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

Sangeetha Muthamilselvan  
SASTRA Deemed University

Abirami Raghavendran  
SASTRA Deemed University

Ashok Palaniappan  
apalania@scbt.sastra.edu  
SASTRA Deemed University

Research Article

Keywords: Colorectal cancer, tumor progression, epigenetics, differential methylation; cancer stages; gene expression; linear modelling; consensus biomarkers, survival analysis, CpG island methylator phenotype

DOI: https://doi.org/10.21203/rs.3.rs-244816/v1

License: © This work is licensed under a Creative Commons Attribution 4.0 International License.  Read Full License
Abstract

Background: Aberrant DNA methylation acts epigenetically to skew the gene transcription rate up or down, with causative roles in the etiology of cancers. However research on the role of DNA methylation in driving the progression of cancers is limited. In this study, we have developed a comprehensive computational framework for the stage-differentiated modelling of DNA methylation landscapes in colorectal cancer (CRC), and unravelled significant stagewise signatures of CRC progression.

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 multiple approaches involving stage-differentiated linear 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). These results were 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 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. Our stage-differentiated analysis yielded the following stage-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 hypermethylated, indicating down-regulation and signifying a CpG island Methylator Phenotype (CIMP) manifestation. A significant prognostic signature consisting of FBN1 and FOXG1 survived all the steps of our analysis pipeline, and represents a novel early-stage biomarker.

Conclusions: We have designed a workflow for stage-differentiated consensus analysis, and identified stage-salient diagnostic biomarkers and an early-stage prognostic biomarker panel. Our studies further yield a novel CIMP-like signature of potential clinical import underlying CRC progression.

Introduction

Colorectal adeno-carcinoma (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 [2,3,4]. 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]. Ablenet 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, IGFL2, 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 tumor behavior corresponds with 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., sounded the advent of methylation-based tests by demonstrating an effective screening method to identify multiple types of cancer based on a 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 investigating stage-salient DMGs using several model-driven approaches, and evaluating their potential as diagnostic and prognostic biomarkers.

Methods

Data Preprocessing:

Processed Level-3 27k CRC methylation data was retrieved from TCGA [25]. By matching on the patient barcode, the updated clinical data (clinical.cases_selected.tar.gz) was obtained from the GDC [26]. 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 (0,20), we used the following function on the β-matrix values, to obtain the M-value matrix [27]: see formula 1 in the supplementary files.
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 ('averep'), 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 σ < 1) across all cases in the M-value matrices. The stages were annotated for both the β-matrix and M-value matrices using the clinical data encoded in the 'Pathologic_stage' attribute. 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 [28] (Table 1).

| TCGA Stage | TNM Classification | Cases |
|------------|--------------------|-------|
| I          | T1N0M0             | 50    |
| II         | -                  | 17    |
| IIa        | T3N0M0             | 64    |
| IIb        | T4aN0M0            | 5     |
| III        | -                  | 16    |
| IIIa       | T1-T2N1/NcM0       | 3     |
| T1N2aM0    |                    |       |
| IIb        | T3-T4aN1/NcM0      | 21    |
| T2-T3N2aM0 |                    |       |
| T1-T2N2bM0 |                    |       |
| IIIc       | T4aN2aM0           | 20    |
| T3-T4bN2bM0|                    |       |
| T4bN1-N2M0 |                    |       |
| IV         | -                  | 35    |
| IVa        | Any-T Any-N M1a    | 1     |
| CONTROL    | -                  | 42    |
| NA         |                    | 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 [29].

Models

(1) Linear model analysis: Linear modelling is essential to identify linear trends in expression across cancer stages and thereby detect stage-sensitive patterns. We used the R package limma [30] for linear modelling of stagewise expression using the complete M-value matrix, with multiple probes per gene (File S1).

(2) Linear modelling with the averep matrix: This is essentially similar to the above model, except that the input is the averep matrix, where each gene is represented by the average M-value across all its probes (File S2). These 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 [31]. 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 phenotype: 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 [32] 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 [33]. 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) [34]. GSEA was used to identify the enrichment of DMPs and DMRs in the MSigDB pathways [35], using the Fisher Exact test calculation with adjusted p-value < 0.05.
(5) Modelling expression from methylation: We used the R package BioMethyl to model the aggregate expression level of a gene from its methylation patterns [36]. 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.

(6) 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_maseqv2_illumina_maseqv2_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.

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 [38]. 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) [39]. 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.

