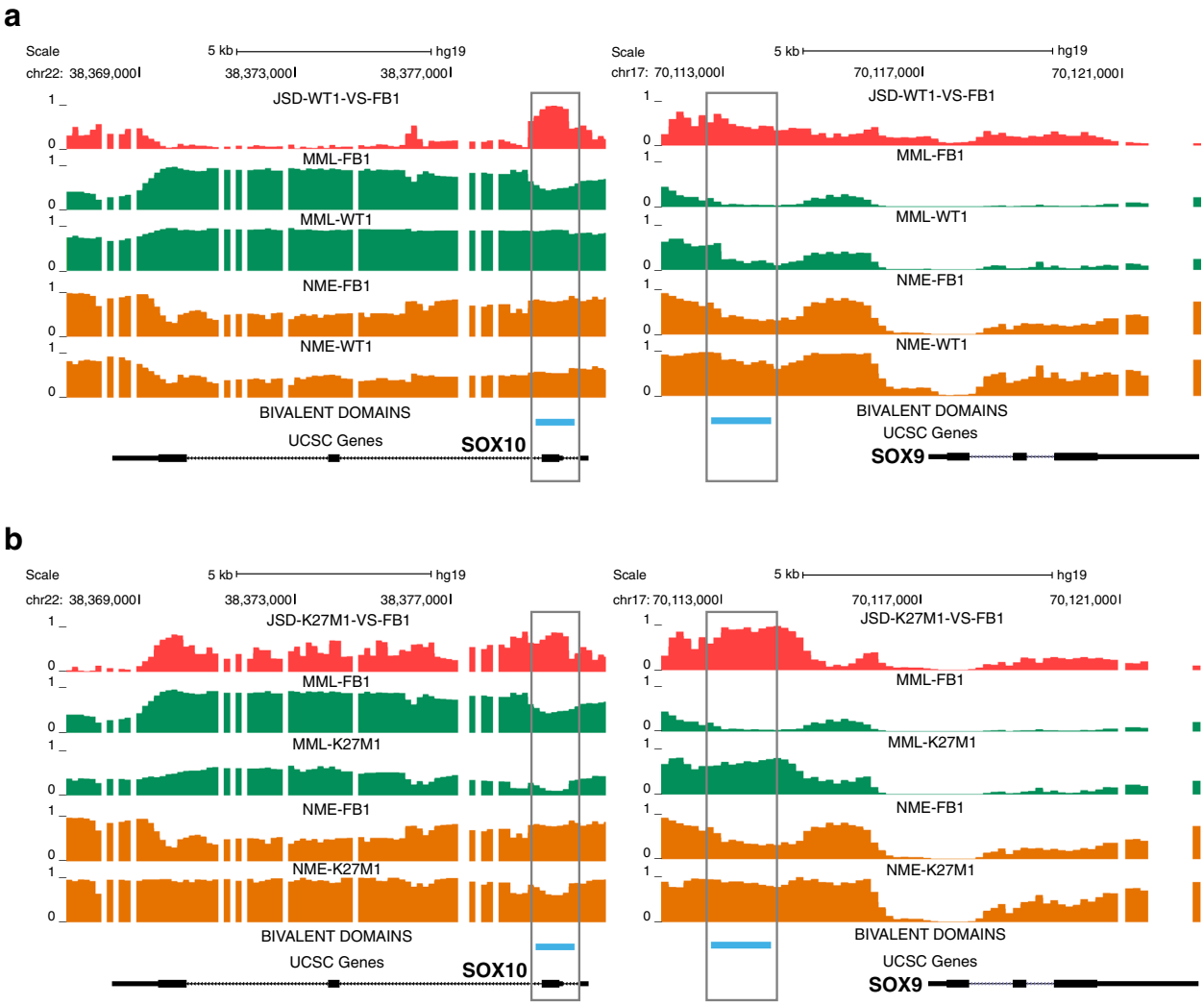

Fig. 5 UCSC genome browser images of JSD, MML, and NME tracks obtained by informME within genomic regions [chr22: 38,367,000–38,381,000] and [chr17: 70,111,800–70,122,800] containing *SOX10* (left) and *SOX9* (right). **a** An H3.3-WT vs. normal fetal brain comparison. **b** A K27M mutant vs. normal fetal brain comparison

H3.3-WT sample), which is now associated with appreciable hypermethylation in the K27M mutant sample (as compared to the small hypermethylation in the H3.3-WT sample), accompanied by a substantial gain in methylation entropy.

