T.C.

DOKUZ EYLÜL UNIVERSITY

İZMİR INTERNATIONAL

BIOMEDICINE AND GENOME

INSTITUTE

ANALYSIS OF LONG NON-CODING RNAS IN EPITHELIAL-MESENCHYMAL TRANSITION AND MESENCHYMAL-EPITHELIAL TRANSITION PROCESSES: COMPARATIVE SYNTENY ANALYSIS BETWEEN HUMAN AND MOUSE GENOMES

ELİF YILDIZ

DEPARTMENT OF BIOMEDICINE AND HEALTH TECHNOLOGIES

MASTER OF SCIENCE THESIS

İZMİR-2024

THESIS CODE: DEU.IBG.2022850020

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

ANALYSIS OF LONG NON-CODING RNAS IN EPITHELIAL-MESENCHYMAL TRANSITION AND MESENCHYMAL-EPITHELIAL TRANSITION PROCESSES: COMPARATIVE SYNTENY ANALYSIS BETWEEN HUMAN AND MOUSE GENOMES

DEPARTMENT OF BIOMEDICINE AND HEALTH TECHNOLOGIES

MASTER OF SCIENCE THESIS

ELİF YILDIZ
Supervisor: Prof. Dr. Gökhan Karakülah
Co-supervisor: Assoc. Prof. Dr. Hani Alotaibi

This work was supported by TÜSEB grant 26563

THESIS CODE: DEU.IBG.2022850020

T.C.
DOKUZ EYLÜL ÜNİVERSİTESİ
İZMİR ULUSLARARASI BİYOTIP VE GENOM ENSTİTÜSÜ

MASTER'S THESIS DEFENSE EXAM REPORT

Elif YILDIZ, who is a student at the Department of Biomedicine And Health Technologies Master's program of the Dokuz Eylul University's Izmir International Biomedicine and Genome Institute, with the student ID 2022850020 succeeded her Master's thesis titled "Analysis of Long Non-Coding Rnas in Epithelial-Mesenchymal Transition and Mesenchymal-Epithelial Transition Processes: Comparative Synteny Analysis Between Human and Mouse Genomes" as a result of her thesis defense on 27/08/24.

JURY HEAD

Prof. Dr. Gökhan KARAKÜLAH

JURRY MEMBER
Asst. Prof. Ayse Banu DEMİR

JURRY MEMBER
Asst.Prof. Yavuz OKTAY

T.C.
DOKUZ EYLÜL ÜNİVERSİTESİ
İZMİR ULUSLARARASI BİYOTIP VE GENOM ENSTİTÜSÜ

YÜKSEK LİSANS TEZ SAVUNMA SINAVI TUTANAĞI

Dokuz Eylül Üniversitesi İzmir Uluslararası Biyotıp ve Genom Enstitüsü, Biyotıp ve Sağlık Teknolojileri Anabilim Dalı, Biyotıp ve Sağlık Teknolojileri Yüksek Lisans Programı 2022850020 numaralı öğrencisi Elif YILDIZ, "Analysis of Long Non-Coding Rnas in Epithelial-Mesenchymal Transition and Mesenchymal-Epithelial Transition Processes: Comparative Synteny Analysis Between Human and Mouse Genomes." konulu Yüksek Lisans tezini 27/08/2024 tarihinde yapılan savunma sınavı sonucunda başarılı / ~~başarısız~~ olmuştur.

BAŞKAN
Prof. Dr. Gökhan KARAKÜLAH

JÜRİ ÜYESİ
Dr. Öğr.Üyesi Ayşe Banu DEMİR

JÜRİ ÜYESİ
Dr. Öğr.Üyesi Yavuz OKTAY

TABLE OF CONTENTS

TABLE OF CONTENTS ..... **IV**
LIST OF TABLES ..... **VII**
LIST OF FIGURES ..... **VIII**
LIST OF ABBREVIATIONS ..... **IX**
ACKNOWLEDGMENTS ..... **XI**
ABSTRACT **1**
ÖZET **2**
1. INTRODUCTION **2**
1.1. Statement and Importance of the Problem **3**
1.2. Aim of Study **4**
1.3. Hypothesis of Study **4**
2. GENERAL INFORMATION **5**
2.1. Epithelial-Mesenchymal Transition and Mesenchymal-Epithelial Transition **6**
2.2. Long Noncoding RNAs (lncRNAs) **9**
2.2.1 General Functions and Mechanisms of Action 9
2.2.2. Role of lncRNAs in EMT and MET 14
2.2.3. Conservation of LncRNAs 16
2.3 Comparative Genomics and Synteny Analysis **17**
2.3.1. Human and Mouse Genome Synteny 17
2.3.2. Applications in lncRNA Research 19
3. MATERIALS AND METHODS **24**
3.1. Type of Study **24**
3.2. Time and Place of Study **24**

3.3. Materials of Study **24**

3.4. Variables of Study **25**

3.5. Tools for Data Collection **25**

3.6 Study Plan and Calendar **25**

3.7 Data Evaluation **26**

3.7.1 RNA-seq Data Quality Control and Adapter Sequence Trimming 26

3.7.2 Short Reads Alignment to Reference Mouse and Human Genome 26

3.7.3 Quantification of LncRNA Expression Levels 26

3.7.4 De novo Transcriptome Assembly 27

3.7.5 Transcriptome Comparison to Reference Transcriptome 27

3.7.6 LncRNA Identification 27

3.7.7 Conservation Study with Synteny Analysis 28

3.8. Limitation of the Study **28**

3.9. Ethics Committee Approval **28**

4. RESULTS **29**

4.1. Alignment and Transcriptome Assembly **29**

4.2. Differential Expression Analysis of Mouse and Human Samples **29**

4.3. Identification and Discovery of LncRNAs **38**

4.4. Synteny Analysis of Human lncRNAs on Mouse Genome **46**

4.4.1. Transcript Structure Validation 51

5. DISCUSSION **55**

6. CONCLUSION AND SUGGESTIONS 60
7. REFERENCES 62
APPENDIX 1. 79
APPENDIX 2. 80

LIST OF TABLES

Table 1. Alignment read of each sample of the EMT-MET experiments..... 29
Table 2. Summary of Gffcompare class codes and transcript classification..... 39
Table3. Summary of comparative analysis results..... 53

LIST OF FIGURES

Figure 1. Mechanism of EMT-MET. 7
Figure 2. MET Mechanism. 8
Figure 3. Classification of lncRNAs based on location and strand orientation. 11
Figure 4. LncRNA interactions with target genes 13
Figure 5. Mechanism of MALAT1 lncRNA. 15
Figure 6. Synteny between human and mouse showing conserved gene arrangements on chromosomes 11 and 19. 18
Figure 7. Ortho2align pipeline. 21
Figure 8. Comprehensive analysis of mouse gene expression in EMT-MET processes. 30
Figure 9. Comprehensive analysis of human MCF10A cell line gene expression in EMT-MET processes. 32
Figure 10. Differential gene expression analysis in human MCF7 cell line. 35
Figure 11. Comparative analysis of mouse previously unannotated and known lncRNAs 41
Figure 12. Comparative analysis of human previously unannotated and known lncRNAs 43
Figure 13. Comparative analysis of previously unannotated lncRNAs in human and mouse ortholog gene regions. 44
Figure 14. Previously unannotated lncRNA comparison in mouse and human ortholog gene regions 45
Figure 15. Identification and validation of potential human orthologous lncRNAs from mouse lncRNAs. 47
Figure 16. Comparative genomic analysis of novel lncRNAs in mouse and human genomes. 48
Figure 17. Treemap visualization of previously unannotated lncRNA categories in mouse and human ortholog gene regions. 50
Figure 18. Chord diagram of overlapped previously unannotated human lncRNAs with previously unannotated mouse lncRNAs. 51

LIST OF ABBREVIATIONS

CPAT: Coding Potential Calculator

E-cadherin: Epithelial Cadherin

EGF: Epidermal Growth Factor

EMT: Epithelial-Mesenchymal Transition

FDR: False Discovery Rate

FGF: Fibroblast Growth Factor

GHM: Gene Homology Matrix

GSEA: Gene Set Enrichment Analysis

HGF: Hepatocyte Growth Factor

HSPs: High-scoring Segment Pairs

IGV: Integrative Genomics Viewer

LncRNA: Long Noncoding RNA

logFC: Log Fold Change

MEG3: Maternally Expressed Gene 3

MET: Mesenchymal-Epithelial Transition

MMPs: Matrix Metalloproteinases

N-cadherin: Neural Cadherin

NGS: Next-generation Sequencing

ORF: Open Reading Frames

PDX: Patient-Derived Xenograft

PI3K: Phosphoinositide 3-kinase

NW: Needleman–Wunsch

PRCs: Polycomb Repressive Complexes

RPKM: Reads Mapped Per Kilobase of Transcript Per Million Mapped Reads

SPRR: Small Proline-rich Proteins

TGFβ: Transforming Growth Factor Beta

TIME: TERT-immortalized microvascular endothelial cells

TPR: True Positive Rate

VEGF: Vascular Endothelial Growth Factor

ACKNOWLEDGMENTS

To everyone who helped with the writing of this thesis, I would like to extend my sincere gratitude:

Firstly, to my academic advisors, Prof. Dr. Gökhan KARAKÜLAH and Assoc. Prof. Dr. Hani ALOTAİBİ, for their patience, mentorship, and the invaluable knowledge they imparted;

To each member of the Bioinformatics Platform, who prioritized resolving every question and confusion i had during my thesis work, just as they have throughout my education, and who showed friendship and solidarity beyond the bounds of academia;

To my family, who never lost their trust in me and offered unwavering support and love during this process, and to my partner, who has always been by my side;

And finally, I'd like to thank TÜBITAK for supporting the mouse gathering data and analysis, and to TÜSEB for providing the funds that allowed the replication and further analysis of these studies on human data.

ANALYSIS OF LONG NON-CODING RNAS IN EPITHELIAL-MESENCHYMAL TRANSITION AND MESENCHYMAL-EPITHELIAL TRANSITION PROCESSES: COMPARATIVE SYNTENY ANALYSIS BETWEEN HUMAN AND MOUSE GENOMES

Elif YILDIZ, İzmir International Biomedicine and Genome Institute, Dokuz Eylül University Health Campus, Balcova 35340 - İzmir / TÜRKİYE

ABSTRACT

Epithelial-Mesenchymal Transition (EMT) and Mesenchymal-Epithelial Transition (MET) are fundamental biological processes involved in development, wound healing, and cancer metastasis, regulated by complex molecular networks, including long non-coding RNAs (lncRNAs). Despite significant advancements, the regulatory roles of lncRNAs in EMT and MET and their conservation across species remain poorly understood. This study aims to identify and characterize human lncRNAs involved in EMT and MET processes from MCF7 and MCF10A cell lines and compare their conservation with previously annotated mouse lncRNAs identified in a prior 219Z034 TÜBİTAK project on NMuMG cells using synteny analysis.

In this study, we identified 719 previously unannotated human lncRNAs. Leveraging insights from our previous work, which identified 6,806 unannotated mouse lncRNAs, we uncovered over 30000 syntenic lncRNAs between human and mouse, comprising both known and newly unannotated lncRNAs. Among these, 116 previously unannotated syntenic lncRNAs were identified as potentially associated with EMT and MET.

This comprehensive analysis of lncRNAs enhances our understanding of their contributions to EMT and MET processes and underscores the importance of conservation across species. These findings pave the way for future research into the therapeutic targeting of lncRNAs in cancer and other diseases involving EMT and MET.

Keywords: MET, EMT, lncRNA, Synteny, Comparative Genomics

EMT VE MET SÜREÇLERİNDE UZUN KODLAMAYAN RNA'LARIN ROLÜNÜN İNCELENMESI: İNSAN VE FARE MODELLERİ ARASINDA SİNTENİ ANALİZİ

Elif YILDIZ, İzmir Uluslararası Biyotıp ve Genom Enstitüsü, Dokuz Eylül Üniversitesi Sağlık Yerleşkesi, Balçova 35340 - İzmir / TÜRKİYE

ÖZET

Epitelyal-Mezenkimal Geçiş (EMT) ve Mezenkimal-Epitelyal Geçiş (MET), gelişim, yara iyileşmesi ve kanser metastazında yer alan temel biyolojik süreçlerdir ve uzun kodlamayan RNA'lar (lncRNA'lar) dahil olmak üzere karmaşık moleküler ağlar tarafından düzenlenir. Önemli ilerlemelere rağmen, lncRNA'ların EMT ve MET'deki düzenleyici rolleri ve türler arasındaki korunumu hâlâ tam olarak anlaşılmamıştır. Bu çalışma, MCF7 ve MCF10A hücre hatlarından elde edilen EMT ve MET süreçlerine dahil olan insan lncRNA'larını tanımlamayı ve karakterize etmeyi, ayrıca NMuMG hücre hatları üzerinde yapılan önceki 219Z034 TÜBİTAK projesinde tanımlanan fare lncRNA'ları ile korunumlarını karşılaştırmayı amaçlamaktadır.

Bu çalışmada, 719 daha önce tanımlanmamış insan lncRNA'sı tespit ettik. Daha önceki çalışmamızda tanımlanan 6806 tanımlanmamış fare lncRNA'sından elde edilen bulguları kullanarak, insan ve fare arasında bilinen ve anotasyonu yapılmamış lncRNA'ları içeren 30000'den fazla sintenik lncRNA keşfettik. Bunlar arasından, daha önceden anotasyounu yapılmamış 116 sintenik lncRNA'nın EMT ve MET ile potansiyel olarak ilişkili olduğu belirlendi.

Bu kapsamlı lncRNA analizi, lncRNA'ların EMT ve MET süreçlerine katkılarını anlamamızı artırmakta ve türler arası korumanın önemini vurgulamaktadır. Bu bulgular, lncRNA'ların kanser ve diğer EMT ve MET süreçlerini içeren hastalıklarda terapötik hedeflenmesi konusunda gelecekteki araştırmalara zemin hazırlamaktadır.

Anahtar kelimeler: MET, EMT, lncRNA, Sinteni Analizi, Karşılaştırmalı Genomik

1. INTRODUCTION

1.1. Statement and Importance of the Problem

Epithelial-Mesenchymal Transition (EMT) and Mesenchymal-Epithelial Transition (MET) are critical events in developmental biology and pathogenesis. EMT signifies a fundamental cellular program where epithelial cells undergo a transformation acquiring mesenchymal traits, thereby facilitating migration and invasion. This process is crucial for development, wound healing, and the progression of diseases, such as cancer, where it aids in metastasis. Traditionally, MET has been viewed merely as the reversible counterpart of EMT, playing a key role in tissue regeneration and organogenesis. However, recent research has shed light on MET's distinct regulation by transcriptional networks, revealing its unique biological significance.

It is now understood that both processes are intricately regulated, involving complex molecular networks, including the participation of non-coding RNAs. Despite the advancements, there are significant knowledge gaps in our comprehensive understanding of these regulatory mechanisms. Long noncoding RNAs (lncRNAs) have become important regulators of gene expression and play a significant role in various cellular processes, such as post-transcriptional processing and transcriptional regulation. Their significance in EMT and MET is highlighted by the fact that their dysregulation has been linked to a number of diseases, including cancer and developmental disorders.

The majority of the human genome consists of non-coding elements, yet elucidating the roles of lncRNAs presents challenges that impede a comprehensive understanding of their functions and the transfer of insights between model organisms. The low conservation of lncRNAs between species is a significant barrier to fully deciphering the complexities of EMT and MET processes.

Comparative genomics has yielded invaluable knowledge into the evolution and functional conservation of genetic components among species, reaching beyond the limits of individual gene regulation. A strong foundation for analyzing the functional significance of individual genomic areas is provided by synteny, which is the preservation of gene order and content across related genomes. However, there is

insufficient data on the comparative aspects of lncRNAs in human and mouse models, which limits our ability to understand their conserved roles and regulatory functions across species.

In this study, a comprehensive analysis of lncRNAs related to the EMT and MET process by specifically comparing synteny among human and mouse genomes was conducted. Firstly, our goal was to identify previously unannotated human lncRNAs that could be related to EMT and MET. Subsequently, it was aimed to find orthologous counterparts within the previously unannotated mouse lncRNAs derived from a previous TÜBİTAK study. These were identified as orthologs following the integration of transcriptomic and genomic data by bioinformatics methods. Filling in knowledge gaps using the comparative approach provides valuable information to shed light on the level of conservation of lncRNAs and thus infer detailed information about lncRNA roles in both development and disease pathogenesis, including wound healing.

1.2. Aim of Study

The primary objective of this study is to discover previously unannotated lncRNAs in human models and to establish their orthologous relationships with mouse lncRNAs. This research is based on the findings of a prior TÜBİTAK study, which identified novel lncRNAs in NMuMG cells. We identified and annotated previously unannotated human lncRNAs related to MCF7 and MCF10A cell lines by creating specific analysis pipelines. Synteny was performed to determine the evolutionary conservation of lncRNAs between different species. This approach aims to describe the functional links between human and mouse lncRNAs and make conserved implications understandable within general biological contexts for an enlarged view regarding their role in cellular procedures and disease mechanisms.

1.3. Hypothesis of Study

The primary hypothesis of the study is that specific lncRNAs serve as pivotal regulators of EMT and MET and can be identified through synteny analysis between human and mouse models. It is proposed that some previously unannotated mouse lncRNAs, identified as significant in the MET process in prior studies, might have conserved orthologs in humans.

The first supporting hypothesis is based on the presence of novel lncRNAs found critical for the progression of EMT and MET using mouse data. Thus, similar novel lncRNAs found in human datasets may also play important roles in these processes. The second supporting hypothesis is that the evolutionary conservation of lncRNAs in both human and mouse genomes reveals their functional importance. The conservation of these lncRNAs suggests they perform critical regulatory functions preserved through selective pressure across species.

2. GENERAL INFORMATION

2.1. Epithelial-Mesenchymal Transition and Mesenchymal-Epithelial Transition

From a single fertilized cell, complex organisms like humans and mice evolve through a series of sophisticated processes, including cellular morphology and differentiation, coordinated by similar molecular cues. Cellular differentiation is how individual cells in a developing embryo specialize into multiple cell types with distinctive morphologies and functions. Complex signaling pathways and transcriptional networks that manage gene expression patterns govern cellular development (Talukdar & Chatterji, 2023) . Cells take on distinct morphologies and functions as they differentiate, which are adapted to their tasks in tissues and organs. The complex remodeling of embryo and organ architecture during gastrulation and organogenesis depends on the EMT and MET, which are phenotypic changes between the epithelial and mesenchymal states (H. Y. Kim, Jackson, & Davidson, 2017). These transitions are also important events for the metastasis of many carcinomas (Their, 2002).

Early embryogenesis researchers first defined EMT as "a program with distinct cellular properties" (D. H. Kim et al., 2017; Li et al., 2021). Recent literature defines it as a process involved in the pathophysiology of cancer, tissue fibrosis, morphogenetic alterations during embryogenesis, adult development, and wound healing processes (Yang et al., 2020).

EMT is a flexible and reversible process consisting of many steps that change the characteristic properties of cells (Kalluri, 2009). During this process, the ability of molecules that enable epithelial cells to hold together decreases. Simultaneously, factors such as mesenchymal cadherins, vimentin, and matrix metalloproteinases (MMPs) become more abundant, which promote cells to lose their interconnection properties, that is, to acquire invasive and migratory abilities by reducing cell-cell adhesion and hindering the organization of the cell layer (J. Jiang et al., 2017). The transformation of epithelial cells into different types occurs under the influence of various signals, such as growth factors (Gonzalez & Medici, 2014). These signals include growth factors such as transforming growth factor beta (TGFβ), epidermal growth factor (EGF), vascular endothelial growth factor (VEGF), and fibroblast growth factor (FGF) (Gugnoni & Ciarrocchi, 2019).

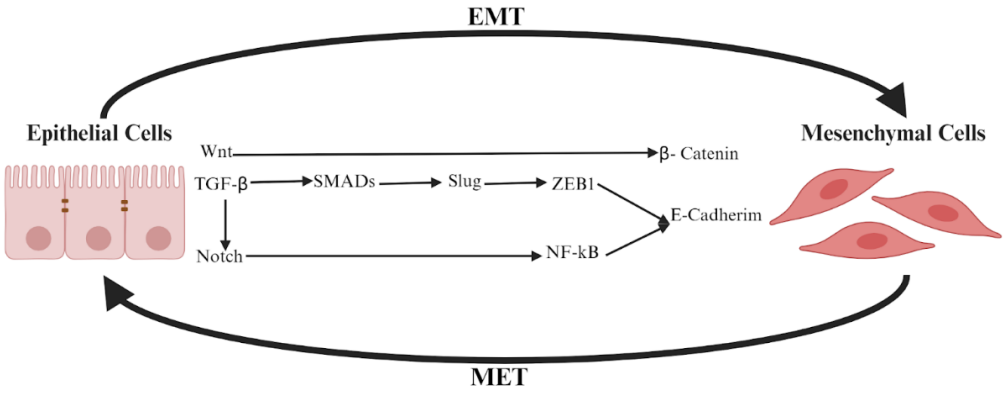

Figure 1. Mechanism of EMT-MET. The figure illustrates Wnt, TGF-β, and Notch signaling pathways involved in EMT.

In particular, TGFβ is the main factor governing the EMT and does so through SMAD-dependent (canonical) or independent (non-canonical) pathways, as shown in Figure 1. TGF-β regulates EMT through the SMAD-mediated pathway by activating transcriptional repressors that control EMT dynamics (J. H. Lee & Massagué, 2022). The phosphorylation and subsequent activation of the SMAD2 and SMAD3 proteins are initiated by the stimulation of TGF-β signaling. Activated SMAD2/3 complexes then migrate to the nucleus and function as transcriptional regulators by forming heteromeric complexes with SMAD4. These SMAD complexes enhance the transcription of significant EMT-associated transcriptional repressors, such as SNAIL, ZEB, and TWIST, within the nucleus (Børretzen et al., 2021). These repressors operate by preventing the expression of E-cadherin, a crucial protein involved in preserving the adhesion of epithelial cells (Hao et al., 2019). Epigenetic and post-translational modifications are triggered by the activation of numerous endogenous signaling pathways, including phosphoinositide 3-kinase (PI3K), wingless/integrated Wnt, Notch, Hedgehog, and specific transcription factors (EMT-TFs) (Lamouille & Derynck, 2011; Song, 2020). These factors combine to profoundly reprogram gene expression in the EMT process.