Results

Linear modelling at the probe-level:

The number of significant genes present in each stage-control pair from the Tier-I contrasts is shown in Figure 1(i). Using the top 100 DM genes of the linear model (given in Supplementary Information S3), we found a clear separation between controls and stage samples (Figure 1(ii)). The top genes in each stage (by adjusted p-value of contrast with control) are shown in Table 2, with |lfc M-value| and inferred regulation status. Figure 2 shows boxplots of stagewise methylation levels for two representative genes: (1) TMEM179, mutations in which could cause MSI [40]; and (2) MEOX2 whose promoter methylation status is a known CRC marker [41]. The top four genes of each stage were used to construct a stagewise methylation heatmap (Figure 3). The stagewise methylation patterns of the top five linear model genes are also shown, in Figure 4. It is notable that a naturally occurring 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 [42]. 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, were used in the consensus analysis.

Table 2. Top ten genes of the linear model at the probe level. 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.
Linear modelling at the gene-level (averep):

The genes with more than one probe were averaged to a single methylation value, which was then further analyzed. The number of genes present in each stage-control pair from the Tier-I contrasts is shown in Figure 5(i). Using the top 100 genes of the linear model (given in Supplementary Information S4), we found a clear separation between controls and stage samples (Figure 5(ii)). 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. Figure 6 shows the boxplots of stagewise methylation levels for two representative genes, NALCN and GLRX. Mutations in NALCN have been reported in sporadic CRC [43]; 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 [44]. The top four genes of each stage were used to construct a stagewise methylation heatmap (Figure 7). The stagewise methylation patterns of the top five linear model genes are also shown, in Figure 8. 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, were used for the consensus analysis.

Table 3. Top ten genes of the linear model at the gene level, using average values of methylation. 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.

| ID               | Stage I lfc (β₁) | Stage II lfc (β₂) | Stage III lfc (β₃) | Stage IV lfc (β₄) | 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             |

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 (Figure 9). The top 100 genes of each stage from this analysis (given in Supplementary Information S5) were used for the consensus investigation.

DMP analysis with ChAMP

The summary features of the β matrix dataset were evaluated using ChAMP (Figure 10). 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 stagell – 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 Supplementary Information S6, 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 Supplementary Information S7. The DMPs and DMRs from the analysis were subjected to GSEA and these results could also be found in Supplementary Information S6. Figure 11 shows representative DMP and DMR plots using MethylMix.

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    |

Methylation and Gene Expression Correlation analysis

Mixture models of genes, indicative of the number of methylation states, were constructed using MethylMix, and the top three genes from an overall cancer vs control comparison are shown in Figure 12. The estimated correlation between the methylation levels and actual gene expression for the same genes is depicted in Figure 13. Genes were differentially methylated and designated as ‘driver’ genes if the p-value of the contrast being studied was significant. The calculated differential methylation (DM) values from stage vs control contrasts ranged from -0.7 to +0.8, and genes were classified as hyper- or hypomethylated based on 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 Supplementary Information S8. Top 100 genes from each comparison were taken forward for the consensus analysis. Certain genes emerged common to all the four comparisons, indicating stage-agnostic differential methylation events. The top such 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 list of 56 stage-agnostic genes arising out of the MethylMix analysis is provided in Supplementary File S9.

BioMethyl analysis

The significant stage-specific DEGs identified by this BioMethyl are shown in Figure 14. Top 100 genes of each stage from this analysis were taken for consensus analysis. The stage-specific genes from this analysis are presented in the Supplementary Information S10.

Stage-salient consensus biomarkers

The top 100 significantly differentially-expressed genes of each stage from all the methods discussed above (collated in Supplementary Information S11) 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 |fM c-M-value| > 0.4 with respect to the other stages, validating their salience. Figures 15,16 represent boxplots of the consensus biomarkers, and Table 5 presents a summary of the consensus analysis. Gene ontology (GO) analysis [45] 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 Supplementary Information S12.

Table 5. Stage-salient biomarkers. The results of the consensus analysis and univariate survival analysis are summarized. All the biomarkers showed hypermethylation, indicating that an inferred common down-regulation of gene expression. Univariate survival analysis does not show significance for NELL1 by itself.
Table 6. GO analysis of stage-salient genes in the order of decreasing significance (i.e., increasing p-value). Ontology: Cellular Compartment (CC), Molecular Function (MF), Biological Process (BP).

| 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       | 0.0006  |
| 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  |

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. 17. 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 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 ≤ 0.02) (Figure 18). Supplementary Information S13 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.

