SCIENTIFIC REPORTS
nature research

OPEN

MeinteR: A framework to prioritize DNA methylation aberrations based on conformational and cis-regulatory element enrichment

Andigoni Malousi[1*], Sofia Kouidou[1], Maria Tsagiopoulou[2], Nikos Papakonstantinou[2], Emmanouil Bouras[3], Elisavet Georgiou[1], Georgios Tzimagiorgis[1] & Kostas Stamatopoulos[2]

DNA methylation studies have been reformed with the advent of single-base resolution arrays and bisulfite sequencing methods, enabling deeper investigation of methylation-mediated mechanisms. In addition to these advancements, numerous bioinformatics tools address important computational challenges, covering DNA methylation calling up to multi-modal interpretative analyses. However, contrary to the analytical frameworks that detect driver mutational signatures, the identification of putatively actionable epigenetic events remains an unmet need. The present work describes a novel computational framework, called MeinteR, that prioritizes critical DNA methylation events based on the following hypothesis: critical aberrations of DNA methylation more likely occur on a genomic substrate that is enriched in cis-acting regulatory elements with distinct structural characteristics, rather than in genomic "deserts". In this context, the framework incorporates functional cis-elements, e.g. transcription factor binding sites, tentative splice sites, as well as conformational features, such as G-quadruplexes and palindromes, to identify critical epigenetic aberrations with potential implications on transcriptional regulation. The evaluation on multiple, public cancer datasets revealed significant associations between the highest-ranking loci with gene expression and known driver genes, enabling for the first time the computational identification of high impact epigenetic changes based on high-throughput DNA methylation data.

Basic and applied research on DNA methylation (DNAm) have been revolutionized with the advent of single-base resolution and genome-wide assays. Along with these advancements, a wide range of bioinformatics methods has been developed to address the computational complexities of high-throughput analyses. These methods are generally classified in two categories: (a) those focusing on the core analysis pipeline that transforms raw array-based[1,2] or sequencing data to DNAm calls[3,4], and (b) those implementing downstream analyses[5–11], e.g. cell mixture proportions, age calculators, differential analysis, visualization, association with gene expression and phenotypic data, pathway enrichment analyses, genomic architecture mappings etc. Furthermore, several frameworks provide comprehensive solutions by either integrating existing tools from both categories in user-friendly pipelines[12–16], or by interpreting DNAm events with respect to the enrichment of diverse types of colocalized regulatory elements[11].

While significant progress has been made on improving DNAm calling methods and enriching the types of interpretative analyses[17], the need to computationally identify the most critical aberrations is still poorly addressed. The main origins of this deficiency are: (a) the plethora of differentially methylated sites (DMS) that are usually identified from high-throughput experiments[18]; (b) the dynamics and tissue specificity of DNAm events; (c) modest interpretability, as aberrant DNAm is usually observed in poorly annotated, non-coding regions[18]; and (d) unlike genomic studies, lack of efficient analytical frameworks that detect critical events[19]. To this end, a computational framework that could address the above issues would constitute a significant advancement in interpretative DNAm analyses.

[1]Lab. of Biological Chemistry, School of Medicine, Aristotle University of Thessaloniki, Thessaloniki, Greece. [2]Institute of Applied Biosciences, Centre for Research and Technology Hellas, Thessaloniki, Greece. [3]Lab. of Hygiene, Social-Preventive Medicine & Medical Statistics, School of Medicine, Aristotle University of Thessaloniki, Thessaloniki, Greece. *email: andigoni@auth.gr

In particular, recent studies highlight the benefits of encompassing DNAm data in computational frameworks that deal with driver event detection. For example, a beta mixture model was proposed for the detection of important methylation-driven genes in cancer by integrating methylome and gene expression data[20,21]. A similar data- driven pathway method identified pan-cancer genes by integrating DNAm, copy number variation and gene expression data[22]. In the same context, a functional interaction network developed to prioritize cancer genes from multi-omics data, including DNAm[23]. The above methods share certain common features: (a) driving events do not solely derive from DNAm profiles; (b) aberrant DNAm is inferred at gene level by averaging DNAm levels at promoters, intronic and exonic regions; and (c) pathway enrichment from each -omics modality is used to classify driver genes.

In addition, these studies do not encounter an important regulator of epigenetic modifications, that is the genomic substrate underlying driver events. The contribution of specific positional (e.g. cis-acting regulatory elements) and compositional features (e.g. CpG islands, k-mers) in DNAm has been highlighted in multiple research studies. For example, a set of cis-acting, methylation-prone and methylation-resistant motifs were identified that increase the predictive power of the DNAm detection methods[24]. Other studies elaborated further on: (a) the role of context-dependent DNAm as instructor for gene regulation[25], (b) associated DNAm with the presence of ENCODE's regulatory elements[26], and (c) developed computational tools that accurately predict DNAm levels, based on context-based features[27–29].

The role of the genomic context is also supported by several studies associating the presence of particular regulatory elements with DNAm. For example, dual-specificity of transcription factors is related with variable binding affinity in methylated and unmethylated forms of a CpG sequence[30,31]. As transcription factor binding sites are identified by position-specific dependencies among nucleotides it is important to specify whether putative bindings are potentially inhibited or promoted when co-localized with methylated CpG sites (CpGs), implying an indirect role of DNAm in regulatory processes[32]. In the same context, DNA sequences that fold into G-quadruplexes were found to be less prone to CpG methylation, while increased DNAm is depleted in these structures[33], and might change protein binding to quadruplex-forming DNA segments during transcriptional regulation, particularly in aging[34,35].

Although less clearly demonstrated, non-canonical hairpin structures formed by palindromic sequences could potentially regulate methylation-mediated processes by becoming resistant to DNAm[36]. Furthermore, short sequences neighboring methylated cytosines in palindromic sequences were found to attract protein-DNA binding[37]. Other DNA helix elements, such as local geometric features (minor/major groove, propeller twist etc.), alter their shapes in the presence of DNAm, and could probably affect protein-DNA binding, subsequently leading to transcriptional activation or silencing[38]. In this context, computational methods could corroborate or deputize limited experimental data by predicting local structural changes of the double helix induced by DNAm that may successively alter protein-DNA binding affinity[39]. DNAm also exhibits distinguishable positional patterns in constitutive and alternative splicing[40–43]. CpGs that are located on the exon-intron junctions exhibit increased DNAm levels, contrary to the neighboring intronic regions[44,45]. Considering these complexities, further analyzed by Machado *et al.*[46], it is evident that the genomic substrate could provide valuable insights in deciphering the impact of aberrant DNAm events.

Herein, we present a computational framework, called MeinteR, that identifies putatively functional DNAm sites based on the following hypothesis: aberrant DNAm that occurs on a genomic substrate enriched in cis-regulatory and conformational elements is more likely to trigger methylation-mediated transcriptional events than differential DNAm observed in genomic "deserts". In this context, MeinteR builds genomic signatures of DMS and identifies critical loci where aberrant DNAm might have a greater effect on phenotype expression, using a linear function of the elements enrichment, called genomic index. With three use cases and extensive comparisons, we show that MeinteR provides an efficient means to decode complex associations between DNAm aberrations and gene expression deregulation.

Results

MeinteR is a computational framework consisting of three modules (Fig. 1). Briefly, the *data preprocessing* module contains functions for loading, validation, reformatting and filtering of DNAm data that are exported from BeadChip arrays or next-generation sequencing platforms. The *feature detection* module implements a set of functions for batch sequence retrieval and enrichment analysis of the incorporated features. Finally, the *signature extraction* module builds genomic signatures of the candidate sites and implements a ranking scheme. The functionality of each module is described in Methods.