Since EMT is not permanent, and during cancer metastasis, the reverse of this process, MET, which stops migration to allow cells to fully settle in distant sites and sustain cell proliferating, plays an equally significant role during cancer metastasis (Gugnoni & Ciarrocchi, 2019). MET was previously regarded as the reversible process of EMT, assuming it has a straightforward regulatory mechanism, like the extraction of substances that induce epithelial-mesenchymal transition (Kalluri & Weinberg, 2009; Sistigu et al., 2017).

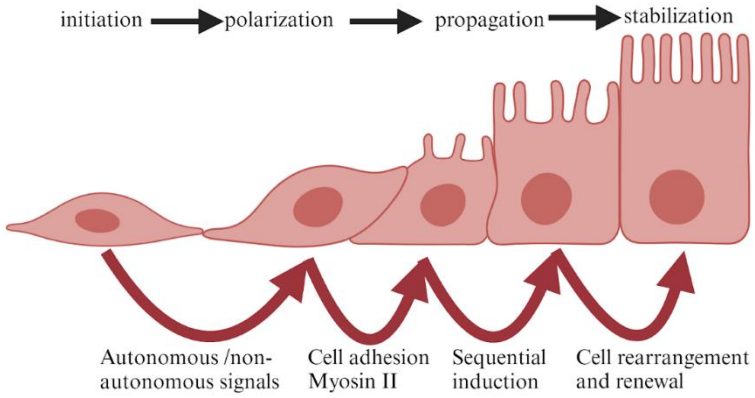

Figure 2. MET mechanism as proposed by Kim, Jackson, and Davidson (2017). The initiation of MET is related to a critical decision modulated by many distinct signals. While exact mechanisms of this decision remain elusive, they are evident in several developmental and pathological contexts such as embryogenesis and cancer metastasis. Successive steps of MET, i.e., polarization, propagation, and stabilization, are intricate processes in which cellular responses to mechanical stimuli, junctional dynamics, and ECM interactions all point toward driving the transition toward an epithelial phenotype, with ramifications for tissue homeostasis and pathology. The explanation provided for each step is derived from their article published in Seminars in Cell & Developmental Biology (H. Y. Kim et al., 2017).

Recent studies have revealed that MET exhibits distinctive and intricate structures, shedding light on its unique functionality and operational mechanisms, wherein transcription factors such as Grhl2/3, Ovol2, and Klf4 are known to play pivotal roles (Alotaibi et al., 2015; Watanabe et al., 2019). As illustrated in Figure 2,

the MET process involves distinct stages such as initiation, polarization, propagation, and stabilization, each driven by a series of complex cellular mechanisms and signaling pathways. MET is part of a process that ensures normality of tissue organization during organ development in embryogenesis and adulthood. Hepatocyte growth factor (HGF), an oncogene, activates a transmembrane tyrosine kinase receptor during the process (Ilangumaran et al., 2016). Various cancer forms have invasive development and progression linked to the HGF/MET signaling pathway. Disruptions in MET have been identified as critical factors in the advancement of cancer, particularly in advanced-stage diseases and in the development of resistance, notably against targeted therapies like trastuzumab (Minuti & Landi, 2015). While recent studies reveal the importance and unique structure of MET, they also reveal the necessity of in-depth studies in this field.

2.2. Long Noncoding RNAs (lncRNAs)

2.2.1 General Functions and Mechanisms of Action

In recent years, belief solely in the functional properties of protein-coding genes or elements has been shattered in light of advancing technologies. Specifically, it was discovered that protein-coding genes account for less than 3% of the human genome through the use of sequencing technologies with high speed and deep sequencing capabilities, such as next-generation sequencing (NGS), along with the utilization of computational methods, highlighted elements such as introns, enhancers, and non-coding RNAs, which constitute the portion outside the coding region and are often referred to as "junk" or "dark matter" (Fagundes et al., n.d.; Hatje et al., 2019; Qin, 2019; Tripathi et al., 2017). Non-coding RNAs, while not directly participating in protein synthesis, assume various regulatory roles by interacting with other molecules in different areas, such as gene expression, epigenetic regulation, and other cellular processes (Chillón & Marcia, 2020). Most non-coding genes can be transcribed into RNAs that do not carry or have little potential for protein synthesis. LncRNAs are RNA molecules that can modify chromatin function, regulate gene expression at both the transcriptional and post-transcriptional stages, alter mRNA stability and translation, and intervene in signaling cascades (Kaikkonen et al., 2011; Ratti et al., 2020). These ncRNAs, characterized by their lengths exceeding 200

nucleotides and lack of protein-coding potential, represent a diverse and heterogeneous group of RNA molecules (Statello et al., 2020).

However, given their variety of functions and regulatory roles in the cell, the description of lncRNAs appears oversimplified. Since NGS technologies have developed such a powerful technique in transcriptomic research as RNA-seq—which enables the investigation of lncRNA expression patterns across a wide range of tissues, cell types, and experimental conditions—they have played a significant role in the accurate identification and quantification of lncRNAs (Choudhury et al., 2024; Satam et al., 2023). RNA-seq improves the knowledge of the non-coding transcriptome by detecting both known and novel lncRNAs using the sequencing depth and precision of NGS tools. Furthermore, *de novo* transcriptome assembly is supported by NGS technologies, which is a crucial step for discovering novel lncRNAs that are absent from existing databases (X. Li et al., 2016). The perspective on non-coding transcriptomes is expanded by the finding of previously unidentified lncRNA transcripts made possible by this NGS approach. Furthermore, a better understanding of the functional relevance of lncRNAs and the unraveling of the complex regulatory networks in which they participate can be achieved by integrating NGS data that illuminates their potential roles in gene regulation, cellular processes, and disease pathogenesis. The extraordinary diversity of lncRNAs in terms of their mode of action and interaction partners makes their classification more difficult than that of protein-coding genes, which have distinct structural and functional patterns guiding their classification.

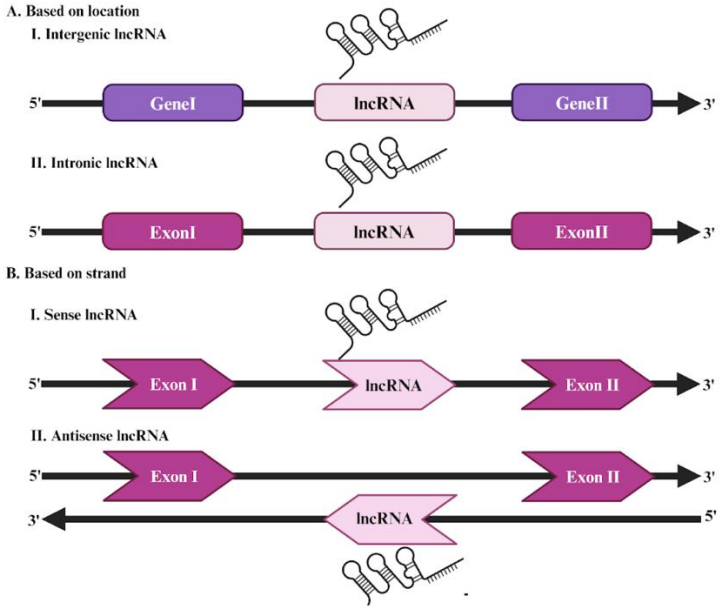

Figure 3. Classification of lncRNAs Based on Location and Strand Orientation. A. LncRNAs are classified based on whether they are transcribed between protein-coding genes or in protein-coding gene introns. B. LncRNA classification based on whether they are transcribed from the same or opposite strand to protein-coding genes

LncRNAs can be classified according to their position and strand orientation in relation to protein-coding genes, as shown in Figure 3. In Figure 3A, Gene I and Gene II are adjacent protein-coding genes between which the intergenic lncRNAs are transcribed (B. Liu et al., 2021). For intronic and strand-specific lncRNAs, Exon I and Exon II are the exons flanking the lncRNA, as depicted in both Panels A and B of Figure 3. Transcription direction is marked by arrows, that denote the sense or antisense direction of transcription of lncRNAs in Figure 3B.

A major challenge in classifying lncRNAs is the variety of functions they perform inside the cell. Interestingly, lncRNAs can function locally at the transcription site and widely throughout the genome, operating in both cis and trans through interactions with other RNA, proteins, and genomic DNA (W. Ma et al., 2021). LncRNAs linked to cancer display many functions in the nucleus via different pathways (Martino et al., 2021). Notably, some lncRNAs are overexpressed in cancer cells, especially in primary and metastatic breast cancer, and are associated with

aggressive disease development (Jin et al., 2021). These lncRNAs include HOTAIR, which is implicated in global epigenetic and transcriptome dysregulation (Fatima et al., 2015). Changes in lncRNAs can also result in gain-of-function activities that disturb the balance of gene expression, the epigenetic landscape, and other critical processes in cancer cells. These effects significantly impair cellular homeostasis and cause the expression of cancer hallmarks (Ahmad et al., 2023).

Furthermore, lncRNAs can act independently of their RNA products by coordinating the formation of massive ribonucleoprotein complexes that affect the transcriptome landscape (Aliperti et al., 2021). These non-product-dependent actions of lncRNAs show that they can act as scaffolds or guides for the recruitment of transcription factors, chromatin-modifying enzymes, and other regulatory proteins. Unlike more traditional methods like microarrays, RNA-seq is a potent NGS technology that offers a thorough profile of RNA and enables accurate quantification (Rao et al., 2019). LncRNAs can be precisely categorized using RNA-seq by considering their orientation and position in relation to neighboring protein-coding genes. The regulatory mechanisms controlling lncRNA function are clarified by this classification. LncRNAs can be categorized into two primary classifications based on their chromosomal position in regard to the closest protein-coding gene: lncRNAs overlap with protein-coding gene loci as genic and intergenic. LncRNAs can be further classified into sense versus antisense subgroups or intronic versus exonic, overlapping an exon by at least 1 bp (Kern et al., 2018), subgroups based on their orientation and position with relative to the overlapped protein-coding gene (Y. Zhao et al., 2020).

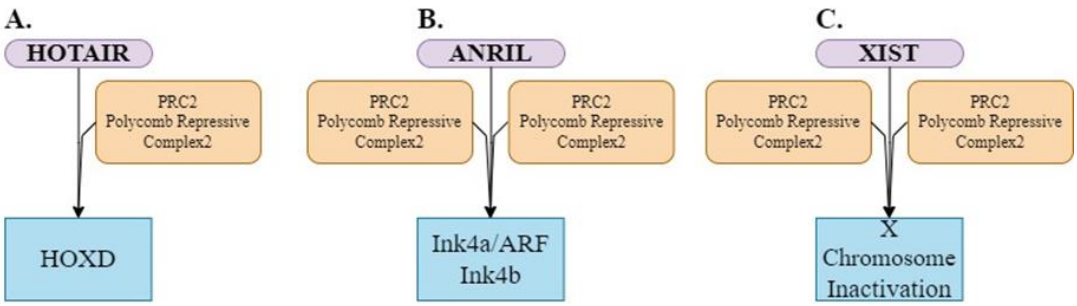

Figure 4. LncRNA interactions with target genes. This figure illustrates the interactions of three lncRNAs, HOTAIR (A), ANRIL (B), and XIST (C), with Polycomb Repressive Complexes (PRCs) and their associated target genes.

The regulatory diversity of lncRNAs is vast and complex in the regulation of transcription and translation through interactions with chromatin-modifying complexes, mRNAs, and transcription factors (Zhang et al., 2019). For example, HOTAIR lncRNA interacts with PRC2, which causes the repression of the HOXD gene region (Figure 4A) (Portoso et al., 2017). Such an interaction leads to the silencing of gene transcription, thus showing how lncRNAs can change the expression of genes by altering chromatin states. Another significant illustration is XIST lncRNA, which controls X chromosomal inactivation. XIST recruits PRC2, leading to a decrease in transcription levels of genes along the X chromosome, thus enabling dosage compensation in females (Figure 4C) (Brockdorff, 2017). ANRIL lncRNA represses transcription of the INK4a/ARF/INK4b gene cluster by recruiting both PRC1 and PRC2 complexes, causing the silencing of these tumor suppressor genes (Figure 4B) (Long et al., 2017).

Since they provide a plethora of information and enable the research of antisense lncRNAs, online databases are essential resources for researching lncRNAs (B. Liu et al., 2021). Two databases stand out in this regard: NONCODE (L. Zhao et al., 2021) and LncRNADisease (Bao et al., 2019). An extensive knowledge database including many ncRNA types is called NONCODE. It makes the study of lncRNA functions in cancer easier by providing a vast collection of human lncRNAs and data on lncRNA-cancer correlations. LncRNADisease focuses on the connections between lncRNAs and diseases, accumulating data supporting these connections through experimental verification. Due to a confidence score system for evaluating

dependability, there is a significant increase in lncRNA-disease links (Biyu et al., 2023; J. Ye et al., 2020).

Although lncRNAs have a wide range of functions, from chromatin organization to gene regulation, it is still very difficult to fully understand these tasks. To completely understand the functional diversity of lncRNAs, computational approaches are limited because the biology of these molecules is complex. The complexities of lncRNA function continue to provide major challenges despite advances in bioinformatics and computational tools, highlighting the necessity for sophisticated approaches to solve their riddles.

2.2.2. Role of lncRNAs in EMT and MET

The mechanisms of EMT and MET are mainly contributed to by lncRNAs, expanding their importance beyond general gene regulation and cellular activities. These alterations are significant in a variety of physiological and pathological circumstances but are particularly crucial in the genesis and spread of cancer.

Within these lncRNAs, one that particularly stands out is MALAT1, which is implicated in several aspects of cancer biology, including cell proliferation, migration, invasion, and metastasis, operating through the regulation of gene expression at both transcriptional and post-transcriptional levels (Arun et al., 2020). Additionally, as seen in Figure 5, MALAT1 regulates apoptosis by interacting with the PI3K/AKT pathway, metastasis by small proline-rich proteins (SPRR), and the cell cycle by miR-328 (Lin et al., 2022; D. Ye et al., 2021a). Specifically, it promotes cell proliferation by interacting with specific proteins and modulating signaling pathways crucial for cell cycle progression. Moreover, MALAT1 enhances cell migration and invasion by regulating genes associated with EMT, a process pivotal for cancer metastasis (Hussein et al., 2024; Thapa et al., 2024). The function classification of lncRNAs in EMT-MET interactions is further complicated by their participation in the formation of dynamic regulatory complexes. Additionally, MALAT1 has been linked to promoting metastasis by influencing the expression of genes involved in cancer cell invasion into surrounding tissues and dissemination to distant sites (Gutschner et al.,

2013). Dysregulation of MALAT1 in cancer cells can instigate aberrant cellular behaviors, contributing to tumor progression and metastasis (D. Ye et al., 2021).

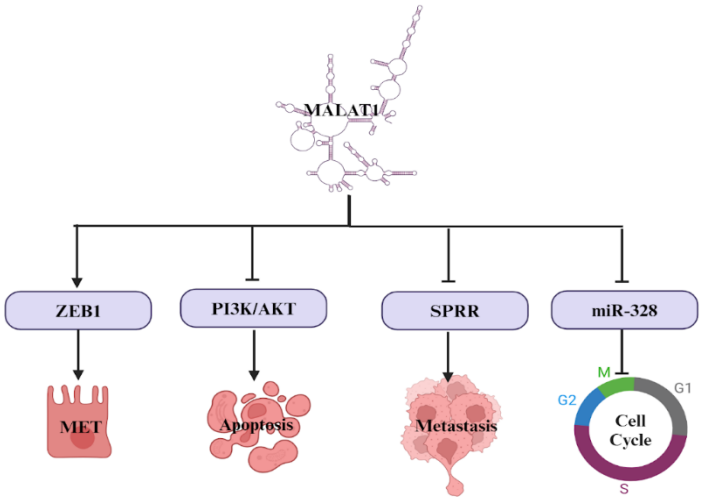

Figure 5. Mechanism of MALAT1 lncRNA. MALAT1 binds to ZEB1, regulating MET, which plays a critical role in the development of cancer and metastasis. Also, it extends its function to modulate the PI3K/AKT pathway, a key regulator of apoptosis, hence controlling cellular survival and proliferation. Furthermore, MALAT1 binds to SPRR, outsourcing cancer cells to faraway organs. In addition, alteration in the expression of miR-328 and provision for cell division and growth control is done by MALAT1.

Furthermore, lncRNAs generated by oncogenic stress frequently display chromatin association and can regulate transcription. For example, LincRNA-Gadd45γ is crucial for activating nearby cell cycle and apoptosis factors, indicating its ability to regulate transcription within the p53 tumor suppressor pathway (Tesfaye et al., 2021). Furthermore, lncRNA-p21 and Pvt1b, two examples of lncRNAs in the p53 transcriptional network that regulate local gene expression, respectively, activate or repress nearby genes in the p53 pathway (Amirinejad et al., 2020; Hall et al., 2015). Additionally, both trans-acting and cis-acting lncRNAs exhibit the mechanism of lncRNA-binding partner sequestration or redistribution in response to changes in lncRNAs, which impacts the epigenetic landscape and balance of gene expression in cancer cells (Winkler & Dimitrova, 2022).

Maternally Expressed Gene 3, MEG3, is a well-researched example of this adaptability, showcasing the multiple regulatory functions typical of numerous lncRNAs (Ghafouri-Fard & Taheri, 2019; Xu et al., 2022). Within the framework of cis-regulation, MEG3 functions by managing imprinting, an occurrence in which the parent of origin influences the expression of a gene. By directly interacting with particular genomic areas and modifying the expression of neighboring genes, MEG3 governs imprinting in cis (Mondal et al., 2015). MEG3 can adjust gene expression patterns in a spatially limited way thanks to this localized control. In contrast, MEG3 regulates the p53 stress response, indicating trans-regulatory actions as well. In this manner, MEG3 affects the expression of genes associated with the p53 pathway at a distance from its transcriptional location (Uroda et al., 2019). MEG3 can affect gene expression throughout the genome by regulating the activity of transcription factors or other regulatory molecules. This can have an impact on cellular processes like apoptosis, cell proliferation, and differentiation (Ferrer & Dimitrova, 2024).

2.2.3. Conservation of LncRNAs

LncRNAs are central to genomic regulation and conserved at multiple levels. There are generally four primary dimensions to the conservation of lncRNA, as argued by Diederichs, 2014: sequence, structure, function, and expression from syntenic loci. Each dimension makes a different contribution and offers different challenges to the lncRNA studies. Sequence conservation of lncRNA is less common than in the coding genes and does so normally only in specific regions (Kurian et al., 2015). This selective conservation thus suggests that only subsets of nucleotides in a lncRNA might be functional and makes it difficult to decide which elements are conserved across species. Indeed, this low sequence similarity often hampers traditional comparative genomic analyses, requiring the design of innovative computational tools and methodologies.

LncRNAs mostly fold into complex secondary and tertiary structures, which play a vital role. Such structures may be conserved, though the nucleotide sequences are quite different in nature, hence underlining the importance of base-pairing properties. However, such structures cannot be easily predicted or compared among species. This is because there are only a few annotated structures of lncRNAs, while

RNA folding itself is a computationally challenging task. Tools for the prediction of secondary structures sometimes lack accuracy; hence, performing studies of structural conservation becomes quite challenging (Chillón & Marcia, 2020).

Despite rapid sequence evolution, lncRNAs seem to conserve their functional roles. Functional conservation can be associated with either sequence or structural conservation, but at the same time, but it can very ocuur independently of these two factors. For example, evidence shows that some lncRNAs associated with development and differentiation, gene regulation, and other functions show functional conservation without sequence conservation (Camilleri-Robles et al., 2022). For example, this functional conservation includes the role of MALAT1 in regulating metastasis-associated genes and NEAT1 in transcriptional regulation and RNA splicing events (Weghorst et al., 2024).

In addition to these aspects, Synteny is the preservation of the arrangement of genes on the chromosomes of different species. In fact, most lncRNAs seem to be transcribed from syntenic loci. Here, genomic context is conserved and defines neighboring gene order. Such a transcriptional context would already be functionally meaningful, even if the sequences of lncRNAs.

2.3 Comparative Genomics and Synteny Analysis

2.3.1. Human and Mouse Genome Synteny

Evolutionary phenomena such as polyploidization, gene duplication or deletion, and adaptive mutation selection are all observed in the genomes of eukaryotic organisms. Within as well as between clades, comparative genomics provides a framework for understanding the dynamics and interactions affecting biological fitness in various species (Paril et al., 2023). The increasing availability of DNA and RNA sequencing has made comprehensive comparative genomics across species possible; nevertheless, the universality and transfer of workflows are hindered by the absence of consistent and reliable research pipelines.

John Renwick utilized the term "synteny" in 1971 to refer to gene loci on the same chromosome, regardless of the presence of genetic linkage evidence to bolster

the association (Renwick, 1971). Synteny, as used in comparative genomics, refers to the conserved gene arrangement on chromosomes that results from a common ancestor shared by related species (Duran et al., 2009). Understanding the evolutionary links of species is aided by synteny analysis, which allows predecessors to exchange information. Based on the known function of a related gene in another species, it helps anticipate the function of a gene in one species. It also contributes to the identification of regulatory elements and non-coding regions. Comparing several species makes it easier to find possible candidate genes for particular diseases in the context of human genetics.