Table 7. Summary of selected multivariate prognostic models. Weight denotes the coefficient in the multivariate model. The ultimate significant signature is highlighted.
**Discussion**

CRC development is due to the accumulation of genetic and epigenetic changes of which DNA methylation is of prime importance. DNA methylation profiles of colorectal cancer have been investigated in several previous studies using various approaches [46, 47]. It is well-known that changes in methylation status correspond with CRC progression [48]. 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 has yielded consensus stage-salient significantly differentially methylated genes, stage-agnostic genes, and their prognostic value. It is significant that none of the stage-salient genes figure as a cancer gene or a hallmark gene in the Cancer Gene Census [49]; HCN1 is notably marked as a candidate cancer gene based on mouse insertional mutagenesis experiments [50]. A discussion of the stage-salient DMGs (Table 5) with respect to the existing literature is provided below.

**Early-stage salient DMGs:**

Promoter hypermethylation of FBN1, a glycoprotein component of calcium-binding extracellular matrix microbrils [51], is a recognized biomarker of CRC [52, 53]. 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 [54]. It is a nodal gene, with connections to oncogenic pathways like WNT pathway in hepatocellular carcinoma [55] and TGF-β pathway in ovarian cancer [56]. 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 [57]. We have identified HCN1 as a stage-III hypermethylated gene, suggesting a loss-of-function mechanism for its tumorigenic potential.

**Stage-IV salient DMGs:**

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 for the hypermethylation of the promotor region of LAMA1 in CRC cases is available [58]. NELL1 is a known tumor suppressor gene [59], whose hypermethylation is associated with poor survival outcomes [60]. 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 [61, 62]. 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, without regard for pathologic stage [63]. FAM123A, also known as AMER2, is associated with microtubule proteins [64], 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 [65, 66, 67]. 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 [68,69]. 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 [70]. CIMP is characterised by hypermethylation of CpG islands surrounding the promoter regions of genes involved in cancer onset and progression [71]. The phenotype is heterogenous with the type of tumor [72] and dependent on definition [73]. In this background, it is less straightforward to interpret the functional importance of hypermethylation of individual genes.
Still it is clear from Table 5 that the stage-salient hypermethylated biomarkers identified in our study could constitute an aggregate novel CIMP. The original CIMP had been associated with advanced T staging (T3/T4) [74], which accords with our finding of five hypermethylated stage IV-salient DMGs. Epigenetic intervention for CIMP-positive cancers has been suggested as a possible treatment strategy [75]. The biomarkers contributing to the putative CIMP were tested with Cox regression and then evaluated independently as well as in combination for prognostic significance. Five of the seven stage-salient genes were prognostically significant in both the Cox univariate model and the Kaplan-Meier analysis (Table 5). A 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.

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 [76,77]. Freitas et al., have validated the performance of a 3-gene biomarker panel for the detection of colorectal cancer irrespective of the molecular subtype [78]. 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 [79]. Hypermethylation in gene body and upstream control regions like enhancers and insulators might affect transcription differently than hypermethylation of promoter regions [80,81]. Further DNA methylation patterns in noncoding RNA genes seem to be important in tumorigenesis and progression [82]. Non-coding RNAs themselves play a significant role in epigenetic modification through the phenomenon of RNA-directed DNA methylation [46]. The nuanced relationship between methylation and gene transcription does urge the application of our results with clinical validation, however consensus study designs such as the one used here suffer less uncertainties with respect to the identified biomarkers. Since methylation mediates a direct epigenetic regulatory mechanism used by all life [83], it is hoped that our studies 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 hitherto unreported in the literature: one stage-I gene (FBN1), one stage-II gene (FOXG1), one stage-III gene (HCN1) and four stage-IV genes (NEL1, ZNF135, FAM123A, LAMA1). Stage-salient genes could serve as diagnostic biomarkers. The top stage-agnostic genes could serve as targets for drug discovery in CRC therapy. All the stage-salient genes were found to be hypermethylated, indicating a novel CIMP-like character possibly promoting epigenetic destabilisation, which in turn would drive the progression of colorectal cancer. Independent prognostic evaluation of the stage-salient genes 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. 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 present approach is extensible to the investigation of epigenomics in other cancers, normal/disease conditions, and developmental biology.

