

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

**IDENTIFICATION AND ANNOTATION OF
PUTATIVE LONG NONCODING RNAS
INVOLVED IN MESENCHYMAL-
EPITHELIAL TRANSITION**

DOĞA ESKİER

DEPARTMENT OF MOLECULAR BIOLOGY AND
GENETICS
MASTER OF SCIENCE THESIS

İZMİR – 2019

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

**IDENTIFICATION AND ANNOTATION OF
PUTATIVE LONG NONCODING RNAS
INVOLVED IN MESENCHYMAL-
EPITHELIAL TRANSITION**

DEPARTMENT OF MOLECULAR BIOLOGY AND
GENETICS

MASTER OF SCIENCE THESIS

DOĞA ESKİER

SUPERVISOR: ASSOC. PROF. GÖKHAN KARAKÜLAH

CO-SUPERVISOR: ASST. PROF. HANI ALOTAIBI

Dokuz Eylül Üniversitesi İzmir Uluslararası Biyotıp ve Genom Enstitüsü Genom Bilimleri ve Moleküler Biyoteknoloji Anabilim Dalı,
Moleküler Biyoloji ve Genetik Yüksek Lisans programı öğrencisi Doğa Eskier
‘IDENTIFICATION AND ANNOTATION OF LONG NONCODING RNAS INVOLVED IN MESENCHYMAL-EPITHELIAL TRANSITION’ konulu
Yüksek Lisans tezini 03 / 01 / 2019 tarihinde başarılı olarak tamamlamıştır.

Doç. Dr. Gökhan KARAKÜLAH
Dokuz Eylül Üniversitesi
BAŞKAN

Dr. Öğr. Üye. Çiğdem Eresen YAZICIOĞLU
Dokuz Eylül Üniversitesi
ÜYE

Dr. Öğr. Üye. Cihangir YANDIM
İzmir Ekonomi Üniversitesi
ÜYE

Dr. Öğr. Üye. Ezgi KARACA
Dokuz Eylül Üniversitesi
YEDEK ÜYE

Dr. Öğr. Üye. Ayşe Banu DEMİR
İzmir Ekonomi Üniversitesi
YEDEK ÜYE

Dokuz Eylül University İzmir International Biomedicine and Genome Enstitute
Department of Genomics and Molecular Biotechnology,
Molecular Biology and Genetics graduate program Master of Science student Doğa
Eskier has successfully completed his Master of Science thesis titled
**‘IDENTIFICATION AND ANNOTATION OF LONG NONCODING RNAS
INVOLVED IN MESENCHYMAL-EPITHELIAL TRANSITION’** on the date
of 03 / 01 / 2019.

Assoc. Prof. Gökhan KARAKÜLAH
Dokuz Eylül University
CHAIR

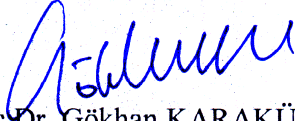
Asst. Prof. Çiğdem Eresen YAZICIOĞLU
Dokuz Eylül University
MEMBER

Asst. Prof. Cihangir Yandım
İzmir University of Economics
MEMBER

Asst. Prof. Ezgi KARACA
Dokuz Eylül University
SUBSTITUTE MEMBER

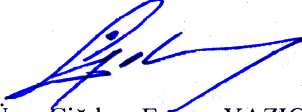
Asst. Prof. Ayşe Banu DEMİR
İzmir University of Economics
SUBSTITUTE MEMBER

Dokuz Eylül Üniversitesi İzmir Uluslararası Biyotıp ve Genom Enstitüsü Genom
Bilimleri ve Moleküler Biyoteknoloji Anabilim Dalı,
Moleküler Biyoloji ve Genetik Yüksek Lisans programı öğrencisi Doğa Eskier
**'IDENTIFICATION AND ANNOTATION OF LONG NONCODING RNAS
INVOLVED IN MESENCHYMAL-EPITHELIAL TRANSITION'** konulu
Yüksek Lisans tezini 03 / 01 / 2019 tarihinde başarılı olarak tamamlamıştır.



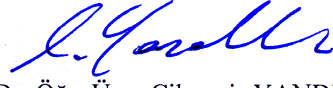
Doç. Dr. Gökhan KARAKÜLAH

BAŞKAN



Dr. Öğr. Üye. Çiğdem Eresen YAZICIOĞLU

ÜYE



Dr. Öğr. Üye. Cihangir YANDIM

ÜYE

Dr. Öğr. Üye. Ezgi KARACA

YEDEK ÜYE

Dr. Öğr. Üye. Ayşe Banu DEMİR

YEDEK ÜYE

Dokuz Eylül University İzmir International Biomedicine and Genome Enstitute
Department of Genomics and Molecular Biotechnology,
Molecular Biology and Genetics graduate program Master of Science student Doğa
Eskier has successfully completed his Master of Science thesis titled
**‘IDENTIFICATION AND ANNOTATION OF LONG NONCODING RNAS
INVOLVED IN MESENCHYMAL-EPITHELIAL TRANSITION’** on the date
of 03 / 01 / 2019.

Dr. Öğr. Üye. Gökhan KARAKÜLAH

CHAIR

Dr. Öğr. Üye. Çiğdem Eresen YAZICIOĞLU

MEMBER

Dr. Öğr. Üye. Cihangir Yandım

MEMBER

Dr. Öğr. Üye. Ezgi KARACA

SUBSTITUTE MEMBER

Dr. Öğr. Üye. Ayşe Banu DEMİR

SUBSTITUTE MEMBER

TABLE OF CONTENTS: **Page Number:**

TABLE OF CONTENTS	i
INDEX OF TABLES	iv
INDEX OF FIGURES	v
LIST OF ABBREVIATIONS	vii
ACKNOWLEDGEMENTS	viii
ABSTRACT	1
ÖZET	2
1. INTRODUCTION AND AIM	3
1.1. Statement and Importance of the Problem	3
1.2. Aim of Study	3
1.3. Hypothesis of Study	3
2. GENERAL INFORMATION	4
2.1. Epithelial-Mesenchymal Transition (EMT)	4
2.2. Mesenchymal-Epithelial Transition (MET)	6
2.3. Long Noncoding RNAs (lncRNAs)	6
2.3.1. <i>lncRNAs in EMT</i>	8
2.3.2. <i>Computational methods for lncRNA discovery and annotation</i>	8
2.3.2.1. <i>Difficulties in experimental approaches in lncRNA research</i>	8
2.3.2.2. <i>Overview of predictive lncRNA annotation methods</i>	9
2.4. Next Generation Sequencing (NGS)	11
2.4.1. <i>Sequencing technologies and their applications</i>	12
2.4.2. <i>Transcriptome research with NGS analysis</i>	14
2.4.2.1. <i>Basic raw data pre-processing and quality assessment</i>	14
2.4.2.2. <i>Alignment of sequencing reads</i>	15

2.4.2.3. Transcriptome assembly and novel transcript discovery	15
2.4.2.4. Transcript / Gene quantification and Differential Expression Analysis	17
2.4.2.5. Construction of co-expression networks for lncRNA research.....	18
3. MATERIALS AND METHODS.....	21
3.1. Type of Study	21
3.2. Time and Place of Study.....	21
3.3. Materials of Study.....	21
3.4. Variables of Study.....	21
3.5. Tools for Data Collection.....	21
3.6. Study Plan and Calendar.....	22
3.7. Data Evaluation	22
3.7.1. RNA-seq run data quality control and adapter sequence removal.....	22
3.7.2. Alignment of short reads to reference mouse genome	22
3.7.3. De novo transcriptome assembly and comparison to reference transcriptome	23
3.7.4. Applying filters to found transcripts to identify putative lncRNAs.....	24
3.7.5. Quantification of lncRNA expression levels.....	25
3.7.6. Weighted gene co-expression network analysis	25
3.7.7. Identification of timepoint specific lncRNAs.....	26
3.8. Limitations of Study	26
3.9. Ethics Committee Approval	26
4. RESULTS	27
4.1. Alignment and transcriptome assembly	27
4.2. Known and previously unannotated lncRNAs.....	28
4.3. Expression profiles of known and previously unannotated lncRNAs.....	31
4.4. Differential expression analysis of total lncRNA	39

4.5. Weighted co-expression network formation and module membership ...	42
4.6. Module-wise gene ontology enrichment.....	43
4.7. Identification of modules with timepoint-specific upregulated average expression.....	44
4.8. Identification of lncRNAs with timepoint-specific upregulation.....	45
4.9. Prediction of potential partners of lncRNAs of interest.....	48
5. DISCUSSION	52
6.CONCLUSION AND SUGGESTIONS.....	55
7. REFERENCES	56
APPENDIX 1.....	66
APPENDIX 2.....	67

INDEX OF TABLES

Table 1. Evaluation of read alignment and transcriptome reconstruction for each sample of the EMT-MET time course experiment..... 27

Table 2. lncRNA genes with high differential expression patterns during EMT-MET 40

Table 3. Transcript count per module divided by gene type (protein-coding, annotated lncRNA, previously unannotated lncRNA, total lncRNA, total gene). 43

Table 4. Gene ontology term enrichments for modules with putative lncRNAs identified as top hub gene..... 48

INDEX OF FIGURES

Fig. 1. Fluorescent microscopy images of epithelial and mesenchymal cells.....	5
Fig. 2. Overview of computational. lncRNA annotation methods.	11
Fig. 3. Overview of Illumina “sequencing-by-synthesis” next-generation sequencing technology.....	13
Fig 4. Genome-guided and reference-free transcriptome assembly methods. ...	16
Fig. 5. Representation of unweighted gene co-expression network formation based on toy data.....	19
Fig. 6. Flowchart of computational RNA-sequencing data processing methods.	24
Fig. 7. Barplots showing count (A) and density (B) of lncRNA genes on canonical mouse chromosomes.....	30
Fig. 8. Barplot representing the exon count distribution of lncRNAs expressed in at least one timepoint during the EMT-MET processes.	30
Fig. 9. Coverage and read mapping visuals of annotated lncRNA, 5430416N02Rik.....	31
Fig. 10. The distribution of gene expressions represented by boxplots.	32
Fig. 11. Heatmap representing the expression variances of all lncRNA transcripts.....	33
Fig. 12. Dendrogram representing clustering of analyzed samples based on their intersample distances.	34
Fig. 13. Principal component analysis plot of variance between analyzed samples.	35
Fig. 14. Expression profile of known lncRNA Malat1.....	36

Fig. 15. Expression profile of previously unannotated lncRNA NH.1987.	36
Fig. 16. Heatmap representing the expression variances of 50 lncRNA transcripts with highest variance.....	37
Fig. 17. Heatmap representing lncRNAs with positive contributions to the epithelial phenotype.....	38
Fig. 18. Circos plot displaying the genomic coordinates and expression patterns of putative and annotated lncRNAs during the EMT-MET experiment.	39
Fig. 19. Dendrogram of all transcripts identified in the samples and their module membership represented by colors.	42
Fig. 20. Count plot showing partial GO term enrichment results for the identified modules.	44
Fig. 21. Line plots showing average (dark line) and individual (light lines) gene expression levels of modules calculated by WGCNA.	45
Fig. 22. Heatmap representing gene expression Z-scores of timepoint specific transcripts.....	46
Fig. 23. Coverage and individual mapped reads representing NH.7997.1 expression in three different timepoints.	47

LIST OF ABBREVIATIONS

DIP-seq:	DNA Immunoprecipitation sequencing
dNTP:	Nucleoside Triphosphate
ddNTP:	Dideoxyribonucleotide Triphosphate
E-cadherin:	Epithelial cadherin
EMT:	Epithelial-Mesenchymal Transition
FPKM:	Fragments mapped Per Kilobase of transcript per Million Mapped reads
iASPP:	inhibitor of Apoptosis-Stimulating Protein of p53
lncRNA:	Long Noncoding RNA
MET:	Mesenchymal-Epithelial Transition
NMuMG:	Normal Murine Mammary Gland
ORF:	Open Reading Frame
PRC2:	Polycomb Repressive Complex 2
RPKM:	Reads mapped Per Kilobase of transcript per Million Mapped reads
SBS:	Sequencing-By-Synthesis
TGFB:	Transforming Growth Factor Beta
TPM:	Transcripts Per Million reads
UPGMA:	Unweighted Pair Group Method with Arithmetic mean
WGCNA:	Weighted Gene Co-expression Network Analysis

ACKNOWLEDGEMENTS

For their academic contributions to the completion of this thesis, I would like to extend my sincerest gratitude to the following individuals and organizations:

Associate Professor Gökhan Karakūlah, for his continued guidance and patience throughout the planning and conducting of this thesis, as well as his mentorship regarding the field of computational biology;

Assistant Professor Hani Alotaibi, for his continued guidance and patience throughout the planning and conducting of this thesis, as well as for paving the way on the knowledgebase of mesenchymal-epithelial transition;

TÜBİTAK, for funding the studies that produced the initial data this thesis is based on (Project #s: 114Z245, 117Z223);

Alotaibi laboratory members, past and present, for their invaluable efforts at identifying regulatory targets of interest and their relationships;

Burcu Şengez, for her work on the NMuMG cell culture, her mentorship of fellow Alotaibi lab members, and producing the RNA-seq samples the analyses of this study is based on;

Deniz Doğan and Tülay Karakulak, for their assistance with concepts of computational biology and the R statistical computing environment;

My family, for their emotional and financial support of 28+ years;

And all members of the İzmir Biomedicine and Genome Center and İzmir International Biomedicine and Genome Institute, past and present, for giving me the opportunity and the environment to conduct my studies.

-

IDENTIFICATION AND ANNOTATION OF PUTATIVE LONG NONCODING RNAS
INVOLVED IN MESENCHYMAL-EPITHELIAL TRANSITION

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

ABSTRACT

Mesenchymal-epithelial transition (MET) is a key process of multicellular organisms, involved in development and wound healing, as well as coopted by cancer metastasis. As recent studies have shown, the regulation of MET is a more involved cellular reprogramming event than removal or inhibition of epithelial-mesenchymal transition (EMT) promoting elements, but the exact mechanics are poorly studied as of yet. Long noncoding RNAs (lncRNAs), RNA molecules that function in biological pathways independently of translation, are known to be involved in cellular reprogramming events. To bolster the limited information available on MET, we applied computational analysis and network construction methods to MET RNA-seq data to identify any previously unannotated lncRNA candidates, and to predict their potential biological functions. As a result, we have identified 608 transcripts as previously unannotated lncRNAs. Furthermore, we have shown that a number of them show meaningful expression patterns, such as timepoint specific expression, or upregulation during MET compared to mesenchymal phenotype. We have also constructed gene co-expression modules to identify the biological niches of the lncRNAs via enrichment of gene ontology terms of previously annotated genes in the modules. We have shown that previously unannotated lncRNAs are likely involved in crucial cellular reprogramming events such as chromatin remodeling or cellular localization.

Keywords: MET, cellular reprogramming, lncRNA, RNA-seq, weighted gene co-expression network analysis

MEZENKİMAL EPİTELYAL DÖNÜŞÜMDE YER ALAN OLASI UZUN PROTEİN
KODLAMAYAN RNALARIN TANIMLANMASI VE ANOTASYONU

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

ÖZET

Mezenkimal epitelyal hücre dönüşümü (MET) gelişim ve yara iyileşmesi gibi süreçlerde yer alan, ve kanser metastazı tarafından ele geçirilen, çok hücreli canlılar için anahtar bir süreçtir. Yakın zamanda yapılan çalışmalar, MET'nin düzenlenmesinin, epitelyal mezenkimal hücre dönüşümünün (EMT) promotörlerinin kaldırılması veya susturulmasından daha ayrıntılı bir hücresel programlama süreci olduğunu göstermiştir, ancak MET düzenlenmesinin mekanikleri henüz ayrıntılı olarak bilinmemektedir. Uzun protein kodlamayan RNalar (lncRNalar), protein translasyonundan bağımsız olarak biyolojik fonksiyonlara sahip RNA molekülleridir. lncRNaların hücresel programlamada yer aldıkları bilinmektedir. MET süreci hakkındaki kısıtlı bilgileri desteklemek amacıyla, bu süreçten elde edilmiş RNA-seq verilerini hesaplama tabanlı analiz ve ağ kurumu yöntemleriyle inceledik ve daha önce anotasyonu yapılmamış lncRNA adaylarını tanımlamaya ve onların biyolojik yollardaki olası görevlerini tahminlemeye çalıştık. Sonuç olarak, 608 transkript daha önce anotasyonu yapılmamış lncRNA olarak tanımlandı. Dahası, bu transkriptlerin bir kısmının zamana özgü ifade veya mezenkimal fenotipe göre MET sürecinde yükselen ifade seviyesi gibi anlamlı ifade değişiklikleri gösterdiği belirtildi. Ayrıca, lncRNaların biyolojik anlamlarını belirlemek için, gen eşifade ağları kuruldu ve modüllerdeki anotasyonu yapılmış genlerin gen ontoloji terimleri zenginleştirildi. Sonuç olarak, daha önce anotasyonu yapılmamış lncRNaların, kromatin düzenlenmesi ve hücresel lokalizasyon gibi önemli hücresel programlama süreçlerinde yer alabilecekleri gösterildi.

Anahtar Sözcükler: MET, hücresel programlama, lncRNA, RNA-seq, ağırlıklı gen eşifade ağ analizi

1. INTRODUCTION AND AIM

1.1. Statement and Importance of the Problem

The two cellular reprogramming events, epithelial-mesenchymal transition (EMT) and mesenchymal-epithelial transition (MET), are complementary, but not opposite processes. These two processes are involved with a variety of events in multicellular organisms, such as wound healing, embryogenesis and gastrulation. In addition, they are also hijacked by tumor cells during neoplasia and metastasis. Initial studies on MET indicated and operated under the assumption that the two processes share a common regulatory network, and MET occurs as a natural outcome of EMT suppression or the removal of EMT inducing factors. Today, we recognize that MET is controlled by a core regulatory network distinct from EMT, which shows that it is a more complex and involved process than previously thought.

To date, the limited number of studies performed on the regulation of MET have focused on protein-coding genes and transcription factors regulating their expression, and there has been no in-depth examination and analysis of the noncoding transcriptome during the process. Given the impact of long noncoding RNAs (lncRNAs) in chromatin remodeling, a key factor of cellular reprogramming events, as well as other modes of gene upregulation and silencing, understanding the changes that occur in lncRNA levels during the MET process can yield invaluable information regarding its regulation.

1.2. Aim of Study

The study aims to determine the potential importance of noncoding transcriptome, specifically lncRNAs, for the MET process, as well as expand on the previously identified transcription network regulating MET. Therefore, we have used transcriptome assembly techniques to identify previously unannotated lncRNAs, and computational analysis methods to annotate the putative roles of the lncRNAs involved in the process.

1.3. Hypothesis of Study

The hypothesis of the study is that clusters of lncRNAs are preferentially expressed and potentially involved in MET, and their predicted functions in this process can be annotated via associations with previously annotated transcripts.

2. GENERAL INFORMATION

2.1. Epithelial-Mesenchymal Transition (EMT)

Cellular morphology and differentiation is an important concept in the study of multicellular organisms, which, during their developmental process, grow from a single cell to a much higher number of specialized cells and the intercellular connections permitting their functions (Frankfurt, 1996; Hindley and Philpott, 2012).

Two major classes of cellular morphology are called epithelial cellular morphology and mesenchymal cellular morphology (Duval and 1844-1907, 1889). These distinct morphologies are established early during the development (Rossant and Tam, 2009), with epithelial cells being polarized and tightly bound to the extracellular matrix as well as surrounding cells, and mesenchymal cells having no polarized cellular directionality, and having the capacity to break down extracellular matrix and maintain mobility within the matrix. However, these established morphologies are not static throughout an organism's lifetime once established (Maccarty and Caylor, 1922; Slack and Tosh, 2001). The ability of a cell to transition between morphologies or lineages is called phenotypical plasticity, which is a vital trait for the survival of multicellular organisms.

Epithelial-mesenchymal transition (EMT) is a process that is a part of cellular plasticity. During EMT, the major phenotypical changes are loss of cellular junctions (Le Bras, Taubenslag and Andl, 2012) and apical-basal polarity (Ozdamar *et al.*, 2005; Aigner *et al.*, 2007; Moreno-Bueno, Portillo and Cano, 2008), reorganization of the cytoskeleton, and increase in cellular motility. Underlying these changes are a number of cellular reprogramming events and alterations to gene expression (Reik, Dean and Walter, 2001). The genes upregulated during EMT are considered to be mesenchymal morphology markers.

One of the key events of EMT is the cleavage and subsequent degradation of epithelial cadherin (E-cadherin) at the plasma membrane (Peinado, Portillo and Cano, 2004). Coupled with the loss of further regular E-cadherin expression and localization of any ectopically expressed E-cadherin to the membrane, this causes a breakdown of cell-cell junctions. Changes to E-cadherin are considered to be the primary marker of a *bona fide* EMT process, as determined by changes in transcript levels (via microarray or RNA sequencing) and cellular localization (via immunostaining, Fig. 1).

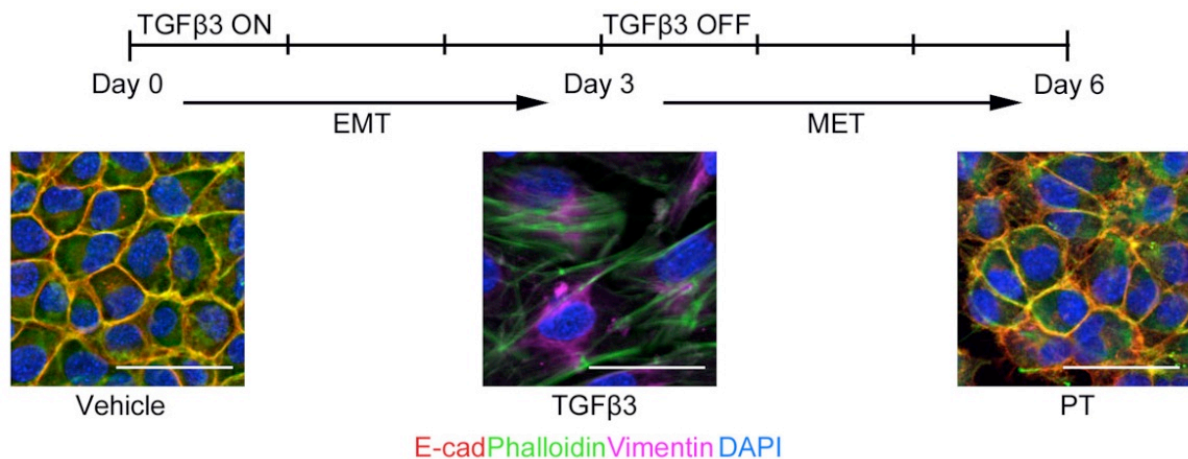


Fig. 1. Fluorescent microscopy images of epithelial and mesenchymal cells.

The images are before TGFβ treatment, 72 hours after TGFβ treatment and 72 hours after TGFβ is removed. Vehicle cells display epithelial morphology, with E-cadherin expressed and localized in the membrane. TGFβ3 cells display mesenchymal morphology after 72 hours of treatment with TGFβ3, with intercellular connections and membrane localized E-cadherin not observed. PT (post-treatment) cells display a return to epithelial phenotype after TGFβ3 is removed from the environment.

