medRxiv preprint doi: https://doi.org/10.1101/2020.09.28.20203539; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

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

Abirami Raghavendran[1&], Sangeetha Muthamilselvan[1&], Ashok Palaniappan[1*]

[1]Department of Bioinformatics, School of Chemical and BioTechnology, SASTRA Deemed University, Thanjavur 613401. India

[*]Corresponding author (apalania@scbt.sastra.edu)

[&]These authors contributed equally

ABSTRACT

Background: Aberrant methylation of DNA acts epigenetically to skew the gene transcription rate up or down. In this study, we have developed a comprehensive computational framework for the stage-specific analysis of methylation patterns in colorectal cancer.

Methods: Combining public-domain methylation and clinical data from TCGA, the methylation β - matrix was converted into M-value matrix, annotated with sample stages and analysed for stage-specific genes using multiple approaches involving stage-differentiated linear modelling of methylation patterns or their correlation with the phenotype and expression patterns. Differentially methylated genes (DMGs) were identified using a contrast against control samples (adjusted p-value <0.001 and |log fold-change of M-value| >2). These results were further filtered using a series of all possible pairwise stage contrasts (p-value <0.05) to obtain stage-specific DMGs. These were then subjected to a consensus analysis, followed by Kaplan–Meier survival analysis to explore the relationship between methylation and prognosis for the consensus stage-salient biomarkers.

Results: We found significant genome-wide changes in methylation patterns in cancer samples relative to controls agnostic of stage. Our stage-differentiated analysis yielded the following stage-salient genes: one stage-I gene ("FBN1"), one stage II specific gene ("FOXG1"), one stage III specific gene ("HCN1") and four stage IV specific genes ("NELL1", "ZNF135", "FAM123A", "LAMA1"), which could be used for stage-specific diagnosis. All the biomarkers were hypermethylated, indicating down-regulation and signifying a CpG island Methylator Phenotype (CIMP) manifestation. A prognostic signature consisting of FBN1 and FOXG1was significantly associated with patient survival (p-value < 0.01) and could be used as a biomarker panel for early-stage CRC prognosis.

Conclusion: Our workflow for stage-differentiated consensus analysis has yielded stage-salient diagnostic biomarkers as well as an early-stage prognostic biomarker panel. In addition, our studies have affirmed a novel CIMP-like signature in colorectal cancer, urging clinical validation.

Supplementary Information:

Supplementary information could be found at **https://doi.org/10.6084/m9.figshare.13013852**

Keywords:

Methylomics; stage-differentiation; colorectal cancer; consensus biomarkers; differential methylation; linear modelling; epigenomic patterns; survival prognosis; CpG island methylator phenotype.

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2020.09.28.20203539; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

INTRODUCTION

Colorectal cancer (CRC) is a clinically important malignant disease with devastating incidence and mortality, claiming the third spot among all cancers globally, only after lung and breast cancers, and accounting for 1.36 million new cases annually [1]. The etiology of CRC involves chromosomal instability (CIN) (involving accumulation of mutation in oncogenes and tumor suppressor genes), microsatellite instability (MSI) (leading to loss of DNA mismatch repair) and CpG island methylator phenotype (CIMP), each of which accounts for 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 severely aid tumor progression in the advanced stages [8]. More recently, Timp et al. found that large hypomethylation blocks (hundreds of kb) are a universal characteristic of colorectal cancers and other solid tumors [9]. DNA hypomethylation could also contribute to tumor initiation and progression by a general increase in genomic instability [10]. DNA hypermethylation could cause a loss of function of tumor suppressor genes, and hypermethylation in the germline could cause heritable loss of gene expression through genomic imprinting. Aberrant DNA hypermethylation of specific CpG islands has been observed to occur in colorectal cancer. The CpG island methylator phenotype (CIMP) was first described in a subset of colorectal cancers in 1999 [11] and was subsequently narrowed down to hypermethylation of five genes *CACNA1G*, *IGF2*, *NEUROG1*, *RUNX3* and *SOCS1* [12]. It is noted that the methylation changes contributing to phenotypic aberrations occur not only in the promoter regions but also in the gene coding regions and intron-exon structures[13-16].The persistence of such modifications throughout the tumor cell lifetime was demonstrated by Lengauer et al. [17], who also showed correlation between genome instability and methylation aberrations, suggesting that methylation abnormalities played a key role in tumorigenic chromosomal segregation processes.

The Cancer Genome Atlas (TCGA) is a comprehensive resource of genome-wide mutation, expression and DNA methylation profiles for 46 types of cancers [18]. Besides the TCGA, the International Human Epigenetic Consortium (IHEC) is specifically devoted to data-driven understanding of the role of epigenomics in normal vs disease states [19]. Methylation patterns constitute an emerging class of promising prognostic factors due to the following reasons: (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 [20]. Early detection may reduce the mortality rate by tailoring the treatment regimen, with fewer side-effects and better patient compliance. Few methylation markers have been previously translated to clinically applicable biomarkers [21]. Chen et al., heralded an era 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 [22]. 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-specific DMGs using several model-driven approaches, and evaluating their prognostic significance.

METHODS

Data Preprocessing:

Processed 27k methylation data (gdac.broadinstitute.org_COADREAD.Merge_methylation_humanmethylation27_jhu_usc_edu_Level_3_within_bioassay_data_set_function_data.Level_3.2016012800.0.0.tar) was retrieved from The Cancer Genome Atlas (TCGA) through firebrowse portal (www.firebrowse.org) [23] and the latest clinical data

medRxiv preprint doi: https://doi.org/10.1101/2020.09.28.20203539; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

(clinical.cases_selected.tar.gz) was obtained from the GDC data portal (https://portal.gdc.cancer.gov/repository) by matching on patient barcode.

The data containing the methylation β-values for each probe in each sample was converted into a matrix with probes as rows and samples 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 [24]:

$$M_i = \log_2\left[\frac{\beta_i}{1-\beta_i}\right] \quad [1]$$

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 ("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 $\sigma < 1$) across all samples 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. Samples with "NA" values for stage information were discarded from the analysis. The sample counts in various stages are represented in Table 1.

Table1. Sample Counts in COADREAD 27kcancer data from TCGA. In the final analysis, the sub-stages were combined into the parent stage.

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

medRxiv preprint doi: https://doi.org/10.1101/2020.09.28.20203539; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

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 [25].

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 limma, we used the R LIMMA package [26] for Linear modelling
Using the complete M-value matrix, with multiple probes per gene (given in S1).
In both the linear models, the control samples contributed to the intercept of the design matrix, while the stages were represented as indicator variables [27]. 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 (described below).
(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 (dataset provided in S2). These alternative representations of the methylation data negotiate a tradeoff with respect to information preservation and interpretability.
(3) Association between methylation status and phentoype: The strength of the association between the methylation levels of CpG sites and the phenotype of interest (CRCstage) could enable the identification of relevant markers. We used the R package CpGassoc [28] to perform this analysis with the stage-annotated β-matrix as input. Five factors were specified (control, stage I, stage II, stage III, stage IV), and the stage-differentiated contrast analysis protocol (given below) was then carried out.
(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 [29]. A mapping of sample IDs with the clinical stage phenotype was provided as an additional input file. The results from this analysis were subjected to the stage-differentiated contrast analysis. 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) [30]. GSEA was used to identify the enrichment of DMPs and DMRs in the MSigDB pathways [31], using the Fisher Exact test calculation with adjusted p-value < 0.05.
(5) Modelling expression from methylation: We used the R BioMethyl package to model a gene's aggregate expression level from its methylation patterns recorded in the β-matrix [32]. The estimated gene expression matrix was then subjected to linear modelling with limma, and subsequently stage-differentiated contrast analysis was carried out.
(6) Correlation between gene methylation and expression: We used MethylMix2.0 to estimate the correlation between the actual methylation and expression patterns of each gene [33]. The expression data for the samples 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 preset p-value (< 0.05).

Stage-differentiated contrast analysis (common to the linear models)

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

medRxiv preprint doi: https://doi.org/10.1101/2020.09.28.20203539; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

(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, aputative DMG identified in Tier I would undergo three inter-stage contrasts in Tier II, to ensure stage-specificity. For example, a stage-II DMG established by Tier I, would undergo the following inter-stage contrasts: Stage-II vs Stage-I, stage-II vs stage-III and stage-II vs stage-IV, before being confirmed as a stage II-specific 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. If three methods agreed for any stage-specific differentially methylated gene DMG, it was adjudged as a consensus stage-salient biomarker.

Survival analysis

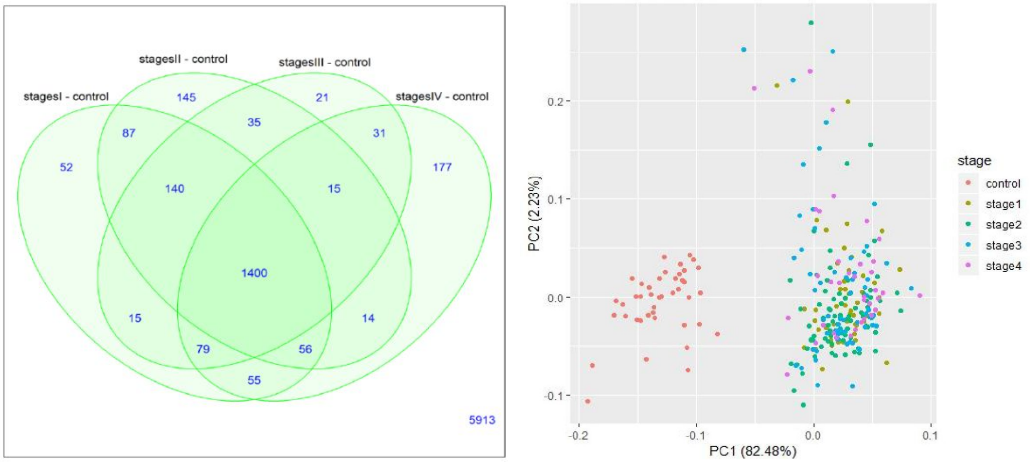
The survival data for each patient was obtained from the following attributes encoded in the clinical data: patient.vital_status, patient.days_to_followup, and patient.days_to _death. The association between consensus stage-specific differentially methylated genes and case overall survival (OS) was evaluated by univariate Cox proportional hazards regression model using the R survival package [34]. 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 patient. Based on this risk score, patients with colorectal cancer were categorized into high and low groups by the optimal cut point determined by the maxstat (maximally selected rank statistic) value [35]. Kaplan-Meier estimation was used to evaluate if the median survival time was significantly different between these two groups. This enabled the assessment of prognostic relevance of the identified consensus stage-specific differentially methylated genes.

