

T.C.

DOKUZ EYLÜL UNIVERSITY

İZMİR INTERNATIONAL BIOMEDICINE AND GENOME INSTITUTE

**ELF3 REGULATED TRANSCRIPTOME
DYNAMICS DURING MESENCHYMAL TO
EPITHELIAL TRANSITION**

HAMDİYE UZUNER

MOLECULAR BIOLOGY AND GENETICS

MASTER OF SCIENCE THESIS

İZMİR-2020

THESIS CODE: 2018850006

T.C.

DOKUZ EYLÜL UNIVERSITY

IZMIR INTERNATIONAL BIOMEDICINE AND GENOME INSTITUTE

**ELF3 REGULATED TRANSCRIPTOME
DYNAMICS DURING MESENCHYMAL TO
EPITHELIAL TRANSITION**

MOLECULAR BIOLOGY AND GENETICS

MASTER OF SCIENCE THESIS

HAMDİYE UZUNER

Supervisor: Assoc. Prof. Dr. Gökhan Karakülah

Co-supervisor: Assist. Prof. Dr. Hani Alotaibi

THESIS CODE: 2018850006

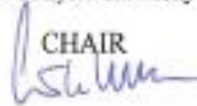
Dokuz Eylül University İzmir International Biomedicine and Genome Institute
Department of Genomics and Molecular Biotechnology, Molecular Biology
and Genetics graduate program Master of Science student Hamdiye Uzuner
has successfully completed her Master of Science thesis titled

**'ELF3 REGULATED TRANSCRIPTOME DYNAMICS DURING
MESENCHYMAL TO EPITHELIAL TRANSITION'** on the date of
26/08/2020.

Assoc. Prof. Gökhan Karakullah

Dokuz Eylül University

CHAIR



Asst. Prof. Cihangir Yandım

İzmir University of Economics

MEMBER



Asst. Prof. Yavuz Oktay

Dokuz Eylül University

MEMBER



Asst. Prof. Athanasia Pavlopoulou

Dokuz Eylül University

SUBSTITUTE MEMBER

Assoc. Prof. Savaş İzzetoğlu

Ege University

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 Hamdiye Uzuner
**‘MEZENKİMEL EPİTEL GEÇİŞ SIRASINDA ELF3 İLE REGÜLE
EDİLEN TRANSKRİPTOM DİNAMİKLERİ’** konulu Yüksek Lisans
tezini 26/08/2020 tarihinde başarılı olarak tamamlamıştır.

Doç. Dr. Gökhan Karakülah

Dokuz Eylül Üniversitesi

BAŞKAN


Dr. Öğr. Üye. Cihangir Yandım

İzmir Ekonomi Üniversitesi

ÜYE


Dr. Öğr. Üye. Yavuz Oktay

Dokuz Eylül Üniversitesi

ÜYE


Dr. Öğr. Üye. Athanasia Pavlopoulou

Dokuz Eylül Üniversitesi

YEDEK ÜYE

Doç. Dr. Savaş İzzetoğlu

Ege Üniversitesi

YEDEK ÜYE

TABLE OF CONTENTS

TABLE OF CONTENTS	i
INDEX OF TABLES.....	iv
INDEX OF FIGURES	v
LIST OF ABBREVIATIONS.....	viii
ACKNOWLEDGEMENTS	ix
ÖZET	10
ABSTRACT	11
INTRODUCTION AND AIM	12
1.1. Statement and Importance of the Problem	12
1.2. Aim of the Study	12
1.3. Hypothesis of the Study.....	12
2. GENERAL INFORMATION.....	12
2.1. Epithelial to Mesenchymal Transition	12
<i>2.1.1. Epithelial cells.....</i>	<i>12</i>
<i>2.1.2. Mesenchymal cells.....</i>	<i>13</i>
<i>2.1.3. Epithelial to mesenchymal transition.....</i>	<i>14</i>
<i>2.1.4. Molecular regulation: Epithelial to mesenchymal transition inducers</i>	<i>15</i>
2.2. Mesenchymal to Epithelial Transition.....	16
<i>2.2.1. Molecular regulation of mesenchymal to epithelial transition</i>	<i>17</i>
<i>2.2.2. Role of E1f3 in transcriptional regulation of mesenchymal to epithelial transition.....</i>	<i>17</i>

2.3. Next Generation Sequencing & Transcriptome Research	18
2.3.1. Next generation sequencing technologies & applications	18
2.3.2. Transcriptome analysis with RNA-sequencing	24
2.3.3. RNA sequencing and data analysis	28
3. MATERIALS AND METHODOLOGY	33
3.1. Type of Study	33
3.2. Time and Place of the Study	33
3.3. Materials of Study.....	33
3.4. Variables of Study.....	34
3.5. Tools for Data Collection.....	34
3.6. Study Plan and Calendar	34
3.7. Data Evaluation.....	35
3.7.1. RNA-seq raw data quality control.....	35
3.7.2. Alignment of reads to reference genome.....	35
3.7.3. Quantification of protein-coding and long non-coding RNA coding genes.....	36
3.7.4. Quantification of repetitive elements	36
3.7.5. Differential expression analysis.....	36
3.7.6. Enrichment analyses and visualizations	37
3.8. Limitations of the study.....	37
3.9. Ethics Committee Approval.....	37
4. RESULTS.....	37
4.1. Preprocessing and alignment of sequencing reads	37
4.2. Overview of gene expression profiling with siElf3 knockdown	39

4.3. Differential expression analysis between Scr and siElf3 samples.....	41
4.4. Overrepresentation test of Gene Ontology terms between conditions.....	48
4.5. KEGG pathway enrichment analysis.....	59
5. DISCUSSION.....	66
6. CONCLUSION AND FUTURE ASPECTS.....	69
7. REFERENCES	70
APPENDIX 1	82
APPENDIX II	83

INDEX OF TABLES

Table 1. Alignment statistics of Scr and siElf3 samples..... 38

Table 2. KEGG pathway enrichment analysis result (GeneRatio: number of input genes that are enriched in the category / total number of input genes that are enriched, BgRatio: total number of genes in that pathway/total number of genes in all pathways)..... 59



INDEX OF FIGURES

Figure 1. Sanger sequencing general workflow	19
Figure 2. Overview of Illumina NGS technology (obtained from An Introduction to Next Generation Sequencing Technology, 01/07/20)	21
Figure 3. General schema of a eukaryotic transcriptome components	26
Figure 4. RNA-seq library preparation overview. Figure adapted from RNA Sequencing And Analysis (Kukurba & Montgomery, 2015) (See Appendix 1)	28
Figure 5. General schema of a typical RNA-seq data analysis experiment. Figure adapted from RNA Sequencing and Analysis (Kukurba & Montgomery, 2015) (see Appendix 1)	31
Figure 6. EMT induction with TGF β 3 treatment which is then followed by MET induction after removal of TGF β 3.	34
Figure 7. Correlation heatmap of expressions belonging to two Scr and to siElf3 samples.	39
Figure 8. Boxplot showing protein coding, lncRNA and repeat elements content distribution... ..	40
Figure 9. Volcano plot showing the distribution of differentially expressed protein-coding genes (NS: Not Significant, log2FC: log2 Fold Change, FDR: False Discovery Rate).....	41
Figure 10. MA plot showing the distribution of differentially expressed protein-coding genes (log2 fold change threshold $\pm$ 0.6, FDR threshold < 0.05; NS: Not Significant, FPKM: Fragments per kilobase of transcript per million mapped reads).	42
Figure 11. Volcano plot showing the distribution of differentially expressed lncRNAs (NS: Not Significant, log2FC: log2 Fold Change, FDR: False Discovery Rate).	44

Figure 12. MA plot showing the distribution of differentially expressed lncRNAs (log2 fold change threshold ± 0.6 , FDR threshold < 0.05 ; NS: Not Significant, FPKM: Fragments per kilobase of transcript per million mapped reads).....	45
Figure 13. Volcano plot showing the distribution of differentially expressed repeat elements. (NS: Not Significant). (NS: Not Significant, log2FC: log2 Fold Change, FDR: False Discovery Rate).	46
Figure 14. MA plot showing the distribution of differentially expressed repeat with log2 fold change and mean expression levels in all samples (log2 fold change threshold ± 0.6 , FDR threshold < 0.05 , NS: Not Significant).....	47
Figure 15. Barplot for top 20 significant simplified GO terms that are enriched for Biological Process. X-axis refers to the number of input genes that are enriched in each category.	48
Figure 16. Dotplot for top 20 significant simplified GO terms that are enriched for Biological Process. (GeneRatio=number of input genes that are enriched in the category / total number of input genes that are enriched).....	49
Figure 17. Barplot for top 20 significant simplified GO terms that are enriched for Molecular Function. X-axis refers to the number of input genes that are enriched in each category.	50
Figure 18. Dotplot for top 20 significant simplified GO terms that are enriched for Molecular Function. (GeneRatio= number of input genes that are enriched in the category / total number of input genes that are enriched).....	51
Figure 19. Barplot for top 20 significant simplified GO terms that are enriched for Cellular Component. X-axis refers to the number of input genes that are enriched in each category.	52
Figure 20. Dotplot for top 20 significant simplified GO terms that are enriched for Cellular Component. (GeneRatio= number of input genes that are enriched in the category / total number of input genes that are enriched).....	53

Figure 21. Barplot showing GO enrichment of 1354 up regulated genes in Biological Process (P-adjust < 0.05).	54
Figure 22. Barplot showing GO enrichment of 1701 down regulated genes in Biological Process (P-adjust < 0.05).	55
Figure 23. Network plot of top three enriched terms in biological process.	57
Figure 24. Network plot of top three enriched terms in cellular component.....	58
Figure 25. DNA replication pathway highlighted with enriched genes; color scale represents logFC distribution.....	62
Figure 26. Cell cycle pathway highlighted with enriched genes; color scale represents logFC distribution.....	63
Figure 27. Mismatch repair pathway highlighted with enriched genes; color scale represents logFC distribution.....	64
Figure 28. Lysosome pathway highlighted with enriched genes; color scale represents logFC distribution.....	65

LIST OF ABBREVIATIONS

NMuMg: Normal Murine Mammary Gland

EMT: Epithelial to Mesenchymal Transition

MET: Mesenchymal to Epithelial Transition

ECM: Extracellular matrix

cDNA: Complementary Deoxyribonucleic acid

LINE: Long interspersed nuclear elements

SINE: Short interspersed nuclear elements

LTR: Long terminal repeat

ACKNOWLEDGEMENTS

I wish to express my deepest gratitude to the following individuals and organizations for their contributions to the completion of this thesis:

Associate Professor Gökhan Karakulah, for his unwavering guidance and invaluable contribution throughout whole course of this thesis as well as his profound belief in my abilities and his relentless support on me,

Assistant Professor Hani Alotaibi, for his precious guidance and support as well as his ingenious suggestions and crediting my work,

TÜBİTAK, for funding this study (Project #: 117Z223),

Ahmet Bursalı, Nazmiye Arslan, Doğa Eskier, Alırıza Arıbaş and other IBG Bioinformatics Platform members, past and present, for being always supportive,

Burcu Şengez, for producing RNA-seq samples of this thesis, and her helpful advices,

My deepest gratitude to my parents Selma and Hamdi Uzuner and my brother Abdullah Uzuner for their endless belief and enduring support on me,

My dearest friends, for never letting me down and their great support and motivation,

And all members of Izmir Biomedicine and Genome Center and Izmir International Genome Institute, past and present, for giving me the opportunity to carry out my studies.

MEZENKİMEL EPİTEL GEÇİŞ SIRASINDA ELF3 İLE REGÜLE EDİLEN TRANSKRİPTOM DİNAMİKLERİ

Hamdiye Uzuner, İzmir Uluslararası Biyotıp ve Genom Merkezi, Dokuz Eylül Üniversitesi Sağlık
Kampüsü, Balçova 35340 – İzmir / TÜRKİYE

ÖZET

Mezenkimal epitel geçiş (MET), içerisinde gelişimsel süreç, kanser gelişimi ve doku rejenerasyonunu barındıran ve birçok fizyolojik süreçte gözlenen anahtar bir biyolojik olaydır. MET regülasyonu, genetik ve epigenetik faktörlere bağlı olarak birçok transkripsiyonel olay sayesinde gerçekleşir. MET, epitel mezenkimal geçiş (EMT) regülasyonunda rol alan faktörlerin ortadan kaldırılması ile tetiklenen bir süreç gibi düşünülse de son yayınlanan çalışmalar MET regülasyonunun bundan ibaret olmadığını göstermektedir. MET tam olarak anlaşılmamış olmasına rağmen, MET regülasyonunda rol alan önemli genler olduğu tahmin edilmektedir. Bunlar arasında, *Elf3*'ün *Grhl3*'ün promotor bölgesi ile ilişki kurarak MET regülasyonunda gerekli olduğu düşünülmektedir. *Grhl3* ise *Cdh1* geni üzerinde bulunan bir cis düzenleyici unsur ile ilişki kurarak E-cadherin lokalizasyonunu etkilediği tahmin edilmektedir. Bu çalışmada, *Elf3* yokluğunda içerisinde *Grhl3* gibi önemli faktörleri barındırabilecek olan kodlayan, kodlamayan ve tekrarlayan (repetitive) transkriptom dinamiklerinde birtakım değişiklikler olacağı öngörülmüştür. Sonuçta, diferansiyel ifade analizi birçok protein kodlayan genin, uzun protein kodlamayan RNA'ların ve tekrar elementlerinin ifade değişimlerini ortaya çıkarmıştır. Diferansiyel olarak ifade edilen genler ile Gen Ontoloji ve KEGG yolak zenginleştirme analizleri yapılmıştır. Son olarak, MET'de *Elf3* yokluğunun, hücre döngüsünde, özellikle DNA replikasyonu, mitotik çekirdek bölünmesi ve kromozomal ayrılımda bozukluklara yol açabileceği gösterilmiştir.

Anahtar kelimeler: Mezenkimal epitel geçiş, MET, Epitel mezenkimal geçiş, EMT, *Elf3*, transkriptom analizi, biyoinformatik

ELF3 REGULATED TRANSCRIPTOME DYNAMICS IN MESENCHYMAL TO EPITHELIAL TRANSITION

Hamdiye Uzuner, Izmir International Biomedicine and Genome Institute, Dokuz Eylül University
Health Campus, Balçova 35340 – Izmir / TURKEY

ABSTRACT

Mesenchymal to epithelial transition (MET) is a key event that is observed in many physiological processes involving development, cancer progression and tissue regeneration. MET regulation is achieved by many transcriptional events that are dependent on genetic and epigenetic factors. According to recent studies, MET regulation is not exactly the opposite of epithelial to mesenchymal transition (EMT) by virtue of removal of the factors involved in the regulation of EMT. Despite the fact that MET is not very well understood, there are some genes that are thought to take important roles in MET regulation. *Elf3* is thought to be essential in the regulation of MET by interacting with the promoter of *Grhl3*. Moreover, *Grhl3* has been reported to act on the cis regulatory element that is found on *Cdh1* and it is thought to be associated with E-cadherin localization. In this study, absence of *Elf3* is hypothesized to cause alterations in transcriptome dynamics, in terms of coding, noncoding and repetitive transcriptome that presumably involve important factors such as *Grhl3*. Therefore, differential expression analysis resulted in alterations in the expressions of various of protein coding genes, long noncoding RNAs and repeat elements. Consequently, enrichment analyses involving Gene Ontology and KEGG pathways for differentially expressed genes showed that the absence of *Elf3* during MET may lead to disruptions in cell cycle, more particularly DNA replication, mitotic nuclear division and chromosomal segregation.

Keywords: Mesenchymal to epithelial transition, MET, Epithelial to mesenchymal transition, EMT, *Elf3*, transcriptome analysis, bioinformatics

INTRODUCTION AND AIM

1.1. Statement and Importance of the Problem

The role and importance of transcription factor Elf3 in Mesenchymal to Epithelial Transition has been shown. Yet, the mechanism of how Elf3 takes part in this biological process is not revealed. In this study, TGF-B treated NMuMg cells at the third day of MET event will be studied by means of investigating transcriptome dynamics of cells in the absence of Elf3. This approach will give insights into the understanding of Elf3 regulation during MET and will help identify potential targets of Elf3 in downstream pathways.

1.2. Aim of the Study

This study aims to investigate the role of Elf3 on transcriptome dynamics during mesenchymal to epithelial transition. Coding (messenger RNAs) and noncoding transcripts that contain long noncoding RNAs and repeat elements are aimed to be analyzed by computational and statistical approaches.

1.3. Hypothesis of the Study

The role of Elf3 is critical in understanding MET regulation and its absence leads to notable changes in transcriptome dynamics in terms of coding and noncoding parts of genome.

2. GENERAL INFORMATION

2.1. Epithelial to Mesenchymal Transition

2.1.1. Epithelial cells

In living organisms, particularly metazoans, adult tissues are known to originate from two basic cell types which are epithelial and mesenchymal cell types (Pérez-Pomares & Muñoz-Chápuli, 2002). These cell types have distinct features and they can revert to each other by a series of events called Epithelial to Mesenchymal Transition (EMT) or Mesenchymal to Epithelial Transition (MET) (Thiery et al., 2009). Epithelial cells are mostly found to be coherently localized and their side by side, sealed appearance is noteworthy for especially low complexity organism

such as amphibians to protect their embryos from external hazards (Shook & Keller, 2003). In developmental processes, acquiring structural integrity to a tissue consisting of only epithelial cells has been a key property in order to provide a mechanical barrier and also a proper way of maintaining a physiologically stable environment for internal cells (Radisky, 2005). The aforementioned strength of epithelia comes from the interaction of consisting cells with each other and the surrounding environment, extracellular matrix (ECM). Epithelial cells also exhibit cellular polarity that consist of apical and basal sites in plasma membrane which contains specific sets of proteins required for the transport of ions and solutes in and out of cells (Yeaman et al., 1999). Many types of cell-cell adhesions can be given as examples to these relations. Gap junctions that transmit chemical messengers into cells through connexin type of integral proteins on the plasma membrane (Churko & Laird, 2011), adherens junctions mainly consisting of cadherins, tight junctions with occludins and claudins and lastly desmosomes maintained by complexes that is consisting of linker proteins, cadherins and intermediate filaments are contained within a typical epithelial cell (Garrod & Chidgey, 2008; Radisky, 2005). Besides, the contact of an epithelial cell with ECM is established with integrin-based proteins contained in basement membrane that is rich in laminins (Nisticò et al., 2012; Radisky, 2005).

2.1.2. Mesenchymal cells

In the early developmental process, while epithelial cells are considered to be patterned covering the outer layer of embryo, future mesenchymal cells are thought to be interior cells that latterly appears in gastrulation (Shook & Keller, 2003). On the contrary to epithelial cells, these cells are more motile, and they have the ability to move individually. Their cell-cell adhesion properties are scarce and there is no sign of apicobasal cellular polarity considering their mobility. Having a three dimensional extracellular matrix, ECM of a mesenchymal cell does not resemble with the ECM of epithelial cells (Thiery et al., 2009). Mesenchymal cells actively produce new components and reorganize their ECM (Radisky, 2005). Thus, a mesenchymal cell's migratory characteristics require the ECM to have distinct features. For instance, matrix metalloproteinases (MMPs) take crucial roles to remodel the ECM so that the mesenchymal cells can migrate through tissues (Radisky, 2005). Epithelial to mesenchymal transition is a complex series of events for an epithelial cell to gain mesenchymal characteristics in order to become a mesenchymal cell. Once

this process ends, maintenance of the mesenchymal phenotype is achieved by a feedback loop that involves various signaling events. As a consequence, ECM of a mesenchymal cell consists of various growth factors including fibroblast, epidermal and hepatocyte growth factor families and TGF β in order to act on other epithelial cells (Radisky, 2005)

2.1.3. Epithelial to mesenchymal transition

Epithelial to mesenchymal transition can be defined as the sum of changes that an immotile epithelial cell bound in a coherent sheet undergoes. These changes are regulated by transcriptional machinery of the cell in order to transform into a different type of cell which has the ability to move in freedom individually, a mesenchymal cell. During gastrulation stage of embryogenesis, EMT is indispensable for the formation of tissues and organs so that epithelial cells should gain migratory characteristics for some purposes. For instance, in the very early stages of embryonic development, epithelial cells undergoes EMT so as to migrate to create mesoderm which is accounted as one of the three germ layers and the responsible one to build muscles and connective tissue in the later development (Technau & Scholz, 2003). In vertebrates, detachment of epithelial cells in the neural crest in terms of delamination also requires transition to mesenchyme (Technau & Scholz, 2003). Other than developmental processes that are documented, there are other types of events which EMT is a strong player. Tissue repair is one of those processes. When there is an injury in skin tissue, EMT processes were observed in keratinocytes thereby exhibiting a physiological response to wound formation (Thiery et al., 2009). Another recapitulation of EMT processes in tissue repair is shown towards the end of ovulation during menstrual cycle of females (Ahmed et al., 2007).

Despite EMT being a part of normal developmental and adult metabolic processes, its imbalanced regulation may lead to unwanted consequences. These include organ fibrosis and metastasis formation as part of cancer progression (Radisky, 2005). Fibrotic tissues exhibit accumulation of collagen in extracellular matrix so that the redundant amount of the protein become harmful for the organ that the tissue belongs to. This process is thought to involve EMT processes (Thiery et al., 2009). The role of EMT in metastatic processes in cancer has capital importance to understand events occurring during metastasis.

Epithelial to mesenchymal transition is regulated by a complex series of events and involves transcriptional regulations and protein modifications. These events may vary from tissue to tissue and cell to cell (Radisky, 2005) nevertheless some key regulations are known so far. As the main idea is disengagement of epithelial cell from its surrounding cells and environment, it must evade cell-cell and cell-ECM adhesive elements that restrain it from moving. In accordance with that purpose, transcription of those elements such as cadherins (specifically E-cadherin), occludins etc. are reduced by means of repression or they are delocalized (Hugo et al., 2007). Therefore, major modifications in the cytoskeletal structure has to be performed and that includes the appearance of stress fibers consisting of actin and myosins and replacement of type of intermediate filaments cytokeratin to vimentin (Hugo et al., 2007). The ultimate ending is to bring in motility to the cell and gain changeability in shape, all of which is carried out by minimizing contact with other cells and the environment, rearranging the cytoskeletal dynamics accordingly and avoiding apicobasal polarity. The resulting mesenchymal cell that lost many of its epithelial characteristics gain a more elongated morphological appearance, looking like a fibroblast cell. Having mentioned before, the production of MMPs also increases in transition to mesenchymal phenotype and the disruption of basement membrane comes with an interstitial ECM that is rich in collagen (Nisticò et al., 2012).

2.1.4. Molecular regulation: Epithelial to mesenchymal transition inducers

Having mentioned about EMT occurring different physiological process from development to tumor progression in living organisms, this process relies on similar regulation mechanisms. These involve either transcriptional or post-transcriptional modifications. Post translational modifications also take part in EMT processes all of which are dependent on genetic and epigenetic factors. Many researches are carried out to study the regulation of EMT in different biological systems and most of them focus on the modulation of E-cadherin on the plasma membrane. Transcription factors play crucial roles to adjust expression of E-cadherin in epithelial cells during EMT. One of them is Snail family of proteins, containing Slug (Snail2) and Snail1 members as Slug has been long known for its importance in mesoderm formation in *Drosophila* development (Alberga et al., 1991; Pérez-Pomares & Muñoz-Chápuli, 2002). Snail is a zinc finger protein known to take role in the transcriptional repression of E-cadherin by acting on E-boxes found in the promoter region of E-cadherin gene structure (Hugo et al., 2007). Transcriptional repression of E-

cadherin then leads to down-regulation of the protein resulting in release of β -catenin in cytoplasm after E-cadherin expression decreases. This trigger mechanism occurs in the early stages of EMT allowing for β -catenin to get involved in Wnt signaling pathway (Hugo et al., 2007). During gastrulation, regulation of Slug expression is maintained by some factors such as bone morphogenetic proteins (BMP) (Sela-Donenfeld & Kalchauer, 1999) and TGF β (Pérez-Pomares & Muñoz-Chápuli, 2002). Therefore, TGF β signaling pathway that is mediated by Ras and Smad2 proteins are of crucial importance in EMT induction (Thiery, 2003). Moreover, receptor tyrosine kinases (RTK) that have many different growth factor ligands such as epidermal growth factor (EGF), hepatocyte growth factor (HGF) and fibroblast growth factor (FGF) are also players of EMT induction as they activate downstream pathways that involve phosphoinositide-3-kinase (PI3K), Ras and mitogen-activated protein kinase (MAPK) pathway resulting in upregulation of Snail1 and Snail2 (Hugo et al., 2007). Other transcriptional repressors of E-cadherin include ZEB1 & ZEB2, E47 (Thiery, 2003) and KLF8 (X. Wang et al., 2011) which act directly on the gene promoter (Thiery et al., 2009). However, Twist, E2-2/Tcf4, FoxC2 and Goosecoid proteins have been shown to downregulate E-cadherin expression levels indirectly (Thiery et al., 2009). Consequently, the factors that are responsible for transcriptional repression of E-cadherin during transition from epithelial to mesenchymal state are contained in a class called “EMT inducers”.