Cellular reprogramming events are maintained and regulated by core gene networks, and many of the genes in such networks share master transcription factors. The core regulating transcription factors of the EMT process include the SNAIL, TWIST and ZEB transcription factors (Lamouille, Xu and Derynck, 2014).

EMT is a major component of cellular plasticity in developmental biology, but it also has key roles in adult organisms. One process where EMT is involved in is wound healing. Early stages of wound healing show events typical of EMT (Kim *et al.*, 2017), and both processes share one possible inducer agent, transforming growth factor beta (TGFβ) (Tan, Olsson and Moustakas, 2015; Gilbert, Vickaryous and Vilorio-Petit, 2016). In both cases, aberrant TGFβ signaling can lead to detrimental conditions, i.e. fibrosis and tumorigenesis.

EMT is a crucial step of tumor growth, invasion and metastasis. There have been 136 publications in the last five years covering the relationship between EMT and carcinogenesis as of May 2018. Two noteworthy reviews on the topic are the 2014 publication by Puisieux *et al.*, which explains the potential pro-oncogenic roles of EMT-inducing transcription factors (Puisieux, Brabletz and Caramel, 2014), and the 2013 publication by Cervantes-Arias *et al.*,

about the EMT as a critical process underlying tumor formation (Cervantes-Arias, Pang and Argyle, 2013).

2.2. Mesenchymal-Epithelial Transition (MET)

Another important part of cellular plasticity, mesenchymal-epithelial transition (MET) is the complementary process to EMT. The phenotypical changes in MET are largely a direct reversal of the EMT process, such as relocalization of E-cadherin to the cytoplasmic membrane, reorganization of the cellular skeleton away from cortical actin stress fibers, formation of cell-matrix adhesions and reassertation of cortical-basal polarity; however, the underlying gene regulation is more complex than the removal of EMT-inducing factors, and not as simple as reverse-EMT (Kim, Jackson and Davidson, 2017).

Previous studies regarding MET as a distinct event than the reversal of EMT are limited in both number and scope. Various studies have examined its importance during the re-epithelialization step of wound repair (Hader, Marlier and Cantley, 2010), stem cell differentiation (Li *et al.*, 2011) and metastasis (Han *et al.*, 2012), but these studies have focused on the phenotypical changes of the cells rather than the epigenetic regulation of the programming. One noteworthy study on such regulation was conducted by Gregoire *et al.* in 2016, which has been limited to the SMART pool siRNA library targeting 729 chromatin-modifying genes in breast cancer (Gregoire *et al.*, 2016).

This lack of in-depth examination makes dedicated studies of MET highly important for accurate understanding of the aforementioned processes.

2.3. Long Noncoding RNAs (lncRNAs)

RNA molecules are one of the primary identifiers of a cell's biological identity, along with DNA molecules and proteins (Sul *et al.*, 2009). The role of proteins in a cell and other biological systems is well-documented, as is the role of genomic DNA and messenger RNA in protein synthesis. As a result, most studies on biological systems focus on the properties of these three classes of molecules. Changes in protein abundance and localization in particular confer a great deal of information.

With the advent of next-generation sequencing (NGS) methods, far more information regarding the content of the genome and the transcriptome is accessible to the scientific community than was with earlier methods. The relative ease, low cost and accessibility of NGS

in biological studies has led to the understanding that the coding genome and transcriptome are only a fraction of a larger, functional whole. Today, we know that the cell expresses a high number of mature transcripts from the genome that are not translated into proteins, yet have functions in the cell. These transcripts are collectively called the noncoding RNAs (Mattick, 2001; Djebali *et al.*, 2012).

One family of noncoding RNAs is named long noncoding RNAs (lncRNAs). The members of this family are currently defined as transcripts of at least 200 nucleotides in length, possessing little to no protein coding ability (Rutenberg-Schoenberg, Sexton and Simon, 2016). Although this description is somewhat arbitrary and still under discussion, as shorter RNA chains have been classified as lncRNAs due to their roles in the cell, and some lncRNA transcripts contain open reading frames that are translated to short-lived peptides, it serves to differentiate them from shorter noncoding RNA families, such as microRNAs and short interfering RNAs, which serve other distinct roles in biological systems (Cech and Steitz, 2014).

Although the molecular functions of lncRNAs that have been discovered to date are not as varied as those of proteins, they still affect a variety of biological processes (Wang and Chang, 2011). One interesting feature of lncRNAs is their involvement in seemingly unconnected, sometimes antagonistic processes. One example of an early discovered lncRNA is Xist, which is known for its dosage compensation effect in mammals, silencing any extra X chromosomes in cells with multiple X chromosomes. It performs this function via polycomb repressive complex recruitment, transcriptionally inactivating the entire chromosome. Aberrant transcriptional activation of silenced X chromosomes via Xist knockout is implicated in certain cancers (Weakley *et al.*, 2011). However, a study by Liang *et al.* in 2017 has shown that actively transcribed Xist can promote tumorigenesis by binding two microRNAs, miR-140 and miR-124, whose downregulation act as potential markers for pancreatic duct adenocarcinoma due to translational repression of inhibitor of apoptosis-stimulating protein of p53 (iASPP) (Liang *et al.*, 2017). The activity of lncRNAs can be classified as *cis*-acting and *trans*-acting, as well as product-dependent and product-independent. Product-independent lncRNAs function through recruitment of the transcription complex, making the chromatin structure around the lncRNA gene locus more accessible, and are *cis*-acting lncRNAs, affecting the expression of genes proximal to its genomic position. Product-dependent lncRNAs can be either *cis*- or *trans*-acting, and can function via nucleic acid binding as well as protein binding (He *et al.*, 2016).

A crucial trait of lncRNAs that makes them of specific interest for the EMT and MET processes is that a majority of lncRNAs are localized in the nucleus, and are involved with events such as chromatin remodeling or gene silencing, which indicates their potential importance in cellular reprogramming events (Bonasio and Shiekhattar, 2014; Zhang *et al.*, 2014).

2.3.1. *lncRNAs in EMT*

There have been previous studies on lncRNAs involved in or partnered with regulators of the EMT process. While these studies are at a relatively primitive stage compared to the studies focused on protein-coding genes, they have still provided a sizable amount of knowledge on the topic (Heery *et al.*, 2017; Liao *et al.*, 2017).

The functions of lncRNAs involved in EMT are largely classifiable into two groups, although the memberships of these groups are not exhaustive. The members of the first group are polycomb recruiters, which recruit the polycomb repressive complex 2 (PRC2) to targeted areas on the genome, silencing the expression of genes in the region (Gupta *et al.*, 2010; Kotake *et al.*, 2011; Beckedorff *et al.*, 2013). The second group comprises competing endogenous RNAs. The members of this group have high numbers of miRNA binding site repeats, acting as preferential binding targets for the miRNAs even in relatively low transcript counts. Their presence in the cell prevents miRNAs from binding to mRNA molecules, allowing their translation to proceed uninhibited (Tay *et al.*, 2011; Pan *et al.*, 2017; Tong *et al.*, 2017).

Depending on the interactions with partner molecules, and the specifics of the EMT process, lncRNAs can act as either pro-EMT or anti-EMT agents, with some examples capable of both promoting and inhibiting EMT in different conditions.

Due to the limited number and scope of studies on MET, there is no comparable body of knowledge on lncRNAs in MET.

2.3.2. *Computational methods for lncRNA discovery and annotation*

2.3.2.1. *Difficulties in experimental approaches in lncRNA research*

One of the key reasons for the slow progress of lncRNA research is the difficulty of applying established experimental methods to lncRNA annotation. lncRNA transcripts are often highly condition- or tissue-specific, and tend to have low expression levels compared to mRNAs even under the conditions they show biological activity (Cabili *et al.*, 2011). In

addition, lncRNA genes exist in higher numbers in mammalian genomes in comparison to protein-coding genes, but relatively few of them are expressed under any given condition (Iyer *et al.*, 2015).

lncRNA genes and protein-coding genes can have overlapping regions, especially in the case of the antisense lncRNAs (Milligan *et al.*, 2016; Raju, Tsinoremas and Capobianco, 2016), making methods such as site-directed mutagenesis infeasible, and associating other factors of the genomic region, such as transcription factor binding or chromatin structure, to the lncRNA instead of the protein-coding gene expression can be challenging.

Furthermore, due to the higher number of unannotated lncRNAs that might be candidates of interest for any biological phenomenon (Quek *et al.*, 2015), and the frequent lack of a direct connection between lncRNA transcript sequence and lncRNA function (Washietl, Kellis and Garber, 2014; Hezroni *et al.*, 2015), methods used for protein-coding gene annotation, such as overexpression, mutagenesis, or gene knockdown, can be costly and time-consuming, and may not yield useful information on individual candidates (Signal, Gloss and Dinger, 2016).

It is therefore invaluable to narrow down the list of lncRNAs of interest as well as their properties before using experimental validation. Computational analysis of a cell's transcriptome can help identify expressed lncRNA transcripts, even at low activity levels. Using the differential expression of such transcripts across a number of conditions provides further details about the lncRNA's biological role in the cellular systems (He *et al.*, 2016; Signal, Gloss and Dinger, 2016).

2.3.2.2. Overview of predictive lncRNA annotation methods

The primary method utilized during lncRNA annotation is differential expression analysis. While it is insufficient by itself for detailed annotation of the genes, it is helpful for zeroing in on candidates, and is often necessary for the more common annotation methods. One such method is the “guilt-by-association” method, which is based on the idea that if two or more genes show similar expression patterns, they are likely to share regulators, and possibly an evolutionarily conserved relationship, or they might have similar functions and pathways (Stuart *et al.*, 2003). Genome-wide clustering of the expression levels of putative lncRNAs and previously annotated genes will result in groups of genes that are enriched for a shared set of biological processes and functions, resulting in revealing information on the possible properties

of the unannotated transcript. In their 2018 article, de Jong *et al.* have described the use of the guilt-by-association method for identification of druggable targets in diffuse large B-cell lymphoma samples, identifying potential partners of CD20 in a genome-wide study in patients who had shown resistance to therapy including monoclonal antibodies targeting CD20 (De Jong *et al.*, 2018).

lncRNAs show much higher specificity compared to other RNAs, either in temporal expression (i.e. expression under specific conditions) or spatial expression (i.e. expression in specific tissues), and this specificity can point towards the biological context of the lncRNAs. In multiple tissue or timepoint experiments, methods such as the tau score can point out the sample showing preferential expression for the transcript of interest, which yields further insight than differential expression alone (Kryuchkova-Mostacci and Robinson-Rechavi, 2017).

The epigenetic status of a lncRNA can also be highly informative, especially with intergenic lncRNAs, dividing them into promoter and enhancer lncRNAs, with the former having enriched H3K4me3 in their promoter region, and the latter having enriched H3K4me1 instead of me3 (Marques *et al.*, 2013).

An overview of the computational methods of lncRNA annotation are available in Fig. 2.

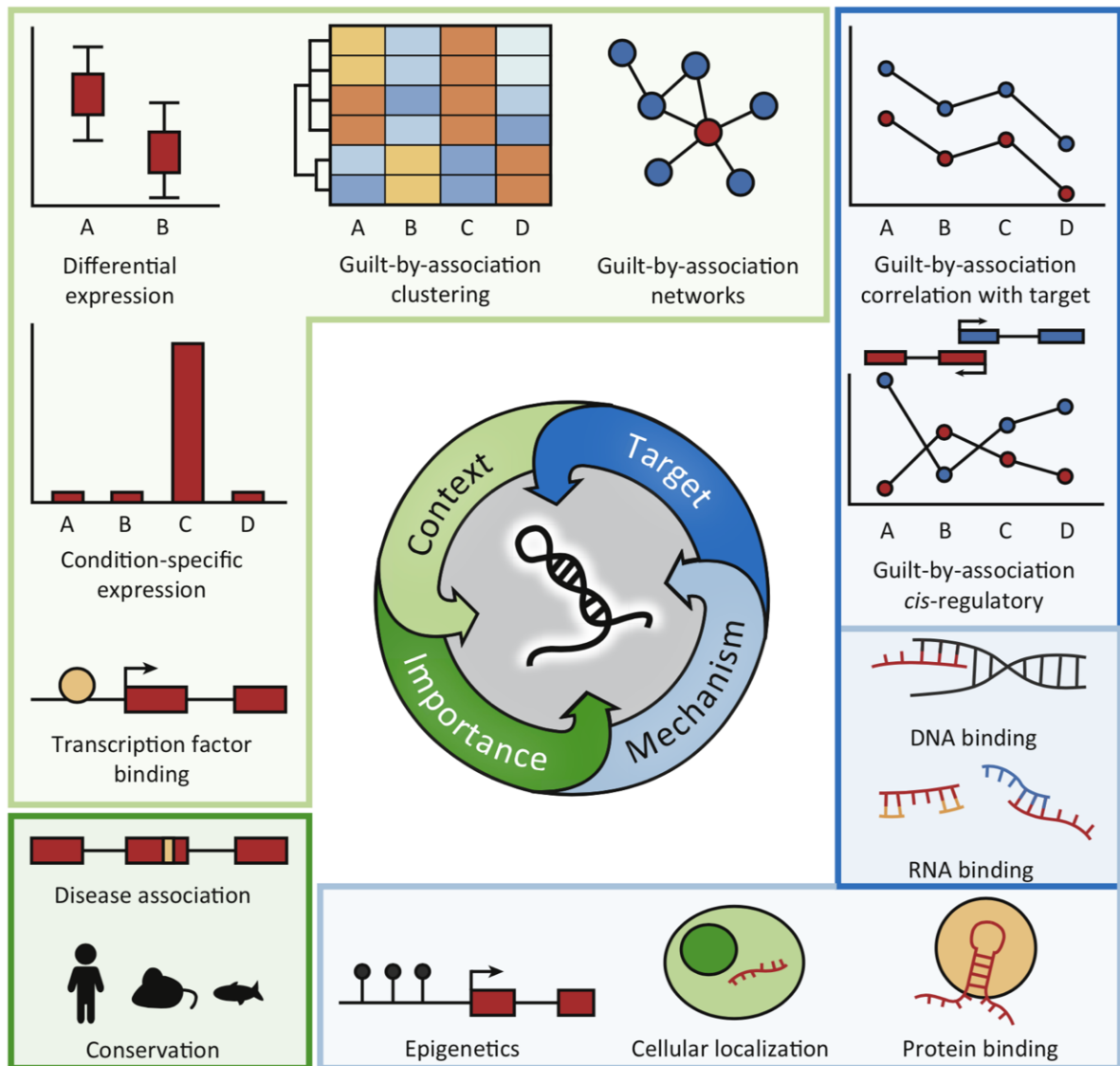


Fig. 2. Overview of computational. lncRNA annotation methods.

Figure adapted from *Computational Approaches for Functional Prediction and Characterisation of Long Noncoding RNAs* (Signal, Gloss and Dinger, 2016), under license number 4363511186842 (see Appendix 1).

2.4. Next Generation Sequencing (NGS)

In nucleotide sequencing, one of the earliest breakthroughs was the advent of Sanger sequencing, described by Frederick Sanger and colleagues in 1977 (Sanger, Nicklen and Coulson, 1977). Based on the DNA polymerase inhibitory activity of dideoxynucleotides (ddNTP) when incorporated in a nucleotide chain in place of unmodified nucleotides (dNTP), Sanger sequencing allows for accurate sequencing of nucleotide chains by the synthesis of

prematurely terminated chains in an environment containing both dNTPs and a specific type of ddNTP, and examining the length of the chains by use of gel electrophoresis to determine the position of sites where the ddNTP substitution occurred.

Later advancements in sequencing technology have greatly increased the efficiency of sequencing in terms of both time and cost. Collectively referred to as “next generation sequencing”, these methods have greatly expanded the potential uses of nucleotide sequencing and the number of applicable systems where they can be utilized. For comparison, the Human Genome Project, which started in 1990, cost 2.7 billion USD and was completed in thirteen years in 2003 (*Human Genome Project Completion: Frequently Asked Questions - National Human Genome Research Institute (NHGRI)*, no date). Today, the same amount of genomic data can be sequenced for less than 2000 USD and takes less than a week to perform, with the numbers rapidly lowering for “production-scale” sequencing (*Specification Sheet: Sequencing*, no date).

2.4.1. Sequencing technologies and their applications

There are multiple NGS approaches available today, based on various chemistries and analysis methods. One of the most widely used methods is the sequencing-by-synthesis (SBS) method utilized by Illumina sequencing technologies. Like Sanger sequencing, it is based on use of terminator nucleotides. However, unlike Sanger sequencing, the terminators used in SBS are reversible and bound to fluorescent labels unique for each base. While new strands of DNA are being synthesized based on the template to be sequenced, the sequencing instrument detects the fluorescent label added to the strand, and the sequence is thus identified (*Specification Sheet: Sequencing*, no date) (Fig. 2).

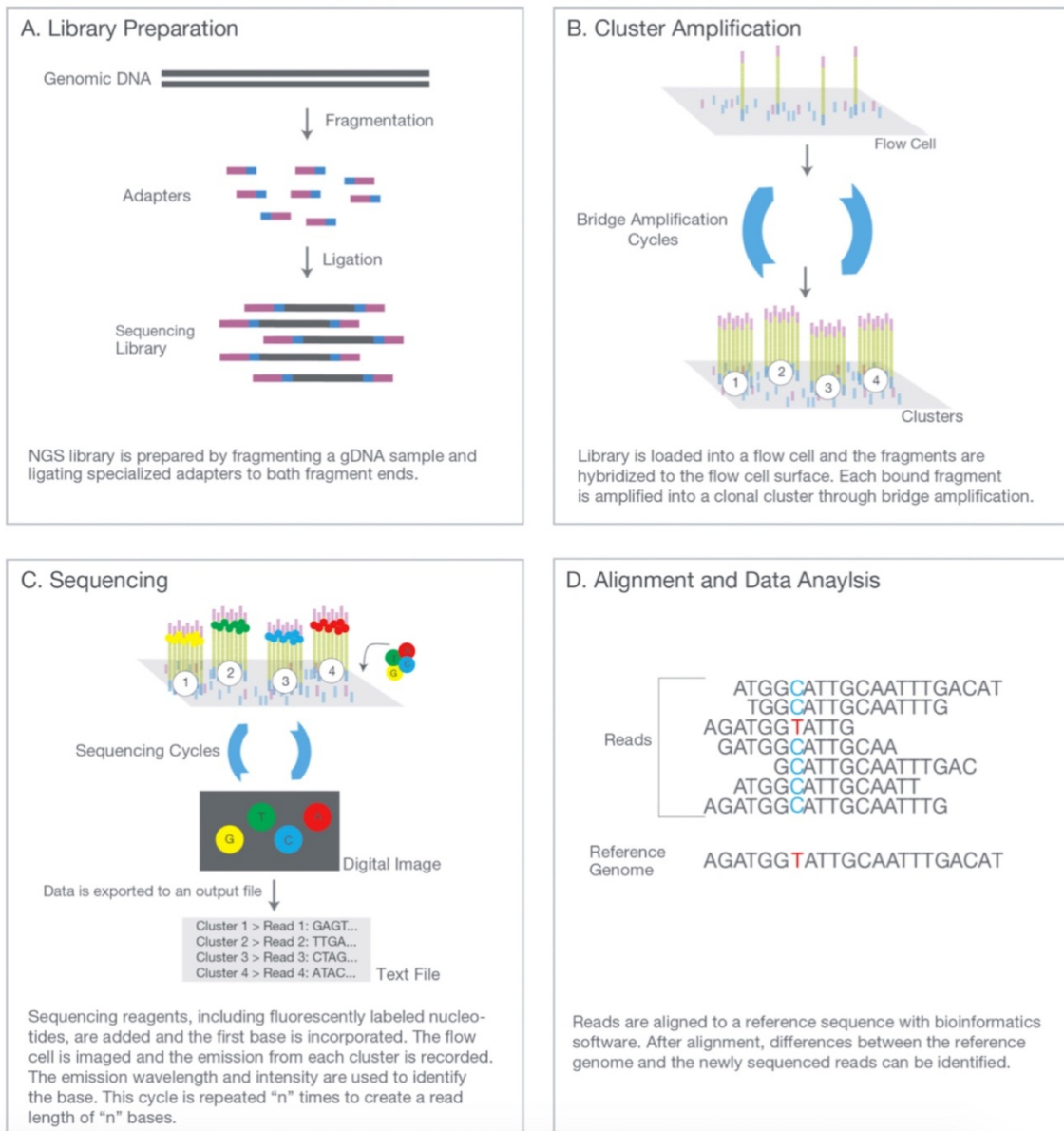


Fig. 3. Overview of Illumina “sequencing-by-synthesis” next-generation sequencing technology.

Image courtesy of Illumina, Inc.’s primer to next-generation sequencing (*An introduction to Next-Generation Sequencing Technology*, no date).

Alternative approaches include ion semiconductor sequencing and pyrosequencing, both based on the detection of the release of radicals from the added base during DNA strand

synthesis (hydrogen ions for ion semiconductor sequencing and pyrophosphate ions for pyrosequencing).

The major advantage of these methods compared to Sanger sequencing is the ability to run sequencing reactions in a massively parallel fashion, greatly reducing the time required for large nucleotide sequences, such as whole genome or transcriptome sequencing.

Different NGS technologies and platforms have various advantages and disadvantages. Among Illumina sequencers, the benchtop MiniSeq system has targeted gene expression profiling as a key application, but it cannot perform whole-transcriptome sequencing, which the benchtop NextSeq series can perform and the production-scale HiSeq series has as a key application (*Sequencing Platforms | Compare NGS platforms (benchtop, production-scale)*, no date).

The earliest application of NGS was as an alternative to Sanger sequencing for whole genome or whole exome sequencing. Today, the use of NGS in genome sequencing is most useful for SNP identification and individual genome sequencing, whether it is the genome of a patient in personalized medicine or a cell line of a previously sequenced model organism. Non-model organism genome sequencing is also performed, although not as frequently as the above uses (da Fonseca *et al.*, 2016).

NGS can also be used to identify epigenetic modifications to a genome. DNA immunoprecipitation sequencing (DIP-seq) can be used to identify sites of cytosine methylation and hemimethylation on the genome (Weber *et al.*, 2005; Shen *et al.*, 2013), while ChIP-seq with antibodies targeted to modified histones can reveal promoter, enhancer or actively transcribed regions (Neff and Armstrong, 2009).

One of the more common uses of NGS in recent years is transcriptome assembly and analysis, supplanting the previous use of microarray for gene expression research. The major advantage of NGS-based transcriptome research is the possibility of identifying novel transcripts and receiving data on the complete transcriptome of the system of interest, instead of only a limited number of previously selected genes (Marques *et al.*, 2013).

2.4.2. Transcriptome research with NGS analysis

2.4.2.1. Basic raw data pre-processing and quality assessment

Raw read data obtained from the sequencing instrument is not suitable to immediate alignment. The presence of low-quality reads, base call errors causing insertions and deletions,

and adapter contamination means that the reads have to undergo pre-processing steps before analysis steps. FastQC is a tool commonly used for quality control on reads obtained from Illumina platforms (Andrews, 2018), but platform-neutral alternatives such as NGSQC are also available (Patel and Jain, 2012). Any reads or bases that would reduce the quality of the sequencing data, as well as adapter sequences introduced during library preparation PCR steps, can be removed with tools including Trimmomatic (Andrews, 2018) or FASTX-Toolkit (Hannon, 2018).

