Epithelial-mesenchymal plasticity through loss of CTCF motif accessibility and protein expression

Kelsey S. Johnson¹, Shaimaa Hussein², Priyanka Chakraborty³, Arvind Muruganantham¹, Sheridan Mikhail¹, Giovanni Gonzalez¹, Shuxuan Song¹, Mohit Kumar Jolly³, Michael J. Toneff⁴, Mary Lauren Benton⁵, Yin C. Lin², Joseph H. Taube^{1,6#}

¹Department of Biology, Baylor University, Waco, TX, USA

²Baylor Institute for Immunology Research, Baylor Scott & White, Dallas, TX, USA

³Centre for BioSystems Science and Engineering, Indian Institute of Science, Bangalore, India

⁴Department of Biology, Widener University, Chester, PA, USA

⁵Department of Computer Science, Baylor University, Waco, TX, USA

⁶Dan L. Duncan Cancer Center, Houston, TX, USA

#Corresponding author: Joseph_Taube@baylor.edu

Corresponding Author email:

Joseph_Taube@baylor.edu

Keywords: EMT, MET, partial EMT, ATAC-seq, chromatin accessibility, CTCF, E-cadherin

Abstract

Epithelial-mesenchymal transition (EMT) and its reversal, mesenchymal-epithelial transition (MET) drive tissue reorganization critical for early development. In carcinomas, processing through EMT, MET or partial states promotes migration, invasion, dormancy, and metastatic colonization. As a reversible process, EMT is inherently regulated at epigenetic and epigenomic

levels. To understand the epigenomic nature of reversible EMT and its partial states, we characterized chromatin accessibility dynamics, transcriptomic output, protein expression, and cellular phenotypes during stepwise reversible EMT. We found that the chromatin insulating protein machinery, including CTCF, is suppressed and re-expressed, coincident with broad alterations in chromatin accessibility, during EMT/MET and is lower in triple-negative breast cancer cell lines with EMT features. Through analysis of chromatin accessibility using ATAC-seq, we identify that early phases of EMT are characterized by enrichment for AP-1 family member binding motifs but also by diminished enrichment for CTCF binding motifs. Through loss-of-function analysis we demonstrate that suppression of CTCF alters cellular plasticity, facilitating entrance into a partial EMT state. These findings are indicative of a role of CTCF and chromatin reorganization for epithelial-mesenchymal plasticity.

Introduction

Epithelial-mesenchymal transition (EMT) is a conserved process that alters the differentiation state of a cell to drive physiological programs such as gastrulation and wound healing. During EMT, epithelial cells alter their gene expression and morphology, lose cell-cell contacts, and adopt a mesenchymal-like state (1). Because this process promotes invasion, intravasation, and resistance to anoikis, EMT is implicated in metastatic tumor cell dissemination (2-4). Recent work has contributed to a revised model of metastasis in which reversal of EMT, mesenchymal-epithelial transition (MET), is necessary for the colonization of cells which arrive at the metastatic site by means of an EMT (2,4-6). The cellular plasticity which is essential for EMT reversal is enabled by changes in genes expression and chromatin accessibility.

EMT propels cells through progressive gene expression changes and phenotypic alterations. The hallmark of EMT is the suppression of genes such as E-cadherin (*CDH1*) and epithelial cell adhesion molecule (*EPCAM*), which can be effected through networks of EMT-transcription factor proteins (EMT-TFs) such as SNAIL (*SNAI1*), SLUG (*SNAI2*), ZEB1, TWIST1, SIX1, SOX10, and FOXC2 (7-11). These transcription regulators are known to act in conjunction with epigenetic regulatory mechanisms such as post-translational modification of histone proteins (12-15) and DNA methylation (16,17). Further key regulatory mechanisms include alternative splicing, microRNAs, and protein translation (18-24). EMT is known to be initiated by microenvironmental signals such as TGF β , EGF, hypoxia, and tissue stiffness (25-28). Conversely, loss of such stimuli can trigger MET resulting in the re-emergence of the epithelial phenotype, re-establishment of cell-cell contacts, a decrease in migratory traits, and re-expression of epithelial-specific transcription factors such as ELF5, GRHL2, and OVOL1/2 (29-31).

Diverse partial- or hybrid-EMT states, which co-exhibit epithelial and mesenchymal traits, have been recently described and are reported to be highly plastic, and, when present in tumor cells, efficiently initiate tumor growth and predict poor patient outcomes (32-41). Distinct hybrid EMT states are likely driven by individual EMT-TFs (42,43), which despite considerable overlap within the EMT-regulatory network, lead to distinct gene expression outcomes (11,44). Specific transcription factor engagement can be potentiated by chromatin looping which is regulated by the CCCTC-binding factors CTCF and BORIS through their interactions with protein complexes that include cohesin and condensin (45). CTCF expression, localization, and DNA-binding activity is critical for cellular differentiation (46), yet transcription factor engagement patterns and the role of CTCF in partial EMT/MET are unclear. Therefore, we conducted phenotypic, transcriptomic, and chromatin-accessibility focused analyses of cells at progressive stages

along the epithelial-mesenchymal plasticity spectrum as well as a functional assessment of the role of CTCF in epithelial-mesenchymal plasticity.

We characterized the genome-wide dynamics of chromatin accessibility at multiple timepoints during EMT and MET including the relationship between chromatin accessibility, gene expression, and cell phenotype. By staggering the time that human mammary epithelial cells (MCF10A) experienced TGF β exposure or withdrawal, we established a series of partial EMT/MET states. Distinct phases of the repression and re-expression of the key EMT marker E-cadherin, was evident in terms of protein localization, protein expression, transcript expression, and locus accessibility. Utilizing the assay for transposase-accessible chromatin with next-generation sequencing (ATAC-seq), we report that EMT imparts a global increase in transposase-accessible chromatin while MET is marked by chromatin compaction.

Transcription factor binding site (TFBS) enrichment analysis points to dynamic engagement of CTCF throughout EMT/MET. Importantly, we find that CTCF knockdown induces partial EMT. Collectively our findings indicate that activation of EMT and MET dramatically reconfigures chromatin accessibility and CTCF is a key modulator of epithelial-mesenchymal plasticity.

Materials & Methods

Biological resources

MCF10A cells were a gift from Dr. Sendurai Mani and were cultured as previously described (47). HMLE cell lines were also a gift from Dr. Sendurai Mani and were cultured as described previously (44). Cells were plated at 10,000 cells/cm² and passaged every other day to maintain consistent densities. For TGF β treatment, media was supplemented with 5 ng/mL recombinant human TGF β -1 (R&D Systems; resuspended in 4 mM HCl, 0.1% BSA). MCF7,

SKBR3, MDA-MB-231, and Hs-578t cell lines (ATCC) and HEK-293T cells, gift from Dr. Jason Herschkowitz, were cultured in DMEM media (Corning) supplemented with 10% FBS (Cytiva), and 1% penicillin/streptomycin (Lonza). Cells were regularly tested for mycoplasma contamination using Plasmotest™ Mycoplasma Detection Kit (InvivoGen).

Viral transduction

FUGW-d2GFP-ZEB1 plasmid was a gift from Dr. Michael Toneff. The pHAGE-CTCF plasmid was a gift from Gordon Mills & Kenneth Scott (Addgene plasmid #116728 <http://n2t.net/addgene:116728>; RRID:Addgene_116728) and pHAGE-GFP was a gift from Jay Shendure (Addgene plasmid #106281; <http://n2t.net/addgene:106281>; RRID:Addgene_106281). Stable reporter cell lines were generated via viral transfection as described(48). pTRIPZ-shCTCF (Horizon Discovery, #RHS4696-200764825), pTRIPZ-shCtrl (Horizon Discovery, #RHS4743), pHAGE-GFP, and pHAGE-CTCF were transfected alongside pCMV-Δ8.2 and pCMV-VSVG using FuGENE HD Transfection reagent (Promega), and DMEM. Viral supernatant was clarified using a 0.45 μm filter before addition to HEK-293T cells. FUGW-d2GFP-ZEB1-GFP⁺ labeled and GFP⁺ pHAGE cells were isolated using FACSMelody™ Cell Sorter (BD Biosciences). pTRIPZ-expressing cells were selected with 0.5 μg/mL puromycin.

Cell fractionation

Subcellular protein fractionation was performed on 5 x 10⁶ cells using the Subcellular Protein Fractionation Kit for Cultured Cells (ThermoScientific) following the manufacturer's protocol. Ten μg total protein was loaded into each well for western blotting.

Western blotting

Cells were harvested and resuspended in RIPA buffer (Alfa Aesar) supplemented with protease and phosphatase inhibitors (ThermoScientific) and incubated in ice for 60 minutes. Lysed cells were centrifuged at 15,000 rcf for 20 minutes at 4°C and the supernatant was isolated. Protein concentrations were determined using a bicinchoninic acid assay (ThermoScientific).

Proteins were separated by a 12% SDS/PAGE gel, transferred onto a 0.45 µm PVDF membrane (ThermoScientific), and probed with the appropriate antibodies. Chemiluminescence signal was obtained using ECL™ Prime (Cytiva) and blot images were captured using the ChemiDoc™ Imaging System (Bio-Rad).

Flow cytometry

At the conclusion of the TGFβ time course, cells were counted and resuspended in 500 µL 1% FBS (Cytiva) in PBS with anti-E-cadherin BB700 (BD Biosciences #745965, 1:100), anti-CD44 BV421 (BioLegend #562890, 1:100), and anti-CD24 (BioLegend #311104, 1:100) and incubated on ice for 90 minutes. Following incubation, cells were pelleted and washed twice with 1% FBS in PBS and subjected to flow cytometry using a FACSMelody™ (BD Biosciences).

Reagents

The following antibodies (in 5% milk in TBST) were used for western blotting: anti-E-cadherin (#14472, monoclonal mouse, 1:1000, Cell Signaling), anti-N-cadherin (#BDB610920, monoclonal mouse, 1:1000, BD Biosciences), anti-Slug (#9585, monoclonal rabbit, 1:500, BD Biosciences), anti-CTCF (#07-729, polyclonal rabbit, 1:1000, Millipore Sigma), anti-BORIS/CTCFL (PA5-97639, polyclonal rabbit, 1:1000, ThermoFisher), anti-SMC3 (#5696, monoclonal rabbit, 1:1000, Cell Signaling), anti-YY1 (#46395S, monoclonal rabbit, 1:1000, Cell Signaling), anti-ZEB1 (21544-1-AP, polyclonal rabbit, 1:2000, ProteinTech), anti-Vimentin (#103661-AP, 1:2000, ProteinTech), anti-Actin (#612656, monoclonal mouse, 1:2000, BD

Biosciences), anti-H3K27me3 (#9733S monoclonal rabbit, Cell Signaling), anti-H3K4me3 (#9751S, monoclonal rabbit, Cell Signaling), anti-H3 (#4499S, monoclonal rabbit, Cell Signaling), anti-RFP (#MA5-15257, 1:2000, ThermoFisher), anti-rabbit HRP-linked IgG secondary (#7074S, 1:2000, Cell Signaling), anti-mouse HRP-linked IgG secondary (#926-80010, 1:2000, Li-Cor).

RNA extraction and qPCR

RNA was extracted from cell cultures using TriZol Reagent (Invitrogen) and isolated following manufacturer guidelines. 500 ng of RNA was used for cDNA synthesis. TaqMan Gene Expression primers (Applied Biosystems) were used for miR-203, miR-200c, and sno-U6. MicroRNA quantitative PCR analyses were performed using TaqMan Gene Expression Master Mix (Applied Biosystems), normalizing to sno-U6. PowerUp SYBR Green Master Mix (Applied Biosystems) and a QuantStudio5 real-time PCR machine (Applied Biosystems) was used for quantitative PCR analyses with four technical replicates per biological replicate, normalizing to *ACTB*. Signal was quantified and normalized using QuantStudio5 software (version 1.5.1) and analyzed using GraphPad Prism (version 8.4.3).