MeinteR is a software package that primarily builds genomic signatures of epigenetic aberrations and identifies critical events in high-throughput datasets. The modularity of the software components enables a wide variety of applications in multiple settings as shown in the following use cases. Use case 1 shows a differential analysis of tumor/normal samples using only G-quadruplexes, and use case 2 builds genomic signatures of tumor/normal samples using the complete set of conformational and cis-regulatory elements. Use case 3 implements the primary goal of the framework that is to export the genomic index of aberrantly methylated sites and to associate the genomic index with differential gene expression. In these use cases, DNAm data are retrieved from Gene Expression Omnibus (GEO)[47] and The Cancer Genome Atlas (TCGA)[48] and for most differential analyses we applied a stringent threshold ($\Delta\beta >= 0.3$, $p < 0.01$ and FDR < 0.01) to avoid the detection of false positive DMS sites and improve the quality of downstream analysis towards biological interpretation of the results. Notably, the objective of these use cases is to provide pre-configured examples on public datasets, rather than to interpret the biological findings. These use cases can be easily adapted to other research applications using custom configurations as regards to the composition of the feature set and weighting scheme. MeinteR provides supplementary functions to further annotate the input data in terms of the CpG/G + C content and β value distribution. These

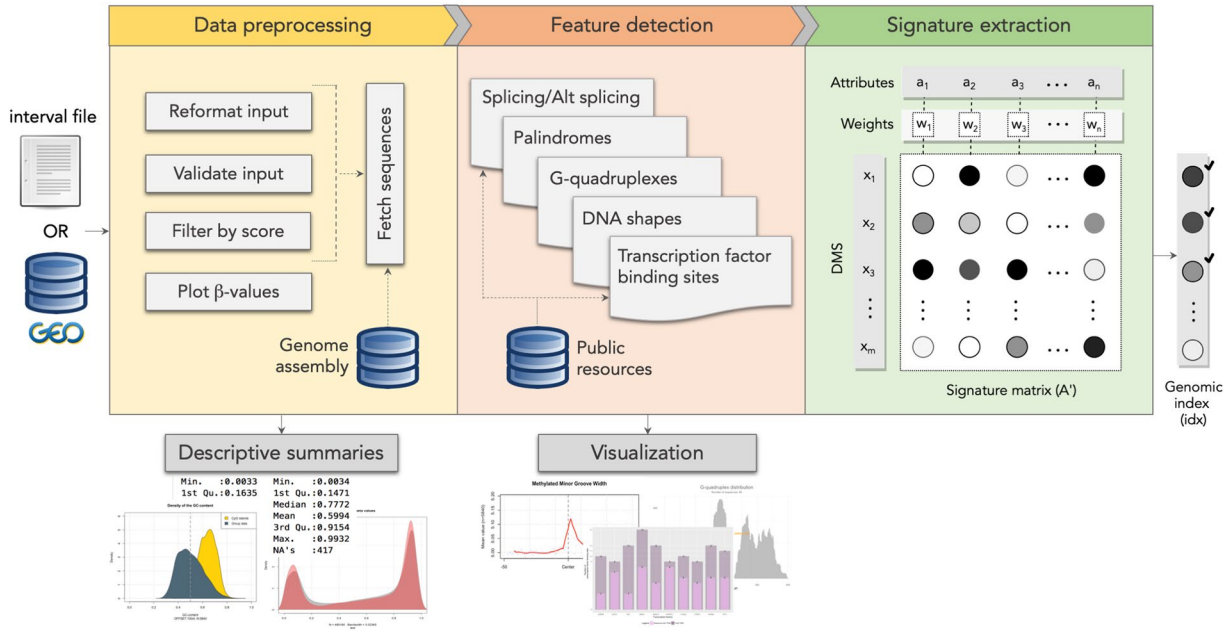


Figure 1. Overview of the MeinteR workflow. Input data, either interval files or GEO data series, are pre-processed and genomic sequences are obtained from human genome assembly (*data preprocessing* panel). Plotting functions and summary statistics supplement the data preprocessing module. Fetched DNA sequences are then analysed with respect to the abundance of the incorporated features through the corresponding MeinteR functions that export tabular and graphical outputs (*feature detection* panel). The identification of transcription factor binding sites and splicing elements is based on reference data that are automatically retrieved from relevant public resources. The last module (*signature extraction* panel) builds a matrix of genomic signatures per CpG site and the genomic index is calculated based on user-defined weighting schemes. The genomic signature of each DNAm is a numerical vector containing the abundance of each feature, multiplied by a user-defined weighting factor. The genomic index is a non-negative real number that is calculated using the linear mixture of the values in the signature vector.

functions, coupled with various filtering parameters embedded in each core function, can be used towards comprehensive and accurate characterization of the input data.

Use Case 1: Genome-wide association of G-quadruplexes with DNAm using public breast cancer datasets. To demonstrate the applicability in revealing associations between DNAm and particular genomic features, we used MeinteR to investigate DNAm resistance in sequences that fold in G-quadruplex structures[33]. First, we downloaded TCGA HumanMethylation450 array data from 91 breast cancer patients with matched primary tumor and normal tissue samples[47]. Then, we calculated the mean DNAm levels per sample group using the beta (β) values of the interrogated CpGs. β values range from 0 to 1 and are calculated by the formula $\beta =$ Intensity of the methylated probe/(Intensity of the unmethylated probe + Intensity of the methylated probe + 100). For each sample group, we identified G-quadruplex structures in sequences centered at unmethylated and methylated sites, with $\beta \leq 0.1$ and $\beta \geq 0.9$, respectively. Batch analysis of normal samples revealed a statistically significant two-fold increase (two-tailed t-test, $p < 0.001$) of G-quadruplex frequency at unmethylated sites compared to methylated sites (Fig. 2A, left-hand side). Similarly, as shown in Fig. 2A (right-hand side), in primary breast tumor samples G-quadruplex-forming sequences are less frequently observed in regions neighboring highly methylated sites (two-tailed t-test, $p < 0.001$).

We further evaluated the propensity of G-quadruplex structures to co-localize with sites that significantly lower their DNAm level in cancer (hypomethylated, DMS^-) and vice versa, sites that exhibit a significant decrease in their DNAm level in cancer (hypermethylated, DMS^+) with ($|\Delta\beta| \geq 0.3$, $p < 0.01$ and FDR < 0.01). Differential analysis of the cancer/normal breast pairs identified 3,981 DMS^+, and 1,869 DMS^-. Both datasets were randomly subsampled to 1,000 DMS and scanned for putative G-quadruplexes. Figure 2B on the left side shows a three-fold increase of G-quadruplex frequency in DMS^+ compared with DMS^- (two-tailed t-test, $p < 0.001$). This observation is inline with the results shown in Fig. 2A, since DMS^+ mostly involves low methylated sites in normal samples due to the bimodal distribution of the β values and vice versa.

To validate these results, we followed the same procedure on HumanMethylation450 DNAm data obtained from 80 breast cancer and 40 normal samples that are deposited to GEO (GSE66695 data series). β values were averaged on each sample group and differentially analyzed, resulting in 293 DMS^+ and 62 DMS^- ($|\Delta\beta| \geq 0.3$, $p < 0.01$, FDR < 0.01). Then, we estimated the frequency of G-quadruplex forming sequences in 100nt regions centered at DMS. The histogram in Fig. 2B (right-hand side) shows that most DMS^- sequences lack G-quadruplex structures (mean G4: 0.597, s.d. = 0.878), while most DMS^+ are co-localized with at least one G-quadruplex structure (mean G4: 1.901, s.d. = 1.462). The difference is statistically significant (two-tailed t-test, $p < 0.001$) and

