

T.C.
DOKUZ EYLÜL UNIVERSITY
İZMİR INTERNATIONAL BIOMEDICINE AND GENOME
INSTITUTE

**REGULATOR: AN R LIBRARY FOR
ANALYSIS OF TRANSPOSABLE
ELEMENTS IN ACCESSIBLE REGIONS**

DOĞA ESKİER

DEPARTMENT OF MOLECULAR BIOLOGY
AND GENETICS

**DOCTOR OF PHILOSOPHY THESIS
İZMİR-2024**

THESIS CODE: DEU.IBG.PhD-2018850025

T.C.
DOKUZ EYLÜL UNIVERSITY
İZMİR INTERNATIONAL BIOMEDICINE AND GENOME
INSTITUTE

**REGULATOR: AN R LIBRARY FOR
ANALYSIS OF TRANSPOSABLE
ELEMENTS IN ACCESSIBLE REGIONS**

DEPARTMENT OF MOLECULAR BIOLOGY
AND GENETICS
DOCTOR OF PHILOSOPHY THESIS

DOĞA ESKİER

Advisor: Prof. Dr. Gökhan Karakölah
Co-advisor: Assoc. Prof. Hani Alotaibi

(This thesis was supported by YÖK 100/2000, TÜBİTAK 2211-A, TÜBİTAK 2250 programmes, and TÜBİTAK grants 121Z132, 122Z886, 120N556.)

THESIS CODE: DEU.IBG.PhD-2018850025

Dokuz Eylül University İzmir International Biomedicine and Genome Institute
Department of Genomics and Molecular Biotechnology, Molecular Biology and
Genetics graduate program Doctor of Philosophy student Doğa Eskier has
successfully completed their Philosophy thesis titled '**REGULATOR: AN R
LIBRARY FOR ANALYSIS OF TRANSPOSABLE ELEMENTS IN
ACCESSIBLE REGIONS**' on the date of **26 / 12 / 2024**.

Prof. Dr. ~~Gökhan~~ Karakölah
Dokuz Eylül University / İzmir International Biomedicine and Genome Institute /
Department of Genomics and Molecular Biotechnology
HEAD OF THE COMMITTEE

Asst. Prof. Dr. ~~Yavuz~~ Oktay
Dokuz Eylül University / Faculty of
Medicine / Department of Medical Biology
MEMBER

Asst. Prof. Dr. Cihangir Yandım
İzmir Economy University / Faculty of
Engineering / Department of Genetics and
Bioengineering
MEMBER

Prof. Dr. Şerif Şentürk
Dokuz Eylül University / İzmir
International Biomedicine and Genome
Institute / Department of Genomics and
Molecular Biotechnology
MEMBER

Prof. Dr. Güneş Özhan
İzmir Institute of Technology / Faculty of
Sciences / Department of Molecular
Biology and Genetics
MEMBER

Prof. Dr. Savaş İzzetoğlu
Ege University / Faculty of Sciences /
Department of Biology
SUBSTITUTE MEMBER

Asst. Prof. Dr. Seyit Kale
İzmir Katip Çelebi University / Faculty of
Medicine / Department of Biophysics
SUBSTITUTE MEMBER

TABLE OF CONTENTS

TABLE OF CONTENTS.....	i
ABBREVIATIONS	iv
LIST OF TABLES.....	v
LIST OF FIGURES.....	vi
ACKNOWLEDGEMENTS	vii
ABSTRACT	1
ÖZET	2
1. Introduction and Aims.....	3
1.1. Statement and Importance of the Problem	3
1.2. Aim of Study	3
1.3. Hypothesis of Study.....	4
2. General Information	5
2.1. Overview of transposable elements.....	5
<i>2.1.1. Classification of TEs</i>	<i>6</i>
<i>2.1.1.1. Mechanisms of transposition</i>	<i>6</i>
<i>2.1.1.2. Structure of TE sequences.....</i>	<i>6</i>
<i>2.1.2. The roles of TEs in genome structure and regulation.....</i>	<i>8</i>
<i>2.1.2.1. Roles of TE in development.....</i>	<i>8</i>
<i>2.1.2.2. Roles of TEs in diseases</i>	<i>9</i>
<i>2.1.4. Regulation of TE activity.....</i>	<i>12</i>
<i>2.1.4.1. Chromatin structure and TE activity</i>	<i>12</i>
<i>2.1.4.2. Small RNAs in TE regulation</i>	<i>13</i>
<i>2.1.4.3. Regulation of TEs during development.....</i>	<i>13</i>

2.1.4.4. <i>Effects of TE dysregulation</i>	14
2.2. Assay for transposase-accessible chromatin using sequencing (ATAC-seq)	15
2.3. Computational analysis of transposable elements	16
2.3.1. <i>Difficulties in experimental approaches to TEs</i>	16
2.3.2. <i>Next generation sequencing in TE analysis</i>	17
2.3.2.1. <i>Assay for transposase accessible chromatin (ATAC) sequencing</i>	18
2.3.2.2. <i>Expression profiling of TEs via RNA-seq</i>	18
2.3.3. <i>Current computational tools for TE analysis</i>	19
2.3.3.1. <i>Tools for identification of regulatory functions of TEs</i>	20
2.3.3.2. <i>Missing perspectives in current tools</i>	20
<u>3. Materials and Methods</u>	<u>22</u>
3.1. Type of study	22
3.2. Time and Place of Study	22
3.3. Materials of Study	22
3.4. Variables of Study	22
3.5. Tools for Data Collection	22
3.6. Study Plan and Calendar	23
3.7. Data Evaluation	23
3.7.1. <i>ATAC-seq run data processing, alignment, and peak identification</i>	23
3.7.2. <i>RNA-seq run data processing, alignment, and analyses</i>	24
3.7.2. <i>Outline of regulaTER requirements</i>	25
3.7.3. <i>Implementation of regulaTER package</i>	25
3.7.4. <i>Function to determine transposable elements enriched in accessible genomic regions</i>	26

3.7.5. Function to determine accessible transposable elements enriched in neighboring regions of differentially expressed genes	26
3.7.6. Function to identify the genomic motifs contributing to the regulatory functions of transposable elements in accessible regions.....	27
3.7.7 Functions for creating graphical representations of the outputs.....	27
3.8. Limitations of Study	28
3.9. Ethics Committee Approval.....	28
<u>4. Results.....</u>	<u>29</u>
4.1. Quality control and interpretation of ATAC-seq reads.....	29
4.2. Characterization of the transcriptome during EMT and MET events and the identification of differentially expressed genes.....	39
4.3. Identification of TEs enriched in accessible regions using the TEAR function of regulaTER.....	41
4.4. Identification of accessible TEs enriched in proximity of differentially expressed genes using the DATE function of regulaTER.....	45
4.5. Motif analysis of accessible TEs enriched near DEGs using the FindMotifs function of regulaTER	48
4.6. Application of regulaTER to independent datasets	50
4.7. Description and potential uses of regulaTER	51
<u>5. Discussion</u>	<u>58</u>
<u>6. Conclusion</u>	<u>64</u>
<u>7. References.....</u>	<u>65</u>
<u>8. Addenda</u>	<u>78</u>
8.1. Ethics committee approval	78
8.2. ARBIS CV	79
8.3. Research article associated with thesis study.....	83

ABBREVIATIONS

ATAC-seq.....	Assay for transposase accessible chromatin using sequencing
DEG	Differentially expressed gene
DNMT	DNA methyltransferase
EMT	Epithelial-mesenchymal transition
ERV	Endogenous retroviral elements
GSEA.....	Gene set enrichment analysis
LINE.....	Long interspersed nuclear element
LTR	Long terminal repeat
MET	Mesenchymal-epithelial transition
ORF.....	Open reading frame
SINE.....	Short interspersed nuclear element
TE	Transposable element
TF	Transcription factor
TFBS	Transcription factor binding site
TSD.....	Target site duplication sequence
TSS	Transcription start site

LIST OF TABLES

Table 1. Study timeline.....	23
Table 2. Evaluation of pre- and post-Trim Galore! ATAC-seq library read counts and percentages.	30
Table 3. Top 10 TE subfamilies enriched in accessible peaks in EMT.	43
Table 4. Top 10 TE subfamilies enriched in accessible peaks in MET.	43
Table 5. Top 10 TE subfamilies enriched near DEGs in EMT.....	46
Table 6. Top 10 TE subfamilies enriched near DEGs in EMT.....	46
Table 7. Comparison of regulaTER with similar tools.	60

LIST OF FIGURES

Figure 1. Structure of different TE classes.	7
Figure 2. Possible mechanisms of TEs in disease.	11
Figure 3. Mapping percentages of ATAC-seq libraries.	29
Figure 4. Fingerprint and read distribution profile plots of ATAC-seq libraries.	34
Figure 5. Plots representing the ataqv quality metrics of the ATAC-seq libraries.	35
Figure 6. Fraction of reads in peak (FRiP) scores of ATAC-seq libraries.	36
Figure 7. Barplot of peak count in annotated genomic regions.	37
Figure 8. Stacked barplot of number of peaks in each sample categorized according to number of samples peak is covered in.	38
Figure 9. Heatmap of key epithelial and mesenchymal marker genes during (A) EMT and (B) MET processes.	40
Figure 10. GSEA plots of EMT hallmark genes during (A) EMT and (B) MET processes.	41
Figure 11. Graphical representations of the subfamily level enrichments of TEs in accessible regions.	44
Figure 12. GO term enrichment dotplot of DEGs near accessible TEs of interest in EMT and MET.	47
Figure 13. UCSC Genome Browser visualization of TF binding to DEG-associated TE loci.	49
Figure 14. Partial screen capture image of the README.md file of regulaTER on GitHub.	51
Figure 15. The TEAR function documentation viewed in Rstudio.	53
Figure 16. The DATE function documentation viewed in Rstudio.	54
Figure 17. The FindMotifs function documentation viewed in Rstudio.	56
Figure 18. The MakePlots function documentation viewed in Rstudio.	57

ACKNOWLEDGEMENTS

Above all, I would like to thank my wonderful wife and child for giving me the strength, patience, and drive to see everything through.

For their various scientific contributions during the writing and completion of this thesis, I hereby extend my sincerest gratitude to the following individuals and organizations:

To my supervisor, Prof. Dr. Gökhan Karakülah for his guidance and mentorship in the field of computational biology and transposable elements, and my co-advisor Dr. Hani Alotaibi for his keen understanding of the epithelial-mesenchymal and mesenchymal-epithelial transitions;

To all members of my thesis monitoring committee and thesis defense jury for their contributions during the discussion sessions to improve quality of this thesis;

To all members of the Karakülah lab on computational sequence analysis, iBG Bioinformatics Platform, and the Alotaibi lab on transcriptional regulation, past and present, for their invaluable work, with special thanks to the following,

Dr. Burcu Şengez for her work in maintaining the cell cultures used to generate the data in this thesis, and her help in understanding the regulation of EMT and MET,

Seray Yetkin, for her work in generating the sequencing libraries used in this thesis,

Nazmiye Odabaşı, for her invaluable help in setting the foundations of regulaTER,

Ahmet Bursalı, for his keen understanding of R and data visualization,

Alırza Arıbaş and Hüseyin Güner, for their vigilant stewardship of all data used in this thesis, and their swift responses to all requests for data analysis tasks;

To Prof. Dr. Esin Firuzan and Dr. Eda Kaleli for all their suggestions and assistance on this thesis and the associated research article;

To all members of the Dokuz Eylül University İzmir International Biomedicine and Genome Institute and the İzmir Biomedicine and Genome Center, past and present, for being valuable colleagues.

I would like to acknowledge all attendees of the EMBO Workshop: the Mobile Genome 2023 for their contributions to TE research, and guiding my understanding of the field,

with special thanks to Dr. Michael Imbeault, Dr. Wayo Matsushima, Prof. Dr. Johan Jakobsson, and Prof. Dr. Eric Miska, for the one-on-one talks which greatly inspired how I interpreted the results of my work.

I would like to thank YÖK 100/2000 (Bioinformatics), TÜBİTAK 2211-A (BİDEB) and 2250 (BİDEB) programs, and TÜBİTAK 3501 (Project Nos: 121Z132, 122Z886, 120N556) grants for supporting me and funding the necessary experiments throughout my PhD studies.

This thesis would not have been possible without the assistance and love of my family members, especially,

My parents, for their protection, moral and financial support, and love over the years, and for letting me face all challenges, no matter how daunting, with confidence,

My grandmother, whose love and compassion and wit helped shape the person I am today, and whose absence I feel keenly every single day,

My wonderful and brilliant wife, Eda Eskier, as her kindness and infinite patience and grace show me how much better the world can be, and her love and understanding grant me the strength and wisdom to bear any burden, and I can only hope that she sees and feels a fraction of my love for her,

And finally, my beautiful shining child, who gives meaning to every moment of every day, whom all my efforts are dedicated to, and who deserves the best of all the worlds I can create. I love you all as deeply and truly as anyone could.

ABSTRACT

REGULATER: AN R LIBRARY FOR ANALYSIS OF TRANSPOSABLE ELEMENTS IN ACCESSIBLE REGIONS

**Doğa Eskier, İzmir International Biomedicine and Genome Institute, Dokuz
Eylul University Health Campus, Balçova 35340 - Izmir / TURKEY**

Transposable elements are repeat elements which can create copies of themselves in the genome, making them a prolific source of genetic variance and novel genomic features. Many transposable elements are enriched for genomic motifs, such as transcription factor binding sites, implicating them in the formation of gene regulatory networks. However, due to the deleterious effects of unregulated transposition, most are repressed, located in heterochromatin regions, and inhibited from such activity. To fully understand the impact of transposable elements on gene regulation, it is necessary to incorporate multiomics data to consider the epigenomic state of transposable elements in a locus-specific manner. As repeat sequences, transposable elements further require the use of specialized computational methods.

Here, I introduce regulaTER, an R library which uses multiomics data to automate many transposable element centric analyses to uncover their roles in gene and genome regulation. regulaTER comprises four functions, and can take a wide range of genomic and transcriptomic data from different sources to identify transposable element enrichment in various genomic contexts. It can further identify their enrichment in potential gene regulatory regions, and find the genomic motifs enriched in transposable element sequences which might drive such regulation. I further validate its use with epithelial-mesenchymal and mesenchymal-epithelial transition models. Finally, I describe the potential impact of B elements / Alu elements on these processes, and their potential to contribute Forkhead transcription factor binding sites for their regulation, and highlighting the capabilities of regulaTER.

Keywords: transposable elements, multiomics, transcriptional regulation, epithelial-mesenchymal transition, mesenchymal-epithelial transition

ÖZET

REGULATER: ULAŞILABİLİR GENOMİK BÖLGELERDEKİ

TRANSPOZONLARIN ANALİZİ İÇİN BİR R PAKETİ

**Doğa Eskier, İzmir Uluslararası Biyotıp ve Genom Enstitüsü, Dokuz Eylül
Üniversitesi Sağlık Kampüsü, Balçova 35340 – İzmir / TÜRKİYE**

Transpozonlar genomda yeni kopyalar oluşturabilen tekrar dizilerdir, ve genetik çeşitlilik ve yeni genomik özelliklere yol açabilirler. Birçok transposon transkripsiyon faktör bağlanma bölgeleri gibi genomik motifler açısından zenginleşmiştir, ve bu şekilde gen düzenleyici ağ oluşumunda yer alırlar. Bununla beraber, düzenlenme dışı transpozisyon hücreye ve organizmaya zarar verebilir. Bu nedenle, transpozonların olası aktiviteleri heterokromatin bölgelerde tutularak ya da başka şekillerde susturularak engellenmektedir. Transpozonların gen düzenleyici katkılarının tam olarak anlaşılması için, transpozonların epigenomik durumunun multiomik veriler aracılığıyla lokus özgü bir şekilde tanımlanması gerekmektedir. Transpozonların tekrar diziler olmaları, aynı zamanda analizleri için özelleşmiş hesaplamalı yöntemler gerektirir.

Bu tezde regulaTER adlı R kütüphanesi tanımlanmaktadır. regulaTER multiomik veriler kullanarak çeşitli transpozon analizlerini otomatize eder, ve gen ve genom düzenleyici işlevlerinin belirlenmesine yardımcı olur. regulaTER kapsamındaki dört fonksiyon, çeşitli genomik bağlamlarda ve gen düzenleyici bölgelerde zenginleşmiş transpozonların tanımlanmasını, ve bu düzenleyici işlevlere katkı sağlayan genomik motiflerin zenginleşmelerinin hesaplanmasını sağlar. Kütüphanenin kullanımının doğrulanması, epitelyal-mezenkimal ve mezenkimal-epitelyal hücre dönüşümü modelleri kullanılarak yapılmıştır. Bu süreçlerde B / Alu transpozonlarının olası katkıları belirlenmiş ve açıklanmıştır. Bu transpozonların üzerindeki Forkhead transkripsiyon fakötrü bağlanma bölgelerinin süreçlerin düzenlenmesine katkıları regulaTER paketinin yetenekleri gösterilmiştir.

Anahtar sözcükler: transpozonlar, multiomiks, transkripsiyonel düzenleme, epitelyal-mezenkimal dönüşüm, mezenkimal-epitelyal dönüşüm

1. Introduction and Aims

1.1. Statement and Importance of the Problem

Transposable elements (TEs) are repeat elements which can create novel copies of themselves in the genome, and thus contribute to genetic and genomic complexity. TEs were previously implicated in various biological processes, such as the formation of gene regulatory networks, as they harbor many genomic motifs which can bind transcription factors to the proximity of genes. As such, it is vital to understand their impact on the genome and the transcriptome. However, dysregulation of TEs and unregulated transposition can cause deleterious mutations, which can lead to diseases such as cancer. To prevent these unwanted effects, TEs are often under repressive epigenomic markers, such as H3K27me3, a repressive histone modification, or DNA methylation. While there have been previous computational tools and databases which analyze the connection of TEs to genes or other genomic regions to elucidate their contributions, these have focused on either the genomic coordinate of the TEs alone, such as GREAM, or their co-expression with genes, such as TEffectR. TEs can contribute to gene regulation independent of their own transcription; however, such regulatory activity is subject to the chromatin state of the TE locus being amenable to transcription factor binding, i.e. in accessible chromatin regions. Extant tools and methods do not allow the incorporation of multiomics data taking into account such chromatin or epigenomic state information. Developing such a method is necessary to further our understanding of TEs.

1.2. Aim of Study

The study aims to understand the gene regulatory impact of TEs using multiomics data that can reveal the genomic and epigenomic context of TE loci. To do so, regulaTER, an R library to calculate the enrichment of TEs and TE-harbored genomic motifs using multiomics data, was created. The study further aims to validate the use of regulaTER by applying its functions to data derived from a biological model of interest, namely

epithelial-mesenchymal and mesenchymal-epithelial transitions, and comparing the findings to previous research in the regulation of these processes.

1.3. Hypothesis of Study

The hypothesis of this study is that TEs contribute to the regulation of nearby genes through genomic motifs, such as transcription factor binding sites, and such regulatory contributions can be identified via enrichment analyses using multiomics data. These analyses can be performed using regulaTER, an R library optimized to perform TE-centric multiomics analyses.



2. General Information

2.1. Overview of transposable elements

Transposable elements (TEs) are repeat genomic elements with the unique ability to relocate themselves through a biological process called transposition. They were originally referred to as “controlling elements” by Barbara McClintock in 1956, where McClintock first described their ability to transpose and regulate the expression of nearby genes without altering the open reading frame of the gene (McClintock, 1956). TEs exist across a wide range of evolutionary complexity, ranging from bacteria to mammals, and comprise at least 45% of the human genome (Lander et al., 2001). Activity level of TEs in extant genomes vary. The TEs in bacterial genomes are often simpler and divided into two categories, insertion sequences (IS), which are shorter and more autonomous (Mahillon and Chandler, 1998), and transposons (Tn), which are more closely related to eukaryotic TEs. Non-animal eukaryotic genomes display high TE activity even after maturation, such as in plants; however, mammalian species, including humans, also have TEs that display active transposition in both somatic and germline cells.

Since their earliest discovery, the understanding of the roles of TEs in a wide range of biological contexts have grown. One of the key ideas in TE research is that, despite their ability to relocate themselves multiple times, their distribution across the genome is not random. Due to the potential deleterious effects of TE insertion or transcription, such as via the disruption of the regulation of other genomic elements, TEs must maintain a strict balance between propagation and silencing (Robillard et al., 2016; Sultana et al., 2017). The most successful TEs often themselves contribute to or make use of the genomic or transcriptomic regulatory mechanisms through the process of exaptation, similar to organisms adapting to a specific ecosystem (Davidson and Britten, 1979; Venner et al., 2009). Due to these features, TE activity and dysregulation has been linked to the maintenance of an organism and its genome, as well as several diseases, which makes them of great interest in genomic research.

2.1.1. Classification of TEs

2.1.1.1. Mechanisms of transposition

TEs are largely divided into two broad classes, based on the method of transposition they use (Wicker et al., 2007). Class I TEs, also known as retrotransposons, use RNA intermediaries which are reverse transcribed into DNA through the reverse transcriptase enzyme, which is then inserted into the genome in another location, creating new copies while leaving the original instance intact. Because it creates novel copies, this process is commonly called “copy and paste” transposition. Class II TEs are DNA transposons, where the original instance is excised from the genome and relocated into another location on the genome. Because this process keeps the number of TE instances the same, it is also called “cut and paste” transposition. In addition to these broad classes, other mechanisms for transposition exist. Helitron DNA transposons, for example, use a mechanism referred to as “peel and paste” (Grabundzija et al., 2016).

2.1.1.2. Structure of TE sequences

Because of the biological activity involved in the transposition of different classes of TEs, TE structures also vary between classes. For example, DNA transposons are built around a core sequence that codes for the transposase enzyme, which is responsible for the excision and reintegration of the TE from and into the genome. This transposase sequence is flanked by two inverted terminal repeats (ITR), which are themselves enclosed by direct repeat sequences (DR). In comparison, RNA transposons display a more varied and directional structure. Among the Class I TEs, the two major subtypes are referred to as autonomous TEs, which contain many of the elements required for their transposition, such as coding for the reverse transcriptase (RT) enzyme, and non-autonomous TEs, which rely on autonomous TEs, such as LINE-1, for successful transposition. The autonomous TEs are also categorized into LTR (long terminal repeat) TEs, which contain terminal repeat sequences similar

to DNA transposons, and non-LTR TEs, such as LINE-1, which lack such sequences (Figure 1) (Kazazian, 2011).

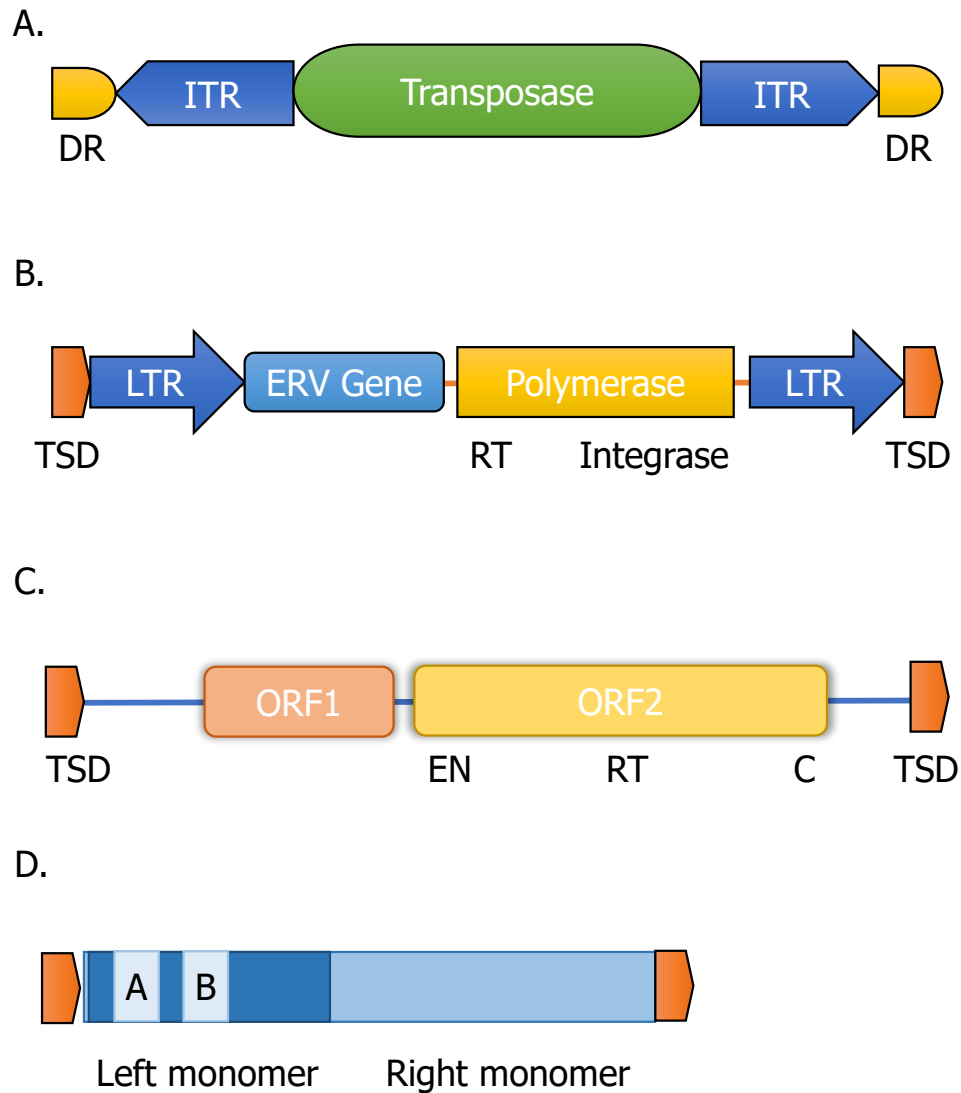


Figure 1. Structure of different TE classes. (A) DNA transposon (~1.2 kb). Yellow sections indicate direct repeats (DR), while blue regions indicate inverted terminal repeats, in opposing directions. Green section encodes for the transposase enzyme. (B) LTR retrotransposon of the endogenous retrovirus (ERV) family (7-9 kb). Orange sections indicate target site duplication sequences. Dark blue arrows indicate long terminal repeats (LTR), in the same direction. Light blue section indicates genes

required for ERV life cycle. Yellow box indicates RNA-dependent DNA polymerase with the reverse transcriptase and integrase domains. (C) LINE-1 retrotransposon (6 kb). The second ORF encodes for endonuclease, reverse transcriptase, and zinc knuckle domains. (D) Alu transposon (300 bp). Both monomers end in poly(A) tails. Left monomer contains A- and B- PolIII promoter regions.

2.1.2. The roles of TEs in genome structure and regulation

TEs contribute to genomic diversity over an evolutionary time period. Examples of such genomic diversity include the formation of novel chromosomal features such as telomerase-independent telomere maintenance in some *Drosophila* species (Pardue and DeBaryshe, 2011), formation of novel genes, both protein coding and noncoding (Naville et al., 2016), and contribution of *cis*-acting regulatory elements (Chuong et al., 2017). In addition to these evolutionary contributions, TE contributions to the genome can be individual- or even tissue-specific. These effects happen via transposition in germline (Hancks and Kazazian, 2016) or somatic cells (Garcia-Perez et al., 2007; Baillie et al., 2011). Such transposition can lead to genetic diseases or cause tumorigenesis (Scott et al., 2016). Although a majority of TEs in mammals do not show active expression or transposition, many TEs yet show intergenerational transposition in both humans, and in model organisms such as *Mus musculus* and *Rattus norvegicus* (Huang et al., 2012).

