PLOS ONE

RESEARCH ARTICLE

# Stage-differentiated ensemble modeling of DNA methylation landscapes uncovers salient biomarkers and prognostic signatures in colorectal cancer progression

**Sangeetha Muthamilselvan**[☯], **Abirami Raghavendran**[☯], **Ashok Palaniappan***

Department of Bioinformatics, School of Chemical and BioTechnology, SASTRA Deemed University, Thanjavur, India

☯ These authors contributed equally to this work.
* apalania@scbt.sastra.edu



OPEN ACCESS

**Citation:** Muthamilselvan S, Raghavendran A, Palaniappan A (2022) Stage-differentiated ensemble modeling of DNA methylation landscapes uncovers salient biomarkers and prognostic signatures in colorectal cancer progression. PLoS ONE 17(2): e0249151. https://doi.org/10.1371/journal.pone.0249151

**Editor:** Surinder K. Batra, University of Nebraska Medical Center, UNITED STATES

**Received:** March 16, 2021

**Accepted:** February 1, 2022

**Published:** February 24, 2022

**Peer Review History:** PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pone.0249151



**Data Availability Statement:** https://doi.org/10.6084/m9.figshare.13013852.

## Abstract

### Background

Aberrant DNA methylation acts epigenetically to skew the gene transcription rate up or down, contributing to cancer etiology. A gap in our understanding concerns the epigenomics of stagewise cancer progression. In this study, we have developed a comprehensive computational framework for the stage-differentiated modelling of DNA methylation landscapes in colorectal cancer (CRC).

### Methods

The methylation β-matrix was derived from the public-domain TCGA data, converted into M-value matrix, annotated with AJCC stages, and analysed for stage-salient genes using an ensemble of approaches involving stage-differentiated modelling of methylation patterns and/or expression patterns. Differentially methylated genes (DMGs) were identified using a contrast against controls (adjusted p-value <0.001 and |log fold-change of M-value| >2), and then filtered using a series of all possible pairwise stage contrasts (p-value <0.05) to obtain stage-salient DMGs. These were then subjected to a consensus analysis, followed by matching with clinical data and performing Kaplan–Meier survival analysis to evaluate the impact of methylation patterns of consensus stage-salient biomarkers on disease prognosis.

### Results

We found significant genome-wide changes in methylation patterns in cancer cases relative to controls agnostic of stage. The stage-differentiated models yielded the following consensus salient genes: one stage-I gene (*FBN1*), one stage-II gene (*FOXG1*), one stage-III gene (*HCN1*) and four stage-IV genes (*NELL1*, *ZNF135*, *FAM123A*, *LAMA1*). All the biomarkers were significantly hypermethylated in the promoter regions, indicating down-regulation of expression and implying a putative CpG island Methylator Phenotype (CIMP) manifestation.

**Funding:** This work has been funded by DST-SERB grant EMR/2017/000470/BBM to A.P. (Department of Science & Technology - Science & Engineering Research Board, Govt. of India). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing interests:** The authors have declared that no competing interests exist.

A prognostic signature consisting of FBN1 and FOXG1 survived all the analytical filters, and represents a novel early-stage epigenetic biomarker / target.

## Conclusions

We have designed and executed a workflow for stage-differentiated epigenomic analysis of colorectal cancer progression, and identified several stage-salient diagnostic biomarkers, and an early-stage prognostic biomarker panel. The study has led to the discovery of an alternative CIMP-like signature in colorectal cancer, reinforcing the role of CIMP drivers in tumor pathophysiology.

## Introduction

Colorectal adenocarcinoma (CRC) is a major malignant disease with devastating incidence and mortality, being the cancer with the third highest global burden of disease, after lung and breast cancers, and accounting for 1.36 million new cases annually [1]. The etiology of CRC involves chromosomal instability (involving accumulation of mutations in oncogenes and tumor suppressor genes), microsatellite instability (MSI) (leading to loss of DNA mismatch repair) and CpG island methylator phenotype (CIMP), observed in nearly 85%, 15% and 10–40% respectively of all reported sporadic cases [2–4]. Epigenetic dysregulation is a key driver of these processes, and DNA methylation is the most important epigenetic modification [5, 6]. DNA hypomethylation could cause gain-of-function of oncogenes [7], and might aid severe tumor progression [8]. It has been found that large hypomethylation blocks are a universal characteristic of colorectal cancers and other solid tumors [9]. Hypomethylation could also contribute to tumor initiation and progression by a general increase in genomic instability [10]. DNA hypermethylation could cause loss-of-function of tumor suppressor genes, and hypermethylation in the germline could cause heritable loss of gene expression through genomic imprinting [11]. Aberrant hypermethylation of specific CpG islands has been observed to occur in colorectal cancer. The CpG island methylator phenotype (CIMP) was originally discovered in a subset of colorectal cancers [12], and subsequently refined to the involvement of five genes *CACNA1G*, *IGF2*, *NEUROG1*, *RUNX3*, and *SOCS1* [13]. Methylation changes contributing to phenotypic aberrations need not be localized to promoter regions but could occur in the gene coding regions and intron-exon structures [14–17]. The persistence of such modifications throughout the tumor cell lifetime has also been demonstrated by Lengauer et al. [18], who showed that methylation aberrations and genome instability were correlated, suggesting a key role for such aberrations in tumorigenic chromosomal segregation processes.

The Cancer Genome Atlas (TCGA) is a comprehensive resource of genome-wide mutation, expression and DNA methylation profiles of 46 different types of cancers [19]. Besides the TCGA, the International Human Epigenetic Consortium is devoted to data-driven understanding of the role of epigenomics in normal vs disease states [20]. Methylation patterns constitute an emerging class of promising prognostic factors mainly due to: (i) the persistence of widespread DNA methylation changes; (ii) the occurrence of such changes much ahead of the consequent changes in gene expression; and (iii) the ability to detect these changes in body fluids and blood plasma [21]. Few methylation markers have been previously translated to clinically applicable biomarkers [22], but it is known that tumorbehavior corresponds with epigenomic changes as reflected in differential DNA methylation [23]. Early detection may reduce the mortality rate via tailored adjustments to the treatment regimen, with the result of

fewer side-effects and better patient compliance. Chen et al., demonstrated a method to screen multiple types of cancer using a methylation-based blood test four years before conventional diagnosis [24]. A consensus approach to identifying significant methylation signatures in each stage of colorectal cancer progression would increase the utility and reliability of putative biomarkers. This motivated our interest in a systematic investigation of stage-salient epigenetic factors using several model-driven approaches, with the main objective of obtaining diagnostic and prognostic biomarkers.

## Methods

### Data preprocessing

Methylation data from 27k assays was used, since it is preferentially enriched in epigenetic profiles in the proximal promoter regions (relative to 450k assays which are enriched in probes in the gene body and intergenic regions) [25]. Processed Level-3 27k CRC methylation data was retrieved from TCGA [26]. All samples in the dataset were processed and submitted by a single organization (namely 05: JHU_USC center), ensuring uniformity in data processing. MBatch analysis yielded low (<0.3) Dispersion Separability Criterion (measured as the ratio of between-batch dispersion to within-batch dispersion), indicating negligible batch effects and obviating the need for batch-correction (https://bioinformatics.mdanderson.org/public-software/mbatch/). The data containing the methylation β-values for each probe in each sample was converted into a matrix with probes as rows and cases as columns. Each probe corresponds to one CpG site in the genome. A single gene may be under the control of multiple epigenetic sites, hence multiple probes may be associated with the same gene. It is noted that multiple probes usually exist for the same gene. The probes which have 'na' values were discarded from the analysis. To transform the range of methylation values from (0,1) to (-∞, +∞), we used the following function on the β-matrix values, to obtain the M-value matrix [27]:

$$M_i = \log_2[\beta_i/(1-\beta_i)] \quad (1)$$

In our study, two M-value matrices were considered: one, where all the probes were used in the analysis; and two, where the probes corresponding to one gene were represented by an average of their values ('avereps'), thus reducing the M-value matrix from a probe:sample matrix to a gene:sample matrix. Further, we filtered out the probes/genes showing little change in methylation (defined as $\sigma < 1$) across all cases in the M-value matrices. The latest clinical data (clinical.cases_selected.tar.gz) was obtained from the GDC by matching on the patient barcode [28]. The stages were annotated for both the β-matrix and M-value matrices using the 'Pathologic_stage' attribute encoded in the clinical data. Cases with unknown stage ('NA' values) were discarded. The stage information was mapped to the American Joint Committee on Cancer (AJCC) Tumor-Node-Metastasis (TNM) classification system [29] (Table 1).

The final β and M-value matrices were subjected to stage-differentiated contrast analysis with a battery of six different methods, described below. All analysis was carried out on R [30].

### Modelling

To compensate for the assumptions specific to individual modelling approaches, an ensemble of models was explored.

**(1) Linear modelling with M-values.** Linear modelling is essential to identify linear trends in expression across cancer stages and thereby detect stage-sensitive patterns. We used

**Table 1. AJCC cancer staging.**

| TCGA Stage | TNM Classification | Cases | |
|---|---|---|---|
| I | T1N0M0 | 50 | |
| II | - | 17 | 86 |
| IIa | T3N0M0 | 64 | |
| IIb | T4aN0M0 | 5 | |
| III | - | 16 | 60 |
| IIIa | T1-T2N1/NcM0 | 3 | |
| | T1N2aM0 | | |
| IIIb | T3-T4aN1/NcM0 | 21 | |
| | T2-T3N2aM0 | | |
| | T1-T2N2bM0 | | |
| IIIc | T4aN2aM0 | 20 | |
| | T3-T4bN2bM0 | | |
| | T4bN1-N2M0 | | |
| IV | - | 35 | 36 |
| IVa | Any-T Any-N M1a | 1 | |
| CONTROL | - | | 42 |
| NA | - | | 1 |

The correspondence between the AJCC staging and the TCGA staging for COADREAD is noted. 'NA' denotes cases where the stage information is unavailable. Sample sizes are successively aggregated to the parent stage.

https://doi.org/10.1371/journal.pone.0249151.t001

the R package limma [31] for linear modelling of stagewise expression using the complete M-value matrix, with multiple probes per gene (S1 Table in S1 Text).

**(2) Linear modelling with avereps matrix.** This is essentially similar to the above model, except that the input was the 'avereps' matrix, where the methylation of each gene was represented by the average of its M-values across all its probes (S2 Table in S1 Text). Such alternative representations of the methylation data negotiate a tradeoff with respect to information loss and interpretability.

In both the linear models, the controls contributed to the intercept of the design matrix, while the stages were represented as indicator variables [32]. The linear fit was subjected to empirical Bayes adjustment to obtain moderated t-statistics. These results were then used for the stage-differentiated contrast analysis

**(3) Association between methylation status and phentoype.** The strength of the association between the methylation levels of CpG sites and the phenotype of interest (CRC-stage) could enable the identification of relevant markers. We used the R package CpGassoc [33] to estimate this association based on ANOVA with multiple hypothesis correction. The β-matrix was used as input, and five factors (control, stage I, stage II, stage III, stage IV) were specified as the target phenotype.

**(4) The Chip Analysis Methylation Pipeline (ChAMP).** The Chip Analysis Methylation Pipeline (ChAMP) integrative analysis suite uses limma to identify differentially methylated probes (DMPs) from the β-matrix [34]. A mapping of sample IDs with the pathological stage phenotype was provided as an additional input file. In addition, the identification of differentially methylated regions (DMRs), consisting of polygenic genomic blocks, was performed using DMRcate in ChAMP (with preset p-value cutoff$<0.05$) [35]. GSEA was used to identify the enrichment of DMPs and DMRs in the MSigDB pathways [36], using the Fisher Exact test calculation with adjusted p-value $< 0.05$.

**(5) Correlation between gene methylation and expression.** We used MethylMix2.0 to estimate the correlation between the methylation and actual expression patterns of each gene [37]. The expression data for the cases of interest were retrieved from TCGA (gdac.broadinstitute.org_COADREAD.Merge_rnaseqv2_illuminaga_rnaseqv2_unc_edu_Level_3_RSEM_genes_data.Level_3.2016012800.0.0.tar.gz). MethylMix was executed with the preset correlation cutoff ($> |0.3|$), and statistical significance was assessed using Wilcoxon Rank Sum test with adj. p-value $< 0.05$.

