Assessment of the transcriptional regulation of EMT and MET through accessible transposable elements

Doğa Eskier[1,2], Seray Yetkin[1,2], Nazmiye Arslan[2], Gökhan Karakülah*[1,2]†, and Hani Alotaibi*[1,2]†

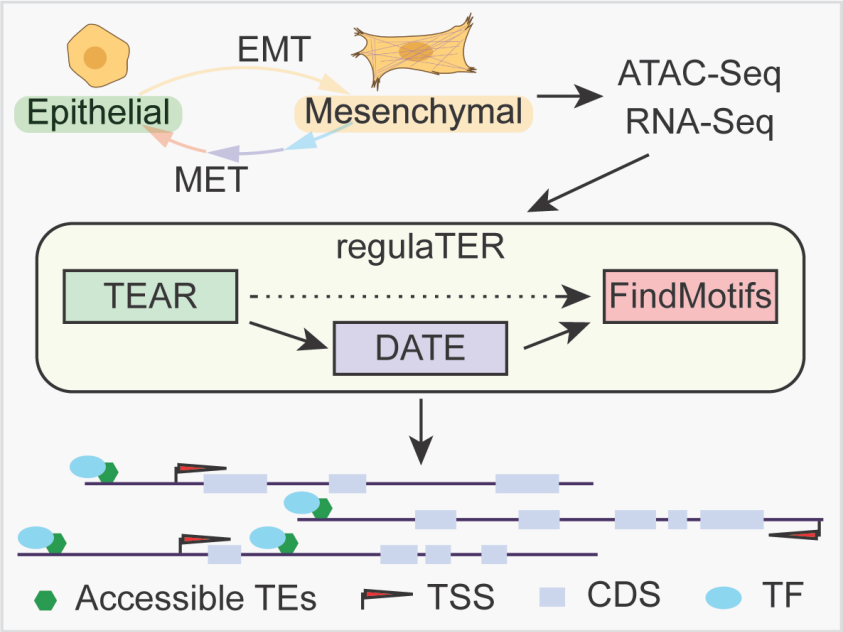
[1] Izmir International Biomedicine and Genome Institute, Dokuz Eylül University, İzmir, 35340, Turkey
[2] İzmir Biomedicine and Genome Center, İzmir, 35340, Turkey

† Joint authors

* To whom correspondence should be addressed. Tel: +902322994157; Email: hani.alotaibi@deu.edu.tr

Correspondence may also be addressed to: Tel: +905335798735; Email: gokhan.karakulah@ibg.edu.tr

Graphical Abstract



Abstract

Gene expression defines the identity of cells and is achieved following multiple levels of regulation at the transcriptional, translational, and a plethora of profound epigenetic mechanisms. Regulation of gene expression by transposable elements (TEs) is well documented. However, comprehensive analysis of the regulatory role of TEs is challenging due to the lack of dedicated analysis approaches to define their contribution. In this study, we sought to understand the contributions of TEs to the epithelial-to-mesenchymal (EMT) and mesenchymal-to-epithelial transition (MET) processes. To do so, we developed regulaTER, an R library that utilizes multi-omics data to predict the regulatory roles of accessible TEs in a phenotype of interest. Our results revealed novel insights into the intricate networks governing the regulation of EMT and MET and showed novel contributions of the MIR and B element subfamilies to the regulation of EMT and MET by the FoxA family of transcription factors. Our tool provides an essential asset for uncovering the impact of accessible TEs on the regulation of gene expression, with high flexibility to perform a range of other TE-centric analyses. regulaTER is publicly available at https://github.com/karakulahg/regulaTER.

Introduction

Transcriptional regulation is a complex process underlying many biological events and one of the key factors of organismal complexity. By varying the expression levels of the genes in the genome, cells with identical genotypes can achieve a wide range of phenotypes, both during development and beyond. Besides transcription factors (TF) and a multitude of epigenetic factors, transposable elements (TEs) are involved in transcriptional regulation (Gebrie 2023). TEs are repeat genomic elements with the unique ability to reposition and create novel copies of themselves in the genome. Initially described by Barbara McClintock in 1956 (McClintock 1956), the impact of TEs on the genome and the cell at large has received more attention in recent years, revealing many of their roles in various biological pathways and processes. TE contributions to genomic complexity can occur via various mechanisms, including the formation of novel chromosomal features (Pardue and DeBaryshe 2011), novel genes (Naville et al. 2016), and regulatory elements (Chuong, Elde, and Feschotte 2017). TEs are implicated in developmental processes, where they contribute to regulatory networks through transcription factor binding sites (TFBSs) for TFs involved in cellular differentiation (Bourque et al. 2008; Chuong et al. 2013). In addition to their roles gained via exaptation, TEs can also influence cellular phenotype via transposition and novel insertion events (Garcia-Perez et al. 2007; Baillie et al. 2011; Hancks and Kazazian 2016), and aberrant TE activity can lead to genetic diseases or tumorigenesis (Kazazian et al. 1988; Rangasamy et al. 2015; Scott et al. 2016; Yandim and Karakulah 2019). Because of the potential deleterious effects of TE activity and insertion, these elements often need to maintain a strict balance between activity and propagation or inhibition (Venner, Feschotte, and Biemont 2009; Robillard et al. 2016; Burgess 2017). Therefore, most organisms have developed multiple mechanisms to restrict TE activity within their genomes, especially transposition and novel insertions, mainly KRAB-ZFPs, which methylate TE sequences in a targeted fashion, and PIWI-interacting RNAs, which silence TEs and prevent mobilization (Aravin, Hannon, and Brennecke 2007; Molaro and Malik 2016; Yang, Wang, and Macfarlan 2017).

As repeat genomic elements with high polymorphic variability, analysis of TEs often requires specialized tools and workflows (Goerner-Potvin and Bourque 2018; Lanciano and Cristofari 2020). Despite the range of tools available, various perspectives are missing in the currently available TE toolbox. For example, while the association of TE-derived TFBSs with gene regulatory networks is well-known, there are currently no tools to identify the contributions of TEs to transcriptomic regulation in a given phenotype. To overcome the current challenges, we developed regulaTER, an R library, to analyze multi-omics data to identify the potential regulatory impact of TEs. RegulaTER utilizes genomic peaks, such as those obtained from ATAC-seq or ChIP-seq libraries, and genes of interest, such as those identified by differential expression analysis of RNA-seq libraries, and accurately identifies TEs with potential role in gene regulation. Here, we describe and validate the functions of regulaTER using an epithelial-mesenchymal transition (EMT) and mesenchymal-epithelial transition (MET) experimental setting. We also provide novel insights into the contribution of TEs in the intricate regulation of EMT and MET.

Materials and methods

Cell culture

To validate the function and outputs of regulaTER, we used the NMuMG (Normal Murine Mammary Gland) epithelial cell line as an EMT and MET model. The cells were maintained in high glucose DMEM containing 10% fetal bovine serum, 100U/ml penicillin, 0.1 mg/ml streptomycin, and 1% non-essential amino acids mixture in an incubator set to 37°C and 10% CO_2. As a supplement, 10 µg/ml insulin (Sigma-Aldrich) freshly added to the medium. EMT was induced by adding 2.5 ng/ml TGFβ3 (PeproTech) for 10 days. MET was induced by the removal of TGFβ3 from the medium. MET was observed for 4 days.

Gene expression analysis

qPCR. RNA was prepared using the Nucleospin RNA II kit (Macherey-Nagel) as recommended by the manufacturer. One microgram of total RNA was used for cDNA synthesis using the Maxima First Strand cDNA Synthesis kit (Thermo Scientific). Primers and probes used for qPCR were previously reported (Alotaibi et al. 2015). Quantitative RT-PCR was performed using the AMPIGENE qPCR probe mix (Enzo), and amplification was performed in the 7500 Real-time PCR system (Applied Biosystems) using the following cycling parameters: 95°C 5min, 40 cycles of 95°C

15s, 60°C 60s. Relative expression to the vehicle control (normalized to Gapdh expression) was calculated using the ΔΔCt method. Data represents at least three biological replicates.
Western Blot. NMuMG cells for each time point were harvested and lysed in RIPA buffer. Protein concentration was determined by the BCA Protein Assay Kit (Takara). Protein samples were then diluted with Laemmli Sample buffer. After electrophoresis, proteins were transferred to a nitrocellulose membrane, blocked for 1 h in blotto (Tris-buffered saline containing 0.5% Tween 20 and 5% nonfat milk powder), and incubated with anti-E-cad antibodies (BD Biosciences) overnight at 4 °C with shaking. Anti-GAPDH (Santa Cruz) was used as a loading control. Proteins were detected with IRDye® 800CW Goat anti-Mouse IgG and IRDye® 680RD Goat anti-Mouse IgG (H + L) antibodies (LI-COR Biosciences) visualized on an Odyssey® CLx Imaging System (LI-COR Biosciences).

Confocal Imaging. Cells were washed with PBS and fixed with 4% formaldehyde (Sigma) for 10 min at room temperature, washed twice with PBS, and permeabilized with 0.25% Triton X-100 (Sigma) for 5 min. After washing twice with PBS, cells were incubated with primary antibodies in PBS with 5% BSA (1:200) for one hour at room temperature. Afterward, Alexa594-conjugated secondary antibody was applied for an additional one hour after washing in PBS. Nuclei were visualized with DAPI (1:1000, Invitrogen) and mounted with Prolong Diamond Antifade Mountant solution (Thermo Scientific). Confocal microscopy was carried out with an LSM 880 microscope equipped with ZEN software (Zeiss). The antibodies used were anti-E-cadherin (BD Bioscience), Alexa Fluor 488 Phalloidin (Invitrogen), DAPI (Invitrogen), and Alexa Fluor 594 goat anti-mouse IgG (Invitrogen)

ATAC-seq

ATAC-seq reactions and library preparation were performed using the ATAC-sequencing Kit (Active Motif), following the manufacturer's protocols. Briefly, the nuclei of approximately 100000 cells were isolated and treated with Tn5 transposase, followed by adaptor ligation and PCR amplification of the libraries. The prepared libraries' quality control was checked using the Agilent Bioanalyzer 2100, and the libraries were sequenced using the Illumina NovaSeq 6000 sequencing platform (acquired by a third-party service provider). We performed the alignment and peak calling analysis of raw ATAC-seq reads using the nf-core atacseq (version 2.1.0) workflow, using the command "nextflow run nf-core/atacseq --input design.csv --genome mm10 -profile docker --save_reference". During the workflow, the reads were aligned to the UCSC Genome Browser mm10 mouse genome assembly using the BWA aligner, and the peaks were called in "broad" format using the MACS2 peak calling software.

RNA-seq

Isolation of RNA samples was performed with the Nucleospin RNA II kit (Macherey-Nagel) as recommended by the manufacturer. RNA-seq libraries were prepared using the Illumina Stranded Total RNA Prep with Ribo-Zero Plus kit following the manufacturer's protocols. Quality control of the prepared libraries was checked using the Agilent Bioanalyzer 2100, and the libraries were sequenced using the Illumina NovaSeq 6000 sequencing platform (acquired by a third-party service provider). The RNA-seq libraries were trimmed of low-quality bases and reads and possible adapter contamination using Trimmomatic (version 0.39) (Bolger, Lohse, and Usadel 2014). The remaining high-quality reads were aligned to the GENCODE GRCm38 (version M21) mouse genome assembly obtained from the GENCODE database, using HISAT2 (version 2.0.5) (Kim, Langmead, and Salzberg 2015), using the "hisat2-build" command to index the genome, and "hisat2 -x -U --dta -S -p 6" to align the reads. The resulting SAM alignment files were converted to binary BAM format using SAMtools utilities (version 1.10) (Danecek et al. 2021) "samtools view" and "samtools sort". The RNA-seq libraries were then quantified using the featureCounts function of the Rsubread R library (Liao, Smyth, and Shi 2019). Transcripts with FPKM values lower than 0.5 in both replicates of all samples were filtered out. We analyzed differential expression using the edgeR library (version 3.32.0) (Robinson, McCarthy, and Smyth 2010) using the count data of remaining transcripts. The count data was normalized using the TMM normalization method and then fit for the generalized linear model (GLM). We considered genes with abs($\log_2$FC) ≥ 1 and adjusted p-value (false discovery rate, FDR) ≤ 0.05 to be differentially expressed genes (DEG). The biological implications of the DEGs were summarized using Gene Set Enrichment Analysis (GSEA) with the GSEA function of the clusterProfiler package (version 3.18.1) (Yu et al. 2012) on

the hallmark genes of Mus musculus, with the line "GSEA(geneList = <named log2FC values>, minGSSize = 15, maxGSSize = 500, pvalueCutoff = 0.05, eps = 0, seed = TRUE, pAdjustMethod = "BH", TERM2GENE = dplyr::select(mm_hallmark_sets, gs_name, gene_symbol))".

