

Table S1. Reported sensitivity and specificity for stage I and stage IV colorectal cancer detection in published studies. N is the number of samples assessed. ND is not disclosed. Grail refers to the Circulating Cell-free Genome Atlas study (NCT02889978; Klein et al, 2021), Freenome to the PREEMPT study (NCT04369053), and Guardant to the ECLIPSE study (Chung et al, 2024).

Reference	Stage I			Stage IV		
	Specificity	Sensitivity	N	Specificity	Sensitivity	N
Grail	99.5%	43.3%	30	99.5%	95.3%	64
Freenome	91.5%	57.1%	ND	ND	100%	ND
Guardant	89.9%	55%	22	89.9%	100%	10

Table S2. Mean AUCs ($\pm$ standard deviation) for stage I CRC versus healthy control models built using all batches compared to models built using all batches except SY01-D894 (three samples).

Model type	All batches	Excluding SY01-D894
5hmC-only	0.55 $\pm$ 0.04	0.51 $\pm$ 0.04
5mC-only	0.69 $\pm$ 0.02	0.80 $\pm$ 0.02
modC	0.66 $\pm$ 0.02	0.81 $\pm$ 0.02
5mC & 5hmC	0.95 $\pm$ 0.01	0.95 $\pm$ 0.01

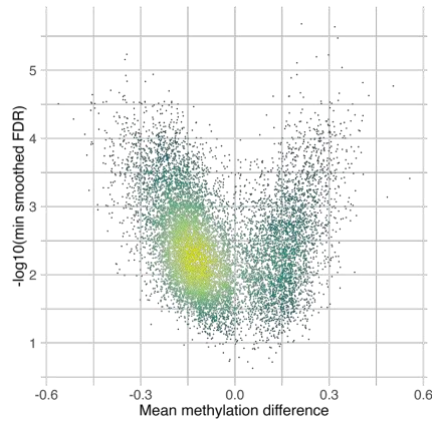


Figure S1. Differentially methylated regions between stage IV colorectal cancer tissue and matched normal tissue. Volcano plot for DMRs between stage IV CRC tissue and adjacent matched normal, using data from TCGA. Color indicates density of points, with yellow used to indicate a higher density.

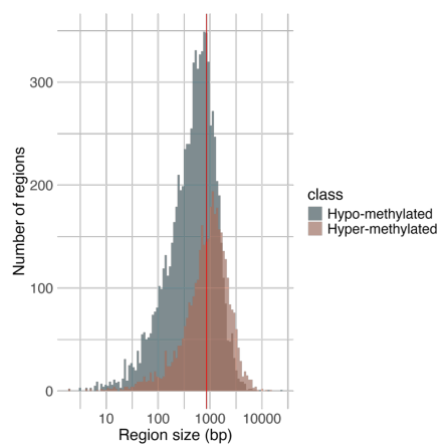


Figure S2. Size distribution of hypo- and hyper-methylated regions in colorectal cancer tissue. Histogram showing the size of hypo- and hyper-methylated DMRs from TCGA tissue data.