The behavior of *SOX10* (but not of *SOX9*) within its associated bivalent domain in the K27M mutant is in agreement with the previously discussed finding suggesting that the K27M mutant is subject to extensive hypomethylation. This observation, together with the fact that K27M mutant pHGG is characterized by a global reduction in the repressive H3K27me3 marks, raises the possibility that expression of *SOX10* is activated in the K27M mutant, a hypothesis that has been recently validated in [66].

Discussion

DNA methylation and its impact on cellular function has become a major focus of biomedical research. This is due to its role as a fundamental epigenetic mechanism in embryonic development and regulation of gene expression, as well as for being an important mediator between environmental risk factors and human diseases [2, 3, 67]. For this reason, several technologies have been developed to provide high throughput measurements of the DNA methylation state genome-wide, with WGBS being currently the best method. Computational analysis of methylation data obtained by WGBS allows extraction of epigenetic information that can be used, in conjunction with other biological data, to better understand the role of epigenetic regulation in health and disease [68].

WGBS data analyzed by software packages, such as informME [7, 8], present biologists with a large volume of epigenetic information that cannot be possibly reviewed and processed in a reasonable time, which poses a serious challenge when working with this type of data. It is therefore critical that, for efficient analysis, we develop computational tools that can direct attention to specific locations in the genome exhibiting statistically significant methylation discordances in test/reference comparisons. While the common practice of identifying DMRs can be helpful in this respect, they can also be difficult to interpret and analyze, especially in cases of widespread epigenomic changes, such as those typically seen in cancer.

From a statistical perspective, a more focused analysis within a set of genomic features of interest can ease some of the previous difficulties and provide higher statistical power. However, as we discussed earlier, methylation analysis is in general hampered by the presence of biological, statistical, and technical variability in the data, whereas a focused approach to this analysis is confounded by the variable length of most genomic features of interest as well as by missing data.

By using basic concepts from information theory and statistics, we have developed in this paper a new computational tool for methylation data analysis that effectively addresses the previous issues in a precise manner. This method passes a set of genomic features of interest to a rigorous statistical machinery that analyzes WGBS data obtained from a test/reference study and returns a list of ranked features together with corresponding p- and q-values.

An important practical question here is how to use our ranked lists for downstream WGBS data analysis. Our results provide methodological guidelines for the case of genes. We can begin with a manual investigation of the highest-ranking genes, which can provide familiarity with the results and leverage biological expertise. We can start at the top and move down the list gene-by-gene, allowing ourselves to adaptively decide how many genes we can afford to analyze using this labor-intensive strategy. However, we can also follow a more automated and unbiased strategy, for example using GO enrichment analysis based on a complete ranked list without having to specify an arbitrary threshold. In general, this should be considered a more preferable use of our ranked lists when comparing with methodologies that require separation of a given ranked list into two unranked categories of "target" and "background" genes. However, we did demonstrate in this paper with our GSEA results that using lists of target/background genes can still produce meaningful results and insights into the data, even when the decision on how many genes to include in the target set is challenging. Clearly, when a small or moderate number of genes are associated with significant q-values, these genes can be used to form a target set, but in the case of widespread epigenetic disruption some judgement and experience is required when selecting the number of genes that should be included in the target set. As a guideline, we have generally found that using around 500 genes in a target set strikes a good balance for GSEA analysis. If the target set is too large, there is too little power to detect enrichment, whereas if the list is made too small, then there may be important highly ranked genes (and thus enrichment categories) that will be missed by the analysis.