**Declarations**

**Ethics approval and consent to participate**

Not applicable. The study uses existing anonymised / de-identified clinical cancer data in the public domain via TCGA / cBio, NCI (USA) portals.

**Consent for publication**

Not applicable.

**Availability of data and materials**

All data generated or analysed during this study are included in this published article [and its supplementary information files]. The data are also available in the Figshare repository, [https://doi.org/10.6084/m9.figshare.13013852](https://doi.org/10.6084/m9.figshare.13013852)

**Competing interests**

The authors declare that they have no competing interests, financial or otherwise.

**Funding**

This work has been funded by DST-SERB grant EMR/2017/000470/BBM (to A.P.).

**Authors' contributions**

A.P. conceived, designed and supervised the work. A.R. and S.M. performed the research; S.M., A.R., and A.P. analyzed the results. A.P. wrote the manuscript.

**Acknowledgements**

We are grateful to the School of Chemical and BioTechnology, SASTRA Deemed University for computing and infrastructure support. We would like to thank Alex Stanley for assistance. This work has been supported in part by DST-SERB grant EMR/2017/000470/BBM (to A.P.). A preliminary version of this
References

1. Ferlay, J. et al. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN. Int J Cancer. 136, 359–386; 10.1002/jic.29210 (2015).
2. Carethers, J. M. & Jung, B.H. Genetics and genetic biomarkers in sporadic colorectal cancer. Gastroenterology. 149(5), 1177–1190; 10.1053/j.gastro.2015.06.047 (2015).
3. Ogino, S. & Goel, A. Molecular classification and correlates in colorectal cancer. J Mol Diagn. 10(1), 13–27; 10.2353/jmdx.2008.070082 (2008).
4. Chen, J. J., Wang, A. Q. & Chen, Q. Q. DNA methylation assay for colorectal carcinoma. Cancer Biol Med. 14(1), 42-49; 10.20892/jissn.2095-3941.2016.0082 (2017).
5. Esteller, M. & Herman, J. G. Cancer as an epigenetic disease: DNA methylation and chromatin alterations in human tumours. J Pathol. 196(1), 1–7; 10.1002/path.1024 (2002).
6. Schnekenburger, M., Florene, C., Dicato, M. & Diederich, M. Epigenetic alterations as a universal feature of cancer hallmarks and a promising target for personalized treatments. Curr Top Med Chem. 16(7), 745–776; 10.2174/1568026615666150825141330 (2016).
7. Feinberg, A. P. & Tycko, B. The history of cancer epigenetics. Nature Reviews Cancer. 4(2), 143–153; 10.1038/nrc1279 (2004).
8. Goel, S. E., Vogelstein, B., Hamilton, S. R. & Feinberg, A. P. Hypomethylation of DNA from benign and malignant human colon neoplasms. Science. 228(4696), 187–90; 10.1126/science.2579435 (1985).
9. Timp, W. et al. Large hypomethylated blocks as a universal defining epigenetic alteration in human solid tumors. Genome Med. 6(8), 61-71; 10.1186/s13073-014-0061-y (2014).
10. Gonzalo, S. Epigenetic alterations in aging. J Appl Physiol. 109(2), 586–597; 10.1152/japplphysiol.00238 (2010).
11. Brittany, L. C. et al. Mother–child transmission of epigenetic information by tunable polymorphic imprinting. Proc Natl Acad Sci USA. 115(51), E11970-E11977; 10.1073/pnas.1815005115 (2018).
12. Toyota, M. & Issa, J. The role of DNA hypermethylation in human neoplasia. Electrophoresis. 21(2), 329-333; 10.1002/(SICI)1522-2683(20000110)21:2<329::AID-ELPS329>3.0.CO;2-9 (2000).
13. Weisenberger, D. J. et al. CpG island methylator phenotype underlies sporadic microsatellite instability and is tightly associated with BRAF mutation in colorectal cancer. Nat Genet. 38(7), 787–793; 10.1038/ng1683 (2006).
14. Lao, V. V. & Grady, W. M. Epigenetics and colorectal cancer. Nat Rev Gastroenterol Hepatol. 8, 686–700; 10.1038/nrgastro.2011.173 (2011).
15. Costello, J. F. et al. Aberrant CpG-island methylation has non-random and tumour-type-specific patterns. Nat Genet. 24(2), 132–138; 10.1038/72785 (2000).
16. Jones, P. A. Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nat Rev Genet. 13(7), 484–492; 10.1038/nrg3230 (2012).
17. Galamb, O. et al. Aberrant DNA methylation of WNT pathway genes in the development and progression of CIMP-negative colorectal cancer. Epigenetics. 11(6), 588–602; 10.1080/15592294.2016.1190894 (2016).
18. Lengauer, C., Kinzler, K. W. & Vogelstein, B. DNA methylation and genetic instability in colorectal cancer cells. Proc Natl Acad Sci USA. 94(6), 2545-2550; 10.1073/pnas.94.6.2545 (1997).
19. Tomczak, K., Czerwińska, P. & Wiznerowicz, M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemp Oncol (Pozn). 19(1A), A68-77; 10.5114/wo.2014.47136 (2015).
20. Stunnenberg, H. G. et al. The International Human Epigenome Consortium: A Blueprint for Scientific Collaboration and Discovery. Cell. 167(5), 1145-1149; 10.1016/j.cell.2016.11.007 (2016).
21. Zitt, M., Zitt, M. & Müller, H. M. DNA methylation in colorectal cancer—impact on screening and therapy monitoring modalities?. Dis Markers. 23(1-2), 51-71; 10.1155/2007/891967 (2007).
22. Draht, M. X. G. et al. Prognostic DNA methylation markers for sporadic colorectal cancer: a systematic review. Clin Epigenet. 10, 35-50; 10.1186/s13148-018-0461-8 (2008).
23. Reyes, H. D. et al. Differential DNA methylation in high-grade serous ovarian cancer (HGSOC) is associated with tumor behavior. Sci Rep. 9, 17996; 10.1038/s41598-019-54401-w (2019).
24. Chen, X. et al. Non-invasive early detection of cancer four years before conventional diagnosis using a blood test. Nat Commun. 11, 3475-3484; 10.1038/s41467-020-17316-z (2020).
25. Broad Institute TCGA Genome Data Analysis Center. Analysis-ready standardized TCGA data from broad GDAC firehose 2016_01_28 run. Broad institute of MIT and Harvard. Dataset: gdac.broadinstitute.org_COADREAD.Merge_methylation_humanmethylation27_jhu_usc_edu_Level_3_within_bioassay_data_set_function_data.Level_3.201 2016 (Last accessed: January 10, 2020).
26. Grossman, R. L. et al. Toward a Shared Vision for Cancer Genomic Data. New England Journal of Medicine. 375(12), 1109-1112; 10.1056/NEJMp1607591 (2016).
27. Du, P. et al. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. BMC Bioinformatics. 11(1), 587-595; 10.1186/1471-2105-11-587 (2010).
28. Amin, M.B. 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 Cancer J Clin.* **67**(2), 93–9, (2017).