2.1.2.1. Roles of TE in development

In addition to their structural roles in the genome, TEs have various key regulatory roles in the genome. Several TE subclasses, including, but not limited to, endogenous retroviral elements (ERVs), long interspersed nuclear elements (LINE), and primate SINE – VNTR – Alus (SVAs) have promoter elements or motifs for transcription factor (TF) binding sites (TFBS) as part of their sequences. TEs which are transposed near gene bodies can lead to gene expression in alternate forms and contexts by contributing these TFBSs to the gene regulatory regions. For example, the *Dicer1* gene in mice is active both during oogenesis and in various cell types, but

an intragenic MT-C ERV transposition that occurs in oocytes leads to the expression of a truncated form of Dicer1. Because this truncated form of the protein lacks the self-inhibitory domain, oocytes show higher Dicer1 enzymatic activity, which is a crucial component of oogenesis. (Flemr et al., 2013). Furthermore, multiple TEs of the same subfamily can contribute to the formation of gene regulatory networks active during cell differentiation and cell fate determination. The ERV subfamily RLTR13, active in trophoblast stem cells, carries TFBSs for Eomes, CDX2, and ELF5, three TFs specific to trophoblast differentiation. RLTR13 contributes to the regulation of more than 100 genes by these TFs, creating a transcriptional network fundamental to the trophoblast differentiation process (Chuong et al., 2013). Similarly, 5-25% of TFBSs for the pluripotency-associated TFs OCT4 and NANOG in human and mouse embryonic stem cells are derived from ERVs (Bourque et al., 2008). Furthermore, the demethylation of LINE-1, one of the few subfamilies of TEs that show active transposition in humans, during the development process leads to transcriptional activation in neural progenitor cells, which is one of the mechanisms that lead to neuron formation (Jönsson et al., 2019)

2.1.2.2. Roles of TEs in diseases

TE activity has been linked to disease for a considerable time, with one of the seminal studies in the field dating back to 1988 (Kazazian et al., 1988; Kazazian, 1998). The functions of TEs in disease occur via multiple mechanisms, which can be grouped under two main groups. The first group is transposition into the gene region, which alters protein structure and gene function. For example, hemophilia A is caused by loss-of-function mutations, usually familial, at the F8 locus on the X chromosome. However, when the genome of a patient with no family history of hemophilia A was examined, a novel LINE-1 insertion was observed in exon 14 of the F8 locus (Kazazian et al., 1988). Transposition can affect not only the exon sequence, but also which exons are spliced in. A splice enhancer sequence located near exon 11 of the CLCN5 is disrupted by the insertion of an Alu element, which causes the exon to be skipped. The resulting protein is enzymatically deficient, which leads to Dent disease (Claverie-

Martín et al., 2005). Similar exon skipping can also be seen as a result of transposition into intronic regions. Certain transpositions can also cause entire gene regions to be deleted. One of the mutations that cause type 1 neurofibromatosis disease is the deletion of the SVA region located somatically in the SUZ12P region, the NF1 gene located in the 3' neighborhood of this region (Vogt et al., 2014). The second class of mechanisms by which transposons cause disease is the change in gene expression as a result of transposition. When C/EPB β , MYC, and E2F1 TF binding sites, which show high activity in MCF7 breast cancer cells, were examined genome-wide, it was observed that these regions were mostly seen together with transposon sites and transposons carrying binding sites were found in important cancer marker mRNA promoter regions (Jiang and Upton, 2019).

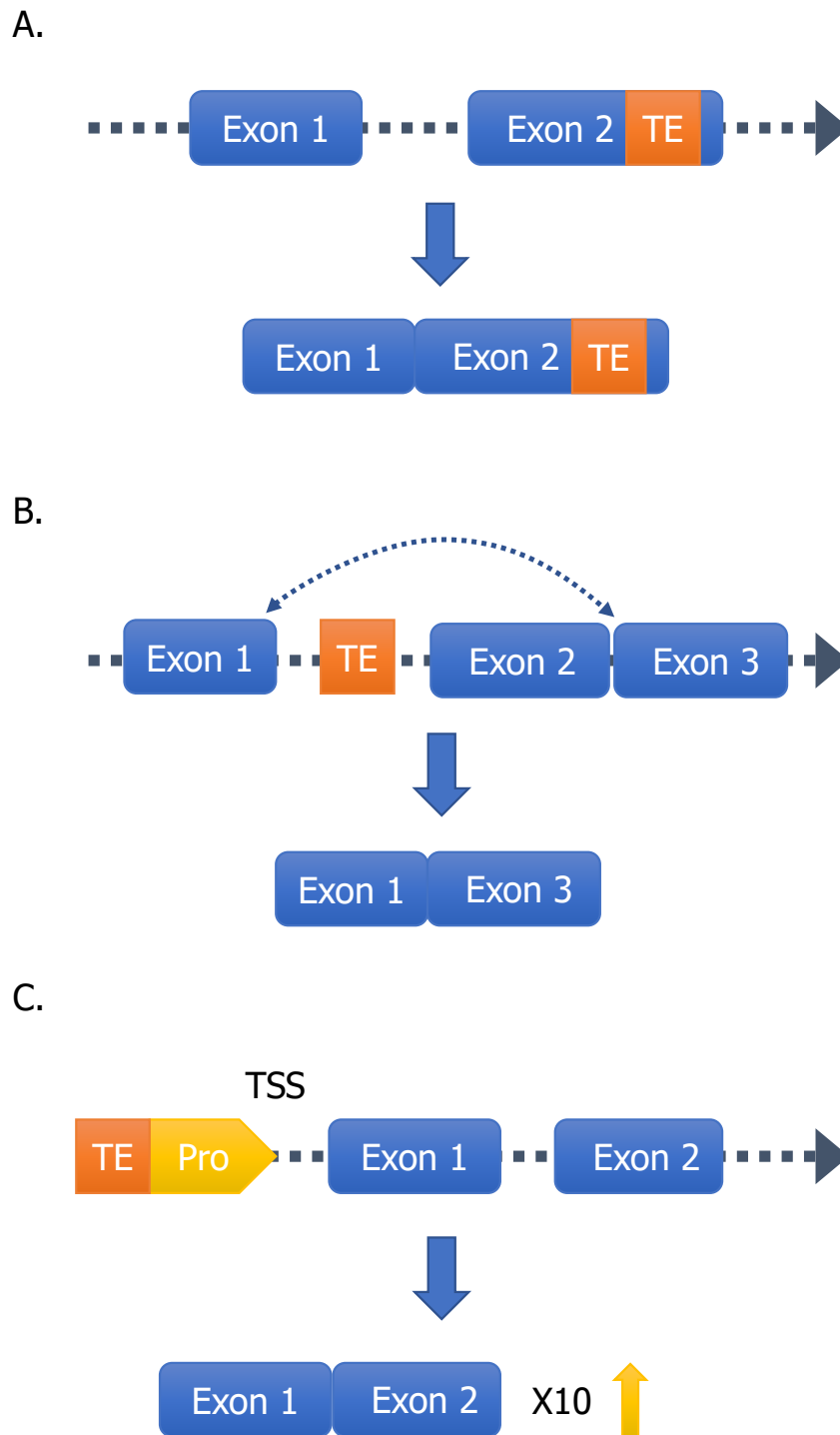


Figure 2. Possible mechanisms of TEs in disease. (A) Insertion into exonic region causes altered gene product by changing exon sequence. (B) Insertion into intronic

region causes altered gene product by exon skipping. (C) Insertion into promoter region causes altered gene expression.

2.1.4. Regulation of TE activity

While TEs form a large part of the genome in many organisms, and retrotransposons have the capacity to create novel copies of themselves, new transposition events do not occur randomly. As previously mentioned, TEs often have to strike a careful balance between expression and repression, which makes them one of the most carefully regulated elements in the genome. Evolutionarily, TE regulation has favored repression (Goodier, 2016), with multiple mechanisms that have developed to restrict transposition (Molaro and Malik, 2016; Yang et al., 2017). Despite the repression-favored regulation of TEs, transposition events do happen, either by escaping regulation, or in a tightly regulated fashion. TE activity is more commonly observed in germline cells, where transposition does not pose a major disadvantage to the cell, and can pass onto future generations (Haig, 2016).

2.1.4.1. Chromatin structure and TE activity

TEs have a strong association with local chromatin structure. As mentioned earlier, TEs often exist in heterochromatin regions, in order to suppress unwanted transposition, which can arise as a result of multiple mechanisms, including CpG methylation or histone modifications. In addition to the H3K27me3 and H3K9me3 markers, two of the most common repressive chromatin markers, other histone markers are also associated with silenced TE regions. H4K20me3, a histone marker often associated with heterochromatin regions in tandem with H3K9me3, is by itself sufficient to maintain LINE-1 methylation in murine cells, via association with a bromo-adjacent homology domain on DNMT1 (Ren et al., 2021). Similarly, evolutionarily young TEs which otherwise show higher activity than other TEs in *Xenopus tropicalis* embryos are marked by H4K20me3-associated silencing. However, the timing and mechanism of H4K20me3 silencing in *Xenopus tropicalis* differs between DNA transposons, which show early stage marking, and retrotransposons, which show increasing levels of silencing over development (van Kruijsbergen et al., 2017). Despite

the frequent presence of heterochromatin in silenced TEs, the relationship is not as straightforward as one might expect. A small group of TEs, including *Sleeping beauty*, *Frog prince*, and *Minos*, with a shared terminal inverted repeat structure, for example, display increased transposition when the local chromatin is condensed by CpG methylation, presumably by bringing several genomic elements required for transposition in close proximity (Jursch et al., 2013).

2.1.4.2. Small RNAs in TE regulation

Small RNAs, including microRNAs (miRNAs), small interfering RNAs (siRNAs), and PIWI-interacting RNAs (piRNAs), have emerged as key players in the regulation of TEs. These small RNAs function as powerful post-transcriptional and transcriptional regulators, suppressing TE activity and maintaining genome integrity. In plants, miRNAs target and downregulate the expression of active TEs, thereby curbing their mobility (Fultz et al., 2015). In animals, the piRNA pathway plays a critical role in TE defense. PIWI proteins and their associated piRNAs act as an adaptive defense system, effectively silencing TEs in the germline and somatic cells to prevent TE mobilization (Aravin et al., 2007). Additionally, siRNAs, which are generated from double-stranded RNA precursors, participate in TE regulation in both plants and animals. In fruit flies, specific small RNA-generating loci can act as master regulators of transposon activity, showcasing the central role of siRNAs in suppressing TEs (Brennecke et al., 2007). These findings show the critical role of small RNAs in the dynamic interplay between host genomes and TEs, enabling the maintenance of genome stability and integrity.

2.1.4.3. Regulation of TEs during development

DNA methylation and associated repression of genomic elements is one of the key factors of TE regulation, with DNMT3C being a crucial methyltransferase for TE repression (Yoder et al., 1997; Barau et al., 2016). However, during the early stages of development, the entire genome undergoes a two-step, passive and active demethylation process (Guibert et al., 2012; Vincent et al., 2013). Therefore, to avoid excess genomic disruption by the TEs, other mechanisms must be deployed to

maintain their regulation. In their 2016 study, Walter *et al.* have studied the chromatin kinetics of cells undergoing dynamic demethylation, and found that TEs can be divided into three categories based on their modes of repression. During early development, TEs can be silenced by H3K27me₃, H3K9me₃, or a combination of the two. This silencing pattern holds true in cells that have adapted to demethylated DNA, which suggests that the mechanisms have are evolutionary responses in pluripotent cells to avoid excessive TE-derived heterogeneity (Walter et al., 2016).

2.1.4.4. Effects of TE dysregulation

As previously mentioned, TE activity is tightly regulated in the genome during the normal functioning of the organism. Failure of such regulation can lead to abnormal TE activity and transposition, which can then cause deleterious effects, such as various diseases. There is considerable information about the relationship between LINE-1 retrotransposons, which show active transposition in humans, and disease and immune response (Rodic, 2018; Zhang et al., 2020). The association between LINE-1 and cancer involves several different mechanisms. One of these mechanisms is increase genomic instability and local mutations following LINE-1 hypomethylation. LINE-1 hypomethylation can be used as a prognostic marker in various cancers such as gastric, lung, liver, and esophageal cancers (Shigaki et al., 2013; Iwagami et al., 2013; Kawano et al., 2014; Lavasanifar et al., 2019). One of the explanations for the association of LINE-1 hypomethylation with cancer is the activation of local oncogenes and abnormal expression of microRNAs, both of which accelerate tumor invasion after hypomethylation. In addition, by localizing to gene regions, LINE-1 can cause an increase of oncogenes in tumor tissue or deletion of tumor suppressor genes. It is known that LINE-1 increases the oncogenic effect of the Myc gene by transposing into its region, as a result of genome-wide hypomethylation in breast cancer cells (Morse et al., 1988). Another function of LINE-1 in the formation and diagnosis of cancer is through the expression of LINE-1 RNA. High expression of the ORF1 region of LINE-1 has been associated with c-MET activation (Ko et al., 2019) and loss of TP53 (Rodić et al., 2014), especially in ovarian cancer. ORF1 expression in cancer shows histological

restrictions. Various cancers, such as kidney, liver, or cervical cancer, exhibit very low ORF1 expression, but esophageal, lung, intestinal, and bladder cancers have high ORF1 expression (Ardeljan et al., 2017).

2.2. Assay for transposase-accessible chromatin using sequencing (ATAC-seq)

Originally described in 2013, ATAC-seq (Buenrostro et al., 2013) is a method for epigenomic analysis, where, instead of individual chromatin markers, such as DNA methylation, or histone modifications, the accessibility of the chromatin is directly assessed using the hyperactive Tn5 transposase (Goryshin and Reznikoff, 1998). In recent years, it has become a pioneering genomic profiling method, leading to advances in understanding gene regulation and chromatin accessibility.

It has several advantages over other methods for identifying accessible genomic regions, such as multidimensionality, and requiring a small cell population (Buenrostro et al., 2015). For example, ATAC-seq can be applied to a population of as few as 500 cells, while methods like FAIRE-seq and DNase-seq require 1 million to 50 million cells. Furthermore, ATAC-seq is much faster, and simpler compared to such methods, with the library preparation experiments lasting only 2-3 hours, compared to 3-4 days for DNase-seq, and having lower room for errors and ease of repeatability if such errors should occur (Sun et al., 2019). Because ATAC-seq directly assesses chromatin accessibility, without relying on specific pre-selected markers as with ChIP-seq, which requires immunoprecipitation of selected DNA-binding proteins, it can also more accurately identify a wider range of genomic regions in a single experiment (Meyer and Liu, 2014). Together, these features have made ATAC-seq the method of choice for investigating chromatin accessibility.

Tn5 is a prokaryotic DNA TE, with a normally low-activity native transposase. Using targeted mutations to increase its activity, the hyperactive Tn5 transposase can single-handedly target all accessible regions in the genome and shear those regions into short fragments, which are then tagged with adapters for sequencing. While the Tn5 TE normally has a specific insertion site sequence, benchmarking studies have

shown that use of Tn5 for fragmentation of accessible genomic regions has very low levels of bias (Adey et al., 2010).

ATAC-seq is a versatile and powerful tool, unlocking new avenues of investigation across various biological contexts. Its ability to identify enhancer regions, promoters, and regulatory elements has offered a comprehensive view of gene regulation networks. For example, during *Drosophila* development, axis determination is active as early as the blastoderm, and helps form germ layers and determine body part placement, using axis patterning enhancers. A study has shown, using ATAC-seq, that these enhancers show regional differences across the body axes, contributing to axis-specific expression of developmental genes (Bozek et al., 2019). Similarly, in *C. elegans*, a species with a highly consistent developmental process across individuals, the use of ATAC-seq have identified multiple novel and active distal enhancer elements that are active during development (Daugherty et al., 2017)

2.3. Computational analysis of transposable elements

2.3.1. Difficulties in experimental approaches to TEs

Frequently, the features of TEs that make them crucial tools in genome maintenance and regulation pose unique challenges in analyzing TE activity. In addition to the traits that plague studies on other non-coding elements in the genome, such as the lack of direct connection between transcription of TEs and their contribution to the genome, and the high number of potential TE targets that make large-scale screening arrays even more challenging than for protein coding genes, TEs also suffer from their nature as repeat elements, as well as their polymorphic nature. Analyses of experimental insertions into the genome, such as by mutagenesis screening TEs, *piggyBac* and *Sleeping Beauty*, often require artificial tags or barcodes to track their insertion and excision into and out of the genome (Ammar et al., 2012; Talluri et al., 2014; Chang et al., 2019). However, natural TE activity does not display such markers for ease of tracking, and identifying that a specific instance of a TE is

responsible for a biological outcome, especially for novel TE-derived polymorphisms, requires considerable time and expense.

2.3.2. Next generation sequencing in TE analysis

As a result of the challenges in experimental methods, computational methods to help identify likely TE targets of interest have flourished. One of the key technologies in the advancement of TE analysis is next generation sequencing (NGS), which allows for large amounts of short DNA or RNA molecules to be sequenced far more quickly than Sanger sequencing. These sequences are then mapped onto a reference genome or transcriptome using a variety of alignment software. Such software has to be selected appropriately for the analysis at hand. For example, splice-unaware aligners, such as Bowtie, are used for alignment of DNA reads, such as from chromatin immunoprecipitation sequencing (ChIP-seq), to the genome, or RNA reads, such as from RNA sequencing, to the transcriptome (Langmead et al., 2009). On the other hand, splice-aware aligners, such as HISAT, are used to align RNA reads from mature transcripts to the genome, by identifying reads that may map onto multiple exons separated by one or more introns (Kim et al., 2015; Pertea et al., 2016). Some aligners, such as Bowtie2, may be used for multiple types of experiments with modifications (Langmead and Salzberg, 2012). Once the reads are mapped to identify their point of origin in the reference sequence, their biological role and function can be further gleaned by downstream analyses.

Despite the constantly growing role of NGS in TE analysis, there are yet challenges to be fully overcome in the field. Because of the relatively short length of NGS reads, correctly identifying the original instance of a read might be difficult for repeat sequences such as TEs. Read alignment software and pipelines originally designed for protein coding genes or other non-repeat elements often discard multi-mapped reads with high ambiguity. In addition, reference genomes often have repeat sequences masked to avoid diluting the mappability of reads (Goerner-Potvin and Bourque, 2018; Lanciano and Cristofari, 2020). Improvements in the field often come from two different approaches. One approach is the application of long-read, or single-

molecule sequencing, to resolve ambiguities caused by short read mapping to repeated sequences in the genome (Chu et al., 2020). Another is the advancement of TE specific aligners and expression quantifying software, such as SalmonTE (Jeong et al., 2018), SQuIRE (Yang et al., 2019), or SoloTE (Rodríguez-Quiroz and Valdebenito-Maturana, 2022), to better utilize the short reads for the task at hand. Many such tools nevertheless can face difficulties in identifying activity in individual instances, and may map reads onto a “metagene” that covers an entire TE subfamily, which will then require further disambiguation.

2.3.2.1. Assay for transposase accessible chromatin (ATAC) sequencing

As mentioned before, TE activity requires accessible chromatin for relevant proteins to bind to the DNA. Such accessible TEs can be identified through the use of ATAC-seq. Previous studies have used extant pipelines to align and evaluate ATAC-seq reads to TE regions, either by itself or in combination with other epigenetic markers (Miao et al., 2020; Lee et al., 2022). However, as apropos as the use of ATAC-seq, a transposase-mediated assay, is for analysis of TEs, there are currently no TE-optimized aligners or workflows for ATAC-seq reads.

2.3.2.2. Expression profiling of TEs via RNA-seq

As some TEs, such as LINE-1 in humans, display an ORF activity, and retrotransposons require an RNA intermediary for transposition, which can be located in the transcriptome, estimating the expression of TEs can be important. RNA-seq is an NGS-based method of estimating the transcription levels of genes and other genomic features, by sequencing the short reads derived from transcript fragments, and using aligners, identifying the genomic source of the transcript. Briefly, there are three main approaches to RNA-seq: genome-guided, transcriptome-guided, and *de novo* reference assembly. In addition, various library preparation methods for RNA-seq exist. For example, poly-A enrichment creates a targeted pool of mature transcripts with a poly-adenylated tail, while rRNA depletion removes highly abundant rRNAs, and frequently similarly abundant tRNA and mtRNA transcripts, for a larger pool that

includes transcripts without a poly-A modification. With the right combination of alignment and library preparation, TE expression can be quantified. However, despite the presence of TE-specialized tools for RNA-seq, there are yet several challenges to consider in TE expression profiling (Treangen and Salzberg, 2012; Goerner-Potvin and Bourque, 2018; O'Neill et al., 2020).

TE expression can be estimated by two different approaches, as described above. The first of them treats the sequence of the TE subfamily as a metagene, and aligns all TE reads to the metagene, to identify the total transcript level of the TE, regardless of instance of origin. The other assigns each read to a specific instance, in order to estimate the local expression for each instance of a TE subfamily. Such instance-based alignment requires more specialized tools, and can be resource intensive, but provides more in-depth data than the alternative. Recently, it has been shown that the inclusion of transcription start site signal profiling can help improve instance-based TE expression analysis (Savytska et al., 2022)t

2.3.3. Current computational tools for TE analysis

Because of the importance of both TEs in biological systems, and computational tools in TE analysis, there are a variety of software and computational workflows for analysis of TEs. In their seminal 2018 review, Goerner-Potvin and Bourque have covered such tools in depth, dividing them into four broad categories (Goerner-Potvin and Bourque, 2018). These categories include TE repositories, which contain information on the TEs found in the genome, such as their classification, sequence, genomic locations, and any TE-derived polymorphisms; tools for TE discovery and annotation, which can be repository-based or *de novo* discovery; tools for TE polymorphism detection, both germline and somatic; and characterization of TEs and their regulatory impact. In addition, they have also covered various general-use tools adapted for TE use. Since then, novel software has been made available for TE analysis. Some examples of such tools include HiC-TE, which examines the 3D conformation of TE loci to more fully identify their interactions with distal or *trans*-acting elements (Lexa et al., 2022); DARTS, which identifies protein-coding TEs, such as LTRs in plants;

RepExpress, which uses RNA-seq reads to quantify TE expression (Stockwell et al., 2021), and dnaPipeTE, an assembly-free TE characterization and quantification tool (Goubert, 2023).

2.3.3.1. Tools for identification of regulatory functions of TEs

As important as tools for annotation and quantification of TEs are, the regulatory roles of TEs are not solely based on their genomic position and expression. Depending on the impact of TEs on the genome and the epigenome, different tools are required to fully characterize them. For example, to annotate gene transcripts with TE-derived sequences, such as TE-derived genes, or genes with TE insertions into their genic regions, several databases are available. The database TranspoGene contains information on TE insertions into protein-coding genes in various reference genomes, including human, mouse, zebrafish, and nematode, while microTranspoGene, its sister database, includes miRNAs derived from TEs (Levy et al., 2008).

As previously mentioned, TEs can also affect the expression levels of other genes without directly affecting their sequence. There are relatively fewer available tools or databases for the estimation of such regulatory impact. GREAM, a web-based tool, predicts gene-regulating TE instances by examining their over- and underrepresentation nearby protein coding genes (Chandrashekar et al., 2015). However, GREAM does not account for polymorphic insertions, or the chromatin state of the TE being suited for such regulatory roles in a given phenotype, and focuses solely on proximity of TE and gene regions in a reference sequence.

2.3.3.2. Missing perspectives in current tools

While our current knowledge of and analysis tools for TEs are still growing, there are yet gaps in our toolkit that need to be filled. Part of the problem facing TE research is the use of general-use tools for TE analysis, which require specific adaptations and are not necessarily optimized for repeat sequences, especially for instance-based analysis. For example, while methylation analysis of specific TE subfamilies or instances is possible through extant tools, there is no software for doing

so on a repeatome-wide scale. Furthermore, epigenome-wide association studies, such as methylation quantitative trait loci analyses, do not cover specific TE instances. A recent tool for single-cell methylation analysis, scTEM-seq, uses the average methylation levels of a subset of TEs, including SINES, as a proxy for global genome methylation, but does not include information on individual TE instances, and it does not include differential methylation analysis (Hunt et al., 2022).

Another gap that needs to be filled is in identification of the regulatory functions of TEs. Currently available tools, such as the previously mentioned GREAM, work from a gene-focused perspective, and do not take into account the requirement of TE instances to be accessible by transcription factors and other proteins to perform enhancer or insulator functions. Therefore, a tool that can evaluate and include the epigenetic state of TEs is necessary to fully appreciate their regulatory impact on the genome and the transcriptome.

3. Materials and Methods

3.1. Type of study

The study pursues an analytical and computational approach.

3.2. Time and Place of Study

The study was conducted at İzmir Biomedicine and Genome Institute, between December 2021 and August 2023.

3.3. Materials of Study

The study was performed using raw RNA-seq and ATAC-seq reads obtained from a time course experiment spanning 10 days of epithelial-mesenchymal transition (EMT) induced by TGBF3 treatment, and 100 hours of mesenchymal-epithelial transition (MET) induced by removal of TGBF3 and replaying. Samples were obtained at the beginning, at three days, and at 10 days of the EMT process (hereafter referred to as Vehicle, MNM3 and MNM10, respectively), as well as 24 hours, 48 hours, 72 hours and 100 hours after the start of the MET process (hereafter referred to as PT24, PT48, PT72, and PT100 respectively). Reads were sequenced from whole transcriptome libraries of the samples depleted for ribosomal RNAs. The reads were produced prior to this study using an Illumina NextSeq 500 platform, using a paired-end sequencing method.

3.4. Variables of Study

The timepoints of the sequencing reads obtained from the NMuMG cell model of EMT and MET, as well as the choice of genome build, are independent variables. The identified accessible genomic region peaks and the expression values of identified transcripts are dependent variables.

3.5. Tools for Data Collection

To prepare the ATAC-seq and RNA-seq libraries for use in regulaTER functions, data collection and processing operations were performed on the high-performance computing cluster of the İzmir Biomedicine and Genome Center Bioinformatics Platform, with a 40-core Intel(R) Xeon(R) Gold 6148 2.40 GHz processor, 384 GB RAM, and 2 TB of disk space per user, running Centos 7 Linux OS(3.10.0-862.2.3.el7.x86_64 xhci-hcd) as the operating system. The implementation of regulaTER was performed on a personal computer with a 2.6 GHz Dual-Core Intel Core i5 processor, 8 GB RAM, and 256 GB of disk space, running macOS version 11.7.9 (20G1426).

3.6. Study Plan and Calendar

Table 1. Study timeline.

Objectives	Time (months)
Establishment of stand-alone regulaTER functions	1-6
Unification of the regulaTER library and its documentation	7-9
Testing of regulaTER functions on EMT-MET data	10-15
Refinement and debugging of regulaTER functions	12-15
Implementation of regulaTER on independent datasets	15-16

3.7. Data Evaluation

3.7.1. ATAC-seq run data processing, alignment, and peak identification

The raw ATAC-seq read data was used as input for the nf-core ATAC-seq pipeline (version 2.1.2) (Harshil Patel et al., 2023), using the command line interface command “nextflow run nf-core/atacseq --input design.csv --genome mm10 -profile docker --save_reference”. The nfcore/atacseq pipeline performs quality control using

the FASTQC tool for checking FASTQ file quality metrics, and Trim Galore! for removal of adapter sequences and low quality ends during preprocessing. The reads are aligned against the University of California, Santa Cruz Genome Browser mouse genome build mm10 using the BWA aligner. The aligned reads are processed for duplicate removal and filtering of reads aligning to the mitochondrial genome or blacklisted regions using the Picard and SAMtools toolkits. Reads are collated into broad genomic peaks using the MACS2 peak calling software. All software and reference files used in the pipeline were obtained and accessed via a Docker container acquired from the nf-core website.

3.7.2. RNA-seq run data processing, alignment, and analyses

The low-quality bases and reads and possible adapter contamination was removed from the RNA-seq libraries using Trimmomatic (version 0.39) (Bolger et al., 2014). Remaining reads were aligned to the GENCODE GRCm38 (version M21) mouse genome assembly using HISAT2 (version 2.0.5) (Kim et al., 2015). The genome was indexed with the "hisat2-build" command, and the reads were aligned with the "hisat2 -x -U --dta -S -p 6" command. The aligned reads were converted to the binary BAM format using SAMtools utilities (version 1.10) (Danecek et al., 2021) "samtools view" and "samtools sort". The gene expression quantification was performed with the featureCounts function of the Rsubread R library (Liao et al., 2019). Transcripts with 0.5 or lower FPKM in at least one replicate in every samples were removed from the expression table. The edgeR library (version 3.32.0) (Robinson et al., 2010) was used to determine differential gene expression with the count data of remaining transcripts, following TMM normalization and fitting the expression values for the generalized linear model (GLM). Genes with $\text{abs}(\log_2\text{FC}) \geq 1$ and adjusted p-value (false discovery rate, FDR) ≤ 0.05 were considered to be differentially expressed genes (DEG). Gene Set Enrichment Analysis (GSEA) was performed on the DEGs with the GSEA function of the clusterProfiler package (version 3.18.1) (Yu et al., 2012) on the Mus musculus EMT hallmark genes, 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))".

3.7.2. Outline of regulaTER requirements

The regulaTER package comprises several functions to analyze repeat information in conjunction with other biological data, including genomic ranges of differentially expressed genes (DEGs), genome-wide motifs, and genomic sequencing peak files, such as those obtained from ATAC-seq reads. Repeat information for a number of organisms and assemblies, in RepeatMasker format, can be downloaded from RepeatMasker's pre-masked species database (<http://pecies/>). Genomic ranges for DEGs should be provided in BED6 format, with the fourth column corresponding to gene name associated with the range, and the sixth column signifying strand information for the gene. Genomic sequencing peak files can be provided in one of narrowPeak, broadPeak, or summits formats, as produced by MACS2 software, and should be annotated via the ChIPseeker package. Genomic motif information is obtained during configuration of the HOMER suite (Heinz et al., 2010).