2.2. Mesenchymal to Epithelial Transition

Mesenchymal to epithelial transition is an inevitable biological process that is followed by epithelial to mesenchymal transition. The role of EMT in normal development, tissue regeneration and cancer progression particularly metastasis have been discussed in the previous sections. However, MET is indispensable for those processes. For instance, tissue and organ formation may require development of specific epithelial cell types during tissue formation after gastrulation which depends on MET processes. While the epithelial cells undergo EMT during the course of metastasis formation as part of carcinogenesis, MET comes next when cells become settled in another tissue that is different from the initiation site. Thus, EMT and MET are complementary processes that are required in normal and abnormal physiological processes (Thiery et al., 2009).

2.2.1. Molecular regulation of mesenchymal to epithelial transition

During MET, mesenchymal cells enter the process of returning to their epithelial characteristics once they acquired. For example, they tend to acquire cell-cell adhesions and gain apicobasal polarity again and lose invasiveness. In fact, restoration of E-cadherin expression is thought to be a hallmark of MET processes (Yao et al., 2011) on the other hand, the reverse applies for EMT. Although the molecular regulation of EMT is a fairly more studied concept, the reverse cannot be said for MET. During MET, the transcriptional repressors of E-cadherin expression (so called EMT inducers) decreases depending on the signaling events. In EMT, transcription factors exhibit their effect on E-cadherin genomic region by binding to E-boxes on the promoter region thus leading to downregulation of E-cadherin. When those factors do not exist any longer, then there will be no repression on the E-cadherin production. Nevertheless, E-cadherin gene promoter is not sufficient for this process alone, as other regulatory sequences are identified (Stemmler et al., 2003, 2005). In a study, deletion of intron 2 of E-cadherin coding gene (*Cdh1*) of mice showed vital role of intron 2 for *Cdh1* expression as it is necessary for the activation and maintenance of its expression (Stemmler et al., 2005). Therefore, two activators, *Hnf4 α* (hepatocyte nuclear factor alpha) and *Grhl3* (grainyhead like transcription factor 3) are reported to act on the referred cis-regulatory element that is found on intron 2 of *Cdh1* (Alotaibi et al., 2015). *Grhl3* is also known to activate the expression of *Hnf4 α* (Sengez et al., 2019). Thus, it can be inferred *Grhl3* is an important transcription factor for the regulation of MET and it is also known to regulate the initiation of MET (Sengez et al., 2019). Other than these factors, there could be other elements that regulate *Cdh1* expression thus regulate MET progression although very little is known about it, currently.

2.2.2. Role of Elf3 in transcriptional regulation of mesenchymal to epithelial transition

Elf3 has been reported to be an important regulator of MET by taking role in the regulation of *Grhl3* by binding to its promoter (Sengez et al., 2019). After *Grhl3* takes action on intronic regulatory sequence of *Cdh1*, it is thought to result in alterations in the localization of E-cadherin (Sengez et al., 2019). Taken all together, *Elf3* is an essential regulator for a normal MET process and its absence may give rise to defects in the transition.

Elf3 (E74 Like ETS Transcription Factor 3) is a transcription factor having almost 5 kb in length and consisting of 8 introns and 9 exons (Tymms et al., 1997) in mice. Being a member of ETS transcription factor family, Elf3 contains a conserved ETS domain for DNA binding (Tymms et al., 1997). The binding sequence of Elf3 is known as 5'-GGA(A/T)-3' and it is known for its role in development and tumorigenesis as common to other members of ETS family (Cooper et al., 2014).

2.3. Next Generation Sequencing & Transcriptome Research

2.3.1. Next generation sequencing technologies & applications

2.3.1.1. Sequencing history: The emergence of next generation sequencing technologies

Forward genetics involves identification of the information contained in DNA is an issue of concern whereas functional interrogation of the information is included in the context of reverse genetics. In context of forward genetics, there has been efforts to identify the sequence of DNA in time, after it has been discovered (Watson & Crick, 1953). The first prevailing method remained more than 30 years was developed by Sanger & colleagues, also known as Sanger sequencing (Sanger & Coulson, 1975). This method requires *in vitro* DNA replication in a reaction mixture containing dNTPs (deoxynucleotide triphosphate), DNA polymerase, primers, a DNA template and lastly ddNTPs (dideoxynucleotide triphosphate) required for chain termination as shown in Figure 1.

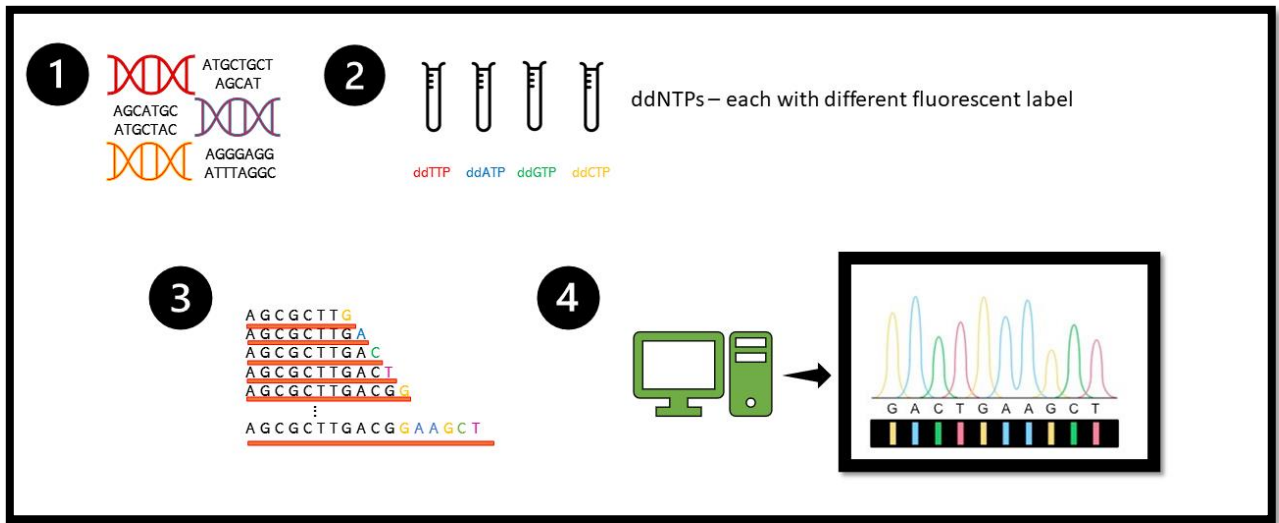


Figure 1. Sanger sequencing general workflow

The ddNTPs having fluorophore labels reveals the DNA sequence that is later being run on a capillary gel electrophoresis (Slatko et al., 2018). Hereafter, computational analysis is carried out and the whole process used to take a lot longer than present-day technologies. The completion of Human Genome Project in 2004 to identify human genome depended on the usage of Sanger sequencing and some developments in technicality of the method (F.S. Collins, E.S. Lander, J. Rogers, 2004).

The advent of newer methods later dominated the Sanger sequencing, also known as first-generation sequencing technology. Developments in the sequencing technologies then led to the emergence of next generation sequencing (NGS) technologies by virtue of need for the sequencing of large genomes with lower cost and higher throughput. NGS introduced massively parallel sequencing at a considerably lower cost especially after researchers have experienced that The Human Genome Project cost almost \$3 billion (Thermes, 2014).

Over time, prevalence of NGS led to its commercialization or vice versa and several companies have adopted different NGS approaches for sequence identification. 454 Life Sciences, also known as Roche developed pyrosequencing method thus paving the way for other initiatives in 2005 (Thermes, 2014). The main principle of this technology is that the addition of pyrophosphates in a chain termination reaction, similar to Sanger sequencing, results in light that is caused by luciferase enzyme activity (Metzker, 2010). Since second generation technologies,

a.k.a. NGS, rely on sequence-based approaches, pyrosequencing determines sequences by means of light intensities which is based on chemical synthesis reactions. Shortly after the release of pyrosequencing, Illumina/Solexa sequencing technology came out in the market in 2006 (Thermes, 2014). Similarly, Illumina sequencing relies on the usage of reversible terminators (fluorescently labeled) that are detectable by the sequencing machine. One year later, Applied Biosystems presented Sequencing by Oligo Ligation Detection (SOLiD). SOLiD and Illumina created greater number of reads in total each having shorter lengths when compared to pyrosequencing technology. The advancements in the sequencing history was then followed by the release of Ion Torrent technology in 2010 (Thermes, 2014). Ion Torrent semiconductor sequencing mainly depended on the detection of protons, released during nucleotide incorporation, by an ion sensor (Thermes, 2014).

All sequencing technologies mentioned here includes a sample preparation step in common together with clonal amplification of genetic material using Polymerase Chain Reaction (PCR). However, while Illumina requires solid phase PCR amplification, 454, Ion Torrent and SOLiD uses emulsion PCR with aid of oil droplets (Metzker, 2010). Yet, in this study, as Illumina technology was utilized, a detailed description to Illumina sequencing technology is given in the next section.

2.3.1.2. Next generation sequencing workflow

A whole typical NGS workflow consists of 4 steps: Sample preparation, amplification, sequencing and lastly data analysis. In context of Illumina technology, the whole process is depicted in Figure 2.

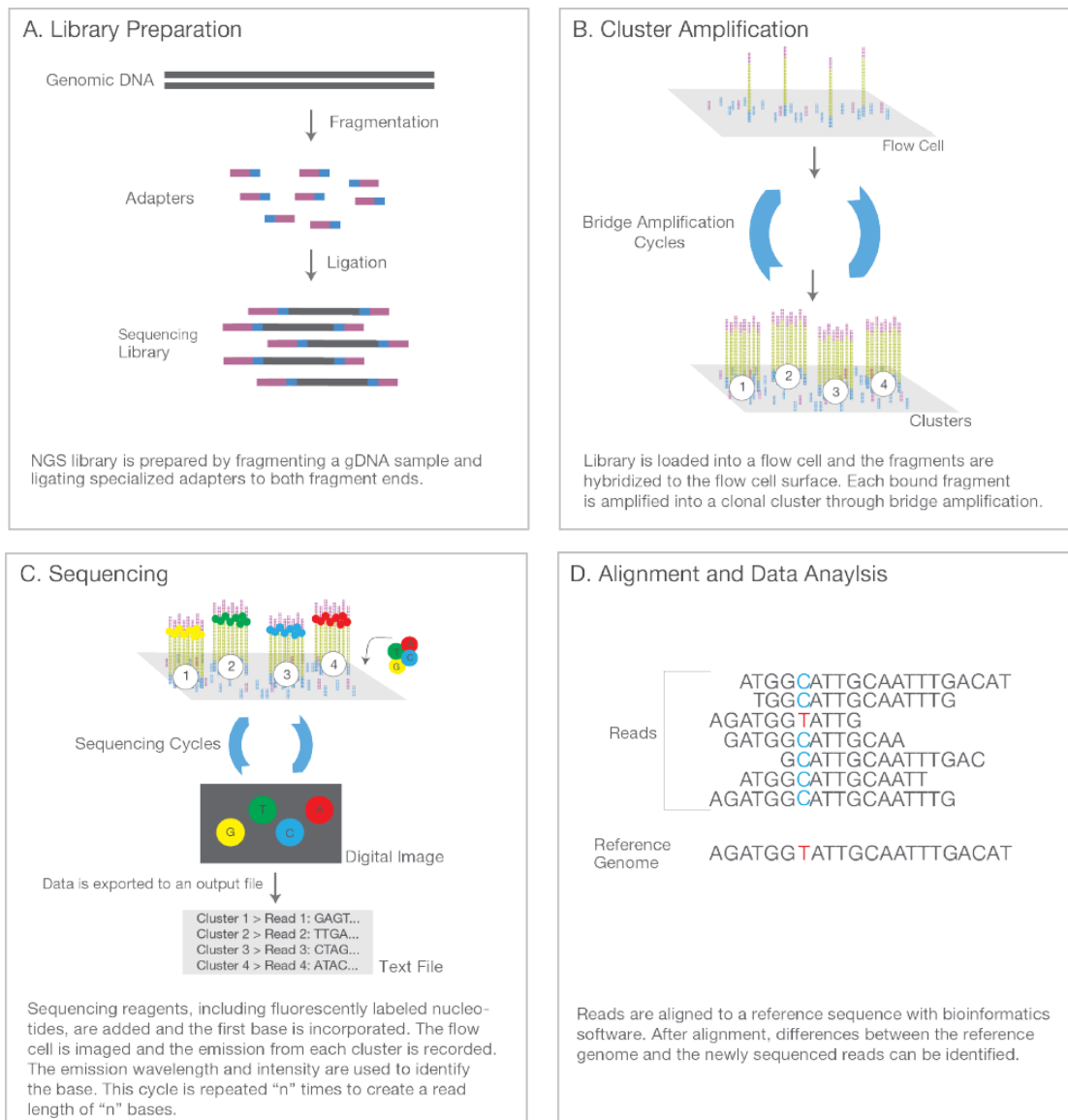


Figure 2. Overview of Illumina NGS technology (obtained from *An Introduction to Next Generation Sequencing Technology*, 01/07/20)

For a summary of NGS steps belonging to Illumina technology; the first step of the procedure is isolating the genetic material. The fragmentation of genetic material is the next step which is required for the adapter binding. Adapters are specific sequences that are required to be ligated to ends of sequences. After library preparation is done, the fragments are required to undergo bridge

amplification by PCR on flow cell surfaces. Once obtained the amplified genetic material, the samples are sequenced by Illumina sequencers having different types such as NextSeq, HiSeq or MiSeq (Source: *Illumina: An Introduction to Next Generation Sequencing Technology*, 01/07/20). The end point to be emphasized for this thesis is what happens after sequencing. The raw data consisting of different combinations of A, C, G and T undergoes a series analysis that requires computational skills and resources depending on the purpose of the study.

2.3.1.3. Next generation sequencing applications

Over the course of years after the emergence of NGS technologies, a burst of massive amount of data has appeared. Therefore, with those data, NGS can provide information for different types of studies depending on the purpose. In this thesis, these are broadly categorized as genome, transcriptome, epigenome and metagenome studies.

Genome Research

NGS can be used to study the somatic and germline mutations with Whole Exome Sequencing (WES) or Whole Genome Sequencing (WGS) to detect disease-causing genes either genome-wide or targeted regions for cancer or inherited type of diseases. WGS aims to target all genomic sequences consisting of exons (1.1%), introns (24%) and intergenic sequences (75%) in human (Craig Venter et al., 2001) whereas WES only targets for exons which makes up a very small portion of human genome. However, it is known that ~85% of known disease causing variants are found in exonic regions (Choi et al., 2009). For example, it has been shown that whole exome sequencing characterized *de novo* mutations that are strongly associated with autism (Sanders et al., 2012). A world-wide consortium (1000 Genomes Project Consortium) aimed to enlighten the effect of human genome variation on phenotype has benefitted extensively by these technologies (Altshuler et al., 2010). Other genome applications of NGS exist in translational research as it is known to be popular in agricultural genetics, forensic genetics and clinical diagnostics (Thermes, 2014)

Transcriptome Research

With the advent of NGS technologies, all types of RNAs with different lengths (messenger RNAs, long non-coding RNAs, microRNAs etc.) can be sequenced with separate protocols and

following analyses. Most of the cases, gene expression profiling is the goal of transcriptome applications which is used to quantify the number of genes or transcripts that are produced in the cell. In the following, differential expression analysis is usually performed to detect genes that are differentially expressed under certain conditions (e.g. drug treatment). Another application of NGS is novel transcript discovery which is a not an easy task taking into consideration of the difficulties of determining splice junctions, transcription start and end sites of transcripts and avoiding false-positives (Conesa et al., 2016). One step ahead of the core mentioned applications of NGS in transcriptome research are comprised of further applications such as functional interrogation of the selected genes. This includes the querying of a number of shortlisted genes in annotation databases to check for any overrepresentation, which can be exemplified as Gene Ontology (Ashburner et al., 2000; Carbon et al., 2019) enrichment or KEGG pathway enrichment analysis (Kanehisa & Goto, 2000). More advanced application of NGS in transcriptome research may include the study of gene fusion discovery which may be difficult due to the fact that fused transcripts are coming from different chromosomes. For example, PTPRZ1-MET fusion transcript in secondary glioblastomas were identified using RNA-seq of 272 glioma samples (Bao et al., 2014). Moreover, if provided sufficient sequence coverage, detection of somatic variation from transcriptome sequencing data is also quite applicable (Piskol et al., 2013) and can be integrated with expression information using only one type of data. Common to any application, analysis of transcriptome sequencing data necessitates careful analysis which will be explained in detail in the next a few pages ahead.

Epigenome Research

Epigenome applications usually refer to the epigenetic marks or chromatin structure which are other types of research topics that NGS has wide applications (Metzker, 2010). Those studies mainly involve the investigation of DNA methylation and histone modifications because they take crucial roles in regulating cellular dynamics by indirectly increasing or decreasing gene expression. Common methods to study epigenome includes Chromatin Immunoprecipitation Sequencing (ChIP-Seq) that employs chromatin immunoprecipitation coupled with NGS to detect binding sequences of histones (Park, 2009), Methyl-seq aiming to detect methylation regions after bisulfite treatment (Ruike et al., 2010) and Assay for Transposase Accessible Chromatin coupled with NGS (ATAC-seq) is employed to detect genome-wide chromatin accessibility (Buenrostro et al., 2015).

For example, using ChIP-seq technology, the role of PHF8 in histone H4 lysine 20 demethylation has been revealed (Liu et al., 2010).

Metagenome Research

The place of NGS in metagenomics is also very crucial for the classification of species in a mixture of genetic material belonging to numerous type of microorganisms (Buermans & den Dunnen, 2014). This is achieved by the brute force approach that aims to sequence any genetic material found in the biological sample by shotgun metagenomic sequencing. Smaller scale applications of metagenomics include 16S rRNA sequencing which employs the usage of hypervariable regions in the 16S rRNA of bacteria (Janda & Abbott, 2007). For example, deep 16S rRNA sequencing of gut microbiota documented the effects of antibiotics on the microorganisms (Dethlefsen et al., 2008). The biggest advantage of NGS in metagenome studies is that NGS does not require labor-intensive culturing procedure which becomes especially hard when it comes to dealing with microorganisms living in extreme conditions.

At the end, it should be noted that integration of different NGS applications may indeed be very powerful for biological research. For example, the combination of variation and expression data by means of DNA and RNA sequencing can be used together for expression quantitative trait loci (eQTL) mapping to investigate disease susceptibility or the genetic basis of some traits such as height (Conesa et al., 2016).

2.3.2. Transcriptome analysis with RNA-sequencing

Transcriptome can be defined as the sum of all transcripts together with their numerical quantities under biological circumstances such as developmental stage or a specific condition (Z. Wang et al., 2009). This definition includes both varying lengths of coding & noncoding RNAs of a cell or tissue. Until this time, there has been efforts to characterize and quantify those transcripts and different types of tools have been developed. These methods are classified into two: hybridization and sequence-based approaches. Microarrays utilize the hybridization-based approach with incubation of cDNA (complementary deoxyribonucleic acid) samples labelled with fluorescent (Z. Wang et al., 2009). Here, microarrays may be custom-made or offered by commercial companies. Microarray technologies have some drawbacks that allowed them to be

dominated by other technologies. This method requires the knowledge of genome prior to studying it. Although inexpensive, high background levels and inadequate resolution are considered other disadvantages of microarrays excluding high-resolution tiling arrays (Z. Wang et al., 2009).

Other high-throughput technologies to quantify transcripts is Next Generation Sequencing technology which has the sequence-based approach. NGS technologies have been used for many years for the characterization of DNAs and RNAs. Quantification of RNAs is possible with NGS technologies being referred to as RNA Sequencing (RNA-seq) which refers to quantifying all types of RNAs which may consist of messenger RNAs (mRNA), long noncoding RNAs. In Figure 2, a general outline of a typical RNA-seq experiment is depicted. Different types of RNAs are subjected to different procedures like Poly-A selection to enrich for mRNAs and ribo-depletion for all types of RNAs except ribosomal RNAs. In the following, coding and noncoding transcriptome in two separate topics are briefly covered.

2.3.2.1. Coding transcriptome

The central dogma of molecular biology states that information kept in DNA molecules are passed onto mRNA molecules which is then used to make proteins that have different particular function in cellular processes (Crick, 1970). According to this understanding, messenger RNAs are called coding transcripts because they are carriers of biological information and they are decoded by translational machinery to be translated into a protein. For eukaryotic cells, messenger RNAs are transcribed in nucleus where DNA is located, and they undergo some modifications during transcription which make them distinguishable afterwards. These modifications include 5' capping, 3' polyadenylation which take crucial roles in mRNA stabilization and transcription termination (Alberts et al, 2014). Besides its structure, a typical mRNA contains important regulatory sequences called 5'- and 3'- UnTranslated Regions (5'- and 3'-UTR) that are required for proper subcellular localization, careful transport of mRNAs together with affecting translation efficiency (Alberts et al, 2014).

2.3.2.2. Noncoding transcriptome

For the recent years, it is now beyond doubt that transcriptome of a eukaryotic cell does not comprise of coding transcripts considering the fact that only 2% of genome is translated into proteins (Dunham et al., 2012). In fact, this shows that transcripts that do not code for proteins may have greater importance compared with coding mRNA transcripts as opposed to the belief that rest 98% percent was considered “junk” once. Therefore, noncoding part of genome give rise to different types of noncoding RNAs that have separate roles in cells. Ribosomal RNA (rRNA) and transfer RNA (tRNA) can be given to noncoding RNAs which are well known for their contribution to mRNA translation and their abundance. On the other hand, noncoding transcripts are broadly categorized into two by their size: long noncoding (lncRNA) and small noncoding RNAs (sncRNA). Small noncoding RNAs have smaller than 200 nucleotides in length and miRNA, siRNA, piRNA, snoRNA and snRNA can be given typical examples of sncRNAs (Diamantopoulos

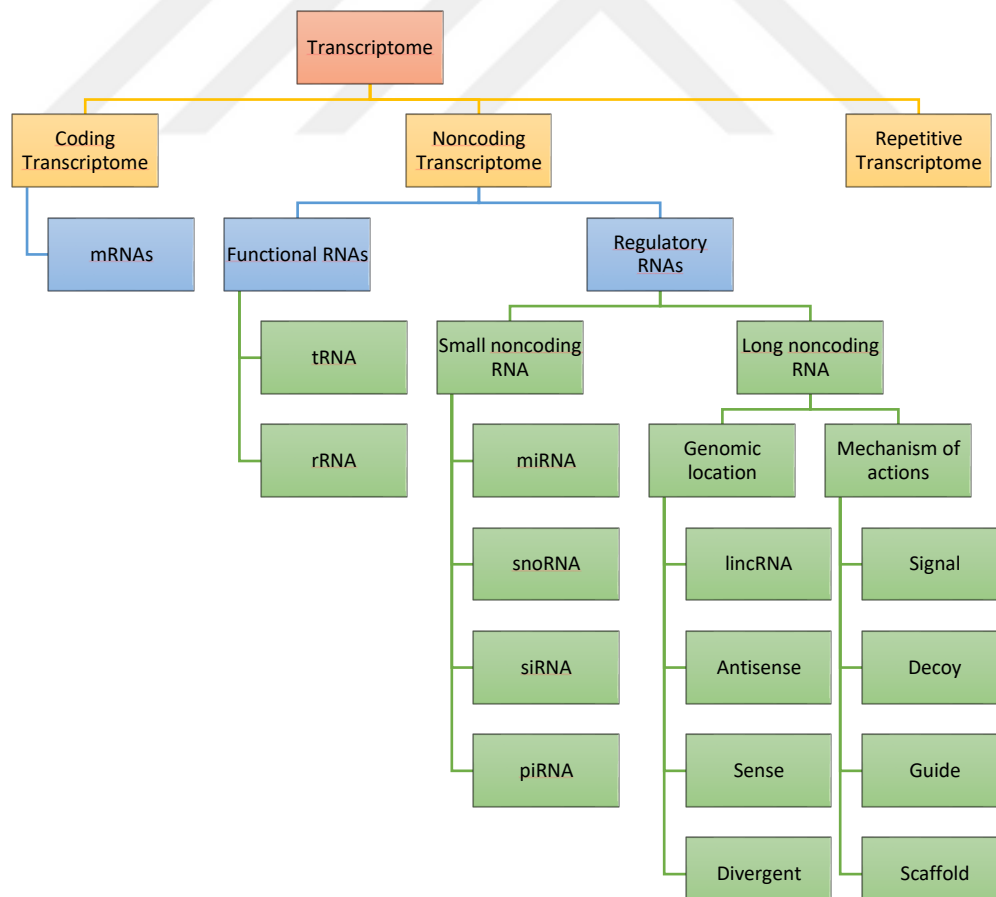


Figure 3. General schema of a eukaryotic transcriptome

et al., 2018). SncRNAs have various functions in gene regulation such as repressing the translation of mRNA by binding to it (miRNA), nucleolytic processing of rRNAs (snoRNAs) or pre-mRNA splicing in spliceosome structure (Diamantopoulos et al., 2018). Other important type of ncRNAs are long noncoding RNAs having more than 200 nts in length. Long noncoding RNAs are special transcripts that their expression is usually observed under stress conditions or they are mostly produced in a tissue specific manner.