29. R Core Team. R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria. URL http://www.R-project.org/ (2013).

30. Ritchie, M. E. et al. Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research.* **43**(7), e47-e59; DOI: 10.1093/nar/gkv007 (2015).

31. Sarathi, A. & Palaniappan, A. Novel significant stage-specific differentially expressed genes in hepatocellular carcinoma. *BMC Cancer.* **19**(1), 663-684; 10.1186/s12885-019-5838-3 (2019).

32. Barfield, T. R. CpGassoc: an R function for analysis of DNA methylation microarray data. *BMC Bioinformatics.* **28**(9), 1280–1281; 10.1093/bioinformatics/bts124 (2012).

33. Morris, T. J. & Butcher, L. M. *Champ: 450k chip analysis methylation pipeline. BMC Bioinformatics.* **30**(3), 428-430; 10.1093/bioinformatics/btt684 (2014).

34. Peters, T. J. De novo identification of differentially methylated regions in the human genome. *Epigenetics & Chromatin,* **8**, 1-16; 10.1186/1756-8935-8-6 (2015).

35. Subramanian, A. et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA.* **102**(43), 15545-15550; 10.1073/pnas.0506580102 (2005).

36. Wang, Y., Franks, M. J., Whitfield, L. M. & Cheng, C. BioMethyl: An R package for Biological Interpretation of DNA Methylation Data. *BMC Bioinformatics.* **35**(19), 3635-3641; 10.1093/bioinformatics/btz137 (2019).

37. Cedoz, L. P., Prunello, M., Brennan, K. & Gevaert, O. MethylMix 2.0: an R package for identifying DNA methylation genes. *BMC Bioinformatics.* **34**(17), 3044-3046; 10.1093/bioinformatics/bty156 (2018).