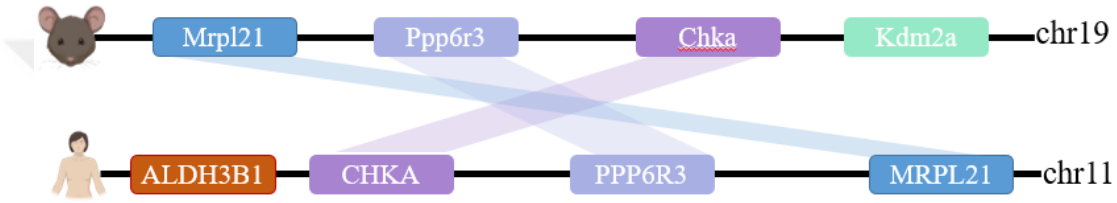

Figure 6. Synteny between human and mouse showing conserved gene arrangements on chromosomes 11 and 19.

For instance, synteny analysis between human and mouse, as illustrated in Figure 6, generally validates the orthologous relationship inferred from sequence similarity or functional conservation.

In addition, synteny analysis can reveal the organization of genes and chromosomes in the genome and provide insights into evolutionary relationships (Zheng et al., 2005). Besides, synteny analysis gives a presentation of the structural architecture of genes and chromosomes and thereby promotes knowledge of evolutionary relationships. Serving as a point of reference in identifying genes that have conserved their relative positions on chromosomes across species, synteny analysis also underlines the evolutionary conservation of genomic architecture. The advancements in NGS technology and the corresponding increase in data, computational analysis capabilities have improved in methodology within conservation studies. Data derived from techniques such as RNA-seq enable the identification of orthologous relationships between organisms by following key steps,

including pairwise genome alignment, synteny block detection, and visualization. In addition, synteny databases may also be used to confirm previously unannotated orthologous elements. Software programs such as BLAST and LASTZ (Harris Robert S., 2007) are used for sequence comparison and whole-genome alignment, while software tools such as Cyntenator (Rödelsperger & Dieterich, 2010) and SyMAP (Soderlund et al., 2011) are applied in the identification of conserved syntenic blocks across multiple species genomes. The outcome can be plotted using tools such as SyntenyPlotteR (Quigley et al., 2023) and Synima (Farrer, 2017), as well as other packages developed for several programming languages.

2.3.2. Applications in lncRNA Research

Synteny analysis, a comparative genomics approach, plays a pivotal role in identifying conserved lncRNAs and elucidating their functions. By examining the conserved arrangement of genes and genomic areas across different species, synteny analysis enables the identification of lncRNAs conserved through evolutionary processes (Bakhtiarizadeh & Salami, 2019; Pegueroles et al., 2019). This conservation suggests functional significance, as essential regulatory elements tend to be maintained across diverse taxa.

Through synteny analysis, researchers can pinpoint lncRNAs that not only share sequence similarity but also retain their genomic context, including neighboring genes and regulatory elements (Mallory & Shkumatava, 2015). This contextual conservation provides insights into the potential regulatory networks and interactions involving lncRNAs. For instance, conserved lncRNAs may be involved in crucial regulatory pathways, influencing gene expression, chromatin structure, and cellular processes (Gandhi et al., 2021).

LncRNAs create a challenge to the identification of conservation areas because of their rapid rates of evolution, variability in length and secondary structure, and lower levels of sequence conservation (Waters et al., 2021). In addition to their complex structure, traditional methods based on genes that code for proteins may not be directly applicable to the study of lncRNA synteny because lncRNAs do not code for proteins. Consequently, the analysis of lncRNAs cannot be done using software

that is normally used to determine the evolutionary links between genes and transcripts that encode proteins.

The computational pipeline developed for the conservation study on lncRNAs, known as lncEVO (Bryzghalov et al., 2021), distinguishes itself from current methods in identifying orthologues of novel lncRNAs through adopting an integrated pipeline approach, encompassing transcriptome assembly, lncRNA prediction, and conservation study modules, facilitating seamless analysis across species from transcript identification to evolutionary conservation assessment. The first module performs transcriptome assembly from paired-end RNA-seq data. Using splice-aware software such as StringTie, RNA-seq reads were aligned to a reference genome (Pertea et al., 2015). The second phase involves predicting lncRNAs using unique criteria distinguishing mRNAs from noncoding RNAs. Based on the sequences of known protein-coding genes, tools such as the Coding Potential Calculator (CPAT) are used to determine if a transcript has the ability to encode proteins based on sequence features (L. Wang et al., 2013). This distinction is important because mRNAs act as templates for protein synthesis, whereas lncRNAs are characterized by their inability to code for proteins (Y. B. Zhao & Krishnan, 2014). In the transcriptome assembly process, lncEvo can efficiently distinguish between mRNAs and noncoding RNAs by evaluating the coding potential of transcripts and identifying those that are unlikely to encode proteins. The last module compares the lncRNA transcriptomes of two target species for conservation research on a genome-wide scale. This implies searching for orthologs, which are lncRNAs in different species that share a common ancestor and have similar functions or sequences, and evaluating the conservation features of novel lncRNAs, which provides insights into the evolutionary dynamics and functional significance of these noncoding transcripts. Furthermore, LncEvo's performance has been validated for detecting conserved lncRNAs between humans and mice, demonstrating its efficacy in efficiently identifying orthologs. Overall, LncEvo is presented as a comprehensive and user-friendly tool for exploring the functional roles of lncRNAs across species, offering valuable insights into evolutionary conservation (Bryzghalov et al., 2021).

Ortho2align presents a distinctive approach to identifying orthologues of novel lncRNAs, differing from existing methods through several key features. Ortho2align combines synteny-based analysis with statistical sequence conservation assessment by taking a sensitive approach rather than searching for a unique lncRNA ortholog, which strategically leverages the syntenic relationships between genomic regions to pinpoint orthologous transcripts (Mylarshchikov & Mironov, 2022) as seen Figure 7. Using liftOver, it first lifts lncRNA coordinates from the genome of the query species to the genome of the subject species. Following that, syntenic areas are built using the lifted coordinates, accounting for permitted duplications and the minimum match criteria. Ortho2align improves computational efficiency and lowers the number of false matches by restricting the search space to the limited number of transcripts encoded within homologous syntenic regions.

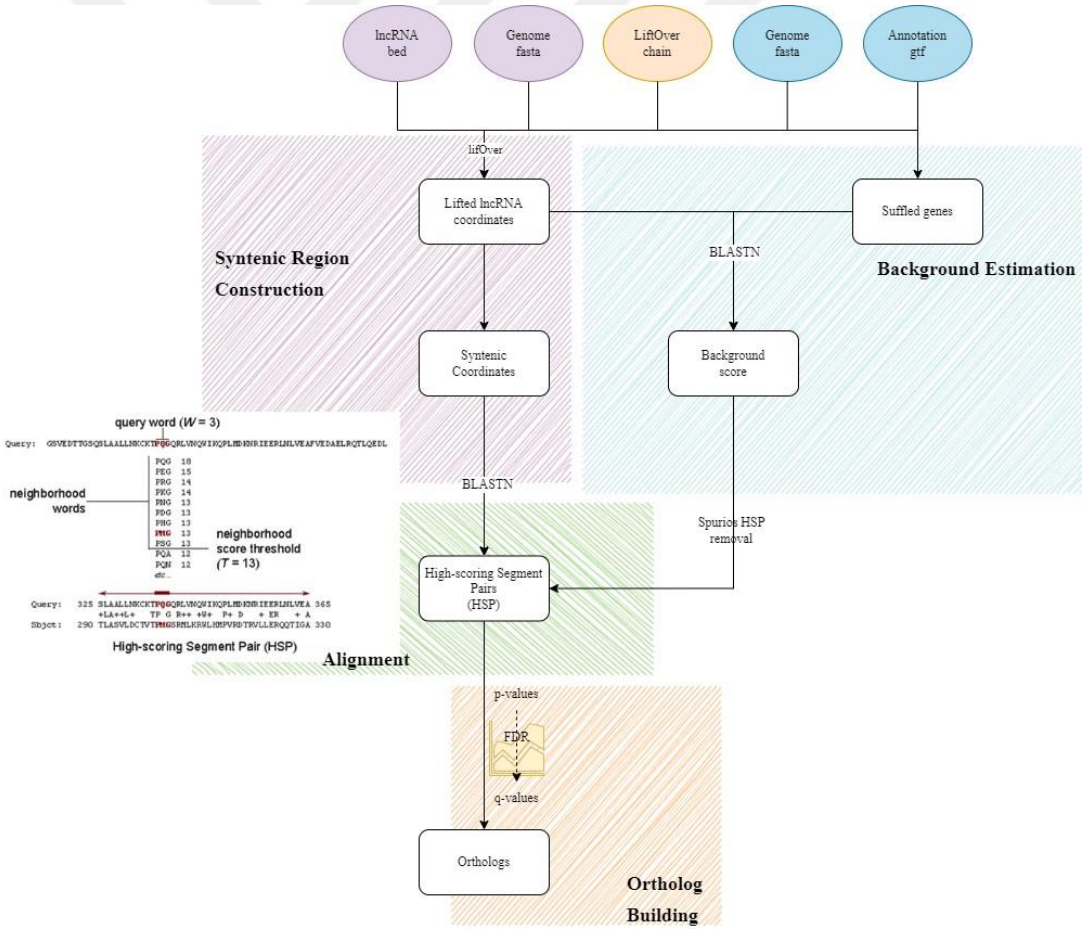

Figure 7. Ortho2align pipeline. Purple ellipses represent query files. Blue ellipses represent subject files.

A set number of nucleotides surrounds the created syntenies, and duplications within a given distance are combined. Flanked lncRNAs are aligned to their syntenic areas using techniques such as BLASTN, which is used with loose parameters. This procedure creates an alignment that yields a list of high-scoring segment pairs (HSPs) that show potential sequence similarities between the subject species genomic regions and the query lncRNAs. Ortholog prediction uses sequence conservation and statistical values of alignments. It gives a comprehensive conservation result using the elevated coordinates of all significant orthologs, non-significant orthologs, and unaligned lncRNAs.The tool considers the alignment scores, sequence similarities, and other alignment parameters to determine the degree of conservation between the lncRNAs across species. The tool uses statistical filtering of HSPs to select orthologues based on their statistical significance, reducing the likelihood of false positives in the predictions. High sequence similarity regions between two sequences aligned during sequence comparison serve as markers of possible homology between genomic areas or gene sequences because of their notable sequence identity and alignment length. The method improves the reliability and precision of orthology predictions by using statistical filtering of HSPs to ensure that only alignments with a high degree of similarity and statistical significance are considered as possible orthologues.

Furthermore, Ortho2align provides optional annotation features for the orthologues that have been found, providing researchers with additional knowledge regarding the conservation status of the loci and the possibility of paralogues. This feature complements orthology investigations and enriches the output. Above all, Ortho2align is an exclusive tool for identifying unannotated orthologues of newly discovered lncRNAs across species for whom there are no lncRNA annotations or consistent experimental evidence. This unique feature sets Ortho2align apart from other tools that focus primarily on cross-species matching of pre-existing lncRNAs.

i-ADHoRe (Proost et al., n.d.), another tool developed for genomic homology and collinear region identification, which refers to regions where gene content and gene order are preserved in the genomes of different species, sheds light on genome evolution and organization. Identifying homology begins by mapping tandem duplicated genes onto a single representative gene, simplifying detection. A Gene

Homology Matrix (GHM) is then generated for each pair of chromosomes or scaffolds, marking homologous gene pairs with dots and displaying collinear regions as diagonals. Statistical evaluation is performed on these diagonals to ascertain their significance as true collinear regions. Significant collinear regions are aligned into profiles, serving as probes for detecting additional collinear regions. This iterative process continues, updating profiles and identifying new collinear regions as long as new homologous regions are found, which are grouped into multiplicons. Profile search, despite its computational cost, enhances the detection of degenerate genomic homology by aligning homologous segments into profiles to uncover more collinear regions. An accurate alignment algorithm, initially utilizing the pairwise Needleman–Wunsch (pNW) algorithm, is crucial for sensitive discovery of collinear (X. Jiang et al., 2017). Based on the scoring scheme, which defines the score for gaps, matches, and mismatches, the algorithm ensures that the aligned sequences have the highest possible similarity score.

No matter which algorithm is used in synteny analysis, the quality of the input files is essential. High-quality annotations are essential when using synteny analysis to find orthologous lncRNAs between species. These annotations provide detailed information about the chromosomal loci of lncRNAs in many species, which makes it easier to identify conserved transcription patterns and the connections between lncRNAs (Wojciech Szczé sniak et al., 2021). Robust annotations allow researchers to construct meaningful comparisons and draw legitimate inferences about the evolutionary conservation of lncRNAs across species, even in cases with limited sequence similarity. The expression of syntenic lncRNAs, which exhibit conserved transcription with low sequence similarity, may have functional importance. This occurrence implies that, in addition to sequence homology, the expression profile itself may have a significant impact on lncRNA function.

3. MATERIALS AND METHODS

3.1. Type of Study

The study adopts an analytical and computational methodology.

3.2. Time and Place of Study

The study was conducted at İzmir Biomedicine and Genome Institute, Bioinformatics Platform (BIP), from July 2023 to June 2024.

3.3. Materials of Study

Raw RNA-seq reads were obtained from the human MCF7 breast cancer cell line, including three samples: uninduced (untreated), induced (doxycycline-treated), which represents EMT, and deinduced (doxycycline treatment removed), which represents MET. Additionally, data were collected from the MCF10A cell line, an immortalized cell line derived from breast adenocarcinoma, with two control samples and two samples treated with TGFB. For mouse lncRNAs used in synteny analysis, data were sourced from NMuMG cells in a time course experiment spanning 72 hours of EMT and 72 hours of MET as part of the TÜBİTAK project which includes four mouse samples: two controls and two TGFB-treated.

In the final stage of the analysis, the following human datasets were used to increase sample number in the study: three control samples from the TERT-immortalized microvascular endothelial cells(TIME) treated similarly to experimental samples which are KSHV-infected samples without actual viral infection from GEO accession number GSE84237 (Sychev et al., 2017), three samples from the HMEC-1 cell line, developed as a model of human dermal microvascular endothelial cells by transfecting with the SV40 large T antigen from GSE244042 (Lisec et al., 2024), and finally, seven wild-type genotype samples of the H1-H9 human embryonic stem cell lines, which were derived from the inner cell mass of blastocysts from GSE203207 (Hossain et al., 2022).

3.4. Variables of Study

The time points of the RNA-seq reads for the cell lines MCF7 and MCF10A and the cell model NMuMG for EMT and MET are independent variables. Furthermore, dependent variables are recorded expression values of the detected transcripts and the undetermined transcripts lncRNAs. In particular, for the MCF7 cell line, the independent variables include the three samples. The independent variables for the MCF10A cell line comprise two control samples and two samples treated with TGFB. The independent variables were measured with the NMuMG cancer model, which is the time course of 72 hours of EMT and 72 hours of MET, with a total of four samples: before and after the end of the study for the two controls, and the two TGFB-treated individuals.

3.5. Tools for Data Collection

A High-Performance Computing cluster, composed of various specialized nodes designed to handle intensive computational tasks efficiently, was utilized in our study. One of the nodes used in this study is Compute Node01, which runs on an Intel(R) Xeon(R) Gold 6148 processor at a speed of 2.40 GHz. It features 40 CPU cores and 80 threads, providing significant parallel processing capabilities. Each user is allocated 2TB of storage in the /archive/username directory. The network infrastructure supporting the compute nodes is based on the Infiniband technology and hence allows a high-speed network at 100 Gbit/s.

3.6 Study Plan and Calendar

Jul- Nov 23	Alignment and transcriptome construction
Jan 24	Identification of previously unannotated lncRNAs from MCF7 and MCF10A cell lines in humans
Mar-May 24	Synteny Analysis
June 24	Thesis writing phase

3.7 Data Evaluation

3.7.1 RNA-seq data quality control and adapter sequence trimming

FastQC was used to do quality filtering on the MCF10A and MCF7 raw sequencing data in order to remove low-quality reads. Adapter sequences and low-quality sections were then eliminated with the Trimmomatic program (Bolger et al., 2014). This procedure aimed to improve the overall quality of sequencing data for use in subsequent analysis.

3.7.2 Short Reads Alignment to Reference Mouse and Human Genome

The genome sequences of GRCh38.p14 in FASTA format and the annotation file in GTF format were obtained from the Gencode website (Frankish et al., 2023). The human genome was indexed using the fast and accurate HISAT2 alignment algorithm (hisat2-build -p) to the obtained genome FASTA files (D. Kim et al., 2019). Subsequently, each file was aligned using the generated index file (hisat2 --dta -x -U -p 6; --dta –S).

Using samtools to convert the resultant SAM files to BAM files and sort them made interacting with high-throughput sequencing data more efficient (samtools view -b -o -@ 6 ; samtools sort -o -@ 6) (H. Li et al., 2009).

3.7.3 Quantification of LncRNA Expression Levels

Using the edgeR and rsubread tools in the R statistical computing environment, differential expression analysis was performed on the transcripts acquired after filtering and the previously unannotated lncRNAs (Liao et al., 2019; Robinson et al., 2010). The filtering parameters included setting a biological coefficient of variation to 0.2 for dispersion estimation and identifying differentially expressed genes based on a log fold change (logFC) threshold greater than 0.6 for upregulated genes and less than -0.6 for downregulated genes, with a false discovery rate (FDR) cutoff of 0.05. Using the Integrative Genomics Viewer (IGV) desktop program, the transcripts from the experiment in which differential expression was observed were shown on the reference chromosomes. To create this graphic, the gene annotations of lncRNA genes from the

Gencode GTF file were compared with the reference genome to confirm their alignment with the lncRNA loci on the reference chromosomes. This procedure aimed to confirm whether the annotated transcripts correspond to the lncRNA loci accurately.

3.7.4 *De Novo* Transcriptome Assembly

The StringTie tool was utilized to assess the RNA-seq alignments produced for each sample in the previous step (Pertea et al., 2015). This technique provides high speed and efficiency in possible transcript assembly. By comparing the BAM files to the ENSEMBL reference, transcripts were produced (stringtie -G -o -p 6). The StringTie merge function (stringtie --merge --fr-G -o -p 6) was used to combine the outputs generated for each sample into a single transcriptome.

3.7.5 Transcriptome Comparison to Reference Transcriptome

The transcriptome is subsequently compared to human reference annotations using GffCompare, which is used for analyzing the transfrags of assemble, and transcripts belonging to the class codes u, x which represent exonic overlap on the opposite strand and unknown, intergenic transcripts, are filtered as potential lncRNA transcripts for future steps.

3.7.6 LncRNA Identification

In order to identify lncRNAs, transcripts with a length of 200 nucleotides or more were first chosen. To find potential coding areas, the transcripts were subsequently processed using the TransDecoder tool to find ORFs with a minimum length of 100 amino acids. Using the CPAT, transcripts with coding potential were evaluated further; those with coding potential values of at least 0.364 for human was eliminated. Following these filtering stages, the transcripts' validity was tested for homology to the Swiss-Prot database using a local Blast tool with an E-value cutoff of 0.01. Comparisons of the genomic context for identified lncRNAs in human and mouse were made using the UCSC Genome Browser (Karolchik et al., 2004). By investigating the alignment, the similarities and differences in genomic context for the lncRNAs between mouse and human are identified, explaining their evolutionary conservation and functional significance.

3.7.7 Conservation Study with Synteny Analysis

The ortho2align tool, which is designed for lncRNA ortholog identification, was used to perform synteny analysis on custom mouse and human GTF files annotated with novel lncRNA information. These custom GTF files were generated to be used as ortho2align input, allowing for the identification of potential orthologous novel lncRNAs between human and mouse. In order to optimize the True Positive Rate (TPR) during the synteny analysis, the min_ratio parameter was set to 0.05 and the p-value threshold parameter to 0.2.

3.8. Limitation of the Study

The results are constrained by the analysis software's algorithms and the transcriptome content of the cell model used. Furthermore, as the results are predictive in nature, they must be confirmed through experimental research in biological systems.

3.9. Ethics Committee Approval

This study commenced following the receipt of ethical approval code 2023-022.

4. RESULTS

4.1. Alignment and Transcriptome Assembly

The alignment files generated by HISAT2 were examined for mapped alignment percentages. The lowest alignment rate, with an average of 90.92%, was determined to be 86.21%, while the highest was found to be 95.06%, confirming that each sample met acceptable standards, which is alignment rates of 80-90% for well-assembled reference genomes and high-quality RNA-seq data. The read counts and percentages, individually analyzed for each sample, have been shown in Table 1.

Table 1. Alignment read of each sample of the EMT-MET experiments.

Sample	Mapped Read Count	Mapped Read Percentage(%)
MCF7_Uninduced	102719897	95.06
MCF7_Induced	99898477	94.19
MCF7_Deinduced	96147550	94.96
MCF10_Control_1	103536843	86.50
MCF10_Control_2	28484126	92.30
TGFB_1	73736751	89.99
TGFB_2	72179986	88.23
Average	6890804	91.60

The alignment statistics presented in Table 1 confirm the high quality of the RNA-seq data, providing an appropriate base for further differential expression and functional analyses.

4.2. Differential expression analysis of mouse and human samples

The differential expression analysis of human cell lines MCF7 and MCF10A and mouse data were already conducted with the TÜBİTAK 219Z034 coded project,

was performed; from that differential expression analysis, some alterations in the lncRNA expression level during EMT-MET transition were determined. Only 5 genes were upregulated in MCF7 samples. On the other hand, MCF10A samples had 585 upregulated genes and 364 genes which were downregulated, while mouse data presented 194 upregulated and 148 downregulated lncRNAs. The figure below shows a heatmap of changes between the control and treatment for the top 30 upregulated genes and the top 30 genes which were downregulated.