**(6) Modelling expression from methylation.** We used the R package BioMethyl to model the aggregate expression level of a gene from its methylation patterns [38]. The gene expression matrix was estimated using the methylation β-matrix and then subjected to linear modelling with limma, followed by stage-differentiated contrast analysis.

## Stage-differentiated contrast analysis

A directed two-tier set of contrasts was performed in limma to drill down to the stage-salient genes:

1. Tier I: Stage-differentiated contrast against controls. Four pairwise contrasts were performed, one for each of the stages I, II, III and IV. To identify reliable DMGs, the following criteria were used: |lfc M-value| >2, and adj. p-value <0.001.
2. Tier II: Inter-stage contrasts. Six pairwise contrasts between the stages (namely: I-II, I-III, I-IV, II-III, II-IV, and III-IV) were performed (p-value for each contrast: <0.05).

To illustrate, a putative DMG identified in Tier I would undergo three inter-stage contrasts in Tier II, to ensure stage-salience. For example, a putative stage-II DMG established by Tier I, would have to pass the following inter-stage contrasts: stage-II vs stage-I, stage-II vs stage-III and stage-II vs stage-IV, for confirmation as stage II-salient DMG.

## Identification of stage-salient biomarkers

Finding the consensus of a set of methods with different algorithms overcomes the biases specific to individual methods, and enables screening out false positives. Consensus was obtained by finding the agreement among the results of the various methods used. At least three methods should agree on a given DMG's stage-salience, for confirmation as *consensus* stage-salient biomarker.

## Survival analysis

The survival data for each case was obtained from the following attributes encoded in the clinical data: patient.vital_status, patient.days_to_followup, and patient.days_to _death. The association between consensus stage-salient DMGs and case overall survival (OS) was evaluated by univariate Cox proportional hazards regression model using the R survival package [39]. This uncovered potential prognostic stage-salient genes from the methylation analysis, using a significance cutoff< 0.05. Such prognostic genes were used as the independent variables in a regression model to estimate the survival risk of each case. Based on this risk score, cases with colorectal cancer were categorized into high and low groups using the optimal cut point determined by the maxstat (maximally selected rank statistic) [40]. Kaplan-Meier estimation was then applied to the median survival times of these two groups for flagging significant differences, providing a prognostic assessment of the biomarkers of interest.

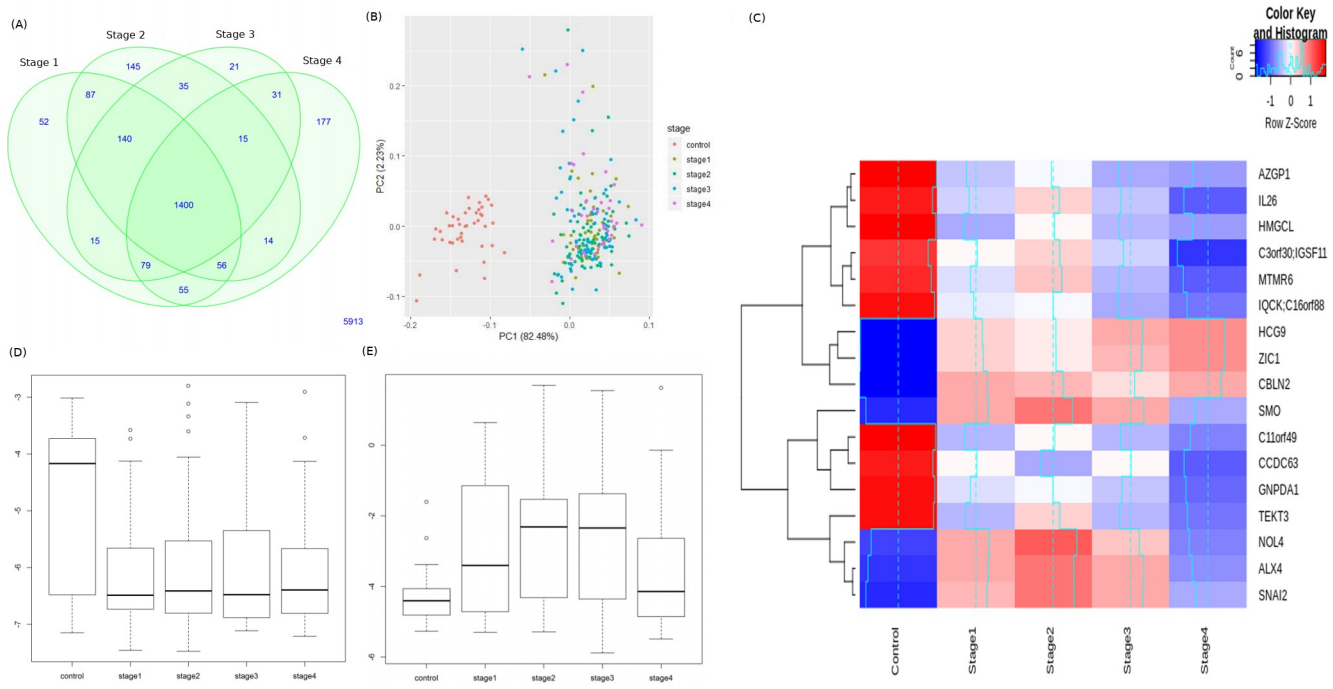


**Fig 1. Linear modelling with M-value matrix, all probes.** (A) Venn distribution of significant DM genes in each stage relative to control. (B) Distribution of samples based on the top two principal components of the top 100 genes shows a clear separation of cancer cases (labelled by stage) from controls. (C) Stagewise methylation portraits of the top four significant stage-specific DMGs. The contrast with the control is especially evident. Also shown are the stagewise methylation levels of (D) TMEM179, and (E) MEOX2.

https://doi.org/10.1371/journal.pone.0249151.g001

## Results

### Linear modelling with M-values (at the probe-level)

The number of significant genes present in each stage-control pair from the Tier-I contrasts is shown in Fig 1A. Using the top 100 DM genes of the linear model (given in S3 File in S1 Text), we found a clear separation between controls and stage samples (Fig 1B). The top genes in each stage (by adjusted p-value of contrast with control) are shown in Table 2, with |lfc M-

**Table 2. Top ten genes of the linear model at the probe level.**

| ID | Stage I lfc ($\beta_1$) | Stage I lfc ($\beta_2$) | Stage III lfc ($\beta_3$) | Stage IV lfc ($\beta_4$) | Adj. P-val | Methylation status |
|---|---|---|---|---|---|---|
| GPR75-ASB3 | 2.28 | 2.19 | 2.16 | 2.32 | 1E-82 | Hyper |
| TM4SF19 | -3.63 | -3.58 | -3.72 | -3.71 | 1E-82 | Hypo |
| CNRIP1 | 2.74 | 2.60 | 2.68 | 2.97 | 1E-78 | Hyper |
| PDE4A | 1.68 | 1.58 | 1.60 | 1.71 | 1E-71 | Hyper |
| KRTAP11-1 | -2.36 | -2.30 | -2.37 | -2.40 | 1E-70 | Hypo |
| ADHFE1 | 3.15 | 2.97 | 3.00 | 3.43 | 1E-69 | Hyper |
| FAM123A | 3.56 | 3.18 | 3.43 | 3.90 | 1E-69 | Hyper |
| KHDRBS2 | 2.30 | 2.16 | 2.10 | 2.34 | 1E-68 | Hyper |
| AJAP1 | 2.52 | 2.44 | 2.46 | 2.64 | 1E-68 | Hyper |
| NALCN | 2.96 | 2.80 | 2.94 | 3.25 | 1E-68 | Hyper |

The log fold-change of M-value of the probe in each stage relative to the controls, followed by p-value adjusted for the false discovery rate, and the methylation status of the gene in the cancer stages with respect to the control.

https://doi.org/10.1371/journal.pone.0249151.t002

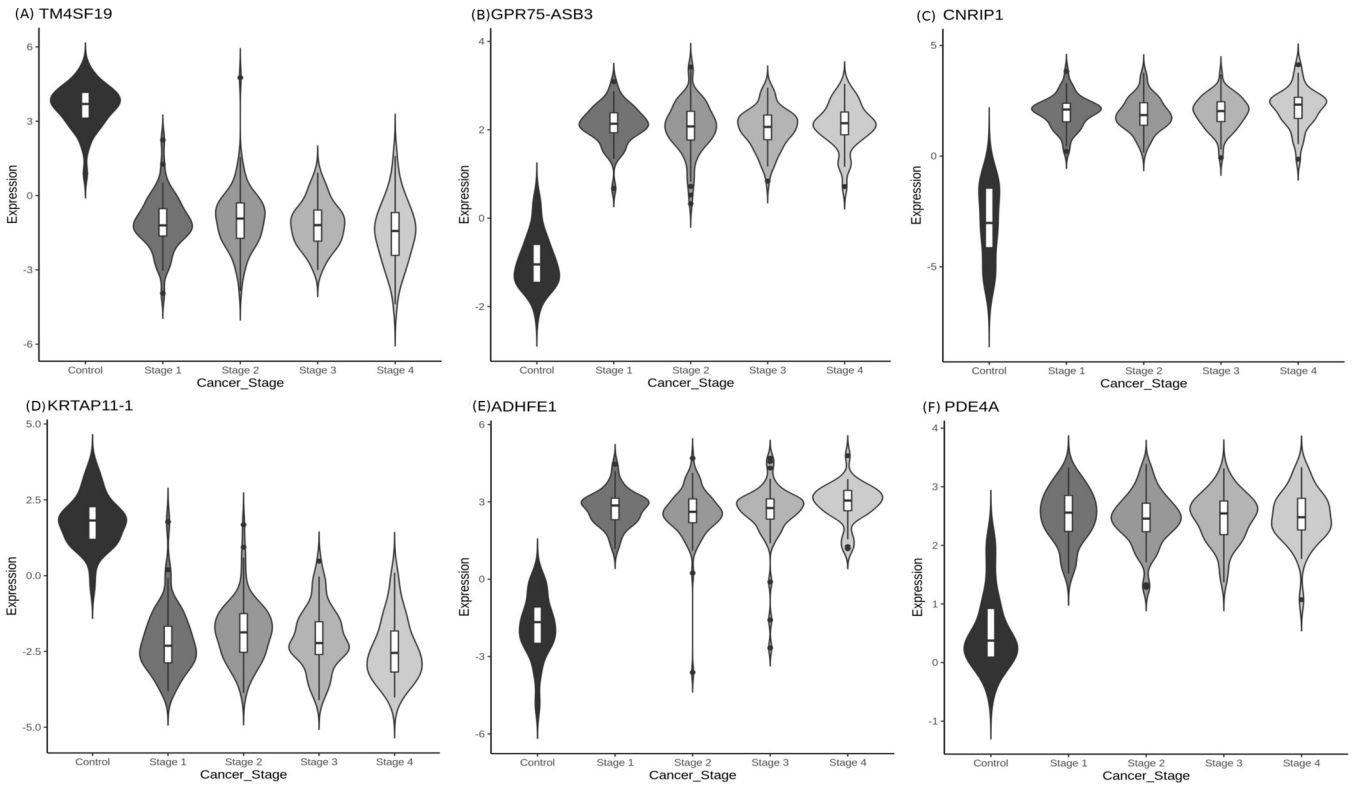


**Fig 2. Top DMGs identified from linear modelling.** (A) GPR75-ASB3, (B) TM4SF19, (C) CNRIP1, (D) KRTAP11-1, (E) ADHFE1 and (F) PDE4A. For each gene, notice that the trend in methylation could be either hyper-or hypo-methylation relative to the control. TM4SF19 and KRTAP11-1 are hypomethylated whereas CNRIP1, GPR75-ASB3, ADHFE1, PDE4A are hypermethylated.

https://doi.org/10.1371/journal.pone.0249151.g002

value| and inferred regulation status. The top four genes of each stage were used to construct a stagewise methylation heatmap (Fig 1C). Fig 1D and 1E show boxplots of stagewise methylation levels for two representative genes: TMEM179, mutations in which could cause MSI [41]; and MEOX2 whose promoter methylation status is a known CRC marker [42]. The stagewise methylation patterns of the top linear model genes are shown in Fig 2. It is notable that a naturally occuring read-through fusion protein GPR75-ASB3 is the top linear model gene with significant differential expression in all stages relative to the control. GPR75-ASB3 is positively differentially expressed in the lung as well as different keratinocyte cell types, and evidence is emerging of its role in other cancers [43]. In this light, GPR75-ASB3 could play a significant role in colorectal cancers which are of epithelial origin. The top 100 significant stage-specific genes, listed in S3 File in S1 Text, were used in the consensus analysis.

