Distinct Cytosine Modification Profiles Define Epithelial-to-Mesenchymal Cell-State Transitions

Min Kyung Lee

Dartmouth College Geisel School of Medicine <https://orcid.org/0000-0001-9558-4819>

Meredith S. Brown

Dartmouth College Geisel School of Medicine

Owen Wilkins

Dartmouth-Hitchcock Norris Cotton Cancer Center

Diwakar R. Pattabiraman (✉ Diwakar.r.pattabiraman@dartmouth.edu)

Dartmouth College Geisel School of Medicine

Brock C. Christensen

Dartmouth College Geisel School of Medicine

Research Article

Keywords: DNA methylation, hydroxymethylation, cytosine modifications, epithelial to mesenchymal transition

Posted Date: August 25th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-823460/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Version of Record: A version of this preprint was published at Epigenomics on April 6th, 2022. See the published version at <https://doi.org/10.2217/epi-2022-0023>.

Abstract

Background: Epithelial-to-mesenchymal transition (EMT) is an early step in the invasion-metastasis cascade, involving progression through a number of cell intermediate states. Due to challenges with isolating intermediate cell states in EMT, genome-wide cytosine modification mechanisms that define transition through EMT states are not completely understood. We measured multiple DNA cytosine methylation modification marks, complemented with chromatin accessibility and gene expression, across clonal populations residing in specific EMT states.

Results: Clones exhibiting intermediate EMT phenotypes demonstrated increased global 5-hydroxymethylcytosine (5hmC), decreased 5-methylcytosine (5mC), and more accessible chromatin. Open chromatin regions containing CpG loci with abundant 5hmC were enriched in motifs of key EMT transcription factors, ZEB1 and Snail. The magnitude of altered gene expression in intermediate cell states was higher for genes both with increased gene promoter 5hmC and differentially accessible chromatin compared with genes that exhibited differentially accessible chromatin alone, implicating functional epigenetic duality in regulation of EMT.

Conclusion: Our results indicate the importance of both distinct and shared epigenetic profiles at the cytosine and chromatin level associated with EMT processes that contribute to gene regulation and which may be targeted to prevent the progression of EMT.

Background

Epithelial-to-mesenchymal transition (EMT) is an early step in the invasion-metastasis cascade, involving progression through a number of cellular states. It is a process by which epithelial cells lose specific properties such as apical-basal polarity, detach from the basement membrane to gain mesenchymal properties such as front-back polarity and motility (1). Rather than being a binary conversion from an epithelial to a mesenchymal state, the EMT encompasses a step-wise progression to a mesenchymal cell state whereby the cells could display intermediate/hybrid phenotypes of both epithelial and mesenchymal cells (2, 3). As metastasis is responsible for the majority of deaths in cancer patients (4, 5), it is critical to understand the molecular underpinnings of EMT.

Cells that reside in an intermediate state display more plasticity than the cells on either ends of the EMT spectrum (3, 6–8). In addition to increased plasticity, intermediate cells have been shown to harbor stem cell characteristics such as self-renewal and increased expression of pluripotent genes (9–11). Although it is evident that there are intermediate phases when transitioning from epithelial to mesenchymal states (12–14), experimental isolation of these specific states has proven challenging. Consequently, the molecular and functional characteristics and of the intermediate states and their contribution to metastasis are poorly understood.

DNA methylation is a well-studied epigenetic mark, mostly known for its role in regulating gene expression. Methylation of cytosines (5-methylcytosine/5mC) can occur in the context of Cytosine-

phosphate-Guanine (CpG) dinucleotides and the reaction is catalyzed by DNA methyltransferase enzymes (DNMTs). Ten eleven translocation (TET) enzymes can oxidize methylcytosine to form 5-hydroxymethylcytosine (5hmC), then 5-formylcytosine (5fC), and finally 5-carboxylcytosine (5caC)(15). Oxidized cytosines can then be deaminated by thymine DNA glycosylase and undergo base excision repair, replaced with cytosine. While around 80% of mammalian CpG dinucleotides are estimated to be methylated (16), hydroxymethylation accounts for a relatively modest proportion of overall cytosine modification and varies greatly with tissue type (17, 18). Although 5hmC levels are low in relation to 5mC in human tissues, it is most highly enriched in brain and breast tissues, relative to other tissue types (19). While a number of studies have shown the importance of DNA methylation in EMT, these studies used traditional bisulfite treatment to measure 5mC, which does not resolve 5hmC (20–25). 5hmC can be estimated from comparing oxidized-bisulfite treatment to bisulfite treated DNA(26), as traditional bisulfite treatment does not distinguish 5mC from 5hmC. In comparison to general repression of transcription from 5mC, 5hmC is positively-associated with transcriptional activity and gene expression (27, 28). If the association is a consequence of passive dilution of 5mC via DNA demethylation, or due to functional actions of 5hmC is yet unclear and is likely context dependent. However, growing evidence suggests 5hmC contributes directly to gene regulation in several specific contexts, aside from its role in DNA demethylation. At the chromatin level, 5hmC has been shown to increase DNA flexibility and mechanical stability, and nucleosome accessibility (29). Transcription factors and their binding sites have been associated with being colocalized with TET and 5hmC (30–33), which provides possible 5hmC mechanism of gene expression regulation through transcription factor recruitment (34).

Although decreased global 5hmC is consistently observed in cancer (35–38), few studies have measured cancer-associated 5hmC changes at nucleotide-resolution. 5hmC maintenance has been associated with protecting against CpG island hypermethylation, which commonly occurs in cancer (39–43). Measures of breast tissue nucleotide-specific 5hmC revealed enrichment within breast-specific enhancers and transcriptionally active chromatin (44). In ER/PR-negative breast cancer particularly, loss of 5hmC is associated with poor prognosis (38). As DNA methylation alterations occur early in breast carcinogenesis and are related with prognosis (45, 46), a better understanding of 5hmC in breast cancer and EMT is needed.

In concert with DNA methylation, chromatin accessibility regulates transcription and cell reprogramming (47). Interactions with different nuclear macromolecules such as transcription factors and histone modifications shape the topology of chromatin (47). Specific chromatin accessibility states have been implicated in regulating EMT. Putative enhancers, defined by promoter-distal H3K27ac and H3K4me1 histone modifications have been shown to recruit key EMT transcription factors such as NF- κ B and AP-1 in epithelial cells in comparison to TGF- β -treated mesenchymal cells (48–50). In addition, motifs of key EMT transcription factors (AP-1, ETS) were enriched in accessible chromatin regions of TGF- β transformed mammary epithelial cells (51). Although transcription factors influencing EMT and metastasis-associated chromatin accessibility have been identified (52–55), gaps in knowledge of chromatin accessibility changes in non-TGF- β -induced EMT cells and cells in EMT intermediate/hybrid

states still remain due to challenges in isolating cells in these states. Moreover, better understanding of the relationship between cytosine modifications and chromatin conformation is needed.

Here, we provide a nucleotide-resolution genome-scale map of cytosine modifications and chromatin accessibility for phenotypes spanning the EMT spectrum. We address gaps in understanding of epigenomic changes in the intermediate/hybrid states during EMT. Using a novel model derived from estrogen receptor/progesterone receptor negative (ER/PR-negative) breast cancer cells to study terminal and intermediate EMT states, we demonstrate substantial differences in the cytosine modifications profiles of cells in intermediate EMT states; particularly, increases in 5hmC enriched in key EMT transcription factor motifs. Further, we utilize novel, integrative multicomponent epigenetic analysis to show cytosine modifications coordinate with chromatin accessibility especially at promoters to regulate transcription.

Methods

Cell culture

Single cell clones, methods of which isolation and characterization are detailed in Brown et al(8) were used. To summarize briefly, six single cell clones were isolated from SUM149PT cells to represent different points of the EMT spectrum. Position on the EMT spectrum was determined by cell morphology, flow cytometry analysis of CD44 and CD104 markers, and mRNA expressions of *ZEB1/2*. Graphic representation of each clones' position on the EMT spectrum can be found in **Supplementary Fig. 1A**.

DNA methylation and hydroxymethylation

DNA conversion and methylation/hydroxymethylation profiling

DNA from each clone of similar passage numbers was extracted using DNeasy Blood and Tissue kit (Catalog ID 69504, Qiagen, Hilden, Germany). DNA was quantified with Qubit 3.0 Fluorometer (Life Technologies, Carlsbad, CA). ~2µg of DNA underwent oxidative-bisulfite conversion to measure both 5mC and 5hmC using the TrueMethyl OxBS Module (Catalog ID 0414 – 32; Nugen, Redwood City, CA). Epigenome-wide DNA methylation profiling was performed using the Infinium MethylationEPIC Bead Chips (Illumina Inc., San Diego, CA) at the Norris Cotton Cancer Center Genomics Shared Resource Core.