2.3.2.3. *Repetitive transcriptome*

Transposable elements (TEs) are mobile genetic elements that “repeat” their sequences along the genome by specific mechanisms. It has been reported that nearly half of the mammalian genome consists of TE contributions (Lander et al., 2001; Waterston et al., 2002). Taking into consideration of their widespread existence in the genome, they have specialized ways to translocate in the genome. These mechanisms of spread results in the classification of TEs to in two classes as Class I and Class II (Wicker et al., 2007). Class I transposons exhibit a “copy and paste” mechanism in which an RNA intermediate is produced to copy the information and then it is converted to DNA back by a special enzyme called reverse transcriptase and finally it is integrated to a different location in the genome (Wicker et al., 2007) Class I transposons resemble retroviruses by means of their translocation, so that they are also known as “retrotransposons” (Bourque et al., 2018). On the other hand, latter class of transposons do not use an RNA as an intermediate element. Class II transposons exhibit a “cut and paste” mechanism since this mechanism involves cutting of DNA sequence in the host site and then integrating it to a new site in the genome (Wicker et al., 2007). TEs seem to be quite responsible for the transcriptome dynamics because they are known to affect the gene expression by means of modifying information in coding and noncoding sequences. For example, they have been reported to be important source of regulatory sequences such as promoters (Jordan et al., 2003) or transcription factor binding sites (Karakulah, 2018). Therefore, in coding transcriptome, they have been shown to be found in mRNA sequences especially untranslated regions (Jasinska & Krzyzosiak, 2004). In the end, the place of TEs in gene expression has been documented several times (Hancks & Kazazian, 2016; Johanning et al., 2017; Yandım & Karakulah, 2019) so that the expression of TEs may give important aspects in the understanding the transcriptome dynamics.

2.3.3. RNA sequencing and data analysis

The raw data coming from sequencers are not suitable for direct use in interpretation. For that purpose, the data undergoes some necessary steps until a meaningful information is extracted. The steps required for a typical RNA-seq data analysis with the purpose of identifying differentially expressed transcripts are gathered in a pipeline consisting of: raw data quality control processing, read alignment to reference genome, gene expression quantification and differential expression analysis.

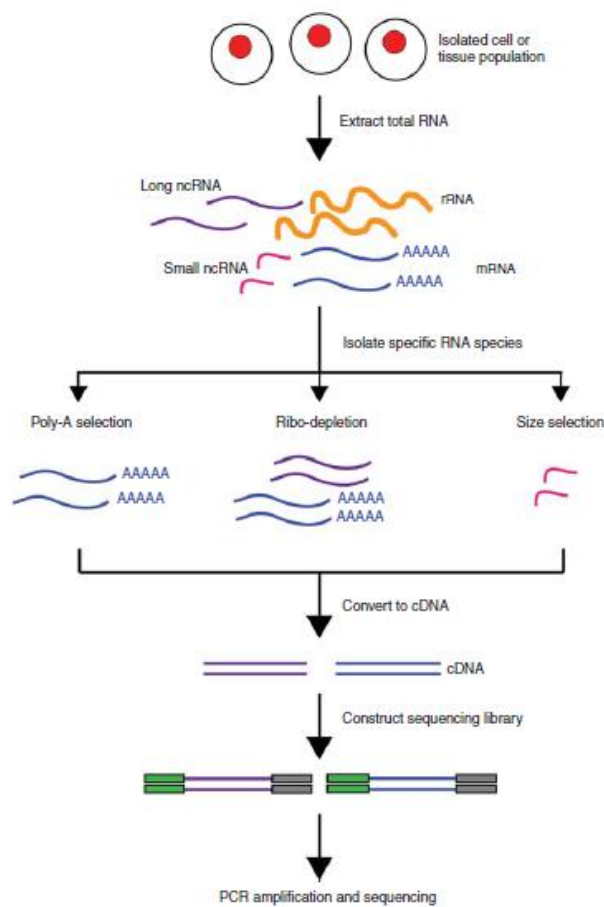


Figure 4. RNA-seq library preparation overview. Figure adapted from *RNA Sequencing And Analysis* (Kukurba & Montgomery, 2015) (See Appendix 1)

2.3.3.1. Raw data quality control processing

Occasionally, raw data may include problematic reads possibly due to problems in sequencing machine, PCR amplification or any type of contamination (i.e. biological

contamination). For that reason, the dataset is routinely checked for its general quality before proceeding to analysis. There is several software that perform quality control of raw reads. NGSQC (Dai et al., 2010) and FASTQC (Andrews, 2015) are tools that can be used for quality control however FASTQC is solely suitable for Illumina reads. Quality control tools give statistics about the sequence quality, overrepresentation of any sequences, duplicated reads, GC content or adapter content (Conesa et al., 2016). The presence of adapter sequences is expected as a result of adapter ligation procedure during library preparation before sequencing. Although some sequencers provide pre-trimmed data nowadays, adapter trimming must be performed in case of adapter content. Likewise, quality trimming is also an application for the presence of low-quality reads after quality check. The tools that are mentioned here for quality check are not used for trimming of raw data, yet other types of software are available for that purpose. Cutadapt (Martin, 2011) can be given as an example to quality or adapter trimming purposes of NGS reads.

2.3.3.2. Read mapping

As soon as clean reads are obtained after a successful quality check processing of raw data, reads are subjected to the key stage of RNA-seq data analysis; read alignment. Read mapping is critical to address the locations of reads on reference genome. Hence, reads are mapped to a reference genome or transcriptome which may differ based on the purpose. Prior to mapping, a reference either should be available on hand for the species of interest or it should be reassembled *de novo* (Z. Wang et al., 2009). Therefore, reference transcriptome is usually preferred if the further processing is only based on the quantification of known transcripts also knowing the fact that it only contains exons let alone introns (I. S. Yang & Kim, 2015). This approach becomes important when choosing the suitable software for read alignment. Read aligners are broadly categorized into two as splice-aware and splice-unaware in consideration of splice junctions (Lowe et al., 2017). A splice-aware read mapper places reads regarding the exon-intron boundary information thus increasing the accuracy of overall alignment (Kukurba & Montgomery, 2015). Some of the splice aware aligners that perform alignments in a spliced manner are STAR (Dobin et al., 2013), TopHat (Trapnell et al., 2009) and HISAT2 (Kim et al., 2015). On the other hand, BWA (Li & Durbin, 2009) and Bowtie (Langmead et al., 2009) are examples of splice-unaware aligners.

Everything aside, there are other factors when selecting the appropriate mapper software. Knowing the fact that alignment is typically the most computational intensive step of high-throughput data analysis experiments, properties like speed, performance and memory utilization of software are considered thoroughly (Kukurba & Montgomery, 2015).

2.3.3.2. *Expression quantification and normalization*

After obtaining sufficient mapping rates (may differ according to the species), the next step of RNA-seq data analysis is the quantification of features. Therefore, there are various approaches for expression quantification. Gene and isoform level quantifications are two approaches that are handled delicately. For gene-level quantification, in some methods, transcript assembly first must be performed and quantification is made with the assembled transcriptome (I. S. Yang & Kim, 2015). As mentioned before, this can be achieved either reference-guided if reference is available or *de novo* (I. S. Yang & Kim, 2015). StringTie (Pertea et al., 2015) can be given as an example for transcript assembly when reference is present, whereas Trinity (Grabherr et al., 2011) is an example for *de novo* transcript assembly (Conesa et al., 2016). After assembling the transcriptome, Cufflinks which depends on maximum likelihood approach (Trapnell et al., 2011) or MISO utilizing Bayesian inference (Katz et al., 2010) are different tools that can be used for estimation of gene expression in the following step (Kukurba & Montgomery, 2015). For the quantification of lowly expressed genes, MISO may perform better estimation of gene expression (Garber et al., 2011). It should be noted that this approach also applies for isoform-level quantification. On the other hand, there exists simpler methods for gene expression quantification. HTSeq-count (Anders et al., 2015) and featureCounts (L. Yang et al., 2014) utilize raw count aggregation approach with the help of a special format file called Gene Transfer Format (GTF). Chromosomal coordinates of exons and genes on genome are found in the GTF file so that the quantifier benefits the information in order to do the estimation of gene expression (Conesa et al., 2016).

Normalization is usually indispensable for interpreting expression quantification due to some reasons. To avoid bias in the experiment, factors like transcript length, library fragment size and read depth must be overcome (Kukurba & Montgomery, 2015). RPKM (reads per kilobase of exon model per million reads) is a normalization method used to normalize the transcript count based on gene length for single end reads whereas FPKM (fragments per kilobase of transcript per million

mapped reads) is analogous to RPKM metric, yet FPKM fits for paired end reads (I. S. Yang & Kim, 2015). It can be concluded that FPKM and RPKM have the same numerical values in the case of single end reads at hand. Lastly, TPM (transcripts per million) is another normalization metric that is developed to overcome statistical bias of aforementioned metrics (Wagner et al., 2012).

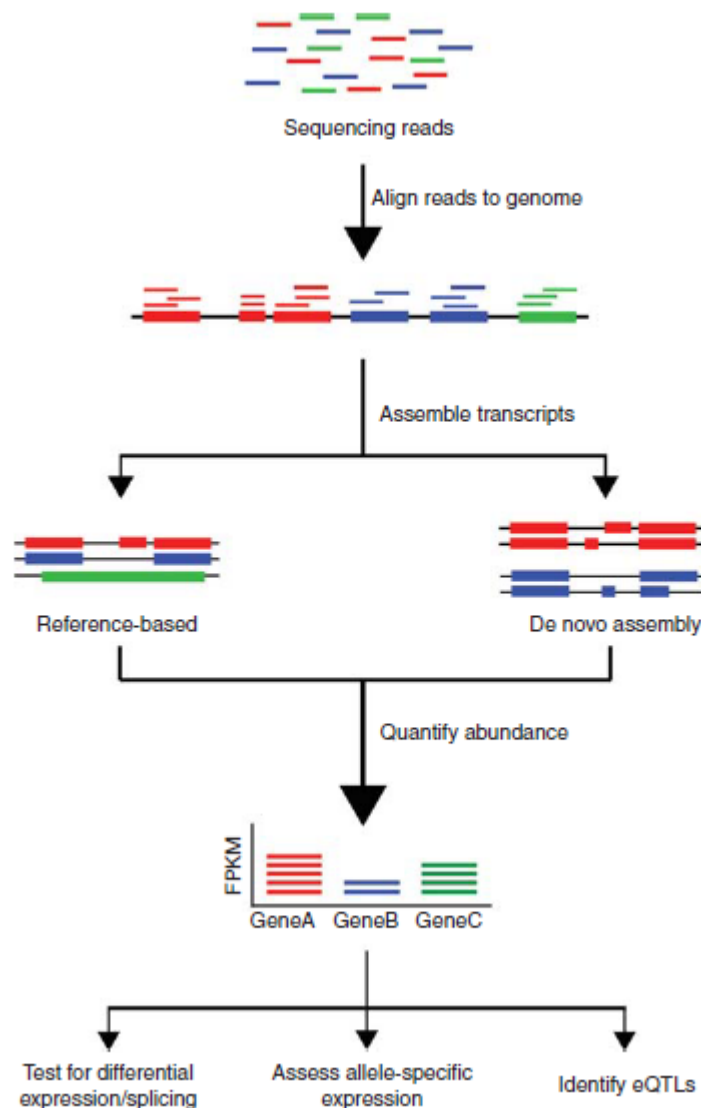


Figure 5. General schema of a typical RNA-seq data analysis experiment. Figure adapted from *RNA Sequencing and Analysis* (Kukurba & Montgomery, 2015) (see Appendix 1)

2.3.3.3. *Differential expression analysis*

Figure 3 finely depicts the general schema of RNA-seq data analysis starting from alignment to differential expression analysis (DEA) including the concepts that are introduced. All parts until differential expression analysis is elucidated in previous parts. Differential expression analysis can basically be described as the comparison of gene expression in different conditions. This analysis can be performed using different methods as Cuffdiff (Trapnell et al., 2010), edgeR (Robinson et al., 2010), DESeq (Anders & Huber, 2010), limma (Ritchie et al., 2015) and many others. These methods utilize different statistical measures as well as different types of normalization however they all take raw count data for the analysis. DESeq and edgeR commonly adopt the hypothesis in the beginning that most genes are not differentially expressed. Yet, DESeq adopts negative binomial distribution approach after normalizing with DESeq sizeFactors between samples. Similarly, edgeR also adopts negative binomial distribution after computing a trimmed mean of M-values (TMM) between samples. Nevertheless, the two approaches depend on exact tests while edgeR, in addition, also applies generalized linear models and quasi-likelihood tests. Cuffdiff also depends on negative binomial distribution and it uses t-test to calculate for the differential expression. On the other hand, Limma has been used mostly for microarray data before however the developers had made it possible to be compatible with RNA-seq data and now it adopts the TMM (trimmed mean of M-values) normalization and then applies empirical bayes method for differential expression testing. All mentioned tools serve the same purpose of calculating the differences in counts of genes between different samples although having small statistical differences that may be suitable for certain set of samples.

In the end, differential expression analysis gives us a list of genes that have a significance level by means of p- and p-adjusted values. P adjustment is usually performed with Benjamini and Hochberg (Benjamini & Hochberg, 1995) equation that applies Fold Discovery Rate (FDR) which takes into consideration of the proportion of false positives among all positives (including both true positives and false positives). After obtaining a list of most significantly expressed genes between samples, they are subjected to further analysis such as Gene Ontology or pathway enrichment analyses to check if they are overrepresented in any Gene Ontology terms or known pathways.

3. MATERIALS AND METHODOLOGY

3.1. Type of Study

This study is based on computational and statistical calculations applied on biological samples.

3.2. Time and Place of the Study

This study was conducted at Izmir International Biomedicine and Genome Institute and Izmir Biomedicine and Genome Center between February 2020 and July 2020.

3.3. Materials of Study

In this study, NMuMg cells were treated with TGF-beta for 72 hours to induce EMT formation. Then, the cells were washed off the TGF-beta in order to initiate the MET process in reverse for another 72 hours. As the cells reach 72 hours after MET initiation, two replicates of cells together with their negative controls were knocked down for Elf3 using the suitable siRNA library (Fig. 6). Two samples belonging to Post-Treatment-72 (PT72) and two samples belonging to scrambled controls were then undergone total RNA isolation. From now on, these samples are referred to as “Scr” and “siElf3” throughout the study. Library preparation for Illumina sequencing of samples were carried out involving the depletion of ribosomes. The constructed whole transcriptome libraries were then sequenced using Illumina HiSeq 2000 platform with single-end reading method. This study starts with the short sequencing reads generated from those samples.

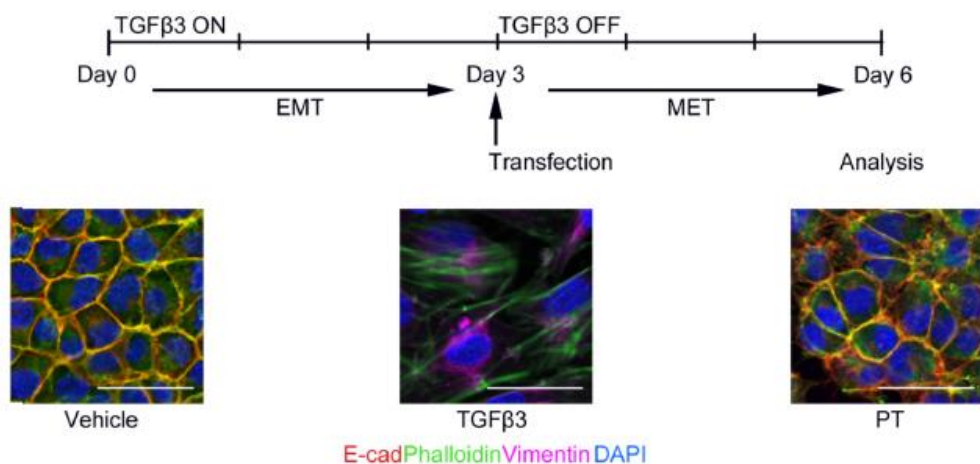


Figure 6. EMT induction with TGF β 3 treatment which is then followed by MET induction after removal of TGF β 3

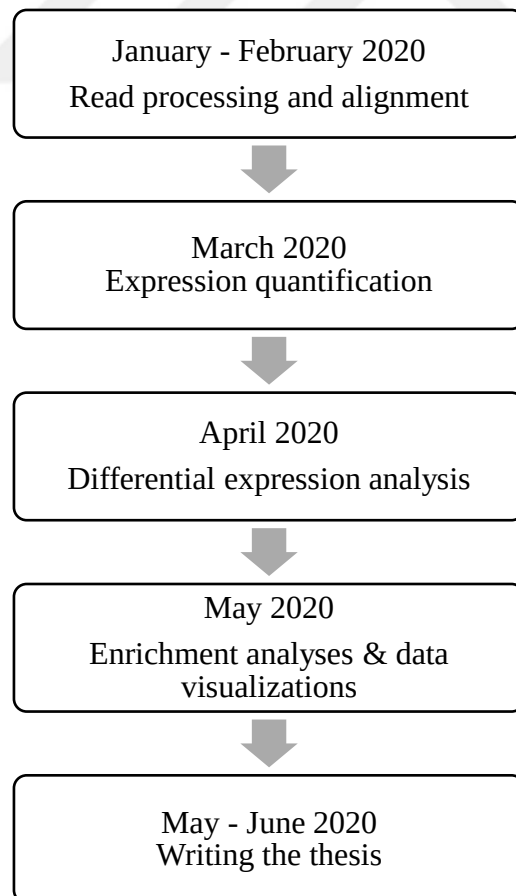
3.4. Variables of Study

Knock-down and control conditions belonging to PT72 of NMuMg cells that were undergone MET process are accounted as independent variables whereas quantified transcripts are dependent variables.

3.5. Tools for Data Collection

HP Z2 G4 WorkStation that contains 32 GB RAM, 2TB hard disk with Ubuntu 18.04 installed and IBG HPC environments have been used for data collection in this study.

3.6. Study Plan and Calendar



3.7. Data Evaluation

3.7.1. RNA-seq raw data quality control

In the beginning, four FASTQ formatted files that contain sequencing information were undergone quality check with FastQC quality control tool in order to see if the samples contain low quality reads or any adapter sequences. As the adapters had to be removed to avoid its contamination out of the study, Cutadapt software was used to eliminate Illumina sequence adapters. At the end of the current quality control stage, clean reads were ready to proceed to further analysis.

3.7.2. Alignment of reads to reference genome

In sequencing data analysis pipelines, analysis-ready reads are required to be aligned to a suitable reference genome build. For that purpose, firstly the reference genome build was selected and determined as Genome Reference Consortium Mouse Build 38 (GRCm38) release M21 that is provided by GENCODE Project. Secondly, index of reference genome was built depending on the selected alignment software. As the reads were being aligned to a reference genome that contains exonic and intronic regions, on the other hand, the reads theoretically did not have intronic regions, thus they had to be aligned to a splice aware aligner which is HISAT2 in this case. The required commands to create a HISAT2 genome index is: `hisat2-build <ref_genome> <genome_index> -p <n_threads>`. The next step was to align the reads to the reference genome having a newly created index. The following commands were used: `hisat2 -p <n_threads> --dta -x <genome_index> -U <fastq> -S <SAM>`. The resulting output of the previous operation is a large SAM (Sequence Alignment Map) file that contains alignment information of reads to the reference genome. The information in this file had to be sorted by their chromosomal coordinates and converted to A compressed file format which is called a Binary Alignment Map (BAM). Samtools software performed those 2 processes in one stage with the following command sequence: `samtools sort -@ <n_threads> -o <BAM> <SAM>`.

3.7.3. Quantification of protein-coding and long non-coding RNA coding genes

In this stage, the resulting BAM files coming from alignment step was subjected to gene quantification. This step involves the usage of a software that counts reads based on their chromosomal coordinates that is attributed to a genetic feature (exon, intron, repeat elements etc. In this section, exon quantification of reads was performed. Here, R statistical computing environment (v3.6.0) was used together with required package, Rsubread. Here, featureCounts() utility that is offered by Rsubread was used so as to count the exonic regions in BAM files. A custom GTF file that was constructed for the MET model of NMuMg cells in the same time-series experiment that these samples are taken from was used as the external annotation input in this study. FeatureCounts options were set to allow the quantification of reads mapping to multi-sites. Using the rpkm() function of another R package “edgeR”, normalized thus RPKM values of quantified transcripts together with raw counts were obtained at the end of this stage.

3.7.4. Quantification of repetitive elements

After counting reads for protein coding and non-coding genes, repeat quantification was also performed. For that purpose, firstly RepeatMasker repeat file belonging to mm10 (also known as GRCm38) released in Dec 2011 was downloaded. Then, using R scripts, only long interspersed nuclear elements (LINEs), short interspersed nuclear elements (SINEs), long terminal repeats (LTRs), DNA transposons and Satellites are included in the final annotation file to be used in FeatureCounts. Afterwards, repeat quantification was performed using FeatureCounts with the specified annotation file without allowing for reads mapping to multi-sites.

3.7.5. Differential expression analysis

After having raw counts and FPKM values, raw counts were filtered based on FPKM values. For that, raw counts of genes or repeats were removed if they had FPKM value of less than 1 in all the samples. To count differences in the expressions of protein-coding, non-coding and repeat elements between scrambled and knock-down samples, differential expression analysis was carried out. The analysis was carried out using custom R scripts that contain steps of TMM normalization, dispersion estimation and performing exact test between groups. EdgeR was used for differential expression analysis. Then, Ensembl gene ids were converted to appropriate gene names.

3.7.6. Enrichment analyses and visualizations

GO enrichment and KEGG pathway enrichment analyses were performed using “enrichGO” and “enrichKEGG” utilities of clusterProfiler, respectively ($p\text{-adjust} < 0.05$). Using the same R package, redundant GO terms were simplified using “simplify” utility which is based on simplification of terms by taking account of GO hierarchy. All visualizations were performed on R, using packages ggplot2, ggpubr, EnhancedVolcano and finally clusterProfiler.

3.8. Limitations of the study

Small number of replicates and samples limit the scope of this study since that prevents further computational analysis such as a correlation analysis between Elf3 and its probable effect on other proteins.

3.9. Ethics Committee Approval

Izmir Biomedicine and Genome Center Noninvasive Researches Ethics Committee Approval carrying protocol number “2020-013” is granted for this thesis (see Appendix II)

4. RESULTS

4.1. Preprocessing and alignment of sequencing reads

Raw FASTQ files containing low-quality reads and adapter sequences were undergone quality check and they were removed from samples. Clean reads were obtained after that process and the data was proceeded to further analysis. In the first place, reads were mapped to reference genome using HISAT2 and the alignment statistics were obtained. Table 1 shows the alignment statistics.

Table 1. Alignment statistics of Scr and siElf3 samples

Samples	Replicate	Overall Alignment Rate	Mapped Read Count	Total Read Count
Scr	1	95.22%	64 522 585	67 761 589
	2	95.28%	62 382 087	65 472 384
siElf3	1	95.04%	55 844 965	58 759 433
	2	95.14%	55 929 625	58 786 656

According to Table 1, it can be seen that overall alignment rate of all samples was around 95% for Scr and siElf3 samples which indicate that almost all reads successfully mapped to somewhere in the mouse reference genome.