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 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 differentially methylated genes identified from the linear model were visualized using boxplots (Supplementary Information S3). Figure 2 represents the boxplots for TMEM179, mutations in which could cause MSI [36], and MEOX2 whose promoter methylation status is a known CRC marker [37].The significant genes for each stage were identified and the first four genes from each stage were represented in the heatmap in figure 3.The top ten genes (ranked by adjusted p-value) in each stage with respect to control are shown in Table 2, along with their |lfc M-valuel and inferred regulation status. Figure 4 shows boxplots of methylation levels. The top 100 genes significant for each stage were used for the consensus analysis and the lists of genes that were specific in each stage was given in S3.

medRxiv preprint doi: https://doi.org/10.1101/2020.09.28.20203539; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .



(i) (ii)

Figure 1. (i).Venn diagram for M-value analysis showing the 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 from the linear model. A clear separation of controls and cancer samples (labelled by stage) could be seen.

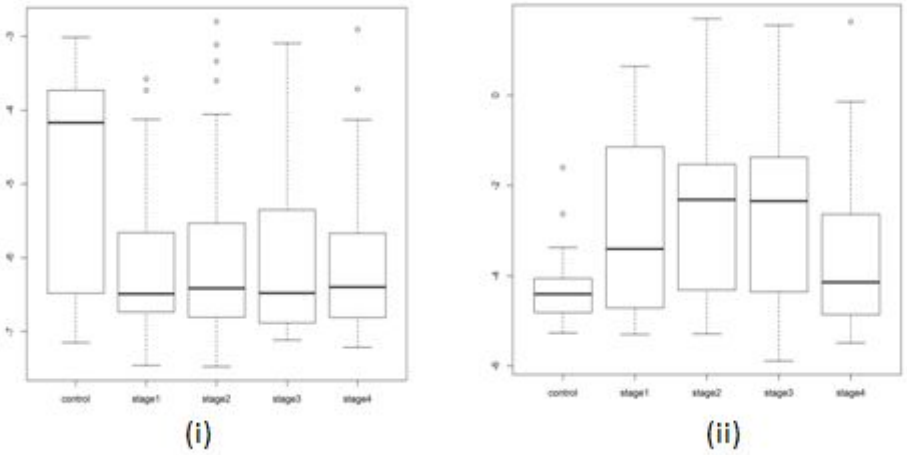


Figure 2. Boxplots for the differentially expressed Methylated genes using M value analysis (i) TMEM179 (ii) MEOX2

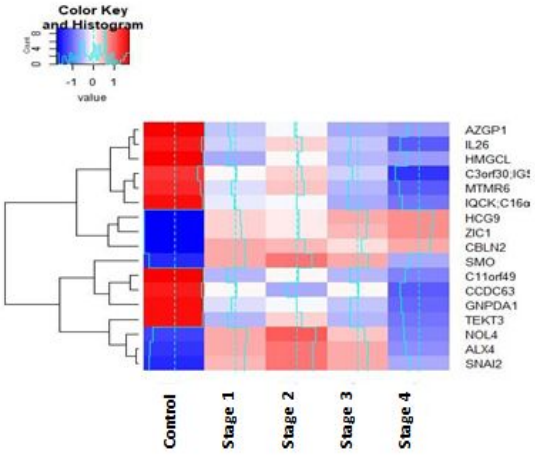


Figure 3. Heatmap for significant Differentially Methylated genes – M value analysis

medRxiv preprint doi: https://doi.org/10.1101/2020.09.28.20203539; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Table 2. Top 10 genes of the linear model at the probe level. The log fold-change of mvalue 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

ID	StageI	StageII	StageIII	StageIV	adj. p-val	Methylation status
LOC100302652;GPR75	2.280584	2.190314	2.159852	2.320709	2.90E-82	Hyper
TM4SF19	-3.62705	-3.57631	-3.72308	-3.71059	4.08E-82	Hypo
CNRIP1	2.743322	2.605539	2.678591	2.974642	5.57E-78	Hyper
NA	1.682123	1.578549	1.599996	1.707521	1.27E-71	Hyper
KRTAP11-1	-2.36005	-2.29961	-2.37583	-2.39962	9.04E-70	Hypo
ADHFE1	3.153667	2.967236	3.000252	3.432456	2.03E-69	Hyper
FAM123A	3.56288	3.181145	3.429552	3.895594	5.54E-69	Hyper
KHDRBS2	2.302345	2.16154	2.103098	2.339009	4.41E-68	Hyper
AJAP1	2.528654	2.438316	2.462037	2.6367	9.36E-68	Hyper
NALCN	2.959115	2.796519	2.949049	3.250277	9.36E-68	Hyper

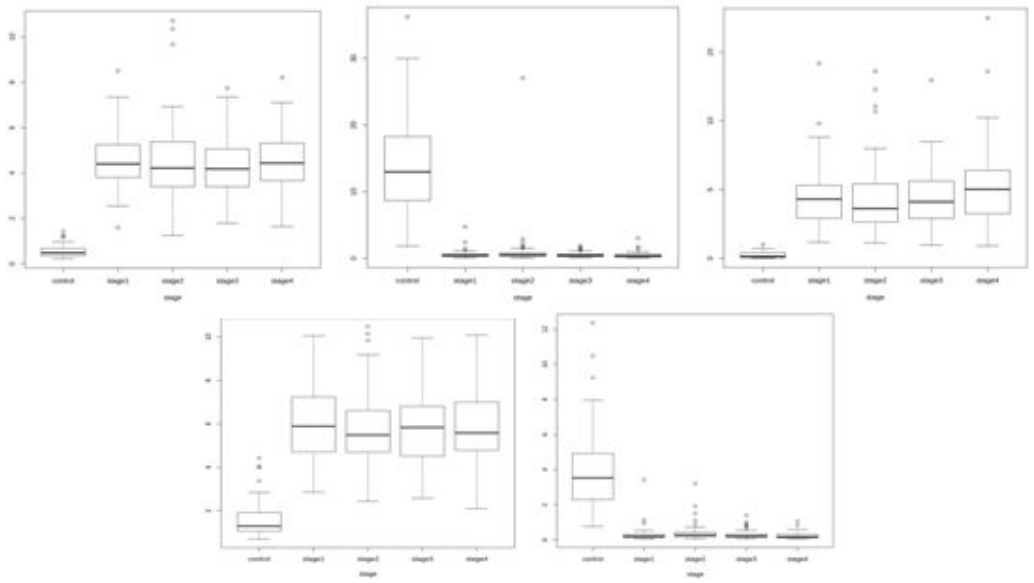


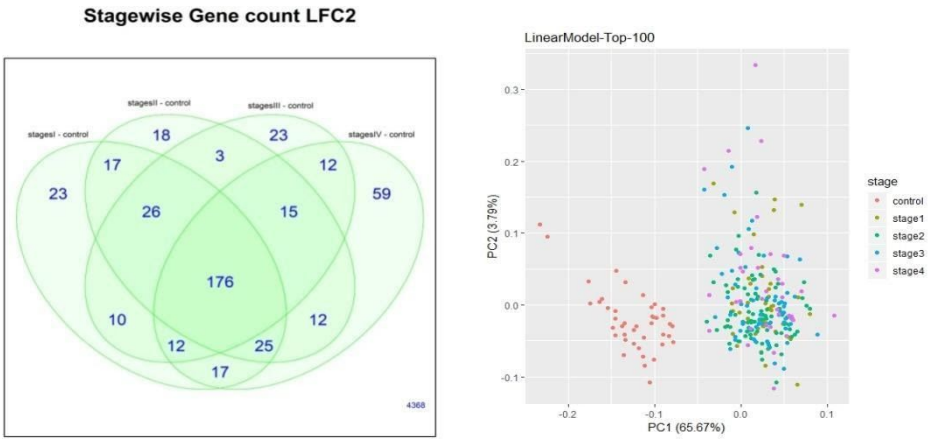
Figure 4. Boxplots of top 5 linear model genes. For each gene, notice that the trend in methylation could be either hypermethylated or hypomethylated relative to the control. For example TM4SF19 and KRTAP11-1 are hypomethylated where as CNRIP1, LOC100302652;GPR75 are hypermethylated.