Quality control and processing

Raw intensity files produced from the MethylationEPIC Bead Chips were preprocessed using the *minfi* R/Bioconductor analysis pipeline (v1.34.0) annotation file version *ilm10b4.hg19*(56, 57). 695 technical probes and 33,360 SNP associated probes were excluded. Quality control was performed using *ENmix* R package. 301,580 probes that failed to meet a detection p-value of 0.00005 in > 30% of the samples and 5% of the CpGs were excluded. High number of CpGs that failed to pass the quality control may have been due to 1) oxidation further damaging the DNA on top of the bisulfite treatment and 2) signal distributions being distorted from the oxidation measurement as the quality control measures were

developed for bisulfite converted DNA. After these exclusions, 545,515 CpGs remained for analysis. The filtered data was then normalized using *preprocessFunnorm* in *minfi* to remove unwanted technical variation.

Annotations of CpGs such as genomic context or relation to CpG Island were provided in the Illumina EPIC B4 manifest and UCSC hg19 reference genome files. “Promoter”, “Intergenic”, “Intron” and “Exon” genomic contexts were defined by finding overlapping genomic regions of the CpGs and each context using the UCSC hg19 reference genome annotation. “DNase hypersensitive site” context was defined by having a record in the “DNase_Hypersensitive_NAME” in the annotation. “Gene body” transcriptional context was defined by having a “Body” in the *UCSC_RefGene_Group*. Likewise, “3’ UTR” and “5’ UTR” regions were defined by having “UTR3” and “UTR5”, respectively, in the *UCSC_RefGene_Group*. Relation to CpG Island were defined by the “Relation_to_UCSC_CpG_Island” in the Illumina EPIC annotation file. If no record of relation to the CpG island was indicated, the CpG was considered to be in the “Open Sea” region. For analysis testing enrichment of CpGs measured on the Illumina EPIC array to ATAC regions, GRCh38 annotation file from Zhou et al was used(58).

CpGs annotated to open chromatin regions were defined by their overlap with open chromatin regions from ATAC-seq data. CpGs were determined to be in enhancers if they were located in distal intergenic regions (within 10–15kbps upstream and downstream of gene) of the ATAC-seq consensus peaks. CpGs were determined to be in open promoters if they were located in promoters of the ATAC-seq consensus peaks.

5hmC estimation

5hmC beta values were estimated using the *fitOxBS* function in the *OxyBS* package(59). Instead of naive subtraction of signals from oxidative-bisulfite treated probes from bisulfite only treated probes, the *OxyBS* package uses maximum likelihood estimation of the signal intensities from the oxidative-bisulfite treated and bisulfite treated DNA from the Illumina EPIC array to determine the parameters for unmethylated, hydroxymethylated and methylated CpGs.

Analysis

Non-negative matrix factorization analyses of 5hmC, 5mC and ATAC-seq data were conducted using *nmf.mnnals* function with $k = 2$ from the *IntNMF* R package(60). Principal component analyses were performed using 5hmC and 5mC beta values using *princomp* function in R. Differential methylation and hydroxymethylated analyses were conducted using *limma* (v3.44.3) and *qvalue* (v2.20.0) R packages in R (v4.0.2)(61, 62). Differentially methylated and hydroxymethylated CpGs were identified by fitting into a linear regression model, testing for differences in beta values CpG-by-CpG in groups of clones based position on the EMT spectrum (distal vs intermediate). Linear regression models were fit by using *lmFit* and *eBayes* functions. E, EM1, M2, and P were considered as distal clones. Intermediate group was comprised of EM2, EM3, M1 clones. The differentially methylated CpGs were deemed to be significant at the q-value threshold of 0.01.

Differentially hydroxymethylated and methylated CpGs were compared to the 545,515 CpGs used in analyses to test for enrichment at specific genomic contexts using Fisher's exact test. Functional significance of these CpGs were assessed using the Genomic Regions Enrichment of Annotations Tool (GREAT)(63).

ATAC-seq and analysis

ATAC-seq and preprocessing

ATAC-seq for 2 replicates per clone was performed as described in Buenrostro et al(64). Similar passage number (+/- 1 passage) of the clones as for the DNA methylation and hydroxymethylation measurements were used. Same processing methods and detailed descriptions can be found in Brown et al(8).

Briefly, ATAC-seq data was then processed using the publicly available ENCODE ATAC-seq pipeline (<https://www.encodeproject.org/pipelines/ENCPL792NWO/>). Illumina adapter and transposase sequences were trimmed using *Cutadapt* (65) (v1.9.1) with parameters “-minimum-length 5 -e 0.1”. Trimmed reads were aligned to hg38 human genome using *Bowtie2* (66) (v2.2.6) in “-local” mode with parameters “-X 2000 -k 2”. Duplicate reads were identified and filtered from final alignments using *MarkDuplicates* (*Picard Tools*(67)). To account for insertion of adapter sequences by the transposase, alignments were converted to tagAlign files and shifted + 4 bp and - 5 bp on the + and - strands, respectively. *MACS2* (68) (v2.1.1) *callpeak* command with parameters “-shift - 75 -extsize 150 -nomodel -keep-dup all -call-summits -p 1.0E-10” were used to call peaks. The peaks were filtered against the ENCODE hg38 blacklist. The Irreproducible Discovery Rate (IDR) method was used to identify a set of reproducible peaks across biological replicates using an IDR threshold of 0.05.

ATAC-seq analysis

Principal component analyses were performed using variance stabilizing transformed ATAC-seq counts using *princomp* function in R. Low level regions were filtered out using *filterByExpr* using *edgeR* (v3.30.3) (69). Testing for differential chromatin accessibility was conducted using *limma*(61). Counts of consensus peaks from ATAC-seq data were voom transformed by sample quality weights using *voomWithQualityWeights* to fit a linear model of the counts. Differentially accessible chromatin in intermediate clones compared to distal clones were identified using a linear model with random effects term for clones using *lmFit* and *eBayes* functions. Consensus correlation (correlation = 0.685) between biological replicates per clone were calculated using *duplicateCorrelation*. Consensus peaks were categorized as differentially accessible under the significance threshold of adjusted p-value < 0.05. Differentially accessible regions' associated genes and genomic contexts were annotated using *TxDb.Hsapiens.UCSC.hg38.knownGene* R annotation file package and the *annotatePeak* function in *ChIPseeker* (v1.24.0) (70, 71). Enriched biological pathways associated with the differentially accessible regions were identified using the *ReactomePA* (v1.32.0) (72).

We tested for over-representation of TF binding site motifs of high 5hmC CpG containing consensus ATAC peaks compared to all ATAC peaks. We scanned these peaks for TF motif occurrences using R-

package *motifmatchr*(73). Position frequency matrices for human TF motifs used as input to *motifmatchr* were downloaded using R-packages *JASPAR2020*(74) and *TFBSTools*(75). Over-represented TF motifs in each peak set were identified through hypergeometric testing using the *phyper* R function, with all peaks identified in that clone used as the background set. TF motifs with an FDR-adjusted hypergeometric *P*-value < 0.05 were deemed as over-represented.

Enrichment of differentially accessible regions in differentially methylated and hydroxymethylated CpGs were tested using Mantel-Haenszel test, adjusting for genomic regions related to CpG islands.

RNA-seq analysis

RNA extraction and preprocessing

RNA was collected using Qiagen RNeasy plus kit (Catalog ID: 74034, Qiagen, Hilden, Germany) and quantified using a NanoDrop (Thermo Fisher Scientific - ND-2000-US-CAN). Same processing methods and detailed descriptions can be found in Brown et al(8).

To summarize, raw single-end RNA-seq data were trimmed of polyA sequences and low quality bases using *Cutadapt* (v2.4)(65). Reads were aligned to human genome hg38 using *STAR* (v 2.7.2b)(76) with parameters “--outSAMattributes NH HI AS NM MD --outFilterMultimapNmax 10 --outFilterMismatchNmax 999 --outFilterMismatchNoverReadLmax 0.04 --alignIntronMin 20 --alignIntronMax 1000000 --alignMatesGapMax 1000000 --alignSJoverhangMin 8 --alignSJDBoverhangMin 1”. Quality of alignments was assessed using *CollectRNASeqMetrics* (*Picard Tools*)(67) and duplicate reads were identified (but retained) with *MarkDuplicates* (*Picard Tools*). Gene-level abundance estimates were generated using *RSEM* (v1.3.2)(77) using the *rsem-calculate-expression* command with the parameters “--strandedness reverse --fragment-length-mean 313 --fragment-length-sd 91”.