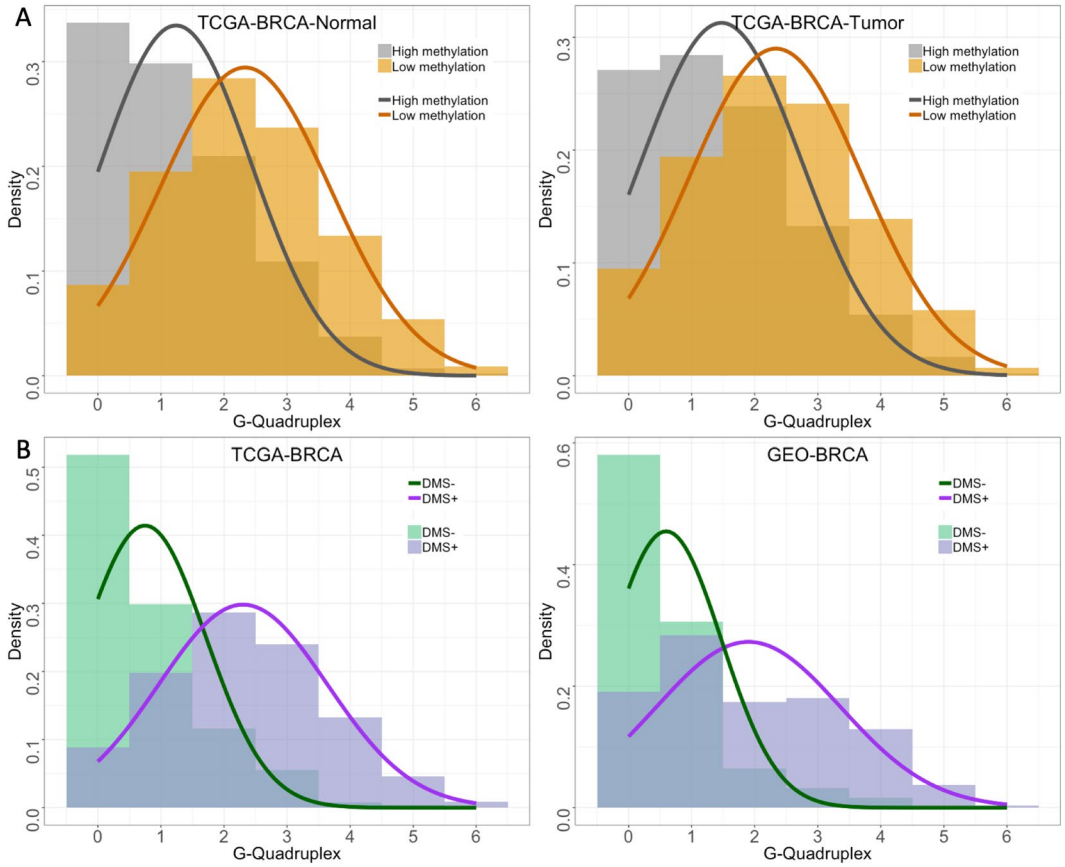


Figure 2. Histograms of the G-quadruplex density in breast cancer. (**A**) Comparison of low and highly methylated sites in matched normal tissue (left) and primary breast tumor (right) retrieved from TCGA. (**B**) Comparison of hypermethylated (DMS^+) and hypomethylated (DMS^-) sites in TCGA and GEO breast cancer samples (left and right panel, respectively). The curves correspond to normal distributions projected over the histograms.

in accordance with the outcome of the TCGA data analysis (Fig. 2B, left). Overall, the results of both breast cancer datasets dictate a protective role of G-quadruplex structures against DNAm that apparently characterizes DNAm patterns independently of the cell type and phenotype.

To further validate these findings, we performed the same analysis on six other cancer types using TCGA datasets. Fig. S1 (Supplementary File 1) shows the results that further validate the same findings. Overall, the results dictate a protective role of G-quadruplex structures against DNAm that apparently characterizes DNAm patterns independently of the cell type and phenotype. This observation needs further elaboration that is beyond the scope of the present work.

Use Case 2: Evaluation of genomic signatures on cancer DNAm profiles. In this use case, we determined the enrichment of the incorporated conformational and regulatory elements, in order to assess the contribution of the genomic substrate to DNAm alterations in cancer. To this end, we built the genomic signatures of nine cancer types using DNAm datasets deposited to GEO[48]. Each cancer dataset contains DNAm data from tumor and normal samples. Table 1 lists the mean values, standard deviations and statistical evaluation of each feature in the hypo- and hypermethylated subsets. In addition, each subset is characterized with respect to the CpG/G + C content in all cancer types. As expected, sites that are hypermethylated in cancer are located in more G + C/CpG-rich regions than DMS^-. Overall, the abundance of DMS varies, depending on the assay. In addition, the number of DMS^+ and DMS^- differs significantly in all cancer types, yet not in the same way. The abundance of G-quadruplex structures in DMS^+ and DMS^- sites (Table 1, *G4* column) is statistically different in eight out of nine cancer types ($p < 0.01$). Similarly, the frequency of palindromic sequences (Table 1, *Pals* column) in DMS^+ and DMS^- sites is significantly different in six out of nine cancer types, while transcription factor binding sites exhibit also important differences. Transcription factor binding sites exhibit statistically significant differences in one out of nine cancer types (Table 1, *TFBS* column), while for conserved human/mouse/rat transcription factors the statistical significance is observed in five cancer types (Table 1, *cTFBS* column). Alternative splicing events exhibit mixed profusion and statistically significant results in two out of six datasets. Among four conformational changes, only minor groove width and propeller twist seem to affect or to be affected by differential DNAm in a small subset of cancer data (Supplementary File 1, Table S1), that is partially consistent with recent non-cancer-specific analyses[39].

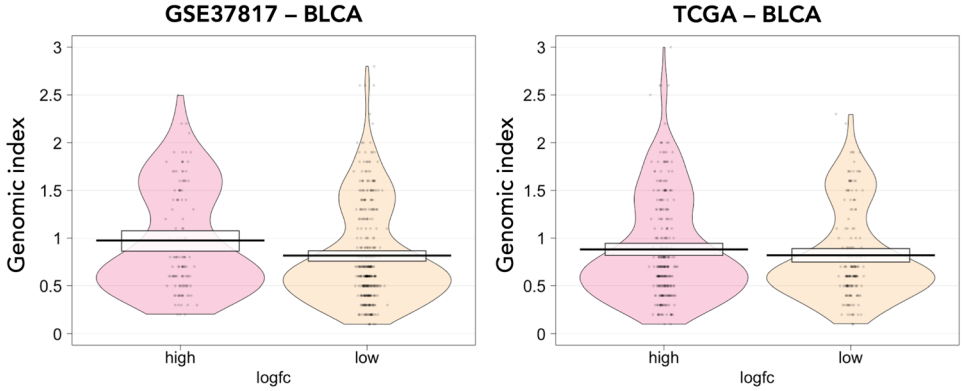
GEO ID	Cancer Data Series	Samples	Assay	DMS$^{(+/-)}$	G + C/ OE CpG content	G4		Pals		Alt. Spl.	TFBS		cTFBS		Ref
						mean(sd)	*p*-val	mean(sd)	*p*-val	*p*-val	mean(sd)	*p*-val	mean(sd)	*p*-val	mean(sd)	
GSE42752	Colorectal adenocarcinoma	22/22	HM450k	2,028	0.68/0.83	2.90 (1.33)	$<10^{-3}$	9.65 (4.63)	**0.002**	—	0.07 (0.25)	0.065	5.96 (5.81)	0.23	0.35 (1.17)	71
				49	0.55/0.52	1.51 (1.10)		7.53 (3.11)			—		9.25 (4.19)		0.2 (0.91)	
GSE54503	Hepatocellular carcinoma	66/66	HM450k	1,227	0.69/0.85	2.87 (1.37)	$<10^{-3}$	10.43 (5.03)	$<10^{-3}$	$<10^{-3}$	0.07 (0.25)	0.143	5.61 (5.07)	$<10^{-3}$	0.38 (1.16)	51
				7,490	0.53/0.52	1.18 (1.19)		6.9 (3.06)			0.01 (0.08)		6.62 (6.52)		0.08 (0.5)	
GSE85464*	Gastric adenocarcinoma	19/19	HM450k	161	0.65/0.78	2.57 (1.37)	$<10^{-3}$	8.95 (4.94)	$<10^{-3}$	—	0.08 (0.27)	**0.002**	5.06 (4.82)	$<10^{-3}$	0.4 (1.13)	72
				353	0.53/0.57	1.25 (1.24)		6.9 (3.4)			—		11.25 (14.29)		0.1 (0.66)	
GSE25093	Head & NeckSC carcinoma	91/18	HM27k	16	0.60/0.68	1.69 (1.13)	0.188	8.12 (4.96)	0.56	0.423	0.06 (0.25)	0.481	12 (12.19)	0.28	0.81 (2.74)	73
				83	0.53/0.4	1.29 (1.04)		7.06 (3.18)			0.02 (0.15)		9.32 (11.9)		0.16 (0.63)	
GSE32866*	Lung adenocarcinoma	28/27	HM27k	175	0.66/0.82	2.45 (1.35)	$<10^{-3}$	9.78 (4.62)	$<10^{-3}$	0.941	0.04 (0.20)	0.443	5.73 (5.1)	0.174	0.63 (1.56)	74
				23	0.53/0.35	1.17 (1.27)		5.26 (2.07)			0.04 (0.21)		9.67 (8.96)		0.13 (0.34)	
GSE37754#	Breast cancer	62/10	HM450k	152	0.59/0.69	1.95 (1.39)	$<10^{-3}$	7.25 (2.72)	0.466	—	0.03 (0.18)	0.546	5.67 (9.54)	0.36	0.24 (0.92)	75
				49	0.48/0.39	1.02 (1.13)		6.9 (3.1)			—		4.83 (3.19)		0.1 (0.42)	
GSE26989	Ovarian cancer	41/10	HM27k	613	0.56/0.49	1.54 (1.34)	**0.001**	7.12 (3.52)	0.099	0.304	0.09 (0.29)	0.751	7.98 (9.56)	**0.008**	0.32 (1.17)	76
				1,160	0.53/0.38	1.31 (1.21)		6.84 (3.11)			0.08 (0.27)		7.57 (7.93)		0.18 (0.69)	
GSE109402	Medulloblastoma	33/5	EPIC	14,120	0.55/0.43	1.22 (1.19)	$<10^{-3}$	7.23 (3.25)	**0.0095**	**0.002**	0.04 (0.19)	0.121	6.1 (5.76)	**0.0099**	0.16 (0.72)	77
				56,444	0.48/0.32	0.95 (1.09)		6.86 (3.09)			0.01 (0.12)		7.83 (8.29)		0.08 (0.53)	
GSE61441	Renal cell carcinoma	46/46	HM450k	86	0.60/0.74	1.97 (1.52)	$<10^{-3}$	8.2 (3.53)	**0.002**	0.36	0.03 (0.18)	0.95	4.64 (2.62)	**0.002**	0.36 (0.94)	78
				129	0.50/0.35	1.15 (1.21)		6.78 (2.97)			0.02 (0.12)		5.69 (5.84)		0.09 (0.64)	