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 differentially methylated genes identified from linear model were visualised using boxplots (Supplementary information S4). Figure 6 shows the boxplots for two such genes, NALCN and GLRX. Mutations in NALCN have been reported in sporadic CRC [38]; here NALCN is seen to be significantly hypermethylated (Figure 6),

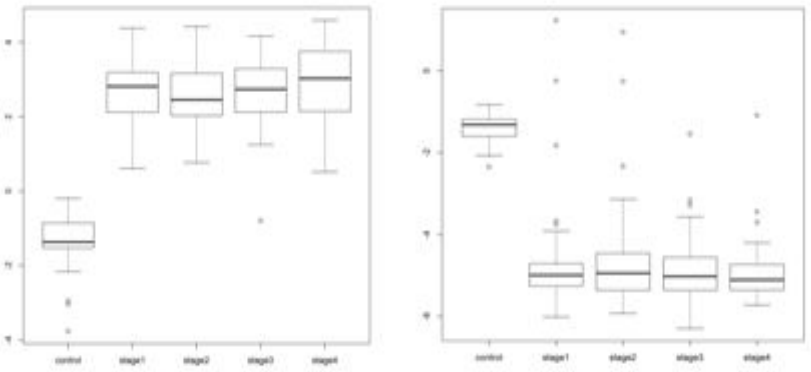
medRxiv preprint doi: https://doi.org/10.1101/2020.09.28.20203539; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

indicating the same outcome (loss of function) could be effected in multiple ways. GLRX is a target of the activating transcription factor MEOX2 [39]. The significant stage-specific genes were identified, and the top four genes based on adj. P-values from each stage were used to construct the heatmap in Figure7. The top ten genes (ranked by adjusted p-value) in each stage with respect to control are shown in Table 3, along with their |lfc M-value| and inferred regulation status. Figure 8 shows boxplots of the methylation levels. It is observed that *LY6H* showed both hypermethylation and hypomethylation when compared to the control samples, indicating the need for experiments to clarify its role in colorectal cancer progression. The top significant 100 genes of each stage were used for consensus analysis and the lists of genes that were specific in each stage was given in S4.



(i) (ii)

Figure 5. (i).Venn diagram for Gene-level analysis showing the 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 from the linear model. A clear separation of controls and cancer samples (labelled by stage) could be seen.



(i) (ii)

Figure 6. Boxplots for the differentially expressed Methylated genes using averep analysis (i) NALCN (ii) GLRX

medRxiv preprint doi: https://doi.org/10.1101/2020.09.28.20203539; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

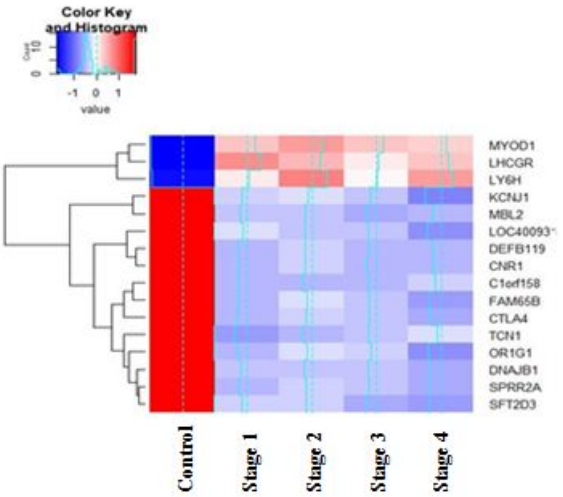


Figure 7. Heatmap for significant Differentially Methylated genes – Gene mapped analysis

Table 3. Top 10 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	StageI	StageII	StageIII	StageIV	adj.P.Val	Methylation status
TM4SF19	-3.62512	-3.576	-3.72406	-3.70707	1.01E-82	Hypo
LOC100302652;GPR75	2.279406	2.186388	2.154401	2.319774	7.39E-82	Hyper
CNRIP1	2.742948	2.602632	2.673634	2.97448	2.74E-77	Hyper
KRTAP11-1	-2.35972	-2.30206	-2.37938	-2.3988	3.65E-70	Hypo
ADHFE1	3.152625	2.962694	2.994224	3.4316	4.02E-69	Hyper
FAM123A	3.561976	3.177355	3.423787	3.894354	1.01E-68	Hyper
AJAP1	2.527459	2.435239	2.458005	2.635741	3.21E-67	Hyper
NALCN	2.957453	2.792623	2.943573	3.248803	1.07E-65	Hyper
IRF4	1.98933	1.824373	1.887686	2.128157	1.07E-65	Hyper
PRKAR1B	3.381341	3.131659	3.244301	3.494884	1.07E-65	Hyper

medRxiv preprint doi: https://doi.org/10.1101/2020.09.28.20203539; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

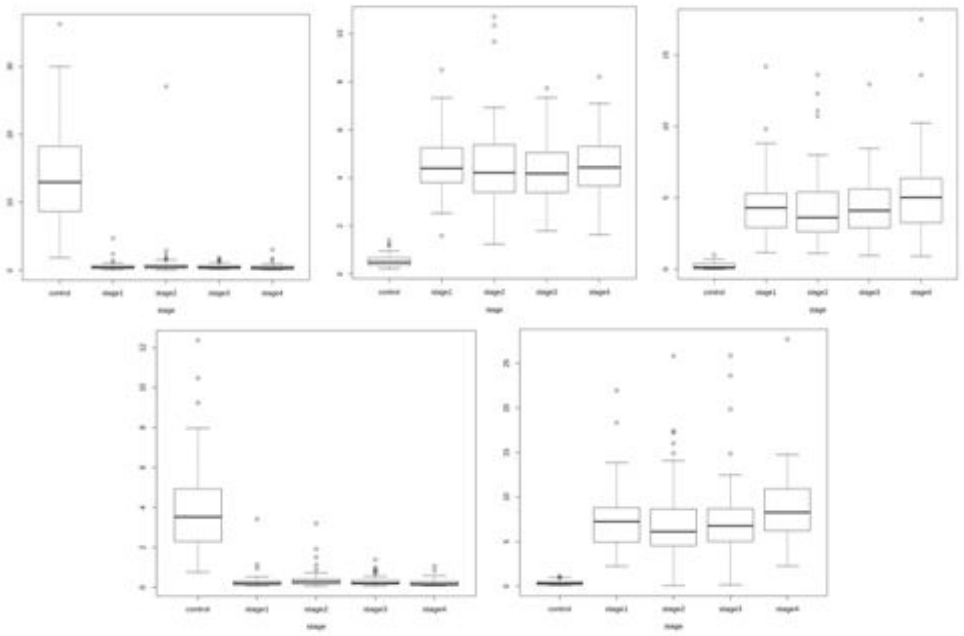


Figure 8. Boxplots of top 5 linear model genes. For each gene, notice that the trend in expression could be either hypermethylated or hypomethylated relative to the control. For example TM4SF19 and KRTAP11-1 are hypomethylated where as LOC100302652;GPR75 , CNRIP1, ADHFE1are hypermathylated

Association with phenotype

The results of the CpGassoc analysis are shown in Figure 9. Significant genes for each stage were identified using the criteria given in Methods (Figure 9). The top 100 genes of each stage from this analysis (given in Supplementary Information S5) were taken for consensus analysis.

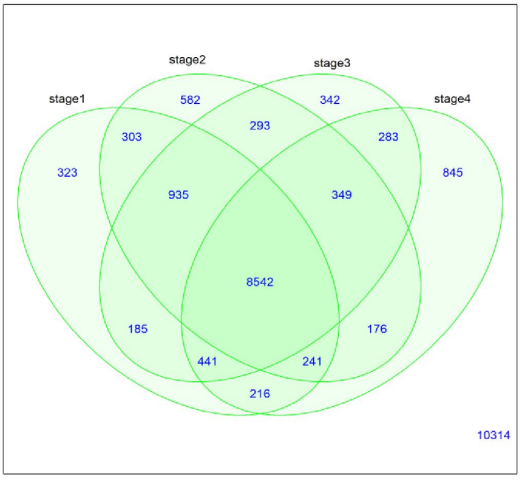


Figure 9. Venn diagram for CpG association analysis showing the distribution of number of significant genes in each stage

DMP analysis with ChAMP

The quality of the data was evaluated using ChAMP and shown in Figure 10.

medRxiv preprint doi: https://doi.org/10.1101/2020.09.28.20203539; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

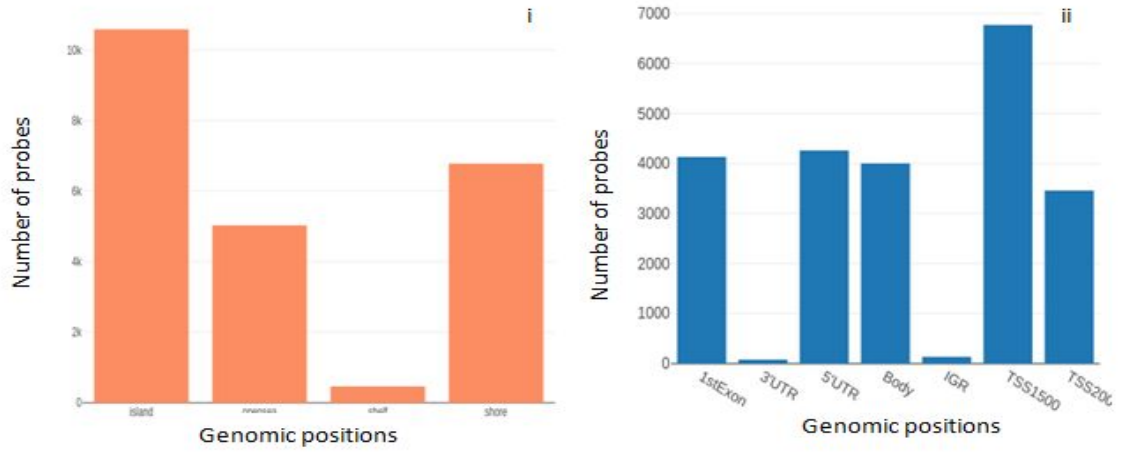


Figure 10. Data distribution (i) based on Genomic position: opensea, shore, island, shelf; (ii) based on gene context: TSS, exon IGR, 5'UTR, 3'UTR (ii) based on TSS region

The DMPs were identified using CHAMP analysis from the β matrix, and illustrated in Figure 11(i). Figure 11(ii) shows a representative DMR plot. In the ChAMP analysis, most of the inter-stage contrasts yielded null results (i.e, no significant genes), except for stageII – stageIV contrast. For this reason, the top 100 DMPs from the stage vs control contrasts were used for consensus analysis directly. Contrasts with significant DMPs were subjected to DMR analysis. The stage-specific DMR regions determined are provided in S6, and summarized in Table 4. It was noted that 31 DMRs were identical across stages (given in S7).The DMPs and DMRs from the analysis were subjected to GSEA and the results of the procedure are provided in S6.