Analysis

Differential gene expression analysis was conducted using *limma* (v1.30.1) R package(78). Low level regions were filtered out using *filterByExpr* using *edgeR* (v3.30.3) (69). Raw counts of RNA-seq data were voom transformed by sample quality weights using *voomWithQualityWeights* to fit a linear model of the counts. Differentially expressed genes in intermediate clones to distal clones were identified using a linear model with random effects term for clones using *lmFit* and *eBayes* functions. Consensus correlation (correlation = 0.549) between biological replicates per clone were calculated using *duplicateCorrelation*. Genes were categorized as differentially expressed under the significance threshold of adjusted p-value < 0.05. Enriched biological pathways associated with the differentially expressed genes were identified using the ReactomePA (v1.32.0) R package(72).

Results

Global distribution of hydroxymethylation and methylation in EMT clones

We utilized a previously derived model of six single-cell clones from SUM149PT, a heterogeneous ER-/PR-inflammatory breast cancer line, that represent phenotypes across the EMT spectrum. The EMT state of each clone was determined by cell morphology, flow cytometry (CD44 and CD104), and immunofluorescence staining for Vimentin/E-cadherin, as well gene expression of canonical EMT markers (*Snai1*, *ZEB1*, *CDH1*, *VIM*, and others), detailed in previous work (8). These clones were ranked as epithelial (E), three distinct intermediates (EM1, EM2, and EM3), two unique mesenchymal-like clones (M1 and M2), and compared here with the parental cell line (P). The relative position of each clone on the EMT spectrum and their representative bright field images is shown in **Supplementary Fig. 1A**. Bright field images indicate more mesenchymal clones (M1, M2) to be more elongated than the more epithelial clones (E, EM1).

We first measured genome-scale cytosine-specific DNA methylation (5mC) and hydroxymethylation (5hmC) levels, using the Illumina EPIC methylation array. As expected (18), a relatively small subset of measured CpGs were hydroxymethylated, with average 5hmC beta values much smaller than that of 5mC across all clones (**Supplementary Fig. 1B, 1C**). Average 5hmC beta values and 5mC beta values were negatively correlated, at marginal significance ($R = -0.72$, $p = 0.071$) with increased global 5hmC and decreased global 5mC abundance in intermediate clones (EM2, EM3, M1; **Supplementary Fig. 1D**). Both 5mC and 5hmC decrease within 1500–2000 bps of transcription start sites (TSS), albeit only slightly for 5hmC (**Supplementary Fig. 1E**).

Genome-wide DNA cytosine modification profiles in EMT clones

To compare genome-scale similarity of DNA methylation profiles among all clones, we compared the 5hmC and 5mC beta values using principal component analysis (PCA). PCA results indicated that 5hmC and 5mC beta values clustered into four distinct groups: intermediate clones 5hmC, intermediate clones 5mC, distal clones 5hmC, and distal clones 5mC (Fig. 1A). EM2, EM3, M1 were defined as intermediate clones and E, EM1, M2, P were defined as distal clones in downstream analyses. Integrative non-negative matrix factorization analyses of 5mC, 5hmC, and chromatin accessibility profiles did not indicate same clustering potentially due to high levels of variance driven by mesenchymal clones, particularly in M2 clone (**Supplementary Fig. 2A, 2B**). Distinct grouping of clones into intermediate and distal was supported by unsupervised hierarchical clustering of the top 5% most variable CpGs (27,276 CpGs) which were chosen based on distribution of variances across CpGs (**Supplementary Fig. 3A, 3B**). Unsupervised clustering identified highly distinct intermediate and distal clone clusters (Fig. 1B, 1C) and highlighted the greater relative abundance of 5hmC in intermediate clones compared to distal clones (**Supplementary Fig. 1C**) at the CpG-specific level.

To determine if overall 5hmC and 5mC abundance was related to expression of cytosine modifying enzymes (DNMTs and TETs), we leveraged RNA-seq to test the correlation of average methylation and gene expression levels. Only *TET1* gene expression was significantly positively correlated with global average 5hmC beta values ($R = 0.86$, $p = 0.024$), and none were correlated with 5mC (**Supplementary Fig. 4A, 4B**). 5hmC and 5mC beta values of DNMT and TET CpGs with unsupervised clustering did not identify extensive variation in cytosine states at cytosine modification enzyme genes (**Supplementary Fig. 2C**). However, a small subset of CpGs (n CpGs = 18 of 241 total), located within TETs (*TET1* = 33%, *TET2* = 28%, *TET3* = 17%), exhibited higher 5hmC in intermediate clones (**Supplementary Fig. 2C**).

Together, these findings suggest there are variable patterns of genome-wide 5hmC and 5mC based on clonal EMT status.

Characterization of high 5hmC CpG loci

To better understand 5hmC at the single nucleotide level, we identified CpGs with the highest 1% 5hmC beta values for each clone (28,750 unique CpGs, Fig. 2A). We termed CpGs with top 1% 5hmC beta values for each clone 'high 5hmC CpGs'. Unsupervised hierarchical clustering of high 5hmC CpGs separated intermediate from distal clones (Fig. 2B). Each distal clone contributed its own unique set of CpGs with high 5hmC (Unique CpGs for E: 4,795 CpGs, EM1: 4,672 CpGs, M2: 4,875 CpGs, P: 4,687 CpGs), while the intermediate clones had similar CpGs with more abundant 5hmC (Unique CpGs for EM2: 959 CpGs, EM3: 1,224 CpGs, M1: 994 CpGs, Fig. 2B, 2C). In the intermediate clones, the genomic distribution of high 5hmC CpGs was significantly enriched for open chromatin regions (as defined by ATAC-seq), enhancers, promoters, 5' UTR, and transcriptional start sites (TSSs) (Fig. 2D, **Supplementary Table 1**). In addition, unlike the distal clones, high 5hmC CpGs in the intermediate clones were enriched in CpG islands (Fig. 2E, **Supplementary Table 2**). Patterns of genomic feature enrichment for the high 5hmC CpG-loci from distal clones were distinct from those of intermediate clones, showing significant enrichment only at enhancer regions. Together, these findings suggest intermediate clones exist in a less repressive epigenetic state compared to distal clones.

To explore potential functional differences in gene regulation between intermediate and distal clones, we identified regions of accessible chromatin using ATAC-seq and conducted transcription factor motif enrichment analysis of accessible chromatin peaks containing at least one high 5hmC CpG each for intermediate clones (8,975 CpGs in 2,616 chromatin accessible peaks) and distal clones (21,022 CpGs in 3,572 chromatin accessible peaks). Motif enrichment analysis found 409 transcription factors (TF) significantly associated with high 5hmC chromatin accessible peaks in intermediate clones compared to 29 TFs in distal clones under the $FDR < 0.05$ threshold (Fig. 2F). Interestingly, in both the intermediate and distal clones, motifs for key EMT transcription factors (ZEB1 and Snail) were enriched among open chromatin regions with high 5hmC CpG, implicating 5hmC in EMT process-associated gene regulation. While not known specifically to play roles in EMT, other TF motifs, particularly motifs of GATA2 and SPI1, were also found to be in open chromatin regions with high 5hmC CpGs.

Pathway analysis using a cis-regulatory region based approach (Genomic Regions Enrichment of Annotations Tool, GREAT(63)) identified that, while high 5hmC CpGs in intermediate clones and distal clones were distinct, some biological pathways associated with each set of high 5hmC CpGs in distal and intermediate clones were similar (**Supplementary Table 3, 4**). The high 5hmC CpGs were highly associated with pattern specification processes (distal clones fold enrichment (FE) = 1.37; intermediate clones FE = 1.53) among other morphogenesis and developmental-related processes. In contrast, GREAT analyses identified separate molecular function (MF) enrichment sets associated with the intermediate and distal 5hmC loci. In distal clones, high 5hmC CpGs were associated with ion channels such as gated channel activity (FE = 1.40) and substrate-specific channel activity (FE = 1.32). In intermediate clones, high 5hmC CpGs were associated with RNA polymerase II transcription factor binding (FE = 1.41). High 5hmC CpGs in intermediate clones' association with RNA polymerase II transcription factor binding further supports the large number of transcription factor motifs found in open chromatin regions with high 5hmC CpGs.