38. Themeau, T. *A Package for Survival Analysis in S.* version 2.38, https://CRAN.R-project.org/package=survival (2015).

39. Kassambara, A., Kosinski, M. & Biecek, M. *survminer: Drawing Survival Curves using 'ggplot2'. R package version 0.4.6.* https://CRAN.R-project.org/package=survminer (2019).

40. Teng, H. et al. Identification of recurrent and novel mutations by whole-genome sequencing of colorectal tumors from the Han population in Shanghai, eastern China. *Mol Med Rep.* **18**(6), 5361-5370; 10.3892/mmr.2018.9563 (2018).

41. De Carvalho, D. D. et al. DNA methylation screening identifies driver epigenetic events of cancer cell survival. *Cancer Cell.* **21**, 655–667; 10.1016/j.ccr.2012.03.045 (2012).

42. Feng, Y. et al. A novel prognostic biomarker for muscle invasive bladder urothelial carcinoma based on 11 DNA methylation signature. *Cancer Biology & Therapy.* **21**(12), 1119-1127; 10.1080/15384047.2020.1833811 (2020).

43. Lin, S. H. et al. The somatic mutation landscape of premalignant colorectal adenoma. *Gut.* **67**(7), 1299-1305; 10.1136/gutjnl-2016-313573 (2018).

44. Atwell, L. L. et al. Epigenetic Regulation by Sulforaphane: Opportunities for Breast and Prostate Cancer Chemoprevention. *Curr Pharmacol Rep.* **1**(2), 102-111; 10.1007/s40495-014-0002-x (2015).

45. Ashburner, M. et al. Gene ontology: tool for the unification of biology. *Nat Genet.* **25**(1), 25-29; 10.1038/75556 (2000).

46. Matzke, M. & Mosher, R. RNA-directed DNA methylation: an epigenetic pathway of increasing complexity. *Nat Rev Genet.* **15**, 394–408; 10.1038/nrg3683 (2014).

47. Kulis, M. & Esteller, M. DNA methylation and cancer. *Adv Genet.* **70**, 27-56; 10.1016/B978-0-12-380866-0.60002-2 (2010).

48. Ashktorob, H. & Brim, H. DNA Methylation and Colorectal Cancer. *Curr Colorectal Cancer Rep.* **10**(4), 425-430; 10.1007/s11198-014-0245-2 (2014).

49. Sondka, Z. et al. The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nat Rev Cancer.* **18**(11), 696-705; 10.1038/s41568-018-0060-1 (2018).

50. Abbott, K. L. et al. The Candidate Cancer Gene Database: a database of cancer driver genes from forward genetic screens in mice. *Nucleic Acids Res.* **43**(Database issue), D844-8; 10.1093/nar/gku770 (2015).

51. Sakai, L. Y., Keene, D. R., Renard, M. & Juli, D. B. FBN1: The disease causing gene for marfan syndrome and other disorders. *Gene.* **591**(1), 279-291; 10.1016/j.gene.2016.07.033 (2016).

52. Guo, Q. et al. Detection of hypermethylated Fibrillin-1 in the stool sample of colorectal cancer patients. *Med. Oncol.* **30**, 695-699; 10.1007/s12032-013-0695-4 (2013).

53. Li, W. H. et al. Detection of SNCA and FBN1 methylation in stool as biomarker for colorectal cancer. *Dis Markers.* **2015**, 655750; 10.1155/2015/655750 (2015).

54. Wu, H. et al. Role and mechanism of FOXG1 in invasion and metastasis of colorectal cancer. *Chinese Journal of Biotechnology.* **34**(5), 752-760; 10.13345/j.issn.1672-728x.2018.06.014 (2018).

55. Zheng, X. 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. *J Exp Clin Cancer Res.* **38**(1), 475; 10.1186/s13046-019-1433-3 (2019).

56. Chan, D. et al. Overexpression of FOXG1 contributes to TGF-β resistance through inhibition of p21WAF1/CIP1 expression in ovarian cancer. *Br J Cancer.* **101**(8), 1433–1443; 10.1038/sj.bjc.6605316 (2009).