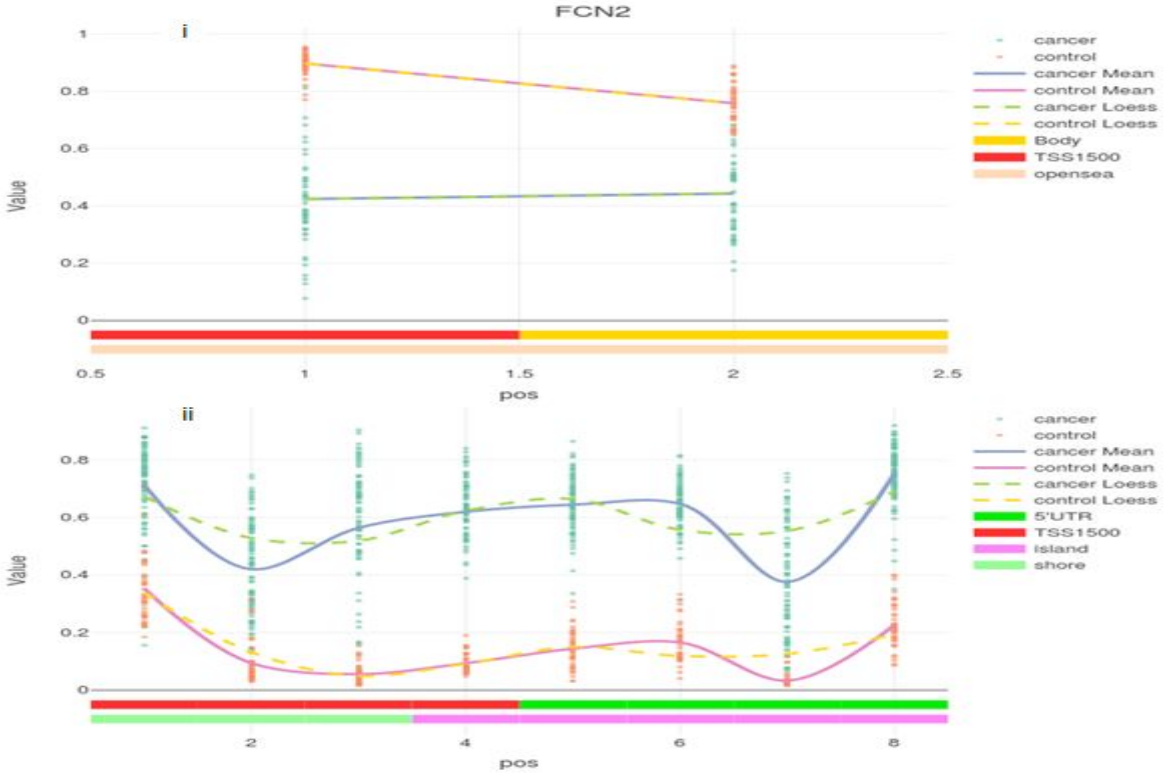


Figure 11. (i) DMP plot of FCN2 for control and stage I illustrating significant hypomethylation (ii) DMR plot EYA4 transcriptional activator for control and stage 1

medRxiv preprint doi: https://doi.org/10.1101/2020.09.28.20203539; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

illustrating significant hypermethylation. Solid lines represent mean values while dashed lines represent the loess.

Table 4. Number of DMRs for each comparison

Comparisons	Number of Differentially Methylated Regions
Control and Stage 1	34
Control and Stage 2	35
Control and Stage 3	36
Control and Stage 4	34
Stage 2 and Stage 4	3

Methylation and Gene Expression Correlation analysis report

Figure 12 illustrates the correlation determined by MethylMix between the methylation levels and gene expression. Mixture models of genes, indicative of the number of methylation states, were constructed using MethylMix (Figure 13). No driver genes were determined for the between stage comparison based on this algorithm. The number of driver genes that were identified in comparisons is represented in Table 5. This package produces results calculating the DM values (differentially methylated values) based on which the hyper or hypomethylated genes were determined. The range of DM values was from -0.7 to 0.8. All the genes which were identified as stage IV specific topped among the fifty genes when sorted using DM values. These results were provided in S8. Top 100 genes from each comparison were taken for consensus analysis. It was seen that 56 genes were common in all the four comparisons; notably, *CCDC88B, C1orf59, CHFR, ZP2, HOXA9, ELF5, FAM50B, MUC17, TBX20, VSIG2* were the top ten common genes. These 56 genes represented stage-agnostic genes which hold immense potential for the treatment of colorectal cancer; they are listed in the Supplementary Information S9.

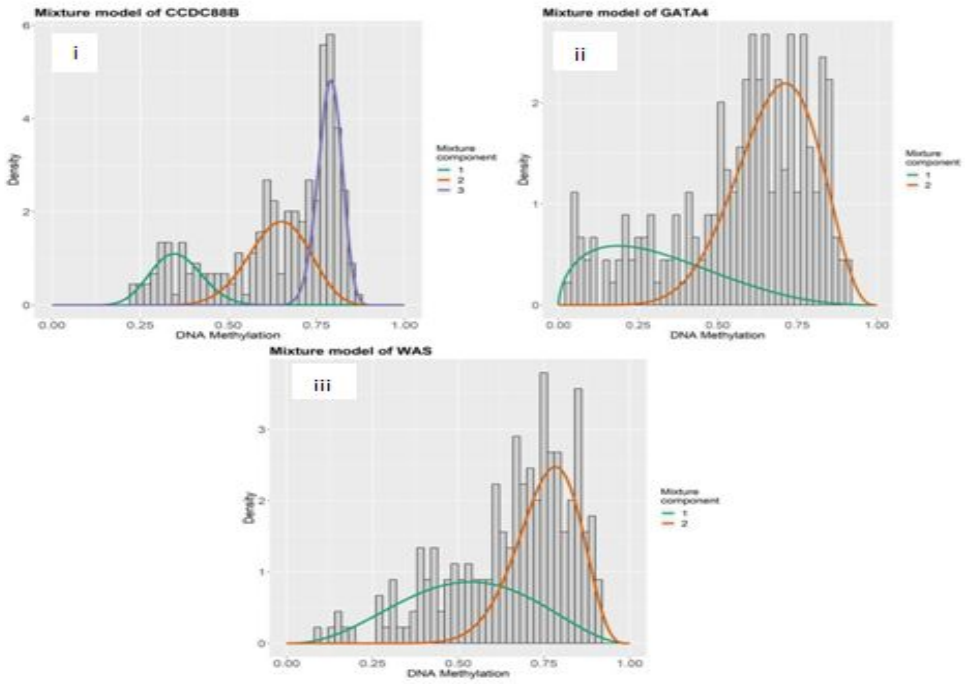


Figure 12. Mixture models of the genes GATA4, CCDC88B, and WAS. The x-axis indicates the degree of methylation, the y-axis indicates the proportion of density at different degrees, and the mixture component curves are density fits of the values.

medRxiv preprint doi: https://doi.org/10.1101/2020.09.28.20203539; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

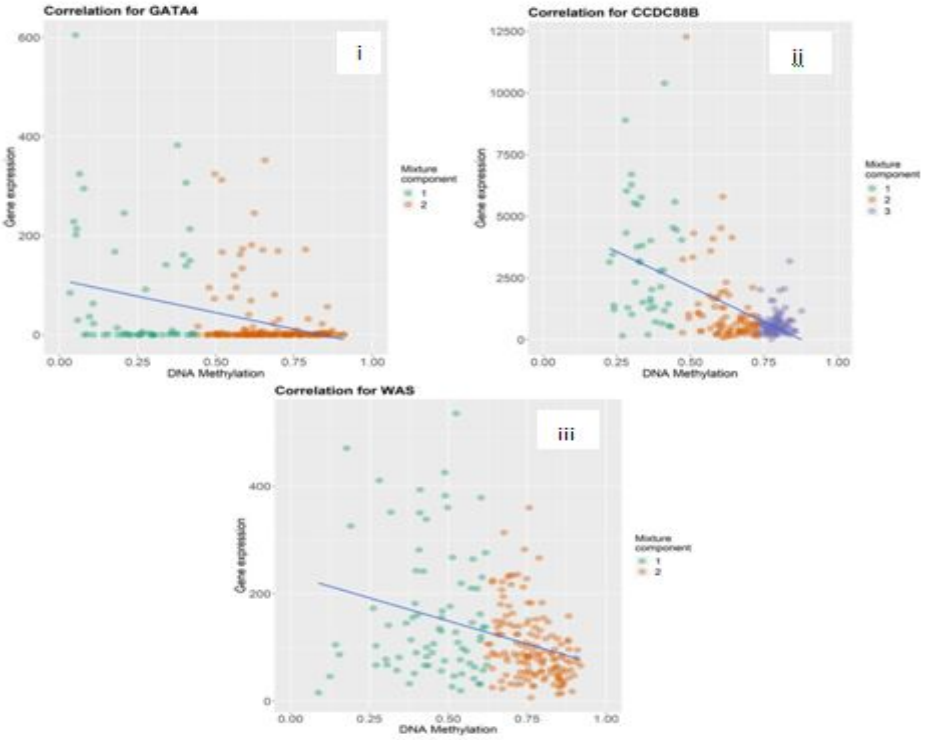


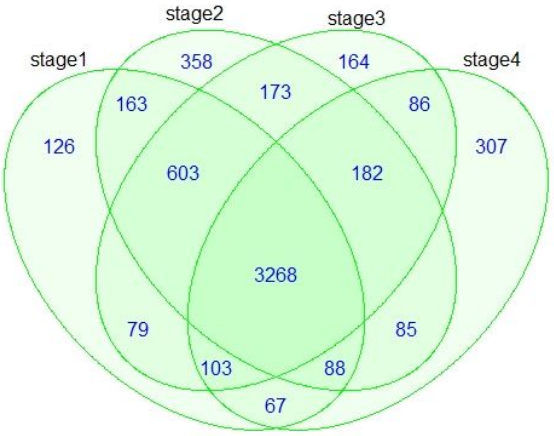
Figure 13. Correlation plots for i. GATA4, ii.CCDC88B, iii.WAS. A negative correlation between methylation and expression is evident, indicating that methylation acts to repress gene transcription, though the strength of the inverse correlation varies from gene to gene. Colour indicates the mixture model fit (cf. Fig. 12).