Table 1. *p*-values of methylation-mediated features in cancer DNAm data obtained from BeadChip GEO data series. DMS^+ and DMS^- columns contain the number of hypermethylated and hypomethylated sites (DMS) respectively, for various cancer datasets (*Cancer Data Series*) and BeadChip assays. Maximum 1,000 sites were analyzed. *p*-values (*p-val* columns) are calculated with t-test and Wilcoxon test for non-normal distributions in addition to the mean values and standard deviations for each feature (*mean(sd)* columns). Column *G + C/OE CpG content* contains the G + C content and the observed/expected (OE) ratio of the CpG frequencies. Putative transcription factor binding sites (*TFBS* column) were identified only in DMS sequences located in promoters. The *samples* column contains the number of tumor/normal samples included from each data series (matched pairs when numbers are identical). *p*-values are not shown when no DMS co-localized with alternative splicing events (*Alt. Spl* column). *$|\Delta\beta| > 0.25$, #transformed M-values, $|\Delta\beta| > 0.20$. *G4*: G-quadruplexes, *Pals*: Palindromes, *cTFBS*: Conserved TFBS, SC: squamous cell.

In this use case, MeinteR automates the investigation of complex associations between epigenomic and genomic features across different cancer types. The results demonstrate a clear association between DNAm profiles and the genomic substrate that should be further elaborated and interpreted in the context of disease-specific analyses, as different outcomes may be attributed to the design, assay, and the experimental protocol of each study.

Use Case 3: Associating genomic DMS signatures with gene expression. The objective of this use case is to appraise the association between the genomic index and differential gene expression. First, we used MeinteR to download a GEO dataset for which both DNAm and gene expression data of the same samples are available. We used a set of 24 non-muscle invasive bladder cancer and matched normal samples (BLCA/GSE37817)[49]. To build expression profiles at gene level we applied GEO2R[48], and calculated the differential expression levels between cancer and control samples, using the binary logarithmic fold change. To analyze DNAm data, we used MeinteR in order to: (a) import β values of all samples, and (b) calculate mean β values per group. DMS were finally mapped to expression data and the level of differential DNAm was correlated with the expression levels of the mapped genes. To validate the results, we performed the same steps on processed TCGA Illumina HiSeq expression data from primary bladder urothelial carcinomas and normal tissues.

Differential DNAm analysis resulted in 1,474 probes that are more frequently located in "open sea" regions and less frequently to CpG islands. Fig. S2 (Supplementary File 1) illustrates the distribution of the genomic index in different regions relative to CpG islands. Probes located in CpG islands exhibit statistically significant differences of the genomic index and increased mean genomic index against all other regions (Shelf: $p\text{-val} = 0.02$, Shore: $p\text{-val} < 0.001$, Open sea: $p\text{-val} < 0.001$). None of the pairwise differences between shelves, shores and open sea probes were found statistically significant. DMS located in CpG islands were further annotated based on their position relative to gene regions. The density plots in Fig. S2 show that most CpG island DMS are located in 5′UTR and first exons, while the probes located in 200nt upstream transcription start sites (TSS200) have the highest mean genomic index. As expected, DMS^+ are more often located in genes that decrease their expression level, while DMS^- are spatially linked with genes that increase their expression with significant statistical difference ($p < 0.001$) for both BLCA datasets (Supplementary File 1, Fig. S3). To demonstrate the relevance of the genomic substrate in prioritizing critical DNAm events, we used MeinteR to assess whether differential gene expression is associated with higher genomic index. First, we calculated the genomic index of all DMS by assigning equal weights to the incorporated feature set. Figure 3 shows that among all aberrantly methylated sites, higher absolute differential expression is observed in sequences with increased genomic index. The differences are statistically significant for both BLCA datasets from TCGA and GEO ($p < 0.05$), implying that the effect of



	GSE37817 – BLCA	TCGA – BLCA
p-value	***p*=0.006**	***p*=0.011**
Mean genomic index/high DEx	0.975	0.898
Mean genomic index/low DEx	0.816	0.680

Figure 3. Smoothed density curves corresponding to the distribution of the genomic index in high differentially expressed genes (DEx) and low DEx in bladder cancer (BLCA). Horizontal black lines correspond to the mean genomic index and white boxes show the 95% bayesian highest density intervals.

differential DNAm in gene expression is probably modulated by the underlying genomic elements involved in the transcriptional regulation. The linear regression models for genomic index and logarithmic fold-change pairs are shown in Fig. S4 (Supplementary File 1).

Comparison with epigenetic markers and driver mutational signatures of hepatocellular carcinoma. To evaluate the competency of our approach, we performed comparisons with software tools and computational methods that identify driver events in two settings: (a) comparison between the genomic index and $\Delta\beta$ values to find the best-fitting metric for identifying epigenetic markers, and (b) correlations of the highly prioritized sites with known mutation-based driver genes and epimarkers.

DNAm data from hepatocellular carcinoma (HCC) were used to assess the overall performance. First, we downloaded TCGA level 3 HumanMethylation450 data from 50 primary HCC and matched normal pairs and built genomic signatures of the 13,153 DMS ($|\Delta\beta| \geq 0.3, p < 0.01$, FDR < 0.01), using equal-weighted attributes. Overlaps with palindromes, G-quadruplexes and conserved transcription factors were analyzed within 100nt region adjacent to each DMS. All sequences were scanned for transcription factors that unveil differential binding on DNAm targets in HepG2 cell line, using the curated data of the MEDReaders database[50]. To validate the results, we additionally used MeinteR to build genomic signatures of 8,717 DMS ($|\Delta\beta| \geq 0.3$, $p < 0.01$, FDR < 0.01) exported from 66 matched HCC and adjacent non-tumor tissues (GSE54503 data series)[51], using the same configuration. The datasets exhibit similar bimodal β value distributions for normal and tumor samples (Fig. 4A). The list of all critical DMS according to our ranking scheme is provided in Supplementary File 2.

To assess the efficiency of our method, we performed comparisons with known HCC markers that have been identified using DNAm data. Specifically, we found two probe-sets corresponding to: (a) 33 high-confidence epimarkers (epiHCC1) predicted by Zheng *et al.*[52], and (b) 109 HCC epimarkers (epiHCC2) identified by Cheng *et al.*[53]. For each reference probe-set, we calculated the enrichment of the genomic substrate using MeinteR. The genomic index of the epiHCC1 and epiHCC2 markers was estimated using the same configuration as for the TCGA-LIHC and GSE54503 datasets and is listed in Supplementary File 3. Figure 4B shows that epiHCC1 and epiHCC2 markers exhibit significantly more enriched genomic substrate (epiHCC1: mean g.index = 1.52, s.d. = 0.57, epiHCC2: mean g.index = 1.62, s.d. = 0.65), compared with the genomic index of all aberrantly methylated sites identified in the TCGA-LIHC and GSE54503 datasets (TCGA-LIHC: mean g.index = 0.77, s.d. = 0.52, GSE54503: mean g.index = 0.85, s.d. = 0.53). The differences of the genomic index levels are statistically significant for the TCGA-LIHC/epiHCC1 ($p = 9.9e\text{-}09$) and GSE54503/epiHCC1 ($p = 1.2e\text{-}07$) pairs and, as expected, not important for the TCGA-LIHC/GSE54503 comparison ($p = 0.31$). In accordance, the genomic index differs significantly for the TCGA-LIHC/epiHCC2 ($p < 2.22e\text{-}16$) and GSE54503/epiHCC2 ($p < 2.22e\text{-}16$) pairs. The genomic index of the epiHCC1 and epiHCC2 markers does not exhibit statistically significant differences ($p = 0.4$).