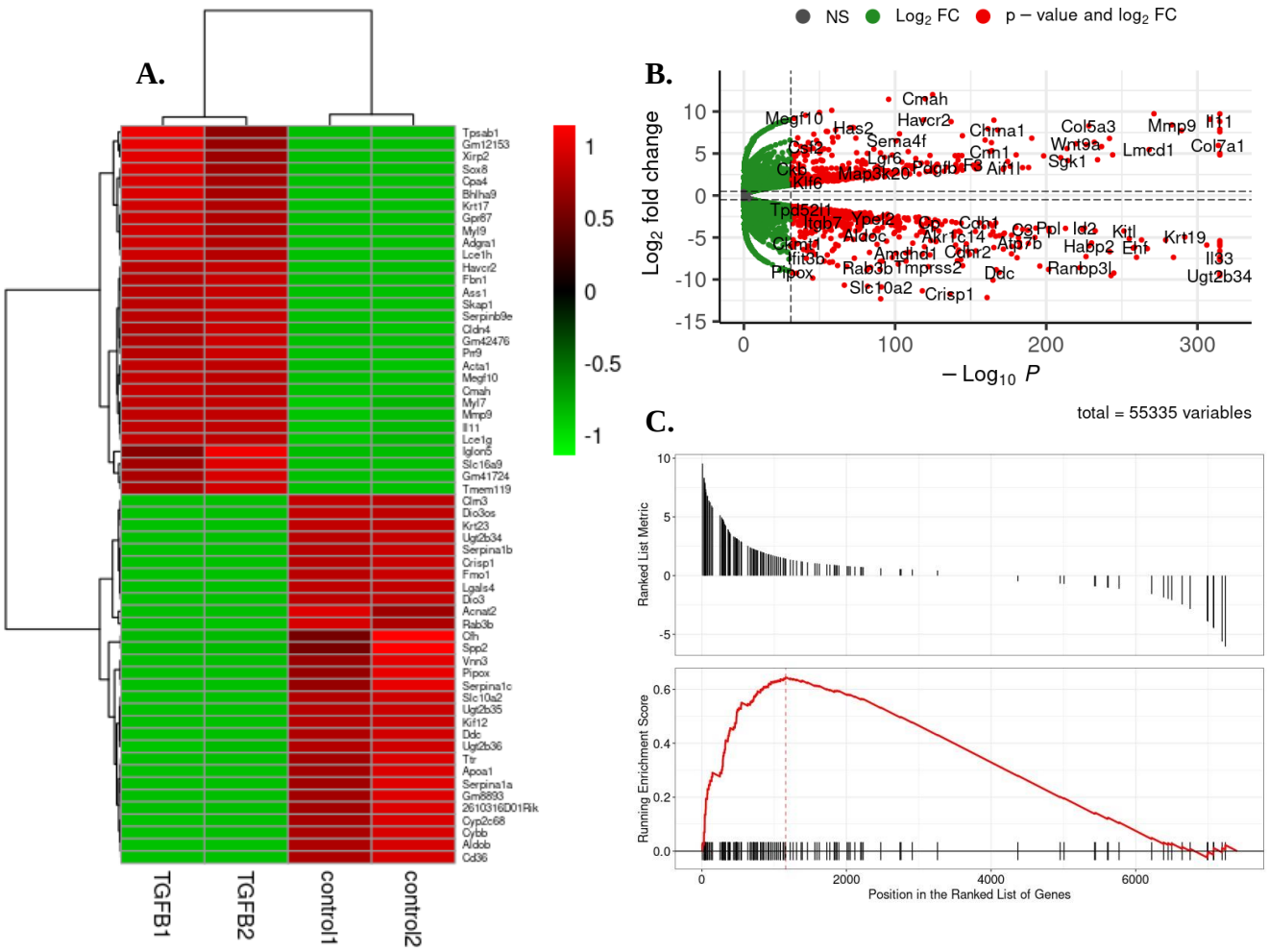

Figure 8. Comprehensive Analysis of Mouse Gene Expression in EMT-MET Processes. (A) Heatmap of Top 30 Up- and Down-regulated Mouse Genes. The heatmap illustrates the differential gene expression in the EMT-MET processes of the mouse. Genes that are upregulated are shown in red, while those that are downregulated are shown in green. (B) A volcano plot of the differential expression analysis shows the top 30 up- and down-regulated mouse genes. Non-significant

results are in gray and labeled "NS". The x-axis (green) represents the Log2FC, indicating the expression difference between two conditions. The y-axis (blue) shows the negative log10 of the p-value, indicating the statistical significance. Significant results (red) meet both p-value and fold change thresholds. (C) GSEA Plot of Mouse Gene Sets for EMT. The GSEA plot consists of two panels: Ranked List Metric (upper) and Running Enrichment Score (lower). The Ranked List Metric panel displays the logFC values for genes ordered by rank, while the Running Enrichment Score panel shows the running enrichment score (ES) across the ranked list**.**

As expected, the two control groups and two TGF-β treated groups represented coherent expression profiles within themselves. The control samples were dominated by red color, which indicates higher expression levels, while the mix of red and green signals in the treated samples represented a distinct change in gene expressions upon TGF-β treatment. Among these differentially expressed genes were 6951 protein-coding genes and 453 lncRNAs. The volcano plot illustrates the distribution of the differentially expressed genes upon TGF-β treatment. From this plot, one can confirm that the different gene expression profiles observed were coherent within the control and TGF-β treated groups and that many changes in gene expression were significant. Clustering is apparent for up and downregulated genes, supporting the experimental finding of distinct gene expression profiles between control and treated samples.

GSEA across these genes revealed that 206 genes are significantly associated with EMT. This revealed that TGF-β treatment has broad effects on multiple pathways and processes. From the data analysis on the EMT phenotype, one gets the GSEA plot showing a high correlation for both up and downregulated genes. Obviously, going back to the running enrichment score plot (Figure 8C), the ES will then be high, meaning these genes are enriched at the top of the ranked list and thus significantly associated with the EMT phenotype.

One may get the implication from putting together information from the heat map, volcano plot, and GSEA plot that there are significant changes in gene expression profiles associated with TGF-β induction in mouse models. Genetic changes induced are not just of statistical significance; they are also of biological relevance to such pathways as EMT and TGF-β signaling. This integrated analysis of the graphical

outputs proposes an informed interpretation of the molecular mechanisms underlying TGF-β-induced EMT processes in mice.

In the human MCF10A cell line samples were used for comparison with previous mouse data, a total of 8120 genes were identified as significantly differentially expressed, indicating that both types of genes were involved in the cellular response to TGF-β treatment, with protein-coding genes being more prevalent.

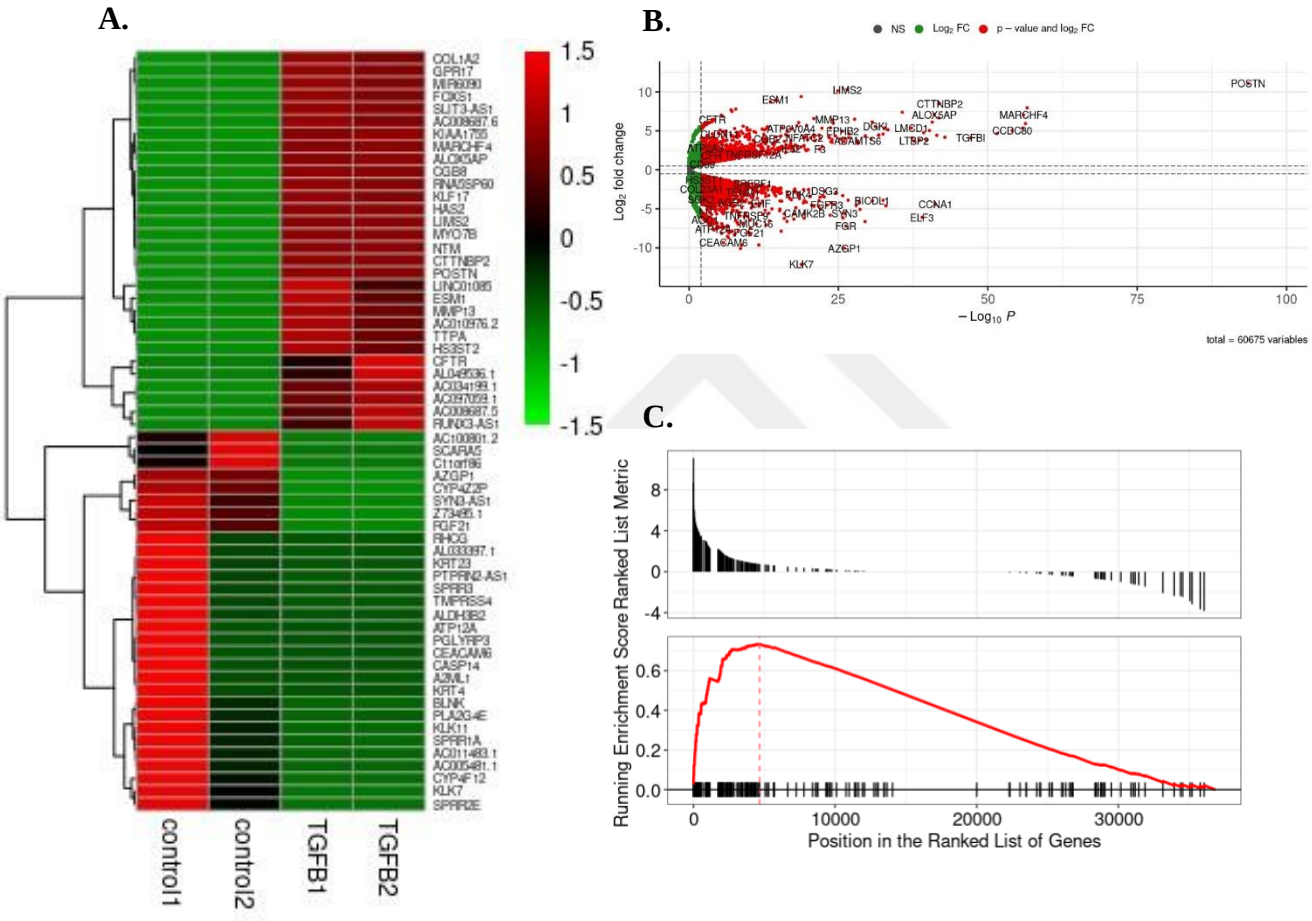

Figure 9. Comprehensive Analysis of Human MCF10A Cell Line Gene Expression in EMT-MET Processes. (A) Heatmap of Top 30 Up- and Down-regulated Human MCF10A Cell Line Genes. The heatmap represents the differential expression of genes in the human MCF10A cell line, with two control groups and two TGF-β treated groups. The differentially expressed genes are shown in varying shades of red (upregulated) and green (downregulated), indicating parallel differential expression in the same genes across the treatment groups. (B) Volcano plot

demonstrating differential expression analysis in the human MCF1A0 cell line. The volcano map depicts the differential expression analysis results, namely the distribution of highly expressed genes (upper right corner) and downregulated genes (upper left corner). This suggests that highly expressed genes are more widely distributed than downregulated ones. (C) Gene Set Enrichment Analysis, GSEA of MCF10A cells. The GSEA plot displays the significant upregulation and downregulation of genes during the EMT process in MCF10A cells. This analysis highlights the considerable number of genes exhibiting changes in expression, emphasizing their involvement in EMT processes.

As illustrated in Figure 9A, differentially expressed genes are expressed significantly differentially between two comparison groups. The dramatic changes in gene expression profiles after TGF-β treatment further underscore the significant effect of this treatment on the EMT process and associated signaling pathways. These differential expression patterns thus indicate that TGF-β treatment induces appreciable changes in gene expression related to these critical biological processes.

The volcano plot from MCF10A data reveals that despite the high number of differentially expressed genes, the number of significantly differentiated genes after the treatment with TGF-β is relatively low (Figure 9B). The patterns of differential expression observed, with a significant number of upregulated genes and a smaller number of downregulated genes, underscore the pivotal role of TGF-β in regulating key biological processes like EMT. The important up-regulation of genes to the right side of this plot suggests the activation of pathways related to EMT and associated signaling. Conversely, the fact that there were significantly downregulated genes, although fewer in number, can be interpreted as the suppression of a subset of genes in response to TGF-β involved in maintaining an epithelial state.

GSEA on differentially expressed genes in the MCF10A human data discovered drastic changes in gene expression during the EMT process in Figure 9C. The running enrichment score for these genes related to EMT proves their extreme enrichment, indicating their essential involvement in the TGF-β-induced EMT process. TGF-β treatment exerts a profound influence on gene expression in human MCF10A cells. The heatmap is able to draw a clear expression profile of coherence

within the control and treated groups, setting them apart and showing the specific effect of TGF-β on gene expression. This is further supported by the volcano plot with regard to the huge up and downregulation of genes with key functions in the EMT process and related pathways. The GSEA plot provides an example of gene sets enriched in this process, further confirming the role of these genes in the mediation of the EMT process. These analyses provide a comprehensive view of the molecular changes induced by TGF-β treatment and underline its key role in driving the EMT process of MCF10A cells.

Using MCF7 cells, differentially expressed genes were identified across uninduced, induced, and deinduced samples treated with doxycycline. A substantial number of 20482 genes were present among these groups.

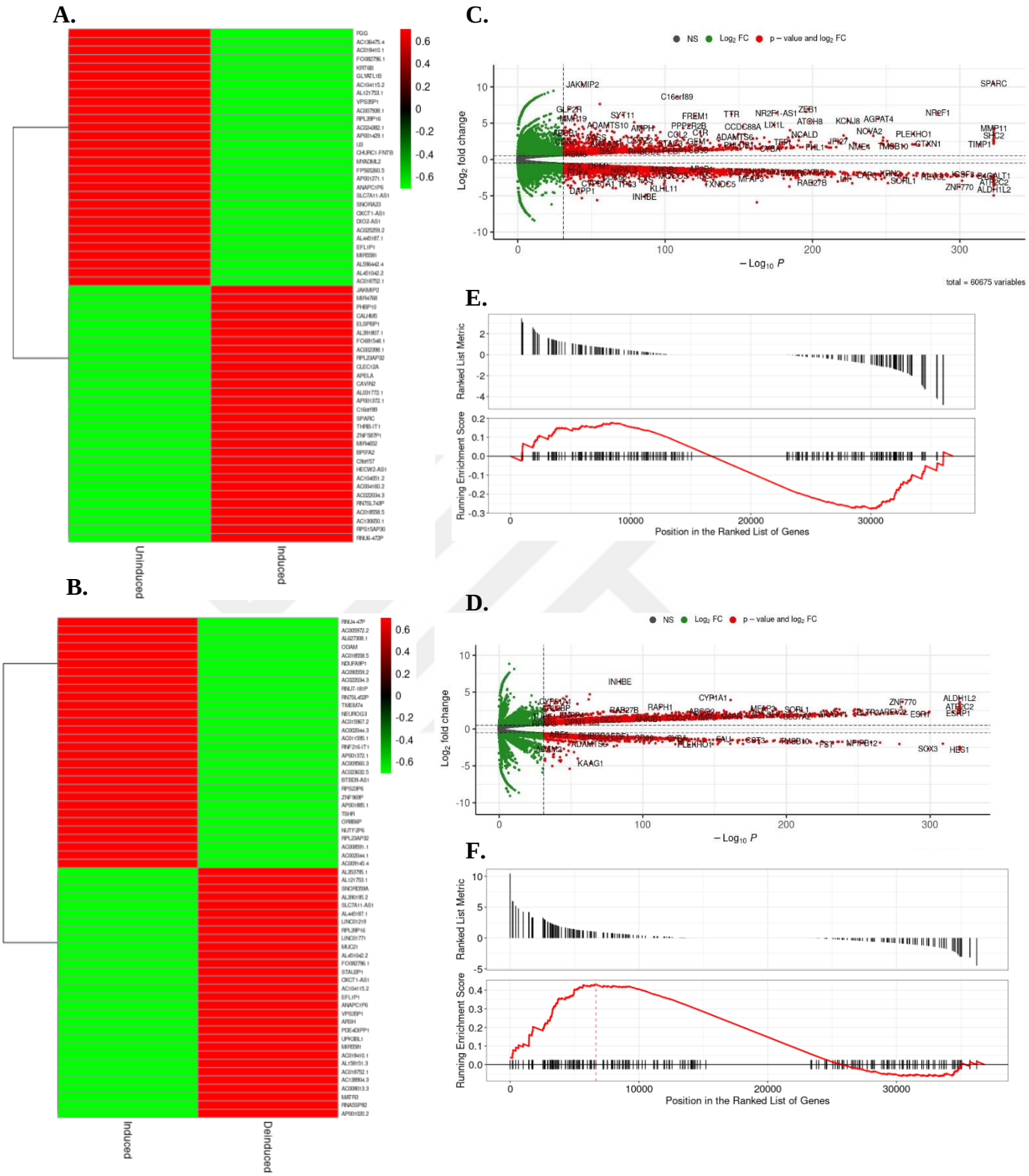

Figure 10. Differential Gene Expression Analysis in Human MCF7 Cell Line. (A) Heatmap of the top 30 downregulated genes in the human MCF7 cell line under three different doxycycline-treated and untreated groups. The heatmap represents differential gene expression, with red indicating upregulation and green indicating

downregulation. The consistent expression profiles within each group illustrate the impact of doxycycline treatment on gene expression. (B) Heatmap showing gene expression profiles in MCF7 cells treated with doxycycline and untreated controls. Differential expression is visualized with red and green colors representing upregulation and downregulation, respectively. (C) Volcano plot depicting differentially expressed genes in untreated (epithelial) vs. doxycycline-treated (mesenchymal) MCF7 cells. This plot highlights the significant changes in gene expression, with red indicating significant results and green indicating non-significant results. (D) Volcano plot comparing gene expression in MCF7 cells before and after doxycycline treatment, illustrating the transitions from untreated (epithelial) to treated (mesenchymal) and back to deinduced (epithelial) states. The significant genes are highlighted, indicating the regulatory shifts due to doxycycline treatment and its removal. (E) GSEA results comparing uninduced and induced) samples. The plot shows upregulated and downregulated genes, with a significant enrichment score indicating key gene sets involved in the transition. (F) GSEA results comparing induced and deinduced samples, demonstrating the reversal of gene expression upon doxycycline removal. The plot reinforces the findings that many genes revert their expression levels, suggesting a reversal of the mesenchymal state.

The heatmap generated from these differentially expressed genes between the uninduced-induced versus induced-deinduced groups indicated that genes, as was expected, were up- and down-regulated in response to doxycycline treatment and its withdrawal (Figure 10A and B). In volcano plots, as seen in Figure 10C and D, obtained with the same data, it was found that differentially expressed genes in the uninduced-induced group had a broader range of logFC compared to genes in the induced-deinduced group. This broader distribution in the uninduced-induced group is probably because of a more vigorous transcriptional response by the cells to induction by doxycycline in the first place, which involves genome-wide changes in gene expression. On the other hand, the induced-deinduced group, which represents doxycycline withdrawal, has a narrower range for logFC. This could be due to gene

expression levels that have become stabilized or normalized upon the return of cells to their baseline state.

Using GSEA of differentially expressed genes, there was already significant expression change from the uninduced to the induced state. Similar to what is shown in Figure 10C and Figure 10D, a number of genes are either up or downregulated, therefore giving a coherent pattern of changes in gene regulation from epithelial to mesenchymal states following doxycycline treatment supporting such observations.

In the MCF7 cell line, the GSEA between induced vs. deinduced samples shows that in the study represented in Figure 10F, there are many genes recovering their expression level after doxycycline removal, which could be a hallmark of reversal of the mesenchymal state. These findings are further reinforced by Figure 10 D and F panels, whereby a significant number of genes change in regulation.

Differential expression analysis across the human MCF7 and MCF10 cell lines, along with mouse data, indicated significant variations in the expression of lncRNAs during the EMT-MET transition. In MCF10A samples, a number of differentially expressed genes were present—585 up- and 364 down-regulated, while only 5 genes were upregulated for the MCF7 samples. Mouse data revealed 194 upregulated and 148 downregulated lncRNAs. Heatmaps and volcano plots of these data sets further supported changes in gene expression under TGF-β treatment, showing a robust response in gene expression following treatment and subsequent normalization upon its removal. The implications of these genes in EMT-related pathways and gene regulation were validated by GSEA with consistent patterns. There were also clear transcriptional changes shown upon both TGF-β and doxycycline treatment in MCF10 and MCF7 cells, respectively, underscoring their responsiveness and relevance of these models for studying EMT-MET transitions.

In summary, variations in gene expression across these three sample groups are overwhelming and consistent, making them appropriate for investigating new lncRNAs related to EMT-MET processes. In the Gene ontology term enrichment analysis, the distribution of differentially expressed genes was examined across various biological processes, including EMT, MET, negative regulation of EMT,

positive regulation of EMT, and positive regulation of mesenchymal cell proliferation. The analysis gave 93 mouse genes involved in EMT, 66 genes of the MCF10 cell line, while that of the MCF7 cell line gave 260 genes. The analysis identified 93 genes involved in EMT from mice, 66 genes from the MCF10 cell line, and 260 genes from the MCF7 cell line.

Furthermore, 32 genes from mice, 30 genes from the MCF10 cell line, and 81 genes from the MCF7 cell line were found to positively regulate EMT. For the negative regulation of EMT, 20 genes from mice, 12 genes from the MCF10 cell line, and 60 genes from the MCF7 cell line were identified. In the MET process, 11 mouse genes, 7 genes from the MCF10 cell line, and 23 genes from the MCF7 cell line were identified. Additionally, 14 genes from mice, 12 genes from the MCF10 cell line, and 39 genes from the MCF7 cell line were found to positively regulate mesenchymal cell proliferation. Among them, 31 common genes related to the EMT process were identified, including TGFB1, NOTCH1, and SMAD7, while 2 genes were found common for MET, namely LIF and CTNNB1. However, none of the genes related to the MET process identified in each dataset reached statistical significance. These analyses represent solid foundational support for investigating the molecular mechanisms driving these critical biological transitions.