4.2. Overview of gene expression profiling with siElf3 knockdown

Expression quantification of each BAM file was performed using FeatureCounts. Raw counts and FPKM (or RPKM) values belonging to 56646 mouse genes were obtained. Figure 7 shows the Spearman correlation between samples. It can be inferred from the figure that all samples correlate with each other however siElf3 and Scr samples show a distinct separation which is an expected consequence. Therefore, replicates belonging to each biological sample resemble each other yet Elf3 replicates show a higher similarity.

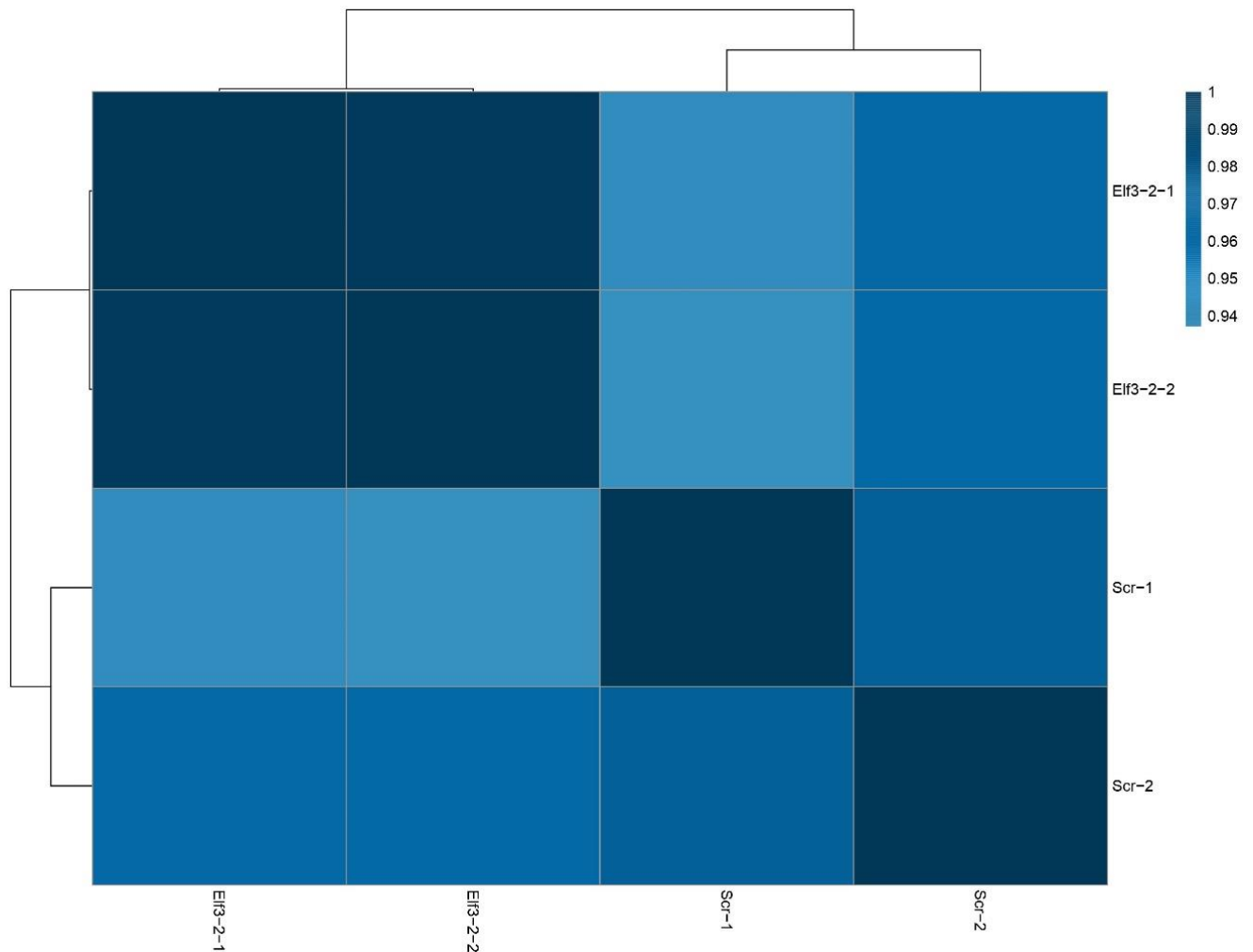


Figure 7. Correlation heatmap of expressions belonging to two Scr and to siElf3 samples.

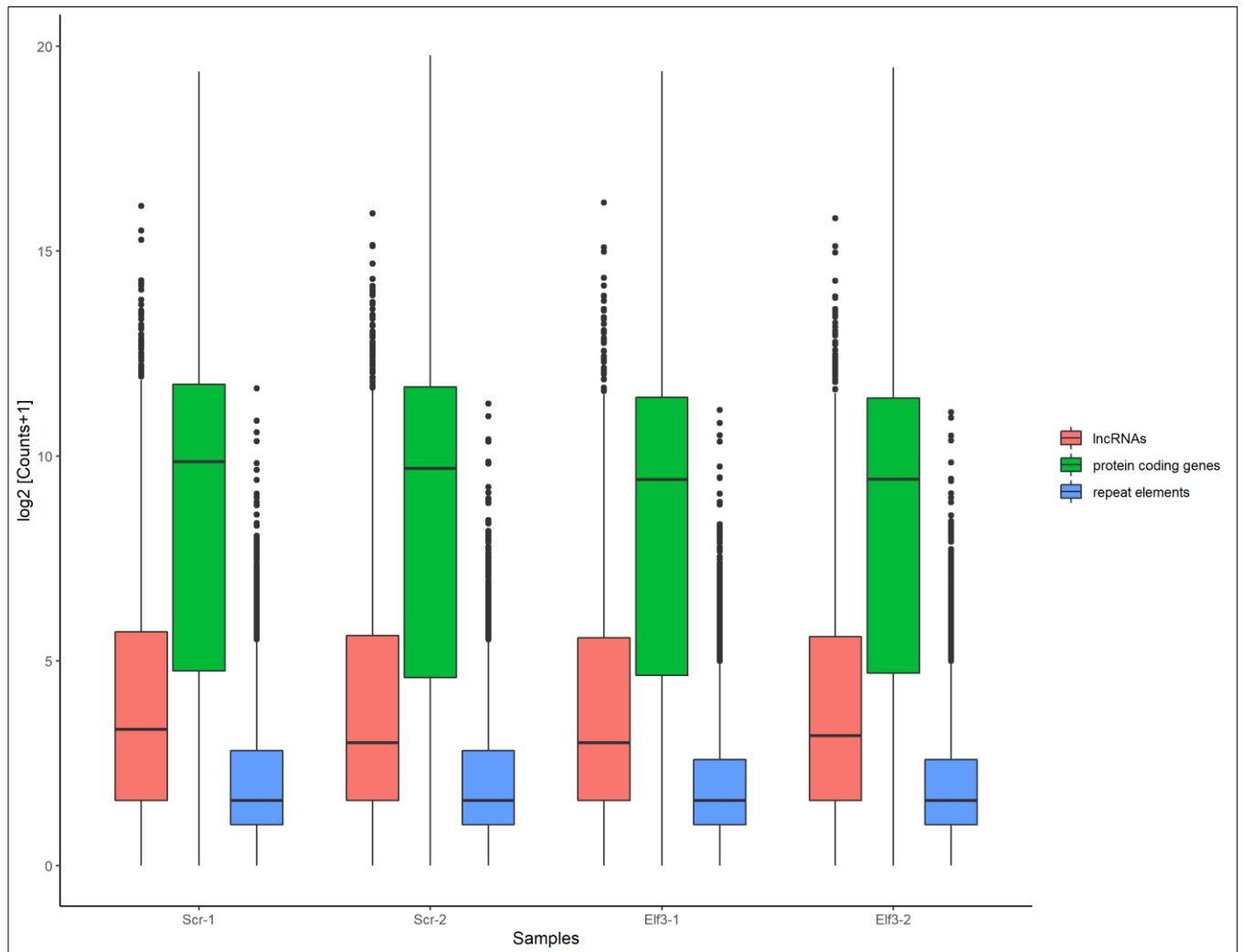


Figure 8. Boxplot showing protein coding, lncRNA and repeat elements content distribution.

Figure 8 shows the expression distribution of the transcriptome by means of protein coding genes, lncRNAs and repeat elements. Here, long noncoding RNAs are confined to the genetic features of lincRNA (long intergenic noncoding RNA) and antisense out of all other annotated biotypes defined in Ensembl annotation. Also, it contains the previously un-annotated lncRNAs as found in the unpublished data. Repeat elements only contain five distinct families as LINE, SINE, DNA, LTR and Satellite. The boxplot shows that all samples exhibit a very similar distribution of genetic features and there is not an outlier sample. It also informs us that the expression of protein coding genes is the highest, which is then followed by the expression of lncRNAs and repeat elements.

4.3. Differential expression analysis between Scr and siElf3 samples

Differential expression analysis between Scr and siElf3 samples resulted in 14130 differentially expressed genes (DEGs) including previously un-annotated lncRNA genes annotated in the previous unpublished study of Alotaibi group. Samples of this study are taken from the referred time-series experiment as described in Materials section. 3223 of the DEGs, having their FDR (False Discovery Rate) less than 0.05 were found to be significant. Of the significant DEGs, 2749 of them had greater than absolute 0.6 log fold change values. The resulting list of genes contained 1208 up regulated and 1541 down regulated genes.

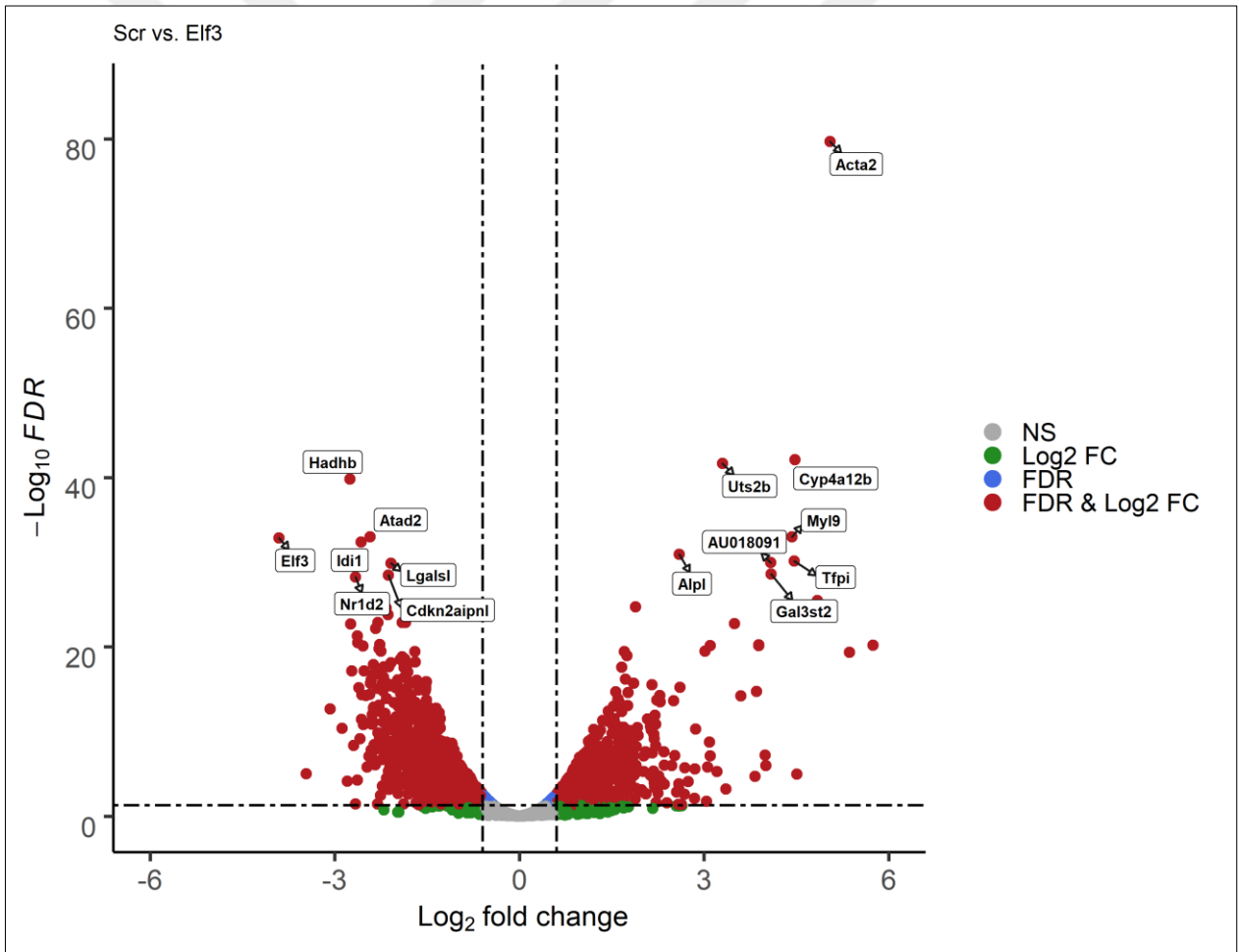


Figure 9. Volcano plot showing the distribution of differentially expressed protein-coding genes (NS: Not Significant, log2FC: log2 Fold Change, FDR: False Discovery Rate).

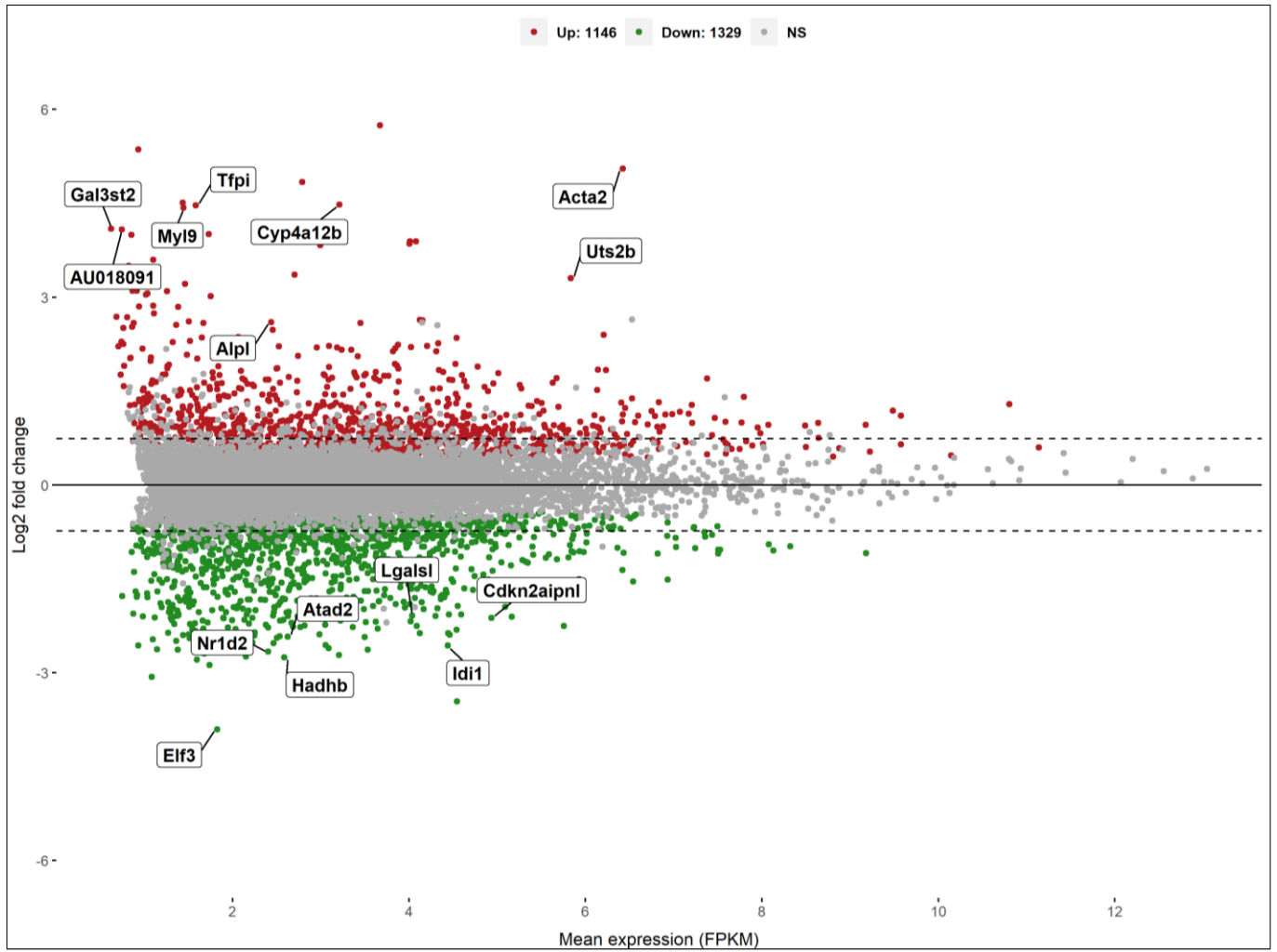


Figure 10. MA plot showing the distribution of differentially expressed protein-coding genes (log₂ fold change threshold ± 0.6 , FDR threshold < 0.05 ; NS: Not Significant, FPKM: Fragments per kilobase of transcript per million mapped reads).

Volcano plot in Figure 9 shows the distribution of differentially expressed protein-coding genes together with log₂ fold change and $-\log_{10}$ FDR values. It is represented by different colors that are separated by cutoffs. The horizontal dashed line separates genes having log₂FC that equals to ± 0.6 which stands for around 1.5-fold change difference between Scr and Elf3 samples. On the other hand, the vertical dashed line separates genes having FDR that equals to 0.05. In the figure, FDR values are given with a negative log₁₀ for a more suitable representation. It is clearly seen from the figure that red circles represent DEGs that are significantly expressed within the fold change cutoff. Therefore, the most upregulated genes are shown in the right side whereas the most downregulated

genes are shown in the left side. According to the volcano plot, the blue and green circles represent DEGs that do not pass the any of the two cutoffs. In the figure, top fifteen most significant DEGs containing Elf3, which is an expected result, are labelled. The volcano plot shows the presence of numerous protein-coding genes such as Elf3, Atad2 and Acta2 that have large fold change values as well as high statistical significance.

In Figure 10, the distribution of protein coding DEGs with their mean expressions in all samples are shown in an MA plot. The threshold in the plot is set as ± 0.6 , resulting in upregulated genes having red color in the top side and down regulated genes having green color in the bottom side. The grey color refers to non-significant genes that are below 0.05 FDR cutoff. Here, again top fifteen most significant DEGs containing Elf3 are labelled. MA plot is an accompanying plot coming after volcano plot since MA plot informs about the mean expression levels of the most upregulated or downregulated genes. It is also successful to show genes having high expression levels and a high fold change but no significance.

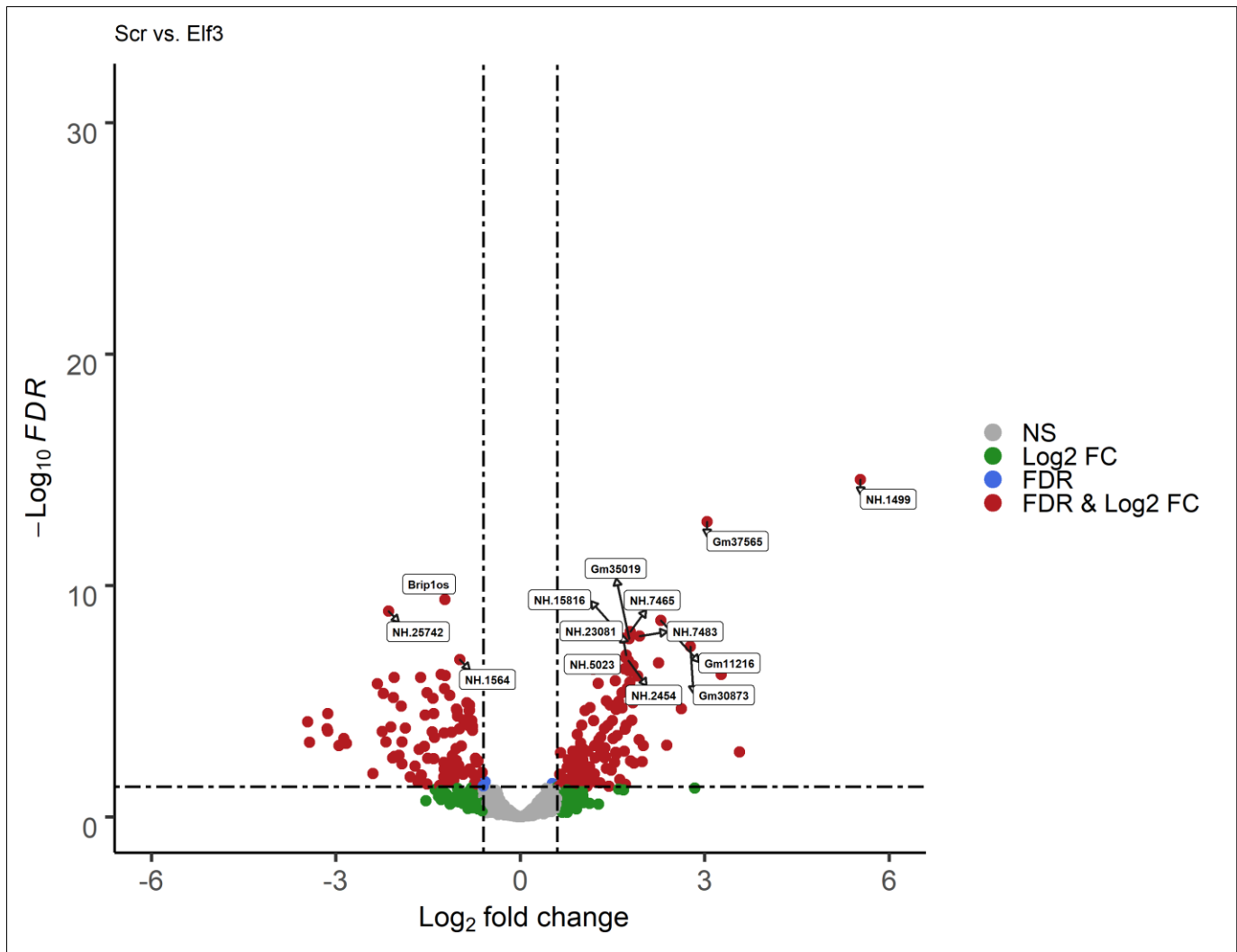


Figure 11. Volcano plot showing the distribution of differentially expressed lncRNAs (NS: Not Significant, log2FC: log2 Fold Change, FDR: False Discovery Rate).