2.4.2.2. *Alignment of sequencing reads*

After the pre-processing steps are complete, the RNA-seq reads can then be aligned to a reference sequence, if one is available (reference-free cases are covered under the next heading). There are two kinds of reference sequences for alignment: reference genomic sequences and reference transcriptomes. The reference of choice informs the alignment tool of choice, as well as how the data should be interpreted. RNA alignments to the genome are performed with gapped aligners, such as TopHat2 (Hannon, 2018), HISAT2 (Kim, Langmead and Salzberg, 2015), or STAR (Hannon, 2018). Different gapped aligners have varying rates of accuracy when the read to be aligned contains splice sites, polymorphisms or indels, and these should be taken into consideration if the research requires variant analysis and discovery.

Alignment to reference transcriptome is generally faster and more accurate than alignment to reference genome, but is only suitable for analysis of known transcripts. Such alignments cannot be used for discovery of novel transcripts or splice junctions, and lack variant analysis power (Garber *et al.*, 2011). When aligning to the transcriptome, ungapped aligners such as Bowtie are used (Hannon, 2018). Multiple mapping reads are more common in alignment to transcriptome as a result of gene isoforms sharing exons, but being represented by separate regions on the transcriptome sequence.

2.4.2.3. *Transcriptome assembly and novel transcript discovery*

When identifying novel transcripts for analysis, it is necessary to assemble the transcriptome of the cell of interest. There are two main approaches to transcriptome assembly. These approaches are classified as “reference-free” (AKA *de novo*) and “genome-guided” (AKA *ab initio*) (Fig. 3).

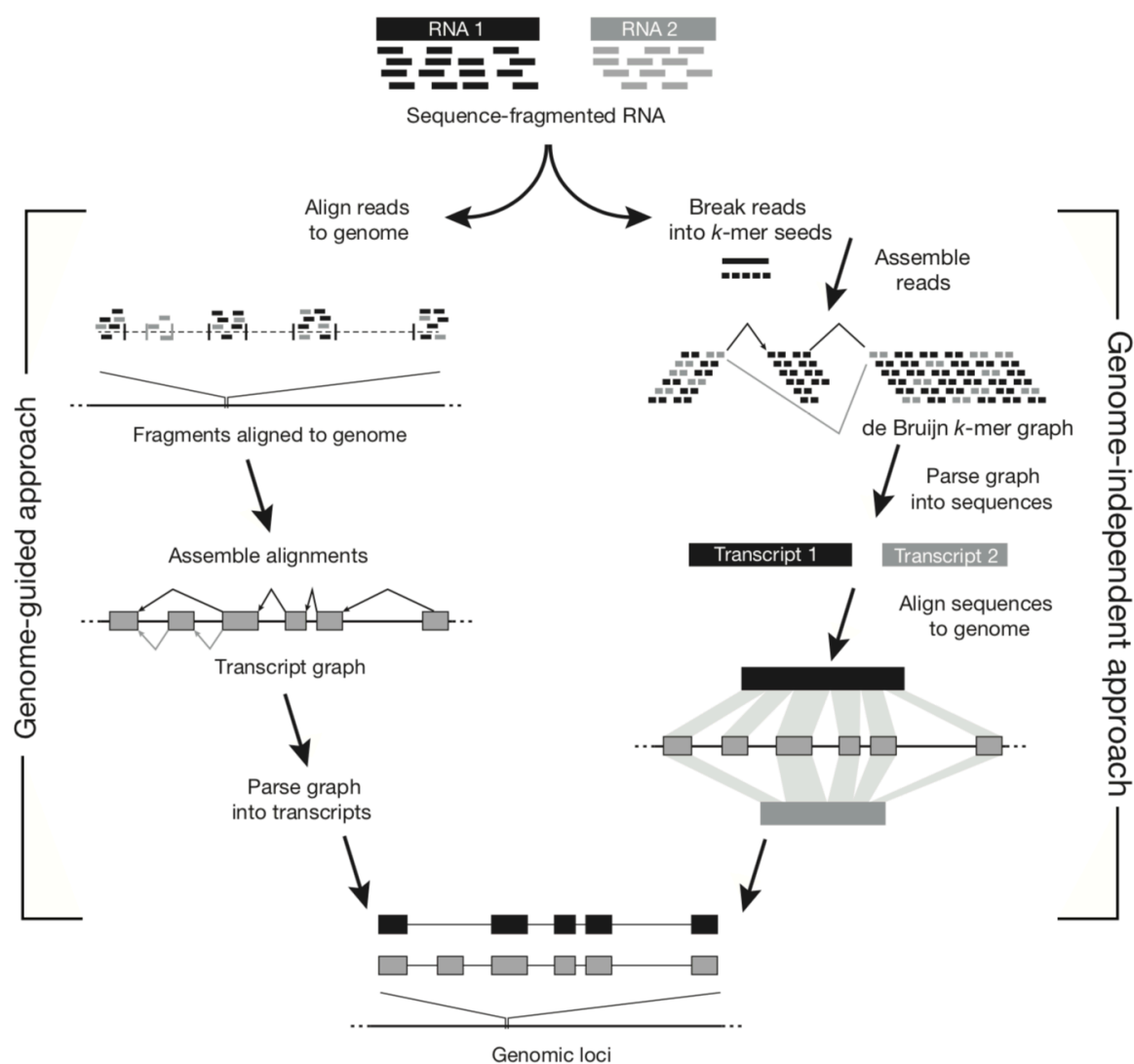


Fig 4. Genome-guided and reference-free transcriptome assembly methods.

Left side of the figure displays genome-guided transcriptome assembly flowchart, while right side of the figure displays reference-free transcriptome assembly. Figure adapted from *Computational methods for transcriptome annotation and quantification using RNA-seq* (Garber *et al.*, 2011) under license number 4363520183733 (See Appendix 2).

In both cases, the reads obtained from RNA-seq, which are often a fraction of a full-length transcript, are assembled into transcripts via overlapping end regions of the reads. Reference-free assembly, as the name indicates, constructs the transcriptome based solely on the RNA reads, often using de Bruijn graphs to identify isoforms of genes and splicing sites (Robertson *et al.*, 2010). While this approach can be used with no information other than the read

sequences, it is computing power intensive and has a higher error rate compared to genome-guided assemblies, producing misleading initial results containing chimera genes, missing alleles, gene fragments, etc., that require a more thorough quality checking process (Cahais *et al.*, 2012). As such, it is usually reserved for working with organisms that lack reliable reference sequences for their genome or transcriptome. When working with established organisms, such as *Mus musculus*, it is preferable to utilize previous sequence assemblies, both genome and transcriptome.

Genome-guided transcriptome assembly, instead of read sequences and de Bruijn graphs, requires alignment files, containing information on the mapping of reads to the genome. When assembling a transcriptome in this fashion, the choice of mapper algorithm is crucial. Mature transcripts found in the cell's transcriptome lack intronic sequences present on the genome, and the 5' end and the 3' end of a single read can map onto two bordering exons of a single intron, possibly several kilobases apart. For accurate mapping of such reads to the genome, we must use gapped aligners, also known as splice-aware or gap-aware, that can recognize potential splice sites to prevent large intronic sequences from creating false negatives when calculating the alignment scores (Garber *et al.*, 2011). Unlike with de Bruijn graphs, the reads are only compared against the single sequence of the reference genome, albeit a very large one, instead of the large numbers of RNA-seq read sequences, with full-length transcript sequences inferred from sections of the genome with multiple overlapping RNA-seq reads mapped during alignment. It is therefore less computationally intensive, and has a lower error rate (Trapnell *et al.*, 2010).

Once the transcriptome is assembled, it can be compared against previous annotated assemblies to locate any identified, previously un-annotated transcript candidates. This candidate pool can be further refined into a pool of putative lncRNAs by applying several filters. These filters include removing any transcripts shorter than 200 nucleotides or possessing open reading frames coding for more than 100 aminoacids, as well as showing homology to known protein domains and housekeeping RNAs (Li *et al.*, 2014; Liao *et al.*, 2017).

2.4.2.4. Transcript / Gene quantification and Differential Expression Analysis

For the majority of novel transcript annotation methods, the expression levels of the transcript in systems of interest must be calculated. The count of reads aligning to a given gene (if aligned to the genome) or transcript (if aligned to the transcriptome) can be potentially

misleading due to various factors. The primary skewing factor is the sequencing depth of the data (Liu *et al.*, 2013; Kim *et al.*, 2015). A uniform sequencing depth is preferable if the sequence data is produced during a single research, but it is not always feasible, especially when working with publicly available data. In addition, the length of the transcript will also affect the count of reads aligned, with longer transcripts naturally having a higher number of reads mapped compared to shorter transcripts, even at similar effective expression levels.

To solve these problems, the expression levels of transcripts are stated in alternative terms such as Fragments Mapped per Kilobase of Transcript per Million Mapped Reads (FPKM), normalizing the numerical values to more accurate representation of the transcript levels. Other normalized values are referred to as Reads Mapped per Kilobase of Transcript per Million Mapped Reads (RPKM) and Transcripts per Million Reads (TPM) (Mortazavi *et al.*, 2008). FPKM and RPKM are largely similar in terms of normalization method. The main difference between the two values arises from single-end read sequencing versus paired-end read sequencing. In paired-end reads, each fragment in FPKM equals the overlapping region of a read pair mapping onto the genome, and as such the FPKM value of any given transcript in paired-end sequencing is expected to be below half of the RPKM value; for single-end sequencing, the two values are expected to be equal (Pachter, 2011).

There is a high number of software available for transcript quantification, such as Cufflinks (Hannon, 2018), RSEM (Hannon, 2018), and StringTie (Pertea *et al.*, 2015). These software are often designed to work optimally with specific aligners or types of aligners, such as HISAT and StringTie (Kim *et al.*, 2016), or RSEM and ungapped aligners such as Bowtie 1. Once the quantification is complete, the resulting data can be analyzed with a number of statistics software or packages, such as DEseq (Love, Huber and Anders, 2014), which works with non-normalized data, or Ballgown (Frazee *et al.*, 2014), which works with normalized data, to identify the differential expression patterns of the transcripts.

2.4.2.5. Construction of co-expression networks for lncRNA research

As previously stated, simple expression quantification and differential expression analysis is generally insufficient for lncRNA analysis. With multiple samples sequenced and quantified, however, genes showing similar expression patterns across the samples can be clustered together, based on the guilt-by-association principle explained above.

To form a co-expression network where the guilt-by-association principle can be applied, it is vital to normalize the fluctuations in expression levels of the transcripts across samples, such as by converting FPKM values into Z-scores, the distance of each sample from the mean in terms of the standard deviation of the transcript expression levels (Zhang and Horvath, 2005). This is to remove any bias that might stem from low levels of fluctuation in otherwise highly expressed genes. Afterwards, the distributions of each transcript can be compared with the rest by calculating the distance or correlation of each pair.

Once the correlation scores are calculated, the genes can be formed into a co-expression network, with edges between the gene nodes based on the calculated scores (Stuart *et al.*, 2003). There are two classes of such co-expression networks: unweighted networks and weighted networks. For unweighted networks, the predicted connections between the genes are identified in a straightforward fashion. As the correlation scores range from -1 to 1, with values further away from 0 (i.e. with a higher absolute value) indicating stronger relationships between the genes, either negative or positive, genes with absolute correlation scores higher than a cut-off value are considered connected (Fig. 3).

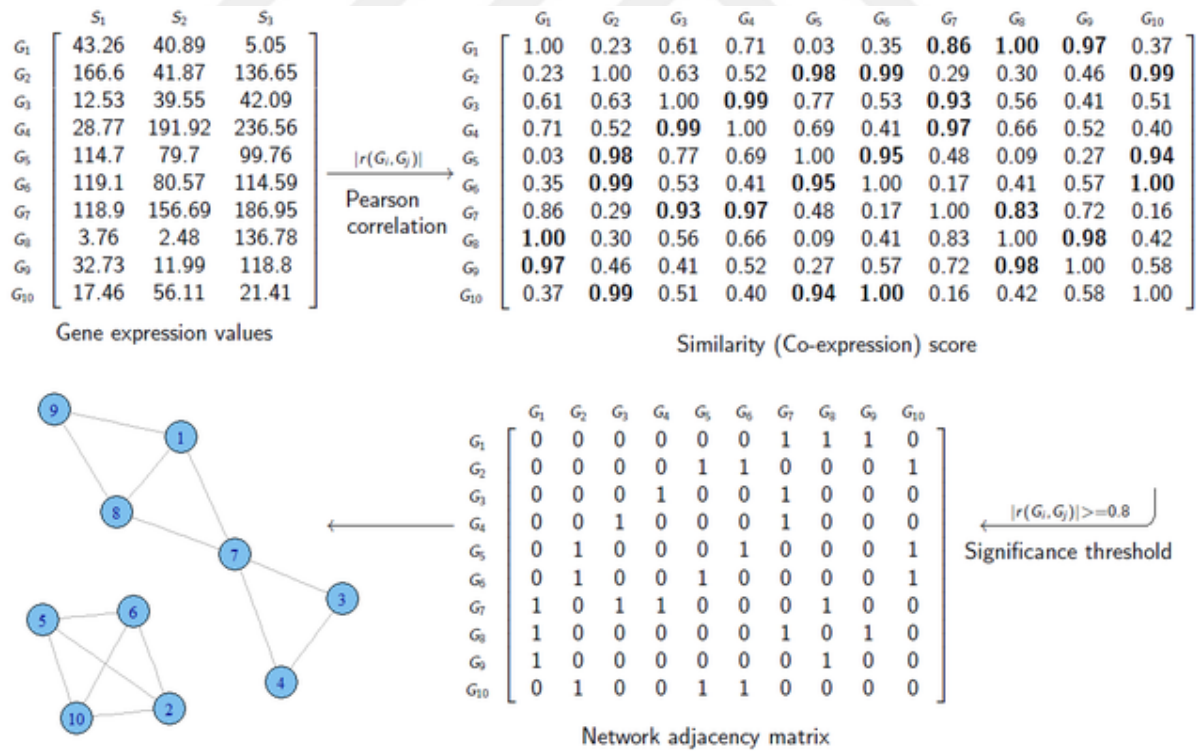


Fig. 5. Representation of unweighted gene co-expression network formation based on toy data.

Figure courtesy of S. Mohammad H. Oloomi, shared under a GNU Free Documentation License.

In comparison to unweighted correlation networks, weighted correlation networks do not implement a hard correlation threshold to determine potential gene relationships. Such weighted correlation networks are based on methods similar to those used in hierarchical clustering (Zhang and Horvath, 2005). One of the software tools used for weighted correlation network construction and analysis is WGCNA (Langfelder and Horvath, 2008), which allows users to group genes into modules based on their co-expression. These modules can then be further analyzed for more in-depth biological information. The major advantage of weighted correlation networks for novel transcript analysis is the possibility of predictive annotation based on module membership. Once the modules in a given transcriptome are identified, members lacking annotations in every module can be studied in terms of the module's overall functional enrichment information based on the previously annotated members, intramodular connectivity, and hub gene identification. One previous study that has included WGCNA as a core method has identified various transcriptional modules correlated with hepatocellular carcinoma using publicly available microarray datasets (Xu *et al.*, 2016).

3. MATERIALS AND METHODS

3.1. Type of Study

The study pursues an analytical and computational approach.

3.2. Time and Place of Study

The study was conducted at İzmir Biomedicine and Genome Institute, between December 2017 and May 2018.

3.3. Materials of Study

The study was performed using raw RNA-seq reads obtained from a time course experiment spanning 72 hours of EMT and 72 hours of MET. The EMT process was induced in NMuMG cells in plated culture using 72 hours of TGFB3 treatment. The MET process was induced by washing TGFB3 from the cell medium and replating the mesenchymal cells. Samples were obtained at the beginning and at the end of the EMT process (hereafter referred to as Vehicle and TGFB72, respectively), as well as 3 hours, 6 hours, 9 hours, 18 hours, 24 hours, 36 hours, 48 hours and 72 hours after the start of the MET process (hereafter referred to as PT3, PT6, PT9, PT18, PT24, PT36, PT48 and PT72, respectively). Reads were sequenced from whole transcriptome libraries of the samples depleted for ribosomal RNAs. The reads were produced prior to this study using an Illumina HiSeq 2500 platform, using a single-end sequencing method.

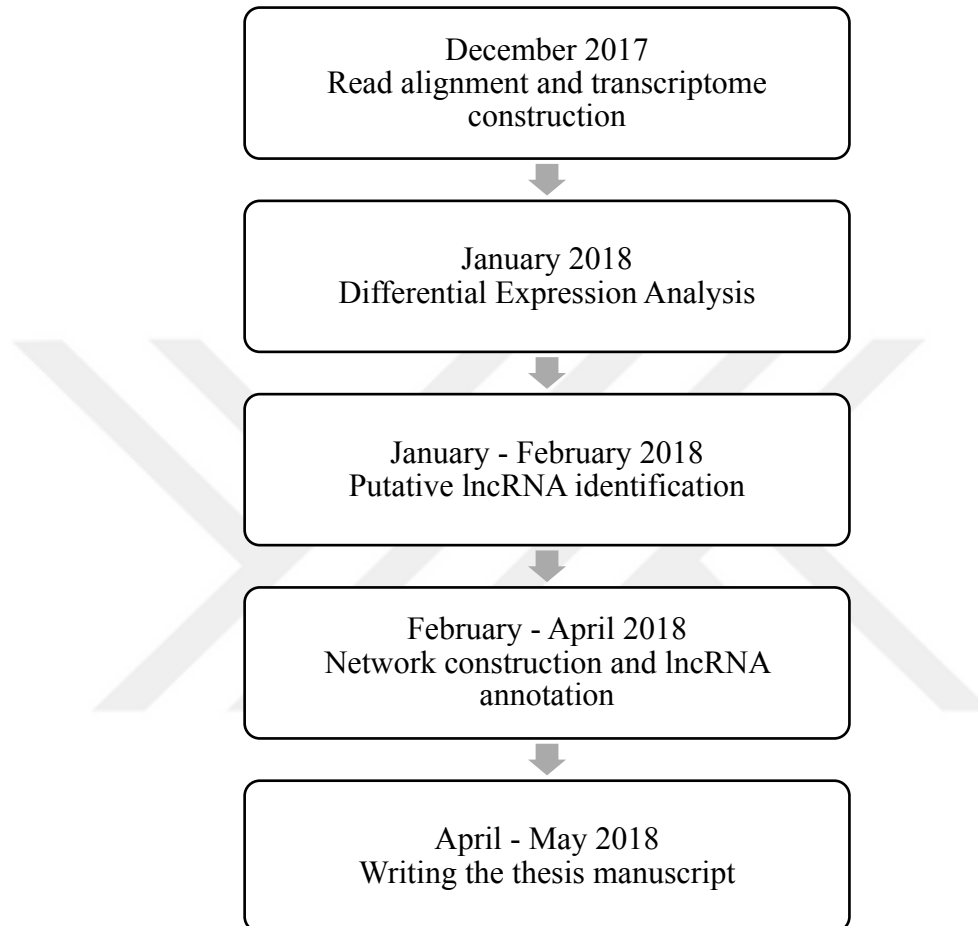
3.4. Variables of Study

The timepoints of the RNA-seq reads obtained from the NMuMG cell model of EMT and MET are independent variables. The expression values of identified transcripts are dependent variables. The transcripts identified as previously unannotated lncRNAs are also dependent variables.

3.5. Tools for Data Collection

All data collection and processing operations were conducted on an Intel® Core™ i7-4820K Processor, 64 GB RAM and 256 GB hard disk computer with Ubuntu release 17.04 operating system installed.

3.6. Study Plan and Calendar



3.7. Data Evaluation

3.7.1. RNA-seq run data quality control and adapter sequence removal

The raw read data obtained was filtered for low-quality reads using the FastQC software, and adapter sequences were removed from the reads with the Trimmomatic software.

3.7.2. Alignment of short reads to reference mouse genome

The whole genome FASTA sequence file of Genome Reference Consortium Mouse Build 38, under the name mm10 as used by the University of California, Santa Cruz, as well as associated reference and index files, were acquired from Illumina's iGenomes reference website. To align reads from RNA-seq to the reference genome, HISAT2, a splice-aware

alignment tool, was selected for its speed, low memory requirements and accuracy (Kim, Langmead and Salzberg, 2015). The mm10 reference genome was indexed from the whole genome FASTA file using hisat2-build indexer (hisat2-build <FASTA file> <index>), and each read file was aligned to the genome using the resulting index and hisat2 single-end alignment algorithm (hisat2 -x <index> -U <reads> --dta -S <SAM file> -p 6), using the default alignment scoring settings. The resulting SAM files were converted to coordinate-sorted BAM files using samtools view and sort utilities (samtools view -b -o <BAM file> -@ 6 <SAM file>; samtools sort -o <sorted BAM file> -@ 6 <BAM file>), and the final sorted BAM file was indexed with use of samtools index (samtools index <sorted BAM file> <BAI index file>).

3.7.3. *De novo transcriptome assembly and comparison to reference transcriptome*

StringTie, a software that can assemble aligned RNA-seq reads into potential transcripts, was used for identification of known and novel transcripts (Pertea *et al.*, 2015). The reads in each alignment file obtained from the previous step was assembled into a GTF file containing information about possible transcripts, including any relevant annotation within the gene information file obtained from iGenomes (stringtie -G <reference GTF> -o <output transcriptome GTF> -p 6 <sorted BAM file>). The resulting transcript files were merged into a non-redundant genome-guided *de novo* transcript assembly using StringTie's merge functionality (stringtie --merge -G <reference GTF> -o <merged transcriptome GTF> -p 6 <list of input transcriptome GTFs>). Using the gffcompare utility (Pertea and Kirchner, 2016), the assembled transcriptome was compared to the reference transcriptome information (gffcompare -r <reference GTF> -G -o <merge prefix> <merged transcriptome GTF>). Of special note were the transcripts marked with class codes *u* and *x*, representing intergenic and antisense transcripts, respectively. The transcript abundances were estimated and written into tables for downstream analysis (stringtie -e -B -p 6 -G <merged transcriptome GTF> -o <output transcriptome GTF> <sorted BAM file>). A flowchart of the computational steps performed up to this point are included in Figure 6.