Cheng *et al.*[53] identified six epiHCC2 markers in four genes (*NEBL, FAM55C, GALNT3*, and *DSE*) that are hypermethylated exclusively in HCC. Interestingly, these HCC-specific diagnostic biomarkers have higher genomic index, compared to all 109 epiHCC2 markers (mean genomic index 2.03 vs.1.62, respectively), and more than two-fold higher genomic index than the average genomic index of all DMS identified in the TCGA/LIHC and GSE54503 datasets. Notably, the epimarkers of both methods would have been missed, if the $\Delta\beta$ level was used for selecting the most critical sites, since the differential DNAm level of the epimarkers (Fig. 4C) is lower (epiHCC1: mean $|\Delta\beta| = 0.17$, s.d. = 0.07, epiHCC2: mean $|\Delta\beta| = 0.26$, s.d. = 0.05) than the threshold commonly

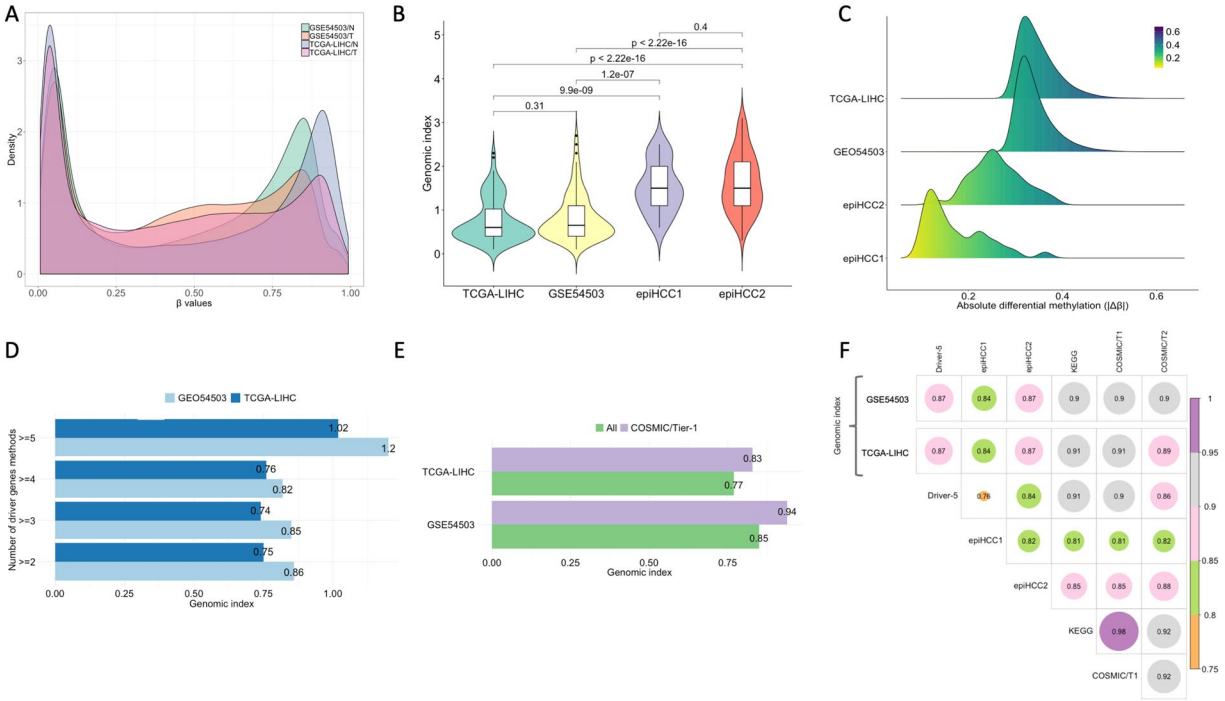


Figure 4. (**A**) Mean β value densities of the tumor(T)/normal(N) samples included in two public HCC datasets (TCGA-LIHC, GSE54503). (**B**) Smoothed density curves of the genomic index distributions and *p*-values (two-tailed t-test) for all pairwise comparisons between TCGA-LIHC/GSE54503 datasets and epimarkers (epiHCC1, epiHCC2). (**C**) Differential β value density of the TCGA-LIHC/GSE54503 datasets and epimarkers (epiHCC1, epiHCC2). (**D**) Mean genomic index of aberrantly methylated sites $|\Delta\beta| \geq 0.3$) in driver genes that have been detected by at least two to five computational methods. (**E**). Mean genomic index of all TCGA-LIHC and GSE54503 differentially methylated sites compared with the genomic index of COSMIC Tier-1 (T1) cancer gene census. (**F**). Semantic similarity plot correlating highly-ranked sites of the TCGA-LIHC and GSE54503 datasets, according to MeinteR, with the COSMIC T1, Tier 2 (T2) Gene Census, KEGG HCC pathway, Driver-5 genes and epigenetic markers (epiHCC1, epiHCC2).

set in cancer-specific analyses $|\Delta\beta| \geq 0.3$). These results further support the relevance of the genomic index as an important criterion to prioritize actionable epigenetic markers compared with the usage of the $\Delta\beta$ level.

Next, we pursued the evaluation of our method with respect to known driver cancer genes, motivated by previous studies that revealed significant associations between mutations in driver cancer genes with DNAm alterations[54–56]. First, we built a set of candidate driver genes based on the number of detection methods that report their causality in HCC, using DriverDB[57]. DriverDB[57] is a database that provides access to common driver genes that are computationally identified by 15 mutation-based methods. We assumed that more accountable driver genes are those reported by multiple driver detection methods and built the genomic signatures of driver genes with different levels of evidence. Accordingly, Driver-5 genes i.e. driver genes shared by at least five methods are more reliable than driver genes detected by two methods. Using these reference mutation-based data, we sought to compare the genomic index of low and high confident gene lists reported for HCC. The analysis of both TCGA-LIHC and GSE54503 datasets shows that driver genes identified by at least five methods (Driver-5) are associated with clearly higher genomic index than those with poorer evidence (Fig. 4D). The same procedure was applied on the gene set that is included in the COSMIC Cancer Gene Census (CGC, Tier 1)[58]. Compared with the entire set of DMS of TCGA-LIHC and GSE54504 datasets, those topologically linked with COSMIC CGC driver genes are on average associated with higher genomic index (Fig. 4E).

Finally, we compared MeinteR with relevant methods that detect driver genes/markers based on mutational and epigenetic HCC signatures. For this analysis, we additionally obtained the list of KEGG genes that are involved in the HCC pathway (hsa05225) and performed pairwise comparisons, in order to estimate the Wang's semantic distance between Disease Ontology terms[59] and the best-max average combination method[60]. The correlation plots in Fig. 4F show the semantic similarity levels of the highly-ranked TCGA-LIHC and GSE54503 genes with known driver genes and epimarkers. The comparison between MeinteR and the consensus mutation-based driver genes (*Driver-5*) shows that MeinteR exhibits similar semantic correlation with the KEGG and COSMIC genes and slightly better semantic correlation with COSMIC Tier 2 genes. These results are obtained using both TCGA and GEO datasets. Finally, the highly-ranked genes prioritized by our method are evidently better correlated with KEGG and COSMIC genes than epiHCC1 and epiHCC2 markers. As expected, the highest correlation level is observed between COSMIC CGC Tier 1 and KEGG pathway genes, while the epiHCC1/Driver-5 comparison has the lowest correlation level.