The implementation of the regulaTER library

The regulaTER tool was created in the R environment (version 4.0.0). We used the biomartr package (version 1.0.2) (Drost and Paszkowski 2017) to import RepeatMasker annotations into the R environment for use in later calculations, GenomicRanges (version 1.42.0) (Lawrence et al. 2013) for processing of genomic peak ranges, and dplyr (version 1.1.1) for easier manipulation of genomic repeat category data (subfamily, family, type). For the calculation of TE enrichment, we used the binom.test and p.adjust functions available as part of the stats package of the base R installation. Genomic peaks were randomized for enrichment analyses using the BEDtools toolkit (version 2.29.0) (Quinlan and Hall 2010). For motif enrichment analyses, we used the HOMER genome analysis software (version 4.10.3) (Heinz et al. 2010) via marge (version 0.0.4.9999), an R environment API for HOMER. The library was documented, version-controlled, and uploaded to the GitHub software repository database via integrated Rstudio (version 2022.12.0) functions. The regulaTER library, as well as more detail on the use of its functions (Supplementary Figure 1).

Identification of TEs Enriched in Accessible Genomic Regions via the Transposable Elements Enriched in Accessible Regions (TEAR) Function. The primary inputs of the TEAR function are the genomic sequencing peak files annotated by ChIPseeker (Yu, Wang, and He 2015), a list of files with genomic coordinates corresponding to ChIPseeker annotations and the sizes of each chromosome and repeat annotations of the genome. In the function, TEs overlapping genomic peaks of interest showing differential accessibility are calculated via the Genomic Ranges package (“observed”). Afterward, the peak regions are randomized across the genome, with each peak being shuffled into a new location sharing the same genomic context, as compatible with the ChIPseeker annotations. Genomic region annotations for this randomization were obtained using the Table Browser utility of UCSC Genome Browser. TEs overlapping the randomized peak regions are then calculated. After the TE-randomized peak overlaps are calculated a number of times, determined by the numberOfShuffle input, the number of these overlaps is averaged (“expected”). To identify whether the “observed” number of TEs overlapping the peaks is significantly higher than the random distribution, we perform a negative binomial test (binom.test(observed, total TEs in genome, expected / total TEs in genome, alternative = "greater"). The p-values are corrected for multiple testing bias using the Benjamini-Hochberg correction method of the p.adjust function. We consider TEs with an FDR ≤ 0.05 and observed value ≥ 10 to be enriched in accessible regions (TEAR). The function performs the enrichment analyses for the TE subfamily, family, and type and outputs a list with three enrichment tables for each category. The TEAR function can accept peak files in narrowPeak, summits, or broadPeak formats generated by MACS2 and annotated by ChIPseeker (“format” parameter).

Identification of Accessible TEs Enriched Near Differentially Expressed Genes via the Differentially Expressed Gene Associated Accessible TEs (DATE) Function. The primary inputs of the DATE function are the TEAR function output, the repeat annotations and peaks used by TEAR, and a custom genomic coordinate file for the genes of interest. In this function, the number of accessible TEs in proximity or within the promoter regions of genes of interest is calculated for each TE subfamily via the Genomic Ranges package (“observed”). Following this calculation, a number of TEs of the same subfamily that do not lie in accessible genomic regions equal to the “observed” count are randomly selected, and these selected TEs are also overlapped with regions neighboring genes or promoter regions a number of times as determined by the user. The average number of such overlaps is calculated (“expected”). The enrichment for these gene-adjacent accessible TEs is calculated using the same negative binomial test parameters as in the TEAR function.

Identification of Genomic Motifs Contributing to Gene Regulatory Functions of Accessible TEs via the FindMotifs function. To identify genomic motifs that might contribute to the regulatory functions of identified TEs, we applied the FindMotifs function using the HOMER genomic sequence analysis tool via marge, an R environment API for HOMER. For the use of FindMotifs with TEAR function output, the background regions are calculated as TEs of the same subfamily outside accessible region peaks. For DATE function output, the background regions are calculated as accessible TEs

outside of gene neighboring regions, i.e. with a distance to the gene region greater than the distance parameter (5000 by default).

GO enrichment analysis

GO term enrichment analysis for TE-associated genes was performed using the enrichGO function of the clusterProfiler library (enrichGO(gene.list, OrgDb = org.Mm.eg.db, "SYMBOL", "BP")) and filtered for level 5 terms using the gofilter function of the same library (gofilter(enrichGO.results, level = 5)). For each TE analyzed, the associated genes were determined using the findOverlaps function of the GenomicRanges library to first identify accessible TEs in the subfamily of interest by finding TE instances overlapping ATAC-seq peaks showing differential accessibility in the sample, then by finding accessible TEs within 5000 nucleotides of upregulated genes (findOverlaps(maxgap = 5000, ignore.strand = TRUE)). The results were visualized using the circle_dat and GOBubble functions of the GOplot library (version 1.0.2) (Walter, Sanchez-Cabo, and Ricote 2015).

Statistical analyses and visualization

Statistical analysis for gene expression analysis was performed using one-way ANOVA and Tukey's multiple comparison tests. For data generated from computational studies, the R statistical computational environment is used for most statistical analyses and visualization purposes. We used pheatmap (version 1.0.12) to visualize the expression z-scores of genes during cellular differentiation using row-wise scaling. Summarized data was visualized using the ggplot2 (version 3.3.2) library.

UCSC Genome Browser images were generated using BigWig files representing each sample; files were uploaded to the UCSC genome browser (GRCm38/mm10), and tracks were visualized using the group-scaled and mean windowing function options. The ReMap track for each transcription factor was added using the same visualization parameters. Sample names and gene IDs were retyped in Adobe Illustrator for clarity and better resolution of the fonts.

Results

Functions of regulaTER

To overcome the limitations while studying the contribution of TEs in gene regulation (stated in the introduction), we developed regulaTER, an R library for the analysis of multi-omics data to identify the potential regulatory impact of TEs. regulaTER utilizes genomic peaks generated from a wide range of experiments, such as ChIP-seq, ATAC-seq, CUT&RUN, and meDIP-seq. Combined with gene expression data from RNA-seq, regulaTER delineates the contribution of TEs in gene regulation. The library has three core functions: **T**ransposable Elements **E**nriched in **A**ccessible **R**egions (TEAR), **D**ifferentially Expressed Gene Associated **A**ccessible **TE**s (DATE), and FindMotifs (Figure 1). The TEAR function primarily determines the enrichment of TEs in genomic peaks, while the DATE function investigates the association of the TEs identified by TEAR with differentially expressed genes. The FindMotifs function calculates or predicts any potential regulatory motifs enriched in a given TE instance and can follow either the TEAR or the DATE function. These functions have been optimized to identify the enrichment of accessible TEs and the transcription factor binding motifs in a given phenotype. Details of the functions are presented in Supplementary Figure 1.

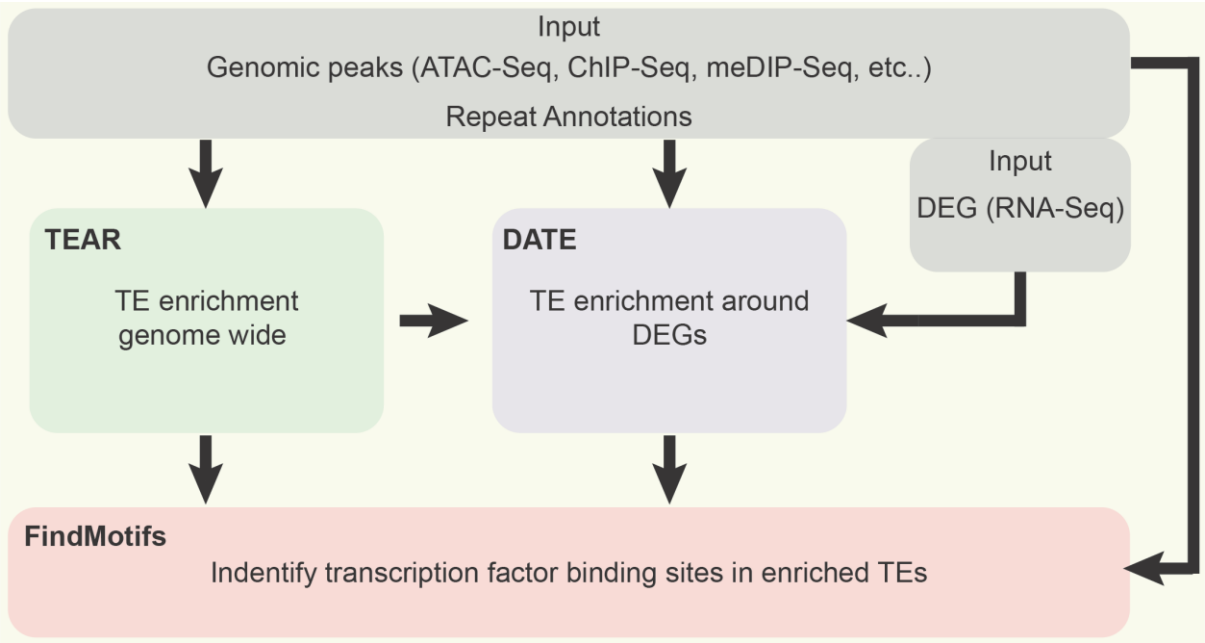


Figure 1. A representation of the combined functions of regulaTER. regulaTER can utilize inputs from sequencing data yielding genomic peaks fed into either TEAR or DATE. While TEAR identifies genome-wide TE enrichments, DATE focuses on TE enrichment around DEGs. FindMotif will then analyze the enrichment of transcription factor binding sites in enriched TEs.

Establishment and validation of EMT and MET cell model to test regulaTER functions

To showcase the functionality of regulaTER, we generated and used omics data representing EMT and MET, which are two biological processes with critical roles during embryonic development, wound healing, and tumorigenesis (Kalluri and Weinberg 2009; Gilbert, Vickaryous, and Viloria-Petit 2016). The advantage of using this model is twofold. First, the transcriptional regulation of the EMT program is well documented, with multiple TEs implicated in its regulation (Apostolou et al. 2015; Reyes-Reyes et al. 2017; Simo-Riudalbas et al. 2022), making it a powerful validator. Second, while EMT and MET are two closely related processes, the regulation of MET is distinct from that of EMT and is poorly understood, with no TEs known to be involved in its regulation, thus rendering MET a powerful model for the discovery of TEs involved in gene regulation (Pei et al. 2019; Yao, Dai, and Peng 2011).

We used the normal murine mouse mammary gland cell line NMuMG to generate expression and chromatin accessibility profiles. The cells were treated with TGFβ3 for 10 days (hereafter called MNM10, MNM as in **M**esenchymal **NM**uMG), and then TGFβ was removed from the medium by three washes with PBS; the cells at this point initiate MET and are incubated in regular culture medium to recover for a period of 4 days (designated PT100, PT as in **P**ost-**T**reatment). Collected samples from control, MNM10, and PT100 cells were used for RNA and protein isolation to validate gene expression changes conferring the quality of EMT and MET. By examining the expression levels of Cdh1 and Cdh2, we confirmed the cadherin switch, a bona fide hallmark of EMT and MET transitions (Figure 2A). Protein levels of E-cadherin showed similar changes to the RNA levels of Cdh1 (Figure 2B). In addition, confocal images revealed morphological changes associated with rearranged actin into stress fibers also validated the characteristic features expected during EMT and MET (Figure 2C).