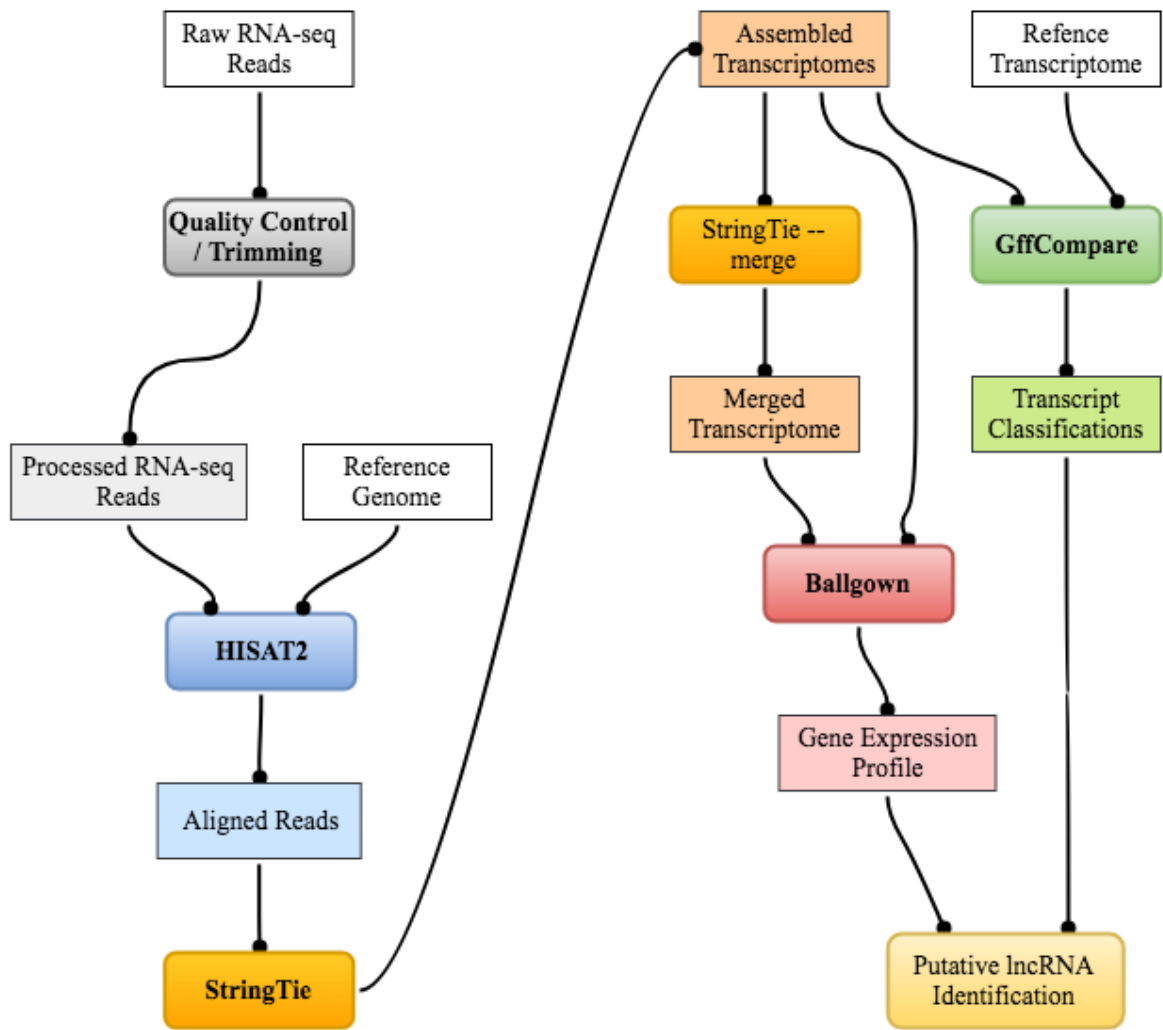


Fig. 6. Flowchart of computational RNA-sequencing data processing methods.

3.7.4. Applying filters to found transcripts to identify putative lncRNAs

The transcripts marked u and x during the comparison to reference annotation were subjected to further filters to determine their potential status as putative lncRNA transcripts. The methodology used was adapted from Li et al.'s 2014 paper (Zhu *et al.*, 2012). To reiterate, the transcripts were first selected for size, choosing only those that are 200 nucleotides or longer in length. Afterwards, the chosen transcripts were checked for open reading frames (ORF) coding for 100+ aminoacids, using TransDecoder, a coding region prediction software (Haas and Papanicolaou, no date). In addition, more than 90% of the proteins found in cells are longer than 100 aminoacids (Frith *et al.*, 2006). Following the ORF filter, the transcripts were checked for homology against the Swiss-Prot database (Bateman *et al.*, 2017) using a locally maintained

BLAST database mirror, using blastx with an E value cutoff of 0.01. Finally, the transcripts were aligned against the Rfam database of RNA families (Kalvari *et al.*, 2018) using Infernal, an RNA alignment inference tool (Nawrocki and Eddy, 2013), to rule out housekeeping RNAs and microRNA precursors.

3.7.5. *Quantification of lncRNA expression levels*

The transcripts identified as putative lncRNAs, as well as the remaining transcripts in the StringTie output, were quantified and checked for differential expression using Ballgown, a transcript differential expression analysis toolkit (Frazee *et al.*, 2014) housed in the R statistical computing environment. The previously annotated transcripts were further classified into protein-coding transcripts, annotated lncRNA transcripts, and other transcripts, using only the former two classes in addition to the putative lncRNA transcripts in further analysis. During this step, any transcripts that could be classified as transcriptional noise were also filtered out of the transcript pool used in downstream steps. For this purpose, transcriptional noise was arbitrarily defined as transcripts that do not show significant expression levels ($\geq .5$ FPKM for lncRNA transcripts, ≥ 1 FPKM for protein-coding transcripts) in both biological replicates of at least one time point.

3.7.6. *Weighted gene co-expression network analysis*

To identify potential partners of the putative lncRNAs, WGCNA, an R package that calculates weighted correlation networks based on gene expression values (Langfelder and Horvath, 2008), was used to analyze the expression data. A dendrogram of the samples was created using the UPGMA (“average”) hierarchical clustering to visualize the distance of samples from the others in terms of their gene expression profiles, and identify any potential outlying samples that could skew downstream analysis steps. All samples were found to be within expected distance of each other. Based on the scale-free topology fit index, a soft-thresholding power of 5 was calculated for use in adjacency matrix formation. The genes were then placed into modules using the blockwiseModules function of WGCNA, based on their adjacency values in the unsigned correlation network. A total of 30 modules were identified, with genes that do not correlate to any modules placed into a 31st module that was not used for further analysis. The functional annotation enrichment of the modules was performed based on

their previously annotated gene members, using the DAVID annotation analysis database (Huang, Sherman and Lempicki, 2009b, 2009a).

3.7.7. Identification of timepoint specific lncRNAs

A combination of ROKU and Tau scores were used to calculate the timepoint specificity of lncRNA expressions (Kryuchkova-Mostacci and Robinson-Rechavi, 2017). The ROKU command of the R package TCC was used for ROKU calculations (Sun *et al.*, 2013), and lncRNAs having a value of 1 in only a single timepoint in the outlier output data frame were selected as being timepoint specifically upregulated. Tau scores for the genes were calculated using a custom script.

3.8. Limitations of Study

The findings are limited to the transcriptomic content of the cell model used, as well as by the algorithms of the analysis software. In addition, any findings are predictive in nature, and will need to be validated in biological systems using experimental studies.

3.9. Ethics Committee Approval

No ethics committee approval was required or requested due to the purely computational scope of the study.

4. RESULTS

4.1. Alignment and transcriptome assembly

Using the flagstat tool of the samtools toolkit, the alignment files produced by HISAT2 were checked for read count of each sample, as well as mapped alignment percentages. The minimum mapped read percentage was 89.78%, within acceptable parameters, with a maximum mapped read percentage of 95.70% and a mean mapped read percentage of 93.42%. The sample-specific assembled transcriptomes were examined for the transcript counts of each sample. As a result, an alignment and transcriptome profile table of the samples was constructed (Table 1).

Table 1. Evaluation of read alignment and transcriptome reconstruction for each sample of the EMT-MET time course experiment.

Timepoint	Replica	Total Read Count	Mapped Read Count	Mapped Read Percentage (%)	Identified Transcript Count
Vehicle	1	52398642	47670200	90.98	27270
	2	65219227	60733347	93.12	26921
TGFB72	1	54055483	51039095	94.42	26296
	2	49087043	45793849	93.29	26882
PT3	1	49184276	46098787	93.73	26264
	2	45094310	42360048	93.94	25845
PT6	1	46558670	43907693	94.31	25514
	2	50351163	47453710	94.25	26969
PT9	1	54743048	52035832	95.05	26742
	2	52603659	47422740	90.15	26074
PT18	1	49406248	46752389	94.63	26421
	2	65219227	60733347	93.12	26921
PT24	1	50080883	47490713	94.83	26296
	2	49087043	45793849	93.29	26882
PT36	1	56713869	54057626	95.32	27106
	2	47103170	43686797	92.75	26889
PT48	1	60911333	57289437	94.05	27212
	2	42842736	41000704	95.70	26400
PT72	1	68654537	61640017	89.78	27308
	2	61034582	55972681	91.71	27193
Average		53517457.45	49946643.05	93.42	26670.25

The merged transcriptome had a total transcript count of 57210 before comparison to the reference mouse genome annotation (UCSC mm10). After the use of gffcompare utility tool, the number was reduced to 56393 transcripts. Given that multiple transcripts can map onto the same gene on the genome, the resulting tmap file was examined in further detail. Of the 56393

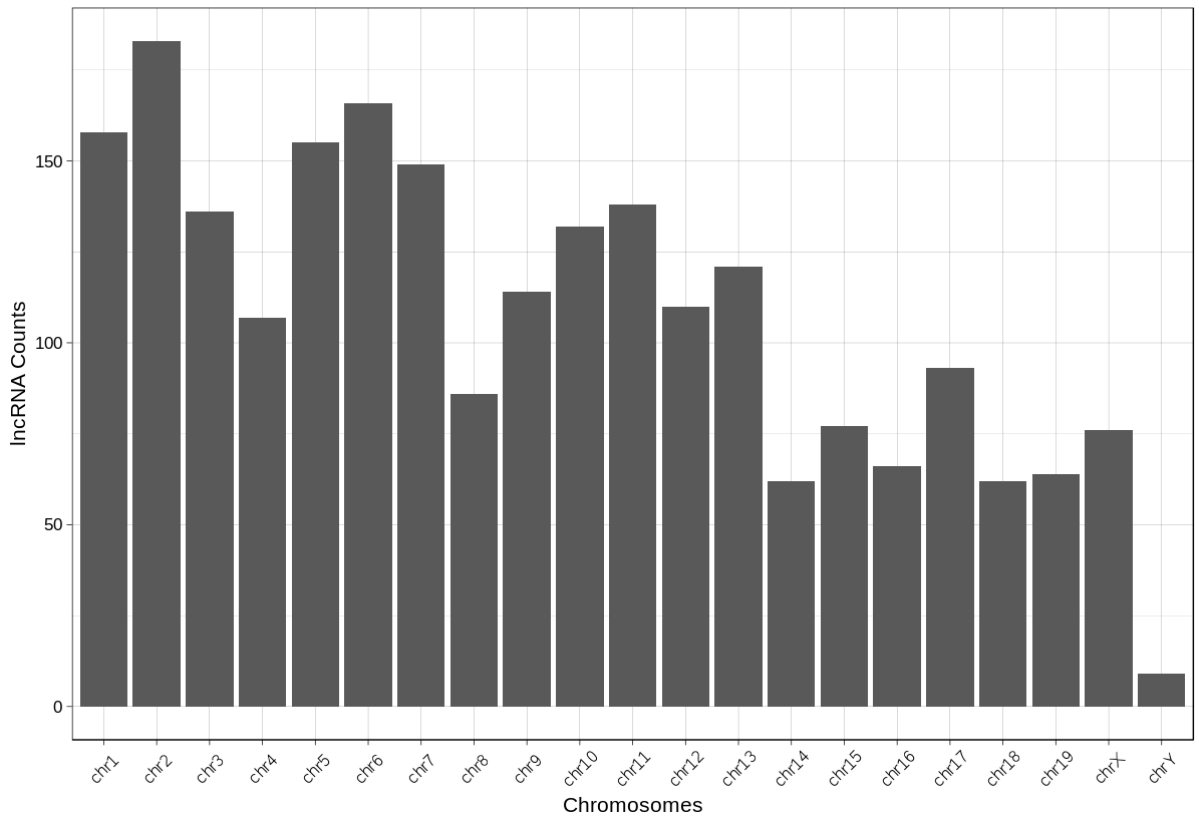
transcripts, 53923 were found to align with a total of 24261 previously annotated genes, both coding and noncoding. The remaining 2470 transcripts were classified into 2359 previously unannotated genes, which are either intergenic in comparison to known genes, or found on the anti-sense strand of one or more previously annotated exons.

4.2. Known and previously unannotated lncRNAs

The 2359 transcripts identified in the previous section were then subjected to further filters, such as size selection and similarity to known housekeeping RNAs, as explained in 3.5.4. As a result, a final pool of 593 putative lncRNA genes coding for 608 transcripts was identified.

The genomic context of the identified lncRNAs, both previously unannotated and known, were subjected to further analysis. The highest number of lncRNA genes was on chromosome 2 (Fig. 7A) while the highest concentration of lncRNA genes (i.e. number of lncRNA genes per hundred million base pairs of chromosome) was in chromosome 11 (Fig. 7B). The number of lncRNA genes per chromosome has a 0.71 Pearson Correlation Coefficient with chromosome size ($p < 0.001$).

A.



B.

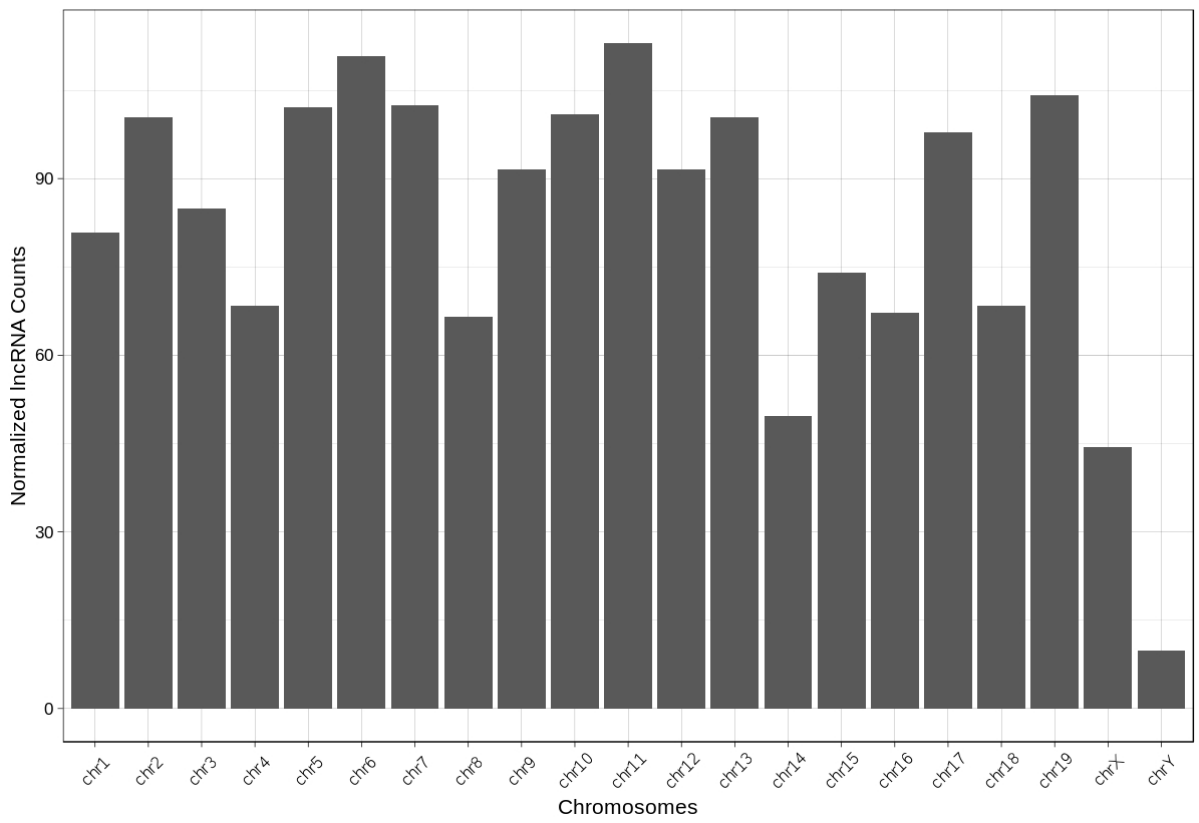


Fig. 7. Barplots showing count (A) and density (B) of lncRNA genes on canonical mouse chromosomes.

“Normalized lncRNA Counts” value represents number of lncRNA genes per hundred million base pairs on chromosome. Number of lncRNA genes per chromosome has a 0.71 Pearson Correlation Coefficient with chromosome size ($p < 0.001$).

In addition to the 608 putative lncRNA transcripts, there were 159 transcripts coding for annotated lncRNAs, for a total of 767. The majority of putative lncRNAs were single-exon transcripts, while the annotated transcripts clustered around 2- and 3-exon transcripts (Fig. 8).

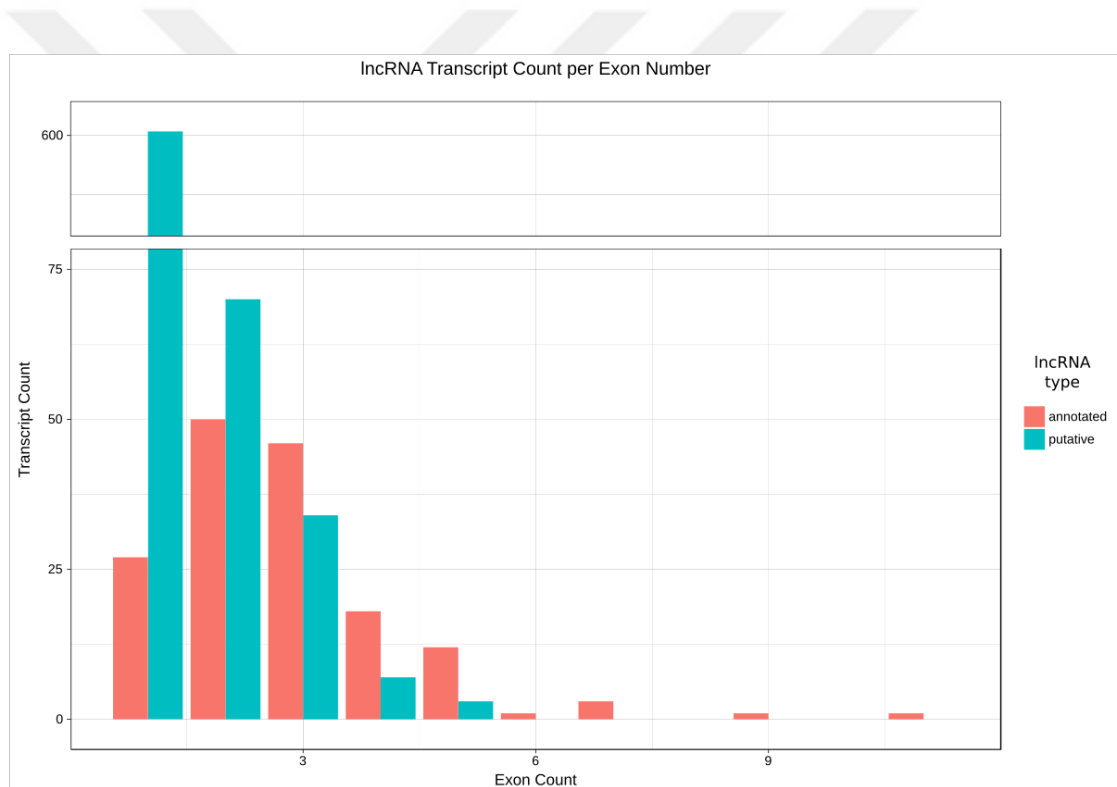


Fig. 8. Barplot representing the exon count distribution of lncRNAs expressed in at least one timepoint during the EMT-MET processes.

Red bars indicate previously annotated lncRNA genes. Turquoise bars indicate putative lncRNA candidates. Majority of high exon count (exon number ≥ 5) lncRNAs have previously been annotated. Largest fraction of previously unannotated genes are single exon transcripts, likely due to difficulty of experimental validation of such genes.

One previously annotated lncRNA, 5430416N02Rik, was visualized on Integrated Genome Viewer to confirm the validity of read mappings (Fig. 9). As the figure indicates, the reads largely map onto exonic sequences, with much lower coverage of intronic sequences, showing that the alignment was largely successful.

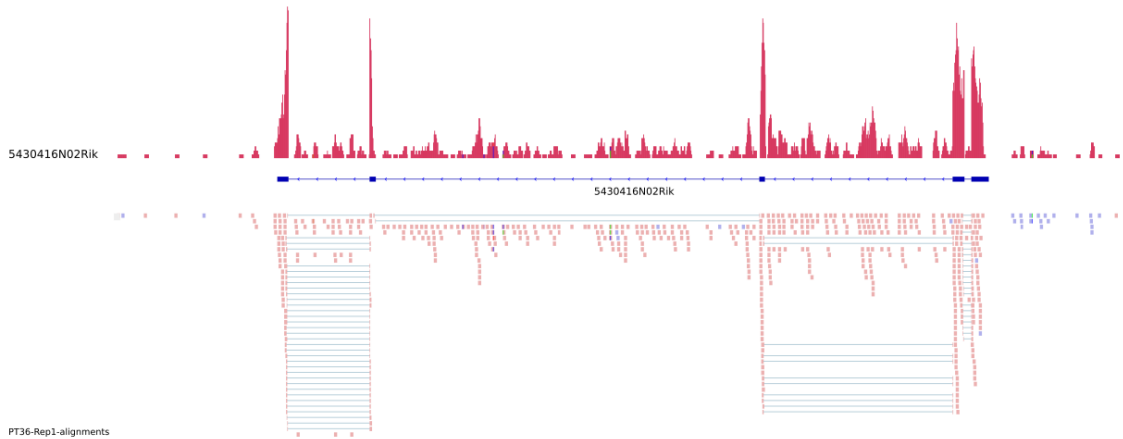


Fig. 9. Coverage and read mapping visuals of annotated lncRNA, 5430416N02Rik.

Upper portion represents coverage, lower portion represents reads. Blue lines on the lower portion represent single reads from transcript covering two exons. Coverage is highest in exonic sequences. Blue lines between ends of sequencing reads align to intronic sequences, showing spliced out sites present on the genome, but absent in mature transcripts.

4.3. Expression profiles of known and previously unannotated lncRNAs

The expression profiles of the samples were visualized, specifically in terms expression level distribution per gene type (Fig 10). As expected, the average expression levels of lncRNA transcripts were lower than protein-coding transcripts. However, given the high number of relatively lowly expressed protein-coding genes in cells, a number of outlier lncRNA transcripts had expressions much higher than the average of protein-coding gene expression.