Finally, the data employed in this paper show that using logitSST regression for approximating the null PDF of our test statistic is superior to a number of alternative regression methods considered. Although we have also found this to be true in other WGBS studies, logitSST may not be the best regression method in general. To deal with the possibility that another regression method may be more preferable in a given study, a user may first consider a set of candidate regression methods, perform the type of "goodness of fit" and model selection analyses reported in this paper, and then replace logitSST with a better fitting model, if necessary.

Conclusions

In this paper, we presented a rigorous approach for computing statistical significance of methylation discordance within genomic features from WGBS data. Our approach uses mutual information, a powerful information-theoretic tool for evaluating dependance of the methylation state on the phenotype, in conjunction with a well-articulated hypothesis testing problem and logitSST regression for estimating the null distribution, to score genomic features in terms of observed methylation discordance in differential studies. We showed that the test statistic associated with this hypothesis testing problem can be evaluated by computing, using informME [7, 8], the JSD between probability distributions of the methylation state within a test and a reference sample. We also suggested effective ways for implementing hypothesis testing in a way that takes into account biological, statistical, and technical variability, as well as variability in feature length and missing data in single-sample and multiple-sample comparisons. This was accomplished by estimating, from available reference data, the null distribution of our test statistic via a novel application of heteroscedastic regression based on generalized additive regression models using penalized splines.

By reanalyzing previously published pHGG data using the proposed method, we obtained results that add credibility to our scheme when ranking genes in terms of their promoters, bodies, or bivalent domains, since these results clearly demonstrate the potential of the method for identifying genes that have been previously reported in the literature to be crucial in glioblastomas. Moreover, our

analysis shows that considering only methylation discordance within promoter regions of genes in a test/reference comparison may not be sufficient for identifying key genes, since methylation differences found within gene bodies and bivalent domains could be critical as well. We therefore believe that our approach can greatly facilitate the exploratory phases of clinically and biologically relevance methylation studies.

We should finally note that the idea of using regression models to build null distributions of test statistics from reference samples while accounting for confounding variables is quite general and can be applied to other problems of computational genomics. In addition to allowing a bioinformatician to handle confounding variables, such as feature length, our approach can naturally handle biological, statistical, and technical sources of variability, which are often overlooked by statistical tests based on theoretical/asymptotic sampling distributions. As such, we believe that the general statistical approach introduced in this paper can find widespread use for analyzing other types of whole-genome sequencing data, such as data obtained using chromatin immunoprecipitation sequencing (ChIP-seq) technologies.

Additional files

Additional file 1: Supplementary material. This file contains additional method descriptions and supplementary figures. (PDF 3994 kb)

Additional file 2: Supplementary tables. This file contains supplementary tables summarizing our results. (XLSX 17,856 kb)

Abbreviations

AIC: Akaike's information criterion; BIC: Bayesian information criterion; CDF: Cummulative distribution function; DMR: Differentially methylated region; FDR: False discovery rate; GB: gene body; GO: Gene ontology; GSEA: Gene set enrichment analysis; GU: Genomic unit; informME: Information-theoretic analysis of methylation; JSD: Jensen-Shannon distance; jsDMR: Jensen-Shannon distance based differentially methylation region; logitSST: Logit skewed Student's *t* distribution; MML: Mean methylation level; MSigDB: Molecular signatures database; NME: Normalized methylation entropy; NPC: Neural progenitor cell; PB: Promoter/body; PDF: Probability density function; PMF: Probability mass function; pHGG: Pediatric high-grade glioma; PR: Promoter region; TSS: Transcription start site; WGBS: Whole genome bisulfite sequencing

Acknowledgements

Not Applicable.