3.7.3. Implementation of regulaTER package

The package regulaTER was written in the R programming language (version 4.0.0), and it utilizes a range of tools for its function. To read RepeatMasker annotations into the R environment and rearrange them for use in further calculations, it uses the biomart package (version 1.0.2). To process genomic sequencing peak intervals, it utilizes the GenomicRanges package (version 1.42.0). The dplyr package (version 1.1.1) is used for ease of manipulation of repeat category data, while the BEDtools software (version 2.29.0) is implemented in the randomized relocation of genomic intervals for enrichment analysis. The binom.test and p.adjust functions of the stats package, a part of the base R installation, are required to calculate enrichment or depletion of peak-associated repeat categories. HOMER (version 4.10.3), a suite of tools for genomic sequence analysis, and marge (version 0.0.4.9999), an API for HOMER providing functions to apply HOMER tools to R objects, are used for motif enrichment analysis of repeats of interest.

3.7.4. Function to determine transposable elements enriched in accessible genomic regions

The function accepts as input a genomic peak file, a repeat information file, and a list of BED files corresponding to genomic coordinates sorted by context, as corresponds to the regions annotated by ChIPseeker, i.e. promoter (< 3kb of transcription start site), 5' UTR exons, 3' UTR exons, coding exons, introns, downstream (< 3kb of transcription end site), and intergenic regions. The function initially checks which repeats contain summit nucleotides within their ranges across the genome, and counts how many instances per repeat type, family, or class are associated with a peak region. It then shuffles the summits across genomic ranges of shared context a number of times determined by the user, and calculates the expected count of TEs associated with accessible genomic regions (TEAR). Afterwards, it uses negative binomial testing to identify TEAR categories that are enriched across the genome, with TEARs displaying higher observed counts than expected with an FDR of 10^{-3} or lower considered to be enriched. For this study, we further only took into consideration TEARs with at least 10 observed instances, to eliminate the biases that can be introduced by such small sample sizes.

3.7.5. Function to determine accessible transposable elements enriched in neighboring regions of differentially expressed genes

The function determines whether TEARs are more likely to be near DEGs, potentially contributing to transcriptomic diversity. For purposes of this thesis, genes identified as differentially expressed during the EMT and MET processes were chosen as genes of interest. The neighboring regions of identified genes (up to X bases upstream of the transcription start site and downstream of the transcription end site) are intersected with the TEAR intervals to identify if the repeats are responsible for any regulatory elements that may affect the expression of the DEGs.

3.7.6. Function to identify the genomic motifs contributing to the regulatory functions of transposable elements in accessible regions

The function takes the binomial test results of either of the previous functions and the ranges of the enriched repeats from the repeat masker file, leveraging the HOMER genomic sequencing analysis software to identify motifs both de novo and previously documented enriched in those ranges, aiming to determine whether repeats that contribute to genomic accessibility affect transcriptional regulation of genes as well, via generating transcription factor binding sites. The function has two separate input requirements to calculate the necessary background genomic regions, depending on whether it is used on the output of the TEAR or DATE function. For the TEAR function, it takes the genomic ranges of the accessible instances of a TE subfamily of interest as the query, and the non-accessible instances of the same subfamily as the background. For the DATE function, it takes the genomic ranges of the accessible TEs of a subfamily of interest found in the DEG promoter regions as the query, and accessible TEs of the same subfamily of interest found outside the promoter regions as the background.

3.7.7 Functions for creating graphical representations of the outputs

The function takes the binomial test results of the TEAR function, a height, and a length for the output plot, and the directory where the resulting plot files should be located, and generates two exploratory plots for each of repeat subfamily, repeat family, and repeat type for quick visualization of the TEAR output. The first plot is a scatter plot, with each TE category represented by a dot, where the X-axis represents the fold enrichment of observed instances of the category versus the expected instances, as calculated by the shuffling step of TEAR, while the Y-axis represents the statistical significance of the enrichment (in $-\log_{10}$ of the false discovery rate). Significant TEs are colored in cyan ($FDR \leq 0.05$), while non-significant TEs are colored red. The second plot is a combined bar and line chart. The x-axis represents enriched TE categories. The left-hand Y-axis corresponds to the bar chart, and represents the total number of observed TE instances in the category. The right-hand Y-axis

corresponds to the line chart, and represents the ratio of observed TE instances to the total number of TEs in the category. The function uses the ggplot2 graphics grammar R library to generate the graphical outputs.

3.8. Limitations of Study

The findings are limited to the interpretation of the ATAC-seq and RNA-seq cell model and the independent datasets used in this study. In addition, any findings are predictive in nature, except where confirmed by previous experimental research, and will need to be validated in biological systems using experimental studies.

3.9. Ethics Committee Approval

The study was approved by Izmir International Biomedicine and Genome Center Ethics Committee for Non-Invasive Research (iBG-GOEK) on the 30th of March, 2023 (Protocol Number: 2022-032).

4. Results

4.1. Quality control and interpretation of ATAC-seq reads

Using the multiqc tool output of the nf-core ATAC-seq pipeline, the produced reads, alignments, and generated peaks of the ATAC-seq libraries were assessed for quantity and quality. Comparing the libraries before and after adapter and low-quality read removal by Trim Galore!, the average percentage of remaining reads in merged libraries was 99.83%, with a replicate minimum of 99.64%, and a maximum of 99.94%, indicating high quality reads overall. Post-trim merged library sizes ranged from 53081171 to 137661787, with a median size of 85768856 (Table 2).

Afterwards, the ATAC-seq reads were aligned to the UCSC Genome Browser mm10 reference genome for *Mus musculus*. The alignment stats were assessed using the samtools flagstats output. All replicates had an alignment rate of over 95%, showing generally favorable alignment rates (Figure 3).

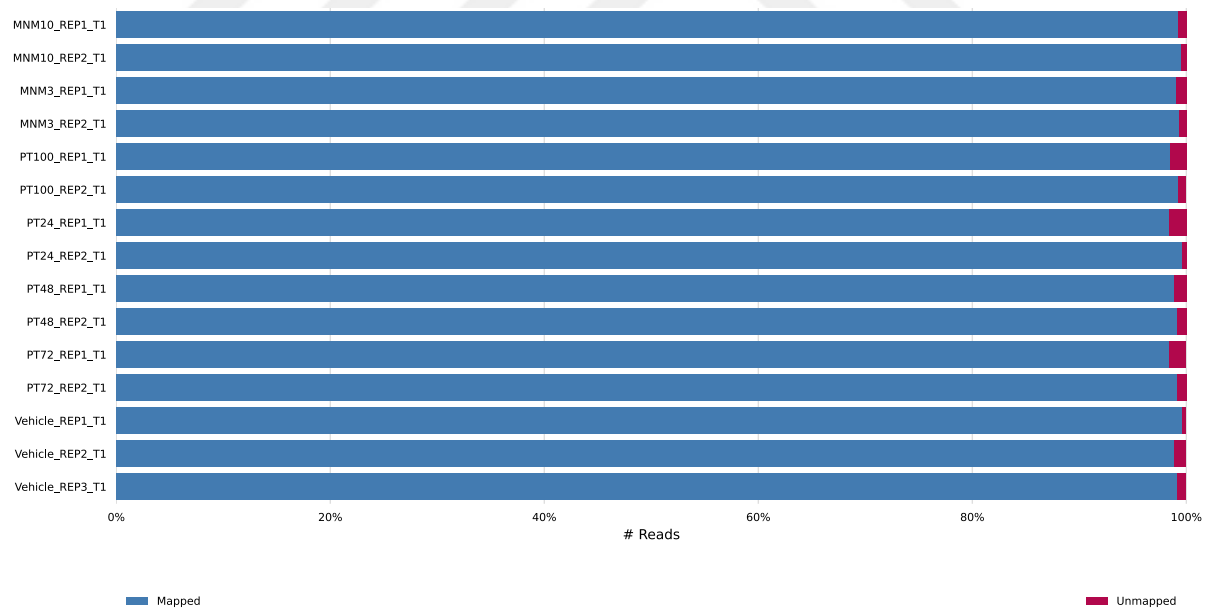


Figure 3. Mapping percentages of ATAC-seq libraries. Blue bars indicate the percent of mapped reads in replicate, while maroon bars indicate the percent of unmapped reads. All bars total to 100% per replicate.

Table 2. Evaluation of pre- and post-Trim Galore! ATAC-seq library read counts and percentages.

Sample	Pre-Trim			Post-Trim			
	Total Reads	%GC	%Deduplicated	Total Reads	%GC	%Deduplicated	% Remaining Reads
Vehicle Rep 1 Forward	8389117	49	75.95025476	8383199	48	78.56081486	99.93%
Vehicle Rep 1 Reverse	8389117	49	74.29388445	8383199	48	77.09073865	99.93%
Vehicle Rep 2 Forward	14492499	48	69.93032749	14439660	48	70.2557839	99.64%
Vehicle Rep 2 Reverse	14492499	50	72.01711064	14439660	49	73.88621847	99.64%
Vehicle Rep 3 Forward	30363526	48	65.60543254	30258312	48	65.92493195	99.65%
Vehicle Rep 3 Reverse	30363526	49	66.90207436	30258312	48	69.16049512	99.65%
MNM3 Rep 1 Forward	5278192	51	76.5837479	5271157	50	78.18091974	99.87%
MNM3 Rep 1 Reverse	5278192	51	80.73454063	5271157	51	82.15092887	99.87%
MNM3 Rep 2 Forward	132556168	50	52.16843199	132390630	49	52.33750319	99.88%
MNM3 Rep 2 Reverse	132556168	50	50.95338811	132390630	50	55.33593542	99.88%
MNM10 Rep 1 Forward	5349490	50	76.76212144	5344555	49	77.8530145	99.91%
MNM10 Rep 1 Reverse	5349490	51	80.94767073	5344555	50	81.91938671	99.91%
MNM10 Rep 2 Forward	128136517	49	51.68935264	128026322	49	51.86912738	99.91%
MNM10 Rep 2 Reverse	128136517	50	51.05609427	128026322	49	54.04431405	99.91%
PT24 Rep 1 Forward	4614105	51	78.37858939	4607160	50	78.98047438	99.85%
PT24 Rep 1 Reverse	4614105	52	85.37550263	4607160	52	85.88531774	99.85%

PT24 Rep 2 Forward	65348308	50	68.1372556	65312038	50	68.26785778	99.94%
PT24 Rep 2 Reverse	65348308	49	58.70902519	65312038	49	60.0318353	99.94%
PT48 Rep 1 Forward	3870169	51	77.36338265	3857351	50	78.68829621	99.67%
PT48 Rep 1 Reverse	3870169	52	81.10779786	3857351	51	82.34926806	99.67%
PT48 Rep 2 Forward	107417415	50	46.93587845	107100088	50	47.23319192	99.70%
PT48 Rep 2 Reverse	107417415	51	47.49909652	107100088	51	50.63930217	99.70%
PT72 Rep 1 Forward	2938204	51	80.27098179	2928944	49	81.22981967	99.68%
PT72 Rep 1 Reverse	2938204	52	84.87836513	2928944	51	85.72880713	99.68%
PT72 Rep 2 Forward	83037543	50	56.43519264	82839912	49	56.70257483	99.76%
PT72 Rep 2 Reverse	83037543	51	58.64393743	82839912	50	61.475814	99.76%
PT100 Rep 1 Forward	4228215	53	74.37573237	4219543	51	76.35938061	99.79%
PT100 Rep 1 Reverse	4228215	53	79.17322241	4219543	53	81.01910124	99.79%
PT100 Rep 2 Forward	76982532	51	58.4601305	76882534	51	58.78412674	99.87%
PT100 Rep 2 Reverse	76982532	52	48.72150981	76882534	51	53.50752075	99.87%
Merged Sample	Pre-Trim Merged Reads		Post-Trim Merged Reads			% Merged Remaining Reads	
Vehicle Total	53245142		53081171			99.69%	
MNM3 Total	137834360		137661787			99.87%	
MNM10 Total	133486007		133370877			99.91%	

PT24 Total	69962413	69919198	99.94%
PT48 Total	111287584	110957439	99.70%
PT72 Total	85975747	85768856	99.76%
PT100 Total	80852701	80739885	99.86%
Average	96091993.4	95928459	99.83%

The aligned read files generated had duplicates, discordant pairs, and unmapped reads removed via samtools and Picard toolkit during the pipeline. Following these filters, the distributions of remaining reads were checked with read fingerprints and read distribution profiles using deeptools generated plots (Figure 4). As expected, accessible regions are highly represented around the TSS sites of genes, especially in the promoter regions, with lower accessibility in gene bodies (Figure 4A). However, the genomic fingerprint of the libraries had a more right-leaning elbow point than the ideal distribution, indicating slight local bias in reads (Figure 4B).



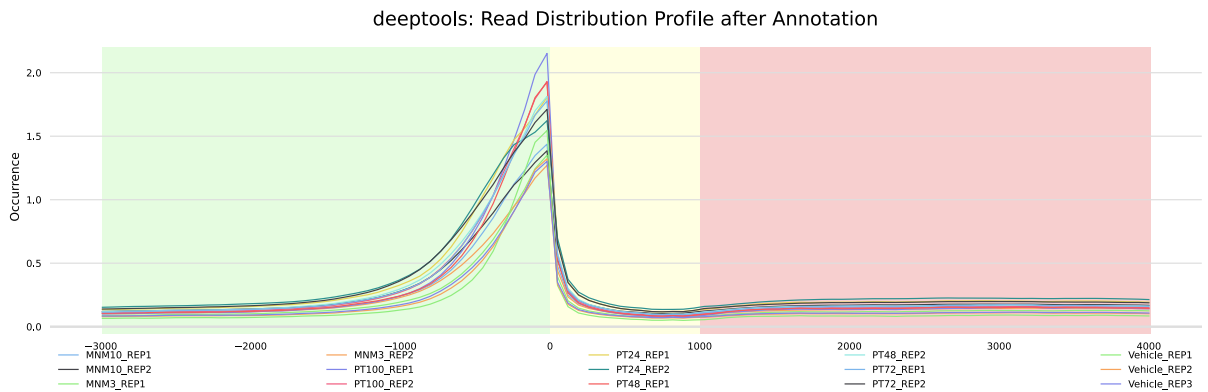
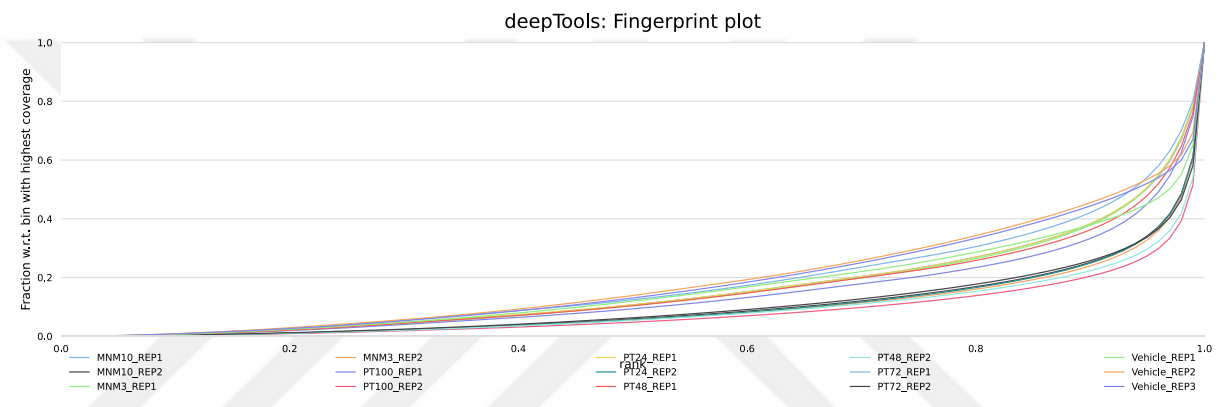
A.**B.**

Figure 4. Fingerprint and read distribution profile plots of ATAC-seq libraries. (A) Fingerprint plot of read coverage across genome. X-axis reflects fraction of genomic bins covered, while Y-axis represents the fraction of mapped reads covering those bins. (B) Read distribution profile nearby known genes. Yellow regions represent transcript coding regions (transcription start site to transcription termination site), scaled to the same length. Green regions indicate upstream regions of genes, while red regions indicate the downstream regions, up to 3 kilobases in both directions. Both green and red regions are unscaled.

To further estimate the qualities of the ATAC-seq libraries, the *ataqv* output of the pipeline was examined (Figure 5). First, we looked at the distribution of fragment lengths for high-quality reads that aligned to the autosomes. The expected distribution of fragment lengths shows a strong peak around 50-100 base pairs, based on the library preparation, with smaller peaks around 150 base pairs, 300 base pairs, and 450

base pairs, where one, two, or three nucleosomes would be bound to the genome. Deviations from the expected distribution was shown in the “Distance from reference FLD” plot. Other than replicate 1 of the Vehicle samples, all libraries showed undertransposition, i.e. reduced Tn5 activity than expected, which creates a lower amount of short peaks, and higher amount of nucleosomal peaks (150-450 base pairs). In contrast, read distributions across the genome, both in genomic territory covered (Figure 4) and the fraction of reads aligning to peak regions (Figure 6) are as expected for all libraries. Together, these data indicate that, despite the under-fragmentation of the genome, the libraries are strongly representative of accessible genomic regions.

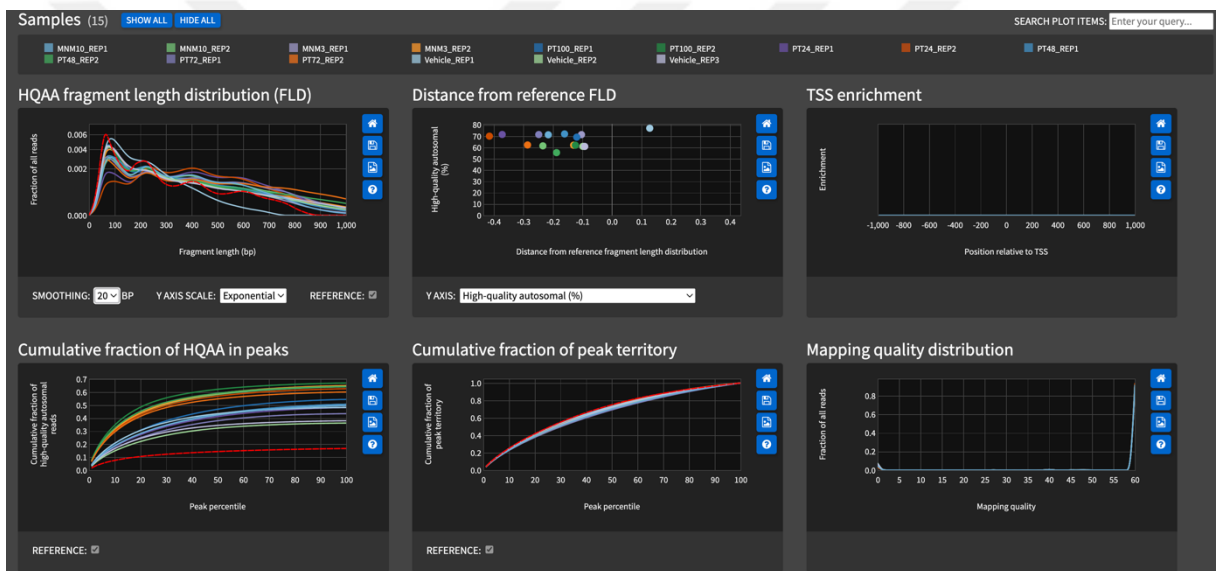


Figure 5. Plots representing the ataqv quality metrics of the ATAC-seq libraries.

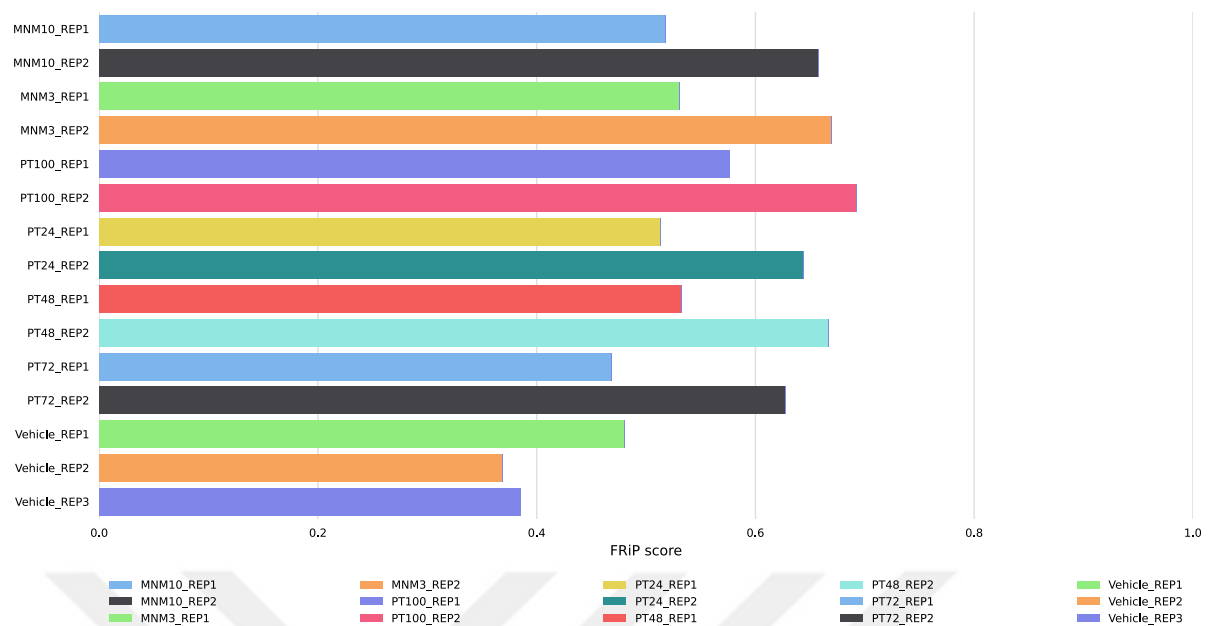


Figure 6. Fraction of reads in peak (FRiP) scores of ATAC-seq libraries. X-axis shows the FRiP score, while Y-axis indicates sample names. All samples have a FRiP score of ≥ 0.3 , indicating ideal fragment distribution in peak regions.

Afterwards, the aligned reads from each replicate of the same sample were merged into single libraries, and the ATAC-seq peaks of each timepoint were calculated using MACS2, with the broad peak setting. A total of 134722 consensus peaks were calculated across all samples. In each sample, a majority of peaks were located in intronic and intergenic regions (Figure 7). Furthermore, TTS sites were consistently far less accessible than both the promoter and exonic regions of genes.

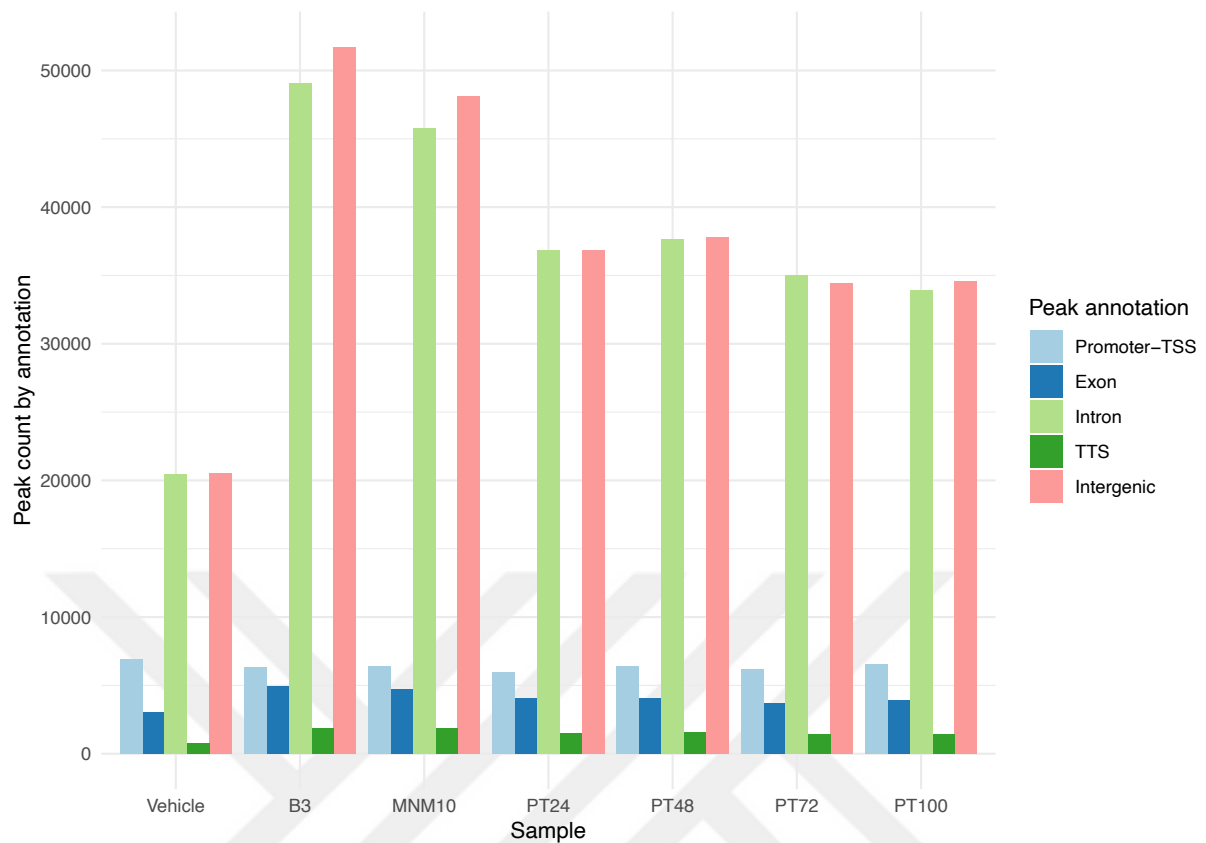


Figure 7. Barplot of peak count in annotated genomic regions. X-axis represents sequenced samples, with each sample divided into five different bars, indicating promoter-TSS (up to 3000 bases upstream of genic region), exon, intron, TTS (up to 3000 bases downstream of genic region) and intergenic peaks, respectively. Y-axis represents number of peaks corresponding to the annotation.

To assess whether the identified peaks could sufficiently serve as accurate phenotypic markers of each timepoint, the number of timepoints each peak was found in was examined. To do so, the file containing the Boolean categorization of consensus peaks was generated by MACS2 was used. Briefly, each peak was given a state of TRUE (i.e. peak had sufficient read coverage) or FALSE (i.e. peak lacked sufficient read coverage) for each sample. For each peak, the number of samples marked TRUE was calculated, and the results were plotted in a stacked bar graph (Figure 8).

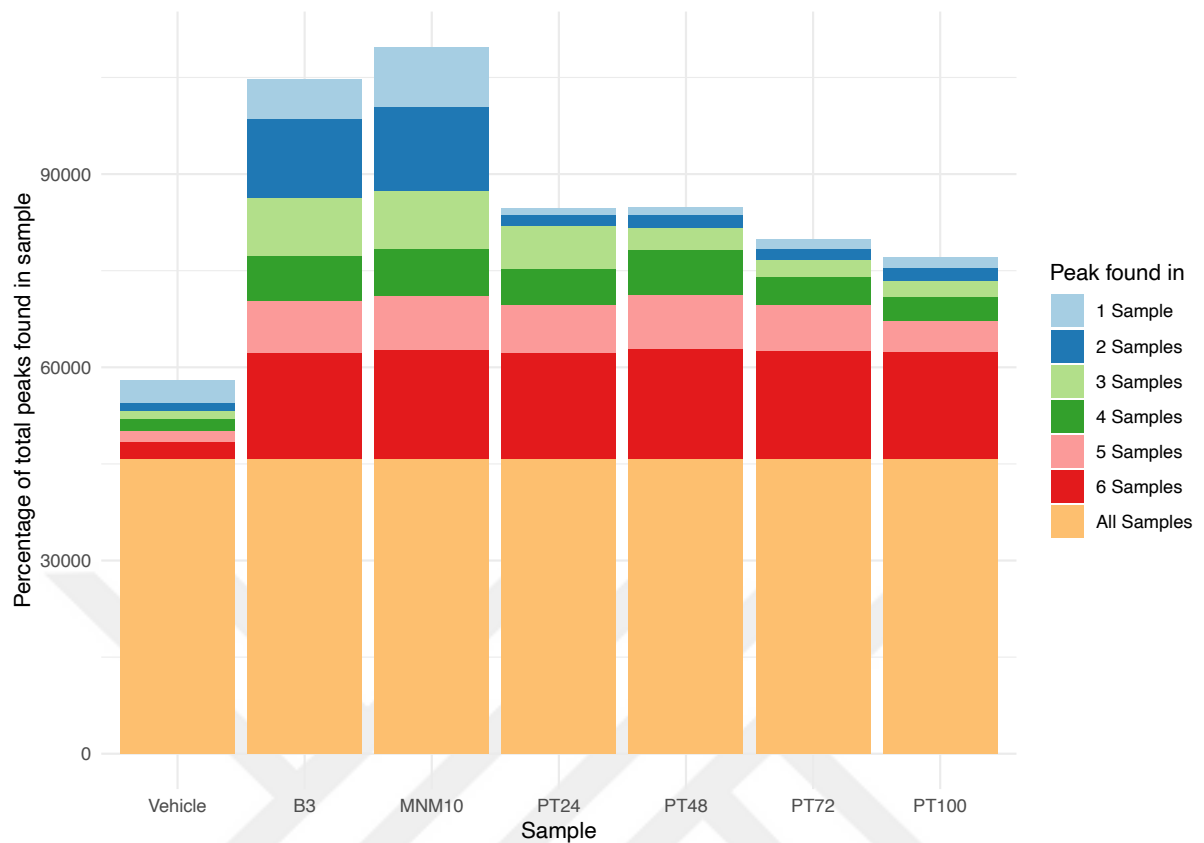


Figure 8. Stacked barplot of number of peaks in each sample categorized according to number of samples peak is covered in. X-axis represents the timepoints of the experiment, while Y-axis represents the total number of peaks with sufficient coverage in sample. Colored sections of the graph correspond to the number of samples the peaks has sufficient coverage. In each stacked bar, the colored sections are ordered in ascending number of samples, from 1 sample at the top (peak is specific to the sample) to all samples at the bottom (peak is constitutively accessible in sequenced samples).