57. Phan, N. N., Huynh, T. T. & Lin, C. Y. Hyperpolarization-activated cyclic nucleotide-gated gene signatures and poor clinical outcome of cancer patient. *Translational clinical research.* **6**(4), 698-708; 10.21037/tcr.2017.07.22 (2017).
58. Ashktorab, H. et al. Toward a comprehensive and systematic methylome signature in colorectal cancers. Epigenetics. 8(8), 807-815; 10.4161/epi.25497 (2013).
59. Mori, Y. et al. A Genome-Wide Search Identifies Epigenetic Silencing of Somatostatin, Tachykinin-1, and 5 Other Genes in Colon Cancer. Gastroenterology. 131(3), 797-808; 10.1053/j.gastro.2006.06.006 (2006).
60. Ma, Z., Williams, M., Cheng, Y. Y. & Leung, K. W. Roles of Methylated DNA Biomarkers in Patients with Colorectal Cancer. Disease Markers. 7, 1-8; 10.1155/2019/2673543 (2019).
61. Fang, S. Q. et al. Combining differential expression and differential coexpression analysis identifies optimal gene and gene set in cervical cancer. J Cancer Res Ther. 14(1), 201-207; 10.4103/0973-1482.199787 (2018).
62. Xi, T. & Zhang, G. Epigenetic regulation on the gene expression signature in esophagus adenocarcinoma. Pathol Res Pract. 213(2), 83-88; 10.1016/j.prp.2016.12.007 (2013).
63. Zhang, H., Sun, X., Lu, Y., Wu, J., & Feng, J. (2020). DNA-methylated gene markers for colorectal cancer in TCGA database. Experimental and therapeutic medicine, 19(4), 3042–3050. https://doi.org/10.3892/etm.2020.8556
64. Siesser, P. F. et al. FAM123A binds to microtubules and inhibits the guanine nucleotide exchange factor ARHGEF2 to decrease actomyosin contractility. Sci Signal. 5(240), ra64; 10.1126/scisignal.2002871 (2012).
65. Lieu, C. H. et al. Comprehensive Genomic Landscapes in Early and Later Onset Colorectal Cancer. Clin Cancer Res. 25(19), 5852-5858; 10.1158/1078-0432.CCR-19-0899 (2019).
66. Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. Nature. 487(7407), 330-337; 10.1038/nature11252 (2012).
67. Sanz-Pamplona, R. et al. Exome Sequencing Reveals AMER1 as a Frequently Mutated Gene in Colorectal Cancer. Clin Cancer Res. 21(20), 4709-4718; 10.1158/1078-0432.CCR-15-0159 (2015).
68. Moore, L. D., Le, T. & Fan, G. DNA methylation and its basic function. Neuropsychopharmacology. 38(1), 23-38; 10.1038/npp.2012.112 (2013).
69. Eden, S. & Cedar, H. Role of DNA methylation in the regulation of transcription. Curr Opin Genet Dev. 4(2), 255-259; 10.1016/s0959-437x(05)80052-8 (1994).
70. Toyota, M. et al. Cpg island methylator phenotype in colorectal cancer. Proc Natl Acad Sci U S A. 96(15), 8681-8686; 10.1073/pnas.96.15.8681 (1999).
71. Jia, M. et al. Different definitions of CpG island methylator phenotype and outcomes of colorectal cancer: a systematic review. Clin Epigenet. 8, 25; 10.1186/s13148-016-0191-8 (2016).
72. Hughes, L. A. E. et al. The CpG Island Methylator Phenotype: What's in a Name? Cancer Research. 73(19), 5858-5868; 10.1158/0008-5472.CAN-12-4306 (2013).
73. Berg, M., Hagland, H.R. & Sereide, 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. 9(1), e86657; 10.1371/journal.pone.0086657 (2014).
74. Advani, S.M. et al. Clinical, Pathological, and Molecular Characteristics of CpG Island Methylator Phenotype in Colorectal Cancer: A Systematic Review and Meta-Analysis. Transl. Oncol. 11, 1188–1201; 10.1016/j.tranon.2018.07.008 (2018).
75. Issa, J. CpG island methylator phenotype in cancer. Nat Rev Cancer. 4, 988-993; 10.1038/nrc1507 (2004).
76. Kerachian, M. A. et al. Crosstalk between DNA methylation and gene expression in colorectal cancer, a potential plasma biomarker for tracing this tumor. Sci Rep. 10, 2813; 10.1038/s41598-020-59690-0 (2020).
77. Gundert, M. et al. Genome-wide DNA methylation analysis reveals a prognostic classifier for non-metastatic colorectal cancer (ProMColl classifier). Gut. 68, 101-110; 10.1136/gutjnl-2017-314711 (2019).
78. Freitas, M. et al. A novel DNA methylation panel accurately detects colorectal cancer independently of molecular pathway. J Transl Med. 16, 45; 10.1186/s12967-018-1415-9 (2018).
79. Vrba, L. & Futscher, B. W. A suite of DNA methylation markers that can detect most common human cancers. Epigenetics. 13(1), 61–72; 10.1080/15592294.2017.1412907 (2018).
80. Ma, X., Wang, Y. W., Zhang, M.Q. & Gazdar, A.F. DNA methylation data analysis and its application to cancer research. Epigenomics. 5(3), 301-316; 10.2217/epi.13.26 (2013).
81. Jones, P. A. The DNA methylation paradox. Trends Genet. 15, 34–37; 10.1016/s0168-9525(98)01636-9 (1999).
82. Ehrlich M. DNA hypermethylation in disease: mechanisms and clinical relevance. Epigenetics. 14(12), 1141-1163; 10.1080/15592294.2019.1638701 (2019).
83. Feng, S. et al. Conservation and divergence of methylation patterning in plants and animals. Proc Natl Acad Sci U S A. 107(19), 8689-8694; 10.1073/pnas.1002720107 (2010).
84. Muthamisellevan, Sangeetha, R., Raghavendran, Abirami, & Palaniappan, Ashok. Stage-differentiated modeling of methylation patterns uncovers salient biomarkers and prognostic signatures in colorectal cancer progression. medRxiv; 10.1101/2020.09.28.20203539 (2020).