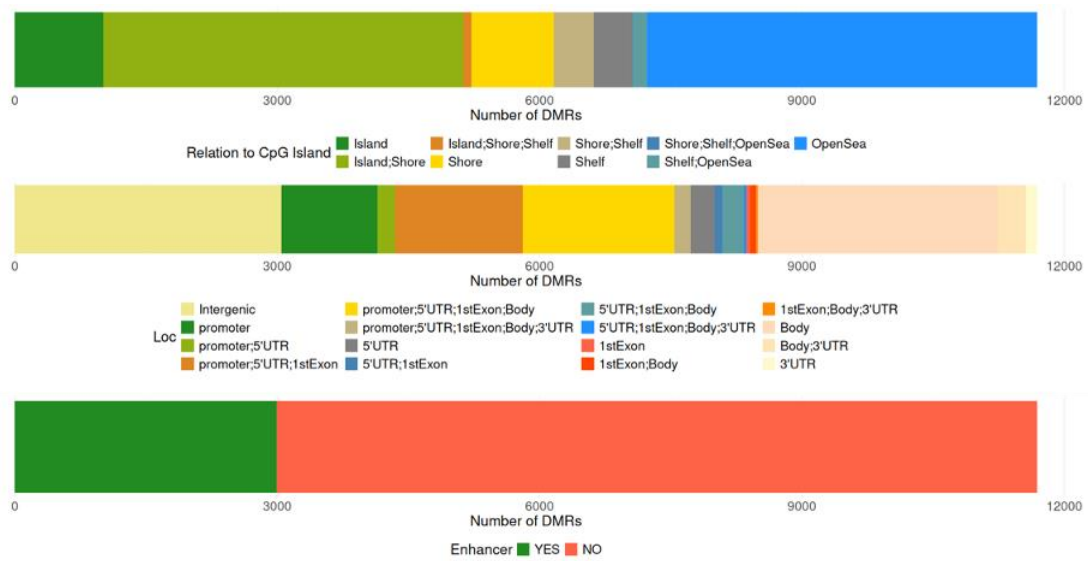


Figure S3. Genomic context of differentially methylated regions in colorectal cancer tissue. Fraction of the 11,691 modC DMRs associated with genomic elements based on their location, regulatory function, and CpG density.

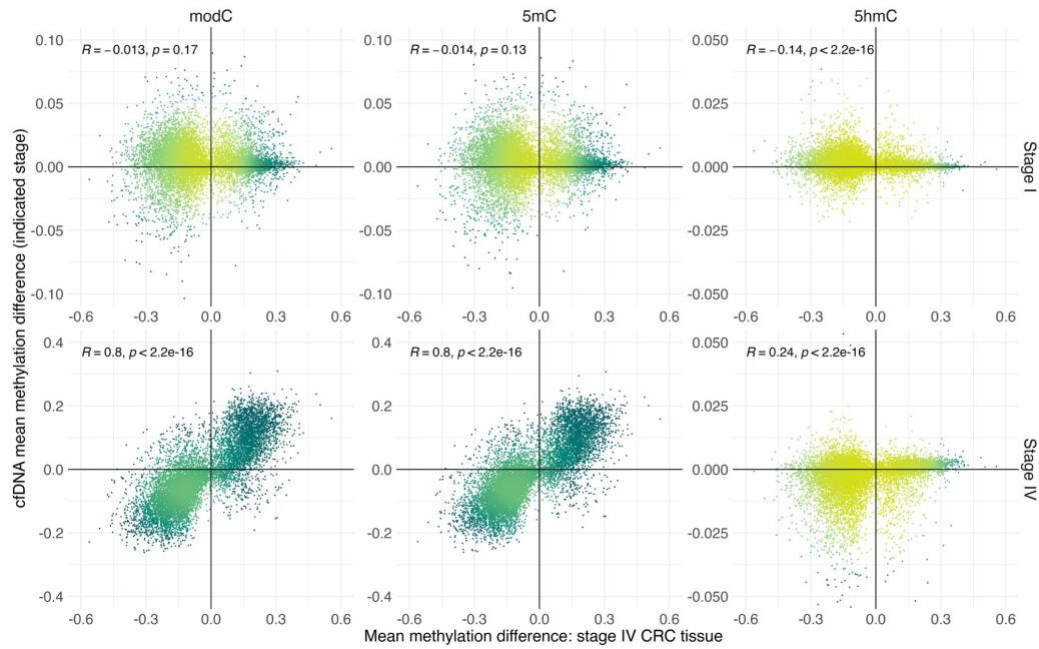


Figure S4. Concordance of methylation differences between tissue and plasma samples. Scatterplots showing mean methylation differences in stage IV tissue (versus adjacent normal tissue) and mean methylation differences in stage I (upper panel) and IV (lower panel) cfDNA versus healthy control cfDNA. Each plot is annotated with the coefficient of determination R^2 .