We then tested the relationship between 5hmC beta values and gene expression levels for the 500 most variably expressed genes with any high 5hmC CpGs, to test the hypothesis that 5hmC was associated with positive transcriptional activity as previous literature have indicated(31, 79, 80). The correlation between 5hmC and their respective gene expression levels were variable. While 54 unique genes (with 77 unique CpGs) had strong correlations in both directions, a majority of the genes did not (**Supplementary Fig. 5A, 5B**). Compared with all high 5hmC CpGs, high 5hmC CpGs significantly correlated with gene expression were enriched in 5' UTR (OR = 3.19, 95% CI = 1.90–5.21), 1st exons (OR = 2.59, 95% CI = 1.34–4.68), and promoter regions within open chromatin (OR = 2.34, 95% CI = 1.43–3.79). High 5hmC CpGs were generally strongly correlated with other high 5hmC CpGs rather than the genes they were located in (**Supplementary Fig. 5C**). Together, these findings suggest promoter-associated 5hmC in accessible chromatin regions may be the most important for transcriptional regulation in comparison to those of other genomic contexts.

Cytosine modifications in key epithelial-to-mesenchymal genes

We next used a candidate gene approach to investigate if EMT-associated 5hmC and 5mC loci distinguished intermediate from distal clones. We performed unsupervised clustering on beta-values of 439 CpGs annotated to epithelial genes (*CDH1*, *CLDN1*, *EPCAM*, *ITGAB4*, *KRT8* and *OCLN*), mesenchymal genes (*CDH2*, *FN1*, *ITGB1*, *MMP19*, *MMP2*, and *VIM*), and transcription factors (*SNAI1*, *SNAI2*, *TWIST1*, *ZEB1* and *ZEB2*). Intermediate clones clustered separately from distal clones for both 5mC- and 5hmC-associated genes, and a subset of CpGs annotated predominantly to epithelial genes (*OCLN*, *CDH1*, *KRT8*, *EPCAM*) had high 5hmC among intermediate clones in cluster #4 (Fig. 3A, 3B) many of which tracked to promoter regions (Fig. 3C). Expression of epithelial genes were higher in more epithelial clones as expected, whereas the more mesenchymal clones had lower expression (Fig. 3D). Together, it suggests potential role of 5hmC in regulating gene expression of epithelial genes during the EMT process.

Differential methylation and hydroxymethylation in intermediate clones

Next, we conducted an epigenome wide association study (EWAS) comparing cytosine modifications at the nucleotide level to identify differential cytosine modifications between intermediate and distal clones. Overall, we identified 17,862 significantly differentially hydroxymethylated CpGs (dhmCpG, FDR < 0.01), between distal and intermediate clones, almost all of which had increased in 5hmC in the intermediate clones (Fig. 4A, **Supplementary Table 5**), including EMT associated genes such as *SNAI1* and *TWIST1*. There were 7,903 significantly differentially methylated CpGs (dmCpG, FDR < 0.01), most of which had decreased in 5mC in intermediate clones (Fig. 4B, **Supplementary Table 6**), including EMT associated cell type markers *CDH1* and *MMP19*. For downstream analyses, dhmCpGs were subset for only CpGs increasing in 5hmC. dmCpGs were subset for only CpGs decreasing in 5mC. Among CpGs with increased 5hmC and decreased 5mC, only 33 CpGs overlapped (Fig. 4C). Expanding to the gene-level, 1,365 genes had both dhmCpGs and dmCpGs among intermediate clones (Fig. 4D). Genomic contexts with enrichment of dhmCpGs were generally depleted among dmCpGs (**Supplementary Fig. 6A, 6B; Supplementary Table 7, 8**). While dhmCpGs were enriched in regulatory regions (open chromatin regions, enhancers, 5'UTR, promoters, TSS1500, TSS200) and in the first exon, dmCpGs were enriched within exons and introns (**Supplementary Fig. 6A, 6B; Supplementary Table 7, 8**).

GREAT analyses revealed that dhmCpGs were associated with fatty acid-related molecular functions (MF), such as peroxisomal fatty-acyl-CoA transporter activity (FE = 19.00) and long-chain fatty acid transporter activity (FE = 7.46), as well as RNA polymerase II transcription factor-related molecular functions such as RNA polymerase II TF sequence-specific DNA binding (FE = 1.18) and RNA polymerase II regulatory region DNA binding (FE = 1.17, Fig. 4E). Similarly, dmCpGs were associated with RNA polymerase II-related molecular functions such as RNA polymerase II transcription coactivator binding (FE = 7.62) and cofactor binding (FE = 6.95, Fig. 4F). Additionally, dmCpGs were associated with metal ion transmembrane activity (FE = 1.44). Other biological processes and cellular components associated with dhmCpGs and dmCpGs can be found in **Supplementary Fig. 7**. Collectively, these results support the role of differential cytosine modifications in RNA polymerase II related regulation of transcription to influence intermediate EMT phenotype.

Furthermore, intermediate clones had 11,237 differentially accessible chromatin regions (DACRs; 4,742 less accessible, 6,495 more accessible) compared to distal clones under the adjusted p-value < 0.05 (Fig. 4G). Interestingly, the differentially accessible regions were associated with processes expected to change during EMT such as extracellular matrix organization/degradation and RHO/RAC1 GTPase cycle which is important in cell cytoskeleton organization and migration (81) (**Supplementary Fig. 8**). The overall more accessible chromatin regions along with increased 5hmC and decreased 5mC in intermediate clones further support the less repressive epigenetic state in intermediate clones.

To determine gene expression changes that may be connected to the identified epigenetic changes, we next conducted differential expression analysis between intermediate and distal clones identifying 442

differentially expressed genes (128 genes with decreased expression and 314 genes with increased expression) in the intermediate clones at the adjusted p-value < 0.05 threshold (**Supplementary Fig. 9A, Supplementary Table 9**). Gene set pathways analysis revealed differentially expressed genes similar to some pathways related to DACRs, such as extracellular matrix organization and assembly of collagen fibrils and other multimeric structures (**Supplementary Fig. 9B**).

To identify epigenetic changes that many occurring at the same genes or genomic coordinates, we extracted CpGs located within DACRs. Among the 93,999 analyzed CpGs found in the consensus ATAC-seq peaks, 8,134 CpGs tracked to DACRs (Fig. 5A), which were not enriched in neither dhmCpGs (OR = 1.07, 95% CI = 0.94–1.22) nor dmCpGs (OR = 1.04, 95% CI = 0.86–1.24). Although differentially modified CpGs were not enriched specifically within genomic locations of DACRs, there were genes that contained both the differentially modified CpGs and DACRs (Fig. 5B).

Given that a total of 3,133 genes had at least two differential epigenetic marks in the intermediate clones, and that promoter specific epigenetic marks appeared to influence gene expression, we compared the magnitude of change in gene expression based on types of promoter epigenetic marks. The absolute log fold change in gene expression was greater if there was at least one differential promoter epigenetic mark compared to no differential epigenetic mark (Fig. 5C). Interestingly, genes with DACRs and increased 5hmC in promoters had greater gene expression changes than those with just DACRs in promoters ($p = 0.009$). However, the same pattern was not found in genes with DACRs and decreased 5mC in promoters compared to just DACRs in promoters ($p = 0.8$). Together these results suggest multiple epigenetic events at promoters may work together regulate transcription.

Discussion

Widely used standard bisulfite conversion used to study DNA methylation is unable to distinguish between 5mC and 5hmC. Using a tandem oxidative-bisulfite treatment approach, we measured both cytosine modifications to understand their unique distribution across distal and intermediate EMT states. The majority of previous studies measuring 5hmC have been limited to global 5hmC levels in tissues of heterogeneous cell types including tumors, where extremely low levels of 5hmC were observed (37, 82, 83). Here, identifying differences in cell state-specific, nucleotide-specific 5hmC is a strength of our approach. The intermediate clones in our EMT model system suggests that genome-wide patterns of hydroxymethylation are associated with specific EMT phenotypes, suggesting a potential role of 5hmC in mediating EMT related processes. Moreover, through multi-component approach of epigenome profiling, we show that EMT phenotypes are underscored by substantial epigenetic differences.