Discussion

The role of DNAm in disease onset and progression has been extensively studied, particularly in cancer (reviewed by Chatterjee *et al.*[19]). Although epigenome aberrations are frequently observed in most cancer types, recent studies have shown that small sets of aberrantly methylated sites are able to discriminate cancer subtypes[54] and to predict drug response[61], posing computationally challenging questions on which of the epigenetic alterations are functionally key events in cancer. MeinteR consolidates our knowledge on methylation-modulating mechanisms enabling, for the first time, the identification of high-impact epigenetic alterations, under the prism of conformational and cis-regulatory element enrichment, quantified by the genomic index as a linear function of the feature abundance. The genomic index does not explicitly dictate the presence of protective regions or regions prone to transcriptional changes, yet higher values imply a greater incidence of methylation-modulated, functional elements, that might play a critical role in transcriptional events. In this context, MeinteR is better described as an approach that implements "upstream" biological interpretation, as it incorporates features associated with potential causes of the DNAm events, as opposed to the "downstream" biological interpretation that quantifies the effect of DNAm events on biological pathways[17].

In comparison with other epigenetic driver detection methods, MeinteR differs in both the research hypothesis and methodology. First, MeinteR identifies the most influential DNAm sites, rather than driver DNAm alterations. The latter imply the presence of causal relationships between driver and passenger epigenetic alterations that are not essentially relevant with the genomic substrate. Second, most computational methods entail the integration of multi-omics data to identify driver events, e.g. gene expression, DNAm and copy number variations[20–22]. MeinteR relies exclusively on DNAm data enabling faster and straight-forward interpretative analyses of high-throughput experiments. The evaluation results demonstrate that aberrant DNAm sites, co-localized with putative conformational and cis-regulatory elements, are better correlated with known cancer drivers, suggesting a potential role in transcriptional regulation with significant diagnostic and therapeutic implications.

MeinteR is an open-source R package, easily applicable to TCGA and GEO data analyses, enabling case-by-case configuration of the incorporated features and weighting scheme. In addition, MeinteR is valuable in improving the accuracy of imputation methods, as it has been shown that the prediction of CpG methylation levels based only on neighboring CpG sites is suboptimal, especially in sparsely assayed genomic regions[26]. Equally important, MeinteR incorporates functions that are time-effectively applied in genome-wide DNAm datasets, with no special hardware requirements. In this respect, our contribution is inline with the FAIR principles[62] (i.e. *Findable* – *Accessible*; as it is publicly available in a reference software repository – *Interoperable*; as it has been implemented in R, an open source environment, and exploits data and software from reference third-party repositories – *Reusable*; since besides its public availability, it is also accompanied with detailed documentation and comprehensive examples of use), fostering transparency and reproducibility of the source code.

Overall, with MeinteR we aim to provide the basis for exploratory and explanatory analyses related with development, aging, cancer and other biological processes and diseases complementing the interpretation of DNAm alterations, beyond local architecture annotations and pathway enrichments and with potential usability in developing predictive models for identifying disease subtypes and response to treatments.

Methods

Module 1: Data preprocessing. MeinteR's functions are applied on bed-formatted chromosomal interval files containing the coordinates of each DMS, and the corresponding score values. These files are retrievable by tools performing differential DNAm analyses, such as limma[9], RnBeads[63], minfi[64], ChAMP[16], Bicycle[65] etc. Alternatively, MeinteR is able to fetch array-based and sequencing-based data from GEO[48] and to automatically build valid data files. In case of array platforms, such as Illumina's BeadChip HumanMethylation27, HumanMethylation450 and MethylationEPIC, MeinteR splits samples in two subsets, according to a predefined annotation file that contains the list of sample identifiers and the corresponding group e.g. normal/tumor, pre-/post-treatment. Then, $\Delta\beta$ values are calculated and valid interval files are generated (Supplementary File 1, Fig. S5). Sequencing data from whole genome bisulfite-sequencing (WGBS), reduced representation bisulfite sequencing (RRBS) and targeted bisulphite-based experiments contain the number of methylated reads and read depth information per CpG site. MeinteR fetches sequencing data for each sample, filters data based on the read depth and chromosome, and builds interval files containing DNAm level per site, as a fraction of cytosine-reporting reads vs. the total number of mapped reads (example usage on WGBS, RRBS data is available on the software's vignette). Besides data loading, the preprocessing module contains a set of functions for the validation of data values and format, M to β value conversion, as well as plotting and filtering functions.

Module 2: Feature detection. MeinteR calculates the abundance of methylation-mediated features in variable-length sequences centered at each CpG target, using a set of functions as described below (Fig. 1).

Transcription factor binding motifs. MeinteR identifies putative binding sites of: (a) conserved transcription factors in human/mouse/rat alignments and (b) human transcription factors included in the JASPAR's core collection (version 2018)[66]. Conserved factors and their ~5.8 million genome-wide binding loci are retrieved from the corresponding track of the UCSC Table Browser. The intersection of the binding loci and the genomic coordinates of the regions flanking each DMS is exported and comparatively visualized against the expected frequency. For the detection of JASPAR's profile matrices, MeinteR uses the scanning algorithm implemented in TFBSTools[67], in order to identify high-scoring matches between transcription factor profile matrices and DMS in user-defined sequence offset. To speed-up the analysis of large datasets, searches can be narrowed-down to promoters or CpG islands. In addition, MeinteR allows users to select a list of transcription factors, among hundreds available, and perform targeted enrichment analyses, excluding the "noisy" binding sites.

Palindromes and G-quadruplex structures. Palindromic sequences are detected using Biostrings[68]. First, MeinteR retrieves sequences of variable length centered at DMS and scans for palindromic regions based on user-defined arm lengths and in-between loop sizes. The identification of potential quadruplex-forming sequences is implemented using the pqsfinder algorithm[69]. As for palindromes, MeinteR retrieves genomic sequences corresponding to the coordinates of each DMS expanded by a user-defined offset and performs batch detection of G-quadruplex structures. The output of both functions includes summary and verbose reports of the detected readouts that are subsequently used to build genomic signatures.

Splice sites and alternative splicing events. MeinteR enables batch analyses of splicing-related events by: (a) detecting putative 5′ and 3′ splice sites, and (b) matching known alternative splicing events to the CpG coordinates. Generally, splice junctions and their short neighboring sequences are characterized by species-specific conserved motifs. In this work, motifs are described as position-specific weight matrices, following the definition of donor and acceptor sites by Shapiro & Senapathy[70]. Then, short sequences adjacent to DMS are searched for these matrices, using the same scanning method that is applied for transcription factor binding site detection[67]. The detection of overlapping alternative isoforms is based on known alternative splicing events available by the UCSC Table Browser. MeinteR calculates the frequency of different alternative splicing events overlapping DMS data and builds graphical reports of the observed and expected frequency in the human reference genome.

Conformational DNA features. To determine putative conformational DNA changes caused by DNAm, MeinteR uses the methyl-DNAshape algorithm[39]. For each DNAm site, the respective function retrieves short sequences of user-defined length adjacent to DMS in batches and uses methyl-DNAshape to calculate minor groove width, roll, propeller and helix twist in the unmethylated and methylated context. The difference between the two states is evaluated and the statistical significance of each DNA shape is calculated using Wilcoxon tests.

Module 3: Signature extraction. The third module aggregates genomic features at each CpG site and: (a) constructs a signature matrix, and (b) performs multi-variate ranking to identify putatively actionable sites as shown below (Fig. 1).

Genomic signature matrix. The genomic signature of each DMS is assembled in a matrix containing at least one of the incorporated genomic features. Let $x_i \in X = \{x_1, x_2, \ldots, x_m\}$ be a set of m DMS sites and $a_i \in A = \{a_1, a_2, \ldots, a_n\}$ is the list of n attributes, i.e. genomic features associated with each CpG site. The attributes are of different scaling and data types and treated accordingly, in order to operate on the same scale. Splicing-related observations i.e. putative donor, acceptor sites and co-localized alternative splicing events are joined into logical values representing the incidence of at least one feature. Similarly, conformational changes are quantified as logical variables that are positive, when at least one DNA shape alteration is statistically significant ($p < 0.05$). Finally, the abundance of G-quadruplex structures, palindromic sequences and transcription factor binding sites (conserved, putative human JASPAR core collection 2018) are normalized to fit (0,1) scale. Overall, depending on the feature a_j, a mapping function m_j is applied for each attribute j, where $a'_j = m_j(a_j)$ and the genomic signature matrix A' is built that tabulates the mapped attribute values a'_{ij} for each x_i.