Nanostring analysis

RNA was extracted from cell cultures using TriZol Reagent (Invitrogen) and isolated following manufacturer guidelines. 100 ng RNA was used as input for probe hybridization to a custom CodeSet. Nanostring nCounter® was used to quantify genes. nSolver 4.0™ software was used to analyze gene counts.

Mammosphere assay

At the conclusion of the time course, 5,000 cells were plated in low-attachment 96-well plates containing 100 μ L MEGM media (without BPE) (Lonza, CC-3150), 20 ng/mL FGF (Sigma Aldrich), 10 ng/mL EGF (Sigma Aldrich), 4 μ g/mL heparin (Sigma Aldrich), and 1% methylcellulose. 25 μ L fresh mammosphere media was added every third day. Spheres were allowed to form for 14 days.

Wound healing assay

Cells were plated to reach 100% confluency at the end of time course. pTRIPZ-shCTCF and pTRIPZ-shCtrl cells were pretreated with 3 μ g/mL doxycycline two days before wounding with a p200 tip. Measurements were made using Nikon NIS Elements Imaging Software (version 4.5) and analyzed using GraphPad Prism (version 8.4.3).

RNA-seq Library Preparation and Sequencing

RNA was extracted from cell cultures using TriZol Reagent (Thermo Fisher) and isolated following manufacturer guidelines. Libraries were prepared using TruSeq Stranded mRNA Library Prep Kit (Illumina).

EMT score calculation

The EMT scores were calculated utilizing the 76-gene expression signature reported (49) and the metric mentioned based on that gene signature (50). For each sample, the score was calculated as a weighted sum of 76 gene expression levels, and weights were measured based on the correlation of a particular gene with *CDH1* expression. The scores were standardized for all the samples in the dataset by subtracting the mean across samples so that the global mean of the score was zero. Negative scores calculated using this method can be interpreted as mesenchymal phenotype and the positive scores as epithelial.

ATAC-seq Library Preparation and Sequencing

ATAC-seq libraries were generated as described (51). Briefly stated, 50,000 cells were centrifuged, resuspended in 50 μ L lysis buffer (10 mM Tris, 10 mM NaCl, 3 mM MgCl₂, and 0.1% IGEPAL CA-630), and centrifuged at 500g for 10 min at 4°C. The pellet was resuspended in transposase reaction mix (25 μ L 2X TD buffer, 2.5 μ L Transposase (Nextera DNA sample preparation kit, Illumina), and 22.5 μ L water and incubated at 37°C for 30 min. Tagmented DNA was purified using MinElute PCR Purification Kit (Qiagen) per manufacturer's instructions. DNA libraries were PCR-amplified using Nextera DNA Sample Preparation Kit (Illumina) using the following PCR conditions: 98°C for 30s, then thermocycling for 98°C for 10s, 63°C for 30s, and 72°C for 1 min for 12 cycles followed by 72°C for 5 min. PCR products were size-selected for 200 to 800 base pair fragments using SPRI-Select Beads (Beckmann-Coulter). ATAC-seq reads were paired-end sequenced using an Illumina NextSeq500.

Computational Resources: ATAC peak abundance and motif analysis

Due to the similarities of both ends, one paired-end read was used for analysis. Adapter sequences were removed using cutadapt (version 1.16.6) and reads were cropped to 30 bp with Trimmomatic (version 0.36.6). ATAC-seq reads were aligned to *hg19* using Bowtie2 (version 2.3.4.2). Mitochondrial, unmapped and random contigs, and ChrY reads were excluded using samtools (version 1.9) (52) to generate filtered bam files for downstream analysis.

Tag directories were generated in HOMER from filtered bam files (53). Peaks from each replicate were called separately using HOMER 'findPeaks [Tag Directory] -style dnase -o auto.' Following, replicate peaks were merged by calling 'mergePeaks -d 300 [Rep 1] [Rep 2]' to determine common peaks. Common peaks were used for downstream analyses.

For Pearson correlation analyses, combined tag directories containing both replicates were used to quantify average peak score. Common merged peaks from each replicate were merged with other time course conditions by calling ‘mergedPeak -d 300 [Peak files].’ Time course peak files were annotated to *hg19* by calling “annotatePeaks.pl [Merged Peak file] hg19 -size 300 -log -d.” Peak scores for each condition were quantile-normalized by the preprocessCore package (version 1.44.0) in R (version 3.5.1). Quantile-normalized Pearson correlation scores were generated by R Base and visualized by the pheatmap R package (version 1.0.12).

Known motif searches were performed using HOMER (53) for 50 base pair regions excluding masked genomic regions by calling ‘findMotifsGenome.pl [Common peaks] hg19 -size 50 -mask.’ Motifs with p-values $> 10^{-12}$ were discarded. Plots representing the number of significant motifs (p-value $\leq 10^{-12}$) and their enrichment were generated using the ggplot2 R package (version 3.2.0).

To identify the differential peaks the HOMER function getDifferentialPeaks was used. This function was used to quickly identify which peaks contain significantly more tags in the target experiment relative to the background experiment. Here, in each pairwise comparison, Vehicle sample was considered as the background sample. The differential peaks were identified based on default parameters i.e., peaks that have 4-fold more tags (sequencing-depth independent) and a cumulative Poisson p-value less than 0.0001 (sequencing-depth dependent). Next, we annotated these differential peaks to genes and obtained the distance of these peaks from the gene transcription start site (TSS). After obtaining the annotated peaks, we looked at the peak distance from TSS together with the fold change in the expression of respective genes. The fold change for a given gene was calculated by simply taking the ratio of expression values in the two samples.

For the GIGGLE Analysis, ATAC-seq BED files were normalized using BEDTools and peaks were called using MACS2. Input peak files were queried by GIGGLE (54) against ~47,000 CistromeDB compiled genome interval files and chromatin accessibility regulators with significant feature overlap to queried files, as quantified by similarity scores, were identified (55,56).

Venn Diagrams and GSEA analysis

Peak similarities were identified using mergePeaks in HOMER and annotated to *hg19* to produce the Entrez IDs for the gene promoters nearest to peaks. GSEA analysis for Molecular Signatures Database (MSigDB) hits was performed for specific peak groups. Because of the abundance of MSigDB hits, our analysis was limited to either unbiased Hallmarks-only gene sets (57), or breast and mammary-specific gene sets (containing keywords “breast” or “mammary”).

ChIP-seq data pre-processing

ChIP-seq bed data files from Fritz et al. (58) were downloaded using SRA Toolkit (version 2.10.5). Before permutation analysis, Bedtools was used to filter out any regions overlapping a list of previously described ENCODE blacklisted and assembly gap regions for ChIP-seq (59).

Peak enrichment analysis

For the provided CTCF ChIP-seq file, we calculated base-pair overlap between the mark and the ATAC peaks. We used a permutation-based technique to determine whether the observed amount of base pair overlap was more than expected by chance. We calculated an empirical p value for the observed amount of overlap by comparing to a null model obtained by randomly shuffling length-matched regions throughout the genome and calculating the amount of base

pair overlap in each permutation. Where relevant, the p-values are adjusted for multiple testing using the Bonferroni correction.

Oligos

The primers used for PCR amplification in this study can be found in Supporting Materials.

Data availability

The ATAC and RNA sequencing data generated in this work have been deposited in NCBI's Gene Expression Omnibus and are accessible through GEO series accession number GSE145851. ATAC-seq files are accessible through the UCSC genome browser [https://genome.ucsc.edu/s/kelsey_johnson1/Reversible%20EMT%20ATAC%2Dseq%20peaks].

Results

TGF β induces EMT phenotypes

MCF10A human epithelial cells, derived from spontaneously-immortalized fibrocystic mammary tissue, (60) are widely utilized as model for epithelial-mesenchymal plasticity (24,61-63). To characterize progressive epithelial-mesenchymal transition (EMT), we subjected MCF10A cells to TGF β treatment and withdrawal for varying durations (64,65). The cells' characteristic epithelial morphology was lost after 2 days of TGF β treatment while a mesenchymal spindle-like morphology emerged at 4 days of TGF β treatment (Figure 1A). Notably, withdrawal following a short-term TGF β treatment (4 days of TGF β) elicited a rapid return to epithelial morphology while 10 days of withdrawal was necessary for cells subjected to long-term TGF β (10 days of TGF β) to resolve back to an epithelial morphology (Figure 1A). To understand the

reversible EMT in the context of partial and full-EMT states, we continued to use short (4 day) and long (10 day) TGF β treatment models.

We next assessed phenotypic alterations associated with EMT. As expected, cells treated with TGF β demonstrated greater migratory capacity than control cells (Figure 1B). This effect is reversible, as TGF β withdrawal suppressed migratory capacity (Figure 1B). Likewise, mammosphere formation capacity, an indicator of stem-like properties imparted through EMT (66), is elevated in TGF β treated cells (Figure 1C). Intriguingly, long-term treatment imparted lower sphere-forming capacity than short-term treatment, yet cells at the initial stage of EMT reversal formed significantly more spheres (Figure 1C). These results demonstrate an enrichment in stem-like properties in partial-EMT states.

To further understand the progressive changes in stem-like phenotype, we next assayed for cell surface expression of E-cadherin, an indicator of epithelial phenotype, as well as CD44 and CD24, for which the combination of CD44^{hi}/CD24^{lo} is indicative of a stem cell-rich sub-population (67). As expected, TGF β treatment decreases the proportion of surface E-cadherin-positive cells (shown in blue) from 93.0% to 34.6% in short-term treated cells and to 17.5% in long-term treated cells (Figure 1D). Withdrawal from short-term TGF β treatment elicited a rapid recovery of surface E-cadherin (93.7% after 4 days withdrawal) while long-term TGF β treatment resulted in a slower recovery of surface E-cadherin expression (83.2% after 4 day withdrawal) (Figure 1D). Interestingly long-term TGF β induces a CD44^{hi}/CD24^{lo} population within the surface E-cadherin^{hi} population which comprises 3.3% of the total population, compared to 0.4% of untreated cells, (Figure 1D, dark blue). TGF β withdrawal facilitates expansion of this population which increases to 24.0% and then 18.1%. The persistence of this population indicates a potential source of stem-like cells with partial-EMT/MET characteristics in cells

undergoing MET. Overall, our TGF β treatment and withdrawal time course reveals asymmetrical acquisition and resolution of EMT phenotypes. In particular, changes in morphology, E-cadherin localization, and migratory capacity are resolved to baseline faster than stemness-associated cell surface markers and sphere-forming ability.

TGF β induces gene expression dynamics

We next evaluated changes in EMT-associated gene expression by RNA-seq, qPCR, and western blotting. To measure broad changes in gene expression, we performed RNA-seq on our short- and long-term reversible EMT models and scored the gene expression based on the epithelial-specific 76-gene score (76GS) (68). Positive 76GS scores correspond to an epithelial-leaning gene expression pattern. TGF β treatment progressively suppresses the 76GS score until TGF β withdrawal whereupon the score progressively recovers (Figure 2A) irrespective of the initial duration of TGF β treatment. Additionally, we assessed progression of EMT-related gene expression using a mammary cell-specific EMT signature (44). Gene expression data is plotted for 50 genes previously shown to be commonly upregulated (EMT-Up) or down-regulated (EMT-Down) by multiple EMT-TFs in HMLE mammary cells (44). In our model, EMT-Up genes increase their expression continuously throughout the time course (Figure 2B) whereas EMT-Down gene expression reaches a minimum level after 2 days of TGF β treatment (Figure 2C). We further analyzed mRNA expression of specific genes through qPCR. *CDH1* (E-cadherin) and *CDH3* (P-cadherin), a proposed partial-EMT marker (69), show progressive downregulation which is durable through short-term TGF β withdrawal (Figure 2D). *CDH2* (N-cadherin) and EMT-TF *SNAI2* (Slug) gene expression showed immediate responsiveness to TGF β treatment and withdrawal (Figure 2D). On the other hand, EMT-TF *ZEB1* gene expression showed delayed upregulation (Figure 2D) while *TWIST1* (Twist) showed no significant change.