Previous work establishing this model system has demonstrated that the intermediate clones represent a population of tumor cells with high tumor-initiating potential(8). We show high 5hmC CpGs association with motifs of key EMT transcription factors which may indicate recruitment of various transcription factors by 5hmC may be a potential mechanism regulating the intermediate clones' high tumor-initial potential. Our results suggest that targeting increases in 5hmC in intermediate cells may impede the

maintenance of this state and/or force lineage commitment, effects that could lead to altered metastatic propensity.

Prior literature has already indicated that DNA methylation states change during TGF- β induced EMT(84). Similarly, our natural (non-TGF- β induced) EMT model suggests that DNA cytosine modifications exhibit altered genome-wide patterns during the EMT process. Our results indicate that these altered patterns may regulate the existence of cells in various EMT states, thereby enabling tumor heterogeneity. Alterations in cytosine modifications and chromatin accessibility towards a less repressive state suggests that the multi-level epigenome is essential in regulating the dynamics of EMT.

Lastly, our study highlights the importance of multicomponent measures of epigenetic states. While the alterations in chromatin accessibility and cytosine modifications may occur in the similar genomic contexts, they do not occur at the exact same genomic location. Our results indicate promoter specific cytosine modification and chromatin accessibility patterns may be essential in regulating gene expression. However, the array-based approach may not have revealed CpG loci in relevant accessible chromatin, a limitation that may be overcome with a whole genome bisulfite, oxidative-bisulfite sequencing approach. It highlights the complex epigenetic landscape that is required in the EMT process.

Conclusion

Our study addresses current gaps that exist in understanding of specific cytosine modifications (5mC and 5hmC) roles in EMT and their associations with other epigenetic changes. Clones exhibiting intermediate EMT phenotypes had distinct, more open epigenetic states with increased 5hmC, decreased 5mC and more accessible chromatin compared to clones exhibiting more distal EMT phenotypes. Open chromatin regions containing CpG loci with abundant 5hmC enriched in motifs of key EMT transcription factors, ZEB1 and Snail, indicate likelihood of multi-component epigenetic regulation during EMT. Epigenetic profiles at the cytosine and chromatin level associated with EMT processes that contribute to gene regulation may be targeted to prevent the progression of EMT.

Abbreviations

5hmC: 5-hydroxymethylcytosine

5mC: 5-methylcytosine

CpG: Cytosine-phosphate-Guanine dinucleotides

DACR: Differentially accessible chromatin regions

DNMT: DNA methyltransferase

dhmCpG: Differentially hydroxymethylated CpGs

dmCpG: Differentially methylated CpGs

EMT: Epithelial-to-mesenchymal transition

ER: Estrogen receptor

EWAS: Epigenome wide association study

FE: Fold enrichment

GREAT: Genomic regions enrichment of annotations tool

OR: Odds ratio

MF: Molecular function

PCA: Principal component analysis

PR: Progesterone receptor

TET: Ten eleven translocation

TF: Transcription factor

TSS: Transcription start site

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The datasets used and/or analysed during the current study are available in GEO repository, under accession number GSE181036 and GSE172613.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by a Prouty Pilot Grant from Friends of the Norris Cotton Cancer Center, funding from The Elmer R. Pfefferkorn & Allan U. Munck Education and Research Fund at the Geisel School of Medicine at Dartmouth and an NCI Cancer Center Support Grant (5P30CA023108-40). This work was supported by the National Institutes of Health 5R00CA201574-05 (to D.R.P.), R01CA216265 (to B.C.C), and R01CA253976 (to B.C.C). O.M.W is supported by Cancer Center Core grant (National Cancer Institute 5P30CA023108-42) and National Institutes of Health 5P20GM130454-02.

Authors' contributions

MKL and MSB carried out the experiments. MSB and DRP conceived the original epithelial-to-mesenchymal transition experimental design. MKL conducted analyses with the help from OMW and BCC. BCC supervised the project. All authors discussed the results and contributed to the final version of the manuscript.

Acknowledgements

Illumina EPIC arrays, ATAC-seq, and RNA-seq experiments were all carried out at Geisel School of Medicine in the Genomics Shared Resource, which was established by equipment grants from the National Institutes of Health and National Science Foundation and is supported in part by a Cancer Center Core Grant (P30CA023108) from the National Cancer Institute.

References

1. Nieto,M.A., Huang,R.Y., Jackson,R.A. and Thiery,J.P. (2016) EMT: 2016. *Cell*, **166**, 21–45.
2. Jolly,M.K., Boareto,M., Huang,B., Jia,D., Lu,M., Onuchic,J.N., Levine,H. and Ben-Jacob,E. (2015) Implications of the hybrid epithelial/mesenchymal phenotype in metastasis. *Front. Oncol.*, **5**, 155.
3. Pastushenko,I., Brisebarre,A., Sifrim,A., Fioramonti,M., Revenco,T., Boumahdi,S., Van Keymeulen,A., Brown,D., Moers,V., Lemaire,S., *et al.* (2018) Identification of the tumour transition states occurring during EMT. *Nature*, **556**, 463–468.
4. Chaffer,C.L. and Weinberg,R.A. (2011) A perspective on cancer cell metastasis. *Science (80-.)*, **331**, 1559–1564.
5. Hanna,D., Rogers,M.S. and Straume,O. (2019) Are 90% of deaths from cancer caused by metastases? *Cancer Med.*, **8**, 5574–5576.

6. Nieto,M.A. (2013) Epithelial plasticity: A common theme in embryonic and cancer cells. *Science (80-)*, **342**.
7. Jolly,M.K., Somarelli,J.A., Sheth,M., Biddle,A., Tripathi,S.C., Armstrong,A.J., Hanash,S.M., Bapat,S.A., Rangarajan,A. and Levine,H. (2019) Hybrid epithelial/mesenchymal phenotypes promote metastasis and therapy resistance across carcinomas. *Pharmacol. Ther.*, **194**, 161–184.
8. Brown,M.S., Abdollahi,B., Wilkins,O.M., Chakraborty,P., Ognjenovic,N.B., Muller,K.E., Kumar Jolly,M., Hassanpour,S. and Pattabiraman,D.R. (2021) Dynamic plasticity within the EMT spectrum, rather than static mesenchymal traits, drives tumor heterogeneity and metastatic progression of breast cancers. *bioRxiv*, 10.1101/2021.03.17.434993.
9. Grosse-Wilde,A., Fouquier D'hroul,A., Mcintosh,E., Ertaylan,G., Skupin,A., Kuestner,R.E., Del Sol,A., Walters,K.-A. and Huang,S. (2015) Stemness of the hybrid Epithelial/ Mesenchymal State in Breast Cancer and Its Association with Poor Survival. *PLoS One*, **10**.
10. Bierie,B., Pierce,S.E., Kroger,C., Stover,D.G., Pattabiraman,D.R., Thiru,P., Donaher,J.L., Reinhardt,F., Chaffer,C.L., Keckesova,Z., *et al.* (2017) Integrin-4 identifies cancer stem cell-enriched populations of partially mesenchymal carcinoma cells. *Proc. Natl. Acad. Sci. U. S. A.*, **114**, E2337–E2346.
11. Ognjenovic,N.B., Bagheri,M., Mohamed,G.A., Xu,K., Chen,Y., Saleem,M.A.M., Brown,M.S., Nagaraj,S.H., Muller,K.E., Gerber,S.A., *et al.* (2020) Limiting Self-Renewal of the Basal Compartment by PKA Activation Induces Differentiation and Alters the Evolution of Mammary Tumors. *Dev. Cell*, **55**, 544-557.e6.
12. Goetz,H., Melendez-Alvarez,J.R., Chen,L. and Tian,X.J. (2020) A plausible accelerating function of intermediate states in cancer metastasis. *PLoS Comput. Biol.*, **16**, e1007682.
13. Hong,T., Watanabe,K., Ha Ta,C., Villarreal-Ponce,A., Nie,Q. and Dai,X. (2015) An Ovol2-Zeb1 Mutual Inhibitory Circuit Governs Bidirectional and Multi-step Transition between Epithelial and Mesenchymal States X (2015) An Ovol2-Zeb1 Mutual Inhibitory Circuit Governs Bidirectional and Multi-step Transition between Epithelial and Mesenchymal States. *PLoS Comput Biol*, **11**, 1004569.
14. Huang,R.Y.J., Wong,M.K., Tan,T.Z., Kuay,K.T., C Ng,A.H., Chung,V.Y., Chu,Y.S., Matsumura,N., Lai,H.C., Lee,Y.F., *et al.* (2013) An EMT spectrum defines an anoikis-resistant and spheroidogenic intermediate mesenchymal state that is sensitive to e-cadherin restoration by a src-kinase inhibitor, saracatinib (AZD0530). *Cell Death Dis.*, **4**, e915–e915.
15. Tahiliani,M., Koh,K.P., Shen,Y., Pastor,W.A., Bandukwala,H., Brudno,Y., Agarwal,S., Iyer,L.M., Liu,D.R., Aravind,L., *et al.* (2009) Conversion of 5-Methylcytosine to 5-Hydroxymethylcytosine in Mammalian DNA by MLL Partner TET1. *Science (80-)*, **324**, 930–5.
16. Janitz,K. and Janitz,M. (2011) Assessing Epigenetic Information. In *Handbook of epigenetics: the new molecular and medical genetics*. Elsevier/Academic Press, Amsterdam; Boston, pp. 173–181.