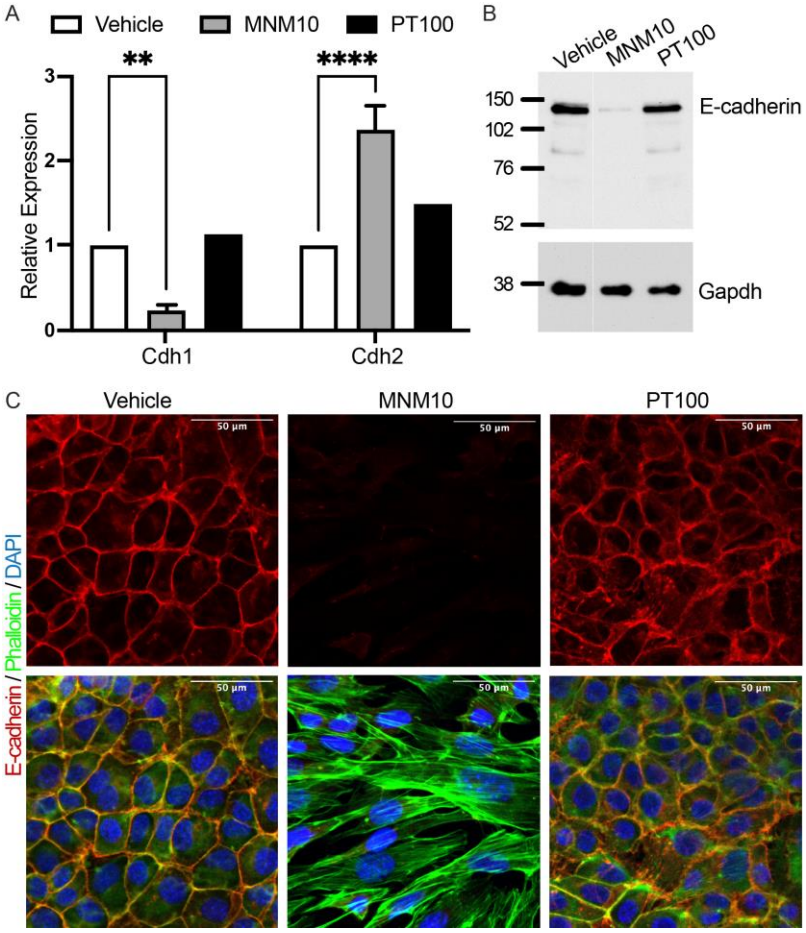


Figure 2. NMuMG cells successfully exhibited EMT and MET in response to TGFβ. A) qPCR analysis of the cadherin switch during EMT shows decreased Cdh1 and increased Cdh2 expression, which is reversed during MET. B) Changes in RNA levels were also reflected at the protein level, showing a decrease of E-cadherin expression during EMT, which is restored during MET. C) Confocal images of NMuMG cells showing the expression and localization of E-cadherin and Actin. Actin was visualized with Phalloidin, and the nuclei were stained with DAPI. The scale bar is 50 µm.

Collected and validated RNA from each sample was then used for RNA sequencing. We identified 3721 differentially expressed genes during EMT (2070 upregulated and 1651 downregulated, abs($\log_2$FC) ≥ 1, FDR ≤ 0.05). On the other hand, 5133 genes were differentially expressed during MET (2929 upregulated and 2204 downregulated, abs($\log_2$FC) ≥ 1, FDR ≤ 0.05) (Supplementary Tables 1 and 2). The changes in the expression profiles corresponded to a typical EMT and MET (Figure 3A and 3B). GSEA confirmed that the differentially expressed genes correlated perfectly with the EMT hallmark gene set (Figure 3C and 3D).

We then prepared ATAC-seq libraries using cells from the same culture plates used for RNA-seq. The changes in gene expression paralleled with the changes in chromatin accessibility as well. For example, the Cdh1 promoter and intronic enhancers showed marked differential accessibility during EMT and MET (Figure 3E). Similarly, the promoter of Cdh2 was more accessible during EMT (Figure 3F). Examining the ATAC-seq data revealed 23636 differentially accessible peaks during EMT and 12129 during MET; 1.2% of these peaks were associated with genes known to be involved in EMT and MET (distance to gene body ≤ 5000 bp).

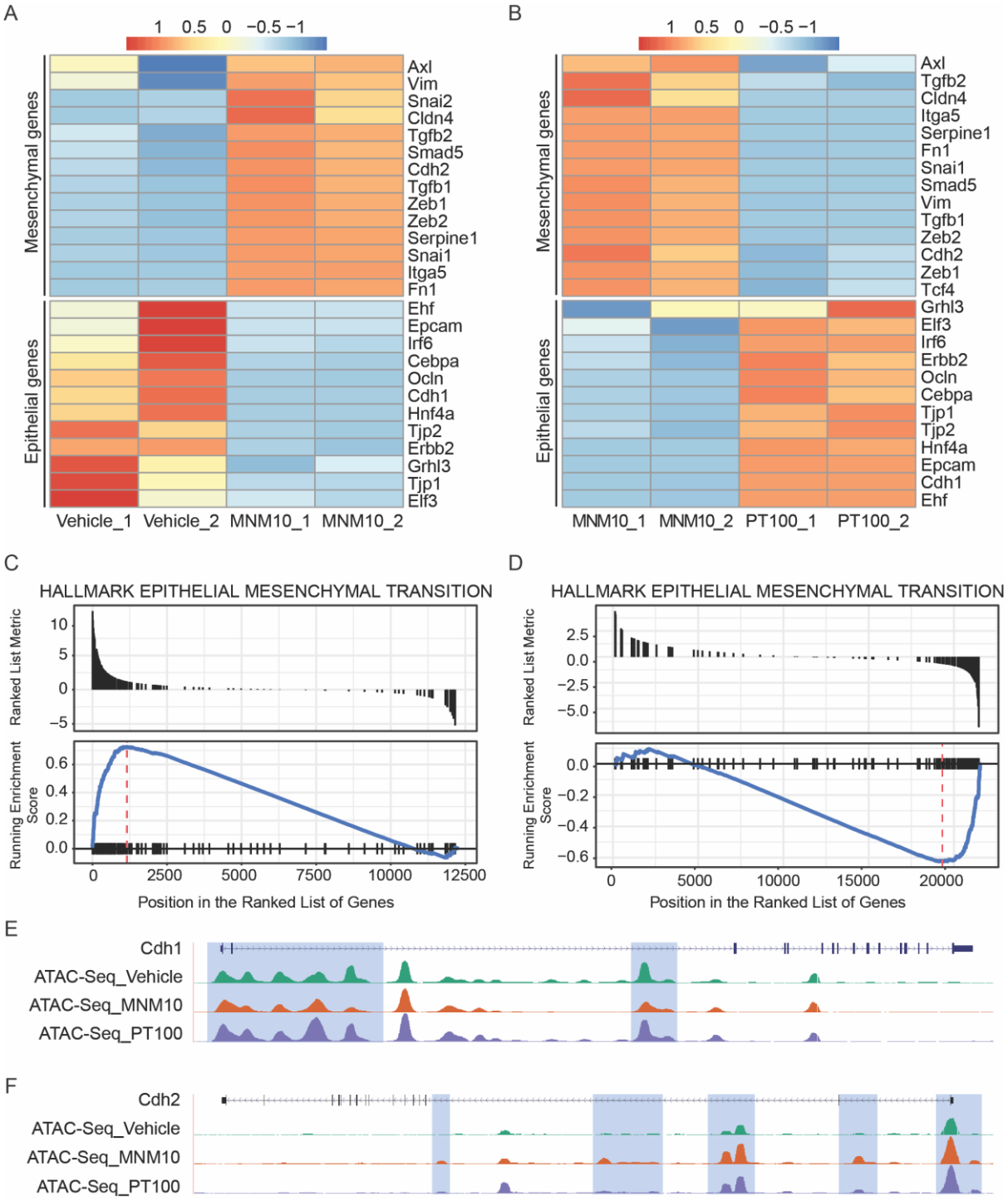


Figure 3. Gene expression analysis and differential chromatin accessibility during EMT and MET. Scaled expression heatmap of EMT markers in the (A) MNM10/Vehicle and (B) PT100/MNM10 comparisons. The colors of the cells indicate the z-score of the gene across the samples. GSEA analysis on the gene set of EMT hallmarks for Mus musculus in the (C) MNM10/Vehicle and (D) PT100/MNM10 comparison. The X-axis indicates the genes as ordered by the GSEA ranking list. The blue line in the bottom panel reflects the enrichment score of each gene. (E and F) UCSG genome browser images showing differential chromatin accessibility for Cdh1 (E) and Cdh2 (F).

MIR is the most significantly enriched TE family during EMT and MET.

To test and verify the TEAR function, we calculated differential peak accessibility of the ATAC-seq data by comparing MNM10 to Vehicle and PT100 to MNM10 using the count data of consensus

peaks as input for the exactTest function of edgeR. Peaks with $\log_2FC \geq 1$ and FDR ≤ 0.05 were considered to have increased accessibility. These peaks were annotated using the ChIPseeker package, which was then used as input for the TEAR function. The peaks were shuffled 10 times across the genome to calculate the expected value, as described in the methods section. As a result, 121 TE subfamilies were found to be enriched in ATAC-seq peaks in the MNM10 sample, while 45 subfamilies were enriched in the PT100 sample; 24 subfamilies were common in both samples (Supplementary Tables 3 and 4). We then examined the top 10 TE subfamilies and families enriched in accessible peaks to identify whether any specific TEs had a potentially significant contribution to the EMT and MET processes. Out of the top 10 subfamilies in EMT and MET, five were common in both processes (RMER6C, MIRb, RMER15, MIR, and RLTR48A). Additionally, the MIR family was the most enriched in EMT and the second most enriched in MET. Among the accessible MIR TEs in EMT, three subfamilies (MIR, MIRb, MIR3) were within 5000 bp of 1423 genes, including Zeb1, a known EMT regulator. Meanwhile, the accessible MIR and MIRb instances in MET were found in the proximity of 343 genes, including Cdh1 (Supplementary Table 5).

B elements are enriched near DEGs in EMT and MET

Following the identification of TEs enriched in genome-wide accessible regions, we examined accessible TE enrichment near DEGs. As DEGs are an important indicator of a given phenotype, we wanted to identify whether accessible TEs were enriched near genes of interest differently than those enriched across the genome. To do so, we used the DATE function, which uses a different enrichment method to accurately identify DEG-associated TEs.

We applied the DATE function using the output of the TEAR function to identify accessible TEs enriched in the neighboring regions of DEGs that show upregulation in either MNM10 or PT100 samples (regionSelection = “gene”, distance = 5000). We identified 32 TE subfamilies with enrichment of accessible instances near the DEGs in both MNM10 and PT100 (Supplementary Tables 6 and 7). We selected the 10 TE subfamilies showing the highest enrichment in both samples for further analysis. Of these 10 subfamilies, 5 were commonly enriched in both conditions, namely B3A, B1_Mus1, ID_B1, B4A, and B3 (Figure 4A). To identify whether the genes showing potential regulation by these TEs significantly impacted the EMT (or the MET) transcriptome, we performed GO enrichment analysis using the upregulated genes neighboring the accessible instances of these TEs. Our results revealed biological pathways associated with EMT and MET (Figure 4B). Notably, response to TGFβ, regulation of substrate adhesion, migration, and wound healing are among the most enriched biological processes identified in EMT. In contrast, MET-related enrichment processes were linked to developmental processes or proliferation. (Supplementary Table 8).