Many EMT-related genes are regulated by translation and are efficacious as proteins (24,70); therefore, we assessed changes in protein expression for EMT markers and EMT-TFs. We observe that TGF β induces suppression of E-cadherin and increases expression of EMT-TFs Slug, Twist, and ZEB1 (Figure 2E). Expression of the EMT-TF Slug mRNA and protein expression peaked early in TGF β treatment while expression of Twist and ZEB1 did not reach its highest until later in the long-term TGF β treatment model (Figure 2E). This successive enrichment of EMT-TFs may be indicative of a progressive activation of EMT-TFs wherein partial-EMT is likely driven by early-EMT TFs like Slug, as has been predicted through a modeling approach (42).

We next examined additional facets of *CDH1* promoter and ZEB1 3'UTR regulatory dynamics in our model. MCF10A cells were transduced with the *CDH1* promoter linked to an RFP-encoding gene, facilitating promoter activity tracking at the single cell level (48). Commensurate with qPCR data, the *CDH1* promoter reporter remained active throughout the short-term TGF β treatment and withdrawal timecourse. However, upon extended treatment, reporter-negative cells outnumbered reporter-positive cells, indicating promoter-induced repression (Supplemental Figure S1). Notably, the repression of the *CDH1* promoter was durable despite ten days of withdrawal. To measure post-transcriptional regulation of *ZEB1* we used a 3'UTR activity assay (48). MCF10A cells were transduced with a GFP-linked *ZEB1* 3'UTR reporter (GFP-Z1) and subjected to TGF β treatment. GFP-Z1 expression within GFP-positive cells became elevated at two days of TGF β treatment (Supplementary Figure S1), agreeing with the detectable ZEB1 protein and increase in *ZEB1* mRNA. Unlike *CDH1*, the ZEB1 3'UTR reporter readout returned to baseline upon TGF β withdrawal (Supplementary Figure S1). Epithelial-specific microRNAs -200c (miR-200c) and -203 (miR-203) are repressed during EMT and have

been reported to target *ZEB1* and *SNAI2* (71,72). As expected, miR-200c showed a significant decrease in expression following 2 days of TGF β treatment while suppression of miR-203 did not reach significance until completions of long-term TGF β treatment (Figures 2F and 2G). TGF β withdrawal elicited re-expression of both microRNAs (Figures 2F and G). Altogether these data are consistent with an early (2 days of TGF β) suppression of miR-200c which may alleviate the repression of ZEB1 mRNA and upregulate ZEB1 protein.

EMT and MET-induced changes in accessible chromatin regions

EMT is accompanied by reversible changes in the epigenome (47). Chromatin accessibility orchestrates dynamic gene expression by exposing or hiding genomic regulatory elements. The systematic coordination of genomic elements produces distinct epigenomic states (73). To uncover the epigenomic basis associated with TGF β -driven reversible EMT, we performed the assay for transposase-accessible chromatin with next-generation sequencing (ATAC-seq) (51).

The extent of chromatin accessibility is highly dynamic across EMT and MET. Long-term TGF β treatment was associated with additional ATAC-seq peaks while withdrawal from TGF β treatment was associated with a loss of ATAC-seq peaks (Figure 3A). Dramatically, long-term treatment led to 50% more peaks than untreated cells (175,103 vs 113,680). This was reversed upon 4 days TGF β withdrawal (10d +TGF β , 4d -TGF β) as the number of peaks dramatically decreased (47,174). Though the number of called peaks fluctuate, the proportion of peaks annotated to untranslated regions (UTR), transcriptional termination sites (TTS), and exons remains stable throughout TGF β treatment (Figure 3B). However, greater variation was evident in the proportion of intergenic and promoter-associated peaks which increase or decrease, respectively, in response to TGF β (Figure 3B). These results suggest a broad effect of TGF β treatment on chromatin structure.

To further determine the genomic distribution of newly-accessible chromatin regions, we interrogated the distance, in base pairs, between annotated ATAC peaks. Generally, TGF β treatment decreases the gap size between peaks (Figure 3C). Long-term TGF β -treated cells had significantly shorter distances between ATAC-seq peaks (median 11,564 bp) compared to control (median 15,348 bp). Because many of the peaks unique to long-term TGF β annotate to intergenic and intronic regions, we hypothesized that TGF β treatment decreased the number of 1 megabase (Mb) regions without a peak—reducing the number of so-called “peak deserts”. We enumerated the number of ATAC-seq peak deserts and found that TGF β treatment decreases the number of peak deserts from 185 in control cells to 99 in long-term TGF β (Figure 3D). To understand how changes in chromatin accessibility associate with changes in transcription, we plotted the fold change in gene expression against distance from TSS for peaks with significantly differential peak intensity between either 2d TGF β (Figure 3Ei) or 10d TGF β (Figure 3Eii) and control. As expected, the strongest gene expression fold change values are associated with ATAC-seq peaks nearest to TSSs. Nevertheless, genes with significant changes in chromatin accessibility up to 500 kb on either side of the TSS also exhibit altered expression (Figure 3E). Overall, these data suggests that EMT increases chromatin accessibility across the genome with TSS distal and proximal changes affecting gene expression.

Given the diverse transcriptional dynamics observed in EMT-related genes, we next examined chromatin accessibility at select genes. We first compared peaks within epithelial-specific genes *CDH1* (E-cadherin), *CDH3* (P-cadherin), and *EPCAM* (epithelial cell adhesion molecule). Accessibility at *CDH1* and *CDH3* remained high throughout short-term TGF β treatment as indicated by minor changes in peak profiles. These results are consistent with *CDH1* relative

expression and total E-cadherin protein expression, suggesting that the lack of chromatin perturbations enabled maintenance of *CDH1* transcription and E-cadherin expression. However, in long-term TGF β , some *CDH1* and *CDH3* peaks began to diminish at 10 days of TGF β and continued to do so despite TGF β withdrawal, highlighting a delayed and possibly stable chromatin alteration following EMT (Supplementary Figure 2A). These results contrast with chromatin accessibility at the EPCAM promoter, where peaks scores are more tightly associated with initial exposure to TGF β , but which fail to return to control levels even after long-term TGF β withdrawal (Supplementary Figure 2A). This difference in the recovery of EPCAM accessibility between short and long-term TGF β exposure is in concordance with new findings from the Scheel lab (74) that show high EPCAM linked with the ability to enter a hybrid EMT state but low EPCAM linked with an irreversible mesenchymal state.

Accessibility patterns at mesenchymal-specific genes *CDH2*, *VIM*, *FN1*, *ZEB1*, *SNAI1*, and *SNAI2* (Supplementary Figures 2B,C) are more responsive to TGF β than epithelial genes—increasing in accessibility and peak intensity following TGF β treatment and resolving to untreated conditions in withdrawal. In particular, there are two regions with dense clusters of peaks near the *FN1* promoter. While peak density at the promoter proximal cluster remains elevated throughout TGF β treatment and withdrawal, peak density at the promoter distal cluster (gray box, Supplementary Figure 2B) corresponds closely to TGF β exposure. Similarly, *SNAI1* and *SNAI2* both show greater changes in peak height at regions distal to the coding sequence or promoter regions (Supplementary Figure 2C).

Chromatin changes occur early in EMT and are mostly reversible

Next, to determine the relationships between timepoints in terms of chromatin accessibility patterns we calculated a correlation coefficient for each pair of samples. Because of the

differences in peak distribution and score among TGF β treated and withdrawn conditions, peaks were merged into 300 bp bins, and peak scores were quantile-normalized, prior to correlation with other time points. Expectedly, untreated and TGF β withdrawn conditions share similarities ($R^2=0.95$ and 0.93 for short-term and long-term, respectively) (Figure 3F). The greatest distinction in chromatin accessibility patterns was between untreated and long-term TGF β treatment ($R^2=0.81$) and untreated and 2 days TGF β ($R^2=0.82$). The lack of similarity in chromatin peaks between untreated and 2d TGF β suggests that chromatin reprogramming occurs very early in EMT induction, preceding many transcriptomic changes. Interestingly, the strong correlation of chromatin accessibility patterns between 2 and 10 days TGF β ($R^2=0.94$) (Figure 3F) suggests that early chromatin accessibility alterations are sustained during EMT induction.

We next asked whether major changes in chromatin accessibility occur at regions with low, moderate, or high accessibility. Mean peak scores (MPS) were generated for pairs of samples by averaging the peak scores for those two samples. MPSs below 3 were considered “low accessibility”, while MPSs above 5 were considered “high accessibility”. We next determined the differential peak scores (DPS) for each pair of samples in order to analyze their change in accessibility. By plotting these two metrics together, we observed that major changes in accessibility (defined as a DPS above 2 or below -2) occur primarily at low and moderately accessible peaks (Figure 3G) but can occur at highly accessible peaks. In particular, a major increase in the accessibility at strong peaks is evident when comparing the long-term TGF β treatment to untreated cells (Figure 3Gi, yellow). This is followed by a major decrease in the accessibility at strong peaks when comparing long-term TGF β treatment to cells that have undergone a full TGF β withdrawal (Figure 3Gii, yellow). These differential peak scores are no longer evident when comparing control to fully withdrawn cells (Figure 3Giii). Analysis of short-

term TGF β treatment indicates a similar trend to a lesser magnitude (Figure 3Giv-vi). These data indicate that long-term TGF β treatment confers a strong increase in accessibility at highly accessible regions, which are likely to directly regulate gene expression, but also at low accessible regions, which may be more indicative of transcriptional noise.

To better understand the phased implementation of EMT-associated chromatin accessibility, we analyzed the sharing of ATAC-seq peaks between timepoints in the long-term treatment model. As expected, many chromatin accessibility regions in untreated cells are retained throughout the time course (n=38,790) (Figure 3H). However, TGF β treatment also led to ATAC-seq peaks which either faded after TGF β withdrawal (n=28,619) or were retained throughout TGF β withdrawal (n=2397) (Figure 3H).

To derive the functional implications of these chromatin accessibility regions, we performed gene set enrichment analysis (GSEA) for the annotated genes closest to differential ATAC-seq peaks (74). Among the long-term TGF β peaks unique to long-term TGF β treatment, Molecular Signatures Database (MSigDB) revealed enrichment of gene sets involved in Hallmark EMT (p=5.5 x 10⁻¹¹), TNF α signaling (p=7.0 x 10⁻¹³), hypoxia (p=2.75 x 10⁻⁶), and TGF β signaling (p=0.00017). This enrichment diminishes following TGF β withdrawal. We were also interested in the overall function of the genes near persistent EMT peaks, and MSigDB hits reveal enrichment of genes involved in EMT (p=1.2 x 10⁻⁶), apical junction assembly (p=0.0010), and mammary stemness genes (p = 1.8 x 10⁻¹⁰) (Supplementary Figure 3).

Dynamic transcription factor engagement during EMT and MET

We next tested the hypothesis that TGF β modulates enrichment of transcription factor binding sites (TFBS) at accessible regions to enable partial- or full-EMT. We used HOMER motif

discovery analysis to determine TFBS enrichment within our ATAC-seq peaks (53). We wished to limit our analysis to global TFBS, so we segmented the peaks to narrow 50 bp regions and identified the top 20 differentially-enriched TFBSs in comparison to untreated conditions. TGF β increases enrichment of AP-1 (17.1% in untreated, 18.9% in short- and 18.8% in long-term) and SMAD3 (0.72% in untreated, 1.53% in short- and 1.04% in long-term) binding motifs in 50 bp segments, and their enrichment resolves to baseline levels during TGF β withdrawal (Figure 4A). To further confirm the level of SMAD protein binding activity in our time course we measured the extent of similarity between the pattern of ATAC-seq peaks and curated patterns of genomic occupancy established for DNA-interacting factors using GIGGLE analysis (54). Limiting our comparisons to genome interval files derived from mammary or breast cancer cells, we observed that similarity scores for SMAD2/3 occupancy patterns diminished over the time course of treatment (Supplemental Figure 4A). Indeed, the SMAD2/3 occupancy patterns yielded the highest similarity scores for 2 days and 4 days of TGF β treatment but not for 10 days of TGF β (Supplementary Figure 4A). Patterns associated with the dimeric AP-1 transcription factor family (FOS, JUN) also yielded high similarity scores for 2 days of TGF β (Supplementary Figure 4B-E), confirming the HOMER analysis (Figure 4A).