Funding

This work was supported by NIH Grants CA054358, HG008529 and NSF Grant CCF-1656201. MAK is a St. Baldrick's Foundation Fellow and a Damon Runyon-Sohn Pediatric Cancer Fellow supported by the Damon Runyon Cancer Research Foundation (DRSG-15P-16). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Availability of data and materials

Source code is provided in three R files: jsGrank.R, runGrank_PromsBods.R, and runGrank_Bivalent.R. The file jsGrank.R is distributed under a GPLv3 license as part of the informME software package and can be downloaded from https://github.com/GarrettJenkinson/informME. The files runGrank_PromsBods.R and runGrank_Bivalent.R, the results of reanalyzing the pHGG data using informME in bigWig format, and the respective "track" files for visualization in the UCSC genome browser, can be downloaded using the following link: http://www.cis.jhu.edu/$\sim$goutsias/data/19-BMCbio.zip.

Authors' contributions

GJ, JA, and JG developed the statistical and computational methods. GJ and JA wrote the computer code and implemented the methods. MAK processed the WGBS data and ran the informME pipeline on APF's cluster allocation. GJ, MAK, JA, APF, and JG interpreted the results. GJ, JA and JG wrote the manuscript with the assistance of MAK and APF. All authors read and approved the final version of the manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Author details

[1]Whitaker Biomedical Engineering Institute, Johns Hopkins University, Baltimore, MD, USA. [2]Center for Epigenetics, Johns Hopkins School of Medicine, Baltimore, MD, USA. [3]Currently with Department of Health Sciences Research, Mayo Clinic, Rochester, MN, USA. [4]Pediatric Oncology, Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University School of Medicine, Baltimore, MD, USA. [5]Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD, USA. [6]Department of Medicine, Johns Hopkins School of Medicine, Baltimore, MD, USA.

Received: 14 September 2018 Accepted: 25 March 2019
Published online: 08 April 2019

References

1. Rakyan VK, Down TA, Balding DJ, Beck S. Epigenome-wide association studies for common human diseases. Nat Rev Genet. 2011;12(8):529–41.
2. Smith ZD, Meissner A. DNA methylation: roles in mammalian development. Nat Rev Genet. 2013;14(3):204–20.
3. Schübeler D. Function and information content of DNA methylation. Nature. 2015;517:321–6.
4. Feinberg AP. The key role of epigenetics in human disease prevention and mitigation. N Engl J Med. 2018;378(14):1323–34.
5. Bock C. Analysing and interpreting DNA methylation data. Nat Rev Genet. 2012;13:705–19.
6. Robinson MD, Kahraman A, Law CW, Lindsay H, Nowicka M, Weber LM, Zhou X. Statistical methods for detecting differentially methylated loci and regions. Front Genet. 2014;5:324.
7. Jenkinson G, Pujadas E, Goutsias J, Feinberg AP. Potential energy landscapes identify the information-theoretic nature of the epigenome. Nat Genet. 2017;49:719–29.
8. Jenkinson G, Abante J, Feinberg AP, Goutsias J. An information-theoretic approach to the modeling and analysis of whole-genome bisulfite sequencing data. BMC Bioinformatics. 2018;19:87.
9. Wu H, Xu T, Feng H, Chen L, Li B, Yao B, Qin Z, Jin P, Conneely KN. Detection of differentially methylated regions from whole-genome bisulfite seqeuencing data without repicates. Nucl Acids Res. 2015;33:141.
10. Park Y, Wu H. Differential methylation analysis for BS-seq data under general experimental design. Bioinformatics. 2016;32:1446–53.
11. Yang X, Han H, De Carvalho DD, Lay FD, Jones PA, Liang G. Gene body methylation can alter gene expression and is a therapeutic target in cancer. Cancer Cell. 2014;26:577–90.
12. Neri F, Rapelli S, Krepelova A, Incarnato D, Parlato C, Basile G, Maldotti M, Anselmi F, Oliviero S. Intragenic DNA methylation prevents spurious transcription initiation. Nature. 2017;543:72–7.
13. Fisher RA. Statistical Methods, Experimental Design, and Statistical Inference. 2nd ed. Oxford, England: Oxford University Press; 1990.

