MethCORR modelling of methylomes from formalin-fixed paraffin-embedded tissue enables characterization and prognostication of colorectal cancer

Trine B. Mattesen, Mads H. Rasmussen, Juan Sandoval, Halit Ongen, Sigrid S. Árnadóttir, Josephine Gladov, Anna Martinez-Cardus, Manuel Castro de Moura, Anders H. Madsen, Søren Laurberg, Emmanouil T. Dermitzakis, Manel Esteller, Claus L. Andersen & Jesper B. Bramsen

Transcriptional characterization and classification has potential to resolve the inter-tumor heterogeneity of colorectal cancer and improve patient management. Yet, robust transcriptional profiling is difficult using formalin-fixed, paraffin-embedded (FFPE) samples, which complicates testing in clinical and archival material. We present MethCORR, an approach that allows uniform molecular characterization and classification of fresh-frozen and FFPE samples. MethCORR identifies genome-wide correlations between RNA expression and DNA methylation in fresh-frozen samples. This information is used to infer gene expression information in FFPE samples from their methylation profiles. MethCORR is here applied to methylation profiles from 877 fresh-frozen/FFPE samples and comparative analysis identifies the same two subtypes in four independent cohorts. Furthermore, subtype-specific prognostic biomarkers that better predicts relapse-free survival (HR = 2.66, 95%CI [1.67–4.22], P value < 0.001 (log-rank test)) than UICC tumor, node, metastasis (TNM) staging and microsatellite instability status are identified and validated using DNA methylation-specific PCR. The MethCORR approach is general, and may be similarly successful for other cancer types.

1Department of Molecular Medicine, Aarhus University Hospital, 8200 Aarhus, Denmark. 2Epigenomic Unit, Health Research Institute La Fe (ISSLaFe), Valencia, Spain. 3Biomarker and precision medicine Unit, Health Research Institute La Fe (ISSLaFe), Valencia, Spain. 4Genetic Medicine and Development, University of Geneva Medical School-CMU, 1 Rue Michel-Servet, 1211 Geneva, Switzerland. 5Badalona Applied Research Group in Oncology (B-ARGO), Germans Trias i Pujol Research Institute (IGTP), Badalona, Barcelona, Catalonia, Spain. 6Medical Oncology Service, Institute Catalan of Oncology (ICO), Badalona, Barcelona, Catalonia, Spain. 7Josep Carreras Leukaemia Research Institute (IJC), Badalona, Barcelona, Catalonia, Spain. 8Department of Surgery, Hospitaalsheden Vest, 7400 Herning, Denmark. 9Colorectal Surgical Unit, Department of Surgery, Aarhus University Hospital, 8200 Aarhus, Denmark. 10Josep Carreras Leukaemia Research Institute (IJC), Badalona, Barcelona, Catalonia, Spain. 11Centro de Investigacion Biomedica en Red Cancer (CIBERONC), Madrid, Spain. 12Institucio Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Catalonia, Spain. 13Physiological Sciences Department, School of Medicine and Health Sciences, University of Barcelona (UB), Barcelona, Catalonia, Spain. These authors contributed equally: Claus L. Andersen, Jesper B. Bramsen email: Cla@clin.au.dk; Bramsen@clin.au.dk
Colorectal cancer (CRC) is a disease with extensive inter-patient heterogeneity, both molecularly and histopathologically, which cannot be resolved by current clinical methods. Despite a continuous refinement of the UICC tumor, node, metastasis (TNM) staging system to measure disease extent and define prognosis, the disease outcome still varies considerably even for patients with the same tumor stage. Therefore, new factors that can more precisely stratify patients into different risk categories are clearly warranted.

Recent attempts to resolve CRC heterogeneity and improve prognostication include molecular subclassification and characterization based on transcriptional profiling2–4. Consensus molecular subtype (CMS) classification stratifies CRC into four subtypes CMS 1–4 with distinct biology and histopathological features5. Still, the CMS taxonomy itself has limited prognostic power for therapeutic decision-making5. To address this, we previously combined transcriptional subtyping with subtype-specific prognostic biomarkers to improve prognostication beyond TNM staging in retrospective cohorts6. This indicated a clinical potential of using molecular classification and subtype-specific biomarkers as a complement to TNM staging for prognostication. Furthermore, it highlighted the importance of archived tumor material for biomarker discovery and pre-clinical validation.

The strategies for transcriptional classification and subtype-specific prognostication were developed by, and still primarily rely on, profiling high-quality RNA purified from fresh-frozen (FF) tissue. However, high-quality RNA is often not recovered from the formalin-fixed, paraffin-embedded (FFPE) tissue that is routinely archived in the clinic. This can preclude confident transcriptional profiling and hereby complicate clinical testing of molecular classification and exploratory analysis in well-annotated, archival FFPE material5–9. The clinical popularity of FFPE tissue is unlikely to change as it forms the basis for histopathological diagnoses and is a convenient, cost-effective preservation method. For wide utility, strategies for molecular classification, characterization, and prognostication should therefore be compatible with FFPE tissue.

Strategies based on DNA rather than RNA profiling may be a way forward. DNA is considered less sensitive to degradation than RNA in FFPE samples10,11 and enzymatic strategies for DNA repair have greatly improved the analysis of FFPE DNA12–15. A strategy for robust analysis of clinical and archival FFPE samples could involve DNA methylation as highly concordant DNA methylation profiles are produced from matched FF and FFPE tissues when using the Illumina Infinium Human Methylation Beadchip technology14,16,17. In addition, many biological traits, such as RNA expression and cell-type identity, are associated with specific and robust DNA methylation patterns in the genome18,19. This suggests that both gene expression and cell-type information may be extracted from DNA methylation profiles of FFPE samples and used for molecular classification and prognostication, as an alternative to RNA profiling. Furthermore, conversion of methylation profiles into a gene-centric expression format would allow molecular analysis of FF and FFPE samples using the plethora of bioinformatics tools, databases, and signatures established for RNA expression analysis.

Motivated by this, we here develop MethCORR, an approach, which identifies genome-wide correlations between gene expression and DNA methylation and use this to obtain gene expression and cell-type information in independent samples from their DNA methylation profiles. In homogenous cell preparations, associations between gene expression and DNA methylation have been observed only for a small fraction of genes when analyzing local promoters, gene bodies, or nearby enhancers20–22. We hypothesize that genome-wide correlation analysis will identify far more associations and that these will include both functional gene-regulatory interactions and indirect associations e.g. between cell-type-specific RNA expression and cell-type-specific DNA methylation. We here show that MethCORR, independent of whether the methylationomes were produced from FFPE or FF tissues, allows expression information to be inferred for a large number of genes (~11,000). Consequently, MethCORR enables a plethora of molecular analyses to be performed on otherwise difficult-to-analyze FFPE tissues e.g. tumor characterization, tumor classification, and interpretation of expression signatures to derive DNA methylation-based biomarkers. Hereby MethCORR also provides a path for improved, subtype-specific prognostication of CRC using clinical FFPE samples.