Multi-variate ranking. Given a signature matrix A', the final step is to rank each DMS x_i based on the genomic index idx_i, with ($idx_i \in \mathbb{N}^+$). For each x_i, idx_i is defined by the weighted sum of all attributes a'_{ij}. For example, if w_j is the positive weight of the j^{th} DMS attribute, x_i is defined by the sum of the weighted attribute values of x_i, i.e.: $idx_i = \sum_{j=1}^{n} w_j a_{ij}$, $i = 1, 2, \ldots, m$. The output data are exported in an m-length vector of genomic index values quantifying the enrichment of the incorporated elements at each CpG site (genomic signature).

Data availability

All data generated or analyzed during this study are included in this published article (and its Supplementary Information Files). MeinteR is an R package available under the GNU General Public Licence v3. The source code and binaries can be found at https://github.com/andigoni/MeinteR. The repository contains also documentation of the respective R package including a manual, a package tutorial with examples and the source code of the three use cases.

Received: 1 July 2019; Accepted: 19 November 2019;
Published online: 16 December 2019

References

1. Teschendorff, A. E. *et al*. A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. *Bioinformatics* **29**, 189–196 (2013).
2. Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E. & Storey, J. D. The SVA package for removing batch effects and other unwanted variation in highthroughput experiments. *Bioinformatics* **28**, 882–883 (2012).
3. Liu, Y., Siegmund, K. D., Laird, P. W. & Berman, B. P. Bis-SNP: Combined DNA methylation and SNP calling for Bisulfite-seq data. *Genome Biol.* **13**, R61 (2012).
4. Krueger, F. & Andrews, S. R. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27**, 1571–1572 (2011).
5. Horvath, S. DNA methylation age of human tissues and cell types. *Genome Biol.* **14** (2013).
6. Houseman, E. A., Molitor, J. & Marsit, C. J. Reference-free cell mixture adjustments in analysis of DNA methylation data. *Bioinformatics* **30**, 1431–1439 (2014).
7. Catoni, M., Tsang, J. M., Greco, A. P. & Zabet, N. R. DMRcaller: a versatile R/Bioconductor package for detection and visualization of differentially methylated regions in CpG and non-CpG contexts. *Nucleic Acids Res.* **46**, e114 (2018).

8. Phipson, B., Maksimovic, J. & Oshlack, A. MissMethyl: An R package for analyzing data from Illumina's HumanMethylation450 platform. *Bioinformatics* **32**, 286–288 (2016).
9. Ritchie, M. E. *et al*. Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
10. Breeze, C. E. *et al*. eFORGE v2.0: updated analysis of cell type-specific signal in epigenomic data. *Bioinformatics* **35**, 4767–4769 (2019).
11. Sheffield, N. C. & Bock, C. LOLA: Enrichment analysis for genomic region sets and regulatory elements in R and Bioconductor. *Bioinformatics* **32**, 587–589 (2016).
12. Müller, F. *et al*. RnBeads 2.0: comprehensive analysis of DNA methylation data. *Genome Biol.* **20** (2019).
13. Preussner, J., Bayer, J., Kuenne, C. & Looso, M. ADMIRE: Analysis and visualization of differential methylation in genomic regions using the Infinium HumanMethylation450 Assay. *Epigenetics and Chromatin* **8** (2015).
14. Min, J. L., Hemani, G., Davey Smith, G., Relton, C. & Suderman, M. Meffil: efficient normalization and analysis of very large DNA methylation datasets. *Bioinformatics* **34**, 3983–3989 (2018).
15. Gorrie-Stone, T. J. *et al*. Bigmelon: Tools for analysing large DNA methylation datasets. *Bioinformatics* **35**, 981–986 (2019).
16. Tian, Y. *et al*. ChAMP: Updated methylation analysis pipeline for Illumina BeadChips. *Bioinformatics* **33**, 3982–3984 (2017).
17. Wang, Y., Franks, J. M., Whitfield, M. L. & Cheng, C. BioMethyl: an R package for biological interpretation of DNA methylation data. *Bioinformatics* **35**, 3635–3641 (2019).
18. Kalari, S. & Pfeifer, G. P. Identification of Driver and Passenger DNA Methylation in Cancer by Epigenomic Analysis. *Adv. Genet.* **70**, 277–308 (2010).
19. Chatterjee, A., Rodger, E. J. & Eccles, M. R. Epigenetic drivers of tumourigenesis and cancer metastasis. *Semin. Cancer Biol.* **51**, 149–159 (2018).
20. Gevaert, O. MethylMix: an R package for identifying DNA methylation-driven genes. *In Bioinformatics* **31**, 1839–41 (2015).
21. Cedoz, P. L., Prunello, M., Brennan, K. & Gevaert, O. MethylMix 2.0: An R package for identifying DNA methylation genes. *Bioinformatics* **34**, 3044–3046 (2018).
22. Champion, M. *et al*. Module Analysis Captures Pancancer Genetically and Epigenetically Deregulated Cancer Driver Genes for Smoking and Antiviral Response. *EBioMedicine* **27**, 156–166 (2018).
23. Dimitrakopoulos, C. *et al*. Network-based integration of multi-omics data for prioritizing cancer genes. *Bioinformatics* **34**, 2441–2448 (2018).
24. Feltus, F. A., Lee, E. K., Costello, J. F., Plass, C. & Vertino, P. M. DNA motifs associated with aberrant CpG island methylation. *Genomics* **87**, 572–579 (2006).
25. Baubec, T. & Schübeler, D. Genomic patterns and context specific interpretation of DNA methylation. *Current Opinion in Genetics and Development* **25**, 85–92 (2014).
26. Zhang, W., Spector, T. D., Deloukas, P., Bell, J. T. & Engelhardt, B. E. Predicting genome-wide DNA methylation using methylation marks, genomic position, and DNA regulatory elements. *Genome Biol.* **16** (2015).
27. Luu, P. L., Schöler, H. R. & Araúzo-Bravo, M. J. Disclosing the crosstalk among DNA methylation, transcription factors, and histone marks in human pluripotent cells through discovery of DNA methylation motifs. *Genome Res.* **23**, 2013–2029 (2013).
28. Kapourani, C. A. & Sanguinetti, G. Higher order methylation features for clustering and prediction in epigenomic studies. *Bioinformatics* **32**, i405–i412 (2016).
29. Lawson, J. T., Tomazou, E. M., Bock, C. & Sheffield, N. C. MIRA: an R package for DNA methylation-based inference of regulatory activity. *Bioinformatics* **34**, 2649–2650 (2018).
30. Hu, S. *et al*. DNA methylation presents distinct binding sites for human transcription factors. *Elife* **2**, e00726 (2013).
31. Zhu, H., Wang, G. & Qian, J. Transcription factors as readers and effectors of DNA methylation. *Nat. Rev. Genet.* **17**, 551–565 (2016).
32. Yin, Y. *et al*. Impact of cytosine methylation on DNA binding specificities of human transcription factors. *Science* **356** (2017).
33. Halder, R. *et al*. Guanine quadruplex DNA structure restricts methylation of CpG dinucleotides genome-wide. *Mol. Biosyst.* **6**, 2439–2447 (2010).
34. Tsukakoshi, K., Saito, S., Yoshida, W., Goto, S. & Ikebukuro, K. CpG methylation changes G-Quadruplex structures derived from gene promoters and interaction with VEGF and SP1. *Molecules* **23**, 1–12 (2018).
35. Malousi, A. *et al*. Age-dependent methylation in epigenetic clock CpGs is associated with G-quadruplex, co-transcriptionally formed RNA structures and tentative splice sites. *Epigenetics* **13**, 808–821 (2018).
36. Allers, T. & Leach, D. R. F. DNA palindromes adopt a methylation-resistant conformation that is consistent with DNA cruciform or hairpin formation in vivo. *J. Mol. Biol.* **252**, 70–85 (1995).
37. Zinoviev, V. V., Yakishchik, S. I., Evdokimov, A. A., Malygin, E. G. & Hattman, S. Symmetry elements in DNA structure important for recognition/methylation by DNA [amino]-methyltransferases. *Nucleic Acids Res.* **32**, 3930–3934 (2004).
38. Lazarovici, A. *et al*. Probing DNA shape and methylation state on a genomic scale with DNase I. *Proc. Natl. Acad. Sci.* **110**, 6376–6381 (2013).
39. Rao, S. *et al*. Systematic prediction of DNA shape changes due to CpG methylation explains epigenetic effects on protein-DNA binding. *Epigenetics and Chromatin* **11** (2018).
40. Maunakea, A. K., Chepelev, I., Cui, K. & Zhao, K. Intragenic DNA methylation modulates alternative splicing by recruiting MeCP2 to promote exon recognition. *Cell Res.* **23**, 1256–1269 (2013).
41. Malousi, A. & Kouidou, S. DNA hypermethylation of alternatively spliced and repeat sequences in humans. *Mol. Genet. Genomics* **287**, 631–642 (2012).
42. Singer, M., Kosti, I., Pachter, L. & Mandel-Gutfreund, Y. A diverse epigenetic landscape at human exons with implication for expression. *Nucleic Acids Res.* **43**, 3498–3508 (2015).
43. Lev Maor, G., Yearim, A. & Ast, G. The alternative role of DNA methylation in splicing regulation. *Trends Genet.* **31**, 274–280 (2015).
44. Gelfman, S., Cohen, N., Yearim, A. & Ast, G. DNA-methylation effect on cotranscriptional splicing is dependent on GC architecture of the exon-intron structure. *Genome Res.* **23**, 789–799 (2013).
45. Malousi, A., Maglaveras, N. & Kouidou, S. Intronic CpG content and alternative splicing in human genes containing a single cassette exon. *Epigenetics* **3**, 69–73 (2008).
46. Machado, A. C. D. *et al*. Evolving insights on how cytosine methylation affects protein-DNA binding. *Brief. Funct. Genomics* **14**, 61–73 (2015).
47. Grossman, R. L. *et al*. Toward a Shared Vision for Cancer Genomic Data. *N. Engl. J. Med.* **375**, 1109–12 (2016).
48. Barrett, T. *et al*. NCBI GEO: Archive for functional genomics data sets - Update. *Nucleic Acids Res.* **41** (2013).
49. Kim, Y. J. *et al*. HOXA9, ISL1 and ALDH1A3 methylation patterns as prognostic markers for nonmuscle invasive bladder cancer: Array-based DNA methylation and expression profiling. *Int. J. Cancer* **133**, 1135–1142 (2013).
50. Wang, G. *et al*. MeDReaders: A database for transcription factors that bind to methylated DNA. *Nucleic Acids Res.* **46**, D146–D151 (2018).
51. Shen, J. *et al*. Exploring genome-wide DNA methylation profiles altered in hepatocellular carcinoma using Infinium HumanMethylation 450 BeadChips. *Epigenetics* **8**, 34–43 (2013).
52. Zheng, Y. *et al*. Genome-wide DNA methylation analysis identifies candidate epigenetic markers and drivers of hepatocellular carcinoma. *Brief. Bioinform.* **19**, 101–108 (2018).