14. Voigt P, Tee W-W, Reinberg D. A double take on bivalent promoters. Gene Dev. 2013;27:1318–38.
15. Cover TM, Thomas JA. Elements of Information Theory. 2nd ed. New York: Wiley; 1991.
16. Lin J. Divergence measures based on the Shannon entropy. IEEE Trans Inform Theory. 1991;37:145–51.
17. Hansen BE. Autoregressive conditional density estimation. Int Econ Rev. 1994;35:705–30.
18. Stasinopoulos DM, Rigby R. Generalized additive models for location scale and shape (GAMLSS) in R. J Stat Softw. 2007;23:1–46.
19. Hossain A, Rigby R, Stasinopoulos M, Enea M. Centile estimation for a proportion response variable. Stat Med. 2016;35:895–904.
20. Rigby R, Stasinopoulos DM. Generalized additive models for location, scale and shape. Appl Statist. 2005;54:507–54.
21. Breitling R, Armengaud P, Amtmann A, Herzyk P. Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. FEBS Lett. 2004;573:83–92.
22. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Statist Soc B. 1995;57: 289–300.
23. Bender S, Tang Y, Lindroth AM, Hovestadt V, Jones DTW, Kool M, Zapatka M, Northcott PA, Sturm D, Wang W, Radlwimmer B, Hojfeldt JW, Truffaux N, Castel D, Schubert S, Ryzhova M, Seker-Cin H, Gronych J, Johann PD, Stark S, Meyer J, Milde T, Schuhmann M, Ebinger M, Monoranu C-M, Ponnuswami A, Chen S, Jones C, Witt O, Collins VP, von Deimling A, Jabado N, Puget S, Grill J, Helin K, Korshunov A, Lichter P, Monje M, Plass C, Cho Y-J, Pfister SM. Reduced H3K27me3 and DNA hypomethylation are major drivers of gene expression in K27M mutant pediatric high-grade gliomas. Cancer Cell. 2013;24(5):660–72.
24. Hovestadt V, Jones DTW, Picelli S, Wang W, Kool M, Northcott PA, Sultan M, Stachurski K, Ryzhova M, Warnatz H-J, Ralser M, Brun S, Bunt J, Jäger N, Kleinheinz K, Erkek S, Weber UD, Bartholomae CC, von Kalle C, Lawerenz C, Eils J, Koster J, Versteeg R, Milde T, Witt O, Schmidt S, Wolf S, Pietsch T, Rutkowski S, Scheurlen W, Taylor MD, Brors B, Felsberg J, Reifenberger G, Borkhardt A, Lehrach H, Wechsler-Reya RJ, Eils R, Yaspo M-L, Landgraf P, Korshunov A, Zapatka M, Radlwimmer B, Pfister SM, Lichter P. Decoding the regulatory landscape of medulloblastoma using DNA methylation sequencing. Nature. 2014;510:537–41.
25. Dunn PK, Smyth GK. Randomized quantile residuals. J Comput Graph Stat. 1996;5:236–44.
26. Eden E, Navon R, Steinfeld I, Lipson D, Yakhini Z. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. BMC Bioinformatics. 2009;10:48.
27. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub T, Lander ES, Mesirov JP. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci. 2005;102(43): 15545–50.