Results

MethCORR infers RNA expression from DNA methylation. Here we developed the MethCORR approach that, by mapping genome-wide correlations between RNA expression and DNA methylation in FF samples, can infer gene expression information in unrelated samples from their DNA methylation profiles. Correlations were identified genome-wide using matching RNA expression and 450K methylation data (methylation β-values) from 394 FF CRC samples of The Cancer Genome Atlas (TCGA) Project, denoted the COREAD cohort (Fig. 1a and Supplementary Fig. 1a; Supplementary Table 1 and Supplementary Data 1). The cohort was divided into two discovery sets (each \( n = 158 \)) in which genome-wide correlation analysis was performed independently and one validation set (\( n = 78 \); Fig. 1a). Our analysis identified positively and negatively expression-correlated CpGs (Spearman’s correlation \( p < 0.01 \)) overlapping in the two discovery sets for 17,776 of 20,530 genes (Fig. 1a). The majority of the genes without expression-correlated CpGs were non-expressed (Supplementary Fig. 1b). To derive gene expression information for these 17,776 genes, we selected up to 200 CpGs whose methylation level were most negatively (≤100 sites) and positively (≤100 sites) correlated with its expression (Fig. 1a). The methylation levels of these expression-correlated CpGs were used to calculate a MethCORR score (MCS) for each gene (formula in Fig. 1b) and simple linear and polynomial regression modeling was used to identify genes with good correlations between MCSs and measured RNA expression (Fig. 1a). Models were established in the discovery sets by ten times tenfold cross validation and selected using root mean square error (RMSE) as a measure of model fit. We found good inter-sample correlations for 16,248 genes in the discovery sets (\( R^2 > 0.16 \)) and confirmed these for 11,222 genes in the validation set (gene model performances in Supplementary Data 2; Supplementary Fig. 1c–e). The 11,222 genes were denoted MethCORR genes and the expression-correlated CpGs of these define the COREAD MethCORR matrix (≤200 CpGs × 11,222 genes; Supplementary Data 3) that was used for calculation of MCSs from DNA methylation profiles of all samples analyzed in this study (Fig. 1c). We also investigated if RNA expression was better modeled using the ≤200 expression-associated CpGs for each gene directly, instead of using MCSs, but found no improvement in overall performance (\( R^2 \) and RMSE; Supplementary Fig. 1f). Similarly, adding age and gender information to MCS-based models did not improve overall performances (Supplementary Fig. 1g). This likely reflect that CRC-induced methylation changes are much greater than the subtler effects of age and gender in normal tissues23. Still, MethCORR captures gender-specific expression by including CpGs located on chromosome X and Y in the MethCORR matrix. Accordingly, known gender-specific RNAs exhibited gender-specific inferred RNA expression (Supplementary Fig. 1h).
Next, we investigated characteristics of the MethCORR genes included in the MethCORR matrix. MethCORR genes exhibited greater variation in RNA expression (Supplementary Fig. 2a), were more frequently dysregulated in cancer vs. normal mucosa (Supplementary Fig. 2b) and encompassed relatively fewer household genes (Supplementary Fig. 2c) than the set of genes not included in the MethCORR matrix. Importantly, the MethCORR genes exhibited the same stroma score distribution as the full set of genes (Supplementary Fig. 2d). This indicates that MethCORR maintains the ability to characterize both traits of the cancer cells and the surrounding stroma. The established MCS regression models were next used to calculate inferred RNA (iRNA) expression for MethCORR genes in the validation samples of the COREAD cohort (set 3) and in an independent Danish CRC cohort, denoted SYSCOL\textsuperscript{3}. We found a high intra-sample correlation between measured RNA and iRNA expression in the COREAD validation samples (median $R^2 = 0.93$ (range = 0.82–0.96); Supplementary Data 4) and SYSCOL samples (median $R^2 = 0.76$ (range = 0.62–0.82); Fig. 1d–e; Supplementary Data 5). To evaluate the robustness of MethCORR to differences between cohorts, we repeated the entire MethCORR discovery and validation process using the SYSCOL cohort to construct a SYSCOL MethCORR matrix, derive MCSs, and to infer iRNA expression (Fig. 1a; Supplementary Data 6–7). Again, we found high intra-sample correlations between observed RNA and iRNA expression (SYSCOL set 3, median $R^2 = 0.92$ (range = 0.87–0.95); COREAD median $R^2 = 0.74$ (range = 0.55–0.82); Fig. 1e; Supplementary Data 4 and 5). We speculated that the moderate decrease in $R^2$ between cohorts was caused by differences in RNA quantification methods rather than the MethCORR approach. In support, comparative analysis of COREAD validation samples using normalized RNA expression data from the UCSC XENA database\textsuperscript{24} and the National Cancer Institute (NCI) genomic database commons (GDC)\textsuperscript{25} confirmed that MethCORR iRNA-RNA correlations were not
lower than if applying two different RNA normalization strategies to the same samples (Supplementary Fig. 2e).

In accordance with the high intra-sample correlations between measured RNA and iRNA expression, we found a good overlap in CMS (84% agreement) and CRC intrinsic subtype (CRIS; 75% agreement) predictions when using the measured RNA or iRNA expression as input (Fig. 1f and Supplementary Fig. 2f).

In situations where high-quality RNA is not obtainable, iRNA expression may provide better estimates of gene expression than RNA sequencing, as even moderate declines in RNA quality can lead to unreliable expression profiles. Indeed, samples with the lowest correlation between measured RNA and iRNA expression had significantly lower RNA quality than high correlation samples (P < 0.0001, Wilcoxon rank sum test; Supplementary Fig. 2g). In contrast, no equivalent drop in 450K methylation data quality was observed (Supplementary Fig. 2g). Compromised RNA quality is inherent to FFPE tissue. In agreement, analysis of nine COREAD samples with available RNA sequencing and 450K methylation profiles of FFPE samples and RNA expression or MCSs in matched fresh-frozen tissue.

Collectively, this showed that MethCORR expression measures (MCSSs and iRNAs) can be inferred from DNA methylation for a large number of genes, even when methylation data are based on FFPE tissue.

MethCORR identifies two subtypes in FF and FFPE cohorts. We next investigated if inferred expression profiles allow uniform subtype discovery and characterization of both FF and FFPE cohorts using bioinformatics strategies normally reserved for FF samples with high-quality RNA expression profiles. As input, we employed MCS profiles as they strengthen the focus on cancer cell-related traits during subtype discovery as compared with RNA and iRNA profiles (Supplementary Fig. 3a, b). Subtype discovery by non-negative matrix factorization (NMF)-based consensus clustering was performed in TNM stage II–III COREAD and SYSCOL samples with available 450K methylation data and in two independent FFPE TMA stage II–III cohorts, denoted FFPE1 and FFPE2 (Supplementary Table 1 and Supplementary Data 12). Our focus was on stage II–III patients, which are most relevant for prognostic biomarker identification due to their heterogeneous prognosis. Two MethCORR subtypes, CRC1 and CRC2, were identified in all four cohorts (Supplementary Fig. 3c) and Submap analysis confirmed the correspondence between the CRC1 and CRC2 subtypes in the different cohorts (Supplementary Fig. 3d; FDR < 0.05). In agreement, samples clustered according to subtype in a PCA of all four CRC cohorts together, irrespectively of their preservation-status type (Supplementary Fig. 3e). We next performed comparative subtype characterization in all cohorts, which indicated that CRC1 and CRC2 differed in terms of DNA methylation, chromosomal instability, and stromal/immune cell activity (Fig. 2a and Supplementary Fig. 3f). These are well-known characteristics for the serrated/microsatellite instability status (MSI) and conventional CRC pathways, respectively, pointing to a biological relevance of the MethCORR subtypes.

Further subtype characterization was performed using pre-ranked gene set enrichment analysis (GSEA). Initially, we investigated if similar gene set enrichments were identified when using MCSs vs. RNA expression as input (Fig. 2b) or when MCSs were derived from FF vs. FFPE samples (Fig. 2c). Indeed, a high concordance was observed between normalized enrichment scores for most gene sets in both situations, supporting that expression-correlated MCSs can substitute RNA expression and enable analysis of FFPE tissue. MCS-based GSEA of each cohort uniformly showed that the CRC1 subtype was enriched in gene sets associated with immune- and stromal processes/cell types such as inflammation, epithelial-mesenchymal transition (EMT), cancer-associated fibroblasts (CAFs), and T/B cells (Fig. 2d and Supplementary Table 3). Furthermore, CRC1 was enriched in gene sets associated with positive MSI-, CIMP-, and serrated CRC-status, whereas CRC2 tumors were enriched in gene sets associated with conventional CRC and a more undifferentiated cell status (Fig. 2d and Supplementary Table 3). Similar results were obtained for the two FF cohorts when using RNA expression as input, rather than MCSs (Fig. 2d). Despite biological differences, no difference in relapse-free survival (RFS) was observed between CRC1 and CRC2 (Fig. 2e).

Collectively, these results demonstrate that MethCORR allows uniform discovery and characterization of biologically relevant CRC subtypes in FF and FFPE samples using well-established bioinformatics tools.