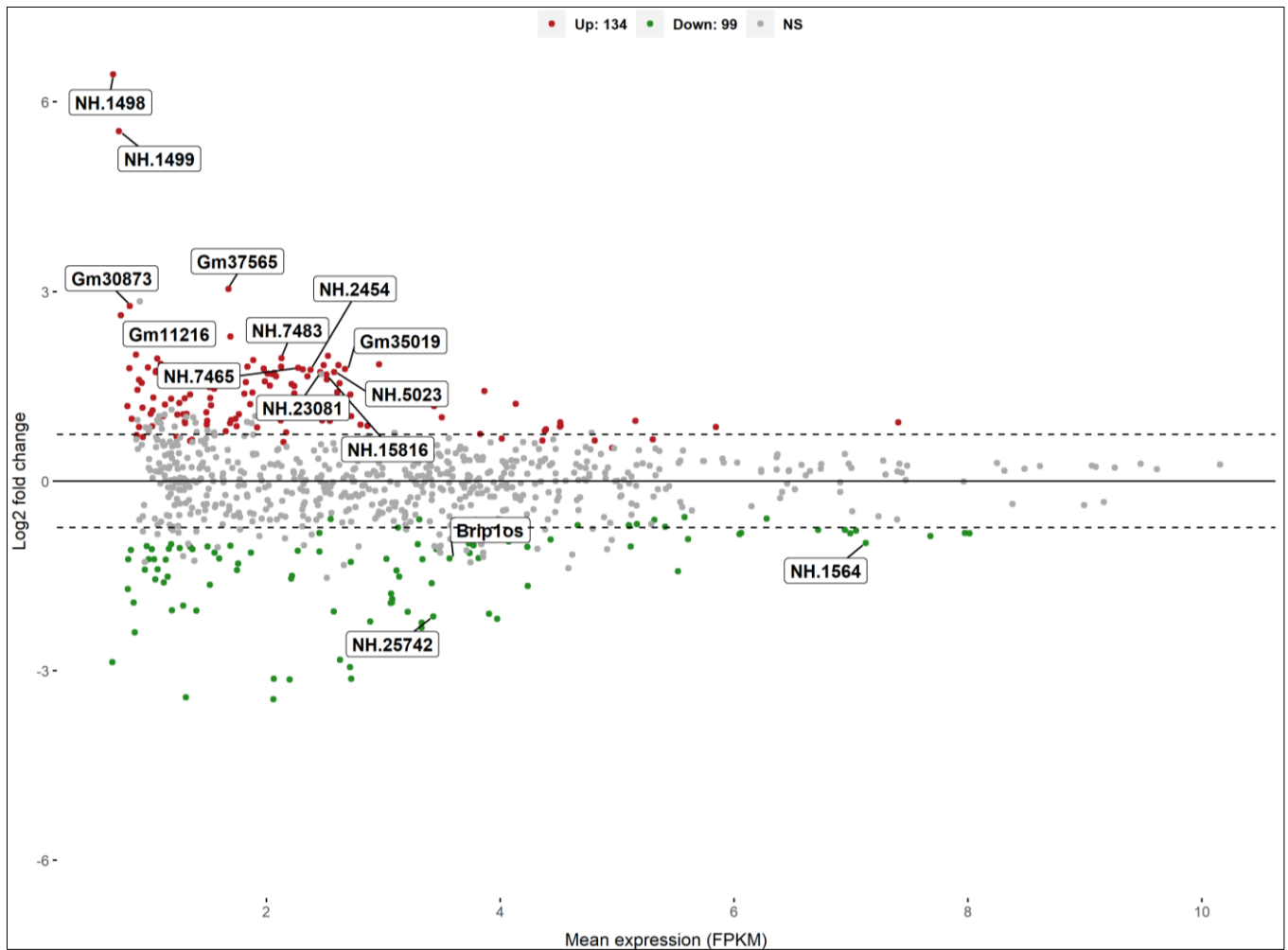


Figure 12. MA plot showing the distribution of differentially expressed lncRNAs (log2 fold change threshold ± 0.6 , FDR threshold < 0.05 ; NS: Not Significant, FPKM: Fragments per kilobase of transcript per million mapped reads).

Figure 11 shows the distribution of differentially expressed lncRNAs. The plot shows us that there are 861 annotated lncRNAs therefore 134 of them are upregulated and 99 of them are downregulated significantly. Note that grey colored circles in MA plot as shown in Figure 12 represent DEGs that remain above 0.05. Top 15 significant DEGs are labelled and 7 of them are seen to be in the class of previously un-annotated lncRNAs.

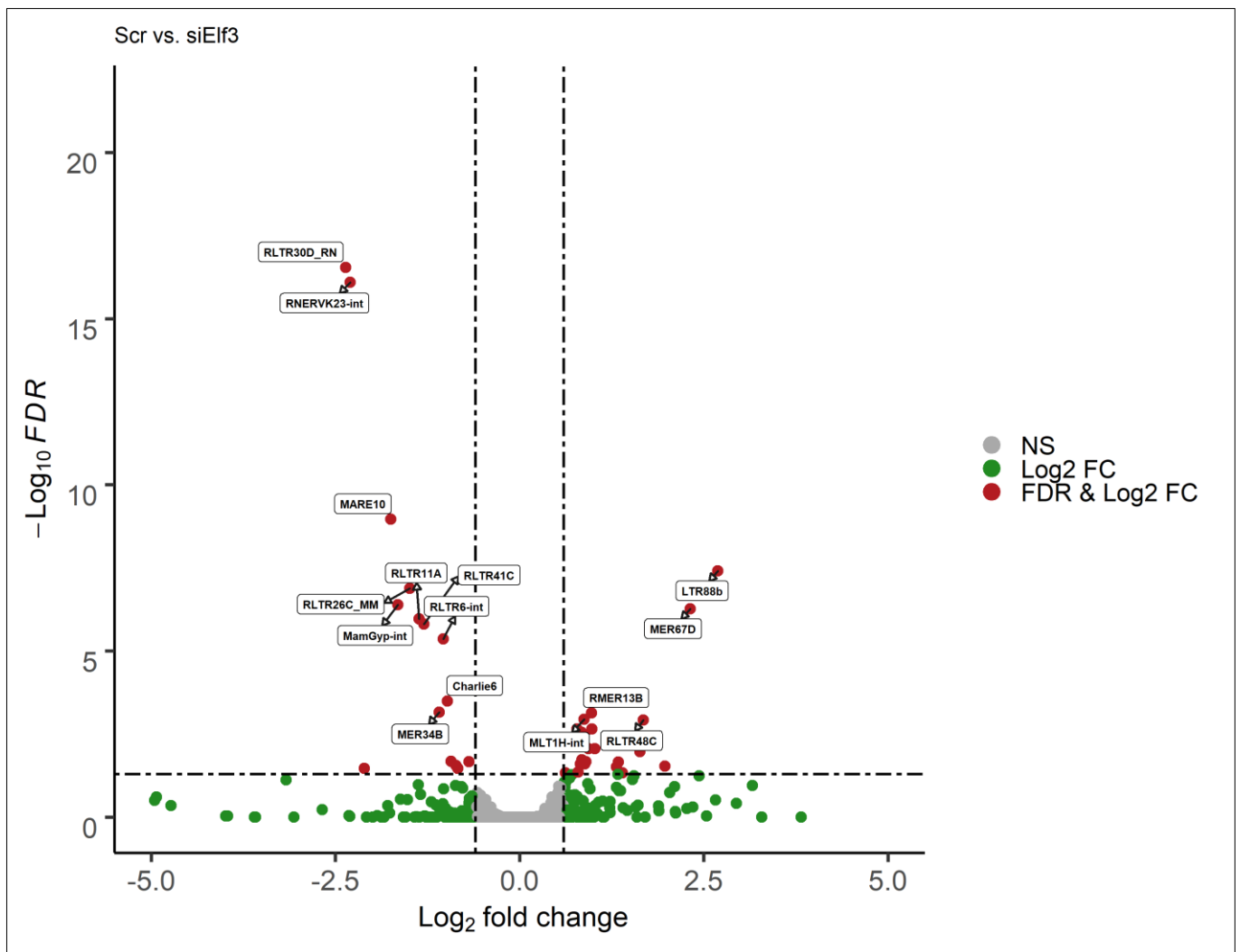


Figure 13. Volcano plot showing the distribution of differentially expressed repeat elements. (NS: Not Significant). (NS: Not Significant, log2FC: log2 Fold Change, FDR: False Discovery Rate).

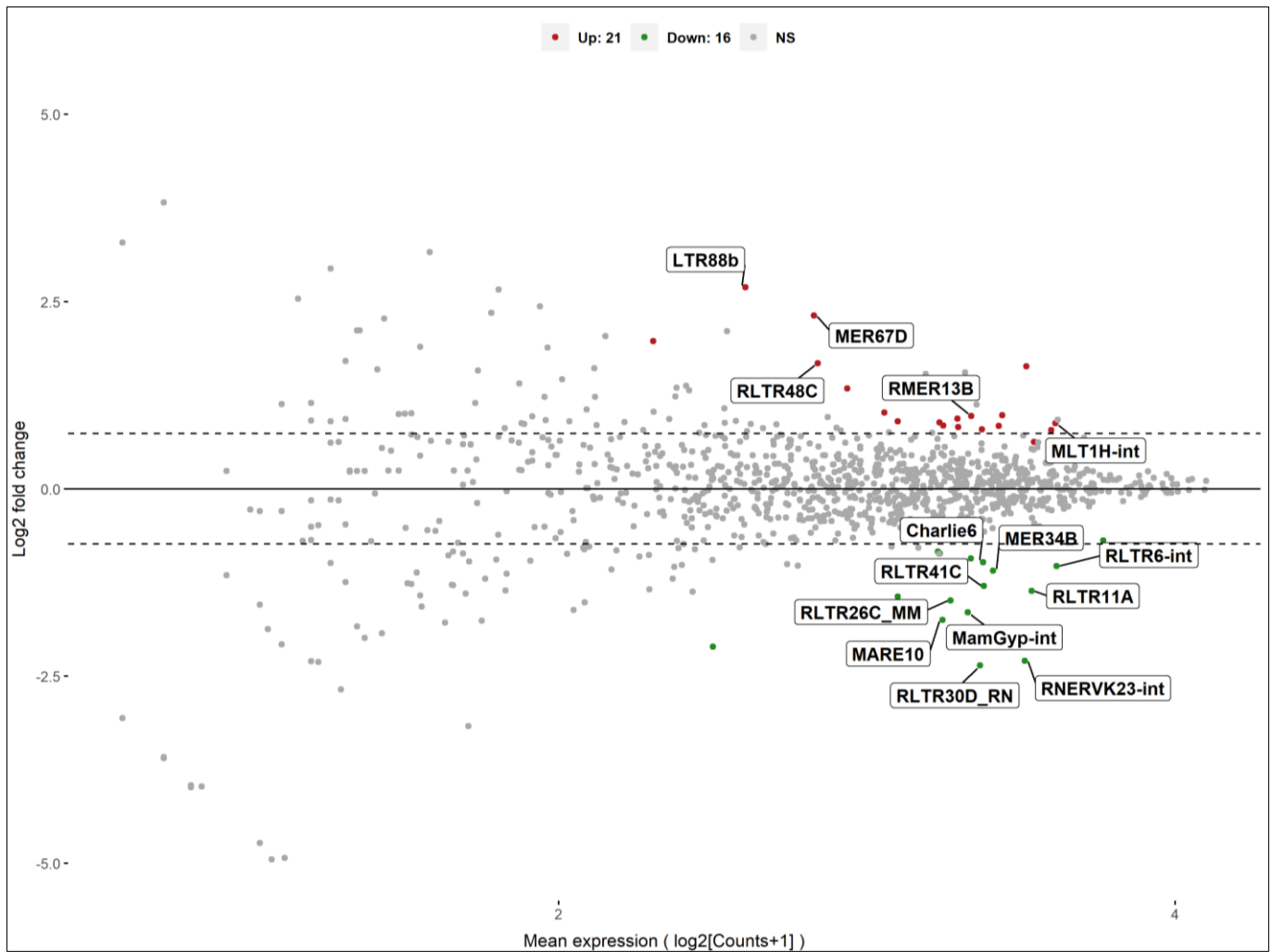


Figure 14. MA plot showing the distribution of differentially expressed repeat with log2 fold change and mean expression levels in all samples (log2 fold change threshold ± 0.6 , FDR threshold < 0.05 , NS: Not Significant).

Volcano plot Figure 13 shows the overall distribution of repeat elements belonging to LINE, SINE, DNA, LTR and Satellite families. It shows 1117 repeat elements that are differentially expressed between conditions. However, according to the MA plot (Figure 14), only 37 repeat elements are significantly regulated, 21 of them being upregulated while 16 of them being downregulated. Top fifteen significantly regulated repeats are labelled containing LTR88b and MER67D that are included in LTR and DNA transposon families, respectively.

4.4. Overrepresentation test of Gene Ontology terms between conditions

13439 DEGs that have smaller than 0.05 FDR value are selected for GO enrichment analysis. 3055 significant DEGs resulted in 352 GO terms for Biological Process (BP), 99 terms for Cellular Component (CC) and 52 terms for Molecular Function (MF), each having the significance below 0.05 for p-adjust value. In order to decrease the redundancy, these terms are simplified using simplify() function ClusterProfiler based on the distance of the terms within the GO hierarchy. Top 20 significant terms are shown in Figure 15, Figure 17 and Figure 19 for BP, MF and CC categories, respectively.

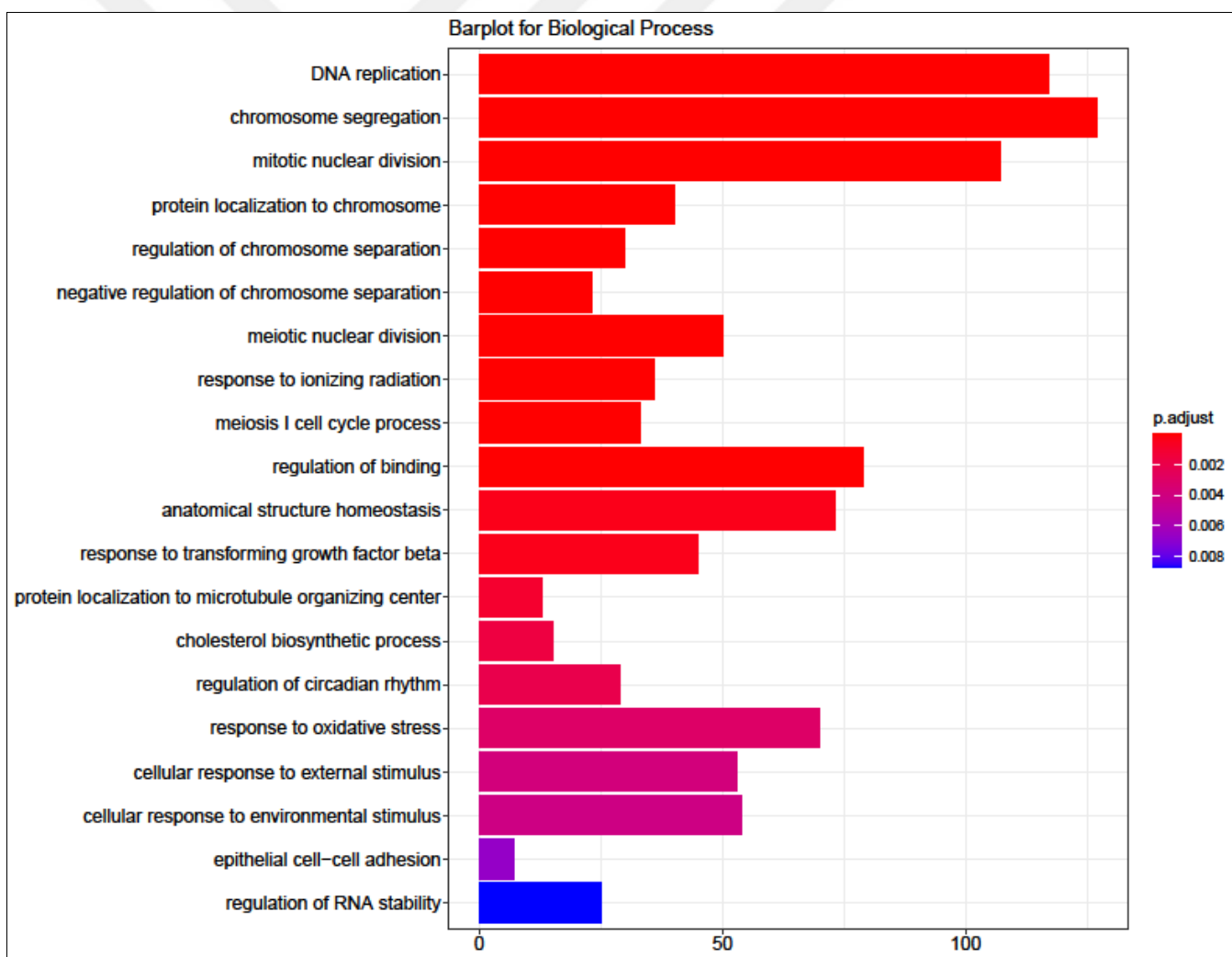


Figure 15. Barplot for top 20 significant simplified GO terms that are enriched for Biological Process. X-axis refers to the number of input genes that are enriched in each category.

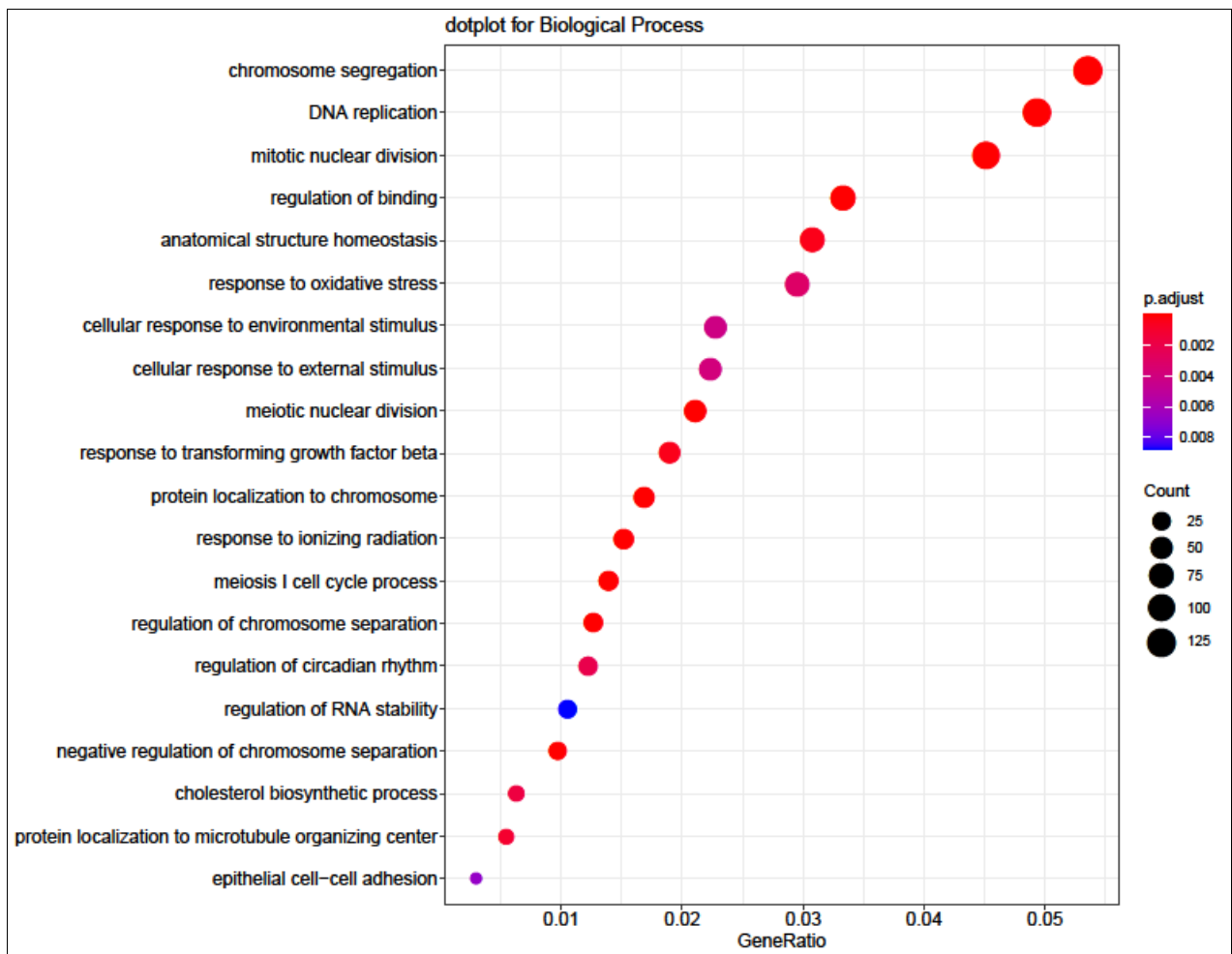


Figure 16. Dotplot for top 20 significant simplified GO terms that are enriched for Biological Process. (GeneRatio=number of input genes that are enriched in the category / total number of input genes that are enriched).

According to Figure 15, DNA replication, chromosome segregation and mitotic nuclear division terms are top three biological processes that are known to be associated with cell cycle. The rest of the terms may be an indicative of a response to a stimulus, siElf4 knockdown in this case. Other enriched terms include terms that are associated with meiosis, protein localization and chromosome separation. The dot plot in Figure 16 confirms the finding in bar plot as the top three GO terms with the highest GeneRatio are the same.

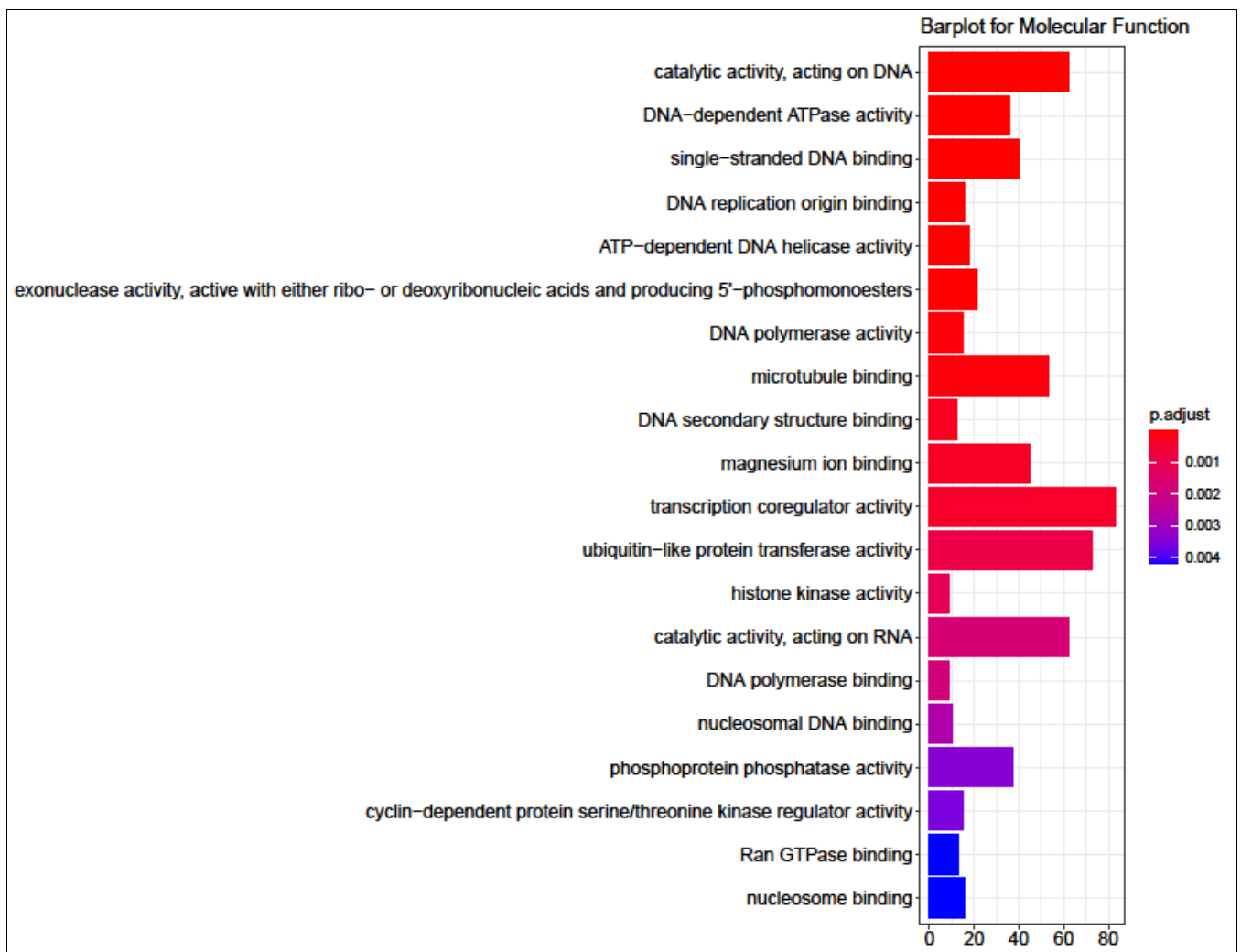


Figure 17. Barplot for top 20 significant simplified GO terms that are enriched for Molecular Function. X-axis refers to the number of input genes that are enriched in each category.