Out of the 134722 consensus peaks identified, a slim plurality of 45787 were identified in all samples, while 25022 were identified in only a single sample, and 28397 of them were identified in two or three samples. Among the peaks with coverage in two or three samples, 18896 of them were located in MNM3 and MNM10 samples, each of them sequenced during the EMT process. 5554 of these peaks were shared with the PT24 sample, sequenced during early MET, indicating that the early MET

samples show distinct similarities to mesenchymal cells in genomic accessibility patterns.

To further detail the changes in coverage patterns in peaks across different samples, differential peak binding analysis was performed, using the consensus peak read counts generated by the nf-core atacseq workflow. Peak counts of different timepoints were compared using the edgeR exactTest function. Peaks with a logFC value ≥ 1 and FDR ≤ 0.05 between consecutively analyzed timepoints were considered to be differentially accessible. Differentially accessible peaks were annotated using the ChIPseeker R library, as compatible with the mm10 genome assembly and gene annotations. For downstream analyses, TE instances located within these differentially accessible peaks were considered to be associated with accessible regions for that timepoint.

4.2. Characterization of the transcriptome during EMT and MET events and the identification of differentially expressed genes

To understand whether the EMT and MET phenotypes and epithelial and mesenchymal morphologies could be successfully recapitulated in the data obtained from the Vehicle, MNM10, and PT100 samples, RNA-seq libraries from those samples were sequenced and analyzed. Key epithelial and mesenchymal marker genes were visualized to highlight the EMT (Figure 9A) and MET (Figure 9B) processes.

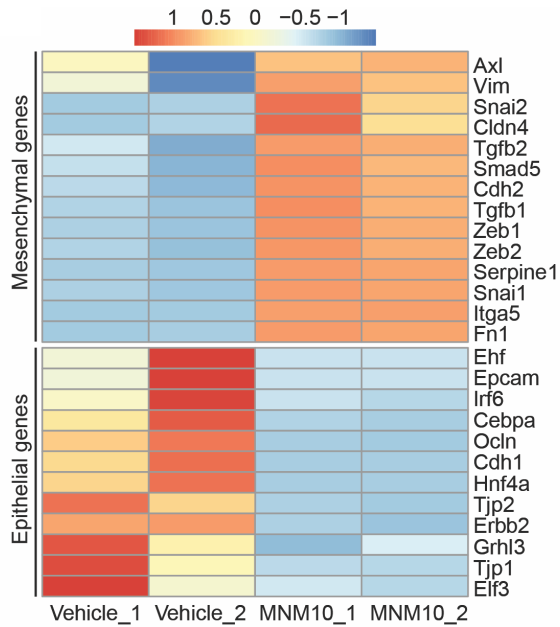
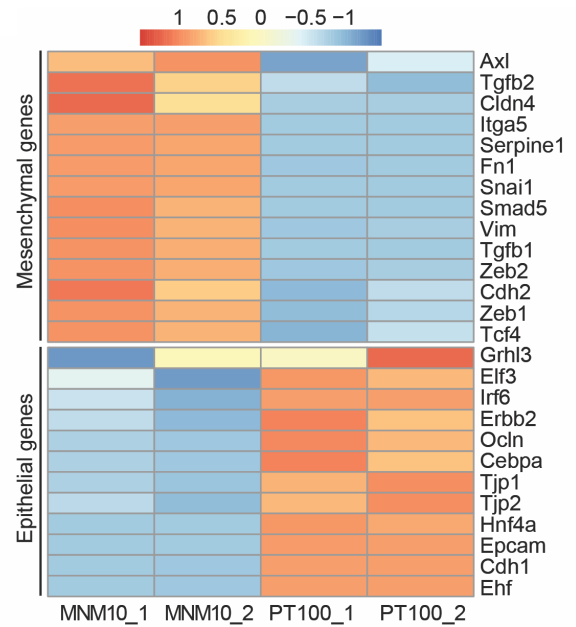
A.**B.**

Figure 9. Heatmap of key epithelial and mesenchymal marker genes during (A) EMT and (B) MET processes. The columns indicate samples, while the rows indicate genes. Cell color indicates the expression level of gene in the sample compared to the mean of all samples shown in the heatmap, with red cells indicating higher-than-average expression, and blue cells indicating lower-than-average expression.

The gene expression data taken from these samples was also used to perform GSEA (Figure 10).

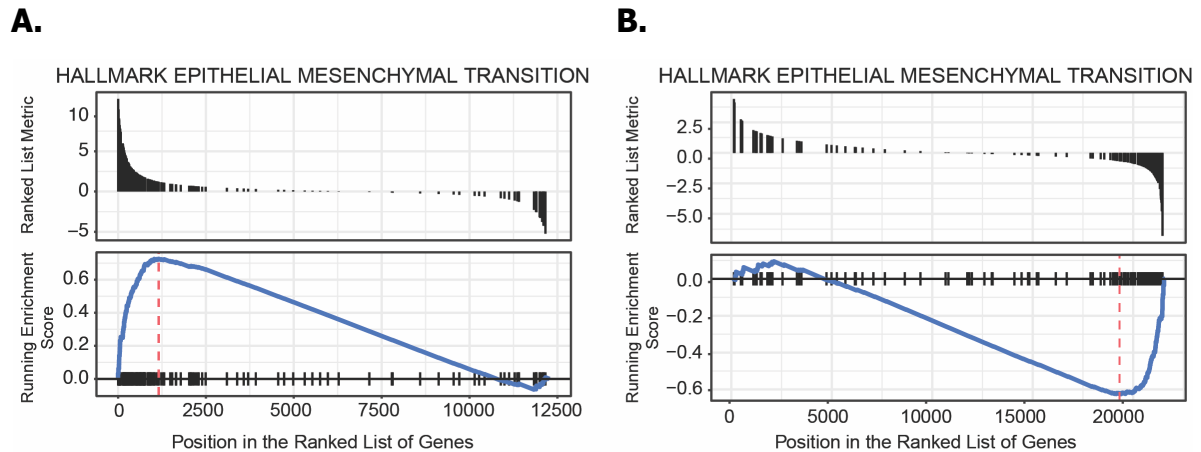


Figure 10. GSEA plots of EMT hallmark genes during (A) EMT and (B) MET processes. The X-axis indicates the genes as ordered by the GSEA ranking list. The blue line in the bottom panel reflects the enrichment score of each gene.

Together, these findings confirmed that the RNA-seq libraries successfully reflected the expected EMT and MET phenotypes, and could be used for downstream analyses including DEGs.

4.3. Identification of TEs enriched in accessible regions using the TEAR function of regulaTER

To identify the enrichment of TEs in accessible regions, the differentially accessible peak regions and the RepeatMasker TE annotations for the mm10 genome were used as inputs for the TEAR function of regulaTER. For this analysis, each peak was shuffled 10 times, as described in subheading 3.7.4. In the MNM10 sample of EMT, 121 TE subfamilies showed such enrichment, while 45 subfamilies showed enrichment in the PT100 sample of MET, with 24 subfamilies shared by both sets. The top 10 subfamilies in each sample were examined in further detail to identify TEs with potentially significant contributions to the EMT and MET processes (Table 3-4). Both top 10 lists had 5 TE subfamilies in common (RMER6C, MIRb, RMER15, MIR, and RLTR48A). Furthermore, subfamilies of the MIR TE family were prominent in both EMT and MET. In addition to the MIRb and MIR found in both sets, MIR3 was among the top 10 enriched TEs in EMT, and accessible instances of these three subfamilies together were located in the proximity of 1423 genes (≤ 5000 bp), including Zeb1, a

known EMT regulator. In comparison, the enriched MIR subfamilies were located near 343 genes in MET, including *Cdh1*, a core marker of the epithelial phenotype.

The full enrichment tables were also visualized using the `MakePlots` function (Figure 11).



Table 3. Top 10 TE subfamilies enriched in accessible peaks in EMT.

Repeat subfamily	Total genomic count	Observed	Expected	FDR	Fold change
RMER6C	4930	176	35	5.59E-64	5.05
MIRb	41669	1051	636	6.43E-51	1.65
RMER15	18012	466	237	6.49E-39	1.97
MIR	38078	865	571	1.01E-29	1.51
RLTR17	1752	91	22	1.75E-27	4.22
ORR1F	22992	488	286	9.40E-27	1.71
RLTR48A	621	57	10	3.77E-24	5.87
MLT1H	2337	94	28	6.42E-22	3.31
MIR3	15244	430	264	3.21E-20	1.63
RLTR30D2_MM	733	50	10	1.01E-18	5.22

Table 4. Top 10 TE subfamilies enriched in accessible peaks in MET.

Repeat subfamily	Total genomic count	Observed	Expected	FDR	Fold change
RLTR48A	621	27	3	2.48E-17	9.00
L2a	23921	282	172	7.74E-15	1.64
L2b	13281	190	110	2.34E-12	1.72
RMER16A2	285	13	1	5.02E-11	10.11
MIR	38078	377	280	1.82E-08	1.35
RMER15	18012	165	104	1.93E-08	1.59
RMER6C	4930	37	13	3.85E-08	2.88
RLTR13C1	321	16	3	9.64E-08	5.60
MIRb	41669	419	324	2.19E-07	1.29
RLTR13D3A1	604	22	6	3.18E-07	3.85

4.4. Identification of accessible TEs enriched in proximity of differentially expressed genes using the DATE function of regulaTER

Following the application of the TEAR function, the enrichment of accessible TEs near DEGs was examined. The DATE function was used, which uses a different enrichment method for more accurate identification of DEG-associated TEs, with the subfamily output of TEAR and the genomic coordinates of upregulated DEGs determined under section 4.2, alongside the 'regionSelection = "gene"' and 'distance = 5000' parameters. As a result, 32 TE subfamilies that show DEG-associated enrichment were identified in both EMT and MET processes. For more in-depth analysis, we selected the 10 TE subfamilies with the highest enrichment scores for both processes (Table 5-6).

Table 5. Top 10 TE subfamilies enriched near DEGs in EMT.

Repeat subfamily	Total genomic count	Observed	Expected	FDR
B4A	107730	147	79	4.147360e-11
B1_Mus1	95439	90	41	2.069845e-10
ID_B1	110809	159	97	3.809269e-08
RSINE1	117117	119	68	1.068608e-07
B3A	93776	96	52	2.307594e-07
B3	148763	138	86	1.144936e-06
MIRb	41669	121	73	1.271693e-06
B1F	45268	69	35	1.867387e-06
ID4_	26391	45	19	2.030584e-06
B1_Mur3	24761	39	16	6.080643e-06

Table 6. Top 10 TE subfamilies enriched near DEGs in EMT.

Repeat subfamily	Total genomic count	Observed	Expected	FDR
B3A	93776	81	32	2.267612e-12
B4	64981	66	27	1.297844e-09
MTD	51250	44	15	7.332148e-09
B1_Mus1	95439	65	29	4.674421e-08
MIRc	22271	44	16	4.674421e-08
ID_B1	110809	123	71	1.048579e-07
B2_Mm2	89526	66	31	2.331149e-07
B4A	107730	77	40	1.000689e-06
B3	148763	95	56	9.591543e-06
L2a	23921	38	16	1.494231e-05

The EMT and MET processes had 5 TE subfamilies in common among their top 10 lists: B1_Mus1, B3, B3A, B4A, and ID_B1. All of these subfamilies are short interspersed nuclear elements (SINE) and members of the B element family, the murine orthologues to the primate-specific Alu elements. To further determine the

potential impact of DEGs near these accessible TEs on the EMT and MET morphology, we used GO term enrichment analysis on upregulated genes in the proximity (≤ 5000 bp) of these five TE subfamilies in EMT and MET (Figure 12).

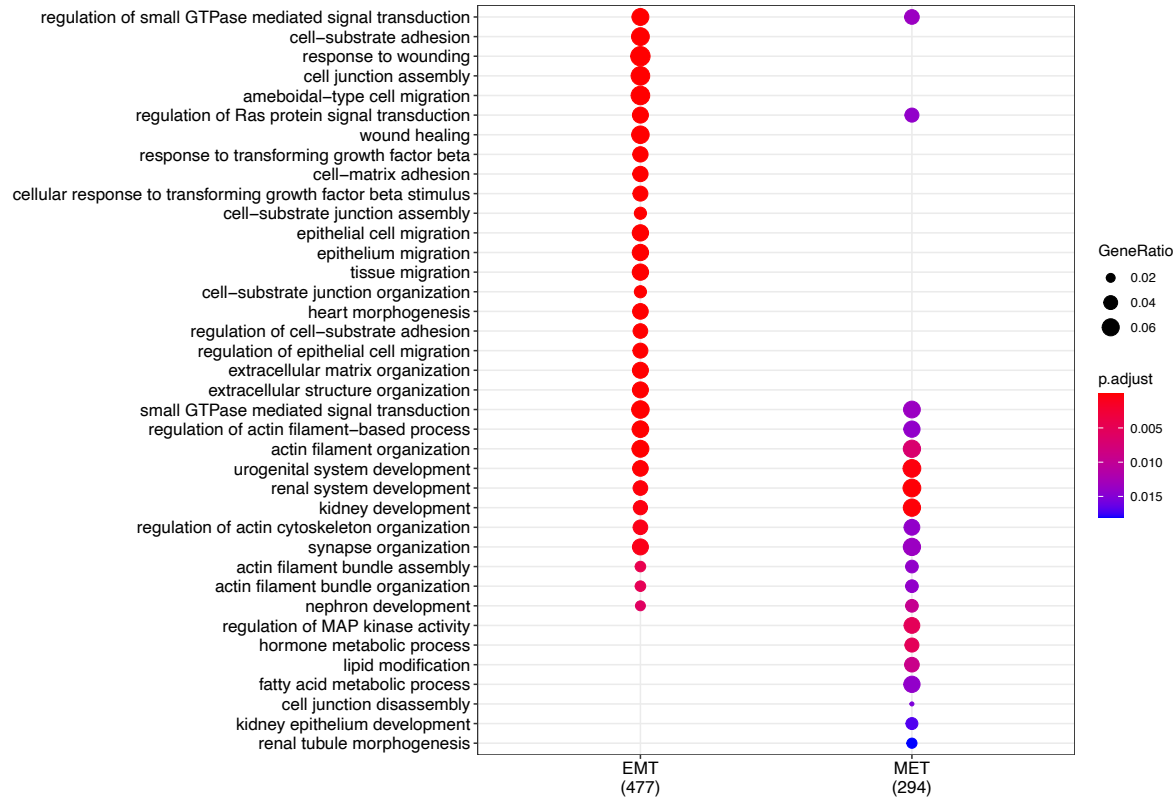


Figure 12. GO term enrichment dotplot of DEGs near accessible TEs of interest in EMT and MET. The X-axis indicates phenotype, while the Y-axis indicates the GO terms. Dot size indicates the ratio of the number of DEGs associated with the term to the total number of term-associated genes, while dot color indicates the adjusted p-value of the enrichment, with red dots indicating higher enrichment than blue dots.

According to these results, the enriched biological pathways were found to be associated with EMT and MET. Notably, response to TGF β , regulation of substrate adhesion, migration, and wound healing, are all terms enriched in EMT, while the biological pathways enriched MET were associated with developmental processes or proliferation. Overall, the highlighted TEs are implicated to take a role in EMT and MET regulation.

4.5. Motif analysis of accessible TEs enriched near DEGs using the FindMotifs function of regulaTER

After the DATE and GO term enrichment analyses were conducted, and the five B element subfamilies described in section 4.4 were implicated in the regulation of EMT and MET, motif analysis was performed to identify the TFs binding to these TE regions which might have active roles in TE-mediated gene regulation. The results showed accessible B1_Mus1 and B4A instances in EMT were enriched for FoxA1 and TEAD binding sites, while ID_B1 loci in MET were enriched for FoxA2 motifs. For further validation, individual differentially accessible loci neighboring DEGs in both processes were visualized using the UCSC Genome Browser. These loci were then compared to ChIP-seq data for FoxA1, TEAD, and FoxA2 TFs obtained from the ReMap 2022 database (Figure 13).

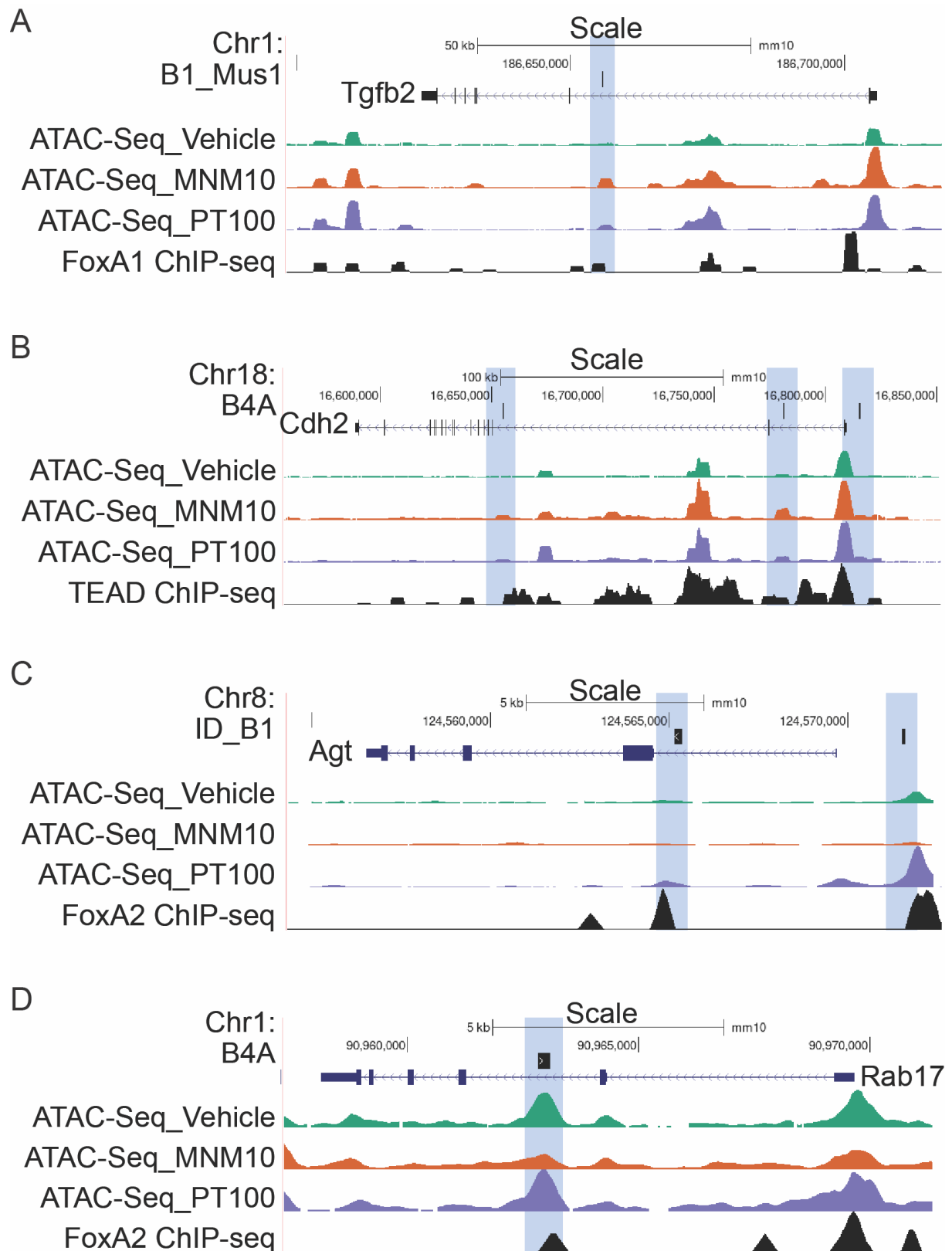


Figure 13. UCSC Genome Browser visualization of TF binding to DEG-associated TE loci.

4.6. Application of regulaTER to independent datasets

To further validate the use of regulaTER in other data types and organisms, analysis results from independently generated datasets were obtained from the Gene Expression Omnibus (GEO) database. The first of these datasets comprised the Foxa2 ChIP-seq (GSE90536) and RNA-seq (GSE90534) results of induced hepatocyte (iHep) differentiation derived from mouse embryonic fibroblasts (MEF) (Horisawa et al., 2020), using the iHep – undifferentiated MEF comparison to determine the DEGs. The second dataset comprised the ATAC-seq (GSE210496) and RNA-seq (GSE210499) results of a human EMT model, TGF β 1 treated MCF10A cell line (Antón-García et al., 2023), using the 72 hours of TGF β 1 – 0 hours of solvent treatment comparison. The peak and RNA-seq count files were directly obtained from GEO.

The results of the TEAR function revealed only 14 TE subfamilies to be enriched in Foxa2 binding sites. The low number of TE subfamilies compared to the EMT – MET data obtained from NMuMg cells is due to the reduced footprint of TF binding compared to accessible chromatin. Among these TEs, L1 retrotransposons and hAT-Charlie DNA transposons displayed higher prominence. Using the DATE function on these samples showed only a single enriched TE subfamily, B4, during iHep differentiation. The FindMotifs results of the B4 subfamily showed enrichment of Oct and Foxa1 binding sites. Both of these TFs have roles in cellular reprogramming, giving weight to their enrichment near DEGs in iHep differentiation.

As an EMT model, the family enrichment results of MCF10A samples largely resembled the NMuMG samples, with MIR and ERV families showing prominence. However, due to the lack of murine-specific B elements in the human genome, the top 10 DATE results showed a large difference, with LINE and MIR representing eight out of 10 subfamilies. However, the remaining 2 subfamilies, as well as a higher proportion of TEs with weaker enrichment, were members of the Alu family of SINEs, the primate orthologues of B elements. The FindMotifs results showed four of these TEs, L2a, L2b, AluJb, and AluY, were enriched for Forkhead transcription factors, including Foxa1, Foxh1, and Foxb1. These results indicate the applicability of regulaTER to a wider range of data types and genomes, and the potential cooperation of TEs and Forkhead

transcription factors in EMT – MET and cellular reprogramming events in general across multiple mammalian species.

4.7. Description and potential uses of regulaTER

The primary use of regulaTER is to identify the enrichment of TE loci of specific subfamilies, families, or classes in specific genomic contexts, and of the genomic motifs located in these TE loci. Using regulaTER, researchers can use many types of genomic and transcriptomic sequencing data to quickly and efficiently determine TEs with potential roles in their biological model of choice, including in biological systems and contexts outside of the scope of this study. This is possible due to the flexible design of individual core regulaTER functions, which there are three of. regulaTER is thoroughly documented and currently accessible on the GitHub database (Figure 14).

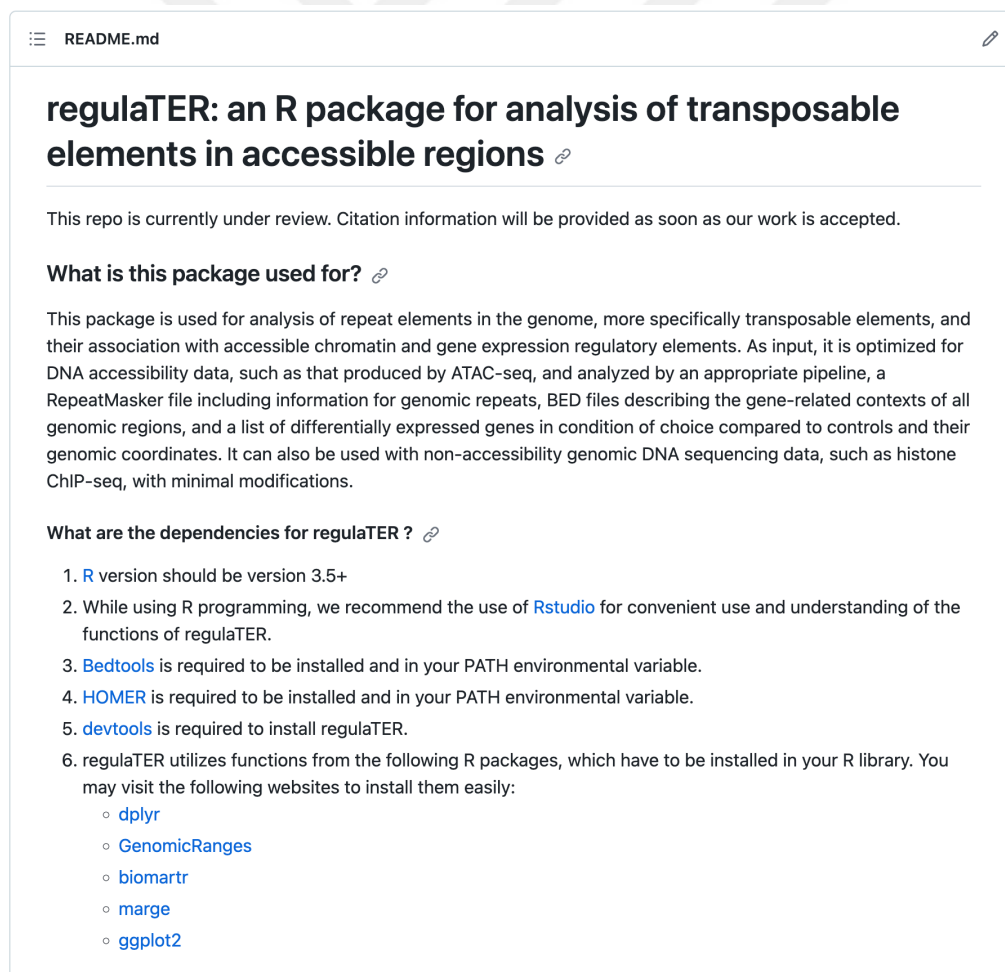


Figure 14. Partial screen capture image of the README.md file of regulaTER on GitHub.

The regulaTER library is primarily based on the enrichment results of TEs in genomic peaks, as determined by the TEAR function. Briefly, TEAR uses negative binomial testing and randomized peak distribution to determine whether the number of TE loci in specific subfamilies, families, and classes found in the peak regions is higher or lower than truly random TE integration. In addition to the ATAC-seq and TF ChIP-seq data used in this study, genomic sequencing data including, but not limited to, histone ChIP-seq, bisulfite sequencing, and MeDIP-seq, can be used to identify the enrichment of specific TEs under specific histone markers (e.g. H3K27ac for active enhancer markers) and methylated DNA regions. Furthermore, as different genomic peaks cover different lengths of DNA, they often require specialized peak calculation methods. For example, MACS2, a popular peak-calling software, can calculate “narrow” peaks, for transcription factor binding sites, methylated regions, and some histone markers, and “broad” peaks, for nucleosome binding or accessible DNA. The TEAR function can accept either type of peak file. Only the summit nucleotide of narrow peaks, where the peak binding is most concentrated, is used to determine the peak occupancy of a TE locus. For broad peaks, the user can determine the amount of expected overlap between the peak region and the TE, which can be at any part of the peak. In cases where the summit nucleotide information is unavailable for narrow peaks, such as for TF ChIP-seq data analyzed by other peak callers, or consensus peaks aggregated from multiple samples, the broad peak mode of the TEAR function can be used.

Additional numerical and statistical parameters used for the TEAR function are described in the regulaTER manual and the function documentation (Figure 15). This level of input flexibility allows the use of regulaTER for TE research outside of the scope or the hypothesis of this study.

Transposable Elements in Accessible Regions - Calculate Enrichment of Repeats Associated with Genomic Peak Regions

Description

Calculates the p-values of peak and repeat associations using negative binomial distribution and multi testing correction, based on genome shuffling of peak regions.