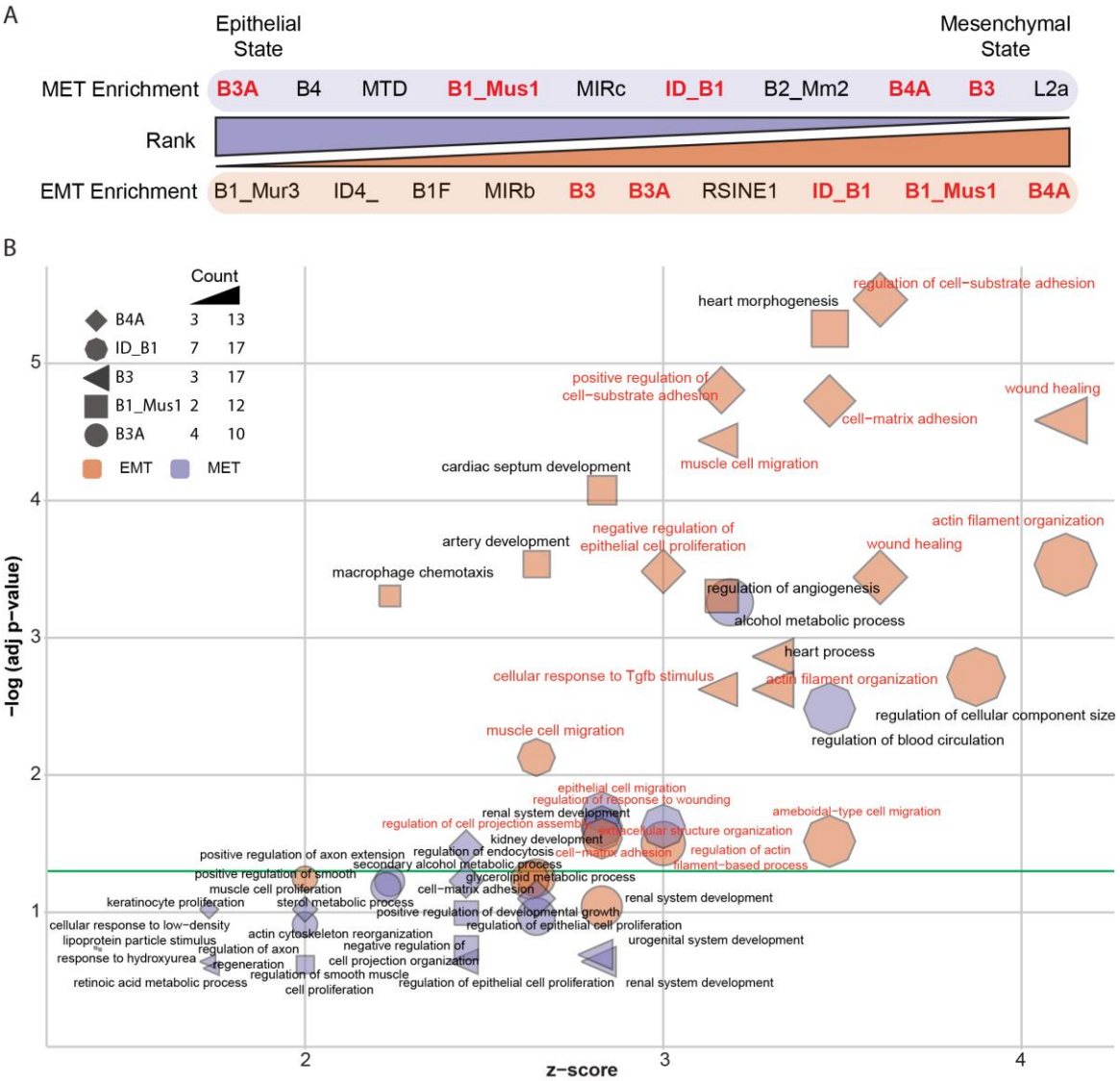


Figure 4. Five commonly enriched TEs show relevance to EMT and MET-related biological processes. (A) Representation of top 10 TE subfamilies enriched in the results of the DATE analysis in EMT (orange) and MET (purple). TE subfamilies are ranked according to their contribution to epithelial (left end of the bar) or mesenchymal (right end of the bar) states. The TE subfamilies written in red and bold text represent the subfamilies commonly enriched in EMT and MET. (B) Modified GO bubble plot of the GO enrichment results for the upregulated genes within 5000 bp of the accessible instances of the B element TE subfamilies indicated. The x-axis indicates the z-score of the term, which indicates whether the genes assigned to the term are more upregulated (z-score > 0) or downregulated (z < 0). The y-axis represents the negative log of the adjusted p-value of the term, with the horizontal green line corresponding to an adjusted p-value of 0.05. The shapes indicate TE subfamilies which the enrichment is calculated for. The color of the shape indicates the process the term is associated with, with orange shapes enriched in EMT, and purple shapes enriched in MET. The size of the shape indicates the number of TE-neighboring upregulated genes in the term. The names written in red indicate terms of interest for the EMT and MET processes.

FoxA1 and FoxA2 confer EMT and MET regulation through TE-contributed binding sites

In order to predict the potential roles of the identified TEs in gene regulation, we implemented the FindMotifs function using the output of the DATE function. The top 10 TEs enriched in EMT and MET were analyzed separately. We examined the FindMotifs results for the five TEs enriched near DEG clusters with significant contributions to EMT or MET. Our examination revealed that the

accessible B1_Mus1 and B4A instances near the EMT DEGs showed enrichment for FoxA1 and TEAD binding sites. To further validate the contribution of these TE-provided motifs to gene regulation, we visualized the upregulated EMT DEGs alongside ATAC-seq peaks and ChIP-seq peaks (ReMap track) in the UCSC Genome Browser. FoxA1 enrichment was observed within an accessible TE in the first intron of Tgfb2 (Figure 5A), while TEAD enrichment was localized to the promoter and 2 introns of Cdh2 (Figure 5B). Examining the motifs within MET-enriched TEs, binding sites for FoxA2 stand out among other transcription factor motifs. FoxA2 motifs were identified in accessible ID_B1 located in the promoter and the first intron of Agt (Figure 5C) and in B4A located in the intron of Rab17 (Figure 5D).

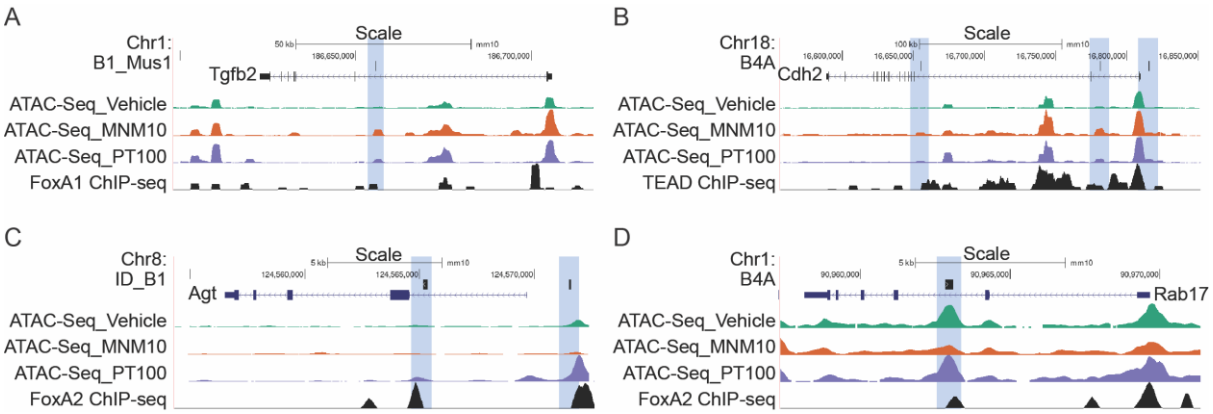


Figure 5. Regulation of EMT and MET by TEAD and the FoxA family of transcription factors through TE-contributed binding sites. UCSC Genome Browser images of (A) Tgfb2 and B1_Mus1, (B) Cdh2 and B4A, (C) Agt and ID_B1, and (D) Rab17 and B4A. Regions highlighted in blue indicate TE instances showing increased differential accessibility in EMT (A-B) and MET (C-D). The green tracks below the gene regions indicate ATAC-seq accessibility scores for the Vehicle sample, the orange tracks indicate accessibility for the MNM10 sample, and the purple tracks indicate the same for the PT100 sample. The black track indicates the ReMap ChIP-seq binding profiles for (A) FoxA1, (B) TEAD, and (C-D) FoxA2 TFs.

Discussion

In this study, we aimed to extend our understanding of the regulatory dynamics governing epithelial-mesenchymal transition (EMT) and mesenchymal-epithelial transition (MET), two fundamental processes in cellular biology that play crucial roles in development, tissue repair, and disease progression, particularly cancer metastasis (Brabletz et al. 2018; Chen, Han, and Pei 2012; Kalluri 2009; Kim et al. 2017; Nieto et al. 2016; Thiery et al. 2009; Yang et al. 2020). We focused on the potential regulatory role of TEs in these processes. To analyze our EMT-MET model, we found the need to automate our analyses due to the difficulties we faced when calculating the enrichment of TEs in accessible genomic regions. Such calculations require shuffling operations that are repeated more times than is feasible to perform manually (Metropolis et al. 1953). Therefore, we developed and applied regulaTER, a unique R library meticulously crafted to dissect the regulatory roles of TEs using multi-omics data.

Previously available methods have failed to meet our analysis criteria for custom multi-omics data. For example, Chandrashekar et al. have published GREAM, a web-based tool to predict TE contributions to gene regulation by identifying their over- or under-representation near a given set of genes (Chandrashekar, Dey, and Acharya 2015). However, GREAM does not consider whether the TE instances in the proximity of a gene are under a suitable chromatin state or epigenetic regulation to infer TF binding. Furthermore, it does not account for polymorphic insertions and can only perform analyses for specific reference genomes available on the web server. In addition, we have previously made available PlanTEnrichment, a tool to identify TEs enriched upstream of a list of given genes in 11 reference plant genomes (Karakulah and Suner 2017), and RTFAdb, a database of computationally predicted associations between TFBSs for 596 TFs identified using

publicly available ChIP-seq data and retrotransposon species in both mouse and human genomes (Karakulah 2018). As with GREAM, however, our previous tools were limited by the number of genomes they could analyze and the genomic peak and chromatin state data they were built to handle. On the other hand, regulaTER can quickly and efficiently perform repeated calculations for any properly formatted input and any given annotated genome without requiring tedious manual interference.

Combining a well-established transcriptomic analysis and a genome-wide accessibility profile, TEAR and DATE functions of regulaTER demonstrated the enrichment of TEs in accessible genomic regions and near differentially expressed genes during EMT and MET. The FindMotifs function predicted the transcription factor bind sites within the identified TEs. Of note, major transcription factors involved in the regulation of EMT were identified, such as FoxA1/2, TEAD2, TCF/LEF, Smad3/4, Twist2, and Zeb2 (Liu et al. 2021; Wang et al. 2013; Chi et al. 2022; Sanchez-Tillo et al. 2011; Park et al. 2008; Zavadil et al. 2004; Yang et al. 2004; Gregory et al. 2008; Sayan 2014). These transcription factors binding sites were in proximity of genes that shape the mesenchymal state during EMT, such as Cdh2, Fn1, Itga5, Wnt2b, Zeb1/2, and Tgfbr1 (Gonzalez and Medici 2014; Huang et al. 2015). These findings provide unprecedented insights into the regulation of EMT through TEs, corroborating previous findings by identifying TEs enriched in accessible genomic regions during EMT; this presents a valuable validation to the capabilities of regulaTER. Several studies suggested the involvement of TEs in EMT and, thus, metastasis progression. One study used a mutagenesis approach to identify TEs involved in the regulation of genes driving metastasis of HCC to facilitate drug identification (Kodama et al. 2016). In another study, LINE-1 elements enhanced the malignancy of human bronchial epithelial cells through interaction and co-regulation with TGFβ1 (Reyes-Reyes et al. 2017). Stemness and cellular plasticity during EMT in colorectal and breast cancer tumorigenesis were also shown to be regulated by the L1 retrotransposon L1-ORF1 (Apostolou et al. 2015). Moreover, identifying genomic insertions of AluSx1, LTR33, L2b, and MLT1F1 led to the activation of transposable element-driven transpochimeric gene transcripts such as POU5F1B, which in turn promoted colorectal cancer growth and metastasis (Simo-Riudalbas et al. 2022).

The validation of regulaTER outcomes for EMT-related TEs lays the groundwork for understanding the intricate regulatory networks governing MET, for which a similar understanding of its regulation is lagging (Alotaibi et al. 2015; Werth et al. 2010; Yang et al. 2020). Although TEs-dependent regulation of MET will not provide exclusive details on how it is regulated, it will add a layer of information (besides TFs and miRNAs) to this complex mechanism (Alotaibi et al. 2015; Li et al. 2010; Sengez et al. 2019; Bullock et al. 2012; Cao et al. 2017; Chen and Zhao 2021; Gregory et al. 2008; Imani et al. 2017). We identified several members of the B family of TEs to be enriched near genes upregulated during MET. These elements provided binding sites for FoxA2, which has been correlated with the inhibition of EMT (Liu et al. 2021; Zhang et al. 2015), indirectly suggesting a role in MET or the maintenance of the epithelial state. Targets of FoxA2 are also linked to EMT and MET regulation. For example, Rab17, a small GTPase of the RAS family, was shown to guard the epithelial state by inhibiting the Stat3/HIF1a/VEGF signaling pathway (Wang et al. 2020). Another target of FoxA2 is angiotensin (Agt), which can inhibit Tgfb1-induced EMT in keratinocytes (Jihu et al. 2024).

In conclusion, we present regulaTER with its ability to comprehensively assess the regulatory potential of TEs, addressing existing limitations in the field. While previous tools have predominantly focused on the association of TE-derived transcription factor binding sites with gene regulatory networks, regulaTER extends this paradigm by enabling the identification of TEs' contributions to transcriptomic regulation within specific phenotypes. This represents a paradigm shift, as it empowers researchers to unravel the nuanced regulatory roles of TEs in driving cellular transitions and phenotypic plasticity.

Data availability

RegulaTER is available on https://github.com/karakulahg/regulaTER. The ATAC-Seq and RNA-seq raw data is available at the Sequence Read Archive with the BioSample accession number PRJNA1129145.

Acknowledgements

We thank Ahmet Bursalı and Alırıza Arıbaş for their assistance with data analysis and visualization.

Author contribution: Doğa Eskier: Conceptualization, Data curation, Formal analysis, Methodology, Software, Validation, Visualization, Writing—original draft, Writing—review & editing. Seray Yetkin: Formal analysis, Investigation, Methodology, Writing—review & editing. Nazmiye Arslan: Conceptualization, Data curation, Formal analysis, Methodology, Software. Hani Alotaibi: Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Project Administration, Resources, Supervision, Visualization, Writing—original draft, Writing—review & editing. Gökhan Karakülah: Conceptualization, Data curation, Formal analysis, Funding acquisition, Methodology, Project Administration, Resources, Software, Validation, Visualization, Writing—original draft, Writing—review & editing.

Funding

This work was supported by the Türkiye Bilimsel ve Teknolojik Araştırma Kurumu (TÜBİTAK) grants 121Z132 to G.K. and 120N556 to H.A.

Conflict of interest statement. None declared.