17. Tost,J. (2010) DNA methylation: An introduction to the biology and the disease-associated changes of a promising biomarker. *Mol. Biotechnol.*, **44**, 71–81.
18. Cui,X.-L., Nie,J., Ku,J., Dougherty,U., West-Szymanski,D.C., Collin,F., Ellison,C.K., Sieh,L., Ning,Y., Deng,Z., *et al.* (2020) A human tissue map of 5-hydroxymethylcytosines exhibits tissue specificity through gene and enhancer modulation. *Nat. Commun.*, **11**, 6161.
19. Nestor,C.E., Ottaviano,R., Reddington,J., Sproul,D., Reinhardt,D., Dunican,D., Katz,E., Dixon,J.M., Harrison,D.J. and Meehan,R.R. (2012) Tissue type is a major modifier of the 5-hydroxymethylcytosine content of human genes. *Genome Res.*, **22**, 467–477.
20. Pistore,C., Giannoni,E., Colangelo,T., Rizzo,F., Magnani,E., Muccillo,L., Giurato,G., Mancini,M., Rizzo,S., Riccardi,M., *et al.* (2017) DNA methylation variations are required for epithelial-to-mesenchymal transition induced by cancer-associated fibroblasts in prostate cancer cells. *Oncogene*, **36**, 5551–5566.
21. Tahara,T., Shibata,T., Okubo,M., Ishizuka,T., Nakamura,M., Nagasaka,M., Nakagawa,Y., Ohmiya,N., Arisawa,T. and Hirata,I. (2014) DNA methylation status of Epithelial-Mesenchymal Transition (EMT) - Related genes is associated with severe clinical phenotypes in Ulcerative Colitis (UC). *PLoS One*, **9**.
22. Carmona,F.J., Davalos,V., Vidal,E., Gomez,A., Heyn,H., Hashimoto,Y., Vizoso,M., Martinez-Cardus,A., Sayols,S., Ferreira,H.J., *et al.* (2014) A comprehensive DNA methylation profile of epithelial-to-mesenchymal transition. *Cancer Res.*, **74**, 5608–5619.
23. Dumont,N., Wilson,M.B., Crawford,Y.G., Reynolds,P.A., Sigaroudinia,M. and Tlsty,T.D. (2008) Sustained induction of epithelial to mesenchymal transition activates DNA methylation of genes silenced in basal-like breast cancers. *Proc. Natl. Acad. Sci. U. S. A.*, **105**, 14867–14872.
24. Rajić,J., Dinić,S., Uskoković,A., Arambašić Jovanović,J., Tolić,A., Đorđević,M., Đorđević,M., Poznanović,G., Mihailović,M., Inic-Kanada,A., *et al.* (2020) DNA methylation of miR-200 clusters promotes epithelial to mesenchymal transition in human conjunctival epithelial cells. *Exp. Eye Res.*, **197**, 108047.
25. Choi,S.K., Pandiyan,K., Eun,J.W., Yang,X., Hong,S.H., Nam,S.W., Jones,P.A., Liang,G. and You,J.S. (2017) Epigenetic landscape change analysis during human EMT sheds light on a key EMT mediator TRIM29. *Oncotarget*, **8**, 98322–98335.
26. Booth,M.J., Ost,T.W.B., Beraldi,D., Bell,N.M., Branco,M.R., Reik,W. and Balasubramanian,S. (2013) Oxidative bisulfite sequencing of 5-methylcytosine and 5- hydroxymethylcytosine. *Nat. Protoc.*, **8**, 1841–1851.
27. Kumar,S., Chinnusamy,V. and Mohapatra,T. (2018) Epigenetics of Modified DNA Bases: 5-Methylcytosine and Beyond. *Front. Genet.*, **9**.
28. Pastor,W.A., Aravind,L. and Rao,A. (2013) TETonic shift: Biological roles of TET proteins in DNA demethylation and transcription. *Nat. Rev. Mol. Cell Biol.*, **14**, 341–356.

29. Ngo,T.T.M., Yoo,J., Dai,Q., Zhang,Q., He,C., Aksimentiev,A. and Ha,T. (2016) Effects of cytosine modifications on DNA flexibility and nucleosome mechanical stability. *Nat. Commun.*, **7**.
30. Stroud,H., Feng,S., Morey Kinney,S., Pradhan,S. and Jacobsen,S.E. (2011) 5-Hydroxymethylcytosine is associated with enhancers and gene bodies in human embryonic stem cells. *Genome Biol.*, **12**.
31. Tekpli,X., Urbanucci,A., Hashim,A., Vågbø,C.B., Lyle,R., Kringen,M.K., Staff,A.C., Dybedal,I., Mills,I.G., Klungland,A., *et al.* (2016) Changes of 5-hydroxymethylcytosine distribution during myeloid and lymphoid differentiation of CD34+ cells. *Epigenetics and Chromatin*, **9**, 21.
32. Costa,Y., Ding,J., Theunissen,T.W., Faiola,F., Hore,T.A., Shliaha,P. V., Fidalgo,M., Saunders,A., Lawrence,M., Dietmann,S., *et al.* (2013) NANOG-dependent function of TET1 and TET2 in establishment of pluripotency. *Nature*, **495**, 370–374.
33. Xiong,J., Zhang,Z., Chen,J., Huang,H., Xu,Y., Ding,X., Zheng,Y., Nishinakamura,R., Xu,G.L., Wang,H., *et al.* (2016) Cooperative Action between SALL4A and TET Proteins in Stepwise Oxidation of 5-Methylcytosine. *Mol. Cell*, **64**, 913–925.
34. He,B., Zhang,C., Zhang,X., Fan,Y., Zeng,H., Liu,J., Meng,H., Bai,D., Peng,J., Zhang,Q., *et al.* (2021) Tissue-specific 5-hydroxymethylcytosine landscape of the human genome. *Nat. Commun.* 2021 121, **12**, 1–12.
35. Jin,S.-G., Jiang,Y., Qiu,R., Rauch,T.A., Wang,Y., Schackert,G., Krex,D., Lu,Q. and Pfeifer,G.P. (2011) 5-Hydroxymethylcytosine Is Strongly Depleted in Human Cancers but Its Levels Do Not Correlate with IDH1 Mutations. *Cancer Res.*, **71**, 7360–7365.
36. Shenoy,N., Bhagat,T.D., Cheville,J., Lohse,C., Bhattacharyya,S., Tischer,A., Machha,V., Gordon-Mitchell,S., Choudhary,G., Wong,L.F., *et al.* (2019) Ascorbic acid–induced TET activation mitigates adverse hydroxymethylcytosine loss in renal cell carcinoma. *J. Clin. Invest.*, **129**, 1612–1625.
37. Yang,H., Liu,Y., Bai,F., Zhang,J.Y., Ma,S.H., Liu,J., Xu,Z.D., Zhu,H.G., Ling,Z.Q., Ye,D., *et al.* (2013) Tumor development is associated with decrease of TET gene expression and 5-methylcytosine hydroxylation. *Oncogene*, **32**, 663–669.
38. Tsai,K.W., Li,G.C., Chen,C.H., Yeh,M.H., Huang,J.S., Tseng,H.H., Fu,T.Y., Liou,H.H., Pan,H.W., Huang,S.F., *et al.* (2015) Reduction of global 5-hydroxymethylcytosine is a poor prognostic factor in breast cancer patients, especially for an ER/PR-negative subtype. *Breast Cancer Res. Treat.*, **153**, 219–234.
39. Skvortsova,K., Masle-Farquhar,E., Luu,P.L., Song,J.Z., Qu,W., Zotenko,E., Gould,C.M., Du,Q., Peters,T.J., Colino-Sanguino,Y., *et al.* (2019) DNA Hypermethylation Encroachment at CpG Island Borders in Cancer Is Predisposed by H3K4 Monomethylation Patterns. *Cancer Cell*, **35**, 297-314.e8.
40. Arechederra,M., Daian,F., Yim,A., Bazai,S.K., Richelme,S., Dono,R., Saurin,A.J., Habermann,B.H. and Maina,F. (2018) Hypermethylation of gene body CpG islands predicts high dosage of functional

oncogenes in liver cancer. *Nat. Commun.*, **9**, 1–16.