Usage

```
TEAR(inputPeakFile, pathList, numberOfShuffle = 1, repeatMaskerFile, format, alternative = "greater", minobserved
```

Arguments

<code>inputPeakFile</code>	GRanges object of peak information in one of narrowPeak, broadPeak, or summits file formats, annotated via ChIPseeker and read using the readPeakFile function, with a minimum of seqnames, ranges, strand required columns, as well as summit (for narrowPeak only) and annotation columns.
<code>pathList</code>	a list of length 7, with the names "Promoter", "Exon", "Intron", "5UTR", "3UTR", "Downstream", and "genomeSizePath", and character values declaring the file paths of BED files identifying genomic regions of respective annotations and chromosome sizes. Refer to ShufflePeaks for further information.
<code>numberOfShuffle</code>	integer determining the number of shuffle operations to be performed for calculation of peak - repeat association enrichment.
<code>repeatMaskerFile</code>	data frame containing repeat information, generated via read_rm .
<code>format</code>	character vector identifying peak file format, one of "narrow", "broad", or "summit", corresponding to ".narrowPeak", ".broadPeak", and "summits.bed" files generated by MACS2, respectively.
<code>alternative</code>	character vector identifying what the null hypothesis of the binomial test is. One of "greater", "less", or "two-sided", or the initial letter of such.
<code>minobserved</code>	integer identifying minimum number of observed peak - repeat associations to be included in the output.
<code>minoverlap</code>	long integer identifying minimum number of overlapping bases between "broad" peak and repeat intervals. Is not called if the format argument is "narrow" or "summit".

Value

A list with three components:

<code>RepeatName</code>	Results of the binom.test as applied to repeat subfamilies, using FDR adjustment of p-values.
<code>RepeatFamily</code>	Results of the binom.test as applied to repeat families, using FDR adjustment of p-values.
<code>RepeatType</code>	Results of the binom.test as applied to repeat classes, using FDR adjustment of p-values.

[Package *regulaTER* version 0.1.0 [Index](#)]

Figure 15. The TEAR function documentation viewed in Rstudio.

Following the use of the TEAR function, the other functions of *regulaTER* can be called using its results as input. The first among these functions is DATE, which calculates the enrichment of accessible TEs, or TEs coinciding with other genomic peaks of choice, in the neighboring or promoter regions of user-defined gene regions. Like TEAR, DATE was originally optimized for ATAC-seq peaks and the neighboring regions of DEGs. However, the enrichment analysis algorithm DATE uses is flexible in the kind of input it will accept. In addition to the data and parameters used by TEAR, DATE also requires the genomic coordinates of the genes of interest (e.g. DEGs), the "regionSelection" parameters, which determines whether the enrichment will be calculated for neighboring regions of the genes, including the gene body itself, or just

the promoter regions, and finally the distance from the gene body or transcription start site the TEs will be considered associated with the genes. DATE is primarily useful for the prediction of TE-derived effects on transcriptional regulation. The underlying enrichment algorithm considers specific TE subfamilies more likely to be conserved near genes compared to intergenic regions if they contribute to the regulatory network for those genes. Unlike TEAR, the DATE function is optional, and does not require inclusion in TE-centric analyses outside of their role in gene regulatory networks. The specific parameters used for DATE are described in the regulaTER manual and function documentation (Figure 16).

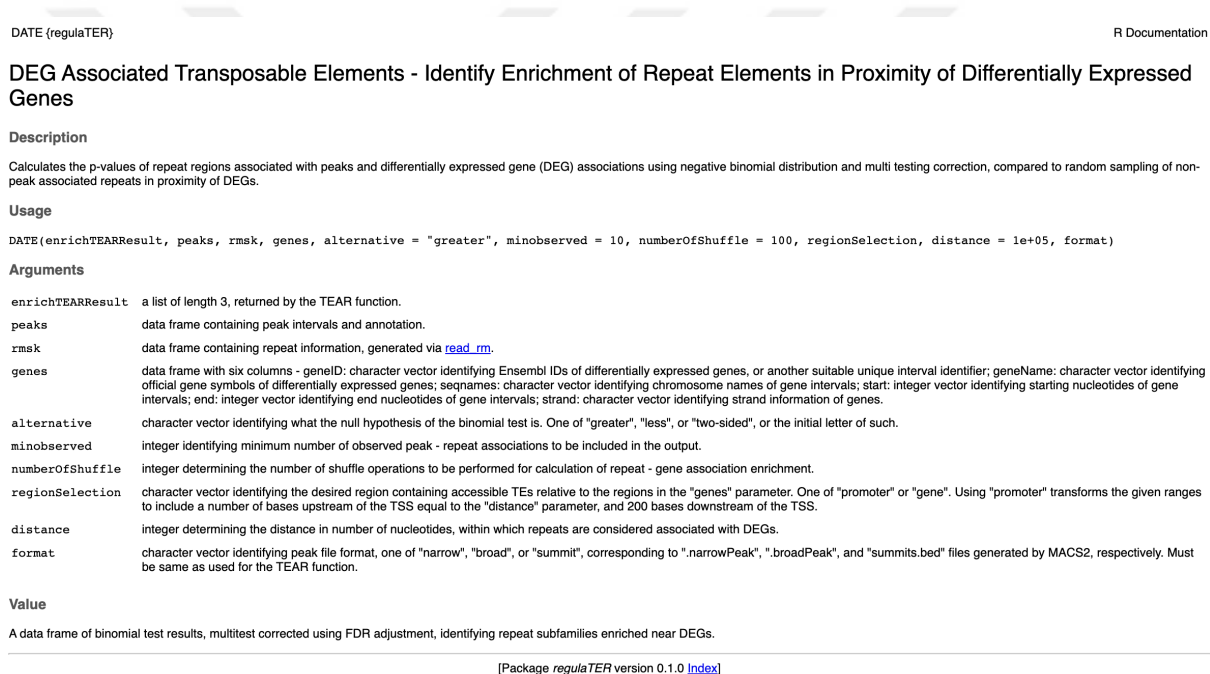


Figure 16. The DATE function documentation viewed in Rstudio.

The second of post-TEAR functions is FindMotifs, which is used to calculate the enrichment of genomic motifs (e.g. TF binding motifs) in the TE loci of interest. This function performs two primary tasks. The first of these tasks is to relay the genomic coordinates of enriched TE loci calculated in the R environment to the HOMER peak analysis and motif enrichment software, using marge, an R-based API for HOMER. The second task is to create a custom background region set to improve the results of the motif enrichment analyses compared to using a whole genome background. As

described under subheading 3.7.6, FindMotifs calculates the background regions differently based on whether the input TE enrichments are generated via TEAR or DATE. For TEAR output, FindMotifs considers whether the motifs located in the sequences of TEs found in genomic peaks are significantly different than the motifs in TE sequences of the same subfamily outside of those genomic peaks. This allows the results to accurately represent whether the identified motifs contribute to the relationship between the TE loci and the genomic peaks, e.g. accessible DNA for ATAC-seq peaks. On the other hand, as DATE considers the relationship between the peak-associated TE loci and genes of interest, when given DATE output, FindMotifs instead calculates peak-associated TEs outside of gene neighboring or promoter regions as background. As a result, the motifs discovered are more likely to be conserved in gene-neighboring regions, potentially due to selective pressure due to their gene regulatory roles, rather than merely being structural components of the TE itself.

Although the use of FindMotifs is described in depth in both the library manual and the function documentation, the users are encouraged to familiarize themselves with the use of both HOMER and marge as well, to better interpret the results (Figure 17).

Find Enriched Genomic Motifs in Repeats Enriched in Promoters using HOMER

Description

Uses HOMER to identify enriched genomic motifs, such as transcription factor binding sites, in provided list of repeats enriched in accessible DEG promoters, and places the results into user provided directory.

Usage

```
FindMotifs(df, repeatMaskerFile, peak, genes, distance, genome, outDir, homerPath, type, format, topRepeats)
```

Arguments

df	data frame of binomial test results, multitest corrected using FDR adjustment, identifying enriched repeat subfamilies, obtained using TEAR or DATE functions.
repeatMaskerFile	data frame containing repeat information, generated via read_rm .
peak	data frame containing peak intervals and annotation.
genes	data frame with six columns - geneID: character vector identifying Ensembl IDs of differentially expressed genes, or another suitable unique interval identifier; geneName: character vector identifying official gene symbols of differentially expressed genes; seqnames: character vector identifying chromosome names of gene intervals; start: integer vector identifying starting nucleotides of gene intervals; end: integer vector identifying end nucleotides of gene intervals; strand: character vector identifying strand information of genes. Should be the same object as used for DATE if type is "linkedRepeats".
distance	integer determining the distance in number of nucleotides, within which repeats are considered associated with DEGs. Should be the same distance as used for DATE if type is "linkedRepeats".
genome	character vector identifying the genome build used for HOMER operations. Usually an organism common name or a genome build name. For details, see the HOMER configuration manual.
outDir	character vector identifying target output directory of motif enrichment analysis results.
homerPath	character vector identifying full path to HOMER binary directory.
type	character vector identifying whether the df input argument is generated via TEAR or DATE. One of "enrichPeak" or "linkedRepeats", for the two functions, respectively. This argument has no default and must be supplied for the function to work.
format	character vector identifying peak file format, one of "narrow", "broad", or "summit", corresponding to ".narrowPeak", ".broadPeak", and "summits.bed" files generated by MACS2, respectively. Only called if "type" is "linkedRepeats". Must be same as used for the DATE function.
topRepeats	boolean value identifying whether only the top ten repeats (TRUE) or all repeats identified in the list (FALSE) should be included in the motif discovery analysis.

Details

A local installation of HOMER genomic sequencing peak analysis software is required for use of the FindMotifs function. HOMER software can be obtained via the instructions on the [HOMER website] (<http://homer.ucsd.edu/homer/>).

Heinz S, Benner C, Spann N, Bertolino E et al. Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. Mol Cell 2010 May 28;38(4):576-589. PMID: 20513432

Users should note that the motif discovery for a single repeat subfamily can take a considerable amount of time, and if "topRepeats = FALSE", it is recommended that users only provide a subset of their input for the "df" argument to analyze selected repeats of interest.

Value

None. Generated output is placed in the directory identified by the outDir parameter. For details of the HOMER software output, please see [HOMER website](#).

[Package regulaTER version 0.1.0 [Index](#)]

Figure 17. The FindMotifs function documentation viewed in Rstudio.

The MakePlots function is the last of the post-TEAR functions. This function is used to generate quick visual representations of the TEAR analyses, and creates two plots for each of the TE subfamily, family, and class level outputs of the function, for a total of six figures. The first of these figures is a scatter plot that shows the statistical significance and the fold change of each TE category (Figure 11A), while the second shows the number of TE loci in genomic peaks, as well as the ratio of peak-associated TEs to the genome-wide number of TEs in the same category (Figure 11B). In addition to the list object containing the TEAR results, the user can also define the directory where the generated plots should be stored, as well as the desired width and height, which will apply to each generated plot (Figure 18).

Create Exploratory Scatter and Bar Plots for Repeat Regions Enriched in Accessible Genomic Peaks

Description

Creates automated plots of significantly repeat regions enriched in accessible genomic peaks identified by the TEAR function, using ggplot2 functions.

Usage

```
MakePlots(enrichTEARResult, outDir, width = 4, height = 4)
```

Arguments

enrichTEARResult a list of length 3, with the names "RepeatName", "RepeatType", and "RepeatFamily", and data frames generated via [binom.test](#), as a result of the [TEAR](#) function.

outDir character vector identifying target output directory of generated plots.

width numeric identifying the width (x-axis length) of the resulting image files.

height numeric identifying the height (y-axis length) of the resulting image files.

Value

None. Generated output is placed in the directory identified by the outDir parameter.

[Package *regulaTER* version 0.1.0 [Index](#)]

Figure 18. The MakePlots function documentation viewed in Rstudio.

While its minimal input requirements makes MakePlots a convenient way to quickly generate exploratory visual representations of the TEAR results, it is not intended to create publication quality figures. Should it be necessary to improve the quality of these plots, the `obsOnTrueMean` and `p.adjust.value` columns of the TEAR results can be used for the scatter plot, while the `observed` and `rmsk` columns are required to replicate the line-and-bar plot.

5. Discussion

TEs hold a great variety of roles in the structural and functional regulation of the genome. Among these roles are the formation of transcriptional regulatory networks, which regulate the expression of multiple genes. TEs have several traits that make them uniquely suited to exapction for use in such regulatory networks. The first of these traits is their ability to create novel copies of themselves elsewhere in the genome. The second is the presence of many genomic motifs, such as promoter elements or TF binding sites, in TE sequences. Together, these traits allow TEs to wire a large number of genes located across the genome to the same regulatory network (Davidson and Britten, 1979; Feschotte, 2008; Bourque et al., 2018; Sun et al., 2018). However, the large copy number of TEs across the genome, their highly polymorphic nature, their various functions in genomic and transcriptomic regulation, and the fact that they are subject to their own strict regulation in most biological contexts pose challenges in uncovering the full impact of TEs in the regulation of gene expression. Furthermore, the repeatome of most organisms have a large number of TE families and subfamilies. Experimentally validating the connection between a phenotype of interest and a single TE subfamily requires considerable time and resources. Therefore, in order to shortlist TEs of interest for research on a biological process or phenotype, use of computational analysis methods using next generation sequencing methods plays a vital role. There are currently a number of computational tools and workflows used in such shortlisting when it comes to identifying the TEs active in gene regulation (Goerner-Potvin and Bourque, 2018). However, many of these tools only take into account the genomic coordinate of the TEs, or their transcript levels, and fail to consider the impact of epigenetic factors or chromatin state on these regulatory activities, or their transcription-independent roles. Therefore, the development of a single computational tool which allows for a range of TE-specialized analyses using integrated multiomics datasets is crucial to achieve a better understanding of TEs in gene regulation. It is with this in mind that regulaTER was developed.

In addition to the main repository, a supplementary repository containing the genomic region files of the hg19 and hg38 human genome assemblies, and the mm10 mouse genome assembly, was created and made available at the regulaTER-regions

repository on GitHub (<https://gER-regions>). Similar genomic annotation files for other genome assemblies can be generated and formatted following the directions provided in the regulaTER documentation.

In this study, ATAC-seq and TF ChIP-seq peaks were used alongside reference gene and repeat annotations for the mm10 mouse and hg38 human genome assemblies. However, the TEAR function accepts a wide range of inputs. For example, it is possible to use non-reference genomes and repeat annotations from any organism, which can include novel polymorphic insertions in individuals or somatic cells. Thus, regulaTER can identify the TE families or subfamilies whose transposition events drive disease formation and progression, such as LINE-1 in cancer (Morse et al., 1988; Rodic, 2018).

The motifs discovered in this study were all calculated using the DATE output due to their expected roles in transcriptional regulation. However, the use of FindMotifs with TEAR output allows the users to answer various TE-related questions. TEs are often under transcriptional repression in most cells. This repression is frequently a result of CpG methylation, or the presence of H3K9me3 (Barau et al., 2016; Walter et al., 2016; He et al., 2019; Chitrakar et al., 2022). Methyl transferase and chromatin remodeler proteins often partner with KRAB-ZFPs, which target specific TE subfamilies (Ecco et al., 2016; Yang et al., 2017). Use of FindMotifs on TEAR results for methylation (e.g. bisulfite-seq, MeDIP-seq) or repressive chromatin (e.g. H3K9me3 ChIP-seq) peaks can help identify which KRAB-ZFPs are targeted to the TE loci in enriched subfamilies.

To highlight the uses and advantages of regulaTER, it is compared to three available computational tools and databases with a high amount of overlap with regulaTER and can help analyze the regulatory impact of TEs. These tools and databases are GREAM (Chandrashekar et al., 2015), TEffectR (Karakulah et al., 2019), and PlanTEEnrichment (Karakulah and Suner, 2017). As every tool requires different types of input and uses different means of computational analysis, a qualitative comparison was used to showcase the benefits of regulaTER (Table 7).

Table 7. Comparison of regulaTER with similar tools.

Software	regulaTER	GREAM	TEffectR	PlanTEenrichment
Software type	R library	Database	R library	Database
Custom genome and repeat annotations	Yes	No	Yes	No
Analyzed species	Custom input	17 mammalian	Custom input	11 plant
Genomic context of TE instances (e.g., methylation, chromatin accessibility, etc.)	Yes	No	No	No
Gene-independent enrichment analysis	Yes	No	No	No
Gene-TE distance	Custom input	Preset options	Custom input	Preset options
Motif enrichment analysis	Compares relevant TE loci to custom background using HOMER	Uses consensus sequence of TE using JASPAR and CLOVER	No	No
Result statistics	TE only	TE or gene	TE and gene	TE only
Orthologous gene set analysis	No	Yes	No	No
Automated GO term enrichment analysis	No	Yes	No	No
TE expression data	No	No	Yes	No

A SWOT analysis highlighted the unique strengths of regulaTER absent in the other tools. The most relevant of these is the use of custom genome and repeat annotation as input for regulaTER, and the use of genomic sequencing peaks to incorporate the locus-level epigenomic states of TE for enrichment analysis. While originally optimized for genomic accessibility data through ATAC-seq, regulaTER can accept multiple types of genomic peak data, including methylation, protein binding, and histone markers (ChIP-seq or CUT&RUN). Use of such multiomics data can narrow down the TE loci and categories to ones directly associated with the phenotype of interest, which is not possible in other available tools.

Another advantage of regulaTER is the ability of the TEAR function to identify TE enrichment in genomic peaks independent of pre-determined gene sets. This obviates the need to perform transcriptome-level analyses and allows researchers to evaluate the regulation of TEs themselves. Finally, regulaTER incorporates a bespoke algorithm to calculate custom background regions for use in the HOMER software for more accurate motif enrichment analyses based on the sequences of TE loci in relevant regions as opposed to the subfamily consensus sequence. These advantages, which are not found in other TE-centric software, reveal that regulaTER fills a previously unmet need in computational TE analysis.

To test the use of regulaTER on biological data, an EMT and MET cell line was used as the biological model of choice. EMT and MET are complementary cellular reprogramming events, with related but distinct transcriptional regulation networks (Kalluri and Weinberg, 2009; Pei et al., 2019). These events are vital to various biological processes, such as development, wound healing, and tumorigenesis (Li et al., 2010; Cervantes-Arias et al., 2013; Kim et al., 2017; Roche, 2018). Previous research has revealed a wealth of information on the intricacies of EMT regulation and the roles of TEs in it (Rangasamy et al., 2015; Di Ruocco et al., 2018; Grillo et al., 2023). However, similar information on the regulation of MET is lacking. The use of an EMT-MET model to test regulaTER functions allows the validation of our findings on EMT data by comparing regulaTER results to previous findings, which in turns reinforces the reliability of the MET results.

In this study, binding sites for TFs with major roles in EMT regulation were identified to be enriched on various B element TEs, such as FoxA1/2 and TEAD2, in proximity to genes that act as markers of the mesenchymal phenotype following EMT, such as *Cdh2*, *Fn1*, *Itga5*, *Wnt2b*, *Zeb1/2*, and *Tgfb1* (Gonzalez and Medici, 2014; Huang et al., 2015). The B elements are a murine-specific TE family. Studies on B elements are relatively scarce, given their relatively low applicability outside of mouse models. However, the B elements are orthologues to the primate-specific Alu family, which are SINEs who show a high degree of cross-species conservation with the B elements in terms of genomic coordinates and function (Tsirigos and Rigoutsos, 2009; Li and Shen, 2023). It is established that overexpression of Alu TEs can trigger EMT (Di Ruocco et al., 2018; Hwang et al., 2019), and they are rich in TF binding sites and CpG nucleotides (Polak and Domany, 2006; Luo et al., 2014). These make them high probability candidates to act during gene and chromatin regulation. However, previous studies have only examined the transcript levels of TEs when considering their function in EMT. The findings of this study reveals the need for further investigation into the EMT regulatory impact of non-transcribed TE loci, especially B elements and Alus.

Multiple studies implicated the role of TEs in EMT and by extension metastasis. Reyes-Ryes et al. explained the enhanced malignancy of human bronchial epithelial cells through interaction and co-regulation of TEs with TGF β 1 (Reyes-Reyes et al., 2017). In a separate study, stemness and cellular plasticity in colorectal and breast cancer tumorigenesis EMT was linked to regulation by the LINE1-ORF1 (Apostolou et al., 2015). Additionally, genomic insertions of AluSx1, LTR33, L2b, and MLT1F1 TEs were shown to cause colorectal cancer growth and metastasis progression via TE-driven transpochimeric transcript activation, such as in POU5F1B (Simó-Riudalbas et al., 2022).

The establishment of the regulaTER pipeline, and its validation using the EMT data in comparison to previous experimental findings, makes it possible to apply the same functions to MET data. As previously mentioned, the information on the regulation of MET is lacking compared to that of EMT. Therefore, the results of MET data analyses can provide clues for future research targets. Such shortlisting can focus on both the TEs identified and their activities in transcriptomic regulation, and the TFs

whose binding sites are enriched in the DEG-associated TEs, as identified by DATE. This information will provide a new avenue of approach to understanding the complexity of MET regulation

In this study, the accessible loci of several B elements were found to be enriched near MET upregulated genes. These accessible loci were enriched for FoxA2 binding sites. FoxA2 has previously been linked to the inhibition of EMT, which is consistent with the GSEA results of the MET transcriptome in Figure 10B, indicating the inhibition of EMT hallmark genes. Targets of FoxA2 are associated with EMT and MET regulation as well. Rab17, which is one of the genes targeted by FoxA2, is responsible for maintaining the epithelial morphology through inhibition of the Stat3/HIF1a/VEGF signaling pathway (Wang et al., 2020). Angiotensin, another target of FoxA2, is known to inhibit Tgfb1-induced EMT in keratinocytes (Jihu et al., 2024).

The use of regulaTER on multiple datasets of EMT and MET revealed a wealth of information regarding the roles of TEs and TE-contributed TFBSs, especially the B and Alu TEs and the Forkhead binding sites harbored by these TEs. The experimental validation of other TEs and TFBSs identified by regulaTER will further our understanding of the regulation of EMT and MET. Additionally, the use of independently produced, publicly available datasets in the study further highlights the general utility of regulaTER to TE analyses in a wider range of biological contexts. Its ease of use, ability to incorporate a wide range of multiomics data, and broad and flexible applicability, make regulaTER an important tool for future TE research, with great potential to save both time and resources by quickly predicting the roles of TEs in regulatory networks.

6. Conclusion

1. An R library, regulaTER, was established to analyze the effects of accessible TEs in gene regulation. The use of regulaTER was documented, and made available for public use.
2. The regulaTER library was tested using multiomics data (ATAC-seq and RNA-seq) obtained during the EMT and MET biological processes.
3. MIR TEs show the highest enrichment in accessible DNA during EMT and MET, and accessible MIR loci are frequently co-located with epithelial marker genes.
4. 5 B element TE subfamilies (B1_Mus1, B3, B3A, B4A, and ID_B1) were enriched in both EMT and MET following the use of the DATE function to establish their association with DEGs. B elements are the murine equivalent of the primate-specific Alu TEs, which are known to be active in EMT induction, and are rich in TF binding sites and CpG dinucleotides. Taken together, this strongly suggests the role of B elements in EMT and MET regulation.
5. Forkhead TF binding site signatures are enriched in DEG-associated accessible TEs during both EMT and MET. Multiple Forkhead TFs, such as FoxA1 and FoxA2, are known to regulate EMT genes.
6. regulaTER uses were validated in biological models from another organism (Homo sapiens) as well as non-ATAC-seq genomic peaks (Foxa2 ChIP-seq) using independently generated datasets. The results further confirmed the potential roles of MIR and B element / Alu TEs in the establishment of accessible chromatin and gene regulation, primarily through the contribution of Forkhead transcription factors, such as FoxA1 and FoxA2.
7. The functions of regulaTER allow for flexible inputs, and can be used to analyze various types of multiomics datasets outside of the scope of this study to answer TE-centric research questions.
8. In future studies, the functional roles of TE targets and their TF partners must be confirmed using experimental methods, such as via CRISPR-based silencing. Furthermore, use of regulaTER with non-reference genome sequences and repeat annotations, as well as with single-cell sequencing data, needs to be validated.

7. References

- Adey A, Morrison HG, Asan, Xun X, Kitzman JO, Turner EH, et al. Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. *Genome Biol* 2010;11:R119. <https://doi.org/10.1186/gb-2010-11-12-r119>.
- Ammar I, Izsvák Z, Ivics Z. The Sleeping Beauty transposon toolbox. *Methods Mol Biol* 2012;859:229–40. https://doi.org/10.1007/978-1-61779-603-6_13.
- Antón-García P, Haghighi EB, Rose K, Vladimirov G, Boerries M, Hecht A. TGFβ1-Induced EMT in the MCF10A Mammary Epithelial Cell Line Model Is Executed Independently of SNAIL1 and ZEB1 but Relies on JUNB-Coordinated Transcriptional Regulation. *Cancers* 2023;15:558. <https://doi.org/10.3390/cancers15020558>.
- Apostolou P, Toloudi M, Chatziioannou M, Kourtidou E, Mimikakou G, Vlachou I, et al. Involvement of retrotransposon L1 in stemness and cellular plasticity. *Cell Communication & Adhesion* 2015;22:1–7. <https://doi.org/10.3109/15419061.2014.970270>.
- Aravin AA, Hannon GJ, Brennecke J. The Piwi-piRNA Pathway Provides an Adaptive Defense in the Transposon Arms Race. *Science* 2007;318:761–4. <https://doi.org/10.1126/science.1146484>.
- Ardeljan D, Taylor MS, Ting DT, Burns KH. The Human Long Interspersed Element-1 Retrotransposon: An Emerging Biomarker of Neoplasia. *Clin Chem* 2017;63:816–22. <https://doi.org/10.1373/clinchem.2016.257444>.
- Baillie JK, Barnett MW, Upton KR, Gerhardt DJ, Richmond TA, De Sapio F, et al. Somatic retrotransposition alters the genetic landscape of the human brain. *Nature* 2011;479:534–7. <https://doi.org/10.1038/nature10531>.
- Barau J, Teissandier A, Zamudio N, Roy S, Nalesso V, Hérault Y, et al. The DNA methyltransferase DNMT3C protects male germ cells from transposon activity. *Science* 2016;354:909–12. <https://doi.org/10.1126/science.aah5143>.
- Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 2014;30:2114–20. <https://doi.org/10.1093/bioinformatics/btu170>.

Bourque G, Burns KH, Gehring M, Gorbunova V, Seluanov A, Hammell M, et al. Ten things you should know about transposable elements. *Genome Biology* 2018;19:199. <https://doi.org/10.1186/s13059-018-1577-z>.

Bourque G, Leong B, Vega VB, Chen X, Lee YL, Srinivasan KG, et al. Evolution of the mammalian transcription factor binding repertoire via transposable elements. *Genome Res* 2008;18:1752–62. <https://doi.org/10.1101/gr.080663.108>.

Bozek M, Cortini R, Storti AE, Unnerstall U, Gaul U, Gompel N. ATAC-seq reveals regional differences in enhancer accessibility during the establishment of spatial coordinates in the *Drosophila* blastoderm. *Genome Res* 2019;29:771–83. <https://doi.org/10.1101/gr.242362.118>.

Brennecke J, Aravin AA, Stark A, Dus M, Kellis M, Sachidanandam R, et al. Discrete small RNA-generating loci as master regulators of transposon activity in *Drosophila*. *Cell* 2007;128:1089–103. <https://doi.org/10.1016/j.cell.2007.01.043>.

Buenrostro J, Wu B, Chang H, Greenleaf W. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. *Curr Protoc Mol Biol* 2015;109:21.29.1-21.29.9. <https://doi.org/10.1002/0471142727.mb2129s109>.

Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. Transposition of native chromatin for multimodal regulatory analysis and personal epigenomics. *Nat Methods* 2013;10:1213–8. <https://doi.org/10.1038/nmeth.2688>.

Cervantes-Arias A, Pang LY, Argyle DJ. Epithelial-mesenchymal transition as a fundamental mechanism underlying the cancer phenotype. *Vet Comp Oncol* 2013;11:169–84. <https://doi.org/10.1111/j.1476-5829.2011.00313.x>.

Chandrashekar DS, Dey P, Acharya KK. 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* 2015;10. <https://doi.org/10.1371/journal.pone.0133647>.

Chang H, Pan Y, Landrette S, Ding S, Yang D, Liu L, et al. Efficient genome-wide first-generation phenotypic screening system in mice using the piggyBac transposon. *Proc Natl Acad Sci U S A* 2019;116:18507–16. <https://doi.org/10.1073/pnas.1906354116>.

Chitrakar A, Noon M, Xiao AZ. Taming the transposon: H3K9me3 turns foe to friend in human development. *Cell Stem Cell* 2022;29:1009–10. <https://doi.org/10.1016/j.stem.2022.06.010>.

Chu C, Zhao B, Park PJ, Lee EA. Identification and Genotyping of Transposable Element Insertions From Genome Sequencing Data. *Curr Protoc Hum Genet* 2020;107:e102. <https://doi.org/10.1002/cphg.102>.