References

Alotaibi, Hani, M. Felicia Basilicata, Huma Shehwana, Tyler Kosowan, Ilona Schreck, Christien Braeutigam, Ozlen Konu, Thomas Brabletz, and Marc P. Stemmler. 2015. 'Enhancer cooperativity as a novel mechanism underlying the transcriptional regulation of E-cadherin during mesenchymal to epithelial transition', *Biochimica Et Biophysica Acta-Gene Regulatory Mechanisms*, 1849: 731-42.

Apostolou, P., M. Toloudi, M. Chatziioannou, E. Kourtidou, G. Mimikakou, I. Vlachou, A. Chlichlia, and I. Papasotiriou. 2015. 'Involvement of retrotransposon L1 in stemness and cellular plasticity', *Cell Commun Adhes*, 22: 1-7.

Aravin, A. A., G. J. Hannon, and J. Brennecke. 2007. 'The Piwi-piRNA pathway provides an adaptive defense in the transposon arms race', *Science*, 318: 761-4.

Baillie, J. K., M. W. Barnett, K. R. Upton, D. J. Gerhardt, T. A. Richmond, F. De Sapio, P. M. Brennan, P. Rizzu, S. Smith, M. Fell, R. T. Talbot, S. Gustincich, T. C. Freeman, J. S. Mattick, D. A. Hume, P. Heutink, P. Carninci, J. A. Jeddeloh, and G. J. Faulkner. 2011. 'Somatic retrotransposition alters the genetic landscape of the human brain', *Nature*, 479: 534-7.

Bolger, A. M., M. Lohse, and B. Usadel. 2014. 'Trimmomatic: a flexible trimmer for Illumina sequence data', *Bioinformatics*, 30: 2114-20.

Bourque, G., B. Leong, V. B. Vega, X. Chen, Y. L. Lee, K. G. Srinivasan, J. L. Chew, Y. Ruan, C. L. Wei, H. H. Ng, and E. T. Liu. 2008. 'Evolution of the mammalian transcription factor binding repertoire via transposable elements', *Genome Res*, 18: 1752-62.

Brabletz, T., R. Kalluri, M. A. Nieto, and R. A. Weinberg. 2018. 'EMT in cancer', *Nat Rev Cancer*, 18: 128-34.

Bullock, M. D., A. E. Sayan, G. K. Packham, and A. H. Mirnezami. 2012. 'MicroRNAs: critical regulators of epithelial to mesenchymal (EMT) and mesenchymal to epithelial transition (MET) in cancer progression', *Biol Cell*, 104: 3-12.

Burgess, Darren J. 2017. 'Technique: CRISPR CAPTURE for multi-omic probing of genomic loci', *Nature Reviews Genetics*, 18: 2017.

Cao, J., J. Liu, J. Long, J. Fu, L. Huang, J. Li, C. Liu, X. Zhang, and Y. Yan. 2017. 'microRNA-23b suppresses epithelial-mesenchymal transition (EMT) and metastasis in hepatocellular carcinoma via targeting Pyk2', *Biomed Pharmacother*, 89: 642-50.

Chandrashekar, D. S., P. Dey, and K. K. Acharya. 2015. 'GREAM: A Web Server to Short-List Potentially Important Genomic Repeat Elements Based on Over-/Under-Representation in Specific Chromosomal Locations, Such as the Gene Neighborhoods, within or across 17 Mammalian Species', *Plos One*, 10: e0133647.

Chen, D., and H. Zhao. 2021. 'The inhibiting effects of microRNA-429 on the progression of pancreatic ductal adenocarcinoma cells by inhibiting epithelial mesenchymal transition', *Am J Transl Res*, 13: 3286-93.

Chen, J., Q. Han, and D. Pei. 2012. 'EMT and MET as paradigms for cell fate switching', *J Mol Cell Biol*, 4: 66-9.

Chi, M., J. Liu, C. Mei, Y. Shi, N. Liu, X. Jiang, C. Liu, N. Xue, H. Hong, J. Xie, X. Sun, B. Yin, X. Meng, and B. Wang. 2022. 'TEAD4 functions as a prognostic biomarker and triggers EMT via PI3K/AKT pathway in bladder cancer', *J Exp Clin Cancer Res*, 41: 175.

Chuong, E. B., N. C. Elde, and C. Feschotte. 2017. 'Regulatory activities of transposable elements: from conflicts to benefits', *Nat Rev Genet*, 18: 71-86.

Chuong, E. B., M. A. Rumi, M. J. Soares, and J. C. Baker. 2013. 'Endogenous retroviruses function as species-specific enhancer elements in the placenta', *Nat Genet*, 45: 325-9.

Danecek, P., J. K. Bonfield, J. Liddle, J. Marshall, V. Ohan, M. O. Pollard, A. Whitwham, T. Keane, S. A. McCarthy, R. M. Davies, and H. Li. 2021. 'Twelve years of SAMtools and BCFtools', *Gigascience*, 10.

Drost, H. G., and J. Paszkowski. 2017. 'Biomartr: genomic data retrieval with R', *Bioinformatics*, 33: 1216-17.

Garcia-Perez, J. L., M. C. Marchetto, A. R. Muotri, N. G. Coufal, F. H. Gage, K. S. O'Shea, and J. V. Moran. 2007. 'LINE-1 retrotransposition in human embryonic stem cells', *Hum Mol Genet*, 16: 1569-77.

Gebrie, A. 2023. 'Transposable elements as essential elements in the control of gene expression', *Mob DNA*, 14: 9.

Gilbert, R. W. D., M. K. Vickaryous, and A. M. Viloria-Petit. 2016. 'Signalling by Transforming Growth Factor Beta Isoforms in Wound Healing and Tissue Regeneration', *J Dev Biol*, 4.

Goerner-Potvin, P., and G. Bourque. 2018. 'Computational tools to unmask transposable elements', *Nat Rev Genet*, 19: 688-704.

Gonzalez, D. M., and D. Medici. 2014. 'Signaling mechanisms of the epithelial-mesenchymal transition', *Sci Signal*, 7: re8.

Gregory, P. A., A. G. Bert, E. L. Paterson, S. C. Barry, A. Tsykin, G. Farshid, M. A. Vadas, Y. Khew-Goodall, and G. J. Goodall. 2008. 'The miR-200 family and miR-205 regulate epithelial to mesenchymal transition by targeting ZEB1 and SIP1', *Nat Cell Biol*, 10: 593-601.

Hancks, D. C., and H. H. Kazazian, Jr. 2016. 'Roles for retrotransposon insertions in human disease', *Mob DNA*, 7: 9.

Heinz, S., C. Benner, N. Spann, E. Bertolino, Y. C. Lin, P. Laslo, J. X. Cheng, C. Murre, H. Singh, and C. K. Glass. 2010. 'Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities', *Mol Cell*, 38: 576-89.

Huang, R. Y., K. T. Kuay, T. Z. Tan, M. Asad, H. M. Tang, A. H. Ng, J. Ye, V. Y. Chung, and J. P. Thiery. 2015. 'Functional relevance of a six mesenchymal gene signature in epithelial-mesenchymal transition (EMT) reversal by the triple angiokinase inhibitor, nintedanib (BIBF1120)', *Oncotarget*, 6: 22098-113.

Imani, S., C. Wei, J. Cheng, M. A. Khan, S. Fu, L. Yang, M. Tania, X. Zhang, X. Xiao, X. Zhang, and J. Fu. 2017. 'MicroRNA-34a targets epithelial to mesenchymal transition-inducing transcription

factors (EMT-TFs) and inhibits breast cancer cell migration and invasion', *Oncotarget*, 8: 21362-79.

Jihu, Y., R. Leng, M. Liu, H. Ren, D. Xie, C. Yao, and H. Yan. 2024. 'Angiotensin (1-7) Inhibits Transforming Growth Factor-Beta1-Induced Epithelial-Mesenchymal Transition of Human Keratinocyte Hacat Cells in vitro', *Clin Cosmet Investig Dermatol*, 17: 1049-58.

Kalluri, R. 2009. 'EMT: when epithelial cells decide to become mesenchymal-like cells', *J Clin Invest*, 119: 1417-9.

Kalluri, R., and R. A. Weinberg. 2009. 'The basics of epithelial-mesenchymal transition', *J Clin Invest*, 119: 1420-8.

Karakulah, G. 2018. 'RTFAdb: A database of computationally predicted associations between retrotransposons and transcription factors in the human and mouse genomes', *Genomics*, 110: 257-62.

Karakulah, G., and A. Suner. 2017. 'PlanTEnrichment: A tool for enrichment analysis of transposable elements in plants', *Genomics*, 109: 336-40.

Kazazian, H. H., Jr., C. Wong, H. Youssoufian, A. F. Scott, D. G. Phillips, and S. E. Antonarakis. 1988. 'Haemophilia A resulting from de novo insertion of L1 sequences represents a novel mechanism for mutation in man', *Nature*, 332: 164-6.

Kim, D. H., T. Xing, Z. Yang, R. Dudek, Q. Lu, and Y. H. Chen. 2017. 'Epithelial Mesenchymal Transition in Embryonic Development, Tissue Repair and Cancer: A Comprehensive Overview', *J Clin Med*, 7.

Kim, D., B. Langmead, and S. L. Salzberg. 2015. 'HISAT: a fast spliced aligner with low memory requirements', *Nat Methods*, 12: 357-60.

Kodama, T., J. Y. Newberg, M. Kodama, R. Rangel, K. Yoshihara, J. C. Tien, P. H. Parsons, H. Wu, M. J. Finegold, N. G. Copeland, and N. A. Jenkins. 2016. 'Transposon mutagenesis identifies genes and cellular processes driving epithelial-mesenchymal transition in hepatocellular carcinoma', *Proc Natl Acad Sci U S A*, 113: E3384-93.

Lanciano, S., and G. Cristofari. 2020. 'Measuring and interpreting transposable element expression', *Nat Rev Genet*, 21: 721-36.

Lawrence, M., W. Huber, H. Pages, P. Aboyoun, M. Carlson, R. Gentleman, M. T. Morgan, and V. J. Carey. 2013. 'Software for computing and annotating genomic ranges', *PLoS Comput Biol*, 9: e1003118.

Li, R., J. Liang, S. Ni, T. Zhou, X. Qing, H. Li, W. He, J. Chen, F. Li, Q. Zhuang, B. Qin, J. Xu, W. Li, J. Yang, Y. Gan, D. Qin, S. Feng, H. Song, D. Yang, B. Zhang, L. Zeng, L. Lai, M. A. Esteban, and D. Pei. 2010. 'A mesenchymal-to-epithelial transition initiates and is required for the nuclear reprogramming of mouse fibroblasts', *Cell Stem Cell*, 7: 51-63.

Liao, Y., G. K. Smyth, and W. Shi. 2019. 'The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads', *Nucleic Acids Res*, 47: e47.

Liu, Y., T. Chen, M. Guo, Y. Li, Q. Zhang, G. Tan, L. Yu, and Y. Tan. 2021. 'FOXA2-Interacting FOXP2 Prevents Epithelial-Mesenchymal Transition of Breast Cancer Cells by Stimulating E-Cadherin and PHF2 Transcription', *Front Oncol*, 11: 605025.

McClintock, B. 1956. 'Controlling elements and the gene', *Cold Spring Harb Symp Quant Biol*, 21: 197-216.

Metropolis, Nicholas, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. 1953. 'Equation of State Calculations by Fast Computing Machines', *The Journal of Chemical Physics*, 21: 1087-92.

Molaro, A., and H. S. Malik. 2016. 'Hide and seek: how chromatin-based pathways silence retroelements in the mammalian germline', *Curr Opin Genet Dev*, 37: 51-58.

Naville, M., I. A. Warren, Z. Haftek-Terreau, D. Chalopin, F. Brunet, P. Levin, D. Galiana, and J. N. Volff. 2016. 'Not so bad after all: retroviruses and long terminal repeat retrotransposons as a source of new genes in vertebrates', *Clin Microbiol Infect*, 22: 312-23.

Nieto, M. A., R. Y. Huang, R. A. Jackson, and J. P. Thiery. 2016. 'Emt: 2016', *Cell*, 166: 21-45.

Pardue, M. L., and P. G. DeBaryshe. 2011. 'Retrotransposons that maintain chromosome ends', *Proc Natl Acad Sci U S A*, 108: 20317-24.

Park, S. M., A. B. Gaur, E. Lengyel, and M. E. Peter. 2008. 'The miR-200 family determines the epithelial phenotype of cancer cells by targeting the E-cadherin repressors ZEB1 and ZEB2', *Genes Dev*, 22: 894-907.

Pei, D., X. Shu, A. Gassama-Diagne, and J. P. Thiery. 2019. 'Mesenchymal-epithelial transition in development and reprogramming', *Nat Cell Biol*, 21: 44-53.