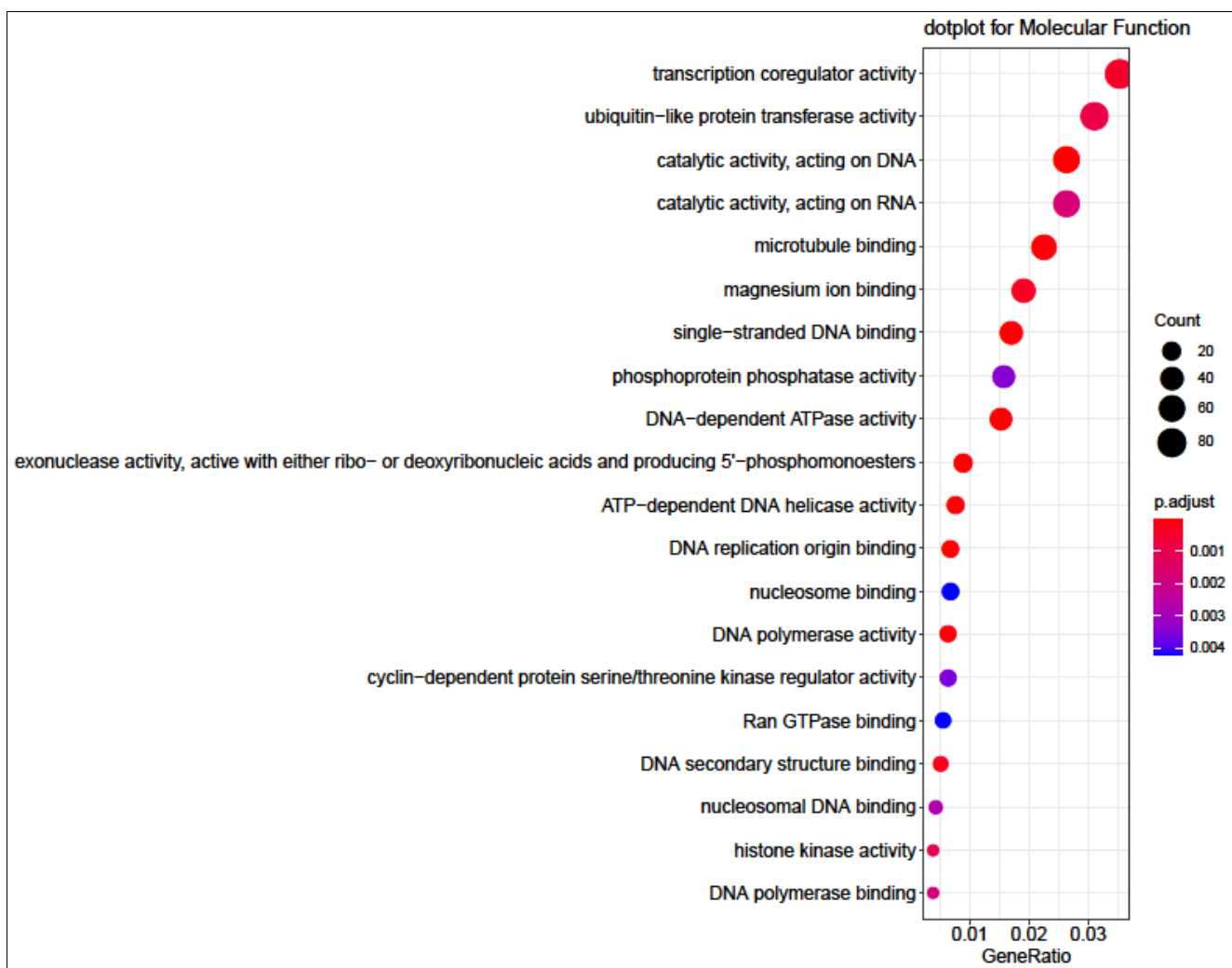


Figure 18. Dotplot for top 20 significant simplified GO terms that are enriched for Molecular Function. (GeneRatio= number of input genes that are enriched in the category / total number of input genes that are enriched).

According to Figure 17, top molecular functions of enriched terms are transcription coregulator activity, ubiquitin like transferase activity and catalytic activity on DNA or RNA. Other enriched terms include microtubule binding, exonuclease activity and single-stranded DNA binding. These findings are also supported by findings shown in dot plot given in Figure 18.

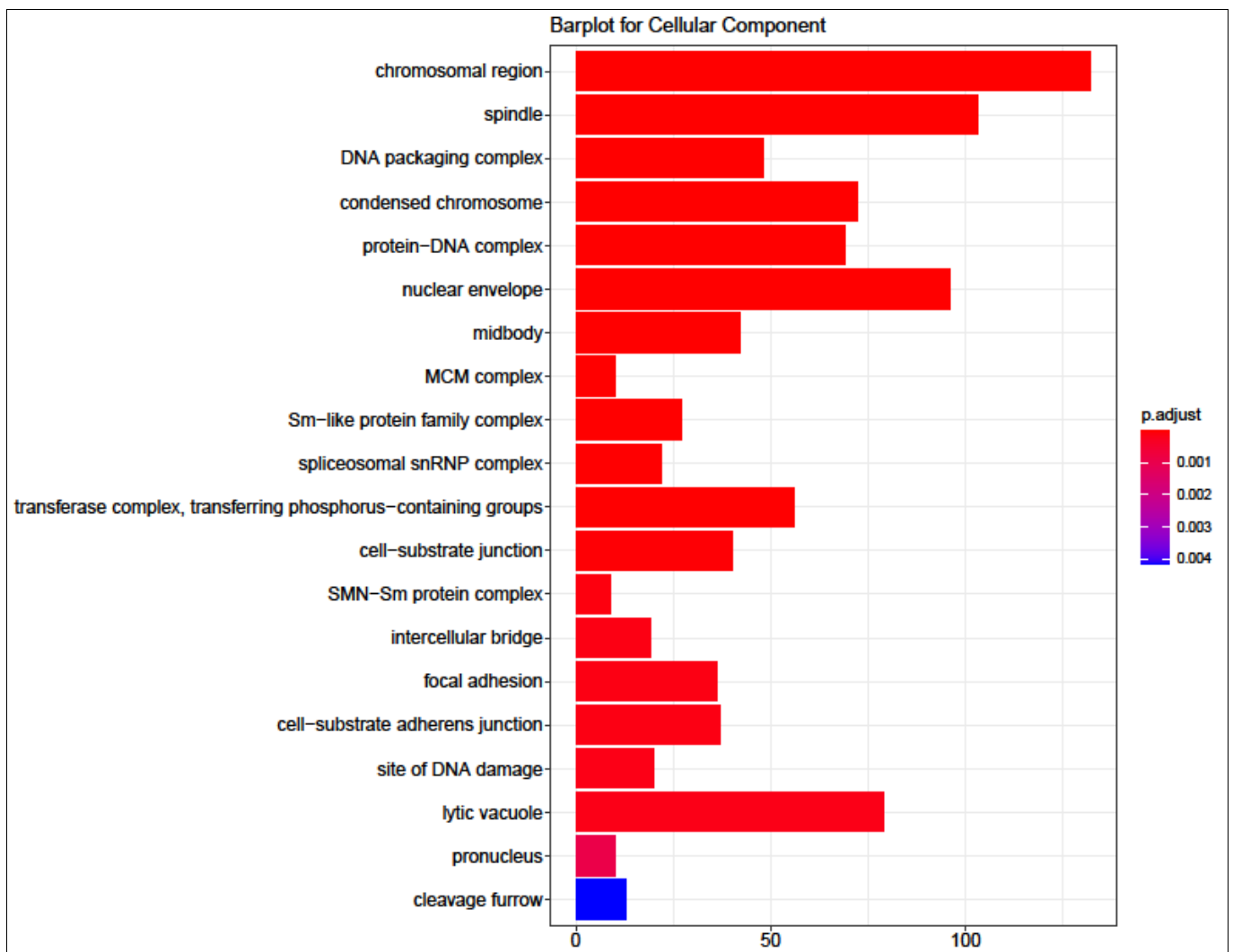


Figure 19. Barplot for top 20 significant simplified GO terms that are enriched for Cellular Component. X-axis refers to the number of input genes that are enriched in each category.

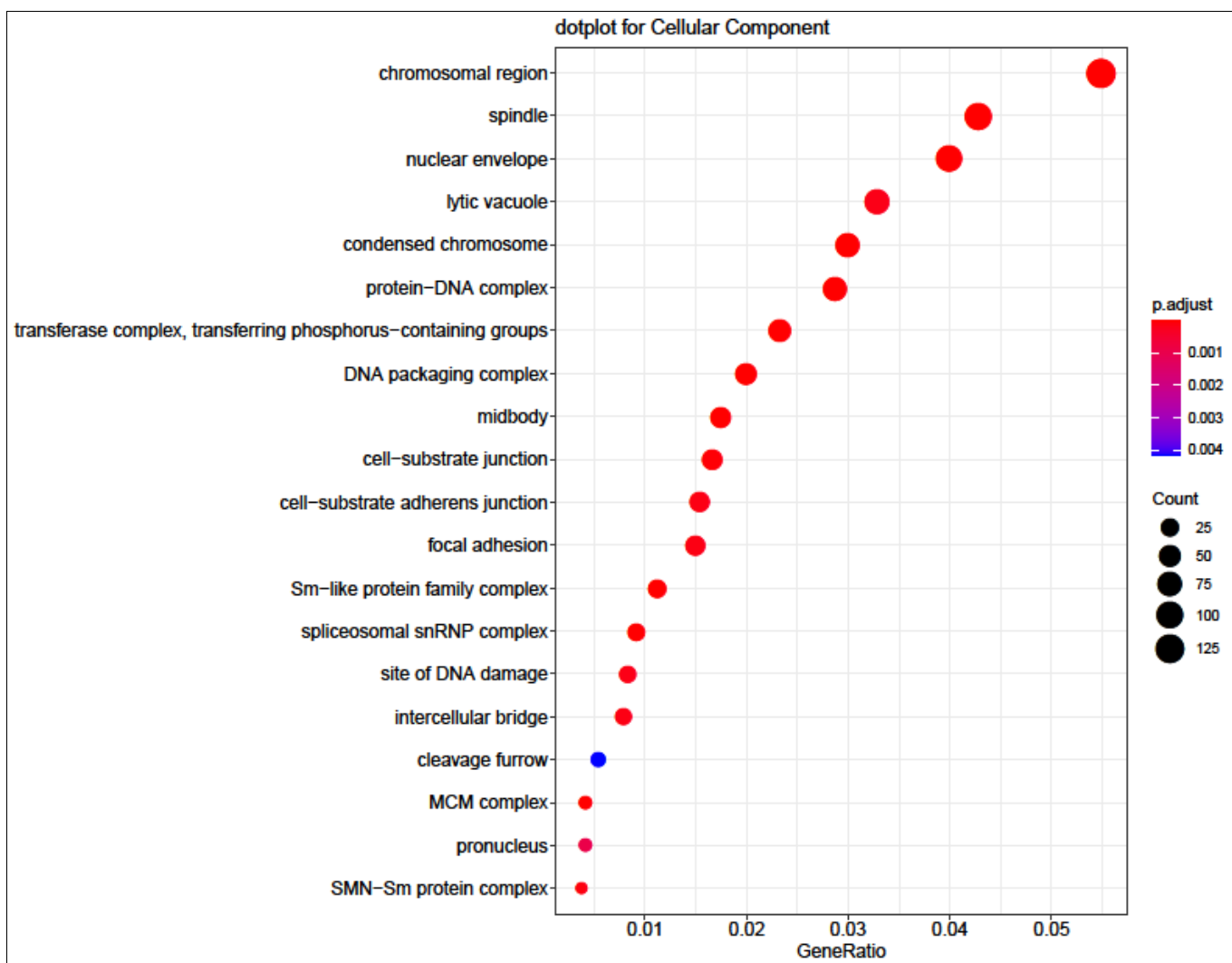


Figure 20. Dotplot for top 20 significant simplified GO terms that are enriched for Cellular Component. (GeneRatio= number of input genes that are enriched in the category / total number of input genes that are enriched).

According to Figure 19, top three enriched terms for Cellular Component are chromosomal region, spindle, and nuclear envelope. Other CC terms include protein-DNA complex, DNA packaging complex and lytic value which are also given in the dot plot shown in Figure 20.

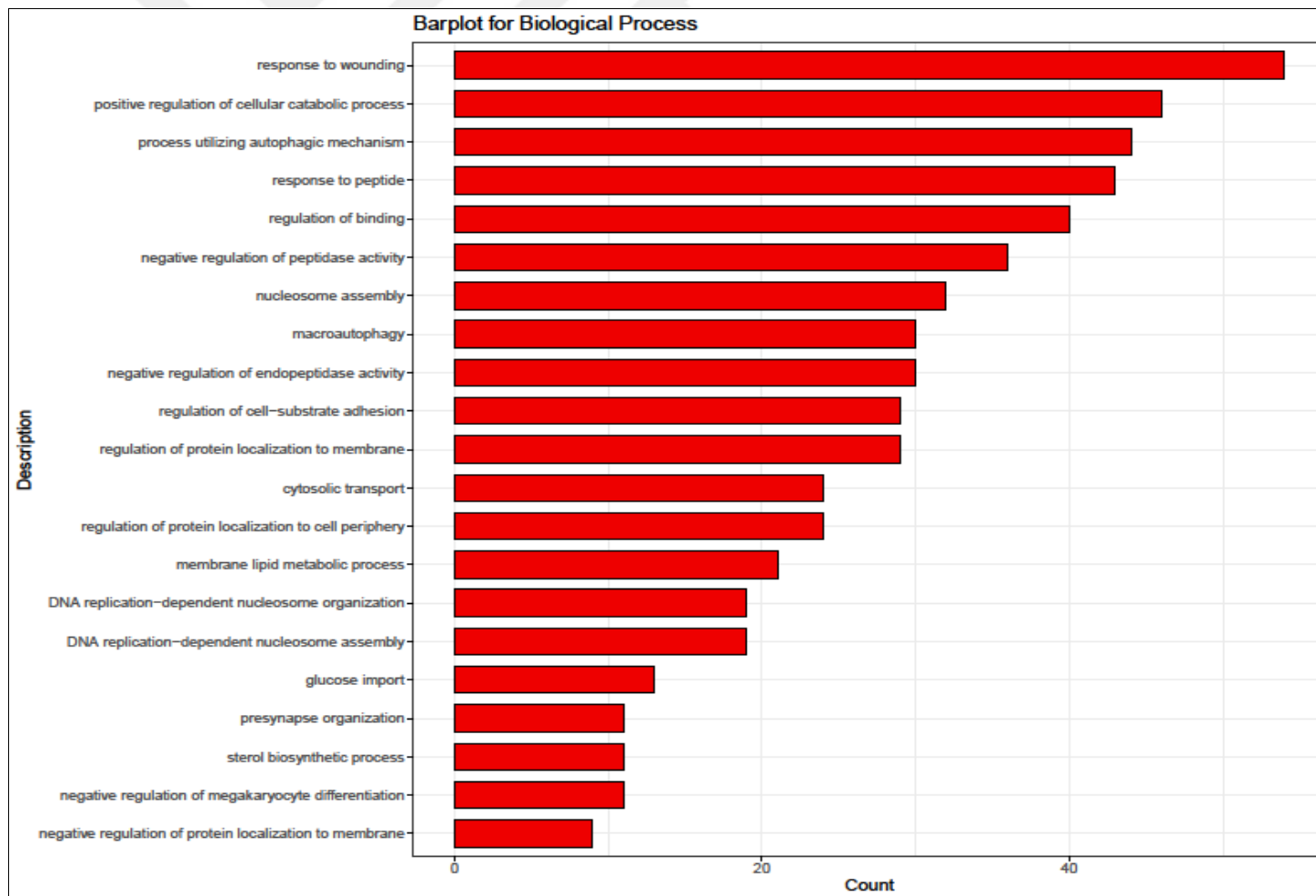


Figure 21. Barplot showing GO enrichment of 1354 up regulated genes in Biological Process (P-adjust < 0.05).

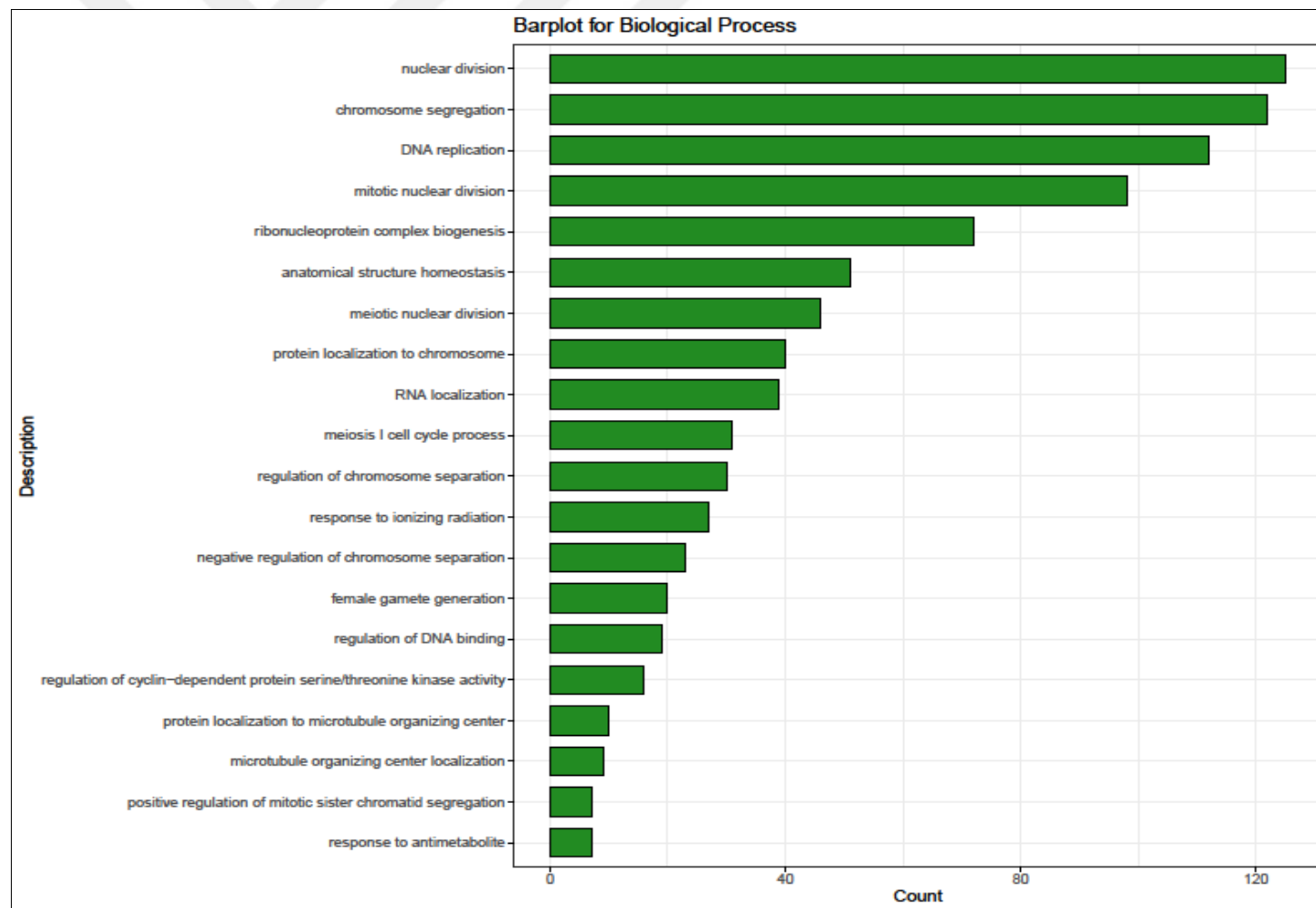


Figure 22. Barplot showing GO enrichment of 1701 down regulated genes in Biological Process (P-adjust < 0.05).

While the previous barplots show GO enrichment of all genes, Figure 21 and Figure 22 shows GO enrichment belonging to only up or down regulated genes. In this case, it can be distinguished that leading categories stay the same. For up regulated genes, top three categories are response to wound healing, positive regulation of cellular catabolic process and process regulating autophagy mechanism. On the other hand, for down regulated genes, top three categories do not differ and have the same cell cycle associated terms which are nuclear division, chromosome segregation and DNA replication. The remainder of the categories are nearly similar as before, especially in the case of down regulated genes. However, different types of categories also appear, such as ribonucleoprotein complex biogenesis or cytosolic transport.

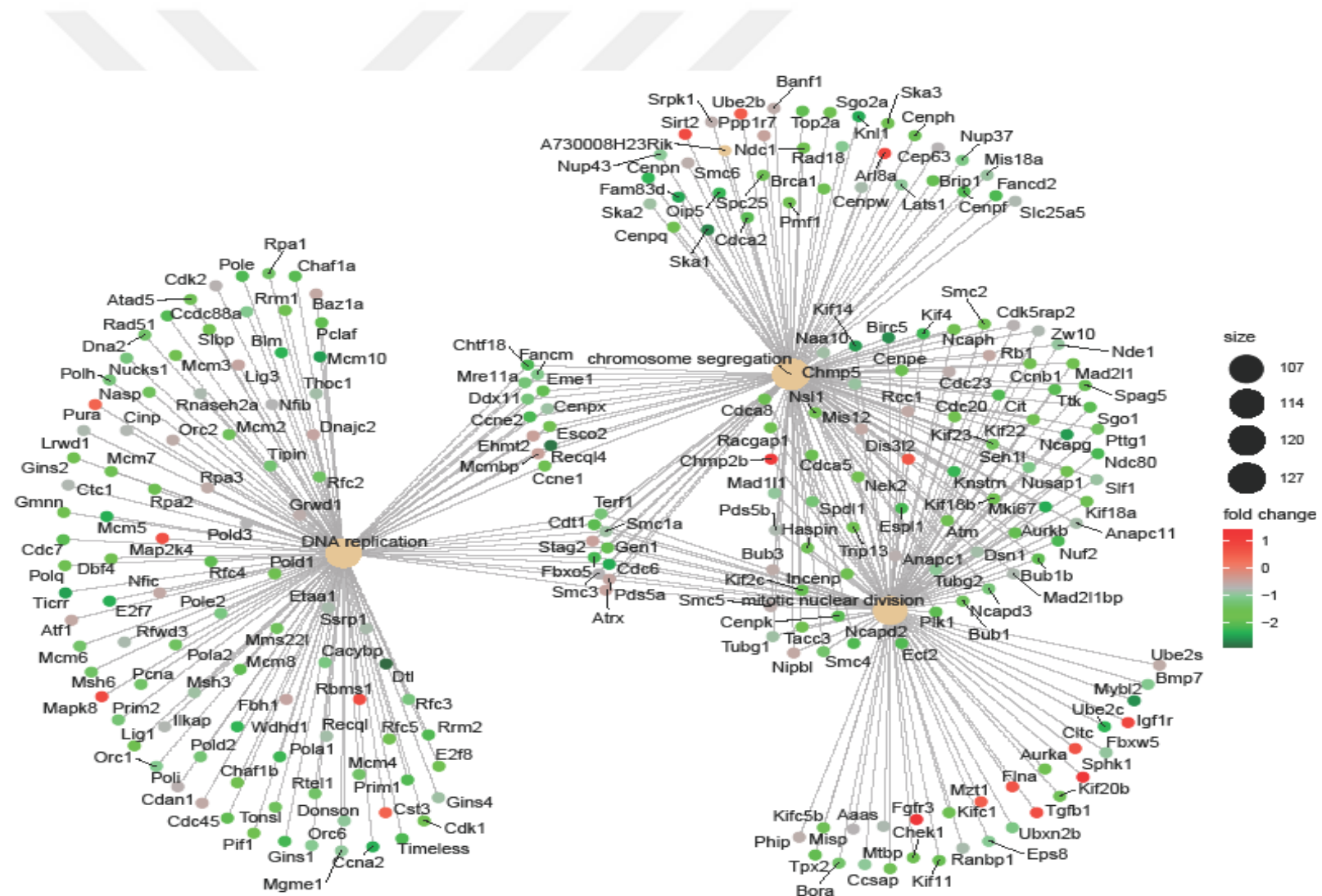


Figure 23. Network plot of top three enriched terms in biological process.

In Figure 23, a gene-concept network is depicted for enriched genes in biological process that are contained in terms of DNA replication, chromosome segregation and mitotic nuclear division. It can be inferred from figure that there is a vast number of genes that are mutually shared among terms. The plot also shows in general that most of the genes are down regulated. Another gene-concept network in Figure 24 is depicted for enrichment of genes in cellular component terms. Similarly, large number of genes are shared between spindle and chromosomal region.

4.5. KEGG pathway enrichment analysis

Differential expression analysis yielded 3055 differentially expressed genes. 2584 of them had absolute logFC above 0.6 and those genes were subjected to overrepresentation test of KEGG pathway enrichment. Based on our analysis results, 26 different KEGG pathways are significantly enriched including DNA replication, cell cycle and mismatch repair pathways (Table 2).