53. Cheng, J. *et al*. Integrative analysis of DNA methylation and gene expression reveals hepatocellular carcinoma-specific diagnostic biomarkers. *Genome Med.* **10**, 42 (2018).
54. Chen, Y. C., Gotea, V., Margolin, G. & Elnitski, L. Significant associations between driver gene mutations and DNA methylation alterations across many cancer types. *PLoS Comput. Biol.* **13**, e1005840 (2017).
55. Tiedemann, R. L. *et al*. Dynamic reprogramming of DNA methylation in SETD2-deregulated renal cell carcinoma. *Oncotarget* **7**, 1927–46 (2016).
56. Turcan, S. *et al*. IDH1 mutation is sufficient to establish the glioma hypermethylator phenotype. *Nature* **483**, 479–83 (2012).
57. Chung, I. F. *et al*. DriverDBv2: A database for human cancer driver gene research. *Nucleic Acids Res.* **44**, D975–9 (2016).
58. Forbes, S. A. *et al*. COSMIC: Somatic cancer genetics at high-resolution. *Nucleic Acids Res.* **45**, D777–D783 (2017).
59. Kibbe, W. A. *et al*. Disease Ontology 2015 update: An expanded and updated database of Human diseases for linking biomedical knowledge through disease data. *Nucleic Acids Res.* **43**, D1071–8 (2015).
60. Wang, J. Z., Du, Z., Payattakool, R., Yu, P. S. & Chen, C.-F. A new method to measure the semantic similarity of GO terms. *Bioinformatics* **23**, 1274–1281 (2007).
61. Xia, X. *et al*. Incorporating methylation genome information improves prediction accuracy for drug treatment responses. *BMC Genet*. **19** (2018).
62. Wilkinson, M. D. *et al*. Comment: The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **3** (2016).
63. Assenov, Y. *et al*. Comprehensive analysis of DNA methylation data with RnBeads. *Nat. Methods* **11**, 1138–1140 (2014).
64. Fortin, J. P., Triche, T. J. & Hansen, K. D. Preprocessing, normalization and integration of the Illumina HumanMethylationEPIC array with minfi. *Bioinformatics* **33**, 558–560 (2017).
65. Graña, O., López-Fernández, H., Fdez-Riverola, F., González Pisano, D. & Glez-Peña, D. Bicycle: A bioinformatics pipeline to analyze bisulfite sequencing data. *Bioinformatics* **34**, 1414–1415 (2018).
66. Khan, A. *et al*. JASPAR 2018: Update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Res.* **46**, D260–D266 (2018).
67. Tan, G. & Lenhard, B. TFBSTools: An R/bioconductor package for transcription factor binding site analysis. *Bioinformatics* **32**, 1555–1556 (2016).
68. Pagès H, Aboyoun P, Gentleman R, DebRoy S Biostrings: Efficient manipulation of biological strings. R package version 2.54.0 (2019).
69. Hon, J., Martínek, T., Zendulka, J. & Lexa, M. pqsfinder: an exhaustive and imperfection-tolerant search tool for potential quadruplex-forming sequences in R. *Bioinformatics* **33**, 3373–3379 (2017).
70. Shapiro, M. B. & Senapathy, P. RNA splice junctions of different classes of eukaryotes: Sequence statistics and functional implications in gene expression. *Nucleic Acids Res.* **15**, 7155–7174 (1987).
71. Naumov, V. A. *et al*. Genome-scale analysis of DNA methylation in colorectal cancer using Infinium HumanMethylation450 BeadChips. *Epigenetics* **8**, 921–934 (2013).
72. Ooi, W. F. *et al*. Epigenomic profiling of primary gastric adenocarcinoma reveals super-enhancer heterogeneity. *Nat. Commun.* **7**, 12983 (2016).
73. Poage, G. M. *et al*. Identification of an epigenetic profile classifier that is associated with survival in head and neck cancer. *Cancer Res.* **72**, 2728–2737 (2012).
74. Selamat, S. A. *et al*. Genome-scale analysis of DNA methylation in lung adenocarcinoma and integration with mRNA expression. *Genome Res.* **22**, 1197–1211 (2012).
75. Terunuma, A. *et al*. MYC-driven accumulation of 2-hydroxyglutarate is associated with breast cancer prognosis. *J. Clin. Invest.* **124**, 398–412 (2014).
76. Campan, M. *et al*. Genome-scale screen for DNA methylation-based detection markers for ovarian cancer. *PLoS One* **6**, e28141 (2011).
77. Rivero-Hinojosa, S. *et al*. Proteomic analysis of Medulloblastoma reveals functional biology with translational potential. *Acta Neuropathol. Commun.* **6**, 48 (2018).
78. Wei, J. H. *et al*. A CpG-methylation-based assay to predict survival in clear cell renal cell carcinoma. *Nat. Commun.* **6**, 8699 (2015).

Acknowledgements

The results shown here are in part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga. This work was supported in part by the Horizon 2020 project MEDGENET funded by the European Union; ERA-NET TRANSCAN-2 JTC 2014, GCH-CLL #143; and, the KRIPIS action, funded by the General Secretariat for Research and Technology of Greece.

Author contributions

A.M. designed and implemented the components of the framework and wrote the manuscript, S.K. contributed to the discussion and to the selection of the incorporated features, M.T. and N.P. contributed to the evaluation of the method, E.B. contributed to the statistical analysis, E.G. and G.T. critically reviewed the manuscript and K.S. was responsible for the coordination and evaluation of this work.

Competing Interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-019-55453-8.

Correspondence and requests for materials should be addressed to A.M.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.