## Linear modelling with avereps matrix (at the gene-level)

The methylation levels of genes with multiple probes were averaged using limma's avereps function, and summarized to one value. The number of genes present in each stage-control pair from the Tier-I contrasts is shown in Fig 3A. Using the top 100 genes of the linear model (given in S4 File in S1 Text), we found a clear separation between controls and stage samples (Fig 3B). The top genes in each stage (by adjusted p-value of contrast with control) are shown in Table 3, with |lfc M-value| and inferred regulation status. The top four genes of each stage were used to construct a stagewise methylation heatmap (Fig 3C). Fig 3D and 3E shows the boxplots of stagewise methylation levels for two representative genes, NALCN and GLRX.

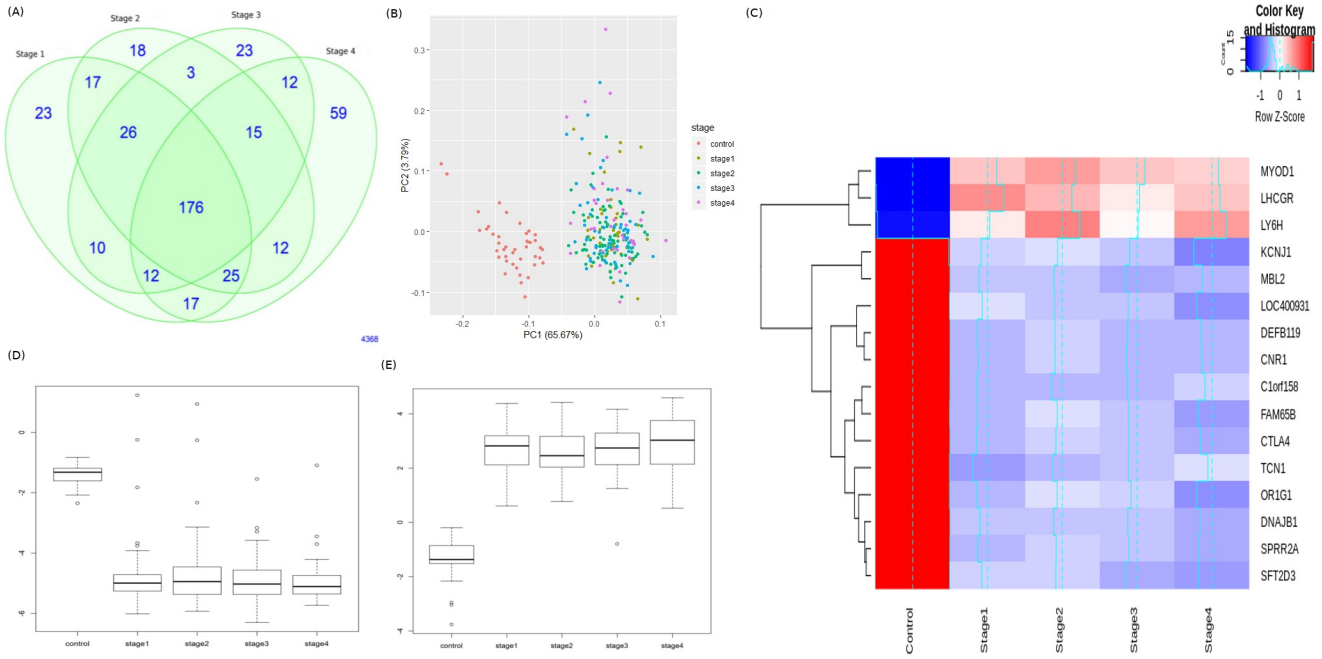


**Fig 3. Linear modelling with M-value matrix, avereps transformation.** (A) Venn distribution of significant DM genes in each stage relative to control. (B) Distribution of samples based on the top two principal components of the top 100 genes shows a clear separation of cancer cases (labelled by stage) and controls. (C) Stagewise methylation portraits of the top four significant stage-specific DMGs. The stark contrast with the control is especially evident. Also shown are the stagewise methylation levels of (D) NALCN, and (E) GLRX.

https://doi.org/10.1371/journal.pone.0249151.g003

Mutations in NALCN have been reported in sporadic CRC [44]; here NALCN is seen to be significantly hypermethylated, indicating the same outcome (loss of function) could be effected in multiple ways. GLRX is a target of the activating transcription factor MEOX2 [45]. It is observed that *LY6H* showed both hypermethylation and hypomethylation when compared to the controls, indicating the role of experimentation necessary to clarify its role in colorectal cancer progression. The top significant 100 genes of each stage, listed in S4 File in S1 Text, were used for the consensus analysis.

**Table 3. Top ten genes of linear modelling with averaging of multiple probes.**

| ID | Stage I lfc ($\beta_1$) | Stage I lfc ($\beta_2$) | Stage III lfc ($\beta_3$) | Stage IV lfc ($\beta_4$) | Adj. P-val | Methylation status |
|---|---|---|---|---|---|---|
| TM4SF19 | -3.63 | -3.57 | -3.72 | -3.71 | 1E-82 | Hypo |
| GPR75-ASB3 | 2.28 | 2.19 | 2.15 | 2.32 | 1E-82 | Hyper |
| CNRIP1 | 2.74 | 2.60 | 2.67 | 2.97 | 1E-77 | Hyper |
| KRTAP11-1 | -2.36 | -2.30 | -2.38 | -2.40 | 1E-70 | Hypo |
| ADHFE1 | 3.15 | 2.96 | 2.99 | 3.43 | 1E-69 | Hyper |
| FAM123A | 3.56 | 3.18 | 3.42 | 3.89 | 1E-68 | Hyper |
| AJAP1 | 2.53 | 2.44 | 2.46 | 2.64 | 1E-67 | Hyper |
| NALCN | 2.96 | 2.79 | 2.95 | 3.25 | 1E-65 | Hyper |
| IRF4 | 1.99 | 1.83 | 1.89 | 2.13 | 1E-65 | Hyper |
| PRKAR1B | 3.38 | 3.13 | 3.24 | 3.50 | 1E-65 | Hyper |

The log fold-change of M-value of the gene in each stage (relative to the control) is given, followed by p-value adjusted for the false discovery rate and the methylation status of the gene in the cancer stages with respect to the control. A consistent methylation pattern is observed for all the top genes.

https://doi.org/10.1371/journal.pone.0249151.t003

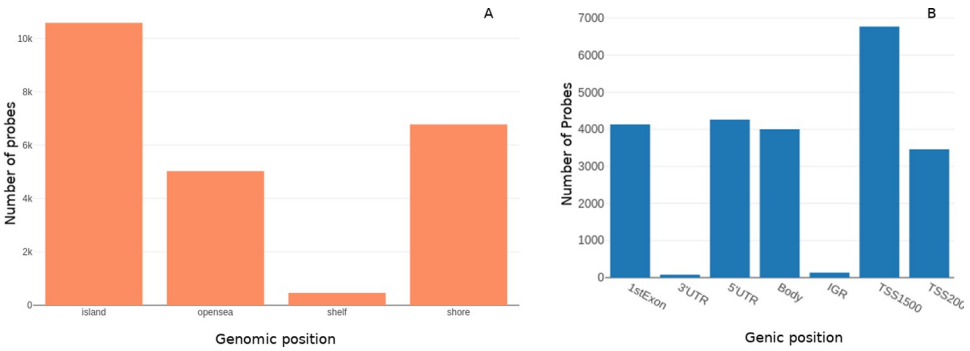


**Fig 4.** Distribution of probes based on (A) genomic position: opensea, shore, island, shelf; (B) gene context: transcription start site (TSS), exons, untranscribed regions (UTRs), and intergenic regions (IGR).

https://doi.org/10.1371/journal.pone.0249151.g004

## Association with phenotype

The ANOVA from CpGassoc yielded p-values and log fold-changes, which were used to identify significant genes for each stage using the criteria given in Methods. The top 100 genes of each stage from this analysis (given in S5 File in S1 Text) were used for the consensus investigation.

## DMP analysis with ChAMP

The summary features of the β matrix dataset were evaluated using ChAMP (Fig 4). The DMPs were identified using ChAMP analysis from the β matrix. All the inter-stage contrasts yielded null results (i.e, no significant genes), except for stageII–stageIV contrast. Due to this, the top 100 DMPs from the stage vs control contrasts were used for the consensus analysis directly. Contrasts that showed significant DMPs were subjected to a further DMR analysis, to enable identification of DM genes. The stage-salient DMR regions (genes) determined are provided in S6 File in S1 Text, and summarized in Table 4. The stage-II vs stage-IV DMR contrast yielded three genes, namely PLAG1, SOCS2, and NNAT. It is observed that these genes might be critical players in the transition to malignancy. Interestingly, some genes were differentially methylated in all the stagewise contrasts with the control; such genes are differentially methylated agnostic of stage and could serve as valuable drug targets for CRC therapy. The top such genes included EYA4, WT1, DCC, RP11, GATA4, MSX1, DLX5, BNC1, WT1-AS, and ZIM2. A total of 31 such genes were identified and tabulated in S7 Table in S1 Text. The DMPs and DMRs from the analysis were subjected to GSEA and these results could also be found in S6 File in S1 Text. Fig 5 shows representative DMP and DMR plots using ChAMP.

**Table 4. Contrast-wise counts of DM probes and DM regions.**

| Contrast | DMPs | DMRs |
|---|---|---|
| Control and Stage 1 | 11045 | 34 |
| Control and Stage 2 | 11254 | 35 |
| Control and Stage 3 | 11254 | 36 |
| Control and Stage 4 | 11108 | 34 |
| Stage 2 and Stage 4 | 404 | 3 |

No DM regions were found for the contrasts not shown, namely the stage-pairs: [(1,2), (1,3), (1,4), (2,3), (3,4)].

https://doi.org/10.1371/journal.pone.0249151.t004

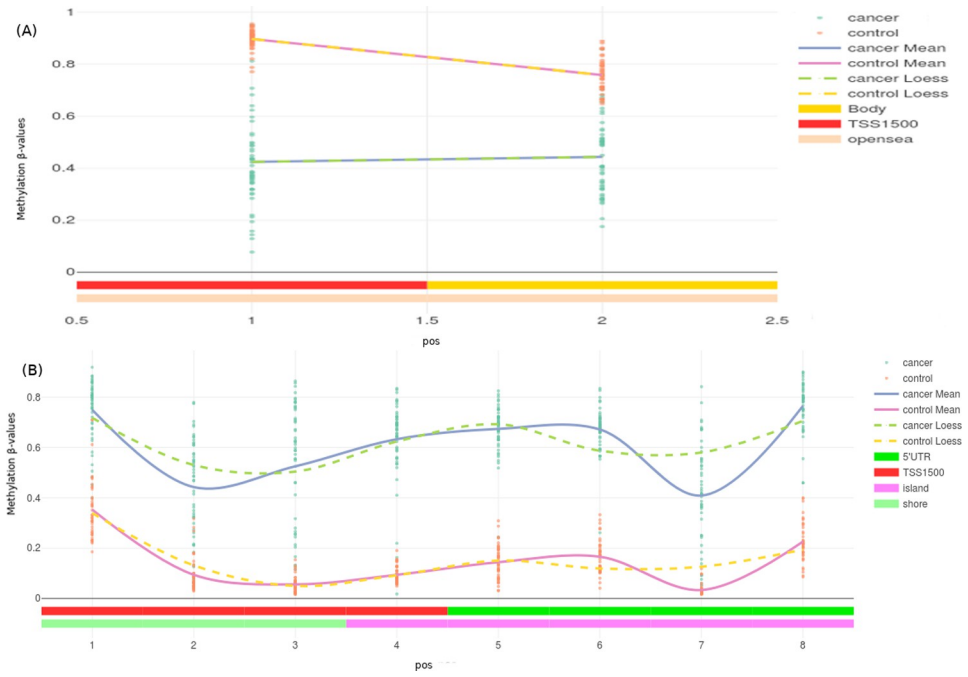


**Fig 5. DMP and DMR plots using ChAMP.** (A) DMP plot of FCN2 for stage-I vs control illustrating significant hypomethylation (B) DMR plot of transcriptional activator EYA4 for stage-I vs control illustrating significant hypermethylation. Solid lines represent mean values while dashed lines represent the loess.