A MethCORR map characterizes CRC subtypes. By analysis of expression-correlated CpGs in the MethCORR matrix, we found

---

**Table 1** $R^2$ and RMSE for intra-sample correlations between MethCORR inferred RNA expression (iRNA), RNA expression, or MCS in FFPE samples and RNA expression or MCS in matched fresh-frozen tissue.

| TCGA COREAD Patient Id | $R^2$ iRNA (FFPE) vs. RNA (FF) | $R^2$ RNA (FFPE) vs. RNA (FF) | $R^2$ MCS (FFPE) vs. MCS (FF) | RMSE iRNA (FFPE) vs. RNA (FF) | RMSE RNA (FFPE) vs. RNA (FF) | RMSE MCS (FFPE) vs. MCS (FF) |
|------------------------|-------------------------------|-------------------------------|--------------------------------|-----------------------------|-------------------------------|-------------------------------|
| Pt. 6650               | 0.94                          | 0.87                          | 1.00                           | 0.47                        | 0.69                          | 0.04                          |
| Pt. 5659               | 0.92                          | 0.74                          | 1.00                           | 0.54                        | 1.08                          | 0.03                          |
| Pt. 5661               | 0.92                          | 0.67                          | 0.99                           | 0.54                        | 1.25                          | 0.03                          |
| Pt. 5665               | 0.91                          | 0.72                          | 0.98                           | 0.54                        | 1.02                          | 0.04                          |
| Pt. 6781               | 0.91                          | 0.69                          | 0.98                           | 0.60                        | 0.82                          | 0.03                          |
| Pt. 6780               | 0.90                          | 0.81                          | 0.99                           | 0.65                        | 1.03                          | 0.04                          |
| Pt. 2684               | 0.88                          | 0.67                          | 0.98                           | 0.65                        | 0.98                          | 0.02                          |
| Pt. 3810               | 0.87                          | 0.70                          | 1.00                           | 0.66                        | 0.98                          | 0.02                          |
| Pt. 5656               | 0.80                          | 0.63                          | 0.98                           | 0.83                        | 1.11                          | 0.07                          |
Signatures Database (MSigDB) v6.1 from a pre-ranked GSEA of CRC1 vs. CRC2 subtypes in the COREAD FF1 cohort using either MCSs (expression using MCSs as input (Table 1). Significance was evaluated by the log-rank test. Right panel: patients with FFPE tumors and good clinical follow-up (The FFPE1 cohort; Supplementary Table 1) were calculated as the 40th percentile of DNA methylation scores were derived for COREAD and FFPE1 samples using GISTIC and EPIC DNA methylation data, respectively, and sample DNA methylation scores were the same two CRC subtypes in fresh-frozen and FFPE cohorts. a Pearson’s r and P value (Wilcoxon rank sum test) is indicated. b, c Scatterplots showing the correlation between normalized enrichment scores (NESs) for ~17 K gene sets of The Molecular Cancer Epithelial cell processes Up in serrated vs. conventional CRC Up in MSI vs. MSSCRC Up in CIMP vs. non-CIMP CRC Hallmark G2M checkpoint Wound healing Hallmark E2F targets Hallmark DNA repair Hallmark MYC targets Hallmark complement Hallmark IFNγ response Undifferentiated cancer Up in WNT pathway Up in colon crypt bottom vs. Top Genes downregulated by APC Up in MSS vs. MSI CRC Up in non-CIMP vs. CIMP CRC MSi MSS MethCORR subtypes

Table showing selected gene sets differentially enriched between FF1 and FF2 patients stratified by MethCORR subtypes. a Main molecular features of the CRC1 and CRC2 MethCORR subtypes in the COREAD FF1 and the FFPE1 cohort (Supplementary Table 1). MSI and MSS status is indicated in black and white. CIN scores were derived for COREAD and FFPE1 samples using GISTIC and EPIC DNA methylation data, respectively, and sample DNA methylation scores were calculated as the 40th percentile of DNA methylation β-values for all CpGs. Stroma- and Immune Scores were generated from MCSs using the ESTIMATE software. Stroma Score Low High Immune Score Low High

Fig. 2 MethCORR based NMF clustering identifies the same two CRC subtypes in fresh-frozen and FFPE cohorts. a Main molecular features of the CRC1 and CRC2 MethCORR subtypes in the COREAD FF1 and the FFPE1 cohort (Supplementary Table 1). MSI and MSS status is indicated in black and white. CIN scores were derived for COREAD and FFPE1 samples using GISTIC and EPIC DNA methylation data, respectively, and sample DNA methylation scores were calculated as the 40th percentile of DNA methylation β-values for all CpGs. Stroma- and Immune Scores were generated from MCSs using the ESTIMATE software. Stroma Score Low High Immune Score Low High

Fig. 2 MethCORR based NMF clustering identifies the same two CRC subtypes in fresh-frozen and FFPE cohorts. a Main molecular features of the CRC1 and CRC2 MethCORR subtypes in the COREAD FF1 and the FFPE1 cohort (Supplementary Table 1). MSI and MSS status is indicated in black and white. CIN scores were derived for COREAD and FFPE1 samples using GISTIC and EPIC DNA methylation data, respectively, and sample DNA methylation scores were calculated as the 40th percentile of DNA methylation β-values for all CpGs. Stroma- and Immune Scores were generated from MCSs using the ESTIMATE software. Stroma Score Low High Immune Score Low High

Fig. 2 MethCORR based NMF clustering identifies the same two CRC subtypes in fresh-frozen and FFPE cohorts. a Main molecular features of the CRC1 and CRC2 MethCORR subtypes in the COREAD FF1 and the FFPE1 cohort (Supplementary Table 1). MSI and MSS status is indicated in black and white. CIN scores were derived for COREAD and FFPE1 samples using GISTIC and EPIC DNA methylation data, respectively, and sample DNA methylation scores were calculated as the 40th percentile of DNA methylation β-values for all CpGs. Stroma- and Immune Scores were generated from MCSs using the ESTIMATE software. Stroma Score Low High Immune Score Low High

Fig. 2 MethCORR based NMF clustering identifies the same two CRC subtypes in fresh-frozen and FFPE cohorts. a Main molecular features of the CRC1 and CRC2 MethCORR subtypes in the COREAD FF1 and the FFPE1 cohort (Supplementary Table 1). MSI and MSS status is indicated in black and white. CIN scores were derived for COREAD and FFPE1 samples using GISTIC and EPIC DNA methylation data, respectively, and sample DNA methylation scores were calculated as the 40th percentile of DNA methylation β-values for all CpGs. Stroma- and Immune Scores were generated from MCSs using the ESTIMATE software. Stroma Score Low High Immune Score Low High
that most CpGs were not located on the same chromosome as the gene they correlate with (Supplementary Fig. 4a). Instead, the most frequently occurring CpGs were located in genomic regions that exhibited great cell-type-specific variation in DNA methylation, as evaluated in 17 tissue types (GSE5019218; Supplementary Fig. 4b). Hence, the MethCORR matrix may help associate gene expression with particular cell types by comparing the methylation pattern of expression-correlated CpGs to known DNA methylation (or DNase I hypersensitivity) profiles of cell monouclones/homogenous cell preparations. Indeed, expression-correlated CpGs for the T-cell-specific CD3 Epsilon (CD3E) gene overlapped with T-cell specific DNase I hypersensitive sites and DNA methylation patterns characteristic of T-cells (Supplementary Fig. 4c, d). Similarly, expression-correlated CpGs for fibroblast activation protein alpha (FAP) and epithelial cellular adhesion molecule (EPCAM) overlapped with patterns characteristic of stromal cells/fibroblasts and intestinal epithelial cells, respectively (Supplementary Fig. 4c, d). We also found that the genes with greatest expression-correlated CpG site overlap with CD3E, FAP, and EPCAM were themselves significantly associated with T-, stromal/fibroblast-, and epithelial-cell activities as evaluated by gene list enrichment analysis30 (Supplementary Fig. 4e; P value < 0.05 by the Enrichr software). This showed that analysis of expression-correlated CpGs help identify clusters of co-expressed genes and link them to particular cell types via comparison to cell-type-specific DNA methylation profiles.

To analyze expression correlations in a genome-wide format, we created a MethCORR map by clustering all MethCORR genes according to their overlap in expression-correlated CpGs (Fig. 3a). Foremost, the map was used to visualize differences between CRC1 and CRC2 by coloring gene nodes according to their difference in median MCS z-score between the subtypes (Δmedian z-score; Fig. 3a). The differences were near-identical for FF and FFPE cohorts (Fig. 3a, b and Supplementary Fig. 5a; Δmedian z-score Pearson’s r range: 0.88–0.97, P value < 10–100, WRS test) and near-identical to a MethCORR map comparing serrated/MSI and conventional adenocarcinomas from the 450K methylation dataset GSE6806031 (Fig. 3c; Δmedian z-score Pearson’s r range: 0.87–94, P value < 10–100, WRS test). Similar results were obtained when the map was overlaid with MethCORR interpretation of a transcriptional gene set de
ing serrated vs. conventional CRC (Supplementary Fig. 5b; Pearson’s r range = 0.94–98, P value < 10–100, WRS test; for comparison to MSI status, CIMP status, CMS- and CRIS-classification status see Supplementary Fig. 5c, d). This suggested that CRC1 and CRC2 subtypes resembles serrated/MSI and conventional carcinomas, respectively. In support, Submap analysis confirmed that CRC1 and CRC2 subtypes from all four cohorts corresponded to the serrated/MSI and conventional subtypes from the GSE68060 dataset31 (Supplementary Fig. 3d). Furthermore, CRC2 encompassed several map regions associated with high CIN scores, whereas CRC1 encompassed a large tumor microenvironment (TME) cluster characterized by genes with high stroma scores, as expected for conventional and serrated/MSI tumor subtypes2,32, respectively (Fig. 3d).