28. Shinawi T, Hill VK, Krex D, Schackert G, Gentle D, Morris MR, Wei W, Cruickshank G, Maher ER, Latif F. DNA methylation profiles of long- and short-term glioblastoma survivors. Epigenetics. 2013;8(2):149–56.
29. Bhargava S, Patil V, Mahalingam K, Somasundaram K. Elucidation of the genetic and epigenetic landscape alterations in RNA binding proteins in glioblastoma. Oncotarget. 2017;8(10):16650–68.
30. Dreyfuss JM, Johnson MD, Park PJ. Meta-analysis of glioblastoma multiforme versus anaplastic astrocytoma identifies robust gene markers. Mol Cancer. 2009;8:71.
31. Khan FH, Pandian V, Ramraj S, Natarajan M, Aravindan S, Herman TS, Aravindan N. Acquired genetic alterations in tumor cells dictate the development of high-risk neuroblastoma and clinical outcomes. BMC Cancer. 2015;15:514.
32. Li CCY, Eaton SA, Young E, Lee M, Shuttleworth R, Humphreys DT, Grau GE, Combes V, Bebawy M, Gong J, Brammah S, Buckland ME, Suter CM. Glioma microvesicles carry selectively packaged coding and non-coding RNAs which alter gene expression in recipient cells. RNA Biol. 2013;10(8): 1333–44.
33. Day BW, Stringer BW, Al-Ejeh F, Ting MJ, Wilson J, Ensbey KS, Jamieson PR, Bruce ZC, Lim YC, Offenhäuser C, Charmsaz S, Cooper LT, Ellacott JK, Harding A, Leveque L, Inglis P, Allan S, Walker DG, Lackmann M, Osborne G, Khanna KK, Reynolds BA, Lickliter JD, Boyd AW. EphA3 maintains tumorigenicity and is a therapeutic target in glioblastoma multiforme. Cancer Cell. 2013;23(2):238–48.
34. Uhlmann K, Rohde K, Zeller C, Szymas J, Vogel S, Marczinek K, Thiel G, Nürnberg P., Laird PW. Distinct methylation profiles of glioma subtypes. Int J Cancer. 2003;106(1):52–9.
35. Ping Y, Deng Y, Wang L, Zhang H, Zhang Y, Xu C, Zhao H, Fan H, Yu F, Xiao Y, Li X. Identifying core gene modules in glioblastoma based on multilayer factor-mediated dysfunctional regulatory networks through integrating multi-dimensional genomic data. Nucleic Acids Res. 2015;43(4):1997–2007.
36. Bax DA, Little SE, Gaspar N, Perryman L, Marshall L, Viana-Pereira M, Jones TA, Williams RD, Grigoriadis A, Vassal G, Workman P, Sheer D, Reis RM, Pearson ADJ, Hargrave D, Jones C. Molecular and phenotypic characterisation of paediatric glioma cell lines as models for preclinical drug development. PLoS ONE. 2009;4(4):5209.
37. Emara M, Turner AR, Allalunis-Turner J. Adult, embryonic and fetal hemoglobin are expressed in human glioblastoma cells. Int J Oncol. 2014;44(2):514–20.
38. Doan NB, Alhajala H, Al-Gizawiy MM, Mueller WM, Rand SD, Connelly JM, Cochran EJ, Chitambar CR, Clark P, Kuo J, Schmainda KM, Mirza SP. Acid ceramidase and its inhibitors: a *de novo* drug target and a new class of drugs for killing glioblastoma cancer stem cells with high efficiency. Oncotarget. 2017;8(68):112662–74.