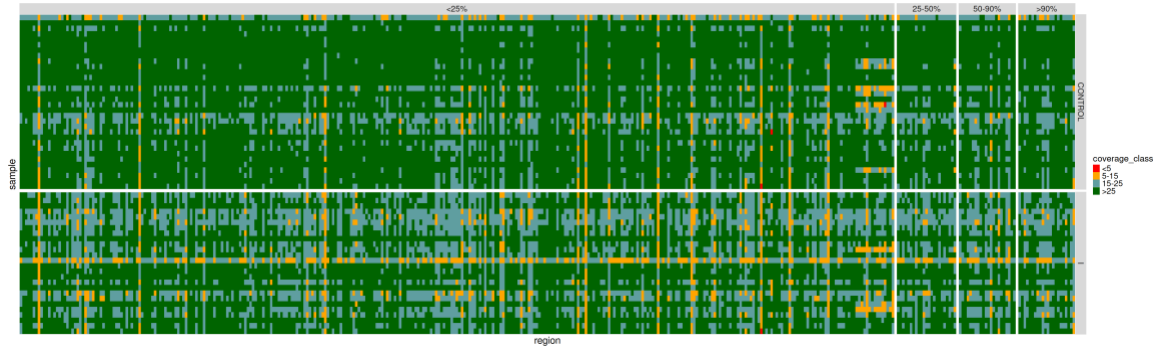


Figure S5. Coverage of differentially methylated regions across study samples. Heatmap showing mean coverage (unique reads per strand) for each DMR (x-axis) across all samples (y-axis). Samples are grouped by healthy control and stage I CRC, and DMRs are ordered by their frequency of inclusion in LOOCV splits. Colour scale represents mean coverage. Coverage was calculated as the mean number of unique reads per strand across all CpGs within each region. Each DMR typically spans multiple CpGs, so the mean modified cytosine fraction for a region reflects numerous measurements. For example, a region with 10 CpGs uniformly covered at 5x per strand yields 100 observations (5x coverage x 10 CpGs x 2 strands).

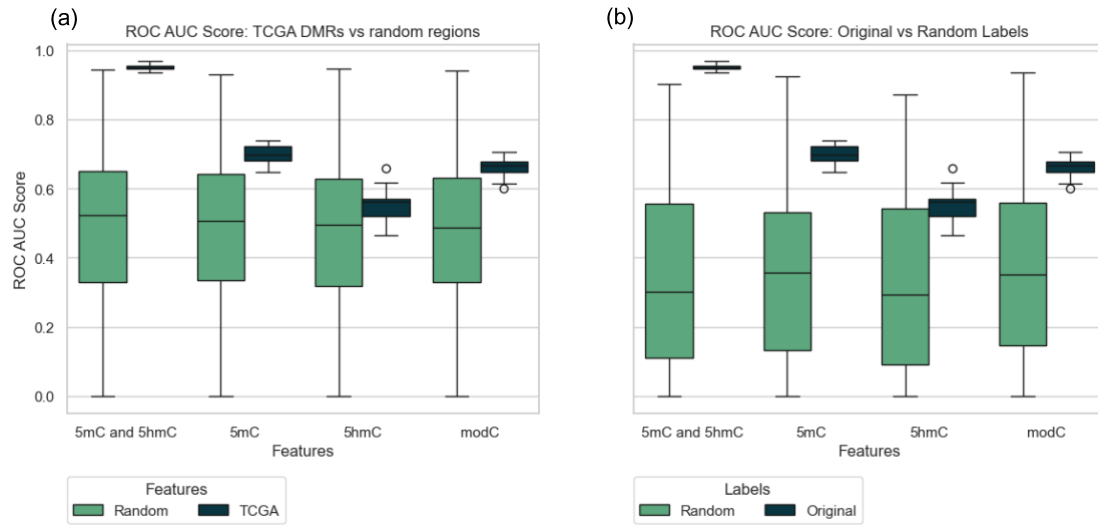


Figure S6. Impact of randomization on model performance. Box plots showing the performance of models built using DMRs identified from TCGA data after random permutation of (a) genomic regions and (b) labels. Box plots display the median and interquartile range (IQR), with whiskers extending to 1.5 x IQR.