Because of the large range in TFBS enrichment amongst motifs, we also calculated the motif percent difference from untreated cells (Figure 4B). While enrichment of most motifs increased following TGF β treatment, we observed that CTCF and BORIS motif enrichment was diminished during the EMT/MET time course, with the notable exception of long-term TGF β cells. Indeed, at two and four days of TGF β treatment CTCF and BORIS motif enrichment exhibited the strongest decline for any TFBS motif. TGF β withdrawal restores these motifs to near-untreated levels.

CTCF (also known as CCCTC-binding factor) and BORIS (also known as CTCF-like or CTCFL) bind to similar cytosine-rich DNA binding motifs and exhibit some overlapping functions. Both CTCF and BORIS establish topologically-associated chromatin domains, regulate genetic imprinting, and modulate gene expression neighborhoods (75,76). Given the unusual motif enrichment pattern, we next confirmed whether ATAC peaks considered to be enriched for the CTCF/BORIS motif overlap with validated CTCF binding sites based on ChIP-seq data in MCF10A cells by Fritz et al., 2017 (58). We observe a 143-fold enrichment for validated CTCF binding sites in our peaks with the CTCF motif ($p=0.001$) compared to just a 25-fold enrichment for validated CTCF binding sites in the set of all ATAC peaks (Figure 4C). These data support the notion that the peaks with a predicted CTCF-binding site correspond to *bona fide* CTCF-bound sites. The average distance between ATAC peaks with CTCF binding motifs remained stable during the initial phases of TGF β treatment but dropped significantly at 10 days of TGF β treatment, indicating that newly formed ATAC peaks with CTCF binding motifs are not clustering near existing peaks. Upon TGF β withdrawal, the average distance between ATAC peaks with CTCF binding motifs stabilized back to untreated levels (Figure 4D). As with the increase in general chromatin accessibility (Figure 3C), the change in the distance between accessible regions containing CTCF motifs was consistent with novel engagement of CTCF bindings sites across the genome rather than in clusters.

Given the loss of CTCF binding sites from accessible chromatin regions during short- but not long-term TGF β treatment, we determined how CTCF or BORIS protein expression was altered over a time course of EMT/MET. Overall, TGF β treatment suppresses expression of CTCF protein while TGF β withdrawal restores expression (Figure 4E). Remarkably, the expression of known CTCF binding partners SMC3 (a member of the cohesin complex) (77) and YY1 (a DNA-binding protein that forms enhancer-associated complexes with CTCF) (78) is suppressed

during TGF β treatment in MCF10A cells (Figure 4E), concurrent with CTCF. BORIS, however, exhibited a less dynamic expression pattern (Supplementary Figure 5A). In order to validate these findings in other models, we observed that TGF β -treated HMLE mammary epithelial have decreased CTCF protein expression (Supplementary Figure 5A), and mesenchymal breast cancer cell lines MDA-MB-231 and Hs578t cells have lower CTCF protein expression than epithelial breast cancer cell lines MCF7 and SKBR3 (Figure 4F).

Considering the TGF β -driven loss of CTCF motif enrichment in ATAC-seq peaks, we next determined if TGF β -driven CTCF protein loss affected the nuclear and chromatin localization of CTCF. We isolated specific subcellular fractions at short- and long-term TGF β conditions and compared the enrichment of CTCF in specific fractions. Diminished CTCF total protein was most evident in the input fraction but also observed in soluble nuclear and chromatin-bound fractions for TGF β -treated MCF10A cells (Figure 4G) and in EMT-positive breast cancer cell lines (Supplementary Figure 5C). Although CTCF protein expression is repressed in nuclear fractions, we observe that CTCF protein is able to stably associate with chromatin in long-term TGF β conditions which may partially explain the increase in CTCF binding motif enrichment in ATAC-seq peaks at that timepoint (Figure 4B). Surprisingly, TGF β treatment increases the proportion of the smaller ~80 kDa CTCF isoform, CTCF-s in nuclear fractions (Figure 4G). CTCF-s is known to antagonize binding of full-length CTCF thus altering the organization of chromatin looping (77). This smaller isoform is also evident in breast cancer cell lines and in HMLE cells exposed to TGF β (Supplementary Figure 5B). These data indicate that expression, localization, and motif exposure of CTCF and other insulator-associated proteins are altered depending on the EMT/MET state.

We next determined if genes nearby ATAC peaks with CTCF motifs were enriched for specific functions. Comparing long-term TGF β treatment and withdrawal, most genes adjacent to CTCF motifs were shared amongst the three conditions (n=4117) however long-term treatment had more unique genes (n=1518) (Supplementary Figure 6A). We performed gene set enrichment analysis (GSEA) for the genes uniquely-accessible in long-term TGF β conditions and found high enrichment in gene sets regulated by chromatin bivalency (nucleosomes marked by both K27me3 and K4me3 histone H3 modifications) and neuronal development (Supplementary Figure 6B). Given the enrichment of CTCF motifs in genes related to histone bivalency, we assessed the global levels of relevant histone modifications (79). Concomitant with CTCF repression patterns, we also observe increases in H3K4me3 and H3K27me3 in TGF β treated conditions (Supplementary Figure 6C).

CTCF knockdown induces epithelial traits and facilitates partial EMT

We next considered the regulatory relationship between CTCF and EMT. By comparing published data (58) with our ATAC-seq data, we note verified CTCF binding sites coincident with CTCF motifs adjacent to key EMT genes. We find that CTCF motifs align to key ATAC peaks nearby *CDH3* and *CDH1* and downstream of *EPCAM* (Supplementary Figure 2A). These sites, in particular, align to peaks that diminish following TGF β treatment. We also note several CTCF binding sites within *CDH2*, *VIM*, upstream of *FN1*, and upstream of *SNAI2* (Supplementary Figures 2B,C). The majority of the aforementioned sites align to peaks that increase following TGF β treatment and diminish after withdrawal.

Given the suppression of CTCF during TGF β -driven EMT and lower expression in mesenchymal breast cancer cell lines, we next investigated the functional role of CTCF expression in EMT. To do so, we generated cell lines with a doxycycline-inducible RFP plus

shRNA targeting *CTCF* and probed for expression of epithelial and mesenchymal markers. Though knockdown of CTCF via doxycycline (doxy) treatment did not change cellular morphology in comparison to vehicle treated cells or doxy-treated non-targeting control (shCtrl) (Figure 5A), we confirmed doxy-induced repression of CTCF (Figure 5B). Unexpectedly, given the loss of CTCF protein during EMT, CTCF knockdown increases E-cadherin protein expression (Figure 5B). CTCF knockdown also alters the mRNA expression of EMT marker genes as noted by up-regulation of *CDH1* and basal epithelial marker *KRT5* (Supplementary Figure 7A). Concordant with CTCF's putative involvement in H3K27me3/H3K4me3-(bivalency) marked genes (Supplementary Figure 6B), the level of H3K27me3 decreases with CTCF knockdown (Figure 5B).

Though we originally expected CTCF knockdown to induce mesenchymal traits, results of our loss-of function experiments led us to postulate that loss of CTCF slows the induction of EMT. To investigate this, we induced CTCF knockdown followed by 2 days of TGF β treatment and assessed epithelial and mesenchymal marker expression. We limited our analysis to 2 days of EMT to capture early response of mesenchymal markers. We found that CTCF knockdown, in addition to elevating E-cadherin transcript expression, also precludes the TGF β -mediated suppression of *CDH1* (Figure 5D). Conversely, CTCF knockdown accelerates the induction of *CDH2* transcriptional upregulation in response to TGF β (Figure 5D). This data implicates CTCF expression in epithelial-mesenchymal plasticity of mammary epithelial cells.

CTCF overexpression induces mesenchymal traits

We next determined the implications of CTCF upregulation on EMT. To do so, we generated *CTCF* and *GFP* over-expressing MCF10A cells. CTCF overexpressing MCF10A cells appear more mesenchymal and spindle-like (Figure 5E), suppress E-cadherin expression while

upregulating vimentin protein expression, as well as the level of H2K27me3 (Figure 5F). CTCF overexpression drives transcript-level changes consistent with EMT including significant gain of *CDH2*, *FN1*, *VIM*, and *ZEB1* (Figure 5G). Moreover, CTCF-overexpressing cell lines had greater migratory capacity than controls (Figure 5H).

In conclusion, we demonstrate that TGF β -induced EMT is accompanied by global chromatin relaxation and that TGF β -withdrawal resolves most but not all alterations in chromatin accessibility. We show that enrichment of CTCF-binding motifs is highly dynamic across EMT/MET. Lastly, we demonstrate that CTCF expression is implicated in epithelial-mesenchymal cell plasticity.

Discussion

EMT and its reversal, MET, are important for normal physiological processes such as development and wound healing. Together, these processes are hypothesized to endow cancer cells with metastatic abilities. Transcriptomic determination of EMT status has linked EMT to poor prognostic characterizations such as claudin-low breast cancer (44), mesenchymal glioblastoma (80), rapamycin-resistance in breast cancer (81), radio-resistance in prostate cancer (82), and immune system suppression (83). Additionally, EMT progresses into and through partial states which can assume various morphologies, gene expression, and E-cadherin profiles and exhibit greater pathogenic properties. Examination of solid tumor models has shown that cells within an intermediate-mesenchymal state are more spheroidogenic and resistant to anoikis (41). Further, cells that express both KRT14 and vimentin disproportionately contribute to metastasis (37). Despite this importance, the factors that mediate partial- and full-EMT states are not well-characterized. In this study, we have exclusively characterized the

chromatin accessibility alterations and gene expression output that occur during distinct states of reversible TGF β -induced EMT.

Herein, we show EMT and MET progress through stage-wise gene expression changes. A majority of gene expression alterations occur within 48 hours of TGF β treatment. TGF β treatment rapidly induces Slug (*SNAI2*) mRNA and protein overexpression, which decreases following additional TGF β treatment into withdrawal (Figure 2D,E). These results suggest that Slug is involved in early EMT induction. *ZEB1* mRNA and protein increase after extended TGF β treatment and the *ZEB1*-3' UTR remains expressed following TGF β withdrawal (Supplementary Figure 1). These data are similar to those recently reported by Jia et al. who find that the *ZEB1* 3'UTR reporter remains expressed despite prolonged TGF β exposure and withdrawal (84). Similarly, our results corroborate the findings by Ye et al. who reported that partial-EMT states are coordinated by Slug while *ZEB1* and *SNAI1* promote a complete-mesenchymal phenotype (85) and data from Addison et al. who show that Slug and *ZEB1* EMT-TFs both contribute to E-cadherin suppression but fail to up-regulate expression of each other (86). Our data reveal distinctions between short- and long-term TGF β treatments, in particular, sequential activation of EMT-TFs.

We also characterized EMT/MET states in terms of surface-localized protein markers. To our surprise, we discovered that E-cadherin is lost from the membrane within 2 days TGF β treatment despite robust total E-cadherin and *CDH1* expression. Further TGF β treatment (4 and 10 days) yields further suppression of surface E-cadherin. Interestingly, TGF β withdrawal following long-term treatment stimulates a return of surface E-cadherin^{hi} populations despite suppressed *CDH1* and total E-cadherin expression. Such a recovery of surface E-cadherin suggests the presence of a cytoplasmic store capable of returning to the membrane without

transcriptional upregulation. This finding is in concert with past studies on E-cadherin recycling (87), including reports of the interaction between E-cadherin and late recycling endosome vesicles via Rab5 and Rab11 (70) and recent studies in MCF10A cells demonstrating that extended TGF β treatment is necessary to induce loss of membrane-localized E-cadherin (65). TGF β withdrawal generates a hybrid E-cadherin^{hi}/CD44^{hi}/CD24^{lo} population, which, given the controversial role of membrane-bound E-cadherin in metastasis (88), may have functional implications in breast cancer progression.