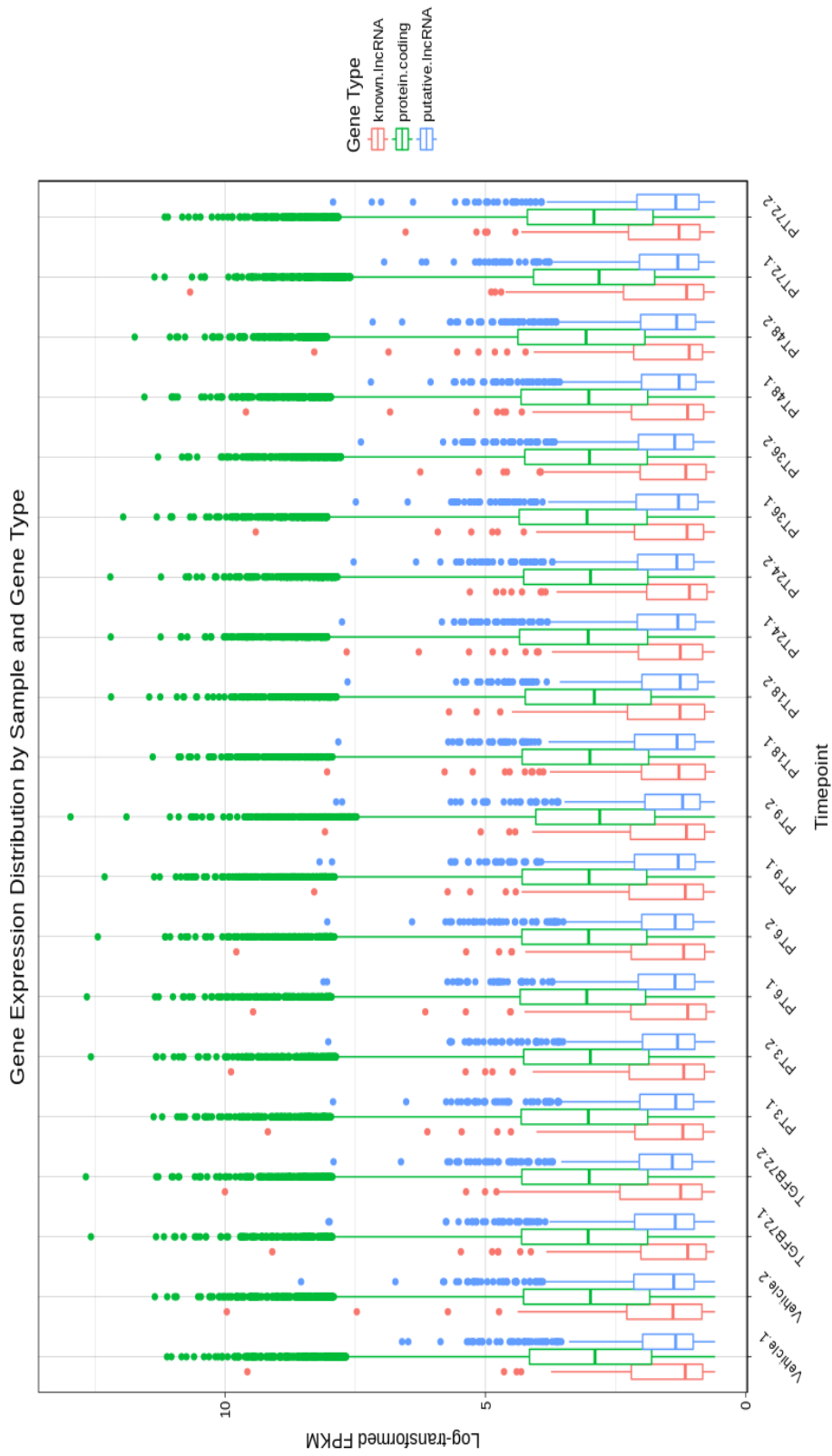


Fig. 10. The distribution of gene expressions represented by boxplots.

Whiskers represent largest and smallest values no further than 1.5 times the interquartile range. Dots represent outliers beyond the whiskers.

To visualize the expression patterns of the transcribed lncRNAs, a heatmap was constructed. As expected, a majority of the lncRNAs displayed dynamic expression patterns across the course of the EMT-MET process (Fig. 11)

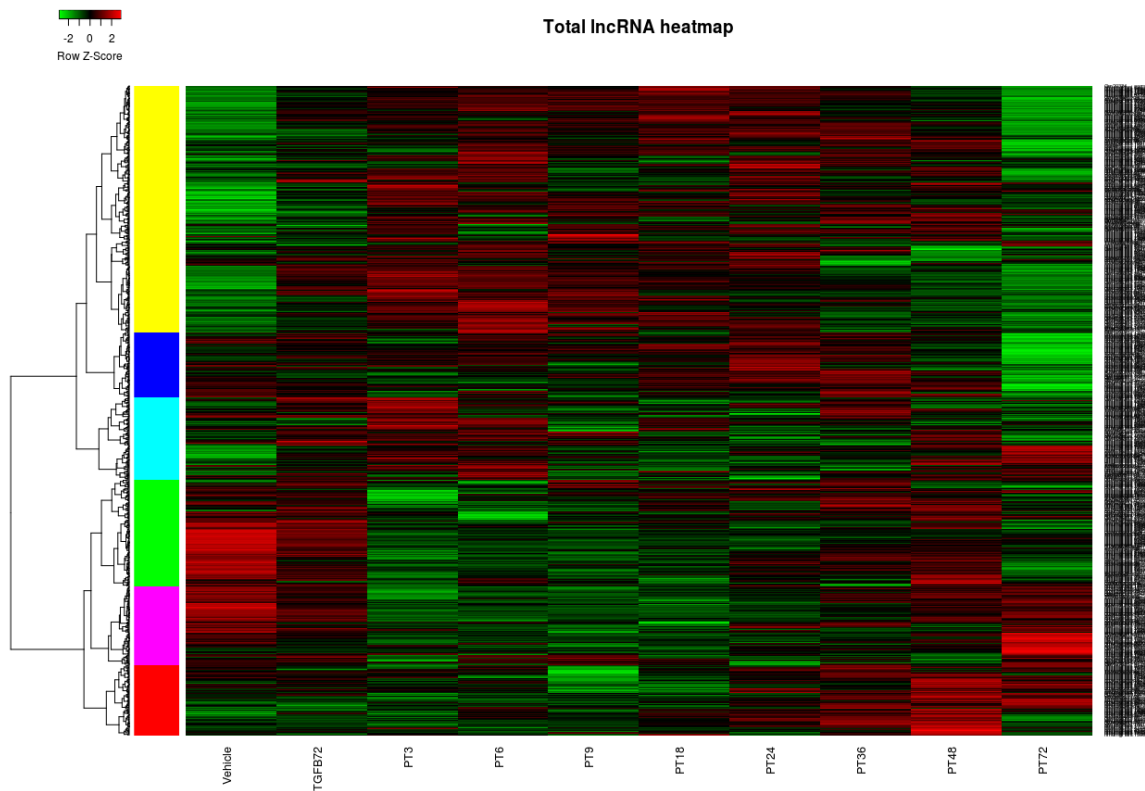


Fig. 11. Heatmap representing the expression variances of all lncRNA transcripts.

Red cells represent high expression compared to other timepoints, while green cells represent low expression. Black cells indicate expression equal to the average of all time points, with bright red and bright green cells being up to 3 standard deviations higher or lower in expression than the mean, respectively. Dendrogram on the left generated by using Ward's clustering.

Using principal component analysis and hierarchical clustering methods, the samples were analyzed in terms of the distance of their expression profiles from each other to identify any possible outliers. The clusters were formed using Euclidean distance.

The dendrogram plot of the hierarchical clustering showed that the second replica of PT9 was a strong outlier among the analyzed samples, with a merge point of 8000+, while the remaining samples had a maximal merge point of 5000+. (Fig. 12). The sample was removed from following analysis steps in order not to skew the results.

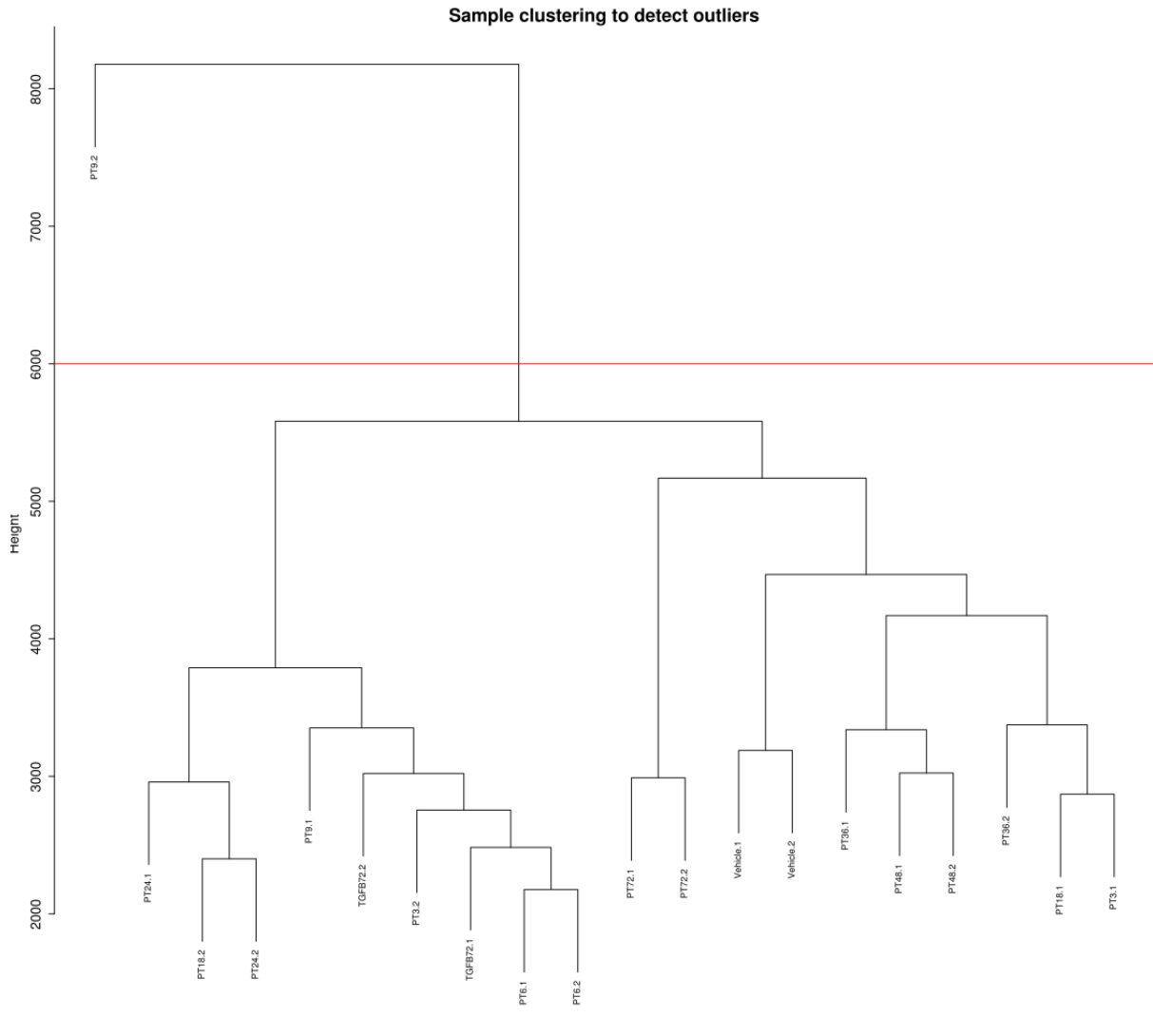


Fig. 12. Dendrogram representing clustering of analyzed samples based on their intersample distances.

Samples were clustered using a hierarchical clustering of their Pearson correlations, using the average clustering method. PT9.2 sample was removed as an outlier using a cutoff excluding any samples with a height greater than 6000, as represented by the red line.

The remaining samples were found to be arranged as expected on the primary principal component, with the timepoints displaying epithelial morphology (Vehicle and PT72) being on the opposite end of PC1 from the timepoint displaying mesenchymal morphology (TGFB72), with the remaining timepoints showing regression towards the epithelial end throughout the course of the MET process (Fig. 13). In the figure, PC1 accounts for the majority of the variance between samples, and 41.14% of the variance is unexplained. The distance between the Vehicle

samples and the PT72 samples likely indicates that the full epithelial morphology and expression profile is not yet present in PT72. The regression of the samples between Vehicle and TGFB72 indicates that the transition process occurs gradually across the 72 hours of the time course experiment, rather than taking place at a single timepoint.

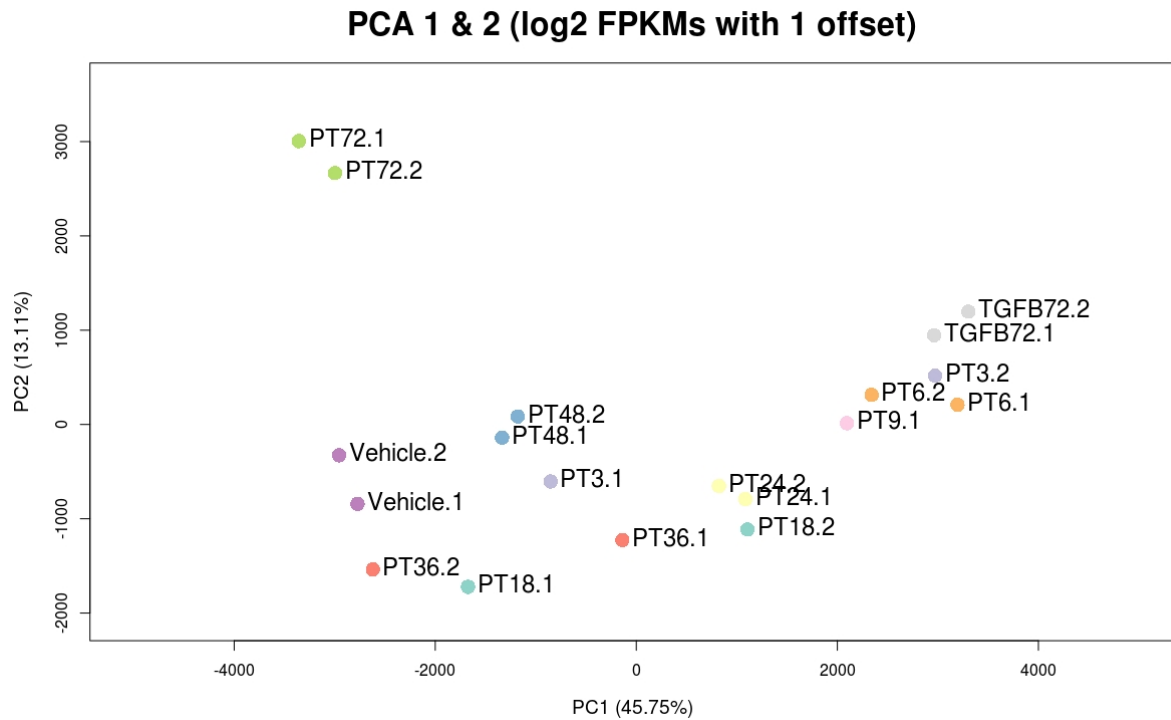


Fig. 13. Principal component analysis plot of variance between analyzed samples.

Epithelial phenotype samples are located on the left (Vehicle.1, Vehicle.2, PT72.1, PT72.2), while mesenchymal phenotype samples are located on the right (TGFB72.1, TGFB72.2). PC1 corresponds to the transcriptomic variance between epithelial and mesenchymal cells, as the largest contributor to total variance. The exact cause of PC2, primarily seen in the gap between Vehicle and PT72 cells, is unknown.

Of the lncRNAs expressed in the samples, the highest expression level belonged to Malat1, with a mean FPKM of 882.03 in the Vehicle samples, and an FPKM of 1634.54 in replica PT72.1 (Fig.14). However, the highest variance in sample averages was in putative lncRNA, NH.1987, with a standard deviation of 4.006, a transcript that is only commonly expressed in both replicas in the PT9 samples (Fig. 15).



Fig. 14. Expression profile of known lncRNA Malat1.

Timepoints represent average expression of both replicas. Points on the line indicate log₂-transformed FPKM values at specific timepoints. Highest expression seen in Vehicle samples, with an average FPKM of 882.03.



Fig. 15. Expression profile of previously unannotated lncRNA NH.1987.

Timepoints represent average expression of both replicas. Points on the line indicate log₂-transformed FPKM values at specific timepoints. Highest mean expression seen in PT9 samples.

Seven of the 50 lncRNAs with highest variance showed high downregulation or lack of expression in a single timepoint (Fig. 16).

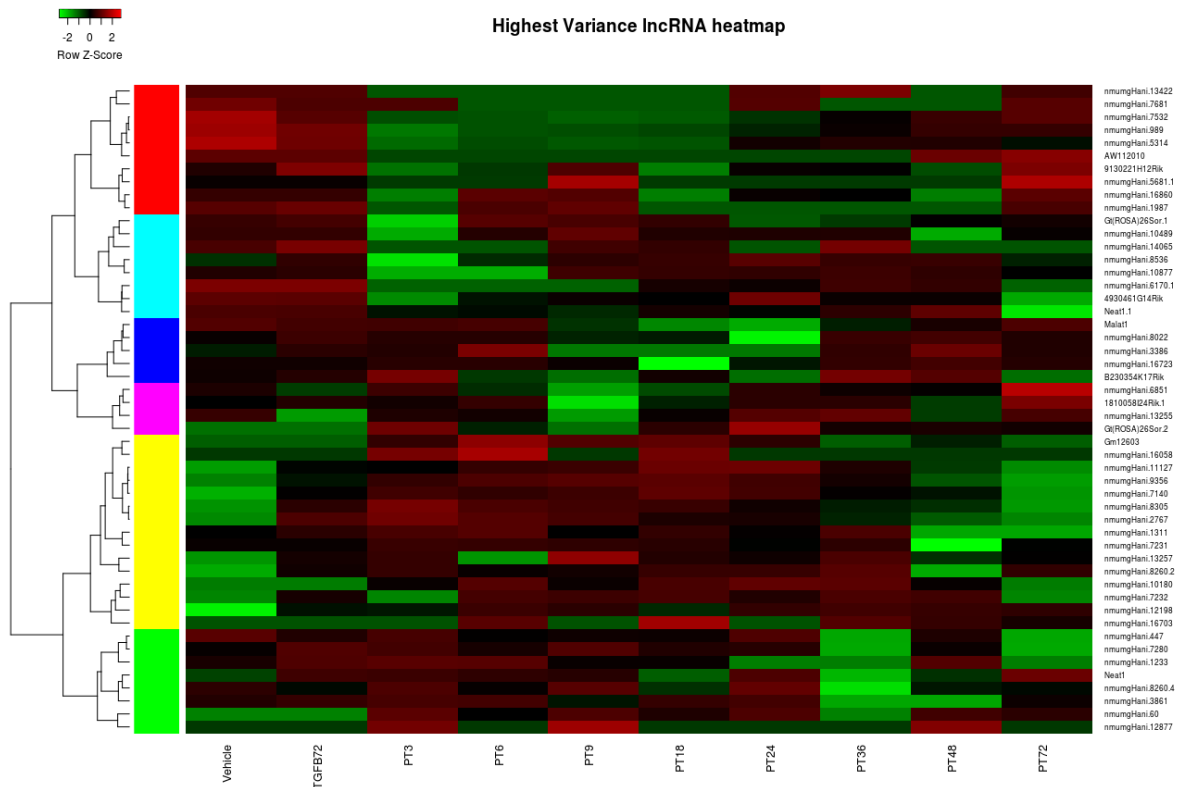


Fig. 16. Heatmap representing the expression variances of 50 lncRNA transcripts with highest variance.

Red cells represent high expression compared to other timepoints, while green cells represent low expression. Black cells indicate expression equal to the average of all time points, with bright red and bright green cells being up to 3 standard deviations higher or lower in expression than the mean, respectively. Dendrogram on the left generated by using Ward's clustering.

To predict lncRNAs that might positively contribute to the maintenance or reacquisition of the epithelial morphology, lncRNAs that show higher than 1.5 fold enrichment at any given timepoint compared to the expression level of TGFB72. A total of 221 lncRNAs were identified as upregulated in Vehicle or MET samples (Fig. 17).

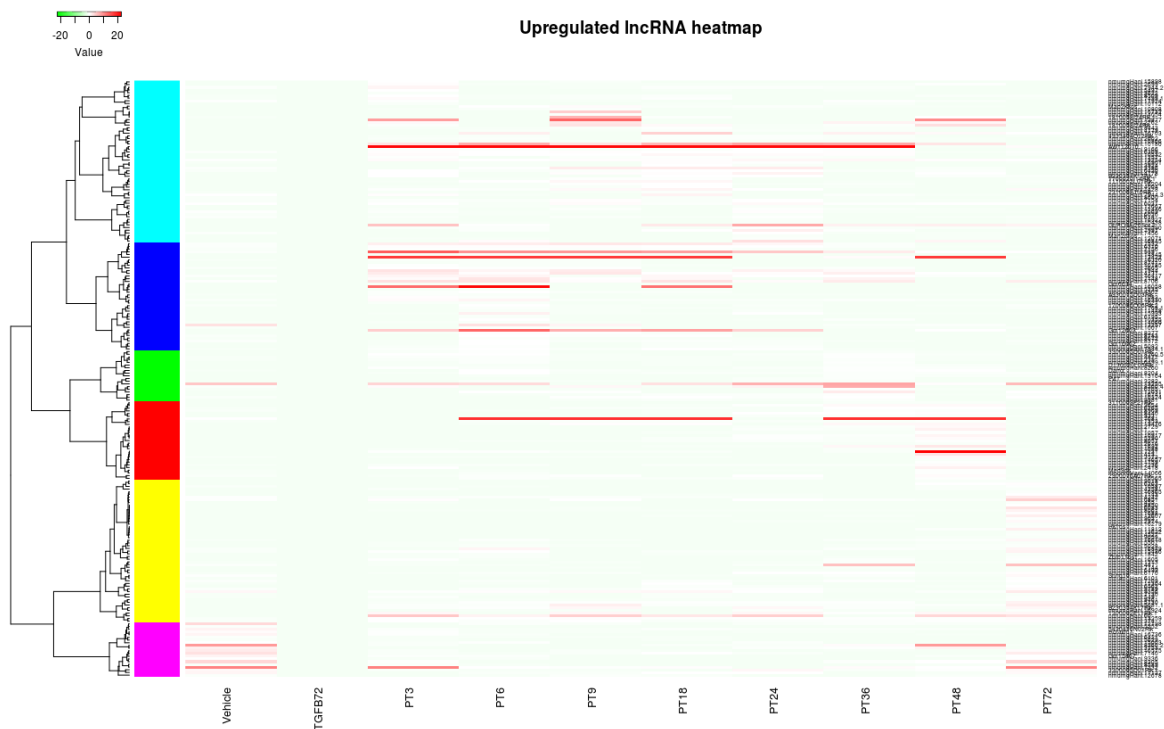


Fig. 17. Heatmap representing lncRNAs with positive contributions to the epithelial phenotype.

Red cells indicate transcripts upregulated in a timepoint compared to the mesenchymal phenotype. Cell values represent \log_2 of the ratio of transcript expression in timepoint compared to the expression of same transcript in TGFB72 samples. White cells indicate expression equal to the TGFB72 samples, with bright red and bright green cells being up to 20 levels of magnitude higher or lower in expression than TGFB72 expression, respectively. Dendrogram on the left generated by using Ward's clustering.

After the expression level analysis of the lncRNA genes were performed, the results were combined with genomic coordinate data to create a Circos plot representing the noncoding transcriptomic landscape of the time course experiment (Fig. 18).