4.3. Identification and Discovery of LncRNAs

In the initial step of our study, de novo analysis utilizing data from both mouse and human samples indicated a significant impact of these lncRNAs on the EMT-MET process. To identify previously unannotated lncRNAs, we employed a variety of computational tools, including TransDecoder, CPAT, gffcompare, and BLAST.

The gffcompare program was used during transcript analysis to classify these transcripts by their class code: =, c, i, j, k, m, n, o, p, s, u, x, y. This enabled further filtering of the transcripts to be done more precisely.

Table 2. Summary of gffcompare class codes and transcript classification.

Class Code	Description	Number of Transcripts	Orientation
=	Exact match of intron chain	230662	sense
c	Query transcript fully contained within a reference transcript	72	sense
i	Reference transcript fully contained within a query transcript	543	unclassified
j	Overlaps with a reference transcript but has at least one novel splice junction	6067	sense
k	Does not overlap with any reference transcripts on the same genomic strand.	6309	sense
m	Multiple transcripts map to the same locus, pointing to possible alternative splicing.	531	unclassified
n	Found in a genomic region not annotated in the reference.	411	unclassified
o	Additional identical strands intersect with the reference exons.	254	unclassified
p	Entirely within a reference intron, pointing to a fragment.	57	intronic
s	Found to overlap with a reference transcript on the opposite strand, pointing to antisense.	7	antisense
u	Intergenic and no supporting evidence for splicing.	542	intergenic
x	Exonic overlap with a reference transcript on the opposite strand.	256	antisense
y	Two exon reads aligning to an intron of a reference gene - a fragment of unspliced pre-mRNA.	22	unclassified

As described in Table 2, the class code u denoting intergenic transcripts, which were not annotated in the reference genome. Transcripts assigned the p class code were characterized as intronic due to localization within the confines of annotated gene introns. Those systemically categorized as =, c, j, and k were classified as sense transcripts for their being, in these cases, either overlapping or falling within exonic regions of an annotated gene on the same strand.

In contrast, those transcripts classified under x and s class codes that overlap exonic regions of known genes on the opposite genomic strand were identified as antisense transcripts. On the other hand, transcripts identified as i, m, o, y, and n categorize more complex or ambiguous relationships with annotated genes, calling for a nuanced means of classification rather than calling them into one category.

As a result of these filtering steps, in addition to the 9963 lncRNAs found in mouse data, 6806 previously unannotated lncRNAs were identified in the mouse genome. As shown in Figure 12A and B, the distribution of previously unannotated lncRNAs differs from the chromosomal distributions of known lncRNAs. The chromosome with the highest number of known lncRNAs is chromosomes 7 and 5, with 616 lncRNAs, whereas the highest number of previously unannotated lncRNAs is found on chromosome 2, with 1291 lncRNAs. Known lncRNAs are least abundant on the X chromosome, with 180 lncRNAs, while previously unannotated lncRNAs are present in very low numbers on the mitochondrial chromosome, with 39 lncRNAs.

Upon examining exon distribution in Figure 12C and D, it was found that in known lncRNAs, 3410 lncRNAs have 2 exons, followed by 1844 lncRNAs with 3 exons. Only one lncRNA was found to have the highest count of exons, with a count of 503. These numbers sharply increase in the case of lncRNAs that were previously unannotated, as 5419 were carrying a single exon, and 921 lncRNAs were found to have 2 exons. Again, only a single lncRNA had the maximum count of exons, 361.

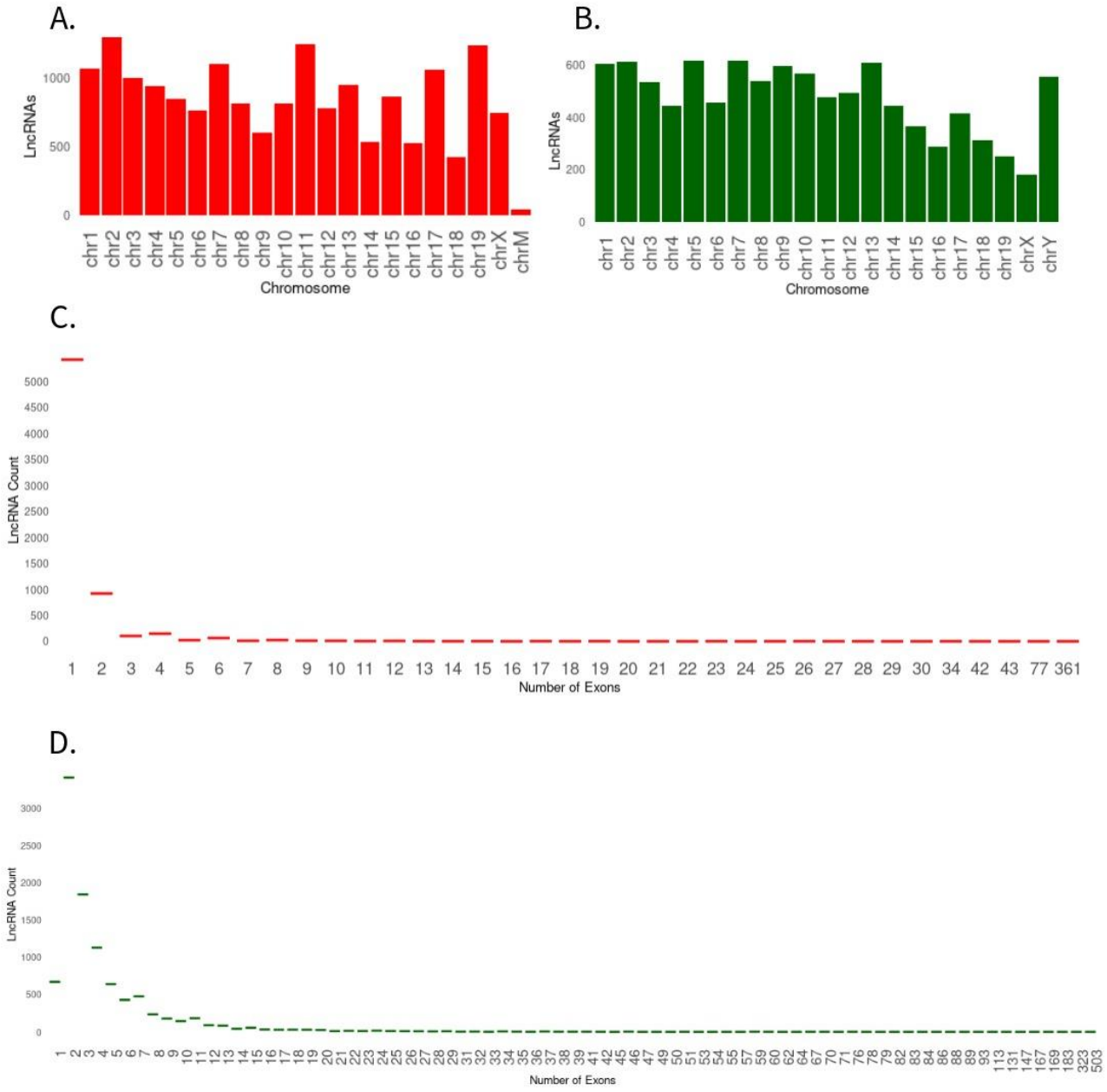

Figure 11. Comparative analysis of mouse previously unannotated and known lncrnas. (A) Barplot representing the counts of previously unannotated lncRNAs in the mouse genome. (B) Barplot illustrating the counts of known lncRNAs in the mouse genome. (C) Distribution of exon counts in mouse previously unannotated lncRNAs in boxplot. (D) Distribution of exon counts in mouse known lncRNAs in boxplot.

After applying the same filtering steps to obtain previously unannotated lncRNAs from samples with human MCF7 and MCF10A cell lines, 719 previously unannotated lncRNAs were obtained from the dataset. Of the previously unannotated lncRNAs, 362 are located on chromosome 1, while the lowest number, 2, is found on the mitochondrial chromosome (Figure 12A). In the reference human genome, the densest population of lncRNAs is on chromosome 1 with 1.590 lncRNAs, while chromosome Y harbors the fewest lncRNAs with 148 as seen in Figure 13B.

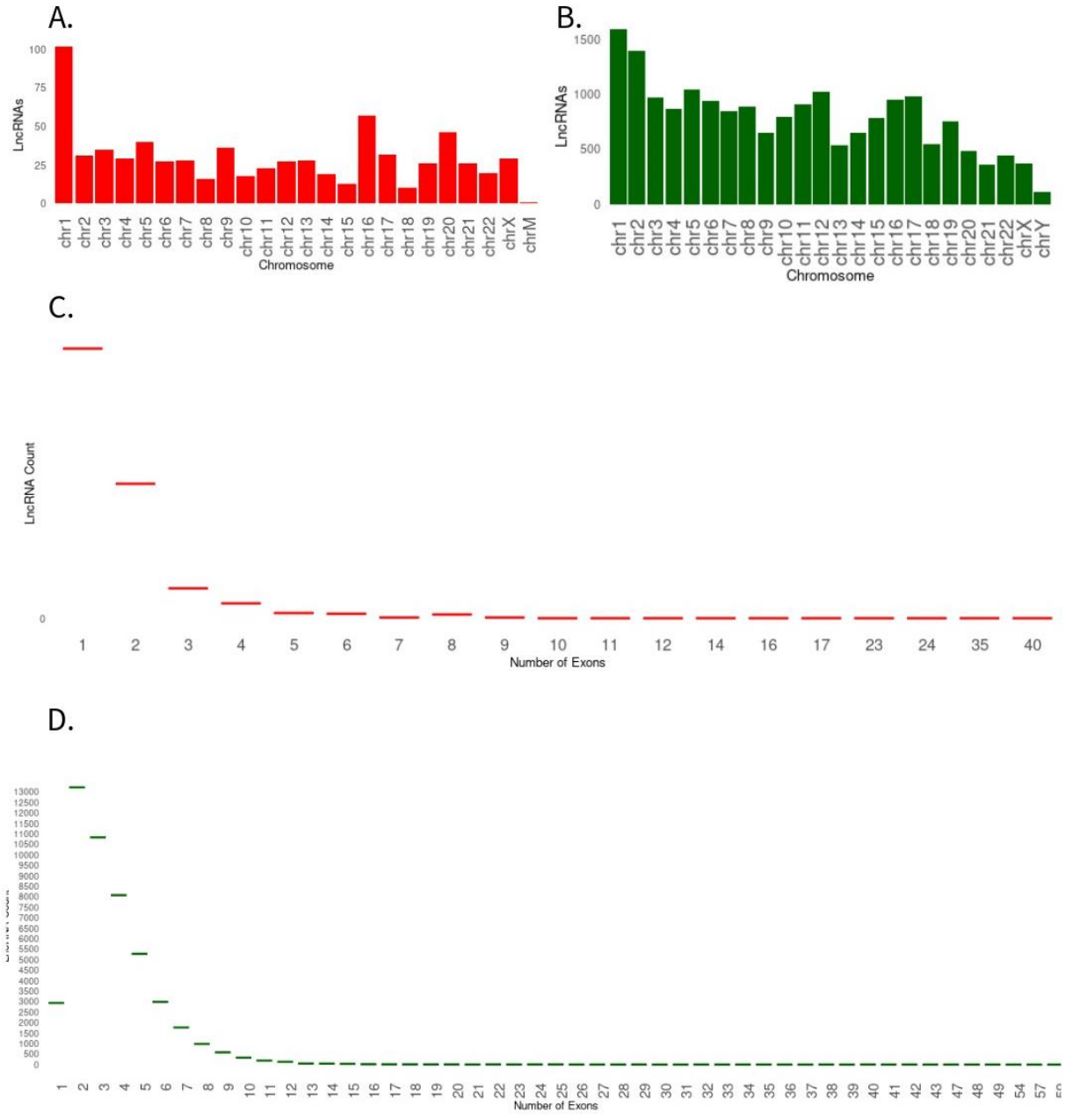

Figure 12. Comparative analysis of human previously unannotated and known lncrnas. (A) The barplot shows the number of hitherto unannotated lncRNAs in the human genome. (B) A barplot depicts the count of known lncRNAs in the human genome**.** (C) Distribution of exon counts in human previously unannotated lncRNAs. (D) Distribution of exon counts in human known lncRNAs.

Among the previously unannotated human lncRNAs, there is a single lncRNA with up to 40 exons, while 362 lncRNAs with a single exon were identified (Fig12C).

In contrast, in the reference genome, although there is an lncRNA with up to 59 exons, the majority is formed by 5687 lncRNAs with 2 exons (Fig12D).

Through filtering and analysis of previously unannotated lncRNAs in both mouse and human, orthologous similarities were examined using the UCSC Genome Browser by selecting specific genes.

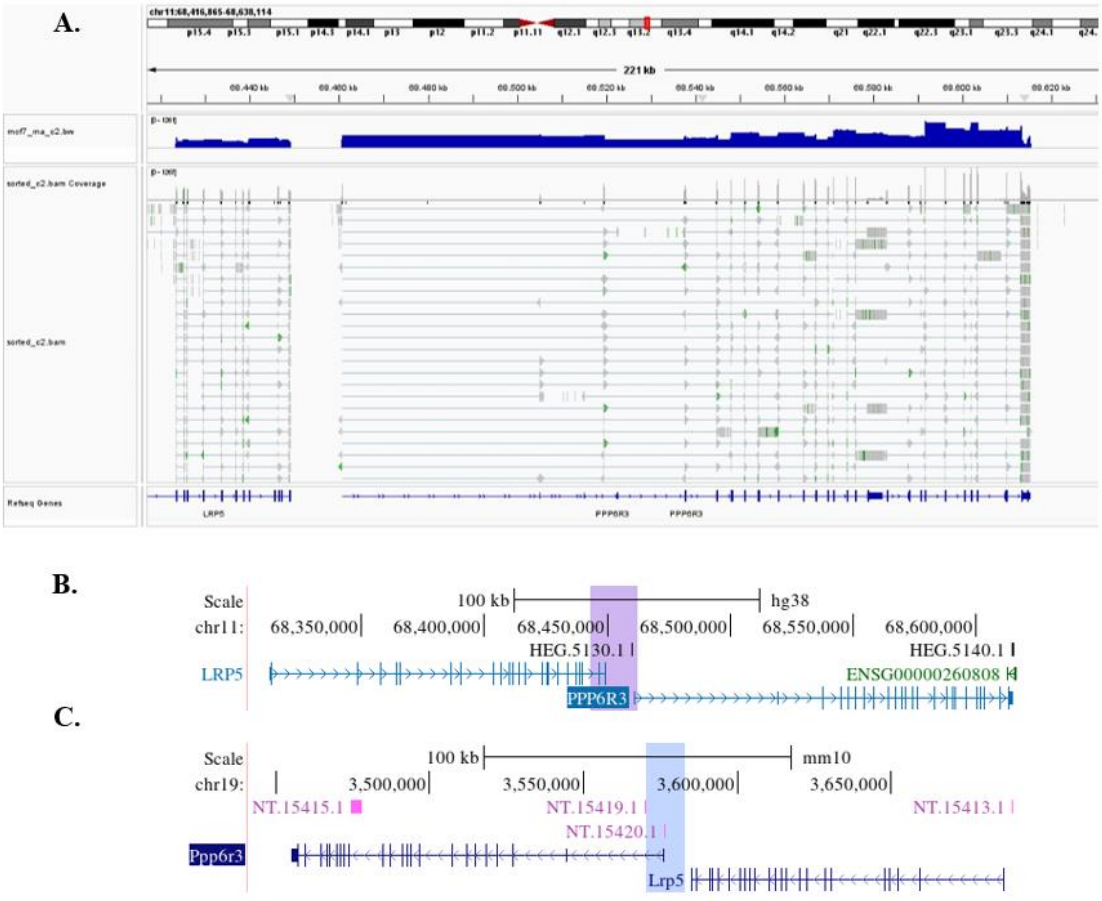

Figure 13. Comparative analysis of previously unannotated lncRNAs in human and mouse ortholog gene regions. (A) IGV view of RNA-seq data for the human samples showing read coverage and alignment in the PPP6R3 locus. (B) UCSC Genome Browser view of the human PPP6R3 and LRP5 gene regions showing the location of previously unannotated lncRNA. (C) UCSC Genome Browser view showing a mouse orthologous gene region for Ppp6r3 and Lrp5, with the comparative location of the previously unannotated lncRNA.

One orthologous pair includes the human PPP6R3 gene located on chromosome 11 and the mouse Ppp6r3 gene located on chromosome 19 (Figure 13).

It was observed that NT.15420.1 is situated downstream (3') of the Ppp6r3 gene in the mouse genome, while the HEG.5130 transcript also maps downstream (3') of the PPP6R3 gene in the human genome. The relative positions downstream of the PPP6R3/Ppp6r3 genes, which are relatively conserved between human and mouse, point toward a potential syntenic relationship. The IGV results provide a detailed read coverage and transcript structure visualization in the human genome. In this IGV track, high coverage of PPP6R3 on chromosome 11 is noted, with representation of multiple transcript isoforms inferred by the multiple RNA-seq reads that align to this region. Notably, the HEG.5130 transcript appears downstream of PPP6R3, in line with its annotated position.

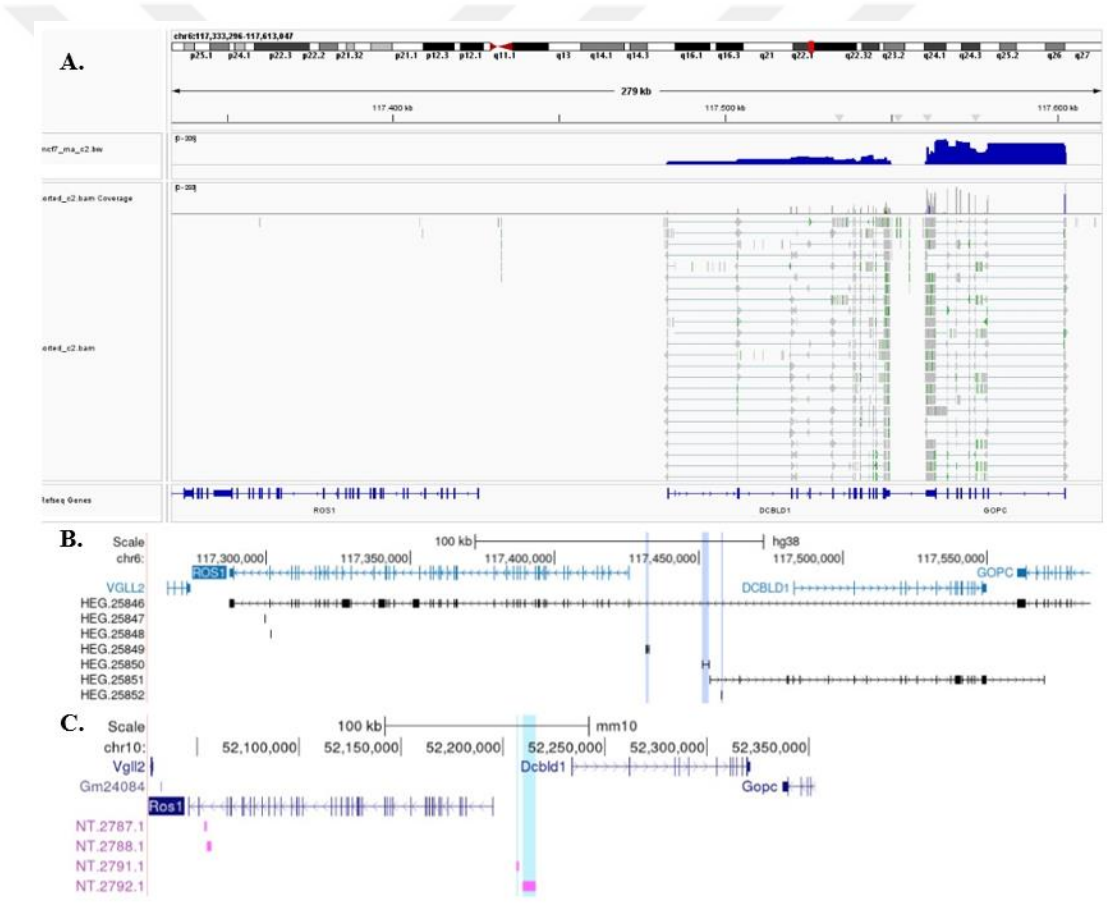

Figure 14. Previously unannotated lncRNA comparison in mouse and human ortholog gene regions. (A) The IGV visualization of the RNA-seq data for the human samples demonstrates the alignment and read coverage of the DCBLD1 and ROS1 loci. (B) The location of previously unannotated lncRNA is displayed in the human

ROS1 and DCBLD1 gene areas shown in the UCSC Genome Browser. (C) A mouse orthologous gene region containing Ros1 and Dcbld1 is displayed in the UCSC Genome Browser, along with the relative placement of the lncRNAs, which were previously unannotated.

As seen in Figure 14, previously unannotated lncRNAs were detected between ROS1 and DCBLD1 genes in both human and mouse genomes, indicating a conserved syntenic block. Also, the IGV track of the human genome provides substantial coverage, and several transcript isoforms align between the ROS1 and DCBLD1 gene regions supported by multiple RNA-seq reads, indicating its expression and potential functional relevance. Those lncRNAs are similarly located with respect to protein-coding genes in both species, indicating potential functional conservation.

This positioning implies that the identified previously unannotated lncRNAs may play significant roles not only in the regulation of EMT and MET processes but also in other biological events. Considering their genomic context and likely interaction with neighboring genes supports the possibility that these lncRNAs may be involved in regulatory networks governing these processes.

4.4. Synteny Analysis of Human lncRNAs on Mouse Genome

Through the analysis of data obtained from the filtering steps, information has been gathered regarding the presence and complex structures of previously unannotated lncRNAs that play a substantial part in EMT-MET. Given this information, a synteny analysis was conducted using the Ortho2align tool, aiming to identify orthology relationships in a genome enhanced with previously unannotated lncRNAs found in both mice and humans. After obtaining potential orthologous human lncRNAs from mouse lncRNAs using the Ortho2align tool, the accuracy of the results was verified using lncRNAs known to be orthologous, such as Dancr, Malat1, and Xist (Figure 15).