Quinlan, A. R., and I. M. Hall. 2010. 'BEDTools: a flexible suite of utilities for comparing genomic features', *Bioinformatics*, 26: 841-2.

Rangasamy, D., N. Lenka, S. Ohms, J. E. Dahlstrom, A. C. Blackburn, and P. G. Board. 2015. 'Activation of LINE-1 Retrotransposon Increases the Risk of Epithelial-Mesenchymal Transition and Metastasis in Epithelial Cancer', *Curr Mol Med*, 15: 588-97.

Reyes-Reyes, E. M., I. Aispuro, M. A. Tavera-Garcia, M. Field, S. Moore, I. Ramos, and K. S. Ramos. 2017. 'LINE-1 couples EMT programming with acquisition of oncogenic phenotypes in human bronchial epithelial cells', *Oncotarget*, 8: 103828-42.

Robillard, E., A. Le Rouzic, Z. Zhang, P. Capy, and A. Hua-Van. 2016. 'Experimental evolution reveals hyperparasitic interactions among transposable elements', *Proc Natl Acad Sci U S A*, 113: 14763-68.

Robinson, M. D., D. J. McCarthy, and G. K. Smyth. 2010. 'edgeR: a Bioconductor package for differential expression analysis of digital gene expression data', *Bioinformatics*, 26: 139-40.

Sanchez-Tillo, E., O. de Barrios, L. Siles, M. Cuatrecasas, A. Castells, and A. Postigo. 2011. 'beta-catenin/TCF4 complex induces the epithelial-to-mesenchymal transition (EMT)-activator ZEB1 to regulate tumor invasiveness', *Proc Natl Acad Sci U S A*, 108: 19204-9.

Sayan, A. E. 2014. 'Tumour-promoting role of EMT-inducing transcription factor ZEB1 in mantle cell lymphoma', *Cell Death Differ*, 21: 194-5.

Scott, E. C., E. J. Gardner, A. Masood, N. T. Chuang, P. M. Vertino, and S. E. Devine. 2016. 'A hot L1 retrotransposon evades somatic repression and initiates human colorectal cancer', *Genome Res*, 26: 745-55.

Sengez, B., I. Aygun, H. Shehwana, N. Toyran, S. Tercan Avci, O. Konu, M. P. Stemmler, and H. Alotaibi. 2019. 'The Transcription Factor Elf3 Is Essential for a Successful Mesenchymal to Epithelial Transition', *Cells*, 8.

Simo-Riudalbas, L., S. Offner, E. Planet, J. Duc, L. Abrami, S. Dind, A. Coudray, M. Coto-Llerena, C. Ercan, S. Piscuoglio, C. L. Andersen, J. B. Bramsen, and D. Trono. 2022. 'Transposon-activated POU5F1B promotes colorectal cancer growth and metastasis', *Nat Commun*, 13: 4913.

Thiery, J. P., H. Acloque, R. Y. Huang, and M. A. Nieto. 2009. 'Epithelial-mesenchymal transitions in development and disease', *Cell*, 139: 871-90.

Venner, S., C. Feschotte, and C. Biemont. 2009. 'Dynamics of transposable elements: towards a community ecology of the genome', *Trends Genet*, 25: 317-23.

Walter, W., F. Sanchez-Cabo, and M. Ricote. 2015. 'GOplot: an R package for visually combining expression data with functional analysis', *Bioinformatics*, 31: 2912-4.

Wang, H., C. A. Meyer, T. Fei, G. Wang, F. Zhang, and X. S. Liu. 2013. 'A systematic approach identifies FOXA1 as a key factor in the loss of epithelial traits during the epithelial-to-mesenchymal transition in lung cancer', *BMC Genomics*, 14: 680.

Wang, M., W. Wang, J. Ding, J. Wang, and J. Zhang. 2020. 'Downregulation of Rab17 promotes cell proliferation and invasion in non-small cell lung cancer through STAT3/HIF-1alpha/VEGF signaling', *Thorac Cancer*, 11: 379-88.

Werth, M., K. Walentin, A. Aue, J. Schonheit, A. Wuebken, N. Pode-Shakked, L. Vilianovitch, B. Erdmann, B. Dekel, M. Bader, J. Barasch, F. Rosenbauer, F. C. Luft, and K. M. Schmidt-Ott. 2010. 'The transcription factor grainyhead-like 2 regulates the molecular composition of the epithelial apical junctional complex', *Development*, 137: 3835-45.

Yandim, C., and G. Karakulah. 2019. 'Dysregulated expression of repetitive DNA in ER+/HER2- breast cancer', *Cancer Genet*, 239: 36-45.

Yang, J., S. A. Mani, J. L. Donaher, S. Ramaswamy, R. A. Itzykson, C. Come, P. Savagner, I. Gitelman, A. Richardson, and R. A. Weinberg. 2004. 'Twist, a master regulator of morphogenesis, plays an essential role in tumor metastasis', *Cell*, 117: 927-39.

Yang, Jing, Parker Antin, Geert Berx, Cédric Blanpain, Thomas Brabletz, Marianne Bronner, Kyra Campbell, Amparo Cano, Jordi Casanova, Gerhard Christofori, Shoukat Dedhar, Rik Derynck, Heide L. Ford, Jonas Fuxe, Antonio García de Herreros, Gregory J. Goodall, Anna-Katerina Hadjantonakis, Ruby Y. J. Huang, Chaya Kalcheim, Raghu Kalluri, Yibin Kang, Yeesim Khew-Goodall, Herbert Levine, Jinsong Liu, Gregory D. Longmore, Sendurai A. Mani, Joan Massagué, Roberto Mayor, David McClay, Keith E. Mostov, Donald F. Newgreen, M. Angela Nieto, Alain Puisieux, Raymond Runyan, Pierre Savagner, Ben Stanger, Marc P. Stemmler, Yoshiko Takahashi, Masatoshi Takeichi, Eric Theveneau, Jean Paul Thiery, Erik W. Thompson, Robert A. Weinberg, Elizabeth D. Williams, Jianhua Xing, Binhua P. Zhou, Guojun Sheng, and E. M. T. International Association On behalf of the. 2020. 'Guidelines and definitions for research on epithelial–mesenchymal transition', *Nature Reviews Molecular Cell Biology*, 21: 341-52.

Yang, P., Y. Wang, and T. S. Macfarlan. 2017. 'The Role of KRAB-ZFPs in Transposable Element Repression and Mammalian Evolution', *Trends Genet*, 33: 871-81.

Yao, D., C. Dai, and S. Peng. 2011. 'Mechanism of the mesenchymal-epithelial transition and its relationship with metastatic tumor formation', *Mol Cancer Res*, 9: 1608-20.

Yu, G., L. G. Wang, Y. Han, and Q. Y. He. 2012. 'clusterProfiler: an R package for comparing biological themes among gene clusters', *OMICS*, 16: 284-7.

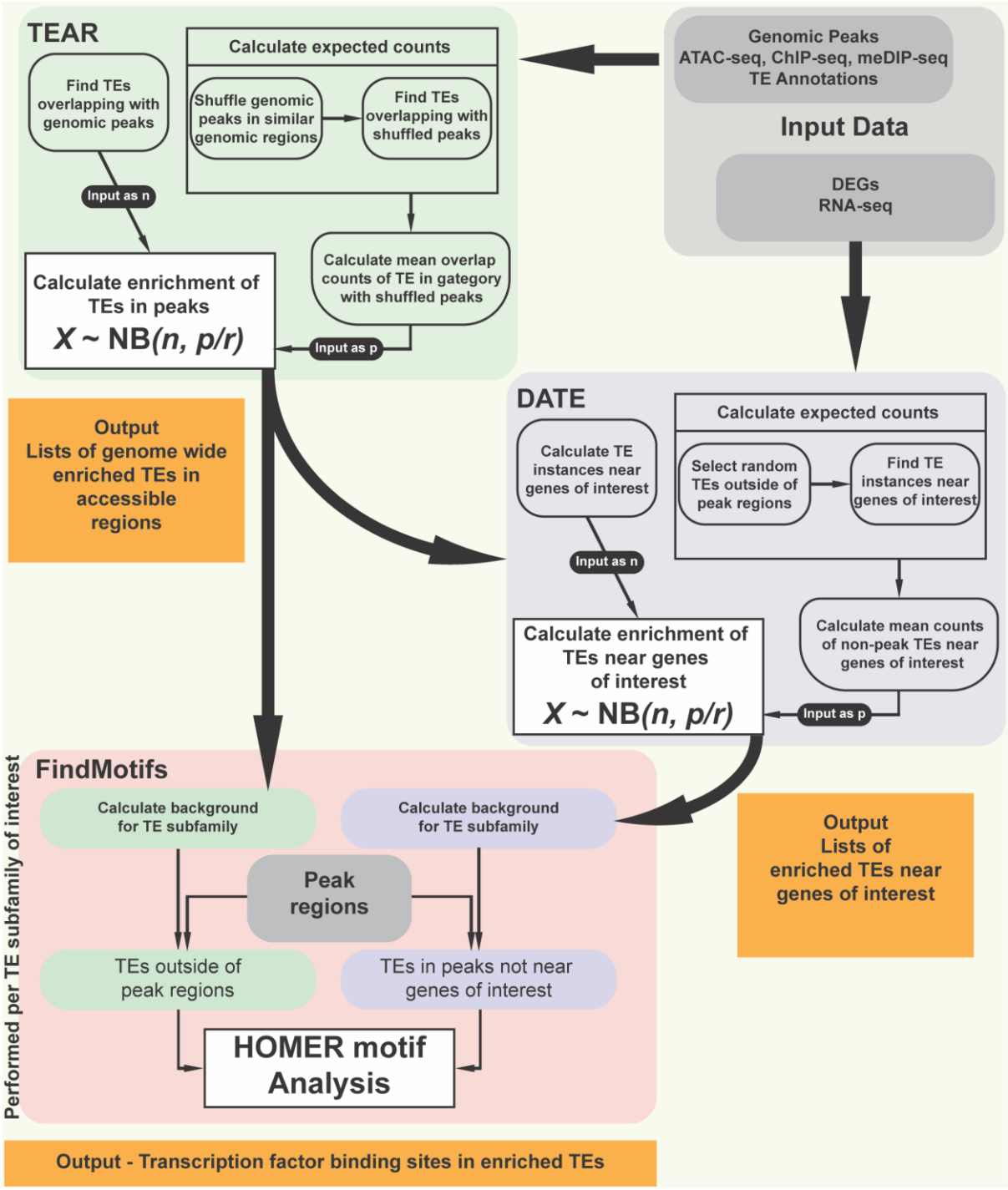
Yu, G., L. G. Wang, and Q. Y. He. 2015. 'ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization', *Bioinformatics*, 31: 2382-3.

Zavadil, J., L. Cermak, N. Soto-Nieves, and E. P. Bottinger. 2004. 'Integration of TGF-beta/Smad and Jagged1/Notch signalling in epithelial-to-mesenchymal transition', *EMBO J*, 23: 1155-65.

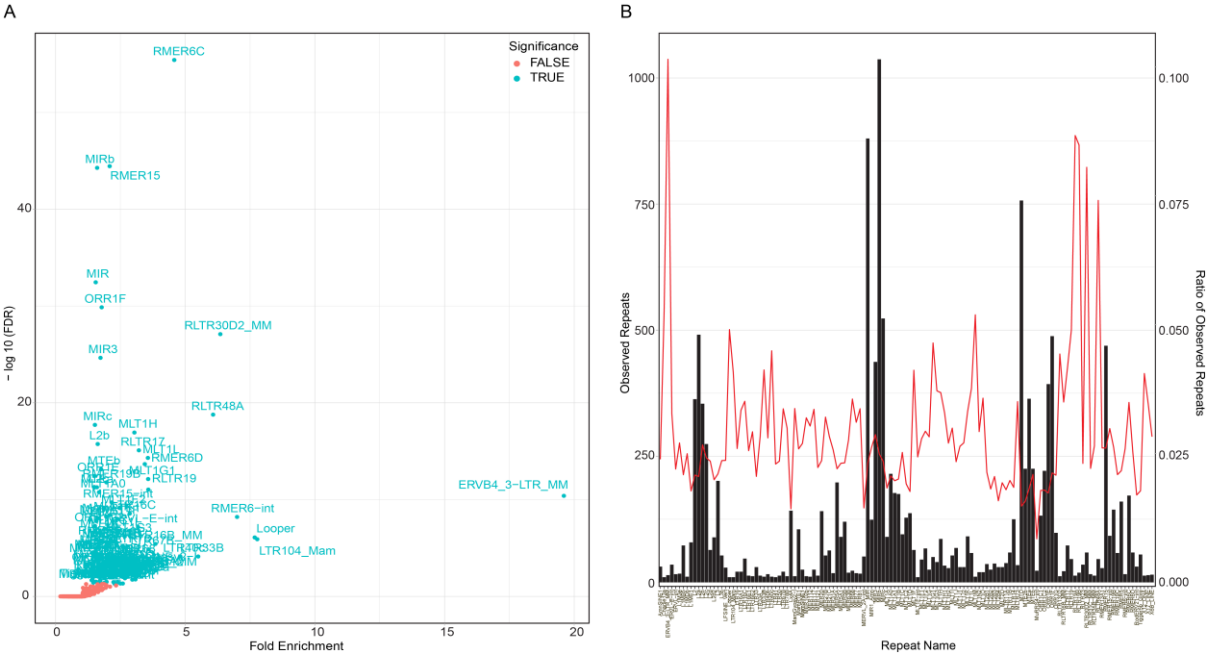
Zhang, Z., C. Yang, W. Gao, T. Chen, T. Qian, J. Hu, and Y. Tan. 2015. 'FOXA2 attenuates the epithelial to mesenchymal transition by regulating the transcription of E-cadherin and ZEB2 in human breast cancer', *Cancer Lett*, 361: 240-50.