Through interrogation of Tn5-accessible chromatin, we identify that TGF β treatment leads to wide-spread alterations in chromatin accessibility. Extensive chromatin alterations occur within 2 days of TGF β treatment—revealing 22,354 more ATAC peaks and sheltering 9,678 peaks. Many newly-accessible regions have low accessibility scores which may contribute to the stochastic transcriptional noise that permits cell plasticity (89). Motif enrichment analyses reveal that TGF β increases the enrichment for AP-1 and SMAD family binding motifs, transcription factors which have been reported to regulate EMT within accessible regions (36,37,62,90,91). Withdrawal from TGF β induces global chromatin constriction, lower motif enrichment, resolving chromatin nearly to the untreated state. These results demonstrate that chromatin is highly-responsive to TGF β treatment.

While most transcription factor motifs either uniformly increase or decrease following TGF β treatment, enrichment of CTCF and BORIS binding elements declines during intermediate states. This suggests a period of chromatin re-shuffling, preparing chromatin for configurations that would promote plasticity through phenotypic states. Notably Pastushenko et al. also identified CTCF as a factor important to full phenotypic states (37). In their investigation, they isolated populations of epithelial (Epcam⁺), intermediate (CD51⁻/CD61⁻), and mesenchymal

(CD106⁺/CD51⁺/CD61⁺) squamous cell carcinoma cells and subjected them to ATAC-seq (37). Motif enrichment analyses reveal that CTCF motifs were highly enriched in cells exhibiting epithelial and mesenchymal, but not intermediate, states.

In our study, we demonstrate that CTCF, a master chromatin organizer, is dynamically expressed in reversible EMT. EMT-inducing signals reduce CTCF expression while MET restores CTCF expression. We also identified that EMT-related genes (*CDH1*, *CDH2*, *SNAI2*, and *ZEB1*) contain nearby CTCF binding sites—many of which demonstrate dynamic peak profiles during the course of reversible EMT. Recently, multiple studies have implicated CTCF binding key features of cancer biology including apoptosis. For example, Kaiser et al. probed transcription factor binding sites across 11 tumor types and identified that CTCF binding sites carried high mutational loads (92). They posited that the mutations alter chromatin landscapes, replication timing, and DNA fidelity within tumors. Further, DNA methylation is well-characterized to affect CTCF binding as methylation at CTCF motifs can drive overexpression of oncogenes such as *PDGFRA* (93). Studies focusing on CTCF protein structure and DNA-binding kinetics have identified an alternatively-spliced isoform, CTCF-s, which lacks the zinc finger domains that associates with cohesin complex proteins (77). CTCF-s competes with canonical CTCF (CTCF-fl) for binding sites, consequently imposing variations in chromatin looping and gene neighborhoods (77). Indeed, we observed consistent loss of the CTCF-s isoform in cells with mesenchymal features (Figure 4E,F,G and Supplementary Figure 5B,C).

Through our investigation of the functional consequences of CTCF gain- and loss-of-function on EMT, we demonstrate that CTCF knockdown enhances the epithelial phenotype and suppresses mesenchymal markers. These findings are concordant with Zhao et al. who found that CTCF knockdown suppresses invasion and migration, proliferation, and ovarian cancer

metastasis (94). Unexpectedly, we found that CTCF suppression also decreases H3K27me3, a histone modification tightly linked to EMT-induced gene bivalency (47). However, the suppression of CTCF paired with the addition of TGF β increases *CDH2* expression without suppressing *CDH1*, suggesting that the withdrawal of CTCF from specific chromatin domains may lead to a partial EMT phenotype. Conversely, constitutive overexpression of CTCF induces a more mesenchymal phenotype including upregulation of mesenchymal markers and increased migratory speed as well as increased global levels of H3K27me3. Given the findings of Zhang et al., who determined that CTCF overexpression is linked to poor prognoses in patients with hepatocellular carcinoma (95), these findings could have considerable relevance to breast cancer. Though beyond the scope of this study, it would be noteworthy to observe how TGF β -induced CTCF suppression affects broad 3D chromatin structure and topological domains.

In conclusion, our study demonstrates that mammary epithelial cells proceed through EMT and MET via progressive and partially reversible chromatin accessibility alterations. We reveal that EMT is marked by global chromatin loosening and MET is marked by chromatin constriction. We show that a subset of differentially-accessible ATAC peaks are enriched for CTCF binding motifs and we demonstrate that CTCF overexpression can promote an EMT-like state in mammary epithelial cells. Collectively our findings indicate that activation of EMT and MET dramatically reconfigures chromatin organization and that CTCF is a key modulator of epithelial-mesenchymal plasticity.

Author Contributions

J.H.T. and Y.C.L. conceived, managed, and arranged funding for the project; J.H.T. and K.S.J. designed the experiments; K.S.J. performed most of the experiments and performed ATAC-

and RNA-seq bioinformatics analysis; S.H., P.C., and M.K.J. provided computational analysis for RNA-seq; M.L.B. and A.M. provided computational analysis for ATAC-seq; K.S.J, G.G., and S.S. assisted with quantitative PCR data; K.S.J. and M.J.T. contributed to the Z-CAD transduced cells used in this study; K.S.J and S.M. performed western blotting; Y.C.L. provided training and guidance with bioinformatic analyses. J.H.T. and K.S.J. wrote the manuscript. All authors read and approved the final draft of the manuscript.

Acknowledgements

We thank Dr. Dwayne Simmons for equipment sharing. We thank Dr. Michelle Nemec and the Baylor University Molecular Biosciences Center for support during the course of this work. We would like to thank Dr. Bernd Zechmann (Center for Microscopy and Imaging, Baylor University, Waco, TX) for technical support during microscopy and image analysis. This work was supported by the Collaborative Faculty Research Investment Program (30300179) and the Susan G. Komen Foundation Career Catalyst Research Grant (CCR18548469) to J.H.T. M.K.J. was supported by Ramanujan Fellowship awarded by Science and Engineering Board (SERB), Department of Science and Technology (DST), and the Government of India (SB/S2/RJN-049/2018).

Competing Interests

The authors declare no competing interests.

Figure Legends

Figure 1. TGFβ treatment and withdrawal elicits phased cell biological changes indicative of EMT-MET. (A) Brightfield photomicrographs of MCF10A cells at indicated

durations; TGF β used at 5 ng/mL Scale bars = 100 μ m. **(B)** Migration capacity was determined using a scratch-wound healing assay. MCF10A cells were treated as indicated prior to replating at confluency. The change in average gap length after 10 hours is reported. Error bars indicate s.e.m. (n = 6). Statistical significance was tested using a one-way ANOVA followed by comparison of each mean to untreated cells using a Dunnet correction for multiple hypothesis testing. **(C)** Mammosphere-formation capacity of cells treated as indicated. Following the conclusion of the time course, TGF β -treated and -withdrawn cells were subjected to mammosphere-promoting conditions for 10 days and mammospheres $\geq$ 50 μ m were counted. Error bars indicate s.d. (n=6). Statistical significance was tested using a one-way ANOVA followed by comparison of each mean to untreated cells using a Dunnet correction for multiple hypothesis testing. **(D)** FACS profiling of CD44, CD24, and surface E-cadherin in MCF10A cells treated as indicated. Cells were categorized as either E-cadherin-high (blue) or -low (pink) and either CD44^{hi}/CD24^{lo} (lighter shading) or non-CD44^{hi}/CD24^{lo} (darker shading). Error bars indicate s.d. (n=3). Statistical significance was tested using a one-way ANOVA followed by comparison of each mean to untreated cells using a Dunnet correction for multiple hypothesis testing. * $p \leq 0.05$, ** $p \leq 0.01$, *** $p < 0.001$, **** $p < 0.0001$

Figure 2. Gene expression dynamics during TGF β treatment and withdrawal. RNA-seq data were analyzed in regard to expression of a **(A)** 76-Gene Epithelial Metric (49), **(B)** Core EMT-Up and **(C)** Core EMT-Down Gene Signatures (44). **(D)** mRNA expression of select epithelial and mesenchymal genes was determined using qPCR, normalized to untreated cells and *ACTB* and is represented as mean and s.d. (n = 3). Statistical significance was tested using a one-way ANOVA followed by comparison of each mean to untreated cells using a Dunnet correction for multiple hypothesis testing. **(E)** Western blot for epithelial (top) and EMT-TFs (bottom) in MCF10A cells treated with TGF β for the indicated durations. Expression of **(F)** miR-200c and **(G)** miR-203 was normalized to sno-U6 and is represented as mean and s.d. Statistical significance was tested using a one-way ANOVA followed by comparison of each mean to untreated cells using a Dunnet correction for multiple hypothesis testing (n=4). $p \leq 0.05$, ** $p \leq 0.01$, *** $p < 0.001$, **** $p < 0.0001$.

Figure 3. Dynamics of EMT- and MET-associated chromatin accessibility. (A) Number and (B) proportion of ATAC-seq peaks according to function of the genomic region and duration of TGFβ treatment and withdrawal. (C) Gap length distribution between ATAC peaks. Center line represents mean, box edges represent 25th and 75th percentiles, whisker ends represent 5th and 95th percentiles. Data were analyzed using two-way ANOVA using Tukey's multiple comparison test vs 0d. ns = not significant, * $p \leq 0.05$, ** $p \leq 0.01$, *** $p \leq 0.001$, **** $p < 0.0001$. (D) Number of gaps between peaks that exceed 1Mb, excluding centrosome sequence. (E) Scatter plot showing fold-change in gene expression (RNA-seq) vs. distance from TSS for specific genes showing differential peak intensity for (i) 2d +TGFβ vs 0d and (ii) 10d +TGFβ vs

0d timepoints. Each dot represents a peak with at least 4-fold more accessibility. **(F)** Pearson's correlation analysis on quantile-normalized log-transformed ATAC peaks. **(G)** Differential accessibility (log₂fold change in reads per 300-bp region) among indicated conditions in long-term (i, ii, iii) and short-term (iv, v, vi) TGF- β treatment models. Colored dots represent peaks with at least 4-fold more accessibility at the indicated conditions. Black = untreated, yellow = 10d +TGF β , purple = 10d +TGF β , 10d -TGF β , blue = 4d +TGF β , and light green = 4d +TGF β , 10d -TGF β . **(H)** Venn diagram representing the overlap between ATAC-seq peaks common among untreated and TGF β -treated conditions.

Figure 4. EMT-driven TFBS accessibility and CTCF expression dynamics. (A) Motif enrichment by percentage of 50-bp targets containing motif of the top 20 differentially-enriched motifs. **(B)** Motif enrichment percent change, compared to 0d, of top 20 differentially-enriched motifs. **(C)** The overlap between ATAC-seq peaks (in total, or with CTCF motifs) and CTCF

ChIP-seq peaks from Fritz et al. (58) was compared to a random distribution model. The observed amount of base pair overlap between the all ATAC-seq peaks and ChIP-seq peaks is 2,758,444 bp, while the expected amount of overlap based on our null distribution is 110,000.50 bp (s.d. = 4894.19 bp). For the ATAC peaks with a CTCF binding motif the enrichment is greater (f.c. = 143, $p = 0.001$). We observe 1,728,041 bp, but expect 12,083.21 bp with a randomly distributed set of genomic regions (s.d. = 1607.12 bp) One thousand permutations were used to generate a p-value which was corrected using Bonferroni method. **(D)** Gap length distribution (in bp) between ATAC peaks containing CTCF motifs (left). Center line represents mean, box edges represent 25th and 75th percentiles, whisker ends represent 5th and 95th percentiles. Data were analyzed using two-way ANOVA using Tukey's multiple comparison test, statistical comparisons shown in table to the right of graph, **** $p < 0.0001$. **(E)** Western blot for CTCF binding partners SMC3 and YY1 during long-term TGF β treatment and withdrawal model. **(F)** Western blot for CTCF and BORIS protein expression in indicated breast cancer cell lines. **(G)** Western blot for CTCF expression in specific subcellular fractions in MCF10A cells treated with 4 or 10 days TGF β in comparison to untreated control. H3 used as reference for quantification.