The MethCORR map characterizes intra-tumor heterogeneity. To investigate the large TME cluster in greater detail and provide insight into sources of CRC heterogeneity, the map was overlaid with MCS z-scores calculated from DNA methylation profiles of epithelial, immune, stem, and mesenchymal cells (primarily cell monouclones; Supplementary Table 4 and Supplementary Data 13). This identified map regions representing CAFs, CD14+ monocytes, CD3+ T cells, and CD19+ B cells among others (Fig. 3e). Again, similar results were obtained when the map was overlaid with MethCORR interpretations of RNA-based biomarkers and signatures defining CAFs, endothelium, myeloid cells, T cells, and B cells (Supplementary Fig. 5e). Hence, the MethCORR map can suggest cell types associated with RNA biomarkers and signatures via comparison to known cell-type-specific methylation profiles.

Based on this, we envisioned that the MethCORR map would visualize and suggest sources of inter-tumor heterogeneity between and within subtypes. CRC heterogeneity can arise from both differences in TME cell composition and in the differentiation status of tumor epithelial cells. For example compared with normal mucosa, CRCs can lose mature enterocyte traits and rather resemble enterocyte precursors, transit amplifying (TA) and stem cells, or undergo EMT2,33,34. Mapping of MCS z-scores from individual tumors revealed inter-tumor heterogeneity in both subtypes. For CRC1, heterogeneity was pronounced in the TME cluster and few samples had a dominant epithelial pattern (Fig. 3f). Three TME patterns were frequently observed, one overlapping with CAF/fibroblast (CAF/fibroblast pattern), another with CD14+ favors and inflammatory cells/plasmacytoid differentiation pattern), and the last with lymphocytes, T cells and B cells (lymphocyte pattern; Fig. 3e–g). This suggested that TME cell composition is a major contributor to intra-subtype heterogeneity in the immune-infiltrated CRC1 subtype. The TME patterns were less dominant among CRC2 samples (Fig. 3h) consistent with CRC2 conventional-like tumors being less immune-infiltrated32 (Fig. 2a, d). Instead, CRC2 heterogeneity was pronounced within epithelial map regions and four patterns were observed (Fig. 3h): Two regions were dominated by signatures of enterocyte precursors and TA cells as estimated by overlapping the map with RNA signatures defining specific differentiation states of intestinal epithelial cells33 (Fig. 3i). A third region overlapped with a mature enterocyte signature characteristic of normal mucosa samples (Fig. 3i and Supplementary Fig. 5f). Finally, an EMT pattern was identified in CRC2 by overlaying the map with the MCSs of Hela cells undergoing EMT35 (Fig. 3i) and GSEA showed enrichment of EMT signatures in the CRC2 samples with this EMT pattern (as compared with an early enterocyte pattern; Supplementary Fig. 5g). Collectively, this suggested that epithelial differentiation status is an important contributor to heterogeneity in the CRC2 subtype. Finally, the above heterogeneity was also identifiable among CRC cell lines and CMS subtypes (Supplementary Fig. 5h, i).

MethCORR interprets prognostic RNA signatures. We next investigated if MethCORR would also help identify DNA methylation-based biomarkers suited for prognostication using FF and FFPE samples. Our strategy was to use the MethCORR map to interpret established, prognostic RNA signatures and suggest cell types associated with tumor aggressiveness, which can be evaluated in DNA samples based on the cell-type specificity of methylation. Analysis of five prognostic signatures, CRC-11336, ColoGuideEx,37, Oncotype DX38, ColoPrint39, and Tian et al.40 showed that MCSs for almost all stromal transcripts were positively correlated with the median MCS for all signatures (Fig. 4a). This suggested that all signatures associated high TME activity with poor prognosis. MethCORR map analysis of the signatures revealed two distinct patterns within the TME cluster: The CRC-113, ColoGuideEX, and Oncotype DX signatures associated with a CAF-like pattern (Figs. 3e, f, and 4b), cancer invasiveness and hepatocyte growth factor (HGF) expression31 (Fig. 4c, d). The ColoPrint and Tian et al. signatures (Fig. 4e) associated with an inflammation/wound healing pattern (Figs. 3e, f, and 4c) encompassing blood platelets, CD14+ monocytes (Fig. 3e), and transforming growth factor beta 1 (TGFβ1) expression (Fig. 4d).
Hence, the prognostic signatures overlapped in predictions, and pointed to CAF or inflammation/wound healing as associated with poor prognosis CRC. We recently reported that subtype-specific RNA signatures can improve prognostication beyond TNM staging in multiple CRC cohorts. Therefore, MethCORR was also used to interpret these subtype-specific prognostic signatures denoted SSC prognosis and CIN prognosis. These are intended for immune-infiltrated/serrated and conventional carcinoma subtypes, which correspond to CRC1 and CRC2 in this study, respectively. MethCORR map analysis suggested that depletion of immune cells, including T cells, was associated with the SSC prognosis signature (Figs. 3e and 4c, f), whereas a CAF...
Fig. 4 The MethCORR map suggests cell types associated with prognostic RNA signatures. a Bar plot showing the fraction of stromal genes (stroma score > 0.5) that have positive or negative cMCSs (correlation to the median MCS) calculated for five prognostic RNA signatures CRC-113, CologuideEx, Oncotype DX, Coloprint, and Tian et al. (see “Methods” for calculation of cMCSs). Stromal transcripts were significantly enriched among positively vs. negatively correlated transcripts for all five signatures (P value < 10^{-100}, Wilcoxon rank sum test). b Magnification of the TME cluster, where genes with the highest cMCSs for the prognostic CRC-113 (left), CologuideEx (middle), and Oncotype DX (right) signatures are highlighted. c Magnification of the TME cluster, where genes with the highest cMCSs for published gene sets defining cytotoxic T-cells (MSigDB M13247 (BioCarta); green), wound healing (MSigDB M11957 (red), or cancer invasiveness (MSigDB M2572 (orange)) are highlighted. d Magnification of the TME cluster, where genes with the highest correlation to the MCS of the IL2 (green), TGFβ1 (red), and HGF (orange) genes are highlighted. e Magnification of the TME cluster, where genes with the highest cMCSs for the prognostic Coloprint (right) and Tian et al. (left) signatures are highlighted. f Magnification of the TME cluster, where genes with the highest &Delta;cMCSs for the prognostic SSC prognosis signature and the CIN prognosis signatures are highlighted. g Magnification of the TME cluster colored according to the gene-specific percentage change in median MCSs between TNM stage I and IV CRCs of the COREAD cohort (left) and SYSCOL cohort (right). h Magnification of the TME cluster, where genes with the highest correlation to the MCS of CD3E (green), PDNP (red), ACTA2 (orange), and HNF4A (blue) are highlighted. i Scatterplot showing the Spearman rho for top CD3E or ACTA2/PDNP-correlated genes to positive relapse recurrence status in the CRC1 and CRC2 subtypes, respectively, in the fresh-frozen FF1–FF2 cohort (left) and FFPE1 (right) cohort. Median correlation is indicated by a black bar.
and EMT pattern was associated with the CIN prognosis signature (Figs. 3e and 4c, g). Furthermore, we compared MCSs for TNM stage I (favorable prognosis) to stage IV tumors (poor prognosis) in the COREAD and SYSCOL cohorts. Here, the relative change in MCSs between TNM stages also pointed to a relative loss of immune cells and increase in CAF content in late-stage, poor prognosis CRC (Fig. 4h). Collectively, the MethCORR analysis of seven published prognostic signatures hereby suggested that poor prognosis is associated with low T-cell content, particularly in the immune-infiltrated CRC1 subtype (Fig. 4f), or high CAF content and inflammation-EMT, particularly in the immune-depleted CRC2 subtype (Fig. 4g). To investigate the predictions of prognostic cell types in our FF and FFPE cohorts, we selected the three biomarkers CD3E, ACTA2, and PDPN. These are well-known markers for T cells, CAF/myofibroblasts, and inflammation-EMT, respectively, and their most closely CpG site-associated genes overlapped with regions highlighted by the prognostic classifiers (compare Fig. 4b, e, f, g, i; Supplementary Fig. 6). Indeed, top CD3E-associated genes negatively correlated with patient recurrence status in the CRC1 subtype and ACTA2/PDPN-associated genes positively correlated to patient recurrence in CRC2 (Fig. 4j).

DNA methylation-based biomarkers for CRC prognostication.