Table 5. Number of driver genes in each comparison

Comparisons	Number of Driver genes
Control and Stage 1	209
Control and Stage 2	441
Control and Stage 3	278
Control and Stage 4	152

BioMethyl results

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.



medRxiv preprint doi: https://doi.org/10.1101/2020.09.28.20203539; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Figure 14. Venn diagram for BioMethyl-based Gene expression modelling showing the distribution of number of significant genes in each stage relative to the control.

Stage-salient biomarkers

The top 100 significantly differentially-expressed genes of each stage from all the methods discussed above (provided in Supplementary Info S11) were used for the consensus determination. Our consensus analysis yielded seven stage-specific DMGs: one stage I specific genes *("FBN1")*, one stage II specific gene *("FOXG1")*, one stage III specific gene *("HCN1")* and four stage IV specific genes *("NELL1", "ZNF135", "FAM123A", "LAMA1")*. It was validated that each of these stage-specific genes presented an |lfc M-value| > 0.4 with respect to the other stages, establishing their stage-salience. Figures 15,16 represent the boxplots for the consensus biomarkers and Table 6 presents a summary of the consensus analysis. Gene ontology analysis of the consensus biomarkers yielded processes related to structural integrity of cell division processes, immunity dysfunction, and cell migration (Table 7) [40]. Detailed GO results are presented in the Supplementary Information S12.

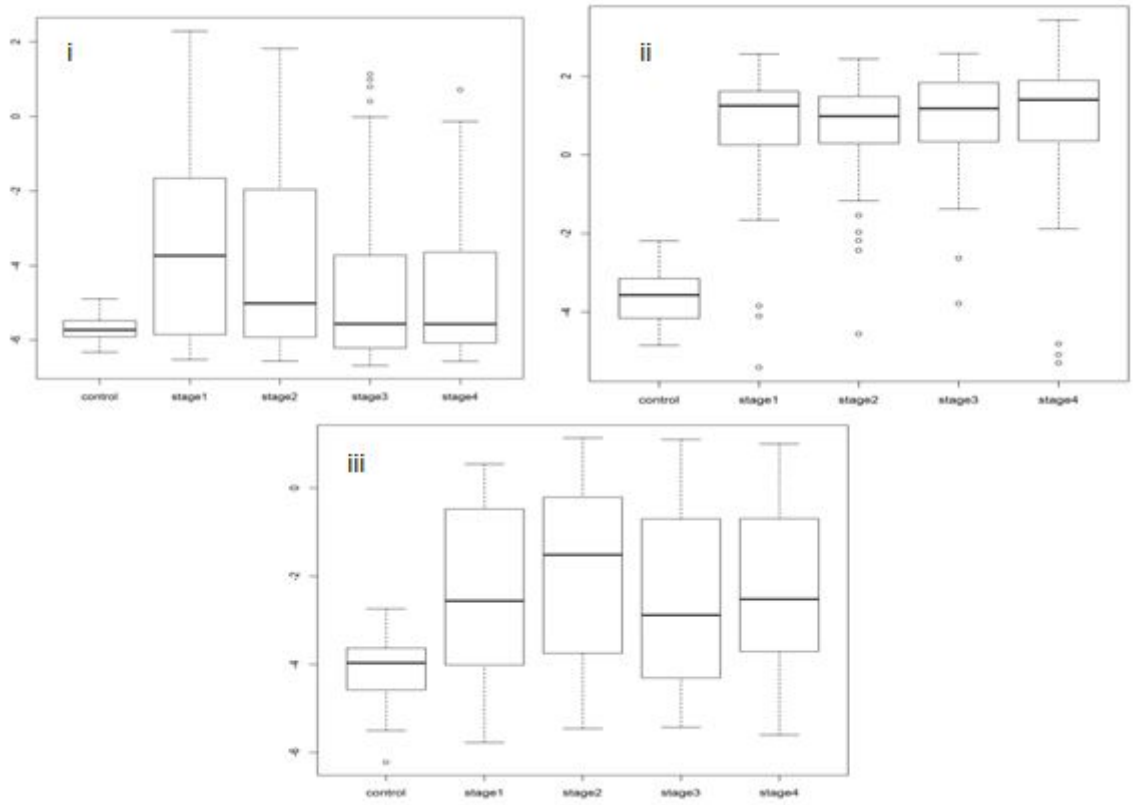


Figure 15. Boxplots for stage-salient genes. (i) Stage I specific Gene FBN1, (ii) Stage II specific Gene – FOXG1, (iii) Stage III specific Gene – HCN1.

medRxiv preprint doi: https://doi.org/10.1101/2020.09.28.20203539; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

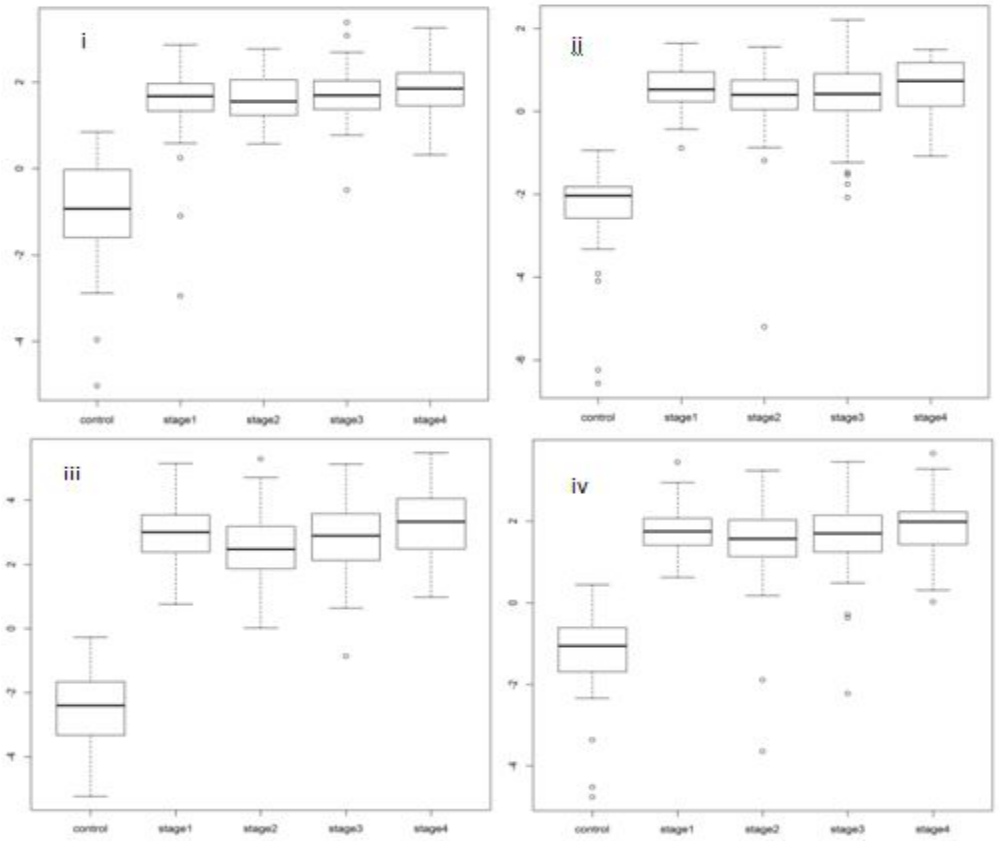


Figure 16. Boxplots for Stage IV specific Genes (i) LAMA1, (ii) NELL1, (iii) FAM123A, (iv) ZNF135.

Table 6. Stage-salient biomarkers. The results of the consensus analysis and univariate survival analysis are summarized.

HGNC ID	Gene Name	Methods providing identical results	Stage salience	Status of methylation	Inferred Effect on gene expression	Pvalue from modelling		p-values from univariate survival analysis	
						M value	Averep	Cox analysis	Kaplan Meier
3603	FBN1	Averep, CHAMP	I	Hypermethylation	Down	0.310	0.040	0.036	0.025
3811	FOXG1	Averep, Mvalue, CHAMP, Methylmix	II	Hypermethylation	Down	3.25E-16	0.003	0.019	0.037
4845	HCN1	Averep, Mvalue, CHAMP	III	Hypermethylation	Down	1.32E-17	0.022	0.031	0.059
7756	NELL1	Mvalue, CHAMP, Methylmix	IV	Hypermethylation	Down	2.49E-68	0.0614	0.283	0.27
12919	ZNF135	Mvalue, CHAMP,	IV	Hypermethylation	Down	1.29E-76	0.0622	0.096	0.084

medRxiv preprint doi: https://doi.org/10.1101/2020.09.28.20203539; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

		Methylmix							
26360	FAM123A	Mvalue, CHAMP, Methylmix	IV	Hypermethylation	Down	5.09E-115	0.0966	0.30	0.28
6481	LAMA1	Mvalue, CHAMP, Methylmix	IV	Hypermethylation	Down	7.57E-86	0.297	0.052	0.051

Table 7. Gene Ontology for Stage specific genes

	Term	Ont	N	DE	P.DE
GO:1990047	spindle matrix	CC	1	1	0.000148
GO:0030109	HLA-B specific inhibitory MHC class I receptor activity	MF	2	1	0.000297
GO:0032396	inhibitory MHC class I receptor activity	MF	4	1	0.000594
GO:0042609	CD4 receptor binding	MF	8	1	0.001187
GO:0032393	MHC class I receptor activity	MF	9	1	0.001336
GO:0050930	induction of positive chemotaxis	BP	11	1	0.001632
GO:0050927	positive regulation of positive chemotaxis	BP	22	1	0.003263
GO:0050926	regulation of positive chemotaxis	BP	23	1	0.003411
GO:0008608	attachment of spindle microtubules to kinetochore	BP	29	1	0.004299
GO:0007094	mitotic spindle assembly checkpoint	BP	30	1	0.004448