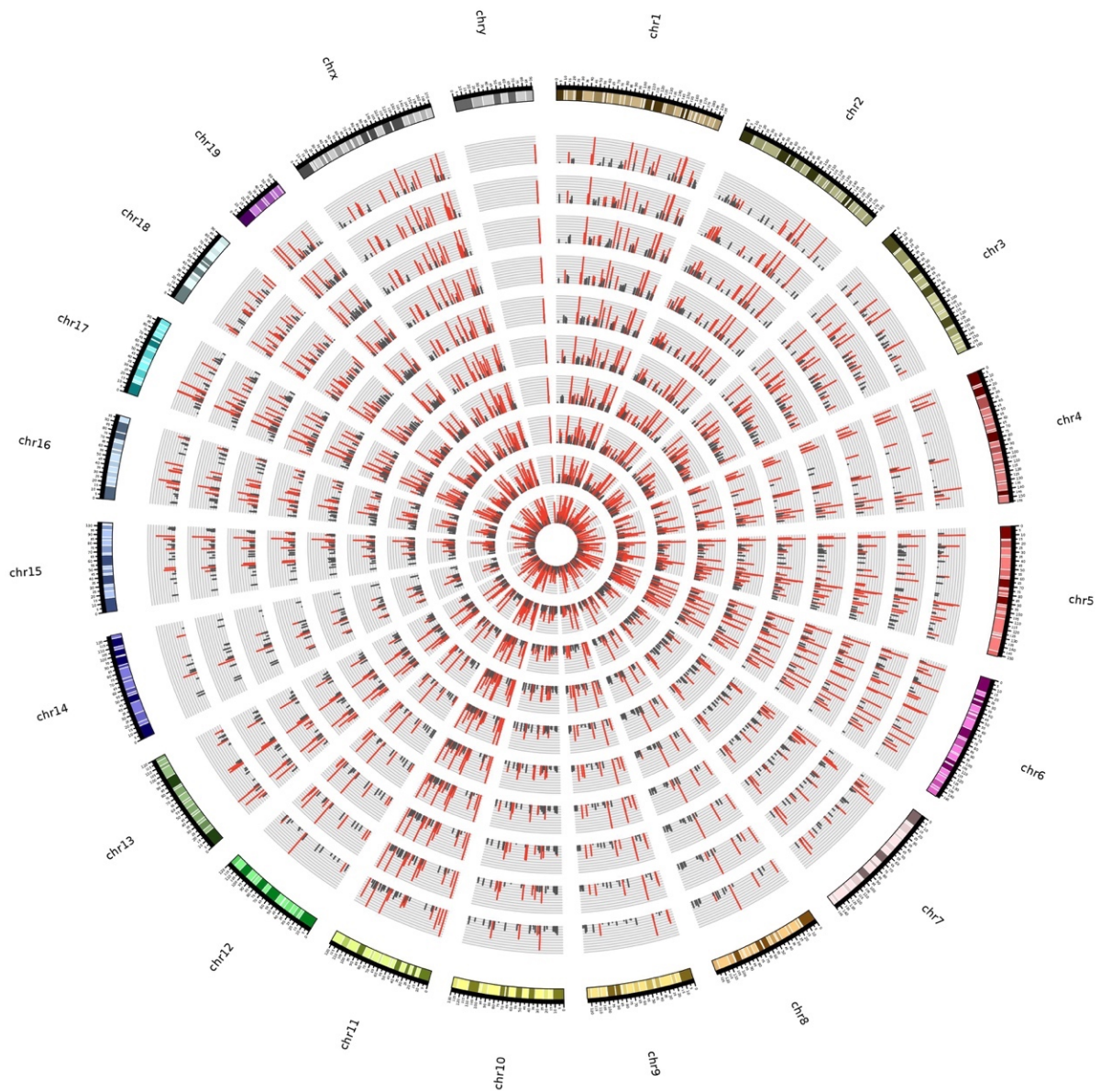


Fig. 18. Circos plot displaying the genomic coordinates and expression patterns of putative and annotated lncRNAs during the EMT-MET experiment.

From the innermost to the outermost (excluding the karyotype bands), the circles correspond to Vehicle, B72, PT3, PT6, PT9, PT18, PT24, PT36, PT48 and PT72. Bars in the inner circles represent lncRNA transcript location and expression level. Red bars represent highly expressed transcripts, while black bands represent lowly expressed transcripts.

4.4. Differential expression analysis of total lncRNA

The lncRNA transcripts were examined with Ballgown's `stattest` command for statistically noteworthy differentially expressed genes. As a result, 75 lncRNA genes were

discovered to have a differential expression pattern with a p- or q-value lower than .001, 70 of which were previously unannotated lncRNAs (Table 2).

Table 2. lncRNA genes with high differential expression patterns during EMT-MET

Gene ID	<i>P</i> value	<i>q</i> value
3110009F21Rik	0.00023877	0.01995703
B230217O12Rik	0.00965488	0.05063333
C330013E15Rik	0.00564824	0.04145879
Gm19589	0.00366885	0.03735155
Gm4961	0.00032066	0.02140844
nmumgHani.10162	0.00225609	0.03222648
nmumgHani.10185	0.00382036	0.03768978
nmumgHani.10257	0.00062999	0.02451894
nmumgHani.10532	0.003843	0.03769602
nmumgHani.10866	0.0086361	0.04819055
nmumgHani.1088	0.00230785	0.03234918
nmumgHani.10917	0.00333275	0.03627914
nmumgHani.11127	0.00565067	0.04145879
nmumgHani.11293	0.00044881	0.02261959
nmumgHani.1132	0.00804142	0.04700911
nmumgHani.12572	0.00938812	0.04990896
nmumgHani.12783	0.00766054	0.04628964
nmumgHani.13622	0.00433424	0.03878368
nmumgHani.13891	0.00077926	0.02554342
nmumgHani.14060	0.00732248	0.04538037
nmumgHani.14694	0.00232237	0.03239031
nmumgHani.15138	0.00672403	0.04413224
nmumgHani.16014	0.00492201	0.03963535
nmumgHani.16378	0.00137129	0.0291501
nmumgHani.1665	0.00313767	0.03563981
nmumgHani.16736	0.00447879	0.03883637
nmumgHani.16741	0.00038249	0.02209387
nmumgHani.1682	0.00230732	0.03234918
nmumgHani.197	0.0017416	0.03050853
nmumgHani.2431	0.00623941	0.04302838
nmumgHani.2616	0.00205647	0.03165211
nmumgHani.272	0.00381136	0.03768978
nmumgHani.2767	0.0030885	0.03534156
nmumgHani.2792	0.00116415	0.02757942
nmumgHani.3282	0.00219682	0.03197819

nmumgHani.3338	0.00309775	0.03538836
nmumgHani.4179	0.00415485	0.03834572
nmumgHani.4470	0.00429371	0.03874546
nmumgHani.4472	2.48E-05	0.01460339
nmumgHani.4795	0.00962549	0.05049209
nmumgHani.4796	0.00029373	0.02048028
nmumgHani.5507	0.00395745	0.03789225
nmumgHani.5626	0.00516445	0.04012289
nmumgHani.6031	0.00536981	0.04074583
nmumgHani.6170	0.00986233	0.05111954
nmumgHani.6185	0.00804736	0.0470241
nmumgHani.6251	0.00704146	0.04477278
nmumgHani.6562	0.00569142	0.04157463
nmumgHani.6735	0.00755311	0.04600549
nmumgHani.7296	0.00228692	0.03229758
nmumgHani.7499	0.00880825	0.04863698
nmumgHani.7532	0.00836554	0.04771724
nmumgHani.7535	0.00525996	0.04041871
nmumgHani.7665	0.00739564	0.04564122
nmumgHani.7799	0.00560837	0.04134993
nmumgHani.7944	0.00787302	0.04670577
nmumgHani.7997	0.00874849	0.0485547
nmumgHani.8206	0.00372536	0.03751077
nmumgHani.8208	0.00215487	0.03192394
nmumgHani.8297	0.00321663	0.03589248
nmumgHani.8305	0.00600731	0.042435
nmumgHani.844	0.000374	0.02209387
nmumgHani.8440	0.00050491	0.02352634
nmumgHani.8452	0.00091292	0.02614542
nmumgHani.8509	0.00397318	0.03790111
nmumgHani.852	0.00389772	0.03779624
nmumgHani.8589	0.00146287	0.02963673
nmumgHani.9018	0.00544646	0.04096799
nmumgHani.9166	0.00014075	0.01808889
nmumgHani.9356	0.00106469	0.0271125
nmumgHani.9690	0.0011702	0.02762549
nmumgHani.9857	0.00639436	0.04347181
nmumgHani.986	0.00936704	0.04987277
nmumgHani.989	0.00977599	0.05095293
nmumgHani.9981	0.0077198	0.0464023

4.5. Weighted co-expression network formation and module membership

Out of 31970 transcripts identified as belonging to protein-coding or lncRNA genes (putative or previously annotated) that showed notable expression in both replicas of at least one timepoint ($\geq .5$ FPKM for lncRNAs, ≥ 1.0 FPKM for protein-coding genes), 11261 of them were removed before clustering due to zero-variance observed in their expression values across samples. The remaining 20709 transcripts were assigned to 30 modules that display internal similarity in terms of their clustering, with 245 of these transcripts not showing strong correlation with the modules and being left out of module membership. An unsigned adjacency network was calculated from the remaining transcripts. `maxBlockSize` was arbitrarily set as 25000, larger than the data set given, to create an unbounded module block size. The soft thresholding power was selected as 5. The minimum module size was set at 30 members (Fig 19). Table 3 shows the gene membership properties of each module.

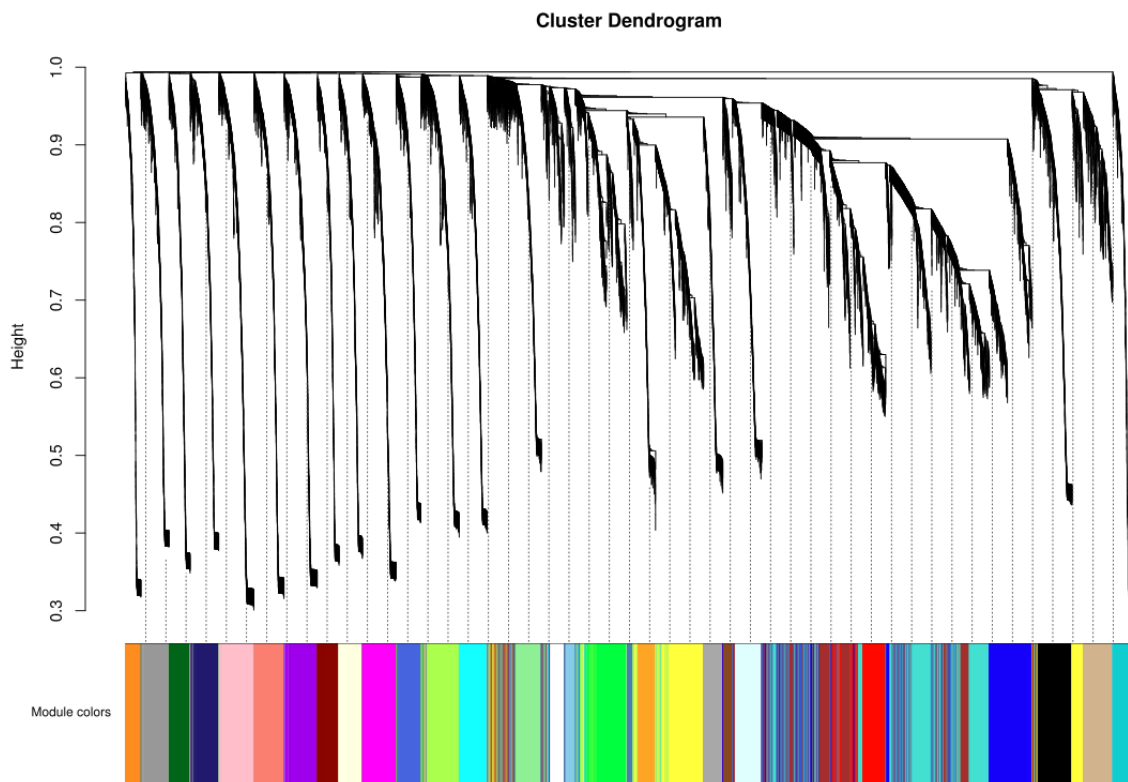


Fig. 19. Dendrogram of all transcripts identified in the samples and their module membership represented by colors.

Each branch in dendrogram represents expression profile of a specific transcript across the timecourse experiment. Module colors are assigned based on similarities between the connections of different genes with neighboring genes using average hierarchical clustering.

Table 3. Transcript count per module divided by gene type (protein-coding, annotated lncRNA, previously unannotated lncRNA, total lncRNA, total gene).

Module	Known lncRNA Count	Putative lncRNA Count	Protein Coding Gene Count	Total lncRNA Count	Total Gene Count
Turquoise	53	126	2412	179	2591
Blue	40	71	1520	111	1631
Brown	43	57	1443	100	1543
Yellow	34	44	1383	78	1461
Green	14	30	823	44	867
Red	11	48	794	59	853
Black	28	24	666	52	718
Pink	22	15	657	37	694
Magenta	9	26	649	35	684
Purple	22	13	635	35	670
Greenyellow	23	20	607	43	650
Tan	20	23	584	43	627
Salmon	21	19	574	40	614
Cyan	18	20	533	38	571
Midnightblue	17	9	533	26	559
Lightcyan	16	21	519	37	556
Grey60	19	7	512	26	538
Lightgreen	14	19	491	33	524
Lightyellow	19	17	448	36	484
Royalblue	18	13	448	31	479
Darkred	18	20	409	38	447
Darkgreen	11	11	412	22	434
Darkturquoise	12	2	408	14	422
Darkgrey	10	8	393	18	411
Orange	11	11	358	22	380
Darkorange	16	4	299	20	319
White	8	10	260	18	278
Skyblue	4	7	185	11	196
Saddlebrown	5	15	150	20	170
Steelblue	3	1	136	4	140
Totals	559	711	19241	1270	20511

4.6. Module-wise gene ontology enrichment

To identify the potential niches of any previously unannotated lncRNAs found in the modules, the annotated members of each module were used for functional annotation

enrichment, with a background list of all mouse genes annotated in the mm10 assembly, and a Bonferroni threshold of $1e-4$. Of particular note were modules enriched for terms relating to chromatin organization and remodeling, cell differentiation and cellular localization (Fig. 20).



Fig. 20. Count plot showing partial GO term enrichment results for the identified modules.

Size of circles indicate genes in that module enriched for the corresponding GO term; fill color indicates false discovery rate, with red indicating low FDR (i.e. higher statistical significance). Plot representing full enrichment results not included due to size constraints.

4.7. Identification of modules with timepoint-specific upregulated average expression

In order to identify peak activity periods for each module and its member genes, the average expression values of the modules were calculated and plotted (Fig. 21). Of interest are modules that display clear peaks at specific timepoints, such as the black module at PT48, and darkred module at PT6.



Fig. 21. Line plots showing average (dark line) and individual (light lines) gene expression levels of modules calculated by WGCNA.

Peaks in the dark lines indicate increased module activity during specific timepoints, with genes in the corresponding module displaying increased or decreased expression in a coregulated manner. Modules with a large number of member genes show largely constant expression values.

4.8. Identification of lncRNAs with timepoint-specific upregulation

To further identify potential relationships between the lncRNAs and MET, the expression values of the genes were used to determine their timepoint-specificity, using the tissue-specific gene identification methods, ROKU and Tau. 116 putative lncRNA transcripts were over-expressed outliers in a single timepoint according to the ROKU method, and 91 lncRNA

transcripts had a Tau score of > 0.6 , with an overlap of 25 transcripts. Out of those 25 transcripts, two putative and two annotated lncRNA transcripts had a Tau score of 1, indicating increased expression exclusive to a single timepoint (Fig. 22).

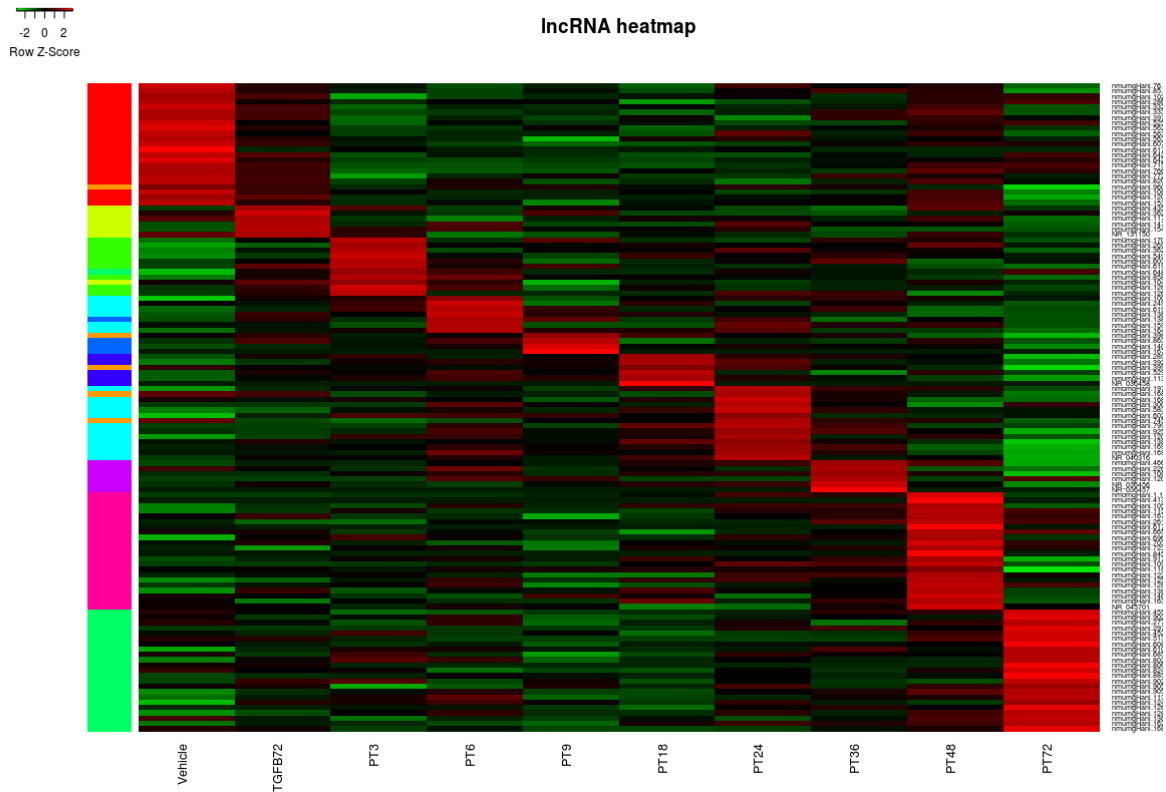


Fig. 22. Heatmap representing gene expression Z-scores of timepoint specific transcripts. Red cells represent high expression compared to other timepoints, while green cells represent low expression. Black cells indicate expression equal to the average of all time points, with bright red and bright green cells being up to 3 standard deviations higher or lower in expression than the mean, respectively. The bar on the left of the heatmap represents clustering of gene expressions, displaying strong correlation with ROKU identified upregulation.

To confirm the validity of the timepoint specific expression predictions, the reads were visualized in IGV. Figure 23 shows the coverage of putative lncRNA NH.7997, which the ROKU method calculated to be specifically expressed at PT24, in three different timepoints (Vehicle, PT24 and PT36), with the exons of the transcript NH.7997.1 highlighted.

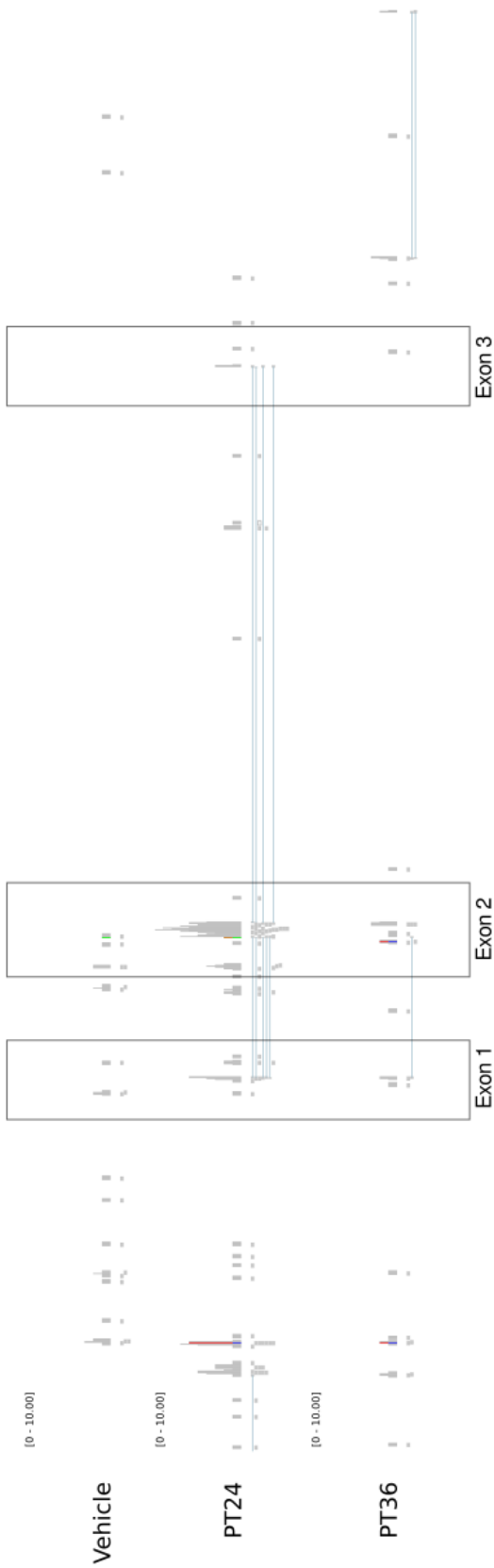


Fig. 23. Coverage and individual mapped reads representing NH.7997.1 expression in three different timepoints.

Boxes highlight the exons of transcript, with highest coverage for each exon found in PT24. Remaining timepoints not shown for image clarity. Exon 2 has highest coverage while NH.7997.1 is transcriptionally active. No RefSeq gene annotations map to the locations of the exons.

4.9. Prediction of potential partners of lncRNAs of interest

Using the expression values of genes, an unsigned network adjacency matrix was calculated for the transcriptome expressed in the samples. Each pair of genes in the matrix had a score between 0 and 1, with 0 indicating no correlation, and values close to 1 indicating strong correlation, either positive, i.e. both genes are upregulated in the same samples, or negative, i.e. one gene is upregulated in the samples that the other is downregulated. Gene pairs with a correlation absolute value of >0.8 were considered connected for a stringent network definition.

The module membership values of the genes were calculated using the signedKME command to identify any possible hub genes. Of these genes, 691 were identified as hub genes, 24 of which were putative lncRNAs, and 19 of which were previously annotated lncRNAs.

In addition, four modules identified were found to have a previously unannotated lncRNA as the top hub gene. These modules were darkturquoise, greenyellow, lightgreen and lightyellow. The enriched GO terms of these modules were largely involved with transcription, translation, protein modification and cellular localization of organelles and macromolecules (Table 4).