Chuong EB, Elde NC, Feschotte C. Regulatory activities of transposable elements: from conflicts to benefits. *Nat Rev Genet* 2017;18:71–86. <https://doi.org/10.1038/nrg.2016.139>.

Chuong EB, Rumi MAK, Soares MJ, Baker JC. Endogenous retroviruses function as species-specific enhancer elements in the placenta. *Nat Genet* 2013;45:325–9. <https://doi.org/10.1038/ng.2553>.

Claverie-Martín F, Flores C, Antón-Gamero M, González-Acosta H, García-Nieto V. The Alu insertion in the CLCN5 gene of a patient with Dent's disease leads to exon 11 skipping. *J Hum Genet* 2005;50:370–4. <https://doi.org/10.1007/s10038-005-0265-5>.

Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. *Gigascience* 2021;10. <https://doi.org/10.1093/gigascience/giab008>.

Daugherty AC, Yeo RW, Buenrostro JD, Greenleaf WJ, Kundaje A, Brunet A. Chromatin accessibility dynamics reveal novel functional enhancers in *C. elegans*. *Genome Res* 2017;27:2096–107. <https://doi.org/10.1101/gr.226233.117>.

Davidson EH, Britten RJ. Regulation of gene expression: possible role of repetitive sequences. *Science* 1979;204:1052–9. <https://doi.org/10.1126/science.451548>.

Di Ruocco F, Basso V, Rivoire M, Mehlen P, Ambati J, De Falco S, et al. Alu RNA accumulation induces epithelial-to-mesenchymal transition by modulating miR-566 and is associated with cancer progression. *Oncogene* 2018;37:627–37. <https://doi.org/10.1038/onc.2017.369>.

Ecco G, Cassano M, Kauzlaric A, Duc J, Coluccio A, Offner S, et al. Transposable Elements and Their KRAB-ZFP Controllers Regulate Gene Expression in Adult Tissues. *Developmental Cell* 2016;36:611–23. <https://doi.org/10.1016/j.devcel.2016.02.024>.

Feschotte C. Transposable elements and the evolution of regulatory networks. *Nat Rev Genet* 2008;9:397–405. <https://doi.org/10.1038/nrg2337>.

Flemr M, Malik R, Franke V, Nejepinska J, Sedlacek R, Vlahovicek K, et al. A retrotransposon-driven dicer isoform directs endogenous small interfering RNA production in mouse oocytes. *Cell* 2013;155:807–16. <https://doi.org/10.1016/j.cell.2013.10.001>.

Fultz D, Choudury SG, Slotkin RK. Silencing of active transposable elements in plants. *Current Opinion in Plant Biology* 2015;27:67–76. <https://doi.org/10.1016/j.pbi.2015.05.027>.

Garcia-Perez JL, Marchetto MCN, Muotri AR, Coufal NG, Gage FH, O'Shea KS, et al. LINE-1 retrotransposition in human embryonic stem cells. *Hum Mol Genet* 2007;16:1569–77. <https://doi.org/10.1093/hmg/ddm105>.

Goerner-Potvin P, Bourque G. Computational tools to unmask transposable elements. *Nat Rev Genet* 2018;19:688–704. <https://doi.org/10.1038/s41576-018-0050-x>.

Gonzalez DM, Medici D. Signaling mechanisms of the epithelial-mesenchymal transition. *Sci Signal* 2014;7. <https://doi.org/10.1126/scisignal.2005189>.

Goodier JL. Restricting retrotransposons: a review. *Mobile DNA* 2016;7:16. <https://doi.org/10.1186/s13100-016-0070-z>.

Goryshin IY, Reznikoff WS. Tn5 in vitro transposition. *J Biol Chem* 1998;273:7367–74. <https://doi.org/10.1074/jbc.273.13.7367>.

Goubert C. Assembly-Free Detection and Quantification of Transposable Elements with dnaPipeTE. *Methods Mol Biol* 2023;2607:25–43. https://doi.org/10.1007/978-1-0716-2883-6_2.

Grabundzija I, Messing SA, Thomas J, Cosby RL, Bilic I, Miskey C, et al. A Helitron transposon reconstructed from bats reveals a novel mechanism of genome shuffling in eukaryotes. *Nat Commun* 2016;7:10716. <https://doi.org/10.1038/ncomms10716>.

Grillo G, Keshavarzian T, Linder S, Arlidge C, Mout L, Nand A, et al. Transposable Elements Are Co-opted as Oncogenic Regulatory Elements by Lineage-Specific Transcription Factors in Prostate Cancer. *Cancer Discovery* 2023;13:2470–87. <https://doi.org/10.1158/2159-8290.CD-23-0331>.

Guibert S, Forné T, Weber M. Global profiling of DNA methylation erasure in mouse primordial germ cells. *Genome Res* 2012;22:633–41. <https://doi.org/10.1101/gr.130997.111>.

Haig D. Transposable elements: Self-seekers of the germline, team-players of the soma. *Bioessays* 2016;38:1158–66. <https://doi.org/10.1002/bies.201600125>.

Hancks DC, Kazazian HH. Roles for retrotransposon insertions in human disease. *Mob DNA* 2016;7. <https://doi.org/10.1186/s13100-016-0065-9>.

Harshil Patel, Espinosa-Carrasco J, Langer B, Ewels P, Nf-Core Bot, Garcia MU, et al. nf-core/atacseq: [2.1.2] - 2022-08-07 2023. <https://doi.org/10.5281/ZENODO.2634132>.

He J, Fu X, Zhang M, He F, Li W, Abdul MMd, et al. Transposable elements are regulated by context-specific patterns of chromatin marks in mouse embryonic stem cells. *Nat Commun* 2019;10:34. <https://doi.org/10.1038/s41467-018-08006-y>.

Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, et al. Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Molecular Cell* 2010;38:576–89. <https://doi.org/10.1016/j.molcel.2010.05.004>.

Horisawa K, Udono M, Ueno K, Ohkawa Y, Nagasaki M, Sekiya S, et al. The Dynamics of Transcriptional Activation by Hepatic Reprogramming Factors. *Molecular Cell* 2020;79:660-676.e8. <https://doi.org/10.1016/j.molcel.2020.07.012>.

Huang CRL, Burns KH, Boeke JD. Active Transposition in Genomes. *Annu Rev Genet* 2012;46:651–75. <https://doi.org/10.1146/annurev-genet-110711-155616>.

Huang RY-J, Kuay KT, Tan TZ, Asad M, Tang HM, Ng AHC, et al. Functional relevance of a six mesenchymal gene signature in epithelial-mesenchymal transition (EMT) reversal by the triple angiokinase inhibitor, nintedanib (BIBF1120). *Oncotarget* 2015;6:22098–113. <https://doi.org/10.18632/oncotarget.4300>.

Hunt KV, Burnard SM, Roper EA, Bond DR, Dun MD, Verrills NM, et al. scTEM-seq: Single-cell analysis of transposable element methylation to link global epigenetic heterogeneity with transcriptional programs. *Sci Rep* 2022;12:5776. <https://doi.org/10.1038/s41598-022-09765-x>.

Hwang YE, Baek YM, Baek A, Kim D-E. Oxidative stress causes Alu RNA accumulation via PIWIL4 sequestration into stress granules. *BMB Rep* 2019;52:196–201. <https://doi.org/10.5483/BMBRep.2019.52.3.146>.

Iwagami S, Baba Y, Watanabe M, Shigaki H, Miyake K, Ishimoto T, et al. LINE-1 hypomethylation is associated with a poor prognosis among patients with curatively resected esophageal squamous cell carcinoma. *Ann Surg* 2013;257:449–55. <https://doi.org/10.1097/SLA.0b013e31826d8602>.

Jeong H-H, Yalamanchili HK, Guo C, Shulman JM, Liu Z. An ultra-fast and scalable quantification pipeline for transposable elements from next generation sequencing data. *Pac Symp Biocomput* 2018;23:168–79.

Jihu Y, Leng R, Liu M, Ren H, Xie D, Yao C, et al. Angiotensin (1-7) Inhibits Transforming Growth Factor- β 1-Induced Epithelial-Mesenchymal Transition of Human Keratinocyte Hacat Cells in vitro. *CCID* 2024;Volume 17:1049–58. <https://doi.org/10.2147/CCID.S441596>.

Jönsson ME, Ludvik Brattås P, Gustafsson C, Petri R, Yudovich D, Piracs K, et al. Activation of neuronal genes via LINE-1 elements upon global DNA demethylation in human neural progenitors. *Nat Commun* 2019;10. <https://doi.org/10.1038/s41467-019-11150-8>.

Jursch T, Miskey C, Izsvák Z, Ivics Z. Regulation of DNA transposition by CpG methylation and chromatin structure in human cells. *Mob DNA* 2013;4:15. <https://doi.org/10.1186/1759-8753-4-15>.

Kalluri R, Weinberg RA. The basics of epithelial-mesenchymal transition. *J Clin Invest* 2009;119:1420–8. <https://doi.org/10.1172/JCI39104>.

Karakulah G, Arslan N, Yandim C, Suner A. TEffectR: an R package for studying the potential effects of transposable elements on gene expression with linear regression model. *PeerJ* 2019;7:e8192. <https://doi.org/10.7717/peerj.8192>.

Karakulah G, Suner A. PlanTEEnrichment: A tool for enrichment analysis of transposable elements in plants. *Genomics* 2017;109:336–40. <https://doi.org/10.1016/j.ygeno.2017.05.008>.

Kawano H, Saeki H, Kitao H, Tsuda Y, Otsu H, Ando K, et al. Chromosomal instability associated with global DNA hypomethylation is associated with the initiation and

progression of esophageal squamous cell carcinoma. *Ann Surg Oncol* 2014;21 Suppl 4:S696-702. <https://doi.org/10.1245/s10434-014-3818-z>.

Kazazian HH. Mobile DNA transposition in somatic cells. *BMC Biology* 2011;9:62. <https://doi.org/10.1186/1741-7007-9-62>.

Kazazian HH. Mobile elements and disease. *Current Opinion in Genetics & Development* 1998;8:343–50. [https://doi.org/10.1016/S0959-437X\(98\)80092-0](https://doi.org/10.1016/S0959-437X(98)80092-0).

Kazazian HH, Wong C, Youssoufian H, Scott AF, Phillips DG, Antonarakis SE. Haemophilia A resulting from de novo insertion of L 1 sequences represents a novel mechanism for mutation in man. *Nature* 1988;332:164–6. <https://doi.org/10.1038/332164a0>.

Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. *Nature Methods* 2015;12:357–60. <https://doi.org/10.1038/nmeth.3317>.

Kim D, Xing T, Yang Z, Dudek R, Lu Q, Chen Y-H. Epithelial Mesenchymal Transition in Embryonic Development, Tissue Repair and Cancer: A Comprehensive Overview. *JCM* 2017;7:1. <https://doi.org/10.3390/jcm7010001>.

Ko E-J, Oh YL, Kim HY, Eo WK, Kim H, Ock MS, et al. Correlation of long interspersed element-1 open reading frame 1 and c-Met proto-oncogene protein expression in ovarian cancer. *Genes Genomics* 2019;41:1293–9. <https://doi.org/10.1007/s13258-019-00858-y>.

van Kruijsbergen I, Hontelez S, Elurbe DM, van Heeringen SJ, Huynen MA, Veenstra GJC. Heterochromatic histone modifications at transposons in *Xenopus tropicalis* embryos. *Dev Biol* 2017;426:460–71. <https://doi.org/10.1016/j.ydbio.2016.08.031>.

Lanciano S, Cristofari G. Measuring and interpreting transposable element expression. *Nat Rev Genet* 2020;21:721–36. <https://doi.org/10.1038/s41576-020-0251-y>.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. *Nature* 2001;409:860–921. <https://doi.org/10.1038/35057062>.

Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 2012;9:357–9. <https://doi.org/10.1038/nmeth.1923>.

Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 2009;10:R25. <https://doi.org/10.1186/gb-2009-10-3-r25>.

Lavasanifar A, Sharp CN, Korte EA, Yin T, Hosseinnjad K, Jortani SA. Long interspersed nuclear element-1 mobilization as a target in cancer diagnostics, prognostics and therapeutics. *Clin Chim Acta* 2019;493:52–62. <https://doi.org/10.1016/j.cca.2019.02.015>.

Lee HJ, Hou Y, Maeng JH, Shah NM, Chen Y, Lawson HA, et al. Epigenomic analysis reveals prevalent contribution of transposable elements to cis-regulatory elements, tissue-specific expression, and alternative promoters in zebrafish. *Genome Res* 2022;32:1424–36. <https://doi.org/10.1101/gr.276052.121>.

Levy A, Sela N, Ast G. TranspoGene and microTranspoGene: transposed elements influence on the transcriptome of seven vertebrates and invertebrates. *Nucleic Acids Res* 2008;36:D47-52. <https://doi.org/10.1093/nar/gkm949>.

Lexa M, Cechova M, Nguyen SH, Jedlicka P, Tokan V, Kubat Z, et al. HiC-TE: a computational pipeline for Hi-C data analysis to study the role of repeat family interactions in the genome 3D organization. *Bioinformatics* 2022;38:4030–2. <https://doi.org/10.1093/bioinformatics/btac442>.

Li R, Liang J, Ni S, Zhou T, Qing X, Li H, et al. A mesenchymal-to-epithelial transition initiates and is required for the nuclear reprogramming of mouse fibroblasts. *Cell Stem Cell* 2010;7:51–63. <https://doi.org/10.1016/j.stem.2010.04.014>.

Li S, Shen X. Long interspersed nuclear element 1 and B1/Alu repeats blueprint genome compartmentalization. *Curr Opin Genet Dev* 2023;80:102049. <https://doi.org/10.1016/j.gde.2023.102049>.

Liao Y, Smyth GK, Shi W. The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. *Nucleic Acids Res* 2019;47:e47. <https://doi.org/10.1093/nar/gkz114>.

Luo Y, Lu X, Xie H. Dynamic Alu methylation during normal development, aging, and tumorigenesis. *Biomed Res Int* 2014;2014:784706. <https://doi.org/10.1155/2014/784706>.

Mahillon J, Chandler M. Insertion sequences. *Microbiol Mol Biol Rev* 1998;62:725–74. <https://doi.org/10.1128/MMBR.62.3.725-774.1998>.

Mcclintock B. Controlling elements and the gene. *Cold Spring Harb Symp Quant Biol* 1956;21:197–216. <https://doi.org/10.1101/sqb.1956.021.01.017>.

Meyer CA, Liu XS. Identifying and Mitigating Bias in Next-Generation Sequencing Methods for Chromatin Biology. *Nat Rev Genet* 2014;15:709–21. <https://doi.org/10.1038/nrg3788>.

Miao B, Fu S, Lyu C, Gontarz P, Wang T, Zhang B. Tissue-specific usage of transposable element-derived promoters in mouse development. *Genome Biology* 2020;21:255. <https://doi.org/10.1186/s13059-020-02164-3>.

Molaro A, Malik HS. Hide and seek: how chromatin-based pathways silence retroelements in the mammalian germline. *Curr Opin Genet Dev* 2016;37:51–8. <https://doi.org/10.1016/j.gde.2015.12.001>.

Morse B, Rotherg PG, South VJ, Spandorfer JM, Astrin SM. Insertional mutagenesis of the myc locus by a LINE-1 sequence in a human breast carcinoma. *Nature* 1988;333:87–90. <https://doi.org/10.1038/333087a0>.

Naville M, Warren IA, Haftek-Terreau Z, Chalopin D, Brunet F, Levin P, et al. Not so bad after all: retroviruses and long terminal repeat retrotransposons as a source of new genes in vertebrates. *Clin Microbiol Infect* 2016;22:312–23. <https://doi.org/10.1016/j.cmi.2016.02.001>.

O'Neill K, Brocks D, Hammell MG. Mobile genomics: tools and techniques for tackling transposons. *Philos Trans R Soc Lond B Biol Sci* 2020;375:20190345. <https://doi.org/10.1098/rstb.2019.0345>.

Pardue M-L, DeBaryshe PG. Retrotransposons that maintain chromosome ends. *Proc Natl Acad Sci USA* 2011;108:20317–24. <https://doi.org/10.1073/pnas.1100278108>.

Pei D, Shu X, Gassama-Diagne A, Thiery JP. Mesenchymal-epithelial transition in development and reprogramming. *Nat Cell Biol* 2019;21:44–53. <https://doi.org/10.1038/s41556-018-0195-z>.

Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nature Protocols* 2016;11:1650–67. <https://doi.org/10.1038/nprot.2016.095>.

Polak P, Domany E. Alu elements contain many binding sites for transcription factors and may play a role in regulation of developmental processes. *BMC Genomics* 2006;7:133. <https://doi.org/10.1186/1471-2164-7-133>.

Rangasamy D, Lenka N, Ohms S, Dahlstrom JE, Blackburn AC, Board PG. Activation of LINE-1 Retrotransposon Increases the Risk of Epithelial-Mesenchymal Transition and Metastasis in Epithelial Cancer. *Curr Mol Med* 2015;15:588–97. <https://doi.org/10.2174/1566524015666150831130827>.

Ren W, Fan H, Grimm SA, Kim JJ, Li L, Guo Y, et al. DNMT1 reads heterochromatic H4K20me3 to reinforce LINE-1 DNA methylation. *Nat Commun* 2021;12:2490. <https://doi.org/10.1038/s41467-021-22665-4>.

Reyes-Reyes EM, Aispuro I, Tavera-Garcia MA, Field M, Moore S, Ramos I, et al. LINE-1 couples EMT programming with acquisition of oncogenic phenotypes in human bronchial epithelial cells. *Oncotarget* 2017;8:103828–42. <https://doi.org/10.18632/oncotarget.21953>.

Robillard É, Le Rouzic A, Zhang Z, Capy P, Hua-Van A. Experimental evolution reveals hyperparasitic interactions among transposable elements. *Proc Natl Acad Sci U S A* 2016;113:14763–8. <https://doi.org/10.1073/pnas.1524143113>.

Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 2010;26:139–40. <https://doi.org/10.1093/bioinformatics/btp616>.

Roche J. The Epithelial-to-Mesenchymal Transition in Cancer. *Cancers (Basel)* 2018;10. <https://doi.org/10.3390/cancers10020052>.

Rodic N. LINE-1 activity and regulation in cancer. *Front Biosci (Landmark Ed)* 2018;23:1680–6. <https://doi.org/10.2741/4666>.

Rodić N, Sharma Reema, Sharma Rajni, Zampella J, Dai L, Taylor MS, et al. Long interspersed element-1 protein expression is a hallmark of many human cancers. *Am J Pathol* 2014;184:1280–6. <https://doi.org/10.1016/j.ajpath.2014.01.007>.

Rodríguez-Quiroz R, Valdebenito-Maturana B. SoloTE for improved analysis of transposable elements in single-cell RNA-Seq data using locus-specific expression. *Commun Biol* 2022;5:1063. <https://doi.org/10.1038/s42003-022-04020-5>.

Savytska N, Heutink P, Bansal V. Transcription start site signal profiling improves transposable element RNA expression analysis at locus-level. *Front Genet* 2022;13:1026847. <https://doi.org/10.3389/fgene.2022.1026847>.

Scott EC, Gardner EJ, Masood A, Chuang NT, Vertino PM, Devine SE. A hot L1 retrotransposon evades somatic repression and initiates human colorectal cancer. *Genome Res* 2016;26:745–55. <https://doi.org/10.1101/gr.201814.115>.

Shigaki H, Baba Y, Watanabe M, Murata A, Iwagami S, Miyake K, et al. LINE-1 hypomethylation in gastric cancer, detected by bisulfite pyrosequencing, is associated with poor prognosis. *Gastric Cancer* 2013;16:480–7. <https://doi.org/10.1007/s10120-012-0209-7>.

Simó-Riudalbas L, Offner S, Planet E, Duc J, Abrami L, Dind S, et al. Transposon-activated POU5F1B promotes colorectal cancer growth and metastasis. *Nat Commun* 2022;13:4913. <https://doi.org/10.1038/s41467-022-32649-7>.

Stockwell PA, Lynch-Sutherland CF, Chatterjee A, Macaulay EC, Eccles MR. RepExpress: A Novel Pipeline for the Quantification and Characterization of Transposable Element Expression from RNA-seq Data. *Curr Protoc* 2021;1:e206. <https://doi.org/10.1002/cpz1.206>.

Sultana T, Zamborlini A, Cristofari G, Lesage P. Integration site selection by retroviruses and transposable elements in eukaryotes. *Nat Rev Genet* 2017;18:292–308. <https://doi.org/10.1038/nrg.2017.7>.

Sun X, Wang X, Tang Z, Grivainis M, Kahler D, Yun C, et al. Transcription factor profiling reveals molecular choreography and key regulators of human retrotransposon expression. *Proc Natl Acad Sci USA* 2018;115. <https://doi.org/10.1073/pnas.1722565115>.

Sun Y, Miao N, Sun T. Detect accessible chromatin using ATAC-sequencing, from principle to applications. *Hereditas* 2019;156:29. <https://doi.org/10.1186/s41065-019-0105-9>.

Talluri TR, Kumar D, Glage S, Garrels W, Ivics Z, Debowski K, et al. Non-viral reprogramming of fibroblasts into induced pluripotent stem cells by Sleeping Beauty and piggyBac transposons. *Biochem Biophys Res Commun* 2014;450:581–7. <https://doi.org/10.1016/j.bbrc.2014.06.014>.

Treangen TJ, Salzberg SL. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat Rev Genet* 2012;13:36–46. <https://doi.org/10.1038/nrg3117>.

Tsirigos A, Rigoutsos I. Alu and B1 Repeats Have Been Selectively Retained in the Upstream and Intronic Regions of Genes of Specific Functional Classes. *PLoS Comput Biol* 2009;5:e1000610. <https://doi.org/10.1371/journal.pcbi.1000610>.

Venner S, Feschotte C, Biémont C. Dynamics of transposable elements: towards a community ecology of the genome. *Trends Genet* 2009;25:317–23. <https://doi.org/10.1016/j.tig.2009.05.003>.

Vincent JJ, Huang Y, Chen P-Y, Feng S, Calvopiña JH, Nee K, et al. Stage-Specific Roles for Tet1 and Tet2 in DNA Demethylation in Primordial Germ Cells. *Cell Stem Cell* 2013;12:470–8. <https://doi.org/10.1016/j.stem.2013.01.016>.

Walter M, Teissandier A, Pérez-Palacios R, Bourc'his D. An epigenetic switch ensures transposon repression upon dynamic loss of DNA methylation in embryonic stem cells. *eLife* 2016;5:e11418. <https://doi.org/10.7554/eLife.11418>.

Wang M, Wang W, Ding J, Wang J, Zhang J. Downregulation of Rab17 promotes cell proliferation and invasion in non-small cell lung cancer through STAT3/HIF-1 α /VEGF signaling. *Thoracic Cancer* 2020;11:379–88. <https://doi.org/10.1111/1759-7714.13278>.

Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, et al. A unified classification system for eukaryotic transposable elements. *Nat Rev Genet* 2007;8:973–82. <https://doi.org/10.1038/nrg2165>.

Yang P, Wang Y, Macfarlan TS. The Role of KRAB-ZFPs in Transposable Element Repression and Mammalian Evolution. *Trends Genet* 2017;33:871–81. <https://doi.org/10.1016/j.tig.2017.08.006>.

Yang WR, Ardeljan D, Pacyna CN, Payer LM, Burns KH. SQuIRE reveals locus-specific regulation of interspersed repeat expression. *Nucleic Acids Res* 2019;47:e27. <https://doi.org/10.1093/nar/gky1301>.

Yoder JA, Walsh CP, Bestor TH. Cytosine methylation and the ecology of intragenomic parasites. *Trends Genet* 1997;13:335–40. [https://doi.org/10.1016/s0168-9525\(97\)01181-5](https://doi.org/10.1016/s0168-9525(97)01181-5).

Yu G, Wang L-G, Han Y, He Q-Y. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS 2012;16:284–7. <https://doi.org/10.1089/omi.2011.0118>.

Zhang X, Zhang R, Yu J. New Understanding of the Relevant Role of LINE-1 Retrotransposition in Human Disease and Immune Modulation. Front Cell Dev Biol 2020;8:657. <https://doi.org/10.3389/fcell.2020.00657>.



8. Addenda

8.1. Ethics committee approval



İZMİR BİYOTIP VE GENOM MERKEZİ
GİRİŞİMSSEL OLMAYAN ARAŞTIRMALAR ETİK KURULU (İBG-GOEK)
KARARI

Toplantı Tarihi : 30/03/2023

Toplantı Günü : Perşembe

Toplantı Sayısı : 4

Toplantı Saati : 14:00

Sayın Doç. Dr. Gökhan KARAKÜLAH,

2022-032 Protokol No'lu; sorumlusu olduğunuz "regulaTER: ulaşılabilir genomik bölgelerdeki transpozonların analizi için bir R paketi" başlıklı araştırmanın uygulanmasında etik açıdan sakınca olmadığına oy birliği ile karar verilmiştir.

Bilgilerinizi ve gereğini rica ederiz.

Prof. Dr. H. Alper BAĞRIYANIK Başkan	Prof. Dr. Hilmi ORHAN Başkan Yardımcısı
Prof. Dr. Sedef AKGÜNGÖR Üye	Prof. Dr. Gülgün OKTAY Üye
Dr. Serap ERKEK ÖZHAN Üye (Katılmadı)	Prof. Dr. Şermin GENÇ Üye
Doç. Dr. Canan Aslı YILDIRIM Üye	Doç. Dr. Can KÜÇÜK Üye (Katılmadı)
Dr. Öğr. Üyesi Yavuz OKTAY Üye (Katılmadı)	

8.2. ARBIS CV

DOĞA ESKİER

Personal Information

Contact Information

Contact Address 1694 Sokak No: 2/1 D: 9
Phone 536-7382646
Email c@gmail.com @gmail.com
Web Page

Education Information

26 February 2019 - Now (5 Year 10 Month)
Doktora, Doktora, DOKUZ EYLÜL ÜNİVERSİTESİ, TÜRKİYE
İZMİR ULUSLARARASI BİYOTİP VE GENOM ENSTİTÜSÜ, MOLEKÜLER BİYOLOJİ VE
GENETİK (DR) (İNGİLİZCE)
Cumulative Grade Point Average: 4.0 / 4.0

01 February 2017 - 01 January 2019 (2 Year)
Yüksek Lisans, Tezli Program, DOKUZ EYLÜL ÜNİVERSİTESİ, TÜRKİYE
İZMİR ULUSLARARASI BİYOTİP VE GENOM ENSTİTÜSÜ, GENOM BİLİMLERİ VE
MOLEKÜLER BİYOTEKNOLOJİ ANABİLİM DALI
Thesis Title: IDENTIFICATION AND ANNOTATION OF PUTATIVE LONG NONCODING
RNAs INVOLVED IN MESENCHYMAL- EPITHELIAL TRANSITION
Thesis Subject: Mezenkimal epitelyal hücre dönüşümünün düzenlenmesinde görev
alan gen düzenleyici ağıın üyeleri olan ve daha önce anotasyonu yapılmamış uzun
protein kodlamayan RNAların belirlenmesi ve olası görevlerinin anotasyonları için
ağırlıklı gen eşifade ağı analizi kullanılması
Date: 2019
Thesis Advisor: GÖKHAN KARAKÜLAH
Thesis Co-Advisor: HANİ ALOTAİBİ
Diploma Number: 201900096
Cumulative Grade Point Average: 3.31 / 4.0

07 February 2013 - 25 February 2019 (6 Year 1 Month)
Yüksek Lisans, Tezli Program, İZMİR YÜKSEK TEKNOLOJİ ENSTİTÜSÜ, TÜRKİYE
MÜHENDİSLİK VE FEN BİLİMLERİ ENSTİTÜSÜ, MOLEKÜLER BİYOLOJİ VE GENETİK
(YL) (TEZLİ)
Cumulative Grade Point Average: 3.21 / 4.0

27 August 2006 - 16 June 2011 (4 Year 10 Month)
Lisans, Anadal/Normal Öğretim, İHSAN DOĞRAMACI BİLKENT ÜNİVERSİTESİ,
TÜRKİYE
FEN FAKÜLTESİ, MOLEKÜLER BİYOLOJİ VE GENETİK PR. (İNGİLİZCE) (TAM
BURSLU)
Diploma Number: 112406226
Cumulative Grade Point Average: 2.75 / 4.0

01 September 2003 - 01 June 2006 (2 Year 10 Month)
MİLLÎ EĞİTİM BAKANLIĞI İL MİLLÎ EĞİTİM MÜDÜRLÜKLERİ VE OKULLAR İZMİR
İLİNDEKİ OKULLAR BORNOVA İLÇESİNDEKİ OKULLAR İZMİR FEN LİSESİ, TÜRKİYE
Cumulative Grade Point Average: 5.0 / 5.0

Experience Information

February 2019 - Now (5 Year 11 Month) (Tam Zamanlı)
MİSAFİR ARAŞTIRMACI, DOKTORA ÖĞRENCİSİ, İZMİR BİYOTİP VE GENOM
MERKEZİ

Foreign Language Information

İNGİLİZCE (Reading: İyi, Writing: İyi, Speaking: İyi)

Scientific Technological Activity Fields

Scientific Technological Activity Field

Temel Bilimler -- Yaşam Bilimleri -- Biyoinformatik -- Biyolojik Modelleme

Temel Bilimler -- Yaşam Bilimleri -- Biyoinformatik -- Biyolojik Veritabanları

Temel Bilimler -- Yaşam Bilimleri -- Moleküler Biyoloji ve Genetik -- Genomiks

Keywords

Hesaplamalı Biyoloji

Biyoinformatik

Transpozonlar

Transkriptom

Fonksiyonel Genomik

Epigenom

Yeni Nesil Dizileme

R&D Competency

Articles

N. E. ESKIER, D. ESKIER, E. FIRUZAN & S. K. UZUNLAR, Physicochemical differences between camelid single-domain antibodies and mammalian antibodies, TURKISH JOURNAL OF BIOLOGY, 2023, 1300-0152, 47, 6, 27.