Figures
Figure 1
Linear modelling with M-value matrix, all probes. (i) Distribution of number of significant genes in each stage relative to the control. (ii) Sample distribution obtained by plotting the first two principal components for the top 100 genes. A clear separation of controls and cancer cases (labelled by stage) could be seen.

Figure 2
Stagewise methylation levels of differentially methylated genes. (i) TMEM179

Figure 3
Stagewise methylation portrait using the top 4 significant stage-specific DMGs identified from the linear model at the probe level. The contrast with the control is especially evident.

Figure 4

Top 5 DMGs of the full linear model: (i) GPR75-ASB3, (ii) TM4SF19, (iii) CNRIP1, (iv) ADHFE1 and (v) KRTAP11-1. For each gene, notice that the trend in methylation could be either hyper- or hypo-methylation relative to the control. In particular, TM4SF19 and KRTAP11-1 are hypomethylated whereas CNRIP1, GPR75-ASB3, PDE4A are hypermethylated.

Figure 5

Linear modelling with M-value matrix, averep. (i) Distribution of number of significant genes in each stage relative to the control. (ii) Sample distribution obtained by plotting the first two principal components of the top 100 genes from the linear model. A clear separation of controls and cancer cases (labelled by stage) could be seen.
Figure 6

Figure 7

Stagewise methylation portrait using the top 4 significant stage-specific DMGs identified from the linear model at the gene level. The contrast with the control is especially outstanding.

Figure 8
Boxplots of top 5 genes of the linear model at the gene level. (i) TM4SF19, (ii) GPR75-ASB3, (iii) CNRIP1, (iv) KRTAP11-1, and (v) ADHFE1. For each gene, notice that the trend in expression could be either hyper- or hypo-methylation relative to the control. In particular, TM4SF19 and KRTAP11-1 are hypomethylated whereas GPR75-ASB3, CNRIP1, ADHFE1 are hypermethylated.

Figure 9
Venn diagram for CpG association analysis showing the distribution of number of significant genes in each stage.

Figure 10
Figure 11

(i) DMP plot of FCN2 for stage-I vs control illustrating significant hypomethylation (ii) 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.

Figure 12

Mixture models of the genes GATA4, CCDC88B, and WAS. The x-axis indicates the degree of methylation; the y-axis represents the frequency of that particular methylation degree; and the mixture component curves represent density fits of the histogram.
Venn diagram for BioMethyl-based Gene expression modelling showing the distribution of number of significant genes in each stage relative to the control.
Figure 15

Boxplot analysis of stage-salient genes. (i) Stage-I Gene FBN1, (ii) Stage-II Gene – FOXG1, (iii) Stage-III Gene – HCN1.

Figure 16
Figure 17

K-M plots for the prognostically significant stage-salient genes. (i) FBN1, (ii) FOXG1, (iii) HCN1, and (iv) LAMA1.

Survival analysis of combination biomarker panels shows significance. (i) Early-stage panel, CRS12; and (ii) Late-stage panel, CRS34.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- formula.docx
- SupplementaryFile.docx