Table 2. KEGG pathway enrichment analysis result (GeneRatio: number of input genes that are enriched in the category / total number of input genes that are enriched, BgRatio: total number of genes in that pathway/total number of genes in all pathways)

ID	Description	GeneRatio	BgRatio	pvalue	p.adjust	Count
mmu03030	DNA replication	27/876	35/8774	7.13E-21	2.20E-18	27
mmu04110	Cell cycle	48/876	123/8774	7.30E-18	1.13E-15	48
mmu03430	Mismatch repair	15/876	22/8774	7.60E-11	7.82E-09	15
mmu04114	Oocyte meiosis	32/876	119/8774	1.10E-07	6.44E-06	32
mmu04142	Lysosome	34/876	131/8774	1.14E-07	6.44E-06	34
mmu03440	Homologous recombination	17/876	41/8774	1.25E-07	6.44E-06	17
mmu03460	Fanconi anemia pathway	19/876	51/8774	1.75E-07	7.72E-06	19

mmu03420	Nucleotide excision repair		16/876	43/8774	1.68E-06	6.50E-05	16
mmu04115	p53 signaling pathway		20/876	71/8774	1.25E-05	0.00043	20
mmu00240	Pyrimidine metabolism		17/876	58/8774	3.17E-05	0.000939	17
mmu04510	Focal adhesion		39/876	201/8774	3.34E-05	0.000939	39
mmu04914	Progesterone-mediated oocyte maturation		22/876	90/8774	5.33E-05	0.001366	22
mmu00100	Steroid biosynthesis		9/876	20/8774	5.75E-05	0.001366	9
mmu04218	Cellular senescence		36/876	185/8774	6.24E-05	0.001378	36
mmu05166	Human T-cell leukemia virus 1 infection		44/876	246/8774	8.36E-05	0.001722	44
mmu03040	Spliceosome		30/876	150/8774	0.000151	0.00291	30
mmu03013	RNA transport		35/876	191/8774	0.00027	0.004909	35
mmu03410	Base excision repair		11/876	34/8774	0.000305	0.005233	11
mmu00983	Drug metabolism - other enzymes		20/876	88/8774	0.000331	0.005385	20
mmu04068	FoxO signaling pathway		26/876	131/8774	0.000461	0.007119	26
mmu00531	Glycosaminoglycan degradation		8/876	21/8774	0.000594	0.008747	8
mmu01524	Platinum drug resistance		17/876	78/8774	0.001486	0.020879	17

mmu01523	Antifolate resistance	9/876	30/8774	0.001957	0.026292	9
mmu00480	Glutathione metabolism	15/876	67/8774	0.002089	0.0269	15
mmu04392	Hippo signaling pathway - multiple species	8/876	27/8774	0.003774	0.046003	8
mmu05222	Small cell lung cancer	18/876	92/8774	0.003871	0.046003	18

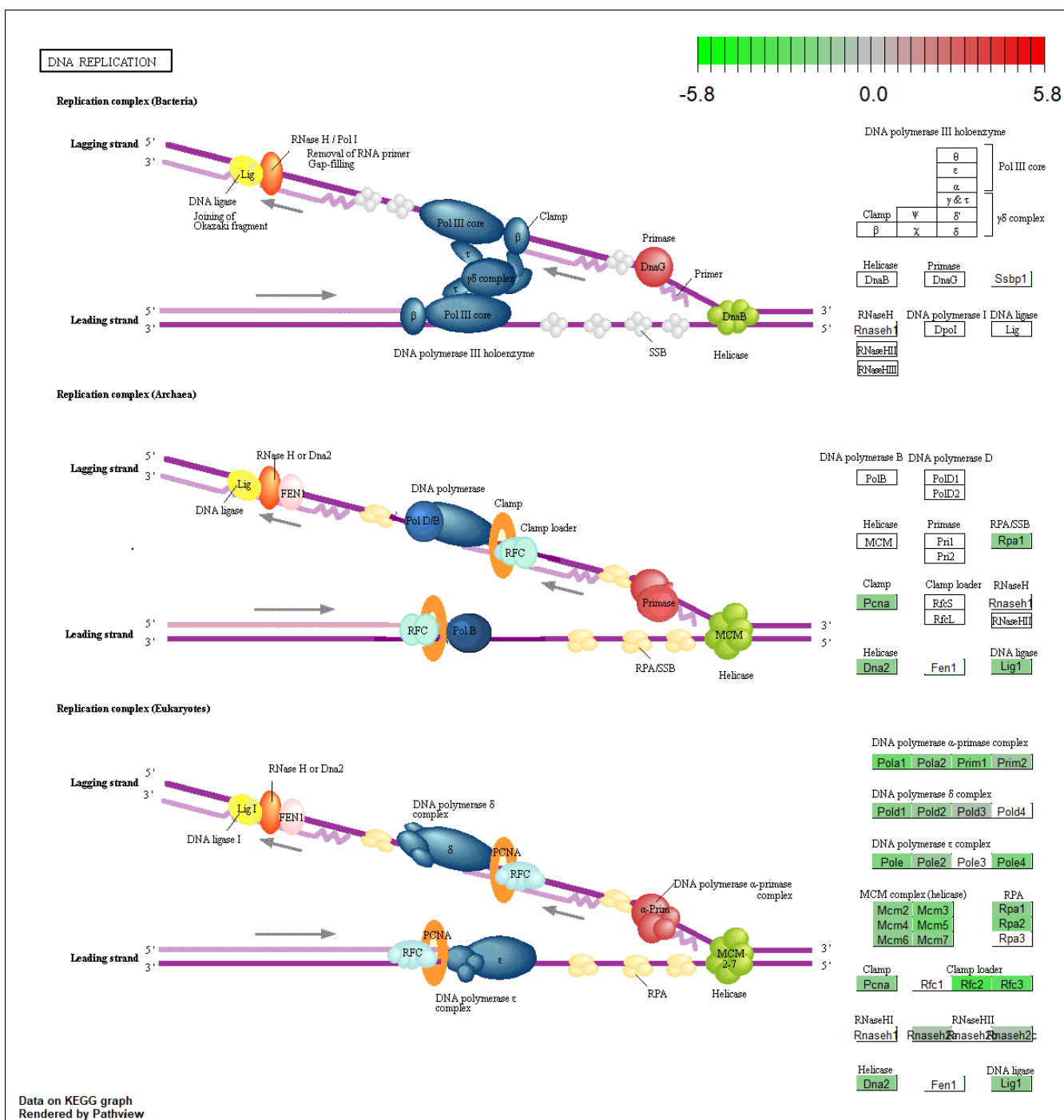


Figure 25. DNA replication pathway highlighted with enriched genes; color scale represents logFC distribution.

Figure 25 shows the DNA replication pathway that is highlighted with enriched genes. It can be inferred from Table 2 that DNA replication pathway contains 35 genes in total and 27 of them are found in our dataset. Moreover, Figure 26 shows the cell cycle pathway that contains 123 genes in total and 48 of them are found in our dataset. Both figures clearly demonstrate that the majority of DEGs are down regulated except a few noticeable genes including Tgfb1, Cdkn1a and Ywhag that are found in cell cycle pathway.

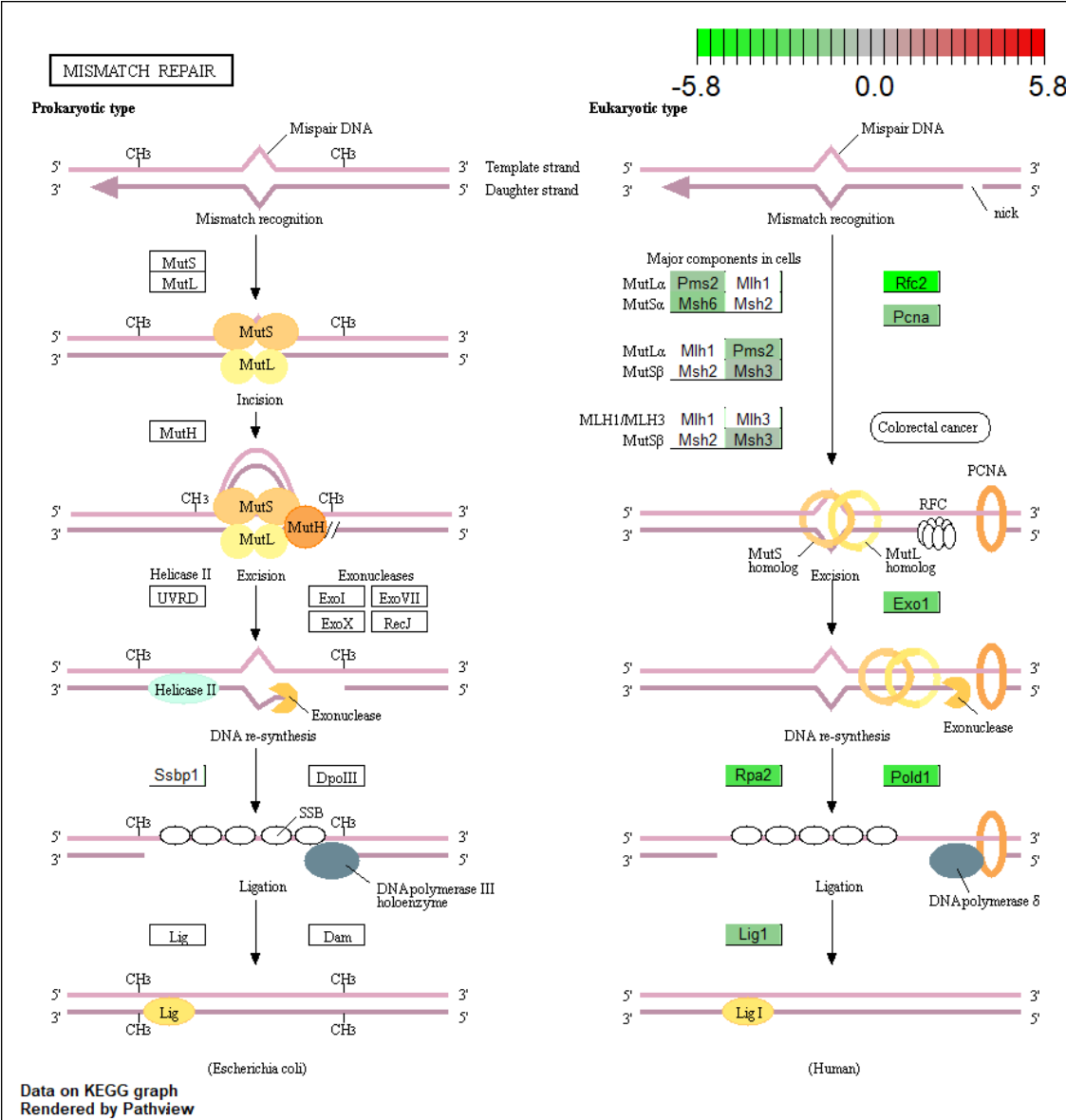


Figure 27. Mismatch repair pathway highlighted with enriched genes; color scale represents logFC distribution.

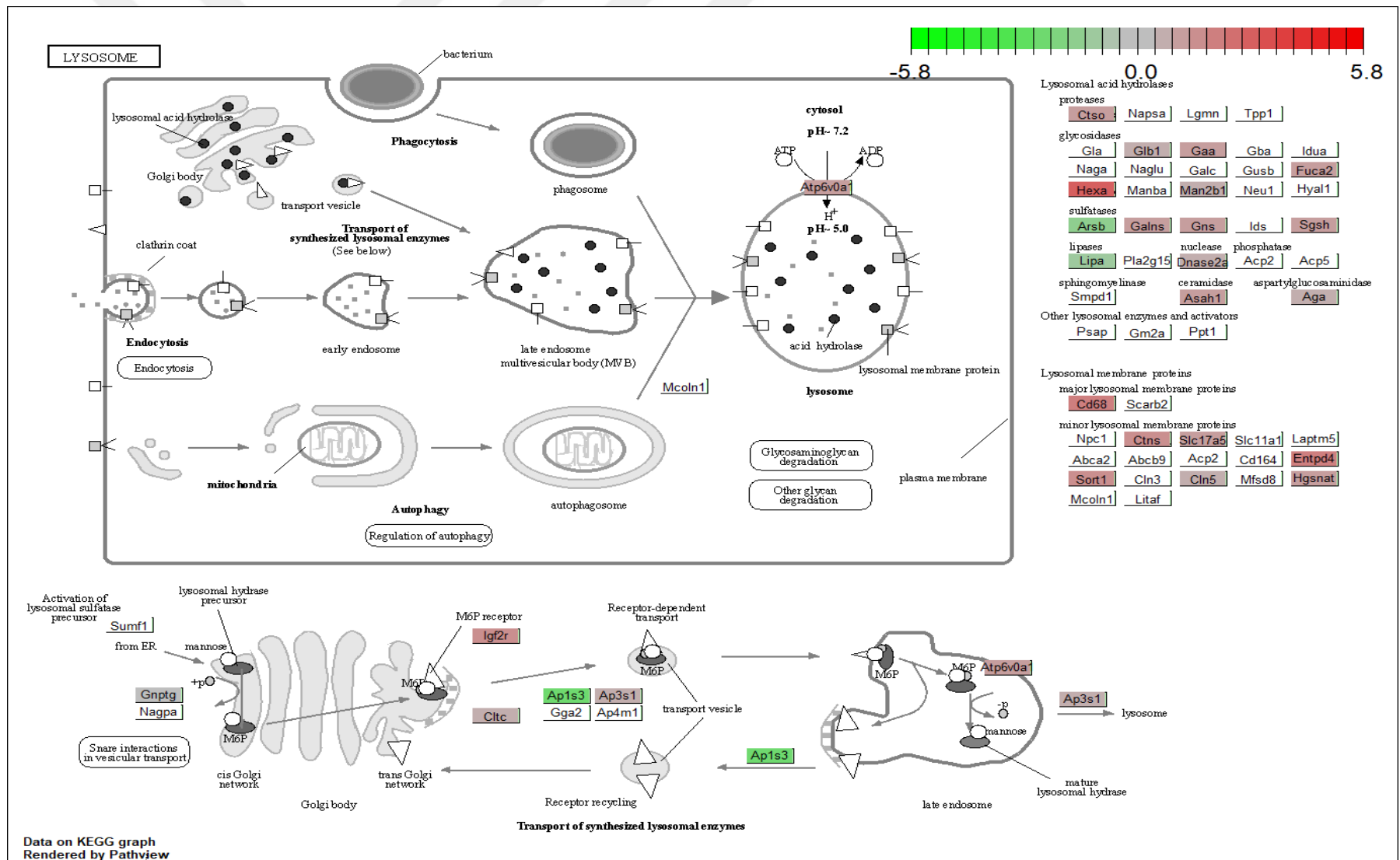


Figure 28. Lysosome pathway highlighted with enriched genes; color scale represents logFC distribution.

KEGG pathway enrichment resulted in the enrichment of two other pathways which are Mismatch Repair and Lysosome as shown in Figure 27 and Figure 28, respectively. Out of 22 genes in mismatch repair, 15 input genes are enriched. Similarly, most of them are consisting of down regulated genes. For lysosome pathway, out of 131 genes defined in the pathway, 34 input genes are enriched. On the contrary to finding in previous pathways, downregulated genes are dominated by upregulated ones in the lysosome pathway.

5. DISCUSSION

Mesenchymal to epithelial transition is a crucial biological process that is in the forefront of various physiological events such as development, tissue regeneration and cancer progression. MET is thought to be the reverse process of EMT, nevertheless very little is known about the molecular regulation of MET. Many players and molecular effectors of MET have been shown that includes Grhl3, Hnf4 and Elf3. E74 Like ETS Transcription Factor 3 is a transcription factor that is known to be essential in mesenchymal to epithelial transition (Sengez et al., 2019). Some of these players have prominent roles in a proper MET event by interacting with Cdh1 that codes for E-cadherin in mice. For example, Hnf4a and Grhl3 act on intron 2 of Cdh1 thus activating the expression of E-cadherin. Moreover, depletion of Elf3 resulted in the absence of E-cadherin at the plasma membrane. Likewise, Elf3 has been also shown to activate the expression of Hnf4a in the same study.

In this project, transcriptome dynamics of NMuMg cells in the absence of Elf3 during mesenchymal to epithelial transition has been studied. For that purpose, Elf3 was knocked-down in PT72 NMuMg cells that were treated with TGF- β beforehand. Transcriptome profiling of siElf3 and Scr shows the explicit separation of siElf3 and Scr samples in the correlation heatmap (Fig. 7).

Transcriptome is studied in the context of protein coding genes, long noncoding RNAs and repeat elements. Long noncoding RNAs are usually expressed under certain conditions and in specific tissues and they have important contributions in cellular processes. One such contribution is accounted in EMT processes involving regulation of some genes. Long noncoding RNAs are known to regulate the expression of certain genes by recruiting Polycomb Repressive Complex 2 (PRC2) to associated sites on chromatin (Beckedorff et al., 2013; Kotake et al., 2011; Pandey et

al., 2008) whereas they may lead to retain translation of mRNAs by preventing them to be suppressed by miRNAs, acting as some sort of molecular sponges (Ge et al., 2019; H. et al., 2019). Although the role of lncRNAs in EMT is appreciated, knowledge of lncRNAs in MET has not been very well studied so far. Investigation of lncRNAs are included in this study to give insights about the regulation of Elf3. For that purpose, ribosomal RNA depletion method was used as library preparation to capture lncRNAs in addition to mRNAs in total transcriptome. Boxplot in Figure 8 shows the expression of lncRNAs come after protein coding genes having a very similar distribution in all samples. It should be noted that transcript quantification was performed using a custom GTF file that contains previously un-annotated lncRNAs that were annotated in a previous study of lab that contains time-series experiments of MET process in NMuMg cells. However, this study has not been published yet. LncRNAs that begin with “NH.” prefix refer to those transcripts.

Differential expression analysis between Scr and siElf3 revealed many differentially expressed genes that codes for both proteins and lncRNAs. Elf3 as a protein coding gene is labelled in both volcano plot (Figure 9) and MA plot (Figure 10) having more than 8-fold downregulation which indicates a successful knockdown of the silencing experiment. Vast amount of lncRNAs are shown to be differentially expressed due to Elf3 knockdown (Figure 11 and Figure 12). Therefore, the prevalence of previously un-annotated lncRNAs among all differentially expressed noncoding genes is an interested finding.

Repeat elements are known to have contribution in gene regulation and they are included in this study to identify their contribution in transcriptome dynamics. LINE, SINE, LTR, DNA transposons and Satellites were quantified using the RepeatMasker annotation. Differential expression analysis of repeats between Scr and siElf3 samples yielded 37 repeats that are differentially expressed upon Elf3 knockdown (Figure 13). Majority of these repeats are included in LTR (e.g. RLTR30D_RN, LTR88, RNERVK23-int) and DNA transposon (e.g. MARE10, MER67D, Charlie6) families. However, CRP1 expression from SINE family of repeats have a 2.47-fold increase and L1MC1 expression from LINE family 1.71-fold increase upon Elf3 knockdown. Alu elements are included in SINE family of repeats and they were shown to be associated with cancer progression by inducing EMT (Di Ruocco et al., 2018). There are very limited number of studies on the contribution of repetitive elements on EMT nor MET. Thus, this

study also gives insights about their potential on gene regulation in MET and it may pave the way to further studies.

GO enrichment analysis of all differentially expressed genes ($FDR < 0.05$) resulted in the enrichment of DNA replication, chromosome segregation and mitotic nuclear division that are top three enriched terms of biological process (Figure 15). According to enrichment in the category of molecular function (Figure 17), transcription coregulator activity, ubiquitin like transferase activity and catalytic activity on DNA or RNA may direct us to possible functions of input DEGs. Therefore, they are also enriched in chromosomal region, spindle, and nuclear envelope as part of cellular component category (Figure 19). GO enrichment of only down regulated genes again results in the enrichment of DNA replication, chromosome segregation and mitotic nuclear division, to appear as top three enriched biological terms (Figure 22). On the other hand, up regulated genes enriched with response to wound healing, positive regulation of cellular catabolic process and process regulating autophagy mechanism (Figure 21). Network plot (Figure 23) of top three enriched terms demonstrates the prevalence of down regulated genes within the enriched categories which confirms the previous finding. It is also worthy of note that the categories share several genes.

KEGG pathway enrichment analysis confirms the finding of GO enrichment analysis. Top two enriched KEGG pathways are DNA replication, cell cycle and mismatch repair (Table 2) which are then followed by several discrete pathways as lysosome, homologous recombination, spliceosome, and small cell lung cancer. Figure 25 illustrates the DNA replication pathway that are enriched with 27 input genes which is quite high considering that the pathway has 35 genes in total. Note that nearly all genes in this pathway are down regulated which is also likewise with cell cycle pathway (Figure 26). Cell cycle pathway is distinctly highlighted with upregulated genes as *Tgfb1*, *Cdkn1a* and *Ywhag*. *Cdkn1a*, Cyclin dependent kinase inhibitor 1a, acts as an inhibitor of cyclin dependent kinases. *Cdkn1a* is also known as p21 and it is known to be a regulator of p53 mediated inhibition of cellular proliferation. Cell cycle pathway demonstrates that upregulation of p21 leads to downregulation Cyclin A1 (*Ccna1*) and Cyclin E1(*Ccne1*) proteins since p21 is shown to inhibit them both in the pathway.

This study illuminates us regarding the transcriptome dynamics that are regulated by Elf3. It directs us to the understanding that Elf3 may be necessary in cell cycle, particularly mitotic nuclear division, and its key events. This finding is supported by GO and KEGG enrichment analyses, however this study needs experimental validations in order to confirm these findings. The major drawback of this study is the sample size. Larger sample size would increase the strength of this study by means of extending its scope and its statistical power. A correlation analysis between expressions of genes and repeat elements would be performed and it would be quite informative to suggest a link about an influence of repeats on gene expressions. Another experiment would be to look for potential Elf3 binding sites if RNA-seq was coupled with a ChIP-Seq technology. Large sample size is also helpful for a wider range of visualizations especially for the projection of gene expressions. Our findings should definitely be confirmed with further wet-lab practices and a larger sample size would lead to supportive conclusions especially if coupled with alternative technologies.

6. CONCLUSION AND FUTURE ASPECTS

In this study, the role of Elf3 in MET regulation was shown by means of altering transcriptome dynamics. It was shown that the absence of Elf3 led to changes in protein-coding and long noncoding RNA gene expression as well as repeat expression using one dataset. Several protein-coding genes and long noncoding RNAs are differentially expressed in response to Elf3 knock-down. Long noncoding RNAs are known to regulate EMT events, yet this study aims to emphasize on their importance in MET regulation. In terms of repeat expression, DNA transposons and LTRs were found to be majorly affected from the knock-down. A few numbers of LINEs and SINEs were also affected. This may point out to the importance of repeat expression for the regulation of MET. GO and KEGG pathway enrichment analyses of differentially expressed genes concentrates on the fact that Elf3 may be important for signaling pathways involved in cell cycle, particularly mitotic nuclear division, and chromosomal segregation. In future researches, the number of replicates should be increased to dive into deeper understanding of Elf3 regulation. In addition, as Elf3 is a transcription factor, other NGS technologies such as ChIP-Seq should be coupled to investigate the place of Elf3 and other transcription factors in MET regulation in future studies.

7. REFERENCES

- Ahmed, N., Thompson, E. W., & Quinn, M. A. (2007). Epithelial–mesenchymal interconversions in normal ovarian surface epithelium and ovarian carcinomas: An exception to the norm. *Journal of Cellular Physiology*, 213(3), 581–588. <https://doi.org/10.1002/jcp.21240>
- Alberga, A., Boulay, J. L., Kempe, E., Dennefeld, C., & Haenlin, M. (1991). The snail gene required for mesoderm formation in *Drosophila* is expressed dynamically in derivatives of all three germ layers. *Development*.
- Alberts et al, B. (2014). The Molecular Biology of the Cell. In *Molecular Biology of the Cell*. <https://doi.org/10.1091/mbc.E14-10-1437>
- Alotaibi, H., Basilicata, M. F., Shehwana, H., Kosowan, T., Schreck, I., Braeutigam, C., Konu, O., Brabletz, T., & Stemmler, M. P. (2015). Enhancer cooperativity as a novel mechanism underlying the transcriptional regulation of E-cadherin during mesenchymal to epithelial transition. *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms*, 1849(6), 731–742. <https://doi.org/10.1016/j.bbagr.2015.01.005>
- Altshuler, D. L., Durbin, R. M., Abecasis, G. R., Bentley, D. R., Chakravarti, A., Clark, A. G., Collins, F. S., De La Vega, F. M., Donnelly, P., Egholm, M., Flicek, P., Gabriel, S. B., Gibbs, R. A., Knoppers, B. M., Lander, E. S., Lehrach, H., Mardis, E. R., McVean, G. A., Nickerson, D. A., ... Peterson, J. L. (2010). A map of human genome variation from population-scale sequencing. *Nature*. <https://doi.org/10.1038/nature09534>
- Anders, S., & Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology*. <https://doi.org/10.1186/gb-2010-11-10-r106>
- Anders, S., Pyl, P. T., & Huber, W. (2015). HTSeq-A Python framework to work with high-throughput sequencing data. *Bioinformatics*. <https://doi.org/10.1093/bioinformatics/btu638>
- Andrews, S. (2015). FASTQC A Quality Control tool for High Throughput Sequence Data. *Babraham Institute*.

- Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., & Sherlock, G. (2000). Gene ontology: Tool for the unification of biology. In *Nature Genetics*. <https://doi.org/10.1038/75556>
- Bao, Z. S., Chen, H. M., Yang, M. Y., Zhang, C. B., Yu, K., Ye, W. L., Hu, B. Q., Yan, W., Zhang, W., Akers, J., Ramakrishnan, V., Li, J., Carter, B., Liu, Y. W., Hu, H. M., Wang, Z., Li, M. Y., Yao, K., Qiu, X. G., ... Jiang, T. (2014). RNA-seq of 272 gliomas revealed a novel, recurrent PTPRZ1-MET fusion transcript in secondary glioblastomas. *Genome Research*. <https://doi.org/10.1101/gr.165126.113>
- Beckedorff, F. C., Ayupe, A. C., Crocci-Souza, R., Amaral, M. S., Nakaya, H. I., Soltys, D. T., Menck, C. F. M., Reis, E. M., & Verjovski-Almeida, S. (2013). The Intronic Long Noncoding RNA ANRASSF1 Recruits PRC2 to the RASSF1A Promoter, Reducing the Expression of RASSF1A and Increasing Cell Proliferation. *PLoS Genetics*. <https://doi.org/10.1371/journal.pgen.1003705>
- Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*. <https://doi.org/10.1111/j.2517-6161.1995.tb02031.x>
- Bourque, G., Burns, K. H., Gehring, M., Gorbunova, V., Seluanov, A., Hammell, M., Imbeault, M., Izsvák, Z., Levin, H. L., Macfarlan, T. S., Mager, D. L., & Feschotte, C. (2018). Ten things you should know about transposable elements. *Genome Biology*. <https://doi.org/10.1186/s13059-018-1577-z>
- Buenrostro, J. D., Wu, B., Chang, H. Y., & Greenleaf, W. J. (2015). ATAC-seq: A method for assaying chromatin accessibility genome-wide. *Current Protocols in Molecular Biology*. <https://doi.org/10.1002/0471142727.mb2129s109>
- Buermans, H. P. J., & den Dunnen, J. T. (2014). Next generation sequencing technology: Advances

- and applications. *Biochimica et Biophysica Acta - Molecular Basis of Disease*, 1842(10), 1932–1941. <https://doi.org/10.1016/j.bbadis.2014.06.015>
- Carbon, S., Douglass, E., Dunn, N., Good, B., Harris, N. L., Lewis, S. E., Mungall, C. J., Basu, S., Chisholm, R. L., Dodson, R. J., Hartline, E., Fey, P., Thomas, P. D., Albou, L. P., Ebert, D., Kesling, M. J., Mi, H., Muruganujan, A., Huang, X., ... Westerfield, M. (2019). The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Research*. <https://doi.org/10.1093/nar/gky1055>
- Choi, M., Scholl, U. I., Ji, W., Liu, T., Tikhonova, I. R., Zumbo, P., Nayir, A., Bakkaloğlu, A., Özen, S., Sanjad, S., Nelson-Williams, C., Farhi, A., Mane, S., & Lifton, R. P. (2009). Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proceedings of the National Academy of Sciences of the United States of America*. <https://doi.org/10.1073/pnas.0910672106>
- Churko, J. M., & Laird, D. W. (2011). Gap Junctions. *Cellular Domains*, 339–347. <https://doi.org/10.1002/9781118015759.ch20>
- Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M. W., Gaffney, D. J., Elo, L. L., Zhang, X., & Mortazavi, A. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biology*, 17(1), 1–19. <https://doi.org/10.1186/s13059-016-0881-8>
- Cooper, C. D. O., Newman, J. A., & Gileadi, O. (2014). Recent advances in the structural molecular biology of Ets transcription factors: Interactions, interfaces and inhibition. *Biochemical Society Transactions*, 42(1), 130–138. <https://doi.org/10.1042/BST20130227>
- Craig Venter, J., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., ... Zhu, X. (2001). The sequence of the human genome. *Science*. <https://doi.org/10.1126/science.1058040>
- Crick, F. (1970). Central dogma of molecular biology. *Nature*. <https://doi.org/10.1038/227561a0>

- Dai, M., Thompson, R. C., Maher, C., Contreras-Galindo, R., Kaplan, M. H., Markovitz, D. M., Omenn, G., & Meng, F. (2010). NGSQC: Cross-platform quality analysis pipeline for deep sequencing data. *BMC Genomics*. <https://doi.org/10.1186/1471-2164-11-S4-S7>
- Dethlefsen, L., Huse, S., Sogin, M. L., & Relman, D. A. (2008). The pervasive effects of an antibiotic on the human gut microbiota, as revealed by deep 16s rRNA sequencing. *PLoS Biology*. <https://doi.org/10.1371/journal.pbio.0060280>
- Di Ruocco, F., Basso, V., Rivoire, M., Mehlen, P., Ambati, J., De Falco, S., & Tarallo, V. (2018). Alu RNA accumulation induces epithelial-to-mesenchymal transition by modulating miR-566 and is associated with cancer progression. *Oncogene*. <https://doi.org/10.1038/onc.2017.369>
- Diamantopoulos, M. A., Tsiakanikas, P., & Scorilas, A. (2018). Non-coding RNAs: the riddle of the transcriptome and their perspectives in cancer. *Annals of Translational Medicine*. <https://doi.org/10.21037/atm.2018.06.10>
- Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*. <https://doi.org/10.1093/bioinformatics/bts635>
- Dunham, I., Kundaje, A., Aldred, S. F., Collins, P. J., Davis, C. A., Doyle, F., Epstein, C. B., Frietze, S., Harrow, J., Kaul, R., Khatun, J., Lajoie, B. R., Landt, S. G., Lee, B. K., Pauli, F., Rosenbloom, K. R., Sabo, P., Safi, A., Sanyal, A., ... Lochovsky, L. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*. <https://doi.org/10.1038/nature11247>
- F.S. Collins, E.S. Lander, J. Rogers, et al. (2004). International Human Genome Sequencing Consortium, Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011), 931–945. <https://doi.org/10.1038/nature03001>
- Garber, M., Grabherr, M. G., Guttman, M., & Trapnell, C. (2011). Computational methods for transcriptome annotation and quantification using RNA-seq. *Nature Methods*, 8(6), 469–477. <https://doi.org/10.1038/nmeth.1613>

- Garrod, D., & Chidgey, M. (2008). Desmosome structure, composition and function. In *Biochimica et Biophysica Acta - Biomembranes* (Vol. 1778, Issue 3, pp. 572–587). Elsevier. <https://doi.org/10.1016/j.bbamem.2007.07.014>
- Ge, X., Li, G. yuan, Jiang, L., Jia, L., Zhang, Z., Li, X., Wang, R., Zhou, M., Zhou, Y., Zeng, Z., Xiang, J., & Li, Z. (2019). Long noncoding RNA CAR10 promotes lung adenocarcinoma metastasis via miR-203/30/SNAI axis. *Oncogene*. <https://doi.org/10.1038/s41388-018-0645-x>
- Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., Di Palma, F., Birren, B. W., Nusbaum, C., Lindblad-Toh, K., ... Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*. <https://doi.org/10.1038/nbt.1883>
- H., D., J., H., K., Z., M., Y., Y., C., C., W., X., C., & M., H. (2019). Activation of LncRNA TINCR by H3K27 acetylation promotes Trastuzumab resistance and epithelial-mesenchymal transition by targeting MicroRNA-125b in breast Cancer 11 Medical and Health Sciences 1112 Oncology and Carcinogenesis. *Molecular Cancer*. <https://doi.org/10.1186/s12943-018-0931-9>
LK
<http://rug.on.worldcat.org/atoztitles/link/?sid=EMBASE&issn=14764598&id=doi:10.1186%2Fs12943-018-0931-9&title=Activation+of+LncRNA+TINCR+by+H3K27+acetylation+promotes+Trastuzuma+b+resistance+and+epithelial-mesenchymal+transition+by+targeting+MicroRNA-125b+in+breast+Cancer+11+Medical+and+Health+Sciences+1112+Oncology+and+Carcinogenesis&stitle=Mol.+Cancer&title=Molecular+Cancer&volume=18&issue=1&spage=&epage=&aulast=Dong&aufirst=Huaying&auinit=H.&aufull=Dong+H.&code>
- Hancks, D. C., & Kazazian, H. H. (2016). Roles for retrotransposon insertions in human disease. In *Mobile DNA*. <https://doi.org/10.1186/s13100-016-0065-9>
- Hugo, H., Ackland, M. L., Blick, T., Lawrence, M. G., Clements, J. A., Williams, E. D., &

- Thompson, E. W. (2007). Epithelial—mesenchymal and mesenchymal—epithelial transitions in carcinoma progression. *Journal of Cellular Physiology*, 213(2), 374–383. <https://doi.org/10.1002/jcp.21223>
- Janda, J. M., & Abbott, S. L. (2007). 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: Pluses, perils, and pitfalls. In *Journal of Clinical Microbiology*. <https://doi.org/10.1128/JCM.01228-07>
- Jasinska, A., & Krzyzosiak, W. J. (2004). Repetitive sequences that shape the human transcriptome. *FEBS Letters*, 567(1), 136–141. <https://doi.org/10.1016/j.febslet.2004.03.109>
- Johanning, G. L., Malouf, G. G., Zheng, X., Esteva, F. J., Weinstein, J. N., Wang-Johanning, F., & Su, X. (2017). Expression of human endogenous retrovirus-K is strongly associated with the basal-like breast cancer phenotype. *Scientific Reports*. <https://doi.org/10.1038/srep41960>
- Jordan, I. K., Rogozin, I. B., Glazko, G. V., & Koonin, E. V. (2003). Origin of a substantial fraction of human regulatory sequences from transposable elements. In *Trends in Genetics*. [https://doi.org/10.1016/S0168-9525\(02\)00006-9](https://doi.org/10.1016/S0168-9525(02)00006-9)
- Kanehisa, M., & Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes - Release 72.1, December 1, 2014 . *Nucleic Acids Res.*
- Karakulah, G. (2018). RTFADB: A database of computationally predicted associations between retrotransposons and transcription factors in the human and mouse genomes. *Genomics*. <https://doi.org/10.1016/j.ygeno.2017.11.002>
- Katz, Y., Wang, E. T., Airolidi, E. M., & Burge, C. B. (2010). Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nature Methods*. <https://doi.org/10.1038/nmeth.1528>
- Kim, D., Langmead, B., & Salzberg, S. L. (2015). HISAT: A fast spliced aligner with low memory requirements. *Nature Methods*. <https://doi.org/10.1038/nmeth.3317>
- Kotake, Y., Nakagawa, T., Kitagawa, K., Suzuki, S., Liu, N., Kitagawa, M., & Xiong, Y. (2011).

- Long non-coding RNA ANRIL is required for the PRC2 recruitment to and silencing of p15 INK4B tumor suppressor gene. *Oncogene*. <https://doi.org/10.1038/onc.2010.568>
- Kukurba, K. R., & Montgomery, S. B. (2015). RNA Sequencing and Analysis. *Cold Spring Harbor Protocols*, 2015(11), pdb.top084970. <https://doi.org/10.1101/pdb.top084970>
- Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., Fitzhugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczký, J., Levine, R., McEwan, P., ... Morgan, M. J. (2001). Initial sequencing and analysis of the human genome. *Nature*. <https://doi.org/10.1038/35057062>
- Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*. <https://doi.org/10.1186/gb-2009-10-3-r25>
- Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. <https://doi.org/10.1093/bioinformatics/btp324>
- Liu, W., Tanasa, B., Tyurina, O. V., Zhou, T. Y., Gassmann, R., Liu, W. T., Ohgi, K. A., Benner, C., Garcia-Bassets, I., Aggarwal, A. K., Desai, A., Dorrestein, P. C., Glass, C. K., & Rosenfeld, M. G. (2010). PHF8 mediates histone H4 lysine 20 demethylation events involved in cell cycle progression. *Nature*. <https://doi.org/10.1038/nature09272>
- Lowe, R., Shirley, N., Bleackley, M., Dolan, S., & Shafee, T. (2017). Transcriptomics technologies. *PLoS Computational Biology*, 13(5), 1–23. <https://doi.org/10.1371/journal.pcbi.1005457>
- Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.Journal*. <https://doi.org/10.14806/ej.17.1.200>
- Metzker, M. L. (2010). Sequencing technologies the next generation. *Nature Reviews Genetics*, 11(1), 31–46. <https://doi.org/10.1038/nrg2626>
- Nisticò, P., Bissell, M. J., & Radisky, D. C. (2012). Epithelial-mesenchymal transition: General

- principles and pathological relevance with special emphasis on the role of matrix metalloproteinases. *Cold Spring Harbor Perspectives in Biology*, 4(2), 1–11. <https://doi.org/10.1101/cshperspect.a011908>
- Pandey, R. R., Mondal, T., Mohammad, F., Enroth, S., Redrup, L., Komorowski, J., Nagano, T., Mancini-DiNardo, D., & Kanduri, C. (2008). Kcnq1ot1 Antisense Noncoding RNA Mediates Lineage-Specific Transcriptional Silencing through Chromatin-Level Regulation. *Molecular Cell*. <https://doi.org/10.1016/j.molcel.2008.08.022>
- Park, P. J. (2009). ChIP-seq: Advantages and challenges of a maturing technology. In *Nature Reviews Genetics*. <https://doi.org/10.1038/nrg2641>
- Pérez-Pomares, J. M., & Muñoz-Chápuli, R. (2002). Epithelial-mesenchymal transitions: A mesodermal cell strategy for evolutive innovation in metazoans. *Anatomical Record*, 268(3), 343–351. <https://doi.org/10.1002/ar.10165>
- Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T. C., Mendell, J. T., & Salzberg, S. L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology*. <https://doi.org/10.1038/nbt.3122>
- Piskol, R., Ramaswami, G., & Li, J. B. (2013). Reliable identification of genomic variants from RNA-seq data. *American Journal of Human Genetics*. <https://doi.org/10.1016/j.ajhg.2013.08.008>
- Radisky, D. C. (2005). Epithelial-mesenchymal transition. *Journal of Cell Science*, 118(19), 4325–4326. <https://doi.org/10.1242/jcs.02552>
- Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*. <https://doi.org/10.1093/nar/gkv007>
- Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford*,

England), 26(1), 139–140. <https://doi.org/10.1093/bioinformatics/btp616>

- Ruike, Y., Imanaka, Y., Sato, F., Shimizu, K., & Tsujimoto, G. (2010). Genome-wide analysis of aberrant methylation in human breast cancer cells using methyl-DNA immunoprecipitation combined with high-throughput sequencing. *BMC Genomics*. <https://doi.org/10.1186/1471-2164-11-137>
- Sanders, S. J., Murtha, M. T., Gupta, A. R., Murdoch, J. D., Raubeson, M. J., Willsey, A. J., Ercan-Sencicek, A. G., Di Lullo, N. M., Parikshak, N. N., Stein, J. L., Walker, M. F., Ober, G. T., Teran, N. A., Song, Y., El-Fishawy, P., Murtha, R. C., Choi, M., Overton, J. D., Bjornson, R. D., ... State, M. W. (2012). De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature*. <https://doi.org/10.1038/nature10945>
- Sanger, F., & Coulson, A. R. (1975). A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of Molecular Biology*. [https://doi.org/10.1016/0022-2836\(75\)90213-2](https://doi.org/10.1016/0022-2836(75)90213-2)
- Sela-Donenfeld, D., & Kalcheim, C. (1999). Regulation of the onset of neural crest migration by coordinated activity of BMP4 and Noggin in the dorsal neural tube. *Development*.
- Sengez, B., Aygün, I., Shehwana, H., Toyran, N., Tercan Avci, S., Konu, O., Stemmler, M. P., & Alotaibi, H. (2019). The Transcription Factor Elf3 Is Essential for a Successful Mesenchymal to Epithelial Transition. *Cells*, 8(8), 858. <https://doi.org/10.3390/cells8080858>
- Shook, D., & Keller, R. (2003). Mechanisms, mechanics and function of epithelial-mesenchymal transitions in early development. *Mechanisms of Development*, 120(11), 1351–1383. <https://doi.org/10.1016/j.mod.2003.06.005>
- Slatko, B. E., Gardner, A. F., & Ausubel, F. M. (2018). Overview of Next-Generation Sequencing Technologies. *Current Protocols in Molecular Biology*, 122(1), 1–11. <https://doi.org/10.1002/cpmb.59>
- Stemmler, M. P., Hecht, A., & Kemler, R. (2005). E-cadherin intron 2 contains cis-regulatory

- elements essential for gene expression. *Development*. <https://doi.org/10.1242/dev.01662>
- Stemmler, M. P., Hecht, A., Kinzel, B., & Kemler, R. (2003). Analysis of regulatory elements of E-cadherin with reporter gene constructs in transgenic mouse embryos. *Developmental Dynamics*, 227(2), 238–245. <https://doi.org/10.1002/dvdy.10301>
- Technau, U., & Scholz, C. B. (2003). Origin and evolution of endoderm and mesoderm. *International Journal of Developmental Biology*, 47(7–8), 531–539. <https://doi.org/10.1387/ijdb.14756329>
- Thermes, C. (2014). Ten years of next-generation sequencing technology. *Trends in Genetics : TIG*, 30(9), 418–426. <https://doi.org/10.1016/j.tig.2014.07.001>
- Thiery, J. P. (2003). Epithelial-mesenchymal transitions in development and pathologies. In *Current Opinion in Cell Biology* (Vol. 15, Issue 6, pp. 740–746). Elsevier Ltd. <https://doi.org/10.1016/j.ceb.2003.10.006>
- Thiery, J. P., Acloque, H., Huang, R. Y. J., & Nieto, M. A. (2009). Epithelial-Mesenchymal Transitions in Development and Disease. *Cell*, 139(5), 871–890. <https://doi.org/10.1016/j.cell.2009.11.007>
- Trapnell, C., Pachter, L., & Salzberg, S. L. (2009). TopHat: Discovering splice junctions with RNA-Seq. *Bioinformatics*. <https://doi.org/10.1093/bioinformatics/btp120>
- Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., Van Baren, M. J., Salzberg, S. L., Wold, B. J., & Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*. <https://doi.org/10.1038/nbt.1621>
- Trapnell, C., Williams, B. a, Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., Salzberg, S. L., Wold, B. J., & Pachter, L. (2011). Transcript assembly and abundance estimation from RNA-Seq reveals thousands of new transcripts and switching among isoforms. *Nature Biotechnology*. <https://doi.org/10.1038/nbt.1621>. Transcript

- Tymms, M. J., Ng, A. Y. N., Thomas, R. S., Schutte, B. C., Zhou, J., Eyre, H. J., Sutherland, G. R., Seth, A., Rosenberg, M., Papas, T., Debouck, C., & Kola, I. (1997). A novel epithelial-expressed ETS gene, ELF3: Human and murine cDNA sequences, murine genomic organization, human mapping to 1q32.2 and expression in tissues and cancer. *Oncogene*, 15(20), 2449–2462. <https://doi.org/10.1038/sj.onc.1201427>
- Wagner, G. P., Kin, K., & Lynch, V. J. (2012). Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory in Biosciences*, 131(4), 281–285. <https://doi.org/10.1007/s12064-012-0162-3>
- Wang, X., Lu, H., Urvalek, A. M., Li, T., Yu, L., Lamar, J., Dipersio, C. M., Feustel, P. J., & Zhao, J. (2011). KLF8 promotes human breast cancer cell invasion and metastasis by transcriptional activation of MMP9. *Oncogene*. <https://doi.org/10.1038/onc.2010.563>
- Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics in Western Equatoria State. *Nature Reviews Genetics*, 10(1), 57.
- Waterston, R. H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J. F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., Antonarakis, S. E., Attwood, J., Baertsch, R., Bailey, J., Barlow, K., Beck, S., Berry, E., Birren, B., Bloom, T., ... Lander, E. S. (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature*. <https://doi.org/10.1038/nature01262>
- Watson, J. D., & Crick, F. H. C. (1953). Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. *Nature*. <https://doi.org/10.1038/171737a0>
- Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J. L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O., Paux, E., SanMiguel, P., & Schulman, A. H. (2007). A unified classification system for eukaryotic transposable elements. In *Nature Reviews Genetics*. <https://doi.org/10.1038/nrg2165>
- Yandım, C., & Karakulah, G. (2019). Expression dynamics of repetitive DNA in early human embryonic development. *BMC Genomics*. <https://doi.org/10.1186/s12864-019-5803-1>

- Yang, I. S., & Kim, S. (2015). Analysis of Whole Transcriptome Sequencing Data: Workflow and Software. *Genomics & Informatics*, 13(4), 119. <https://doi.org/10.5808/gi.2015.13.4.119>
- Yang, L., Smyth Gordon K, & Wei, S. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*.
- Yao, D., Dai, C., & Peng, S. (2011). Mechanism of the mesenchymal-epithelial transition and its relationship with metastatic tumor formation. *Molecular Cancer Research*, 9(12), 1608–1620. <https://doi.org/10.1158/1541-7786.MCR-10-0568>
- Yeaman, C., Grindstaff, K. K., & Nelson, W. J. (1999). New perspectives on mechanisms involved in generating epithelial cell polarity. *Physiological Reviews*, 79(1), 73–98. <https://doi.org/10.1152/physrev.1999.79.1.73>

APPENDIX 1

**Cold Spring Harbor Laboratory Press
One Bungtown Road
Cold Spring Harbor, NY 11724, USA**

June 26, 2020

Hamdiye Uzuner
Izmir International Biomedicine and Genome Institute
Dokuz Eylül University
Turkey

Dear Hamdiye,

In response to your request regarding permission to use the following material:

Items: Figs 1 and 2, pp. 3 and 10
Title: Cold Spring Harbor Protocols
Article: RNA Sequencing and Analysis
Authors: Kimberly R. Kukurba, Stephen B. Montgomery
Copyright date: 2015

To be used in:

Title: Masters thesis: Elf3 regulated transcriptome dynamics during
mesenchymal to epithelial transition
Institution: Izmir International Biomedicine and Genome Institute, Dokuz Eylül
University
Requester: Hamdiye Uzuner

I confirm that we are willing to grant non-exclusive rights in the English language, subject to the following conditions:

1. Full acknowledgment shall be made to the source and correct references given.
2. The permission of the author is obtained, where possible.
3. Permission is limited to the proposed use, on a non-exclusive, **one-time-only basis**, with distribution rights in the English language only.

*Please note that we do not grant blanket permissions for future uses or translations. If you need any of the aforementioned, please reapply and include details of usage.

By Carol C. Brown Date: 6/26/20
For Cold Spring Harbor Laboratory Press, 500 Sunnyside Blvd., Woodbury NY 11797, USA

Tel. (516) 422-4038/ e-mail: brown@cshl.edu / Fax (516) 422-4095

APPENDIX II



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

Toplantı Tarihi : 24/04/2020 **Toplantı Günü** : Cuma
Toplantı Sayısı : 5 **Toplantı Saati** : 14:30

Sayın Gökhan KARAKÜLAH,

2020-013 Protokol No'lu; sorumlusu olduğunuz "Mezenkimal Epitel geçiş sırasında Elf3 ile regüle edilen transkriptom dinamikleri" başlıklı araştırmanın uygulanmasında etik açıdan sakınca olmadığına oy birliği ile karar verilmiştir.

Bilgilerinizi ve gereğini rica ederiz.

Prof. Dr. H. Alper BAĞRIYANIK
Başkan

Pandemi süresince online ortamda gerçekleştirilen toplantımızda alınan kararlar tek imzalı olarak düzenlenmektedir. Pandemi sona erdikten sonra ıslak imzalı karar belgesi teslim edilecektir.