39. Lin B, Lee H, Yoon J-G, Madan A, Wayner E, Tonning S, Hothi P, Schroeder B, Ulasov I, Foltz G, Hood L, Cobbs C. Global analysis of H3K4me3 and H3K27me3 profiles in glioblastoma stem cells and identification of SLC17A7 as a bivalent tumor suppressor gene. Oncotarget. 2015;6(7):5369–81.
40. Men C, Chai H, Song X, Li Y, Du H, Ren Q. Identification of DNA methylation associated gene signatures in endometrial cancer via integrated analysis of DNA methylation and gene expression systematically. J Gynecol Oncol. 2017;28(6):83.
41. Pongor L, Kormos M, Hatzis C, Pusztai L, Szabo A, Gyorffy B. A genome-wide approach to link genotype to clinical outcome by utilizing next generation sequencing and gene chip data of 6697 breast cancer patients. Genome Med. 2015;7:104.
42. Xu K, Qiu C, Pei H, Mehmood M, Wang H, Li L, Xia Q. Homeobox B3 promotes tumor cell proliferation and invasion in glioblastoma. Oncol Lett. 2018;15(3):3712–8.
43. Costa BM, Smith JS, Chen Y, Chen J, Phillips HS, Aldape KD, Zardo G, Nigro J, James CD, Fridlyand J, Reis RM, Costello JF. Reversing HOXA9 oncogene activation by PI3K inhibition: epigenetic mechanism and prognostic significance in human glioblastoma. Cancer Res. 2010;70(2): 453–62.
44. 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(18):3015–24.
45. Margueron R, Reinberg D. The Polycomb complex PRC2 and its mark in life. Nature. 2011;469(7330):343–9.
46. Brinkman AB, Gu H, Bartels SJ, Zhang Y, Matarese F, Simmer F, Marks H, Bock C, Gnirke A, Meissner A, Stunnenberg HG. Sequential ChIP-bisulfite sequencing enables direct genome-scale investigation of chromatin and DNA methylation cross-talk. Genome Res. 2012;22(6):1128–38.
47. Boyd AW, Bartlett PF, Lackmann M. Therapeutic targeting of EPH receptors and their ligands. Nat Rev Drug Discov. 2013;13(1):39–62.
48. Nagaraja S, Vitanza NA, Woo PJ, Taylor KR, Liu F, Zhang L, Li M, Meng W, Ponnuswami A, Sun W, Ma J, Hulleman E, Swigut T, Wysocka J, Tang Y, Monje M. Transcriptional dependencies in diffuse intrinsic pontine glioma. Cancer Cell. 2017;31(5):635–52.
49. Rheinbay E, Suvà ML, Gillespie SM, Wakimoto H, Patel AP, Shahid M, Oksuz O, Rabkin SD, Martuza RL, Rivera MN, Louis DN, Kasif S, Chi AS, Bernstein BE. An aberrant transcription factor network essential for Wnt signaling and stem cell maintenance in glioblastoma. Cell Rep. 2013;3: 1567–79.
50. Zong Z, Pang H, Yu R, Jiao Y. PCDH8 inhibits glioma cell proliferation by negatively regulating the AKT/GSK3β/β-catenin signaling pathway. Oncol Lett. 2017;14(3):3357–62.
51. Mock A, Geisenberger C, Orlik C, Warta R, Schwager C, Jungk C, Dutruel C, Geiselhart L, Weichenhan D, Zucknick M, Nied AK, Friauf S, Exner J, Capper D, Hartmann C, Lahrmann B, Grabe N, Debus J, von Deimling A,