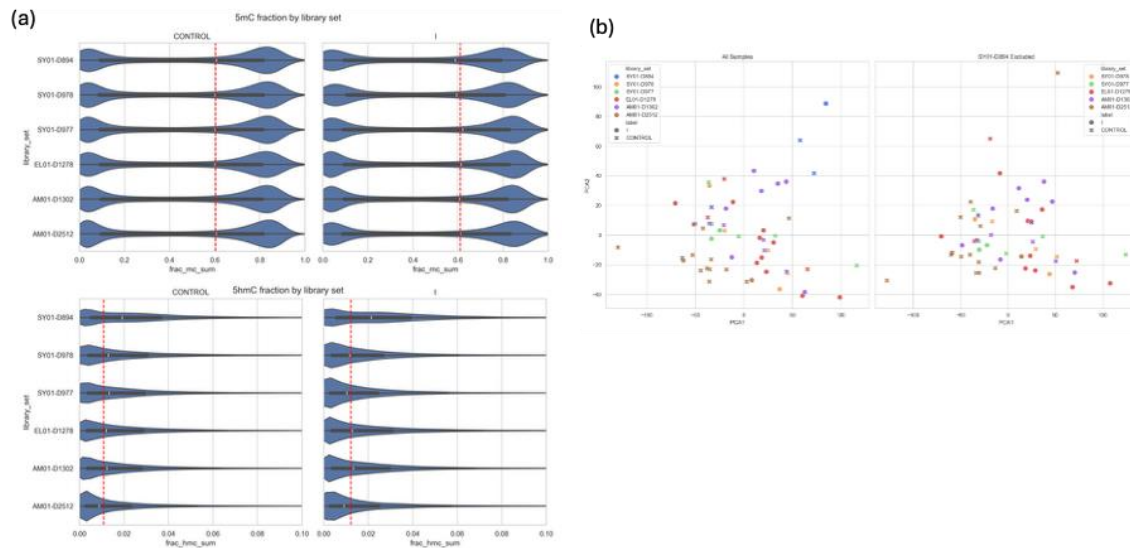


Figure S7. Assessment of batch effects. (a) Violin plots showing the fraction of 5mC or 5hmC in each batch of samples for controls (left facet) and stage I CRC (right facet). The red dotted lines indicate the mean methylation across all batches. (b) PCA plots of 5hmC fraction in all samples (left facet) and all samples excluding the SY01-D894 batch (right facet). Each sample is coloured by its batch. Symbol shape indicates whether the sample was from a healthy control or stage I CRC patient.

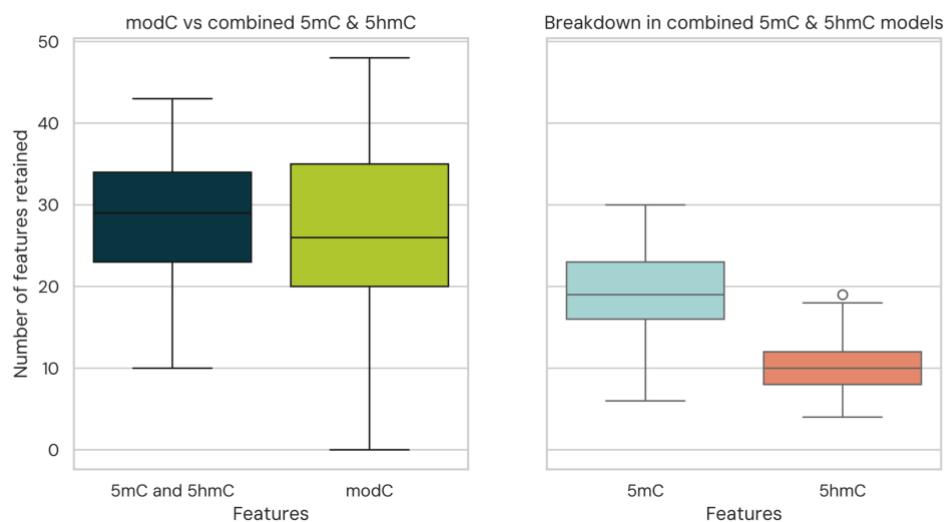


Figure S8. Feature selection patterns in methylation-based models. (a) Numbers of features selected by the LASSO classifier when using modC or the combination of 5mC and 5hmC as features. **(b)** Number of 5mC and 5hmC features selected in models trained on the combined 5mC and 5hmC feature set. Each box plot represents the distribution across 25 random seeds and 58 cross-validation splits. Box plots display the median and interquartile range (IQR), with whiskers extending to 1.5 x IQR.