https://doi.org/10.1371/journal.pone.0249151.g005

## Methylation and expression correlation analysis

Differential methylation (DM) calculated from stage vs control contrasts ranged from -0.7 to +0.8, and genes could be hyper- or hypo-methylated based on the sign of the DM value. There were 209, 441, 275, and 134 driver genes in each of the contrasts with the controls (stage-I, stage-II, stage-III and stage-IV, respectively). All between-stages contrasts yielded null DM genes. The results from this analysis, including driver genes for all the contrasts, are provided in S8 File in S1 Text. It is notable that the top genes from an overall cancer vs control comparison included GATA4, CCDC88B, and WAS. Top 100 genes from each comparison with the controls were taken forward for consensus analysis. Certain genes emerged common to all the four comparisons with the controls, thereby suggesting stage-agnostic differential methylation events. The top such stage-agnostic differentially methylated genes included *CCDC88B*, *C1orf59*, *CHFR*, *ZP2*, *HOXA9*, *ELF5*, *FAM50B*, *MUC17*, *TBX20*, and *VSIG2*. Stage-agnostic genes hold promise as therapeutic targets for the treatment of colorectal cancer; the complete set of 56 stage-agnostic genes identified in this analysis is provided in S9 Table in S1 Text. Mixture models of genes, indicative of the number of methylation states, were constructed using MethylMix, and illustrated for a few stage-IV driver genes in Fig 6. The estimated correlation between the methylation levels and actual gene expression for the same genes shows the inverse relationship between methylation and gene expression, thereby highlighting the effect of epigenetic events (Fig 6).

## BioMethyl analysis

The significant stage-specific DEGs identified by BioMethyl are shown in UpSet plot [46] (Fig 7), and provided in S10 File in S1 Text. Top 100 genes of each stage from this analysis were taken forward for consensus analysis.

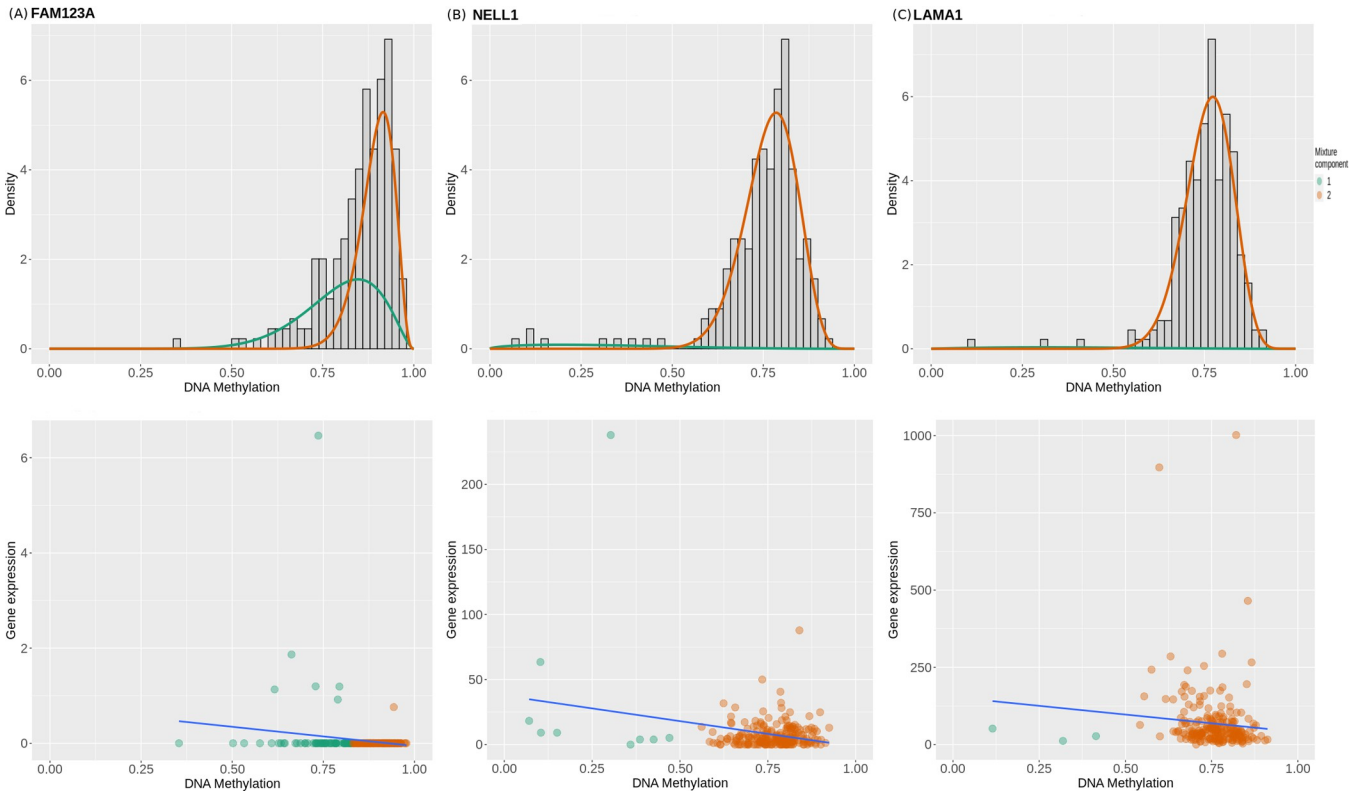


**Fig 6.** Mixture models and Correlation plots for (A) FAM123A, (B) LAMA1 and (C) NELL1. The x-axis indicates the level of methylation (in terms of β values); y-axis, the frequency. Mixture component curves represent density fits of the histogram. A negative correlation between methylation and expression is evident, indicating that methylation acts to repress gene transcription, though the strength of the inverse correlation varies from gene to gene. olour indicates the mixture model fit.

https://doi.org/10.1371/journal.pone.0249151.g006

## Stage-salient consensus biomarkers

The top 100 significantly differentially-expressed genes of each stage from all the methods discussed above (collated in S11 File in S1 Text) were used for the consensus determination. The consensus analysis yielded seven stage-salient DMGs: one stage-I gene *(FBN1)*, one stage-II gene *(FOXG1)*, one stage-III gene *(HCN1)* and four stage-IV genes *(NELL1*, *ZNF135*, *FAM123A*, *LAMA1)*. Each of these stage-salient genes presented an |lfc M-value| $> 0.4$ with respect to the other stages, validating their salience. Fig 8 represents violin plots of the consensus biomarkers, and Table 5 presents a summary of the consensus analysis. Gene ontology (GO) analysis [47] of the consensus biomarkers yielded processes related to structural integrity of cell division processes, immunity dysfunction, and cell migration (Table 6). Detailed GO results are presented in the S12 File in S1 Text.

## Survival analysis

We constructed independent prognostic models of the stage-salient genes and identified the prognostically significant biomarkers as FBN1, FOXG1, HCN1, and LAMA1. The corresponding univariate Kaplan-Meier plots are shown in Fig 9. Rational combinations of stage-salient genes, termed ColoRectal cancer Signatures (CRS), were modelled using multivariate Kaplan-Meier regression, to yield a risk score. Risk scores were then used to estimate survival-effect significance, as described in Methods. The results of this exercise are summarised in Table 7. We found that CRS12 signature (consisting of FBN1 and FOXG1) yielded significant risk

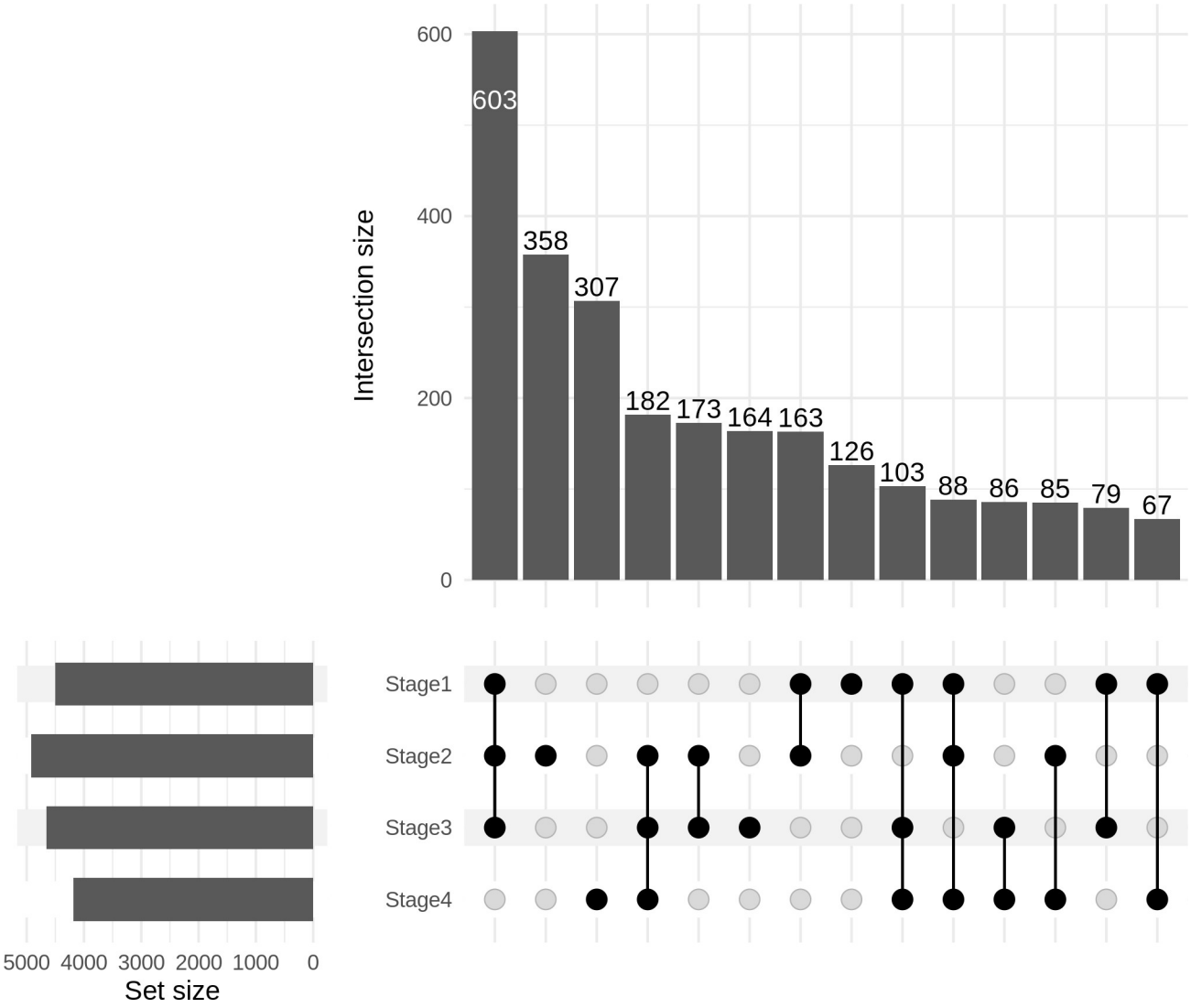


**Fig 7. UpSet plot of BioMethyl-based stagewise gene expression modelling.** The intersection of all stages yielded 3268 genes, which represent consistently differentially regulated genes.

https://doi.org/10.1371/journal.pone.0249151.g007

scores in the multivariate Kaplan-Meier analysis, and both CRS12 and CRS34 (consisting of HCN1, NELL1, ZNF135, FAM123A, LAMA1) were significant in estimating overall survival (prognosis p-value $\leq 0.02$) (Fig 10). S13 File in S1 Text provides survival plots of all possible signatures. At the end of our analysis pipeline, CRS12 passed all the filters and emerged as a significant early-stage panel for CRC prognosis.