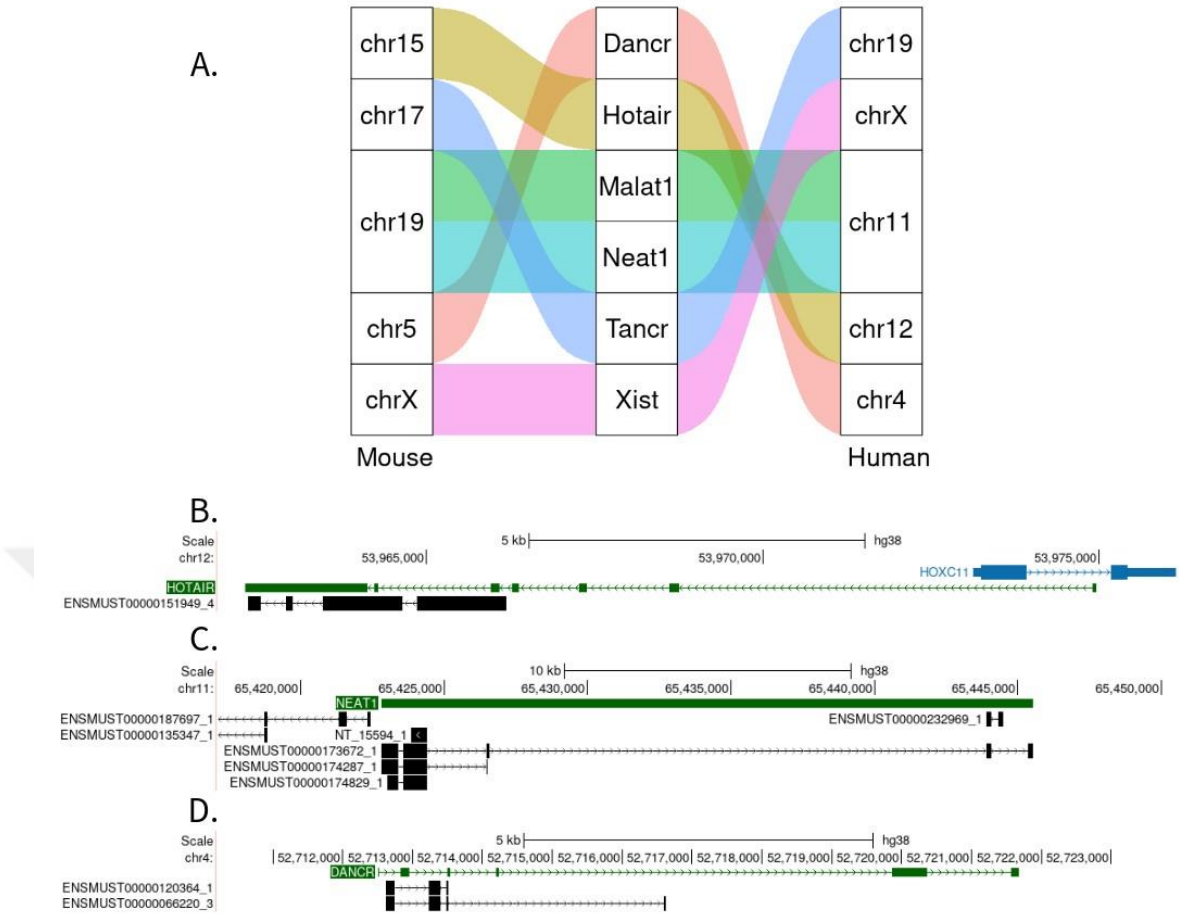

Figure 15. Identification and Validation of Potential Human Orthologous lncRNAs from Mouse lncRNAs. (A) Sankey diagram shows the alignment and mapping of mouse lncRNAs to potential human orthologs identified by Ortho2align. (B) Genome Browser View displays the genomic location and structure of the human ortholog of the mouse lncRNA ENSMUST00000159194.4 identified as HOTAIR. (C) Genome Browser View shows the genomic position of the human orthologous mouse lncRNA ENSMUST00000187897.1 corresponding to NEAT1. (D) Genome Browser View of DANCR, which is the human ortholog of the mouse lncRNA ENSMUST00000102360.1. Genome browser images of Fig14B, C, and D are detailed views of the genomic location and structure of the human orthologs corresponding to mouse lncRNAs: ENSMUST00000159194.4 (HOTAIR), ENSMUST00000187897.1 (NEAT1), and ENSMUST00000102360.1 (DANCR). Although the genomic context is from the human genome, these names of transcripts originated from the mouse

lncRNAs, thus fully consistent with the output of Ortho2align; it takes mouse input and reports human potential orthologs.

Novel lncRNAs NT_20227 and NT_25341 were previously identified as potential contributors to the MET process in mouse (Özkan, A., 2024). In the current study, such lncRNAs were selected for visualization of the novel lncRNAs potentially active in the EMT-MET process in search of the potential orthologous regions between mouse and human genome.

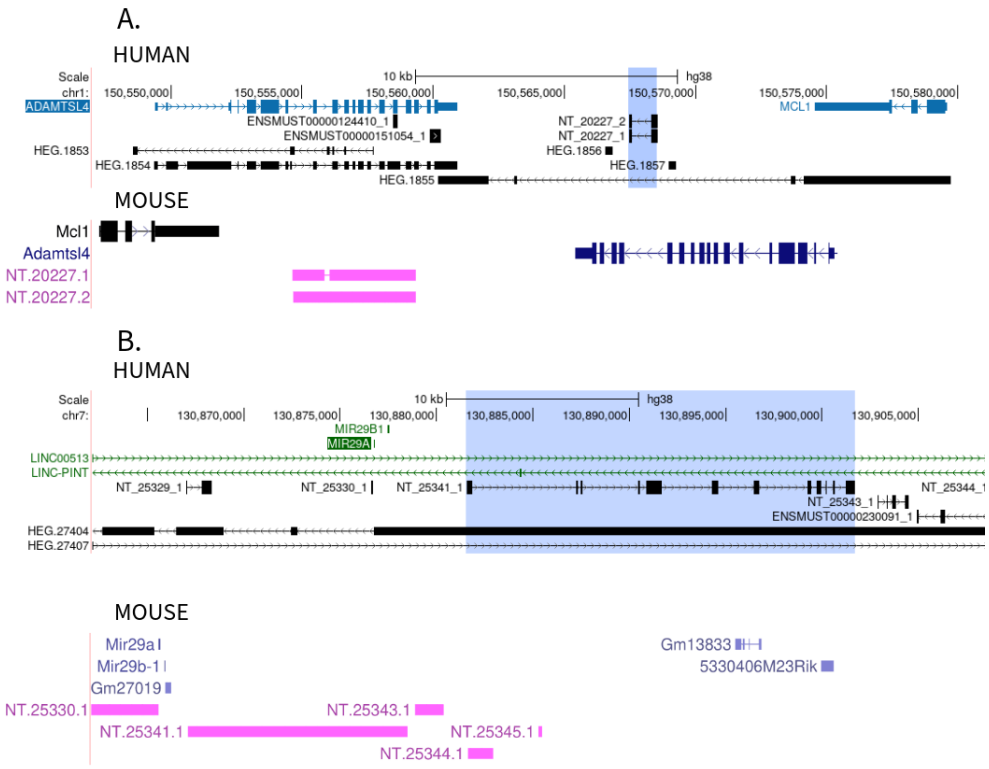

Figure 16. Comparative Genomic Analysis of Novel lncRNAs in Mouse and Human Genomes. (A) Genomic localization of the novel mouse lncRNA NT_20227 in the mouse genome and its previously unannotated ortholog and overlapping novel human lncRNAs in the human genome. (B) Localization of the novel mouse lncRNA NT_25341 and human corresponding regions, including novel human lncRNAs and their potential orthologs.

Panel A of Figure 16 shows NT_20227 in the mouse genome, located between the genes Mcl1 and Adamts4. The orthologous human genome region contains orthologous transcripts of NT_20227 but also novel human lncRNAs such as HEG.1853, HEG.1855, HEG.1856, and HEG.1857, evidencing conserved functional elements. The NT_20227 is placed upstream of ADAMTS4 and downstream of MCL1 in both genomes, pointing to a conserved syntenic block. Figure 17B displays the genomic location of NT_25341 in the mouse genome, flanked by Mir29a and Gm27019. The human genome contains the potential ortholog NT_25341 and novel human lncRNAs HEG.27404 and HEG.27407. NT_25341 is upstream to MIR29A and overlaps with LINC-PINT, which indicates a complex regulatory landscape of coding and non-coding elements. Both NT_20227 and NT_25341 have multiple exons suggestive of complex splicing patterns. Within the mouse genome, NT_20227 is a span of exons in an intergenic region between Mcl1 and Adamts4, where the corresponding orthologous regions in humans also have similar exon structures. Likewise, NT_25341 in the mouse genome has multiple exons upstream of Mir29a, whereby corresponding orthologues in humans display comparable exon arrangements.

The input mouse genome lncRNA file contained 9963 known and 6806 previously unannotated lncRNAs. In the mouse lncRNA input file, the highest number of lncRNAs was found on chromosome 11, with 2821 lncRNAs, whereas the mitochondrial chromosome had the fewest lncRNAs, with 12. From the input set of previously unannotated and known mouse lncRNAs, 30024 lncRNAs demonstrated orthology with the human genome. This human genome dataset has been enhanced with previously unannotated human lncRNAs, facilitating the identification of these orthologous relationships.

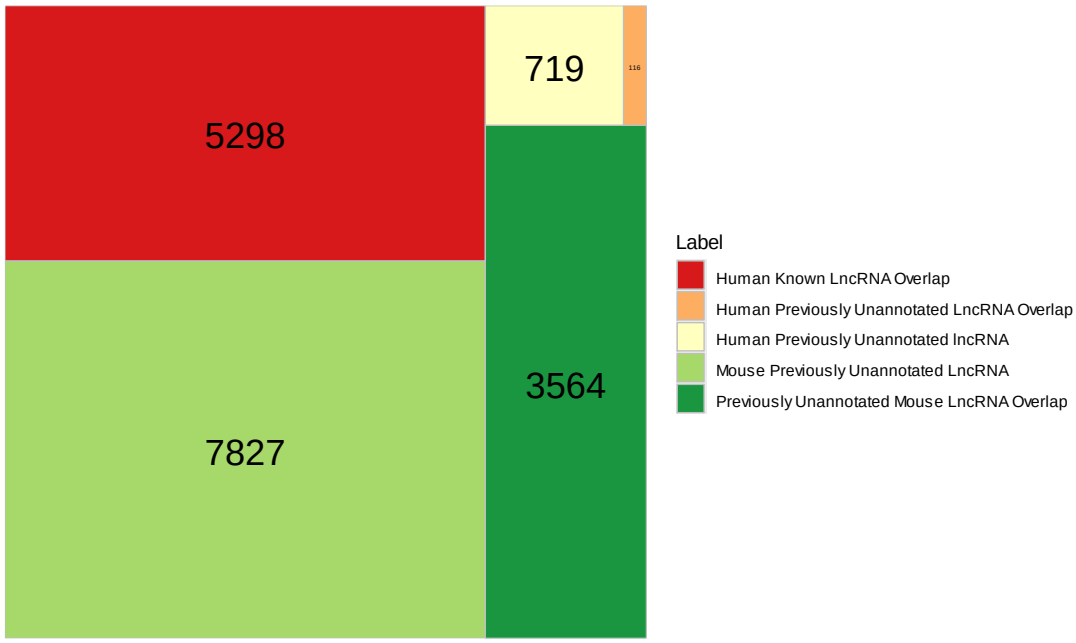

Figure 17. Treemap visualization of previously unannotated lncRNA categories in mouse and human ortholog gene regions.

As seen in Figure 17, Potential orthologous regions overlap with 49781 transcripts in the human genome enhanced with previously unannotated lncRNAs. Among these transcripts, 116 overlap with previously unannotated lncRNAs, and 5298 overlap with known human lncRNAs.

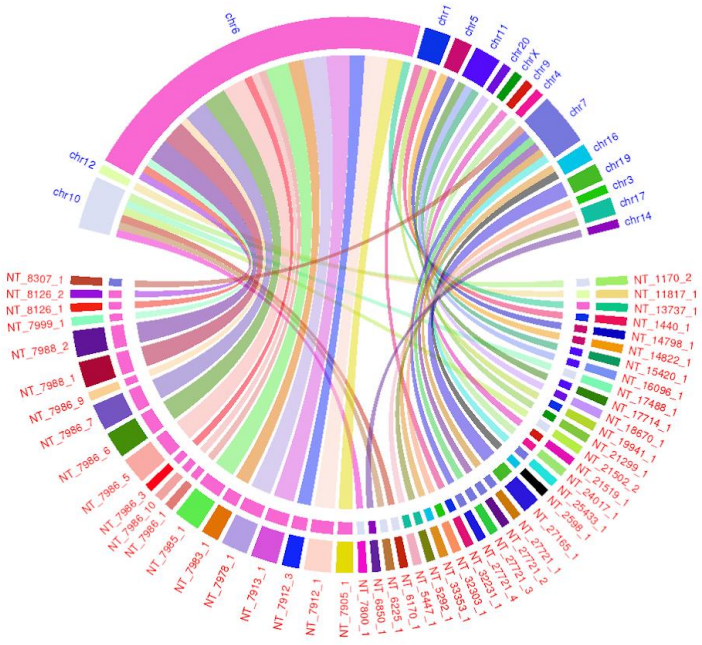

Figure 18. Chord diagram of overlapped previously unannotated human lncRNAs with previously unannotated mouse lncRNAs. The diagram highlights conserved synteny patterns between mouse and human genomes, which are structured with previously unannotated lncRNA containing mouse chromosomes on one side (blue) and previously unannotated human lncRNAs on the opposite side (red).

The highest number of overlaps with previously unannotated lncRNAs occurs on chromosome 6, as seen in Figure 18, with 49 overlaps, while the lowest number of overlaps, with just one, occurs on chromosome 15.

4.4.1. Transcript Structure Validation

Despite obtaining the structures and orthological information of the previously unannotated lncRNAs in the human and mouse datasets, the methods used in these analytical steps were modified to examine the changes in results. As a result, supplementing the human dataset with additional samples increased the number of previously unannotated lncRNAs to 1712. In this analysis, the highest number of previously unannotated lncRNAs, 579, were found on chromosome 2, while the mitochondrial chromosome had the lowest count with only 1 lncRNA. The fact that 1146 lncRNAs possess only a single exon indicates that the increase in sample size did

not affect the read depth. Nonetheless, one lncRNA with 15 exons raised the maximum exon count above 8.

Using synteny analysis, a set of 31698 orthologs was yielded containing previously unannotated human lncRNA orthologs of 4302 genes. The overlap of the positions mapped back to the human genome enhanced with previously unannotated human lncRNAs resulted in 54971 overlapped regions. Of these overlaps, previously unannotated mouse lncRNAs numbered 3713, with an additional known lncRNAs. Of these overlaps, 126 matches were previously unannotated human lncRNAs, while 5908 were known lncRNAs. There were 116 known lncRNAs related to these overlaps in EMT and 12 in MET.

In addition to increasing the number of human samples, the analysis steps were also attempted by merging the fastq files of these samples to analyze them as a single sample. This approach resulted in a dramatic increase in the number of previously unannotated lncRNAs, reaching 16735. The highest number of previously unannotated lncRNAs was again found on chromosome 1, while no previously unannotated lncRNAs were found on the mitochondrial chromosome. Examination of the exons of these lncRNAs revealed that 6942 of them are single-exon. Moreover, there is one lncRNA with 17 exons.

When synteny analysis was performed using Ortho2align, the total number of orthologs identified decreased to 33099, and the number of previously unannotated mouse lncRNAs to 4229. When these positions were mapped onto the human genome, enhanced with previously unannotated human lncRNAs, a total of 5406 overlaps were detected. Among these overlaps, 3638 previously unannotated mouse lncRNAs were present along with additional known lncRNAs. Notably, 118 previously unannotated human lncRNAs were matched, and 5555 known human lncRNAs were identified. Of the known lncRNAs, 115 are involved in the EMT and 12 in the MET.

Table3. Summary of Comparative Analysis Results.

	Initial Sample Group	Expanded Sample Group	Merged Sample Group
Previously unannotated LncRNA	719	1712	16735
Ortho2align Output			
Total orthologs	30024	31698	30999
Previously unannotated human lncRNA orthologs	4112	4302	4229
Human Genome Overlapping			
Total overlaps	49781	54971	51406
Overlaps with known human lncRNAs	5298	5908	5555
Overlaps with previously unannotated human lncRNAs	116	126	118
No overlapped	594	615	600

When examining the findings from different sample groups in Table 3, the initial increase in sample number did not considerably improve the count of previously unannotated human lncRNAs. However, depending on changes in the sample processing, a dramatic increase in the number of previously unannotated lncRNAs was observed. Despite this increase in the total number of samples, the potential orthologs between human and mouse did not change significantly. When the synteny results intersect with the previously unannotated lncRNA-enhanced human genome, the number of overlaps with previously unannotated human lncRNAs remained consistent, ranging between 116 and 126. Furthermore, the analysis of known human

lncRNAs in relation to the overlapped regions showed that the number of lncRNAs involved in EMT and MET was similar across all groups.

5. DISCUSSION

In recent years, studies have revealed that lncRNAs, previously overshadowed by protein-coding elements, play significant roles in various key research areas such as cancer, development, and especially EMT-MET processes, interacting with different elements (Gholami et al., 2023; Jia et al., 2022). Although the discovery of lncRNAs dates back to the early 1990s and their functional properties have been found to be similar between humans and mice, the progress in discovering and functionally annotating lncRNAs has not occurred as rapidly as for protein-coding elements (J. T. Lee & Bartolomei, 2013; Lyon, 1992). One of the main reasons these studies have not progressed rapidly is the low sequence conservation of lncRNAs, coupled with their variable expression even within the same biological process of the same species (Ruiz-Orera & Albà, 2019). This diversity and variability increase difficulties for both in-depth investigations with respect to lncRNAs and for the understanding of conserved lncRNAs across species. Nevertheless, recent advances in comparative transcriptomics methods have shown that at least a subset of lncRNAs exhibits conserved functions across species (Breschi et al., 2017). Although this inherent lncRNA variability does not intuitively suggest a conserved pattern of existence, many of these ncRNAs have been shown to be in fact syntenically conserved across species; therefore, their genomic context and neighboring genes may provide clues to functional roles (Bhutkar et al., 2008; Ghanam et al., 2022; Kemkemer et al., 2009). Searching for syntenic regions has become an alternative approach that could allow the search for lncRNAs that are likely to be conserved regulators of EMT and MET by regulating the expression of nearby genes involved in signaling pathways related to cell adhesion, migration, and epithelial-mesenchymal transitions (Akhmetkaliyev et al., 2023; Gonzalez & Medici, 2014).

As with lncRNAs, MET has classically been considered simply the reverse of EMT; its specific signaling pathways and effects are largely unexplored. The role of lncRNAs in EMT-MET processes is commonly investigated, focusing on EMT alone, totally omitting MET, and exploring biological processes in this limited framework (D. H. Kim et al., 2017). Therefore, understanding the elements and processes that lncRNAs interact with is essential for comprehending their biological activity.

This study emphasizes identifying previously unannotated lncRNAs that may be active in the EMT-MET process using human and mouse RNA-seq data from previous studies. To provide sufficient depth for accurate transcriptome assembly, we aimed for a sequencing depth that is consistent with existing norms. For typical eukaryotic transcriptomes, a depth of 30-50 million reads per sample is typically recommended (Conesa et al., 2016; Raghavan et al., n.d.; Z. Wang et al., 2009). Given the complexity of the human and mouse genomes, particularly in capturing the diversity of lowly expressed elements such as noncoding RNAs and alternative splicing process, a high sequencing depth was required (Babarinde & Hutchins, 2021). Our samples have fulfilled this goal of making accurate detection not only for full-length transcripts but also for the other elements, thus allowing significant lncRNAs identification across species comprehensively. The study aims to identify conserved lncRNAs across the human and mouse genomes. Utilizing the expression data, differentially expressed genes between sample groups, for which changes can be observed among the different treatment methods and EMT-MET, will be retained as having active genes in these processes. The computational analyses filtered these identified lncRNAs for elements like ORFs typical of protein-coding transcripts and assessed orthology between human and mouse genomes thereafter.

Through this analysis, in the MCF7 and MCF10A cell lines, 719 previously unannotated lncRNAs were identified, establishing the existence of a significant number of lncRNAs that had not been characterized in MCF7 and MCF10A cell lines. Notably, many of the newly identified human lncRNAs consist of single exons. This observation may be influenced by the possibility that the current genome annotation does not capture the full range of transcript isoforms, which could contribute to the discovery of novel single-exon transcripts (P. Liu et al., 2020; Wan et al., 2011). As researchers continue to apply more advanced and comprehensive transcriptome analysis techniques, a greater number of these simpler structures are becoming evident (J. Ma et al., 2022). Furthermore, alternative splicing mechanisms, including splicing events that exclude introns for some genes to yield these single exon forms, might further add to the predominance of single-exon transcripts (Gill et al., 2023; Sturgeon & Gardiner, 2012). Further synteny analysis of the human genome, enhanced with these identified lncRNAs, compared to the mouse genome, which contains a total of

15511 lncRNAs—6806 of which were previously unannotated—led to the identification of 30024 previously unannotated orthologous transcripts. Among these, 116 possible orthologs overlapped with the previously unannotated human lncRNA transcripts, further supporting the strong conservation of these orthologs between human and mouse. These findings further support the presence of lncRNAs taking role in EMT-MET processes and underscore the conservation of these potential transcripts between species with close evolutionary relationships. During conservation analysis, clear changes in differential expression profiles of EMT versus MET and their co-inclusion within the same genes further support orthological relationships. However, no single tool has yet been defined that provides a golden ratio for synteny analysis of non-coding elements and closely linked species (D. Liu et al., 2018).