Table 4. Gene ontology term enrichments for modules with putative lncRNAs identified as top hub gene

Class	Rank	Data Set ID	Data Set Name	P Value
darkturquoise	54	GO:0006464	cellular protein modification process	9.05E-17
darkturquoise	55	GO:0036211	protein modification process	9.05E-17
darkturquoise	72	GO:0034654	nucleobase-containing compound biosynthetic process	3.93E-15
darkturquoise	75	GO:0019219	regulation of nucleobase-containing compound metabolic process	4.28E-15
darkturquoise	76	GO:0018130	heterocycle biosynthetic process	6.95E-15
darkturquoise	77	GO:0031326	regulation of cellular biosynthetic process	7.59E-15
darkturquoise	78	GO:0019438	aromatic compound biosynthetic process	8.12E-15
darkturquoise	80	GO:0009889	regulation of biosynthetic process	1.37E-14
darkturquoise	81	GO:0051171	regulation of nitrogen compound metabolic process	1.66E-14
darkturquoise	82	GO:1901362	organic cyclic compound biosynthetic process	2.21E-14
darkturquoise	83	GO:0010556	regulation of macromolecule biosynthetic process	2.61E-14
darkturquoise	84	GO:0006950	response to stress	2.94E-14
darkturquoise	86	GO:2000112	regulation of cellular macromolecule biosynthetic process	3.78E-14
darkturquoise	89	GO:0010468	regulation of gene expression	8.93E-14
darkturquoise	90	GO:0006351	transcription, DNA-templated	1.06E-13
darkturquoise	91	GO:0048869	cellular developmental process	1.14E-13
darkturquoise	93	GO:0097659	nucleic acid-templated transcription	1.25E-13

darkturquoise	95	GO:0032774	RNA biosynthetic process	1.45E-13
darkturquoise	99	GO:0051252	regulation of RNA metabolic process	2.01E-13
darkturquoise	101	GO:0006355	regulation of transcription, DNA-templated	2.11E-13
darkturquoise	102	GO:1903506	regulation of nucleic acid-templated transcription	2.50E-13
darkturquoise	103	GO:2001141	regulation of RNA biosynthetic process	2.69E-13
darkturquoise	108	GO:0048731	system development	1.16E-12
darkturquoise	109	GO:0030154	cell differentiation	1.55E-12
darkturquoise	114	GO:0051246	regulation of protein metabolic process	1.73E-11
darkturquoise	120	GO:0032268	regulation of cellular protein metabolic process	4.57E-11
darkturquoise	121	GO:0033036	macromolecule localization	4.96E-11
greenyellow	135	GO:0051128	regulation of cellular component organization	2.76E-16
greenyellow	136	GO:0044248	cellular catabolic process	6.15E-16
greenyellow	142	GO:0009056	catabolic process	3.23E-15
greenyellow	145	GO:0033554	cellular response to stress	4.82E-15
greenyellow	146	GO:0032879	regulation of localization	5.62E-15
greenyellow	151	GO:0031325	positive regulation of cellular metabolic process	4.12E-14
greenyellow	152	GO:0032268	regulation of cellular protein metabolic process	4.39E-14
greenyellow	153	GO:0098609	cell-cell adhesion	5.12E-14
greenyellow	154	GO:0051239	regulation of multicellular organismal process	8.29E-14
greenyellow	155	GO:0007155	cell adhesion	1.05E-13
greenyellow	158	GO:0022610	biological adhesion	1.33E-13
greenyellow	159	GO:0010604	positive regulation of macromolecule metabolic process	1.39E-13
greenyellow	162	GO:0051246	regulation of protein metabolic process	1.66E-13
greenyellow	164	GO:0050793	regulation of developmental process	1.68E-13
greenyellow	165	GO:0065009	regulation of molecular function	2.03E-13
greenyellow	166	GO:0002376	immune system process	2.29E-13
greenyellow	167	GO:0010033	response to organic substance	2.61E-13
greenyellow	168	GO:0042592	homeostatic process	2.66E-13
greenyellow	169	GO:1901564	organonitrogen compound metabolic process	3.00E-13
greenyellow	170	GO:0033036	macromolecule localization	3.00E-13
greenyellow	171	GO:1902578	single-organism localization	3.63E-13
greenyellow	172	GO:1901575	organic substance catabolic process	4.83E-13
greenyellow	173	GO:0043933	macromolecular complex subunit organization	4.85E-13
greenyellow	174	GO:0009968	negative regulation of signal transduction	4.95E-13
greenyellow	175	GO:1902531	regulation of intracellular signal transduction	5.23E-13
greenyellow	176	GO:0070727	cellular macromolecule localization	5.33E-13
greenyellow	177	GO:0006928	movement of cell or subcellular component	5.93E-13
greenyellow	184	GO:0010648	negative regulation of cell communication	1.64E-12
greenyellow	185	GO:0023057	negative regulation of signaling	1.79E-12
greenyellow	186	GO:0034613	cellular protein localization	2.09E-12
greenyellow	187	GO:0048585	negative regulation of response to stimulus	2.42E-12

greenyellow	189	GO:0051649	establishment of localization in cell	2.95E-12
greenyellow	190	GO:0040011	locomotion	3.03E-12
greenyellow	191	GO:0031324	negative regulation of cellular metabolic process	3.30E-12
greenyellow	192	GO:0009653	anatomical structure morphogenesis	3.32E-12
greenyellow	194	GO:0051173	positive regulation of nitrogen compound metabolic process	5.71E-12
greenyellow	195	GO:0009891	positive regulation of biosynthetic process	6.53E-12
greenyellow	196	GO:0048878	chemical homeostasis	7.09E-12
greenyellow	198	GO:0006366	transcription from RNA polymerase II promoter	7.73E-12
greenyellow	199	GO:0044085	cellular component biogenesis	9.08E-12
greenyellow	200	GO:2000026	regulation of multicellular organismal development	1.06E-11
greenyellow	202	GO:0008104	protein localization	1.25E-11
greenyellow	203	GO:0009892	negative regulation of metabolic process	1.44E-11
greenyellow	204	GO:0010941	regulation of cell death	1.74E-11
greenyellow	205	GO:0016477	cell migration	3.20E-11
greenyellow	206	GO:0043067	regulation of programmed cell death	4.56E-11
greenyellow	207	GO:0012501	programmed cell death	4.86E-11
greenyellow	209	GO:0018193	peptidyl-amino acid modification	4.98E-11
greenyellow	210	GO:0048870	cell motility	5.32E-11
greenyellow	211	GO:0051674	localization of cell	5.32E-11
greenyellow	212	GO:0031328	positive regulation of cellular biosynthetic process	6.64E-11
greenyellow	213	GO:0072359	circulatory system development	8.42E-11
greenyellow	214	GO:0009605	response to external stimulus	8.85E-11
greenyellow	215	GO:0008219	cell death	9.06E-11
greenyellow	216	GO:0006915	apoptotic process	1.20E-10
lightgreen	109	GO:0006366	transcription from RNA polymerase II promoter	3.16E-16
lightgreen	110	GO:0006357	regulation of transcription from RNA polymerase II promoter	5.46E-16
lightgreen	115	GO:0010557	positive regulation of macromolecule biosynthetic process	1.32E-14
lightgreen	116	GO:0045935	positive regulation of nucleobase-containing compound metabolic process	3.62E-14
lightgreen	117	GO:0051173	positive regulation of nitrogen compound metabolic process	3.89E-14
lightgreen	118	GO:0051254	positive regulation of RNA metabolic process	8.46E-14
lightgreen	119	GO:0045893	positive regulation of transcription, DNA-templated	1.65E-13
lightgreen	120	GO:1903508	positive regulation of nucleic acid-templated transcription	1.65E-13
lightgreen	121	GO:1902680	positive regulation of RNA biosynthetic process	1.78E-13
lightgreen	122	GO:0009891	positive regulation of biosynthetic process	2.31E-13
lightgreen	123	GO:0045944	positive regulation of transcription from RNA polymerase II promoter	5.55E-13
lightgreen	124	GO:0031328	positive regulation of cellular biosynthetic process	6.40E-13
lightgreen	125	GO:0032879	regulation of localization	1.28E-12
lightgreen	127	GO:0010628	positive regulation of gene expression	1.65E-12
lightgreen	129	GO:0006464	cellular protein modification process	2.86E-12
lightgreen	130	GO:0036211	protein modification process	2.86E-12
lightgreen	132	GO:0031325	positive regulation of cellular metabolic process	5.43E-12

lightgreen	136	GO:0048583	regulation of response to stimulus	8.31E-12
lightgreen	138	GO:0010604	positive regulation of macromolecule metabolic process	1.86E-11
lightgreen	143	GO:0009893	positive regulation of metabolic process	4.81E-11
lightgreen	144	GO:1902578	single-organism localization	5.11E-11
lightgreen	145	GO:0065008	regulation of biological quality	5.16E-11
lightgreen	146	GO:0044710	single-organism metabolic process	5.78E-11
lightgreen	147	GO:0044765	single-organism transport	6.39E-11
lightgreen	148	GO:0006396	RNA processing	8.42E-11
lightgreen	150	GO:0009653	anatomical structure morphogenesis	1.33E-10
lightgreen	154	GO:0071702	organic substance transport	1.48E-10
lightgreen	155	GO:0035556	intracellular signal transduction	1.49E-10
lightyellow	70	GO:0006464	cellular protein modification process	1.77E-15
lightyellow	71	GO:0036211	protein modification process	1.77E-15
lightyellow	81	GO:0051252	regulation of RNA metabolic process	7.98E-15
lightyellow	89	GO:0044281	small molecule metabolic process	6.47E-14
lightyellow	96	GO:2000112	regulation of cellular macromolecule biosynthetic process	5.82E-13
lightyellow	98	GO:0006351	transcription, DNA-templated	1.11E-12
lightyellow	99	GO:0097659	nucleic acid-templated transcription	1.32E-12
lightyellow	100	GO:0032774	RNA biosynthetic process	1.53E-12
lightyellow	102	GO:0010556	regulation of macromolecule biosynthetic process	1.70E-12
lightyellow	105	GO:0031326	regulation of cellular biosynthetic process	2.33E-12
lightyellow	106	GO:0009889	regulation of biosynthetic process	4.07E-12
lightyellow	108	GO:1901564	organonitrogen compound metabolic process	7.17E-12
lightyellow	109	GO:0006355	regulation of transcription, DNA-templated	7.54E-12
lightyellow	112	GO:1903506	regulation of nucleic acid-templated transcription	8.91E-12
lightyellow	115	GO:2001141	regulation of RNA biosynthetic process	9.58E-12
lightyellow	130	GO:0065008	regulation of biological quality	1.42E-10

5. DISCUSSION

The importance of individual lncRNAs for cellular reprogramming events, both artificially induced and topically present in organisms, as well as the maintenance of cellular phenotypes, has been established since 2013 (Pádua Alves *et al.*, 2013; Payer *et al.*, 2013). However, the number of studies on transcriptome-wide analysis of lncRNAs in cellular reprogramming is quite lacking. As of May 2018, using “RNA, Long Noncoding” as a major MeSH topic filter in a literature search, there are 170 articles covering cell differentiation, 116 articles covering epithelial-mesenchymal transition, and 13 articles covering cellular reprogramming, with one article shared under both cell differentiation and EMT. Adding in filters focused on transcriptome profiling or whole exome sequencing brings the total number down to 15. The literature search also reveals that there have been no published studies regarding the noncoding transcriptome of cells undergoing mesenchymal-epithelial transition.

lncRNAs do not work alone in biological pathways. Much like proteins, their functions in the cell are informed by their interactions with other biological molecules and structures, as well as genomic loci (Marchese, Raimondi and Huarte, 2017). Therefore, it is possible, and indeed crucial, to identify potential partners of lncRNA transcripts in networks of biological activity and regulation. For a large number of product-dependent lncRNAs, construction of such networks might be the only available method of annotation, due to the poorly understood connection between lncRNA sequence and structure and lncRNA function (Necsulea *et al.*, 2014; Hezroni *et al.*, 2015).

Recent findings on MET and EMT suggest that regulation of gene expression via noncoding genetic elements, such as enhancers, play a major role in such cellular reprogramming events (Alotaibi *et al.*, 2015; Schnappauf *et al.*, 2016). Given the established associations of enhancer elements and lncRNAs in transcriptional regulation, whether the lncRNA regions acting as enhancers via product-independent transcription (Aune *et al.*, 2017; Fanucchi and Mhlanga, 2017), or binding to enhancer regions on the genome (Soibam, 2017), it is an important avenue of research that requires further exploration.

This is the first known study to analyze the noncoding transcriptome during the MET process with this level of detail. As dedicated studies of MET are a very recent phenomenon, integrating information on lncRNAs to our growing body of knowledge at such an early stage will be beneficial for future research.

The highlight of this study is the application of the transcript abundance estimation and lncRNA identification pipelines to previously produced RNA-seq data. While the reads were originally produced for the purposes of studying the transcription factors active in MET regulation, the use of rRNA depletion as the library selection method, instead of other common methods such as polyA selection, allows reads originating from lncRNAs, which often lack features that are present in protein-coding transcripts, to be included in the library, making the identification of novel lncRNA transcripts possible. Furthermore, previous computational analysis of the data in the context of transcription factors has been consistent with features of the MET regulatory network that had been identified using experimental methods and microarray data by Alotaibi *et al.* (unpublished data).

The findings of this study confirm the suspected presence of lncRNAs during the MET process and their potential partnerships with known transcription factors. Furthermore, they are in line with the multistage progression of cellular reprogramming events, shown by the high number of lncRNAs that display notable expression levels only in a limited number of timepoints, instead of across all samples. This supports the idea that MET is a full-fledged cellular programming event, and cannot be confined to the realm of being defined as “anti-EMT”, as was previously indicated by Kim *et al.* (Kim, Jackson and Davidson, 2017).

The scarcity of lncRNAs expressed during the MET process, their relatively low levels of expression compared to the coding transcriptome, outliers notwithstanding, and the number of lncRNAs that display significant expression spikes or drops in at most two timepoints is consistent with the body of knowledge on lncRNAs.

It is worth noting that the correlation between chromosome size and the number of lncRNA loci on the chromosome, while consistent, is not a universal trait of lncRNAs. Mammalian genomes usually have uniform distribution of lncRNA genes across the genome (Liu *et al.*, 2017), but other organisms, such as plants, can have zero or negative correlation between lncRNA count and chromosome size (Li *et al.*, 2014).

One unexpected finding of the study is the low number of modules enriched for chromatin remodeling. This finding can be explained by the remaining modules operating in the cytoplasm and cytoplasmic organelles, as well as around the membrane. This is consistent with earlier findings about the changes in cytoskeleton, and E-cadherin expression and localization, which is affected by multiple cytoplasmic factors during its downregulation in EMT (Peinado, Portillo and Cano, 2004; Le Bras, Taubenslag and Andl, 2012). Due to the relative minority of such

modules, their importance to the MET process must be carefully examined. Another possibility is that modules not enriched for terms integral to cellular reprogramming include genes that are involved in the basal metabolism of the cell, but nevertheless show varying expression levels as influenced by factors such as the specific stage of the cell cycle the cells are in, or extracellular conditions such as cell confluence and signals received from surrounding cells. Genes showing zero or near-zero variance across all samples would not be included in the WGCNA analysis, due to the inability of the algorithm to calculate their correlations with other genes, thus the regulation of housekeeping genes would already be excluded from the modules.

In addition, the findings of this process are limited to a single inducer of MET in only one cell model. Previous EMT studies, such as the 2016 study by Liao *et al.* have shown that the expression of lncRNAs during a single process, even in the same organism, can be highly variable, depending on the inducing signal used, such as TGF β induction versus Snail overexpression, or the epithelial cell type of origin, such as two different non-tumorigenic human mammary epithelial cell lines, HMLE and MCF10A (Liao *et al.*, 2017). Follow-up studies need to be performed on different MET transcriptomes to hone in on the core lncRNA regulators and enhancers of the universal MET network.

It bears repeating that the findings of this study are computational, and have not been confirmed with experimental approaches, such as PCR, as of yet. While the findings have a high probability of being accurate, the presence of the transcripts in the transcriptomes of any timepoint samples suggested by the previous RNA-seq analyses have to be confirmed empirically. Possible filters to apply when selecting any candidates would include the lncRNA transcript being expressed in multiple subsequent timepoints, to ensure that highly transiently expressed lncRNAs are not missed; high intramodular connectivity, to increase the likelihood of any functional modifications to transcript structure or expression level would significantly affect the outcome of the MET process; and multiple exon transcripts with splice sites detected by HISAT2 and StringTie, to minimize the likelihood of reads misaligning to a region in the genome rather than accurately reflecting the biological conditions of the transcriptome.

Once experimental validation studies are performed, future directions in the field will likely require confirmation of the partners of discovered lncRNAs, as well as their method of function, such as competing endogenous RNAs being analyzed for the miRNAs they act as sponges for.

6.CONCLUSION AND SUGGESTIONS

- A total of 593 previously unannotated lncRNA genes were found to be expressed during MET.
- Of these lncRNAs, 70 were found to be differentially expressed during MET.
- Among the putative lncRNAs, 116 were found to be expressed in a timepoint specific manner in EMT and MET, with two of these 116 genes having zero expression in other timepoints.
- Gene co-expression network of 20709 transcripts was constructed to identify gene modules and gene expression correlations. Thirty modules were constructed as a result. Putative lncRNA content of these modules was identified, and the previously annotated module members were used for GO term enrichment to assign biological context to the putative lncRNAs. Four modules identified were discovered to have a putative lncRNA as the gene with highest degree of intramodular connectivity, defining them as hub genes of great import in the modules. Among the genes identified as putative lncRNAs, 24 were identified to have a KME of > 0.8 , marking them as potential hub genes in the MET process as a whole.
- Modules with enriched GO terms of importance to MET and other cellular reprogramming, such as chromatin remodeling, were identified.
- It is apparent that lncRNA involvement in the MET process is significant, and given the high number of previously unannotated lncRNA transcripts expressed during MET, further research on these lncRNAs is necessary.
- Putative lncRNAs showing high correlation with previously identified protein-coding genes involved in MET regulation, such as *Cebpa* or *Nfya1*, were identified, but not yet examined in detail. Information on their genomic or transcriptomic context, such as distance to and identity of neighboring genes, or expression patterns during the MET process, need to be analyzed further.
- In future research, a number of putative lncRNAs of highest possible involvement in MET, as indicated by gene connectivity and module GO term enrichment, must be selected, in order to validate their transcription during MET using further experimental methods, as well as identify their direct biological partners or genomic binding sites.

7. REFERENCES

- Aigner, K. *et al.* (2007) 'The transcription factor ZEB1 (δ EF1) promotes tumour cell dedifferentiation by repressing master regulators of epithelial polarity', *Oncogene*, 26(49), pp. 6979–6988. doi: 10.1038/sj.onc.1210508.
- Alotaibi, H. *et al.* (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), pp. 731–742. doi: 10.1016/j.bbagr.2015.01.005.
- An introduction to Next-Generation Sequencing Technology* (no date). Available at: www.illumina.com/technology/next-generation-sequencing.html (Accessed: 29 May 2018).
- Andrews, S. (2018) *Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data*. Available at: <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/> (Accessed: 5 June 2018).
- Aune, T. M. *et al.* (2017) 'Expression of long non-coding RNAs in autoimmunity and linkage to enhancer function and autoimmune disease risk genetic variants', *Journal of Autoimmunity*, 81, pp. 99–109. doi: 10.1016/j.jaut.2017.03.014.
- Bateman, A. *et al.* (2017) 'UniProt: The universal protein knowledgebase', *Nucleic Acids Research*. Oxford University Press, 45(D1), pp. D158–D169. doi: 10.1093/nar/gkw1099.
- Beckedorff, F. C. *et al.* (2013) 'The Intronic Long Noncoding RNA ANRASSF1 Recruits PRC2 to the RASSF1A Promoter, Reducing the Expression of RASSF1A and Increasing Cell Proliferation', *PLoS Genetics*. Edited by J. T. Lee, 9(8), p. e1003705. doi: 10.1371/journal.pgen.1003705.
- Bonasio, R. and Shiekhattar, R. (2014) 'Regulation of transcription by long noncoding RNAs.', *Annual review of genetics*. NIH Public Access, 48, pp. 433–55. doi: 10.1146/annurev-genet-120213-092323.
- Le Bras, G. F., Taubenslag, K. J. and Andl, C. D. (2012) 'The regulation of cell-cell adhesion during epithelial-mesenchymal transition, motility and tumor progression.', *Cell adhesion & migration*. Taylor & Francis, 6(4), pp. 365–73. doi: 10.4161/cam.21326.
- Cabili, M. N. *et al.* (2011) 'Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses', *Genes & Development*, 25(18), pp. 1915–1927. doi: 10.1101/gad.17446611.
- Cahais, V. *et al.* (2012) 'Reference-free transcriptome assembly in non-model animals from

- next-generation sequencing data', *Molecular Ecology Resources*, 12(5), pp. 834–845. doi: 10.1111/j.1755-0998.2012.03148.x.
- Cech, T. R. and Steitz, J. A. (2014) 'The noncoding RNA revolution-trashing old rules to forge new ones.', *Cell*. Elsevier, 157(1), pp. 77–94. doi: 10.1016/j.cell.2014.03.008.
- Cervantes-Arias, A., Pang, L. Y. and Argyle, D. J. (2013) 'Epithelial-mesenchymal transition as a fundamental mechanism underlying the cancer phenotype', *Veterinary and Comparative Oncology*. Wiley/Blackwell (10.1111), pp. 169–184. doi: 10.1111/j.1476-5829.2011.00313.x.
- Djebali, S. *et al.* (2012) 'Landscape of transcription in human cells', *Nature*. Nature Publishing Group, 489(7414), pp. 101–108. doi: 10.1038/nature11233.
- Duval, M. and 1844-1907 (1889) 'Atlas d'embryologie'. G. Masson. Available at: <http://agris.fao.org/agris-search/search.do?recordID=US201300442359> (Accessed: 17 May 2018).
- Fanucchi, S. and Mhlanga, M. M. (2017) 'Enhancer-Derived lncRNAs Regulate Genome Architecture: Fact or Fiction?', *Trends in Genetics*. Elsevier Current Trends, 33(6), pp. 375–377. doi: 10.1016/J.TIG.2017.03.004.
- da Fonseca, R. R. *et al.* (2016) 'Next-generation biology: Sequencing and data analysis approaches for non-model organisms', *Marine Genomics*. The Authors, 30, pp. 3–13. doi: 10.1016/j.margen.2016.04.012.
- Frankfurt, U. (1996) 'Cell Growth and Differentiation From the Perspective of Dynamical Organization of Cellular and Subcellular Processes', *Science*, 64(1), pp. 55–79.
- Frazeo, A. C. *et al.* (2014) 'Flexible analysis of transcriptome assemblies with Ballgown', *bioRxiv*, p. 003665. doi: 10.1101/003665.
- Frith, M. C. *et al.* (2006) 'The abundance of short proteins in the mammalian proteome', *PLoS Genetics*, 2(4), pp. 515–528. doi: 10.1371/journal.pgen.0020052.
- Garber, M. *et al.* (2011) 'Computational methods for transcriptome annotation and quantification using RNA-seq', *Nature Methods*. Nature Publishing Group, pp. 469–477. doi: 10.1038/nmeth.1613.
- Gilbert, R., Vickaryous, M. and Vitoria-Petit, A. (2016) 'Signalling by Transforming Growth Factor Beta Isoforms in Wound Healing and Tissue Regeneration', *Journal of Developmental Biology*, 4(2), p. 21. doi: 10.3390/jdb4020021.
- Gregoire, J. M. *et al.* (2016) 'Identification of epigenetic factors regulating the mesenchyme to epithelium transition by RNA interference screening in breast cancer cells', *BMC Cancer*. BMC

- Cancer, 16(1), pp. 1–11. doi: 10.1186/s12885-016-2683-5.
- Gupta, R. A. *et al.* (2010) ‘Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis’, *Nature*, 464(7291), pp. 1071–1076. doi: 10.1038/nature08975.
- Haas, B. and Papanicolaou, A. (no date) *TransDecoder*. Available at: <https://github.com/TransDecoder/TransDecoder/wiki>.
- Hader, C., Marlier, A. and Cantley, L. (2010) ‘Mesenchymal-epithelial transition in epithelial response to injury: The role of Foxc2’, *Oncogene*. Nature Publishing Group, 29(7), pp. 1031–1040. doi: 10.1038/onc.2009.397.
- Han, X. *et al.* (2012) ‘Silencing SOX2 induced mesenchymal-epithelial transition and its expression predicts liver and lymph node metastasis of CRC patients’, *PLoS ONE*, 7(8). doi: 10.1371/journal.pone.0041335.
- Hannon, G. (2018) *FASTX-Toolkit*. Available at: http://hannonlab.cshl.edu/fastx_toolkit/ (Accessed: 5 June 2018).
- He, C. *et al.* (2016) ‘Systematic Characterization of Long Noncoding RNAs Reveals the Contrasting Coordination of Cis- and Trans-Molecular Regulation in Human Fetal and Adult Hearts.’, *Circulation. Cardiovascular genetics*. American Heart Association, Inc., 9(2), pp. 110–8. doi: 10.1161/CIRCGENETICS.115.001264.
- Heery, R. *et al.* (2017) ‘Long non-coding RNAs: Key regulators of epithelial-mesenchymal transition, tumour drug resistance and cancer stem cells’, *Cancers*, 9(4), pp. 1–48. doi: 10.3390/cancers9040038.
- Hezroni, H. *et al.* (2015) ‘Principles of Long Noncoding RNA Evolution Derived from Direct Comparison of Transcriptomes in 17 Species’, *Cell Reports*. Cell Press, 11(7), pp. 1110–1122. doi: 10.1016/J.CELREP.2015.04.023.
- Hindley, C. and Philpott, A. (2012) ‘Co-ordination of cell cycle and differentiation in the developing nervous system.’, *The Biochemical journal*. Portland Press Ltd, 444(3), pp. 375–82. doi: 10.1042/BJ20112040.
- Huang, D. W., Sherman, B. T. and Lempicki, R. A. (2009a) ‘Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists.’, *Nucleic acids research*. Oxford University Press, 37(1), pp. 1–13. doi: 10.1093/nar/gkn923.
- Huang, D. W., Sherman, B. T. and Lempicki, R. A. (2009b) ‘Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources’, *Nature Protocols*, 4(1), pp. 44–57. doi: 10.1038/nprot.2008.211.