## Discussion

CRC development is due to the accumulation of genetic and epigenetic changes of which DNA methylation is of paramount importance. DNA methylation profiles of colorectal cancer have been investigated in several previous studies using various approaches [48, 49]. It is well-known that changes in methylation status correspond with CRC progression [50]. Here we have designed a comprehensive approach to systematically analyze stage-differentiated DNA methylation patterns in colorectal cancer and their relationship to patient survival. Our study

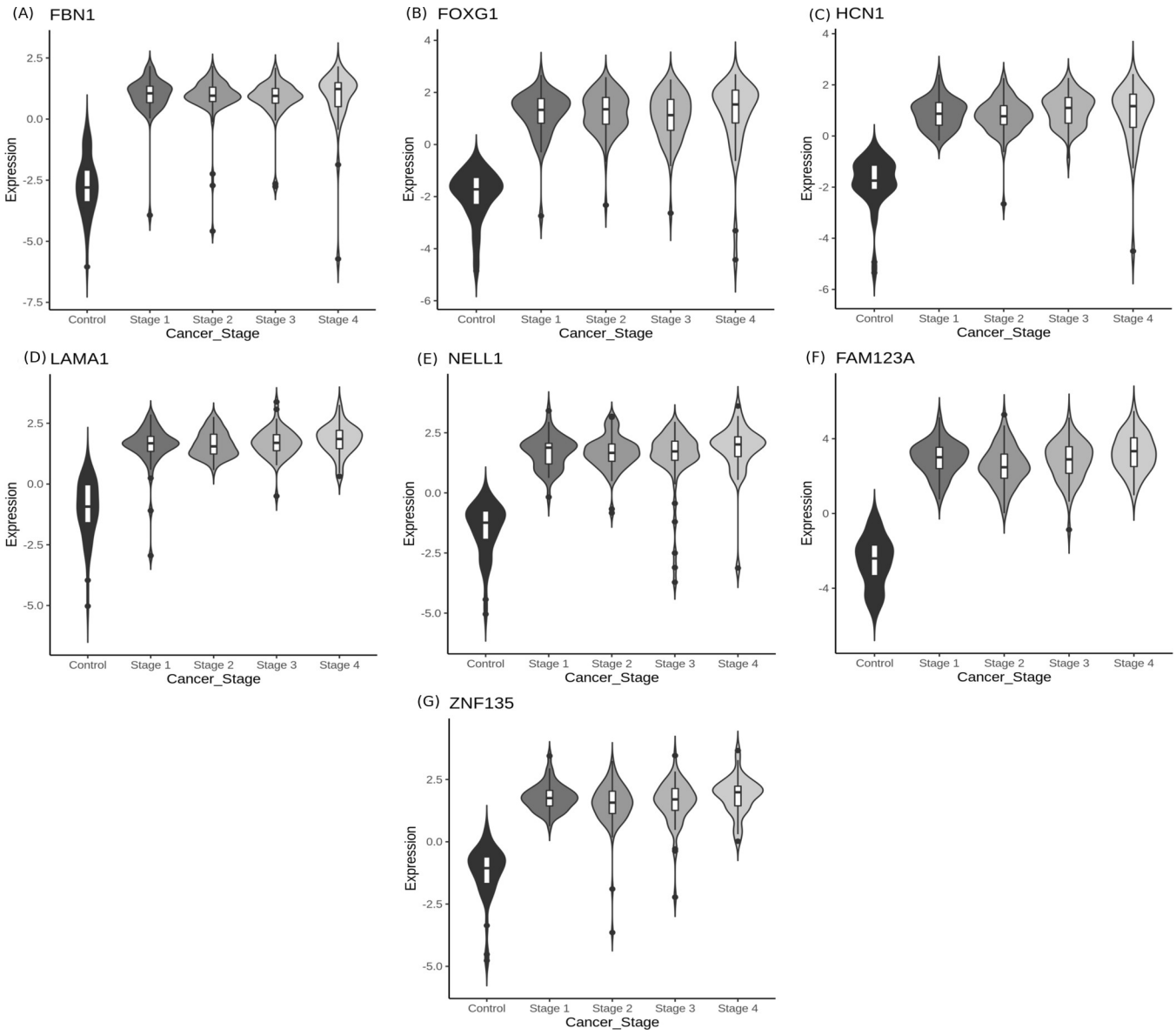


**Fig 8. Violin plots of stage-salient genes.** (A) Stage-I Gene FBN1, (B) Stage-II Gene–FOXG1, (C) Stage-III Gene–HCN1 and Stage-IV genes (D) LAMA1, (E) NELL1, (F) FAM123A, (G) ZNF135.

https://doi.org/10.1371/journal.pone.0249151.g008

has yielded consensus stage-salient significantly differentially methylated genes, and evaluated their prognostic value. Corollary insights obtained in the course of our investigations, such as stage-agnostic genes, have been documented, and would also be of interest to researchers in the field. It is significant that none of the stage-salient genes figure as a cancer gene or a hallmark gene in the Cancer Gene Census [51]; HCN1 is notably marked as a candidate cancer gene based on mouse insertional mutagenesis experiments [52]. The dominant differentially methylated CpG site in all the stage-salient genes is located within the core / proximal promoter regions (Table 8). Mixture models of methylation levels of stage-salient genes, along with their inverse correlation to corresponding expression levels are shown in Fig 11, and unambiguously establish the epigenetic impact of the changes in methylation. Our findings are

**Table 5. Stage-salient biomarkers.**

| HGNC ID | Gene Name | Methods in agreement | Salience | Meth. status | Statistical significance | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | M value | Avereps | Cox analysis | Kaplan Meier |
| 3603 | FBN1 | Avereps, ChAMP | I | Hyper | 0.310 | 0.040 | 0.036 | 0.025 |
| 3811 | FOXG1 | Mvalue, Avereps, ChAMP, Methylmix | II | Hyper | 1E-16 | 0.003 | 0.019 | 0.037 |
| 4845 | HCN1 | Mvalue, Avereps, ChAMP | III | Hyper | 1E-17 | 0.022 | 0.031 | 0.059 |
| 7756 | NELL1 | Mvalue, Avereps, ChAMP | IV | Hyper | 1E-68 | 0.061 | 0.283 | 0.27 |
| 12919 | ZNF135 | Mvalue, ChAMP, Methylmix | IV | Hyper | 1E-76 | 0.062 | 0.096 | 0.084 |
| 26360 | FAM123A | Mvalue, ChAMP, Methylmix | IV | Hyper | 1E-115 | 0.097 | 0.30 | 0.28 |
| 6481 | LAMA1 | Mvalue, ChAMP, Methylmix | IV | Hyper | 1E-86 | 0.297 | 0.052 | 0.051 |

The results of the consensus analysis and univariate survival analysis are summarized. All the biomarkers showed hypermethylation, reflecting downregulation of gene expression.

https://doi.org/10.1371/journal.pone.0249151.t005

further discussed in the context of the existing literature, and lead us to detect a strange CpG island methylator phenotype (CIMP) signature in colorectal cancer.

## Stage-salient DMGs

Promoter hypermethylation of FBN1, a glycoprotein component of calcium-binding extracellular matrix microfibrils [53], is a recognized biomarker of CRC [54, 55]. Our analysis supports this literature, while pinpointing the stage I-salience in its action. FOXG1 is well-known as an etiological factor in certain neurological disorders and plays a role in the epithelial-mesenchymal transition of CRC cells (a key hallmark of cancer progression), and is known to be overexpressed in CRC cases [56]. It is a nodal gene, with connections to oncogenic pathways like WNT pathway in hepatocellular carcinoma [57] and TGF-β pathway in ovarian cancer [58] Interestingly, FOXG1 was found to be a hypermethylated stage-II salient gene. HCN1, coding for hyperpolarization-activated cyclic nucleotide-gated channel subunits, is associated with low survival rates in breast, brain, and colorectal cancer [59]. We have identified HCN1 as a stage-III hypermethylated gene, suggesting a loss-of-function mechanism for its tumorigenic potential.

Our study has provided clear evidence that hypermethylation of LAMA1 (which codes for α-laminin of the extracellular matrix) is a stage IV-specific signature. Experimental evidence

**Table 6. GO analysis of stage-salient genes in the order of decreasing significance (i.e, increasing p–value).**

| GO ID | Term | Ontology | p-value |
|---|---|---|---|
| GO:1990047 | Spindle matrix | CC | 0.0001 |
| GO:0030109 | HLA-B specific inhibitory MHC class I receptor activity | MF | 0.0003 |
| GO:0032396 | Inhibitory MHC class I receptor activity | MF | 0.006 |
| GO:0042609 | CD4 receptor binding | MF | 0.0012 |
| GO:0032393 | MHC class I receptor activity | MF | 0.0013 |
| GO:0050930 | Induction of positive chemotaxis | BP | 0.0016 |
| GO:0050927 | Positive regulation of positive chemotaxis | BP | 0.0033 |
| GO:0050926 | Regulation of positive chemotaxis | BP | 0.0034 |
| GO:0008608 | Attachment of spindle microtubules to kinetochore | BP | 0.0043 |
| GO:0007094 | Mitotic spindle assembly checkpoint | BP | 0.0044 |

Ontology could be Cellular Compartment (CC), Molecular Function (MF), or Biological Process (BP).

https://doi.org/10.1371/journal.pone.0249151.t006

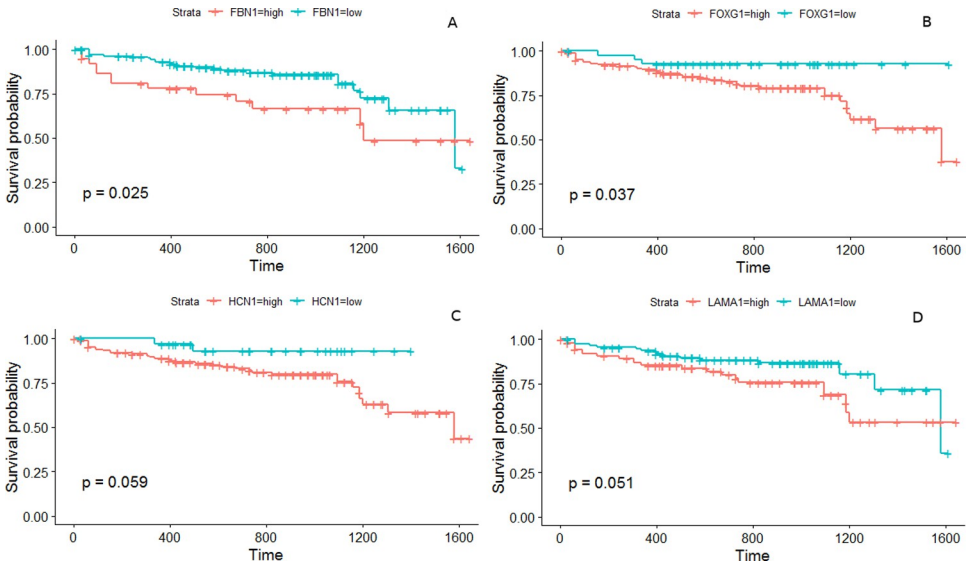


**Fig 9. K-M plots for the prognostically significant stage-salient genes.** (A) FBN1, (B) FOXG1, (C) HCN1, and (D) LAMA1.

https://doi.org/10.1371/journal.pone.0249151.g009

for the hypermethylation of the promoter region of LAMA1 in CRC cases is available [60]. NELL1 is a known tumor suppressor gene [61], whose hypermethylation is associated with poor survival outcomes [62]. Here it is found to be a stage IV-specific hypermethylated gene, resonating with the above findings. ZNF135 is a zinc-finger protein involved in regulation of cell morphology and cytoskeletal organizations. Its expression and epigenetic regulation have been reported to be key in cancers of the cervix and esophagus, respectively [63, 64]. Here we have found that epigenetic silencing of ZNF135 is a key feature of stage-IV CRC. It is interesting that another member of the zinc-finger protein family, ZNF726, has been recently identified as the only methylated gene significantly associated with OS in patients with CRC,