In the context of identifying orthologous lncRNAs between human and mouse, the choice of computational tools is crucial due to the unique challenges as their rapid evolution, variability in length, and lower sequence conservation presented by lncRNAs.Even though synteny analysis of the previously unannotated lncRNAs was performed with Ortho2align, other tools such as lncEVO, i-ADHoRe, and LincOFinder were considered for evaluation where each tool has its strengths and limitations.

lncEVO is one such comprehensive pipeline for identifying conserved lncRNAs. It has modules specifically dedicated to transcriptome assembly, prediction of lncRNA, and conservation studies, with a primary focus on sequence and coding potential (Bryzghalov et al., 2021). While highly proficient at distinguishing lncRNAs from other RNA types, such as mRNAs, using tools like CPAT, lncEVO still suffers from intrinsic limitations due to its reliance on sequence conservation and coding potential and frequently misses orthologues that are primarily conserved for their genomic context rather than their actual sequence. Additionally, lncEVO does not integrate with other tools, hindering its utility in broader analyses. i-ADHoRe excels in identifying genomic homology and collinear regions, with detailed insights on synteny relationships. It utilizes a GHM and profile search to identify collinear regions, which is remarkably accurate for protein-coding genes (Proost et al., n.d.). However, this approach is less effective for non-coding elements like lncRNAs, as its algorithms

are more tailored to protein-coding genes. LincOFinder represents a significant advancement by overcoming the limitations of sequence-based approaches and offering a powerful tool for uncovering conserved lncRNAs (Herrera-Úbeda et al., 2019). However, in our research, LincOFinder presented limitations due to its design for identifying conserved lincRNAs over long evolutionary distances. This approach reduces its precision for species with close evolutionary relationships, such as human and mouse. Additionally, LincOFinder uses a fixed number of upstream and downstream genes to define syntenic blocks, which may not account for larger or more complex syntenic regions.

While tools like lncEVO, i-ADHoRe, and LincOFinder offer valuable capabilities for studying lncRNA evolution, Ortho2align's synteny-based approach, versatility, and ability to identify unannotated orthologues make it a more suitable choice for the analysis of human and mouse lncRNAs. Future enhancements, such as integrating sophisticated statistical methods and machine learning algorithms, could further augment Ortho2align's effectiveness in this field.

One of the major drawbacks of this research is the limited number of processes examined using the MCF7 and MCF10A sample groups. Future studies should include an increased number of samples for validation purposes. Additionally, since lncRNAs exhibit lower expression compared to protein-coding elements, analyses using samples believed to yield higher reads during alignment would show a more homogeneous distribution of transcripts' exon numbers (Statello et al., 2020). However, this study has laid a strong foundation for accurate transcript discovery by utilizing high-quality RNA-Seq data, with sufficient sequencing depth and coverage to capture low-abundance transcripts, such as non-coding elements. Sophisticated computational tools were employed to accurately assemble and quantify transcripts. These tools integrate both alignment-based and de novo assembly approaches, allowing for the effective reconstruction of transcript structures (Harris Robert S., 2007; X. Li et al., 2016). These of algorithms may be further empowered by machine learning and network flow models to optimize splicing and thereby form further steps toward the improvement in the identification of novel transcripts (Baek et al., 2018). To enhance the accuracy of transcript and conserved lncRNA discovery, we combine data from

various conditions, providing contextual insights that strengthen the discovery of transcript structures. Although these computational predictions are a good starting point, they must be experimentally validated through methods like RT-PCR and Northern blotting that will validate the presence and structure of the predicted transcripts for biological relevance in our findings (Mao et al., 2021).

Furthermore, comprehensive annotation of the discovered transcripts is essential to understanding their biological significance. This includes investigating expression patterns across different tissues or developmental stages. These transcripts may require functional assays to find out about their cellular roles. Although transcript discovery protocols are constantly improving, especially as sequencing technologies and computational methods evolve, we currently use state-of-the-art approaches to address the complexities of transcriptomes, such as alternative splicing and post-transcriptional modifications. Collaboration within the scientific community and the use of shared databases will continue to improve the discovery and validation of novel transcript structures. As part of this, the functions of previously unannotated orthologous lncRNAs should be further investigated, once experimentally validated, to determine their roles in biological processes.

6. CONCLUSION AND SUGGESTIONS

•This analysis in MCF7 and MCF10A cell lines resulted in the identification of 719 previously unannotated lncRNAs, which indicate the existence of a considerable number of lncRNAs that were previously uncharacterized in these lines of MCF7 and MCF10A cells.

•Synteny analysis of the human genome was increased with currently unannotated lncRNAs to the total 9963 mouse lncRNAs; among them are 6806 previously unannotated, we are able to identify a total of 30024 previously unannotated orthologous transcripts.

•Of the possible orthologs, 116 overlapped with human lncRNA transcript previously unannotated, further enhancing the excellent conservation of the orthologs found between human and mouse.

•By incorporating additional SRA samples into the MCF7-MCF10A data, 5566 previously unannotated lncRNAs were identified, thereby expanding the dataset and enhancing the robustness of the analysis.

•Synteny analysis with the enhanced human genome by additional SRA samples resulted in the discovery of 31698 potential orthologous transcripts, thus further confirming the identification of a high number of conserved lncRNAs.

•Among these identified potential orthologous regions, 126 were found to overlap with the previously unannotated human lncRNAs, providing further evidence of conservation and potential functional significance.

•Analyzing the merged SRA samples and MCF7-MCF10A data as a single fastq file led to the identification of 16735 previously unannotated lncRNAs, significantly increasing the number of lncRNAs identified and providing a comprehensive view of lncRNA diversity.

•Synteny analysis with these previously unannotated lncRNAs and the enhanced human genome identified 30999 potential orthologous transcripts.

•Among these potential orthologous regions, 12807 overlap with previously unannotated human lncRNAs.

•The involvement of these orthologous lncRNAs in EMT-MET processes should be experimentally validated. For this purpose, quantitative PCR may confirm the expression levels of the found previously unannotated lncRNAs.

•Functional analysis using knockdown or overexpression experiments can be applied to define the participation of mentioned orthologous ncRNAs in EMT-MET regulation.

7. REFERENCES

Ahmad, M., Weiswald, L. B., Poulain, L., Denoyelle, C., & Meryet-Figuiere, M. (2023). Involvement of lncRNAs in cancer cells migration, invasion and metastasis: cytoskeleton and ECM crosstalk. *Journal of Experimental & Clinical Cancer Research 2023 42:1*, *42*(1), 1–14. https://doi.org/10.1186/S13046-023-02741-X

Akhmetkaliyev, A., Alibrahim, N., Shafiee, D., & Tulchinsky, E. (2023). EMT/MET plasticity in cancer and Go-or-Grow decisions in quiescence: the two sides of the same coin? *Molecular Cancer 2023 22:1*, *22*(1), 1–16. https://doi.org/10.1186/S12943-023-01793-Z

Aliperti, V., Skonieczna, J., & Cerase, A. (2021). Long Non-Coding RNA (lncRNA) Roles in Cell Biology, Neurodevelopment and Neurological Disorders. *Non-Coding RNA*, *7*(2). https://doi.org/10.3390/NCRNA7020036

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 (BBA) - Gene Regulatory Mechanisms*, *1849*(6), 731–742. https://doi.org/10.1016/J.BBAGRM.2015.01.005

Amirinejad, R., Rezaei, M., & Shirvani-Farsani, Z. (2020). An update on long intergenic noncoding RNA p21: A regulatory molecule with various significant functions in cancer. *Cell and Bioscience*, *10*(1), 1–9. https://doi.org/10.1186/S13578-020-00445-9/FIGURES/5

Arun, G., Aggarwal, D., & Spector, D. L. (2020). MALAT1 Long Non-Coding RNA: Functional Implications. *Non-Coding RNA*, *6*(2). https://doi.org/10.3390/NCRNA6020022

Babarinde, I. A., & Hutchins, A. P. (2021). *LncGSEA: a versatile tool to infer lncRNA associated pathways from large-scale cancer transcriptome sequencing data*. https://doi.org/10.1186/s12864-022-08717-z

Baek, J., Lee, B., Kwon, S., & Yoon, S. (2018). LncRNAnet: long non-coding RNA identification using deep learning. *Bioinformatics*, *34*(22), 3889–3897. https://doi.org/10.1093/BIOINFORMATICS/BTY418

Bakhtiarizadeh, M. R., & Salami, S. A. (2019). Identification and Expression Analysis of Long Noncoding RNAs in Fat-Tail of Sheep Breeds. *G3 Genes|Genomes|Genetics*, *9*(4), 1263–1276. https://doi.org/10.1534/G3.118.201014

Bao, Z., Yang, Z., Huang, Z., Zhou, Y., Cui, Q., & Dong, D. (2019). LncRNADisease 2.0: an updated database of long non-coding RNA-associated diseases. *Nucleic Acids Research*, *47*(Database issue), D1034. https://doi.org/10.1093/NAR/GKY905

Bhutkar, A., Schaeffer, S. W., Russo, S. M., Xu, M., Smith, T. F., & Gelbart, W. M. (2008). Chromosomal rearrangement inferred from comparisons of 12 drosophila genomes. *Genetics*, *179*(3), 1657–1680. https://doi.org/10.1534/genetics.107.086108

Biyu, H., GuangWen, T., Ming, Z., Lixin, G., & Mengshan, L. (2023). A lncRNA-disease association prediction model based on the two-step PU learning and fully connected neural networks. *Heliyon*, *9*(7), e17726. https://doi.org/10.1016/J.HELIYON.2023.E17726

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114. https://doi.org/10.1093/BIOINFORMATICS/BTU170

Børretzen, A., Gravdal, K., Haukaas, S. A., Mannelqvist, M., Beisland, C., Akslen, L. A., & Halvorsen, O. J. (2021). The epithelial-mesenchymal transition regulators Twist, Slug, and Snail are associated with aggressive tumour features and poor

outcome in prostate cancer patients. *Clinical Research J Pathol Clin Res*, *7*, 253–270. https://doi.org/10.1002/cjp2.202

Breschi, A., Gingeras, T. R., & Guigó, R. (2017). Comparative transcriptomics in human and mouse. *Nature Reviews. Genetics*, *18*(7), 425. https://doi.org/10.1038/NRG.2017.19

Brockdorff, N. (2017). Polycomb complexes in X chromosome inactivation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1733), 663–676. https://doi.org/10.1098/RSTB.2017.0021

Bryzghalov, O., Makałowska, I., & Szcześniak, M. W. (2021). lncEvo: automated identification and conservation study of long noncoding RNAs. *BMC Bioinformatics*, *22*(1), 1–14. https://doi.org/10.1186/S12859-021-03991-2/FIGURES/7

Camilleri-Robles, C., Amador, R., Klein, C. C., Guigó, R., Corominas, M., & Ruiz-Romero, M. (2022). Genomic and functional conservation of lncRNAs: lessons from flies. *Mammalian Genome*, *33*(2), 328. https://doi.org/10.1007/S00335-021-09939-4

Chillón, I., & Marcia, M. (2020a). The molecular structure of long non-coding RNAs: emerging patterns and functional implications. *Critical Reviews in Biochemistry and Molecular Biology*, *55*(6), 662–690. https://doi.org/10.1080/10409238.2020.1828259/ASSET/890BBE92-5667-45C5-867E-5ADD97147142/ASSETS/IMAGES/IBMG_A_1828259_F0005_C.JPG

Chillón, I., & Marcia, M. (2020b). The molecular structure of long non-coding RNAs: emerging patterns and functional implications. *Critical Reviews in Biochemistry and Molecular Biology*, *55*(6), 662—690. https://doi.org/10.1080/10409238.2020.1828259

Choudhury, S., Rathore, A. S., & Raghava, G. P. S. (2024). Compilation of resources on subcellular localization of lncRNA. *Frontiers in RNA Research*, *2*, 1419979. https://doi.org/10.3389/FRNAR.2024.1419979

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M. W., Gaffney, D. J., Elo, L. L., Zhang, X., & Mortazavi, A. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biology*, *17*(1). https://doi.org/10.1186/S13059-016-0881-8

Diederichs, S. (2014). The four dimensions of noncoding RNA conservation. *Trends in Genetics : TIG*, *30*(4), 121–123. https://doi.org/10.1016/J.TIG.2014.01.004

Duran, C., Edwards, D., & Batley, J. (2009). Genetic Maps and the Use of Synteny. *Methods in Molecular Biology*, *513*, 41–55. https://doi.org/10.1007/978-1-59745-427-8_3

Fagundes, N. J. R., Bisso-Machado, R., Figueiredo, P. I. C. C., Varal, M., & Zani, A. L. S. (n.d.). *What We Talk About When We Talk About "Junk DNA."* https://doi.org/10.1093/gbe/evac055

Farrer, R. A. (2017). Synima: A Synteny imaging tool for annotated genome assemblies. *BMC Bioinformatics*, *18*(1), 1–4. https://doi.org/10.1186/S12859-017-1939-7/FIGURES/1

Fatima, R., Akhade, V. S., Pal, D., & Rao, S. M. (2015). Long noncoding RNAs in development and cancer: potential biomarkers and therapeutic targets. *Molecular and Cellular Therapies 2015 3:1*, *3*(1), 1–19. https://doi.org/10.1186/S40591-015-0042-6

Ferrer, J., & Dimitrova, N. (2024). Transcription regulation by long non-coding RNAs: mechanisms and disease relevance. *Nature Reviews. Molecular Cell Biology*, *25*(5), 396–415. https://doi.org/10.1038/S41580-023-00694-9

Frankish, A., Carbonell-Sala, S., Diekhans, M., Jungreis, I., Loveland, J. E., Mudge, J. M., Sisu, C., Wright, J. C., Arnan, C., Barnes, I., Banerjee, A., Bennett, R., Berry, A., Bignell, A., Boix, C., Calvet, F., Cerdán-Velez, D., Cunningham, F.,

Davidson, C., … Flicek, P. (2023). GENCODE: reference annotation for the human and mouse genomes in 2023. *Nucleic Acids Research*, *51*(D1), D942–D949. https://doi.org/10.1093/NAR/GKAC1071

Gandhi, S., Witten, A., De Majo, F., Gilbers, M., Maessen, J., Schotten, U., de Windt, L. J., & Stoll, M. (2021). Evolutionarily conserved transcriptional landscape of the heart defining the chamber specific physiology. *Genomics*, *113*(6), 3782–3792. https://doi.org/10.1016/J.YGENO.2021.09.002

Ghafouri-Fard, S., & Taheri, M. (2019). Maternally expressed gene 3 (MEG3): A tumor suppressor long non coding RNA. *Biomedicine & Pharmacotherapy = Biomedecine & Pharmacotherapie*, *118*. https://doi.org/10.1016/J.BIOPHA.2019.109129

Ghanam, A. R., Bryant, W. B., & Miano, J. M. (2022). Of mice and human-specific long noncoding RNAs. *Mammalian Genome*, *33*(2), 281. https://doi.org/10.1007/S00335-022-09943-2

Gholami, M., Klashami, Z. N., Ebrahimi, P., Mahboobipour, A. A., Salehi Farid, A., Vahidi, A., Zoughi, M., Asadi, M., & Amoli, M. M. (2023). Metformin and long non-coding RNAs in breast cancer. *Journal of Translational Medicine*, *21*, 155. https://doi.org/10.1186/s12967-023-03909-x

Gill, M. E., Rohmer, A., Erkek-Ozhan, S., Liang, C. Y., Chun, S., Ozonov, E. A., & Peters, A. H. F. M. (2023). De novo transcriptome assembly of mouse male germ cells reveals novel genes, stage-specific bidirectional promoter activity, and noncoding RNA expression. *Drug Metabolism and Disposition*, *33*(12), 2060–2078. https://doi.org/10.1101/gr.278060.123

Gonzalez, D. M., & Medici, D. (2014). Signaling mechanisms of the epithelial-mesenchymal transition. *Science Signaling*, *7*(344), re8. https://doi.org/10.1126/SCISIGNAL.2005189

Gugnoni, M., & Ciarrocchi, A. (2019). Long Noncoding RNA and Epithelial Mesenchymal Transition in Cancer. *International Journal of Molecular Sciences*, *20*(8). https://doi.org/10.3390/IJMS20081924

Gutschner, T., Hämmerle, M., Eißmann, M., Hsu, J., Kim, Y., Hung, G., Revenko, A., Arun, G., Stentrup, M., Groß, M., Zörnig, M., MacLeod, A. R., Spector, D. L., & Diederichs, S. (2013). The noncoding RNA MALAT1 is a critical regulator of the metastasis phenotype of lung cancer cells. *Cancer Research*, *73*(3), 1180–1189. https://doi.org/10.1158/0008-5472.CAN-12-2850

Hall, J. R., Messenger, Z. J., Tam, H. W., Phillips, S. L., Recio, L., & Smart, R. C. (2015). Long noncoding RNA lincRNA-p21 is the major mediator of UVB-induced and p53-dependent apoptosis in keratinocytes. *Cell Death & Disease*, *6*(3), e1700. https://doi.org/10.1038/CDDIS.2015.67

Hao, Y., Baker, D., & Dijke, P. Ten. (2019). TGF-β-Mediated Epithelial-Mesenchymal Transition and Cancer Metastasis. *International Journal of Molecular Sciences*, *20*(11). https://doi.org/10.3390/IJMS20112767

Harris Robert S. (2007). *IMPROVED PAIRWISE ALIGNMENT OF GENOMIC DNA*.

Hatje, K., Mühlhausen, S., Simm, D., & Kollmar, M. (2019). The Protein-Coding Human Genome: Annotating High-Hanging Fruits. *BioEssays : News and Reviews in Molecular, Cellular and Developmental Biology*, *41*(11). https://doi.org/10.1002/BIES.201900066

Herrera-Úbeda, C., Barba, M. M., Pérez, E. N., Gravemeyer, J., Albuixech-Crespo, B., Wheeler, G. N., & Garcia-Fernàndez, J. (2019). Microsyntenic Clusters Reveal Conservation of lncRNAs in Chordates Despite Absence of Sequence Conservation. *Biology*, *8*(3). https://doi.org/10.3390/BIOLOGY8030061

Hossain, M. E., Cevallos, R. R., Zhang, R., & Hu, K. (2022). Attenuating iPSC reprogramming stress with dominant-negative BET peptides. *IScience*, *26*(1). https://doi.org/10.1016/J.ISCI.2022.105889

Hussein, M. A., Valinezhad, K., Adel, E., & Munirathinam, G. (2024). MALAT-1 Is a Key Regulator of Epithelial-Mesenchymal Transition in Cancer: A Potential Therapeutic Target for Metastasis. *Cancers*, *16*(1). https://doi.org/10.3390/CANCERS16010234

Ilangumaran, S., Villalobos-Hernandez, A., Bobbala, D., & Ramanathan, S. (2016). The hepatocyte growth factor (HGF)-MET receptor tyrosine kinase signaling pathway: Diverse roles in modulating immune cell functions. *Cytokine*, *82*, 125–139. https://doi.org/10.1016/J.CYTO.2015.12.013

Jia, Z., An, J., Liu, Z., & Zhang, F. (2022). Non-Coding RNAs in Colorectal Cancer: Their Functions and Mechanisms. In *Frontiers in Oncology* (Vol. 12). Frontiers Media S.A. https://doi.org/10.3389/fonc.2022.783079

Jiang, J., Wang, K., Chen, Y., Chen, H., Nice, E. C., & Huang, C. (2017). Redox regulation in tumor cell epithelial–mesenchymal transition: molecular basis and therapeutic strategy. *Signal Transduction and Targeted Therapy*, *2*, 17036. https://doi.org/10.1038/SIGTRANS.2017.36

Jiang, X., Fu, X., Dong, G., & Li, H. (2017). *Research on Pairwise Sequence Alignment Needleman-Wunsch Algorithm*. 1041–1046. https://doi.org/10.2991/ICMMCCE-17.2017.187

Jin, H., Du, W., Huang, W., Yan, J., Tang, Q., Chen, Y., & Zou, Z. (2021). lncRNA and breast cancer: Progress from identifying mechanisms to challenges and opportunities of clinical treatment. *Molecular Therapy. Nucleic Acids*, *25*, 613. https://doi.org/10.1016/J.OMTN.2021.08.005

Kaikkonen, M. U., Lam, M. T. Y., & Glass, C. K. (2011). Non-coding RNAs as regulators of gene expression and epigenetics. *Cardiovascular Research*, *90*(3), 430–440. https://doi.org/10.1093/CVR/CVR097

Kalluri, R. (2009). EMT: when epithelial cells decide to become mesenchymal-like cells. *The Journal of Clinical Investigation*, *119*(6), 1417–1419. https://doi.org/10.1172/JCI39675

Kalluri, R., & Weinberg, R. A. (2009). The basics of epithelial-mesenchymal transition. *The Journal of Clinical Investigation*, *119*(6), 1420–1428. https://doi.org/10.1172/JCI39104

Karolchik, D., Hinricks, A. S., Furey, T. S., Roskin, K. M., Sugnet, C. W., Haussler, D., & Kent, W. J. (2004). The UCSC Table Browser data retrieval tool. *Nucleic Acids Research*, *32*(Database issue), D493. https://doi.org/10.1093/NAR/GKH103

Kemkemer, C., Kohn, M., Cooper, D. N., Froenicke, L., Högel, J., Hameister, H., & Kehrer-Sawatzki, H. (2009). Gene synteny comparisons between different vertebrates provide new insights into breakage and fusion events during mammalian karyotype evolution. *BMC Evolutionary Biology*, *9*(1). https://doi.org/10.1186/1471-2148-9-84