To derive DNA methylation biomarkers for the above prognostic cell types we exploited the cell type-specificity of DNA methylation. Comprehensive comparison of multiple cell types identified low methylation of CpGs within the CD3E, ACTA2, and PDPN promoter as biomarkers for T cells, CAFs/myofibroblasts, and inflammation-EMT, respectively (Fig. 5a; Supplementary Data 13). Indeed, analysis of promoter CpGs in CRC samples showed that high methylation of the CD3E promoter, reflecting low levels of T-cell infiltration, associated with significantly poorer RFS in CRC1 in both FF and FFPE cohorts (Fig. 5b). In addition, low ACTA2/PDPN promoter methylation, reflecting high CAF/EMT levels, associated with poor RFS in CRC2 (Fig. 5b). The biomarkers were superior predictors of RFS as compared with TNM staging and MSI status (Fig. 5c, Supplementary Fig. 7a, b), and the biomarkers were only prognostic within the intended subtype (Supplementary Fig. 7c). Finally, to provide a cost-effective alternative to genome-wide methylome analysis, we evaluated CD3E, ACTA2, and PDPN promoter methylation using quantitative methylation-specific PCR (Q MSP) assays. In addition, a QMSP assay targeting the HNF4A promoter was included for CRC subtyping; HNF4A is upregulated in CRC2 (Fig. 4i) and correspondingly, its promoter is less methylated in CRC2 (Fig. 5a). We applied our four biomarker assays to FFPE1 cohort samples, stratified patients into CRC1 and CRC2 using the HNF4A QMSP assay (Fig. 5d), and used CD3E and ACTA2/PDPN assays as prognostic biomarkers in CRC1 and CRC2. RFS analysis confirmed that the QMSP assays allowed subtype-specific prognostication using FFPE samples (Fig. 5e and Supplementary Fig. 7d).

Discussion

We here introduce MethCORR as an approach for uniform molecular analysis of FF and FFPE samples based on DNA methylation profiling. MethCORR allows inference of expression information from DNA methylation for a large number of genes (>11,000; Fig. 1). The inferred expression profiles support identical subtype discovery, characterization, and prognostication in FF and FFPE cohorts (Figs. 2–5). Notably, MethCORR allows three layers of information to be extracted from a DNA methylation array experiment, namely an inferred gene expression profile, a DNA methylation profile and a chromosome copy-number profile, calculated from the methylation array signal intensity. This improves cost-effectiveness and makes MethCORR attractive for analysis of archival FFPE material, where RNA profiling can be difficult. The MethCORR concept bears resemblance to transcriptome-wide association studies, where gene expression is correlated to genetic variation. However, MethCORR allows the expression of many more genes to be modeled, which indicates that gene expression is stronger associated with DNA methylation than genetic variation.

The high number of MethCORR genes with inferred expression may be surprising, as several previous studies reported more infrequent correlations, when investigating associations between gene expression and methylation at local enhancers, promoters, and gene bodies. MethCORR instead performs correlation analysis genome-wide and hereby identify far more associations from which expression information can be inferred. Indeed, expression-correlated CpGs were often located far from the gene locus, in regions with cell-type-specific methylation (Supplementary Fig. 4). Hence, MethCORR benefits from associating cell-type-specific gene expression with cell-type-specific DNA methylation patterns to infer expression information for many genes, even if associations are not functionally linked. Such indirect associations are expected in heterogenous cancer samples, which vary in their content of cancerous and non-cancerous cell types. Support for a genome-wide correlation strategy is also found in two previous studies, which on a smaller scale, performed RNA expression-correlation analysis with more distant located CpGs. However, these studies only included ~500 CpG sites distributed across the genome compared with 480,000 sites utilized in MethCORR, and consequently found much fewer strong correlations.

MethCORR introduces an expression-correlated measure, the MCS, which enabled identification of the same two CRC subtypes in all four cohorts analyzed, and this independent of the analyzed tissue being FF or FFPE. The subtypes resemble the two major carcinogenesis pathways described in CRC that are characterized by epithelial-cell hyper-methylation or chromosomal instability (Figs. 2 and 3). We speculate that MethCORR identified these well-established carcinogenesis pathways due to the relative emphasis of MCSs on cancer epithelial traits over stroma-related traits. Also, we observed higher correlations between MCSs profiles for matched FF and FFPE biopsies taken from the same tumor than between RNA and iRNA profiles (Table 1). We therefore speculate that MCS-based characterization and subtyping is more independent of sample preservation type, which now require further testing.

MethCORR also introduces a map that visualizes genome-wide associations between gene expression and DNA methylation in CRC (Fig. 3). We envision that MethCORR map analysis may provide a framework for more detailed characterization of FF and archival FFPE samples than categorical subtyping alone, e.g., to reveal cellular sources of inter-tumor heterogeneity (Fig. 3). In particular, we illustrated that the MethCORR map can help identify cell types associated with RNA signatures (Figs. 3 and 4) and hereby help to derive DNA methylation-based biomarkers suitable for FFPE samples (Fig. 5). Our MethCORR map analysis of several prognostic RNA signatures (Fig. 4) showed that they all predicted cancer aggressiveness to be associated with cell types within the TME: In particular, a high CAF content, inflammation-associated EMT, and low T-cell content were associated with poor prognosis (Fig. 4). This agrees with clinically promising biomarkers such as the Immunoscore and Tumor-Stroma Ratio. Our analysis of CRC subtype-specific prognostic RNA signatures offered additional resolution: the T-cell content was primarily prognostic within the immune-infiltrated CRC1 subtype.
whereas CAF-content/inflammation-EMT was only prognostic in the less immune-infiltrated CRC2 subtype (Fig. 5). This supported our previous observations of subtype-specific prognostic biomarkers\(^7\). To aid further testing of subtype-specific prognostication, we established four simple QMSP assays for cost-efficient CRC subtyping and prognostication. The application of the four QMSP assays in CRC samples confirmed and reproduced the RFS analysis derived from the more costly DNA methylome profiles (Fig. 5). Collectively, this illustrates the ability of MethCORR to help derive DNA methylation biomarkers from transcriptional signatures by extracting cell-type information from their expression-correlated CpGs.
Finally, MethCORR can provide high-quality gene expression measures in samples with poor RNA quality, such as archival FFPE samples for which confident RNA profiling is challenging. Our analysis of matched FFPE and FF tissue showed that RNA expression profiles from FFPE tissue resembled the RNA-sequencing profiles of the FF tissue better than RNA-sequencing profiles of the FFPE tissue. In PCA, matched FFPE iRNA and FF RNA-sequencing profiles clustered sample wise, while matched RNA-sequencing profiles of FFPE and FF tissue clustered according to preservation type. Preservation type-dependent clustering of FFPE and FF RNA-sequencing profiles have been reported previously, even in studies that report very high correlation between RNA-sequencing profiles of matched FFPE and FF samples. We acknowledge that recent studies focusing on newly produced FFPE samples with optimal fixation and short storage time have reported improved correlations between matched FFPE and FF RNA-sequencing profiles. However, such samples are not standard in the clinical FFPE archives. A large study, focusing on clinical FFPE samples, stored for many years, showed that gene expression quantification was achieved in only 60% of samples and that correlation between biological replicates was very variable.

The robustness of MethCORR likely reflects that the Illumina Infinium HumanMethylation microarray produces highly concordant results in FFPE and FF samples when using DNA restoration for FFPE samples (Supplementary Fig. 2h). Furthermore, the DNA methylation \( \beta \)-values are calculated as the ratio between methylated and unmethylated CpG sites at a given genomic position. Hence, although a genomic region is affected by degradation, the ratio between the methylated and unmethylated fragments (i.e., the DNA methylation \( \beta \)-value) would expectably be robust. In contrast, RNA profiling is highly affected by RNA degradation and the RNA quality obtainable from FFPE is often compromised. In agreement, tumor samples with the lowest correlation between iRNA and measured RNA expression had lower RNA quality scores than samples with high correlations, whereas 450K methylation data quality did not differ (Supplementary Fig. 2g). This suggest that expression profiling of FF samples is influenced by even slight RNA degradation, as reported previously.

In conclusion, DNA methylation profiling and MethCORR analysis enables reliable and robust gene expression estimates to be obtained from clinical samples with compromised RNA quality. Furthermore, MethCORR data can be used to obtain clinically relevant information on tumor subtypes, cellular heterogeneity, and to develop prognostic biomarkers. Consequently, MethCORR represents an effective mean to unlock the unique and extensive resource of FFPE tissues in the pathology archives. We envision that MethCORR in the future will be established for many other cancer types.