Figure 5. CTCF suppression and overexpression modulates epithelial and mesenchymal gene expression. (A) Representative images of TRIPZ-non-targeting control (shCtrl) and TRIPZ-shRNA against *CTCF* (shCTCF) MCF10A cells with and without 2 days 3.0 μg/mL doxycycline (doxy) treatment. Scale bars = 200 μm. (B) Western blot for indicated proteins in shCtrl- and shCTCF-expressing cells. (C) Gene expression measured by qPCR (normalized to *ACTB*) for indicated genes in shCtrl and shCTCF MCF10A cell lines with and without doxycycline (n = 3). (D) Gene expression measured by qPCR (normalized to *ACTB*) for shCTCF-expressing cells with or without 2d TGFβ treatment (n = 3). (E) Representative images of pHAGE-GFP (GFP) and pHAGE-CTCF-expressing MCF10A cells. Scale bars = 100 μm. (F) Western blots for indicated proteins in GFP and CTCF-expressing cells. (G) Gene expression measured by qPCR (normalized to *ACTB*) for indicated genes in GFP and CTCF-

expressing cells (n = 3). **(H)** Scratch assay for GFP- and CTCF-expressing cells. The change in average gap length after 4, 6, and 22 hours is reported. Error bars indicate s.e.m., (n = 4). Statistical significance was tested using a two-tailed Student's t-test. * $p \leq 0.05$, ** $p \leq 0.01$, *** $p < 0.001$, **** $p < 0.0001$

REFERENCES

1. Nieto, M.A., Huang, R.Y., Jackson, R.A. and Thiery, J.P. (2016) EMT: 2016. *Cell*, **166**, 21-45.
2. Tran, H.D., Luitel, K., Kim, M., Zhang, K., Longmore, G.D. and Tran, D.D. (2014) Transient SNAIL1 expression is necessary for metastatic competence in breast cancer. *Cancer research*, **74**, 6330-6340.
3. Tsai, Jeff H., Donaher, Joana L., Murphy, Danielle A., Chau, S. and Yang, J. (2012) Spatiotemporal Regulation of Epithelial-Mesenchymal Transition Is Essential for Squamous Cell Carcinoma Metastasis. *Cancer cell*, **22**, 725-736.
4. Xu, Y., Lee, D.K., Feng, Z., Xu, Y., Bu, W., Li, Y., Liao, L. and Xu, J. (2017) Breast tumor cell-specific knockout of Twist1 inhibits cancer cell plasticity, dissemination, and lung metastasis in mice. *Proceedings of the National Academy of Sciences of the United States of America*, **114**, 11494-11499.
5. Tsai, J.H., Donaher, J.L., Murphy, D.A., Chau, S. and Yang, J. (2012) Spatiotemporal regulation of epithelial-mesenchymal transition is essential for squamous cell carcinoma metastasis. *Cancer cell*, **22**, 725-736.
6. Revenco, T., Nicodeme, A., Pastushenko, I., Sznurkowska, M.K., Latil, M., Sotiropoulou, P.A., Dubois, C., Moers, V., Lemaire, S., de Maertelaer, V. *et al.* (2019) Context Dependency of Epithelial-to-Mesenchymal Transition for Metastasis. *Cell Rep*, **29**, 1458-1468 e1453.
7. De Craene, B. and Berx, G. (2013) Regulatory networks defining EMT during cancer initiation and progression. *Nature reviews. Cancer*, **13**, 97-110.
8. Dravis, C., Chung, C.Y., Lytle, N.K., Herrera-Valdez, J., Luna, G., Trejo, C.L., Reya, T. and Wahl, G.M. (2018) Epigenetic and Transcriptomic Profiling of Mammary Gland Development and Tumor Models Disclose Regulators of Cell State Plasticity. *Cancer cell*, **34**, 466-482 e466.
9. Dravis, C., Spike, B.T., Harrell, J.C., Johns, C., Trejo, C.L., Southard-Smith, E.M., Perou, C.M. and Wahl, G.M. (2015) Sox10 Regulates Stem/Progenitor and Mesenchymal Cell States in Mammary Epithelial Cells. *Cell Rep*, **12**, 2035-2048.
10. Hollier, B.G., Tinnirello, A.A., Werden, S.J., Evans, K.W., Taube, J.H., Sarkar, T.R., Sphyris, N., Shariati, M., Kumar, S.V., Battula, V.L. *et al.* (2013) FOXC2 Expression Links Epithelial-Mesenchymal Transition and Stem Cell Properties in Breast Cancer. *Cancer research*, **73**, 1981-1992.
11. Stemmler, M.P., Eccles, R.L., Brabletz, S. and Brabletz, T. (2019) Non-redundant functions of EMT transcription factors. *Nature cell biology*, **21**, 102-112.
12. Malouf, G.G., Taube, J.H., Lu, Y., Roysarkar, T., Panjarian, S., Estecio, M.R., Jelinek, J., Yamazaki, J., Raynal, N.J., Long, H. *et al.* (2013) Architecture of epigenetic reprogramming following Twist1-mediated epithelial-mesenchymal transition. *Genome biology*, **14**, R144.
13. McDonald, O.G., Wu, H., Timp, W., Doi, A. and Feinberg, A.P. (2011) Genome-scale epigenetic reprogramming during epithelial-to-mesenchymal transition. *Nat Struct Mol Biol*, **18**, 867-874.
14. Weyemi, U., Redon, C.E., Choudhuri, R., Aziz, T., Maeda, D., Boufraqueh, M., Parekh, P.R., Sethi, T.K., Kasoji, M., Abrams, N. *et al.* (2016) The histone variant H2A.X is a regulator of the epithelial-mesenchymal transition. *Nature communications*, **7**.
15. Skrypek, N., Goossens, S., De Smedt, E., Vandamme, N. and Berx, G. (2017) Epithelial-to-Mesenchymal Transition: Epigenetic Reprogramming Driving Cellular Plasticity. *Trends Genet*, **33**, 943-959.
16. Lombaerts, M., van Wezel, T., Philippo, K., Dierssen, J.W., Zimmerman, R.M., Oosting, J., van Eijk, R., Eilers, P.H., van de Water, B., Cornelisse, C.J. *et al.* (2006) E-cadherin transcriptional

- downregulation by promoter methylation but not mutation is related to epithelial-to-mesenchymal transition in breast cancer cell lines. *Br J Cancer*, **94**, 661-671.
17. Vrba, L., Jensen, T.J., Garbe, J.C., Heimark, R.L., Cress, A.E., Dickinson, S., Stampfer, M.R. and Futscher, B.W. (2010) Role for DNA Methylation in the Regulation of miR-200c and miR-141 Expression in Normal and Cancer Cells. *PloS one*, **5**.
18. Brown, R.L., Reinke, L.M., Damerow, M.S., Perez, D., Chodosh, L.A., Yang, J. and Cheng, C. (2011) CD44 splice isoform switching in human and mouse epithelium is essential for epithelial-mesenchymal transition and breast cancer progression. *The Journal of clinical investigation*, **121**, 1064-1074.
19. Zhang, H., Brown, R.L., Wei, Y., Zhao, P., Liu, S., Liu, X., Deng, Y., Hu, X., Zhang, J., Gao, X.D. *et al.* (2019) CD44 splice isoform switching determines breast cancer stem cell state. *Genes & development*, **33**, 166-179.
20. Gregory, P.A., Bracken, C.P., Bert, A.G. and Goodall, G.J. (2008) MicroRNAs as regulators of epithelial-mesenchymal transition. *Cell cycle*, **7**, 3112-3118.
21. Ma, L., Young, J., Prabhala, H., Pan, E., Mestdagh, P., Muth, D., Teruya-Feldstein, J., Reinhardt, F., Onder, T.T., Valastyan, S. *et al.* (2010) miR-9, a MYC/MYCN-activated microRNA, regulates E-cadherin and cancer metastasis. *Nature cell biology*, **12**, 247-U252.
22. Taube, J.H., Malouf, G.G., Lu, E., Sphyris, N., Vijay, V., Ramachandran, P.P., Ueno, K.R., Gaur, S., Nicoloso, M.S., Rossi, S. *et al.* (2013) Epigenetic silencing of microRNA-203 is required for EMT and cancer stem cell properties. *Scientific reports*, **3**, 2687.
23. Hussey, G.S., Chaudhury, A., Dawson, A.E., Lindner, D.J., Knudsen, C.R., Wilce, M.C.J., Merrick, W.C. and Howe, P.H. (2011) Identification of an mRNP Complex Regulating Tumorigenesis at the Translational Elongation Step. *Molecular Cell*, **41**, 419-431.
24. Chaudhury, A., Cheema, S., Fachini, J.M., Kongchan, N., Lu, G., Simon, L.M., Wang, T., Mao, S., Rosen, D.G., Ittmann, M.M. *et al.* (2016) CELF1 is a central node in post-transcriptional regulatory programmes underlying EMT. *Nature communications*, **7**, 13362.
25. Kim, K., Lu, Z. and Hay, E.D. (2002) Direct evidence for a role of beta-catenin/LEF-1 signaling pathway in induction of EMT. *Cell Biol Int*, **26**, 463-476.
26. Leight, J.L., Wozniak, M.A., Chen, S., Lynch, M.L. and Chen, C.S. (2012) Matrix rigidity regulates a switch between TGF-beta1-induced apoptosis and epithelial-mesenchymal transition. *Molecular biology of the cell*, **23**, 781-791.
27. Miettinen, P.J., Ebner, R., Lopez, A.R. and Derynck, R. (1994) TGF-beta induced transdifferentiation of mammary epithelial cells to mesenchymal cells: involvement of type I receptors. *J Cell Biol*, **127**, 2021-2036.
28. Yang, M.H., Wu, M.Z., Chiou, S.H., Chen, P.M., Chang, S.Y., Liu, C.J., Teng, S.C. and Wu, K.J. (2008) Direct regulation of TWIST by HIF-1alpha promotes metastasis. *Nature cell biology*, **10**, 295-305.
29. Chakrabarti, R., Hwang, J., Andres Blanco, M., Wei, Y., Lukacisin, M., Romano, R.A., Smalley, K., Liu, S., Yang, Q., Ibrahim, T. *et al.* (2012) Elf5 inhibits the epithelial-mesenchymal transition in mammary gland development and breast cancer metastasis by transcriptionally repressing Snail2. *Nature cell biology*, **14**, 1212-1222.
30. Ray, H.J. and Niswander, L.A. (2016) Grainyhead-like 2 downstream targets act to suppress epithelial-to-mesenchymal transition during neural tube closure. *Development*, **143**, 1192-1204.
31. Roca, H., Hernandez, J., Weidner, S., McEachin, R.C., Fuller, D., Sud, S., Schumann, T., Wilkinson, J.E., Zaslavsky, A., Li, H. *et al.* (2013) Transcription factors OVOL1 and OVOL2 induce the mesenchymal to epithelial transition in human cancer. *PloS one*, **8**, e76773.