M. KURUS, S. AKBARI, D. ESKIER, A. BURSALI, K. ERGIN, E. ERDAL & G. KARAKULAH, Transcriptome Dynamics of Human Neuronal Differentiation From iPSC, FRONTIERS IN CELL AND DEVELOPMENTAL BIOLOGY, 2021, 2296-634X, 9, 14.

N. KOCHAN, D. ESKIER, A. SUNER, G. KARAKULAH & Y. OKTAY, Different selection dynamics of S and RdRp between SARS-CoV-2 genomes with and without the dominant mutations, INFECTION GENETICS AND EVOLUTION, 2021, 1567-1348, 91, 8.

D. ESKIER, E. AKALP, O. DALAN, G. KARAKULAH & Y. OKTAY, Current mutome of SARS-CoV-2 in Turkey reveals mutations of interest, TURKISH JOURNAL OF BIOLOGY, 2021, 1300-0152, 45, 1, 10.

D. ESKIER & G. KARAKULAH, In Silico Identification of Stress-Associated Transposable Elements in <i>Arabidopsis thaliana</i> Using Public Transcriptome Data, PLANT TRANSPOSABLE ELEMENTS, 2021, 1064-3745, 2250, 16.

D. ESKIER, A. SUNER, G. KARAKULAH & Y. OKTAY, Mutation density changes in SARS-CoV-2 are related to the pandemic stage but to a lesser extent in the dominant strain with mutations in spike and RdRp, PEERJ, 2020, 2167-8359, 8, 15.

D. ESKIER, G. KARAKULAH, A. SUNER & Y. OKTAY, RdRp mutations are associated with SARS-CoV-2 genome evolution, PEERJ, 2020, 2167-8359, 8, 14.

D. ESKIER, A. SUNER, Y. OKTAY & G. KARAKULAH, Mutations of SARS-CoV-2 nsp14 exhibit strong association with increased genome-wide mutation load, PEERJ, 2020, 2167-8359, 8, 14.

D. ESKIER & A. NALBANT ALDANMAZ, Genes Associated With T Helper 17 Cell Differentiation And Function., FRONTIERS IN BIOSCIENCE-LANDMARK, 2016, 1093-9946, 1, 8, 427-35.

TÜBİTAK Scholarships and Fundings

Project Info

122Z886, Mezenkimal Epitelial Dönüşüm Sürecinde Rol Alan Mikrorna'ların Tanımlanması Ve Fonksiyonel Analizleri, Burslu, Yürürlükte, KBAG - Ki?mya Bi?yoloji? Araştırma Destek Grubu, 1001 - Araştırma, ARDEB, Project Participation/Leave Dates: 4/4/23 - (), Project Start/End Dates: 3/15/23 - 3/15/26.

219Z034, Mezenkimal Epitelial Dönüşüm Sürecinde Kritik Olan Uzun Kodlamayan RNA'ların İşlevsel Analizleri, BİDEB Burslusu, Sonuçlandı, KBAG - Ki?mya Bi?yoloji? Araştırma Destek Grubu, 1001 - Araştırma, ARDEB, Project Participation/Leave Dates: 5/30/22 - 4/4/23, Project Start/End Dates: 7/15/20 - 7/15/23.

117Z223, Mezenkimal Epitelial Dönüşüm (MET) Sürecinde Epigenetik Yeniden Programlamayı Etkileyen Transkripsiyon Faktörlerinin Belirlenmesi, Burslu, Sonuçlandı, KBAG - Ki?mya Bi?yoloji? Araştırma Destek Grubu, 1001 - Araştırma, ARDEB, Project Participation/Leave Dates: 12/1/17 - 1/3/19, Project Start/End Dates: 10/1/17 - 10/1/20.

BİDEB Supports

DOĞA ESKIER, Etkinlik Destekleri ve Eğitim Bursları Müdürlüğü, 2250-Lisansüstü Bursları Performans Programı, Destek Sona Erdi, 2022 - 1.

DOĞA ESKIER, Etkinlik Destekleri ve Eğitim Bursları Müdürlüğü, 2250-Lisansüstü Bursları Performans Programı, Destek Sona Erdi, 2023 - 1.

DOĞA ESKIER, Etkinlik Destekleri ve Eğitim Bursları Müdürlüğü, 2250-Lisansüstü Bursları Performans Programı, Destek Sona Erdi, 2023 - 2.

DOĞA ESKİER, Etkinlik Destekleri ve Eğitim Bursları Müdürlüğü, 2250-Lisansüstü Bursları Performans Programı, Destek Sona Erdi, 2024 - 1.

DOĞA ESKİER, Etkinlik Destekleri ve Eğitim Bursları Müdürlüğü, 2250-Lisansüstü Bursları Performans Programı, Destekleniyor, 2024 - 2.

DOĞA ESKİER, Etkinlik Destekleri ve Eğitim Bursları Müdürlüğü, 2205-Lisans Burs Programı, Bursu Kesildi, 2006, 10/1/06 - 6/1/10.

DOĞA ESKİER, Etkinlik Destekleri ve Eğitim Bursları Müdürlüğü, 2211-A Genel Yurt İçi Doktora Burs Programı, Bursu Kesildi, 2019 - 1, 3/1/19 - 2/28/23.

DOĞA ESKİER, Etkinlik Destekleri ve Eğitim Bursları Müdürlüğü, 2224-A Yurt Dışı Bilimsel Etkinliklere Katılma Desteği Programı, Destek Sona Erdi, 2023 - 4, 11/8/23 - 11/11/23.

Number of Panelist/Observer/Reporter Jobs

Role	ARDEB	BİDEB	SCIENCE SOCIETY	UIDB	TEYDEB	Total
Refereeing/Panelist/Outer Consultant Count	0	0	0	0	0	0
Moderatorships Count	0	0	0	0	0	0
Number of Observer/Consultant Jobs	0	0	0	0	0	0
Number of Reporter Jobs	0	0	0	0	0	0
Not Attended Refereeing/Panelist/Outer	0	0	0	0	0	0

8.3. Research article associated with thesis study

Research Article



Exploring Regulatory Roles of Transposable Elements in EMT and MET through Data-Driven Analysis: Insights from regulaTER

Doğ'a Eskier¹, Seray Yetkin¹, Nazmiye Arslan², Gökhan Karakulah^{1,2,*}, and Hani Alotaibi^{1,2,*}

¹ - Izmir International Biomedicine and Genome Institute, Dokuz Eylül University, İzmir 35340, Turkey

² - Izmir Biomedicine and Genome Center, İzmir 35340, Turkey

Correspondence to Gökhan Karakulah and Hani Alotaibi: Izmir International Biomedicine and Genome Institute, Dokuz Eylül University, İzmir 35340, Turkey. gdeu.edu.tr (G. Karakulah), h@deu.edu.tr (H. Alotaibi)
<https://doi.org/10.1016/j.jmb.2024.168887>

Editor: Rita Casadio

Abstract

Gene expression is regulated at the transcriptional and translational levels and a plethora of epigenetic mechanisms. Regulation of gene expression by transposable elements is well documented. However, a comprehensive analysis of their regulatory roles is challenging due to the lack of dedicated approaches to define their contribution. Here, we present regulaTER, a new R library dedicated to deciphering the regulatory potential of transposable elements in a given phenotype. regulaTER utilizes a variety of genomics data of any origin and combines gene expression level information to predict the regulatory roles of transposable elements. We further validated its capabilities using data generated from an epithelial-mesenchymal and mesenchymal-epithelial transition cellular model. regulaTER stands out as an essential asset for uncovering the impact of transposable elements on the regulation of gene expression, with high flexibility to perform a range of transposable element-focused analyses. Our results also provided insights on the contribution of the MIR and B element subfamilies in regulating EMT and MET through the FoxA transcription factor family. regulaTER is publicly available and can be downloaded from <https://github.com/karakulah/regulaTER>.

© 2024 Elsevier Ltd. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

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 many epigenetic factors, transposable elements (TEs) are involved in transcriptional regulation.¹ 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 1956,² 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 forming novel chromosomal features, novel genes, and regulatory elements.³ TEs are implicated in developmental processes, where they contribute to regulatory networks through transcription factor binding sites (TFBSs) for TFs involved in cellular differentiation.^{1,4} In addition to their roles

0022-2836/© 2024 Elsevier Ltd. All rights are reserved, including those for text and data mining, AI training, and similar technologies.
Journal of Molecular Biology 437 (2025) 168887

gained via exaptation, TEs can also influence cellular phenotype via transposition and novel insertion events, and aberrant TE activity can lead to genetic diseases or tumorigenesis.^{5,6} Because of the potential deleterious effects of TE activity and insertion, these elements often must maintain a strict balance between activity and propagation or inhibition.⁷ 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.^{8–10}

As repeat genomic elements with high polymorphic variability, analysis of TEs often requires specialized tools and workflows.^{11,12} Despite the range of tools available, various perspectives are missing in the current 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

The implementation of the regulaTER library

The regulaTER tool was created in the R environment (version 4.0.0). We used the biomart package (version 1.0.2)¹³ to import RepeatMasker annotations into the R environment for use in later calculations, GenomicRanges (version 1.42.0)¹⁴ 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)¹⁵. For motif enrichment analyses, we used the HOMER genome analysis software (version 4.10.3)¹⁶ 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, are illustrated in 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,¹⁷ 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 number of Shuffle 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

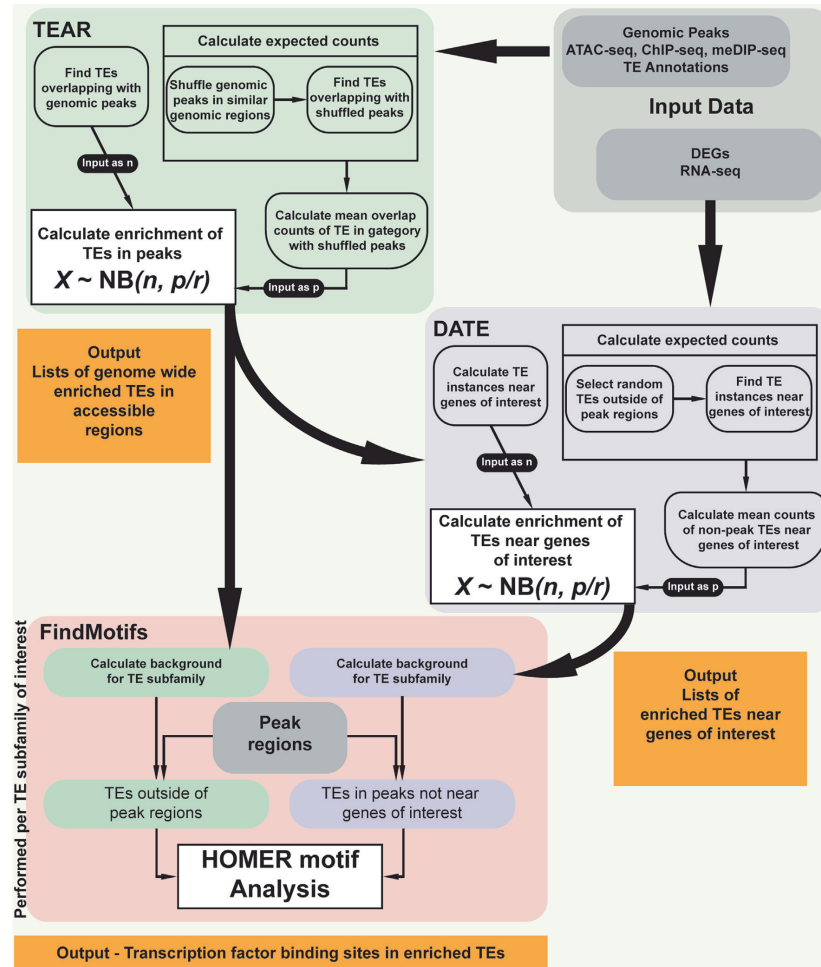


Figure 1. Detailed outline of the regulaTER functions and workflow. The gray boxes represent primary user-defined inputs (genomic peak data, DEG data). The green boxes indicate information regarding the TEAR function, while the purple boxes indicate information regarding the DATE function. The pink box 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 is repeated a number of times as determined by the user.

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).

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% CO₂. As a supplement, 10 µg/ml insulin (Sigma-Aldrich) was freshly added to the medium. EMT was induced by adding 2.5 ng/ml TGFβ3 (PeproTech) and progressed 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.¹⁸ 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 5 min, 40 cycles of 95 °C 15 s, 60 °C 60 s. 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, 250 ng/ml) overnight at 4 °C with shaking. Anti-GAPDH (Santa Cruz, 50 ng/ml) was used as a loading control. Detection was done by incubation with horse-radish peroxidase-conjugated secondary antibodies (1:10,000, Dianova), ECL prime

reagent (GE Healthcare), and exposure to X-ray films.

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-seq Kit (Active Motif), following the manufacturer's protocols. Briefly, the nuclei of approximately 100,000 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).¹⁹ The remaining high-quality reads were aligned to the GENCODE GRCh38 (version M21) mouse genome assembly obtained from the GENCODE database, using HISAT2 (version 2.0.5)²⁰, 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)²¹ “samtools view” and “samtools sort”. The RNA-seq libraries were then quantified using the featureCounts function of the Rsubread R library²². 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)²³ 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 $\text{abs}(\log_2\text{FC}) \geq 1$ and adjusted p -value (false discovery rate, $\text{FDR}) \leq 0.05$ 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)²⁴ 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))”.

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).²⁵

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 (GRCh38/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 organism-independent 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: Transposable Elements Enriched in Accessible Regions (TEAR), Differentially Expressed Gene Associated Accessible TEs (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. An overview of the functions is presented in Figure 1.

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.²⁶ 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,^{27–29} 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 discovering TEs involved in gene regulation.^{30,31}

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 Mesenchymal NMuMG), and then TGFβ3 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 4 days (designated PT100; PT as in Post-Treatment).

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 (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, validating the characteristic features expected during EMT and MET (Figure 2C).

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, $\text{abs}(\log_2\text{FC}) \geq 1$, $\text{FDR} \leq 0.05$). On the other hand, 5133 genes were differentially expressed during MET (2929 upregulated and 2204 downregulated, $\text{abs}(\log_2\text{FC}) \geq 1$, $\text{FDR} \leq 0.05$) (Tables S1 and S2). The changes in the expression profiles corresponded to a typical EMT and MET (Figure 3A and B). GSEA confirmed that the differentially expressed genes correlated perfectly with the EMT hallmark gene set (Figure 3C and D).

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 23,636 differentially accessible peaks during EMT and 12,129 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).

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

The generated ATAC-seq and RNA-seq data were used as inputs for regulaTER, and following the procedure summarized in the methods section (and detailed in the user manual), we

implemented the different functions of regulaTER as outlined in Figure 1. 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 $\log_2\text{FC} \geq 1$ and $\text{FDR} \leq 0.05$ 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 (Tables S3 and S4). Additionally, 21 families were enriched during EMT, 12 were enriched in MET, and 11 were common. Among these enriched families, the representation of accessible TE loci versus the entire genome ranged from 3.4% to 0.6% (Figure S1). When we tested the relationship between both the number and the ratio of observed accessible TEs to enrichment score using Spearman ranked test, we found a strong positive correlation for the number of observed repeats (0.79 and 0.66, respectively, for EMT and MET) but no correlation for the ratio of observed repeats (0.07 and 0), indicating that high representation of accessible loci in a TE family is not an indicator of enrichment. 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) (Tables 1 and 2). 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 (Table S5).

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.

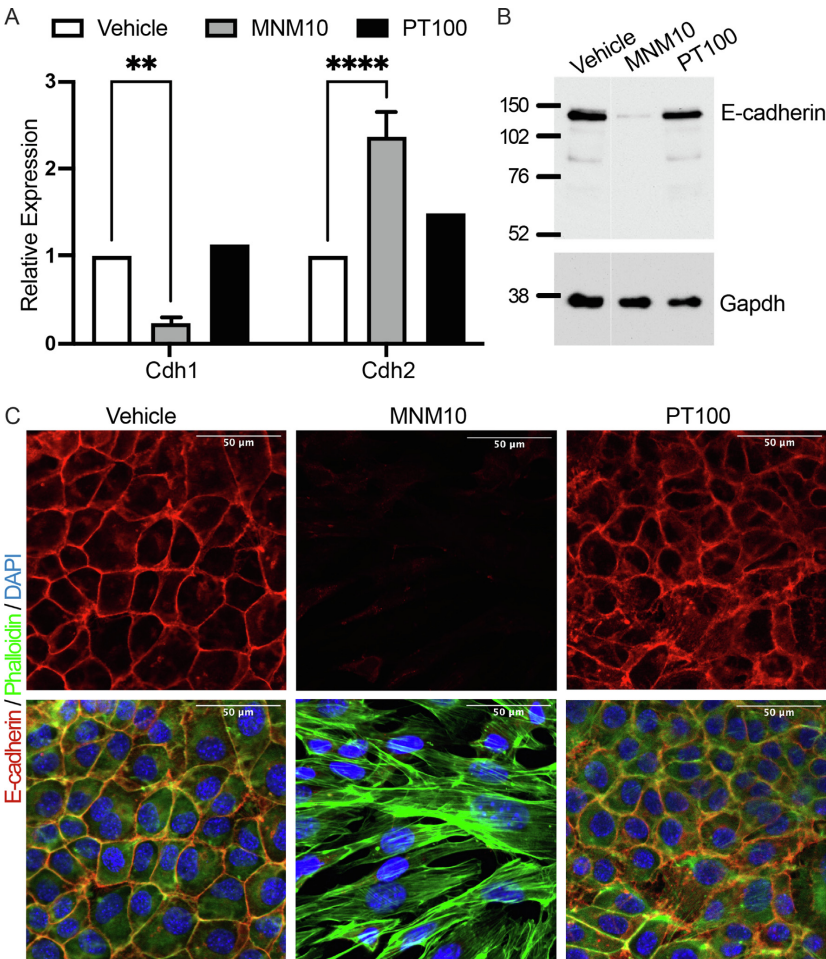


Figure 2. NMuMG cells successfully exhibited EMT and MET in response to TGFβ3. (A) qPCR analysis of the cadherin switch during EMT shows decreased Cdh1 and increased Cdh2 expression, which is reversed during MET. (B) Changes in RNA levels were also reflected at the protein level, showing a decrease of 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.

We applied the DATE function using the output of the TEAR 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 (Tables S6 and S7). 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

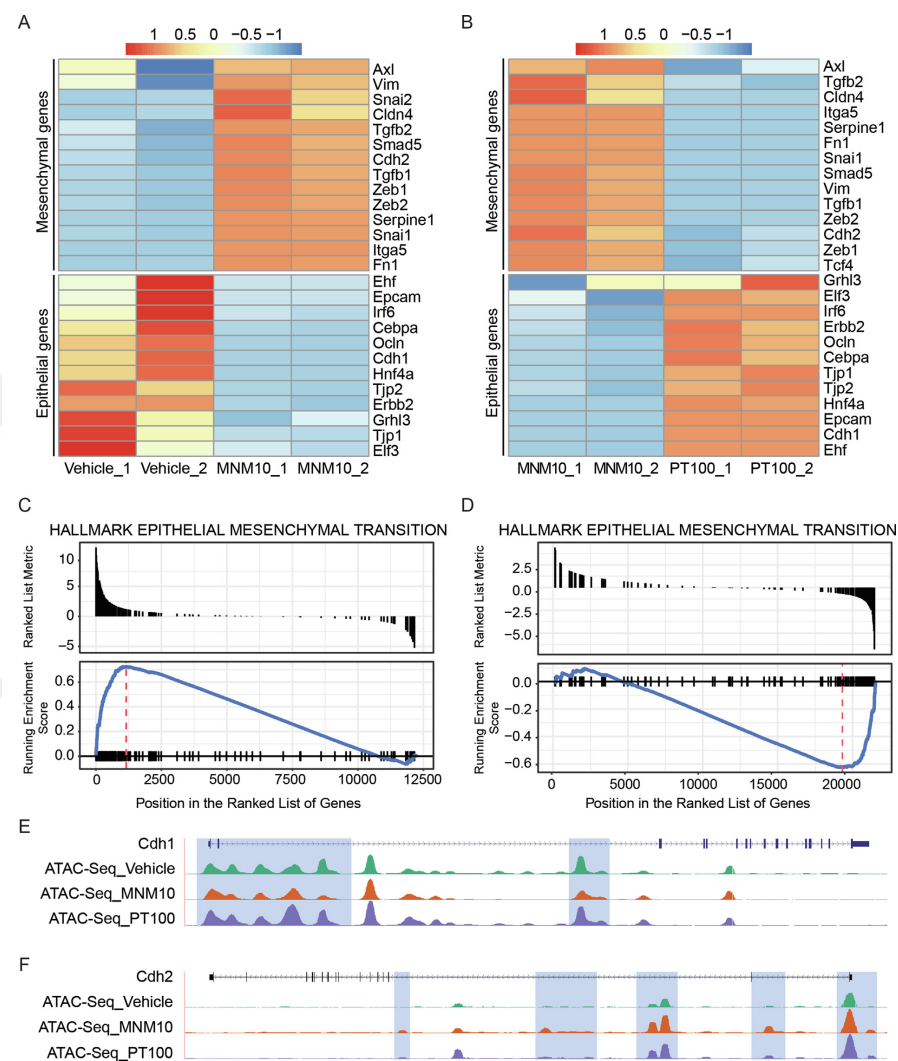


Figure 3. Gene expression analysis and differential chromatin accessibility during EMT and MET. Scaled expression heatmap of EMT markers in the (A) MNM10/Vehicle and (B) PT100/MNM10 comparisons. The colors of the cells indicate the z-score of the gene across the samples. GSEA analysis on the gene set 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).

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

Table 1 Top 10 TE subfamilies enriched in accessible peaks in EMT.

Repeat subfamily	Total genomic count	Observed	Expected	FDR	Fold change
RMER6C	4930	176	35	5.59E-64	5.05
MIRb	41,669	1051	636	6.43E-51	1.65
RMER15	18,012	466	237	6.49E-39	1.97
MIR	38,078	865	571	1.01E-29	1.51
RLTR17	1752	91	22	1.75E-27	4.22
ORR1F	22,992	488	286	9.40E-27	1.71
RLTR48A	621	57	10	3.77E-24	5.87
MLT1H	2337	94	28	6.42E-22	3.31
MIR3	15,244	430	264	3.21E-20	1.63
RLTR30D2_MM	733	50	10	1.01E-18	5.22

Table 2 Top 10 TE subfamilies enriched in accessible peaks in MET.

Repeat subfamily	Total genomic count	Observed	Expected	FDR	Fold change
RLTR48A	621	27	3	2.48E-17	9.00
L2a	23,921	282	172	7.74E-15	1.64
L2b	13,281	190	110	2.34E-12	1.72
RMER16A2	285	13	1	5.02E-11	10.11
MIR	38,078	377	280	1.82E-08	1.35
RMER15	18,012	165	104	1.93E-08	1.59
RMER6C	4930	37	13	3.85E-08	2.88
RLTR13C1	321	16	3	9.64E-08	5.60
MIRb	41,669	419	324	2.19E-07	1.29
RLTR13D3A1	604	22	6	3.18E-07	3.85

(Figure 4B–C and Figure S2). Notably, response to TGFβ3, regulation of substrate adhesion, migration, and wound healing are among the most enriched biological processes identified in EMT. Upregulated genes (such as Acta2, Itgb5, and Timp1) relevant to these processes also showed marked chromatin accessibility (Figure S3). In contrast, MET-related enrichment processes were linked to developmental processes or proliferation (Table S8).

TE-contributed binding sites for FoxA1 and FoxA2 are enriched near differentially expressed genes related to EMT and MET

In order to predict the potential roles of the identified TEs in gene regulation, we applied 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 among 66 and 53 enriched genomic motifs, respectively. 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, among 58 enriched motifs (Figure 5C), and in B4A located in the intron of Rab17, among 46 motifs (Figure 5D). The entire known motif enrichment results for B1_Mus1 and B4A in EMT and ID_B1 and B4A in MET are presented in Supplemental Tables S9–S12.

Application of regulaTER to independent datasets shows wide applicability

To understand how regulaTER functions would perform when applied to independently produced datasets, as well as other organisms or library types, we obtained FoxA2 ChIP-seq (GSE90536) and RNA-seq (GSE90534) results of induced hepatocyte (iHep) differentiation derived from mouse embryonic fibroblasts (MEF),³² and ATAC-seq (GSE210496) and RNA-seq (GSE210499) results of a human EMT model, TGFβ1 treated MCF10A cell line.³³ The peak files for both the ChIP-seq and ATAC-seq libraries, as well as the transcript counts for the RNA-seq libraries, were obtained from the Gene Expression Omnibus database and used as inputs for regulaTER analysis. The FoxA2 binding peaks in the induced hepatocyte cells were primarily enriched for L1 retrotrans-

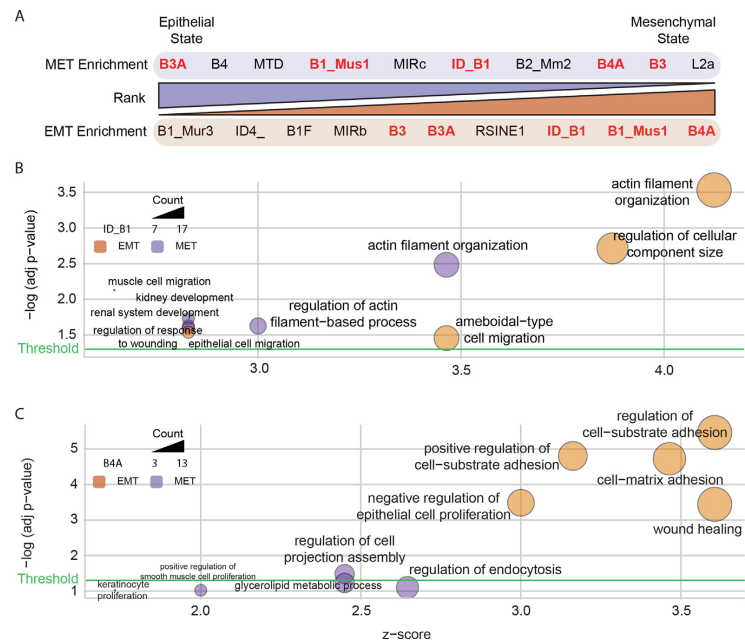


Figure 4. Five commonly enriched TEs show relevance to EMT- and MET-related biological processes. (A) The top 10 TE subfamilies enriched after the DATE analysis in EMT (orange) and MET (purple). TE subfamilies are ranked according to their contribution to epithelial (left end of the bar) or mesenchymal (right end of the bar) states. The TE subfamilies in red and bold text represent those commonly enriched in EMT and MET. (B–C) Modified GO bubble plots of the GO enrichment results for the upregulated genes within 5000 bp of the accessible instances of the B element TE subfamilies ID_B1 (B) and B4A (C). 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 color of the bubbles indicates the process the term is associated with, with orange shapes enriched in EMT and purple shapes enriched in MET. The shape's size indicates the number of TE-neighboring upregulated genes in the term.

posons and hAT-Charlie DNA transposons across the genome (Table S13). Due to the smaller footprint of FoxA2 binding compared to accessible chromatin, a smaller number of TEs showed enrichment compared to our data, with only 14 enriched TE subfamilies. Similarly, the DATE results revealed only the B4 subfamily as enriched near DEGs in the iHep-MEF comparison, further highlighting their potential role in cell reprogramming (Table S14). The DEG-associated B4 loci that bind FoxA2 were enriched in Oct and FoxA1 binding sites (Table S15), which are TFs with functions in cellular reprogramming.