Methods

CRC patient cohorts. The COREAD cohort encompasses mucosa and UICC TNM stage I–IV CRC samples collected as part of TCGA project. All information regarding COREAD samples including processed DNA methylation data, RNA expression data, gene-level copy-number data, and clinical patient information (phenotype) were acquired via the UCSC XENA Public Data Hubs and the GDC Data Portal. The SYSCOL and FPPE1 cohorts were acquired from the CRC biobank at the Department of Molecular Medicine, Aarhus University Hospital, Denmark. SYSCOL samples were collected at hospitals in the central region of Jutland, Denmark from 1999–2013. The FPPE1 cohort encompasses CRC samples from the prospective study COLOFOI, collected at hospitals in the central region of Jutland, Denmark. None of the patients received neoadjuvant therapy. The tumors were histologically classified and staged according to the UICC TNM staging system. Cancer cell percentage was evaluated individually by two trained researchers, and when necessary, tumor biopsies were macroscopically trimmed to enrich the fraction of neoplastic cells. The SYSCOL and COLOFOI study was conducted in accordance with Danish law and is approved by local institutional review boards and ethical committees and written informed consent was obtained from all patients. The FPPE2 cohort (IDIBELL) encompasses 56 samples collected at Medical Oncology Service of ICO Badalona-Germans Trias i Pujol Research Institute (IGTP), Spain. None of the patients received neoadjuvant therapy. The tumors were histologically classified and staged according to the UICC TNM staging system. Cancer cell percentage was evaluated individually by two trained researchers, and when necessary, tumor biopsies were macroscopically trimmed to enrich the fraction of neoplastic cells. Patients were followed according to the national clinical guidelines and written informed consent was obtained from all patients. Clinical information regarding the COREAD, SYSCOL, COLOFOI, and IDIBELL cohort samples is presented in Supplementary Table 1.

DNA methylation data. FF tumors from the SYSCOL cohort were macromixed to enrich the fraction of neoplastic cells and DNA was extracted using the Puregene DNA purification kit (Gentra Systems). Integrity of the DNA from FFPE was assessed by melting curves and short storage time have reported improved correlations between matched FFPE and FF RNA-sequencing profiles. However, such samples are not standard in the clinical FFPE archives. A large study, focusing on clinical FFPE samples, stored for many years, showed that gene expression quantification was achieved in only 60% of samples and that correlation between biological replicates was very variable. The robustness of MethCORR likely reflects that the Illumina Infinium HumanMethylation microarray produces highly concordant results in FFPE and FF samples when using DNA restoration for FFPE samples (Supplementary Fig. 2h). Furthermore, the DNA methylation \( \beta \)-values are calculated as the ratio between methylated and unmethylated CpG sites at a given genomic position. Hence, although a genomic region is affected by degradation, the ratio between the methylated and unmethylated fragments (i.e., the DNA methylation \( \beta \)-value) would expectably be robust. In contrast, RNA profiling is highly affected by RNA degradation and the RNA quality obtainable from FFPE is often compromised. In agreement, tumor samples with the lowest correlation between iRNA and measured RNA expression had lower RNA quality scores than samples with high correlations, whereas 450K methylation data quality did not differ (Supplementary Fig. 2g). This suggest that expression profiling of FF samples is influenced by even slight RNA degradation, as reported previously.

In conclusion, DNA methylation profiling and MethCORR analysis enables reliable and robust gene expression estimates to be obtained from clinical samples with compromised RNA quality. Furthermore, MethCORR data can be used to obtain clinically relevant information on tumor subtypes, cellular heterogeneity, and to develop prognostic biomarkers. Consequently, MethCORR represents an effective mean to unlock the unique and extensive resource of FFPE tissues in the pathology archives. We envision that MethCORR in the future will be established for many other cancer types.

Methods

CRC patient cohorts. The COREAD cohort encompasses mucosa and UICC TNM stage I–IV CRC samples collected as part of TCGA project. All information regarding COREAD samples including processed DNA methylation data, RNA expression data, gene-level copy-number data, and clinical patient information (phenotype) were acquired via the UCSC XENA Public Data Hubs and the GDC Data Portal. The SYSCOL and FPPE1 cohorts were acquired from the CRC biobank at the Department of Molecular Medicine, Aarhus University Hospital, Denmark. SYSCOL samples were collected at hospitals in the central region of Jutland, Denmark from 1999–2013. The FPPE1 cohort encompasses CRC samples from the prospective study COLOFOI, collected at hospitals in the central region of Jutland, Denmark. None of the patients received neoadjuvant therapy. The tumors were histologically classified and staged according to the UICC TNM staging system. Cancer cell percentage was evaluated individually by two trained researchers, and when necessary, tumor biopsies were macroscopically trimmed to enrich the fraction of neoplastic cells. The SYSCOL and COLOFOI study was conducted in accordance with Danish law and is approved by local institutional review boards and ethical committees and written informed consent was obtained from all patients. The FPPE2 cohort (IDIBELL) encompasses 56 samples collected at Medical Oncology Service of ICO Badalona-Germans Trias i Pujol Research Institute (IGTP), Spain. None of the patients received neoadjuvant therapy. The tumors were histologically classified and staged according to the UICC TNM staging system. Cancer cell percentage was evaluated individually by two trained researchers, and when necessary, tumor biopsies were macroscopically trimmed to enrich the fraction of neoplastic cells. Patients were followed according to the national clinical guidelines and written informed consent was obtained from all patients. Clinical information regarding the COREAD, SYSCOL, COLOFOI, and IDIBELL cohort samples is presented in Supplementary Table 1.

DNA methylation data. FF tumors from the SYSCOL cohort were macromixed to enrich the fraction of neoplastic cells and DNA was extracted using the Puregene DNA purification kit (Gentra Systems). Integrity of the DNA from FFPE was assessed by melting curves and short storage time have reported improved correlations between matched FFPE and FF RNA-sequencing profiles. However, such samples are not standard in the clinical FFPE archives. A large study, focusing on clinical FFPE samples, stored for many years, showed that gene expression quantification was achieved in only 60% of samples and that correlation between biological replicates was very variable. The robustness of MethCORR likely reflects that the Illumina Infinium HumanMethylation microarray produces highly concordant results in FFPE and FF samples when using DNA restoration for FFPE samples (Supplementary Fig. 2h). Furthermore, the DNA methylation \( \beta \)-values are calculated as the ratio between methylated and unmethylated CpG sites at a given genomic position. Hence, although a genomic region is affected by degradation, the ratio between the methylated and unmethylated fragments (i.e., the DNA methylation \( \beta \)-value) would expectably be robust. In contrast, RNA profiling is highly affected by RNA degradation and the RNA quality obtainable from FFPE is often compromised. In agreement, tumor samples with the lowest correlation between iRNA and measured RNA expression had lower RNA quality scores than samples with high correlations, whereas 450K methylation data quality did not differ (Supplementary Fig. 2g). This suggest that expression profiling of FF samples is influenced by even slight RNA degradation, as reported previously.

In conclusion, DNA methylation profiling and MethCORR analysis enables reliable and robust gene expression estimates to be obtained from clinical samples with compromised RNA quality. Furthermore, MethCORR data can be used to obtain clinically relevant information on tumor subtypes, cellular heterogeneity, and to develop prognostic biomarkers. Consequently, MethCORR represents an effective mean to unlock the unique and extensive resource of FFPE tissues in the pathology archives. We envision that MethCORR in the future will be established for many other cancer types.
data from 394 samples of the COAD and READ cohorts (COREAD) of the TCGA project, acquired in normalized format via the UCSC XENA Public Data Hubs (Supplementary Data 2). The analysis was performed using log2(FFPKM + 1) normalized RNA expression values for all available 20,530 RNAs and DNA methylation beta-values for the 396,065 CpGs, where beta-values were provided by the XENA Public Data Hubs54. This analysis generated the COREAD MethCORR matrix (Supplementary Data 3) that is used for calculation of MCSs throughout the manuscript, unless otherwise indicated and modeling metric is reported in Supplementary Data 4. Second, the MethCORR approach was applied to RNA-sequencing (20,336 RNAs) and HM-450K DNA methylation profiles (485,512 CpGs) from 314 samples of the SYSCOL cohort (Supplementary Data 5–7) with the aim to validate the performance of the MethCORR approach in an independent cohort. Third, the MethCORR approach was applied to 405 TCGA COREAD samples using expression (17,611 RNAs, these were selected from the original dataset of 60,483 transcripts as they overlap with the RNAs included in the UCSC XENA RNA dataset) and DNA methylation data (395,011 CpGs) acquired via the NCI GDC25 (Supplementary Data 8). This analysis was performed to investigate the impact of RNA normalization methods on MethCORR performance (modeling metrics in Supplementary Data 9 and 10) and to generate a GDC data based MethCORR matrix that was used for analysis of the TCGA FFPE samples included in this study, as data from these FFPE samples were also acquired via the GDC database (Supplementary Data 11).