Human Genome Project Completion: Frequently Asked Questions - National Human Genome Research Institute (NHGRI) (no date). Available at: <https://www.genome.gov/11006943/human-genome-project-completion-frequently-asked-questions/> (Accessed: 23 May 2018).

Iyer, M. K. *et al.* (2015) 'The landscape of long noncoding RNAs in the human transcriptome', *Nature Genetics*, 47(3), pp. 199–208. doi: 10.1038/ng.3192.

De Jong, M. R. W. *et al.* (2018) 'Identification of relevant drugable targets in diffuse large B-cell lymphoma using a genome-wide unbiased CD20 guilt-by association approach', *PLoS ONE*, 13(2), pp. 1–18. doi: 10.1371/journal.pone.0193098.

Kalvari, I. *et al.* (2018) 'Rfam 13.0: Shifting to a genome-centric resource for non-coding RNA families', *Nucleic Acids Research*. Oxford University Press, 46(D1), pp. D335–D342. doi: 10.1093/nar/gkx1038.

Kim, D. *et al.* (2016) 'Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown', *Nature Protocols*. Nature Publishing Group, 11(9), pp. 1650–1667. doi: 10.1038/nprot.2016-095.

Kim, D. *et al.* (2017) 'Epithelial Mesenchymal Transition in Embryonic Development, Tissue Repair and Cancer: A Comprehensive Overview', *Journal of Clinical Medicine*, 7(1), p. 1. doi: 10.3390/jcm7010001.

Kim, D., Langmead, B. and Salzberg, S. L. (2015) 'HISAT: A fast spliced aligner with low memory requirements', *Nature Methods*, 12(4), pp. 357–360. doi: 10.1038/nmeth.3317.

Kim, H. Y., Jackson, T. R. and Davidson, L. A. (2017) 'On the role of mechanics in driving mesenchymal-to-epithelial transitions', *Seminars in Cell & Developmental Biology*. Academic Press, 67, pp. 113–122. doi: 10.1016/J.SEMCDB.2016.05.011.

Kim, K. *et al.* (2015) 'Effect of Next-Generation Exome Sequencing Depth for Discovery of Diagnostic Variants.', *Genomics & informatics*. Korea Genome Organization, 13(2), pp. 31–9. doi: 10.5808/GI.2015.13.2.31.

Kotake, Y. *et al.* (2011) 'Long non-coding RNA ANRIL is required for the PRC2 recruitment to and silencing of p15 INK4B tumor suppressor gene', *Oncogene*, 30(16), pp. 1956–1962. doi: 10.1038/onc.2010.568.

Kryuchkova-Mostacci, N. and Robinson-Rechavi, M. (2017) 'A benchmark of gene expression tissue-specificity metrics', *Briefings in Bioinformatics*, 18(2), pp. 205–214. doi:

10.1093/bib/bbw008.

Lamouille, S., Xu, J. and Derynck, R. (2014) 'Molecular mechanisms of epithelial–mesenchymal transition', *Nature Reviews Molecular Cell Biology*. Nature Publishing Group, 15(3), pp. 178–196. doi: 10.1038/nrm3758.

Langfelder, P. and Horvath, S. (2008) 'WGCNA: An R package for weighted correlation network analysis', *BMC Bioinformatics*, 9. doi: 10.1186/1471-2105-9-559.

Li, B. *et al.* (2011) 'Evidence for mesenchymal-epithelial transition associated with mouse hepatic stem cell differentiation', *PLoS ONE*, 6(2). doi: 10.1371/journal.pone.0017092.

Li, L. *et al.* (2014) 'Genome-wide discovery and characterization of maize long non-coding RNAs', *Genome Biology*. BioMed Central, 15(2), p. R40. doi: 10.1186/gb-2014-15-2-r40.

Liang, S. *et al.* (2017) 'The lncRNA XIST interacts with miR-140 / miR-124 / iASPP axis to promote pancreatic carcinoma growth', 8(69), pp. 113701–113718.

Liao, J. Y. *et al.* (2017) 'Deep sequencing reveals a global reprogramming of lncRNA transcriptome during EMT', *Biochimica et Biophysica Acta - Molecular Cell Research*, 1864(10), pp. 1703–1713. doi: 10.1016/j.bbamcr.2017.06.003.

Liu, X. F. *et al.* (2017) 'An atlas and analysis of bovine skeletal muscle long noncoding RNAs', *Animal Genetics*. Wiley/Blackwell (10.1111), 48(3), pp. 278–286. doi: 10.1111/age.12539.

Liu, Y. *et al.* (2013) 'Evaluating the impact of sequencing depth on transcriptome profiling in human adipose.', *PloS one*. Public Library of Science, 8(6), p. e66883. doi: 10.1371/journal.pone.0066883.

Love, M. I., Huber, W. and Anders, S. (2014) 'Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2', *Genome Biology*. BioMed Central, 15(12), p. 550. doi: 10.1186/s13059-014-0550-8.

Maccarty, W. C. and Caylor, H. D. (1922) 'METAPLASIA IN OVARIAN DERMOIDS AND CYSTADENOMAS: REPORT OF THREE CASES.', *Annals of surgery*. Lippincott, Williams, and Wilkins, 76(2), pp. 238–45. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/17864684> (Accessed: 17 May 2018).

Marchese, F. P., Raimondi, I. and Huarte, M. (2017) 'The multidimensional mechanisms of long noncoding RNA function.', *Genome biology*. BioMed Central, 18(1), p. 206. doi: 10.1186/s13059-017-1348-2.

Marques, A. C. *et al.* (2013) 'Chromatin signatures at transcriptional start sites separate two equally populated yet distinct classes of intergenic long noncoding RNAs', *Genome Biology*.

- BioMed Central, 14(11), p. R131. doi: 10.1186/gb-2013-14-11-r131.
- Mattick, J. S. (2001) 'Non-coding RNAs: the architects of eukaryotic complexity', *EMBO reports*, 2(11), pp. 986–991. Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1084129/pdf/kve230.pdf> (Accessed: 20 May 2018).
- Milligan, M. J. *et al.* (2016) 'Global Intersection of Long Non-Coding RNAs with Processed and Unprocessed Pseudogenes in the Human Genome', *Frontiers in Genetics*, 7, p. 26. doi: 10.3389/fgene.2016.00026.
- Moreno-Bueno, G., Portillo, F. and Cano, A. (2008) 'Transcriptional regulation of cell polarity in EMT and cancer', *Oncogene*. Nature Publishing Group, 27(55), pp. 6958–6969. doi: 10.1038/onc.2008.346.
- Mortazavi, A. *et al.* (2008) 'Mapping and quantifying mammalian transcriptomes by RNA-Seq', *Nature Methods*. Nature Publishing Group, 5(7), pp. 621–628. doi: 10.1038/nmeth.1226.
- Nawrocki, E. P. and Eddy, S. R. (2013) 'Infernal 1.1: 100-fold faster RNA homology searches', *Bioinformatics*, 29(22), pp. 2933–2935. doi: 10.1093/bioinformatics/btt509.
- Necsulea, A. *et al.* (2014) 'The evolution of lncRNA repertoires and expression patterns in tetrapods', *Nature*. Nature Publishing Group, 505(7485), pp. 635–640. doi: 10.1038/nature12943.
- Neff, T. and Armstrong, S. A. (2009) 'Chromatin maps, histone modifications and leukemia', *Leukemia*. Nature Publishing Group, 23(7), pp. 1243–1251. doi: 10.1038/leu.2009.40.
- Ozdamar, B. *et al.* (2005) 'Regulation of the polarity protein Par6 by TGFbeta receptors controls epithelial cell plasticity.', *Science (New York, N.Y.)*. American Association for the Advancement of Science, 307(5715), pp. 1603–9. doi: 10.1126/science.1105718.
- Pachter, L. (2011) 'Models for transcript quantification from RNA-Seq'. doi: 10.1038/nbt.162.
- Pádua Alves, C. *et al.* (2013) 'Brief Report: The lincRNA Hotair Is Required for Epithelial-to-Mesenchymal Transition and Stemness Maintenance of Cancer Cell Lines', *STEM CELLS*. Wiley-Blackwell, 31(12), pp. 2827–2832. doi: 10.1002/stem.1547.
- Pan, C. *et al.* (2017) 'Long Noncoding RNA FAL1 Promotes Cell Proliferation, Invasion and Epithelial-Mesenchymal Transition Through the PTEN/AKT Signaling Axis in Non-Small Cell Lung Cancer.', *Cellular physiology and biochemistry : international journal of experimental cellular physiology, biochemistry, and pharmacology*. Karger Publishers, 43(1), pp. 339–352. doi: 10.1159/000480414.

- Patel, R. K. and Jain, M. (2012) 'NGS QC Toolkit: A Toolkit for Quality Control of Next Generation Sequencing Data', *PLoS ONE*. Edited by Z. Liu. Public Library of Science, 7(2), p. e30619. doi: 10.1371/journal.pone.0030619.
- Payer, B. *et al.* (2013) 'Tsix RNA and the Germline Factor, PRDM14, Link X Reactivation and Stem Cell Reprogramming', *Molecular Cell*. Cell Press, 52(6), pp. 805–818. doi: 10.1016/J.MOLCEL.2013.10.023.
- Peinado, H., Portillo, F. and Cano, A. (2004) 'Transcriptional regulation of cadherins during development and carcinogenesis', *International Journal of Developmental Biology*, 48(5–6), pp. 365–375. doi: 10.1387/ijdb.041794hp.
- Pertea, G. and Kirchner, R. (2016) *The GffCompare Utility*. Available at: <http://ccb.jhu.edu/software/stringtie/gffcompare.shtml> (Accessed: 5 June 2018).
- Pertea, M. *et al.* (2015) 'StringTie enables improved reconstruction of a transcriptome from RNA-seq reads', *Nature Biotechnology*, 33(3), pp. 290–295. doi: 10.1038/nbt.3122.
- Puisieux, A., Brabletz, T. and Caramel, J. (2014) 'Oncogenic roles of EMT-inducing transcription factors', *Nature Cell Biology*. Nature Publishing Group, 16(6), pp. 488–494. doi: 10.1038/ncb2976.
- Quek, X. C. *et al.* (2015) 'lncRNADB v2.0: expanding the reference database for functional long noncoding RNAs', *Nucleic Acids Research*. Oxford University Press, 43(D1), pp. D168–D173. doi: 10.1093/nar/gku988.
- Raju, H. B., Tsinoremas, N. F. and Capobianco, E. (2016) 'Emerging Putative Associations between Non-Coding RNAs and Protein-Coding Genes in Neuropathic Pain: Added Value from Reusing Microarray Data', *Frontiers in Neurology*, 7, p. 168. doi: 10.3389/fneur.2016.00168.
- Reik, W., Dean, W. and Walter, J. (2001) 'Epigenetic reprogramming in mammalian development.', *Science (New York, N.Y.)*. American Association for the Advancement of Science, 293(5532), pp. 1089–93. doi: 10.1126/science.1063443.
- Robertson, G. *et al.* (2010) 'De novo assembly and analysis of RNA-seq data', *Nature Methods*. Nature Publishing Group, 7(11), pp. 909–912. doi: 10.1038/nmeth.1517.
- Rossant, J. and Tam, P. P. L. (2009) 'Blastocyst lineage formation, early embryonic asymmetries and axis patterning in the mouse', *Development*, 136(5), pp. 701–713. doi: 10.1242/dev.017178.
- Rutenberg-Schoenberg, M., Sexton, A. N. and Simon, M. D. (2016) 'The Properties of Long

Noncoding RNAs That Regulate Chromatin', *Annual Review of Genomics and Human Genetics*, 17(1), pp. 69–94. doi: 10.1146/annurev-genom-090314-024939.

Sanger, F., Nicklen, S. and Coulson, A. R. (1977) 'DNA sequencing with chain-terminating inhibitors', *Proceedings of the National Academy of Sciences*, 74(12), pp. 5463–5467. doi: 10.1073/pnas.74.12.5463.

Schnappauf, O. *et al.* (2016) 'Enhancer decommissioning by Snail1-induced competitive displacement of TCF7L2 and down-regulation of transcriptional activators results in EPHB2 silencing', *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, 1859(11), pp. 1353–1367. doi: 10.1016/j.bbagr.2016.08.002.

Sequencing Platforms | Compare NGS platforms (benchtop, production-scale) (no date). Available at: <https://www.illumina.com/systems/sequencing-platforms.html> (Accessed: 14 May 2018).

Shen, L. *et al.* (2013) 'Genome-wide analysis reveals TET- and TDG-dependent 5-methylcytosine oxidation dynamics', *Cell*. Elsevier Inc., 153(3), pp. 692–706. doi: 10.1016/j.cell.2013.04.002.

Signal, B., Gloss, B. S. and Dinger, M. E. (2016) 'Computational Approaches for Functional Prediction and Characterisation of Long Noncoding RNAs', *Trends in Genetics*. Elsevier Ltd, 32(10), pp. 620–637. doi: 10.1016/j.tig.2016.08.004.

Slack, J. M. . and Tosh, D. (2001) 'Transdifferentiation and metaplasia — switching cell types', *Current Opinion in Genetics & Development*. Elsevier Current Trends, 11(5), pp. 581–586. doi: 10.1016/S0959-437X(00)00236-7.

Soibam, B. (2017) 'Super-lncRNAs: identification of lncRNAs that target super-enhancers via RNA:DNA:DNA triplex formation', *RNA*, 23(11), pp. 1729–1742. doi: 10.1261/rna.061317.117.

Specification Sheet: Sequencing (no date). Available at: <https://www.illumina.com/content/dam/illumina-marketing/documents/products/datasheets/datasheet-hiseq-x-ten.pdf> (Accessed: 23 May 2018).

Stuart, J. M. *et al.* (2003) 'A gene-coexpression network for global discovery of conserved genetic modules', *Science*, 302(5643), pp. 249–255. doi: 10.1126/science.1087447.

Sul, J.-Y. *et al.* (2009) 'Transcriptome transfer produces a predictable cellular phenotype.', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 106(18), pp. 7624–9. doi: 10.1073/pnas.0902161106.

- Sun, J. *et al.* (2013) 'TCC: An R package for comparing tag count data with robust normalization strategies', *BMC Bioinformatics*. BMC Bioinformatics, 14(1), p. 1. doi: 10.1186/1471-2105-14-219.
- Tan, E. J., Olsson, A. K. and Moustakas, A. (2015) 'Reprogramming during epithelial to mesenchymal transition under the control of TGFbeta', *Cell Adh Migr*, 9(3), pp. 233–246. doi: 10.4161/19336918.2014.983794.
- Tay, Y. *et al.* (2011) 'Coding-independent regulation of the tumor suppressor PTEN by competing endogenous mRNAs.', *Cell*. Elsevier, 147(2), pp. 344–57. doi: 10.1016/j.cell.2011.09.029.
- Tong, L. *et al.* (2017) 'MTDH promotes glioma invasion through regulating miR-130b-ceRNAs', *Oncotarget*. Impact Journals, 8(11), pp. 17738–17749. doi: 10.18632/oncotarget.14717.
- Trapnell, C. *et al.* (2010) 'Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation', *Nature Biotechnology*. Nature Publishing Group, 28(5), pp. 511–515. doi: 10.1038/nbt.1621.
- Wang, K. C. and Chang, H. Y. (2011) 'Molecular mechanisms of long noncoding RNAs.', *Molecular cell*. Elsevier, 43(6), pp. 904–14. doi: 10.1016/j.molcel.2011.08.018.
- Washietl, S., Kellis, M. and Garber, M. (2014) 'Evolutionary dynamics and tissue specificity of human long noncoding RNAs in six mammals.', *Genome research*. Cold Spring Harbor Laboratory Press, 24(4), pp. 616–28. doi: 10.1101/gr.165035.113.
- Weakley, S. M. *et al.* (2011) 'Expression and function of a large non-coding RNA gene XIST in human cancer', *World Journal of Surgery*, 35(8), pp. 1751–1756. doi: 10.1007/s00268-010-0951-0.
- Weber, M. *et al.* (2005) 'Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells', *Nature Genetics*. Nature Publishing Group, 37(8), pp. 853–862. doi: 10.1038/ng1598.
- Xu, X. *et al.* (2016) 'Transcriptional modules related to hepatocellular carcinoma survival: coexpression network analysis', *Frontiers of Medicine*, 10(2), pp. 183–190. doi: 10.1007/s11684-016-0440-4.
- Zhang, B. and Horvath, S. (2005) 'A General Framework for Weighted Gene Co-Expression Network Analysis', *Statistical Applications in Genetics and Molecular Biology*. De Gruyter, 4(1). doi: 10.2202/1544-6115.1128.

Zhang, K. *et al.* (2014) 'The ways of action of long non-coding RNAs in cytoplasm and nucleus', *Gene*. Elsevier, 547(1), pp. 1–9. doi: 10.1016/J.GENE.2014.06.043.

Zhu, Q.-H. *et al.* (2012) 'Molecular Functions of Long Non-Coding RNAs in Plants', *Genes*, 3(4), pp. 176–190. doi: 10.3390/genes3010176.



APPENDIX 1.

6/7/2018

RightsLink Printable License

ELSEVIER LICENSE TERMS AND CONDITIONS

Jun 07, 2018

This Agreement between Mr. Doğa Eskier ("You") and Elsevier ("Elsevier") consists of your license details and the terms and conditions provided by Elsevier and Copyright Clearance Center.

License Number	4363511186842
License date	Jun 07, 2018
Licensed Content Publisher	Elsevier
Licensed Content Publication	Trends in Genetics
Licensed Content Title	Computational Approaches for Functional Prediction and Characterisation of Long Noncoding RNAs
Licensed Content Author	Bethany Signal,Brian S. Gloss,Marcel E. Dinger
Licensed Content Date	Oct 1, 2016
Licensed Content Volume	32
Licensed Content Issue	10
Licensed Content Pages	18
Start Page	620
End Page	637
Type of Use	reuse in a thesis/dissertation
Portion	figures/tables/illustrations
Number of figures/tables/illustrations	1
Format	both print and electronic
Are you the author of this Elsevier article?	No
Will you be translating?	No
Original figure numbers	1
Title of your thesis/dissertation	IDENTIFICATION AND ANNOTATION OF PUTATIVE LONG NON-CODING RNAS INVOLVED IN MESENCHYMAL-EPITHELIAL TRANSITION
Expected completion date	Jul 2018
Estimated size (number of pages)	70
Requestor Location	Mr. Doğa Eskier 1671 sokak No: 153/1 D:7 Karşıyaka, İzmir 35530 Turkey Attn: Mr. Doğa Eskier
Publisher Tax ID	GB 494 6272 12
Total	0.00 EUR
Terms and Conditions	

INTRODUCTION

1. The publisher for this copyrighted material is Elsevier. By clicking "accept" in connection with completing this licensing transaction, you agree that the following terms and conditions

<https://s100.copyright.com/CustomAdmin/PLF.jsp?ref=8659a0d9-d433-40b6-b128-d3c76beca72a>

1/5

APPENDIX 2.

6/7/2018

RightsLink Printable License

SPRINGER NATURE LICENSE TERMS AND CONDITIONS

Jun 07, 2018

This Agreement between Mr. Doğa Eskier ("You") and Springer Nature ("Springer Nature") consists of your license details and the terms and conditions provided by Springer Nature and Copyright Clearance Center.

License Number	4363520183733
License date	Jun 07, 2018
Licensed Content Publisher	Springer Nature
Licensed Content Publication	Nature Methods
Licensed Content Title	Computational methods for transcriptome annotation and quantification using RNA-seq
Licensed Content Author	Manuel Garber, Manfred G Grabherr, Mitchell Guttman, Cole Trapnell
Licensed Content Date	May 27, 2011
Licensed Content Volume	8
Licensed Content Issue	6
Type of Use	Thesis/Dissertation
Requestor type	academic/university or research institute
Format	print and electronic
Portion	figures/tables/illustrations
Number of figures/tables/illustrations	1
High-res required	no
Will you be translating?	no
Circulation/distribution	<501
Author of this Springer Nature content	no
Title	IDENTIFICATION AND ANNOTATION OF PUTATIVE LONG NON-CODING RNAs INVOLVED IN MESENCHYMAL-EPITHELIAL TRANSITION
Instructor name	Gökhan Karakölah
Institution name	İzmir International Biomedicine and Genome Institute
Expected presentation date	Jul 2018
Portions	Figure 2a on page 472
Requestor Location	Mr. Doğa Eskier 1671 sokak No: 153/1 D:7 Karşıyaka, İzmir 35530 Turkey Attn: Mr. Doğa Eskier
Billing Type	Invoice
Billing Address	Mr. Doğa Eskier 1671 sokak No: 153/1 D:7

<https://s100.copyright.com/CustomAdmin/PLF.jsp?ref=8b7a5e22-d4aa-40c1-aa97-933028eb4007>

1/3