The MCF10A human cell line EMT samples, in line with our murine EMT model, showed enrichment of MIR and ERV families in accessible regions, although diverged in enriched subfamilies (Table S16). DATE results displayed even greater

differences, with four different L2 subfamilies and four different MIR subfamilies showing the highest enrichment (Table S17). Interestingly, two Alu subfamilies were also in the top 10 of the DATE results, with a higher proportion of Alus among TEs showing less significant enrichment. As Alu TEs are the primate orthologues of murine B elements,³⁴ and given the ability of Alu insertions to alter nearby gene expression,³⁵ this suggests a potential role of Alu / B element and MIR TEs in EMT regulation shared between both species. In the FindMotifs results, L2a and L2b both showed enrichment for FoxA1, while AluJb and AluY were enriched for other Forkhead TF binding motifs, namely Foxh1 and Foxb1. Taken together, our findings point towards the potential impact of TEs on EMT and MET, as well as cellular differentiation, through their contributions of Forkhead binding

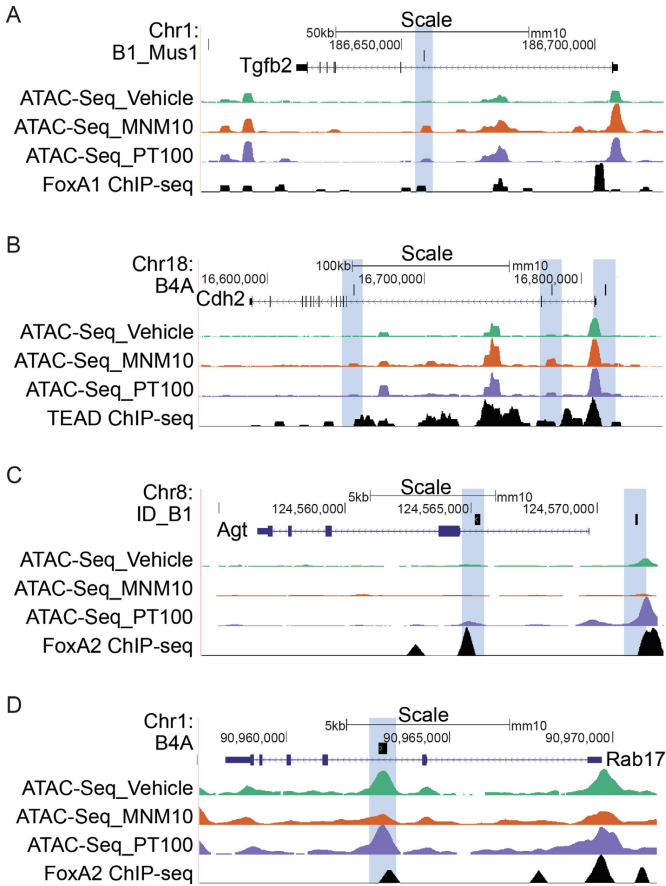


Figure 5. ReMap ChIP-seq data for TEAD and the FoxA family of transcription factors colocate with the enriched accessible TEs. UCSC Genome Browser images of (A) Tgfb2 and B1_Mus1, (B) Cdh2 and B4A, (C) Agt and ID_B1, and (D) Rab17 and B4A. Regions highlighted in blue indicate TE instances showing increased 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.

sites to multiple genes across multiple species. Overall, this demonstrated the applicability of regulaTER in utilizing different omics data from different species.

Comparison to similar analysis methods

To highlight the uses and advantages of regulaTER, we compared it to similar analysis tools and databases. We selected the three

available tools with a high amount of overlap with our library used to calculate the potential effects of TEs on the regulation of nearby genes, namely GREAM,³⁶ TEffectR,³⁷ and PlanTEEnrichment.³⁸ Due to the different input types and computational methods utilized by each tool, a quantitative comparison was unfeasible. Therefore, we compared each tool's capabilities and potential uses (Table 3). Overall, regulaTER has several unique strengths that were not found in other tools. Of highest rele-

vance is the ability of regulaTER to accept both custom genomes and repeat annotations and its capacity to accept genomic sequencing peaks to consider the epigenomic state of individual TE loci in the enrichment analysis. These genomic sequencing peaks can consist of various types of omics data, such as methylation, genomic accessibility, protein binding, and histone marks (ChIP-seq or CUT&RUN). Incorporating such multi-omics data into TE analyses allows researchers to identify the TE loci relevant to the phenotype of interest. This perspective is unavailable in other tools, such as GREAM, which calculates enrichment using all TE loci, regardless of whether their chromatin state is compatible with gene regulatory functions.

RegulaTER can also perform TE enrichment in genomic peaks independent of gene sets using the TEAR function. Finally, regulaTER uses the HOMER software and custom background region calculation based on the multi-omics input to identify motif enrichment in relevant TE loci rather than based on the consensus sequence of the TE subfamily. In comparison, GREAM only uses a limited number of preset genomes and TE annotations for its enrichment analyses. TEeffectR allows for custom genome and repeat annotations but cannot include genomic peak information or perform gene-independent enrichment analysis. While the PlanTEEnrichment database has the lowest functionality and output detail, it covers plant species not included in the GREAM database, which can also be used for regulaTER analyses. This comparison shows that regulaTER fills a vital niche in TE analysis tools that are not covered by other similar software.

Discussion

In this study, we aimed to extend our understanding of the regulatory dynamics governing EMT and MET, two fundamental processes in cellular biology that play crucial roles in development, tissue repair, and disease progression, particularly cancer metastasis.^{26,39,40} 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.⁴¹ 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.³⁶ 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. Other tools, such as TEeffectR³⁷ and TEProf2,⁴² have examined the effects of TE transcription, whether on their own or as part of TE-chimeric transcripts of other genes. However, these tools depend on TE-derived transcriptome reads and cannot identify the regulatory effects of TEs occurring independently of TE transcription. In addition, we have previously made available PlanTEEnrichment, a tool to identify TEs enriched upstream of a list of given genes in 11 reference plant genomes,³⁸ 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.⁴³ 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, including non-TE repeats included in the RepeatMasker format, such as simple repeats, tRNAs, and endogenous retroviruses.

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 binding 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.^{44–52} These transcription factor binding sites were in proximity to genes that shape the mesenchymal state during EMT, such as Cdh2, Fn1, Itga5, Wnt2b, Zeb1/2, and Tgfb1.^{53,54} 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 regulating genes driving HCC metastasis to facilitate drug identification.⁵⁵ In another study, LINE-1 elements enhanced the malignancy of human bronchial epithelial cells through interaction and co-regulation with TGFβ1.²⁸ Stemness and cellular plasticity during EMT in colorectal and breast can-

Table 3 Comparison of regulaTER with similar tools.

Software	regulaTER	GREAM	TEffectR	PlanTEEnrichment
Software type	R library	Database	R library	Database
Custom genome and repeat annotations	Yes	No	Yes	No
Analyzed species	Custom input	17 mammalian	Custom input	11 plant
Genomic context of TE instances (e.g., methylation, chromatin accessibility, etc.)	Yes	No	No	No
Gene-independent enrichment analysis	Yes	No	No	No
Gene-TE distance	Custom input	Preset options	Custom input	Preset options
Motif enrichment analysis	Compares relevant TE loci to custom background using HOMER	Uses consensus sequence of TE using JASPAR and CLOVER	No	No
Result statistics	TE only	TE or gene	TE and gene	TE only
Orthologous gene set analysis	No	Yes	No	No
Automated GO term enrichment analysis	No	Yes	No	No
TE expression data	No	No	Yes	No

cer tumorigenesis were also shown to be regulated by the L1 retrotransposon L1-ORF1.²⁷ 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.²⁹

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.^{18,40,56} Although the 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.^{18,51,57–62} 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,^{44,63} 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.⁶⁴ Another target of FoxA2 is angiotensin (Agt), which can inhibit Tgfb1-induced EMT in keratinocytes.⁶⁵

Conclusions

regulaTER offers a powerful tool for comprehensively assessing the regulatory roles of transposable elements (TEs) in gene regulation, overcoming existing limitations in the field. By applying regulaTER to multi-omics data from EMT and MET processes, we identified significant

contributions of the MIR and B element TE families, particularly in relation to FoxA and TEAD TFBSs near key regulatory genes. These findings underscore the nuanced regulatory roles of TEs in driving cellular transitions and phenotypic plasticity and highlight regulaTER's ability to extend beyond traditional tools, enabling the identification of TEs' contributions to transcriptomic regulation within specific phenotypes. We also complemented our results with independently produced datasets using other genomic peak types or species, further highlighting both the flexibility of regulaTER and the validity of our findings. In addition, the consistency of our predictions with previous experimental data showcases regulaTER as a robust tool for identifying future research targets. This advancement represents a significant step in unraveling the complex interactions between TEs and gene regulation, offering new insights into their impact on cellular behavior.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

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

CRedit authorship contribution statement

Doğa Eskier: Writing – review & editing, Writing – original draft, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Seray Yetkin:** Writing – original draft, Resources, Methodology, Investigation, Formal analysis. **Nazmiye Arslan:** Software, Methodology, Formal analysis, Data curation, Conceptualization. **Gökhan Karakulah:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Hani Alotaibi:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Funding

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

DATA AVAILABILITY

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

DECLARATION OF COMPETING INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We thank Ahmet Bursalı and Alinza Arbaş for their assistance with data analysis and visualization.

Appendix A. Supplementary material

Supplementary material to this article can be found online at <https://doi.org/10.1016/j.jmb.2024.168887>.

Received 24 September 2024;
Accepted 28 November 2024;
Available online 2 December 2024

Keywords:

transposable elements;
gene regulation;
regulaTER;
EMT;
MET

References

- Gebrie, A., (2023). Transposable elements as essential elements in the control of gene expression. *Mob. DNA* **14**, 9. <https://doi.org/10.1186/s13100-023-00297-3>.
- McClintock, B., (1956). Controlling elements and the gene. *Cold Spring Harb. Symp. Quant. Biol.* **21**, 197–216. <https://doi.org/10.1101/sqb.1956.021.01.017>.
- Liao, X., Zhu, W., Zhou, J., Li, H., Xu, X., Zhang, B., Gao, X., (2023). Repetitive DNA sequence detection and its role in the human genome. *Commun. Biol.* **6**, 954. <https://doi.org/10.1038/s42003-023-05322-y>.
- Fueyo, R., Judd, J., Feschotte, C., Wysocka, J., (2022). Roles of transposable elements in the regulation of mammalian transcription. *Nature Rev. Mol. Cell Biol.* **23**, 481–497. <https://doi.org/10.1038/s41580-022-00457-y>.
- Bhat, A., Ghatage, T., Bhan, S., Lahane, G.P., Dhar, A., Kumar, R., Pandita, R.K., Bhat, K.M., Ramos, K.S., Pandita, T.K., (2022). Role of transposable elements in genome stability: implications for health and disease. *Int. J. Mol. Sci.* **23** <https://doi.org/10.3390/ijms23147802>.
- Yandim, C., Karakulah, G., (2019). Dysregulated expression of repetitive DNA in ER+/HER2- breast cancer. *Cancer Genet.* **239**, 36–45. <https://doi.org/10.1016/j.cancergen.2019.09.002>.
- Yoth, M., Jensen, S., Brasset, E., (2022). The intricate evolutionary balance between transposable elements and their host: who will kick at goal and convert the next try? *Biology (Basel)* **11** <https://doi.org/10.3390/biology11050710>.
- Aravin, A.A., Hannon, G.J., Brennecke, J., (2007). The Piwi-piRNA pathway provides an adaptive defense in the transposon arms race. *Science* **318**, 761–764. <https://doi.org/10.1126/science.1146484>.
- Molaro, A., Malik, H.S., (2016). Hide and seek: how chromatin-based pathways silence retroelements in the mammalian germline. *Curr. Opin. Genet. Dev.* **37**, 51–58. <https://doi.org/10.1016/j.gde.2015.12.001>.
- Yang, P., Wang, Y., Macfarlan, T.S., (2017). The role of KRAB-ZFPs in transposable element repression and mammalian evolution. *Trends Genet.* **33**, 871–881. <https://doi.org/10.1016/j.tig.2017.08.006>.
- Goerner-Potvin, P., Bourque, G., (2018). Computational tools to unmask transposable elements. *Nature Rev. Genet.* **19**, 688–704. <https://doi.org/10.1038/s41576-018-0050-x>.
- Lanciano, S., Cristofari, G., (2020). Measuring and interpreting transposable element expression. *Nature Rev. Genet.* **21**, 721–736. <https://doi.org/10.1038/s41576-020-0251-y>.
- Drost, H.G., Paszkowski, J., (2017). Biomart: genomic data retrieval with R. *Bioinformatics* **33**, 1216–1217. <https://doi.org/10.1093/bioinformatics/btw821>.
- Lawrence, M., Huber, W., Pages, H., Abouyoun, P., Carlson, M., Gentleman, R., Morgan, M.T., Carey, V.J., (2013).

- Software for computing and annotating genomic ranges. *PLoS Comput. Biol.* **9**, <https://doi.org/10.1371/journal.pcbi.1003118> e1003118.
15. Quinlan, A.R., Hall, I.M., (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842. <https://doi.org/10.1093/bioinformatics/btq033>.
 16. Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., 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. <https://doi.org/10.1016/j.molcel.2010.05.004>.
 17. Yu, G., Wang, L.G., He, Q.Y., (2015). ChIPseeker: an R/bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics* **31**, 2382–2383. <https://doi.org/10.1093/bioinformatics/btv145>.
 18. Alotaibi, H., Basilicata, M.F., Shehwana, H., Kosowan, T., Schreck, I., Braeutigam, C., Konu, O., Brabletz, T., Stemmler, M.P., (2015). Enhancer cooperativity as a novel mechanism underlying the transcriptional regulation of E-cadherin during mesenchymal to epithelial transition. *Biochimica et Biophysica Acta-Genes Regulat. Mech.* **1849**, 731–742. <https://doi.org/10.1016/j.bbaggm.2015.01.005>.
 19. Bolger, A.M., Lohse, M., Usadel, B., (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120. <https://doi.org/10.1093/bioinformatics/btu170>.
 20. Kim, D., Langmead, B., Salzberg, S.L., (2015). HISAT: a fast spliced aligner with low memory requirements. *Nature Methods* **12**, 357–360. <https://doi.org/10.1038/nmeth.3317>.
 21. Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R.M., Li, H., (2021). Twelve years of SAMtools and BCFtools. *GigaScience* **10** <https://doi.org/10.1093/gigascience/giab008>.
 22. Liao, Y., Smyth, G.K., Shi, W., (2019). The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. *Nucleic Acids Res.* **47**, e47.
 23. Robinson, M.D., McCarthy, D.J., Smyth, G.K., (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140. <https://doi.org/10.1093/bioinformatics/btp616>.
 24. Yu, G., Wang, L.G., Han, Y., He, Q.Y., (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* **16**, 284–287. <https://doi.org/10.1089/omi.2011.0118>.
 25. Walter, W., Sanchez-Cabo, F., Ricote, M., (2015). GOplot: an R package for visually combining expression data with functional analysis. *Bioinformatics* **31**, 2912–2914. <https://doi.org/10.1093/bioinformatics/btv300>.
 26. Yao, W., Wang, Z., Ma, H., Lin, Y., Liu, X., Li, P., He, X., (2024). Epithelial-mesenchymal plasticity (EMP) in wound healing: exploring EMT mechanisms, regulatory network, and therapeutic opportunities. *Heliyon* **10**, <https://doi.org/10.1016/j.heliyon.2024.e34269> e34269.
 27. Apostolou, P., Toloudi, M., Chatzioannou, M., Kourtidou, E., Mimikakou, G., Vlachou, I., Chlichlia, A., Papasotiriou, I., (2015). Involvement of retrotransposon L1 in stemness and cellular plasticity. *Cell Commun. Adhes.* **22**, 1–7. <https://doi.org/10.3109/15419061.2014.970270>.
 28. Reyes-Reyes, E.M., Aispuro, I., Tavera-García, M.A., Field, M., Moore, S., Ramos, I., Ramos, K.S., (2017). LINE-1 couples EMT programming with acquisition of oncogenic phenotypes in human bronchial epithelial cells. *Oncotarget* **8**, 103828–103842. <https://doi.org/10.18632/oncotarget.21953>.
 29. Simo-Riuadbas, L., Offner, S., Planet, E., Duc, J., Abrami, L., Dind, S., Coudray, A., Coto-Llerena, M., Ercan, C., Piscuoglio, S., Andersen, C.L., Bramsen, J.B., Trono, D., (2022). Transposon-activated POU5F1B promotes colorectal cancer growth and metastasis. *Nature Commun.* **13**, 4913. <https://doi.org/10.1038/s41467-022-32649-7>.
 30. Pei, D., Shu, X., Gassama-Diagne, A., Thiery, J.P., (2019). Mesenchymal-epithelial transition in development and reprogramming. *Nature Cell Biol.* **21**, 44–53. <https://doi.org/10.1038/s41556-018-0195-z>.
 31. Yao, D., Dai, C., Peng, S., (2011). Mechanism of the mesenchymal-epithelial transition and its relationship with metastatic tumor formation. *Mol. Cancer Res.* **9**, 1608–1620. <https://doi.org/10.1158/1541-7786.MCR-10-0568>.
 32. Horisawa, K., Udono, M., Ueno, K., Ohkawa, Y., Nagasaki, M., Sekiya, S., Suzuki, A., (2020). The dynamics of transcriptional activation by hepatic reprogramming factors. *Mol. Cell* **79**, 660–676.e668. <https://doi.org/10.1016/j.molcel.2020.07.012>.
 33. Anton-Garcia, P., Haghighi, E.B., Rose, K., Vladimirov, G., Boeries, M., Hecht, A., (2023). TGFβ1-induced EMT in the MCF10A mammary epithelial cell line model is executed independently of SNAIL1 and ZEB1 but relies on JUNB-coordinated transcriptional regulation. *Cancers (Basel)* **15** <https://doi.org/10.3390/cancers15020558>.
 34. Tsigos, A., Rigoutsos, I., (2009). Alu and b1 repeats have been selectively retained in the upstream and intronic regions of genes of specific functional classes. *PLoS Comput. Biol.* **5**, <https://doi.org/10.1371/journal.pcbi.1000610> e1000610.
 35. Payer, L.M., Steranka, J.P., Kryatova, M.S., Grillo, G., Lupien, M., Rocha, P.P., Burns, K.H., (2021). Alu insertion variants alter gene transcript levels. *Genome Res.* **31**, 2236–2248. <https://doi.org/10.1101/gr.261305.120>.
 36. Chandrashekar, D.S., Dey, P., Acharya, K.K., (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**, <https://doi.org/10.1371/journal.pone.0133647> e0133647.
 37. Karakulah, G., Arslan, N., Yandim, C., Suner, A., (2019). TEffector: an R package for studying the potential effects of transposable elements on gene expression with linear regression model. *PeerJ* **7**, <https://doi.org/10.7717/peerj.8192> e8192.
 38. Karakulah, G., Suner, A., (2017). PlanTEncrichment: a tool for enrichment analysis of transposable elements in plants. *Genomics* **109**, 336–340. <https://doi.org/10.1016/j.ygeno.2017.05.008>.
 39. Brabletz, T., Kalluri, R., Nieto, M.A., Weinberg, R.A., (2018). EMT in cancer. *Nature Rev. Cancer* **18**, 128–134. <https://doi.org/10.1038/nrc.2017.118>.
 40. Yang, J., Antin, P., Bex, G., Blanpain, C., Brabletz, T., Bronner, M., Campbell, K., Cano, A., Casanova, J., Christofori, G., Dedhar, S., Derynck, R., Ford, H.L., Fuxe, J., García de Herreros, A., Goodall, G.J., Hadjantonakis, A.-K., Huang, R.Y.J., Kalchauer, C., Kalluri, R., Kang, Y., Khew-Goodall, Y., Levine, H., Liu, J., Longmore, G.D.,

- Mani, S.A., Massagué, J., Mayor, R., McClay, D., Mostov, K.E., Newgreen, D.F., Nieto, M.A., Puisieux, A., Runyan, R., Savagner, P., Stanger, B., Stemmler, M.P., Takahashi, Y., Takeichi, M., Theveneau, E., Thiery, J.P., Thompson, E.W., Weinberg, R.A., Williams, E.D., Xing, J., Zhou, B.P., Sheng, G., On behalf of the, E. M. T. I. A., (2020). Guidelines and definitions for research on epithelial-mesenchymal transition. *Nature Rev. Mol. Cell Biol.* **21**, 341–352. <https://doi.org/10.1038/s41580-020-0237-9>.
41. Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., Teller, E., (1953). Equation of state calculations by fast computing machines. *J. Chem. Phys.* **21**, 1087–1092. <https://doi.org/10.1063/1.1699114>.
42. Shah, N.M., Jang, H.J., Liang, Y., Maeng, J.H., Tzeng, S. C., Wu, A., Basri, N.L., Qu, X., Fan, C., Li, A., Katz, B., Li, D., Xing, X., Evans, B.S., Wang, T., (2023). Pan-cancer analysis identifies tumor-specific antigens derived from transposable elements. *Nature Genet.* **55**, 631–639. <https://doi.org/10.1038/s41588-023-01349-3>.
43. Karakulah, G., (2018). RTFADB: A database of computationally predicted associations between retrotransposons and transcription factors in the human and mouse genomes. *Genomics* **110**, 257–262. <https://doi.org/10.1016/j.ygeno.2017.11.002>.
44. Liu, Y., Chen, T., Guo, M., Li, Y., Zhang, Q., Tan, G., Yu, L., Tan, Y., (2021). FOXA2-interacting FOXP2 prevents epithelial-mesenchymal transition of breast cancer cells by stimulating E-Cadherin and PHF2 transcription. *Front. Oncol.* **11**, <https://doi.org/10.3389/fonc.2021.605025> 605025.
45. Wang, H., Meyer, C.A., Fei, T., Wang, G., Zhang, F., Liu, X.S., (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. <https://doi.org/10.1186/1471-2164-14-680>.
46. Chi, M., Liu, J., Mei, C., Shi, Y., Liu, N., Jiang, X., Liu, C., Xue, N., Hong, H., Xie, J., Sun, X., Yin, B., Meng, X., Wang, B., (2022). TEAD4 functions as a prognostic biomarker and triggers EMT via PI3K/AKT pathway in bladder cancer. *J. Exp. Clin. Cancer Res.* **41**, 175. <https://doi.org/10.1186/s13046-022-02377-3>.
47. Sanchez-Tillo, E., de Barrios, O., Siles, L., Cuatrecasas, M., Castells, A., Postigo, A., (2011). beta-catenin/TCF4 complex induces the epithelial-to-mesenchymal transition (EMT)-activator ZEB1 to regulate tumor invasiveness. *PNAS* **108**, 19204–19209. <https://doi.org/10.1073/pnas.1108977108>.
48. Park, S.M., Gaur, A.B., Lengyel, E., Peter, M.E., (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. <https://doi.org/10.1101/gad.1640608>.
49. Zavadil, J., Cermak, L., Soto-Nieves, N., Bottinger, E.P., (2004). Integration of TGF-beta/Smad and Jagged1/Notch signalling in epithelial-to-mesenchymal transition. *EMBO J.* **23**, 1155–1165. <https://doi.org/10.1038/sj.emboj.7600069>.
50. Yang, J., Mani, S.A., Donaher, J.L., Ramaswamy, S., Itzykson, R.A., Come, C., Savagner, P., Gitelman, I., Richardson, A., Weinberg, R.A., (2004). Twist, a master regulator of morphogenesis, plays an essential role in tumor metastasis. *Cell* **117**, 927–939. <https://doi.org/10.1016/j.cell.2004.06.006>.
51. Gregory, P.A., Bert, A.G., Paterson, E.L., Barry, S.C., Tsykin, A., Farshid, G., Vadas, M.A., Khew-Goodall, Y., Goodall, G.J., (2008). The miR-200 family and miR-205 regulate epithelial to mesenchymal transition by targeting ZEB1 and SIP1. *Nature Cell Biol.* **10**, 593–601. <https://doi.org/10.1038/ncb1722>.
52. Sayan, A.E., (2014). Tumour-promoting role of EMT-inducing transcription factor ZEB1 in mantle cell lymphoma. *Cell Death Differ.* **21**, 194–195. <https://doi.org/10.1038/cdd.2013.178>.
53. Gonzalez, D.M., Medici, D., (2014). Signaling mechanisms of the epithelial-mesenchymal transition. *Sci. Signal.* **7**, re8. <https://doi.org/10.1126/scisignal.2005189>.
54. Huang, R.Y., Kuay, K.T., Tan, T.Z., Asad, M., Tang, H.M., Ng, A.H., Ye, J., Chung, V.Y., Thiery, J.P., (2015). Functional relevance of a six mesenchymal gene signature in epithelial-mesenchymal transition (EMT) reversal by the triple angiokinase inhibitor, nintedanib (BIBF1120). *Oncotarget* **6**, 22098–22113. <https://doi.org/10.18632/oncotarget.4300>.
55. Kodama, T., Newberg, J.Y., Kodama, M., Rangel, R., Yoshihara, K., Tien, J.C., Parsons, P.H., Wu, H., Finegold, M.J., Copeland, N.G., Jenkins, N.A., (2016). Transposon mutagenesis identifies genes and cellular processes driving epithelial-mesenchymal transition in hepatocellular carcinoma. *PNAS* **113**, E3384–E3393. <https://doi.org/10.1073/pnas.1606876113>.
56. Werth, M., Walentin, K., Aue, A., Schonheit, J., Wuebben, A., Pode-Shakked, N., Vilianovitch, L., Erdmann, B., Dekel, B., Bader, M., Barasch, J., Rosenbauer, F., Luft, F.C., Schmidt-Ott, K.M., (2010). The transcription factor grainyhead-like 2 regulates the molecular composition of the epithelial apical junctional complex. *Development* **137**, 3835–3845. <https://doi.org/10.1242/dev.055483>.
57. Li, R., Liang, J., Ni, S., Zhou, T., Qing, X., Li, H., He, W., Chen, J., Li, F., Zhuang, Q., Qin, B., Xu, J., Li, W., Yang, J., Gan, Y., Qin, D., Feng, S., Song, H., Yang, D., Zhang, B., Zeng, L., Lai, L., Esteban, M.A., Pei, D., (2010). A mesenchymal-to-epithelial transition initiates and is required for the nuclear reprogramming of mouse fibroblasts. *Cell Stem Cell* **7**, 51–63. <https://doi.org/10.1016/j.stem.2010.04.014>.
58. Sengez, B., Aygun, I., Shehwana, H., Toyran, N., Tercan Avci, S., Konu, O., Stemmler, M.P., Alotaibi, H., (2019). The transcription factor Elf3 is essential for a successful mesenchymal to epithelial transition. *Cells* **8** <https://doi.org/10.3390/cells8080858>.
59. Bullock, M.D., Sayan, A.E., Packham, G.K., Mirnezami, A. H., (2012). MicroRNAs: critical regulators of epithelial to mesenchymal (EMT) and mesenchymal to epithelial transition (MET) in cancer progression. *Biol. Cell* **104**, 3–12. <https://doi.org/10.1111/boc.201100115>.
60. Cao, J., Liu, J., Long, J., Fu, J., Huang, L., Li, J., Liu, C., Zhang, X., Yan, Y., (2017). microRNA-23b suppresses epithelial-mesenchymal transition (EMT) and metastasis in hepatocellular carcinoma via targeting Pyk2. *Biomed. Pharmacother.* **89**, 642–650. <https://doi.org/10.1016/j.biopha.2017.02.030>.
61. Chen, D., Zhao, H., (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**, 3286–3293.
62. Imani, S., Wei, C., Cheng, J., Khan, M.A., Fu, S., Yang, L., Tania, M., Zhang, X., Xiao, X., Zhang, X., Fu, J., (2017). MicroRNA-34a targets epithelial to mesenchymal transition-inducing transcription factors (EMT-TFs) and inhibits breast cancer cell migration and invasion.

- Oncotarget* **8**, 21362–21379. <https://doi.org/10.18632/oncotarget.15214>.
63. Zhang, Z., Yang, C., Gao, W., Chen, T., Qian, T., Hu, J., Tan, Y., (2015). FOXA2 attenuates the epithelial to mesenchymal transition by regulating the transcription of E-cadherin and ZEB2 in human breast cancer. *Cancer Lett.* **361**, 240–250. <https://doi.org/10.1016/j.canlet.2015.03.008>.
 64. Wang, M., Wang, W., Ding, J., Wang, J., Zhang, J., (2020). Downregulation of Rab17 promotes cell proliferation and invasion in non-small cell lung cancer through STAT3/HIF-1alpha/VEGF signaling. *Thorac Cancer* **11**, 379–388. <https://doi.org/10.1111/1759-7714.13278>.
 65. Jihu, Y., Leng, R., Liu, M., Ren, H., Xie, D., Yao, C., Yan, H., (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**, 1049–1058. <https://doi.org/10.2147/CCID.S441596>.