**Table 7. Summary of selected multivariate prognostic models.**

| Signature | Stages | Biomarker | Weight | P-value | |
|---|---|---|---|---|---|
| | | | | Multivariate model | Prognosis |
| CRS12 | I+II | FBN1 | -0.62 | **0.015** | **0.005** |
| | | FOXG1 | -1.05 | | |
| CRS34 | III+IV | NELL1 | 0.1 | 0.172 | 0.02 |
| | | ZNF135 | -0.21 | | |
| | | FAM123A | -0.23 | | |
| | | LAMA1 | -0.39 | | |
| | | HCN1 | -1.1 | | |
| CRS234 | II+III+IV | FOXG1 | -0.99 | 0.0877 | 0.032 |
| | | HCN1 | -1.07 | | |
| | | NELL1 | -0.10 | | |
| | | ZNF135 | -0.22 | | |
| | | FAM123A | -0.37 | | |
| | | LAMA1 | -0.27 | | |

Weight denotes the coefficient in the multivariate model. The ultimate significant signature is highlighted.

https://doi.org/10.1371/journal.pone.0249151.t007

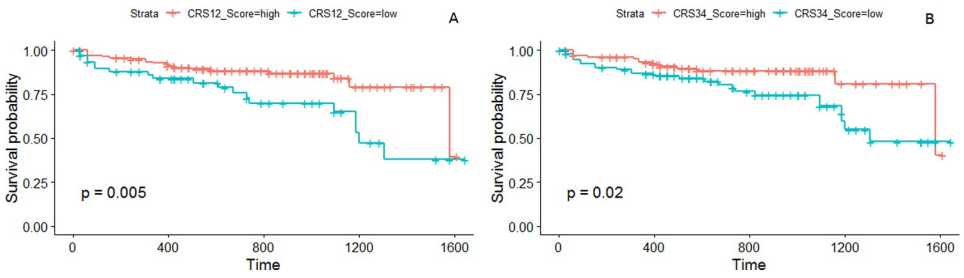


**Fig 10. Survival analysis of combination biomarker panels shows significance.** (A) Early-stage panel, CRS12; and (B) Late-stage panel, CRS34.

https://doi.org/10.1371/journal.pone.0249151.g010

without regard for pathologic stage [65]. FAM123A, also known as AMER2, is associated with microtubule proteins [66], and is a paralog of the well-documented FAM123B, a tumor-suppressor whose loss-of-function by mutation, methylation and copy-number aberrations is known to play a pivotal role in colorectal cancer, especially in older patients [67–69]. It is significant that our study has uncovered FAM123A as a hypermethylated stage IV-specific DMG, signalling the need for experimental investigations. There is very little literature on the cancer significance of any of the above stage-salient genes, marking our findings as novel and important in the context of gaps in our knowledge.

## Putative CIMP signature

Aberrant methylation of CpG promoter regions causes stable repression of transcription leading to gene-silencing [70, 71]. In the context of tumorigenic processes, this is likely to lead to loss-of-function of tumor-suppressor genes. Multiple CpG islands might be methylated simultaneously in some cancers, paving the way for CpG island methylator phenotype (CIMP), first discovered in colorectal cancer [72]. CIMP is characterised by hypermethylation of CpG islands surrounding the promoter regions of genes involved in cancer onset and progression [73]. The phenotype is heterogenous with the type of tumor [74] and dependent on definition [75]. Table 8 suggests that the stage-salient hypermethylated biomarkers identified in our study are components constituting an aggregate novel CIMP, and there is preliminary experimental evidence in this direction. Earlier studies have identified LAMA1 as a CIMP panel constituent [50, 60]. FBN1 has been used as an epigenetic biomarker in diagnostic panels associated with CIMP-positive tumors [54, 76]. While this paper was under review, FAM123A has been used in a five marker panel to detect stage-IV CRC using blood samples [77]. The original CIMP had been associated with advanced T staging (T3/T4) [78], which accords with

**Table 8. Location of the major DM CpG site in stage-salient genes.**

| Stage-salient gene | DM CpG site | Distance to TSS | Location in the promoter region |
|---|---|---|---|
| FBN1 | cg18671950 | 146 | Proximal |
| FOXG1 | cg10300684 | 36 | Core |
| HCN1 | cg06498267 | 298 | Proximal |
| NELL1 | cg17371081 | 179 | Proximal |
| ZNF135 | cg16638540 | 144 | Proximal |
| FAM123A | cg22029275 | 73 | Core |
| LAMA1 | cg07846220 | 133 | Proximal |

All the hypermethylated CpG sites of stage-salient DMGs were found in the core/proximal promoter regions.

https://doi.org/10.1371/journal.pone.0249151.t008

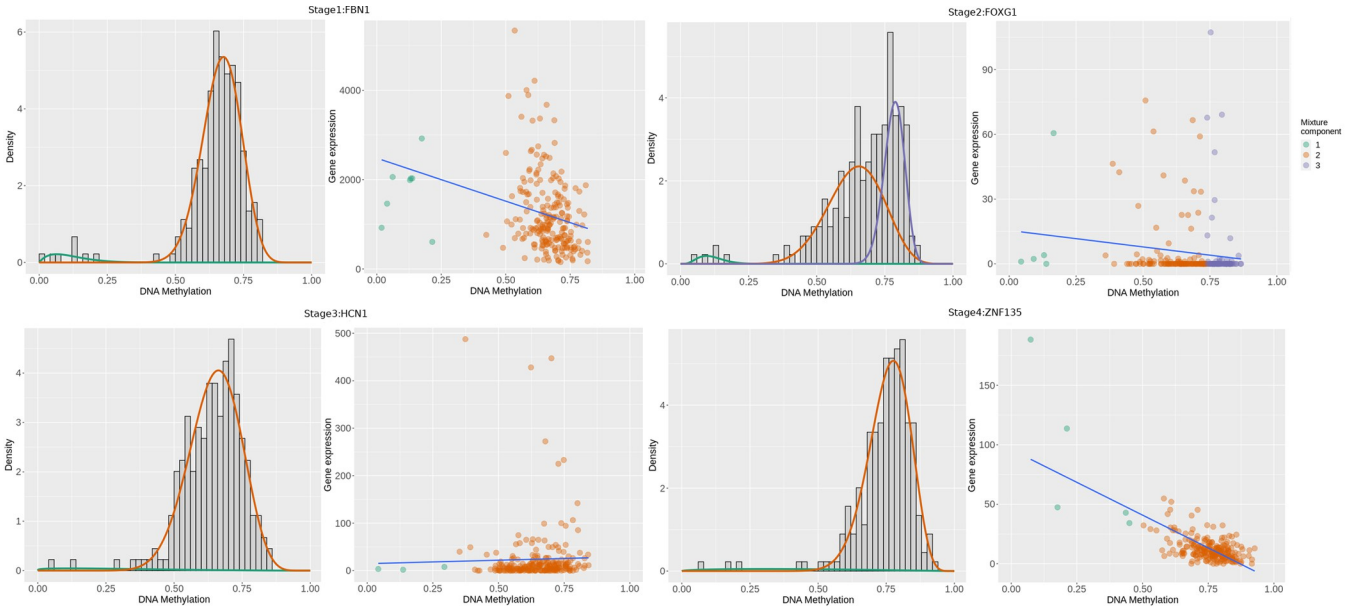


**Fig 11. Mixture models and correlation plots of stage-salient genes.** Shown are FBN1, FOXG1, HCN1, and ZNF135. Two mixture components are seen for FBN1, HCN1, and ZNF135, and three for FOXG1. A strong inverse correlation exists for all genes, except HCN1. Other stage-salient genes are shown in Fig 6.

https://doi.org/10.1371/journal.pone.0249151.g011

our finding of four hypermethylated stage IV-salient DMGs. The biomarkers from our study contributing to the putative CIMP were tested with a standard survival analysis workflow yielding significant prognostication power for five of the seven stage-salient genes (Table 5). A Cox multivariate analysis of biomarker panels uncovered two signatures, an early-stage CRS12, and a late-stage CRS34 that might be prognostically valuable. In particular, CRS12 (composed of FBN1 and FOXG1) suggests a significant early-stage biomarker panel (p-value $< 0.01$) for the effective prognosis and stage-sensitive detection of colorectal cancer.

Diagnostic biomarkers that are also superior in prognostication power imply methylation events that are vital to tumor-specific pathophysiology. This suggests future directions for therapeutic intervention. Epigenetic intervention for CIMP-positive cancers has been advanced as a possible treatment strategy [79]. The alternative CIMP-like biomarkers could serve to stratify the cancer subtype, thereby facilitating precision medicine. The current standard of CRC screening is colonoscopy, an invasive method with a significant rate of complications. A non-invasive method based on molecular diagnostics would improve patient satisfaction and efficiency. Several studies have been conducted to identify and/or validate biomarkers for CRC diagnosis. It is recognized that DNA methylation patterns could serve as valid biomarker candidates [80, 81]. Freitas et al., have validated the performance of a 3-gene biomarker panel for the detection of colorectal cancer irrespective of the molecular subtype [82]. However optimal stage-salient epigenetic biomarkers have not yet been reported. Using hypermethylated DNA patterns as cancer markers offers the advantage of providing small targets with high concentrations of CpG for assays, useful for the design of analytical amplicons [83]. Hypermethylation in the gene body and upstream control regions like enhancers and insulators might affect transcription differently than hypermethylation of promoter regions [84, 85]. Further DNA methylation patterns in noncoding RNA genes seem to be important in tumorigenesis and progression [86]. Non-coding RNAs themselves play a significant role in epigenetic modification through the phenomenon of RNA-directed DNA methylation [48]. The nuanced relationship between methylation and gene transcription signals the need for clinical validation of our

results, however ensemble approaches such as the one used here suffer less uncertainties with respect to translation of the identified biomarkers. Since methylation mediates a direct epigenetic regulatory mechanism used by all life [87], it is hoped that the workflow herein designed would advance our understanding of the complex effects of methylation events, patterns, and landscapes in different settings, including in the developmental stages of life.

## Conclusion

We have developed a comprehensive computational framework for the consensus identification of stage-differentiated significant differentially methylated genes, and evaluation of their prognostic significance. Our analysis has yielded seven stage-salient genes, all hypermethylated in the promoter regions and relatively unreported in the literature: one stage-I gene *(FBN1)*, one stage-II gene *(FOXG1)*, one stage-III gene *(HCN1)* and four stage-IV genes *(NELL1*, *ZNF135*, *FAM123A*, *LAMA1)*. Stage-salient genes could serve as diagnostic biomarkers, and their concordant hypermethylation would signal a distinct CIMP-like character possibly promoting epigenetic destabilisation, which in turn would drive the progression of colorectal cancer. These findings lend further evidence to CIMP drivers of colorectal cancer and point more generally to a pervasive role for these aberrations in tumor biology that remains to be discovered. Independent prognostic evaluation of the stage-salient markers yielded significance for FBN1 and FOXG1. Survival analysis of biomarker signatures composed of the stage-salient genes yielded a significant early-stage panel consisting of FBN1 and FOXG1. Our studies have also spawned secondary results such as stage-agnostic genes that could serve as targets for drug discovery in CRC therapy. Consensus approaches, like the one used here, are more reliable, and the epigenetic biomarkers identified in our study could potentially advance the accurate early detection of colorectal cancers, their treatment and prognostic evaluation. The methods are extendable to the investigation of epigenomics in other cancers, normal/disease conditions, and developmental biology.

## Supporting information

**S1 Text.**
(TXT)

## Acknowledgments

We are grateful to the School of Chemical and Biotechnology, SASTRA Deemed University for computing and infrastructure support.