Identification of RNA expression-correlated CpG sites. The CRC cohort was divided into two discovery sets (sets 1–2, each encompassing 40% of samples), whereas a third set was reserved for independent validation (set 3, 20% of the samples; Fig. 1a and Supplementary Data 1, 6, and 8). Genome-wide correlations (Spearman) between the expression of each of the RNAs (log2(FFPKM + 1)) and the DNA methylation beta-value of each CpG site were calculated independently in discovery sets 1 and 2 using the publicly available R function “corr.” All non-significant correlation pairs were discarded (Spearman’s correlation P value < 0.01). The remaining expression-correlated CpGs were ranked by their Spearman’s rho in each discovery set and next by their rank sum within discovery sets 1 and 2 to identify top common expression-correlated CpGs. From these lists of ranked CpGs specific for each RNA, we selected up to 100 CpGs whose methylation beta-value was most positively or negatively correlated with its expression resulting in lists of ≤200 RNA expression-correlated CpGs for each RNA (depending on the number of expression-correlated CpGs in the ranked lists). To ensure analysis robustness, especially in FFPE samples, we included all CpGs in lists that had a detection value > 0.05 (ChAMP package57) in ≥5% samples in either the SYSCOL, FFPE1, or FFPE2 cohort. Top ranking CpGs for all analyzed genes for the TCGA COREAD cohort (datasets acquired via the UCSC XENA Public Data Hubs) can be found in Supplementary Data 3.

Calculation of MethCORR scores. For each sample we used the methylation beta-values of the top ≤200 RNA expression-correlated CpGs (for each gene) to calculate a MCS for all genes with both positively and negatively expression-correlated CpGs using the formula:

\[
\text{MCS} = \frac{1}{2 \cdot N} \sum_{i=1}^{200} \beta_{\text{value pos. corr. CpG probe}} + \sum_{i=1}^{200} (1 - \beta_{\text{value neg. corr. CpG probe}})
\]

The MCS formula calculates the average methylation value of the expression-correlated CpG sites specific for each gene. Unless otherwise indicated, the COREAD MethCORR matrix encompassing expression-correlated CpGs for 11,222 genes (Supplementary Data 3; MethCORR genes) was used for calculation of MCSs throughout the manuscript. The use of the MSC formula above and the MethCORR matrix provided in Supplementary Data 3 allow calculation of MCSs from DNA methylation beta-values of any relevant 450K CRC data set choice of.

Modeling and inferring of RNA expression from MCSs. We modelled the relationship between MCSs and RNA expression for each gene in the discovery samples (set 1 + 2; Fig. 1A) using both simple linear (RNA = β0 + β1 × MCS) and polynomial regression models (RNA = β0 + β1 × MCS + β2 × MCS2...+ β8 × MCS8; n = 2–4). The Caret R package58 was used to perform modeling by 10 × 10-fold cross validation and we used the average RMSE to select the best model for each gene. As performances were highly similar for simple linear and polynomial models for most genes, we only selected polynomial models if a 25% relative decrease in RMSE values were observed over simple linear models. Model performances were independently validated in validation set 3 (Supplementary Data 2, 7, and 8) with both models (linear and polynomial) in both sets (set 1 + 2) and validation (set 3) were regarded as MethCORR genes and included in the MethCORR matrix (Supplementary Data 3), whereas genes with poorer performing models were excluded. For MethCORR genes we inferred RNA (iRNA) expression for each gene in each sample using the MCS as input in the gene-specific models and the sample-specific iRNA expression values provided in Supplementary Data 2, which allow calculation of RNA profiles from MCSs for any relevant 450K CRC data set choice of.

Establishment and analysis of a MethCORR map. The MethCORR map for the COREAD cohort was created by clustering MethCORR genes according to their overlap in expression-correlated CpGs using Cytoscope V3.2.059 and the application EnrichmentMap80 (Jaccard + Overlap filtering cutoff 0.126). Only CpGs with negatively expression-correlated CpGs from the MethCORR matrix were used for identifying the overlap given that inclusion of all expression-correlated CpGs in a single map would complicate interpretation as genes with opposite expression-correlated CpGs would be clustered together. The initial significant CpG overlap to other genes are not included in the graphical representation of the MethCORR map for visual simplicity. For interpretation, the MethCORR map was overlain with several data types including external DNA methylation data, transcriptionally defined marker genes, gene sets, and signatures. To visualize these diverse data types using the MethCORR map, four types of scores were established as follows:

For DNA methylation datasets (450K/EPIC arrays), MCSs were first calculated for all samples and two types of scores were used for map visualization. The difference in median MCS z-scores (Δmedian MCS z-score) was used to visualize scores between subtypes encompassing multiple samples (such as between MethCORR subtypes, CMS subtypes, CRIS subtypes, MSI vs. MSS tumors etc.) whereas MCS z-scores were used for visualization of differences between individual samples within a cohort. MCS z-scores were calculated for each gene within each investigated cohort by subtracting the cohort mean from an individual sample MCS and dividing the difference by the cohort standard deviation. E.g. for analysis of inter-tumor heterogeneity, MCS z-scores were calculated for each within the whole COREAD FF1 cohort. For analysis of the cellular composition of the TME cluster, MCS z-scores were calculated from a collection of cell types with available 450K data downloaded from either Marmal-aid81, Gene Expression Omnibus (GEO)82, or Array express (see Supplementary Table 4 and Supplementary Data 13 for details of included samples; before calculation of MCS z-scores across all sample types the median MCSs were calculated for similar sample types, such as technical replicates).

For transcriptionally defined marker genes, gene sets, and signatures, two types of scores were used for map visualization depending on the data format. For simple gene sets and RNA signatures, defined by only one gene list (e.g., either up or downregulated RNAs), a correlation to median MCS (cMCS) was calculated for each MethCORR gene. The cMCSs were calculated as the average Pearson correlation between the median MCS of the gene set and the MCS of each MethCORR gene within the FF1, FF2, FFPE1, or FFPE2 cohorts. For complex gene sets/signatures, defined by two gene lists (e.g., of both up and downregulated genes), a correlation to median MCS difference score (ΔcMCS) was instead calculated for each MethCORR gene. The ΔcMCSs were calculated by subtracting the cMCSs for the downregulated gene set from the cMCSs for the upregulated gene set (ΔcMCS = cMCSup - cMCSdown) for each gene. For visualization, MethCORR map gene nodes were colored according to these MCS z-scores, ΔMCS z-scores, cMCS, and ΔcMCS as indicated in the text. For map visualization of published prognostic signatures, cMCS were calculated for the five general (non subtype-specific) signatures (CRC-11386, ColoGeneEx137, Oncotype DX87, ColoPrim88, and Tian et al.89), as they are simple lists of RNAs associated with poor prognosis CRC (only recurrence score genes from the Oncotype DX panel were analyzed, whereas treatment genes were excluded). For theCRC subtype-specific SSC prognostic and CIN prognostic signatures ΔcMCS were calculated, as they are complex signatures encompassing lists of RNAs with high and low expression in aggressive CRC.

NMF-based consensus clustering and Submap analysis. NMF consensus clustering was performed using the R-package NMFH90 with MCSs as input. The number of classes was determined by the first distinctive reduction in the cophenetic score and silhouette consensus score91 and samples were classified according to consensus class. The similarity of independent subtype predictions was analyzed using the GenePattern Submap module (v3.140) using pairwise comparisons of MCSs and the following settings: nums. marker genes = 50, number permutations = 1000, weighted score type = no, null distribution = equal. A false discovery rate (FDR) P value < 0.05 was used as significance cutoff (provided by the Submap software91).

CMS and CRIS subtype classification. CMS classification was performed with the R-package CMScalifier using the single sample method and nearest CMS as predicted subtype92. RNA expression or iRNA expression were used as input, as indicated in the text. CRIS classification was performed using the R-package CRIScaler provided by Isella et al.4 using RNA expression or iRNA expression as input, as indicated in the text.