Survival analysis:

We constructed independent prognostic models as well as combined models known Colorectal cancer Signatures (CRS) obtained by integrating the stage-salient genes. The results of the survival analysis are summarised in Table 8. The Kaplan-Meier plots for the stage salient genes are shown in Fig. 17. From the K-M analysis, we found that CRS12 and CRS34 signatures were effective in discriminating cases with significantly worse OS ($P \leq 0.02$) based on the estimated signature scores (Figure 18). Survival plots for all possible signatures are provided in the Supplementary Information S13.

medRxiv preprint doi: https://doi.org/10.1101/2020.09.28.20203539; this version posted September 29, 2020. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

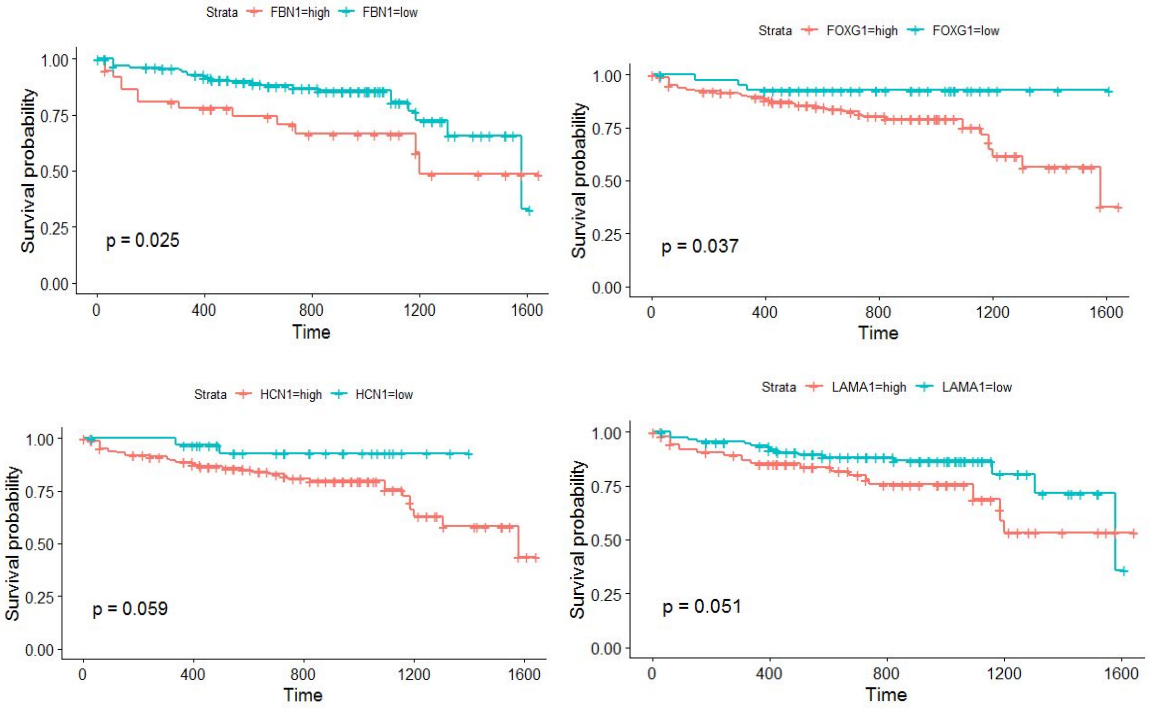


Figure 17. K-M plots for the prognostically significant stage-salient genes. (A) FBN1, (B) FOXG1, (C) HCN1, and (D) LAMA1.

Table 8. Coefficients based on cox regression analysis

Signature names	Stages	Methylated Genes	Coeff	Multivariate P value	Prognosis value
CRS12	Stage I+II	FBN1	-0.6224	0.01526	0.005
		FOXG1	-1.0549		
CRS34	Stage III+IV	NELL1	0.09638	0.172	0.02
		ZNF135	-0.21216		
		FAM123A	-0.22547		
		LAMA1	-0.39093		
		HCN1	-1.101045		
CRS4	Stage IV	NELL1	0.0984	0.3978	0.12
		ZNF135	-0.24114		
		FAM123A	-0.05873		
		LAMA1	-0.51838		
CRS234	Stage II+III+IV	FOXG1	-0.99151	0.0874	0.032
		HCN1	-1.07172		
		NELL1	-0.09822		
		ZNF135	-0.21515		
		FAM123A	-0.37211		
		LAMA1	-0.27019		
CRSall	STAGE I+II+III+IV	FBN1	-0.4346	0.9875	0.041
		FOXG1	-0.9381		
		HCN1	-1.0038		
		NELL1	0.2649		
		ZNF135	0.2210		
		FAM123A	-0.3165		
		LAMA1	-0.1817		

medRxiv preprint doi: https://doi.org/10.1101/2020.09.28.20203539; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

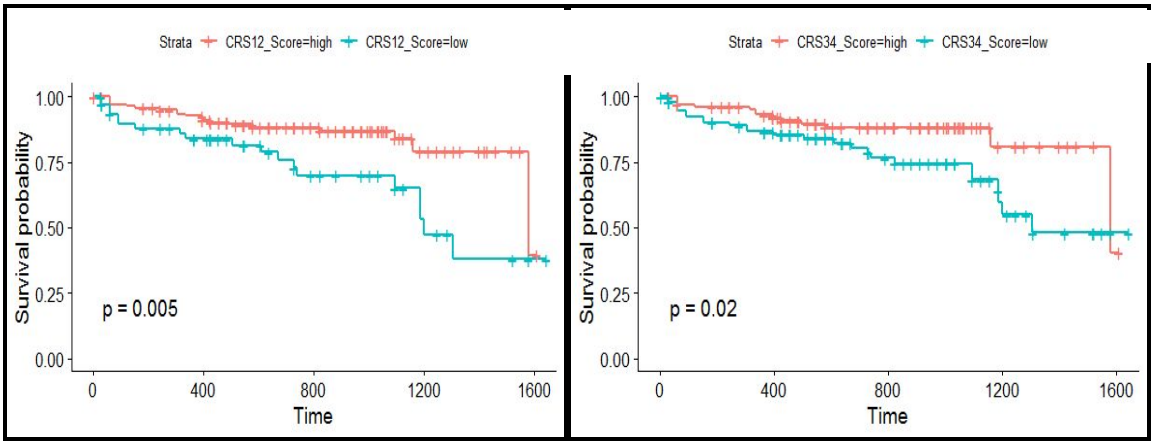


Figure 18. Survival analysis for the combination biomarker panels shows significance. (A) Early-stage panel; and (B) Late-stage panel.

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 [41, 42]. It is well-known that changes in methylation status correspond with CRC progression [43]. 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 and evaluated their prognostic value. A total of seven genes were identified by at least two methods, and of those, six were identified by at least three methods, the exception being the stage I-specific FBN1. None of the stage-salient genes is included as a cancer gene or hallmark gene in the Cancer Gene Census [44], while HCN1 alone is reported as a candidate cancer gene based on mouse insertional mutagenesis experiments [45]. Below, a discussion of all the stage-salientDMGs (Table 6) is provided with respect to the existing literature.

Early-stage salient DMGs:

Promoter hypermethylation of FBN1, a glycoprotein component of calcium-binding extracellular matrix microfibrils [46], is a recognized biomarker of CRC [47, 48]. Our analysis supports this literature, while pinpointing the stage I-specificity 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 patients [49]. It is a nodal gene, with connections to oncogenic pathways like WNT pathway in hepatocellular carcinoma [50] and TGF-β pathway in ovarian cancer [51]. Interestingly, FOXG1 was found to be a hypermethylated stage-II specific gene. HCN1, coding for hyperpolarization-activated cyclic nucleotide-gated channel is associated with low survival rates in breast, brain, and colorectal cancer [52]. 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 clearly indicated 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 promoter region of LAMA1 in CRC patients is available [53]. NELL1 is a known tumor suppressor gene [54], whose hypermethylation is associated with

medRxiv preprint doi: https://doi.org/10.1101/2020.09.28.20203539; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

poor survival outcomes [55]. Our results that NELL1 is a stage IV-specific hypermethylated gene resonates with the above findings. ZNF135 is involved in regulation of cell morphology and cytoskeletal organizations, and its expression and epigenetic regulation have been reported to be key in cancers of the cervix and esophagus, respectively [56, 57]. Here we have found that epigenetic silencing of ZNF135 is a key feature of stage-IV CRC. FAM123A, also known as AMER2, is associated with microtubule proteins [58], and is a lesser known cousin of FAM123B, a tumor-suppressor whose loss-of-function by mutation, methylation and copy-number aberrations is known play pivotal roles in colorectal cancer, especially in older patients [59,60,61]. 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.

Aberrant methylation of CpG promoter regions causes stable repression of transcription leading to gene-silencing [62,63]. 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 [64]. CIMP is characterised by hypermethylation of CpG islands surrounding the promoter regions of genes involved in cancer onset and progression [65]. The phenotype is heterogenous with the type of tumor [66] and dependent on definition [67]. In this background, it is less straightforward to interpret the functional importance of hypermethylation of individual genes. Still it is clear from Table 6 that the stage-salient hypermethylated biomarkers identified in our study could constitute an aggregate novel CIMP. The CIMP has been associated with advanced T staging (T3/T4) [68], 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 [69].