## Author Contributions

**Conceptualization:** Ashok Palaniappan.

**Data curation:** Abirami Raghavendran.

**Formal analysis:** Sangeetha Muthamilselvan, Abirami Raghavendran.

**Funding acquisition:** Ashok Palaniappan.

**Investigation:** Sangeetha Muthamilselvan, Ashok Palaniappan.

**Methodology:** Abirami Raghavendran, Ashok Palaniappan.

**Project administration:** Ashok Palaniappan.

**Resources:** Ashok Palaniappan.

**Software:** Sangeetha Muthamilselvan, Abirami Raghavendran, Ashok Palaniappan.

**Supervision:** Ashok Palaniappan.

**Validation:** Sangeetha Muthamilselvan, Ashok Palaniappan.

**Visualization:** Sangeetha Muthamilselvan, Abirami Raghavendran, Ashok Palaniappan.

**Writing – original draft:** Sangeetha Muthamilselvan, Ashok Palaniappan.

**Writing – review & editing:** Ashok Palaniappan.

## References

1. Ferlay J, Soerjomataram I, Dikshit R, Eser S, Mathers C, Rebelo M, et al. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. International journal of cancer. 2015; 136(5). https://doi.org/10.1002/ijc.29210 PMID: 25220842
2. Carethers JM, Jung BH. Genetics and Genetic Biomarkers in Sporadic Colorectal Cancer. Gastroenterology. 2015; 149(5). https://doi.org/10.1053/j.gastro.2015.06.047 PMID: 26216840
3. Ogino S, Goel A. Molecular classification and correlates in colorectal cancer. The Journal of molecular diagnostics: JMD. 2008; 10(1). https://doi.org/10.2353/jmoldx.2008.070082 PMID: 18165277
4. Chen J-J, Wang A-Q, Chen Q-Q. DNA methylation assay for colorectal carcinoma. Cancer biology & medicine. 2017; 14(1). https://doi.org/10.20892/j.issn.2095-3941.2016.0082 PMID: 28443202
5. Esteller M, Herman JG. Cancer as an epigenetic disease: DNA methylation and chromatin alterations in human tumours. The Journal of pathology. 2002; 196(1). https://doi.org/10.1002/path.1024 PMID: 11748635
6. Schnekenburger M, Florean C, Dicato M, Diederich M. Epigenetic alterations as a universal feature of cancer hallmarks and a promising target for personalized treatments. Current topics in medicinal chemistry. 2016; 16(7). https://doi.org/10.2174/1568026615666150825141330 PMID: 26303418
7. Feinberg AP, Tycko B. The history of cancer epigenetics. Nature reviews Cancer. 2004; 4(2). https://doi.org/10.1038/nrc1279 PMID: 14732866
8. Goelz SE, Vogelstein B, Hamilton SR, Feinberg AP. Hypomethylation of DNA from benign and malignant human colon neoplasms. Science. 1985; 228(4696). https://doi.org/10.1126/science.2579435 PMID: 2579435
9. Timp W, Bravo HC, McDonald OG, Goggins M, Umbricht C, Zeiger M, et al. Large hypomethylated blocks as a universal defining epigenetic alteration in human solid tumors. Genome Medicine. 2014; 6 (8):1–11. https://doi.org/10.1186/s13073-014-0061-y PMID: 25191524
10. Gonzalo S. Epigenetic alterations in aging. Journal of applied physiology 2010; 109(2).
11. Carpenter BL, Zhou W, Madaj Z, DeWitt AK, Ross JP, Grønbæk K, et al. Mother–child transmission of epigenetic information by tunable polymorphic imprinting. Proc Natl Acad Sci U S A. 2018; 115(51): E11970–E11977. https://doi.org/10.1073/pnas.1815005115 PMID: 30509985
12. Toyota M, Issa JP. The role of DNA hypermethylation in human neoplasia. Electrophoresis. 2000; 21 (2). https://doi.org/10.1002/(SICI)1522-2683(20000101)21:2<329::AID-ELPS329>3.0.CO;2-9 PMID: 10675010
13. Weisenberger DJ, Siegmund KD, Campan M, Young J, Long TI, Faasse MA, et al. CpG island methylator phenotype underlies sporadic microsatellite instability and is tightly associated with BRAF mutation in colorectal cancer. Nature genetics. 2006; 38(7). https://doi.org/10.1038/ng1834 PMID: 16804544
14. Lao VV, Grady WM. Epigenetics and colorectal cancer. Nature reviews Gastroenterology & hepatology. 2011; 8(12). https://doi.org/10.1038/nrgastro.2011.173 PMID: 22009203
15. Costello JF, Frühwald MC, Smiraglia DJ, Rush LJ, Robertson GP, Gao X, et al. Aberrant CpG-island methylation has non-random and tumour-type-specific patterns. Nature genetics. 2000; 24(2). https://doi.org/10.1038/72785 PMID: 10655057
16. Jones PA. Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nature reviews Genetics. 2012; 13(7). https://doi.org/10.1038/nrg3230 PMID: 22641018
17. Galamb O, Kalmár A, Péterfia B, Csabai I, Bodor A, Ribli D, et al. Aberrant DNA methylation of WNT pathway genes in the development and progression of CIMP-negative colorectal cancer. Epigenetics. 2016; 11(8). https://doi.org/10.1080/15592294.2016.1190894 PMID: 27245242
18. Lengauer C, Kinzler KW, Vogelstein B. DNA methylation and genetic instability in colorectal cancer cells. Proc Natl Acad Sci USA. 1997; 94(6). https://doi.org/10.1073/pnas.94.6.2545 PMID: 9122232

19. Tomczak K, Czerwińska P, Wiznerowicz M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemporary oncology (Pozn). 2015; 19(1A). https://doi.org/10.5114/wo.2014.47136 PMID: 25691825

20. Stunnenberg HG, Consortium IHE, Hirst M. The International Human Epigenome Consortium: A Blueprint for Scientific Collaboration and Discovery. Cell. 2016; 167(5).

21. Zitt M, Zitt M, Müller HM. DNA methylation in colorectal cancer—impact on screening and therapy monitoring modalities? Disease markers. 2007; 23(1–2). https://doi.org/10.1155/2007/891967 PMID: 17325426

22. Draht MXG, Goudkade D, Koch A, Grabsch HI, Weijenberg MP, Engeland Mv, et al. Prognostic DNA methylation markers for sporadic colorectal cancer: a systematic review. Clinical epigenetics. 2018; 10: 35. https://doi.org/10.1186/s13148-018-0461-8 PMID: 29564023

23. Reyes HD, Devor EJ, Warrier A, Newtson AM, Mattson J, Wagner V, et al. Differential DNA methylation in high-grade serous ovarian cancer (HGSOC) is associated with tumorbehavior. Scientific reports. 2019; 9(1): 17996. https://doi.org/10.1038/s41598-019-54401-w PMID: 31784612

24. Chen X, Gole J, Gore A, He Q, Lu M, Min J, et al. Non-invasive early detection of cancer four years before conventional diagnosis using a blood test. Nature communications. 2020; 11(1): 3475. https://doi.org/10.1038/s41467-020-17316-z PMID: 32694610

25. Bibikova M, Le J, Barnes B, Saedinia-Melnyk S, Zhou L, Shen R et al. Genome-wide DNA methylation profiling using Infinium® assay. Epigenomics. 2009; 1(1):177–200. https://doi.org/10.2217/epi.09.14 PMID: 22122642

26. Analysis-ready standardized TCGA data from broad GDAC firehose 2016_01_28 run. Broad institute of MIT and Harvard. Dataset: [Internet]. 2013 [cited January 20]. Available from: gdac.broadinstitute.org_COADREAD.Merge_methylation_humanmethylation27_jhu_usc_edu_Level_3_within_bioassay_data_set_function_data.Level_3.2016012800.0.0.tar.

27. Du P, Zhang X, Huang C-C, Jafari N, Kibbe WA, Hou L, et al. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. BMC Bioinformatics. 2010; 11(1):1–9. https://doi.org/10.1186/1471-2105-11-587 PMID: 21118553

28. Grossman RL, Heath AP, Ferretti V, Varmus HE, Lowy DR, Kibbe WA, et al. Toward a Shared Vision for Cancer Genomic Data. The New England Journal of Medicine. 2016; 375(12): 1109–12. https://doi.org/10.1056/NEJMp1607591 PMID: 27653561

29. Amin MB, Greene FL, Edge SB, Compton CC, Gershenwald JE, Brookland RK, et al. The Eighth Edition AJCC Cancer Staging Manual: Continuing to build a bridge from a population-based to a more "personalized" approach to cancer staging. CA: a cancer journal for clinicians. 2017; 67(2): 93–99. https://doi.org/10.3322/caac.21388 PMID: 28094848

30. R Core Team. R: A language and environment for statistical computing. 2020. URL https://www.R-project.org.

31. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic acids research. 2015; 43(7): e47. https://doi.org/10.1093/nar/gkv007 PMID: 25605792

32. Sarathi A, Palaniappan A. Novel significant stage-specific differentially expressed genes in hepatocellular carcinoma. BMC Cancer. 2019; 19(1):1–22. https://doi.org/10.1186/s12885-018-5219-3 PMID: 30606139

33. Barfield RT, Kilaru V, Smith AK, Conneely KN. CpGassoc: an R function for analysis of DNA methylation microarray data. Bioinformatics. 2012; 28(9): 1280–1. https://doi.org/10.1093/bioinformatics/bts124 PMID: 22451269

34. Morris TJ, Butcher LM, Feber A, Teschendorff AE, Chakravarthy AR, Wojdacz TK, et al. ChAMP: 450k Chip Analysis Methylation Pipeline. Bioinformatics. 2014; 30(3): 428–30. https://doi.org/10.1093/bioinformatics/btt684 PMID: 24336642

35. Peters TJ, Buckley MJ, Statham AL, Pidsley R, Samaras K, Lord RV, et al. De novo identification of differentially methylated regions in the human genome. Epigenetics & chromatin. 2015; 8: 6. https://doi.org/10.1186/1756-8935-8-6 PMID: 25972926

36. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA. 2005; 102(43): 15545–50. https://doi.org/10.1073/pnas.0506580102 PMID: 16199517

37. Cedoz P-L, Prunello M, Brennan K, Gevaert O. MethylMix 2.0: an R package for identifying DNA methylation genes. Bioinformatics. 2018; 34(17): 3044–46. https://doi.org/10.1093/bioinformatics/bty156 PMID: 29668835

38. Wang Y, Franks JM, Whitfield ML, Cheng C. BioMethyl: an R package for biological interpretation of DNA methylation data. Bioinformatics. 2019; 35(19): 3635–41. https://doi.org/10.1093/bioinformatics/btz137 PMID: 30799505

39. Therneau TM, Lumley T. Package 'survival'. R Top Doc. 2015; 128(10):28–33.

40. Kassambara A, Kosinski M, Biecek P, Fabian S. Package 'survminer'. Drawing Survival Curves using 'ggplot2'(R package version 03 1). 2017.