Stroma, CIN, DNA methylation, and ESTIMATE scores. Stroma scores for each gene (fraction of reads of murine origin) was acquired from Isella et al.48. Genes with stroma scores < 0.5 were considered stromal genes, whereas genes with stroma scores > 0.1 were considered epithelial cancer genes. For the COREAD cohort, representative sample-specific CMS scores were established from the gene-level copy-data number (GISTIC2 analysis) available at the UCSC XENA Public Data Hubs84. The gene CIN scores were defined for each gene as the standard deviation of the
GISTIC2 copy-numbers of all samples within the COREAD cohort. The sample CIN scores were defined as the standard deviation of GISTIC copy-number scores across genes within a sample, calculated for each sample within the COREAD cohort. For non-COREAD cohort samples (without GISTIC data) CIN scores were derived from copy-number data extracted from the HM–450K/EPIC methylation BeadChips using the chAMP.CNA module of the ChAMP R-package27. Here, the sample CIN score was defined as the mean interquartile range of the copy-numbers for all chromosomal segments (seg-mean) covered by at least 25 Illumina probes (num. probes). The sample DNA methylation score for each sample was defined as the 40th percentile of DNA methylation β-values of all CpG sites common to all four CRC cohorts. ESTIMATE Stroma scores and Immune scores were calculated for the R-package parameters and MCSs as input. Household gene status was defined as genes included in the list of housekeeping genes20 available at [https://www.tau.ac.il/~eileis/HKG/].

**Gene set enrichment analysis.** Pre-ranked GSEA was performed using the GSEA 3.0 tool29 using default settings. Genes were pre-ranked according to their Spearman correlation of their MCS to CRC1 subtype status and gene sets were considered significantly up- or downregulated for FDR q values <0.05 (provided by the Bioconductor Software28). The Molecular Signatures Database (MsigDB) gene set collection v6.1 was used with the addition of custom gene sets (Supplementary Table 3).

**Immunohistochemistry.** Immunohistochemical staining of CRC tissue sections were acquired from the Human Protein Atlas21 [https://www.proteinatlas.org/]. The following antibody and tissue sections were chosen (available from v8.proteinatlas.org): ACTA2 (antibody: CAB019417; Pt. 2017, Pt. 2151, Pt. 3074), PDN (antibody: HPAA07534; Pt. 201, Pt. 1958, Pt. 3264), CD3E (antibody: PA043955; Pt. 4724, Pt. 5005, Pt. 4448, Pt. 5004), HNF4A (antibody: CAB019417; Pt. 2017, Pt. 2151, Pt. 3074).

**Identification of cell-type-specific DNA methylation.** Genomic regions with cell-type-specific DNA methylation was identified by comparing multiple cell types with available 450K analysis downloaded from either Marmal-aid64, GEO65, or Array express72. Median MCSs were calculated for similar sample types, such as technical replicates prior to analysis (see Supplementary Data 13 for details). For selection of cell-type-specific methylation markers, only CpG probes with a ≥0.3 lower methylation β-value in the intended cell type, as compared with other relevant cell types, were selected. The following genes/CpG probes were included here: CD3E/cpg24612198, ACTA2/cpg0999048, PDN/cpg1556936, HNF4A/cpg0664063.

**Quantitative methylation-specific PCR.** QMSP was performed using DNA priers specific for unmethylated CD3E, ACTA2, PDN, and HNF4A gene promoter regions (See Supplementary Table 5 for primer sequences). BS conversion was performed with the EZ DNA Methylation–Direct Kit (Zymo research) according to the manufacturers protocol. QMSP was performed using the Via qPCR system (Applied Biosystems). Biomarker assay reactions were carried out in triplicate in a final volume of 6 µl and contained 2.5 µl TaqMan® Universal PCR Master Mix, No AmpErase® UNG (Applied Biosystems), 0.15 µl of 20 pmol/µl forward and reverse primer, 0.2 µl of 5 pmol/µl hydrolysis probe, 0.125 µl TEMase hot start DNA polymerase, 0.4 µl of 12.5 µmol/µl dNTP mix, 0.475 µl H2O, and 2 µl of 2.5 ng/µl BS treated DNA template. AluC4A reference gene reactions were carried out in triplicate in a final volume of 6 µl and contained 2.5 µl TaqMan® Universal PCR Master Mix, No AmpErase® UNG (Applied Biosystems), 0.2 µl of 25 pmol/µl forward and reverse primer, 0.1 µl of 17 pmol/µl hydrolysis probe, 0.125 µl TEMase hot start DNA polymerase, 0.4 µl of 12.5 µmol/µl dNTP mix, 0.475 µl H2O, and 2 µl of 2.5 ng/µl BS treated DNA template. QMSP reactions were mixed in MicroAmp Optical 384 Well reaction Plates (Applied Biosystem) and run on the ViiA™ 7 Real-Time PCR system (Applied Biosystems) with the following PCR program: denaturation at 95 ℃ for 10 min, followed by 40 cycles at 95 ℃, 1 min, and 60 ℃, 1 min. The Via qPCR TM software was used for evaluation of the fluorescence signals and the ΔCT was calculated by the use of the reference gene AluC4A. Subtyping was performed using ΔCTuniv as a marker for CRC2 and the average of ΔCTCD3E and ΔTCPDN as a marker for CRC1 (high stromal/immune cell infiltration). Samples with a ΔCTHNF4A/ΔCTCD3E + ΔTCPDN/ΔCTaverage ≥ 0.85 were defined as CRC2 samples. CT values were measured three times for each sample (technical replicates).

**Statistical analysis and RFS analysis.** Unless otherwise noted, statistical significance of differences between groups was determined using a non-parametric WRS test. During Submap analysis26 and pre-ranked GSEA29 a FDR-corrected P value < 0.05 was considered significant and P values were provided by the corresponding software. During gene list enrichment analysis an adjusted P value < 0.05 was considered significant (provided by the Enrichr software). During eQFDR analysis a Q-value < 0.05 was considered significant (provided by the eQFDR software). RFS analysis was performed in UICC TNM stage II–III samples with good clinical annotation and follow-up (Supplementary Table 1). The inclusion criteria were as follow: A minimum of 2 years follow-up and survival after tumor resection, no local recurrence of the disease, no other cancer within 3 years, and no synchronous cancers. RFS was measured from date of surgery to verified first radiologic recurrence (distant) and was censored at the last follow-up or death. The following average normalized β-value cutoffs were used for the CD3E, ACTA2, and PDN CpG probes to stratify patients into high- and low relapse risk groups: β- values ≤ 0.1. The following average ΔCT cutoffs were used for the CD3E, ACTA2, and PDNP QMSP biomarker assays to stratify patients into high- and low relapse risk groups: ΔCTCD3E ≥ 19.5, ΔTCPDN < 13.5. Survival analysis was performed using the Kaplan–Meier method with the Stata/IC 14.2 StataCorp) software. Significance was evaluated by log-rank test of equality. Cox proportional hazards regression analysis was used to assess the impact of MethCORR risk groups, TNM stage, and MSI status on RFS. The proportional hazard assumption was tested by a global test of the Schoenfeld residuals.

**Code availability.** R codes for calculation of MCSs and RNA profiles are available upon request.
Acknowledgements

This research is supported by grants from the European Commission FP7 project SYSCOL (U7E-SYSCOL-258236), the Novo Nordisk Foundation (NNF16OOC023182), the Danish National Advanced Technology Foundation (056-2010-1), the John and Birthe Meyer Foundation, the Danish Council for Independent Research (Medical Sciences) (DFF – 0602-02128B, DFF – 4183-00619, DFF – 7016-00332R), the Danish Council for Strategic Research (1309-00069), the Danish Cancer Society (R40-A1965_11_S2, R56-A3110-12-S2, R107-A7035, R113-A8520), the National Cancer Institute of the National Institutes of Health (R01 CA207467), the Aage and Johanne Louis-Hansen’s Foundation (17-2-0457), Dansk Kraftforskningsfond (DKF-2017-26 - (26)), the Knud and Edith Eriksen’s Memorial Foundation, the Neye Foundation, and the Manufacturer Einar Willumsen’s Memorial Foundation (6000073). The Danish Cancer Biobank is acknowledged for biological material. We thank P. Celis, L. Nielsen, L. Kjeldsen, B. Devantie, B. Trolle, S. Moran, D. Garcia, and C. Arribas for their technical support. The results published here are in part based upon data generated by the TCGA Research Network [https://cancergenome.nih.gov/].

Author contributions

T.B.M., C.L.A., and J.B.B. designed the experiments. T.B.M., M.H.R., J.S., H.O., S.S.A., J.G., A.M.G., M.C.M., A.H.M., S.L., E.T.D., M.E., C.L.A., and J.B.B. performed the experiments and included patients. T.B.M., M.H.R., C.L.A., and J.B.B. analyzed and interpreted the data. T.B.M., C.L.A., and J.B.B. drafted the manuscript. All authors reviewed and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41467-020-16000-6.

Correspondence and requests for materials should be addressed to C.L.A. or J.B.B.

Peer review information Nature Communications thanks the anonymous reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission information is available at http://www.nature.com/reprints

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

© The Author(s) 2020