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 6). 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 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 [70,71]. Freitas et al., have validated the performance of a 3-gene biomarker panel for the detection of colorectal cancer irrespective of the molecular subtype [72]. However optimal stage-salient epigenetic biomarkers have not yet been

medRxiv preprint doi: https://doi.org/10.1101/2020.09.28.20203539; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

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 subsequent design of analytical amplicons[73].Hypermethylation in gene body and upstream control regions like enhancers and insulators might affect transcription differently than hypermethylation of promoter regions [74.75]. Further DNA methylation in noncoding RNA genes seem to be important in tumor formation and progression [76]. Non-encoding RNAs themselves play a significant role in epigenetic modification through the phenomenon of RNA-directed DNA methylation [77].The nuanced relationship between methylation and gene transcription does urge the interpretation of our results with caution, contingent on experimental validation, however we would reiterate the efficacy of our study design in teasing out valid epigenetic biomarkers. Since methylation is a direct, ubiquitous and effective mechanism of epigenetic regulation used by plants and animals [78], it is hoped that our studies would further advance our understanding of the complex nuanced effects of methylation events and patterns, including in the developmental stages of life.

CONCLUSION

We have developed a stage-differentiated comprehensive computational framework for the consensus identification of significant differentially methylated genes. Our analysis has yielded seven stage-salient genes, all hitherto unreported in the literature: one stage I-specific gene *("FBN1")*, one stage II-specific gene *("FOXG1")*, one stage III-specific gene *("HCN1")* and four stage IV-specific genes *("NELL1", "ZNF135", "FAM123A", "LAMA1")*.All the stage-salient genes were found to be hypermethylated, indicating a novel CIMP-like character possibly promoting epigenetic destabilisation that merits further investigations. Prognostic evaluation of biomarker signatures composed of these stage-salient genes yielded the highest significance for a panel of the four stage IV-specific genes. The robust consensus approaches employed here would increase the reliability of identified epigenetic biomarkers both for diagnostic and prognostic purposes, and greatly advance the early detection of cancers and their treatment. Our approach is extendable to the investigation of epigenomics in other cancers, normal/disease conditions, and developmental biology.

ACKNOWLEDGMENTS

We are grateful to the School of Chemical and BioTechnology, SASTRA Deemed University for research, computing and infrastructure support. A.P. would like to acknowledge funding from DST-SERB grant EMR/2017/000470/BBM.

COMPETING INTERESTS

The authors declare no competing interests, financial or otherwise.

AUTHOR CONTRIBUTIONS

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

medRxiv preprint doi: https://doi.org/10.1101/2020.09.28.20203539; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

REFERENCES

1. Ferlay J, Soerjomataram I, Dikshit R, Eser S, Mathers C, Rebelo M, Parkin DM, Forman D, Bray F. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN. Int J Cancer. 2015;136:359–386. DOI: 10.1002/ijc.29210.
2. Carethers JM, Jung BH. Genetics and genetic biomarkers in sporadic colorectal cancer. Gastroenterology. 2015;149(5):1177–1190. DOI:10.1053/j.gastro.2015.06.047
3. Ogino S, Goel A. Molecular classification and correlates in colorectal cancer. J Mol Diagn. 2008;10(1):13–27. DOI:10.2353/jmoldx.2008.070082.
4. Chen JJ, Wang AQ, Chen QQ. DNA methylation assay for colorectal carcinoma. Cancer Biol Med. 2017;14(1):42-49. doi:10.20892/j.issn.2095-3941.2016.0082
5. Esteller M, Herman JG. Cancer as an epigenetic disease: DNA methylation and chromatin alterations in human tumours. J Pathol. 2002;196(1):1–7. DOI:10.1002/path.1024.
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. Curr Top Med Chem. 2016;16(7):745–76. DOI: 10.2174/1568026615666150825141330.
7. Feinberg, A. P., & Tycko, B. (2004). The history of cancer epigenetics. Nature Reviews Cancer, 4(2), 143–153. doi:10.1038/nrc1279Goelz SE, Vogelstein B, Hamilton SR, Feinberg
8. AP. Hypomethylation of DNA from benign and malignant human colon neoplasms. Science. 1985;228(4696):187–90. DOI: 10.1126/science.2579435.Timp W, Bravo HC, McDonald
9. OG, Goggins M, Umbricht C, Zeiger M, Feinberg AP, Irizarry RA. Large hypomethylated blocks as a universal defining epigenetic alteration in human solid tumors. Genome Med. 2014;6(8):61. DOI: 10.1186/s13073-014-0061-y.
10. Gonzalo S. Epigenetic alterations in aging. J Appl Physiol. 2010;109(2):586–597. DOI: 10.1152/japplphysiol.00238.
11. Toyota M, Issa J. The role of DNA hypermethylation in human neoplasia. Electrophoresis. 2000;21(2):329-333. DOI: 10.1002/(SICI)1522-2683(20000101)21:2<329::AID-ELPS329>3.0.CO;2-9.
12. Weisenberger DJ et al., CpG island methylator phenotype underlies sporadic microsatellite instability and is tightly associated with BRAF mutation in colorectal cancer. Nat Genet. 2006;38(7):787–793. DOI: 10.1038/ng1834.
13. Lao VV, Grady WM. Epigenetics and colorectal cancer. Nat Rev Gastroenterol Hepatol. 2011;8:686–700. DOI: 10.1038/nrgastro.2011.173.
14. Costello JF, Fruhwald MC, Smiraglia DJ, Rush LJ, Robertson GP, Gao X, Wright FA, Feramisco JD, Peltomaki P, Lang JC et al. Aberrant CpG-island methylation has non-random and tumour-type-specific patterns. Nat Genet. 2000;24(2):132–138. DOI: 10.1038/72785.
15. Jones PA. Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nat Rev Genet. 2012;13(7):484–492. DOI: 10.1038/nrg3230.
16. Galamb O, Kalmar A, Peterfia B, Csabai I, Bodor A, Ribli D, Krenacs T, Patai ÁV, Wichmann B, Bartak BK, et al. Aberrant DNA methylation of WNT pathway genes in the development and progression of CIMP-negative colorectal cancer. Epigenetics. 2016;11(8):588–602. DOI: 10.1080/15592294.2016.1190894.

medRxiv preprint doi: https://doi.org/10.1101/2020.09.28.20203539; this version posted September 29, 2020. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

17. Lengauer.C, Kinzler K.W, and Vogelstein B. DNA methylation and genetic instability in colorectal cancer cells.PNAS. 1997;94(6):2545-2550. DOI:10.1073/pnas.94.6.2545
18. 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; 2016
19. Hendrik G. Stunnenberg et al. The International Human Epigenome Consortium: A Blueprint for Scientific Collaboration and Discovery. Cell.2016;167(5):1145-1149. DOI:10.1016/j.cell.2016.11.007
20. Zitt M, Zitt M, Müller HM. DNA methylation in colorectal cancer--impact on screening and therapy monitoring modalities?. Dis Markers. 2007;23(1-2):51-71. DOI:10.1155/2007/891967
21. Draht, M.X.G., Goudkade, D., Koch, A. et al. Prognostic DNA methylation markers for sporadic colorectal cancer: a systematic review. Clin Epigenet 10, 35 (2018). DOI:10.1186/s13148-018-0461-8
22. Chen, X., Gole, J., Gore, A. et al. Non-invasive early detection of cancer four years before conventional diagnosis using a blood test. Nat Commun 11, 3475 (2020). https://doi.org/10.1038/s41467-020-17316-z
23. 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; 2016
24. Du.P, Zhang.X, Huang.C.C, Jafari.N, Kibbe.A.W, Hou.L, Lin.M.S. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. BMC Bioinformatics. 2010;11(1):587. DOI: 10.1186/1471-2105-11-587.
25. R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.
26. Ritchie, M. E., et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research. 2015;43(7):e47. DOI: 10.1093/nar/gkv007.
27. Sarathi, A. and A. Palaniappan. Novel significant stage-specific differentially expressed genes in hepatocellular carcinoma. BMC Cancer. 2019;19(1):663. DOI: 10.1186/s12885-019-5838-3.
28. Barfield T,R. CpGassoc: an R function for analysis of DNA methylation microarray data. BMC Bioinformatics. 2012;28(9):1280–1281 DOI:10.1093/bioinformatics/bts124.
29. Morris, T. J., Butcher, L. M..*Champ*: 450k chip analysis methylation pipeline. BMC Bioinformatics. 2014;30(3):428-430. DOI: 10.1093/bioinformatics/btt684.
30. Peters TJ, Buckley MJ, Statha*m AL,* Pidsley R, Samaras K, Lord RV, Clark SJ, Molloy PL (2015). "De novo identification of differentially methylated regions in the human genome." Epigenetics & Chromatin, 8, 6. http://www.epigeneticsandchromatin.com/content/8/1/6.
31. Subramanian, A., et al. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. PNAS 102 (43): 15545-15550; DOI: 10.1073/pnas.0506580102
32. Wang Y, Franks M.J, Whitfield L.M, and Cheng C. BioMethyl: An R package for Biological Interpretation of DNA Methylation Data. BMC Bioinformatics. 2019;35(19):3635-3641 DOI:10.1093/bioinformatics/btz137.

medRxiv preprint doi: https://doi.org/10.1101/2020.09.28.20203539; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