41. Gonzalez-Zulueta,M., Bender,C.M., Yang,A.S., Nguyen,T., Beart,R.W., Van Tornout,J.M. and Jones,P.A. (1995) Methylation of the 5' CpG Island of the p16/CDKN2 Tumor Suppressor Gene in Normal and Transformed Human Tissues Correlates with Gene Silencing. *Cancer Res.*, **55**.
42. Herman,J.G., Latif,F., Weng,Y., Lerman,M.I., Zbar,B., Liu,S., Samid,D., Duan,D.S.R., Gnarr,J.R., Linehan,W.M., *et al.* (1994) Silencing of the VHL tumor-suppressor gene by DNA methylation in renal carcinoma. *Proc. Natl. Acad. Sci. U. S. A.*, **91**, 9700–9704.
43. Jin,C., Lu,Y., Jelinek,J., Liang,S., Estecio,M.R.H., Barton,M.C. and Issa,J.P.J. (2014) TET1 is a maintenance DNA demethylase that prevents methylation spreading in differentiated cells. *Nucleic Acids Res.*, **42**, 6956–6971.
44. Wilkins,O.M., Johnson,K.C., Houseman,E.A., King,J.E., Marsit,C.J. and Christensen,B.C. (2020) Genome-wide characterization of cytosine-specific 5-hydroxymethylation in normal breast tissue. *Epigenetics*, **15**, 398–418.
45. Fleischer,T., Frigessi,A., Johnson,K.C., Edvardsen,H., Touleimat,N., Klajic,J., Riis,M.L., Haakensen,V.D., Wärnberg,F., Naume,B., *et al.* (2014) Genome-wide DNA methylation profiles in progression to in situ and invasive carcinoma of the breast with impact on gene transcription and prognosis. *Genome Biol.*, **15**, 435.
46. Johnson,K.C., Koestler,D.C., Fleischer,T., Chen,P., Jenson,E.G., Marotti,J.D., Onega,T., Kristensen,V.N. and Christensen,B.C. (2015) DNA methylation in ductal carcinoma in situ related with future development of invasive breast cancer. *Clin. Epigenetics*, **7**, 75.
47. Klemm,S.L., Shipony,Z. and Greenleaf,W.J. (2019) Chromatin accessibility and the regulatory epigenome. *Nat. Rev. Genet.* 2018 204, **20**, 207–220.
48. Cieřlik,M., Hoang,S.A., Baranova,N., Chodaparambil,S., Kumar,M., Allison,D.F., Xu,X., Jacob Wamsley,J., Gray,L., Jones,D.R., *et al.* (2013) Epigenetic coordination of signaling pathways during the epithelial-mesenchymal transition. *Epigenetics and Chromatin*, **6**.
49. L,B., S,M.-M., I,C., J,N., A,G.-C., S,C,H., A,E., M,M., H,B. and EF,W. (2015) Fra-1/AP-1 induces EMT in mammary epithelial cells by modulating Zeb1/2 and TGFβ expression. *Cell Death Differ.*, **22**, 336–350.
50. Hardy,K., Wu,F., Tu,W., Zafar,A., Boulding,T., McCuaig,R., Sutton,C.R., Theodoratos,A. and Rao,S. (2016) Identification of chromatin accessibility domains in human breast cancer stem cells. *Nucleus*, **7**, 50–67.
51. Arase,M., Tamura,Y., Kawasaki,N., Isogaya,K., Nakaki,R., Mizutani,A., Tsutsumi,S., Aburatani,H., Miyazono,K. and Koinuma,D. (2017) Dynamics of chromatin accessibility during TGF- β- induced EMT of Ras-transformed mammary gland epithelial cells. *Sci. Reports* 2017 71, **7**, 1–14.

52. Denny,S.K., Yang,D., Chuang,C.H., Brady,J.J., Lim,J.S.S., Grüner,B.M., Chiou,S.H., Schep,A.N., Baral,J., Hamard,C., *et al.* (2016) Nfib Promotes Metastasis through a Widespread Increase in Chromatin Accessibility. *Cell*, **166**, 328–342.
53. Xin,L., Zhao,R., Lei,J., Song,J., Yu,L., Gao,R., Ha,C., Ren,Y., Liu,X., Liu,Y., *et al.* (2019) SND1 acts upstream of SLUG to regulate the epithelial–mesenchymal transition (EMT) in SKOV3 cells. *FASEB J.*, **33**, 3795–3806.
54. Frey,W.D., Chaudhry,A., Slepicka,P.F., Ouellette,A.M., Kirberger,S.E., Pomerantz,W.C.K., Hannon,G.J. and dos Santos,C.O. (2017) BPTF Maintains Chromatin Accessibility and the Self-Renewal Capacity of Mammary Gland Stem Cells. *Stem Cell Reports*, **9**, 23–31.
55. Stowers,R.S., Shcherbina,A., Israeli,J., Gruber,J.J., Chang,J., Nam,S., Rabiee,A., Teruel,M.N., Snyder,M.P., Kundaje,A., *et al.* (2019) Matrix stiffness induces a tumorigenic phenotype in mammary epithelium through changes in chromatin accessibility. *Nat. Biomed. Eng.* 2019 312, **3**, 1009–1019.
56. Aryee,M.J., Jaffe,A.E., Corrada-Bravo,H., Ladd-Acosta,C., Feinberg,A.P., Hansen,K.D. and Irizarry,R.A. (2014) Minfi: A flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics*, **30**, 1363–1369.
57. Hansen,K. (2017) IlluminaHumanMethylationEPICanno.ilm10b4.hg19: Annotation for Illumina’s EPIC methylation arrays. *R Packag. version 0.6.0*.
58. Zhou,W., Laird,P.W. and Shen,H. (2017) Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip probes. *Nucleic Acids Res.*, **45**, e22.
59. Houseman,E.A., Johnson,K.C. and Christensen,B.C. OxyBS: estimation of 5-methylcytosine and 5-hydroxymethylcytosine from tandem-treated oxidative bisulfite and bisulfite DNA. 10.1093/bioinformatics/btw158.
60. Chalise,P. and Fridley,B.L. (2017) Integrative clustering of multi-level ‘omic data based on non-negative matrix factorization algorithm. *PLoS One*, **12**, e0176278.
61. Ritchie,M.E., Phipson,B., Wu,D., Hu,Y., Law,C.W., Shi,W. and Smyth,G.K. (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.*, **43**, e47.
62. Storey,J.D., Bass,A.J., Dabney,A., Robinson,D. and Warnes,G. (2019) qvalue: Q-value estimation for false discovery rate control.
63. McLean,C.Y., Bristor,D., Hiller,M., Clarke,S.L., Schaar,B.T., Lowe,C.B., Wenger,A.M. and Bejerano,G. (2010) GREAT improves functional interpretation of cis-regulatory regions. *Nat. Biotechnol.*, **28**, 495–501.
64. Buenrostro,J.D., Wu,B., Chang,H.Y. and Greenleaf,W.J. (2015) ATAC-seq: A method for assaying chromatin accessibility genome-wide. *Curr. Protoc. Mol. Biol.*, **2015**, 21.29.1-21.29.9.