Supplementary data

Supplementary Data are available at NAR online.



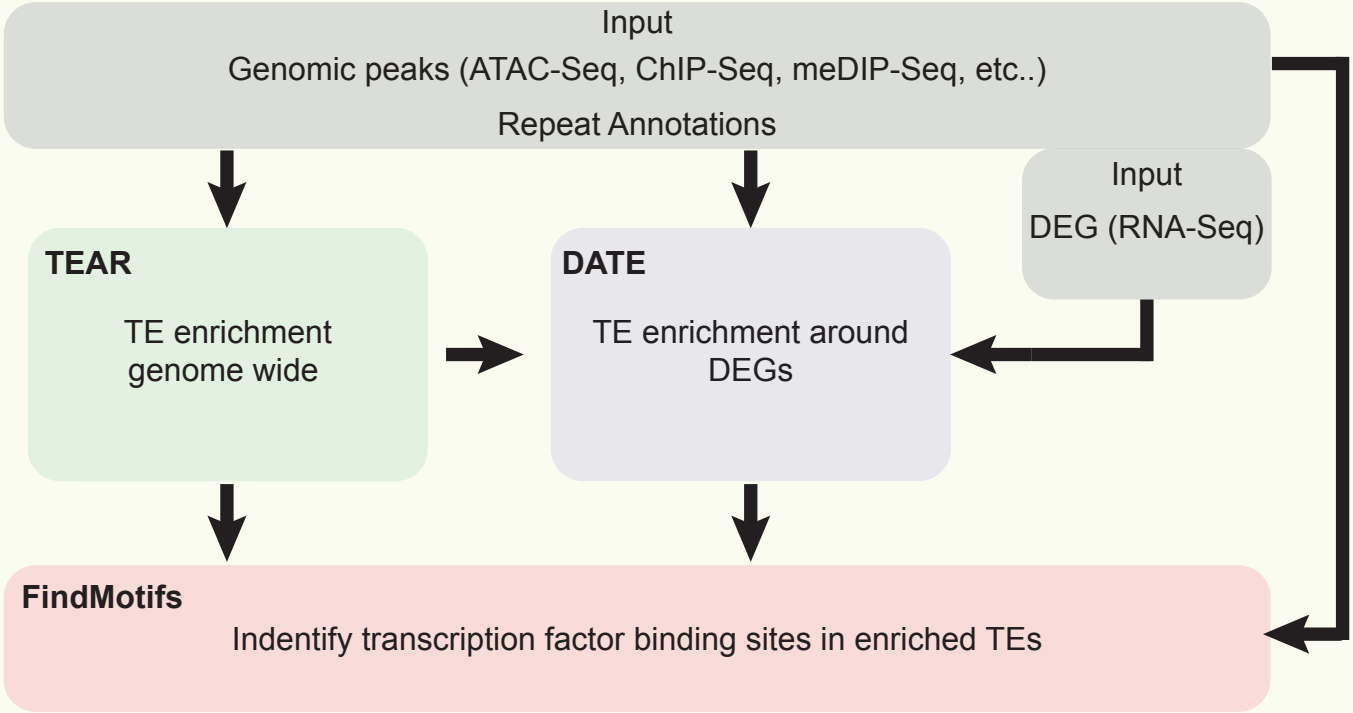
Supplementary Figure 1. Detailed outline of the regulaTER functions and workflow. The gray boxes represent primary user defined inputs (genomic peak data, DEG data). The green boxes indicate information regarding the TEAR function, while the purple boxes indicate information regarding the DATE function. The pink box represents the FindMotifs function. Orange boxes display the primary output of each function. Boxes with headers within the TEAR and DATE function descriptions each indicate a Monte Carlo simulation, which are repeated a number of times as determined by the user.

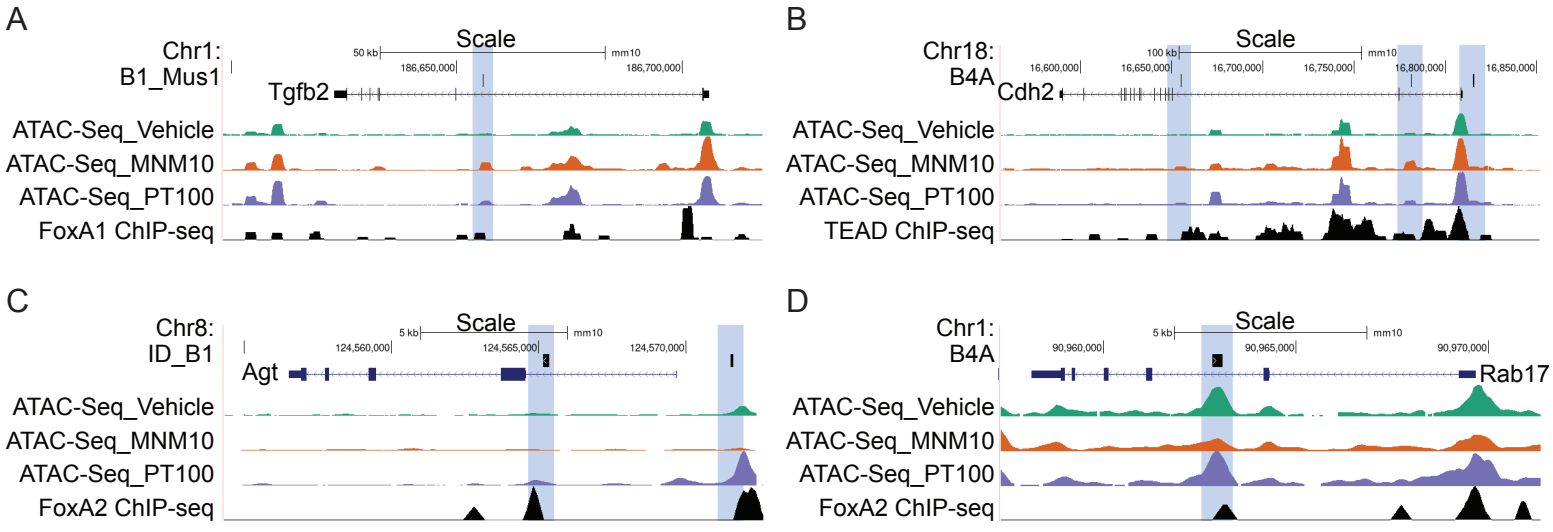


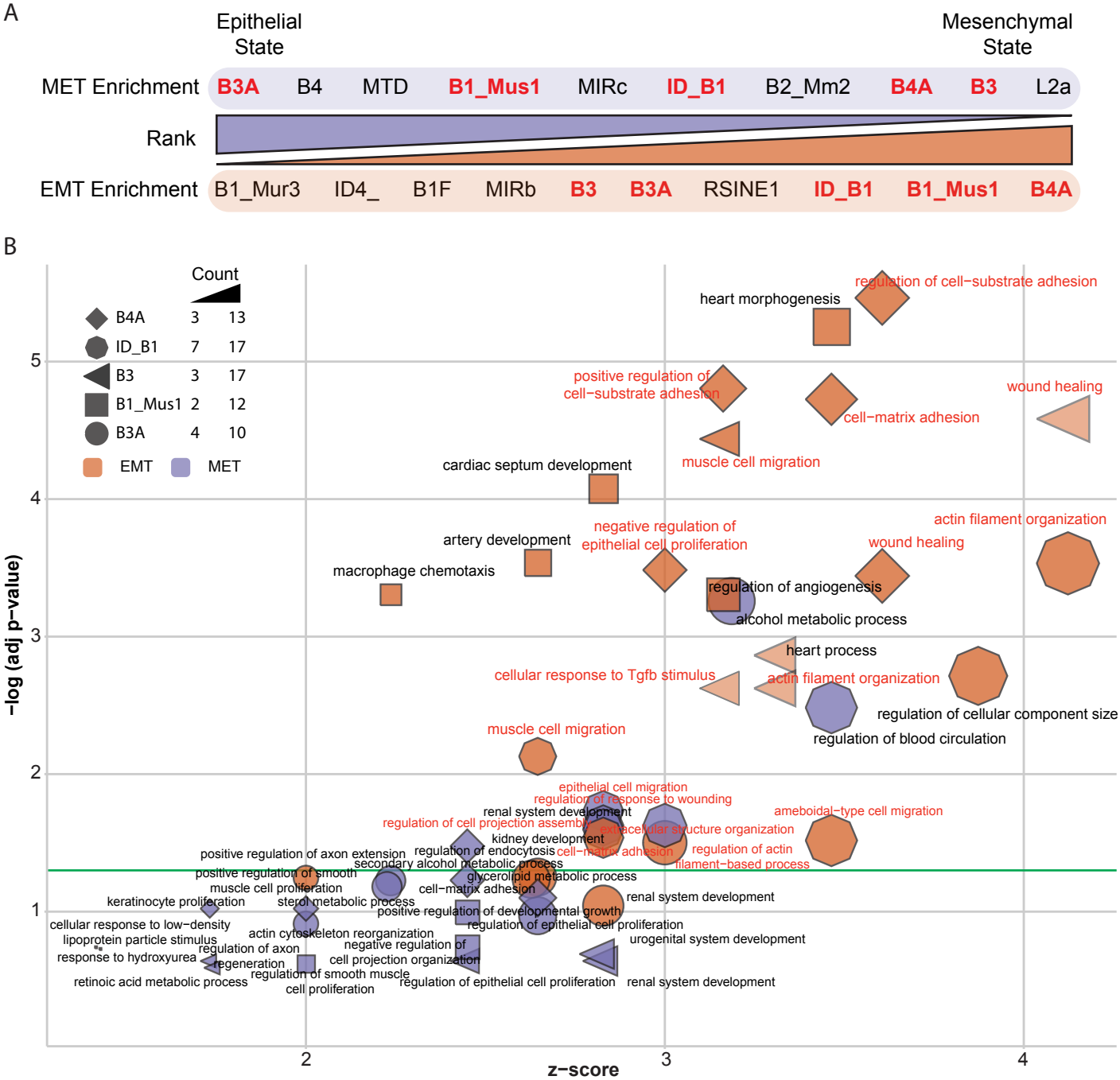
Supplementary Figure 2. The outputs of MakePlots function. (A) Exploratory scatterplot generated using TEAR function output. The X-axis shows the ratio of TEs observed to overlap peak regions to the expected number of such TEs. In contrast, the Y-axis shows the negative log value of the FDR score of the negative binomial testing. Blue dots indicate TEs with statistically significant enrichment (FDR ≤ 0.05). Red dots indicate TEs without significant enrichment. The FDR values of all TEs with an observed value of ≤ 10 or an observed/expected ratio of ≤ 1 are displayed as 1. (B) Exploratory bar-and-line graph generated using TEAR function output. The X-axis represents TEs. The left-hand Y-axis is associated with the bars and indicates the number of observed TEs overlapping peak regions. The right-hand Y-axis is associated with the line and indicates the ratio of observed accessible TEs to the total number of the same TEs in the genome.