33. Cedoz.L.P, Prunello.M, Brennan.K, Gevaert.O. MethylMix 2.0: an R package for identifying DNA methylation genes. BMC Bioinformatics. 2018;34(17):3044-3046. DOI:10.1093/bioinformatics/bty156.
34. Therneau T (2015). *A Package for Survival Analysis in S*. version 2.38, https://CRAN.R-project.org/package=survival.
35. Kassambara.A, Kosinski.M and Biecek.M (2019).survminer: Drawing Survival Curves using 'ggplot2'. R package version 0.4.6. https://CRAN.R-project.org/package=survminer.
36. Teng H, Gao R, Qin N, 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. 2018;18(6):5361-5370. doi:10.3892/mmr.2018.9563
37. De Carvalho DD, Sharma S, You JS, Su SF, Taberlay PC, et al. (2012) DNA methylation screening identifies driver epigenetic events of cancer cell survival. Cancer Cell 21:655–667 DOI:10.1016/j.ccr.2012.03.045
38. Lin SH, et al.. The somatic mutation landscape of premalignant colorectal adenoma. Gut. 2017gutjnl-2016-313573. 10.1136/gutjnl-2016-313573
39. Atwell LL, Beaver LM, Shannon J, Williams DE, Dashwood RH, Ho E. Epigenetic Regulation by Sulforaphane: Opportunities for Breast and Prostate Cancer Chemoprevention. Curr Pharmacol Rep. 2015;1(2):102-111. doi:10.1007/s40495-014-0002-x
40. Ashburner et al. Gene ontology: tool for the unification of biology. Nat Genet. May 2000;25(1):25-9.DOI: 10.1038/75556
41. Matzke, M., Mosher, R. RNA-directed DNA methylation: an epigenetic pathway of increasing complexity. Nat Rev Genet 15, 394–408 (2014). DOI:10.1038/nrg3683
42. Kulis M, Esteller M. DNA methylation and cancer. Adv Genet. 2010;70:27-56. DOI:10.1016/B978-0-12-380866-0.60002-2
43. Ashktorab H, Brim H. DNA Methylation and Colorectal Cancer. Curr Colorectal Cancer Rep. 2014;10(4):425-430. DOI:10.1007/s11888-014-0245-2
44. Sondka Z, Bamford S, Cole CG, Ward SA, Dunham I, Forbes SA. The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. Nat Rev Cancer. 2018;18(11):696-705. DOI: 10.1038/s41568-018-0060-1.
45. Abbott KL, Nyre ET, Abrahante J, Ho YY, Isaksson Vogel R, Starr TK. The Candidate Cancer Gene Database: a database of cancer driver genes from forward genetic screens in mice. Nucleic Acids Res. 2015;43(Database issue):D844-8. DOI: 10.1093/nar/gku770.
46. Sakai, Lynn Y, Keene, Douglas R, Renard, Maryolyn, De Blacker, juli, FBN1:The disease causing gene for marfan syndrome and other disorders, Gene.2016;591(1):279-291. DOI: 10.1016/j.gene.2016.07.033.
47. Guo Q, Song Y, Zhang H, Wu X, Xia P, Dang C. Detection of hypermethylated Fibrillin-1 in the stool sample of colorectal cancer patients. Med. Oncol, 2013;30:695. DOI:10.1007/s12032-013.0695-4.
48. Li W H, Zhang H, Guo Q et al., Detection of SNCA and FBN1 methylation in stool as biomarker for colorectal cancer. Dis Markers.2015;2015:657570. DOI: 10.1155/2015/657570.
49. Wu.H, Qian.C, Liu.C, Xiang.J, Ye.D, Zhang.Z, and Zhang.X. Role and mechanism of FOXG1 in invasion and metastasis of colorectal cancer. Chinese Journal of Biotechnology. 2018;34(5):752-760. DOI: 10.13345/j.cjb.170389.
50. Zheng, X., Lin, J., Wu, H. *et al.* Forkhead box (FOX) G1 promotes hepatocellular carcinoma epithelial-Mesenchymal transition by activating Wnt signal through

medRxiv preprint doi: https://doi.org/10.1101/2020.09.28.20203539; this version posted September 29, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

forming T-cell factor-4/Beta-catenin/FOXG1 complex. J Exp Clin Cancer Res 2019; 38(1):475 DOI:10.1186/s13046-019-1433-3.

51. Chan, D., Liu, V., To, R. *et al.* Overexpression of FOXG1 contributes to TGF-β resistance through inhibition of p21WAF1/CIP1 expression in ovarian cancer. Br J Cancer. 2009;101(8):1433–1443. DOI:10.1038/sj.bjc.6605316.
52. 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. 2017;6(4):698-708 DOI: 10.21037/tcr.2017.07.22.
53. Ashktorab.H, Rahi.H, Wansley.D et al,. Toward a comprehensive and systematic methylome signature in colorectal cancers. Epigenetics. 2013:8(8):807-815, DOI: 10.4161/epi.25497.
54. Mori.Y, Cai.K, Cheng.Y 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. DOI:10.1053/j.gastro.2006.06.006.
55. Ma.Z, Williams.M, Cheng.Y.Y, and Leung.K.W. Roles of Methylated DNA Biomarkers in Patients with Colorectal Cancer. Disease Markers. 2019;7:1-8 DOI:10.1155/2019/2673543.
56. Fang SQ, Gao M, Xiong SL, Chen HY, Hu SS, Cai HB. Combining differential expression and differential coexpression analysis identifies optimal gene and gene set in cervical cancer. J Cancer Res Ther. 2018;14(1):201-207. DOI:10.4103/0973-1482.199787
57. Xi T, Zhang G. Epigenetic regulation on the gene expression signature in esophagus adenocarcinoma. Pathol Res Pract. 2017;213(2):83-88. DOI:10.1016/j.prp.2016.12.007
58. Siesser PF, Motolese M, Walker MP, et al. FAM123A binds to microtubules and inhibits the guanine nucleotide exchange factor ARHGEF2 to decrease actomyosin contractility. Sci Signal. 2012;5(240):ra64. DOI:10.1126/scisignal.2002871
59. Lieu, C. H., et al. Comprehensive Genomic Landscapes in Early and Later Onset Colorectal Cancer. 2019: 25(19): 5852-5858. DOI: 10.1158/1078-0432.CCR-19-0899
60. Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. Nature. 2012;487(7407):330-337. Published 2012 Jul 18. DOI:10.1038/nature11252
61. Sanz-Pamplona R et al,. Exome Sequencing Reveals AMER1 as a Frequently Mutated Gene in Colorectal Cancer. Clin Cancer Res. 2015;21(20):4709-4718. DOI:10.1158/1078-0432.CCR-15-0159
62. Moore LD, Le T, Fan G. DNA methylation and its basicfunction.*Neuropsychopharmacology*.2013;38(1):23-38. doi:10.1038/npp.2012.112
63. Eden S, Cedar H. Role of DNA methylation in the regulation of transcription. *Curr Opin Genet Dev*. 1994;4(2):255-259. DOI:10.1016/s0959-437x(05)80052-8
64. Toyota M, Ahuja N, Ohe-Toyota M, Herman JG, Baylin SB, Issa JP. CpG island methylator phenotype in colorectal cancer. Proc Natl Acad Sci U S A. 1999;96(15):8681-8686. DOI:10.1073/pnas.96.15.8681
65. Jia, M., Gao, X., Zhang, Y. et al. Different definitions of CpG island methylator phenotype and outcomes of colorectal cancer: a systematic review. Clin Epigenet. 2016; 8:25. DOI:10.1186/s13148-016-0191-8
66. Hughes, L. A. E., et al. The CpG Island Methylator Phenotype: What's in a Name? Cancer Research. 2013;73(19): 5858-5868.DOI: 10.1158/0008-5472.CAN-12-4306
67. Berg M, Hagland HR, Søreide K (2014) Comparison of CpG Island Methylator Phenotype (CIMP) Frequency in Colon Cancer Using Different Probe- and

medRxiv preprint doi: https://doi.org/10.1101/2020.09.28.20203539; this version posted September 29, 2020. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Gene-Specific Scoring Alternatives on Recommended Multi-Gene Panels. PLoS ONE 9(1): e86657. DOI:10.1371/journal.pone.0086657

68. Advani, S.M.; Advani, P.; DeSantis, S.M.; Brown, D.; VonVille, H.M.; Lam, M.; Loree, J.M.; Mehrvarz Sarshekeh, A.; Bressler, J.; Lopez, D.S.; et al. Clinical, Pathological, and Molecular Characteristics of CpG Island Methylator Phenotype in Colorectal Cancer: A Systematic Review and Meta-Analysis. Transl. Oncol. 2018, 11, 1188–1201. DOI:10.1016/j.tranon.2018.07.008
69. Issa, J. CpG island methylator phenotype in cancer. Nat Rev Cancer 4, 988–993 (2004). DOI:10.1038/nrc1507
70. Kerachian, M.A., Javadmanesh, A., Azghandi, M. et al. Crosstalk between DNA methylation and gene expression in colorectal cancer, a potential plasma biomarker for tracing this tumor. Sci Rep 10, 2813 (2020).DOI:10.1038/s41598-020-59690-0
71. Gundert M, Edelmann D, Benner A, et al. Genome-wide DNA methylation analysis reveals a prognostic classifier for non-metastatic colorectal cancer (ProMCol classifier).Gut 2019;68:101-110. DOI:10.1136/gutjnl-2017-314711
72. Freitas, M., Ferreira, F., Carvalho, S. et al. A novel DNA methylation panel accurately detects colorectal cancer independently of molecular pathway. J Transl Med 16, 45 (2018). DOI:10.1186/s12967-018-1415-9
73. Vrba L, Futscher BW. A suite of DNA methylation markers that can detect most common human cancers.Epigenetics. 2018;13(1):61–72. DOI:10.1080/15592294.2017.1412907
74. Ma X, Wang YW, Zhang MQ, Gazdar AF. DNA methylation data analysis and its application to cancer research. *Epigenomics*. 2013;5(3):301-316. DOI:10.2217/epi.13.26
75. Jones, P. A. The DNA methylation paradox. Trends Genet. 15, 34–37 (1999).DOI:10.1016/s0168-9525(98)01636-9
76. Ehrlich M. DNA hypermethylation in disease: mechanisms and clinical relevance.Epigenetics.2019;14(12):1141-1163.DOI:10.1080/15592294.2019.1638701
77. Matzke, M., Mosher, R. RNA-directed DNA methylation: an epigenetic pathway of increasing complexity. Nat Rev Genet 15, 394–408 (2014). DOI:10.1038/nrg3683
78. Feng S, Cokus SJ, Zhang X, et al. Conservation and divergence of methylation patterning in plants and animals. Proc Natl Acad Sci U S A. 2010;107(19):8689-8694. DOI:10.1073/pnas.1002720107