Kern, C., Wang, Y., Chitwood, J., Korf, I., Delany, M., Cheng, H., Medrano, J. F., Van Eenennaam, A. L., Ernst, C., Ross, P., & Zhou, H. (2018). Genome-wide identification of tissue-specific long non-coding RNA in three farm animal species 06 Biological Sciences 0604 Genetics. *BMC Genomics*, *19*(1), 1–14. https://doi.org/10.1186/S12864-018-5037-7/FIGURES/4

Kim, D. H., Xing, T., Yang, Z., Dudek, R., Lu, Q., & Chen, Y. H. (2017). Epithelial Mesenchymal Transition in Embryonic Development, Tissue Repair and Cancer: A Comprehensive Overview. *Journal of Clinical Medicine*, *7*(1). https://doi.org/10.3390/JCM7010001

Kim, D., Paggi, J. M., Park, C., Bennett, C., & Salzberg, S. L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology 2019 37:8*, *37*(8), 907–915. https://doi.org/10.1038/s41587-019-0201-4

Kim, H. Y., Jackson, T. R., & Davidson, L. A. (2017). On the role of mechanics in driving mesenchymal-to-epithelial transitions. *Seminars in Cell & Developmental Biology*, *67*, 113–122. https://doi.org/10.1016/J.SEMCDB.2016.05.011

Kurian, L., Aguirre, A., Sancho-Martinez, I., Benner, C., Hishida, T., Nguyen, T. B., Reddy, P., Nivet, E., Krause, M. N., Nelles, D. A., Esteban, C. R., Campistol, J. M., Yeo, G. W., & Belmonte, J. C. I. (2015). Identification of novel long noncoding RNAs underlying vertebrate cardiovascular development. *Circulation*, *131*(14), 1278–1290. https://doi.org/10.1161/CIRCULATIONAHA.114.013303

Lamouille, S., & Derynck, R. (2011). Emergence of the phosphoinositide 3-kinase-Akt-mammalian target of rapamycin axis in transforming growth factor-β-induced epithelial-mesenchymal transition. *Cells, Tissues, Organs*, *193*(1–2), 8–22. https://doi.org/10.1159/000320172

Lee, J. H., & Massagué, J. (2022). TGF-β in developmental and fibrogenic EMTs. *Seminars in Cancer Biology*, *86*(Pt 2), 136–145. https://doi.org/10.1016/J.SEMCANCER.2022.09.004

Lee, J. T., & Bartolomei, M. S. (2013). X-Inactivation, Imprinting, and Long Noncoding RNAs in Health and Disease. *Cell*, *152*(6), 1308–1323. https://doi.org/10.1016/J.CELL.2013.02.016

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078. https://doi.org/10.1093/BIOINFORMATICS/BTP352

Li, X., Kong, Y., Zhao, Q., Li, Y.-Y., & Hao, P. (2016). De novo assembly of transcriptome from next-generation sequencing data. *Quantitative Biology*, *4*, 94–105. https://api.semanticscholar.org/CorpusID:9229389

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 Research*, *47*(8). https://doi.org/10.1093/NAR/GKZ114

Lin, X., Huang, X., Wang, L., & Liu, W. (2022). The long noncoding RNA MALAT1/microRNA-598-3p axis regulates the proliferation and apoptosis of

retinoblastoma cells through the PI3K/AKT pathway. *Molecular Vision*, *28*, 269. /pmc/articles/PMC9514550/

Lisec, B., Bozic, T., Santek, I., Markelc, B., Vrecl, M., Frangez, R., & Cemazar, M. (2024). Characterization of two distinct immortalized endothelial cell lines, EA.hy926 and HMEC-1, for in vitro studies: exploring the impact of calcium electroporation, Ca2+ signaling and transcriptomic profiles. *Cell Communication and Signaling*, *22*(1), 1–24. https://doi.org/10.1186/S12964-024-01503-2/FIGURES/10

Liu, B., Xiang, W., Liu, J., Tang, J., Wang, J., Liu, B., Long, Z., Wang, L., Yin, G., & Liu, J. (2021). The regulatory role of antisense lncRNAs in cancer. *Cancer Cell International 2021 21:1*, *21*(1), 1–15. https://doi.org/10.1186/S12935-021-02168-4

Liu, D., Hunt, M., & Tsai, I. J. (2018). Inferring synteny between genome assemblies: A systematic evaluation. *BMC Bioinformatics*, *19*(1), 1–13. https://doi.org/10.1186/S12859-018-2026-4/FIGURES/9

Liu, P., Soukup, A. A., Bresnick, E. H., Dewey, C. N., & Keles, S. (2020). PRAM: A novel pooling approach for discovering intergenic transcripts from large-scale RNA sequencing experiments. *Genome Research*, *30*(11), 1655–1666. https://doi.org/10.1101/GR.252445.119/-/DC1

Long, Y., Wang, X., Youmans, D. T., & Cech, T. R. (2017). How do lncRNAs regulate transcription? *Science Advances*, *3*(9). https://doi.org/10.1126/SCIADV.AAO2110

Lyon, M. F. (1992). SOME MILESTONES IN THE HISTORY OF X-CHROMOSOME INACTIVATION. In *Annu. Rev. Genet* (Vol. 26). www.annualreviews.org.

Ma, J., Wu, J. Y., & Zhu, L. (2022). Detection of orthologous exons and isoforms using EGIO. *Bioinformatics*, *38*(19), 4474–4480. https://doi.org/10.1093/bioinformatics/btac548

Ma, W., Liu, Y., Ma, H., Ren, Z., & Yan, J. (2021). Cis-acting: A pattern of lncRNAs for gene regulation in induced pluripotent stem cells from patients with Down syndrome determined by integrative analysis of lncRNA and mRNA profiling data. *Experimental and Therapeutic Medicine*, *22*(1). https://doi.org/10.3892/ETM.2021.10133

Mallory, A. C., & Shkumatava, A. (2015). LncRNAs in Vertebrates: Advances and Challenges. *Biochimie*, *117*, 3. https://doi.org/10.1016/J.BIOCHI.2015.03.014

Mao, M., Xue, Y., He, Y., Zhou, X., Hu, H., Liu, J., Feng, L., Yang, W., Luo, J., Zhang, H., Li, X., & Ma, J. (2021). Validation of Reference Genes for Quantitative Real-Time PCR Normalization in Ananas comosus var. bracteatus During Chimeric Leaf Development and Response to Hormone Stimuli. *Frontiers in Genetics*, *12*, 716137. https://doi.org/10.3389/FGENE.2021.716137/BIBTEX

Martino, M. De, Esposito, F., & Pallante, P. (2021). Long non-coding RNAs regulating multiple proliferative pathways in cancer cell. *Translational Cancer Research*, *10*(6), 3140. https://doi.org/10.21037/TCR-21-230

Minuti, G., & Landi, L. (2015). MET deregulation in breast cancer. *Annals of Translational Medicine*, *3*(13), 181. https://doi.org/10.3978/J.ISSN.2305-5839.2015.06.22

Mondal, T., Subhash, S., Vaid, R., Enroth, S., Uday, S., Reinius, B., Mitra, S., Mohammed, A., James, A. R., Hoberg, E., Moustakas, A., Gyllensten, U., Jones, S. J. M., Gustafsson, C. M., Sims, A. H., Westerlund, F., Gorab, E., & Kanduri, C. (2015). MEG3 long noncoding RNA regulates the TGF-β pathway genes through formation of RNA–DNA triplex structures. *Nature Communications 2015 6:1*, *6*(1), 1–17. https://doi.org/10.1038/ncomms8743

Mylarshchikov, D. E., & Mironov, A. A. (2022). ortho2align: a sensitive approach for searching for orthologues of novel lncRNAs. *BMC Bioinformatics*, *23*(1), 1–14. https://doi.org/10.1186/S12859-022-04929-Y/FIGURES/5

Paril, J., Zare, T., & Fournier-Level, A. (2023). Compare_Genomes: A Comparative Genomics Workflow to Streamline the Analysis of Evolutionary Divergence Across Eukaryotic Genomes. *Current Protocols*, *3*(8), e876. https://doi.org/10.1002/CPZ1.876

Pegueroles, C., Iraola-Guzmán, S., Chorostecki, U., Ksiezopolska, E., Saus, E., & Gabaldón, T. (2019). Transcriptomic analyses reveal groups of co-expressed, syntenic lncRNAs in four species of the genus Caenorhabditis. *RNA Biology*, *16*(3), 320. https://doi.org/10.1080/15476286.2019.1572438

Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T. C., Mendell, J. T., & Salzberg, S. L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology*, *33*(3), 290–295. https://doi.org/10.1038/NBT.3122

Portoso, M., Ragazzini, R., Brenčič, Ž., Moiani, A., Michaud, A., Vassilev, I., Wassef, M., Servant, N., Sargueil, B., & Margueron, R. (2017). PRC2 is dispensable for HOTAIR-mediated transcriptional repression. *The EMBO Journal*, *36*(8), 981–994. https://doi.org/10.15252/EMBJ.201695335

Proost, S., Fostier, J., De Witte, D., Dhoedt, B., Demeester, P., Van De Peer, Y., & Vandepoele, K. (n.d.). *i-ADHoRe 3.0-fast and sensitive detection of genomic homology in extremely large data sets*. https://doi.org/10.1093/nar/gkr955

Qin, D. (2019). Next-generation sequencing and its clinical application. *Cancer Biology & Medicine*, *16*(1), 4. https://doi.org/10.20892/J.ISSN.2095-3941.2018.0055

Quigley, S., Damas, J., Larkin, D. M., & Farré, M. (2023). syntenyPlotteR: a user-friendly R package to visualize genome synteny, ideal for both experienced and novice bioinformaticians. *Bioinformatics Advances*, *3*(1). https://doi.org/10.1093/BIOADV/VBAD161

Raghavan, V., Kraft, L., Mesny, F., & Rigerte, L. (n.d.). A simple guide to de novo transcriptome assembly and annotation. *Briefings in Bioinformatics*, *2022*(2), 1–30. https://doi.org/10.1093/bib/bbab563

Rao, M. S., Van Vleet, T. R., Ciurlionis, R., Buck, W. R., Mittelstadt, S. W., Blomme, E. A. G., & Liguori, M. J. (2019). Comparison of RNA-Seq and Microarray Gene Expression Platforms for the Toxicogenomic Evaluation of Liver From Short-Term Rat Toxicity Studies. *Frontiers in Genetics*, *9*(JAN). https://doi.org/10.3389/FGENE.2018.00636

Ratti, M., Lampis, A., Ghidini, M., Salati, M., Mirchev, M. B., Valeri, N., & Hahne, J. C. (2020). MicroRNAs (miRNAs) and Long Non-Coding RNAs (lncRNAs) as New Tools for Cancer Therapy: First Steps from Bench to Bedside. *Targeted Oncology*, *15*(3), 261–278. https://doi.org/10.1007/S11523-020-00717-X

Renwick, J. H. (1971). The Rhesus syntenic group in man. *Nature*, *234*(5330), 475. https://doi.org/10.1038/234475A0

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*(1), 139–140. https://doi.org/10.1093/BIOINFORMATICS/BTP616

Rödelsperger, C., & Dieterich, C. (2010). CYNTENATOR: Progressive Gene Order Alignment of 17 Vertebrate Genomes. *PLoS ONE*, *5*(1), 8861. https://doi.org/10.1371/JOURNAL.PONE.0008861

Ruiz-Orera, J., & Albà, M. M. (2019). Conserved regions in long non-coding RNAs contain abundant translation and protein–RNA interaction signatures. *NAR Genomics and Bioinformatics*, *1*(1). https://doi.org/10.1093/nargab/lqz002

Satam, H., Joshi, K., Mangrolia, U., Waghoo, S., Zaidi, G., Rawool, S., Thakare, R. P., Banday, S., Mishra, A. K., Das, G., & Malonia, S. K. (2023). Next-Generation Sequencing Technology: Current Trends and Advancements. *Biology*, *12*(7). https://doi.org/10.3390/BIOLOGY12070997

Sistigu, A., Di Modugno, F., Manic, G., & Nisticò, P. (2017). Deciphering the loop of epithelial-mesenchymal transition, inflammatory cytokines and cancer immunoediting. *Cytokine & Growth Factor Reviews*, *36*, 67–77. https://doi.org/10.1016/J.CYTOGFR.2017.05.008

Soderlund, C., Bomhoff, M., & Nelson, W. M. (2011). SyMAP v3.4: a turnkey synteny system with application to plant genomes. *Nucleic Acids Research*, *39*(10), e68–e68. https://doi.org/10.1093/NAR/GKR123

Song, J. (2020). Targeting epithelial-mesenchymal transition pathway in hepatocellular carcinoma. *Clinical and Molecular Hepatology*, *26*(4), 484. https://doi.org/10.3350/CMH.2020.0220

Statello, L., Guo, C. J., Chen, L. L., & Huarte, M. (2020). Gene regulation by long non-coding RNAs and its biological functions. *Nature Reviews Molecular Cell Biology 2020 22:2*, *22*(2), 96–118. https://doi.org/10.1038/s41580-020-00315-9

Sturgeon, X. H., & Gardiner, K. J. (2012). RCDA: A highly sensitive and specific alternatively spliced transcript assembly tool featuring upstream consecutive exon structures. *Genomics*, *100*(6), 357–362. https://doi.org/10.1016/J.YGENO.2012.08.004

Sychev, Z. E., Hu, A., DiMaio, T. A., Gitter, A., Camp, N. D., Noble, W. S., Wolf-Yadlin, A., & Lagunoff, M. (2017). Integrated systems biology analysis of KSHV latent infection reveals viral induction and reliance on peroxisome mediated lipid metabolism. *PLoS Pathogens*, *13*(3). https://doi.org/10.1371/JOURNAL.PPAT.1006256

Tesfaye, E., Martinez-Terroba, E., Bendor, J., Winkler, L., Olivero, C., Chen, K., Feldser, D. M., Zamudio, J. R., & Dimitrova, N. (2021). The p53 transcriptional response across tumor types reveals core and senescence-specific signatures modulated by long noncoding RNAs. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(31). https://doi.org/10.1073/PNAS.2025539118/-/DCSUPPLEMENTAL

Thapa, R., Afzal, O., Afzal, M., Gupta, G., Bhat, A. A., Hassan almalki, W., Kazmi, I., Alzarea, S. I., Saleem, S., Arora, P., Singh, S. K., & Dua, K. (2024). From LncRNA to metastasis: The MALAT1-EMT axis in cancer progression. *Pathology - Research and Practice*, *253*, 154959. https://doi.org/10.1016/J.PRP.2023.154959

Tripathi, R., Chakraborty, P., & Varadwaj, P. K. (2017). Unraveling long non-coding RNAs through analysis of high-throughput RNA-sequencing data. *Non-Coding RNA Research*, *2*(2), 111. https://doi.org/10.1016/J.NCRNA.2017.06.003

Uroda, T., Anastasakou, E., Rossi, A., Teulon, J. M., Pellequer, J. L., Annibale, P., Pessey, O., Inga, A., Chillón, I., & Marcia, M. (2019). Conserved Pseudoknots in lncRNA MEG3 Are Essential for Stimulation of the p53 Pathway. *Molecular Cell*, *75*(5), 982. https://doi.org/10.1016/J.MOLCEL.2019.07.025

Wan, Y., Kertesz, M., Spitale, R. C., Segal, E., & Chang, H. Y. (2011). Understanding the transcriptome through RNA structure. *Nature Reviews. Genetics*, *12*(9), 641–655. https://doi.org/10.1038/NRG3049

Wang, L., Park, H. J., Dasari, S., Wang, S., Kocher, J. P., & Li, W. (2013). CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic Acids Research*, *41*(6), e74. https://doi.org/10.1093/NAR/GKT006

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews. Genetics*, *10*(1), 57. https://doi.org/10.1038/NRG2484

Watanabe, K., Liu, Y., Noguchi, S., Murray, M., Chang, J. C., Kishima, M., Nishimura, H., Hashimoto, K., Minoda, A., & Suzuki, H. (2019). OVOL2 induces mesenchymal-to-epithelial transition in fibroblasts and enhances cell-state reprogramming towards epithelial lineages. *Scientific Reports*, *9*(1). https://doi.org/10.1038/S41598-019-43021-Z

Waters, E., Pucci, P., Hirst, M., Chapman, S., Wang, Y., Crea, F., & Heath, C. J. (2021). HAR1: An insight into lncRNA genetic evolution. *Epigenomics*, *13*(22), 1831–1843. https://doi.org/10.2217/EPI-2021-0069

Weghorst, F., Torres Marcén, M., Faridi, G., Lee, Y. C. G., & Cramer, K. S. (2024). Deep Conservation and Unexpected Evolutionary History of Neighboring lncRNAs MALAT1 and NEAT1. *Journal of Molecular Evolution*, *92*(1), 30–41. https://doi.org/10.1007/S00239-023-10151-Y

Winkler, L., & Dimitrova, N. (2022). A mechanistic view of long noncoding RNAs in cancer. *Wiley Interdisciplinary Reviews. RNA*, *13*(3). https://doi.org/10.1002/WRNA.1699

Wojciech Szczé sniak, M., Regina Kubiak, M., Wanowska, zbieta, Makalowska, I., & Wojciech Szczé, M. (2021). Comparative genomics in the search for conserved long noncoding RNAs. *Essays in Biochemistry*, *65*, 741–749. https://doi.org/10.1042/EBC20200069

Xu, J., Wang, X., Zhu, C., & Wang, K. (2022). A review of current evidence about lncRNA MEG3: A tumor suppressor in multiple cancers. *Frontiers in Cell and Developmental Biology*, *10*. https://doi.org/10.3389/FCELL.2022.997633

Ye, D., Deng, Y., & Shen, Z. (2021a). *The Role and Mechanism of MALAT1 Long Non-Coding RNA in the Diagnosis and Treatment of Head and Neck Squamous Cell Carcinoma*. https://doi.org/10.2147/OTT.S317234

Ye, D., Deng, Y., & Shen, Z. (2021b). The Role and Mechanism of MALAT1 Long Non-Coding RNA in the Diagnosis and Treatment of Head and Neck Squamous Cell Carcinoma. *OncoTargets and Therapy*, *14*, 4127. https://doi.org/10.2147/OTT.S317234

Ye, J., Li, H., Wei, J., Luo, Y., Liu, H., Zhang, J., & Luo, X. (2020). Risk Scoring System based on lncRNA Expression for Predicting Survival in Hepatocellular Carcinoma with Cirrhosis. *Asian Pacific Journal of Cancer Prevention : APJCP*, *21*(6), 1787. https://doi.org/10.31557/APJCP.2020.21.6.1787

Zhang, X., Wang, W., Zhu, W., Dong, J., Cheng, Y., Yin, Z., & Shen, F. (2019). Mechanisms and Functions of Long Non-Coding RNAs at Multiple Regulatory Levels. *International Journal of Molecular Sciences*, *20*(22). https://doi.org/10.3390/IJMS20225573

Zhao, L., Wang, J., Li, Y., Song, T., Wu, Y., Fang, S., Bu, D., Li, H., Sun, L., Pei, D., Zheng, Y., Huang, J., Xu, M., Chen, R., Zhao, Y., & He, S. (2021). NONCODEV6: an updated database dedicated to long non-coding RNA annotation in both animals and plants. *Nucleic Acids Research*, *49*(D1), D165. https://doi.org/10.1093/NAR/GKAA1046

Zhao, Y. B., & Krishnan, J. (2014). mRNA translation and protein synthesis: an analysis of different modelling methodologies and a new PBN based approach. *BMC Systems Biology*, *8*(1), 25. https://doi.org/10.1186/1752-0509-8-25/TABLES/1

Zhao, Y., Teng, H., Yao, F., Yap, S., Sun, Y., & Ma, L. (2020). Challenges and Strategies in Ascribing Functions to Long Noncoding RNAs. *Cancers 2020, Vol. 12, Page 1458*, *12*(6), 1458. https://doi.org/10.3390/CANCERS12061458

Zheng, X. H., Lu, F., Wang, Z. Y., Zhong, F., Hoover, J., & Mural, R. (2005). Using shared genomic synteny and shared protein functions to enhance the identification of orthologous gene pairs. *Bioinformatics (Oxford, England)*, *21*(6), 703–710. https://doi.org/10.1093/BIOINFORMATICS/BTI045

APPENDIX 1.

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

Toplantı Tarihi	: 28/07/2023	**Toplantı Günü**	: Cuma
Toplantı Sayısı	: 8	**Toplantı Saati**	: 10:00

Sayın Doç. Dr. Gökhan KARAKULAH,

2023-022 Protokol No'lu; sorumlusu olduğunuz "Epitel-Mezenkimal Geçiş ve Mezenkimal-Epitel Geçiş Süreçlerinde Uzun Kodlanmayan RNA'ların İncelenmesi: İnsan ve Fare Genomları Arasında Sinteni Analizi" 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** (Katılmadı)
Prof. Dr. Sedef AKGÜNGÖR **Üye** (Katılmadı)	**Prof. Dr. Gülgün OKTAY** **Üye**
Dr. Serap ERKEK ÖZHAN **Üye**	**Prof. Dr. Şermin GENÇ** **Üye**
Doç. Dr. Canan Aslı YILDIRIM **Üye** (Katılmadı)	**Doç. Dr. Can KÜÇÜK** **Üye**
Dr. Öğr. Üyesi Yavuz OKTAY **Üye**	

APPENDIX 2.

CURRİCULUM VİTAE

Name	Elif YILDIZ
ID	306
Birth Year	18
Correspondence Address	Dokuz Eylül University State Conservatory, İzmir Biomedicine and Genome Center, 35340 İzmir
Phone	+99
e-mail	kgmail.com

EDUCATIONAL INFORMATION

Country	**University**	**Academic Unit**	**Department**	**Degree**	**Graduation Year**
Türkiye	Izmir Institute of Technology	Faculty of Science	Molecular Biology and Genetics	Bachelor`s Degree	2022
Türkiye	Dokuz Eylul University	Izmir International Biomedicine and Genome Institute	Biomedicine and Health Technologies	Master`s Degree	2024