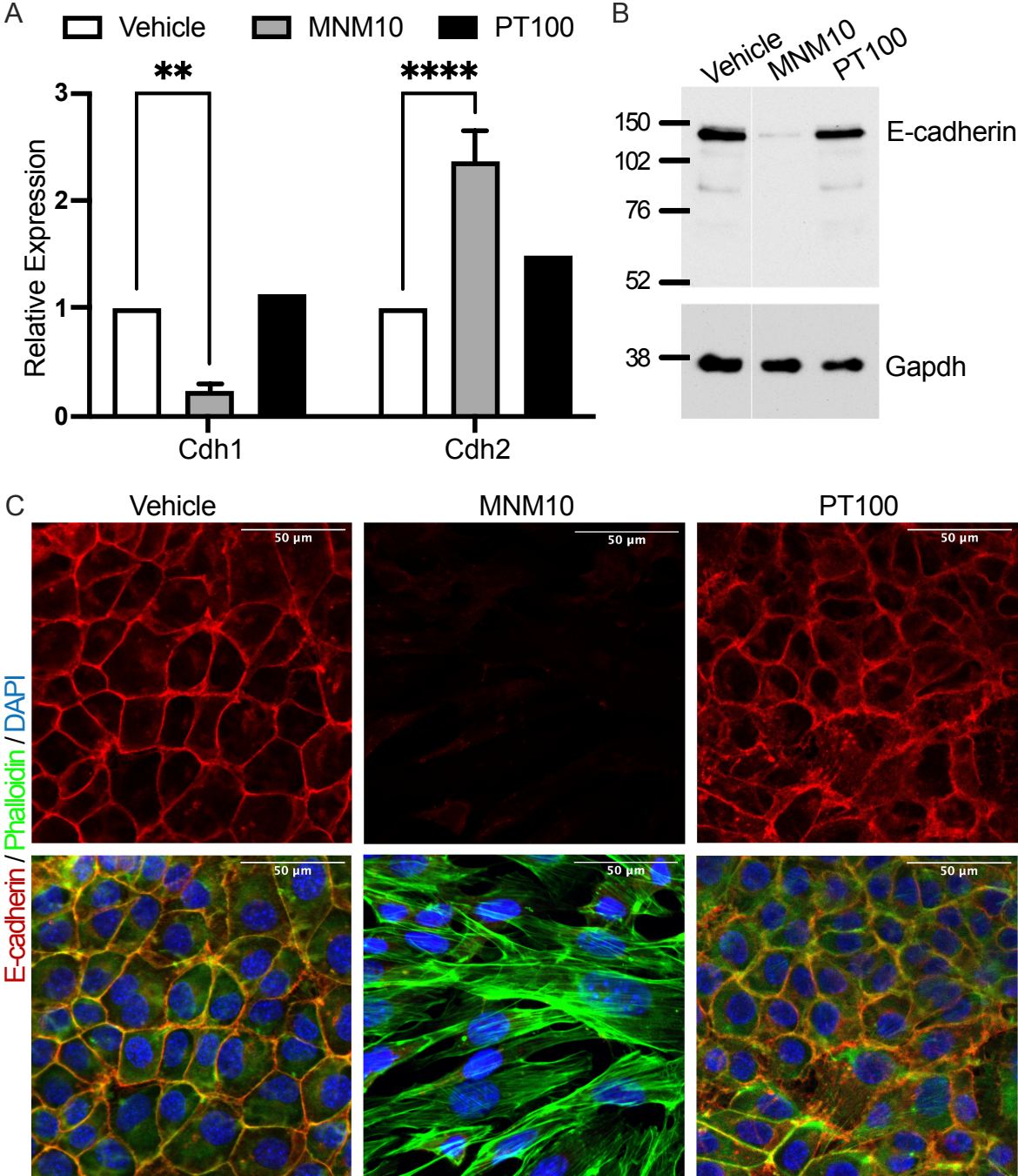
Suppl. Methods:

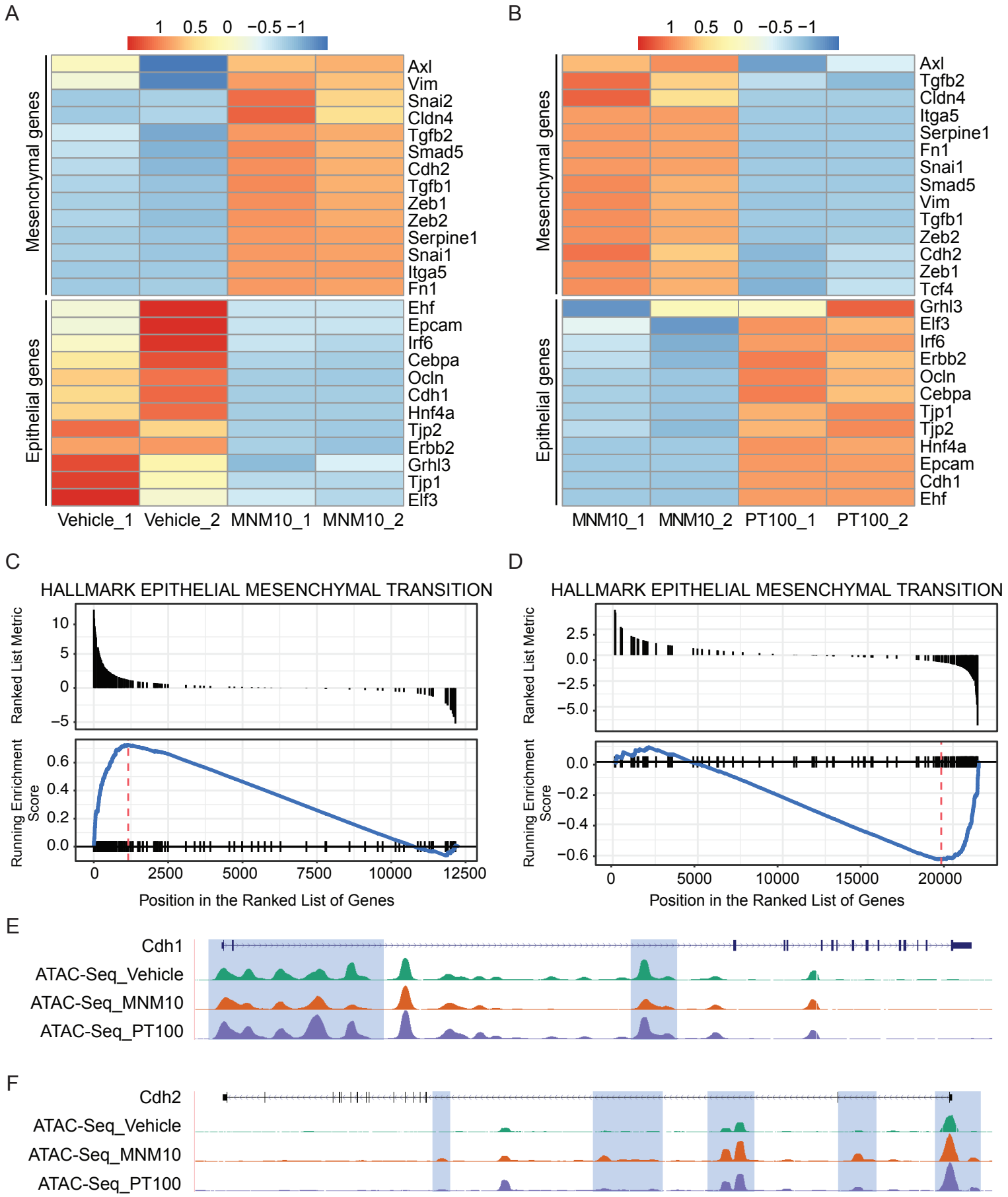
Exploratory Plot Generation for TEAR Output. For each repeatType, repeatFamily, and repeatName object generated by the TEAR function, exploratory scatterplots and bar-and-line graphs are generated via a wrapper for ggplot2 (version 3.3.2) (Supplementary Figure 2). For the scatterplot, we used the obsOnTrueMean value of each TE as the X-axis and the -log(p.adjust.value) as the Y-axis. For the bar-and-line graphs, the left-hand Y-axis observed repeats bars is plotted using the observed value of each TE, while the right-hand Y-axis ratio of observed repeats line is plotted using the ratio of the observed value to the rmsk value.


Input
Genomic peaks (ATAC-Seq, ChIP-Seq, meDIP-Seq, etc..)
Repeat Annotations
Input
DEG (RNA-Seq)
TEAR
TE enrichment genome wide
DATE
TE enrichment around DEGs
FindMotifs
Indentify transcription factor binding sites in enriched TEs


A
Chr1:
B1_Mus1
Scale
50 kb
mm10
186,650,000
186,700,000
Tgfb2
ATAC-Seq_Vehicle
ATAC-Seq_MNM10
ATAC-Seq_PT100
FoxA1 ChIP-seq
B
Chr18:
B4A
Scale
100 kb
mm10
16,600,000
16,650,000
16,700,000
16,750,000
16,800,000
16,850,000
Cdh2
ATAC-Seq_Vehicle
ATAC-Seq_MNM10
ATAC-Seq_PT100
TEAD ChIP-seq
C
Chr8:
ID_B1
Scale
5 kb
mm10
124,560,000
124,565,000
124,570,000
Agt
ATAC-Seq_Vehicle
ATAC-Seq_MNM10
ATAC-Seq_PT100
FoxA2 ChIP-seq
D
Chr1:
B4A
Scale
5 kb
mm10
90,960,000
90,965,000
90,970,000
Rab17
ATAC-Seq_Vehicle
ATAC-Seq_MNM10
ATAC-Seq_PT100
FoxA2 ChIP-seq


A
Epithelial State
Mesenchymal State
MET Enrichment
B3A
B4
MTD
B1_Mus1
MIRc
ID_B1
B2_Mm2
B4A
B3
L2a
Rank
EMT Enrichment
B1_Mur3
ID4_
B1F
MIRb
B3
B3A
RSINE1
ID_B1
B1_Mus1
B4A
B
Count
B4A 3 13
ID_B1 7 17
B3 3 17
B1_Mus1 2 12
B3A 4 10
EMT
MET
−log (adj p−value)
z−score
regulation of cell−substrate adhesion
heart morphogenesis
positive regulation of cell−substrate adhesion
wound healing
cell−matrix adhesion
muscle cell migration
cardiac septum development
actin filament organization
negative regulation of epithelial cell proliferation
wound healing
artery development
macrophage chemotaxis
regulation of angiogenesis
alcohol metabolic process
heart process
cellular response to Tgfb stimulus
actin filament organization
regulation of cellular component size
regulation of blood circulation
muscle cell migration
epithelial cell migration
regulation of response to wounding
renal system development
ameboidal−type cell migration
regulation of cell projection assembly
extracellular structure organization
kidney development
regulation of endocytosis
cell−matrix adhesion
regulation of actin filament−based process
positive regulation of axon extension
secondary alcohol metabolic process
positive regulation of smooth
glycerolipid metabolic process
muscle cell proliferation
cell−matrix adhesion
keratinocyte proliferation
sterol metabolic process
renal system development
positive regulation of developmental growth
cellular response to low−density lipoprotein particle stimulus
regulation of epithelial cell proliferation
actin cytoskeleton reorganization
urogenital system development
regulation of axon
negative regulation of
response to hydroxyurea
regeneration
cell projection organization
regulation of smooth muscle
retinoic acid metabolic process
cell proliferation
regulation of epithelial cell proliferation
renal system development


A
Vehicle
MNM10
PT100
**
****
Relative Expression
0
1
2
3
Cdh1
Cdh2
B
Vehicle
MNM10
PT100
150
102
76
52
38
E-cadherin
Gapdh
C
Vehicle
MNM10
PT100
50 µm
E-cadherin / Phalloidin / DAPI


A
1 0.5 0 −0.5 −1
Mesenchymal genes
Axl
Vim
Snai2
Cldn4
Tgfb2
Smad5
Cdh2
Tgfb1
Zeb1
Zeb2
Serpine1
Snai1
Itga5
Fn1
Epithelial genes
Ehf
Epcam
Irf6
Cebpa
Ocln
Cdh1
Hnf4a
Tjp2
Erbb2
Grhl3
Tjp1
Elf3
Vehicle_1 Vehicle_2 MNM10_1 MNM10_2
B
1 0.5 0 −0.5 −1
Mesenchymal genes
Axl
Tgfb2
Cldn4
Itga5
Serpine1
Fn1
Snai1
Smad5
Vim
Tgfb1
Zeb2
Cdh2
Zeb1
Tcf4
Epithelial genes
Grhl3
Elf3
Irf6
Erbb2
Ocln
Cebpa
Tjp1
Tjp2
Hnf4a
Epcam
Cdh1
Ehf
MNM10_1 MNM10_2 PT100_1 PT100_2
C
HALLMARK EPITHELIAL MESENCHYMAL TRANSITION
Ranked List Metric
10
5
0
−5
Running Enrichment Score
0.6
0.4
0.2
0.0
0 2500 5000 7500 10000 12500
Position in the Ranked List of Genes
D
HALLMARK EPITHELIAL MESENCHYMAL TRANSITION
Ranked List Metric
2.5
0.0
−2.5
−5.0
Running Enrichment Score
0.0
−0.2
−0.4
−0.6
0 5000 10000 15000 20000
Position in the Ranked List of Genes
E
Cdh1
ATAC-Seq_Vehicle
ATAC-Seq_MNM10
ATAC-Seq_PT100
F
Cdh2
ATAC-Seq_Vehicle
ATAC-Seq_MNM10
ATAC-Seq_PT100