32. Bierie, B., Pierce, S.E., Kroeger, C., Stover, D.G., Pattabiraman, D.R., Thiru, P., Liu Donaher, J., Reinhardt, F., Chaffer, C.L., Keckesova, Z. *et al.* (2017) Integrin-beta4 identifies cancer stem cell-enriched populations of partially mesenchymal carcinoma cells. *Proceedings of the National Academy of Sciences of the United States of America*, **114**, E2337-E2346.
33. Grosse-Wilde, A., Fouquier d'Herouel, 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**, e0126522.
34. Jolly, M.K., Boareto, M., Huang, B., Jia, D., Lu, M., Ben-Jacob, E., Onuchic, J.N. and Levine, H. (2015) Implications of the Hybrid Epithelial/Mesenchymal Phenotype in Metastasis. *Front Oncol*, **5**, 155.
35. McFaline-Figueroa, J.L., Hill, A.J., Qiu, X., Jackson, D., Shendure, J. and Trapnell, C. (2019) A pooled single-cell genetic screen identifies regulatory checkpoints in the continuum of the epithelial-to-mesenchymal transition. *Nat Genet*, **51**, 1389-1398.
36. Meyer-Schaller, N., Cardner, M., Diepenbruck, M., Saxena, M., Tiede, S., Luond, F., Ivanek, R., Beerenwinkel, N. and Christofori, G. (2019) A Hierarchical Regulatory Landscape during the Multiple Stages of EMT. *Developmental Cell*, **48**, 539-+.
37. 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.
38. Cook, D.P. and Vanderhyden, B.C. (2020) Context specificity of the EMT transcriptional response. *Nature communications*, **11**.
39. Puram, S.V., Tirosh, I., Parikh, A.S., Patel, A.P., Yizhak, K., Gillespie, S., Rodman, C., Luo, C.L., Mroz, E.A., Emerick, K.S. *et al.* (2017) Single-Cell Transcriptomic Analysis of Primary and Metastatic Tumor Ecosystems in Head and Neck Cancer. *Cell*, **171**, 1611-1624 e1624.
40. George, J.T., Jolly, M.K., Xu, S.N., Somarelli, J.A. and Levine, H. (2017) Survival Outcomes in Cancer Patients Predicted by a Partial EMT Gene Expression Scoring Metric. *Cancer research*, **77**, 6415-6428.
41. Huang, R.Y., Wong, M.K., Tan, T.Z., Kuay, K.T., 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 & disease*, **4**, e915.
42. Subbalakshmi, A.R., Sahoo, S., Biswas, K. and Jolly, M.K. (2021) A Computational Systems Biology Approach Identifies SLUG as a Mediator of Partial Epithelial-Mesenchymal Transition (EMT). *Cells, tissues, organs*, 1-14.
43. Bocci, F., Tripathi, S.C., Mercedes, S.A.V., George, J.T., Casabar, J.P., Wong, P.K., Hanash, S.M., Levine, H., Onuchic, J.N. and Jolly, M.K. (2019) NRF2 activates a partial epithelial-mesenchymal transition and is maximally present in a hybrid epithelial/mesenchymal phenotype. *Integr Biol-Uk*, **11**, 251-263.
44. Taube, J.H., Herschkowitz, J.I., Komurov, K., Zhou, A.Y., Gupta, S., Yang, J., Hartwell, K., Onder, T.T., Gupta, P.B., Evans, K.W. *et al.* (2010) Core epithelial-to-mesenchymal transition interactome gene-expression signature is associated with claudin-low and metaplastic breast cancer subtypes. *Proceedings of the National Academy of Sciences of the United States of America*, **107**, 15449-15454.
45. Ghirlando, R. and Felsenfeld, G. (2016) CTCF: making the right connections. *Genes & development*, **30**, 881-891.

46. Arzate-Mejia, R.G., Recillas-Targa, F. and Corces, V.G. (2018) Developing in 3D: the role of CTCF in cell differentiation. *Development*, **145**.
47. Taube JH, S.N., Johnson KS, Reisenauer KN, Nesbit TA, Joseph R, Vijay GV, Sarkar TR, Bhangre NA, Song JJ, Chang JT, Lee MG, Soundararajan R, Mani SA. (2017) The H3K27me3-demethylase KDM6A is suppressed in breast cancer stem-like cells, and enables the resolution of bivalency during the mesenchymal-epithelial transition. *Oncotarget*.
48. Toneff, M.J., Sreekumar, A., Tinnirello, A., Hollander, P.D., Habib, S., Li, S., Ellis, M.J., Xin, L., Mani, S.A. and Rosen, J.M. (2016) The Z-cad dual fluorescent sensor detects dynamic changes between the epithelial and mesenchymal cellular states. *BMC Biol*, **14**, 47.
49. Byers, L.A., Diao, L., Wang, J., Saintigny, P., Girard, L., Peyton, M., Shen, L., Fan, Y., Giri, U., Tumula, P.K. *et al.* (2013) An epithelial-mesenchymal transition gene signature predicts resistance to EGFR and PI3K inhibitors and identifies Axl as a therapeutic target for overcoming EGFR inhibitor resistance. *Clinical cancer research : an official journal of the American Association for Cancer Research*, **19**, 279-290.
50. Guo, C.C., Majewski, T., Zhang, L., Yao, H., Bondaruk, J., Wang, Y., Zhang, S., Wang, Z., Lee, J.G., Lee, S. *et al.* (2019) Dysregulation of EMT Drives the Progression to Clinically Aggressive Sarcomatoid Bladder Cancer. *Cell Rep*, **27**, 1781-1793 e1784.
51. Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y. and Greenleaf, W.J. (2013) Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature methods*, **10**, 1213-+.
52. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. and Genome Project Data Processing, S. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078-2079.
53. Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H. and Glass, C.K. (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell*, **38**, 576-589.
54. Layer, R.M., Pedersen, B.S., DiSera, T., Marth, G.T., Gertz, J. and Quinlan, A.R. (2018) GIGGLE: a search engine for large-scale integrated genome analysis. *Nature methods*, **15**, 123-+.
55. Zheng, R., Wan, C., Mei, S., Qin, Q., Wu, Q., Sun, H., Chen, C.H., Brown, M., Zhang, X., Meyer, C.A. *et al.* (2019) Cistrome Data Browser: expanded datasets and new tools for gene regulatory analysis. *Nucleic acids research*, **47**, D729-D735.
56. Mei, S., Qin, Q., Wu, Q., Sun, H., Zheng, R., Zang, C., Zhu, M., Wu, J., Shi, X., Taing, L. *et al.* (2017) Cistrome Data Browser: a data portal for ChIP-Seq and chromatin accessibility data in human and mouse. *Nucleic acids research*, **45**, D658-D662.
57. Liberzon, A., Birger, C., Thorvaldsdottir, H., Ghandi, M., Mesirov, J.P. and Tamayo, P. (2015) The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst*, **1**, 417-425.
58. Fritz, A.J., Ghule, P.N., Boyd, J.R., Tye, C.E., Page, N.A., Hong, D., Shirley, D.J., Weinheimer, A.S., Barutcu, A.R., Gerrard, D.L. *et al.* (2018) Intranuclear and higher-order chromatin organization of the major histone gene cluster in breast cancer. *Journal of Cellular Physiology*, **233**, 1278-1290.
59. Amemiya, H.M., Kundaje, A. and Boyle, A.P. (2019) The ENCODE Blacklist: Identification of Problematic Regions of the Genome. *Scientific reports*, **9**, 9354.
60. Soule, H.D., Maloney, T.M., Wolman, S.R., Peterson, W.D., Jr., Brenz, R., McGrath, C.M., Russo, J., Pauley, R.J., Jones, R.F. and Brooks, S.C. (1990) Isolation and characterization of a spontaneously immortalized human breast epithelial cell line, MCF-10. *Cancer research*, **50**, 6075-6086.

61. Gilles, C., Polette, M., Zahm, J.M., Tournier, J.M., Volders, L., Foidart, J.M. and Birembaut, P. (1999) Vimentin contributes to human mammary epithelial cell migration. *Journal of Cell Science*, **112**, 4615-4625.
62. Zhang, J., Tian, X.J., Zhang, H., Teng, Y., Li, R., Bai, F., Elankumaran, S. and Xing, J. (2014) TGF-beta-induced epithelial-to-mesenchymal transition proceeds through stepwise activation of multiple feedback loops. *Sci Signal*, **7**, ra91.
63. Liu, Y., Xue, M., Du, S., Feng, W., Zhang, K., Zhang, L., Liu, H., Jia, G., Wu, L., Hu, X. *et al.* (2019) Author Correction: Competitive endogenous RNA is an intrinsic component of EMT regulatory circuits and modulates EMT. *Nature communications*, **10**, 5344.
64. Xu, J., Lamouille, S. and Derynck, R. (2009) TGF-beta-induced epithelial to mesenchymal transition. *Cell Res*, **19**, 156-172.
65. Deshmukh, A.P., Vasaikar, S.V., Tomczak, K., Tripathi, S., den Hollander, P., Arslan, E., Chakraborty, P., Soundararajan, R., Jolly, M.K., Rai, K. *et al.* (2021) Identification of EMT signaling cross-talk and gene regulatory networks by single-cell RNA sequencing. *Proceedings of the National Academy of Sciences of the United States of America*, **118**.
66. Dontu, G., Abdallah, W.M., Foley, J.M., Jackson, K.W., Clarke, M.F., Kawamura, M.J. and Wicha, M.S. (2003) In vitro propagation and transcriptional profiling of human mammary stem/progenitor cells. *Genes & development*, **17**, 1253-1270.
67. Stingl, J., Eirew, P., Ricketson, I., Shackleton, M., Vaillant, F., Choi, D., Li, H.I. and Eaves, C.J. (2006) Purification and unique properties of mammary epithelial stem cells. *Nature*, **439**, 993-997.
68. Chakraborty, P., George, J.T., Tripathi, S., Levine, H. and Jolly, M.K. (2020) Comparative Study of Transcriptomics-Based Scoring Metrics for the Epithelial-Hybrid-Mesenchymal Spectrum. *Front Bioeng Biotech*, **8**.
69. Ribeiro, A.S. and Paredes, J. (2014) P-Cadherin Linking Breast Cancer Stem Cells and Invasion: A Promising Marker to Identify an "Intermediate/Metastable" EMT State. *Front Oncol*, **4**, 371.
70. Aiello, N.M., Maddipati, R., Norgard, R.J., Balli, D., Li, J., Yuan, S., Yamazoe, T., Black, T., Sahmoud, A., Furth, E.E. *et al.* (2018) EMT Subtype Influences Epithelial Plasticity and Mode of Cell Migration. *Dev Cell*, **45**, 681-695 e684.
71. Bracken, C.P., Gregory, P.A., Kolesnikoff, N., Bert, A.G., Wang, J., Shannon, M.F. and Goodall, G.J. (2008) A double-negative feedback loop between ZEB1-SIP1 and the microRNA-200 family regulates epithelial-mesenchymal transition. *Cancer research*, **68**, 7846-7854.
72. Wellner, U., Schubert, J., Burk, U.C., Schmalhofer, O., Zhu, F., Sonntag, A., Waldvogel, B., Vannier, C., Darling, D., zur Hausen, A. *et al.* (2009) The EMT-activator ZEB1 promotes tumorigenicity by repressing stemness-inhibiting microRNAs. *Nature cell biology*, **11**, 1487-1495.
73. Klemm, S.L., Shipony, Z. and Greenleaf, W.J. (2019) Chromatin accessibility and the regulatory epigenome. *Nature reviews. Genetics*, **20**, 207-220.
74. Eichelberger, L., Saini, M., Moreno, H.D., Klein, C., Bartsch, J.M., Falcone, M., Reitberger, M., Espinet, E., Vogel, V., Graf, E. *et al.* (2020) Maintenance of epithelial traits and resistance to mesenchymal reprogramming promote proliferation in metastatic breast cancer. *bioRxiv*, 2020.2003.2019.998823.
75. Debruyne, D.N., Dries, R., Sengupta, S., Seruggia, D., Gao, Y., Sharma, B., Huang, H., Moreau, L., McLane, M., Day, D.S. *et al.* (2019) BORIS promotes chromatin regulatory interactions in treatment-resistant cancer cells. *Nature*, **572**, 676-680.
76. Braccioli, L. and de Wit, E. (2019) CTCF: a Swiss-army knife for genome organization and transcription regulation. *Essays Biochem*, **63**, 157-165.