Popanda O, Plass C, Unterberg A, Abdollahi A, Schmezer P, Herold-Mende C. LOC283731 promoter hypermethylation prognosticates survival after radiochemotherapy in IDH1 wild-type glioblastoma patients. Int J Cancer. 2016;139(2):424–32.
52. O'Loghlen A, Martin N, Krusche B, Pemberton H, Alonso MM, Chandler H, Brookes S, Parrinello S, Peters G, Gil J. The nuclear receptor NR2E1/TLX controls senescence. Oncogene. 2015;34:4069–77.
53. Qi XW, Zhang F, Wu H, Liu JL, Zong BG, Xu C, Jiang J. Wilms' tumor 1 (WT1) expression and prognosis in solid cancer patients: a systematic review and meta-analysis. Sci Rep. 2015;5:8924.
54. Crespo I, Tão H, Nieto AB, Rebelo O, Domingues P, Vital AL, Patino M, Barbosa MCML, Oliveira CR, Orfao A, Tabernero MD. Amplified and homozygously deleted genes in glioblastoma: Impact on gene expression levels. PLoS ONE. 2012;7(9):46088.
55. Liu Y, Hu H, Wang K, Zhang C, Wang Y, Yao K, Yang P, Han L, Kang C, Zhang W, Jiang T. Multidimensional analysis of gene expression reveals TGFB1I1-induced EMT contributes to malignant progression of astrocytomas. Oncotarget. 2014;5:12593–606.
56. Syed P, Gupta S, Choudhary S, Pandala NG, Atak A, Richharia A, Manubhai KP, Zhu H, Epari S, Noronha SB, Moiyadi A, Srivastavaa S. Autoantibody profiling of glioma serum samples to identify biomarkers using human proteome arrays. Sci Rep. 2015;5:13895.
57. Ray SK. The transcription regulator krüppel-like factor 4 and its dual roles of oncogene in glioblastoma and tumor suppressor in neuroblastoma. Immunopathol Dis Therap. 2016;7(1-2):127–39.
58. Etcheverry A, Aubry M, de Tayrac M, Vauleon E, Boniface R, Guenot F, Saikali S, Hamlat A, Riffaud L, Menei P, Quillien V, Mosser J. DNA methylation in glioblastoma: impact on gene expression and clinical outcome. BMC Genomics. 2010;11:701.
59. Acanda de la Rocha AM, López-Bertoni H, Guruceaga E, González-Huarriz M, Martínez-Vélez N, Xipell E, Fueyo J, Gomez-Manzano C, Alonso MM. Analysis of SOX2-regulated transcriptome in glioma stem cells. PLoS ONE. 2016;11(9):0163155.
60. Yoshino A, Ogino A, Yachi K, Ohta T, Fukushima T, Watanabe T, Katayama Y, Okamoto Y, Naruse N, Sano E, Tsumoto K. Gene expression profiling predicts response to temozolomide in malignant gliomas. Int J Oncol. 2010;36(6):1367–77.
61. Karnati HK, Panigrahi M, Shaik NA, Greig NH, Bagadi S, Kamal MA, Kapalavayi N. Down regulated expression of Claudin-1 and Claudin-5 and up regulation of β-catenin: association with human glioma progression. CNS Neurol Disord Drug Targets. 2014;13(8):1413–26.
62. Wang L, He S, Yuan J, Mao X, Cao Y, Zong J, Tu Y, Zhang Y. Oncogenic role of SOX9 expression in human malignant glioma. Med Oncol. 2012;29(5):3484–90.
63. Filbin MG, Tirosh I, Hovestadt V, Shaw ML, Escalante L, Mathewson ND, Neftel C, Frank N, Pelton K, Hebert CM, Haberler C, Yizhak K, Gojo J, Egervari K, Mount C, van Galen P, Bonal DM, Nguyen QD, Beck A, Sinai C, Czech T, Dorfer C, Goumnerova L, Lavarino C, Carcaboso AM, Mora J, Mylvaganam R, Luo CC, Peyrl A, Popović M, Azizi A, Batchelor TT, Frosch MP, Martinez-Lage M, Kieran MW, Bandopadhayay P, Beroukhim R, Fritsch G, Getz G, Rozenblatt-Rosen O, Wucherpfennig KW, Louis DN, Monje M, Slavc I, Ligon KL, Golub T, Regev A, Bernstein BE, Suvà ML. Developmental and oncogenic programs in H3K27M gliomas dissected by single-cell RNA-seq. Science. 2018;360(6386):331–5.
64. Gao J, Zhang JY, Li YH, Ren F. Decreased expression of SOX9 indicates a better prognosis and inhibits the growth of glioma cells by inducing cell cycle arrest. Int J Clin Exp Pathol. 2015;8(9):10130–8.
65. Wang Z, Xu X, Liu N, Cheng Y, Jin W, Zhang P, Wang X, Yang H, Liu H, Tu Y. SOX9-PDK1 axis is essential for glioma stem cell self-renewal and temozolomide resistance. Oncotarget. 2017;9(1):192–204.
66. Nomura M, Mukasa A, Nagae G, Yamamoto S, Tatsuno K, Ueda H, Fukuda S, Umeda T, Suzuki T, Otani R, Kobayashi K, Maruyama T, Tanaka S, Takayanagi S, Nejo T, Takahashi S, Ichimura K, Nakamura T, Muragaki Y, Narita Y, Nagane M, Ueki K, Nishikawa R, Shibahara J, Aburatani H, Saito N. Distinct molecular profile of diffuse cerebellar gliomas. Acta Neuropathol. 2017;134(6):941–56.
67. Bergman Y, Cedar H. DNA methylation dynamics in health and disease. Nat Struct Mol Biol. 2013;20:274–81.
68. Teschendorff AE, Relton CL. Statistical and integrative system-level analysis of DNA methylation data. Nat Rev Genet. 2018;19(3):129–47.