41. Teng H, Gao R, Qin N, Jiang X, Ren M, Wang Y, et al. Identification of recurrent and novel mutations by whole-genome sequencing of colorectal tumors from the Han population in Shanghai, eastern China. Molecular medicine reports. 2018; 18(6): 5361–70. https://doi.org/10.3892/mmr.2018.9563 PMID: 30365144

42. Carvalho DDD, Sharma S, You JS, Su S-F, Taberlay PC, Kelly TK, et al. DNA methylation screening identifies driver epigenetic events of cancer cell survival. Cancer cell. 2012; 21(5): 655–667. https://doi.org/10.1016/j.ccr.2012.03.045 PMID: 22624715

43. Feng Y, Jiang Y, Feng Q, Xu L, Jiang Y, Meng F, et al. A novel prognostic biomarker for muscle invasive bladder urothelial carcinoma based on 11 DNA methylation signature. Cancer biology & therapy. 2020; 21(12): 1119–27. https://doi.org/10.1080/15384047.2020.1833811 PMID: 33151129

44. Lin S-H, Raju GS, Ye CHY, Gu J, Chen J-S, Hildebrandt MAT, et al. The somatic mutation landscape of premalignant colorectal adenoma. Gut. 2018; 67(7): 1299–1305. https://doi.org/10.1136/gutjnl-2016-313573 PMID: 28607096

45. Atwell LL, Beaver LM, Shannon J, Williams DE, Dashwood RH, Ho E. Epigenetic Regulation by Sulforaphane: Opportunities for Breast and Prostate Cancer Chemoprevention. Current pharmacology reports. 2015; 1(2): 102–111. https://doi.org/10.1007/s40495-014-0002-x PMID: 26042194

46. Lex A, Gehlenborg N, Strobelt H, Vuillemot R, Pfister H. UpSet: Visualization of Intersecting Sets. IEEE Trans Vis Comput Graph. 2014; 20(12): 1983–92. https://doi.org/10.1109/TVCG.2014.2346248 PMID: 26356912

47. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature genetics. 2000; 25(1): 25–9. https://doi.org/10.1038/75556 PMID: 10802651

48. Matzke MA, Mosher RA. RNA-directed DNA methylation: an epigenetic pathway of increasing complexity. Nature reviews Genetics. 2014; 15(6): 394–408. https://doi.org/10.1038/nrg3683 PMID: 24805120

49. Kulis M, Esteller M. DNA methylation and cancer. Advances in genetics. 2010; 70: 27–56. https://doi.org/10.1016/B978-0-12-380866-0.60002-2 PMID: 20920744

50. Ashktorab H, Brim H. DNA Methylation and Colorectal Cancer. Current colorectal cancer reports. 2014; 10(4): 425–30. https://doi.org/10.1007/s11888-014-0245-2 PMID: 25580099

51. Sondka Z, Bamford S, Cole CG, Ward SA, Dunham I, Forbes SA. The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. Nature reviews Cancer. 2018; 18(11): 696–705. https://doi.org/10.1038/s41568-018-0060-1 PMID: 30293088

52. Abbott KL, Nyre ET, Abrahante J, Ho Y-Y, Vogel RI, Starr TK. The Candidate Cancer Gene Database: a database of cancer driver genes from forward genetic screens in mice. Nucleic acids research. 2015; 43(Database issue): D844–8. https://doi.org/10.1093/nar/gku770 PMID: 25190456

53. Sakai LY, Keene DR, Renard M, Backer JD. FBN1: The disease-causing gene for Marfan syndrome and other genetic disorders. Gene. 2016; 591(1): 279–291. https://doi.org/10.1016/j.gene.2016.07.033 PMID: 27437668

54. Guo Q, Song Y, Zhang H, Wu X, Xia P, Dang C. Detection of hypermethylated fibrillin-1 in the stool samples of colorectal cancer patients. Medical oncology. 2013; 30(4): 695. https://doi.org/10.1007/s12032-013-0695-4 PMID: 23963856

55. Li W-h, Zhang H, Guo Q, Wu X-d, Xu Z-s, Dang C-x, et al. Detection of SNCA and FBN1 methylation in the stool as a biomarker for colorectal cancer. Disease markers. 2015; 2015: 657570. https://doi.org/10.1155/2015/657570 PMID: 25802477

56. Wu H, Qian C, Liu C, Xiang J, Ye D, Zhang Z, et al. Role and mechanism of FOXG1 in invasion and metastasis of colorectal cancer. Chinese journal of biotechnology. 2018; 34(5): 752–60. https://doi.org/10.13345/j.cjb.170389 PMID: 29893083

57. Zheng X, Lin J, Wu H, Mo Z, Lian Y, Wang P, et al. Forkhead box (FOX) G1 promotes hepatocellular carcinoma epithelial-Mesenchymal transition by activating Wnt signal through forming T-cell factor-4/Beta-catenin/FOXG1 complex. Journal of experimental & clinical cancer research. 2019; 38(1): 475. https://doi.org/10.1186/s13046-019-1433-3 PMID: 31771611

58. Chan DW, Liu VWS, To RMY, Chiu PM, Lee WYW, Yao KM, et al. Overexpression of FOXG1 contributes to TGF-beta resistance through inhibition of p21WAF1/CIP1 expression in ovarian cancer. British journal of cancer. 2009; 101(8): 1433–43. https://doi.org/10.1038/sj.bjc.6605316 PMID: 19755996

59. Phan NN, Huynh TT, Lin Y-C. Hyperpolarization-activated cyclic nucleotide-gated gene signatures and poor clinical outcome of cancer patient. Translational Cancer Research:. 2017; 6(4). Available from: https://tcr.amegroups.com/article/view/15057/html. https://doi.org/10.21037/tcr.2017.07.22

60. Ashktorab H, Rahi H, Wansley D, Varma S, Shokrani B, Lee E, et al. Toward a comprehensive and systematic methylome signature in colorectal cancers. Epigenetics. 2013; 8(8): 807–15. https://doi.org/10.4161/epi.25497 PMID: 23975090

61. Mori Y, Cai K, Cheng Y, Wang S, Paun B, Hamilton JaP, et al. A genome-wide search identifies epigenetic silencing of somatostatin, tachykinin-1, and 5 other genes in colon cancer. Gastroenterology. 2006; 131(3): 797–808. https://doi.org/10.1053/j.gastro.2006.06.006 PMID: 16952549

62. Ma Z, Williams M, Cheng YY, Leung WK. Roles of Methylated DNA Biomarkers in Patients with Colorectal Cancer. Disease markers. 2019; 2019: 2673543. https://doi.org/10.1155/2019/2673543 PMID: 30944663

63. Fang S-Q, Gao M, Xiong S-L, Chen H-Y, Hu S-S, Cai H-B. Combining differential expression and differential coexpression analysis identifies optimal gene and gene set in cervical cancer. Journal of cancer research and therapeutics. 2018; 14(1): 201–7. https://doi.org/10.4103/0973-1482.199787 PMID: 29516986

64. Xi T, Zhang G. Epigenetic regulation on the gene expression signature in esophagus adenocarcinoma. Pathology, research and practice. 2017; 213(2): 83–88. https://doi.org/10.1016/j.prp.2016.12.007 PMID: 28049580

65. Zhang H, Sun X, Lu Y, Wu J, Feng J. DNA-methylated gene markers for colorectal cancer in TCGA database. Experimental and therapeutic medicine. 2020; 19(4): 3042–50. https://doi.org/10.3892/etm.2020.8565 PMID: 32256791

66. Siesser PF, Motolese M, Walker MP, Goldfarb D, Gewain K, Yan F, et al. FAM123A binds to microtubules and inhibits the guanine nucleotide exchange factor ARHGEF2 to decrease actomyosin contractility. Science signaling. 2012; 5(240): ra64. https://doi.org/10.1126/scisignal.2002871 PMID: 22949735

67. Lieu CH, Golemis EA, Serebriiskii IG, Newberg J, Hemmerich A, Connelly C, et al. Comprehensive Genomic Landscapes in Early and Later Onset Colorectal Cancer. Clinical cancer research. 2019; 25 (19): 5852–5858. https://doi.org/10.1158/1078-0432.CCR-19-0899 PMID: 31243121

68. Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. Nature. 2012; 487(7407): 330–7. https://doi.org/10.1038/nature11252 PMID: 22810696

69. Sanz-Pamplona R, Lopez-Doriga A, Paré-Brunet L, Lázaro K, Bellido F, Alonso MH, et al. Exome Sequencing Reveals AMER1 as a Frequently Mutated Gene in Colorectal Cancer. Clinical cancer research. 2015; 21(20): 4709–18. https://doi.org/10.1158/1078-0432.CCR-15-0159 PMID: 26071483

70. Moore LD, Le T, Fan G. DNA methylation and its basic function. Neuropsychopharmacology: official publication of the American College of Neuropsychopharmacology. 2013; 38(1): 23–38. https://doi.org/10.1038/npp.2012.112 PMID: 22781841

71. Eden S, Cedar H. Role of DNA methylation in the regulation of transcription. Current opinion in genetics & development. 1994; 4(2): 255–9. https://doi.org/10.1016/s0959-437x(05)80052-8 PMID: 8032203

72. Toyota M, Ahuja N, Ohe-Toyota M, Herman JG, Baylin SB, Issa JP. CpG island methylator phenotype in colorectal cancer. Proc Natl Acad Sci USA. 1999; 96(15): 8681–6. https://doi.org/10.1073/pnas.96.15.8681 PMID: 10411935

73. Jia M, Gao X, Zhang Y, Hoffmeister M, Brenner H. Different definitions of CpG island methylator phenotype and outcomes of colorectal cancer: a systematic review. Clinical epigenetics. 2016; 8: 25. https://doi.org/10.1186/s13148-016-0191-8 PMID: 26941852

74. Hughes LAE, Melotte V, Schrijver Jd, Maat Md, Smit VTHBM, Bovée JVMG, et al. The CpG island methylator phenotype: what's in a name? Cancer research. 2013; 73(19): 5858–68. https://doi.org/10.1158/0008-5472.CAN-12-4306 PMID: 23801749

75. Berg M, Hagland HR, Søreide K. Comparison of CpG island methylator phenotype (CIMP) frequency in colon cancer using different probe- and gene-specific scoring alternatives on recommended multi-gene panels. PloS one. 2014; 9(1): e86657. https://doi.org/10.1371/journal.pone.0086657 PMID: 24466191

76. Lind GE, Danielsen SA, Ahlquist T. et al. Identification of an epigenetic biomarker panel with high sensitivity and specificity for colorectal cancer and adenomas. Mol Cancer. 2011; 10: 85. https://doi.org/10.1186/1476-4598-10-85 PMID: 21777459

77. Cho NY, Park JW, Wen X, Shin YJ, Kang JK, Song SH et al. Blood-Based Detection of Colorectal Cancer Using Cancer-Specific DNA Methylation Markers. Diagnostics (Basel). 2020; 11(1):51. https://doi.org/10.3390/diagnostics11010051 PMID: 33396258

78. Advani SM, Advani P, Brown SMDe, VonVille HM, Lam M, Loree JM, et al. Clinical, Pathological, and Molecular Characteristics of CpG Island Methylator Phenotype in Colorectal Cancer: A Systematic

Review and Meta-analysis. Translational oncology. 2018; 11(5): 1188–1201. https://doi.org/10.1016/j.tranon.2018.07.008 PMID: 30071442

79. Issa J-P. CpG island methylator phenotype in cancer. Nature reviews Cancer. 2004; 4(12): 988–93. https://doi.org/10.1038/nrc1507 PMID: 15573120

80. Kerachian MA, Javadmanesh A, Azghandi M, Shariatpanahi AM, Yassi M, Davodly ES, et al. Crosstalk between DNA methylation and gene expression in colorectal cancer, a potential plasma biomarker for tracing this tumor. Scientific reports. 2020; 10(1): 2813. https://doi.org/10.1038/s41598-020-59690-0 PMID: 32071364

81. Gündert M, Edelmann D, Benner A, Jansen L, Jia M, Walter V, et al. Genome-wide DNA methylation analysis reveals a prognostic classifier for non-metastatic colorectal cancer (ProMCol classifier). Gut. 2019; 68(1): 101–110. https://doi.org/10.1136/gutjnl-2017-314711 PMID: 29101262

82. Freitas M, Ferreira F, Carvalho S, Silva F, Lopes P, Antunes L, et al. A novel DNA methylation panel accurately detects colorectal cancer independently of molecular pathway. Journal of translational medicine. 2018; 16(1): 45. https://doi.org/10.1186/s12967-018-1415-9 PMID: 29486770

83. Vrba L, Futscher BW. A suite of DNA methylation markers that can detect most common human cancers. Epigenetics. 2018; 13(1): 61–72. https://doi.org/10.1080/15592294.2017.1412907 PMID: 29212414

84. Ma X, Wang Y-W, Zhang MQ, Gazdar AF. DNA methylation data analysis and its application to cancer research. Epigenomics. 2013; 5(3): 301–16. https://doi.org/10.2217/epi.13.26 PMID: 23750645

85. Jones PA. The DNA methylation paradox. Trends in genetics. 1999; 15(1): 34–7. https://doi.org/10.1016/s0168-9525(98)01636-9 PMID: 10087932

86. Ehrlich M. DNA hypermethylation in disease: mechanisms and clinical relevance. Epigenetics. 2019; 14 (12): 1141–63. https://doi.org/10.1080/15592294.2019.1638701 PMID: 31284823

87. Feng S, Cokus SJ, Zhang X, Chen P-Y, Bostick M, Goll MG, et al. Conservation and divergence of methylation patterning in plants and animals. Proc Natl Acad Sci USA. 2010; 107(19): 8689–94. https://doi.org/10.1073/pnas.1002720107 PMID: 20395551