77. Li, J., Huang, K., Hu, G., Babarinde, I.A., Li, Y., Dong, X., Chen, Y.S., Shang, L., Guo, W., Wang, J. *et al.* (2019) An alternative CTCF isoform antagonizes canonical CTCF occupancy and changes chromatin architecture to promote apoptosis. *Nature communications*, **10**, 1535.
78. Weintraub, A.S., Li, C.H., Zamudio, A.V., Sigova, A.A., Hannett, N.M., Day, D.S., Abraham, B.J., Cohen, M.A., Nabet, B., Buckley, D.L. *et al.* (2017) YY1 Is a Structural Regulator of Enhancer-Promoter Loops. *Cell*, **171**, 1573-+.
79. Meissner, A., Mikkelsen, T.S., Gu, H., Wernig, M., Hanna, J., Sivachenko, A., Zhang, X., Bernstein, B.E., Nusbaum, C., Jaffe, D.B. *et al.* (2008) Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature*, **454**, 766-770.
80. Zarkoob, H., Taube, J.H., Singh, S.K., Mani, S.A. and Kohandel, M. (2013) Investigating the link between molecular subtypes of glioblastoma, epithelial-mesenchymal transition, and CD133 cell surface protein. *PloS one*, **8**, e64169.
81. Holder, A.M., Akcakanat, A., Adkins, F., Evans, K., Chen, H.Q., Wei, C.M., Milton, D.R., Li, Y.S., Do, K.A., Janku, F. *et al.* (2015) Epithelial to mesenchymal transition is associated with rapamycin resistance. *Oncotarget*, **6**, 19500-19513.
82. Stark, T.W., Hensley, P.J., Spear, A., Pu, H., Strup, S.S. and Kyprianou, N. (2017) Predictive value of epithelial-mesenchymal-transition (EMT) signature and PARP-1 in prostate cancer radioresistance. *Prostate*, **77**, 1583-1591.
83. Thompson, J.C., Hwang, W.T., Davis, C., Deshpande, C., Jeffries, S., Rajpurohit, Y., Krishna, V., Smirnov, D., Verona, R., Lorenzi, M.V. *et al.* (2020) Gene signatures of tumor inflammation and epithelial-to-mesenchymal transition (EMT) predict responses to immune checkpoint blockade in lung cancer with high accuracy. *Lung Cancer*, **139**, 1-8.
84. Jia, W., Deshmukh, A., Mani, S.A., Jolly, M.K. and Levine, H. (2019) A possible role for epigenetic feedback regulation in the dynamics of the epithelial-mesenchymal transition (EMT). *Phys Biol*, **16**, 066004.
85. Ye, X., Tam, W.L., Shibue, T., Kaygusuz, Y., Reinhardt, F., Ng Eaton, E. and Weinberg, R.A. (2015) Distinct EMT programs control normal mammary stem cells and tumour-initiating cells. *Nature*, **525**, 256-260.
86. Addison, J.B., Voronkova, M.A., Fugett, J.H., Lin, C.C., Linville, N.C., Trinh, B., Livengood, R.H., Smolkin, M.B., Schaller, M.D., Ruppert, J.M. *et al.* (2021) Functional Hierarchy and Cooperation of EMT Master Transcription Factors in Breast Cancer Metastasis. *Mol Cancer Res*.
87. Le, T.L., Yap, A.S. and Stow, J.L. (1999) Recycling of E-cadherin: a potential mechanism for regulating cadherin dynamics. *J Cell Biol*, **146**, 219-232.
88. Padmanaban, V., Krol, I., Suhail, Y., Szczerba, B.M., Aceto, N., Bader, J.S. and Ewald, A.J. (2019) E-cadherin is required for metastasis in multiple models of breast cancer. *Nature*, **573**, 439-444.
89. Pujadas, E. and Feinberg, A.P. (2012) Regulated noise in the epigenetic landscape of development and disease. *Cell*, **148**, 1123-1131.
90. 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-beta-induced EMT of Ras-transformed mammary gland epithelial cells. *Scientific reports*, **7**, 1166.
91. Latil, M., Nassar, D., Beck, B., Boumahdi, S., Wang, L., Brisebarre, A., Dubois, C., Nkusi, E., Lenglez, S., Checinska, A. *et al.* (2017) Cell-Type-Specific Chromatin States Differentially Prime Squamous Cell Carcinoma Tumor-Initiating Cells for Epithelial to Mesenchymal Transition. *Cell Stem Cell*, **20**, 191-204 e195.
92. Kaiser, V.B., Taylor, M.S. and Semple, C.A. (2016) Mutational Biases Drive Elevated Rates of Substitution at Regulatory Sites across Cancer Types. *PLoS genetics*, **12**, e1006207.

93. Flavahan, W.A., Drier, Y., Liao, B.B., Gillespie, S.M., Venteicher, A.S., Stemmer-Rachamimov, A.O., Suva, M.L. and Bernstein, B.E. (2016) Insulator dysfunction and oncogene activation in IDH mutant gliomas. *Nature*, **529**, 110-+.
94. Zhao, L., Yang, Y., Yin, S., Yang, T., Luo, J., Xie, R., Long, H., Jiang, L. and Zhu, B. (2017) CTCF promotes epithelial ovarian cancer metastasis by broadly controlling the expression of metastasis-associated genes. *Oncotarget*, **8**, 62217-62230.
95. Zhang, B., Zhang, Y., Zou, X., Chan, A.W., Zhang, R., Lee, T.K., Liu, H., Lau, E.Y., Ho, N.P., Lai, P.B. *et al.* (2017) The CCCTC-binding factor (CTCF)-forkhead box protein M1 axis regulates tumour growth and metastasis in hepatocellular carcinoma. *J Pathol*, **243**, 418-430.

Epithelial-mesenchymal plasticity through loss of CTCF motif accessibility and protein expression

Kelsey S. Johnson¹, Shaimaa Hussein², Priyanka Chakraborty³, Arvind Muruganantham¹, Sheridan Mikhail¹, Giovanni Gonzalez¹, Shuxuan Song¹, Mohit Kumar Jolly³, Michael J. Toneff⁴, Mary Lauren Benton⁵, Yin C. Lin², Joseph H. Taube^{1,6#}

¹Department of Biology, Baylor University, Waco, TX, USA

²Baylor Institute for Immunology Research, Baylor Scott & White, Dallas, TX, USA

³Centre for BioSystems Science and Engineering, Indian Institute of Science, Bangalore, India

⁴Department of Biology, Widener University, Chester, PA, USA

⁵Department of Computer Science, Baylor University, Waco, TX, USA

⁶Dan L. Duncan Cancer Center, Houston, TX, USA

#Corresponding author: Joseph_Taube@baylor.edu

Corresponding Author email:

Joseph_Taube@baylor.edu

Keywords: EMT, MET, partial EMT, ATAC-seq, chromatin accessibility, CTCF, E-cadherin

Supplementary Figure 1. TGFβ treatment suppresses *CDH1* promoter reporter activity and increases *ZEB1* 3'UTR reporter activity. (A) FACS profiling of *ZEB1* 3'UTR reporter and *CDH1* promoter reporter activity. FACS quantification for Z1-GFP reporter (green panel), *CDH1*-RFP reporter (pink panel).

Supplementary Figure 2. Dynamics of TGF β -induced and TGF β -withdrawn chromatin accessibility. ATAC-seq peak profiles of select (A) epithelial (*CDH3*, *CDH1*, and *EPCAM*), (B) mesenchymal (*VIM*, *CDH2*, and *FN1*), and (C) EMT-TF (*ZEB1*, *SNAI1*, and *SNAI2*) genes at indicated short-term (top) and long-term (bottom) TGF β treatment models (n=2). Black rectangles below genes refer to ENCODE CTCF ChIP-seq peaks in human mammary epithelial cells (HMEC).

Table 1

Hallmark Gene Sets	
1	UV Response DN
2	Estrogen Response Early
3	Epithelial mesenchymal transition
4	TNF α Signaling via NF κ B
5	Apoptosis
6	IL2 STAT5 Signaling
7	Hypoxia
8	Mitotic Spindle
9	TGF β Signaling
10	Androgen Response
11	Inflammatory Response
12	Hedgehog Signaling
13	Apical Junction
14	p53 Pathway
15	Xenobiotic Metabolism
16	KRas Signaling Up

Table 2

Breast and Mammary-Specific Gene Sets		Study
A	Mammary Stem Cell Up	Lim et al., 2010
B	Luminal vs Mesenchymal Down	Charafe et al., 2006
C	Luminal vs Basal Down	Charafe et al., 2006
D	Pubertal Breast 4/5wk Up	McBryan et al., 2007
E	Basal vs Luminal	Farmer et al., 2005
F	Luminal B Down	Smid et al., 2008
G	Pubertal Breast 3/4wk Up	McBryan et al., 2007
H	Basal Up	Smid et al., 2008
I	Invasive Breast Cancer Down	Poola et al., 2005
J	Breast Cancer Relapse in Bone Down	Smid et al., 2008
K	Ductal Invasive Up	Schuetz et al., 2006
O	Basal Down	Smid et al., 2008
L	Medullary vs Ductal Breast Cancer Down	Bertucci et al., 2006
M	Luminal Mature Down	Lim et al., 2010
N	Ductal Invasive Down	Schuetz et al., 2006
P	ESR1 Laser Up	Yang et al., 2006
Q	Copy Number Down	Climent et al., 2007
R	Luminal vs Basal Up	Charafe et al., 2006

Supplementary Figure 3. EMT, mammary basal cell, and stemness gene sets are enriched in TGF β -induced and persistent peaks. (Table 1) Number key for GSEA MSigDB hits for Hallmark gene sets. **(Table 2)** Letter key for GSEA MSigDB hits for breast and mammary-specific gene sets. **(A)** GSEA MSigDB hits (based on number key in Table 1) for ATAC peaks in Hallmarks gene sets (left) and hits (based on letter key in Table 2) for breast and mammary-specific gene sets (right) in long-term TGF β -induced and -withdrawn conditions. Gray = 0d; yellow = 10d +TGF β ; purple = 10d +TGF β , 10d -TGF β . **(B)** GSEA MSigDB hits for genes annotated to ATAC peaks with persistent chromatin alterations following TGF β treatment and withdrawal in Hallmarks gene sets (top) and breast and mammary-specific gene sets (bottom)

Supplementary Figure 4. GIGGLE similarity scores of probable chromatin accessibility regulators at various stages of the EMT-MET spectrum in breast-specific cell lines. Peaks enriched at the indicated timepoints vs untreated cells were compared to genome interval files. **(A)** SMAD2/3-associated genome interval files show a diminished score as EMT progresses. **(B-E)** All scoring genome interval files for the indicated timepoint comparisons are shown. Data points refer to independent genome interval files queried by GIGGLE.

Supplementary Figure 6. Genes near accessible CTCF motifs are enriched for bivalent, neuronal, and signal transduction genes. (A) Venn diagram representing the number of genes adjacent to accessible CTCF motifs at the indicated timepoints. (Table S6) GSEA MSigDB gene sets determined to be highly-enriched in accessible CTCF motifs. (B) GSEA MSigDB enrichment (based on number key in Table S6) for top-10 enriched gene sets by condition circled in venn diagram to the right. (C) Western blot for histone modification changes in treated MCF10A cells.

Supplementary Figure 7. Change in gene expression upon CTCF manipulation in MCF10A cells. (A) Gene expression (normalized z-scores) was analyzed via Nanostring for indicated genes in shCtrl and shCTCF MCF10A cell lines with and without doxycycline. **(B)** Gene expression (normalized z-scores) was analyzed via Nanostring for indicated genes in GFP and CTCF overexpressing MCF10A cell lines.