65. Martin,M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, **17**, 10.
66. Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–359.
67. Picard Tools - by Broad Institute (2018) *Github*.
68. Zhang,Y., Liu,T., Meyer,C.A., Eeckhoute,J., Johnson,D.S., Bernstein,B.E., Nussbaum,C., Myers,R.M., Brown,M., Li,W., *et al.* (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol.*, **9**, R137.
69. Robinson,M.D., McCarthy,D.J. and Smyth,G.K. (2009) edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.
70. Maintainer,B. and Team,B. (2019) TxDb.Hsapiens.UCSC.hg38.knownGene: Annotation package for TxDb object(s). *R Packag. version 3.4.6*.
71. Yu,G., Wang,L.-G. and He,Q.-Y. (2015) ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics*, **31**, 2382–2383.
72. Yu,G. and He,Q.Y. (2016) ReactomePA: An R/Bioconductor package for reactome pathway analysis and visualization. *Mol. Biosyst.*, **12**, 477–479.
73. Schep,A. (2020) motifmatchr: Fast Motif Matching in R.
74. Fornes,O., Castro-Mondragon,J.A., Khan,A., Van Der Lee,R., Zhang,X., Richmond,P.A., Modi,B.P., Correard,S., Gheorghe,M., Baranašić,D., *et al.* (2020) JASPAR 2020: Update of the open-Access database of transcription factor binding profiles. *Nucleic Acids Res.*, **48**, D87–D92.
75. Tan,G. and Lenhard,B. (2016) TFBSTools: an R/bioconductor package for transcription factor binding site analysis. *Bioinformatics*, **32**, 1555–1556.
76. Dobin,A., Davis,C.A., Schlesinger,F., Drenkow,J., Zaleski,C., Jha,S., Batut,P., Chaisson,M. and Gingeras,T.R. (2013) Sequence analysis STAR: ultrafast universal RNA-seq aligner. **29**, 15–21.
77. Li,B. and Dewey,C.N. (2011) RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, **12**, 323.
78. Love,M.I., Huber,W. and Anders,S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, 550.
79. Rodríguez-Aguilera,J.R., Ecsedi,S., Goldsmith,C., Cros,M.-P., Domínguez-López,M., Guerrero-Celis,N., Pérez-Cabeza de Vaca,R., Chemin,I., Recillas-Targa,F., Chagoya de Sánchez,V., *et al.* (2020) Genome-wide 5-hydroxymethylcytosine (5hmC) emerges at early stage of in vitro differentiation of a putative hepatocyte progenitor. *Sci. Reports 2020 101*, **10**, 1–16.

80. Lin,I.-H., Chen,Y.-F. and Hsu,M.-T. (2017) Correlated 5-Hydroxymethylcytosine (5hmC) and Gene Expression Profiles Underpin Gene and Organ-Specific Epigenetic Regulation in Adult Mouse Brain and Liver. *PLoS One*, **12**.
81. Parri,M. and Chiarugi,P. (2010) Rac and Rho GTPases in cancer cell motility control. *Cell Commun. Signal.* 2010 81, **8**, 1–14.
82. Nestor,C.E., Ottaviano,R., Reddington,J., Sproul,D., Reinhardt,D., Dunican,D., Katz,E., Dixon,J.M., Harrison,D.J. and Meehan,R.R. (2012) Tissue type is a major modifier of the 5-hydroxymethylcytosine content of human genes. *Genome Res.*, **22**, 467–477.
83. Lian,C.G., Xu,Y., Ceol,C., Wu,F., Larson,A., Dresser,K., Xu,W., Tan,L., Hu,Y., Zhan,Q., *et al.* (2012) Loss of 5-hydroxymethylcytosine is an epigenetic hallmark of Melanoma. *Cell*, **150**, 1135–1146.
84. Gong,F., Guo,Y., Niu,Y., Jin,J., Zhang,X., Shi,X., Zhang,L., Li,R., Chen,L. and Ma,R.Z. (2017) Epigenetic silencing of TET2 and TET3 induces an EMT-like process in melanoma. *Oncotarget*, **8**, 315–328.

Figures

Figure 1

Figure 1

Distal and intermediate clones have distinct methylation and hydroxymethylation profiles. A) Results from principal component analysis of 5hmC (blue) and 5mC (yellow) beta values. Points are colored by modification. Heatmap of unsupervised clustering of the top 5% (27,276 CpGs) most variable B) 5hmC and C) 5mC CpGs. Color scale ranges from yellow (low beta value) to blue (high beta value). Horizontal tracking bars indicate clones and position on the EMT spectrum.

Figure 2

While distal clones have distinct highly hydroxymethylated CpGs, intermediate clones share highly hydroxymethylated CpGs. A) Ordered distribution of median 5hmC beta values per clone. Each colored, dashed line indicates the 5hmC beta value at the top 1% 5hmC beta value percentile. B) Heatmap of 5hmC beta values at the top 1% 5hmC CpGs per clone (28,750 CpGs). Horizontal tracking bars indicate clones and position on the EMT spectrum. Vertical tracking bars indicate genomic context annotations of

CpGs. Vertical tracking bars indicate genomic context annotations of CpGs. DHS: DNase hypersensitive sites; ReltolIsland - Relation to CpG islands; S_Shelf - South shelf; S_Shore - South shore; N_Shore - North shore; N_Shelf - North shelf. C) Proportion of unique CpGs contributed to the top 1% high 5hmC CpG set for each clone. Enrichment of top 1% 5hmC CpGs at different D) genomic contexts and E) regions in relation to CpG islands for distal clones (21,022 unique CpGs) and intermediate clones (8,975 unique CpGs). Odds ratios calculated using Fisher's exact test. Table of odds ratio, confidence intervals and p-values found in ~Supplementary Table 1 and 2. F) Transcription factors with enriched motifs in open chromatin regions with high 5hmC CpGs compared to open chromatin regions without high 5hmC CpGs separately for intermediates and distal clones. Vertical tracking bars indicate genomic context annotations of CpGs. DHS: DNase hypersensitive sites; ReltolIsland - Relation to CpG islands; S_Shelf - South shelf; S_Shore - South shore; N_Shore - North shore; N_Shelf - North shelf.

Figure 3

A

B

C

D

Figure 3

Intermediate clones have higher hydroxymethylation among epithelial genes. A) Heatmap of unsupervised clustering of 5hmC and 5mC in a set of 231 CpGs within epithelial (CDH1, CLDN1, EPCAM, ITGB4, KRT8, and OLCN), mesenchymal (CDH2, FN1, ITGB1, MMP19, MMP2, and VIM), and transcription factors (SNAI1, SNAI2, TWIST1, ZEB1, and ZEB2). Vertical tracking bars indicate DNA modification, clones, and position on the EMT spectrum. Horizontal tracking bars indicate EMT marker group

(epithelial, mesenchymal, transcription factors) and hierarchical clustering group from when height = 1.6. B) Proportions of the genes annotated to the 40 CpGs in cluster #4 of the hierarchical clustering from the heatmap. C) Enrichment of genomic contexts of the CpGs in hierarchical cluster 4. 40 CpGs in cluster 4 were compared to all 231 CpGs in EMT-related genes using Fisher's test. D) Average gene expression of epithelial genes (CDH1, EPCAM, KRT8, OCLN) for each clone. Error bars represent standard deviations.

Figure 4

Differential 5hmC CpGs are distinct from the differential 5mC CpGs. Volcano plots indicating A) 17,862 significantly differentially hydroxymethylated CpGs and B) 7,903 significantly differentially methylated CpGs under FDR Q-value of 0.01, in intermediate clones in comparison to distal clones. Red dashed lines indicate the $-\log_{10}(\text{p-value})$ at FDR q-value of 0.05 and 0.01. Venn diagrams comparing C) dhmCpGs vs dmCpGs and D) genes annotated to dhmCpGs vs genes annotated to dmCpGs. dhmCpGs were subset for only CpGs increasing in 5hmC. dmCpGs were subset for only CpGs decreasing in 5mC. Molecular functions associated with E) dhmCpGs and F) dmCpGs. G) Volcano plot of differentially accessible chromatin regions (DACRs) in intermediate clones compared to distal clones under adjusted p-value < 0.05.

Figure 5

Figure 5

Multiple epigenetic alterations in the promoters of the same gene are associated with greater change in gene expression. A) Heatmap of unsupervised clustering of 5hmC and 5mC at 1,432 CpGs located within differentially accessible regions. Horizontal tracking bars indicate each cytosine modification beta values, clones, and EMT spectrum position groups. Vertical tracking bars indicate direction of differentially accessible chromatin, CpG location in relation to CpG islands and dhmcP/dmCpG status. B) Venn

diagram of genes annotated to DACRs, dhmCpGs, and dmCpGs. C) Absolute values of log fold changes in gene expression in intermediate clones from distal clones for each differentially expressed gene categorized by epigenetic event (dhmCpG, dmCpG, DACR) in the promoters. Comparisons in means for different epigenetic events tested using Wilcoxon rank sum test.

Supplementary Files

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

- [EMTManuscriptSupplementaryTablesAug2021.xlsx](#)
- [EMTManuscriptSupplementaryFiguresAug2021.pdf](#)